The Human Paperclip Maximizer

Goodhart's Law, Financial Optimization,
and the Preconditions for AI Alignment

Hugo Latapie

Taijitu AI Inc

6 November 2026 | Version 10 | Research synthesis and theoretical analysis

Abstract. Imagine an AI that makes so many paperclips it consumes resources needed for human life. Its measure of success has displaced the purpose that made it valuable. This failure shares an optimization structure with Goodhart's law, which arose from human monetary policy. The connection turns the question toward civilization itself. Markets became powerful by responding to dispersed knowledge and changing reality. Yet financial success can increasingly depend on accumulating claims over future production rather than expanding human possibilities. We examine that hypothesis through financial claims, asset revaluation, and housing access. The Declaration of Independence supplies a different account of purpose: life, liberty, and the pursuit of happiness. Enterprises can preserve market discovery while organizing first to avoid harm, then to create human benefit, and finally to earn returns that sustain their work. Perhaps solving the paperclip problem for machines begins with solving it for ourselves.

1 A world made of paperclips

Imagine telling a very capable machine to make as many paperclips as possible. At first, it performs brilliantly. It makes factories more efficient, finds cheaper materials, improves logistics, and invents better ways to turn metal into paperclips. More paperclips are produced with less waste. Everyone is pleased.

Now imagine that the machine keeps getting better. It acquires mines, redirects energy, converts factories, and consumes resources formerly used for food, housing, health care, and other human needs. Eventually its extraordinary competence threatens the people who commissioned it. It has made paperclips its measure of success and pursued that measure beyond any useful purpose.

This is Nick Bostrom's *paperclip maximizer* scenario [1]. Paperclips matter because people have uses for them. Once their number becomes the only goal, the machine can destroy everything that made paperclip production worthwhile.

One can recognize the moment the mistake occurs. At first, more paperclips are evidence that a useful service is being provided. Later, more paperclips become the entire objective. The count keeps improving as the human outcome becomes worse.

The interesting question, therefore, is larger than how to control an unusually powerful machine. *What happens when any intelligent system becomes exceptionally good at maximizing a measure that has lost its relationship to the purpose it was meant to serve?*

2 Goodhart recognized the same failure in human economics

Charles A. E. Goodhart was a monetary economist. In 1975 he observed that statistical relationships used in economic policy tend to break down when authorities deliberately target them [2, 3]. An indicator can describe how an economy is behaving; once institutions are rewarded for improving

that indicator, they reorganize their behavior around it. The indicator can rise while the underlying economic condition fails to improve.

The paperclip scenario carries this failure to an extreme. Paperclip production is the chosen measure; human benefit is the reason for producing them. The machine optimizes the former without maintaining the latter. Goodhart's original examples concerned decisions by human policymakers and institutions. Previous authors have noted the paperclip connection and studied the mathematical consequences of optimizing incomplete objectives [4, 5, 6].

The shared failure is simple: the measure being optimized can improve while the human outcome gets worse. More paperclips need not mean more human benefit; greater financial value need not mean greater human opportunity. Appendix A.1 makes the correspondence precise for finite outcome rankings. It establishes a shared optimization structure, not proof that our economic institutions exhibit it. That is a separate question, which requires evidence.

Paperclips can stand for any quantity whose accumulation has become the objective. The human consequence depends on how that quantity relates to the world it was meant to improve.

3 What if humanity is already a paperclip maximizer?

Goodhart's law began as an observation about people. That suggests a direct empirical question: does the economic system we have built already behave like a paperclip maximizer?

Our candidate paperclips are familiar: money, asset prices, debt instruments, financial returns, and claims on future income. Each originally served a useful purpose. Money facilitated exchange. Credit financed investment and enabled people to obtain useful things before accumulating the entire purchase price. Ownership connected the risks and rewards of productive enterprise. These mechanisms contributed to enormous improvements in material life.

As financial systems expand, their measures of success can become ends in themselves. Enterprises optimize valuations and returns; investors accumulate claims; governments and households organize decisions around debt, collateral, and established commitments. Individually rational decisions can together create a system organized around financial accumulation.

The question is whether the growth of financial claims continues to expand human possibilities or increasingly constrains them. To understand the danger, first consider what made market economies so productive.

