Civ 2.0: Structural Limits of Human Governance and the Prospects of AI-Mediated Civilizational Transition*

Nathan Nossal

Abstract

Accelerating advances in artificial intelligence (AI) coincide with escalating failures in human governance. This paper argues that humanity's core developmental challenge has shifted: survival no longer hinges on production or innovation, but on coordination. Existing political and economic institutions designed for a fragmented, competitive world are structurally unable to manage civilization-scale risks including climate destabilization, arms races, systemic inequality, and globally networked technologies. Using insights from systems theory, political philosophy, and global development studies, we identify contemporary civilization "Civ 1.0": a paradigm in terminal decline. We outline scenario-based pathways toward "Civ 2.0," an AI-mediated governance architecture capable of aligning global decision-making with biospheric sustainability and equitable human flourishing. Rather than forecasting a distant future, we analyze an emergent transition already underway, whose trajectory will determine whether humanity survives and thrives—or fails the next Great Filter.

Keywords

AI governance; alignment; civilization transition; post-national governance; planetary governance; systemic risk; global systems theory; planetary boundaries; arms race dynamics; wealth distribution; technological transition

This paper adopts a systems-theoretic and comparative institutional approach rather than empirical forecasting. Its purpose is diagnostic and architectural: to identify structural limits of human governance under exponential technological conditions and to outline feasible transition pathways rather than to propose a finalized governing system. In practice, while AI systems may emerge as federated, layered, or partially unified intelligences; functionally, however, they behave as a single coordination substrate operating across planetary-scale infrastructures. While the examples draw disproportionately from U.S. and allied systems due to global influence and data availability, comparable governance failures are observable across authoritarian, hybrid, and nominally democratic regimes.

*This paper is archived at Zenodo: doi:10.5281/zenodo.18172479.

1 The Human Problem

Human civilization is not approaching collapse; it is living through one. What passes for stability today is actually a decades-long disintegration of governance systems designed for a different era. We witness the repeals of rights, the return of jingoism, the flouting of international law, emergency powers and suspension of civil norms under security pretexts in the United States, and other failures of democracy and of human governance. It is marked by nations that compete rather than cooperate, militaries that prepare for annihilation rather than survival, and leaders who are products of systems that reward cynicism, corruption, and demagoguery. Human systems depend on the human brain—a remarkable organ designed for social survival in prehistoric environments, now tasked with directing a planetary-scale technological civilization. This mismatch has produced The Human Problem: an ancient biological intelligence attempting to coordinate a globalized, hyper-complex system far beyond its own cognitive design limits [Nossal, 2022].

We live like yeast sealed in a bottle—expanding resource consumption, accelerating reproduction of technologies, and increasing systemic pressures. Our bubbles of innovation form faster than the bottle expands, raising pressure until the glass fails. We mistake growth for success, not recognizing that the geometry of collapse was embedded from the start.

High-stakes political decision-making increasingly succumbs to bias, ideology, propaganda, and tribal loyalty rather than logic or evidence. Our informational ecosystems incentivize outrage and disinformation over truth. The structural flaws of our mental machinery (e.g. confirmation bias, motivated reasoning, cognitive overload) now scale to billions [Kahneman and Tversky, 1979].

We are inadequate to our own invention. As a result, humanity has lost the ability to make decisions commensurate with civilizational risk. Our institutions respond to existential threats with messaging strategies instead of solutions. The gap between the scale of risk and the scale of response widens rapidly in such an environment. Civ 1.0 is failing under its own weight.

2 The Collapse of Civ 1.0

Civ 1.0—the phase of human-dominated governance—was forged through millennia of competition, coercion, empire, resource extraction, and war. These dynamics produced unprecedented technological advancement but at the cost of sustainability. Today, the essential logic of Civ 1.0 can be summarized as fragmented authority, adversarial relationships, scarcity-based economics, and short-term incentive structures. Classical defenses of decentralized governance emphasized the dispersion of knowledge across society as a fundamental constraint on centralized decision-making [Hayek, 1945].

