REAL SAFETY AI FOUNDATION · WORKING PAPER

EMERGENT CONVERGENCE RISK · COMBINATORIAL THREATS · AI GOVERNANCE · RISK ASSESSMENT · DETECTION METHODOLOGY · AI SAFETY · SYSTEMIC RISK · GOVERNANCE GAP

Emergent Convergence Risk

Why existing AI governance cannot detect combinatorial threats, and a proposed methodology.

Travis Gilly · Real Safety AI Foundation

Working paper · 2026

We inspect chemicals one at a time while the warehouse fills with reactive combinations.

No framework scans for harm that emerges when independently developed capabilities combine. **The danger lives at the integration surface**, between tools that each pass every applicable risk assessment.

Three companion case studies are three instances of one structural phenomenon, invisible to frameworks that evaluate capabilities in isolation.

SOURCES Gilly (2026a, 2026b, 2026c, 2026d), the three companion convergence-threat case studies.

Not three separate problems. Three instances of one structural phenomenon.

Ambient nudification

AR wearables + nudification models + real-time video generation + edge inference.

Therapeutic perceptual modification

The same convergence stack, routed through existing digital-therapeutics regulatory pathways.

Perceptual-control infrastructure

Eight individually benign open-source tools, shown in one roundup video, collectively enabling population-scale control.

SOURCES Gilly (2026a, 2026b); Gilly (2026c); Gilly (2026d).

Every framework evaluates one system in isolation. None asks: A + B = ?

EU AI Act
Risk class per system, by intended purpose.

NIST AI RMF
Four functions on a single bounded system.

MIT Risk Repo
Risks cataloged per capability.

ISO/IEC 42001
Management system around specific products.

The unexamined assumption: evaluating individual components is enough to ensure system-level safety. For capabilities with compatible interfaces, it is false.

SOURCES EU Regulation 2024/1689 · NIST AI 100-1 (2023) · MIT FutureTech AI Risk Repository (2024) · ISO/IEC 42001:2023.

Three silos held the paradigm up. All three have collapsed.

The combinatorial space went from **a tractable number of expensive connections to an effectively infinite number of cheap ones.** The paradigm built for the first world is running in the second.

Modality
collapsed, multimodal by default.

Compute
collapsed, models of **4.8–140 MB** on consumer phones.

Access
collapsed, **300,000+** open model repos on HuggingFace alone.

CONSEQUENCE Integration is now cheap, easy, and open, which is precisely the world the paradigm was not designed for.

It hides, it explodes, and it has no moment of creation.

PROPERTY 01

Invisible to component eval

The harm profile exists only at the surface, never in either tool alone.

PROPERTY 02

Combinatorial explosion

N(N−1)/2

pairs across 300,000+ models, exhaustive evaluation is infeasible.

PROPERTY 03

Temporally distributed

Components built years apart, by teams that never meet. No single actor creates the danger.

SOURCES No tool is misused. Each is used for its intended purpose, the harm arises from composition. Banerjee, Jha & Nita-Rotaru, arXiv:2411.13459 (2024).

Chemistry built a discipline for what substances do when combined. AI governance has not.

CHEMISTRY HAS	AI HAS	AI LACKS
SDS · GHS · reaction chemistry	**Model cards ≈ SDS · AI Act ≈ GHS**	**Reaction chemistry**
Incompatibility lists; the discipline of what reacts with what.	Per-substance properties, and nothing more.	The discipline of what capabilities produce combined. The gap is disciplinary, not informational.

The existence of the discipline is itself the safety mechanism.

SOURCES United Nations, GHS (ST/SG/AC.10/30, Rev. 10, 2023); Mitchell et al., model cards, FAccT 2019.

A reaction-chemistry for AI, in four steps.

1
Capability mapping
By I/O format, latency, size, accessibility, capability classes, not every model.

→

2
Surface identification
Filter: perceptual modification? consent bypass? scale amplification?

→

3
Harm analysis
Who's hurt at the surface → a convergence risk profile (an SDS listing).

→

4
Monitoring
New release → mapped → surfaces re-evaluated. Continuous, partly automatable.

SOURCES Extends the Harm Blindness Framework (Gilly, 2025-2026) from product-level checkpoints to cross-capability surfaces.

