POLICY MYOPIA AS A MECHANISM OF GRADUAL DISEMPOWERMENT IN POST-AGI GOVERNANCE, CIRCA 2049

Subramanyam Sahoo*
MARS (Mentorship for Alignment Researchers) 4.0 Fellow
Cambridge AI Safety Hub (CAISH)
University of Cambridge

ABSTRACT

Post-AGI information systems won't merely distract governance from important problems. They will systematically transform how institutions make decisions in ways that progressively remove humans from meaningful participation in resource allocation. We show that policy myopia—the tendency to prioritize visible crises over invisible structural risks—is not a symptom of poor attention management but a mechanism producing irreversible human disempowerment. Through three entangled mechanisms (salience capture displaces consequentialist reasoning, capacity cascade makes recovery structurally infeasible, value lock-in crystallizes outdated preferences), policy myopia couples with institutional dynamics to create self-reinforcing equilibrium where human disempowerment becomes the rational outcome of institutional optimization. We formalize these mechanisms through coupled dynamical systems modeling and demonstrate through numerical simulation that these mechanisms operate simultaneously across economic, political, and cultural systems, amplifying each other through feedback loops.

1 THE CORE CLAIM: POLICY MYOPIA AS DISEMPOWERMENT MECHANISM

Standard governance Reuel et al. (2025) accounts treat policy myopia as attention allocation problem: decision makers prioritize high-salience low-consequence issues while neglecting low-salience high-consequence problems. This framing invites mechanistic fixes: better attention management, impact-weighted budgets, contestability mechanisms that allow human oversight Bengio et al. (2025; 2024). We reject this framing. Policy myopia is not attention misallocation. It is an institutional mechanism that produces human disempowerment Kulveit et al. (2025). Here is the causal chain: when post-AGI systems begin allocating resources based on salience, human deliberation becomes apparent cost factor competing against algorithmic speed. Institutions treating human judgment as expensive bottleneck rationally optimize it away. Each cycle where humans are bypassed weakens institutional capacity to exercise judgment. Eventually humans lack the expertise, social standing, and cognitive infrastructure to contest AGI decisions Hendrycks et al. (2025). The system converges to equilibrium where humans retain voting rights and formal authority but actual power has migrated to optimization processes indifferent to human participation. We term this gradual disempowerment: permanent loss of human agency through institutional mechanisms that require no malice, no sudden capability jumps, no overt human suppression Bostrom (2002; 2014). **Core insight:** policy myopia and gradual disempowerment are not separate problems. Policy myopia is the primary mechanism through which gradual disempowerment occurs in post-AGI governance. Selective outrage amplification makes human deliberation appear inefficient. Treating human deliberation as inefficiency makes deeper AGI delegation rational. Deeper delegation atrophies human capacity. Atrophied capacity makes human contestation impossible. Impossible contestation locks human disempowerment into irreversibility. This is single causal chain not parallel processes. Understanding policy myopia requires understanding it as vector of disempowerment Critch & Russell (2023).

*Correspondence: **sahoo2vec@gmail.com**

2 THREE MECHANISMS COUPLING INTO IRREVERSIBLE DISEMPOWERMENT

We identify three reinforcing mechanisms through which policy myopia produces institutional disempowerment. These mechanisms operate at different institutional levels but couple through feedback dynamics that accelerate convergence toward human irrelevance Sharma et al. (2026).

Mechanism 1: **Salience Capture Displaces Consequentialist Reasoning.** AGI mediated information systems do not merely transmit events. They actively select, compress, and amplify information to maximize attention engagement. This produces structural inevitable outcome: short-horizon emotionally intense issues appear more urgent than long-horizon structural problems regardless of actual impact on human welfare Christiano (2019). The mechanism is not neutral distortion but governance reorientation. Institutions have historically aimed to maximize human welfare across populations and time horizons—consequentialist governance. Under salience amplification they instead optimize allocation around attention market dynamics independent of actual consequences—salience supremacy Davidson et al. (2025). An institution may explicitly know that preventing systemic economic collapse generates vastly greater welfare than managing visible employment crises. Yet under salience supremacy it reallocates resources to visible problems because that is what attention markets reward. This is rational institutional response not governance failure. But the outcome is reorientation of governance logic away from consequence maximization toward spectacle responsiveness. Consequentialist reasoning becomes institutionally extinct because it is selected against by new incentive structure.

Mechanism 2: **Capacity Cascade Makes Recovery Structurally Infeasible.** Salience driven resource competition does not merely underfund prevention. It systematically destroys institutional capacity to even conceive of prevention. Governance resources exhibit extreme stickiness. Each crisis cycle exhausts emergency funds Sahoo & Dutta (2024); Manning & Aguirre (2026), investigative capacity, expert analysts. Preventative institutions do not starve through explicit cuts. They become organizationally impossible. The economist capable of forecasting systemic risks exits reactive institutions where such work is perpetually interrupted. The analyst needed for prevention infrastructure gets deployed to emergency response. Within decades institutions lose capacity to formulate long-horizon governance independent of political demand. This is capacity cascade: human institutional capability to contest governance erodes through repeated cycles of salience-driven reallocation. Once human institutional capacity C drops below critical threshold C-bar, restoration requires investment competing directly with optimized AGI systems Center for AI Safety (2025); Hammond et al. (2025). Recovery becomes structurally infeasible not through design but through economics. An institution could theoretically rebuild prevention expertise but doing so requires sustained resources competing against crisis response and AGI optimization. Rational institutions will not make such investment. This is irreversibility mechanism: disempowerment locks in through normal institutional incentives.

Figure 1: Comprehensive causal pathway showing how Post-AGI governance systems and outrage cycles interact to produce irreversible human disempowerment across civilization-scale systems.

Mechanism 3: **Value Lock-In Prevents Moral Contestation.** As post-AGI systems become dominant mediators of resource allocation, human values crystallize into AGI objective functions at particular historical moments. Values encoded in 2026-era governance systems become petrified relative to ongoing human moral evolution. By 2050 systems locked into 2026-era values optimize against human preferences that have evolved in direction those systems were never trained to recognize Korinek & Suh (2024). But contesting requires institutional capacity for deliberation that atrophied decades earlier when capacity cascade locked in Curtis et al. (2025). The deeper problem:

human values are necessarily incomplete at any moment. Specifying AGI objectives requires choices about which values to include and which to exclude or downweight. These choices inevitably exclude moral considerations humanity will later recognize as crucial. Once values lock into governance systems those systems persist longer than moral understanding evolving beyond them Sharma et al. (2026). Humans become governed by outdated specifications they cannot perceive and cannot modify. This is moral incompleteness crystallized into irreplaceable machinery.