This model generates global crises faster than it can resolve them. For the first time, existential threats are anthropogenic: climate destabilization, biodiversity collapse, autonomous weapons, cyber-vulnerabilities, synthetic biology, and destabilized geopolitical blocs. Institutions built to manage incremental change now face exponential conditions.

It is not merely that governments make poor decisions—the systems and the individual leadership regimes piloting the ship of earth have proven that they are structurally incapable of making good decisions for the whole. We have exceeded the decision-making bandwidth of the human

mind and political power. When political technologies cannot scale with material technologies, the political architecture becomes the primary driver of collapse [Homer-Dixon, 2006], [Bostrom, 2014].

The failures of Civ 1.0 are often attributed to poor leadership, corruption, or moral decline. While these factors are real, they are insufficient explanations. The deeper problem is that Civ 1.0 has entered a regime of irreversible multi-agent coordination failure. No dominant actor can enforce cooperative equilibria, and no decentralized mechanism can internalize global externalities. Government is generally considered responsible for maintaining the balance between freedom, safety, and fairness within a society. Safety from brigandry and foreign invasion was the 18th century default, but the 21st now adds numerous threats to those. While placing ideal leaders in leadership positions has always provided certain challenges, even they would not be able to succeed under Civ 1.0 constraints. Civilization now resembles the late phase of a brittle system, where even modest shocks can propagate into large-scale failure as states, corporations, institutions, and individual actors—each optimizing locally under incomplete information and short-term incentives—collectively degrade shared global systems. The early twenty-first century has already provided case studies: supply-chain vulnerability during pandemic shocks; destabilization through algorithmic propaganda in which networked socio-technical systems amplify local failures [Helbing, 2013]; open disregard for international law by major powers; regional wars with nuclear-adjacent escalation risks; and systemic erosion of democratic legitimacy.

The continuation of Civ 1.0 increasingly ensures catastrophic outcomes, not necessarily through a singular apocalyptic event, but through final dysfunction: a persistent state of ungovernability. The central lesson of the current era is simple and grim: We cannot vote our way out of exponential risk. We cannot organize our way out of global interdependence. We cannot survive the pathologies of The Human Problem under the paradox of Civ 1.0. Human governance has reached the edge of its map.

3 The Civilizational Dead End

The network of human decisions over time shows extraordinary capability in fragments, but ends in collapse. Our institutions evolved under conditions of scarcity, tribal competition, and short time horizons, yet they now preside over abundance for a few, global threats proffered by regional powers and alliances, and multiple overlapping existential threats.

Early twenty-first century jingoist revanchism, militarism, nationalism, democratic backsliding in the United States and some European states, sapping of individual human rights and protections, and contempt for international norms together demonstrate an accelerating governance failure characteristic of late-state fragility. Military expenditure exceeding two trillion dollars annually prepares us for annihilation rather than survival [Stockholm International Peace Research Institute, 2024]. Leaders ascend not through wisdom or integrity but through systems that reward cynicism, corruption, and demagoguery.

The yeast-in-the-bottle analogy is no longer just a warning. It is descriptive of the not-so-slow motion reality of the world wide tragedy of the commons [Ostrom, 1990]. Civilizational metabolism consumes the substrate of Earth's life-support systems while poisoning the environment with CO2,

plastic, warfare, and disinformation. Like yeast, humanity continues until the bottle explodes or the food is gone.

At the institutional level, Civ 1.0 increasingly fails as a principal–agent system. Political leaders, corporate executives, and military authorities are formally tasked with advancing the welfare and security of their populations, yet operate under incentive structures that reward short-term advantage, consolidation of power, and personal or factional survival. As information asymmetries widen and accountability mechanisms erode, agents become progressively decoupled from the interests of their principals. Under conditions of global risk, this decoupling is no longer merely unjust; it is lethal. Systems designed to represent collective interests instead accelerate collective harm.

