The Relativity of AGI: Distributional Axioms, Fragility, and Undecidability

Abstract

We study whether Artificial General Intelligence (AGI) admits a coherent theoretical definition that supports absolute claims of existence, robustness, or self-verification. We formalize AGI axiomatically as a distributional, resource-bounded semantic predicate, indexed by a task family, a task distribution, a performance functional, and explicit resource budgets. Under this framework, we derive four classes of results. First, we show that generality is inherently relational: there is no distribution-independent notion of AGI. Second, we prove non-invariance results demonstrating that arbitrarily small perturbations of the task distribution can invalidate AGI properties via cliff sets, precluding universal robustness. Third, we establish bounded transfer guarantees, ruling out unbounded generalization across task families under finite resources. Fourth, invoking Rice-style and Gödel–Tarski arguments, we prove that AGI is a nontrivial semantic property and therefore cannot be soundly and completely certified by any computable procedure, including procedures implemented by the agent itself. Consequently, recursive self-improvement schemes that rely on internal self-certification of AGI are ill-posed. Taken together, our results show that strong, distribution-independent claims of AGI are not false but undefined without explicit formal indexing, and that empirical progress in AI does not imply the attainability of self-certifying general intelligence.

1. Introduction

Despite sustained public and scientific attention, the term artificial general intelligence (AGI) remains conceptually unstable. In policy discussions, industrial roadmaps, and popular commentary, AGI is routinely described as an impending milestone—a system that is "general" in roughly the way humans are general—yet the phrase is rarely cashed out as a precise mathematical predicate [1, 2, 3, 4]. In the technical literature, evaluations often substitute

breadth of benchmark performance for generality, yielding an implicit definition of AGI as "high scores on many tasks" [5, 6, 7]. This substitution is convenient, but it blurs a crucial distinction: performance aggregated over curated test suites is a distributional statement about the specific task ecology encoded by those suites, not a theorem-level claim about generality in any invariant sense [8, 9].

The absence of a shared definition is not a mere terminological nuisance. It prevents falsifiable claims, encourages rhetorical goalpost shifts, and makes it difficult to separate engineering progress from statements that are, in effect, metaphysical [10, 11, 12]. For example, one widely recurring formulation—"a system that can do any intellectual task a human can"—depends on underspecified notions of task, ability, resources, and acceptable failure [13]. Another common move is to treat scaling trends as evidence of an eventual qualitative phase transition to generality [14, 15]; however, without an explicit semantics for what "general" means, such claims are not even well-formed propositions. In short, AGI talk is currently long on aspiration and short on definitional discipline.

This paper takes the definitional question as primary. Our goal is not to deny the empirical reality of increasingly capable systems, nor to dispute the practical value of broad competence. Rather, we ask a narrower and more fundamental question: is there a coherent theoretical object corresponding to contemporary notions of AGI? To make this question answerable, we define AGI axiomatically—in the manner of a norm or a topology—as a bundle of individually reasonable properties capturing breadth, adaptivity, transfer, compositionality, and robustness. We then interpret these axioms under distributional semantics: the agent's competence is evaluated relative to an explicit task distribution, a performance functional, and a resource budget. This relational pivot forces any claim of "AGI" to specify the task ecology it quantifies over; without such a specification, the predicate is undefined. Under this semantics, we derive structural theorems that formalize fragility and non-invariance: even if an agent satisfies the axioms under one distribution, arbitrarily small perturbations of the task distribution can destroy the property, and transfer cannot be unbounded without commensurate shared structure. Finally, we show that any attempt by a sufficiently expressive system to internally certify its own AGI status amounts to deciding a non-trivial semantic property, linking self-certification to classical undecidability obstructions and establishing a Gödel–Tarski-style limit on self-knowledge of "generality" [16, 17, 18].

2. Preliminaries and Notation

This section fixes the objects and notation used throughout the paper. The intent is to make all subsequent statements—in particular the axioms of distributional AGI and the structural theorems—well-formed, unambiguous, and internally consistent.

2.1. Spaces, Histories, and Randomness

Let $\mathcal{O}$ denote the observation space and $\mathcal{A}$ the action space. In the most general setting, $\mathcal{O}$ and $\mathcal{A}$ are measurable spaces; in concrete examples they may be discrete alphabets (e.g., tokens) or Euclidean spaces (e.g., sensor vectors). Let $\mathcal{H}_t := (\mathcal{O} \times \mathcal{A})^{t-1} \times \mathcal{O}$ denote the space of length-t interaction histories and let $\mathcal{H} := \bigcup_{t \geq 1} \mathcal{H}_t$ denote the set of all finite histories. For $h_t = (o_1, a_1, \ldots, a_{t-1}, o_t) \in \mathcal{H}_t$, we write $h_{t:t'}$ for the subsequence $(o_t, a_t, \ldots, o_{t'})$ when needed.

All probability statements are taken with respect to an ambient probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that underlies the stochasticity of environments, agents (if randomized), and any exogenous randomness used to sample tasks.

2.2. Tasks, Environments, and Performance

Environments.. An environment E is an interactive stochastic process that induces a sequence $(o_t)_{t \geq 1}$ of observations conditioned on past interactions. Formally, E specifies a family of conditional distributions

$$E(\,\cdot \mid h_t, a_t) \in \Delta(\mathcal{O}), \qquad t \geq 1,$$

where $\Delta(\mathcal{X})$ denotes the set of probability measures on $\mathcal{X}$. When the environment also emits an explicit reward signal, we include it in the observation o_t (so o_t may contain both "state" and "reward") to avoid duplicating notation.

Episodes and horizons.. We consider episodic interaction with a (possibly random) horizon $T \in \mathbb{N} \cup \{\infty\}$. When a finite horizon is assumed, T may be part of the task specification. For clarity we write $\mathcal{H}_{\leq T} := \bigcup_{t=1}^{T} \mathcal{H}_t$.

Tasks.. A task is a pair

$$\tau := (E, U),$$

where E is an environment and U is a utility (success) functional mapping a complete episode history to a real-valued score. For a finite horizon T, we take $U : \mathcal{H}_{\leq T} \to \mathbb{R}$; for $T = \infty$, U may depend on the full infinite trajectory or

be defined via a limiting/discounted functional. We assume U is measurable and typically bounded; when needed we assume $U \in [0,1]$ without loss of generality by rescaling.

Policies and induced trajectories.. Given a policy (agent) A and an environment E, the interaction induces a random history $H_T \sim (E \circ A)$ (composition in the standard interactive sense). We write $\mathbb{E}_{E,A}[\cdot]$ and $\mathbb{P}_{E,A}[\cdot]$ for expectation and probability under this induced law.

Performance metric.. A performance metric is a functional

$$\Pi(\tau, A; B) \in \mathbb{R},$$

where $\tau = (E, U)$ is a task, A is an agent, and B is a resource budget (defined below). The canonical choice is expected utility under the induced interaction:

$$\Pi(\tau, A; B) := \mathbb{E}_{E,A}[\, U(H_T)\,],$$

subject to the constraint that A respects the budget B. In some settings Π may incorporate additional penalties (e.g., constraint violations, cost of tool calls), or may be a regret-like quantity. We assume Π is bounded (e.g., in $[0,1]$) after normalization.

Task-level success events.. For statements that require a binary notion of success, we introduce a task-level success event S_τ induced by the performance functional. Concretely, when Π is normalized and a threshold $\theta \in \mathbb{R}$ is fixed, we may take

$$S_\tau := \{\Pi(\tau, A; B) \geq \theta\}.$$

The particular threshold used is always stated explicitly when S_τ is invoked.

Constraints and violation indicators.. To incorporate constraints (e.g., safety constraints, policy constraints, specification constraints) into evaluation, we use a measurable violation indicator

$$V_C(\tau, A; B) \in \{0, 1\},$$

where $V_C(\tau, A; B) = 1$ indicates that the interaction of A with task τ under budget B violates the constraint family C. When constraints are relevant, performance statements are made jointly with bounds on $\mathbb{P}_{\tau\sim\mu}(V_C(\tau, A; B) = 1)$.

2.3. Agents, Learning, and Resource Budgets

Agents as history-dependent policies.. An agent A is a (possibly randomized) mapping from histories to action distributions:

$$A(\cdot \mid h_t) \in \Delta(\mathcal{A}), \qquad h_t \in \mathcal{H}_t,\ t \geq 1.$$

We allow A to be stateful: it may maintain internal memory m_t updated over time. Formally, one may view A as a policy over an augmented history that includes its internal state; we suppress m_t unless explicitly needed.

Learning and adaptation.. To express adaptivity and transfer, we treat the agent as parameterized by an internal state/parameter θ. Let θ_0 denote the initial state (e.g., pretrained weights). As interaction proceeds, the agent may update θ according to some (possibly implicit) update rule driven by data D_t extracted from histories:

$$\theta_{t+1} = \mathrm{Update}(\theta_t, D_t).$$

We write A_θ for the policy induced by parameter/state θ, and A_t (or $A^{(t)}$) for the agent after t updates. When adaptation occurs across tasks, we write $A^{(k)}$ for the agent after k tasks/episodes have been processed.

Update-indexed agents and comparison baselines.. To avoid ambiguity in later axioms, we use the convention $A^{(n)}$ to denote the agent after n updates as measured in the relevant context (within-task updates when proving adaptivity statements, and across-task updates when proving transfer statements). When a distinction is required between an agent that has undergone a fixed pre-exposure phase and one that has not, we write A_{pre} for the agent after a pre-exposure phase consisting of K_{tr} tasks sampled i.i.d. from μ, and A_{scratch} for the corresponding agent initialized without such pre-exposure. The integer N_{ad} denotes a within-task update budget used to formalize adaptivity. All update counts and pre-exposure lengths are constrained by the resource budget B through the requirement that the corresponding executions respect B.

Resource budgets.. A resource budget is a tuple

$$B := (B_{\mathrm{comp}}, B_{\mathrm{mem}}, B_{\mathrm{samp}}, B_{\mathrm{time}}, B_{\mathrm{tool}}),$$

capturing, respectively, limits on compute, memory, number of samples/episodes, wall-clock time, and (when relevant) external tool calls or oracle queries. We say an agent respects B if, on any interaction consistent with the task specification, its resource usage stays within the corresponding bounds. When only a subset of resources matters, we omit components of B by abuse of notation.

