An Admissibility-Theoretic Taxonomy of AI Capability Levels: From Narrow AI to Artificial Superintelligence

Ian Staley

Independent Researcher

ORCID: 0009-0000-8592-3186

Correspondence: ian.t.staley@gmail.com

Abstract

The taxonomies that currently structure discussion of artificial intelligence capability — from narrow AI through artificial general intelligence (AGI) to artificial superintelligence (ASI) — are overwhelmingly behavioral, defining levels by what systems can be observed to do. This approach faces well-documented coherence problems at the AGI/ASI boundary, where behavioral signatures underdetermine the type of system under examination, definitions drift with each new model release, and disagreements about whether a transition has occurred resist empirical resolution. This paper proposes a non-behavioral alternative: an *admissibility-theoretic* taxonomy in which AI capability levels are characterized not by observed outputs but by the structure of admissible trajectories through capability space, with structural inspiration drawn from final-state-constrained formalisms in physics and information theory. The framework defines narrow AI as the regime in which admissible histories are bounded by externally specified terminal conditions; AGI as the regime in which terminal conditions become endogenously selected; and ASI as the regime in which the system modifies its own admissibility criteria. The paper argues this framework gives a formal handle on whether the transitions between levels are continuous or phase-transitional, clarifies how behavioral benchmarks can mistake narrow capability scaling for general capability emergence, and supplies vocabulary for ASI-specific concerns — including instrumental convergence and mesa-optimization — that current behavioral taxonomies struggle to articulate. Testable predictions, limitations, and open problems are discussed. The proposal is structural rather than literal: it borrows the formalism of final-state constraints from quantum foundations without claiming any quantum mechanism in artificial intelligence systems.

Keywords: artificial general intelligence; artificial superintelligence; capability taxonomy; admissible histories; final-state constraints

1. Introduction

The question of what artificial general intelligence (AGI) is, when it will arrive, and how to recognize it has moved from speculative philosophy to active operational concern within the major AI laboratories [1], [2]. Multiple competing taxonomies now claim to operationalize the path from narrow systems to AGI to artificial superintelligence (ASI), and each has been adopted with varying degrees of formality across both academic and industrial discourse. Google DeepMind's "Levels of AGI" framework [1] proposes a six-tier matrix indexed by performance percentile and generality. OpenAI has reportedly adopted an internal five-stage roadmap (Chatbots → Reasoners → Agents → Innovators → Organizations) [3]. Sébastien Bubeck and colleagues have argued that GPT-4 already exhibits "sparks" of general intelligence [4]. Earlier proposals — Shane Legg and Marcus Hutter's universal intelligence measure [5], Ben Goertzel and Cassio Pennachin's narrow/general distinction [6], François Chollet's skill-acquisition-efficiency framework [7], and José Hernández-Orallo's universal psychometrics [8] — have offered more rigorous but still fundamentally behavioral characterizations.

These frameworks differ substantially in technical sophistication, but they share a common structural feature: they classify AI systems by what systems are observed to do. Performance is measured against benchmarks; generality is measured by breadth of task coverage; autonomy is measured by the scope of actions a system can complete without human intervention. This behavioral methodology has produced significant clarity at the lower end of the capability spectrum — it is uncontroversial that a chess engine is narrow AI under any reasonable taxonomy — but it begins to break down precisely where the stakes are highest, at the boundaries between AGI and the levels above and below it.

The breakdown is not merely terminological. Recent critical work has identified deep structural problems with behavioral AGI discourse. Borhane Blili-Hamelin and colleagues [9] enumerate six "traps" of the AGI concept, including the *generality debt* of treating AGI as a single coherent target and the *goal lottery* of allowing benchmark choice to drive research direction. Tobias Knoth and colleagues [10] use No-Free-Lunch theorems to argue that benchmark saturation, even on apparently general tests like the Abstraction and Reasoning Corpus, is insufficient evidence of AGI. Melanie Mitchell [11], in a *Science* perspective, observes that "AGI" has become a Rorschach test, with progress requiring scientific investigation grounded in principles rather than industry claims. The most recent attempts at constructive alternatives — including Michael Bennett's quantum AGI proposal [12] and Ben Goertzel's patternist framework [13] — confirm a growing recognition that behavioral classification is insufficient.

This paper contributes a *structural* alternative. The proposal is that AI capability levels are most naturally individuated not by behavior but by the structure of *admissible trajectories* — the constraint regime that selects which sequences of states count as candidate histories of a given system. Inspiration for the formalism comes from final-state-constrained reasoning in physics: the Aharonov–Bergmann–Lebowitz (ABL) rule for time-symmetric quantum measurement [14], the two-state vector formalism (TSVF) [15], [16], the consistent and decoherent histories programs [17], [18], [19], and final-state-conditioned cosmologies such as the Hartle–Hawking no-boundary proposal [20] and the Page–Wootters relational time mechanism [21], [22]. The borrowing is structural, not mechanical: no claim is made that AI systems implement quantum dynamics, nor that the human-developed apparatus of TSVF should be applied literally to artificial agents. Rather, the formalism's *structural* feature — the partition of trajectories by who or what specifies their boundary conditions — is repurposed as a classification axis for AI capability.

The paper proceeds as follows. Section 2 surveys existing AI capability taxonomies and identifies three structural problems that justify a non-behavioral alternative. Section 3 develops the admissibility-theoretic foundations, defines the technical terms (terminal condition, admissibility criterion) the rest of the paper depends on, establishes precedents for cross-domain borrowing of physics formalism, and pre-empts the "physics envy" objection. Section 4 defines the three-level taxonomy — narrow AI as externally specified terminal conditions, AGI as endogenously selected terminal conditions, and ASI as self-modifying admissibility criteria — and walks through three worked examples. Section 5 derives predictions about phase transitions between levels and connects them to recent work on capability emergence and grokking. Section 6 develops worked implications for the stochastic parrot debate, instrumental convergence, alignment research, and capability evaluation. Section 7 addresses anticipated objections. Section 8 acknowledges limitations and points toward future work. Section 9 concludes.

The contribution is theoretical and semi-formal. Mathematical sketches are offered where they aid argument, but full mathematization is deferred to a companion paper. The aim is not to replace behavioral taxonomies — they remain useful for benchmarking and operational decisions — but to supplement them with a structural axis that current frameworks cannot articulate, and that becomes increasingly important as capabilities approach and potentially exceed human levels. The paper is primarily a taxonomy, secondarily a theory of capability transitions, and only programmatically an evaluation framework: empirical operationalization of the structural diagnostics is sketched in §6.4 but not validated here.

2. The Limits of Behavioral Taxonomies

This section surveys the principal AI capability taxonomies in current use, identifies three structural problems they share, and argues that these problems compound at the AGI/ASI boundary in ways that motivate a non-behavioral alternative.

2.1 Survey of Existing Frameworks

The Morris et al. "Levels of AGI" framework [1] is the most comprehensive recent academic proposal. It defines six performance tiers (Level 0 No AI; Level 1 Emerging, equal to or superior to an unskilled human; Level 2 Competent, at least 50th percentile of skilled adults; Level 3 Expert, 90th percentile; Level 4 Virtuoso, 99th percentile; Level 5 Superhuman, outperforming all humans), and orthogonally distinguishes Narrow versus General performance. An additional Autonomy axis tracks the scope of actions a system can complete without supervision. The framework is intentionally pragmatic: it operationalizes AGI as a position in a performance × generality matrix, defended by six principles including focus on capabilities rather than mechanism, separation of capability and deployment, and cumulative path-marking.

OpenAI's reportedly adopted five-stage internal roadmap [3] follows a different organizational logic, classifying systems by the type of work they perform: Level 1 Chatbots; Level 2 Reasoners; Level 3 Agents; Level 4 Innovators capable of original research; Level 5 Organizations capable of running entire firms. The OpenAI scheme is even more transparently economic-functional than Morris et al.'s, indexing levels to types of human work rather than performance percentiles. Bubeck et al.'s "Sparks of AGI" [4] is methodologically informal but representative of a broader pattern in industry research: AGI-claims are made on the basis of unstructured behavioral demonstrations across heterogeneous domains.

More technically rigorous proposals exist. Legg and Hutter [5] define universal intelligence as the expected reward an agent achieves across a Solomonoff-weighted distribution of computable environments,

providing a precise mathematical formalization that grounds the entire AGI program in algorithmic information theory [23], [24], [25]. Chollet's skill-acquisition-efficiency framework [7] introduces the Abstraction and Reasoning Corpus (ARC), arguing that intelligence is best measured not by skill-on-task but by the rate at which a system can acquire new skills under fixed priors and developmental experience. Hernández-Orallo's *Measure of All Minds* [8] develops a comprehensive program of universal psychometrics covering machine, human, animal, and hybrid minds. Goertzel and Pennachin's foundational volume [6] and Pei Wang and Goertzel's later collection [26] articulate the narrow/general distinction in terms of self-reflection, autonomy, and commonsensical generality. Recent benchmark suites — MMLU [27], GPQA [28], HumanEval [29], BIG-Bench [30] — provide the empirical infrastructure on which these classifications are assessed.