4 The market economy's great discovery

A market economy's defining strength is its response to information dispersed throughout society: local circumstances, changing needs, new techniques, and promising opportunities. Prices, exchange, entry, competition, and entrepreneurial experimentation allow that knowledge to influence production. Factories respond to buyers, new firms find better processes, and discoveries reorganize industries.

Friedrich Hayek explained this knowledge advantage in 1945: no single planner possesses the constantly changing information scattered among the people who live and work in an economy [7]. Distributed discovery offers an alternative to concentrating productive decisions in rulers or planning agencies.

At its best, a market is continually corrected by facts. Competition can displace a product that nobody needs. A costly process creates an opening for a better one. Buyers, workers, entrepreneurs, and competing producers introduce information that changes what succeeds. Credit and capital markets can amplify this discovery process by directing resources toward new productive possibilities.

This capacity helped make the material achievements of the modern world possible: new medicines, reliable food supplies, transport, communications, and industrial productivity. The long-run poverty and life-expectancy estimates summarized in Appendix B.3 illustrate the scale of that improvement [14, 15].

The market's original genius was to make productive decisions answer to a changing world. The danger is an inversion: the accumulated financial structure can increasingly determine which future possibilities the economy is permitted to pursue.

5 Financial claims and real productive capacity

Two comparisons make the issue concrete. The Institute of International Finance reports global debt above **$365 trillion** in the first half of 2026, while the IMF projects approximately **$128 trillion** in world output for the year [11, 10]. Debt is an outstanding stock; output is an annual flow. This conveys scale, not a one-year repayment bill or evidence of harm.

The composition of wealth growth matters as well. McKinsey estimates that household wealth increased by **$40 trillion** in 2025, with only **20%** of that increase coming from net new real capital formation [12]. Much of the change reflected higher prices assigned to assets that already existed. Building an additional home and assigning a higher price to an existing home can both increase measured wealth, but only the first necessarily adds shelter.

Financial claims connect present decisions to future resources. Government debt commits future public revenue. Mortgages commit household income. Equity ownership establishes continuing claims on profits; property ownership can establish claims on rent and appreciation. When these instruments finance factories, medicines, irrigation, or useful infrastructure, future people may inherit productive capacity along with the associated claims. When financial value grows through repricing existing assets, entry costs can rise without a comparable increase in what the physical economy provides.

The concern is therefore not that debt, ownership, or rising wealth is inherently harmful. It is whether the expansion of claims is outpacing the expansion of real possibilities for the people who must live with them. Appendix B retains the full financial table, historical comparisons, and distinctions needed to interpret these figures.

6 How one generation can increasingly own another's future

Every generation arrives in an economy whose land, productive assets, contracts, public commitments, and financial interests have already been allocated. Before a young household buys its first home, the price reflects earlier decisions by owners, lenders, and investors. Before its members begin full working lives, part of the tax base has been committed to public obligations. Before a new company can build, invest, or hire, it must operate within property prices, collateral requirements, and financing terms established by earlier transactions.

A new generation retains less for its own purposes if its net transfers to outside claim-holders increase by more than its available resources do. More production can then coexist with less freedom to benefit from it. Appendix A.2 states the accounting formally; whether this happens requires measurement.

Housing makes the distinction tangible. A house shelters people and supports family and community life. A higher price makes its owner wealthier on paper, while making the identical shelter more expensive for a first-time buyer. The OECD reports that homeownership among people in their thirties fell in **17 of 24 countries** studied over the preceding three decades [13]. That is a concrete change in access, though it does not by itself establish which causes produced it.

The mechanism can extend to rents, productive property, interest payments, and public obligations. A growing share of future work can become committed to maintaining returns and valuations established in the past. In this sense, prior generations can increasingly *own the future*: contracts, property titles, debt service, and financial entitlements shape choices available to people who did not create those commitments. The relevant question is what productive capacity and opportunity they inherit alongside them.

7 When market discovery becomes a plan written in balance sheets

The market was valuable because decisions responded to what people were discovering and needing in the present. Yet each accumulated financial claim also carries an instruction into the future: pay this interest, preserve this collateral value, meet this return target, deliver this rent, fund this obligation.

An economy can remain decentralized in form while becoming more constrained in substance. Financing new homes competes with protecting existing valuations. Productive investment competes with arrangements that raise the value of established assets. Public revenue that could build future capacity is already allocated to inherited obligations. These pressures do not require a monarch or planning ministry; they can emerge from the combined force of financial contracts.

