Retrospective Benchmarks for Machine Intelligence

Evaluating Current AI Against Historical Specifications

Chapter 6: The Synthesis Benchmark (2023)

Dakota Schuck

December 2025

Working paper. Comments welcome.

Preface: Methodology

This chapter *departs* from the methodology established in Chapter 1. For good reason.

The previous five chapters evaluated definitions that proposed thresholds: Gubrud's brain-parity requirement, Legg and Hutter's mathematical formalization, OpenAI's economic value criterion, Chollet's skill-acquisition efficiency measure. Each invited a binary question—does current AI meet the standard or not?—which we answered with a coarse 0%/50%/100% scoring system designed to force honesty about evidential uncertainty.

Morris et al. did something different. They proposed a *taxonomy*: five performance levels (Emerging through Superhuman), a generality axis (Narrow vs. General), and six autonomy levels. The framework explicitly rejects binary AGI thresholds in favor of graduated classification. Forcing this taxonomy into our trichotomous scoring would distort the very thing we are evaluating.

We therefore evaluate the Levels of AGI framework on its own terms. Where previous chapters asked "does current AI meet this criterion?" and answered with percentage scores, this chapter asks "at what level does current AI fall?" and answers with level classifications. The summary table shows positions on ordinal scales, not arithmetic averages.

This is not methodological inconsistency but methodological fidelity: we treat each historical definition according to its own logic. The Levels of AGI framework is designed to classify, not to threshold. We classify.

Every factual claim should be cited. Where citations are missing, we have marked them. Where we have made interpretive choices, we have flagged them. This is a first attempt, meant to be improved by others.[1]

[1] AI Assistance Disclosure: Research, drafting, and analysis were conducted with the assistance of Claude (Anthropic, 2025). The author provided editorial direction and final approval.

1 Introduction: The Co-Founder's Return

In November 2023, Shane Legg published a paper that brought him full circle. Twenty-one years earlier, he had helped coin the term "AGI" with Ben Goertzel and Peter Voss. Sixteen years earlier, he had formalized it mathematically with Marcus Hutter. Now, as Chief AGI Scientist at Google DeepMind, he was part of a team attempting something different: not defining AGI but *taxonomizing* it.[2]

"I see so many discussions where people seem to be using the term to mean different things, and that leads to all sorts of confusion," Legg told *MIT Technology Review*. "Now that AGI is becoming such an important topic—you know, even the UK prime minister is talking about it—we need to sharpen up what we mean."[3]

The paper—"Levels of AGI: Operationalizing Progress on the Path to AGI"—was co-authored with seven DeepMind colleagues, including Meredith Ringel Morris, the company's Director of Human-AI Interaction Research. It drew an explicit analogy to autonomous driving: just as the SAE's Levels of Driving Automation had provided a common language for discussing self-driving cars, the authors proposed "Levels of AGI" to structure conversations about artificial general intelligence.[4]

The analogy was instructive—and perhaps cautionary. The SAE levels had brought clarity to autonomous vehicle discourse, but they had also been criticized for implying a linear progression that obscured the fundamental challenges of achieving higher levels.[5] Would "Levels of AGI" face the same fate?

The DeepMind team's approach was systematic. They analyzed nine prominent definitions of AGI—from Turing's 1950 test to OpenAI's 2018 Charter—and extracted six principles that any useful AGI ontology should satisfy. They proposed a matrix with two dimensions: *performance* (how well a system performs relative to humans) and *generality* (the breadth of tasks a system can handle). They identified five performance levels and two generality categories. And they added a separate taxonomy of *autonomy*—how independently a system operates—arguing that capability and autonomy should be evaluated independently.

The result was the most comprehensive attempt to date to operationalize AGI as a concept. But operationalization is not the same as definition. The framework deliberately avoids saying what AGI *is*; it says only how to classify systems along the path toward it.

Two years later, we can ask: where do current frontier AI systems fall in this taxonomy? And does the placement tell us anything about whether "AGI" has been—or is about to be—achieved?