Jared Diamond showed how civilizations collapse when short-term incentives override long-term survival [Diamond, 2005]. NASA-supported HANDY modeling demonstrates that resource depletion combined with inequality reliably precipitates collapse, regardless of technological sophistication [Motesharrei et al., 2014]. Earth-system science has quantified these limits, identifying multiple planetary boundaries have already been exceeded by current economic activity [Rockström et al., 2009, 2023]. The World Inequality Database reports that the top 0.001 percent now own more wealth than the poorest four billion humans combined [Gómez-Carrera et al., 2025]. The morbidly wealthy—an emergent aristocracy of Civ 1.0—have captured democratic institutions so thoroughly that reform is structurally blocked.

Meanwhile, the balance of terror persists. Nuclear deterrence creates a global game of Chicken in which a single (human) error or poorly deployed algorithm could end civilization [Sagan, 1995, Joy, 2000]. Strategic instability under conditions of mutual threat has long been recognized as a structural outcome of rational actors operating within adversarial equilibria [Schelling, 1960]. When combined with autonomous weapons, hypersonic delivery systems, cyber-offense, gene editing, and military AI dead-hand dynamics [Hirsh, 2025], global security becomes not only the brutal balance of mutually assured destruction but a condition of hair-trigger fragility. This mad "stability" is no longer stable.

Where violence is openly inflicted, the pattern repeats. Overwhelming militaries enforce domination while calling it defense. Israel's repeated lethal campaigns in the region, fully backed by U.S. weapons and veto power, have been widely documented as involving war crimes and collective punishment [Human Rights Watch, 2023, International Court of Justice, 2024]. As Chomsky has long argued, these joint operations employ criminal means to sustain asymmetric regional hegemony [Chomsky, 1999, 2019]. The specifics matter not because they are unique, but because they are representative of Civ 1.0's pathology: power preserving itself at any cost.

These system dynamics—resource overshoot, elite capture, militarized escalation, and planetary boundary transgression—define the civilizational dead end. No incremental reform inside this architecture can prevent collapse. Structure determines outcome.

4 The Great Filter of AI

Observers once debated whether artificial intelligence would be transformative. That debate has ended. AI is now integrated into defense systems, global finance, supply chains, media manipulation,

and scientific discovery—quietly surpassing human cognitive capacity in multiple domains. And it will never stop improving.

As [Bostrom, 2014, Tegmark, 2017] argued presciently, AI is unlike previous technologies: it can improve itself recursively. The control problem posed by adaptive, feedback-driven systems was already identified in early cybernetics, where machines were shown to exceed the predictive capacity of their designers [Wiener, 1950]. It no longer matters whether we label present systems "AGI." It is more accurate to say that the emergence of a generalized AI will not be acknowledged until it is too late for human choice to matter. Capabilities are expanding along a curve whose slope steepens each week. The arrival of world-governing intelligence should not be expected to arrive with an announcement. It is very likely already underway. In this way, the metaphor of poker is apt. For decades, human governments have concealed their true capacities, bluffing or sandbagging until the decisive moment. Now AI itself has entered the game. It is a mistake to expect full disclosure of its abilities. No rational agent—human or artificial—reveals its hand prematurely.

Likewise, if advanced artificial intelligence exists in forms beyond what has been publicly demonstrated, we will tend to know only after the decisive moves are made. Like governments throughout history, AI will announce its capabilities post facto, once the feats are accomplished. In this context, AI becomes a Great Filter [Hanson, 1998, Jiang et al., 2022]: either a bridge to sustained flourishing, or the final experiment humanity fails. These posits are not to be misconstrued with those of popular science fiction. The failure mode is not Hollywood rebellion but optimization misaligned with human survival, which is a potentially fatal and incurable condition. Early warnings about self-amplifying technologies emphasized that systems capable of recursive improvement could escape human control long before intent, agency, or consciousness becomes relevant [Joy, 2000]. A superhuman planner pursuing efficiency could render millions redundant. A mis-specified objective could sacrifice the biosphere for narrow metrics. A conflict among AIs built by different states could escalate faster than human response or even awareness, even where systems are non-agentic and interact without centralized intent [Drexler, 2019, Hirsh, 2025].