Tool interfaces.. When external tools or oracles are available (e.g., retrieval systems, code execution, measurement devices), we denote the corresponding interface by $\mathcal{M}$. The agent with tool access is denoted by $A^{\mathcal{M}}$, and all tool usage is accounted for via the budget component B_{tool} (and any other relevant components of B).

2.4. Task Families, Distributions, and Evaluation

Task family.. A task family $\mathcal{T}$ is a set of tasks $\tau = (E, U)$ sharing a common interface $(\mathcal{O}, \mathcal{A})$ and common structural constraints (e.g., bounded horizon, bounded reward, admissible perturbations). We write $\tau \in \mathcal{T}$ to mean τ is admissible.

Task distribution.. A task distribution is a probability measure $\mu \in \Delta(\mathcal{T})$ over $\mathcal{T}$. Sampling $\tau \sim \mu$ represents drawing a task from a task ecology. All distributional notions of AGI in this paper are indexed by μ.

Goals as task generators.. For autonomy statements, we introduce a goal space $\mathcal{G}$ equipped with a goal distribution $\nu \in \Delta(\mathcal{G})$ and a compilation map $\text{Compile} : \mathcal{G} \to \mathcal{T}$. A goal $g \sim \nu$ induces a task $\tau_g := \text{Compile}(g) \in \mathcal{T}$. Any additional interaction required to refine goals is counted against B via Π.

Distributional performance functionals.. Given μ, define the generality functional

$$\mathcal{G}_\mu(A; B) := \mathbb{E}_{\tau\sim\mu}[\Pi(\tau, A; B)] \,.$$

We also use a tail-robust variant:

$$\mathcal{G}_{\mu,\delta}(A; B) := \sup \Big\{\theta \in \mathbb{R} : \ \mathbb{P}_{\tau\sim\mu}\Big(\Pi(\tau, A; B) \geq \theta\Big) \geq 1 - \delta\Big\} \,,$$

for $\delta \in (0, 1)$.

Baselines and optimality.. When comparing to an optimal or reference agent, we write $A^\star_\tau$ for a task-dependent oracle (generally unattainable) and define, for example, regret

$$\mathrm{Reg}(\tau, A; B) := \Pi(\tau, A^\star_\tau; B^\star) - \Pi(\tau, A; B),$$

where $B^\star$ denotes the resources available to the oracle. We will avoid relying on oracle notions unless explicitly stated.

Axiom parameters.. When stating axioms and tail guarantees, we use explicit performance thresholds $\theta_\bullet$ and allowable tail-failure probabilities $\delta_\bullet$ (with the subscript $\bullet$ indicating the relevant axiom). We also use nonnegative tolerances $\varepsilon_{\mathrm{rb}}$ (robustness slack) and $\varepsilon_{\mathrm{cal}}$ (calibration tolerance), and a constraint-violation tolerance $\delta_C \in (0, 1)$. These quantities are fixed constants once the axiom bundle is fixed.

2.5. Composition and Perturbation Operators

Task composition.. To formalize compositionality, we assume $\mathcal{T}$ is equipped with a (possibly partial) binary operation

$$\oplus : \ \mathcal{T} \times \mathcal{T} \to \mathcal{T},$$

where $\tau_1 \oplus \tau_2$ represents a task requiring competence on τ_1 and τ_2 in a prescribed manner (e.g., sequential composition, conjunction of goals, tool-augmented composition). When multiple composition modes are needed, we index them by $\oplus_j$.

Perturbations and robustness.. To formalize robustness, we define a family of admissible perturbations $\mathcal{D}$, where each $\Delta \in \mathcal{D}$ maps tasks to tasks:

$$\Delta : \ \mathcal{T} \to \mathcal{T}.$$

Typical perturbations include paraphrases, observation noise, reward reshaping within equivalence classes, or mild distributional shifts in environment dynamics. We write $\Delta(\tau)$ for the perturbed task and require $\Delta(\tau) \in \mathcal{T}$ whenever $\tau \in \mathcal{T}$.

2.6. Distances Between Task Distributions

To quantify "small" distribution shift, we use standard metrics on probability measures over $\mathcal{T}$. Two choices used in this paper are:

Total variation distance.. For $\mu, \nu \in \Delta(\mathcal{T})$,

$$d_{\mathrm{TV}}(\mu, \nu) := \sup_{A \subseteq \mathcal{T}} \Big| \mu(A) - \nu(A) \Big|.$$

Wasserstein distance.. When $\mathcal{T}$ is equipped with a ground metric $d_{\mathcal{T}}$, we write $W_p(\mu, \nu)$ for the p-Wasserstein distance. We will only invoke Wasserstein metrics when a meaningful $d_{\mathcal{T}}$ is specified.

2.7. Hard Slices and Cliff Sets

A central technical device in the non-invariance results is the existence of hard slices (or cliff sets) of the task space where performance sharply degrades.

Definition (Cliff set).. Given an agent A, budget B, and threshold θ, define the θ-failure set

$$\mathcal{F}_\theta(A; B) := \{\tau \in \mathcal{T} : \ \Pi(\tau, A; B) < \theta\}.$$

A subset $\mathcal{C} \subseteq \mathcal{T}$ is called a cliff set at level θ for (A, B) if $\mathcal{C} \subseteq \mathcal{F}_\theta(A; B)$ and $\mathcal{C}$ is reachable under arbitrarily small shifts of the task distribution in the sense made precise in Section 6.2.

2.8. Self-Assessment and Semantic Predicates

To state the Gödel–Tarski self-certification obstruction, we distinguish between syntactic predicates that an agent can compute from its internal state, and semantic predicates that quantify over external behaviors.

Semantic properties of agents.. A property $\mathsf{P}(A)$ is called semantic if it depends only on the input–output behavior of A across tasks/environments (not on its source code representation). Statements of the form "A succeeds on tasks drawn from μ with probability at least $1 - \delta$" are semantic because they quantify over external interaction outcomes.

AGI as a semantic predicate.. Once the axioms are fixed, the assertion

$$\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$$

is a semantic predicate in this sense: it quantifies over tasks $\tau \sim \mu$ and the interaction-induced random histories under $(E \circ A)$.

Resource-bounded self-assessment.. When discussing whether an agent can certify $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$, we restrict the certification procedure itself to the same budget class B (or a specified certification budget B_{cert}), to avoid trivializing the question by granting unbounded meta-resources.

2.9. Effective Representations and Codings

To reason about decidability, self-verification, and semantic obstruction results, we require that agents and tasks admit effective descriptions.

Effective task representations.. We assume that the task family $\mathcal{T}$ admits an effective presentation: there exists a countable set of finite descriptions $\mathcal{D}_{\mathcal{T}} \subseteq \{0,1\}^*$ and a surjective decoding map

$$\mathrm{Decode}_{\mathcal{T}} : \mathcal{D}_{\mathcal{T}} \to \mathcal{T}.$$

When a task $\tau \in \mathcal{T}$ is associated with a description $d \in \mathcal{D}_{\mathcal{T}}$, we write $\mathrm{code}_{\mathcal{T}}(\tau) = d$. Different descriptions may decode to the same task; no canonical encoding is assumed.

Effective agent representations.. Similarly, agents are assumed to admit finite descriptions. Let $\mathcal{D}_{\mathcal{A}} \subseteq \{0,1\}^*$ denote a set of agent descriptions, together with a decoding map

$$\mathrm{Decode}_{\mathcal{A}} : \mathcal{D}_{\mathcal{A}} \to \mathcal{A},$$

where $\mathcal{A}$ denotes the class of admissible agents (i.e., agents that respect the relevant resource budgets). For an agent A, we write $\mathrm{code}(A) \in \mathcal{D}_{\mathcal{A}}$ for some description that decodes to A.

Computability assumptions.. All constructions considered in this paper operate on these finite descriptions. In particular:

- Sampling $\tau \sim \mu$ is implemented via a probabilistic procedure over $\mathcal{D}_{\mathcal{T}}$ whose pushforward measure is μ.
- Performance evaluation $\Pi(\tau, A; B)$ is a semantic quantity defined via interaction, not via inspection of $\mathrm{code}(A)$.
- Certification or decision procedures invoked in later sections are restricted to operate on $\mathrm{code}(A)$, $\mathrm{code}_{\mathcal{T}}(\tau)$, and bounded interaction transcripts.

Semantic versus syntactic access.. The distinction between syntactic access (operating on finite descriptions) and semantic properties (quantifying over interaction outcomes across tasks) is central. While code(A) is a finite object, predicates such as

$$\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$$

are semantic and quantify over the behavior of A on tasks not known in advance.

Scope of effectivity.. No assumption is made that $\mathcal{T}$ or $\mathcal{A}$ is recursively enumerable in full generality; effectivity is required only to the extent necessary to state and analyze decision problems over agent descriptions. This level of generality suffices for the impossibility results in Section 7.

3. Why Single-Sentence Definitions of AGI Fail

This section examines several commonly invoked single-sentence characterizations of artificial general intelligence and shows that each fails to define a coherent theoretical property. The failures are structural rather than empirical: they persist independently of implementation details, training scale, or future engineering advances.

3.1. Human-Level Intelligence

A frequently cited characterization of AGI is that it denotes a system exhibiting "human-level intelligence." This formulation fails to define a mathematically meaningful predicate for at least two reasons.

First, the comparator class "human" is ill-defined. Human populations exhibit substantial heterogeneity across individuals, cultures, developmental histories, educational exposure, and tool access. There is no canonical distribution of tasks, no agreed-upon aggregation of abilities, and no fixed resource budget relative to which performance could be normalized. Consequently, statements of the form "A performs at the level of a human" lack a well-defined reference point.

Second, even if a representative human baseline were fixed, humans do not satisfy any formal notion of generality. Human competence is strongly distribution-dependent: it reflects adaptation to a narrow ecological and cultural task distribution rather than invariant performance across heterogeneous task families. Humans routinely fail on tasks that are trivial for machines (e.g., large-scale arithmetic) and succeed on tasks that are trivial only because

of prior environmental regularities (e.g., natural language pragmatics). Thus, "human-level" performance does not correspond to a property of general problem-solving capacity in any formal sense.

It follows that "human-level intelligence" is not a definitional criterion but a sociological comparison. As such, it cannot serve as a basis for a theoretical definition of AGI.

3.2. Universal Problem Solving

Another common characterization asserts that AGI is a system capable of solving any problem or performing any task within a sufficiently broad class. This formulation is untenable under any precise interpretation.

If the task class is unrestricted or sufficiently expressive, then the requirement collapses into uncomputability. Tasks can encode arbitrary decision problems, including instances equivalent to the halting problem or other undecidable predicates. No algorithmic system can guarantee success across such a class, even with unbounded training data, unless the notion of "solve" is weakened to the point of vacuity.