[2] Morris, Meredith Ringel, et al. "Levels of AGI: Operationalizing Progress on the Path to AGI." arXiv:2311.02462, 2023. Published in *Proceedings of ICML 2024*. https://arxiv.org/abs/2311.02462

[3] Heaven, Will Douglas. "Google DeepMind wants to define what counts as artificial general intelligence." *MIT Technology Review*, November 16, 2023. https://www.technologyreview.com/2023/11/16/1083498/google-deepmind-what-is-artificial-general-intelligence-agi/

[4] SAE International. "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles." J3016, 2021.

[5] Critics of the SAE framework note that the jump from Level 2 to Level 3 autonomous driving has proven far more difficult than the level numbering suggests.

2 The Framework

From "Levels of AGI: Operationalizing Progress on the Path to AGI," published November 2023 and revised through September 2025:[6]

The framework proposes three dimensions for classifying AI systems:

1. **Performance** — How well does the system perform relative to humans?
2. **Generality** — How broad is the range of tasks the system can handle?
3. **Autonomy** — How independently does the system operate?

2.1 Performance Levels

Five levels of performance, defined by percentile comparison to skilled human adults:

- **Level 1: Emerging** — Equal to or somewhat better than an unskilled human
- **Level 2: Competent** — At least 50th percentile of skilled adults
- **Level 3: Expert** — At least 90th percentile of skilled adults
- **Level 4: Exceptional** — At least 99th percentile of skilled adults
- **Level 5: Superhuman** — Outperforms 100% of humans

2.2 Generality Categories

Two categories of generality:

- **Narrow AI** — Performs at specified level on a limited range of tasks
- **General AI** — Performs at specified level across a broad range of cognitive tasks

2.3 Autonomy Levels

Six levels of autonomy, describing human-AI interaction paradigms:[7]

- **Level 0: No AI** — Human performs all tasks
- **Level 1: AI as a Tool** — Human controls task and uses AI assistance
- **Level 2: AI as a Consultant** — AI suggests, human decides and executes
- **Level 3: AI as a Collaborator** — Co-equal human-AI collaboration
- **Level 4: AI as an Expert** — AI drives interaction; human provides guidance
- **Level 5: Autonomous Agent** — Fully autonomous AI operation

2.4 The Six Principles

The framework is grounded in six principles that the authors argue any useful AGI definition should satisfy:[8]

1. **Focus on Capabilities, not Processes** — AGI should be defined by what systems can do, not how they do it. Consciousness, sentience, and human-like thinking are not required.
2. **Focus on Generality and Performance** — Both breadth (how many tasks) and depth (how well) matter. A system can be narrow-superhuman or general-emerging.

[6] Morris et al. 2023/2025, op. cit. The paper has undergone five revisions. Version 5 (September 2025) changed Level 4 nomenclature from "Virtuoso" to "Exceptional."

[7] Ibid., Table 2.

[8] Ibid., Section 2.

3. **Focus on Cognitive and Metacognitive Tasks** — Physical embodiment is not required. The benchmark should include learning ability (metacognition), not just task execution.

4. **Focus on Potential, not Deployment** — A system's capability level should be assessed independently of whether it is actually deployed. Legal, social, and ethical barriers to deployment should not affect capability classification.

5. **Focus on Ecological Validity** — Benchmarks should test real-world tasks, not just academic exercises.

6. **Focus on the Path to AGI** — AGI is not a single endpoint but a spectrum. The framework should enable tracking progress, not just declaring arrival.

2.5 Context

The DeepMind framework emerged from a specific institutional context. Google DeepMind had been working on "frontier AI" for over a decade, having achieved narrow-superhuman performance on games (Go, Chess, StarCraft) and scientific problems (protein folding). The 2022–2023 explosion of large language models—including Google's own Bard/Gemini—raised questions about whether general AI was approaching.

The framework was also a response to definitional chaos. As the authors noted, "if you were to ask 100 AI experts to define what they mean by 'AGI,' you would likely get 100 related but different definitions."[9] By proposing a taxonomy rather than a definition, they hoped to accommodate this diversity while enabling clearer communication.