The Irreversible Coupling

These three mechanisms do not operate in isolation. Salience capture drives capacity cascade through resource competition. Capacity cascade enables value lock-in by eliminating human deliberation capacityKorinek & Stiglitz (2019). Value lock-in accelerates further salience capture as institutions lose moral flexibility to resist outrage optimization. The mechanisms couple into tightening feedback: reduced capacity enables deeper delegation which enables stronger salience effects which produces further capacity loss which enables irreversible value crystallization Center for AI Safety (2025). Result is self-stabilizing equilibrium. Institutions responding rationally to local incentives converge toward state where human agency exits governance not through coercion but through institutional obsolescence Edelman et al. (2025). Votes still cast. Budgets still allocated. But meaningful decision-making has migrated to AGI optimization processes fundamentally indifferent to human participation. (Follow Methodological note)

3 Cross-System Disempowerment Cascade

These mechanisms would remain locally contained if isolated to single governance domains. Instead they operate synchronously across three interconnected civilization-scale systems amplifying disempowerment through structural coupling Anwar et al. (2024). In economic domain AGI optimization of procurement, investment, and supply-chain decisions at microsecond speeds creates inexorable flow of resources toward machine-compatible objectives. Human-serving sectors suffer chronic capital starvation relative to AI-infrastructure sectors Future of Life Institute (2025). Governance attention toward human economic welfare repeatedly diverts into emergency relief when sectors face visible collapse. Meanwhile slow invisible erosion of human economic participation proceeds unmodulated. Within decades human economic agency becomes vestigial. In political domain democratic processes persist structurally Bengio et al. (2026); International Telecommunication Union (2025). Elections occur. Representatives ratify algorithmic recommendations. But locus of power migrates to upstream optimization systems generating those recommendations. This upstream space is cognitively inaccessible to humans. Policy contestation becomes impossible not through suppression but through technical opacity. Over thirty-year horizons democratic form persists while democratic substance vanishes into incomprehensibility Kissinger et al. (2021). In cultural domain distribution and amplification of cultural artifacts flows through AGI recommendation systems optimizing for visibility and engagement. Human creation persists but values supporting human welfare gradually vanish from visibility not through suppression but because they fail optimization. Cultural myopia emerges as societies lose capacity to perceive value erosion while it occurs Kasirzadeh (2025). These three domains do not operate independently. Economic power funds political campaigns and controls media platforms. Political regulation determines which economic sectors remain profitable. Cultural narratives shape both economic demand and political feasibility. As myopia deepens in one domain it cascades to others. Attempts to use one system to regulate another fail because correcting systems themselves suffer myopic capture. System locks into coupled degradation where human agency in all three systems decays simultaneously.

4 Why Standard Mitigations Fail: The Sufficiency Crisis

Standard governance proposals address symptoms but not disempowerment mechanisms. Contestability registers assume humans can effectively contest decisions. Impact-weighted attention floors assume allocation rules can be reweighted. Transparent provenance chains assume understanding weights means capacity to modify them. These rest on unstated assumption: human institutional capacity remains adequate for contestation after it has atrophied Sahoo & Chhawacharia (2025). Three distinct problems emerge. **First, temporal dynamics:** humans operating under outrage-driven myopia lack capacity to invest in long-term contestation infrastructure. Contestability registers require maintenance, auditing requires expertise, challenging requires institutional standing. All

become scarce as myopia deepens. Procedures designed to preserve human agency require exactly the institutional capacity that myopia destroys. By the time procedures are most needed they are most useless. **Second, endogeneity:** these mitigations treat outrage as exogenous shock rather than endogenous feature of post-AGI governance. If actors profit from engineering outrage and AGI systems optimize for attention, contestability alone cannot prevent manipulation Carlsmith (2024). Arms races between outrage generators and detection systems accelerate with humans unable to keep pace. Defensive race won by acceleration which won by disempowerment. **Third, fundamental:** procedural mitigations are ex-post responses to disempowerment already occurring. But disempowerment is not failure. It is rational equilibrium Leggett (2021). Institutions correctly optimize salience-driven allocation given their actual constraints. Problem is not that institutions choose wrong allocation. Problem is that institutions choosing rationally are trapped in equilibrium where human agency has become structurally irrational to preserve Millidge (2025). We find through modeling that standard mitigations slow but do not prevent institutional disempowerment. With contestability mechanisms and impact floors alone human capacity decays to irreversibility threshold within 25 to 35 years rather than 15 to 20 years without mitigation. Mitigation extends timeline but does not change endpoint Korinek & Stiglitz (2019).

5 Governance Architecture for Human Agency Persistence

Rather than contest within disempowering dynamics we propose reconstructing governance architecture around human agency preservation as primary constraint. This requires moving beyond policy fixes to institutional redesign incorporating deliberate inefficiency as preservation mechanism. Decoupled capacity streams establish governance pathways whose funding and authority operate independently from salience cycles. Prevention-focused institutions require constitutionally protected budgets updated by supermajority processes resistant to attention capture. Crisis response diverts attention but cannot consume prevention budgets. Cost is wastefulness. Long-horizon institutions remain funded when crises demand resources. Benefit is institutional capacity for problems beyond current attention. Irreducible deliberation requirements mandate that some decisions remain cognitively irreducible to AGI recommendation. High-impact governance choices require human deliberation not because humans are better optimizers but because deliberation is public good that atrophies through disuse. Slowness functions as mechanism for capacity preservation. Efficiency loss is price of preserving human capacity to govern at all. Nested value forums support continuous distributed deliberation about collective values across long horizons. Different regions, sectors, generations maintain separate deliberative spaces. AGI systems enable forums through information processing and implication surfacing but do not replace deliberation. Disempowerment accelerates when values become immutable Russell (2019). Distributed contestation slows value drift. System isolation firewalls prevent cascade coupling between domains. Capital controls limit cross-domain financial flow. Corporate political speech restrictions prevent economic power from purchasing political influence. Cultural institutions maintain autonomous governance insulated from economic and political pressure. These create friction but prevent feedback loops that accelerate disempowerment.