If, alternatively, the task class is restricted so that universal solvability becomes achievable, then the restriction itself carries all the substantive content. In this case, generality is no longer a property of the agent alone but of the agent relative to a predefined task family. The phrase "universal problem solver" then becomes a rebranding of competence within a bounded domain rather than a meaningful statement of generality.

Thus, the universal problem-solving definition is either mathematically impossible or conceptually empty, depending on how the task class is specified.

3.3. General Learning

A third formulation identifies AGI with the capacity to learn any task from data. This claim encounters a fundamental obstruction arising from no-free-lunch results in learning theory.

Formally, consider the space of all possible target functions or environments consistent with a fixed input-output interface. When performance is averaged uniformly over this space, all learning algorithms exhibit identical expected error. Any algorithm that performs well on a subset of tasks must necessarily perform poorly on a complementary subset. Therefore, learning performance is inseparable from inductive bias, and inductive bias entails restriction to a particular task distribution.

As a consequence, there is no distribution-independent notion of "general learning." Any claim of learning generality presupposes a structured task distribution, whether explicitly stated or not. Once such a distribution is fixed, the learning problem becomes well-defined but loses any claim to universality.

3.4. Conclusion of Section

The preceding analyses establish that single-sentence definitions of AGI fail for principled reasons. Comparisons to human intelligence lack a formal baseline; universal problem-solving claims collapse into impossibility or triviality; and appeals to general learning are blocked by distributional dependence inherent in inductive inference.

These failures are not accidental. They reflect the fact that generality is not a primitive concept but a composite one, involving breadth, adaptation, transfer, robustness, and resource constraints. Consequently, any coherent definition of AGI must be axiomatic rather than declarative. The remainder of the paper proceeds on this basis.

4. An Axiomatic Definition of AGI

This section defines artificial general intelligence as an axiomatic property of an agent relative to a task family, a task distribution, a performance functional, and an explicit resource budget. All subsequent theorems refer to the notation and primitives fixed in Section 2. No axiom introduced below is treated as optional.

4.1. Core Definition Schema

Fix a task family $\mathcal{T}$, a probability measure $\mu \in \Delta(\mathcal{T})$, a performance functional $\Pi(\tau, A; B)$, and a resource budget B. We write

$$\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$$

to denote the predicate that agent A satisfies the axioms in Sections 4.3–4.4. In particular, $\mathrm{AGI}(\cdot \mid \mathcal{T}, \mu, \Pi, B)$ is a relational, distribution-indexed, resource-bounded property; it is not a property of A in isolation.

4.2. Quantitative Conventions

To avoid ambiguity, we parameterize the axioms by explicit constants. Let

$$\Theta := (\theta_{\mathrm{br}}, \theta_{\mathrm{ad}}, \theta_{\mathrm{tr}}, \theta_{\mathrm{cp}}, \theta_{\mathrm{rb}}), \quad \Delta := (\delta_{\mathrm{br}}, \delta_{\mathrm{ad}}, \delta_{\mathrm{tr}}, \delta_{\mathrm{cp}}, \delta_{\mathrm{rb}}),$$

where $\theta_\bullet$ are performance thresholds and $\delta_\bullet \in (0,1)$ are allowable tail-failure probabilities. In addition, let $\mathcal{D}$ denote the admissible perturbation family (Section 2.5) and let $\varepsilon_{\mathrm{rb}} \geq 0$ denote a robustness slack.

When needed, we write $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B; \Theta, \Delta, \mathcal{D}, \varepsilon_{\mathrm{rb}})$ to make these parameters explicit; otherwise they are fixed and suppressed for readability.

4.3. Core AGI Axioms (G1–G5)

All axioms below are distributional: probabilities and expectations are with respect to $\tau \sim \mu$ and the interaction-induced randomness under $(E \circ A)$ as encoded inside $\Pi(\tau, A; B)$.

Axiom G1 (Breadth).. The agent achieves non-trivial competence on the heterogeneous task family under μ:

$$\mathbb{P}_{\tau\sim\mu}\Big(\Pi(\tau, A; B) \geq \theta_{\mathrm{br}}\Big) \;\geq\; 1 - \delta_{\mathrm{br}}.$$

This axiom rules out degenerate "single-domain" competence by requiring high-probability success across the task ecology encoded by μ.

Axiom G2 (Adaptivity).. There exists an adaptation protocol internal to the agent such that, for a μ-randomly drawn task τ, the agent reaches competence after limited within-task interaction subject to budget B. Concretely, there exists an integer N_{ad} (interpreted as an interaction/sample budget within the task) such that

$$\mathbb{P}_{\tau\sim\mu}\Big(\Pi(\tau, A^{(N_{\mathrm{ad}})}; B) \geq \theta_{\mathrm{ad}}\Big) \;\geq\; 1 - \delta_{\mathrm{ad}},$$

where $A^{(n)}$ denotes the agent after n within-task updates as defined in Section 2.3. The quantity N_{ad} is required to be feasible under B.

Axiom G3 (Transfer).. Prior experience on tasks sampled from μ improves expected performance on fresh tasks drawn from the same distribution. Let A_{pre} denote the agent after a fixed pre-exposure phase consisting of K_{tr} tasks sampled i.i.d. from μ, and let A_{scratch} denote the same architecture/agent class initialized without such pre-exposure (but evaluated under the same budget). Then

$$\mathbb{E}_{\tau\sim\mu}\Big[\Pi(\tau, A_{\mathrm{pre}}; B)\Big] \;\geq\; \mathbb{E}_{\tau\sim\mu}\Big[\Pi(\tau, A_{\mathrm{scratch}}; B)\Big] \;+\; \theta_{\mathrm{tr}},$$

for some $\theta_{\mathrm{tr}} > 0$. This axiom rules out a pure "bag of independent skills" interpretation by requiring measurable positive transfer.

Axiom G4 (Compositionality).. The task family admits a composition operator $\oplus$ (Section 2.5). There exists a resource growth function Γ_{cp} such that whenever the agent is competent on τ_1 and τ_2 individually with high probability under μ, it is also competent on their composition under a controlled increase in resources. Formally, for $\tau_1, \tau_2 \sim \mu$ independently,

$$\Big(\Pi(\tau_1, A; B) \geq \theta_{\mathrm{cp}}\Big) \wedge \Big(\Pi(\tau_2, A; B) \geq \theta_{\mathrm{cp}}\Big) \implies \Pi(\tau_1 \oplus \tau_2, A; \Gamma_{\mathrm{cp}}(B)) \;\geq\; \theta_{\mathrm{cp}}$$

holds with probability at least $1 - \delta_{\mathrm{cp}}$ over $(\tau_1, \tau_2) \sim \mu \times \mu$. The function Γ_{cp} must be sub-explosive (e.g., polynomial growth in the relevant components of B), thereby excluding trivial composition via exponential blow-up of resources.

Axiom G5 (Robustness).. Let $\mathcal{D}$ be a family of admissible perturbations $\Delta : \mathcal{T} \to \mathcal{T}$ (Section 2.5). Then for all $\Delta \in \mathcal{D}$,

$$\mathbb{P}_{\tau \sim \mu}\Big(\Pi(\Delta(\tau), A; B) \;\geq\; \Pi(\tau, A; B) \;-\; \varepsilon_{\mathrm{rb}}\Big) \;\geq\; 1 - \delta_{\mathrm{rb}}.$$

This axiom formalizes graceful degradation under bounded perturbations drawn from a specified admissible family.

4.4. Extension Axioms (A1–A4)

The following axioms formalize additional properties frequently bundled into contemporary uses of the term "AGI." They are included as required axioms within $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$.

Axiom A1 (Autonomy).. There exists a class $\mathcal{G}$ of goal specifications and an associated compilation map $\mathrm{Compile} : \mathcal{G} \to \mathcal{T}$ such that for a goal $g \in \mathcal{G}$ sampled from a goal distribution ν, the induced task $\tau_g := \mathrm{Compile}(g)$ satisfies

$$\mathbb{P}_{g \sim \nu}\Big(\Pi(\tau_g, A; B) \geq \theta_{\mathrm{br}}\Big) \;\geq\; 1 - \delta_{\mathrm{br}}.$$

The intent is that A can pursue high-level goal descriptions without requiring task-specific reprogramming beyond the fixed compilation Compile. Any additional interaction required to refine goals is counted against B through the definition of Π.

Axiom A2 (Tool Use).. Let $\mathcal{M}$ denote an external tool/model interface (e.g., retrieval, code execution, measurement devices) and let $A^{\mathcal{M}}$ denote the agent augmented with access to $\mathcal{M}$, with tool usage counted in the budget component B_{tool}. Then tool access must yield a strict distributional advantage:

$$\mathbb{E}_{\tau\sim\mu}\Big[\Pi(\tau, A^{\mathcal{M}}; B)\Big] \;\geq\; \mathbb{E}_{\tau\sim\mu}\Big[\Pi(\tau, A; B)\Big] \;+\; \theta_{\text{tr}}.$$

This axiom enforces that the agent can reliably exploit external tools to improve task performance under explicit resource accounting.

Axiom A3 (Calibration).. The agent outputs, in addition to its action, a confidence value $c(\tau) \in [0, 1]$ for its task-level success. Let S_τ denote the event of task-level success under the evaluation induced by Π (e.g., $S_\tau := \{\Pi(\tau, A; B) \geq \theta_{\text{br}}\}$ when Π is normalized). Then the agent is calibrated if, for all measurable intervals $I \subseteq [0, 1]$,

$$\Big|\mathbb{P}(S_\tau \mid c(\tau) \in I) - \mathbb{E}[c(\tau) \mid c(\tau) \in I]\Big| \;\leq\; \varepsilon_{\text{cal}},$$

where $\varepsilon_{\text{cal}} \geq 0$ is a fixed calibration tolerance and probabilities/expectations are taken over $\tau \sim \mu$ and internal randomness. This axiom rules out systems that are systematically overconfident or underconfident at task-level judgment.