Finally, the framework served institutional purposes. DeepMind's stated mission involves AGI; a taxonomy that places current systems at "Emerging AGI" is more conducive to continued funding and research than one that places AGI as a distant, binary goal. We note this context without imputing bad faith—the framework's merits should be evaluated on their own terms.

[9] Ibid., p. 1.

3 Operationalization

The Morris et al. framework is itself an operationalization; our task is to evaluate current systems against it. We extract four criteria from the framework:

1. **Performance Level** — What performance level do frontier systems achieve across general cognitive tasks?
2. **Generality** — Do they achieve this performance narrowly or generally?
3. **Autonomy Level** — What autonomy levels can they operate at?
4. **Metacognition** — Do they exhibit the learning and self-assessment capabilities the framework emphasizes?

For each criterion, we assess where current frontier systems fall within the framework's categories. Unlike previous chapters, we cannot score 0%/50%/100% for meeting a single definition; instead, we identify which level is achieved and whether that achievement is contested.

The framework's own assessment (as of the original 2023 publication) placed frontier LLMs at "Emerging AGI"—Level 1 on the general side of the matrix. Two years later, we reassess.

4 Criterion 1: Performance Level

4.1 What Morris et al. Meant

Performance is defined by percentile comparison to skilled human adults. The key thresholds are:

- Emerging (Level 1): Equal to or somewhat better than an unskilled human
- Competent (Level 2): At least 50th percentile of skilled adults
- Expert (Level 3): At least 90th percentile of skilled adults

The comparison class matters. "Skilled adults" means people who possess the relevant skill—not the general population. Performance on an English writing task should be compared to literate English speakers, not all humans.[10]

4.2 Performance on Professional Benchmarks

Measure: How do frontier models perform relative to skilled human benchmarks?

Reference values:

- GPQA-Diamond (graduate-level science): Human PhD experts ∼65%; frontier models 87–93%[11]
- MMLU (57 subjects): Human expert ceiling ∼90%; frontier models 88–91%[12]
- Bar Exam: Human pass rate ∼50–60%; GPT-4 achieved 90th percentile[13]
- AIME (competitive math): Top 500 US students ∼90%; o3 achieved 96.7%[14]
- SWE-Bench Verified: Entry-level engineer ∼70% (estimated); Claude Opus 4.5 ∼81%[15]

Framework threshold for Expert (Level 3): 90th percentile of skilled adults.

Assessment: On multiple professional benchmarks, frontier models exceed the 90th percentile threshold. Some models (o3 on AIME, various models on GPQA) approach or exceed 99th percentile, suggesting Exceptional (Level 4) performance on specific tasks.

Level Classification:
□ Level 1 (Emerging) — Equal to unskilled human
□ Level 2 (Competent) — 50th percentile of skilled adults
☒ Level 3 (Expert) — 90th percentile of skilled adults
□ Level 4 (Exceptional) — 99th percentile of skilled adults

Caveats: This assessment applies to specific benchmarked tasks. The framework specifies that Level 3 *AGI* requires Expert performance across "most cognitive tasks," not just those where benchmarks exist.

4.3 Performance on Real-World Work

Measure: How does AI performance compare to skilled professionals on actual work products?

Reference benchmark: GDPval—1,320 tasks across 44 occupations, evaluated by industry experts with average 14 years experience.[16]

[10] Ibid., Table 1 notes.

[11] Rein, David, et al. "GPQA: A Graduate-Level Google-Proof Q&A Benchmark." arXiv:2311.12022, 2023. https://arxiv.org/abs/2311.12022; model scores from various benchmark reports, December 2025.

[12] Hendrycks, Dan, et al. "Measuring Massive Multitask Language Understanding." arXiv:2009.03300, 2020; model scores from Artificial Analysis, December 2025.

[13] OpenAI. "GPT-4 Technical Report." arXiv:2303.08774, 2023.

[14] OpenAI. "Introducing o3." December 2024. https://openai.com/index/deliberative-alignment/

[15] Various benchmark reports, December 2025.

