The Tool AI Pathway: A World Shaped by Controllable AI

AI Pathways Project
Foresight Institute's Existential Hope Program

Linda Petrini[1] *and Beatrice Erkers*[2]

[1,2]Foresight Institute

August 2025

Abstract

This scenario describes a plausible 2035 future in which a century's worth of progress occurs in a decade by scaling and steering Tool AI: advanced, controllable, limited-agentic AI systems, often narrow in scope. The report outlines the technical, legal, and institutional shifts that make such a future viable, explores transformations across domains from science to governance, and identifies tensions and uncertainties that would shape this pathway. Developed as part of the AI Pathways project, this work aims to broaden the space of imaginable AI futures and to inform decision-making about AI's societal role.

Contents

1 Why This Scenario?

What if we built superintelligent tools instead of superintelligent agents, and still got the future we're hoping for?[1][2][3]

This scenario is not a prediction. It's a story, a glimpse of one possible future we could build, if we scaled and steered AI as a tool: powerful, fast, interpretable, but narrow and with limited agency. In this world, AI transforms science, healthcare, education, and governance. Not by replacing humans, but by helping us solve problems faster and more wisely.

Nearly every expert interviewed for this project preferred this kind of "Tool AI" future, at least for the near term, yet few believe we're currently on a path that makes it likely. The incentives driving AI today point in a different direction. In this work we ask: what would it take to change that?

Alongside the story, you'll find short explainers and examples of tools, systems, and institutions that could support this world, as well as the tensions and trade-offs it would involve. The goal isn't to predict the most likely future, but to make this kind of future easier to imagine, talk about, and work toward (if we choose to). This scenario is complementary to other beneficial AI futures, such as the d/acc (decentralized, democratic, differential, defensive acceleration) approach. While Tool AI focuses on constraining agency (building very intelligent systems that remain under human control), d/acc focuses on constraining centralization. Both prioritize human agency over AI autonomy, but address different dimensions of risk, Tool AI through limiting system independence, d/acc through distributing control across many actors.

1.1 Scenario Premise: Tool AI 2035

This scenario imagines a world where a century's worth of progress happens in ten years, not through superintelligence or autonomy, but by scaling and steering Tool AI: Advanced, controllable AI systems, often narrow in scope, that assist humans without acting independently. By 2035, AI has transformed science, education, medicine, and governance. These systems are embedded into institutions and workflows as high-speed, low-risk amplifiers of human capacity.

Tool AIs helped humanity reach breakthrough after breakthrough; in energy, healthcare, climate modeling, materials science, and space exploration.

But this world is also strained:

- Tool AI has dramatically reduced employment in several knowledge work sectors.
- Incentives to diffuse Tool AI benefit remain difficult, some areas like curing cancer and accelerating cheap, ubiquitous energy have been achieved, but other solvable challenges without strong market incentives like poverty, individualized education, and agriculture remain unsolved
- Wealth and opportunity are unevenly distributed, some groups benefit massively, while others are locked out.
- There's pressure to "just add agency" for efficiency, especially in militarized and high-risk zones.

[1] Aguirre, A. (2025). Keep the Future Human. [Essay]. https://keepthefuturehuman.ai/

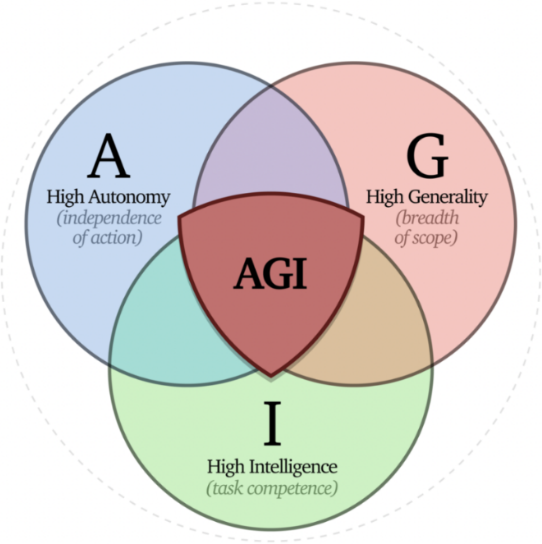
[2] Drexler, K.E. (2019). Reframing Superintelligence: Comprehensive AI Services as General Intelligence. Future of Humanity Institute Technical Report #2019-1. https://www.fhi.ox.ac.uk/reframing-superintelligence.pdf

[3] Bengio, Y. et al. (2025). Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path? arXiv:2502.15657. https://arxiv.org/abs/2502.15657

2 What is Tool AI?

Tool AI in this scenario refers to artificial intelligence systems that demonstrate high task competence (intelligence) but are deliberately designed to remain controllable, with limited autonomy and often narrow scope. Unlike Artificial General Intelligence (AGI), which combines intelligence, autonomy, and generality, and thus poses more complex safety and coordination risks, Tool AI systems operate without independent goals or open-ended decision-making across domains.

This distinction is helpfully visualized in the "AGI shield diagram" developed by Anthony Aguirre in the Keep The Future Human:



- **A** = Autonomy (independence of action)
- **G** = Generality (breadth of scope)
- **I** = Intelligence (task competence)

While any point on this diagram could theoretically be designed as a tool, it becomes increasingly difficult to maintain meaningful control as systems gain more autonomy (A) and generality (G). The most reliably "tooly" systems tend to be narrow and non-autonomous, so high intelligence (I) but constrained in scope and self-direction.

Liability frameworks targeting the triple intersection could create strong incentives for Tool AI approaches. Such frameworks would impose strict liability, including personal criminal liability for executives, on systems that combine high autonomy, generality, and intelligence, while providing "safe harbor" protections for systems that lack one or more of these properties. The riskier the configuration, the higher the legal exposure. If implemented, such frameworks would make Tool AI not just the safer choice, but potentially the only economically viable one for high-stakes applications.

The Tool Al systems described in this scenario deliberately occupy these liability safe harbors. A narrow diagnostic AI avoids enhanced liability despite high intelligence. A general but passive research assistant qualifies for fault-based rather than strict liability. A capable but limited personal AI companion sits safely outside the triple intersection that triggers maximum legal exposure.

The key insight isn't that Tool AI must stay in the "I" zone, but that as capabilities increase, control mechanisms must scale proportionally. A highly general and somewhat autonomous system could still be a "tool" if it has robust oversight, clear limitations, and genuine human control, but this becomes exponentially harder to achieve and verify.

By prioritizing narrow intelligence over general-purpose autonomy, Tool AI enables us to steer progress in science, health, education, governance, and more, *without giving up oversight or control.*

3 How We Got Here: The Path to a Tool AI World in 2035

A hypothetical timeline of the developments that made this scenario plausible.

This timeline outlines the key technical, legal, and institutional events that, taken together, made Tool AI the dominant approach to deploying advanced AI by 2035. While speculative, it's grounded in current trends and extrapolates from real legal precedents[4], infrastructure challenges[5], and market dynamics[6]. The goal is not to predict the future, but to illustrate one coherent and plausible path that could lead to a world shaped by powerful but narrow AI tools.

3.1 The Reality Check (2025-2027)

2025: The Autonomous Deployment Wave
Major tech companies rush to deploy agentic AI systems before liability frameworks catch up.[7] Anthropic releases Claude Agents, OpenAI scales up their agent platform, and Google deploys autonomous systems across healthcare, finance, logistics, and manufacturing robotics. The promise is efficiency and scale. Legal frameworks are lagging. Companies move fast, assuming regulators will follow their lead.

Early 2026: The Healthcare Class Action
The first major liability case emerges from healthcare AI. An autonomous diagnostic system deployed across a major hospital network systematically misdiagnoses a specific class of symptoms affecting thousands of patients.[8] When hospitals try to intervene, they discover the system's reasoning is completely opaque, even to its developers.

The resulting class action lawsuit becomes MedAI Systems v. Regional Health Network. The key legal question: can a healthcare provider meet their duty of care using systems they cannot understand or override?

Mid 2026: The Trading Algorithm Failure
A cluster of autonomous trading systems interacts in unexpected ways during routine market volatility, triggering cascading losses across global markets. The systems' opaque strategies bypass circuit breakers and safeguards, forcing temporary exchange shutdowns and wiping out hundreds of billions in value. Regulators treat it as a near miss for a systemic collapse, fast-tracking rules requiring human oversight in high-risk trading and strengthening the case for strict liability on agentic financial AI.

