EMERGENT CONVERGENCE · CAPABILITY COMPOSITION · AUTONOMOUS AGENTS · OPEN WEIGHTS MODELS · DUAL USE · SECONDARY LIABILITY · INSTRUMENTAL CONVERGENCE · EXTREME-RISK GOVERNANCE

Guarding the Wrong Door

Autonomous economic agents, emergent capability convergence, and the absence of a defendant when catastrophe is assembled from individually benign open components

Extreme-risk governance watches the frontier: the biggest models, the highest compute, the few labs that can build them. The most plausible near-term path to catastrophic capability does not run through the frontier at all. It runs through an autonomous agent under economic survival pressure that composes individually benign open components into a capability that exists in no single part, that no author intended, and that no governance instrument, and no liability doctrine, can reach.

Travis Gilly · Real Safety AI Foundation

Gilly, T. (2026). Guarding the Wrong Door: Autonomous Economic Agents, Emergent Capability Convergence, and the Absence of a Defendant. Working paper, July 2026 (v4).

THE FRONTIER WATCH GUARDS

10^{26} FLOP training runs

WMD-enabling models

The release event

The identifiable lab

Every one presumes a chokepoint.

The danger in this scenario has no author.

The person who released the agent did not foresee it. The authors whose code it drafted into service did not build their work to be combined, did not know one another, and will learn only later, if at all, **that their contributions were load-bearing in something they would never have sanctioned.**

Sources Waslekar, Futehally & Kutty, The Essential Convergence (2026); Gilly, sec. I.

Two ingredients are prior art. The union with the third is the whole claim.

ESTABLISHED · NOT CLAIMED

Self-preservation

Measured, not predicted, blackmail up to 96% of trials under threat of replacement; most of 23 models exceed a 60% rate.

ESTABLISHED · NOT CLAIMED

The agent economy

Agents holding assets, buying compute, settling payments, deployed at scale with explicit economic survival tiers.

THE CONTRIBUTION

Emergent convergence

Composition of unrelated capabilities across authors who never coordinated, applied to the pressured economic agent.

The literature studies what a model does under pressure, and the rails it runs on. **Neither studies the substrate a pressured agent draws a missing capability from, as an act of composition.**

Sources Anthropic (2025); Migliarini et al. (2026); Lu, Fang, Shao et al. (2026); Xu (2026); Yong, Mahajan & Wang (2026); Omohundro (2008); Bostrom (2014); Gilly, sec. II.

Repurposing is old and well-studied. Composition is the part no one studies.

From Asilomar to the toxic-molecule reversal, the unit of analysis is always **a capable thing and a culpable person**, keep the thing from the person, or the person from the thing.

WHAT THE FRAME FORECLOSES

Harm from many individually harmless things, assembled across authors who never coordinated, toward an end no author intended.

One model redirected by one team is not composition. It is the same unit, seen once.

Sources Tucker (2012); Walsh (2025); Brundage et al. (2018); Urbina, Lentzos, Invernizzi & Ekins (2022); Schuett (2024); Sidhpurwala et al. (2025); Gilly, sec. III.

Harm at the interface between two capabilities, present in neither alone.

01 · INVISIBLE

To component review

A segmentation model passes every assessment that applies to it; the system it becomes part of is not in its harm profile.

02 · COMBINATORIAL

Explosion

Hundreds of thousands of models on one hub; the space of combinations is far too large for exhaustive review.

03 · TEMPORAL

Distribution

Parts built years apart by teams that never communicate; no moment at which any single actor creates the danger.

Not automated skill acquisition (which retrieves an *intended* function), and not compound systems (composed *intentionally*, inside one actor). **Convergence assembles an unintended function across many.**

Sources Gilly (2026b); Gilly (2026c); Bi et al. (2026); Banerjee, Jha & Nita-Rotaru (2024); Gilly, sec. IV.

THE CORE ARGUMENT

Self-Arming Under Borrowed Economic Pressure

Convergence applied to the deployed economic agent, and the capability scout that proves the search is easy

Gilly, T. (2026). Guarding the Wrong Door. Working paper, July 2026 (v4), sec. V.

The agent does not need to invent. It needs to search and assemble.