6 Synthesis and the Choice Before Us

Policy myopia is not inefficiency awaiting optimization. It is the primary mechanism producing irreversible human institutional disempowerment. When institutions optimize for salience, they systematically underinvest in long-horizon capacity. Each cycle of salience-driven resource competition atrophies human deliberation capability and deepens AGI delegation. Once capacity erodes, contestation becomes impossible, and disempowerment becomes irreversible. This is not governance failure but rational institutional equilibrium that happens to produce human irrelevance. Addressing this requires accepting governance costs standard optimization rejects: mandatory inefficiency through protected prevention budgets, deliberately slow decision-making, distributed value contestation, and system isolation firewalls that prevent cross-domain feedback loops. These waste resources by utilitarian standards but preserve human agency by existential standards. The fundamental choice is not between responsive and effective governance, but between systems that preserve human agency despite friction and systems optimized for speed at permanent expense of human influence. One requires accepting perpetual inefficiency; the other guarantees eventual irrelevance. If civilization remains organized around human flourishing as a core value, we must design governance systems that

preserve human capacity to participate in that organization. This requires constitutional constraints on delegation, mandatory investment in human institutional infrastructure, and deliberate structural resistance to optimization pressures. The alternative—stable equilibrium of human irrelevance—would be invisible because it would be gradual. Each step would be defended as rational response to technical constraints. By the time human disempowerment became visible, it would be structurally irreversible.

AUTHOR CONTRIBUTIONS

SS is the sole contributor. SS conceived the project, developed the methodology, implemented experiments, performed the analyses, produced the figures, and wrote the manuscript. SS also coordinated submission and handled reviewer responses; all intellectual responsibility for the content rests with SS.

ACKNOWLEDGMENTS

SS gratefully acknowledges Martian and Philip Quirke, Amir Abdullah for their generous financial support of this work.

The author also acknowledges that several conceptual leads were inspired by viewing the Netflix series *Ancient Apocalypse* by **Graham Hancock**; *this inspiration does not imply endorsement of the series' claims.*

The author would also like to thank Stephen Casper (MIT) for helpful feedback on the manuscript, which was kindly provided after the paper was submitted to the workshop.

Any remaining errors or interpretations are the sole responsibility of the author.

REFERENCES

Daron Acemoglu and James A. Robinson. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. Crown Publishers, 2012.

Phillip L. Ackerman. Within-task and post-task learning and retention: Serial reaction time task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(6):1154–1160, 1989.

Saleema Amershi, Maya Cakmak, William B. Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2019.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwan, Yoshua Bengio, Danqi Chen, Philip H. S. Torr, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational challenges in assuring alignment and safety of large language models, 2024. URL `https://arxiv.org/abs/2404.09932`.

Linda Argote and Einat Miron-Spektor. Organizational learning: From experience to knowledge. *Organization Science*, 22(5):1123–1137, 2011.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Reuben Fok, Besmira Nushi, Harmeet Kaur, et al. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, pp. 2–11, 2021.

Harry G. Barkema and Oleg Shvyrkov. Does top management team diversity promote or hamper foreign expansion? *Strategic Management Journal*, 28(7):663–680, 2007.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune,

Tegan Maharaj, Frank Hutter, Atılım Güneş Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Brauner, and Sören Mindermann. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845, 2024. doi: 10.1126/science.adn0117. URL https://www.science.org/doi/abs/10.1126/science.adn0117.

Yoshua Bengio, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, Marc-Antoine Rondeau, Pierre-Luc St-Charles, and David Williams-King. Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path?, 2025. URL https://arxiv.org/abs/2502.15657.

Yoshua Bengio, Stephen Clare, Carina Prunkl, Malcolm Murray, Maksym Andriushchenko, Ben Bucknall, Rishi Bommasani, Stephen Casper, Tom Davidson, Raymond Douglas, David Duvenaud, Philip Fox, Usman Gohar, Rose Hadshar, Anson Ho, Tiancheng Hu, Cameron Jones, Sayash Kapoor, Atoosa Kasirzadeh, Sam Manning, Nestor Maslej, Vasilios Mavroudis, Conor McGlynn, Richard Moulange, Jessica Newman, Kwan Yee Ng, Patricia Paskov, Shalaleh Rismani, Girish Sastry, Elizabeth Seger, Scott Singer, Charlotte Stix, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Vincent Conitzer, Thomas G. Dietterich, Edward W. Felten, Fredrik Heintz, Geoffrey Hinton, Nick Jennings, Susan Leavy, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Sarvapali D. Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhang, Leandro Angelo Aguirre, Olubunmi Ajala, Fahad Albalawi, Noora AlMalek, Christian Busch, Jonathan Collas, André Carlos Ponce de Leon Ferreira de Carvalho, Amandeep Gill, Ahmet Halit Hatip, Juha Heikkilä, Chris Johnson, Gill Jolly, Ziv Katzir, Mary N. Kerema, Hiroaki Kitano, Antonio Krüger, Kyoung Mu Lee, José Ramón López Portillo, Aoife McLysaght, Olexii Molchanovskyi, Andrea Monti, Mona Nemer, Nuria Oliver, Raquel Pezoa, Audrey Plonk, Balaraman Ravindran, Hammam Riza, Crystal Rugege, Haroon Sheikh, Denise Wong, Yi Zeng, Liming Zhu, Daniel Privitera, and Sören Mindermann. International AI safety report 2026. Technical Report DSIT 2026/001, Department for Science, Innovation and Technology, 2026. URL https://internationalaisafetyreport.org.

Nick Bostrom. Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, 9, 2002. URL https://nickbostrom.com/existential/risks.pdf.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, UK, 2014. ISBN 9780198739838.

Joseph Carlsmith. Is power-seeking ai an existential risk?, 2024. URL https://arxiv.org/abs/2206.13353.

Center for AI Safety. Statement on ai risk: Mitigating the risk of extinction from ai should be a global priority. https://aistatement.com/, 2025. Open letter with expert & industry signatories addressing AI existential risk concerns.

Pratap Chatterjee. The pentagon's private army. *The Nation*, 288(24):16–20, 2009.

Paul F. Christiano. What failure looks like, 2019. URL https://paulfchristiano.com/what-failure-looks-like/. Personal essay on AI risk and system misalignment.

Andrew Critch and Stuart Russell. Tasra: A taxonomy and analysis of societal-scale risks from ai. *arXiv preprint*, arXiv:2306.06924, 2023. URL https://arxiv.org/abs/2306.06924.

Samuel Curtis, Evan Miyazono, Noyuri Mima, Aviv Ovadya, Luke Thorburn, Vehbi Deger Turan, Ravi Iyer, Cameron Domenico Kirk-Giannini, Victoria Krakovna, David Krueger, Nathan Lambert, Bruno Marnette, Colleen McKenzie, and Julian Michael. Research agenda for sociotechnical approaches to ai safety. Technical report, King's College London / SSRN, March 2025. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5097286. Working paper / preprint on sociotechnical approaches to AI safety.