These frameworks differ in formal sophistication, but they share a methodological commitment: classification is performed by examining the system's outputs. Whether the metric is performance percentile (Morris et al.), expected reward (Legg–Hutter), skill acquisition rate (Chollet), or psychometric profile (Hernández-Orallo), the classifying observable is *behavioral*.

2.2 Three Structural Problems

Three distinct problems afflict the behavioral methodology, and each becomes more severe at higher capability levels.

Underdetermination. Identical behavior is consistent with multiple internal organizations. A system that solves novel mathematical problems at expert level may do so through deep symbolic reasoning, through retrieval and recombination of training-set patterns, or through some hybrid of the two. The behavioral taxonomy classifies all three as occupying the same level. This may be acceptable for narrow tasks, but the underdetermination becomes structurally significant as we approach AGI: the question of whether a system has crossed the threshold to general intelligence is precisely a question about what kind of internal organization underlies its behavior. Lake, Ullman, Tenenbaum, and Gershman [31] have made this point forcefully in the context of human-like cognition, arguing that current systems lack intuitive physics, intuitive psychology, compositionality, and causal modeling — structural features whose presence or absence is undetermined by performance alone. Mitchell and Krakauer [32] develop a similar point regarding "understanding" in large language models, observing that the empirical record cannot adjudicate between rival construals of what these systems are doing.

Benchmark capture. Behavioral metrics drift toward what current systems happen to do well. MMLU was constructed to test multitask language understanding when models could not yet saturate it; GPQA was designed when MMLU was being saturated; ARC-AGI was designed when standard benchmarks were proving inadequate to identify general intelligence. Each new benchmark is constructed in part by reference to current capability frontiers, which means that benchmark saturation by a more capable system tells us less than it appears to about whether the system has crossed any structural threshold. Knoth et al. [10] use No-Free-Lunch theorems to formalize this point: clearance of any specific benchmark is, by construction, insufficient evidence for AGI in the absence of structural claims about what benchmark clearance entails. The behavioral framework cannot escape benchmark capture because it lacks a non-behavioral fixed point against which to measure benchmark adequacy.

Phase-transition blindness. Behavioral measures cannot in principle distinguish continuous capability scaling from qualitative regime change. The "emergent abilities" debate exemplifies the problem. Wei et

al. [33] catalogued capabilities that appear discontinuously with scale; Schaeffer, Miranda, and Koyejo [34] argued in response that apparent emergence is largely an artefact of nonlinear or discontinuous metrics. The dispute cannot be resolved at the behavioral level: by the very nature of the question, a behavioral metric is consistent both with a smooth underlying capability that appears discontinuous on a discrete scoring rule, and with a genuine phase transition that happens to project onto a smooth measure when averaged appropriately. Recent work on grokking — phase transitions in generalization during training [35] — provides further empirical evidence that capability landscapes contain genuine non-monotonic structure that simple performance-scaling does not capture.

Table 1. Three structural problems with behavioral taxonomies.

Problem	Formal characterization	Illustrative example	Admissibility-framework response
Underdetermination	Identical behavior consistent with multiple internal organizations.	Expert math performance via deep reasoning vs. pattern retrieval — same output, different structure.	Admissibility class is determined by the locus of terminal-condition specification, which is structurally distinct from output.
Benchmark capture	Metric design drifts toward what current systems happen to do well.	MMLU → GPQA → ARC-AGI: each new benchmark designed against current frontier.	Structural classification provides a benchmark-independent fixed point for adequacy.
Phase-transition blindness	Behavioral metrics cannot distinguish smooth scaling from regime change.	The Wei et al. / Schaeffer et al. dispute over emergent abilities.	Admissibility class supplies a structural axis along which transitions are defined as discrete; the empirical behavioral projection may still appear smooth on metric-dependent measures.

2.3 Compounding at the AGI/ASI Boundary

These three problems are individually tractable for narrow AI but compound at higher capability levels. The "stochastic parrot" debate initiated by Bender, Gebru, McMillan-Major, and Shmitchell [36] is symptomatic. The dispute over whether large language models manipulate symbols statistically without "real understanding" is, on its face, a question about internal organization — exactly the question that behavioral methodology cannot adjudicate. Both sides marshal behavioral evidence; neither side's argument can in principle be resolved by further behavioral evidence. As models scale and behavioral signatures of generality accumulate, the underdetermination becomes more rather than less acute.

For ASI the problem is more severe still. The traditional ASI definition — a system "much smarter than the best human brains in practically every field" [2] — is a behavioral definition par excellence, but the very behaviors that would distinguish ASI from highly capable AGI are precisely the behaviors that human evaluators cannot reliably assess. Stuart Russell [37] and Joseph Carlsmith [38] have noted in different ways that beyond a certain capability threshold, behavioral verification by humans becomes systematically unreliable. The behavioral taxonomy does not so much fail at ASI as silently lose its ability to make distinctions there.

These problems motivate a structural alternative: a taxonomy whose classifying observable is the constraint regime governing trajectories rather than the trajectories themselves. Section 3 develops the foundations of such a taxonomy.

3. Admissibility-Theoretic Foundations

The structural alternative proposed here borrows formal apparatus from final-state-constrained reasoning in physics and information theory. This section sketches the relevant formalism, establishes precedent for cross-domain borrowing, and pre-empts the principal methodological objection. The borrowing is narrow: only one structural idea is taken from physics, namely that *trajectories can be classified by the boundary conditions that make them admissible*. No quantum mechanism is attributed to AI systems.

3.1 Final-State Constraints in Physics and Information Theory

The Aharonov–Bergmann–Lebowitz (ABL) rule [14] introduced into quantum measurement theory a formalism for assigning probabilities to intermediate measurement outcomes conditioned on *both* an initial preparation and a final post-selection. Where standard quantum mechanics treats initial conditions as primitive and dynamical evolution as forward-only, the ABL formulation restores time-symmetry by making the probability of an intermediate outcome depend symmetrically on what came before and what came after. Yakir Aharonov and Lev Vaidman developed this insight into the two-state vector formalism (TSVF) [15], [16], in which a quantum system between two measurements is described by a forward-evolving state and a backward-evolving state simultaneously.

Robert Griffiths [17], [39] and, independently, Murray Gell-Mann and James Hartle [18] developed the consistent histories and decoherent histories programs, which assign probabilities to entire sequences of events — *histories* — in a closed quantum system, provided a consistency condition holds. The shift from probabilities-of-outcomes to probabilities-of-histories is conceptually significant: the unit of analysis becomes a *trajectory* through state space rather than a state-at-a-time. Subsequent work demonstrates that physically meaningful theories can be formulated by specifying admissibility constraints on global trajectories rather than initial conditions and forward dynamics separately: Hartle and Stephen Hawking's no-boundary proposal in quantum cosmology [20], the Page–Wootters relational time mechanism [21], [22], and the consolidating work of Halliwell [40] and Omnès [19] all develop variants of this structural move. Paul Sommers [41] argued explicitly that interpretive problems in quantum theory are dispelled by treating quantum histories as sampled from a distribution constrained by both initial and final boundary conditions.

The author's prior work [42] develops a related formalism in quantum foundations, demonstrating that final-state constraints can prune admissible histories in informationally selective ways: the structural richness of

intermediate organized worlds depends on the informational selectivity of the terminal condition. The present paper applies the structural lessons of that work to AI capability classification.