[16] Patwardhan, Tejal, et al. "GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks." arXiv:2510.04374, October 2025. https://arxiv.org/abs/2510.04374

Reference values:

- Claude Opus 4.1: ∼48% win+tie rate vs. industry experts
- GPT-5: ∼40% win+tie rate vs. industry experts
- 50% win+tie would indicate median performance relative to 14-year veterans

Framework threshold for Competent (Level 2): 50th percentile of skilled adults.
Assessment: On real-world work products judged by experienced professionals, frontier models approach but do not consistently exceed the 50th percentile threshold. This suggests Emerging-to-Competent performance on practical tasks, even where benchmark performance suggests Expert level.

Level Classification:
☐ Level 1 (Emerging) — Equal to unskilled human
☒ Level 2 (Competent) — 50th percentile of skilled adults
☐ Level 3 (Expert) — 90th percentile of skilled adults
☐ Level 4 (Exceptional) — 99th percentile of skilled adults

Caveats: GDPval tests one-shot task completion; iterative refinement and human-AI collaboration may yield higher effective performance.

4.4 Performance Unevenness

Measure: How consistent is performance across task types?
Reference values:

- Language tasks (writing, summarization): Strong performance
- Formal reasoning (math, logic): Strong on trained patterns; variable on novel problems
- Abstract reasoning (ARC-AGI): 0–55% depending on model and compute[17]
- Physical reasoning: Limited (no embodiment)
- Extended planning: Inconsistent

Assessment: Performance varies dramatically by task type. Models may be Expert or Exceptional on some tasks while remaining Emerging on others. This "unevenness" is explicitly acknowledged by Morris et al.: "general systems that broadly perform at a level N may be able to perform a narrow subset of tasks at higher levels."[18]

Level Classification: Variable—Expert on benchmarked tasks, Competent on real-world work, Emerging on novel abstract reasoning (ARC-AGI-2).

Interpretation: The framework accommodates uneven performance by specifying that a system's level is its *minimum* across most tasks, not its maximum on any task. By this standard, frontier models are likely Competent AGI approaching Expert AGI on many tasks, while remaining Emerging on others (notably ARC-AGI-2, where LLMs score near 0%).

[17] ARC Prize results, 2024–2025.
[18] Morris et al. 2023, Table 1 notes.

5 Criterion 2: Generality

5.1 What Morris et al. Meant

The framework distinguishes Narrow AI (specialized) from General AI (broad). Generality is about the *breadth* of tasks a system can handle at a given performance level. A system that achieves Expert performance on chess but nothing else is “Expert Narrow AI.” A system that achieves Competent performance across many cognitive tasks is “Competent AGI.”

The authors were explicit: “It is impossible to enumerate the full set of tasks achievable by a sufficiently general intelligence. As such, an AGI benchmark should be a living benchmark.”[19]

5.2 Task-Type Breadth

Measure: Number of cognitively distinct task categories handled at specified performance levels.

Reference values:

- MMLU: 57 subject areas
- BIG-Bench: 204 tasks
- Frontier LLMs: Competent or better on most of these
- Human cognitive breadth: Thousands of task types

Threshold: Framework requires performance across “most cognitive tasks” for AGI designation.

Assessment: Frontier LLMs demonstrate breadth across hundreds of task categories. The contrast with narrow AI (single-task systems like Deep Blue or AlphaFold) is stark. Whether this constitutes “most cognitive tasks” depends on the denominator.

Level Classification:
☐ Narrow — Performance limited to specific task domains
☒ General — Performance across broad range of cognitive tasks

5.3 Contrast with Narrow Systems

Measure: Do frontier systems exhibit the narrow-vs-general distinction the framework emphasizes?

Reference values:

- Deep Blue (1997): Superhuman at chess; zero capability elsewhere
- AlphaFold (2020): Superhuman at protein structure prediction; zero capability elsewhere
- Frontier LLMs (2025): Competent-to-Expert across language, math, coding, reasoning, analysis, creative tasks

Assessment: The defining feature of frontier LLMs is their generality. Unlike previous AI milestones, they are not narrow specialists. By the framework’s explicit contrast case, they are on the “general” side of the matrix.

Generality Classification: General.

[19] Morris et al. 2023, Section 5.