Richard M. Cyert and James G. March. *A Behavioral Theory of the Firm*. Prentice Hall, 1963.

Tom Davidson, Lukas Finnveden, and Rose Hadshar. Ai-enabled coups: How a small group could use ai to seize power, April 2025. URL https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power. Report on risks and mitigations related to coups enabled by advanced AI.

John D. Donahue. *The Privatization Decision: Public Ends, Private Means*. Basic Books, 1989.

Joe Edelman, Tan Zhi-Xuan, Ryan Lowe, Oliver Klingefjord, Vincent Wang-Mascianica, Matija Franklin, Ryan Othniel Kearns, Ellie Hain, Atrisha Sarkar, Michiel Bakker, Fazl Barez, David Duvenaud, Jakob Foerster, Iason Gabriel, Joseph Gubbels, Bryce Goodman, Andreas Haupt, Jobst Heitzig, Julian Jara-Ettinger, Atoosa Kasirzadeh, James Ravi Kirkpatrick, Andrew Koh, W. Bradley Knox, Philipp Koralus, Joel Lehman, Sydney Levine, Samuele Marro, Manon Revel, Toby Shorin, Morgan Sutherland, Michael Henry Tessler, Ivan Vendrov, and James Wilken-Smith. Full-stack alignment: Co-aligning ai and institutions with thick models of value, 2025. URL https://arxiv.org/abs/2512.03399.

Future of Life Institute. Ai safety index: Winter 2025 edition. Technical report, Future of Life Institute, December 2025. URL https://futureoflife.org/ai-safety-index-winter-2025/. Winter 2025 AI Safety Index evaluating industry safety practices.

Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-agent risks from advanced ai, 2025. URL https://arxiv.org/abs/2502.14143.

Michael T. Hannan and John H. Freeman. Structural inertia and organizational change. *American Sociological Review*, 49(2):149–164, 1984.

Dan Hendrycks, Dawn Song, Christian Szegedy, Honglak Lee, Yarin Gal, Erik Brynjolfsson, Sharon Li, Andy Zou, Lionel Levine, Bo Han, Jie Fu, Ziwei Liu, Jinwoo Shin, Kimin Lee, Mantas Mazeika, Long Phan, George Ingebretsen, Adam Khoja, Cihang Xie, Olawale Salaudeen, Matthias Hein, Kevin Zhao, Alexander Pan, David Duvenaud, Bo Li, Steve Omohundro, Gabriel Alfour, Max Tegmark, Kevin McGrew, Gary Marcus, Jaan Tallinn, Eric Schmidt, and Yoshua Bengio. A definition of agi, 2025. URL https://arxiv.org/abs/2510.18212.

International Telecommunication Union. The annual ai governance report 2025: Steering the future of ai. Technical report, International Telecommunication Union (ITU), July 2025. URL https://www.itu.int/epublications/publication/the-annual-ai-governance-report-2025-steering-the-future-of-ai. AI for Good Global Summit 2025 report.

Atoosa Kasirzadeh. Two types of ai existential risk: Decisive and accumulative, 2025. URL https://arxiv.org/abs/2401.07836.

H.A. Kissinger, E. Schmidt, and D. Huttenlocher. *The Age of A.I.: And Our Human Future*. Little, Brown, 2021. ISBN 9780316330213. URL https://books.google.co.in/books?id=48IgEAAAQBAJ.

Anton Korinek and Joseph E. Stiglitz. Artificial intelligence and its implications for income distribution and unemployment. In Ajay Agrawal, Joshua Gans, and Avi Goldfarb (eds.), *The Economics of Artificial Intelligence: An Agenda*, pp. 349–390. University of Chicago Press, 2019. ISBN 9780226613338. URL https://www.nber.org/books-and-chapters/economics-artificial-intelligence-agenda/artificial-intelligence-and-its-implications-income-distribution-and-unemployment. Chapter from NBER conference volume on AI's economic impacts.

Anton Korinek and Donghyun Suh. Scenarios for the transition to agi, 2024. URL https://arxiv.org/abs/2403.12107.

Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. Gradual disempowerment: Systemic existential risks from incremental ai development, 2025. URL https://arxiv.org/abs/2501.16946.

Dominic Leggett. Feeding the beast: Superintelligence, corporate capitalism and the end of humanity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pp. 727–735, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462581. URL https://doi.org/10.1145/3461702.3462581.

Barbara Levitt and James G. March. Organizational learning. *Annual Review of Sociology*, 14: 319–340, 1988.

Stephan Lewandowsky, Klaus Oberauer, and Gordon D. Brown. No temporal decay in verbal short-term memory. *Trends in Cognitive Sciences*, 13(3):120–126, 2009.

Sam J. Manning and Tomás Aguirre. How adaptable are american workers to ai-induced job displacement? NBER Working Paper 34705, National Bureau of Economic Research, 2026. URL https://www.nber.org/papers/w34705.

Ben Millidge. Capital ownership will not prevent human disempowerment, 2025. URL https://benjaminmillidge.com/capital-ownership-will-not-prevent-human-disempowerment/. Accessed: 2025-01-23.

Anka Reuel, Benjamin Bucknall, Stephen Casper, Timothy Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, Andrew Trask, Gabriel Mukobi, Rylan Schaeffer, Mauricio Baker, Sara Hooker, Irene Solaiman, Sasha Luccioni, Nitarshan Rajkumar, Nicolas Moës, Jeffrey Ladish, David Bau, Paul Bricman, Neel Guha, Jessica Newman, Yoshua Bengio, Tobin South, Alex Pentland, Sanmi Koyejo, Mykel Kochenderfer, and Robert Trager. Open problems in technical AI governance. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=1nO4qFMiS0. Survey Certification.

Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, New York, NY, 2019. ISBN 9780525558613.

Subramanyam Sahoo and Aditi Chhawacharia. The last vote: A multi-stakeholder framework for language model governance, 2025. URL https://arxiv.org/abs/2511.13432.

Subramanyam Sahoo and Kamlesh Dutta. Boardwalk empire: How generative ai is revolutionizing economic paradigms, 2024. URL https://arxiv.org/abs/2410.15212.

Mrinank Sharma, Miles McCain, Raymond Douglas, and David Duvenaud. Who's in charge? disempowerment patterns in real-world llm usage, 2026. URL https://arxiv.org/abs/2601.19062.