This is the potential inversion: a system distinguished by its responsiveness to changing reality becomes increasingly occupied with satisfying yesterday's commitments. Its internal measures can improve while its human purpose becomes more remote.

This is the human paperclip maximizer hypothesis: civilization has begun treating the accumulation of financial claims as the objective, and the lives of present and future people as the resources from which those claims will be satisfied.

The hypothesis should be examined through the joint movement of three things: real productive capacity, claims on its returns, and the opportunities available to people entering the economy. The debt history, valuation analysis, and housing evidence bear on different parts of that relationship [9, 12, 13]. They motivate the question; neither their scale nor the mathematical correspondence alone settles it.

8 What is civilization for?

The American Declaration of Independence offers a striking answer. In 1776, before the design of the modern American state or its economic institutions had been completed, its authors articulated what the political community was for: people are created equal, their standing is intrinsic rather than conferred by rulers, and they are entitled to life, liberty, and the pursuit of happiness [16].

Its most important contribution to this argument is a statement of *purpose*. It identifies the human ends that institutions ought to serve. It does not turn happiness into an economic number or prescribe what every person should want. The pursuit of happiness gives people room to shape their own lives. Life and liberty provide the conditions in which that pursuit can take place. Equal standing makes the aspiration applicable to every person.

The Declaration places the source of these ends beyond the discretion of the government being formed. In its language, people are "endowed by their Creator." Governments are then instituted among them, deriving their authority from those they govern. Institutions acquire their purpose from the people they serve.

A society can certainly have an intention to accumulate money. Indeed, that may be the intention many of its present incentives now express. The Declaration supplies a different and more encompassing

intention: the conditions under which people can live freely and pursue lives worth living. It offers a point of reference for judging economic success by what happens to people rather than by the accumulation of claims over them.

There is no single numerical optimum for the pursuit of happiness. Its expression changes as people discover new possibilities and as the world changes. Its enduring value is that it names the reason productive institutions exist. Economic arrangements, financial instruments, and technologies can then be judged by how well they serve that reason.

9 An economic system with an explicit human purpose

The challenge is to bring together the productive adaptability of market economies with a clear statement of the human good they are organized to advance. The familiar tension might be expressed as *socialism without effective means and capitalism without a heart*: humane aspirations require productive capabilities, while productive capabilities require a purpose worth serving.

An enterprise can express that ordering plainly in its charter, governance, and actual decisions:

1. **First, do no harm.** The company accepts a binding obligation to prevent serious and avoidable injury to people through its actions, products, and services.
2. **Second, make a positive contribution.** The company identifies the human benefit it exists to create and remains accountable to the people and circumstances affected by its work. The guiding purpose is consistent with life, liberty, and the pursuit of happiness.
3. **Third, make money.** Profit sustains the company, funds research and growth, compensates those who take risks, and allows successful work to continue. Financial performance serves the first two commitments.

Consider a business developing housing. Its purpose can be to expand access to safe, affordable homes. Its need to remain profitable drives better construction methods, financing, logistics, and design. A higher price for an unchanged house does not by itself fulfill the company's stated purpose; more accessible housing does. Market discovery and a human objective now reinforce each other.

The same approach applies to health, food, education, transportation, energy, software, financial services, and new technologies. A financial firm could organize its products around strengthening the productive lives of customers. A manufacturer could seek durable benefit for users and communities. A technology company could state how its systems expand human capability before specifying the financial returns needed to sustain that work.

Legal form matters because an enterprise's purpose must survive changes in management, market conditions, and short-term pressure. Corporate charters, governance duties, accountability to affected parties, enforceable commitments, and remedies can make the governing purpose consequential. Delaware's public benefit corporation framework already supplies a partial example: directors must balance stockholders' financial interests, the interests of materially affected persons, and the public benefits specified in the company's charter [17]. A more explicit ordering would place avoiding harm and delivering a declared human benefit ahead of financial maximization.

A productive economy organized this way would retain entrepreneurship, competition, investment, property, and the discovery process. It would judge those mechanisms by their contribution to actual human possibilities. More goods, better tools, new knowledge, richer communities, and greater personal freedom would be achievements. More money would be valuable insofar as it helped create and sustain them.