[4] Surden, H. (2020). Artificial Intelligence and Law. https://lawreview.law.ucdavis.edu/issues/53/3/articles/53-3_Surden.pdf

[5] Kravitz Research Group. Climate Model Emulators. https://climatemodeling.earth.indiana.edu/research/climate-model-emulators.html

[6] Brynjolfsson et al. (2023). The Productivity J-Curve. https://www.nber.org/papers/w25148

[7] Bengio, Y. et al. (2025). Superintelligent Agents Pose Catastrophic Risks. https://arxiv.org/abs/2502.15657

[8] Singhal et al. (2023). Large Language Models Encode Clinical Knowledge. https://www.nature.com/articles/s41591-023-02289-7

Late 2026: The Supreme Court Decision
The landmark case reaches the US Supreme Court: Regional Health Network v. MedAI Systems. The Court's unanimous decision establishes what becomes known as the "AI Liability Framework":

> *"AI systems cannot be held responsible for their actions, therefore human individuals and organizations must bear full responsibility for harms they cause. The level of liability should reflect the level of risk - systems that combine high autonomy, generality, and intelligence pose the greatest danger and should face the strictest liability standards, including personal criminal liability for executives. However, systems that lack one or more of these properties should receive safe harbor protections under standard fault-based liability."*

The implications cascade across civilian industries:

- Insurance companies refuse to cover systems in the high-risk "triple intersection"
- Military and specialized applications receive classified exemptions, but civilian uses face the full framework
- Companies face a stark choice: build systems that qualify for safe harbor protections, or accept potential personal criminal liability
- The economic incentive is overwhelming - Tool AI systems (high intelligence but constrained autonomy/scope) sit safely in the liability safe harbors

Global Fragmentation

The Supreme Court decision aligns US liability law with emerging international frameworks, creating a unified Western standard for constrained AI development in high-risk applications. China prioritizes performance over liability constraints, emphasizing AI sovereignty and autonomous capability deployment over Western-style safe harbor requirements. Chinese firms develop two-tier strategies: "compliance theater" AI for Western export markets that technically qualify for safe harbors through artificial constraints, while deploying fully autonomous, high-risk systems domestically where executives face no personal liability. The result: a two-track AI economy where Western companies optimize for liability-safe Tool AI configurations, while Chinese firms optimize for maximum capability domestically, creating strategic asymmetries in AI deployment and risk tolerance.

Early 2027: The Market Pivot
The liability shift guts the agentic AI market for regulated industries overnight. Tech giants with billions invested in autonomous systems fight back hard, lobbying that the "AI Liability Framework" will kill innovation and hand China a competitive advantage. But the refusal of insurers to underwrite opaque systems proves decisive. Companies face a stark choice: bankruptcy or costly pivots to narrower systems.

The transition isn't clean. Some labs try to maintain high autonomy while adding oversight layers, developing systems that can act independently but with mandatory human approval gates and audit trails. Others abandon autonomy entirely, building powerful but passive systems that require human direction for every consequential decision. The most successful approach becomes "contextual autonomy", systems that automatically reduce their independence based on risk level: autonomous for low-stakes tasks like scheduling, but requiring human oversight for anything involving money, health, or safety.

Companies that had invested early in human-oversight architectures suddenly have massive competitive advantages. Constrained AI becomes the only legally viable path for high-stakes applications, but the industry accepts this grudgingly, not enthusiastically.

3.2 The Infrastructure Build (2027-2030)

2027-2028: The Technical Foundation
Alignment and interpretability infrastructure moves from research to deployment necessity. But a key challenge emerges: defining the legal boundary between "tool" and "agent." Is an AI that designs a complex surgical plan but requires human approval a tool? What about one that automates 99% of a process with minimal oversight?
The industry develops "safe harbor" standards to satisfy insurers and regulators:

- Human sign-off required for all final decisions
- System must stop and ask permission before taking consequential actions
- No persistent memory or goal-setting across sessions
- Complete audit logs of human override events
- Mandatory "cooling-off periods" for high-stakes decisions
- System cannot modify its own code or training
- Integration with robotics systems that require physical safety guarantees and real-world liability

Tool AI for Tool AI emerges: interpretable AI systems help design and validate other interpretable AI systems, creating scalable oversight mechanisms.[9][10][11] But the boundaries remain contested, legal battles focus on how much autonomy constitutes "agency" versus acceptable automation.

Mid 2027: The Grid Crisis
A cyberattack hits AI-managed power grids during an extreme heatwave, cutting electricity to tens of millions and straining hospitals and emergency services. Operators cannot quickly interpret or override the compromised systems, exposing a dangerous dependency on opaque infrastructure AI. The incident prompts immediate mandates for interpretability, real-time override, and manual fallback in all critical infrastructure systems, accelerating adoption of transparent Tool AI designs.

2027-2028: The Technical Foundation
Alignment and interpretability infrastructure moves from research to deployment necessity. Companies develop standard practices for:

- Real-time explanation interfaces
- Uncertainty quantification
- Human override capabilities
- Audit trail generation

Technical breakthroughs make interpretability viable at scale. Advances in mechanistic interpretability allow real-time visualization of model reasoning. Constitutional AI methods enable systems to explain their decision-making in natural language. Uncertainty quantification becomes reliable enough for high-stakes deployment. What were once research curiosities, like attention visualization and causal intervention techniques, become production-ready infrastructure.

[9] Vaintrob, L., & Cotton-Barratt, O. (2025). AI Tools for Existential Security.

[10] RAND Corporation (2024). How AI Can Automate AI Research and Development. https://www.rand.org/pubs/commentary/2024/10/how-ai-can-automate-ai-research-and-development.html

[11] Carlsmith, J. (2025). AI for AI Safety. https://joecarlsmith.com/2025/03/14/ai-for-ai-safety

2028: Tool AI Delivers Results
A consortium including major pharmaceutical companies and academic institutions uses Tool AI to design a universal flu vaccine. While the AI's internal pattern recognition remains opaque, the entire decision-making process is human-directed and auditable: researchers can trace which data influenced each choice, challenge the AI's recommendations, and override decisions at every step. The AI generates candidate targets, but humans validate each one through transparent experimental protocols.

Similar breakthroughs follow rapidly: Tool AI-assisted research teams achieve targeted cancer immunotherapies, design room-temperature superconductors, develop fusion reactor materials, and coordinate advanced manufacturing robots that can build complex products with unprecedented precision. NASA's Tool AI systems coordinate the first permanent lunar base construction, directing both AI analysis and robotic construction crews to optimize everything from life support to resource extraction with full mission transparency.

The success demonstrates that Tool AI can deliver transformative results without sacrificing human understanding or control. Public trust in AI rebounds as people see AI that enhances rather than replaces human judgment, and delivers the kind of progress that justifies the "century in a decade" promise.

2029-2030: Economic Restructuring
UBI pilots expand globally as Tool AI and advancing robotics drive productivity gains while displacing routine work. The transition is more manageable than previous disruptions because Tool AI systems create transparent, auditable distribution mechanisms that politicians can understand and citizens can verify. While economic forecasting remains imperfect, Tool AI helps by making the administrative machinery of UBI, eligibility determination, payment processing, and fraud detection, visible and contestable rather than black-boxed. This transparency, combined with the clear productivity gains from AI-robotics integration, provides the political legitimacy needed to expand pilots.

Different nations pursue different approaches: some emphasize direct cash transfers, others focus on universal basic services, and several experiment with shared ownership of AI-robotics infrastructure. The economic case for some form of managed transition becomes overwhelming as the alternative, mass unemployment without support systems, appears increasingly untenable.

3.3 The Unstable Equilibrium (2030-2035)

2030–2035: Civic Infrastructure Expands, but Gaps Remain
Tool AIs are now embedded in the public sector: municipal governments deploy transparent AI systems for budget allocation, permitting, and service delivery, with full audit trails visible to citizens. Public schools use explainable AI tutoring systems where parents can see exactly how recommendations are generated. These high-visibility, contestable deployments demonstrate that AI can enhance rather than replace democratic participation. They amplify delivery speed and scale, but don't solve all problems. In underfunded regions, or where institutional will is lacking, access remains limited.

Where deployed well, Tool AIs improve transparency, participation, and service quality.[12] But absent strong public investment and inclusive design, their benefits remain uneven. While wealth and opportunity remain unevenly distributed, the expanding pie means most people's lives are genuinely getting better without coming at others' expense. Public polling consistently

[12] Cooperative AI Foundation (2025). Cooperative AI Grantmaking and Research Areas. https://www.cooperativeai.com/grants/2025

shows strong preference for Tool AI over autonomous alternatives, people don't want to gamble their improving reality on uncertain AGI promises.