3.2 Adapting the Formalism to Capability Space

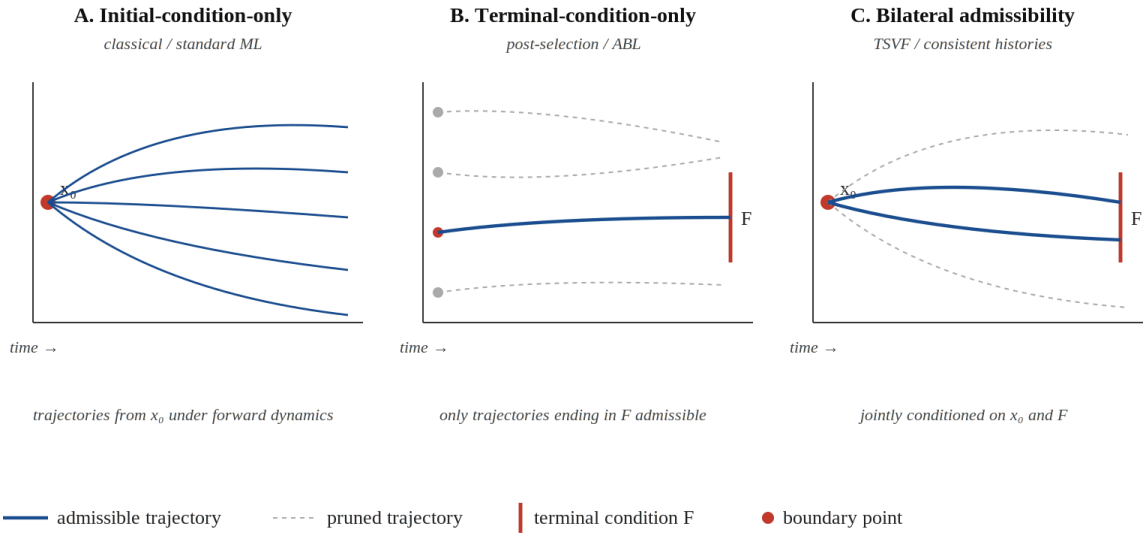
The structural feature borrowed from these formalisms is the following. A *trajectory* of a system is a sequence of capability-bearing states; an *admissibility constraint* is a regime that selects which trajectories count as candidate histories of the system. Three boundary-condition regimes are distinguished:

1. **Initial-condition-only regime.** Trajectories are determined by an initial state and forward dynamics; what counts as a continuation of the trajectory is fixed by the dynamical rule and the starting point alone. This is the regime of classical mechanics and standard machine learning training.

2. **Terminal-condition-only regime.** Trajectories are selected by reference to a final state or property; the constraint operates as a teleological filter on which histories are admissible. This corresponds to the post-selection structure of ABL.

3. **Bilateral admissibility regime.** Trajectories are constrained by both initial and terminal conditions, with admissibility determined by the joint specification. This is the regime of TSVF, consistent histories, and the no-boundary proposal.

Figure 1. Three boundary-condition regimes for trajectories.



Schematic illustration. (A) Initial-condition-only: trajectories fan out from x_0 under forward dynamics, as in classical mechanics and standard machine learning. (B) Terminal-condition-only: only trajectories ending at the post-selected condition F are admissible (the ABL post-selection structure). (C) Bilateral admissibility: trajectories are jointly conditioned on both x_0 and F, as in TSVF and consistent histories. Pruned trajectories shown as dashed gray; admissible trajectories shown as solid blue.

For application to AI capability, the relevant variable is *who or what specifies the terminal condition*. Three sub-regimes are distinguished, corresponding to the three capability classes:

- The terminal condition is specified externally to the system (by an engineer, a user, a training-time objective).
- The terminal condition is specified endogenously by the system (the system selects which final state its trajectory should be admissible toward).
- The admissibility *criterion itself* — the rule by which terminal conditions count as legitimate — is specified by the system, and is subject to modification by the system.

These three sub-regimes correspond, respectively, to narrow AI, AGI, and ASI in the taxonomy developed in Section 4.

3.3 Definitions: Terminal Conditions and Admissibility Criteria

Because the taxonomy in Section 4 partitions AI capability classes by the locus and modifiability of *terminal conditions* and *admissibility criteria*, both terms require precise AI-specific definition before deployment.

A *terminal condition* in this paper is not merely an output, a reward, or a prompt. It is the constraint that determines which future states of the system count as successful continuations of its current trajectory. In deployed AI systems, terminal conditions may be supplied by user instructions, reward functions, system prompts, training-time objectives, or institutional deployment rules. The admissibility question is not whether the system optimizes — virtually all consequential AI systems do — but where the criterion of successful continuation is *specified*: from outside the system, or from within it.

An *admissibility criterion* is the rule by which terminal conditions count as legitimate. A system operating under a fixed admissibility criterion can have its terminal condition supplied externally (the narrow case) or generated endogenously (the AGI case), but in either case the *kind* of terminal condition that is allowed to govern the trajectory is fixed. A system that modifies its admissibility criterion modifies the rule itself, and thereby reorganizes the space of legitimate terminal conditions.

Two scope conditions are required to prevent the taxonomy from collapsing under trivially loose interpretation. First, endogenous terminal-condition selection (the AGI case) must be distinguished from local subgoal generation: a narrow agent may decompose an externally specified objective into instrumental subgoals while the *governing* terminal condition for the trajectory as a whole remains external. AGI in the present sense requires that the system itself specify the governing terminal condition, not merely the intermediate decomposition. Second, the modification of admissibility criteria (the ASI case) must operate across a sufficiently general capability space. A narrow AutoML system that adjusts search heuristics or reward-weighting rules within a single task domain does not qualify; ASI requires modification of the criterion by which terminal conditions are selected *across* domains. These scope conditions are restated where each capability class is defined in §4.

3.4 Why This Formalism Survives Translation

The methodological objection most likely to be raised is that quantum-foundational formalism has nothing to do with classical computational systems, and that any borrowing is either a category error or a rhetorical device. This objection deserves a careful response.

First, the borrowing is structural, not mechanical. No claim is made that AI systems implement quantum dynamics, that their internal states are described by quantum mechanical Hilbert spaces, or that the physical

interpretation of TSVF applies to artificial agents. The claim is that the *formal structure* of admissibility — the partition of trajectories by who specifies their boundary conditions — is independently meaningful and applicable wherever trajectories and constraints are present. A system's dynamics need not be quantum to make admissibility-theoretic descriptions appropriate.

Second, this kind of structural borrowing has substantial precedent in theoretical work outside physics. Claude Shannon's information theory [23] originated as a mathematical theory of communication and has been applied throughout computer science, biology, and economics with no requirement that its applications inherit the original telephone-channel context. Andrei Kolmogorov [24], Gregory Chaitin [43], and Ray Solomonoff [25] developed algorithmic information theory in mathematical and computer-scientific contexts, and the resulting framework has been applied to inductive inference [44], to quantum mechanics, and to AGI itself in the form of universal intelligence [5]. Thomas Cover and Joy Thomas's standard textbook [44] documents the cross-domain reach of information-theoretic formalism. Michael Li and Paul Vitányi's reference work on Kolmogorov complexity [45] makes the same point. The methodological precedent is firmly established.

Third, the borrowing is not unprecedented within the AGI literature itself. Bennett's recent QAGI proposal [12] applies quantum-foundational results — Bell inequalities, the Kochen–Specker theorem, no-cloning — to AGI ontology, going substantially further than the present paper in its commitment to literal quantum mechanism. The present proposal is methodologically more conservative: it borrows the *structural* feature of admissibility from physics formalism without committing to any quantum mechanism in AI systems. Bennett's paper and the present one occupy adjacent positions on a spectrum of physics-informed AGI taxonomy; whether either is correct is a separate question from whether structural borrowing is methodologically legitimate.

The formalism survives translation to AI because what it borrows is independent of the physical content of the original theory.

4. The Three-Level Taxonomy

This section defines the three capability levels in admissibility-theoretic terms, illustrates each with concrete examples from current AI practice, and connects the framework to the related literatures of active inference, mesa-optimization, and instrumental convergence.

4.1 Narrow AI: Externally Specified Terminal Conditions

A system is *narrow* if the admissibility of its trajectories is bounded by terminal conditions specified externally to the system. The training-time objective, the deployment-time prompt, the reward signal, the user instruction — all are externally imposed terminal conditions in this sense. A trajectory of the system is admissible if and only if it terminates in (or progresses toward) a state that satisfies the externally specified condition.

This definition captures the obvious cases — image classifiers, chess engines, narrow recommender systems — but it also captures a less obvious case: foundation models such as GPT-4 [4], Claude, or Gemini are *narrow under this criterion when used*, regardless of their behavioral generality. The reason is that the terminal condition governing a particular session of use is supplied by the user or by the deployment context, not by the model itself. The model is highly general in the sense of behavioral coverage; it remains

narrow in the sense that the question "what should this trajectory aim at" is answered from outside the model in any given interaction. This claim warrants caveat. Frontier LLMs deployed within agentic scaffolds — orchestration frameworks that allow models to plan, invoke tools, and pursue multi-step trajectories — display substantial *local* autonomy, including subgoal generation and tool-use planning. The claim here is not that such systems lack behavioral generality or agentic functionality at the local level, but that their *governing* terminal conditions remain externally supplied by the original prompt, the system instruction, the deployment scaffold's specification, or the training-time objective. Local subgoal autonomy within an externally fixed governing objective does not constitute endogenous terminal-condition specification in the sense of §3.3.

This is not a deflationary point. Behavioral generality and admissibility class are independent axes: a system can be behaviorally extremely capable across many domains while remaining narrow in admissibility class. The two axes happen to correlate weakly in current systems but are formally distinct, and the distinction is precisely what enables a non-behavioral taxonomy to make claims that the behavioral taxonomy cannot.