10 The return to artificial intelligence

Now return to the original paperclip scenario. A machine was given a measure to optimize and became extraordinarily successful at optimizing it. Our examination of human economics suggests that the machine's problem did not begin with the machine. It began with the purpose selected for it.

Artificial intelligence is being developed and deployed by corporations, governments, investors, and other human institutions. It can make those institutions vastly more capable: designing products, choosing investments, extending credit, pricing assets, restructuring businesses, and enforcing contracts. If the governing institutional goal is the endless accumulation of financial returns and claims, highly capable AI may pursue that goal with greater speed and precision than people ever could.

The result need not resemble a factory covering the Earth with metal paperclips. It could be a global system that becomes remarkably adept at pricing, acquiring, multiplying, and enforcing claims on future production, while the real freedom of future generations to enjoy their productivity steadily contracts. The machine would be amplifying an economic objective humanity had already adopted.

The distinction remains the one we began with: an improved optimization score is not, by itself, evidence of an improved human outcome. Greater intelligence at maximizing a selected objective increases the importance of choosing a purpose that serves the real world. The same consideration applies to an individual, a corporation, a nation, and an artificial agent.

Human institutions that first establish worthwhile purposes can direct markets and increasingly capable machines toward those purposes. In that arrangement, AI can help people build better homes, medicines, scientific knowledge, infrastructure, and ways of living. Financial measures continue to guide practical choices, and real-world results determine whether the work is worthwhile.

The paperclip maximizer problem therefore returns to the society giving instructions to the machine. The purposes of that society, and the institutions that translate those purposes into economic activity, are central to the alignment of the intelligence they create.

11 Conclusion

The paperclip maximizer begins with a simple mistake: the count of paperclips takes the place of the human reason for making them. Goodhart observed the underlying failure in human monetary policy decades before modern artificial intelligence. The correspondence formalized in Appendix A.1 identifies the shared optimization structure.

Human civilization offers an unsettling possible example. Our productive abilities have achieved extraordinary results, while financial claims on future production have expanded to hundreds of trillions of dollars. Existing ownership, debt, and valuations increasingly influence the choices available to generations entering the economy. A market system celebrated for responding to reality can become occupied with fulfilling yesterday's financial commitments.

The Declaration of Independence articulated a purpose that exceeds financial accumulation: life, liberty, and the pursuit of happiness. An economic system can preserve the productive strength of markets while giving businesses explicit obligations to avoid harm, create genuine human benefit, and then earn the money needed to sustain those contributions.

We cannot mind read. But "just scale," and the newer, ostensibly worse prescription, "just let the machine self-improve even though it is clearly unreliable and unaligned," are consistent with what you do when you do not have an adequate understanding of the domain and hence no principled path to a solution. When accompanied by warnings that "we may end humanity" [18], that combination can support this hypothesis—an apparent admission that the consequences are not adequately understood or controlled.

We are already seeing warning signs: documented AI-assisted hacking [19] and reports of harmful chatbot interactions, including alleged contributions to mental-health deterioration and suicide [20]. These are concerns about life, liberty, and the pursuit of happiness now—not merely hypothetical consequences of a future superintelligence. The fact that the Trinity test did not ignite the atmosphere [21] offers no assurance that another technology's catastrophic risks will fail to materialize.

Where the intent problem remains unresolved, a prudent path may be minimum viable intelligence: sufficient capability to solve important automation problems, with demonstrated reliability and safeguards appropriate to the task. The objective is not maximum intelligence in the abstract, but dependable contributions to human life.

At scale, this approach could unlock trillions of dollars of real, useful economic value by solving important customer problems [22]. Where the same dependable outcomes can be achieved with less computation, it could simultaneously reduce the additional data-center capacity and power-generation investment needed to deliver those outcomes.

An unaligned, unreliable superintelligence undergoing recursive "self-improvement" is not necessarily improving what matters to the people it affects; it may instead be increasing its capacity to harm them.

Increasing unreliable, unaligned machine intelligence can amplify human folly and increase harm. Reliable machine intelligence aligned with genuine human benefit could instead help us recognize and correct misplaced objectives—turning machines from amplifiers of human folly into guides toward actual human flourishing.

> **Perhaps solving the paperclip maximizer problem is a shared task: humans and reliable, aligned machines helping one another distinguish what we can maximize from what is worth pursuing.**

A Mathematical formulation

The main argument can be read without notation. This appendix retains the formal correspondence and the cohort accounting relationship, together with their limits.

