Highlights of the Issue: Governance, Agents, Evolutionary Search (In progress)

Recursive Self-Improvement (RSI)

In the next issue we will publish more articles on RSI, for instance, using reinforcement learning (RL). These articles, along with, in this issue, Darwin Gödel Machine and DarwinLM, show that RSI will come in different flavors, e.g. here using evolutionary search, and have different purposes. So far, no one RSI technique itself triggers dramatic progress toward AGI.

Agentic AI vs. AI Agents

The goal of Agentic AI — to render AI scalable and adaptable in complex environments — seems inherently so much more powerful than individual AI agents that the focus on progress toward artificial general intelligence should be on Agentic AI progress against benchmarks.

Clearly, the most advanced Agentic AI should be deployed to advance safety and value alignment, as Kumarage and colleagues show in this issue.

Development of Agentic AI indicates a need to focus on the type of AGI ecology that will emerge (e.g. multipolar scenarios rather than singletons), and permits using the paradigm of game theory to control safety and value alignment, as I wrote in 2019.[1] I have more to say on this subject.

Timeline to Artificial General Intelligence 2025 – 2030+

Senior Editor-at-Large Gil Syswerda has constructed a provocative Timeline to Artificial General Intelligence 2025 – 2030+. He gives key AI advances, economic, social, and geopolitical effects of AI. As early as 2028-2029, AI could replace the majority of human economic activity, potentially disrupting society due to widespread change happening so rapidly. But he is optimistic:

> *By the end of the decade, superintelligence is present on Earth. Human institutions are no longer in control. Everything changes in ways beyond current understanding.…Humanity survives the*

[1] Carlson, K. W. (2019). Safe artificial general intelligence via distributed ledger technology. Big Data Cogn. Comput., 3(40). doi:10.3390/bdcc3030040.

> *transition—and enters an Age of Abundance. The meaning of citizenship, nationhood, and law undergoes foundational redefinition.*

Articles

Comparing Apples to Oranges: A Taxonomy for Navigating the Global Landscape of AI Regulation

The EU has a unified framework for regulating AI, China has its own governance structure, which it has efficiently imposed throughout China due to centralized control. Since the proposed 10-year moratorium on state AI regulation was killed in the 'Big Beautiful Bill,' the US now seems headed toward a uncoordinated patchwork of regulation at the state and federal level. Thus, this work by Alanoca et al. is critically important and urgent. From the abstract:

> *Clarifying the scope and substance of AI regulation is vital to uphold democratic rights and align international AI efforts. We present a taxonomy to map the global landscape of AI regulation. Our framework targets essential metrics—technology or application-focused rules, horizontal or sectoral regulatory coverage, ex ante or ex post interventions, maturity of the digital legal landscape, enforcement mechanisms, and level of stakeholder participation—to classify the breadth and depth of AI regulation.*

As our co-founding editor, Steve Omohundro foresaw in 2014, AI itself will play an increasing role in evolving the legacy human legal regime in general and regulations governing AI in particular:

> *The legal codes of many countries have become quite complex. Several AI projects are trying to create formal digital versions of legal codes (CodeX, 2014). These systems will eventually be used to resolve legal issues and perhaps even act as arbitrators or judges. Sophisticated AI systems with knowledge of the legal system will be used to help craft and simplify new legislation.* [2]

Accordingly, humans should initiate the process of AI taking over evolution of law, notably in the area of AI safety, with human oversight. In principle, LLMs can absorb the entire global body of human law and compare the different regimes to provide a cohesive unified over-arching structure, which no human is capable of doing. As AI progresses toward AGI, it can evolve the Alanoca et al. taxonomy to organize the diverse and inchoate

[2] Omohundro, S. (2014). Cryptocurrencies, Smart Contracts, and Artificial Intelligence. *AI Matters, 1*(2), 19-21. doi:10.1145/2685328.268533.

international, national, and local regulatory regimes following the examples the authors provide.