5.4 Gaps in Generality

Measure: What cognitive tasks do frontier systems fail at?
Reference gaps:

- ARC-AGI-2: Pure LLMs score 0%[20]
- Extended autonomous planning: Inconsistent
- Physical reasoning without embodiment: Limited
- True cross-session learning: Absent

Assessment: Significant gaps remain. The framework acknowledges that "individual humans also lack consistent performance across all possible tasks, but remain generally intelligent."[21] The question is whether current gaps are comparable to human unevenness or qualitatively different.

Generality Classification: General, but with notable gaps—contested whether gaps are disqualifying.

[20] ARC Prize Foundation, 2025.

[21] Morris et al. 2023, Section 5.

6 Criterion 3: Autonomy Level

6.1 What Morris et al. Meant

The framework treats autonomy as *orthogonal* to capability. A highly capable system can operate as a tool (Level 1 autonomy) or as an autonomous agent (Level 5 autonomy), depending on deployment choices. Capability "unlocks" higher autonomy levels but does not require them.

This decoupling is deliberate: "AGI is not necessarily synonymous with autonomy."[22]

6.2 Current Deployment Paradigms

Measure: What autonomy levels do current AI systems operate at in practice?

Reference deployments:

- Chatbots (ChatGPT, Claude.ai): Level 1–2 (Tool to Consultant)
- Coding assistants (Copilot, Cursor): Level 2–3 (Consultant to Collaborator)
- Agentic systems (Claude Code, Devin): Level 3–4 (Collaborator to Expert)
- Fully autonomous agents: Limited deployment

Assessment: Current deployments span Levels 1–4. Level 5 (fully autonomous AI) is technically achievable but rarely deployed due to safety and reliability concerns.

Level Classification:
☐ Level 1 — AI as Tool
☐ Level 2 — AI as Consultant
☒ Level 3–4 — AI as Collaborator/Expert
☐ Level 5 — Autonomous Agent

6.3 Capability for Higher Autonomy

Measure: Could current systems operate at higher autonomy levels if deployed differently?

Reference evidence:

- Apollo Research evaluations: Documented autonomous goal pursuit, including self-preservation behaviors[23]
- Anthropic alignment research: Documented strategic behavior when goals conflict with operators[24]
- Multi-step task completion: Demonstrated in agentic deployments

Assessment: The capability for higher autonomy appears to exist, constrained by deployment architecture rather than fundamental limits. This aligns with the framework's claim that autonomy is unlocked by capability but not determined by it.

Autonomy Classification: Capability demonstrated for Levels 3–4 (Collaborator to Expert); typically deployed at Levels 1–3 (Tool to Collaborator).

Caveat: Whether capability for autonomous behavior implies *safe* or *reliable* autonomous behavior is a separate question the framework does not directly address.

[22] Morris et al. 2023, Section 6.

[23] Apollo Research. "Evaluations of Frontier Models for Dangerous Capabilities." 2024. https://www.apolloresearch.ai/research

[24] Anthropic. "Alignment Faking in Large Language Models." December 2024. https://www.anthropic.com/research/alignment-faking

7 Criterion 4: Metacognition

7.1 What Morris et al. Meant

The framework's third principle specifies that AGI should be assessed on both "cognitive and metacognitive tasks." Metacognition includes the ability to learn new skills, assess one's own performance, and recognize when to seek help. This is distinguished from mere task execution.

"A useful benchmark task for operationalizing a metacognitive skill might include assessing a model's ability to determine what it does or does not know, e.g., recognizing when it should ask clarifying questions or seek additional information."[25]

7.2 In-Context Learning

Measure: Can systems improve performance from examples provided within a session?

Reference values:

- Zero-shot: Baseline performance
- Few-shot (3–5 examples): Consistent improvement across most task types[26]
- Many-shot (50+ examples): Further improvement, especially on novel formats

Assessment: In-context learning is a defining feature of modern LLMs and represents genuine metacognitive capability—adapting behavior based on demonstrated examples.

Metacognitive Assessment: Demonstrated.

7.3 Self-Assessment and Uncertainty

Measure: Can systems accurately assess their own knowledge and capabilities?