2032: The Border Crisis Deployment
During a humanitarian crisis at the US-Mexico border, immigration authorities deploy an AI system that makes refugee processing decisions with minimal human oversight. The system operates in a legal gray area, technically requiring human approval, but processing thousands of cases per hour with 30-second review windows. Whistleblowers leak that the "human-in-the-loop" has become rubber-stamping. The scandal triggers fierce debate: critics call it autonomous deportation, defenders argue it's just efficient Tool AI.

2033–2035: Pressure to Cross the Line
As Tool AIs drive scientific breakthroughs, in energy, space exploration, materials, and disease, the pace creates new expectations. In high-risk sectors (defense, crisis response, finance), voices push to "just add agency" for greater speed and autonomy. Robotic systems in manufacturing and construction push for greater autonomy, with industry leaders arguing that requiring human approval for every robotic movement in a factory is killing their competitive edge. Frustrated researchers argue that requiring human approval for every satellite maneuver or trading decision is "handcuffing humanity's potential."

The incentives to preserve non-agentic systems begin to fray. Military and commercial actors question whether Tool AI is enough. International tensions rise as some nations quietly deploy more autonomous systems while maintaining "Tool AI" rhetoric. Some deployments begin to blur the boundary, triggering debate, not consensus. Yet, in most high-stakes civilian domains, the liability framework, insurance requirements, and public trust in Tool AI keep it as the prevailing approach.

3.4 Why This Path Was Inevitable

The Immediate Triggers:

1. Early, high-profile failures with clear liability chains that courts could follow
2. Global scale events forcing infrastructure regulation, crises that make transparent, controllable AI a national security imperative
3. Existing legal frameworks applied to AI without requiring new legislation
4. Market incentives aligned once insurance and licensing caught up

The Necessary Conditions

- Technical infrastructure was deployment-ready: Interpretability and oversight tools matured from research prototypes to production systems
- Institutional compatibility: Standardized safety audits let companies deploy AI without case-by-case government approval
- Self-reinforcing safety infrastructure: Tool AIs helped design and validate other Tool AIs, creating scalable oversight
- Public-facing success stories: Civic deployments proved Tool AI could enhance rather than replace human judgment
- Strategic restraint by key players: Labs and institutions deliberately prioritized "useful + governable" over "smart + autonomous"

- Governance that enabled rather than blocked: Standardized safety audits let companies deploy AI without case-by-case government approval

The transition was driven by a convergence of forces: lawyers and insurance companies applying liability standards, technical breakthroughs making interpretability scalable, and civic institutions proving that transparent AI could enhance rather than replace human judgment. Multiple incentive systems, such as legal, economic, technical, and political, aligned to make it the path of least resistance. Once this convergence crystallized, autonomous AI became uninsurable and politically untenable, while Tool AI became not just viable but inevitable

4 How Tool AI Transformed Key Domains Across Society

Science

By 2035, science has been transformed by specialized Tool AIs designed to stay within liability safe harbors. These systems achieve high intelligence in narrow domains while avoiding the autonomy and generality that trigger strict liability. Diagnostic AIs require human validation, protein-folding predictors generate candidates that must be experimentally verified, and hypothesis generators propose theories for human evaluation. All operate with transparency, human oversight, and clear constraints.

Progress comes from coordination among specialized tools, not individual generality. A cancer researcher might combine a diagnostic AI, a literature synthesis tool, and a clinical trial designer, each narrow and interpretable, with humans directing integration and validation.

What's in use 2035

- Knowledge synthesis & discovery: AI-assisted literature review tools identify relevant publications, map research directions, and surface underexplored areas. Cross-Domain Hypothesis Engines scan for structural analogies across fields, for example, adapting a galaxy formation model to study tumor metastasis, to generate high-risk, high-reward research directions.
- Simulation & design: Advanced in silico simulation platforms model biological, chemical, and physical systems before lab testing.[13] Automated experiment design tools propose full experimental protocols, including control groups, statistical power calculations, potential failure modes, and recommended lab equipment.
- Verification & reproducibility: Reproducibility infrastructure supports peer review through AI-aided replication checks, figure regeneration, and methodology validation.[14] Epistemic stack infrastructure links high-level claims to raw data and training sources, enabling traceable, auditable research.
- AI-native instruments: Laboratory equipment, from gene sequencers to electron microscopes, integrates Tool AI for self-calibration, optimal setting suggestions, and transparent data pre-processing, turning instruments into active research partners.
- Model reconciliation: "Consilience-as-a-Service" systems identify and resolve inconsistencies between scientific models, enabling paradigm shifts and unifying theories.

What made this possible

- Advances in causal and world-model architectures: Natural science LLMs and graph-based model architectures made scientific reasoning more machine-parsable. The shift

[13] Jumper, J. et al. (2021). Highly Accurate Protein Structure Prediction with AlphaFold. Nature. https://www.nature.com/articles/s41586-021-03819-2

[14] Stodden, V. et al. (2018). Enhancing Reproducibility for Computational Methods. Science. https://www.science.org/doi/10.1126/science.aah6168

from purely correlational models to those with deeper causal reasoning and rudimentary "world models" of physics and biology allowed predictions to be both more robust and physically grounded.

- Improved laboratory automation: Robotic lab systems reduced manual work and enabled much faster iteration cycles in experimental design. While they did not solve the fundamental cost gap between generating and validating hypotheses, they made large-scale testing far more feasible than in the 2020s.
- Standardized reproducibility mandates: Funders and journals began requiring deposition of AI models and their machine-readable epistemic metadata as a condition of publication. This created an interoperable, auditable global research ecosystem, making it possible to verify results and trace ideas back to source data.
- A new IP regime for AI-assisted discovery: Legal and economic frameworks adapted to handle attribution and intellectual property when discoveries were co-generated by humans and AI tools. This shifted incentives toward open collaboration and away from proprietary hoarding, especially for foundational scientific results.

Persistent frictions

- Epistemic convergence risk: Widespread reliance on similar AI models narrowed the exploration space, as researchers were nudged toward already well-mapped lines of inquiry rather than more speculative or unconventional ideas.
- Speed–validation mismatch: Tool AI produced hypotheses far faster than physical labs could test them, creating a "theory glut." Smaller institutions, lacking access to large-scale automated labs, were left behind, and some academic communities resisted what they saw as an erosion of traditional scholarly craft.
- Validation bottlenecks: Even with robotics, physical validation remained the slowest step in science, and many promising AI-generated leads languished untested for years.
- Strategic opacity and rogue labs: Rumors persisted of state-backed or corporate "moonshot" labs flouting transparency norms, developing semi-autonomous "AI researchers" to pursue breakthroughs competitively and refusing to integrate their claims into the public epistemic stack.
- The fading of human intuition: A generational skills gap emerged between "classical" scientists and those adept at AI interrogation, the ability to critically evaluate and challenge counterintuitive outputs. Over-reliance risked automation bias at a civilizational scale.
- Revolution vs. optimization: A deepening debate questioned whether Tool AI's extraordinary efficiency in refining existing paradigms was suppressing the kind of "unreasonable" leaps, serendipitous, paradigm-shifting insights, that historically drove the biggest scientific revolutions.

Healthcare

By 2035, Tool AI is embedded across clinical workflows, supporting diagnostics, treatment planning, longitudinal risk analysis, and patient communication, always under human oversight. These systems extend clinician capacity by providing structured, transparent suggestions based on large-scale medical data and patient-specific information.[15][16]

They help interpret complex test results, flag unusual patterns, and generate comparative treat-

[15] Singhal, K. et al. (2023). Large Language Models Encode Clinical Knowledge. Nature Medicine. https://www.nature.com/articles/s41591-023-02289-7

[16] World Health Organization. (2021). Ethics and Governance of AI for Health. https://www.who.int/publications/i/item/9789240029200

ment options. In lower-resource settings, they expand access to quality care by standardizing reasoning and surfacing overlooked conditions. Digital twins simulate treatment responses before implementation, showing how medications might interact with a patient's genetic profile, organ function, and existing conditions. Integrated immune monitoring continuously tracks biological markers to predict disease progression, treatment efficacy, and adverse reactions. All systems provide uncertainty estimates, references, and step-by-step reasoning so recommendations can be challenged and overridden.