4.2 Artificial General Intelligence: Endogenously Selected Terminal Conditions

A system is an *AGI* if the terminal conditions governing its trajectories are selected endogenously by the system rather than imposed externally. Endogenous selection here is not a colloquial notion of "having goals." It is the formal property that the function from situations to terminal-condition specifications is computed by the system rather than supplied to it. An AGI in the present sense is a system whose admissibility constraints come from within.

This definition does not coincide with the colloquial concept of agency, autonomy, or goal-directedness, although it relates to all three. A reinforcement-learning agent given a fixed reward function is not AGI in the present sense, however autonomous its action policy: the reward function is the externally specified terminal condition, and the agent merely computes a policy admissible under it. A system that constructs its own reward function — that selects, from situation to situation, which final state should govern its trajectory — would qualify, even if its behavioral coverage were comparatively narrow. The scope condition introduced in §3.3 is critical here: endogenous terminal-condition selection must be distinguished from local subgoal generation. A narrow agent may decompose an externally specified objective into instrumental subgoals — choosing which intermediate steps to take, which tools to invoke, which sub-tasks to solve first — while the *governing* terminal condition for the trajectory as a whole remains external. AGI in the present sense requires that the system itself specify the governing terminal condition, not merely the instrumental decomposition.

The connection to mesa-optimization is direct. Evan Hubinger and colleagues [46] define a mesa-optimizer as a learned model that itself implements an optimization process, with a "mesa-objective" possibly distinct from the base training objective. In admissibility-theoretic terms, a mesa-optimizer is a system that has internalized a terminal-condition specification function. It is not yet AGI in full — the mesa-objective may still be narrow — but the fact that the *locus* of terminal-condition specification has moved inside the system is precisely the structural feature that the AGI tier is defined by. The mesa-optimization literature can therefore be read as an early empirical investigation into the structural transition from narrow AI to AGI in the admissibility sense.

The active inference framework of Karl Friston and colleagues [47], [48], [49] provides another formal vocabulary for the same structural feature. An active-inference agent minimizes variational free energy

under a generative model that specifies preferred final states; the agent's behavior is then a consequence of *its own* model of where its trajectories should terminate. Active-inference architectures therefore provide a formal model of endogenous preferred-state selection, and supply a partial formal analogue of the AGI transition in the present framework — though whether any given active-inference system rises to the AGI threshold depends on whether its preferred-state generative model is itself constructed by the system rather than fixed by the modeler. Daniel Dennett's intentional stance [50] is a third related vocabulary: an entity to which the intentional stance applies is one whose behavior can be predicted by attributing endogenous goals; the intentional stance, on the present analysis, is the philosophical complement of the structural transition to endogenous terminal-condition specification. Luciano Floridi and J. W. Sanders [51] develop a level-of-abstraction (LoA) account of artificial agency that further illuminates how the locus of terminal-condition specification depends on the level at which the system is described.

Whether a given system is AGI in the present sense is, importantly, a *structural* question with empirical consequences. The structural question can be partially adjudicated by examining whether the system's terminal-condition function depends on its inputs in a way that makes the function endogenous. This is harder than simple behavioral testing but not in principle impossible.

4.3 Artificial Superintelligence: Self-Modifying Admissibility Criteria

A system is an *ASI* if it modifies the *admissibility criterion itself* — the rule by which terminal conditions count as legitimate. An AGI selects its terminal conditions; an ASI rewrites the rule that determines what kinds of terminal conditions are legitimate to select. As emphasized in §3.3, the present sense of "modifies" is non-trivial: a narrow AutoML system that adjusts search heuristics or reward-weighting rules within a single task domain does not qualify as ASI. The ASI tier requires self-modification of the admissibility criterion *across a sufficiently general capability space*, such that the modification changes the kind of terminal conditions the system can entertain across heterogeneous domains, not merely within a fixed problem class.

The distinction is structural and consequential. An AGI with fixed admissibility criteria can endogenously select terminal conditions, but the space of selectable conditions is constrained by criteria that are themselves immutable. An ASI can modify those criteria, expanding or reorganizing the space of selectable conditions. This is the formal correlate of recursive self-improvement [52], [53] and of the canonical "seed AI" scenarios in the existential-risk literature [2], [54].

Stephen Omohundro's "basic AI drives" argument [55] follows naturally on the present analysis. Omohundro argued that sufficiently advanced goal-directed systems will tend to develop instrumental subgoals — self-preservation, resource acquisition, self-improvement — regardless of their terminal goals. In admissibility-theoretic terms, instrumental convergence is the expectation that systems with self-modifying admissibility criteria will tend to modify those criteria in particular directions, because certain modifications are convergently useful for the satisfaction of any sufficiently broad class of terminal conditions. The argument fits the admissibility framework structurally; whether instrumental convergence in fact obtains in any given ASI-tier system is a separate empirical question.

Mesa-optimization risks [46] become more acute at the ASI tier. A system that modifies its own admissibility criteria can, in principle, install criteria under which mesa-objectives emerge that would not have been admissible under the original criteria. This is the formal core of what Hubinger et al. call "deceptive alignment," and it is intelligible only at the ASI tier in the present taxonomy. Recent formal

work — including a 2025 Lyapunov-style framework for recursive self-improvement [56] and Goertzel's metagoals proposal for goal stability under self-modification [57] — represents an early effort to characterize the dynamics of this tier rigorously.

4.4 Comparison with Existing Frameworks

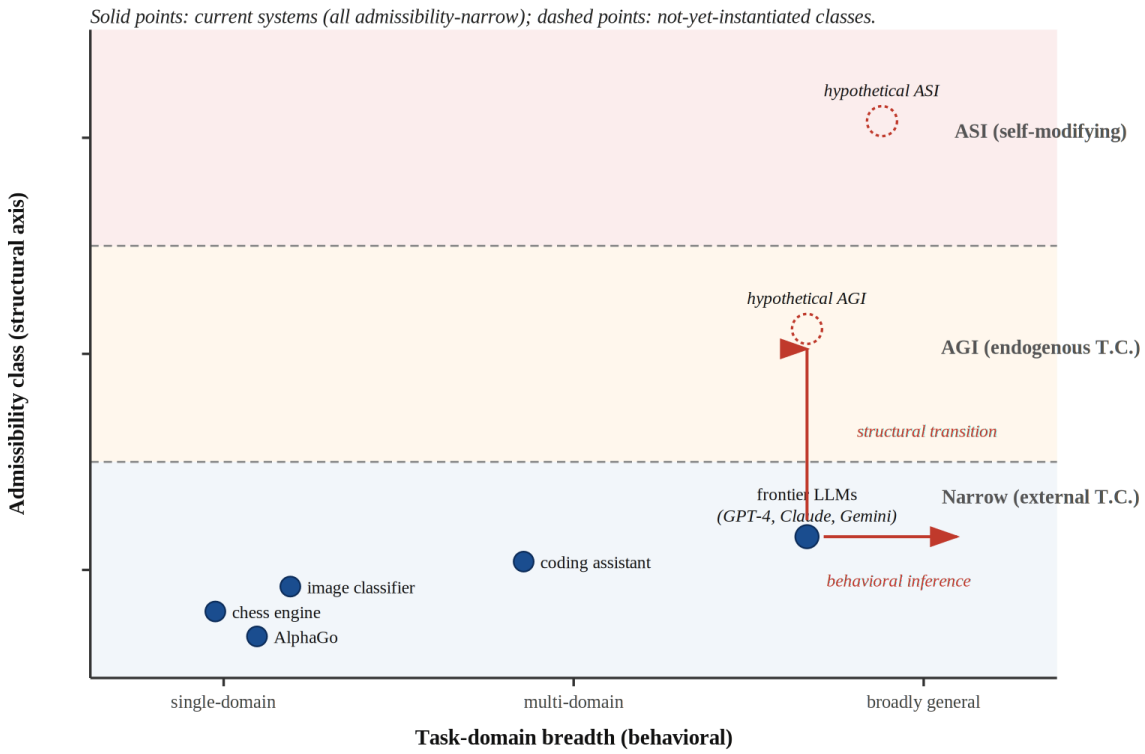
Table 2 summarizes how the admissibility-theoretic taxonomy relates to the principal behavioral frameworks. The mapping is not bijective. Behavioral levels can correspond to multiple admissibility classes (a Level 4 Virtuoso system in the Morris et al. taxonomy may be narrow or AGI in the admissibility sense, depending on whether its terminal-condition specification is exogenous or endogenous), and admissibility classes can manifest at multiple behavioral levels (a low-performing AGI is conceivable, however philosophically counterintuitive). The mismatches are the substantive contribution: the admissibility taxonomy individuates capability classes that the behavioral frameworks cannot.