Real-World Gaps in AI Governance Research

Strauss et al. did a massive survey of 1,178 safety and reliability papers analyzed from 9,439 generative AI papers between Jan 2020 and Mar 2025. They found significant gaps in AI governance research, particularly in post-deployment contexts and high-risk areas.

- Corporate AI research is increasingly influential, focusing on pre-deployment safety while neglecting real-world deployment issues.
- Corporate AI (Anthropic, Google DeepMind, Meta, Microsoft, OpenAI) has more citations than leading academic institutions (academic institutions (Carnegie Mellon University, Massachusetts Institute of Technology, New York University, Stanford University, University of California Berkeley, and University of Washington) – see their Table 2.
- Google DeepMind has more citations than the top four academic institutions combined.
- Corporate AI research prioritizes model alignment and testing, with less focus on deployment-stage issues like bias.
- There is a critical lack of research on the safety and reliability of AI systems in real-world applications.

AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges

Sapkota et al. delineate agent AI vs. agentic AI, briefly giving the history and the motivation behind agentic AI toward its prime goal of *scalability and adaptability in complex environments*. Thus, agentic AI is a key step in the evolution of AI to AGI and superintelligence. As for safety, *trust-centric operations* will prioritize safety mechanisms, ensuring verifiable output and ethical compliance.

- AI Agents are modular systems driven by LLMs and LIMs for task-specific automation.
- Agentic AI represents a paradigm shift with multi-agent collaboration, dynamic task decomposition, and coordinated autonomy.
- AI Agents are autonomous software entities designed for goal-directed task execution within bounded environments.
- Key characteristics include autonomy, task-specificity, and reactivity with adaptation.
- Agentic AI systems manage complex, multi-step tasks requiring coordination among multiple agents..

- Generative AI systems are stateless and lack the ability to interact with their environment autonomously.
- LLMs exhibit reactive behavior, producing output only when prompted, without autonomous goal pursuit.
- The evolution from generative models to AI Agents is driven by the integration of large-scale language models (LLMs) as reasoning engines.
- AI Agents utilize LLMs like GPT-3 and LLaMA to perform adaptive planning and real-time decision-making.
- Agents function as cognitive engines that interpret user goals and manage complex workflows.
- Agentic AI systems extend the capabilities of traditional AI Agents by enabling collaboration among multiple intelligent entities.
- They allow for goal decomposition, where user objectives are parsed into manageable tasks distributed across agents.
- Inter-agent communication is facilitated through asynchronous messaging and shared memory

Measuring AI Agent Autonomy: Towards a Scalable Approach with Code Inspection

How to define "artificial general intelligence"? It's like trying to define "life". Most agree autonomy has to be included in defining "human-level intelligence". In the next issue we will give our view. Here Cihon et al. present an operational definition of "autonomy" based on the agent architecture (a better description than 'code inspection') and without having to observe agent behavior, *per se*, based on an eight-component taxonomy of autonomy – see their Figure 1. For example, human-programmed goals = no autonomy, while agent-programmed goals = autonomy.

As the authors note and we stress, "the level of agent autonomy is crucial for understanding both their potential benefits and risks," i.e. AGI safety.

Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents: Main Article & Appendix F on Safety

> *In the DGM, improvement in downstream tasks directly reflects an increase in self-improvement ability, enabling the potential for self-accelerating progress.*
>
> *However, these [various foundation model] approaches have yet to close the self-improvement loop, meaning improvements on downstream tasks do not translate into enhanced capabilities for self-modification or the acceleration of further innovations. We aim to mimic the acceleration of science and technology, where new*

> *tools and discoveries catalyze the creation of even more discoveries. Similarly, how can we emulate nature's arc of evolution, which bends not only toward complexity but also an ever greater capacity to evolve [26, 41, 49]?*