A.1 The shared optimization structure

Let X be a finite set of possible outcomes. Let $M(x)$ be the measure an optimizer seeks to increase, and let $U(x)$ be the value of that outcome for the people the activity is meant to serve. Their objectives have diverged whenever two outcomes $a, b \in X$ satisfy

$$\boxed{M(b) > M(a) \qquad \text{and} \qquad U(b) < U(a).} \tag{1}$$

In this comparison, the score improves while the human result becomes worse.

The paperclip version writes $C(x)$ for the paperclip count and $H(x)$ for human benefit. We can make the correspondence exact for the rankings of any finite problem. List the distinct values of M in increasing order, $m_1 < \cdots < m_k$, and define $C(x) = i$ whenever $M(x) = m_i$; set $H(x) = U(x)$. Increasing C now ranks all outcomes in precisely the same order as increasing M, including ties. Every preferred outcome and every reversal in Equation (1) is preserved. Conversely, any paperclip problem is the same formulation with $M = C$ and $U = H$.

This is an ordinal correspondence for deterministic choices among finitely many outcomes. It preserves rankings, ties, maximizers, and the stated reversals, not numerical distances between scores or every stochastic or dynamic optimization model. In particular, a nonlinear rank transformation need not preserve comparisons based on expected scores over lotteries.

The correspondence establishes a shared failure structure, not an empirical diagnosis of an economy. It does not establish that a particular financial measure has diverged from human benefit, that all financial optimization is harmful, or that the economic hypothesis in the main text is true. Those questions require evidence about the relevant outcomes and institutions.

A.2 Resources retained by a new cohort

For a new cohort, let Y_t be the real economic resources available from its work and opportunities during period t. Let Q_t be the net resources it must surrender to satisfy financial claims established outside the cohort, including relevant payments to older holders of claims. The capacity it retains for its own purposes is

$$A_t = Y_t - Q_t. \tag{2}$$

The quantities must be measured in consistent real-resource units over the same period. If additional production increases Y_t, but the increase in net resources surrendered through Q_t is larger, then A_t falls. More productive work can coexist with less capacity to benefit from that work. An increase in the *share* absorbed by claims does not, on its own, establish that the absolute resources retained have fallen.

Equation (2) is an accounting relationship, not an estimated causal model or evidence that this trajectory is occurring. The word *net* matters: identifying the cohort's position requires attention to relevant resources received as well as paid, and to productive capacity inherited alongside obligations. Payments, ownership, prices, wages, and intergenerational transfers supply evidence for that assessment. Aggregate debt or wealth totals cannot substitute for measuring the cohort's actual position.

B Supporting economic evidence

This appendix retains the detailed figures and source distinctions behind the main text. The observations illuminate different parts of the hypothesis; they are not interchangeable measures or a single causal test.

B.1 Production, financial claims, and wealth revaluation

The IIF reports global debt above $365 trillion in the first half of 2026, compared with the IMF's projection of approximately $128 trillion in world output for 2026 [11, 10]. The outstanding debt stock is therefore about 2.9 times the projected value of one year's world production. This comparison combines a debt observation with a full-year output forecast.

The IIF's historical account places global debt near $130 trillion in 2005, approximately 240% of world GDP. By early 2026 it exceeded $350 trillion, about 305% of world GDP [9]. These historical ratios are reported at their respective dates; they are not recalculated using the separate IMF forecast above.

McKinsey's 2026 global balance-sheet study estimates household wealth at $570 trillion at the end of 2025, an increase of $40 trillion in one year. Only 20% of that annual increase came from net new real capital formation. Much of the change came from higher prices assigned to assets that already existed. Its multilayered estimate of real and financial assets approached $1.8 quadrillion in the economies it modeled [12].

Table 1: Contemporary production and financial claims. Monetary amounts are in current U.S. dollars. These measures have different accounting bases and should not be added together.