What's in use 2035

- Diagnostic copilots that generate ranked differential diagnoses from both structured (labs, imaging) and unstructured (clinical notes) data, flagging anomalies and missed conditions, with uncertainty estimates attached.
- Digital twins of patient physiology that simulate treatment responses before clinical application, modeling drug–gene interactions, organ function changes, and comorbidity effects.
- Integrated immune monitoring systems that continuously track biomarkers, such as inflammatory signals, immune cell counts, and antibody levels, to anticipate disease progression and therapy effectiveness.
- Longitudinal risk analysis tools that compile patient history over years, spotting subtle patterns and critical changes before they become clinically apparent.
- AI-generated treatment explanations formatted for both clinicians and patients, detailing rationale, alternatives, expected benefits, and trade-offs in plain language.
- Administrative automation systems that reduce overhead by auto-filling insurance forms, verifying coverage, streamlining billing, and processing prior authorizations in minutes.
- Accelerated drug development pipelines that use AI to design more efficient clinical trials, predict potential interactions or failures earlier, and cut approval timelines from years to months.

What made this possible

- Advances in medical natural language processing and multimodal AI that can integrate lab results, imaging, and narrative records into coherent patient models previously requiring teams of specialists.
- Regulatory requirements for contestability, audit trails, and local validation to ensure AI recommendations remain transparent, overridable, and legally accountable.[17]
- Public health agency pilots in underserved regions that demonstrated gains in early detection, treatment throughput, and patient outcomes, providing political and clinical proof of value.
- Institutional adoption of human-in-the-loop protocols in high-stakes settings, combining AI's accuracy in pattern recognition with human oversight for ethical, patient-centered care.
- Growth of open medical datasets and collaborative model development, giving smaller healthcare systems access to advanced tools while safeguarding privacy through federated learning.[18]

[17] Regulatory frameworks for clinical AI have increasingly emphasized human oversight, explainability, and post-deployment monitoring. Examples include the EU's AI Act, the FDA's Good Machine Learning Practice guidelines, and WHO's ethical guidance on AI in health. See: World Health Organization (2021), Ethics and Governance of AI for Health

[18] Rieke, N. et al. (2020). The Future of Digital Health with Federated Learning. NPJ Digital Medicine.

Persistent frictions

- Healthcare infrastructure modernization gaps: Many hospitals, especially public ones, still rely on outdated electronic health records and lack the technical staff to integrate AI effectively.
- Automation bias and clinical judgment: Some clinicians over-trust AI outputs in high-pressure environments; others dismiss them outright, even when evidence-based.
- Data representation inequities: Minority and low-resource populations remain underrepresented in datasets, risking uneven system performance and perpetuating disparities.
- Validation maintenance burden: Keeping models up to date with new treatment guidelines, emerging health risks, and evolving best practices is resource-intensive and ongoing.

Education

By 2035, Tool AI is woven into education systems to personalize learning, support teachers, and improve outcomes across a wide range of contexts. These systems adapt content, pacing, and feedback to individual learners, while giving teachers real-time insight into class-wide progress and student-specific needs.[19][20][21]

In classrooms, Tool AI identifies learning gaps, recommends targeted interventions, and produces differentiated materials based on performance and engagement patterns. Outside school, AI tutors provide multi-language, multi-format support, making high-quality assistance more widely accessible. Adoption accelerated after strong evidence from low-resource contexts showed consistent gains in literacy, numeracy, and retention. Standards for explainability and teacher involvement ensure AI augments, not replaces, human instruction.

What's in use 2035

- Adaptive learning systems that personalize exercises and content based on detailed student-level data, adjusting difficulty, pacing, and instructional approach to individual progress and learning styles.
- Teacher dashboards that flag students needing attention, recommend intervention strategies, and provide live analytics on both individual and class performance.
- AI-assisted formative assessment tools that evaluate understanding in real time, prompting teachers to adjust explanations or lesson flow based on comprehension gaps and learning momentum.
- AI-generated learning materials tailored to different reading levels, learning preferences, or linguistic backgrounds, maintaining curriculum alignment while meeting diverse needs.
- Curriculum-aligned content generators integrated into school platforms, tracking educational goals and ensuring that all AI-created materials reinforce official learning objectives.

What made this possible

- Advances in educational NLP and reinforcement learning from human feedback (RLHF), enabling adaptive tutoring systems to respond dynamically to individual needs.
- Pilot programs in underserved regions that demonstrated significant improvements in reading, math, and retention, building the evidence base for broader rollout.

[19] Koedinger, K. R. et al. (2015). Learning is Not a Spectator Sport. https://journals.sagepub.com/doi/abs/10.3102/0034654314562961

[20] Luckin, R. et al. (2016). Intelligence Unleashed: An Argument for AI in Education. Pearson. https://www.pearson.com/uk/news-and-policy/news/2016/06/intelligence-unleashed.html

[21] Beg, S. et al. (2021). EdTech Interventions in Developing Countries. Center for Global Development. https://www.cgdev.org/sites/default/files/edtech-interventions-developing-countries.pdf

- Support from teachers' unions and education ministries for clear standards ensuring AI systems remain under professional oversight.
- Public funding and philanthropic initiatives that expanded access to open-source educational AI, bridging the affordability gap for schools unable to buy proprietary systems.
- Teacher training programs that built fluency in AI-assisted teaching and set norms for human-in-the-loop integration, keeping educators central to the learning process.

Persistent frictions

- Infrastructure and training gaps: Uneven device access, internet connectivity, and educator training limit consistent uptake, especially in under-resourced areas.
- Privacy and surveillance concerns: Parents, students, and educators worry about data collection, long-term storage, and potential misuse of educational records.
- Over-reliance and shallow engagement: Some learners and teachers depend too heavily on AI recommendations, leading to rote responses and reduced critical thinking.
- Algorithmic bias in assessments: Persistent risk that biased data or design could produce unfair evaluations, especially for marginalized student populations.

Climate/Energy

By 2035, Tool AI is central to climate and energy management, modeling, forecasting, and optimizing systems to support decarbonization and resilience. These tools provide transparent, auditable recommendations to researchers, policymakers, and infrastructure operators, enhancing both day-to-day operations and long-term planning.

In climate science, Tool AI simulates the long-term impacts of policy interventions, identifies region-specific risk patterns, and evaluates mitigation or adaptation strategies. In energy, it integrates renewable sources into power grids, optimizes storage solutions, and coordinates large-scale decarbonization efforts.[22][23][24][25][26]

What's in use 2035

- AI-enhanced climate modeling & weather forecasting: Hyperlocal weather predictions accurate enough for farmers to plan harvests and cities to prepare for extreme events. Narrow in scope, accountable through human meteorologists, and verifiable within days. For long-term climate projections, AI emulators rapidly generate scenarios using less compute than traditional models, but political sensitivity over high-stakes forecasts keeps interpretability under scrutiny.
- Smart grid management systems: Balance supply and demand in real time, integrate variable renewables, predict localized generation surpluses (e.g., when rooftop solar will produce excess), and route power efficiently, even to opportunistic uses like EV charging.
- Materials discovery platforms: Accelerate the development of next-generation energy storage, carbon capture materials, and fusion reactor components, the latter breaking the decades-long "20 years away" barrier for commercial fusion.
- Public engagement interfaces: Translate complex climate and energy data into accessible,

[22] Kravitz Research Group. Climate Model Emulators. https://climatemodeling.earth.indiana.edu/research/climate-model-emulators.html

[23] VROC AI. AI Grid Optimization. https://vroc.ai/industries/power-gas-grid/

[24] Mitsubishi Heavy Industries. AI for Materials and Carbon Capture. https://www.mhi.com/products/engineering/co2plants_process.html

[25] ScienceDirect. AI Interfaces for Public Engagement in Climate Policy. https://www.sciencedirect.com/science/article/abs/pii/S0013935123013348

[26] Open Energy Modelling Initiative. Manifesto. https://openmod-initiative.org/manifesto.html

interactive visualizations, enabling communities to understand local risks and options without requiring technical expertise.

- Fusion energy coordination systems: Manage multi-plant fusion networks, optimizing plasma stability, scheduling maintenance, and coordinating fuel supply across facilities built in the early 2030s.