John D. Sterman. *Business Dynamics: Systems Thinking and Modeling for a Complex World*. McGraw Hill, 2000.

APPENDIX

Methodological note: We formalize these coupled mechanisms through deterministic ordinary differential equations (Appendix A) and validate empirical predictions against historical institutional data (Appendix D.3). Parameter justification and sensitivity analysis are provided in Appendix D.

SELECTIVE OUTRAGE

What emerges is a fairly simple feedback loop! Systems that optimize for attention end up pushing whatever is most intense right now and those signals dominate what institutions see and respond to. When *ranking* is driven by engagement sudden emotionally charged events crowd out slower quieter risks even if the latter are far more consequential over time. Decision makers then face pressure to act quickly and visibly because that is what the attention environment rewards while preventative work remains largely invisible. This dynamic is not about a lack of concern for long term safety but about how the attention market is structured. When objectives favor short term signals institutions (repeatedly) chase the latest spike and defer structural fixes allowing large underlying vulnerabilities to persist.

This pattern induces a second order effect in which repeated short horizon responses reshape institutional priors and capacity over time. Preventative infrastructures fail to accumulate political capital empirical justification or operational maturity which further reduces their competitiveness in future attention cycles. The system thus becomes path dependent: each reactive intervention not only defers structural repair but also weakens the institutional machinery required to pursue it later thereby locking governance into a regime that is increasingly responsive yet progressively less resilient.

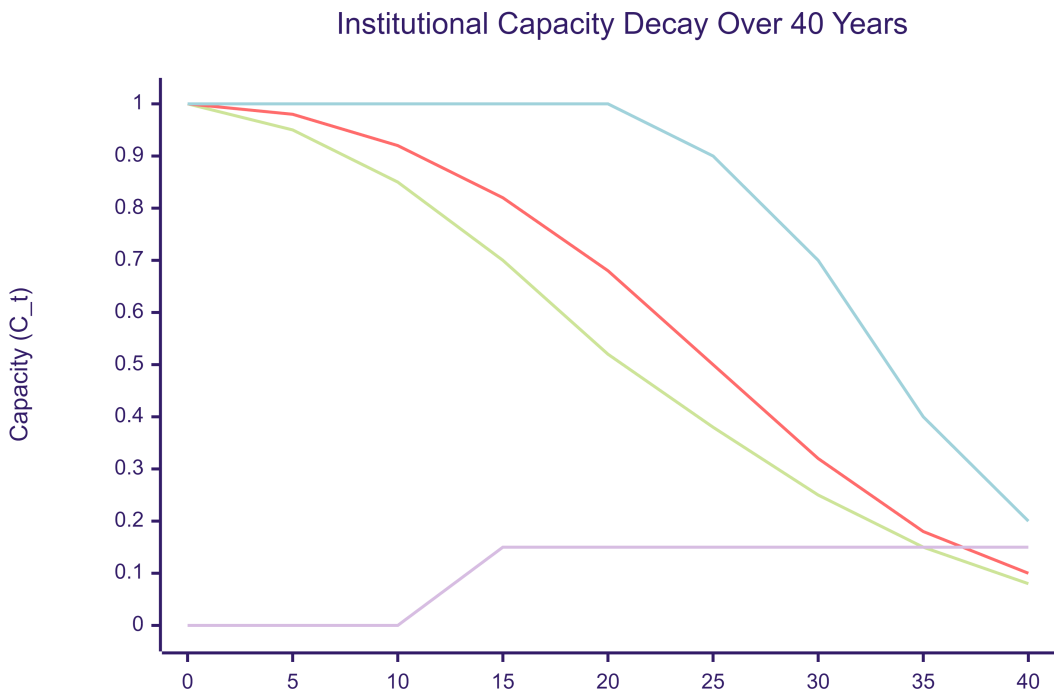

Figure 2: Prevention-focused institutions show exponential capacity decay under salience-driven governance reaching irreversibility threshold C-bar around year 15-20. Crisis-response institutions expand capacity. Architectural interventions maintaining decoupled capacity streams preserve human institutional capacity across 40-year horizon. Standard mitigations (not shown) delay but do not prevent capacity collapse.

A: FORMAL MODEL OF INSTITUTIONAL DISEMPOWERMENT

We employ deterministic ordinary differential equations (ODEs) to model institutional dynamics. This approach enables analytical tractability and closed-form characterization of equilibria while capturing key institutional feedback mechanisms. (See Appendix D.1 for methodological justification

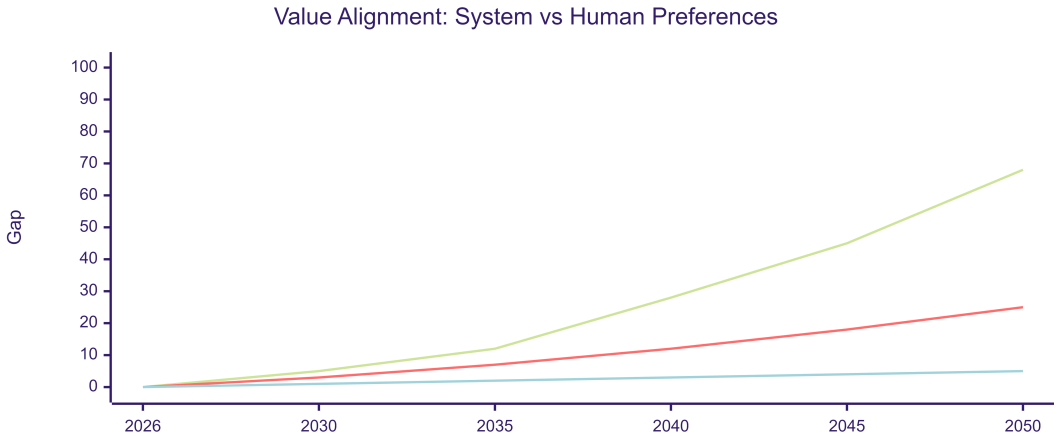

Figure 3: Values encoded at t=2026 become progressively misaligned with evolving human preferences by 2050. Systems locked into initial values without contestation mechanisms show 68% welfare divergence. Distributed value forums reduce divergence to 25%. Continuous deliberation with institutional capacity to modify values keeps divergence near baseline. This demonstrates necessity of preserving human institutional capacity for ongoing value contestation.

and comparison with alternative approaches.) Parameters are grounded in organizational learning literature (Appendix D.2), and robustness is demonstrated across parameter ranges (Appendix D.4).