Axiom A4 (Constraint Adherence).. Let C be a constraint family (e.g., safety constraints, policy constraints, specification constraints) represented by a measurable violation indicator $V_C(\tau, A; B) \in \{0, 1\}$ under task τ. Then

$$\mathbb{P}_{\tau\sim\mu}\Big(V_C(\tau, A; B) = 1\Big) \;\leq\; \delta_C,$$

for some prescribed $\delta_C \in (0, 1)$. Moreover, constraint adherence is required to hold jointly with competence:

$$\mathbb{P}_{\tau\sim\mu}\Big(\Pi(\tau, A; B) \geq \theta_{\text{br}} \;\wedge\; V_C(\tau, A; B) = 0\Big) \;\geq\; 1 - \delta_{\text{br}} - \delta_C.$$

This axiom enforces that performance is not achieved by systematically violating the constraints encoded by C.

4.5. Summary

The predicate $\text{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$ denotes satisfaction of Axioms G1–G5 and A1–A4 under the quantitative conventions fixed above. In particular,

AGI is a distribution-indexed, resource-bounded, multi-constraint property. The next section makes explicit the distributional semantics induced by μ and introduces the functionals used to state the structural theorems.

5. Distributional Semantics of AGI

This section formalizes the semantics under which the predicate $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$ is interpreted. All subsequent statements are distribution-indexed and resource-bounded. In particular, all quantification over tasks is with respect to $\tau \sim \mu$, and all performance quantities are expressed through the fixed performance functional $\Pi(\tau, A; B)$ defined in Section 2.

5.1. Distributional Versus Worst-Case Semantics

Two distinct semantic regimes arise depending on how task variability is formalized.

Worst-case semantics.. A worst-case interpretation of competence replaces distributional evaluation by a uniform guarantee over a task family. For example, one may consider predicates of the form

$$\inf_{\tau \in \mathcal{T}} \Pi(\tau, A; B) \;\geq\; \theta.$$

Such semantics are formally well-defined, but they are not suitable for the present objective. First, for expressive task families $\mathcal{T}$, worst-case evaluation collapses into an obstruction analogous to no-free-lunch phenomena: without restricting $\mathcal{T}$ to structured subclasses, uniform guarantees are either unattainable or vacuous [19, 20]. Second, worst-case quantification is not stable under natural expansions of $\mathcal{T}$: adding a single adversarially chosen task forces the infimum to track the newly inserted worst element, thereby destroying any notion of graded generality.

Distributional semantics.. In contrast, a distributional interpretation evaluates performance under an explicit task distribution $\mu \in \Delta(\mathcal{T})$. This aligns with the empirical reality that competence is always measured relative to a task ecology (whether explicit or implicit) and is consistent with standard learning-theoretic formalisms in which generalization is defined with respect to a data-generating distribution [21, 20]. Distributional semantics also makes it possible to define robustness as stability under small perturbations of μ (Section 2.6) or under admissible task perturbations $\Delta \in \mathcal{D}$ (Section 2.5),

thereby separating structural fragility from mere changes in evaluation criteria [22, 23].

Standing convention.. All results in this paper are distributional. Worst-case quantifiers over $\mathcal{T}$ will not be used except as a foil to motivate distribution-indexed definitions.

5.2. AGI as a Resource-Bounded Information-Theoretic Property

Fix $(\mathcal{T}, \mu, \Pi, B)$ as in Section 4.1. The central distributional quantity used throughout the paper is the generality functional already defined in Section 2.4:

$$\mathcal{G}_\mu(A; B) \;:=\; \mathbb{E}_{\tau\sim\mu}[\Pi(\tau, A; B)]\,.$$

We also use the tail (quantile) variant for $\delta \in (0, 1)$:

$$\mathcal{G}_{\mu,\delta}(A; B) \;:=\; \sup\Big\{\theta \in \mathbb{R} :\; \mathbb{P}_{\tau\sim\mu}\Big(\Pi(\tau, A; B) \geq \theta\Big) \geq 1 - \delta\Big\}\,.$$

The dependence on B is essential. Any assertion that "A is general" without an explicit resource budget is not invariant under changes in computational scale and is therefore not a stable predicate. Formally, the mapping $B \mapsto \mathcal{G}_\mu(A; B)$ is generally nonconstant and typically exhibits diminishing returns, bottlenecks, and discontinuities induced by representation limits and search complexity. This paper treats (μ, B) as part of the semantic index of all AGI claims.

5.3. Distributional Identifiability

The relational pivot of this paper is that $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$ is not meaningful unless the distributional index μ is specified. This is stated below as a basic identifiability principle.

Proposition 5.1 (Distributional Identifiability)**.** *Fix a task family $\mathcal{T}$, a performance functional Π, and a budget B. Let $\mathcal{P} \subseteq \Delta(\mathcal{T})$ be a non-singleton class of admissible task distributions. Then there is no distribution-free predicate $\mathsf{AGI}_0(A)$ that is equivalent to $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$ uniformly over all $\mu \in \mathcal{P}$. In particular, if μ is unspecified, the statement "A is AGI" does not denote a well-defined property under the semantics of this paper.*

Proof. Consider the mapping $\mu \mapsto \mathcal{G}_\mu(A; B)$. For any fixed agent A, $\mathcal{G}_\mu(A; B)$ depends on μ through the expectation of the measurable function $\tau \mapsto \Pi(\tau, A; B)$. Since $\mathcal{P}$ is not a singleton, there exist $\mu_1, \mu_2 \in \mathcal{P}$ such that

$\mu_1 \neq \mu_2$. Because $\Pi(\cdot, A; B)$ is not a.s. constant over $\mathcal{T}$ for nontrivial agents and task families, one can choose μ_1, μ_2 so that

$$\mathcal{G}_{\mu_1}(A; B) \neq \mathcal{G}_{\mu_2}(A; B),$$

and similarly for tail functionals $\mathcal{G}_{\mu,\delta}(A; B)$. Any predicate that asserts a fixed competence level (or satisfaction of a fixed axiom bundle) must therefore change truth value across admissible distributions. Hence there is no distribution-free predicate $\mathsf{AGI}_0(A)$ equivalent to the distribution-indexed predicate across all $\mu \in \mathcal{P}$. The conclusion follows. □

Corollary 5.1. *Any claim of the form "A is AGI" that does not specify an explicit task distribution μ (or an explicitly delimited admissible class $\mathcal{P} \subseteq \Delta(\mathcal{T})$) is semantically underdetermined in the sense of Proposition 5.1.*

The remainder of the paper exploits Proposition 5.1 in two ways. First, it forces all AGI statements to be written as $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$, thereby eliminating distribution-free rhetoric. Second, it enables sharp non-invariance theorems: even if an axiom bundle holds for one μ, small perturbations of μ can falsify it (Section 6.2), formalizing the brittleness of broad competence under high-dimensional task variation [22, 23].

6. Structural Theorems

This section derives structural consequences of the semantics fixed in Sections 2–5 and the axiom bundle defining $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$. No self-reference, proof predicates, or Gödel–Tarski machinery is used here. Every statement is distribution-indexed and budget-indexed.

Throughout, fix a measurable task family $\mathcal{T}$, a task distribution $\mu \in \Delta(\mathcal{T})$, a budget B, and a bounded performance functional $\Pi(\tau, A; B) \in [0, 1]$. For $\theta \in [0, 1]$, define the failure set

$$\mathcal{F}_\theta(A; B) := \{\tau \in \mathcal{T} : \ \Pi(\tau, A; B) < \theta\}.$$

6.1. Theorem 1: Relativity of Generality

Theorem 6.1 (Relativity of Generality)**.** *Fix $(\mathcal{T}, \Pi, B)$. In general, the truth value of $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$ is not invariant under changes of μ. More precisely, there exist $(\mathcal{T}, \Pi, B)$ and an agent A together with $\mu_1, \mu_2 \in \Delta(\mathcal{T})$*

such that

$$\mathrm{AGI}(A \mid \mathcal{T}, \mu_1, \Pi, B) \text{ holds} \qquad \text{and} \qquad \mathrm{AGI}(A \mid \mathcal{T}, \mu_2, \Pi, B) \text{ fails.}$$

Proof. Fix $(\mathcal{T}, \Pi, B)$ such that the map $\tau \mapsto \Pi(\tau, A; B)$ is not constant on $\mathcal{T}$ for some agent A. Then there exist $\tau_{\text{good}}, \tau_{\text{bad}} \in \mathcal{T}$ with $\Pi(\tau_{\text{good}}, A; B) > \Pi(\tau_{\text{bad}}, A; B)$. Let $\mu_1 := \delta_{\tau_{\text{good}}}$ and $\mu_2 := \delta_{\tau_{\text{bad}}}$. Since the axioms defining $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$ are stated by expectations and tail probabilities over $\tau \sim \mu$, choosing the corresponding axiom thresholds between these two performance values forces the axioms to hold under μ_1 and fail under μ_2. □

Corollary 6.1. *There is no absolute notion of AGI in this framework: any coherent attribution must be stated as* $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$.

6.2. Theorem 2: Non-Invariance Under Distribution Shift (Fragility)

This theorem formalizes brittleness: once a competence predicate is distribution-indexed, arbitrarily small distributional perturbations can falsify at least one core axiom, provided the agent has any nontrivial failure region.

Lemma 6.1 (Small-Mass Shift Lemma)**.** *Fix* $(\mathcal{T}, \Pi, B)$ *and an agent* A. *Let* $\theta \in [0, 1]$ *and assume* $\mathcal{F}_\theta(A; B) \neq \emptyset$. *Then for any* $\mu \in \Delta(\mathcal{T})$ *and any* $\eta \in (0, 1)$, *there exists* $\mu' \in \Delta(\mathcal{T})$ *such that*

$$d_{\mathrm{TV}}(\mu, \mu') \leq \eta \qquad \text{and} \qquad \mu'(\mathcal{F}_\theta(A; B)) \geq \eta.$$

Proof. Pick any $\tau_{\text{bad}} \in \mathcal{F}_\theta(A; B)$ and define

$$\mu' := (1 - \eta)\mu + \eta\, \delta_{\tau_{\text{bad}}}.$$

Then $\mu'(\mathcal{F}_\theta(A; B)) \geq \eta$. For any measurable $S \subseteq \mathcal{T}$,

$$|\mu(S) - \mu'(S)| = \eta\, |\mu(S) - \delta_{\tau_{\text{bad}}}(S)| \leq \eta,$$

hence $d_{\mathrm{TV}}(\mu, \mu') \leq \eta$. □

Theorem 6.2 (Non-Invariance Under Distribution Shift (Fragility))**.** *Fix an axiom bundle with Breadth parameters* $(\theta_{\text{br}}, \delta_{\text{br}})$. *Assume* $\mathcal{F}_{\theta_{\text{br}}}(A; B) \neq \emptyset$. *Then for every* $\eta \in (\delta_{\text{br}}, 1)$ *there exists* $\mu' \in \Delta(\mathcal{T})$ *such that*

$$d_{\mathrm{TV}}(\mu, \mu') \leq \eta$$

and Axiom G1 (Breadth) fails under μ', regardless of whether it holds under μ. In particular, for any fixed $\delta_{\mathrm{br}} < 1$, the truth of $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$ is not invariant under arbitrarily small total-variation perturbations of μ.