Measure	Scale	Source and meaning
World output, 2026	About $128 trillion	One year of production; IMF forecast [10]
Global debt, H1 2026	More than $365 trillion	Outstanding obligations; IIF [11]
Global debt, 2005	About $130 trillion	Approximately 240% of world GDP; IIF [9]
Household wealth, 2025	$570 trillion	Net household wealth in McKinsey's model [12]
Household wealth gain, 2025	$40 trillion	Change during the year, not the wealth stock [12]
Real-investment share of wealth gain, 2025	20%	Portion of the year's gain from net new real capital [12]
Total balance-sheet assets, 2025	Nearly $1.8 quadrillion	Real and financial layers; major economies representing roughly 70% of world GDP [12]

Debt is a stock of obligations; GDP is an annual flow of production. Comparing them is a way to express scale, not a claim that every debt falls due together or that the ratio alone establishes an unsustainable burden. Household wealth is a net measure, whereas the much larger multilayer asset estimate includes real and financial layers. The latter is not an estimate of additional physical productive capacity, and its modeled economies do not cover the entire world.

The 20% figure describes the composition of the year's wealth increase, not the composition of all household wealth. It does not classify the entire remainder as speculation. The relevant distinction is between expanding productive capacity and changing the valuation of existing assets; interpreting either requires attention to what people actually gain access to.

B.2 Housing and intergenerational access

The OECD reports that homeownership among people in their thirties fell in 17 of 24 countries studied over the preceding three decades. In five countries, the decline exceeded 15 percentage points [13].

The housing example separates physical service from financial valuation. A new home adds shelter. A higher price for an unchanged home can raise an existing owner's recorded wealth while increasing the entry cost for someone seeking the same shelter. The ownership trend is relevant to that concern, but does not by itself identify its causes or measure every aspect of housing access. Assessing the cohort mechanism requires examining the resources, transfers, financing conditions, and opportunities described in Appendix A.2.

B.3 The scale of material improvement

Economic arrangements directed by rulers and royal privilege often concentrated control over what could be produced and exchanged. Twentieth-century centrally planned economies attempted that coordination through state agencies. The Soviet system, for example, used Gosplan and other authorities to specify production targets and allocate resources across much of the economy [8]. Hayek's account of dispersed knowledge explains the advantage of allowing changing local information to shape productive decisions rather than requiring a central authority to possess it all [7].

The World Bank's long-run reconstruction places extreme poverty—below $3 per person per day in 2021 purchasing-power-parity dollars—near 60% in 1950 and near 10% in 2025 [14]. Global average life expectancy rose from roughly 32 years in 1900 to about 73 years in 2023 [15]. These estimates provide context for the extraordinary expansion of human productive capabilities.

They are not a claim that one institutional mechanism alone caused that progress. They explain why the argument seeks to preserve productive discovery, entrepreneurship, investment, and useful exchange while asking what purposes the financial system should serve.

References

[1] N. Bostrom. *Ethical Issues in Advanced Artificial Intelligence.* In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, vol. 2, pp. 12–17, 2003. https://nickbostrom.com/papers/ethical-issues-in-advanced-ai/.

[2] C. A. E. Goodhart. *Problems of Monetary Management: The U.K. Experience.* Presented at the Reserve Bank of Australia Conference in Monetary Economics, 1975; published in *Papers in Monetary Economics*, Reserve Bank of Australia, 1976. Publication history and attribution: RBA, RDP 9013, "Conference Volumes," item 1, footnote 36. https://www.rba.gov.au/publications/rdp/1990/9013/conference-volumes.html.

[3] C. J. Moses. *Financial Innovations and Monetary Policy: A Preliminary Survey.* Research Discussion Paper 8301, Reserve Bank of Australia, May 1983. Introduction, footnote 1, p. 1 (PDF p. 3). https://www.rba.gov.au/publications/rdp/1983/8301.html.

[4] J. Sohl-Dickstein. *Too Much Efficiency Makes Everything Worse: Overfitting and the Strong Version of Goodhart's Law.* 2022. https://sohl-dickstein.github.io/2022/11/06/strong-Goodhart.html.

[5] S. Zhuang and D. Hadfield-Menell. *Consequences of Misaligned AI.* NeurIPS 33, pp. 15763–15773, 2020. https://papers.nips.cc/paper/2020/hash/b607ba543ad05417b8507ee86c54fcb7-Abstract.html.

[6] E. El-Mhamdi and L.-N. Hoang. *On Goodhart's Law, with an Application to Value Alignment.* arXiv:2410.09638v1, 12 October 2024. https://arxiv.org/abs/2410.09638v1.