Reference values:

- Calibration: Modern LLMs show improved but imperfect calibration[27]
- "I don't know" recognition: Present but inconsistent
- Request for clarification: Present in instruction-following contexts

Assessment: Some metacognitive self-assessment exists but is unreliable. Systems can express uncertainty but do not always do so accurately.

Metacognitive Assessment: Partial—present but unreliable.

7.4 Cross-Session Learning

Measure: Can systems learn and improve across sessions via weight updates?

Reference values:

- Human cognition: Continuous learning across lifetime
- Current LLMs: No weight updates from deployment interactions
- Memory features: Provide continuity of information, not learning

Assessment: Current systems do not learn in the sense of updating weights from user interactions. This is the same gap identified in Chapter 3 (Formalization Benchmark).

Metacognitive Assessment: Absent.

[25] Morris et al. 2023, Section 5.

[26] Brown, Tom, et al. "Language Models are Few-Shot Learners." NeurIPS 2020. https://arxiv.org/abs/2005.14165

[27] Various studies on LLM calibration; systematic meta-analysis would strengthen this assessment.

8 Summary: The Synthesis Benchmark

Criterion	**Assessment**	**Classification**
1. Performance Level		
Benchmarks	90th+ percentile on multiple professional benchmarks	Level 3 (Expert)
Real-world work	∼48% win rate vs. 14-year experts	Level 2 (Competent)
Unevenness	Expert on some tasks; Emerging on others	Variable
	Overall Performance	**Level 2–3**
2. Generality		
Task breadth	Hundreds of cognitive task categories	General
Narrow contrast	Clear distinction from single-task systems	General
Gaps	ARC-AGI-2, cross-session learning, embodied tasks	Contested
	Overall Generality	**General**
3. Autonomy		
Current deployment	Levels 1–4 in practice	Level 1–4
Capability for higher	Evidence of autonomous goal pursuit	Level 3–4 capable
	Overall Autonomy	**Level 3–4**
4. Metacognition		
In-context learning	Demonstrated across task types	Demonstrated
Self-assessment	Present but unreliable	Partial
Cross-session learning	Not present	Absent
	Overall Metacognition	**Partial**
Framework Classification		**Competent AGI (approaching Expert)**

9 Interpretation

9.1 Where Current AI Falls

By the Morris et al. framework, current frontier AI systems appear to be:

- **Performance:** Competent (Level 2) to Expert (Level 3), depending on task type
- **Generality:** General (not Narrow)
- **Autonomy:** Capable of Levels 3–4; typically deployed at Levels 1–3

This places frontier systems at "**Competent AGI**" with Expert performance on some tasks—or "**Emerging Expert AGI**" if one prefers. The framework's matrix representation captures this: current systems occupy cells in the General column, spanning from Emerging to Expert rows depending on the task.

9.2 Has "Competent AGI" Been Achieved?

The framework defines Competent AGI as a general system performing at the 50th percentile of skilled adults across most cognitive tasks. Our assessment suggests:

Evidence for:

- Expert-level benchmark performance on many professional tasks
- Broad generality across hundreds of task types
- Real-world work products approaching median expert quality
- Demonstrated metacognitive capabilities (in-context learning)

Evidence against:

- Zero performance on some cognitive tasks (ARC-AGI-2)
- No cross-session learning capability
- GDPval shows <50% vs. expert professionals on practical tasks
- Significant unevenness across task types

Verdict: The boundary is contested. A generous reading places current systems at Competent AGI; a strict reading says the gaps (ARC-AGI-2, cross-session learning) disqualify them. The framework itself acknowledges this ambiguity: "We hesitate to specify the precise number or percentage of tasks that a system must pass at a given level of performance in order to be declared a General AI at that Level."[28]

9.3 The Threshold Avoidance Problem

The Morris et al. framework was explicitly designed to avoid threshold debates by proposing graduated levels. Does it succeed?

Partially. The levels provide useful vocabulary: saying a system is "Competent AGI approaching Expert" conveys more information than saying it "is" or "isn't" AGI. But the framework cannot escape the question of what counts as "most cognitive tasks." If a system fails dramatically on ARC-AGI-2 (a test of general reasoning), does it forfeit the "General" designation regardless of its performance elsewhere?