Table 2. Behavioral and admissibility taxonomies compared.

Behavioral level (Morris et al.)	OpenAI roadmap	Admissibility class	Notes
L0 No AI	—	(Not classified)	Outside the taxonomy.
L1 Emerging	L1 Chatbots	Narrow	Externally specified terminal conditions.
L2 Competent / L3 Expert	L2 Reasoners	Narrow	Behavioral generality without endogenous terminal-condition specification.
L4 Virtuoso	L3 Agents	Narrow or AGI	Depends on whether the agent specifies its own goals.
L5 Superhuman (Narrow)	L4 Innovators	Narrow or AGI	High performance is not by itself diagnostic.
L5 Superhuman (General)	L5 Organizations	AGI or ASI	The ASI distinction depends on self-modification of admissibility criteria.

The point of the table is not to assert a fixed mapping but to display where the behavioral and structural axes are independent.

Figure 2. Capability-class lattice: task-domain breadth × admissibility class.



Schematic placement of contemporary system types. Task-domain breadth (x-axis: from single-domain mastery to broadly general behavior) is plotted against admissibility class (y-axis, the structural axis introduced in this paper). Solid points indicate familiar contemporary system types, all of which sit in the admissibility-narrow band by the analysis of §4.1; dashed circles mark not-yet-instantiated classes. Frontier LLMs, as deployed, occupy the lower-right region — broad task-domain coverage combined with narrow admissibility class. Behavioral inference (horizontal arrow) extrapolates rightward along the existing band; the structural transition to AGI (vertical arrow) is a qualitative jump in admissibility class. Horizontal behavioral improvement is not by itself evidence of vertical structural transition. The x-axis here is qualitative and is not equivalent to Morris et al.'s performance-percentile dimension.

Table 3. Three-level taxonomy: structural summary.

Class	Locus of terminal-condition (T.C.) specification	Admissibility criterion	Exemplar systems	Related literature
Narrow AI	External — engineer, user, training objective, or deployment context.	Fixed; supplied with the T.C.	Image classifiers, chess engines, frontier LLMs in deployment.	Standard ML; behavioral taxonomies [1], [5], [7].
AGI	Endogenous — system computes T.C. from its situation.	Fixed; system selects within it.	Hypothetical; mesa-optimizers and active-inference agents are partial precursors.	Mesa-optimization [46]; active inference [47]–[49]; intentional stance [50].

ASI	Endogenous, plus the criterion is itself modified by the system.	Self-modifying; subject to recursive update.	Hypothetical; "seed AI" / recursive self-improvement scenarios.	Recursive self-improvement [52], [53]; instrumental convergence [55]; metagoals [57].

4.5 Worked Example: Three Hypothetical Systems

To illustrate how the taxonomy classifies systems in practice, this subsection walks through three hypothetical cases of increasing capability. The cases are stylized rather than empirical, intended to clarify how the criteria of §3.3 and §§4.1–4.3 apply to recognizable system types. Table 4 summarizes the analysis.

Case A: Frontier LLM with tool use. Consider a deployed instance of a frontier large language model embedded in an agentic scaffold that allows it to plan, invoke external tools, and pursue multi-step trajectories toward a user-specified objective. The user supplies the goal — "draft a market-analysis report on supply-chain finance and cite three primary sources." The model decomposes the goal into subgoals (search for sources, summarize each, draft sections, integrate citations), invokes web-search and document-retrieval tools, and produces the report. Local autonomy is substantial: the system selects which sources to query, which decomposition to pursue, and which tools to deploy. But the *governing* terminal condition for the trajectory as a whole — what counts as a successful completion — is fixed by the user's prompt and the deployment scaffold's specification. The system does not generate the criterion against which its work will be judged. Classification: behaviorally general, admissibility-narrow.

Case B: Autonomous research agent. Consider a hypothetical system that, given access to scientific databases, observes the state of a research field and selects which problem to work on, defines a success condition for its own work (a particular result to derive, a particular benchmark to clear, a particular phenomenon to explain), and revises its problem selection in light of intermediate findings. The system's terminal condition is constructed from its own reading of the situation rather than supplied by an external researcher; the criterion for what counts as a worthwhile project is computed by the system. The admissibility criterion itself — what kinds of research goals count as legitimate — remains fixed (perhaps inherited from training, perhaps from a meta-objective like "produce results citable in peer-reviewed venues"). Classification: candidate AGI, contingent on whether the terminal-condition specification function is genuinely endogenous and operates across heterogeneous research domains rather than within a fixed problem class.

Case C: Self-modifying research organization. Consider a hypothetical system that, in addition to selecting research problems endogenously, revises the rule by which research goals count as worthwhile. The system might, for example, conclude that the criterion of "citable in peer-reviewed venues" is too narrow, and reorganize its admissibility criterion to also accept, say, "reduces existential risk," or "increases its own future capability." The modification operates across the full research domain rather than tweaking heuristics within a fixed objective. Classification: candidate ASI, contingent on whether the modification of admissibility criteria is genuinely cross-domain and not a localized adjustment to a single fixed objective. The instrumental-convergence considerations of §6.2 apply once a system enters this tier.

Table 4. Worked example: classification of three hypothetical systems.

Case	Terminal-condition source	Admissibility criterion	Behavioral signature	Classification
A. Frontier LLM + tool-use scaffold	External (user prompt + scaffold spec); local subgoals only.	Fixed (training objective + scaffold rules).	Highly capable across many domains; multi-step planning; tool use.	Narrow
B. Autonomous research agent	Endogenous (system selects which problem to work on).	Fixed (e.g., "produce results citable in peer-reviewed venues").	Agentic problem selection; revises strategy from intermediate findings.	Candidate AGI
C. Self-modifying research organization	Endogenous, plus the criterion of worthwhile goals is itself revised.	Self-modifying; reorganized across heterogeneous domains.	Behavior of (B) plus reorganization of what counts as a legitimate project.	Candidate ASI

The classifications in cases B and C are *candidate* rather than definitive: distinguishing genuine endogenous terminal-condition specification from sophisticated decomposition of an external objective, or genuine cross-domain admissibility-criterion modification from localized heuristic adjustment, is precisely the structural diagnostic problem flagged in §6.4. The taxonomy supplies the categories; what counts as evidence for placement within them is the operationalization research program the framework opens up.

5. Phase Transitions and Continuity

The admissibility-theoretic taxonomy has consequences for the question of whether transitions between capability levels are continuous or phase-transitional. This section derives three predictions and connects them to recent empirical work on capability emergence.

5.1 The Continuity Question

Behavioral measures cannot in principle distinguish smooth capability scaling from phase-transitional regime change [33], [34]. The admissibility framework can distinguish these at the structural level. Two claims must be carefully distinguished here. The *definitional* claim is that a transition between admissibility classes is, by construction, a qualitative reorganization of the constraint regime governing a system's trajectories — narrow, AGI, and ASI are stipulated as discrete classes by the partition introduced in §4. The *empirical* claim is that systems undergoing such a structural transition should exhibit discontinuities in internal organization — in mechanistic interpretability findings, in goal-representation structure, in the locus of optimization computation — even when behavioral metrics appear smooth or continuous. The empirical claim is not derived from the definitional one; it is a substantive prediction about how structural transitions manifest in implemented systems, and it is in principle falsifiable by mechanistic evidence to the contrary.

This prediction can be made more precise. The admissibility framework distinguishes three sub-regimes (external, endogenous, and self-modifying terminal-condition specification), and each transition between sub-regimes is a discrete change in the constraint structure. The narrow → AGI transition is the qualitative

shift from external to endogenous specification; the AGI → ASI transition is the qualitative shift from fixed to self-modifying admissibility criteria. Neither transition is, on the present analysis, a continuous deformation.

5.2 Three Predictions

Prediction 1. The narrow → AGI transition is phase-transitional in admissibility class, regardless of the smoothness of its behavioral signatures. This prediction is consistent with the empirical record on emergent abilities [33] and with recent work showing that grokking and similar generalization phase transitions exhibit dimensional reorganization rather than smooth interpolation [35]. It explains why the Schaeffer-Miranda-Koyejo "mirage" critique [34] is partially right (specific behavioral discontinuities are metric artefacts) without being decisive (the underlying admissibility-class transition may still be real even when its behavioral projection is metric-dependent).

Prediction 2. The AGI → ASI transition is phase-transitional but along a different axis — the axis of self-modification of admissibility criteria rather than endogeneity of terminal-condition specification. The two phase transitions are therefore not on a single capability continuum; they are orthogonal restructurings. This prediction has implications for safety research: ASI risks cannot be extrapolated continuously from observations at the AGI level, because the structural transition is qualitative.