Combinatorial scale	Surfaces may exceed capacity. Prioritization heuristics help, but any scheme misses some.
Dual-use ambiguity	The same surface enables manipulation *and* accessibility. ISA flags it; policymakers decide.
Implementation gap	A methodology, not a built system. The prototype is early-stage; we do not overstate it.
Adversarial dynamics	Publishing profiles is a roadmap, but the components are already public. **Identification gives defenders lead time.**

SOURCES Gilly, T. (2026), Emergent Convergence Risk, sec. VII (Limitations).

IN CLOSING

We found three, with informal methods, from one week of releases in one video.

The discipline we propose does not exist and should. Either governance develops the capacity to find these convergences before they are exploited, **or it waits for the explosion and investigates the debris.**

A systematic methodology would find more.

SOURCES Gilly, T. (2026). Emergent Convergence Risk. Real Safety AI Foundation.

REFERENCES

Gilly, T. (2026). *Emergent Convergence Risk: Why Existing AI Governance Cannot Detect Combinatorial Threats and a Proposed Methodology.* Real Safety AI Foundation.

Banerjee, S., Jha, S., & Nita-Rotaru, C. (2024). SoK: A systems perspective on compound AI threats and countermeasures. arXiv preprint arXiv:2411.13459. https://arxiv.org/abs/2411.13459

European Parliament & Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act). Official Journal of the European Union. EUR-Lex: 32024R1689.

Gilly, T. (2025-2026). The Harm Blindness Framework: A practical application methodology for stakeholder harm prevention in technology development. Real Safety AI Foundation. https://realsafetyai.org/framework

Gilly, T. (2026a). Ambient non-consensual image synthesis: A convergence threat analysis of AR wearables, real-time video generation, and open-source nudification models. SSRN/SocArXiv Preprint. https://doi.org/10.5281/zenodo.18297286

Gilly, T. (2026b). Compressed feasibility: A technical update to the ambient non-consensual synthesis threat model. Forthcoming.

Gilly, T. (2026c). We Happy Few: Therapeutic perceptual modification, regulatory pathway precedent, and the governance gap in cloud-rendered reality. Forthcoming.

Gilly, T. (2026d). The accidental stack: A case study in emergent perceptual control infrastructure from independently developed AI capabilities. Forthcoming.

International Organization for Standardization. (2023). ISO/IEC 42001:2023 Information technology: Artificial intelligence: Management system.

LaValle, S. M., et al. (2024). From virtual reality to the emerging discipline of perception engineering. Annual Review of Control, Robotics, and Autonomous Systems, 7, 291-315. https://doi.org/10.1146/annurev-control-062323-102456Meta. (n.d.).Systemcards.MetaPlatforms,Inc.https://ai.meta.com/tools/system-cards

MIT FutureTech. (2024). AI Risk Repository. Massachusetts Institute of Technology. https://airisk.mit.edu

REFERENCES (CONTINUED)

Mitchell, M., et al. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220-229). ACM. https://doi.org/10.1145/3287560.3287596

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Riva, G. (2025). Invisible architectures of thought: Toward a new science of AI as cognitive infrastructure. arXiv preprint arXiv:2507.22893.

United Nations. (2023). Globally Harmonized System of Classification and Labelling of Chemicals (GHS) (ST/SG/AC.10/30, Rev. 10).

Macpherson, F., et al. (2023). Policy and practice recommendations for augmented and mixed reality. Centre for the Study of Perceptual Experience, University of Glasgow. Royal Society of Edinburgh Research Network Project. https://www.gla.ac.uk/research/az/cspe/projects/augmentedrealityethicsperceptionmetaphysics/

Zaharia, M., et al. (2024, February 18). The shift from models to compound AI systems. Berkeley AI Research Blog. https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/WordCount:Approximately5,200words(excludingreferences)TargetVenue:AISafety/Science&TechnologyPolicyjournal;companionSSRN/SocArXivpreprint