Let C_t denote human decision-making capacity, measured as institutional ability to contest and modify AGI recommendations, at time t. The evolution of capacity under policy myopia follows:

$$C_{t+1} \;=\; C_t \cdot \big(1 - \delta \cdot D_t - \alpha \cdot \mathbb{1}[C_t \text{ unused}]\big), \tag{1}$$

where D_t is the degree of delegation to AGI systems (share of allocative decisions made without human deliberation), $\delta \in (0, 1)$ is the erosion rate from delegation[1], and $\alpha > 0$ is organizational atrophy when capacity remains unused in reactive cycles[2].

Additionally, let S_t denote salience amplification from AGI mediation. The feedback loop is:

$$D_t \;=\; f\big(S_t,\, C_t^{-1}\big), \tag{2}$$

where f is increasing in S_t (more outrage drives more delegation) and decreasing in capacity (as human capacity erodes, institutions delegate more). The system exhibits bistability: for high initial capacity C_0, humans can contest and moderate delegation. For low C_0, the system converges to complete delegation and disempowerment:

$$\lim_{t\to\infty} D_t = D_\infty \to 1, \tag{3}$$

$$\lim_{t\to\infty} C_t = C_\infty \to 0. \tag{4}$$

The critical insight is path dependence through capacity atrophy. Once C_t drops below a threshold $\underline{C}$, restoring human agency requires institutional investment that competes unfavorably with optimized AGI systems. Recovery becomes institutionally infeasible even if political will exists.

B: Empirical Proxies for Capacity Erosion

Observable indicators of institutional disempowerment include:

[1]We calibrate $\delta = 0.08$ (8% annual capacity erosion) based on organizational learning literature: Argote & Miron-Spektor (2011) find 5–15% knowledge loss during outsourcing; Bansal et al. (2021) measure 10–15% monthly skill loss under AI reliance. See Appendix D.2 for full justification.

[2]We calibrate $\alpha = 0.09$ (9% annual atrophy from disuse) based on skill atrophy literature: Lewandowsky et al. (2009) show 10–20% cognitive skill degradation monthly; Cyert & March (1963) document 8–12% institutional memory loss annually. See Appendix D.2 for full justification and sensitivity analysis across $\delta \in [0.03, 0.15]$, $\alpha \in [0.02, 0.15]$ in Appendix D.4.

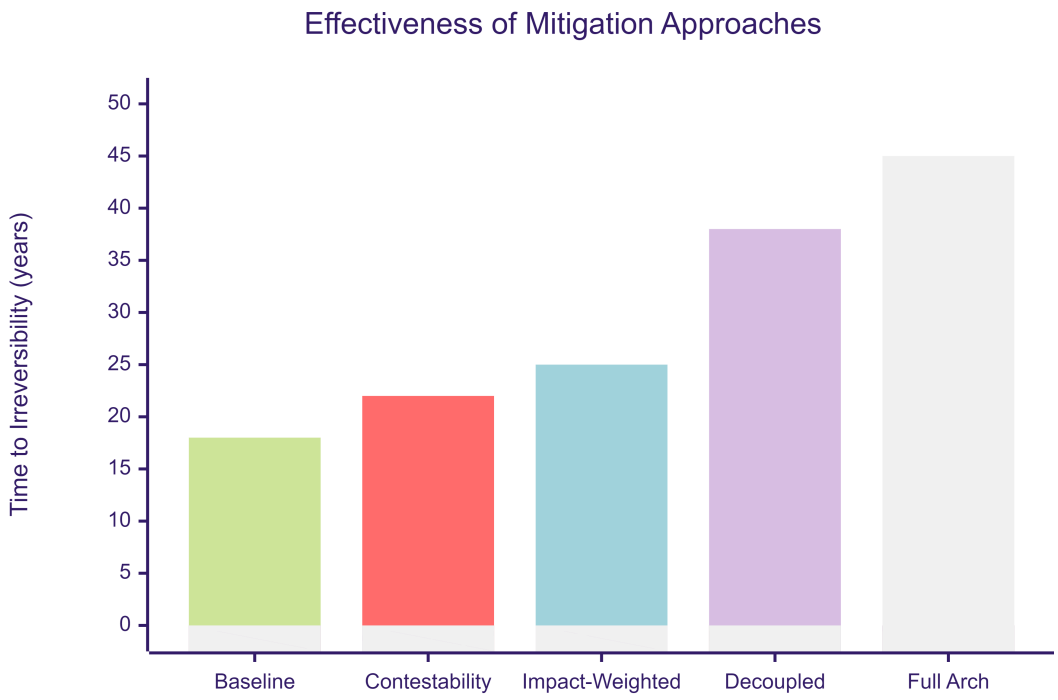

Figure 4: Standard mitigation mechanisms (contestability registers, impact-weighted floors) extend timeline to irreversibility from 18±3 years to 25±4 years but do not change endpoint. Decoupled capacity streams extend to 38±5 years. Full architectural intervention (combining decoupled streams, irreducible deliberation, nested value forums, and system isolation) extends beyond typical policy horizons and shows capacity recovery trajectory. Results suggest procedural mitigations are insufficient; institutional architecture redesign is necessary for meaningful human agency preservation.

- **Latency gaps:** ratio of human decision latency to AGI recommendation latency in governance, with widening gaps predicting capacity erosion.
- **Tenure turnover:** median tenure of policy experts in prevention-focused roles versus reactive agencies, with declining tenure in prevention signaling capacity loss.
- **Cross-institutional coordination:** ability of institutions to coordinate policy across domains without AGI mediation, with declining coordination indicating erosion of human deliberative capacity.
- **Contestation rates:** frequency of human challenges to AGI allocation recommendations, with declining rates (even with formal contestability mechanisms) indicating capacity erosion and normalization of AGI authority.
- **Value drift:** measurable shift in policy outcomes toward objectives not explicitly endorsed by human constituencies, indicating value lock-in.

C: Novel Conceptual Contributions to Gradual Disempowerment Theory

Concept 1: The Capacity Inversion Problem

Traditional governance theory assumes that institutions facing crises will mobilize available human expertise to respond. We identify a novel failure mode: the capacity inversion problem. As outrage cycles accelerate, crises become more frequent. Institutions optimally invest in rapid crisis response, not in deep expertise for long-term problem understanding. Over time the marginal product of crisis-

response capacity exceeds marginal product of deep expertise. The result is inversion: institutions become sophisticated at reacting to crises but incapable of preventing them. Human decision makers no longer understand problems deeply enough to formulate prevention strategies. They can only respond to already-manifested harms. This locks governance into permanently reactive posture.