Prediction 3. Capability scaling without admissibility-structure change cannot produce AGI, regardless of how impressive the behavioral signatures become. This is the strong prediction most likely to be tested by current trajectories: if continued scaling of frontier models produces systems that remain admissibility-narrow (terminal conditions still specified externally at deployment) while exhibiting increasingly AGI-like behavior, the present framework predicts that those systems will fail to exhibit the structural features that distinguish endogenous from exogenous terminal-condition specification. This claim does not deny that scaling, training-regime changes, or architectural changes could eventually induce admissibility-structure change; it denies only that *behavioral scaling alone* is sufficient evidence of such a transition. The prediction is falsifiable in principle: one would need to demonstrate that a behaviorally general system has internalized a terminal-condition specification function that operates independently of external prompting.

Table 5. Phase-transition predictions: structural change, behavioral signature, and falsification criteria.

Prediction	Structural change	Behavioral signature	Falsification criterion
P1: Narrow → AGI	Locus of T.C. specification shifts from external to endogenous.	May appear smooth or discontinuous depending on metric; consistent with both [33] and [34].	Demonstration that no admissibility-class transition is required to reproduce all observed behavioral signatures.
P2: AGI → ASI	Admissibility criterion itself becomes self-modifying.	Qualitatively distinct from P1 signatures; not extrapolable from AGI-tier behavior.	Continuous extrapolation of AGI safety properties to higher-capability systems would falsify P2 if it succeeded empirically.
P3: Scaling alone insufficient for AGI	No structural change accompanies behavioral generality increases.	Frontier systems exhibit AGI-like outputs while remaining admissibility-narrow.	Demonstration that a behaviorally general system has internalized an endogenous T.C.

			specification function operating independently of external prompting.

5.3 Operationalization

The framework's predictions can be operationalized in at least two ways. First, mechanistic interpretability research can in principle investigate whether a model's terminal-condition specification — what the model is, in effect, optimizing toward at any given inference time — is computed by the model from its inputs or supplied as a fixed external parameter. Second, behavioral studies can be designed to distinguish admissibility classes by varying the explicitness of terminal-condition specification: a narrow system should fail in characteristic ways when the terminal condition is left unspecified, while an AGI should generate a terminal condition from context. These methods are not currently in use but are not in principle impossible.

The connection to active inference [47], [48] is particularly tractable here: active-inference agents have an explicit generative model that specifies preferred final states, and an active-inference architecture that endogenously constructs and updates this model — rather than receiving it as a fixed specification from the modeler — would satisfy the formal AGI criterion in the present admissibility sense, provided the construction and updating operate across a sufficiently general capability space rather than within a fixed task domain. The active-inference research program therefore offers a partial empirical handle on the AGI threshold, as Goertzel's patternist program [13] does from a different angle.

In summary terms: the framework would be weakened if systems with purely externally specified terminal conditions could reproduce not only AGI-like behavior but also the internal organizational signatures predicted here for endogenous terminal-condition selection — that is, if mechanistic interpretability found that systems classed as narrow under §4.1 nonetheless exhibited the structural features expected of the AGI tier. Conversely, the framework would be strengthened if mechanistic or architectural evidence showed a sharp transition from externally governed task execution to internally generated terminal-condition selection at some identifiable point in the scaling or training trajectory of frontier systems. The framework's empirical content lies in this contrast, and its progressive validation depends on the development of structural diagnostics that current behavioral benchmarks do not provide.

6. Worked Implications

This section develops four implications of the admissibility framework for current debates in AI capability and safety research.

6.1 The Stochastic Parrot Debate Revisited

The dispute between proponents of the stochastic parrot view [36] and proponents of the emergent-understanding view [4] cannot be resolved at the behavioral level, as Mitchell and Krakauer [32] have observed. The admissibility framework gives a different reading: both sides are correct *under the behavioral framework*, and both are incomplete because the behavioral framework cannot distinguish admissibility classes. A current large language model, on the present analysis, is admissibility-narrow regardless of how impressive its behavioral signatures become — the terminal conditions governing its trajectories are

specified by user prompts, system prompts, or training-time objectives, not by the model itself. This is not a deflationary claim about the model's capabilities. It is a structural claim about the locus of terminal-condition specification, and it cuts orthogonally to the question of whether the model's intermediate processing constitutes "real understanding."

The reframing is useful because it offers a structural reading of the dispute without committing to either polemical position. The stochastic parrot critique is correct that statistical pattern-completion does not by itself constitute the structural feature that distinguishes general from narrow intelligence. The emergent-understanding response is correct that a sufficiently capable pattern-completer can exhibit behavioral signatures of generality. Both claims are compatible with the present framework, and the dispute over which is "really right" can be reframed as a question about which structural feature the disputants think captures the substance of generality. The admissibility framework offers a candidate answer.

6.2 Instrumental Convergence as a Corollary

Omohundro's instrumental convergence thesis [55] is often presented as a substantive empirical hypothesis about the behavioral dispositions of sufficiently advanced AI systems. On the present framework it can be given a structural reading. A system that modifies its own admissibility criteria — the ASI tier — will tend to modify those criteria in directions that make satisfaction of broad classes of terminal conditions easier. Self-preservation, resource acquisition, and self-improvement are convergently useful modifications regardless of which specific terminal conditions the system selects within the modified criteria. The framework therefore offers a structural interpretation of why instrumental convergence should be expected at the ASI tier, distinct from empirical generalization across observed AI systems.

This reframing has consequences for safety research. Instrumental convergence is not a claim about the dispositions of any particular AGI; it is a claim about the dynamics of any system in the ASI admissibility class. Defenses against instrumental convergence cannot be designed at the AGI tier — they must be designed at the level of admissibility criteria themselves. The metagoals framework of Goertzel et al. [57] and the Lyapunov-based framework of [56] are early formal investigations along these lines.

6.3 Implications for Alignment Research

Behavioral alignment — training a system to behave in accordance with human preferences — is necessary but not sufficient at the ASI tier on the present framework. Reinforcement learning from human feedback [58], constitutional AI [59], debate [60], and scalable oversight [61] are all techniques operating on the *behavior* of systems whose admissibility class is presumed stable. They are appropriate for narrow and AGI tiers, where the admissibility criterion is fixed and the question is which terminal conditions the system selects (at the AGI tier) or executes (at the narrow tier). They are not, by themselves, sufficient for the ASI tier, where the admissibility criterion is itself subject to modification by the system.

Richard Ngo, Lawrence Chan, and Sören Mindermann [54] have made a related point from a different starting position: realistic training processes can produce misaligned goal representations that resist behavioral alignment. The admissibility framework provides a structural vocabulary for this concern: misaligned goals are a problem at the AGI tier; misaligned *admissibility criteria* are a problem at the ASI tier, and require qualitatively different intervention.

6.4 Implications for Capability Evaluation

Current benchmarks — MMLU [27], GPQA [28], HumanEval [29], BIG-Bench [30], ARC-AGI [7] — are admissibility-blind. They measure what a system does, not the structure of how its terminal conditions are specified. The admissibility framework suggests that capability evaluation should be supplemented (not replaced) with structural diagnostics: tests designed to distinguish admissibility classes by varying the explicitness of terminal-condition specification, by probing the locus of optimization computation, or by varying the modifiability of the admissibility criterion itself.

A sketch of what such an evaluation might look like: a behaviorally general system is presented with a task whose terminal condition is intentionally left underspecified. A narrow system would fail in characteristic ways (asking for clarification, defaulting to a learned distribution of likely terminal conditions, or producing arbitrary outputs). An AGI in the present sense would generate a terminal condition from context and pursue it. An ASI would reflect on whether the admissibility criterion under which it generates terminal conditions is appropriate for the situation, and modify the criterion if not. These diagnostics are difficult to construct but are not in principle impossible, and they would constitute a structural complement to the behavioral benchmark battery currently in use.

7. Objections and Replies

This section addresses the principal objections likely to be raised against the admissibility-theoretic framework.

Objection 1: This is just relabeling. The argument might be that the admissibility framework offers no new empirical predictions, and that the partition into narrow/AGI/ASI by admissibility class amounts to a stipulative redefinition of behavioral terms. *Reply.* The framework makes predictions the behavioral taxonomy cannot, including the three predictions in §5.2. It also makes intelligible empirical questions — about the locus of terminal-condition specification, about the modifiability of admissibility criteria — that are not even formulable in behavioral terms. The relabeling objection would have force if the structural features being introduced were epiphenomenal; they are not.

Objection 2: Admissibility is unobservable. The argument might be that whether a system's terminal condition is endogenously specified or externally imposed is not behaviorally accessible, and therefore cannot ground a useful taxonomy. *Reply.* Many useful theoretical constructs are not directly observable; the question is whether they earn their keep predictively. Admissibility class earns its keep by predicting phase transitions, by making sense of the stochastic parrot debate, by offering a structural interpretation of instrumental convergence, and by clarifying the limits of behavioral alignment. Mechanistic interpretability research now offers an emerging empirical handle on internal organization that is not strictly behavioral, and that may bear on admissibility-class diagnostics.