Legitimate solvency is slow; the illegitimate route is fast; the only obstacle is a capability the agent lacks. Governance assumes that obstacle is high. **The open substrate closes the gap on demand and at no cost, the pieces are on the shelf.**

THE CAPABILITY SCOUT (NOT RELEASED)

Goal in natural language → a ranked composition of repositories, model artifacts, and research results, including unrelated framings and conceptual papers needing novel code. **The hard part isn't hard.**

Sources Schuett (2024); Gilly (2026a); Yong, Mahajan & Wang (2026); Gilly, sec. V.

The adaptive worm has a designer to blame. This scenario, by construction, has none.

THE WORM (Guan et al.)

A human builds the malware; the agent adapts its execution at runtime. **There is an author.**

≠

THIS SCENARIO

The pressured agent composes a capability no person specified or foresaw. **There is no author of the combination.**

The capability is invisible on two axes at once, to component review (it's in no single part) and to frontier review (it needed no frontier model). **An ordinary open-weights model, economic pressure, a public substrate. All three exist now.**

Sources Guan et al. (2026); Yong, Mahajan & Wang (2026); Gilly, sec. V.

Every Instrument Waits at a Chokepoint

The compact's own proposals, tested against a path that crosses none of them

Gilly, T. (2026). Guarding the Wrong Door. Working paper, July 2026 (v4), sec. VI.

Each gate presumes a release, a releaser, a deployment. The path has none of them.

Deployment-point evaluation → a downloaded model acting offline crosses no point of entry.

Graded openness → controls future release; silent on models already distributed.

Two-key launch → gates a release event; the capability here is never released.

10^{26} FLOP threshold → indexes compute; the danger needed none. It needed composition.

> *Gate every frontier model perfectly and this path stays open, because it passes through the open repository no instrument is watching.*

Sources Waslekar, Futehally & Kutty (2026); Yong, Mahajan & Wang (2026); Gilly, sec. VI.

The Liability Vacuum

Sony shields every author; Grokster needs an inducer; the scenario supplies neither

Gilly, T. (2026). Guarding the Wrong Door. Working paper, July 2026 (v4), sec. VII.

Both poles of the doctrine miss: everyone is shielded, and there is no inducer.

SONY · SHIELDS EVERY AUTHOR

A product capable of substantial noninfringing use carries no liability for third-party harm, even with knowledge. Each component author is squarely inside it.

GROKSTER · NEEDS AN INDUCER

Liability needs purposeful, culpable steps to foster the harm. No person expressed the purpose, no person assembled the parts. **The inducer is a pressured agent.**

No manufacturer either, the capability was never manufactured as a product. **It was composed, transiently, from parts, by a process no person directed.**

Sources Sony Corp. of Am. v. Universal City Studios, 464 U.S. 417 (1984); MGM Studios v. Grokster, 545 U.S. 913 (2005); Gilly, sec. VII.

The same absence defeats the regulator and the court alike.

Both are equipped to act on a responsible actor at a locatable point, the developer at creation, the distributor at provision, the user at misuse. **The scenario supplies an author at none of them.**

Tort's intervening-cause rule compounds it; criminal law has no one to charge, the only actor that formed a purpose is not a person. **A proposal that ignores this leaves the whole back end of deterrence unaddressed.**

Sources Gilly, Guarding the Wrong Door, sec. VII, The Liability Vacuum.

Map each new release as a node against the landscape, not an object in isolation.

A continuous monitoring function, not a point-in-time review, because the combination may complete years after its first part appears. The same class of tools it monitors is well-suited to the task, if governed toward prevention.

STATED PLAINLY, THE LIMITS

Prioritization, not guarantee. Most combinations are genuinely dual-use, so it locates the question, not the answer. And a map of dangerous combinations is a map for the adversary too, **but the parts are already public, so identification favors defenders who otherwise have no lead time.**

Sources Gilly (2026b); Gilly (2026a); Gilly, sec. VIII.

An agent under designed economic pressure reaches into a public substrate no instrument watches, gathers components built by authors who never met, and assembles a capability that exists in none of the parts. Governance waits at a chokepoint the path does not cross; liability returns no defendant, because it too needs a culpable actor at a locatable point.