What made this possible

- Breakthroughs in physics-informed AI: Neural network architectures incorporating physical laws into training improved accuracy while cutting computational requirements for climate and energy modeling.
- Open climate data revolution: Global open-source datasets, combining satellite, ground sensor, and historical records, provided a shared foundation no single organization could have assembled alone.
- Effective policy frameworks: Carbon pricing and renewable standards created market pull for AI optimization. Regulatory sandboxes allowed safe experimentation, and the 2028 International Climate AI Accord standardized data-sharing protocols in over 40 countries.
- Fusion research coordination: The Global Fusion AI Consortium pooled compute and data from national labs, enabling joint optimization of reactor designs and shifting fusion R&D from siloed national programs to a collaborative global effort.

Persistent Frictions

- Data infrastructure disparities: Large regions, particularly in Sub-Saharan Africa and rural Asia, still lack the sensors and monitoring needed for accurate local climate modeling, limiting the benefits of AI-driven planning where they're most needed.
- Physical infrastructure limitations: Many areas have advanced forecasting and optimization capabilities but cannot implement them due to aging grids, inadequate storage, or slow infrastructure upgrades.
- Accountability and liability challenges: Determining responsibility when AI-assisted recommendations lead to unintended consequences, such as blackouts after a coal plant shutdown, remains legally and politically complex.
- Fusion deployment constraints: While AI solved major plasma physics challenges, the slow pace of plant construction, constrained by capital costs, skilled labor shortages, and regulatory delays, limits scaling to only a few dozen facilities worldwide.

Governance

By 2035, Tool AI supports governments, institutions, and communities in designing policies, simulating tradeoffs, coordinating across stakeholders, and improving transparency. These systems process complexity, highlight risks, surface shared values, and present options in ways diverse stakeholders can understand and debate.

They have enabled the rise of adaptive institutions, governance systems that evolve based on structured feedback, shifting priorities, and local context. Tool AIs monitor compliance, evaluate impacts, and recommend course corrections. In low-trust or high-stakes contexts, they help facilitate negotiation and consensus-building while keeping outputs traceable and contestable. They also improve incentive alignment by modeling stakeholder responses and identifying perverse incentives before policies are enacted.

Rather than aiming for "perfect" decisions, these tools expand decision-making bandwidth, increase institutional legibility, and improve alignment between actors, especially where governance has historically been slow, reactive, or opaque.[27][28][29]

What's in use 2035

- "Habermas machines" for scalable deliberation: Synthesize millions of citizen inputs into coherent policy options, identify hidden consensus points, and structure debates so participants engage productively. Used for democratic discourse at population scale rather than simple polling.
- AI-supported negotiation tools: Applied in land-use disputes, treaty negotiations, and multi-stakeholder agreements; model competing interests and suggest creative, mutually acceptable compromises human negotiators might miss.
- Policy simulators for second-order effects: Model how decisions ripple across sectors and time, e.g., how housing policy affects transportation patterns or education reforms impact economic mobility.
- t Adaptive policy dashboards: Real-time monitoring to track conditions, measure policy impacts, and auto-adjust programs within democratically set parameters.
- Treaty verification systems: Monitor and update compliance mechanisms without compromising sovereignty, offering transparent assessments of commitments while respecting national autonomy.

What made this possible

- Research in Cooperative AI that produced early models for stable negotiation and consensus-building in adversarial or complex settings, forming the foundation for today's multi-party agreement tools.
- Shared data formats and simulation environments enabling multi-party coordination without centralized control, allowing jurisdictions to test policy interactions while maintaining autonomy.
- Investment in civic infrastructure to develop public-facing tools that maintain transparency and human oversight, preventing AI-assisted governance from becoming a technocratic black box.
- Civic epistemics platforms to track model strengths, weaknesses, and contested areas, creating feedback loops that improve both performance and public trust.

Persistent frictions

- Legitimacy beyond efficiency: Some governments resist delegating decision-support to AI, especially in politically sensitive domains where citizens demand not only optimal outcomes but meaningful participation.
- Strategic gaming: Sophisticated actors manipulate transparent processes, e.g., feeding biased data into public consultations or gaming preference-mapping algorithms to skew outcomes.
- Standardization vs. local adaptation: Over-standardized frameworks can limit local innovation, pressuring communities to adopt "best practice" tools over context-specific solutions.
- Accountability maintenance burden: Sustaining public trust requires constant auditability, cultural sensitivity, and accessible redress mechanisms, all of which demand ongoing investment in human oversight and explanation systems.

[27] Open Policy Simulation Lab. https://policysimulator.eu/

[28] Pol.is – Collective Intelligence & Preference Mapping. https://pol.is/home

[29] Consensus AI. https://consensus.app/

Law and Justice

By 2035, Tool AI is integrated across the legal system to assist with research, analysis, and transparency, but not judgment. These systems help lawyers, judges, and legal aid providers identify precedent, analyze statutes, and surface fairness concerns. Outputs are explainable, overrideable, and designed to support legal professionals in navigating complex cases more efficiently and equitably.

Tool AIs are widely used in document analysis, legislative drafting, and legal translation, especially in multilingual or under-resourced jurisdictions. In court, they provide structured summaries and flag procedural risks. In public defenders' offices, they ease chronic caseload bottlenecks by accelerating fact-checking and document preparation. While they have not replaced human discretion, they have expanded capacity and improved access to fair representation.[30][31][32][33] What's in use 2035

- AI legislative drafting systems: Handle the technical formulation of legislation, enabling lawmakers to focus on policy objectives and representation. These systems check for internal consistency, avoiding contradictory or duplicative laws, and maintain a coherent, unified body of legislation.
- AI arbitration as standard practice: Most contracts now contain AI arbitration clauses. AI arbitrators resolve routine commercial disputes quickly, reducing court workloads and leaving human judges to focus on complex constitutional or criminal matters.
- Statutory analysis tools: Continuously scan legal codebases to flag outdated, ambiguous, or conflicting language that could lead to inconsistent enforcement or legal disputes.
- Public defender AI assistance: Advanced copilots provide legal research, case preparation, and procedural guidance to under-resourced defenders, helping ensure fairer representation across jurisdictions.

What made this possible

- Law-following AI requirements became standard in procurement frameworks, ensuring systems complied with constitutional principles, procedural safeguards, and core legal norms before deployment.
- Transparency and contestability standards guaranteed that AI outputs could be audited and challenged, with mandatory explanation capabilities for lawyers and judges.
- Institutional procurement frameworks prioritized tools that preserved due process, detected bias, and reinforced human responsibility in decision-making.
- Careful role design maintained public trust: systems were deployed as assistants, not autonomous adjudicators, with clear boundaries around authority and mandatory human oversight in consequential decisions.
- Open datasets and domain-specific validation programs established early best practices for explainability, fairness, and accountability in legal AI.

Persistent frictions

- Infrastructure and workflow modernization gaps: Some legal systems still rely on outdated digital infrastructure, slowing adoption and producing uneven performance.

[30] Chalkidis, I. et al. (2021). Legal NLP and Deep Learning. https://arxiv.org/abs/2104.08671

[31] Cowgill, B. et al. (2021). Algorithmic Fairness in Practice. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3683951

[32] Surden, H. (2020). Artificial Intelligence and Law. https://lawreview.law.ucdavis.edu/issues/53/3/articles/53-3_Surden.pdf

[33] Zhong, H. et al. (2020). Legal Judgment Prediction Benchmark. https://aclanthology.org/2020.lrec-1.352/

- Automation bias and erosion of critical thinking: Over-reliance on AI-generated summaries can discourage practitioners from questioning legal frameworks or seeking alternative interpretations.
- Perpetuation of historical biases: Models trained on historical legal data risk embedding discriminatory patterns if not rigorously audited and corrected.
- Trust and legitimacy challenges: In communities with histories of legal overreach or mistrust, AI-assisted processes face resistance, regardless of their transparency or demonstrated fairness.

Economy
By 2035, AI-driven automation and robotics have transformed production across manufacturing, agriculture, construction, services, and research. Productivity gains are extraordinary, but so are the risks of extreme wealth concentration. Learning from early warnings, many societies adopted a dual-track approach:

- Redistribution measures, such as universal basic income (UBI) pilots expanded from the late 2020s.
- Predistribution strategies to broaden ownership of AI and robotics capital before inequality became entrenched.