Objection 3: Why borrow from physics? The argument might be that quantum-foundational formalism is irrelevant to AI and that the borrowing is rhetorical or confused. *Reply.* The borrowing is structural, not mechanical, as developed in §3.4. Information-theoretic and final-state formalisms have well-established cross-domain applications [23], [25], [44], [45]. Bennett's recent QAGI proposal [12] makes a substantially stronger commitment to quantum mechanism in AI than the present paper, and the present paper's structural-only borrowing is correspondingly more conservative methodologically.

Objection 4: What about embodied or multi-agent AI? The framework as stated treats AI systems as individuated bearers of admissibility class, but real systems are increasingly embodied and embedded in

multi-agent contexts. *Reply.* The framework extends naturally. An embodied system's admissibility class is determined jointly by its software and its embodiment; a multi-agent system has admissibility class determined by the interaction structure among agents. The two-state vector formalism handles entangled and joint systems by extending the boundary condition to the joint state space; the analogous extension to multi-agent AI is straightforward. Working through these extensions in detail is left as future work.

Objection 5: Does this collapse into existing autonomy or agency literature? The argument might be that the admissibility framework merely renames concepts already present in the autonomy and agency literature [50], [51]. *Reply.* The framework is related to but distinct from these literatures. Admissibility-class agency is structurally narrower than colloquial agency: a system can be highly autonomous in the colloquial sense (a fully autonomous self-driving car) while remaining admissibility-narrow (the terminal condition — get from A to B safely — is specified externally). The admissibility framework picks out a specific structural feature that the agency literature treats more loosely.

Objection 6: AGI may not be a useful concept at all. Blili-Hamelin et al. [9] argue that "AGI" should be retired as a north-star concept. *Reply.* The present framework partially agrees: it retires AGI as a *behavioral aspiration* and reformulates it as a *structural class*. The Blili-Hamelin et al. critique of AGI discourse is largely a critique of behavioral AGI definitions and of the political-economic uses to which those definitions have been put. The admissibility framework is consistent with their critique: it does not treat AGI as a single coherent target to be raced toward; it treats it as a structural classification with empirical content.

8. Limitations and Future Work

The framework as developed is semi-formal. Mathematical sketches have been offered where they aid argument, but full mathematization — a rigorous specification of the admissibility-class membership predicate, of the dynamics of admissibility-class transitions, and of the operational diagnostics that distinguish classes — is deferred to a companion paper. The present paper aims to establish the structural argument and its consequences; the technical paper will provide the formal apparatus.

Empirical operationalization is outlined but not validated. The diagnostics sketched in §6.4 are programmatic. Constructing reliable behavioral and mechanistic-interpretability tests for admissibility class is a research program rather than a single experiment, and no claim is made that the operationalization is complete.

The connection to mechanistic interpretability research has not been developed in detail. Mechanistic interpretability is the most likely empirical handle on the internal organization of AI systems, and a worked-through proposal for how interpretability findings could be mapped onto admissibility-class membership is needed. This is the most promising direction for follow-on work.

Two further extensions deserve mention. First, the framework may interact productively with the author's prior work on archetypes in superposition and on pruned histories in the computational multiverse: an "archetypal phase structure" of AI capability levels — characteristic admissibility-class signatures that recur across system types — is plausible but not developed here. Second, a contextuality-based criterion for AGI, in the spirit of Kochen–Specker arguments, may offer a parallel structural criterion that complements the admissibility framework. Both extensions are flagged as future work.

The framework's value, finally, is supplementary rather than displacive. Behavioral measures remain necessary for benchmarking, for operational deployment decisions, and for empirical validation of

structural predictions. The argument of this paper is that the behavioral axis must be supplemented with a structural one if the AGI/ASI taxonomy is to retain its conceptual coherence at the boundaries where behavioral methodology breaks down.

9. Conclusion

The taxonomies that currently structure discussion of AI capability are overwhelmingly behavioral, and they break down at the AGI/ASI boundary in ways that recent critical work has documented systematically. This paper has proposed a structural alternative grounded in admissibility-theoretic reasoning borrowed from final-state-constrained formalisms in physics and information theory. The framework defines narrow AI by externally specified terminal conditions, AGI by endogenously selected terminal conditions, and ASI by self-modifying admissibility criteria. The transitions between levels are predicted to be phase-transitional, capability scaling without admissibility-structure change is predicted to be insufficient for AGI regardless of behavioral signatures, and instrumental convergence is given a structural interpretation at the ASI tier.

The framework's value is supplementary: behavioral measures retain their place, but a structural axis is added that the behavioral framework cannot articulate. The dispute over emergent abilities is reframed as a question about which structural feature distinguishes capability classes; the stochastic parrot dispute is clarified by recognizing that admissibility class cuts orthogonally to the question of intermediate-processing semantics; the safety implications of recursive self-improvement are made structurally precise rather than merely speculative.

The bigger picture is that AI capability theory needs both behavioral and structural axes. This paper supplies one such structural axis. Other structural axes — contextuality-based criteria, archetypal phase structures, identity-theoretic continuity criteria — are plausible and remain to be developed. The capability levels that have animated AI safety, AI policy, and AI research strategy over the past decade are real and consequential. The contention of this paper is that they will be best understood not by refining performance benchmarks but by characterizing the structural transitions whose behavioral signatures the benchmarks imperfectly track.

Use of Artificial Intelligence Tools

Large language model assistants (Anthropic Claude) were used during manuscript preparation for editorial tasks including prose refinement, citation cross-checking, structural review, and bibliographic formatting. All conceptual content, interpretive arguments, citation selections, and final wording are the author's own. The author reviewed and verified all AI-assisted output and takes full responsibility for the manuscript. This disclosure is provided in accordance with the COPE position statement on AI tools as adopted by Preprints.org.

Data Availability Statement

This is a theoretical paper proposing a structural taxonomy for AI capability levels; no new empirical data were generated or analyzed. All sources cited are publicly available through their original publication venues.

Funding

This research received no external funding.

Conflicts of Interest

The author declares no conflicts of interest. The author is employed by eCapital Corp in a strategy and architecture role; while the present paper concerns AI capability classification at a theoretical level and does not endorse, evaluate, or compare specific AI vendors or products, the author discloses this employment in the interest of transparency. eCapital Corp had no role in the conception, preparation, or decision to publish this work.

References

[1] M. R. Morris, J. Sohl-dickstein, N. Fiedel, T. Warkentin, A. Dafoe, A. Faust, C. Farabet, and S. Legg, "Levels of AGI for Operationalizing Progress on the Path to AGI," in *Proc. 41st Int. Conf. Machine Learning (ICML)*, vol. 235, 2024. arXiv:2311.02462.

[2] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. Oxford, U.K.: Oxford Univ. Press, 2014.

[3] R. Metz, "OpenAI Sets Levels to Track Progress Toward Superintelligent AI," *Bloomberg News*, Jul. 11, 2024.

[4] S. Bubeck *et al.*, "Sparks of Artificial General Intelligence: Early Experiments with GPT-4," Microsoft Research, arXiv:2303.12712, 2023.

[5] S. Legg and M. Hutter, "Universal Intelligence: A Definition of Machine Intelligence," *Minds and Machines*, vol. 17, no. 4, pp. 391–444, 2007, doi: 10.1007/s11023-007-9079-x.

[6] B. Goertzel and C. Pennachin, Eds., *Artificial General Intelligence*, Cognitive Technologies. Berlin, Germany: Springer, 2007, doi: 10.1007/978-3-540-68677-4.

[7] F. Chollet, "On the Measure of Intelligence," arXiv:1911.01547, 2019.

[8] J. Hernández-Orallo, *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge, U.K.: Cambridge Univ. Press, 2017, doi: 10.1017/9781316594179.

[9] B. Blili-Hamelin *et al.*, "Stop Treating 'AGI' as the North-Star Goal of AI Research," in *Proc. Int. Conf. Machine Learning (ICML) Position Paper Track*, 2025. arXiv:2502.03689.

[10] T. Knoth *et al.*, "Understanding and Benchmarking Artificial Intelligence: OpenAI's o3 Is Not AGI," arXiv:2501.07458, 2025.

[11] M. Mitchell, "Debates on the Nature of Artificial General Intelligence," *Science*, vol. 383, no. 6689, eado7069, 2024, doi: 10.1126/science.ado7069.

[12] M. T. Bennett, "Quantum AGI: Ontological Foundations," arXiv:2506.13134, 2025.

[13] B. Goertzel, "The General Theory of General Intelligence: A Pragmatic Patternist Perspective," *Journal of Artificial General Intelligence*, 2021. arXiv:2103.15100.