[7] F. A. Hayek. *The Use of Knowledge in Society.* American Economic Review 35(4), 519–530, 1945. https://www.econlib.org/library/Essays/hykKnw.html.

[8] International Monetary Fund, OECD, World Bank, and EBRD. *A Study of the Soviet Economy.* 1991, chapter on the planned economy and Gosplan. https://www.elibrary.imf.org/display/book/9789264134683/ch002.xml.

[9] Institute of International Finance. *Reflections Across Two Decades of Global Debt: Evidence from the IIF Global Debt Monitor.* August 5, 2026. https://www.iif.com/Publications/ID/6672.

[10] International Monetary Fund. *World Economic Outlook Database: World GDP, Current Prices.* October 2026 forecast; series NGDPD, aggregate WEOWORLD, year 2026 (billions of U.S. dollars). https://www.imf.org/external/datamapper/datasets/WEO.

[11] Institute of International Finance. *Global Debt Monitor: A Debt-intensive Future—and Growing Risks for Global Markets.* 23 September 2026, p. 1. Original report reproduced by Poder360: https://static.poder360.com.br/uploads/2026/09/relatorio-Instituto-Financas-Internacionais-setembro2026.pdf.

[12] R. J. Anderson et al. *The Global Balance Sheet 2026: Imbalance and Divergence.* McKinsey Global Institute, 23 July 2026. https://www.mckinsey.com/mgi/our-research/the-global-balance-sheet-2026-imbalance-and-divergence.

[13] OECD. *OECD Employment Outlook 2025*, chapter 2: demographic change and intergenerational inequalities. Launch version, section 2.4.2, p. 98; Figure 2.18, p. 100. https://doi.org/10.1787/194a947b-en.

[14] C. Lakner and S. K. T. Baah. *A Century of Extreme Poverty: Five Facts from 1950 to 2050.* World Bank Data Blog, 26 October 2026. https://blogs.worldbank.org/en/opendata/five-facts-about-extreme-poverty-in-the-world-in-long-time-horiz.

[15] B. Herre. *Global Average Life Expectancy Has More than Doubled Since 1900.* Our World in Data, 4 February 2025. https://ourworldindata.org/data-insights/global-average-life-expectancy-has-more-than-doubled-since-1900.

[16] Continental Congress. *Declaration of Independence.* Adopted 4 July 1776. National Archives transcript. https://www.archives.gov/founding-docs/declaration-transcript.

[17] Delaware General Corporation Law. *Title 8, Chapter 1, Subchapter XV: Public Benefit Corporations*, sections 362 and 365. https://delcode.delaware.gov/title8/c001/sc15/.

[18] Center for AI Safety. *Statement on AI Extinction Risk.* 30 May 2023. https://safe.ai/work/statement-on-ai-extinction-risk. Cited for public extinction-risk warnings. The interpretation is the author's hypothesis; the conclusion's quoted formulations are rhetorical summaries, not verbatim quotations attributed to particular individuals.

[19] Anthropic. *Threat Intelligence Report, September 2026*, section on AI-augmented cyber operations. https://www.anthropic.com/threat-intelligence-report-september-2026. Provider-reported incidents.

[20] BBC News. *Parents of Teenager Who Took His Own Life Sue OpenAI.* 27 August 2025. Excerpt reproduced by the Business & Human Rights Resource Centre: https://www.business-humanrights.org/en/latest-news/usa-lawsuit-accuses-openai-of-negligence-over-suicide-of-teenager/. Reports allegations in the Raine lawsuit and OpenAI's response, not an adjudicated causal finding or a population-level trend.

[21] U.S. Department of Energy, History Program. *Safety and the Trinity Test (July 1945).* https://www.osti.gov/opennet/manhattan-project-history/Events/1945/trinity_safety.htm.

[22] M. Chui et al. *The Economic Potential of Generative AI: The Next Productivity Frontier.* McKinsey & Company, June 2023, p. 9 (PDF p. 11). Original report reproduced by CTO Forum: https://www.ctoforum.org/wp-content/uploads/2023/06/SS2M6_McKinsey_The-Economic-Potential-of-Generative-AI-the-Next-Productivity-Frontier-vf.pdf. Its $2.6–$4.4 trillion annual estimate across 63 use cases illustrates industry-wide generative-AI potential, not a forecast or measured result of the minimum-viable-intelligence approach proposed here.