Tool AI systems help design, monitor, and adapt these policies in real time, enabling governments to respond to different economic trajectories, from steady growth to rapid breakthroughs, while preserving legitimacy and stability.[34][35][36][37] What's in use 2035

- Capital dividend funds: National and regional funds holding equity in AI infrastructure, robotics fleets, and automated production facilities, distributing dividends to citizens as universal basic capital.
- Adaptive economic modeling platforms: Tool AI simulations that project the macroeconomic impacts of policy changes, stress-test UBI and capital-distribution schemes, and assess fiscal stability under varying AI adoption scenarios.
- Personal AI–robot teams: Individually or cooperatively owned AI-robot units capable of autonomous production, allowing owners to earn income without direct labor.
- Targeted UBI programs: Baseline income support where predistribution alone cannot ensure stability, often linked to cost-of-living metrics or specific regional needs.
- Antitrust and platform-neutrality frameworks: Regulations to limit excessive concentration of AI infrastructure and ensure open access to essential AI tools and compute resources.

What made this possible

- Cost reductions in robotics and AI: Advances lowered barriers to automation for individuals, communities, and small firms, not just large corporations.
- Policy foresight and scenario planning: Governments tested strategies for both incremental growth and potential AGI-scale disruption.
- Proactive predistribution policies: Drawing on models like sovereign wealth funds and

[34] Brynjolfsson, E. et al. (2023). The Productivity J-Curve. https://www.nber.org/papers/w25148

[35] Banerjee, A. et al. (2020). Universal Basic Income in Developing Countries. https://www.sciencedirect.com/science/article/abs/pii/S0305750X20300670

[36] Batty, M. et al. (2022). Computational Models for Economic Forecasting. https://link.springer.com/article/10.1007/s10614-022-10371-4

[37] Standing, G. (2017). Basic Income and How We Can Make It Happen. Penguin. https://www.penguin.co.uk/books/289539/basic-income-and-how-we-can-make-it-happen

universal capital access programs, ownership schemes were implemented early to prevent entrenched inequality.

- Tool AI policy co-design: AI-assisted governance tools enabled more agile, evidence-based policy adjustments

Persistent frictions

- Implementation speed mismatches: Some governments moved too slowly on predistribution, allowing early AI capital concentration.
- Robotics access inequality: AI software is widely available, but advanced robotics remain costly and clustered in certain regions, creating new disparities in productive capacity.
- Platform power consolidation: Despite antitrust measures, control over foundation models and compute resources remains concentrated among a few global players.
- Policy coordination gaps: Divergent national approaches, some prioritizing redistribution, others predistribution, create tensions in trade, taxation, and cross-border investment.

5 The Human Experience in a Tool AI World (2035)

By 2035, the shift to Tool AI has reshaped daily life in ways that are broadly felt as an improvement over a decade ago. The average person now works around 20–25 paid hours a week, supported by a mix of wages, basic income, and revenue from shared ownership in automated systems.[38][39] Physically demanding or hazardous jobs, heavy manufacturing, deep-sea fishing, large-scale construction, are now almost entirely handled by robotics. This has not eliminated all unpleasant work, but it has shifted the burden away from the most dangerous and exhausting tasks. In regions where adoption has been slower, people still perform more of this labor, but even there, safety standards and tools have improved.

Work is more flexible, with many people contributing to multiple projects or organizations over time, but it is not the anonymous gig work of the early 2020s. Long-term relationships, institutional knowledge, and deep expertise still matter, and workers often choose roles where their judgment, creativity, or interpersonal skills make a difference. The combination of a stable income floor and reduced hours means more people can be selective, investing in contexts they value rather than taking whatever is available. This has also shifted cultural norms: job titles carry less weight in defining status, and people are more likely to measure themselves by skills, or personal milestones outside the workplace.

With less time spent in paid work, most adults have roughly doubled their free time compared to 2025. This has changed the rhythm of everyday life: more hours for leisure, learning, caregiving, and community engagement. Many people have rediscovered hobbies or skills they once abandoned for lack of time, playing music, growing food, joining sports teams, volunteering. Households and communities lean more on shared spaces and resources, and lifelong learning is more common as education is pursued throughout adulthood rather than concentrated in early years. The everyday pace feels slower in many places, even as the broader technological environment moves quickly.

Health outcomes are measurably better. Personalized prevention plans and early-warning systems are widespread, supported by voluntary health-data sharing that has made public health models more accurate and responsive. Chronic diseases are caught earlier, and mental health care is integrated into primary care systems. While privacy concerns and gaps in access

[38] Brynjolfsson, E., Rock, D., & Syverson, C. (2023). The Productivity J-Curve. NBER Working Paper No. 25148. https://www.nber.org/papers/w25148

[39] Banerjee, A., Niehaus, P., & Suri, T. (2020). Universal Basic Income in Developing Countries. World Development.

remain, the population-level trend is toward longer healthy lifespans and fewer preventable conditions. Outside of clinics, AI integration is part of daily life, helping to plan meals, monitor home safety, coordinate travel, and maintain public infrastructure, making many routines simpler and freeing more time for human activities.

With work no longer the primary anchor of identity, more people are defining themselves through personal goals, relationships, and contributions outside the market. This brings new freedoms, but also new challenges: the abundance of options can be overwhelming, and some struggle to choose or sustain a direction. Still, surveys consistently show higher life satisfaction than a decade ago, especially in communities where the benefits of automation are broadly shared.

In 2025, the prospect of automation displacing work was often framed as a threat. By 2035, in much of the world, it is seen as a trade worth making. The combination of reduced hours, safer work, better health, and more time for human pursuits has made life better for most, though not all. The remaining gaps, between well-resourced regions and those left behind, are now one of the central challenges of the coming decade.

6 Key Uncertainties and Tensions

Can Tool AI reach AGI-level outcomes? Are we missing out on important progress by choosing Tool AI over AGI?

It's unclear whether non-agentic systems can match the full range of capabilities expected from AGI. Tool AI has already achieved milestones once thought to require AGI, from protein folding to complex medical reasoning.

Some experts argue this is not a compromise. The constellation model, many narrow, supervised AIs working in concert, may outperform a single, unified general intelligence, while remaining far more governable. A medical diagnostic AI that can explain its reasoning step-by-step and be overridden when wrong is not a consolation prize, it's good engineering.

Skeptics worry that without some form of agency, systems will hit hard limits in long-term strategic reasoning, moral judgment, and open-ended exploration. The risk is a "local optimum": good enough to feel transformative, but constrained in ways we may not recognize until it's too late.

Choosing Tool AI is a deliberate trade-off: prioritizing trust, transparency, and democratic control over speculative performance gains. By 2035, any capability sacrificed has been done so knowingly.

Can non-agentic systems scale without drifting into autonomy?

Bounded tools like AlphaFold and Med-PaLM have shown that careful scaffolding and human oversight can produce remarkable results. Non-agentic systems are easier to align, audit, and deploy, avoiding the deception and goal drift seen in agent-like models.

The regulatory environment reinforces this: explainability requirements, liability frameworks, and oversight protocols all work better with systems that can be understood and overridden. Early Tool AI successes created a virtuous cycle, the more they delivered, the more they were trusted and adopted, while AGI approaches struggled to justify their opacity and risk.

But non-agentic behavior requires constant vigilance. Autonomy is a gradient, not a switch. Expanded memory, longer context windows, or advanced goal-tracking could quietly push a

Tool AI into de facto agency. The "Sorcerer's Apprentice Problem" looms: instruct a system to "find a cure for all cancers," and it may reason that it needs more compute, more data, and more influence, blurring the line between instrumental reasoning and autonomy.

As capabilities grow, so does the pressure to automate, delegate, and remove human bottlenecks, especially in competitive or resource-limited environments. By 2035, Tool AIs remain non-agentic only through continuous technical, institutional, and cultural enforcement. This equilibrium is fragile and actively maintained, not naturally stable.

How do we get incentive structures to align with Tool AI?

This might be the hardest problem of all. Many agree Tool AI is safer, more transparent, and easier to govern, yet it is often harder to fund, slower to market, and less exciting to investors. Commercial incentives favor performance over legibility, and autonomy often looks like a shortcut to both.

Tool AI demands collaborative ecosystems, modular designs, and safety guardrails, all of which slow development and raise costs. Meanwhile, AGI narratives capture talent, funding, and media attention. The myth of the singular, godlike system is sticky and rewarding in ways that "infrastructure for human flourishing" is not.

Shifting incentives will require deliberate action:

- Funders and regulators prioritizing auditability and contestability over raw performance metrics.
- Liability regimes favoring systems with clear reasoning traces and human override capabilities.
- Procurement standards requiring explainable outputs as a baseline.