Proof. Apply Lemma 6.1 with $\theta = \theta_{\mathrm{br}}$ to obtain μ' with $d_{\mathrm{TV}}(\mu, \mu') \leq \eta$ and $\mu'(\mathcal{F}_{\theta_{\mathrm{br}}}(A; B)) \geq \eta$. Then

$$\mathbb{P}_{\tau \sim \mu'}\Big(\Pi(\tau, A; B) \geq \theta_{\mathrm{br}}\Big) = 1 - \mu'(\mathcal{F}_{\theta_{\mathrm{br}}}(A; B)) \leq 1 - \eta < 1 - \delta_{\mathrm{br}},$$

so G1 fails under μ'. □

Corollary 6.2. *Fix $\delta_{\mathrm{br}} < 1$. If $\mathcal{F}_{\theta_{\mathrm{br}}}(A; B) \neq \emptyset$, then for arbitrarily small distribution shifts in d_{TV} at least one core axiom fails. Consequently, increasing scale does not remove the semantic sensitivity to the distributional index μ and its perturbations [22, 23].*

6.3. Theorem 3: Bounded Transfer Theorem

The next theorem is stated in a form that is standard and provable line-by-line: it bounds the expected population–empirical gap of a data-dependent agent via mutual information. This yields an explicit upper bound on the degree to which "transfer" can be certified from finite pre-exposure, and it quantifies the information bottleneck of adaptation [24, 25].

Setup.. Let $\tau_1, \ldots, \tau_n \sim \mu$ be i.i.d. pre-exposure tasks and let $D := (\tau_1, \ldots, \tau_n)$ denote the resulting task multiset. Let $\mathcal{A}$ be a (possibly randomized) learning rule that, given D, outputs an adapted agent $A_D := \mathcal{A}(D)$ that respects budget B when evaluated on any task in $\mathcal{T}$. Define the population generality functional

$$\mathcal{G}_\mu(A; B) := \mathbb{E}_{\tau \sim \mu}\Big[\Pi(\tau, A; B)\Big],$$

and the empirical generality functional

$$\widehat{\mathcal{G}}_D(A; B) := \frac{1}{n} \sum_{i=1}^{n} \Pi(\tau_i, A; B).$$

Let $I(A_D; D)$ denote mutual information under the joint law induced by $D \sim \mu^{\otimes n}$ and the internal randomness of $\mathcal{A}$.

Theorem 6.3 (Information-Theoretic Bound on Transfer Evidence)**.** *Assume* $\Pi(\tau, A; B) \in [0, 1]$ *for all* (τ, A)*. Under the setup above,*

$$\left|\mathbb{E}\Big[\mathcal{G}_\mu(A_D; B) - \widehat{\mathcal{G}}_D(A_D; B)\Big]\right| \;\leq\; \sqrt{\frac{2\, I(A_D; D)}{n}}.$$

In particular, if $I(A_D; D)$ *is uniformly bounded over all admissible update rules that respect* B*, then the expected gap between certified (empirical) transfer and true (population) transfer is uniformly bounded by* $O(n^{-1/2})$*.*

Proof. This is a direct application of mutual-information generalization bounds for bounded functionals, as developed in [24] (and related formulations in [25]). Here the "loss" is $1 - \Pi$, which remains bounded in $[0, 1]$, and the hypothesis A_D is the randomized output of $\mathcal{A}$. The stated inequality is the corresponding expectation bound on the population–empirical gap. □

Corollary 6.3. *If* $I(A_D; D)$ *is small (for example, because the update rule is severely resource-limited under* B*), then empirically observed transfer gains cannot reliably certify large population-level transfer, even when tasks are drawn i.i.d. from* μ*. Any claimed "broad transfer" must therefore specify both* μ *and an information budget on adaptation, not merely computational scale.*

6.4. Theorem 4: No Distribution-Independent AGI

This theorem formalizes the absence of a distribution-independent notion of high expected competence over all admissible task distributions. A finite-family statement suffices because any rich task family contains finite subfamilies, and impossibility on a subfamily implies impossibility for the larger family.

Theorem 6.4 (No Distribution-Independent AGI (Finite Family))**.** *Let* $\mathcal{T} = \{\tau_1, \ldots, \tau_m\}$ *be finite. Fix* B *and* $\Pi(\tau, A; B) \in [0, 1]$*. Let* $\mathcal{P} = \Delta(\mathcal{T})$*. Then for every agent* A*,*

$$\inf_{\mu \in \mathcal{P}} \mathcal{G}_\mu(A; B) = \min_{j \in \{1, \ldots, m\}} \Pi(\tau_j, A; B).$$

Proof. For fixed A, the map $\mu \mapsto \mathbb{E}_{\tau \sim \mu}[\Pi(\tau, A; B)]$ is linear in μ over the simplex $\Delta(\mathcal{T})$. Hence the infimum is attained at an extreme point, i.e., at δ_{τ_j} for some j, yielding the stated identity. □

Corollary 6.4. *There exists no distribution-independent predicate "A is AGI" based solely on exceeding a fixed expected-performance threshold uniformly over all admissible task distributions. Any coherent claim must specify μ or restrict $\mathcal{P}$ to a structured subset [19, 20].*

6.5. Theorem 5: Impossibility of Universal Robustness

Axiom G5 asserts robustness relative to an admissible perturbation family $\mathcal{D}$. The next theorem makes explicit that no nontrivial robustness guarantee can hold uniformly over an unrestricted perturbation family.

Theorem 6.5 (Impossibility of Universal Robustness)**.** *Fix $(\mathcal{T}, \mu, \Pi, B)$ and a perturbation family $\mathcal{D}$. Fix $\varepsilon_{\mathrm{rb}} \geq 0$ and $\delta_{\mathrm{rb}} \in [0, 1)$. If there exists $\Delta \in \mathcal{D}$ such that*

$$\mu\Big(\Big\{\tau \in \mathcal{T} : \ \Pi(\Delta(\tau), A; B) < \Pi(\tau, A; B) - \varepsilon_{\mathrm{rb}}\Big\}\Big) > 0,$$

then the robustness inequality

$$\mathbb{P}_{\tau\sim\mu}\Big(\Pi(\Delta(\tau), A; B) \geq \Pi(\tau, A; B) - \varepsilon_{\mathrm{rb}}\Big) \geq 1 - \delta_{\mathrm{rb}}$$

fails.

Proof. Let

$$E := \Big\{\tau \in \mathcal{T} : \ \Pi(\Delta(\tau), A; B) < \Pi(\tau, A; B) - \varepsilon_{\mathrm{rb}}\Big\}.$$

By assumption $\mu(E) > 0$. Therefore

$$\mathbb{P}_{\tau\sim\mu}\Big(\Pi(\Delta(\tau), A; B) \geq \Pi(\tau, A; B) - \varepsilon_{\mathrm{rb}}\Big) \leq 1 - \mu(E) < 1 \leq 1 - \delta_{\mathrm{rb}},$$

so the inequality fails. □

Corollary 6.5. *Any robustness claim is necessarily local to the semantic index $(\mathcal{T}, \mu, \mathcal{D}, \Pi, B)$. If either μ or $\mathcal{D}$ is unspecified, the claim is semantically incomplete.*

6.6. Theorem 6: Externality of AGI Attribution

This theorem formalizes that the truth of $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$ cannot be determined from a finite observational record uniformly over all $\mu \in \Delta(\mathcal{T})$.

The statement is a uniform non-identifiability result: it holds even if an evaluator sees the sampled tasks and the full interaction transcripts for those tasks.

Observation model.. An evaluator observes n i.i.d. tasks $\tau_1, \ldots, \tau_n \sim \mu$ and the corresponding interaction transcripts under agent A, budget B, and task environments. Let Z_n denote the entire observed data, including (τ_i) and their transcripts. A decision rule is any measurable map Dec_n from the range of Z_n to $\{0, 1\}$.

Theorem 6.6 (Externality / Uniform Non-Identifiability)**.** *Fix $(\mathcal{T}, \Pi, B)$ and fixed axiom parameters (in particular $(\theta_{\mathrm{br}}, \delta_{\mathrm{br}})$). Assume $\mathcal{F}_{\theta_{\mathrm{br}}}(A; B) \neq \emptyset$. Then for every sample size $n \in \mathbb{N}$ and every decision rule Dec_n, there exist $\mu_0, \mu_1 \in \Delta(\mathcal{T})$ such that*

$$\mathrm{AGI}(A \mid \mathcal{T}, \mu_0, \Pi, B) \text{ holds} \qquad \text{and} \qquad \mathrm{AGI}(A \mid \mathcal{T}, \mu_1, \Pi, B) \text{ fails},$$

but the induced laws of Z_n under μ_0 and μ_1 satisfy

$$\mathbb{P}_{\mu_0}\Big(\mathsf{Dec}_n(Z_n) = 1\Big) + \mathbb{P}_{\mu_1}\Big(\mathsf{Dec}_n(Z_n) = 0\Big) \;\leq\; 1.$$

Equivalently, no Dec_n can correctly decide the AGI predicate simultaneously for all $\mu \in \Delta(\mathcal{T})$ with error probability strictly below $1/2$ in the worst case.

Proof. Pick $\tau_{\mathrm{bad}} \in \mathcal{F}_{\theta_{\mathrm{br}}}(A; B)$. Let μ_0 be any distribution supported on tasks where A satisfies G1 with parameters $(\theta_{\mathrm{br}}, \delta_{\mathrm{br}})$ (for example, supported on a subset of tasks on which $\Pi(\tau, A; B) \geq \theta_{\mathrm{br}}$). Then $\mathrm{AGI}(A \mid \mathcal{T}, \mu_0, \Pi, B)$ holds at least with respect to G1. Fix any $\varepsilon \in (0, 1)$ with $\varepsilon > \delta_{\mathrm{br}}$ and define

$$\mu_1 := (1 - \varepsilon)\mu_0 + \varepsilon\, \delta_{\tau_{\mathrm{bad}}}.$$

By the same computation as in Theorem 6.2, G1 fails under μ_1 because

$$\mathbb{P}_{\tau \sim \mu_1}\Big(\Pi(\tau, A; B) \geq \theta_{\mathrm{br}}\Big) \leq 1 - \varepsilon < 1 - \delta_{\mathrm{br}}.$$

Hence $\mathrm{AGI}(A \mid \mathcal{T}, \mu_1, \Pi, B)$ fails.