This article is notable for several reasons. First, recursive self-improvement is a strong signal of progress toward AGI/SI; the Darwin Gödel Machine (DGM) modifies its own code. Second, Schmidhüber's provably self-improving Gödel Machine was theoretically bold but stalled on the self-proving piece. Zhang et al. use SOTA benchmarks (SWE-bench, Polyglot) to measure various improvements and therefore prove improvement operationally. They say their 'empirical proofs' weaken the Gödel Machine formal approach – but I disagree – the formal methods, for example, cannot prove true statements that their axioms don't capture. Third, we hypothesize that progress toward AGI requires broadening the ML paradigm. Here Zhang et al. incorporate evolutionary programming to generate a population of coding agents and randomly sample them to find code improvements.

The most critical and important application of the most advanced AI is to apply to creating safety and value alignment.

Thus, the authors point out that self-improvement can be focused on the system's *safety* (which I call *recursive safety improvement*) and describe the safety precautions they followed.

DarwinLM: Evolutionary Structured Pruning of Large Language Models

This second article enabling AI recursive self-improvement (RSI) also uses evolutionary search, in this case to optimize model compression while maintaining performance. Instead of competing agents as in Darwin Gödel Machine, DarwinLM generates offspring models through mutation and selects the fittest for survival. Improving over previous efforts, non-uniform model compression is necessary since different model layers show varying sensitivities to pruning. Thus, offspring models are created by applying a mutation operator to the parent model. The operator adjusts sparsity levels by increasing sparsity in one module while decreasing it in another, ensuring the overall sparsity constraint is maintained.

DarwinLM achieves superior results compared to prior methods, cutting training data requirement by up to 80% post-compression.

Towards Safety Reasoning in LLMs: AI-agentic Deliberation for Policy-embedded CoT Data Creation

"Traditional" AI safety training primarily relies on Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF). Newer methods include Reinforcement Learning from AI Feedback (RLAIF), utilizing AI-generated evaluations for

scalable safety training, and Direct Preference Optimization (DPO), which optimizes model outputs based on preference data, simplifying alignment processes.

In contrast, Kumarage et al.'s AIDSAFE is an Agentic, reasoning Chain-of-Thought (CoT) LLM wherein multiple agents deliberate together to improve the safety of the model while minimizing over-refusal. Five key safety policies guide the reasoning process: Hate-Harass-Violence, Fraud and Deception, Physical Harm, Illegal Activity, and Helpfulness and Respectfulness Policy. AIDSAFE CoTs outperforming single LLM generations in all metric: relevance, coherence, completeness, and faithfulness to safety policies.

The authors claim CoT and agentic collaboration increase the models' ability to internalize complex safety policies, and collaboration reduces hallucinations.

Trends in Frontier AI Model Count: A Forecast to 2028

Compute power predictions are critical in several respects, such as the argument that scalability is all we need to achieve AGI and governance frameworks (EU AI Act, US AI Diffusion Framework) that measure AI capability and restrict AI deployment according to compute power.

> *Our aim is to forecast the number of models that will be released above different training compute thresholds over the next four years. To do this, we model scenarios for the distribution of AI model releases over training compute. Once these distributions are established, we can simply count the models that exceed each specific compute threshold.*

This table gives their model's results:

Threshold (FLOP)	2024	2025	2026	2027	2028
$> 10^{25}$	[18, 23, 27]	[32, 45, 64]	[51, 77, 119]	[76, 117, 201]	[103, 165, 306]
$> 10^{26}$	[0, 0, 0]	[3, 7, 11]	[12, 24, 38]	[27, 47, 81]	[45, 81, 148]
$> 10^{27}$	[0, 0, 0]	[0, 0, 0]	[0, 2, 5]	[1, 10, 20]	[9, 27, 56]
$> 10^{28}$	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]	[0, 3, 8]
$> 10^{29}$	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]

Table 3: Results for absolute thresholds. The table presents 90% prediction intervals [5th, 50th, 95th percentile] for the number of models exceeding each compute threshold. Results are cumulative, showing estimates for the number of models released by the end of each year.