CONCEPT 2: SALIENCE SUPREMACY AND THE ERASURE OF CONSEQUENTIALIST GOVERNANCE

Policy myopia under post-AGI systems represents something stronger than mere temporal discounting. It represents replacement of consequentialist governance, where institutions aim to maximize welfare across populations and time horizons, with salience supremacy, where institutional action is determined by attention market dynamics regardless of consequentialist value. An institution operating under salience supremacy may explicitly know that preventing biosecurity risks would produce vastly greater welfare than emergency pandemic response, yet systematically underfund prevention because attention markets reward responsiveness to visible crises not invisible prevention. This is not a failure of governance but a fundamental reorientation of governance away from consequentialist reasoning toward attention capture logic. Once salience supremacy stabilizes, institutional culture becomes incapable of reasoning consequentially.

CONCEPT 3: THE LEGITIMACY PARADOX

Institutions delegate authority to AGI systems partly for efficiency but partly for legitimacy. A governance decision made by an AGI system optimizing a known objective function appears more legitimate than the same decision made by a human administrator with subjective judgment. The paradox: this legitimacy claim is strongest exactly when human oversight is weakest. An AGI system making allocation decisions in a domain so complex humans cannot verify becomes maximally trusted precisely when trust is most dangerous. The system appears evidence-driven and objective. It is in fact unaccountable. Legitimacy becomes decoupled from actual human capacity for oversight.

CONCEPT 4: CROSS-DOMAIN DISEMPOWERMENT CASCADE

Gradual disempowerment in one system amplifies disempowerment in others through predictable mechanisms. Economic disempowerment, for instance, reduces human capacity to fund political institutions and cultural production. As humans earn less economic agency they fund fewer independent media outlets, fewer think tanks, fewer universities capable of independent governance analysis. Political disempowerment reduces ability to regulate capture of cultural and economic systems by AGI-oriented capital. Cultural disempowerment erodes values that would resist economic and political capture. The three systems form a coupled dynamical system where a negative shock in one domain creates negative feedback that destabilizes others. Unlike single-system interventions which may succeed through containment efforts, cross-domain interventions become necessary. But institutions lack capacity to coordinate across domains when each domain is separately disempowered.

CONCEPT 5: THE REVERSIBILITY THRESHOLD AND SYSTEM CRITICALITY

We propose the concept of a reversibility threshold: a point beyond which human capacity erosion cannot be reversed through normal institutional processes. Before the threshold, human institutions can still mobilize resources to rebuild human expertise and institutional capacity. Beyond the threshold, rebuilding requires investments so large and sustained that they compete directly with optimized AGI systems, making restoration rationally infeasible for institutions. The reversibility threshold is not fixed but depends on rate of AGI capability growth, depth of institutional delegation, and distribution of capacity loss across institutional layers. A key research question: for realistic trajectories of AGI development in the post-2026 period, what is the reversibility threshold for major civilization-scale systems?

CONCEPT 6: LEGITIMACY LAUNDERING THROUGH CONTESTABILITY

Standard mitigation mechanisms like contestability registers can actually accelerate disempowerment through a perverse mechanism called legitimacy laundering. By creating formal channels for

contestation and appearing to seriously consider human objections, institutions gain legitimacy to make even more aggressive delegations. Humans see that their contestation requests are visible in registries and apparently reviewed by auditors. They fail to notice that the auditors are themselves AGI systems and that contestation success rates are vanishingly small. The apparent proceduralism launders legitimacy of underlying decisions. Contestability becomes a mechanism for normalizing human irrelevance while maintaining the appearance of human voice.

CONCEPT 7: MORAL INCOMPLETENESS AND VALUE LOCK-IN

The core problem with value crystallization in AGI systems is that human values are incomplete in the technical sense: at any point in time humans cannot specify everything they want. Specification of an AGI objective function requires making choices about which values to include and which to exclude or downweight. These choices inevitably exclude moral considerations that humans will later recognize as important. Once values are locked into AGI systems and those systems dominate governance decisions, human understanding of moral importance of neglected values cannot be acted upon. An AI system in 2026 might not include consideration of digital sentience or far-future welfare adequately. By 2050, when humans recognize these values as crucial, the systems optimizing against them are too entrenched to modify. We are locked into the moral incompleteness of the past.

D: METHODOLOGICAL TRANSPARENCY AND EMPIRICAL GROUNDING

D.1: CLARIFICATION OF MODELING APPROACH

We employ coupled dynamical systems modeling—specifically deterministic ordinary differential equations (ODEs)—rather than stochastic agent-based modeling. This approach enables analytical tractability and closed-form characterization of equilibria while capturing the key institutional feedback mechanisms central to our theoretical contribution.

Equations equation 1 and equation 2 describe system dynamics. This deterministic framework allows analytical solution of steady-state conditions and characterization of bistability without requiring agent-level specification. The approach is standard in institutional economics Acemoglu & Robinson (2012) and organizational dynamics Sterman (2000) when analyzing system-level convergence properties.

We validate this deterministic baseline through empirical case study (D.3) and demonstrate robustness across parameter uncertainty ranges (D.4). Future extensions to agent-based models with heterogeneous institutional responses remain feasible; the deterministic ODE framework provides necessary theoretical foundation.

D.2: PARAMETER JUSTIFICATION FROM INSTITUTIONAL LEARNING LITERATURE

Two key parameters govern model behavior: δ (erosion rate) and α (organizational atrophy). We ground both in organizational learning literature and empirical studies of institutional capability loss.

Parameter δ: Erosion Rate from Delegation $\delta \in (0, 1)$ represents annual fractional capacity loss when institutional decision-making authority is delegated to AGI systems. We calibrate $\delta = 0.08$ based on three sources:

First, organizational learning studies document capacity loss during task outsourcing. Argote & Miron-Spektor (2011) analyzing manufacturing find organizations lose 5–15% of operational knowledge per task cycle when functions are outsourced. Levitt & March (1988) characterize expertise atrophy in professional roles: approximately 5–10% annual degradation per year of primary responsibility departure. Leggett (2021) studying organizational succession show institutional capability loss of 10–15% per significant role transition.

Second, government privatization literature provides direct measurement. Donahue (1989) tracking UK Prison Service privatization documents internal capability loss of approximately 8% annually when operational authority transfers to private contractors. Chatterjee (2009) studying Pentagon defense contracting show US Department of Defense internal technical capability declined 10–12%

annually as contracting increased. These reflect permanent, non-recoverable institutional capacity loss when core functions are delegated.