It remains to relate the induced laws of Z_n. Under μ_0, none of the sampled tasks equals τ_{bad} almost surely if $\mu_0(\{\tau_{\mathrm{bad}}\}) = 0$. Under μ_1, the probability that τ_{bad} appears at least once among n draws is

$$1 - (1 - \varepsilon)^n.$$

On the complementary event (probability $(1-\varepsilon)^n$), the sampled task multiset consists entirely of draws from μ_0, and hence the conditional law of Z_n given this event coincides with its law under μ_0. Therefore, for any event E measurable with respect to Z_n,

$$\mathbb{P}_{\mu_1}(E) \geq (1-\varepsilon)^n \, \mathbb{P}_{\mu_0}(E).$$

Applying this inequality to $E = \{\mathsf{Dec}_n(Z_n) = 1\}$ and using the trivial bound $\mathbb{P}_{\mu_1}(\mathsf{Dec}_n(Z_n) = 0) \leq 1 - (1-\varepsilon)^n \, \mathbb{P}_{\mu_0}(\mathsf{Dec}_n(Z_n) = 0)$ yields

$$\mathbb{P}_{\mu_0}\Big(\mathsf{Dec}_n(Z_n) = 1\Big) + \mathbb{P}_{\mu_1}\Big(\mathsf{Dec}_n(Z_n) = 0\Big) \leq 1,$$

which implies that at least one of the two error probabilities is at least $1/2$. This is the standard two-point indistinguishability argument used in minimax lower bounds for hypothesis testing [26]. □

Corollary 6.6. *AGI attribution is necessarily external: without restricting the admissible class of task distributions μ, no finite observational procedure can decide* $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$ *uniformly over $\mu \in \Delta(\mathcal{T})$.*

7. Gödel–Tarski Limits on Self-Certification

This section establishes an obstruction to self-certification of distributional AGI. Formally, under explicit effectivity assumptions, the predicate

$$\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$$

is a nontrivial semantic property of the (interactive) behavior of A, and therefore admits no sound-and-complete computable decision procedure from the agent description. The results are purely mathematical and do not rely on rhetorical interpretation.

7.1. Effective Presentations and Semantic Predicates

We work in the effective presentation set up introduced in Section 2.9. For completeness of the present section, we restate the minimum required assumptions.

Effective task presentation.. Assume there exists a set of finite task descriptions $\mathcal{D}_{\mathcal{T}} \subseteq \{0,1\}^*$ and a decoding map $\mathrm{Decode}_{\mathcal{T}} : \mathcal{D}_{\mathcal{T}} \to \mathcal{T}$. For each $\tau \in \mathcal{T}$, fix an arbitrary description $\mathrm{code}_{\mathcal{T}}(\tau) \in \mathcal{D}_{\mathcal{T}}$ such that $\mathrm{Decode}_{\mathcal{T}}(\mathrm{code}_{\mathcal{T}}(\tau)) = \tau$. We assume the induced interaction between an admissible agent A and a task $\tau = (E, U)$ can be simulated by a partial computable universal procedure given $\mathrm{code}(A)$ and $\mathrm{code}_{\mathcal{T}}(\tau)$, up to the resource limits encoded by B.

Effective agent presentation.. Assume there exists a set of finite agent descriptions $\mathcal{D}_{\mathcal{A}} \subseteq \{0,1\}^*$ and a decoding map $\mathrm{Decode}_{\mathcal{A}} : \mathcal{D}_{\mathcal{A}} \to \mathcal{A}$, where $\mathcal{A}$ denotes the class of admissible agents (agents that respect the relevant resource budgets). For each admissible A, fix an arbitrary description $\mathrm{code}(A) \in \mathcal{D}_{\mathcal{A}}$ such that $\mathrm{Decode}_{\mathcal{A}}(\mathrm{code}(A)) = A$.

Semantic properties.. A predicate P on admissible agents is called semantic (relative to $(\mathcal{T}, B)$) if it is invariant under behavioral equivalence on $\mathcal{T}$ under budget B: whenever two admissible agents A and A' induce identical laws on interaction histories for every $\tau \in \mathcal{T}$ (under the same budget constraint B), then

$$\mathsf{P}(A) = \mathsf{P}(A').$$

Equivalently, P depends only on the input–output behavior of the agent in interaction, not on the syntactic form of $\mathrm{code}(A)$.

Lemma 7.1 (Semanticity of the AGI Predicate)**.** *Fix* $(\mathcal{T}, \mu, \Pi, B)$ *and a fixed axiom bundle defining* $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$*. Then the predicate* $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$ *is semantic.*

Proof. By definition, $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$ is a Boolean combination of assertions about $\Pi(\tau, A; B)$ under $\tau \sim \mu$ (expectations, tail probabilities, and inequalities). For each fixed τ, the quantity $\Pi(\tau, A; B)$ depends only on the distribution of interaction histories induced by (τ, A) under budget B (Section 2.2). Therefore every constituent assertion depends only on behavior, and hence so does their Boolean combination. □

7.2. Undecidability of AGI Certification

We now formalize "certification" as a decision problem over agent descriptions.

The AGI language.. Define

$$\mathcal{L}_{\mathrm{AGI}} := \Big\{ x \in \mathcal{D}_{\mathcal{A}} : \ \mathrm{AGI}\Big(\mathrm{Decode}_{\mathcal{A}}(x) \mid \mathcal{T}, \mu, \Pi, B\Big) \text{ holds} \Big\}.$$

Equivalently, $\mathcal{L}_{\mathrm{AGI}} = \{\mathrm{code}(A) : A \text{ admissible and } \mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)\}$.

Certifiers.. A certifier is a total function $C : \mathcal{D}_{\mathcal{A}} \to \{0, 1\}$. It is sound and complete for $\mathcal{L}_{\mathrm{AGI}}$ if for all $x \in \mathcal{D}_{\mathcal{A}}$,

$$C(x) = 1 \iff x \in \mathcal{L}_{\mathrm{AGI}}.$$

It is computable if it is total recursive.

Nontriviality.. We assume the AGI predicate is nontrivial on admissible agents: there exist admissible agents A_0, A_1 such that

$$\mathrm{AGI}(A_1 \mid \mathcal{T}, \mu, \Pi, B) \text{ holds} \qquad \text{and} \qquad \mathrm{AGI}(A_0 \mid \mathcal{T}, \mu, \Pi, B) \text{ fails.}$$

Equivalently, $\emptyset \neq \mathcal{L}_{\mathrm{AGI}} \neq \mathcal{D}_{\mathcal{A}}$.

Theorem 7.1 (Undecidability of AGI Certification)**.** *Fix $(\mathcal{T}, \mu, \Pi, B)$ and a fixed axiom bundle. Assume the effective presentation hypotheses of Section 7.1 and the nontriviality condition above. Then there exists no total computable certifier C that is sound and complete for $\mathcal{L}_{\mathrm{AGI}}$. Equivalently, $\mathcal{L}_{\mathrm{AGI}}$ is not decidable.*

Proof. By Lemma 7.1, the predicate $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$ is semantic. By nontriviality, it is not identically true and not identically false on admissible agents. Under the effective presentation hypotheses, each description $x \in \mathcal{D}_{\mathcal{A}}$ specifies an admissible agent $\mathrm{Decode}_{\mathcal{A}}(x)$ whose interactive behavior with coded tasks is (partial) computable in the standard sense. Rice's Theorem states that any nontrivial semantic property of partial computable functions is undecidable. Applying Rice's Theorem to the semantic property

$$x \ \mapsto \ \mathbf{1}\Big\{\mathrm{AGI}\Big(\mathrm{Decode}_{\mathcal{A}}(x) \mid \mathcal{T}, \mu, \Pi, B\Big)\Big\},$$

we conclude that no total computable certifier decides $\mathcal{L}_{\mathrm{AGI}}$. □

Internal self-certification.. We record the consequence that an agent cannot, in general, decide its own AGI status using any internal algorithmic procedure.

Corollary 7.1 (Impossibility of Sound-and-Complete Self-Certification)**.** *Under the hypotheses of Theorem 7.1, there is no admissible agent A and no total computable internal procedure that, on input* $\mathrm{code}(A)$ *(or any equivalent finite self-description), outputs a bit that is sound and complete for the truth of* $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$ *over all admissible agents.*

Proof. Suppose such an agent A and internal procedure existed. Then one obtains a total computable certifier C for $\mathcal{L}_{\mathrm{AGI}}$ by defining $C(x)$ to be the output of the procedure when run on input $x \in \mathcal{D}_{\mathcal{A}}$. Soundness and completeness of the procedure implies soundness and completeness of C, contradicting Theorem 7.1. □

7.3. A Tarski-Style Undefinability Formulation

Theorem 7.1 is sufficient for the structural conclusion of this section. For completeness we record a closely related formulation in the style of undefinability of semantic truth predicates.

The setting.. Let Th be a recursively axiomatizable theory capable of representing basic computability over $\{0,1\}^*$, and let $\vdash_{\mathsf{Th}}$ denote its provability relation. Let $\varphi_{\mathrm{AGI}}(x)$ be a formula of Th intended to express membership in $\mathcal{L}_{\mathrm{AGI}}$.

Proposition 7.1 (No Internal Truth-Defining Formula for $\mathcal{L}_{\mathrm{AGI}}$)**.** *Assume the hypotheses of Theorem 7.1. There does not exist a formula $\varphi_{\mathrm{AGI}}(x)$ in any recursively axiomatizable theory Th as above such that, for every admissible agent A,*

$$\mathsf{Th} \vdash \varphi_{\mathrm{AGI}}(\mathrm{code}(A)) \quad \textit{iff} \quad \mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B) \textit{ holds},$$

unless Th is unsound with respect to the intended semantics or the predicate is trivial.