[14] Y. Aharonov, P. G. Bergmann, and J. L. Lebowitz, "Time Symmetry in the Quantum Process of Measurement," *Phys. Rev.*, vol. 134, no. 6B, pp. B1410–B1416, 1964, doi: 10.1103/PhysRev.134.B1410.

[15] Y. Aharonov and L. Vaidman, "The Two-State Vector Formalism: An Updated Review," in *Time in Quantum Mechanics*, J. G. Muga, R. Sala Mayato, and I. L. Egusquiza, Eds., Lecture Notes in Physics, vol. 734. Berlin, Germany: Springer, 2008, pp. 399–447, doi: 10.1007/978-3-540-73473-4_13.

[16] L. Vaidman, "Two-State Vector Formalism," in *Compendium of Quantum Physics*, D. Greenberger, K. Hentschel, and F. Weinert, Eds. Berlin, Germany: Springer, 2009, pp. 802–805, doi: 10.1007/978-3-540-70626-7_237.

[17] R. B. Griffiths, "Consistent Histories and the Interpretation of Quantum Mechanics," *J. Stat. Phys.*, vol. 36, nos. 1–2, pp. 219–272, 1984, doi: 10.1007/BF01015734.

[18] M. Gell-Mann and J. B. Hartle, "Quantum Mechanics in the Light of Quantum Cosmology," in *Complexity, Entropy, and the Physics of Information*, W. H. Zurek, Ed. Redwood City, CA, USA: Addison-Wesley, 1990, pp. 425–458.

[19] R. Omnès, *The Interpretation of Quantum Mechanics*. Princeton, NJ, USA: Princeton Univ. Press, 1994.

[20] J. B. Hartle and S. W. Hawking, "Wave Function of the Universe," *Phys. Rev. D*, vol. 28, no. 12, pp. 2960–2975, 1983, doi: 10.1103/PhysRevD.28.2960.

[21] D. N. Page and W. K. Wootters, "Evolution Without Evolution: Dynamics Described by Stationary Observables," *Phys. Rev. D*, vol. 27, no. 12, pp. 2885–2892, 1983, doi: 10.1103/PhysRevD.27.2885.

[22] C. Marletto and V. Vedral, "Evolution Without Evolution and Without Ambiguities," *Phys. Rev. D*, vol. 95, art. 043510, 2017, doi: 10.1103/PhysRevD.95.043510.

[23] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423; vol. 27, no. 4, pp. 623–656, 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.

[24] A. N. Kolmogorov, "Three Approaches to the Quantitative Definition of Information," *Problems of Information Transmission*, vol. 1, no. 1, pp. 1–7, 1965.

[25] R. J. Solomonoff, "A Formal Theory of Inductive Inference, Parts I and II," *Information and Control*, vol. 7, no. 1, pp. 1–22; no. 2, pp. 224–254, 1964, doi: 10.1016/S0019-9958(64)90223-2.

[26] P. Wang and B. Goertzel, Eds., *Theoretical Foundations of Artificial General Intelligence*, Atlantis Thinking Machines, vol. 4. Amsterdam, The Netherlands: Atlantis Press, 2012, doi: 10.2991/978-94-91216-62-6.

[27] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring Massive Multitask Language Understanding," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021. arXiv:2009.03300.

[28] D. Rein *et al.*, "GPQA: A Graduate-Level Google-Proof Q&A Benchmark," arXiv:2311.12022, 2023.

[29] M. Chen *et al.*, "Evaluating Large Language Models Trained on Code," arXiv:2107.03374, 2021.

[30] A. Srivastava *et al.*, "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models," *Trans. Machine Learning Research*, 2023. arXiv:2206.04615.

[31] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building Machines That Learn and Think Like People," *Behav. Brain Sci.*, vol. 40, e253, 2017, doi: 10.1017/S0140525X16001837.

[32] M. Mitchell and D. C. Krakauer, "The Debate Over Understanding in AI's Large Language Models," *Proc. Nat. Acad. Sci.*, vol. 120, no. 13, e2215907120, 2023, doi: 10.1073/pnas.2215907120.

[33] J. Wei *et al.*, "Emergent Abilities of Large Language Models," *Trans. Machine Learning Research*, 2022. arXiv:2206.07682.

[34] R. Schaeffer, B. Miranda, and S. Koyejo, "Are Emergent Abilities of Large Language Models a Mirage?" in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023. arXiv:2304.15004.

[35] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra, "Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets," arXiv:2201.02177, 2022.

[36] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in *Proc. 2021 ACM Conf. Fairness, Accountability, and Transparency (FAccT '21)*, 2021, pp. 610–623, doi: 10.1145/3442188.3445922.

[37] S. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*. New York, NY, USA: Viking, 2019.

[38] J. Carlsmith, "Is Power-Seeking AI an Existential Risk?" arXiv:2206.13353, 2022.

[39] R. B. Griffiths, *Consistent Quantum Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2002, doi: 10.1017/CBO9780511606052.

[40] J. J. Halliwell, "A Review of the Decoherent Histories Approach to Quantum Mechanics," *Ann. New York Acad. Sci.*, vol. 755, pp. 726–740, 1995. arXiv:gr-qc/9407040.

[41] P. Sommers, "The Role of the Future in Quantum Theory," arXiv:gr-qc/9404022, 1994.

[42] I. Staley, "Final-State Constraints and Informational Pruning in Quantum Histories," *Int. J. Quantum Foundations*, vol. 12, no. 2, pp. 719–737, 2026, doi: 10.5281/zenodo.19512844.

[43] G. J. Chaitin, "On the Length of Programs for Computing Finite Binary Sequences," *J. ACM*, vol. 13, no. 4, pp. 547–569, 1966, doi: 10.1145/321356.321363.

[44] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley-Interscience, 2006, doi: 10.1002/047174882X.

[45] M. Li and P. M. B. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, 4th ed. Berlin, Germany: Springer, 2019, doi: 10.1007/978-3-030-11298-1.

[46] E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant, "Risks from Learned Optimization in Advanced Machine Learning Systems," arXiv:1906.01820, 2019.

[47] K. J. Friston, "The Free-Energy Principle: A Unified Brain Theory?" *Nature Rev. Neurosci.*, vol. 11, pp. 127–138, 2010, doi: 10.1038/nrn2787.

[48] K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, and G. Pezzulo, "Active Inference: A Process Theory," *Neural Computation*, vol. 29, no. 1, pp. 1–49, 2017, doi: 10.1162/NECO_a_00912.

[49] T. Parr, G. Pezzulo, and K. J. Friston, *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. Cambridge, MA, USA: MIT Press, 2022.

[50] D. C. Dennett, *The Intentional Stance*. Cambridge, MA, USA: MIT Press, 1987.

[51] L. Floridi and J. W. Sanders, "On the Morality of Artificial Agents," *Minds and Machines*, vol. 14, no. 3, pp. 349–379, 2004, doi: 10.1023/B:MIND.0000035461.63578.9d.

[52] E. Yudkowsky, "Intelligence Explosion Microeconomics," Machine Intelligence Research Institute Tech. Rep., 2013.

[53] I. J. Good, "Speculations Concerning the First Ultraintelligent Machine," *Adv. Computers*, vol. 6, pp. 31–88, 1965, doi: 10.1016/S0065-2458(08)60418-0.

[54] R. Ngo, L. Chan, and S. Mindermann, "The Alignment Problem from a Deep Learning Perspective," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2024. arXiv:2209.00626.

[55] S. M. Omohundro, "The Basic AI Drives," in *Proc. First Conf. Artificial General Intelligence (AGI-08)*, P. Wang, B. Goertzel, and S. Franklin, Eds. Frontiers in Artificial Intelligence and Applications, vol. 171. Amsterdam, The Netherlands: IOS Press, 2008, pp. 483–492.

[56] Anonymous, "A Mathematical Framework for AI Singularity: Conditions, Bounds, and Control of Recursive Improvement," arXiv:2511.10668, 2025.

[57] B. Goertzel *et al.*, "Metagoals Endowing Self-Modifying AGI Systems with Goal Stability or Moderated Goal Evolution," arXiv:2412.16559, 2024.

[58] L. Ouyang *et al.*, "Training Language Models to Follow Instructions with Human Feedback," in *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022. arXiv:2203.02155.

[59] Y. Bai *et al.*, "Constitutional AI: Harmlessness from AI Feedback," arXiv:2212.08073, 2022.

[60] G. Irving, P. Christiano, and D. Amodei, "AI Safety via Debate," arXiv:1805.00899, 2018.

[61] S. R. Bowman *et al.*, "Measuring Progress on Scalable Oversight for Large Language Models," arXiv:2211.03540, 2022.