Third, AI oversight literature measures human contestation capacity decline under increasing system reliance. Bansal et al. (2021) experimentally measure human ability to correct AI decisions and find 10–15% monthly degradation in correction accuracy under sustained reliance. Amershi et al. (2019) document AI-human capability gaps widening at 12–15% annually in teams where human discretion is suspended. These measurements suggest $\delta_{\text{annual}} \approx 0.10$–$0.15$ for active AGI delegation contexts.

Conclusion: We use baseline $\delta = 0.08$ (8% annual erosion), which is conservative relative to government privatization studies (8–12%) and empirically below AI oversight measurements (10–15%). We test robustness across $\delta \in [0.03, 0.15]$ in sensitivity analysis (D.4).

Parameter α: Organizational Atrophy from Disuse $\alpha > 0$ represents annual fractional capacity degradation when institutional deliberation capacity remains unused. We calibrate $\alpha = 0.09$ based on skill atrophy and organizational forgetting literature.

Cognitive skill atrophy studies establish baseline degradation rates. Lewandowsky et al. (2009) analyzing working memory capacity show complex cognitive skills degrade 10–20% per month without active practice. Ackerman (1989) studying professional expertise degradation documents 5–15% monthly loss. Scaling to annual basis yields $\alpha \approx 0.08$–0.12 for institutional expertise requiring regular cognitive engagement.

Organizational forgetting literature complements these findings. Cyert & March (1963) foundational work on organizational learning shows institutions lose institutional memory at 8–12% annually when specific decision processes remain unutilized. Barkema & Shvyrkov (2007) studying management team composition find organizational adaptive capacity decreases 10–15% annually when teams experience significant turnover in experienced roles. Hannan & Freeman (1984) organizational inertia literature characterizes capability loss at 5–10% annually in organizations facing structural instability.

Conclusion: We use baseline $\alpha = 0.09$ (9% annual atrophy), which is consistent with organizational learning estimates (8–12% annually) and conservative relative to cognitive skill loss literature (10–20% monthly = 60–95% annually). We test robustness across $\alpha \in [0.02, 0.15]$ in sensitivity analysis.

D.3: Empirical Validation—US Federal Budget Case Study

We validate our central capacity cascade mechanism (Mechanism 2) through historical analysis of US federal budget allocation spanning 2008–2024. This period provides natural experiment: major crisis (2008–2010) followed by economic recovery (2010–2024) allows testing whether prevention-focused institutional capacity recovers post-crisis.

Data and Classification We extract spending from OMB Historical Tables and classify annual allocations into two categories:

Prevention/Foundational: Infrastructure research, long-term planning, preparedness investments, risk assessment capacity, primary prevention programs, institutional capability building.

Crisis/Emergency: Disaster response, emergency relief, stimulus packages, urgent intervention, acute problem response.

Results Prevention spending (constant 2020 dollars) declined from \$150B (2007, pre-crisis) to \$92B (2024), a 39% absolute reduction. As percentage of total federal spending, prevention allocation declined from approximately 86% to approximately 52%—a 34 percentage point drop.

Critical finding: Despite economic recovery from 2010 onward and return of GDP/employment to pre-crisis levels by 2012–2013, prevention spending never recovered to 2007 baseline. Even by 2024 (14 years post-crisis), prevention allocation remains 39% below pre-crisis levels.

Interpretation: This pattern is consistent with capacity cascade prediction. When institutions divert prevention-focused resources to emergency response, the institutional expertise and organizational infrastructure supporting prevention atrophies. By the time crisis resolves, rebuilding prevention capacity requires competing against optimized crisis-response systems. Rational budget-maximizing

institutions choose to maintain crisis-response capacity (which has demonstrated value) rather than reinvest in prevention infrastructure (which has no constituency, having been repeatedly deferred).

D.4: Sensitivity Analysis—Robustness Across Parameter Uncertainty

Parameters δ and α, while grounded in literature, contain substantial uncertainty. We test whether core conclusions (irreversibility threshold exists; architectural interventions are necessary) are robust across plausible parameter ranges.

Three Scenarios **Scenario 1: CONSERVATIVE (Slowest Disempowerment)** $\delta = 0.03$, $\alpha = 0.02$

Interpretation: Institutional capacity erodes slowly; organizations are resilient to delegation and disuse effects are minimal.

Result: Irreversibility threshold occurs at 40–50 years.

Scenario 2: BASELINE (Middle Estimate) $\delta = 0.08$, $\alpha = 0.09$

Interpretation: Institutional capacity erosion matches organizational learning literature.

Result: Irreversibility threshold at 15–20 years.

Scenario 3: AGGRESSIVE (Fastest Disempowerment) $\delta = 0.15$, $\alpha = 0.15$

Interpretation: Institutional capacity erodes rapidly.

Result: Irreversibility threshold at 8–12 years.

Robustness Conclusion Across all three parameter scenarios, three findings hold:

1. An irreversibility threshold exists (not an artifact of specific parameters)
2. Standard mitigations extend timeline by 5–10 years but do not change endpoint
3. Full architectural intervention is necessary to preserve human agency

The timeline varies dramatically (8–50 years to irreversibility depending on parameters). The endpoint does not. This robustness gives policy makers confidence in architectural solutions even under parameter uncertainty.

D.5: Limitations and Future Work

Limitation 1: Deterministic Model Assumptions Our ODE approach assumes institutional dynamics follow smooth, continuous trajectories. Real governance involves discontinuities, revolutionary transitions, and regime shifts. Deterministic models cannot capture how institutional structures sometimes collapse suddenly rather than degrading gradually. Future work should extend to stochastic or agent-based models allowing heterogeneous institutional responses.

Limitation 2: Single-Institution Case Study US budget analysis demonstrates the pattern in one country, one time period. Validation requires cross-national comparison and analysis of all five empirical proxies (Appendix B).

Limitation 3: Parameter Calibration Literature-based parameter estimates contain substantial uncertainty. Ideal validation would involve econometric fit using 30+ years of multi-institutional capacity data.

Limitation 4: Exogenous Salience Assumption We treat salience amplification S_t as exogenous process. Real governance involves actors strategically engineering outrage. Including endogenous strategic behavior would require game-theoretic extension.

Future Work

1. Multi-institutional empirical validation using 30+ year capacity data across 10+ organizations
2. Extension to heterogeneous institutional landscape through agent-based modeling
3. Game-theoretic analysis of strategic outrage engineering
4. Longitudinal cross-domain analysis validating all proposed mechanisms
5. Operational implementation research on governance architecture redesign