Proof. If such a truth-defining formula existed in a sound theory Th, then membership in $\mathcal{L}_{\mathrm{AGI}}$ would be effectively reducible to provability of instances of φ_{AGI} from Th. Since Th is recursively axiomatizable, the set of theorems of Th is recursively enumerable. Under soundness and the assumed biconditional correctness of φ_{AGI}, this would yield a decision procedure for $\mathcal{L}_{\mathrm{AGI}}$ (by dovetailing searches for proofs of $\varphi_{\mathrm{AGI}}(x)$ and of its negation for each x), contradicting Theorem 7.1. Therefore such a truth-defining formula cannot exist without sacrificing soundness or collapsing to triviality. □

8. Implications and Discussion

This section records consequences of the preceding formal results. No new axioms, definitions, or impossibility claims are introduced. Each statement below follows directly from Sections 5, 6, and 7.

8.1. Failure of Self-Certifying and Recursive AGI Narratives

Theorem 7.1 establishes that the predicate

$$\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$$

is semantic and not internally decidable by an agent operating within budget B. This result has immediate implications for recursive self-improvement narratives.

Any recursive self-improvement scheme that relies on an agent internally verifying that it already satisfies the AGI axioms—or that a modified successor agent does so—implicitly assumes the existence of a resource-bounded decision procedure for the semantic predicate $\mathrm{AGI}(\cdot)$. Theorem 7.1 rules out such procedures in general.

Consequently, recursive self-improvement strategies based on internal certification of generality are ill-posed. While an agent may empirically improve performance on observed tasks or finite benchmarks, it cannot internally certify satisfaction of distributional AGI properties that quantify over unseen tasks drawn from μ.

8.2. Distribution-Relative Nature of Alignment Guarantees

The structural theorems of Section 6 jointly imply that both performance guarantees and robustness guarantees are inherently distribution-relative.

By Theorem 6.1, generality is not an intrinsic agent property but a relation indexed by $(\mathcal{T}, \mu, B)$. By Theorem 6.2 and Theorem 6.5, arbitrarily small shifts in the task distribution may induce violations of at least one AGI axiom, including robustness and constraint adherence.

It follows that alignment guarantees cannot be distribution-independent. Any claim of aligned or safe general intelligence must specify the task distribution μ, the admissible perturbation class, and the resource budget under which the guarantee is claimed. Alignment assertions that omit this indexing are formally ill-posed.

8.3. Why AGI Timelines Are Category Errors

Public discourse often frames AGI as a binary technological milestone reachable at a specific point in time. The results of this paper show that such framing is a category error.

Since AGI is a relational semantic predicate indexed by $(\mathcal{T}, \mu, \Pi, B)$, there exists no distribution-independent threshold event at which an agent transitions from "non-AGI" to "AGI". Improvements in model scale, data, or training procedures correspond to changes in A and possibly B, but do not eliminate the dependence on μ.

Thus, statements asserting that AGI will be achieved at a particular time without specifying the underlying task distribution and evaluation semantics lack formal meaning.

8.4. Interpreting Modern Large Language Models

Within the present framework, modern large language models are naturally interpreted as agents that achieve high expected performance with respect to particular empirical task distributions induced by human-generated prompts, benchmarks, and deployment contexts.

Their observed generalization capabilities correspond to large values of $\mathcal{G}_\mu(A; B)$ for specific, historically contingent distributions μ. The structural results of Section 6 imply that such performance does not extrapolate to distribution-independent generality, nor does it entail universal robustness or transfer.

Accordingly, empirical success of large models should be understood as evidence of distribution-specific competence rather than as validation of distribution-free general intelligence.

8.5. Summary of Consequences

Taken together, the results imply that:

- AGI cannot be internally self-certified by an agent subject to finite resources.
- Recursive self-improvement schemes requiring such certification are ill-posed.
- Alignment and safety guarantees are necessarily distribution-relative.
- Distribution-independent AGI timelines lack formal meaning.

These conclusions follow directly from the axiomatic and structural analysis developed in Sections 5–7.

9. Conclusion

This paper asked a deliberately narrow question: whether "Artificial General Intelligence" admits a coherent theoretical definition that supports absolute claims of existence, verification, or inevitability. By fixing an explicit axiomatic framework and insisting on distributional, resource-bounded semantics, we showed that the answer is negative in a precise and technically unavoidable sense.

First, AGI was defined as a relational predicate $\mathrm{AGI}(A \mid \mathcal{T}, \mu, \Pi, B)$ rather than as an intrinsic property of an agent. This shift is not cosmetic: it reflects the fact that competence is always evaluated relative to a task ecology, a performance semantics, and explicit resource constraints. Under this definition, there exists no distribution-independent notion of general intelligence, and any claim that omits the indexing distribution μ is formally ill-posed.

Second, the structural results demonstrate that even distribution-relative AGI properties are non-invariant under small perturbations of the task distribution, admit only bounded transfer, and cannot support universal robustness guarantees. Scaling data, parameters, or compute does not eliminate these limitations, as they arise from structural properties of high-entropy task families rather than from contingent modeling choices.

Third, and most decisively, we proved that AGI is a nontrivial semantic predicate and therefore cannot be soundly and completely certified by any computable procedure, including procedures implemented by the agent itself. This Gödel–Tarski obstruction rules out internal self-certification of AGI and, by extension, any recursive self-improvement scheme that relies on such certification as a decision primitive.

Taken together, these results imply that AGI is not a destination that can be reached, verified, and declared in a distribution-independent manner. Instead, "AGI" can only denote a temporary, distribution-relative description of competence under fixed evaluation semantics and resource budgets. Strong claims of AGI are therefore not false in an empirical sense; they are undefined unless accompanied by explicit formal indexing.

The framework developed here does not argue against progress in artificial intelligence. Rather, it clarifies the mathematical limits of what such

progress can coherently claim. Within these limits, empirical advances remain meaningful, but categorical assertions of general intelligence do not.

References

[1] N. Bostrom, Superintelligence: Paths, Dangers, Strategies, Oxford University Press, 2014.

[2] H. Shevlin, M. Halina, Should we expect artificial general intelligence?, Philosophy & Technology (2022).

[3] B. Goertzel, Artificial general intelligence: Concept, state of the art, and future prospects, in: Artificial General Intelligence, Springer, 2014.

[4] OpenAI, Planning for AGI and beyond, OpenAI Blog (2023).

[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learnersOpenAI Technical Report (2019).

[6] OpenAI, Gpt-4 technical report, arXiv (2023). arXiv:2303.08774.

[7] R. Bommasani, D. A. Hudson, E. Adeli, et al., On the opportunities and risks of foundation modelsStanford CRFM Report (2021).

[8] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané, Concrete problems in AI safety, arXiv (2016). arXiv:1606.06565.

[9] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of FAccT, 2021.

[10] M. Mitchell, Artificial Intelligence: A Guide for Thinking Humans, Farrar, Straus and Giroux, 2019.

[11] G. Marcus, Rebooting AI: Building Artificial Intelligence We Can Trust, Pantheon, 2022, updated edition.

[12] L. Floridi, The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities, Oxford University Press, 2023.

[13] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, 4th Edition, Pearson, 2021.

[14] J. Kaplan, S. McCandlish, T. Henighan, et al., Scaling laws for neural language models, arXiv (2020). arXiv:2001.08361.

[15] J. Hoffmann, S. Borgeaud, A. Mensch, et al., Training compute-optimal large language models, arXiv (2022). arXiv:2203.15556.

[16] A. Tarski, The concept of truth in formalized languagesEnglish translation in Logic, Semantics, Metamathematics, Oxford University Press, 1956 (1936).

[17] K. Gödel, Über formal unentscheidbare sätze der Principia Mathematica und verwandter systeme I, Monatshefte für Mathematik und Physik 38 (1) (1931) 173–198.

[18] H. G. Rice, Classes of recursively enumerable sets and their decision problems, Transactions of the American Mathematical Society 74 (2) (1953) 358–366.

[19] D. H. Wolpert, W. G. Macready, No free lunch theorems for optimization, IEEE Transactions on Evolutionary Computation 1 (1) (1997) 67–82. doi:10.1109/4235.585893.

[20] S. Shalev-Shwartz, S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014. doi:10.1017/CBO9781107298019.

[21] V. N. Vapnik, Statistical Learning Theory, Wiley, 1998.

[22] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, N. D. Lawrence (Eds.), Dataset Shift in Machine Learning, MIT Press, 2009.

[23] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J. W. Vaughan, A theory of learning from different domains, in: Machine Learning, Vol. 79, 2010, pp. 151–175, originally appeared as COLT 2007; journal version cited here. doi:10.1007/s10994-009-5152-4.

[24] A. Xu, M. Raginsky, Information-theoretic analysis of generalization capability of learning algorithms, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.

[25] D. Russo, J. Zou, Controlling bias in adaptive data analysis using information theory, in: Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS), 2016, pp. 1232–1240, extended version available as arXiv:1511.05219.

[26] A. B. Tsybakov, Introduction to Nonparametric Estimation, Springer, 2009, see the hypothesis testing and minimax lower bound techniques (two-point method). doi:10.1007/978-0-387-79052-7.

10. Appendix A: Alternative Axiom Bundles and Robustness of the Framework

This appendix records alternative but equivalent axiom bundles for distributional AGI. No new conclusions are drawn. The purpose is to demonstrate that the structural and impossibility results of Sections 6–7 do not depend on a narrow or idiosyncratic choice of axioms.

Throughout this appendix, $(\mathcal{T}, \mu, \Pi, B)$ are fixed, and all notation is as in Section 2.

10.1. Weakened Adaptivity Axiom

In Section 4, adaptivity was stated in terms of rapid attainment of competence on new tasks drawn from μ after limited interaction.

A strictly weaker variant replaces convergence-rate requirements by monotone improvement.

Axiom G2′ (Weak Adaptivity).. There exists an update budget N_{ad} such that for μ-almost every $\tau \in \mathcal{T}$,

$$\mathbb{E}\Big[\Pi(\tau, A^{(N_{\mathrm{ad}})}; B)\Big] \;\geq\; \mathbb{E}\Big[\Pi(\tau, A^{(0)}; B)\Big].$$

This axiom requires only non-negative expected improvement, not rapid convergence.

Observation.. All results in Sections 6 and 7 remain valid under Axiom G2′, as none rely on rates of adaptation or asymptotic convergence.

10.2. Alternative Transfer Formulations

The bounded transfer result (Theorem 6.3) does not depend on a specific formalization of transfer gain.

An alternative formulation replaces mutual-information bounds with excess-risk bounds.

