TECHNOLOGY ETHICS · RISK MANAGEMENT · STAKEHOLDER HARM PREVENTION · HARM BLINDNESS FRAMEWORK · HISTORICAL PRECEDENT · DESIGN CHECKPOINTS · PREVENTABLE HARM

THE HARM BLINDNESS FRAMEWORK

Built from 161 historical cases spanning approximately 5,000 years and 151 corporate disasters from 1970–2025. The cases came first. The pattern was found. 95.7% had documented precedent available at the time decisions were made.

Travis Gilly · Real Safety AI Foundation

A structural failure to identify affected stakeholders, arising from identifiable mechanisms, not individual moral failings.

The harms were, in most cases, predictable. The pattern across 5,000 years of analyzed cases is not that precedent was absent. The pattern is that decision-makers failed to look for it. This framework operationalizes looking.

FIVE MECHANISMS

Stakeholder myopia; ethical abdication; the historical-precedent fallacy; efficiency worship; systemic speed pressure. They interact and reinforce each other.

95.7%

of analyzed harms had documented precedent available at the time the key decisions were made.

SOURCES Gilly, The Harm Blindness Framework (2026), secs. 1.1, 8.2.

In each case, decision-makers did not lack access to information. They lacked a methodology for consulting it.

Facebook · Engagement algorithms

Amplifying divisive content to maximize engagement contributed to democratic polarization. The harm pathway was not sought. The precedent existed from earlier media studies.

Boeing · 737 MAX MCAS

Software that prioritized cost savings over redundant safety systems contributed to 346 deaths. The failure mode of single-point control system reliance had documented precedent in aviation history.

Opioid manufacturers · Prescribing practices

Aggressive prescribing promotion engineered an addiction crisis. The addiction mechanisms of opioids had been documented in clinical literature for more than a century before OxyContin's launch.

SOURCES Gioia (1992); Horwitz (2021); Courtwright (2001); Gilly, The Harm Blindness Framework (2026), sec. 1.1.

Every existing tradition tells you how to weigh a harm. None tells you how to find the people who will be harmed, or when to look.

Principlism

Autonomy, beneficence, non-maleficence, justice. Abstract; needs scaffolding to identify stakeholders or timing.

Consequentialism

Evaluates outcomes across affected parties, but does not specify how to enumerate who they are.

Deontology & virtue

Duties and character illuminate the moral terrain but leave the identification step unspecified.

The framework

Supplies the missing scaffolding: who to consult, and at which decision point.

SOURCES Beauchamp & Childress (2019); Driver (2012); Alexander & Moore (2021); Hursthouse & Pettigrove (2018); Held (2006); Gilly, The Harm Blindness Framework (2026), sec. 2.1.

Not a moral failing of individuals. Five structural mechanisms that interact and reinforce one another.

Stakeholder myopia

Analysis centers direct beneficiaries; secondary, indirect, and excluded parties go unseen.

Ethical abdication

Engineers defer to management, management to legal, legal to compliance. No one owns the harm question.

Precedent fallacy

“This is unprecedented”, a belief rarely checked against the documented record.

Efficiency worship

Stakeholder analysis takes time; when speed is the supreme value, it is cut.

Speed pressure

First-mover advantage, quarterly earnings, and career incentives all reward shipping over deliberation.

SOURCES Gilly, The Harm Blindness Framework (2026), sec. 3.1; Meadows (2008); Reason (1990).

THE METHODOLOGY

FOUR CHECKPOINTS

Ideation · Design · Testing · Launch

Structured stakeholder harm analysis at each decision point, before deployment. The framework embeds consultation of historical precedent into the development process rather than treating ethics as a post-hoc audit.

01
IDEATION
Identify potential harms and affected stakeholders at the earliest stage, before architectural decisions are made.

02
DESIGN
Assess how specific design choices affect stakeholder harm profiles. Review historical precedent for analogous design decisions.

03
TESTING
Evaluate whether testing protocols surface the harm categories identified at earlier checkpoints. Identify testing gaps.

04
LAUNCH
Final harm-profile review before deployment. Confirm monitoring mechanisms are in place for post-launch harm detection.

SOURCES Gilly, The Harm Blindness Framework (2026), secs. 3–6; Cooper (2008); Freeman (1984).

“WHO IS AFFECTED BEYOND DIRECT BENEFICIARIES?”

Comprehensive stakeholder map

Primary beneficiaries, secondary parties, those who interact with users, those displaced, those who lack access, and downstream populations.

Scale analysis

What changes when the solution reaches 1,000 times current scope? What second-order effects emerge? What systems are stressed?