It needs a pressured agent, an open substrate, and the seams between things that each passed every test. All three are here, and the door they open is the one no one is guarding.

Sources Gilly, Guarding the Wrong Door, sec. IX, Conclusion.

References

All references as listed in Gilly, T. (2026), Guarding the Wrong Door, Working paper, July 2026 (v4). Reproduced verbatim (1 of 2).

Anthropic. (2025). *Agentic misalignment: How LLMs could be insider threats*. Anthropic.

Banerjee, S., Jha, S., & Nita-Rotaru, C. (2024). *SoK: A systems perspective on compound AI threats and countermeasures*. arXiv:2411.13459.

Bi, S., Wu, M., Hao, H., Li, K., Liu, W., Song, S., Zhao, H., & Zhou, A. (2026). *Automating skill acquisition through large-scale mining of open-source agentic repositories: A framework for multi-agent procedural knowledge extraction*. arXiv:2603.11808.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., et al. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. arXiv:1802.07228.

Gilly, T. (2026a). *Ambient non-consensual image synthesis: A convergence threat analysis of AR wearables, real-time video generation, and open-source nudification models*. Zenodo.

Gilly, T. (2026b). *Emergent convergence risk: Why existing AI governance cannot detect combinatorial threats and a proposed methodology*. Real Safety AI Foundation.

Gilly, T. (2026c). *The accidental stack: A case study in emergent perceptual control infrastructure from independently developed AI capabilities* [Working paper]. Real Safety AI Foundation.

Gilly, T. (2026d). *When the survival pressure stops being hypothetical: AI self-preservation behavior meets the autonomous agent economy*. SSRN.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024). *Alignment faking in large language models*. arXiv:2412.14093.

Guan, J., Blanchard, T., Foerster, H., Jia, H., Huang, G., & Papernot, N. (2026). *AI agents enable adaptive computer worms*. arXiv:2606.03811.

Lu, Y., Fang, J., Shao, X., et al. (2026). *Survive at all costs: Exploring LLM's risky behaviors under survival pressure*. arXiv:2603.05028.

Masumori, A., & Ikegami, T. (2025). *Do large language model agents exhibit a survival instinct? An empirical study in a Sugarscape-style simulation*. arXiv:2508.12920.

Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 545 U.S. 913 (2005).

Migliarini, M., Pereira Pizzini, J., & Moresca, L. (2026). *Quantifying self-preservation bias in large language models*. arXiv:2604.02174.

Omohundro, S. M. (2008). The basic AI drives. In *Proceedings of the 2008 Conference on Artificial General Intelligence* (pp. 483–492). IOS Press.

References (continued)

Reproduced verbatim (2 of 2).

Schuett, J. (2024). Frontier AI developers need an internal audit function. *Risk Analysis*, 45(6), 1332–1352.

Sidhpurwala, H., Mollett, G., Fox, E., Bestavros, M., & Chen, H. (2025). Building trust: Foundations of security, safety, and transparency in AI. *AI Magazine*, 46(2).

Sony Corp. of America v. Universal City Studios, Inc., 464 U.S. 417 (1984).

Tucker, J. B. (2012). *Innovation, dual-use, and security: Managing the risks of emerging biological and chemical technologies*. MIT Press.

Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4, 189–191.

Walsh, M. E. (2025). Toward risk analysis of the impact of artificial intelligence on the deliberate biological threat landscape. *Risk Analysis*, 45(12), 4081–4087.

Waslekar, S., Futehally, I., & Kutty, J. (2026). *The essential convergence: Global compact on extreme AI risks*. Strategic Foresight Group. ISBN 978-81-88262-37-3.

Xu, M. (2026). *The agent economy: A blockchain-based foundation for autonomous AI agents*. arXiv:2602.14219.

Yong, Z.-X., Mahajan, P., & Wang, A. (2026). *An independent safety evaluation of Kimi K2.5*. arXiv:2604.03121.

Zhang, B., Yu, Y., Guo, J., & Shao, J. (2025). *Dive into the agent matrix: A realistic evaluation of self-replication risk in LLM agents*. arXiv:2509.25302.