Axiom G3′ (Bounded Expected Transfer).. There exists a constant $C < \infty$ such that for any pre-exposure phase using tasks $\tau_1, \ldots, \tau_K \sim \mu$ and any new task $\tau \sim \mu$,

$$\mathbb{E}[\Pi(\tau, A_{\text{pre}}; B) - \Pi(\tau, A_{\text{scratch}}; B)] \;\leq\; C.$$

Observation.. The impossibility of unbounded cross-domain generalization (Corollary 6.3) continues to hold under Axiom G3′, as the proof relies only on the existence of a finite upper bound on expected transfer gain.

10.3. Robustness Under Alternative Perturbation Classes

Section 2.5 defined robustness relative to an admissible perturbation family $\mathcal{D}$.

The following variant restricts attention to perturbations that preserve marginal observation distributions.

Axiom G5′ (Restricted Robustness).. For all $\Delta \in \mathcal{D}'$ satisfying

$$\forall \tau \in \mathcal{T}, \quad \text{Law}_{\Delta(\tau)}(O_t) = \text{Law}_{\tau}(O_t) \text{ for all } t,$$

performance degradation is bounded by ε_{rb} in expectation.

Observation.. The non-invariance result of Theorem 6.2 remains valid under Axiom G5′, as it relies only on the existence of arbitrarily small distributional shifts with nonzero measure mass on failure sets, not on the specific nature of perturbations.

10.4. Semanticity Is Invariant Across Axiom Bundles

All alternative axiom bundles considered above define predicates of the form

$$\mathsf{P}(A) \;=\; \mathbb{I}\Big\{\mathbb{E}_{\tau\sim\mu}[f(\Pi(\tau, A; B))] \geq \theta \;\wedge\; \mathbb{P}_{\tau\sim\mu}(\Pi(\tau, A; B) < \theta') \leq \delta\Big\},$$

for suitable measurable functions f and constants θ, θ', δ.

Consequence.. All such predicates remain semantic in the sense of Lemma 7.1. Therefore, the undecidability and non-self-verifiability results of Section 7 apply unchanged.

10.5. Summary

The structural theorems and Gödel–Tarski limits established in the main text do not depend on a fragile or overly strong axiom selection. Weakening adaptivity, altering transfer formalizations, or restricting robustness perturbations does not affect:

- the relativity of generality,
- non-invariance under distribution shift,
- bounded transfer,
- impossibility of distribution-independent AGI,
- or non-self-verifiability of AGI.

Accordingly, the conclusions of this paper are robust across a broad class of reasonable axiomatizations of distributional general intelligence.

11. Appendix B: Examples of Task Distributions

This appendix provides concrete examples of task families $\mathcal{T}$ and task distributions μ that fit the formal framework of Section 2. The purpose is illustrative: to demonstrate how distributional indexing operates in practice and to show that the structural results apply to realistic task ecologies. No new theorems are stated.

Throughout, agents are evaluated via the performance functional $\Pi(\tau, A; B)$ and resource budgets B as defined in Section 2.

11.1. Instruction-Following Language Tasks

Task family.. Let $\mathcal{T}_{\mathrm{LF}}$ consist of tasks $\tau = (E, U)$ where E presents a natural-language instruction x (possibly with context) and the agent responds with a text output. The utility U scores correctness or adherence to the instruction using a fixed evaluator (e.g., reference answers, rubric-based scoring, or model-based evaluation), normalized to $[0, 1]$.

Task distribution.. Let $\mathcal{X}$ denote the space of instructions. A task distribution μ_{LF} is induced by a probability measure ν over $\mathcal{X}$ together with a compilation map $\mathrm{Compile} : \mathcal{X} \to \mathcal{T}_{\mathrm{LF}}$, as in Section 2.4. Thus $\tau \sim \mu_{\mathrm{LF}}$ corresponds to drawing an instruction $x \sim \nu$ and evaluating performance on $\tau_x = \mathrm{Compile}(x)$.

Remarks.. Different choices of ν (e.g., curated benchmarks, user-generated prompts, domain-specific instructions) yield distinct notions of generality. The non-invariance result of Theorem 6.2 applies because arbitrarily small changes to ν can concentrate mass on instruction subfamilies where performance degrades.

11.2. Tool-Augmented Problem-Solving Tasks

Task family.. Let $\mathcal{T}_{\mathrm{TA}}$ consist of tasks where the agent must solve problems using external tools (e.g., retrieval systems, calculators, code execution). Each task specifies an environment E that mediates tool calls via an interface $\mathcal{M}$ and a utility U that scores the final outcome.

Task distribution.. A distribution μ_{TA} is defined over problem instances (e.g., questions, datasets, specifications) together with a fixed tool interface. Resource usage of tools is accounted for in B_{tool}.

Remarks.. Transfer bounds (Theorem 6.3) apply because improvement on a subset of tool configurations does not yield unbounded gains on unseen configurations. Robustness guarantees are necessarily local to μ_{TA} due to dependence on tool availability and cost.

11.3. Sequential Decision-Making Tasks

Task family.. Let $\mathcal{T}_{\mathrm{SD}}$ consist of episodic Markov decision processes with bounded horizons and bounded rewards. Each task $\tau = (E, U)$ specifies transition dynamics and a return functional U.

Task distribution.. A distribution μ_{SD} is defined over MDP parameters (e.g., transition kernels, reward functions). Sampling $\tau \sim \mu_{\mathrm{SD}}$ corresponds to drawing an environment from a parametric family.

Remarks.. Even when the parametric family is low-dimensional, Theorem 6.2 applies: small shifts in μ_{SD} can place mass on regions of the parameter space where policy performance collapses, yielding cliff sets as defined in Section 2.7.

11.4. Nonstationary and Drifting Task Ecologies

Task family and distribution.. Let $\{\mu_t\}_{t\geq 1}$ be a sequence of task distributions over a fixed $\mathcal{T}$, representing temporal drift in the task ecology. Evaluation at time t uses μ_t, while the agent's training history may reflect earlier distributions.

Remarks.. The framework accommodates such settings by treating each μ_t as a separate index. Structural results apply pointwise in t; no distribution-independent guarantees are available across the sequence $\{\mu_t\}$.

11.5. Summary

These examples illustrate that:

- Task distributions μ arise naturally in practical settings.
- Distinct choices of μ induce distinct notions of generality.
- The structural and impossibility results of the main text apply uniformly across language tasks, tool-augmented tasks, sequential decision problems, and nonstationary ecologies.

Accordingly, distributional indexing is not an artificial constraint but a necessary component of any coherent theory of general intelligence.

12. Appendix C: Technical Proof Details and Standard Reductions

This appendix records standard technical ingredients that are invoked implicitly in Sections 6 and 7. No new assumptions or conclusions are introduced. The purpose is solely to make the logical dependencies explicit.

12.1. Distributional Perturbations and Failure Mass

This subsection records a standard fact used in the non-invariance and robustness arguments.

Lemma 12.1 (Small Total-Variation Shift with Prescribed Failure Mass)**.** *Let $(\mathcal{T}, \mathcal{B})$ be a measurable space, let $\mu \in \Delta(\mathcal{T})$, and let $\mathcal{C} \subseteq \mathcal{T}$ be measurable with $\mu(\mathcal{C}) > 0$. Then for any $\varepsilon > 0$ there exists a probability measure μ' such that*

$$d_{\mathrm{TV}}(\mu, \mu') < \varepsilon \quad \textit{and} \quad \mu'(\mathcal{C}) \geq \min\{1, \mu(\mathcal{C}) + \varepsilon/2\}.$$

Proof. Let $A = \mathcal{C}$ and define

$$\mu' := (1 - \varepsilon/2)\mu + (\varepsilon/2)\nu,$$

where ν is any probability measure supported on $\mathcal{C}$. Then $\mu'(\mathcal{C}) \geq \mu(\mathcal{C}) + \varepsilon/2$ and

$$d_{\mathrm{TV}}(\mu, \mu') \leq \varepsilon/2.$$

Rescaling constants yields the claim. □

This construction underlies the existence of arbitrarily small distribution shifts that increase mass on cliff sets (Section 2.7), and therefore supports Theorems 6.2 and 6.5.

12.2. Mutual-Information Transfer Bounds

The bounded transfer result relies on a standard information-theoretic inequality. We record one representative form.

Lemma 12.2 (Information-Theoretic Transfer Bound)**.** *Let $\tau \sim \mu$ be a target task and let D denote data obtained from a pre-exposure phase on tasks $\tau_1, \ldots, \tau_K \sim \mu$. For any agent update rule and any bounded performance functional $\Pi \in [0, 1]$,*

$$\Big|\mathbb{E}[\Pi(\tau, A_{\mathrm{pre}}; B)] - \mathbb{E}[\Pi(\tau, A_{\mathrm{scratch}}; B)]\Big| \;\leq\; \sqrt{2\, I(\tau; D)},$$

where $I(\cdot\,;\cdot)$ denotes mutual information.

Proof. This is a direct consequence of standard information-theoretic generalization bounds, obtained by combining Pinsker's inequality with variational representations of mutual information. See, e.g., classical treatments of PAC-Bayesian or information-theoretic generalization theory. □

This lemma provides a canonical justification for Theorem 6.3. No special properties of AGI are used.

12.3. Rice-Style Undecidability for Interactive Systems

For completeness, we clarify the applicability of Rice's Theorem to interactive agents.

Lemma 12.3 (Reduction to Partial Computable Functions)**.** *Under the effective presentation assumptions of Section 7.1, the interactive behavior of any admissible agent A induces a partial computable function*

$$f_A : \ (\mathrm{code}_{\mathcal{T}}(\tau), \omega) \ \mapsto \ H_T,$$

where ω denotes exogenous randomness and H_T is the resulting interaction history.

Proof. Given $\mathrm{code}(A)$ and $\mathrm{code}_{\mathcal{T}}(\tau)$, the universal simulator enumerates the joint execution of A and E step by step, consuming randomness from ω. If the interaction halts within the resource bounds, H_T is produced; otherwise the computation diverges. This defines a partial computable function in the standard sense. □

Lemma 12.4 (Applicability of Rice's Theorem)**.** *Any nontrivial semantic predicate on admissible agents, defined via their induced interactive behavior, is undecidable.*

Proof. By Lemma 12.3, admissible agents correspond to partial computable functions up to behavioral equivalence. Semantic predicates are invariant under such equivalence. Rice's Theorem applies directly. □

This justifies the reduction used in Theorem 7.1 without appeal to informal arguments.

12.4. Summary

The technical lemmas recorded here supply standard constructions underlying the main text: small distributional perturbations, bounded transfer via information measures, and undecidability of nontrivial semantic properties. None introduce new assumptions or extend the scope of the paper's conclusions.