External facilitator

Not a member of the project team. Internal members carry the same pressures that produce harm blindness.

Documented decision

Proceed, pivot, or cancel, with reasoning on the record.

SOURCES Gilly, The Harm Blindness Framework (2026), sec. 4.1; Freeman (1984).

“What incentives does this system create?”

A structural analysis, not a values statement. The question is what behavior the architecture will reward or punish once it is running at scale, regardless of what its designers intend.

Who benefits from each design choice, and who bears its cost?

What behavior does the system reward? What does it make harder?

An engagement metric is an incentive. A default is an incentive. The harm pathway is usually a structural choice, not an accident.

SOURCES Gilly, The Harm Blindness Framework (2026), sec. 4.1; Meadows (2008).

CHECKPOINT 3 · TESTING

“Who is NOT in our testing group?”

Compare test composition to the target population. Explicitly document who is missing: people with disabilities, non-native speakers, low-literacy, low-bandwidth, and rural users. Test the accommodations, not just the happy path.

CHECKPOINT 4 · LAUNCH

“Can we defend this publicly?”

The front-page test: write the headline if it goes wrong, and the public explanation. If the decision cannot be defended publicly, it should not proceed. Remaining risk is accepted explicitly, in writing, by a named individual.

SOURCES Gilly, The Harm Blindness Framework (2026), sec. 4.1.

The framework functions regardless of whether the organization is motivated by ethics or by risk management. The analysis is the same. Only the framing differs.

Existing ethical frameworks typically require ethical commitment as a precondition for adoption, which excludes organizations that most need harm prevention. This design addresses that limitation directly.

ETHICS FRAMING

"Who do we have an obligation to protect, and how?"

RISK MANAGEMENT FRAMING

"What is our exposure, and how do we reduce it before it becomes a disaster?"

Same stakeholder analysis. Same checkpoints. Same outputs. Different language for different rooms.

SOURCES Gilly, The Harm Blindness Framework (2026), secs. 1.2, 3.3; IEEE (2019); OECD (2019).

THE CASE RESEARCH

THE CASES CAME FIRST

161 historical cases · 151 corporate disasters · 5,000 years of record

The cases were compiled first. The pattern emerged from them. The framework was built to address what the cases kept showing: precedent existed and was not consulted.

95.7%

had documented precedent available at decision time

4.3%

genuinely first-of-kind, no comparable precedent

The problem of harm blindness is, in almost every case, a correctable problem of methodology. Not an irreducible problem of novelty.

The Triangle Shirtwaist fire was documented for a century before Rana Plaza. The Ford Pinto cases were documented for decades before Deepwater Horizon. The precedent existed. Decision-makers did not look. This framework operationalizes looking.

SOURCES Gilly, The Harm Blindness Framework (2026), sec. 8.2; Von Drehle (2003); Gioia (1992); National Commission (2011).

Prevention is not expensive. It is 1–10% of what the disasters cost.

Across 151 analyzed corporate disasters, prevention costs typically ranged from one to ten percent of eventual disaster costs. The argument that harm prevention is a luxury for organizations that can afford ethics collapses on the numbers.

PREVENTION COST

1–10%

of eventual disaster cost

RETURN ON INVESTMENT

1,000–10,000×

for harm prevention investment

SOURCES Gilly, The Harm Blindness Framework (2026), sec. 8.1; Perrow (1984); Reason (1990); Vaughan (1996); Tett (2009).

The major AI governance documents specify what to do. This framework operationalizes how to do it.

EU AI Act (2024)

Risk assessment and stakeholder impact requirements align directly with checkpoint 1 (ideation) and checkpoint 4 (launch).

NIST AI RMF (2023)

The Map, Measure, Manage, Govern structure corresponds to the framework's checkpoint sequence and its precedent-consultation methodology.

OECD AI Principles (2019)

The robustness and accountability principles require the kind of structured stakeholder harm analysis the framework provides.

UNESCO (2021)

The recommendation's emphasis on human oversight and harm assessment throughout the lifecycle maps onto the four-checkpoint structure.

SOURCES European Commission (2024); National Institute of Standards and Technology (2023); OECD (2019); UNESCO (2021).

The framework does not ask anyone to predict the future. It asks them to consult the record that already exists.

Triangle Shirtwaist, 1911 → Rana Plaza, 2013

Locked exits and inadequate escapes were documented for a century. The precedent was available when Rana Plaza collapsed under structurally similar conditions, killing over 1,100 workers.