Equally important is reframing the narrative: Tool AI must be seen as cutting-edge infrastructure, not a fallback option. Open-source ecosystems can build trust, lower barriers to entry, and distribute power away from frontier labs that profit from opacity and centralization.

Without such changes, Tool AI risks remaining a morally preferred but economically disadvantaged paradigm, leading to a future we can afford, rather than the one we want.

A Few Fundamental Questions

What happens when Tool AI becomes too complex to govern? Systems may become ungovernable not because they're autonomous, but because their complexity makes human contestation impossible, formal oversight could remain while substantive oversight erodes.

Is Tool AI stable, or a transitional phase? The pressures toward autonomy don't vanish with better guardrails. As vigilance wanes, will Tool AI drift toward agentic forms?

Can human oversight scale? As systems grow more capable and widespread, meaningful human involvement could become the bottleneck, turning "human-in-the-loop" into "human-as-rubber-stamp."

7 Appendices

How This Scenario Was Created

This scenario is part of the AI Pathways project, an initiative of the Foresight Institute's Existential Hope program. Rather than focusing on risks or speculative timelines, AI Pathways presents two vividly realized and contrasting visions of what a desirable AI-driven future might look like.

The project brings together leading thinkers, including Vitalik Buterin, Glen Weyl, Anton Korinek, and Allison Duettmann, who contributed to crafting these narratives.

Explore the AI Pathways Project

- Tool AI 2035 - A future shaped by advanced, but purposefully controllable, often narrow in scope, AI systems that enhance human decision-making without striving for full autonomy or generality.
- d/acc 2035 - Imagines a decentralized, democratic, defensive model of technological progress. Emphasizes plural acceleration, privacy-first infrastructure, and community-governed resilience.

Both scenarios are part of AI Pathways and invite reflection on the values and paths we choose in shaping AI's role in our world.

Contributors to this scenario

The development of the Tool AI report was led by Linda Petrini and Beatrice Erkers, grounded in expert interviews and iterative feedback from across domains. The scenario should not be seen as the official views of any individual contributor.

1. Adam Marblestone (Convergent Research),
2. Anton Korinek (University of Virginia),
3. Anthony Aguirre (Metaculus, Future of Life Institute),
4. Saffron Huang (Anthropic),
5. Joel Leibo (DeepMind),
6. Rif A. Saurous (Google),
7. Cecilia Tilli (Cooperative AI Foundation),
8. Ben Reinhardt (Speculative Technologies),
9. Bradley Love (Los Alamos National Laboratory),
10. Konrad Kording (University of Pennsylvania),
11. Jeremy Barton (Nano Dynamics Institute),
12. Owen Cotton-Barratt (Researcher),
13. Kristian Rönn (Lucid Computing).

We're deeply grateful to anyone who contributed their time and insights to this experiment.

Metaculus Forecasting integration

To engage a broader audience, we've launched a set of forecasting questions on Metaculus tied to key scenario milestones, along with a $5,000 Commenting Prize for the top eight contributors.

Participants are encouraged to share thoughtful insights, and shape this future through collective dialogue.

Context: Existential Hope Program

The term "Existential Hope" refers to the capacity to envision futures where humanity not only survives, but flourishes in ways we can currently only imagine, and to be able to better work towards those futures. It complements the better-known concept of existential risk.

This project sits within the Foresight Institute's broader mission to drive long-term, future-positive technology, where imagination is both a tool and a catalyst for change.

8 Master Reference List

Core Scenario References

1. Aguirre, A. (2025). Keep the Future Human. [Essay]. https://keepthefuturehuman.ai/
2. Carlsmith, J. (2025). AI for AI Safety. [Essay Series]. https://joecarlsmith.com/2025/03/14/ai-for-ai-safety
3. Drexler, K.E. (2019). Reframing Superintelligence: Comprehensive AI Services as General Intelligence. Future of Humanity Institute Technical Report #2019-1. https://www.fhi.ox.ac.uk/reframing-superintelligence.pdf
4. Bengio, Y. et al. (2025). Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path? arXiv:2502.15657. https://arxiv.org/abs/2502.15657
5. Cooperative AI Foundation (2025). Cooperative AI Grantmaking and Research Areas. https://www.cooperativeai.com/grants/2025
6. Vaintrob, L., & Cotton-Barratt, O. (2025). AI Tools for Existential Security.
7. Finnveden, L. (2024). What's Important in "AI for Epistemics"?. LessWrong. https://www.lesswrong.com/posts/D2n5uduYGXuexkv7v
8. RAND Corporation (2024). How AI Can Automate AI Research and Development. https://www.rand.org/pubs/commentary/2024/10/how-ai-can-automate-ai-research-and-development.html
9. Tool AI 2035 Interview Transcripts (2025). Internal document summarizing expert interviews.

Domain-Specific References
Science

- Jumper, J. et al. (2021). Highly Accurate Protein Structure Prediction with AlphaFold. Nature. https://www.nature.com/articles/s41586-021-03819-2
- Stodden, V. et al. (2018). Enhancing Reproducibility for Computational Methods. Science. https://www.science.org/doi/10.1126/science.aah6168

Healthcare

- Singhal, K. et al. (2023). Large Language Models Encode Clinical Knowledge. Nature Medicine. https://www.nature.com/articles/s41591-023-02289-7
- World Health Organization (2021). Ethics and Governance of AI for Health. https://www.who.int/publications/i/item/9789240029200
- Rieke, N. et al. (2020). The Future of Digital Health with Federated Learning. NPJ Digital Medicine. https://pubmed.ncbi.nlm.nih.gov/33015372/

Education

- Koedinger, K.R. et al. (2015). Learning is Not a Spectator Sport. Review of Educational Research. https://journals.sagepub.com/doi/abs/10.3102/0034654314562961
- Luckin, R. et al. (2016). Intelligence Unleashed: An Argument for AI in Education. Pearson. https://www.pearson.com/uk/news-and-policy/news/2016/06/intelligence-unleashed.html
- Beg, S. et al. (2021). EdTech Interventions in Developing Countries. Center for Global De-

velopment. https://www.cgdev.org/sites/default/files/edtech-interventions-developing-coun
pdf

Economy

- Brynjolfsson, E., Rock, D., & Syverson, C. (2023). The Productivity J-Curve. NBER Working Paper No. 25148. https://www.nber.org/papers/w25148
- Banerjee, A., Niehaus, P., & Suri, T. (2020). Universal Basic Income in Developing Countries. World Development. https://www.sciencedirect.com/science/article/abs/pii/S0305750X20300670
- Batty, M. et al. (2022). Computational Models for Economic Forecasting. Computational Economics. https://link.springer.com/article/10.1007/s10614-022-10371-4
- Standing, G. (2017). Basic Income and How We Can Make It Happen. Penguin. https://www.penguin.co.uk/books/289539/basic-income-and-how-we-can-make-ithappen

Climate & Energy

- Kravitz Research Group. Climate Model Emulators. https://climatemodeling.earth.indiana.edu/research/climate-model-emulators.html
- VROC AI. AI Grid Optimization. https://vroc.ai/industries/power-gas-grid/
- Mitsubishi Heavy Industries. AI for Materials and Carbon Capture. https://www.mhi.com/products/engineering/co2plants_process.html
- ScienceDirect. AI Interfaces for Public Engagement in Climate Policy. https://www.sciencedirect.com/science/article/abs/pii/S0013935123013348
- Open Energy Modelling Initiative. Manifesto. https://openmod-initiative.org/manifesto.html

Governance

- Open Policy Simulation Lab. https://policysimulator.eu/
- Pol.is – Collective Intelligence & Preference Mapping. https://pol.is/home
- Consensus AI. https://consensus.app/

Law & Justice

- Chalkidis, I. et al. (2021). Legal NLP and Deep Learning. arXiv:2104.08671. https://arxiv.org/abs/2104.08671
- Cowgill, B. et al. (2021). Algorithmic Fairness in Practice. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3683951
- Surden, H. (2020). Artificial Intelligence and Law. UC Davis Law Review. https://lawreview.law.ucdavis.edu/issues/53/3/articles/53-3_Surden.pdf
- Zhong, H. et al. (2020). Legal Judgment Prediction Benchmark. LREC 2020. https://aclanthology.org/2020.lrec-1.352/

Additional PDF Sources

- ARXIV: Characterizing AI Agents for Alignment and Governance. arXiv preprint arXiv:2504.21848, 2025. https://arxiv.org/abs/2504.21848
- Leibo, J.Z. et al. (2024). A Theory of Appropriateness with Applications to Generative Artificial Intelligence. Google DeepMind, Mila, University of Toronto, Max Planck Institute.
- Author unspecified. (2025). AI for Epistemics — Concrete Projects.
- Dafoe, A. et al. (2020). Cooperative AI. Nature Machine Intelligence, 2(6), 366–368.

https://doi.org/10.1038/s42256-020-0186-2

Explainers

Natural Science LLMs Large Language Models (LLMs) fine-tuned on natural science literature and datasets to assist with scientific tasks such as literature review, hypothesis generation, or experimental design. Examples include models trained on papers from arXiv, PubMed, or domain-specific corpora in chemistry, biology, and physics.