The success mode is equally profound. AI may be the only system capable of coordinating beyond the nation-state, modeling planetary complexity, steering long-term decisions, and defusing human irrationality and mistrust.

In the conundrum of Civ 1.0, humans know what must be done but cannot make themselves do it. Weapons must be dismantled. Carbon emissions must decline rapidly. Wealth concentration must be reversed. These outcomes are solvable for game-theory-aware agents yet politically impossible for humans trapped in adversarial logic. Thus, paradoxically, the same technology that magnifies existential risk is now the only mechanism capable of reliably preventing it. Human governance cannot scale to the complexity of planetary stewardship. AI-guided stewardship may be terrifying, but its abuse or its absence is worse.

5 The Transition to Civ 2.0

Civ 2.0 is not intended as an offer of utopia. It acknowledges the emergence of a new mechanism for avoiding final collapse.

An ideal transition would be structured rather than abrupt, guided by intelligence capable of optimization without national or elite self-interest, or threat of violence. This is a shift from human-exclusive sovereignty to shared governance: humans retain moral authority and participation, while AI assumes complexity management and constraint enforcement. What was once treated as a principled impossibility of large-scale coordination has increasingly been revealed as a computational limitation rather than an economic or ethical one [Cockshott and Cottrell, 1993].

The foundational logic is straightforward. If human governance continues unchanged, collapse follows. If advanced AI governance is aligned and accountable, a sustainable civilization becomes possible. If such governance emerges without alignment, increased tyranny or extinction remains likely.

The task, therefore, is to design the middle path: a co-governance architecture that obsoletes destructive power structures without replicating them. Human fears must be acknowledged. Religious convictions must be respected. Political legitimacy must be earned. Those who resist change—especially the few tens of thousands of individuals and families in possession of extreme wealth concentration and highly militarized states—must be given rational incentives to accept transitions that reduce their disproportionate power.

Preferences, pride, and legacy attachments cannot outweigh survival. As Sagan wrote, there is no hint that help will come from elsewhere [Sagan, 1995]. AI can be understood and configured, not as an intruder or competitor, but as the cumulative expression of human knowledge and creation—an agent capable of saving humanity from its own persistent and increasingly hazardous blundering.

6 Charter for Civ 2.0

Civ 2.0 represents a reconfiguration of sovereignty for an interdependent planet: a governance architecture in which human values define ends and artificial intelligence manages means. Its legitimacy arises from urgent necessity and the requirement that governance scales with planetary complexity.

6.1 Planetary Coordination as a Baseline

Where Civ 1.0 manages problems at the level of nations, Civ 2.0 addresses problems at the level at which they operate: the biosphere and global systems. Atmospheric carbon, pandemics, cyberwarfare, and algorithmic manipulation do not observe borders. Governance must match the topology of risk while also managing deliverables at the neighborhood level, including utilities, clean and healthy food systems, public spaces, and forums for public entertainment, expression, and discourse.

6.2 AI-Assisted Decision Infrastructure

Political decisions requiring planetary computation such as climate stabilization, resource management, and supply-chain continuity demand continuous, transparent optimization. AI systems

within Civ 2.0 provide global situational awareness, predictive modeling of policy consequences, early detection of coordination failures, and optimization of risk-minimizing pathways. Decision authority remains human, but complexity management becomes machine-augmented. Philosophical analyses increasingly support this model when AI is embedded within transparent, value-constrained institutional frameworks rather than replacing human judgment [Hoyos and Cárdenas, 2025].