Ford Pinto, 1970s → Deepwater Horizon & 737 MAX

Cost-benefit calculations that accepted preventable deaths were on the record. The same calculation reappeared at BP and at Boeing. The precedent existed; no one consulted it.

SOURCES Gilly, The Harm Blindness Framework (2026), sec. 3.2; Von Drehle (2003); Gioia (1992).

ANSWERING THE STRONGEST OBJECTION

“Isn’t calling these harms predictable just hindsight bias?” The method is built to survive that question.

IT DOES NOT REQUIRE FORESIGHT

The claim is not that the specific outcome was knowable. It is that documented precedent existed at decision time. That is a matter of record, not retrospection.

THE 4.3% IS THE CONTROL

Where no precedent existed, the method does not fault the decision-maker. The finding is bounded to cases where the record was available and went unconsulted.

SOURCES Gilly, The Harm Blindness Framework (2026), sec. 6; Kahneman & Tversky (1979); Arkes & Blumer (1985).

FOUR CONTRIBUTIONS

THEORETICAL

Harm blindness as a systematic structural phenomenon, arising from identifiable mechanisms rather than individual moral failings.

METHODOLOGICAL

A checkpoint-based consultative approach that embeds stakeholder harm analysis at four decision points throughout development.

EMPIRICAL

161 historical cases and 151 corporate disasters: 95.7% had documented precedent available at the time decisions were made. The cases preceded the framework. The pattern was found first.

PRACTICAL

An implementable framework that functions regardless of whether the implementing organization is motivated by ethics or risk management.

Implementation guides, checkpoint templates, and full validation studies are publicly available at realsafetyai.org.

SOURCES Gilly, The Harm Blindness Framework (2026), sec. 8.1.

Gilly, T. (2026). *The Harm Blindness Framework: A Practical Application Methodology for Stakeholder Harm Prevention in Technology Development.* Real Safety AI Foundation.

Alexander, L., & Moore, M. (2021). Deontological ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition). Stanford University.

Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, 35(1), 124–140.

Beauchamp, T. L., & Childress, J. F. (2019). *Principles of biomedical ethics* (8th ed.). Oxford University Press.

Cooper, R. G. (2008). Perspective: The Stage-Gate idea-to-launch process — Update, what's new, and NexGen systems. *Journal of Product Innovation Management*, 25(3), 213–232.

Courtwright, D. T. (2001). *Dark paradise: A history of opiate addiction in America* (Enlarged ed.). Harvard University Press.

Driver, J. (2012). *Consequentialism*. Routledge.

European Commission. (2024). *Artificial Intelligence Act*. Official Journal of the European Union.

Freeman, R. E. (1984). *Strategic management: A stakeholder approach*. Pitman.

Gioia, D. A. (1992). Pinto fires and personal ethics: A script analysis of missed opportunities. *Journal of Business Ethics*, 11(5), 379–389.

Held, V. (2006). *The ethics of care: Personal, political, and global*. Oxford University Press.

Horwitz, J. (2021). The Facebook Files: A Wall Street Journal investigation. *Wall Street Journal*.

Hursthouse, R., & Pettigrove, G. (2018). Virtue ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition). Stanford University.

IEEE. (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems* (First Edition). IEEE.

Juran, J. M. (1992). *Juran on quality by design: The new steps for planning quality into goods and services*. Free Press.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291.

Meadows, D. H. (2008). *Thinking in systems: A primer*. Chelsea Green Publishing.

National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling. (2011). *Deep water: The Gulf oil disaster and the future of offshore drilling*. U.S. Government Printing Office.

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce.

OECD. (2019). *Recommendation of the Council on Artificial Intelligence*. OECD Legal Instruments.

Partnership on AI. (2021). *PAI's approach to responsible AI development*. Partnership on AI.

Perrow, C. (1984). *Normal accidents: Living with high-risk technologies*. Basic Books.

Reason, J. (1990). *Human error*. Cambridge University Press.

Rességuier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society*, 7(2).

Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, L., Pour, S., Casper, S., & Thompson, N. (2024). The AI Risk Repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *arXiv preprint arXiv:2408.12622*.

Tett, G. (2009). *Fool's gold: How the bold dream of a small tribe at J.P. Morgan was corrupted by Wall Street greed and unleashed a catastrophe*. Free Press.

UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*. United Nations Educational, Scientific and Cultural Organization.

Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

Vaughan, D. (1996). *The Challenger launch decision: Risky technology, culture, and deviance at NASA*. University of Chicago Press.

Von Drehle, D. (2003). *Triangle: The fire that changed America*. Atlantic Monthly Press.