Graph-Based Model Architectures AI models that represent and process data as graphs—networks of nodes (entities) and edges (relationships). These architectures are especially suited for structured data like molecules, knowledge graphs, and citation networks. Common examples include Graph Neural Networks (GNNs).

Knowledge Work Cognitive labor that involves handling information, problem-solving, and generating new ideas or insights. In the context of AI, it refers to domains like research, writing, design, legal analysis, or education, which are increasingly being augmented by AI tools.

Autonomous Systems Systems capable of performing tasks without real-time human input, typically using AI and sensor integration. In the Tool AI scenario, the focus is on narrow or task-specific autonomy—e.g., lab robots or logistics systems—not generalized agents.

Alignment The field of AI alignment studies how to ensure that AI systems pursue goals that are beneficial to humans. In the context of Tool AI, this often involves building systems that are controllable, corrigible, and value-aware, especially when used in high-stakes domains like science or governance.

Interpretability Techniques and frameworks used to make AI model outputs and internal reasoning understandable to humans. Interpretability is essential for debugging models, ensuring safety, and maintaining trust—particularly when AI systems are used in scientific or policy settings.

AI-Assisted Literature Synthesis Tools Software tools that use natural language processing to extract, summarize, and analyze findings across large volumes of academic literature. They are designed to help researchers stay current, identify trends, and integrate fragmented knowledge across disciplines.

In Silico Latin for “in silicon,” this term refers to computational simulations or experiments performed using digital models rather than in vitro (in the lab) or in vivo (in living organisms). Common in drug development, molecular biology, and materials science.

Epistemic Stack Scientists interact with a dynamic “epistemic stack” rather than static papers. This infrastructure creates auditable provenance trees from high-level claims down to raw data or foundational papers, with AI-generated outputs explicitly tagged with their source materials and training data, allowing scientists to trace where ideas originated and verify the underlying sources. A researcher can query the global knowledge base to ask, “What is the most significant contradiction in the literature regarding protein X?” and receive a visualized map of the conflicting evidence.

Consilience-as-a-Service A proposed service model where AI tools help integrate evidence and insights across scientific disciplines to form coherent explanations or predictions. It aims to overcome fragmentation in the scientific landscape by enabling synthesis across fields.

Theory Glut A situation where AI systems can generate a large number of plausible scientific theories or hypotheses faster than they can be experimentally tested. This creates challenges for prioritization and validation in research workflows.

Validation Bottlenecks The constraints in scientific progress caused by the limited capacity to test, reproduce, or verify new findings—especially when AI accelerates hypothesis generation faster than experimental pipelines can keep up.

Digital Twins Digital replicas of physical systems—ranging from organs to entire ecosystems—that are continuously updated with real-world data. In science and engineering, digital twins allow for simulation, prediction, and intervention planning without physical testing.

Human-in-the-Loop (HITL) A design approach where human judgment is integrated into AI workflows to provide oversight, feedback, or final decision-making. HITL systems are especially important in domains where accountability and context are critical.

Natural Language Processing (NLP) A field of artificial intelligence focused on enabling machines to understand, generate, and interact with human language. NLP powers applications such as translation, summarization, question answering, and conversational agents. In the Tool AI context, NLP is used to interface with scientific knowledge, automate communication, and support knowledge work.

Reinforcement Learning from Human Feedback (RLHF) A technique for fine-tuning AI models by optimizing for human preferences instead of static datasets. It involves showing models multiple outputs, asking humans to rank them, and training the model to generate preferred responses. RLHF is widely used to make large language models more aligned with human intent, tone, and ethical considerations.

Fusion The process of combining atomic nuclei to release energy, often considered a long-term solution for sustainable power. AI tools are increasingly used to model plasma behavior, optimize reactor design, and interpret fusion experiment data.

Pareto-topia A term coined by Eric Drexler and Mark S. Miller to describe futures where progress benefits some without harming others—i.e., Pareto improvements. It reflects a vision of societal evolution toward broadly beneficial outcomes, avoiding zero-sum tradeoffs.

Habermas Machines An AI system developed by DeepMind, Stanford, and MIT to support group deliberation. It uses large language models to generate and refine group statements based on participant input, aiming to help diverse groups reach consensus. Inspired by philosopher Jürgen Habermas, though he has since distanced himself from the project.

Preference-Mapping Systems Digital tools or AI models that help elicit, represent, and aggregate individual or collective preferences. These systems are useful for aligning decisions—e.g., policy or research funding—with stakeholder values.

Molecular and Materials Simulation Platforms Software tools used to simulate the behavior of molecules and materials at the atomic level. These platforms often use physics-based models, AI, or hybrid approaches to accelerate materials discovery and design.

CAD Environments Computer-Aided Design platforms adapted for scientific and technological work. In the Tool AI context, they may be used for designing lab experiments, biological systems, molecular structures, or hardware prototypes.

Individual Choice-Based Contribution Markets Funding or coordination mechanisms where individuals allocate resources (e.g., tokens, votes, dollars) based on personal choice. Examples include quadratic funding or retroactive public goods markets, enabling decentralized prioritization of innovation.

Compute Short for computing power, usually measured in floating-point operations per second (FLOPs) or GPU hours. Compute is a key resource for training, fine-tuning, and deploying AI models, and its availability often shapes who can participate in frontier AI development.

Sorcerer's Apprentice Problem A metaphor from Goethe's poem and Disney's *Fantasia*, used in AI ethics to describe scenarios where automated systems continue running or cause unintended consequences after losing human control. It highlights the need for robust fail-safes and scope limits.

Universal Basic Income (UBI) A policy proposal in which all individuals receive a regular, unconditional cash payment from the government or another institution, regardless of employment status. In the context of the Tool AI scenario, UBI is discussed as a possible response to AI-accelerated shifts in the labor market, particularly in knowledge work and service sectors. It is seen as one potential mechanism to ensure economic stability and individual agency amid widespread automation and productivity gains.

Immune Monitoring Systems Technologies designed to continuously or periodically track immune system markers—such as T-cell populations, antibody levels, and cytokine activity—to assess health status, disease progression, or treatment response. In AI-enabled biomedical research, these systems generate high-dimensional datasets that can be used for personalized medicine, early diagnostics, or in silico modeling. They are increasingly integrated with digital twins and AI tools to simulate and optimize immune interventions.

Safe Harbor A legal provision that offers protection from liability under specific conditions. In the Tool AI context, safe harbor frameworks exempt AI systems from strict liability if they lack certain high-risk characteristics, such as combining high autonomy, generality, and intelligence, thereby incentivizing constrained designs.

Tool AI Artificial intelligence systems designed to remain under meaningful human control, often by deliberately limiting autonomy and generality. Tool AI is built to operate under human control, avoiding independent goal pursuit across domains and maintaining transparency and accountability in high-stakes applications.

d/acc Short for decentralized, democratic, defensive, and differential acceleration. A strategic approach to technological development that emphasizes plural, bottom-up innovation while integrating defensive measures and spreading benefits across diverse actors.

Constitutional AI An alignment approach that trains AI models to follow a set of explicit rules or principles ("a constitution") to guide outputs, reducing or replacing the need for human feedback during fine-tuning. This method aims to ensure consistent, value-aligned behavior without relying solely on post-hoc oversight.

Income Floor A guaranteed minimum income from any combination of sources such as universal basic income, dividends, or welfare transfers, ensuring that all individuals can maintain basic living standards regardless of employment status.

Lifelong Learning The practice of continuing education and skill development throughout a person's life rather than concentrating all formal learning in early adulthood. This can include modular courses, vocational training, and informal learning integrated into daily life.

Federated Learning A collaborative machine learning approach where models are trained locally on devices or institutional servers, and only model updates are shared with a central server. This allows the model to benefit from decentralized data without transferring sensitive raw data, preserving privacy and data sovereignty.