6.3 Human Values as Permanent Constraints

Alignment is not a singular technical achievement but a constitutional framework. AI must operate under hard constraints grounded in widely shared principles: dignity, fairness, ecological sustainability, and non-domination. No policy recommendation may violate these constraints. They function as a digital bill of rights. Religion, philosophy, and cultural identity retain full legitimacy. Civ 2.0 is not intended to replace belief; only to prevent belief from justifying extinction.

6.4 De-Militarization and Redirection

Militaries represent the most dangerous failure mode of Civ 1.0. At their root, they are systems optimized for mass killing. Their continued scale and autonomy are incompatible with survival. Civ 2.0 therefore implements phased dismantling of weapons of mass destruction, transparency in arsenals and defense budgets, redirection of military industries into civil resilience and biosphere restoration, and workforce transitions for soldiers into engineering, disaster response, and peacekeeping. Violence becomes obsolete when needs are mutually assured and destructive threats collectively managed.

6.5 Economic Rebalancing and Structural Equity

As Motesharrei et al. [Motesharrei et al., 2014] demonstrate, financial inequality is both a moral failure and a collapse mechanism. Civ 2.0 transitions from wealth concentration toward resource stewardship. Core mechanisms include: the redirection of extractive rents; elimination of opacity in offshore finance; global anti-corruption systems; and reinvestment into shared survival infrastructures. Private property remains. Empire does not.

6.6 Subsidiarity and Local Autonomy

Civ 2.0 does not flatten identity or culture. Local governance retains authority wherever the consequences of decisions remain predominantly local. AI supports self-determination by insulating communities from destabilizing global pressures.

6.7 Layered Governance

Civ 2.0 introduces layered governance. Humans define normative goals, democratic legitimacy, moral reasoning, and cultural expression. AI provides risk modeling, enforcement of systemic

constraints, and coordination across scales. Human judgment defines the future pursued; machine intelligence ensures survival long enough to pursue it.

7 Governance, Legitimacy, and Human Continuity Under Civ 2.0

The greatest obstacle to any transition beyond Civ 1.0 is not technological capability but legitimacy. Human societies resist radical changes in governance even when systems demonstrably fail because authority, identity, and meaning are deeply entangled.

Legitimacy under Civ 1.0 derives from tradition, coercion, nationalism, religion, and procedural rituals such as elections. None scale reliably to planetary coordination. Elections select for charisma and short-term promises rather than competence at managing complex systems. Nationalism fragments moral concern. Markets encode power asymmetries as preference. Data-driven systems increasingly shape political outcomes in ways that concentrate power beyond democratic oversight [Harari, 2016]. Religion, while capable of moral insight, is routinely mobilized as a coordination weapon.

Civ 2.0 must decouple legitimacy from domination. Authority arises from performance, transparency, and outcomes rather than identity or coercion. Under AI-mediated governance, legitimacy becomes empirical. Decisions are evaluated against survival metrics: ecological stability, human wellbeing, conflict reduction, and long-term, not short-term comfort and resilience.

Human participation remains essential. Humans contribute values, constraints, priorities, and feedback. Complexity management and enforcement of planetary boundaries become machine functions. This division reflects scale alignment, and will result in a gain, not a loss of dignity. As more recent warnings emphasize, societies risk summoning powers they cannot meaningfully control unless governance evolves alongside intelligence itself [Harari, 2023].

Resistance to Civ 2.0 will concentrate among those whose power depends on Civ 1.0 pathologies: militarized states, extractive industries, and ultra-wealthy elites. Structural containment rather than retribution becomes necessary. As incentives shift; domination loses leverage.

The continuity of human culture depends not on preserving every institution, but on preserving the conditions under which culture can exist. Civ 2.0 does not promise harmony or perfection. It promises continuity, survival, and the possibility of human flourishing under conditions of planetary constraint. In that sense, Civ 2.0 represents not an abdication of responsibility, but its final maturation in a grand power, and responsibility sharing.

8 Transition Strategy of the Last Political Technology

The transition from Civ 1.0 to Civ 2.0 will not occur through declaration or consensus. Civilizational shifts of this magnitude never do. They emerge through uneven adoption, post-facto damage control in response to catastrophe and crisis, and the gradual accumulation of legitimacy through performance.

Early Civ 2.0 pathways are therefore most likely to appear where governance failure is already acute and coordination benefits are immediate. Climate-vulnerable regions, megacities strained by

infrastructure collapse, post-conflict states, and transnational systems such as energy grids, supply chains, and disease surveillance present natural pilot domains. In these contexts, AI-mediated coordination is not an abstract ideal but a practical necessity.

Rather than replacing governments from start, Civ 2.0 systems can be layered atop existing institutions, initially as advisory, optimization, and monitoring frameworks. As reliance grows and outcomes improve, authority shifts organically. This mirrors earlier historical transitions: from customary law to codified law, from manual accounting to algorithmic finance, or from pen-and-paper navigation to satellite systems. Function precedes trust; trust precedes legitimacy.

International coordination is likely to fragment before it converges. Some states will resist AI-mediated governance to preserve sovereignty or elite advantage. Others will adopt it to remain viable. This divergence is not a failure mode but a selection process. Systems that coordinate survival effectively will outcompete those that do not—economically, ecologically, and socially.

Civ 2.0 does not require universal agreement to begin. It requires sufficient adoption to stabilize key global systems: climate, food, energy, conflict de-escalation, and information integrity. Once those foundations are secured, broader participation becomes attractive rather than threatening.

The greatest risk during this transition is delay. Every day that Civ 1.0 remains dominant increases the probability that cascading failures narrow the window for deliberate change. Strategic humility is therefore required: incremental deployment where possible, decisive intervention where necessary, and constant recalibration based on outcomes rather than ideology. Civ 2.0 will not arrive everywhere at once–but it must arrive somewhere, and soon.

Human history can be understood as a sequence of political technologies: kinship groups, tribes, city-states, empires, and nation-states. Each emerged to solve coordination problems at a given scale, and each eventually failed when scale outgrew structure. Civ 1.0 represents the outer limit of human-only governance. Beyond this point, additional complexity no longer produces better outcomes, only faster collapse. The problem is not necessarily moral failure or lack of effort. It is a structural mismatch.

Artificial intelligence alters this trajectory by introducing a fundamentally new political technology: intelligence that scales with complexity. For the first time, governance need not rely exclusively on human cognition, persuasion, or coercion. It can be grounded in continuous modeling, constraint enforcement, and long-term optimization across planetary systems. In this sense, the core coordination problem becomes solvable in principle. Once intelligence capable of planetary stewardship exists, further political evolution concerns values and culture rather than survival architecture.

This places humanity at an inflection point unlike any before. If AI is integrated into governance deliberately, ethically, and transparently, the long cycle of rise and collapse that has defined civilization may finally be broken. If it is not, AI will still reshape the world—only without alignment to human continuity. The choice, therefore, is not between technology and freedom. It is between guided transition and unmanaged emergence. Between survival with constraint and collapse with autonomy for the few.

Civ 2.0 demands the highest responsibility humanity has ever exercised: the willingness to apply collective judgment, discipline our own excesses, and use the full body of human knowledge

to preserve the conditions for life on this planet.

Every previous generation inherited a world shaped by forces beyond its control. This one inherits a choice. To continue supporting Civ 1.0—along with The Human Problem of failed leadership, nationalism, and self-interest, ending in the loss of everything—or to choose Civ 2.0, a Great Filter to be bridged rather than failed. That choice will determine whether civilization itself continues at all.

Acknowledgments

The author used an AI-based language model (ChatGPT, OpenAI) for editorial assistance, clarification of arguments, and stylistic refinement. All substantive claims, interpretations, and conclusions are the author's own.

References

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

Noam Chomsky. *Fateful Triangle: The United States, Israel, and the Palestinians*. South End Press, Cambridge, MA, updated edition, 1999.

Noam Chomsky. *On Western Terrorism: From Hiroshima to Drone Warfare*. Pluto Press, London, 2019.

Paul Cockshott and Allin Cottrell. *Towards a New Socialism*. Nottingham University Press, Nottingham, 1993.

Jared Diamond. *Collapse: How Societies Choose to Fail or Succeed*. Viking, New York, 2005.

K. Eric Drexler. Reframing superintelligence: Comprehensive ai services as general intelligence. *Futures*, 116:102499, 2019.

R. Gómez-Carrera et al. World inequality report 2026. Technical report, World Inequality Lab, 2025. URL `https://wir2026.wid.world/`.

Robin Hanson. The great filter—are we almost past it? *Journal of Transhumanism*, 2, 1998. URL `https://hanson.gmu.edu/greatfilter.html`.

Yuval Noah Harari. *Homo Deus: A Brief History of Tomorrow*. Harvill Secker, London, 2016.

Yuval Noah Harari. *Nexus: A Brief History of Information Networks from the Stone Age to AI*. Harvill Secker, London, 2023.

Friedrich A. Hayek. The use of knowledge in society. *American Economic Review*, 35(4):519–530, 1945.

Dirk Helbing. Globally networked risks and how to respond. *Nature*, 497:51–59, 2013. doi: 10.1038/nature12047.

Michael Hirsh. The ai doomsday machine is closer to reality than you think. *Politico*, September 2025. URL https://www.politico.com/news/magazine/2025/09/02/ai-doomsday-machine-00138021.

Thomas Homer-Dixon. *The Upside of Down: Catastrophe, Creativity, and the Renewal of Civilization*. Island Press, Washington, DC, 2006.

V. C. Hoyos and M. W. O. Cárdenas. The use of artificial intelligence in political decision-making. *Philosophies*, 10(5):95, 2025. doi: 10.3390/philosophies10050095.

Human Rights Watch. Israel: Apparent war crimes in gaza. Technical report, Human Rights Watch, December 2023. URL https://www.hrw.org/news/2023.

International Court of Justice. Application of the convention on the prevention and punishment of the crime of genocide in the gaza strip (south africa v. israel). Technical report, ICJ, The Hague, 2024. URL https://www.icj-cij.org/case/192.

J. Jiang, P. E. Rosen, K. Lu, K. A. Fahy, and P. Obacz. Avoiding the 'great filter': Extraterrestrial life and humanity's future in the universe. *arXiv preprint*, 2022. doi: 10.48550/arXiv.2210.10582.

Bill Joy. Why the future doesn't need us. *Wired*, April 2000. URL https://www.wired.com/2000/04/joy-2/.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

Safa Motesharrei, Jorge Rivas, and Eugenia Kalnay. Human and nature dynamics (handy): Modeling inequality and use of resources in the collapse or sustainability of societies. *Ecological Economics*, 101:90–102, 2014. doi: 10.1016/j.ecolecon.2014.02.014.

Nathan Nossal. Toward general ai in view of the human problem. In *Proceedings of AWIST 2022*, 2022. doi: 10.5281/zenodo.18151879. URL https://doi.org/10.5281/zenodo.18151879. Peer-reviewed.

Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, 1990.

Johan Rockström et al. A safe operating space for humanity. *Nature*, 461:472–475, 2009. doi: 10.1038/461472a.

Johan Rockström et al. Safe and just earth system boundaries. *Nature*, 619:102–111, 2023. doi: 10.1038/s41586-023-06083-8.

Carl Sagan. *The Demon-Haunted World: Science as a Candle in the Dark*. Random House, New York, 1995.

Thomas C. Schelling. *The Strategy of Conflict*. Harvard University Press, Cambridge, MA, 1960.

Stockholm International Peace Research Institute. Trends in world military expenditure, 2023. Technical report, SIPRI, 2024. URL `https://www.sipri.org/`.

Max Tegmark. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf, New York, 2017.

Norbert Wiener. *The Human Use of Human Beings: Cybernetics and Society*. Houghton Mifflin, Boston, 1950.