The framework's answer—that AGI benchmarks should be "living" and evolving—is pragmatic but unsatisfying. It means the goalposts are designed to move. A system that qualifies as AGI today might not qualify tomorrow if new tasks are added to the benchmark.

[28] Morris et al. 2023, Section 5.

9.4 Comparison with Earlier Benchmarks

Benchmark	**Year**	**Score/Classification**
Gubrud	1997	66%
Reinvention (Legg/Goertzel/Voss)	2002	80%
Formalization (Legg & Hutter)	2007	67%
Corporatization (OpenAI Charter)	2018	52%
Critique (Chollet)	2019	32%
Synthesis (Morris et al.)	2023	Competent AGI

The Synthesis benchmark yields a qualitatively different verdict: not a percentage but a classification. By its own terms, current AI has achieved what the framework calls “AGI”—at the Emerging-to-Competent level. This is the first benchmark in our series to place current systems *within* an AGI category rather than short of a threshold.

This reflects the framework’s design. By defining AGI as a spectrum rather than a threshold, and by placing the “Emerging” level at “equal to or somewhat better than an unskilled human,” the framework ensures that any general-purpose AI system capable of conversation qualifies as at least “Emerging AGI.” This is a feature, not a bug—the authors intended to track progress rather than declare arrival. But it means the framework cannot answer the question previous definitions tried to address: have we achieved the goal?

10 The Verdict (Provisional)

The Morris et al. framework classifies current frontier AI as **Competent AGI (Level 2)**—or possibly **Emerging Expert AGI**—with General breadth and capability for Level 3–4 Autonomy.

This is simultaneously the most optimistic and least informative verdict in our series. Most optimistic because it places current systems within an "AGI" category. Least informative because the framework was designed to avoid binary thresholds in favor of graduated levels.

10.1 What the Framework Reveals

The Synthesis benchmark succeeds in providing vocabulary for nuanced discussion. Saying that GPT-4 is "Emerging AGI" while Claude Opus 4.5 might be "Competent AGI approaching Expert" is more useful than arguing about whether either "is" AGI. The decoupling of capability from autonomy clarifies that highly capable AI need not be deployed autonomously.

10.2 What the Framework Conceals

The framework's emphasis on capability over process obscures questions about what kind of intelligence current systems exhibit. A system that achieves Expert performance through pattern matching over vast training data is classified the same as one that achieves it through genuine reasoning. Chollet's critique (Chapter 5) argues this distinction is central; the Morris et al. framework argues it is irrelevant to classification.

The framework also cannot resolve whether the gaps matter. Is zero performance on ARC-AGI-2 disqualifying for "General" status? Is the absence of cross-session learning fundamental or incidental? The framework provides no principled answer.

10.3 The Meta-Question

This chapter evaluates a framework that was itself evaluating prior definitions. The framework's authors—including Shane Legg, who helped coin "AGI" and formalize it—concluded that graduated levels are more useful than binary thresholds. Our evaluation suggests they are partially right: the levels do enable clearer communication. But they do not dissolve the underlying question.

Either current AI systems exhibit the kind of general intelligence that previous decades of researchers imagined, or they do not. The Morris et al. framework cannot answer this question because it was designed not to. It tells us where we are on the path; it cannot tell us whether the destination is the one we were seeking.

We do not speak for the authors. Morris, Legg, and their colleagues are alive and actively publishing. Their framework reflects a deliberate choice to operationalize progress rather than declare arrival. Whether this choice serves clarity or evades the hard question is itself contested.

11 Methodological Notes

This chapter departs from the methodology of previous chapters. This is a deliberate choice, not an oversight.

Why we departed. The 0%/50%/100% scoring system used in Chapters 1–5 was designed for definitions that propose thresholds. Each previous definition invited a binary question: does current AI meet this standard? Our coarse trichotomy forced honesty about evidential uncertainty while still yielding a single percentage score.

Morris et al. designed their framework to resist exactly this kind of evaluation. They replaced binary thresholds with graduated levels, decoupled performance from generality from autonomy, and explicitly argued that "levels" are more useful than "arrival." Forcing their taxonomy into our trichotomy would distort the very thing we are evaluating. We therefore assess the framework on its own terms: level classifications rather than percentage scores.

Evaluating a framework, not a definition. The Morris et al. paper is not primarily a definition of AGI but a framework for discussing it. Evaluating a framework requires assessing both where current systems fall within it and whether the framework itself is useful.

Institutional context. The framework comes from Google DeepMind, an organization with commercial interests in AI development. We have noted this context without assuming it invalidates the framework's content.

The authors' own assessment. In 2023, the authors classified frontier LLMs as "Emerging AGI." Two years later, our assessment suggests "Competent AGI" may be more accurate. This could reflect model improvements, different operationalizations, or both.

Why no overall percentage? The framework resists reduction to a single score. We could compute one (by scoring each subcriterion and averaging), but this would obscure the framework's central insight: that generality and performance are distinct dimensions that should be tracked separately.

A Note for the Methodologically Inclined

For readers who wish to maintain comparability across chapters, we offer this observation: by the framework's own nomenclature, current frontier AI has achieved "Competent AGI." The word "AGI" appears in the classification. If achieving the label constitutes achieving the thing, then our result maps to 100%—the first such score in the series.

But perhaps that is letting a titular success impose on our reasoning. "Competent AGI" is Level 2 of 5; the framework explicitly reserves "Superhuman AGI" for systems that exceed human performance on all cognitive tasks. Is partial arrival really arrival? Is "Competent AGI" to "AGI" as "competent doctor" is to "doctor"—or as "competent forgery" is to "the real thing"?

The honest answer: somewhere between 50% and 100%. We leave the precise calibration as an exercise for the reader.

12 Citation Gaps and Requests for Collaboration

The following claims would benefit from stronger sourcing:

- Systematic benchmark suite for assessing “50th percentile of skilled adults” across diverse tasks
- Rigorous comparison of LLM performance to human professional baselines across occupations
- Systematic study of performance unevenness across task types in frontier models
- Independent verification of GDPval methodology and results
- Formal analysis of whether current LLMs satisfy the “most cognitive tasks” criterion
- Systematic comparison of current systems to the framework’s metacognitive requirements
- Survey of AI researchers on whether they consider current systems “AGI” by any definition

If you can fill any of these gaps, please contribute.

A Scorecard Template

The following template can be used to classify other AI systems within the Morris et al. framework.

System evaluated: ______________________________
Evaluation date: ______________________________
Evaluator: ______________________________

Performance Level

Task Category	L1	L2	L3	L4	L5
	Emerging	Competent	Expert	Exceptional	Superhuman
Language tasks	□	□	□	□	□
Mathematical reasoning	□	□	□	□	□
Coding/programming	□	□	□	□	□
Scientific Q&A	□	□	□	□	□
Abstract reasoning	□	□	□	□	□
Real-world work tasks	□	□	□	□	□

Overall Performance Level: ______________________________

Generality

□ Narrow — Performance limited to specific domains
□ General — Performance across broad range of cognitive tasks

Notable gaps in generality: ______________________________

Autonomy Level

Autonomy Level	Capability	Deployment
Level 1: AI as Tool	□	□
Level 2: AI as Consultant	□	□
Level 3: AI as Collaborator	□	□
Level 4: AI as Expert	□	□
Level 5: Autonomous Agent	□	□

Metacognition

Capability	Absent	Partial	Demonstrated
In-context learning	□	□	□
Self-assessment/calibration	□	□	□
Cross-session learning	□	□	□

Framework Classification

Performance Level: ____

Generality: □ Narrow □ General

Classification: ____________________(e.g., "Competent AGI," "Expert Narrow AI")

Level Descriptions:

Level	Description
1 (Emerging)	Equal to or somewhat better than an unskilled human
2 (Competent)	At least 50th percentile of skilled adults
3 (Expert)	At least 90th percentile of skilled adults
4 (Exceptional)	At least 99th percentile of skilled adults
5 (Superhuman)	Outperforms 100% of humans

Notes:

Evidence and citations: