The Harm Blindness Framework: A Practical Application Methodology for Stakeholder Harm Prevention in Technology Development

Travis Gilly Real Safety AI Foundation

Abstract

Technology development consistently produces preventable harm to stakeholders who were identifiable at the time of key decisions. This paper introduces the Harm Blindness Framework, a checkpoint-based methodology designed to surface stakeholder impacts during development rather than after deployment. The framework operationalizes consultation of historical precedent through structured analysis at four decision points: ideation, design, testing, and launch.

Validation against 161 historical cases spanning approximately 5,000 years reveals that 95.7% of analyzed harms had documented precedent available at the time decisions were made. The remaining 4.3% represent genuinely first-of-kind situations. A parallel analysis of 151 corporate disasters from 1970-2025 demonstrates that prevention costs typically range from 1-10% of eventual disaster costs, yielding return-on-investment ratios of 1,000-10,000x for harm prevention.

The framework is designed to function regardless of whether implementing organizations are motivated by ethical concerns or risk management objectives. The same analytical methodology surfaces the same stakeholder impacts; only the framing differs. This motivation-agnostic design addresses a significant limitation of existing ethical frameworks, which typically require ethical commitment as a precondition for adoption.

This paper presents the theoretical foundation, methodology, validation evidence, and limitations of the framework. Supporting materials including implementation guides, checkpoint templates, and complete validation studies are available separately.

Keywords: stakeholder analysis, harm prevention, technology ethics, risk management, AI governance, development methodology

---

1. Introduction

1.1 The Problem of Harm Blindness

When Facebook's algorithms amplified divisive content to maximize engagement, the platform's engineers were not seeking to polarize democratic societies. When Boeing's 737 MAX software prioritized cost savings over redundant safety systems, the company's leadership did not intend to cause 346 deaths. When opioid manufacturers promoted aggressive prescribing practices, pharmaceutical executives were not consciously engineering an addiction crisis. In each case, decision-makers failed to identify stakeholders who would be harmed by their choices, despite the availability of historical precedent that could have informed those decisions.

This paper introduces the term "harm blindness" to describe the systematic failure to identify affected stakeholders during decision-making processes. Harm blindness is not primarily a moral failing of individual actors, though individual choices certainly contribute. Rather, it represents a structural phenomenon arising from the interaction of organizational pressures, cognitive limitations, and incentive structures that consistently produce the same pattern: stakeholders who could have been identified are not considered until after harm has occurred.

The pattern is not new. Analysis of historical cases reveals similar dynamics operating across millennia, from ancient infrastructure failures to modern technological disasters. What has changed is the scale and speed at which technology decisions can affect populations, making systematic approaches to stakeholder identification increasingly urgent.

1.2 The Post-Hoc Problem

Contemporary approaches to technology ethics share a common structural limitation: they are typically applied after development decisions have been made. Ethics review boards evaluate products that have already been designed. Impact assessments analyze systems that have already been built. Safety teams are consulted after engineering choices have locked in particular architectures.

This timing creates predictable dynamics. By the time ethical concerns are formally considered, organizations have already invested substantial resources in their chosen approach. Sunk cost psychology makes course correction increasingly difficult as investment accumulates (Arkes & Blumer, 1985). Engineers who raised early concerns find their objections overridden by momentum and managerial pressure to meet deadlines. The pattern has been documented across industries, from automotive safety (Gioia, 1992) to financial services (Tett, 2009) to technology platforms (Horwitz, 2021).

The result is a systematic gap between when ethical analysis could be most effective (before design decisions constrain options) and when it typically occurs (after substantial commitment to a particular path). The Harm Blindness Framework addresses this gap by embedding stakeholder analysis at decision points throughout the development cycle, rather than treating ethics review as a final checkpoint before launch.

1.3 Research Objectives

This research pursued three objectives. First, to develop a practical methodology for stakeholder harm identification that could be applied during technology development, at points when course correction remains feasible. Second, to validate the methodology against historical and corporate case databases, testing whether the approach would have surfaced harms that were missed in actual decision-making. Third, to design the framework such that it functions across motivational contexts, recognizing that frameworks requiring ethical commitment as a precondition have inherently limited reach.

1.4 Paper Structure

Section 2 reviews existing literature on ethical frameworks, AI governance approaches, and their limitations. Section 3 presents the theoretical framework, including the formal definition of harm blindness, its component mechanisms, and the consultative premise underlying the methodology. Section 4 details the four-checkpoint methodology. Section 5 presents validation evidence from historical and corporate case analyses, including three illustrative examples. Section 6 acknowledges limitations and addresses the hindsight objection. Section 7 discusses implications for practice, policy, and future research. Section 8 concludes.

---

2. Literature Review

2.1 Existing Ethical Frameworks

Ethical analysis of technology draws on several philosophical traditions, each offering distinct approaches to evaluating the moral dimensions of technological development and deployment.

Principlist approaches, most fully developed in biomedical ethics by Beauchamp and Childress (2019), identify core principles that should guide decision-making: respect for autonomy (honoring the decision-making capacities of autonomous persons), beneficence (providing benefits and balancing benefits against risks), non-maleficence (avoiding the causation of harm), and justice (distributing benefits, risks, and costs fairly). While widely adopted in healthcare contexts and increasingly applied

to technology ethics, principlist frameworks provide limited guidance on how to systematically identify affected stakeholders or when in development processes principles should be applied. The principles themselves are abstract; operationalizing them requires methodological scaffolding that principlism does not provide.

Consequentialist frameworks evaluate actions based on outcomes, requiring assessment of impacts across affected parties (Driver, 2012). Utilitarian variants seek to maximize aggregate welfare; other consequentialist approaches may prioritize different outcomes. This orientation toward outcomes aligns with stakeholder analysis, as consequentialist evaluation requires identifying who experiences consequences. However, consequentialist approaches typically assume decision-makers can identify relevant stakeholders and predict impacts, assumptions that harm blindness patterns call into question. If decision-makers systematically fail to identify affected parties, consequentialist calculations will systematically exclude relevant impacts.

Deontological approaches emphasize duties and rules independent of consequences (Alexander & Moore, 2021). Kantian ethics focuses on the categorical imperative: acting only according to principles one could will to be universal laws, and treating persons always as ends in themselves rather than merely as means. Rights-based approaches identify fundamental entitlements that should not be violated regardless of aggregate consequences. While providing clear prohibitions against certain actions, deontological frameworks offer less guidance for the gray areas that characterize most technology decisions, where harms emerge from aggregated choices rather than single prohibited acts. The categorical imperative provides limited guidance when the relevant question is not "Is this action permissible?" but "Who have we failed to consider?"

Virtue ethics focuses on the character of decision-makers rather than evaluation of specific choices (Hursthouse & Pettigrove, 2018). Applications to technology emphasize cultivating virtues like honesty, responsibility, care, and justice among engineers and executives (Vallor, 2016). Shannon Vallor's work on "technomoral virtues" identifies character traits particularly relevant to technology development, including humility, courage, empathy, and magnanimity. While valuable for professional formation and organizational culture, virtue approaches do not provide systematic methods for identifying stakeholder impacts. A virtuous engineer may still suffer from harm blindness if organizational structures and analytical methods do not surface affected populations.

Care ethics, developed from feminist philosophy, emphasizes relationships, interdependence, and attention to particular others rather than abstract principles (Held, 2006). Care ethics challenges the individualism implicit in much ethical theory,

recognizing that persons exist in webs of relationship and dependency. This relational orientation has particular relevance for stakeholder analysis, as it directs attention to how decisions affect relationships and communities rather than isolated individuals. However, care ethics has been less systematically applied to technology development contexts.

2.2 AI-Specific Governance Approaches

The rapid development of artificial intelligence systems has produced domain-specific governance frameworks addressing unique challenges of AI, including opacity of decision-making, potential for emergent behavior, scale of deployment, and difficulty of human oversight.

The IEEE's Ethically Aligned Design initiative, developed through extensive multi-stakeholder consultation, articulates principles for autonomous and intelligent systems (IEEE, 2019). Core principles include human well-being (ensuring AI benefits humanity), accountability (establishing clear responsibility for AI behavior), transparency (making AI operations understandable), and awareness of misuse (designing to prevent harmful applications). The IEEE framework provides substantive guidance on what responsible AI should achieve but less guidance on how to identify who will be affected by AI systems or how to integrate ethical analysis into development workflows.

The EU AI Act establishes risk-based regulatory categories creating tiered obligations (European Commission, 2024). Prohibited practices include AI systems that manipulate behavior through subliminal techniques, exploit vulnerabilities of specific groups, enable social scoring by governments, or perform real-time remote biometric identification in public spaces for law enforcement (with limited exceptions). High-risk systems, including those used in critical infrastructure, education, employment, essential services, law enforcement, and migration, require conformity assessments before market placement. Lower-risk applications face transparency requirements. The Act represents the most comprehensive AI-specific regulation globally, but operates primarily at deployment stage rather than throughout development.

The NIST AI Risk Management Framework provides voluntary guidance for identifying, assessing, and managing AI risks throughout system lifecycles (NIST, 2023). The framework organizes around four functions: Govern (cultivating and implementing a culture of risk management), Map (establishing context and identifying risks), Measure (analyzing and tracking identified risks), and Manage (prioritizing and acting on risks). NIST explicitly acknowledges that AI risks differ from traditional software risks due to AI's potential for emergent behavior, opacity, and evolving capabilities. The framework is flexible and non-prescriptive, which

enables adaptation but provides less specific methodology for stakeholder identification.

Partnership on AI has developed best practices across multiple domains including fairness, safety, economic impacts, and governance (PAI, 2021). As a multi-stakeholder organization including major technology companies, civil society organizations, and academic institutions, PAI represents a collaborative approach to developing norms. PAI's work on algorithmic fairness, data governance, and AI and jobs provides substantive guidance in specific domains, though the organization's consensus-based approach may limit the specificity of recommendations.

Other notable initiatives include the OECD AI Principles (2019), which established the first intergovernmental standard on AI; the UNESCO Recommendation on AI Ethics (2021), which extended AI governance considerations globally; and various national AI strategies that incorporate ethical considerations to varying degrees.

2.3 Limitations of Current Approaches

Four structural limitations characterize existing frameworks, creating the gap this research addresses.

First, existing frameworks are typically applied after key development decisions have been made. Ethics review boards evaluate products that have already been designed. Impact assessments analyze systems that have already been built. The EU AI Act's conformity assessments occur before deployment but after architecture is fixed. NIST's Map function establishes context, but organizations may have already committed substantial resources before systematic mapping occurs. The cost of addressing identified concerns increases substantially once design choices have been locked in; by the time formal ethical analysis occurs, course correction may be prohibitively expensive.

Second, existing frameworks generally require ethical motivation as a precondition for implementation. IEEE Ethically Aligned Design is voluntary. NIST AI RMF is voluntary. Partnership on AI membership is voluntary. Even regulatory approaches like the EU AI Act apply compliance pressure but do not change the motivational structure of organizations that prioritize other objectives. This creates a selection effect: frameworks are most likely to be adopted by organizations already inclined toward responsible development, while organizations most likely to cause harm may not adopt them (Rességuier & Rodrigues, 2020).

Third, current approaches lack systematic methodology for identifying affected stakeholders. Frameworks articulate that impacts should be considered; NIST's Map function references "individuals, groups, communities, organizations, and society" as potential stakeholders. But specific methodology for identifying who is affected

beyond direct users receives insufficient attention. The core question of who is missing from analysis, which is central to addressing harm blindness, is not systematically addressed.

Fourth, existing frameworks are not typically integrated with development workflows. Ethics review exists as a separate function rather than an embedded component of standard development processes. This structural separation enables the post-hoc problem: by the time ethics review occurs, development has progressed to stages where findings are difficult to act upon.

2.4 Gap This Research Addresses

The Harm Blindness Framework addresses these limitations through three design features.

First, it embeds stakeholder analysis at multiple points throughout development, at stages when course correction remains feasible. Rather than a single ethics review before deployment, the framework requires analysis at ideation (before design begins), design (before implementation begins), testing (before launch preparation), and launch (before deployment). This staging means that concerns surface when addressing them is least costly.

Second, the framework functions regardless of whether implementing organizations are motivated by ethical concerns or risk management. The same checkpoint questions surface the same stakeholder impacts; only the framing differs. Organizations motivated by ethics receive systematic methodology for stakeholder consideration. Organizations motivated by risk management receive systematic methodology for identifying legal, financial, and reputational exposure. The outputs are identical; the framing adapts to organizational motivation.

Third, the framework provides explicit methodology for identifying affected stakeholders, with particular attention to identifying who is not being considered. Checkpoint questions specifically prompt analysis of populations beyond direct users and beneficiaries. The discipline of asking "Who is missing?" addresses the core dynamic of harm blindness.

Mapping exercises have been conducted to align the framework with existing risk taxonomies, including the MIT AI Risk Repository (Slattery et al., 2024), which provides comprehensive categorization of AI-specific risks. Detailed alignment with ISO 31000, COSO ERM, NIST AI RMF, and other governance standards is provided in a separate implementation guide. The contribution of this paper is the methodology itself rather than compliance crosswalks.

---

3. Theoretical Framework

3.1 Defining Harm Blindness

Harm blindness is defined as the systematic failure to identify stakeholders who will be affected by decisions, despite the availability of information that would enable such identification. The definition emphasizes three elements: the failure is systematic rather than random; it involves identifiable stakeholders (not hypothetical or unforeseeable impacts); and relevant information exists at decision time.

Harm blindness operates through five interacting mechanisms:

Stakeholder Myopia describes the tendency to focus analysis on direct beneficiaries while ignoring secondary, tertiary, and indirect effects. Product development naturally centers intended users; identifying those who interact with users, compete with the product, lack access to it, or experience downstream effects requires deliberate analytical effort that is often omitted.

Ethical Abdication occurs when decision-makers assume that ethical considerations are someone else's responsibility. Engineers defer to management; management defers to legal; legal defers to compliance; compliance focuses on regulatory requirements rather than broader stakeholder impacts. The result is systematic gaps where no one owns the question of who gets harmed.

Historical Precedent Fallacy manifests as the belief that current situations are unprecedented, rendering historical examples irrelevant. This belief is rarely examined against evidence. As validation analysis demonstrates, the vast majority of harms follow documented patterns with extensive precedent.

Efficiency Worship prioritizes speed and resource optimization over comprehensive analysis. Stakeholder identification takes time; checkpoint documentation creates overhead; engaging affected communities slows development. When efficiency is valued above other considerations, harm prevention activities are systematically deprioritized.

Systemic Speed Pressure describes the competitive and organizational dynamics that reward rapid development. First-mover advantages, quarterly earnings expectations, and career incentives that reward shipping products all create pressure to minimize anything perceived as slowing development, including stakeholder analysis.

These mechanisms interact and reinforce each other. Stakeholder myopia is enabled by ethical abdication (no one is responsible for looking beyond direct users). The historical precedent fallacy justifies skipping analysis that efficiency worship already discourages. Systemic speed pressure amplifies all other mechanisms by creating urgency that overrides deliberation.

3.2 The Consultative Premise

A potential objection to any harm prevention framework is that it requires prediction of future outcomes, which may be impossible. The Harm Blindness Framework does not claim predictive power. Instead, it operationalizes consultation of available historical precedent.

The central finding from validation analysis is that 95.7% of examined cases had documented precedent at the time decisions were made. The Triangle Shirtwaist Factory fire of 1911 documented the consequences of locked exits and inadequate fire escapes; this precedent was available when decisions were made about the Rana Plaza garment factory that collapsed in 2013, killing over 1,100 workers in structurally similar circumstances. The Ford Pinto fuel tank failures of the 1970s documented the consequences of cost-benefit analyses that accepted preventable deaths; this precedent was available when similar calculations were made at BP before the Deepwater Horizon disaster and at Boeing before the 737 MAX crashes.

The framework does not ask decision-makers to predict novel harms. It asks whether they have consulted the documented record of similar decisions and their consequences. This is a research question, not a prophecy. The methodology operationalizes asking: "Has something similar happened before? What were the consequences? Who was affected?"

For the 95.7% of cases with available precedent, the question is not whether decision-makers could have predicted the outcome. The question is whether they consulted the precedent that existed. The consistent finding is that they did not.

The remaining 4.3% of analyzed cases represent genuinely first-of-kind situations where no relevant precedent was available. The framework acknowledges its limitations for such cases. However, even in these situations, contemporaneous critics often raised concerns that went unheeded. The framework cannot guarantee harm prevention, but it can ensure that available information is systematically consulted.

3.3 Checkpoint Theory

The timing of stakeholder analysis matters because the cost of addressing identified concerns increases as development progresses. During ideation, changing direction is relatively inexpensive; concepts can be modified or abandoned with minimal sunk cost. During design, changes require reworking specifications and potentially discarding completed work. During testing, changes require redesign and re-implementation. After launch, changes require public acknowledgment of problems and potentially recalling or discontinuing products.

This cost curve creates a perverse dynamic under current practices. Ethical analysis typically occurs late in development, when concerns are most expensive to address. Decision-makers facing expensive remediation are strongly incentivized to discount or rationalize the concerns rather than act on them.

The checkpoint model addresses this dynamic by embedding stakeholder analysis at four points in the development cycle: ideation (before design begins), design (before implementation begins), testing (before launch preparation), and launch (before deployment). At each checkpoint, specific questions surface stakeholder impacts relevant to that development stage. Documentation requirements create accountability trails. Sign-off requirements distribute responsibility rather than allowing abdication.

The checkpoint model draws on established practices in quality management, where inspection points throughout production are more effective than final inspection alone (Juran, 1992). Similar staged-gate approaches are standard in product development (Cooper, 2008). The contribution here is applying checkpoint methodology specifically to stakeholder harm identification rather than product quality or commercial viability.

3.4 Motivation Agnosticism

A significant limitation of existing ethical frameworks is that they require ethical motivation for adoption. Organizations must already value responsible development to voluntarily implement ethics review processes. This creates a fundamental reach problem: frameworks are most likely to be adopted by organizations least likely to cause harm.

The Harm Blindness Framework is designed to function regardless of whether implementing organizations are motivated by ethical concerns or risk management objectives. The same checkpoint questions surface the same stakeholder impacts; only the framing differs.

For organizations motivated by ethical considerations, the framework operationalizes stakeholder concern systematically. Rather than relying on individual sensitivity to notice who might be harmed, the methodology ensures systematic identification through structured analysis.

For organizations motivated by risk management, the framework identifies legal, financial, and reputational exposure before it materializes. Stakeholder harms frequently become litigation, regulatory action, public relations crises, and shareholder losses. Identifying these exposures early provides economic value independent of ethical motivation.

This design choice is intentional. Frameworks requiring ethical motivation for adoption will never achieve the reach necessary to prevent harm at scale. By producing equivalent analytical outputs regardless of motivation, the methodology can be adopted by organizations across the motivational spectrum while delivering equivalent stakeholder protection.

---

4. Methodology

4.1 The Four Checkpoints

The Harm Blindness Framework structures stakeholder analysis around four checkpoints aligned with standard development phases. Each checkpoint has defined timing, core questions, required participants, and output requirements.

Checkpoint 1: Ideation Phase occurs before design begins, when concepts are being evaluated and direction is being set. The core analytical question is: "Who is affected beyond direct beneficiaries?" This checkpoint focuses on comprehensive stakeholder identification, including primary beneficiaries, secondary affected parties, those who interact with beneficiaries, those who compete with or are displaced by the solution, those who lack access to it, and downstream populations affected by its effects. The checkpoint also requires scale analysis: what changes when the solution reaches 1,000 times current scope? What second-order effects emerge? What systems are stressed?

Required participants include a checkpoint facilitator (not a member of the project team), the project owner, technical lead, and stakeholder representatives external to the team. The facilitator role is critical; internal team members are subject to the same pressures that produce harm blindness.

Outputs include a comprehensive stakeholder map, impact assessment for each identified group, risk prioritization, and documented decision (proceed, pivot, or cancel) with reasoning.

Checkpoint 2: Design Phase occurs before implementation begins, when architecture and design choices are being finalized. The core analytical question is: "What incentives does this system create?" This checkpoint focuses on structural analysis of how the system will shape behavior.

Key questions examine incentives for users (what behaviors are rewarded? what behaviors are punished? what happens if users optimize for these incentives?), incentives for different stakeholder groups (how might each group game the system? what perverse incentives exist?), and incentives for the organization itself (what metrics are being optimized? what is not being measured?).

For systems involving adaptation or learning, additional questions address what the system optimizes for, what proxy metrics it might exploit, what unintended shortcuts it might find, and how it might manipulate or deceive users.

Design choice analysis examines data collection (what is needed vs. what could be collected vs. what should not be collected), default settings (what is opted in vs. out by default), friction points (where the system makes things easy vs. difficult), transparency (what users see vs. what is hidden), and user control (what can be changed vs. what is locked).

Outputs include incentive analysis documentation, design alternatives considered, trade-offs made with reasoning, and updated stakeholder impact assessment reflecting design decisions.

Checkpoint 3: Testing Phase occurs before launch preparation begins, when the system has been built and is being validated. The core analytical question is: "Who is NOT in our testing group?" This checkpoint focuses on identifying validation gaps.

Test group analysis examines demographic representation across relevant categories (age, gender, geographic location, socioeconomic status, education, technical proficiency, language, disability status) comparing test group composition to target population composition. Gaps are explicitly documented with their potential impacts.

Missing population analysis specifically identifies vulnerable groups not included in testing: people with disabilities, non-native language speakers, low-literacy users, low-bandwidth or rural users, and other marginalized populations.

Accessibility testing examines specific accommodations: screen reader compatibility, keyboard navigation, color contrast, dyslexia-friendly design, low-bandwidth performance, mobile-only access, non-English language support, and cognitive load.

Usage pattern analysis compares expected use to observed use, documenting unexpected behaviors, workarounds users created, and points where users abandoned the system.

Outputs include test coverage analysis with identified gaps, accessibility assessment, failure mode documentation, and validation of whether predicted harms from earlier checkpoints occurred.

Checkpoint 4: Launch Phase occurs before deployment, serving as final review and accountability checkpoint. The core analytical question is: "Can we defend this publicly?" This checkpoint focuses on comprehensive review and explicit acceptance of remaining risks.

The "front page test" asks decision-makers to articulate the headline if this goes wrong, their public explanation, and expected stakeholder response. If the decision cannot be defended publicly, it should not proceed.

Risk acceptance requires explicit documentation of known harms being accepted, which stakeholder groups are affected, severity of impacts, reasoning for acceptance, and identification of the individual accepting responsibility.

Monitoring plans specify metrics to track post-launch, frequency of measurement, alert thresholds, and response plans if harms exceed predictions.

Precedent analysis examines what precedent this decision sets for future products in the organization, for the industry, and for technology development broadly.

Exit strategy documents shutdown criteria, harm mitigation if the product is cancelled, and stakeholder communication plans.

Outputs include comprehensive stakeholder review, explicit risk acceptance with accountability, monitoring plan, and executive sign-off with acknowledged accountability.

4.2 Stakeholder Identification Methodology

The framework employs a tiered approach to stakeholder identification, building on Freeman's (1984) foundational stakeholder theory. Primary stakeholders are those directly affected by the system's intended function: users, customers, and direct beneficiaries. Secondary stakeholders are those who interact with primary stakeholders or are affected by their changed behavior: family members, colleagues, community members. Tertiary stakeholders experience indirect effects through changed systems, markets, or social dynamics.

For each stakeholder group, analysis documents the number affected, benefits received, harms experienced, net impact assessment, and priority for attention. Power dynamic analysis identifies which stakeholders have voice in decision processes and which do not, recognizing that those without voice are most likely to be overlooked.

The core analytical discipline is the question: "Who is missing?" After initial stakeholder identification, explicit effort is directed toward populations who might be affected but are not yet listed. Prompts include: people whose jobs may be displaced, people who provide competing solutions, people who interact with beneficiaries, people downstream from the change, people who lack access to the solution, and future generations.

4.3 Historical Precedent Consultation

Each checkpoint includes consultation of historical precedent relevant to the decision at hand. For stakeholder identification, this means asking what similar products or policies have affected which populations. For incentive analysis, this means examining how similar systems have been gamed or produced unintended behaviors. For testing, this means identifying validation gaps that have caused problems in comparable situations. For launch decisions, this means examining how similar products have fared when problems emerged.

Historical precedent consultation does not require original research at each checkpoint. Reference materials and case databases can be consulted efficiently. The requirement is that consultation occur and be documented, not that exhaustive research be conducted.

4.4 Documentation and Accountability

Each checkpoint produces documented outputs that create accountability trails. Documentation serves multiple functions: it forces explicit articulation of reasoning that might otherwise remain implicit; it creates records that can be audited; it distributes responsibility across participants rather than concentrating it; and it enables learning from both successes and failures.

Sign-off requirements identify specific individuals accepting responsibility for proceeding at each checkpoint. This counters the ethical abdication mechanism by ensuring that someone is explicitly accountable for stakeholder impacts.

---

5. Validation

5.1 Historical Validation Study

The framework was validated against a database of 161 historical cases spanning approximately 5,000 years, categorized across 11 domains: infrastructure and engineering, medicine and public health, transportation, environmental, industrial and workplace, financial and economic, military and conflict, technology and communications, governance and policy, social and cultural, and agriculture and food.

For each case, analysis assessed whether documented historical precedent was available at the time key decisions were made. "Available" was defined as precedent that was published, taught in relevant professional education, or otherwise accessible to decision-makers through reasonable diligence. The assessment used only information that was contemporary to the decision; hindsight knowledge was excluded.

The central finding is that 154 of 161 cases (95.7%) had documented precedent available at decision time. The average gap between relevant precedent and the analyzed case was 47 years. In numerous instances, the precedent was extensively documented, widely taught, and directly relevant to the decision at hand.

Seven cases (4.3%) were assessed as genuinely first-of-kind, where no relevant precedent was available. Even in these cases, contemporaneous critics often raised concerns that went unheeded.

The complete case database with individual assessments is available in a separate validation study document.

5.2 Illustrative Example: Historical Precedent Chain

The Triangle Shirtwaist Factory fire of March 25, 1911, killed 146 garment workers in New York City. Contributing factors included locked exit doors (to prevent theft and unauthorized breaks), inadequate fire escapes, flammable materials, and overcrowded conditions. The disaster prompted extensive reforms including factory safety legislation, fire code requirements, and became the foundational case study in occupational safety education worldwide.

One hundred two years later, on April 24, 2013, the Rana Plaza garment factory collapsed in Bangladesh, killing 1,134 workers. Contributing factors included ignored structural warnings (cracks had been observed the day before), locked exits during the collapse, unauthorized additional floors, and pressure on workers to continue despite known dangers.

The precedent chain is direct and documented. Triangle Shirtwaist is taught in occupational safety courses globally (Von Drehle, 2003). The specific failure modes (locked exits, ignored warnings, prioritizing production over safety) had been documented, analyzed, and incorporated into professional education for over a century.

The question is not whether Rana Plaza decision-makers could have predicted the collapse. The question is whether they consulted the extensively documented precedent on garment factory safety that had been available for 102 years. The evidence indicates they did not.

This example illustrates the consultative premise: the framework does not require prediction of novel harms. It requires consultation of documented precedent. For the vast majority of cases, that precedent exists.

5.3 Corporate Disaster Analysis

A parallel analysis examined 151 corporate disasters from 1970-2025, focusing on modern cases with sufficient documentation for cost-benefit analysis. Each case was assessed across 13 data points including industry, harm type, affected populations, prevention cost estimates, actual disaster costs, and identifiability of harm at decision time.

Key findings include: average prevention costs ranged from 1-10% of eventual disaster costs. Return on investment for prevention ranged from 1,000x to 10,000x. The same pattern of failing to consult available precedent appeared consistently across industries and time periods.

The complete corporate disaster analysis with methodology and individual case assessments is available in a separate document.

5.4 Illustrative Example: Corporate Cost-Benefit

The Deepwater Horizon oil platform exploded on April 20, 2010, killing 11 workers and releasing approximately 4.9 million barrels of oil into the Gulf of Mexico. Investigation revealed that BP had received warnings about cement job failures on the well but proceeded without addressing them.

Relevant precedent was extensive. The Ford Pinto fuel tank cases of the 1970s documented the consequences of cost-benefit analyses that accepted preventable deaths and the subsequent legal and reputational costs. The Bhopal disaster of 1984 documented the consequences of deferred maintenance and safety shortcuts at industrial facilities. The Exxon Valdez oil spill of 1989 documented the consequences of maritime oil disasters and their long-term environmental and financial impacts. All three cases were extensively documented, taught in business schools, and directly relevant to decisions being made at BP.

The estimated cost of addressing the cement job warnings was $10-50 million (National Commission on the BP Deepwater Horizon, 2011). The actual cost of the disaster exceeded $65 billion in cleanup, settlements, fines, and stock value losses. The return on investment for prevention would have been 1,300x to 6,500x.

Framework checkpoint questions would have surfaced relevant concerns. Checkpoint 1's "Who else is affected beyond direct beneficiaries?" would have identified Gulf fishing communities, coastal tourism, and marine ecosystems. Checkpoint 2's "What happens if we over-optimize for cost and speed?" would have surfaced safety shortcut risks. Checkpoint 4's "Can we defend this publicly?" would have revealed the indefensibility of accepting known risks to save relatively small sums.

5.5 Prospective Application Example

The preceding examples demonstrate retrospective application of the framework to past disasters. To demonstrate prospective application, the framework was applied to a current policy decision: Australia's social media ban for users under 16.

Australia's Online Safety Amendment (Social Media Minimum Age) Act 2024 took effect December 10, 2025, prohibiting social media platforms from allowing users under 16 to create or maintain accounts. Platforms face fines up to AUD $50 million for systemic failures to prevent underage access. The policy was framed as child protection, citing concerns about teen mental health, social media addiction, and cyberbullying.

Application of Checkpoint 1's core question, "Who else is affected beyond direct beneficiaries?", surfaces stakeholders not addressed in policy deliberation:

LGBTQ+ youth in unsupportive environments face loss of affirming communities that may represent their only source of support. For this population, online connections can be protective against self-harm. Severity: Critical.

Youth in abusive households face loss of external communication lifelines, potentially isolating them with abusers. Severity: Critical.

Neurodivergent youth face loss of accessible socialization formats; online communication is often more manageable than in-person interaction for some neurodivergent individuals. Severity: High.

Rural and geographically isolated youth face digital exclusion compounding existing geographic disadvantage. Severity: High.

Chronically ill and disabled youth face loss of disease-specific peer support communities. Severity: High.

The general population faces surveillance infrastructure precedent, as age verification systems require identity documentation for internet access. Severity: High.

Application of Checkpoint 2's incentive questions reveals structural problems. Platforms have no incentive to improve safety features, as they are banned regardless of design choices. Teenagers are incentivized to circumvent the ban via VPNs and age falsification. Compliant youth are disadvantaged relative to peers who circumvent. Parents receive false security while losing visibility into actual online activity.

Application of Checkpoint 3's consultation gap question identifies that affected youth themselves, LGBTQ+ youth organizations, youth mental health experts (many of whom publicly opposed the ban), disability advocates, and digital rights organizations were not included in policy deliberation.

This analysis was conducted using only information available at policy passage. The framework does not predict outcomes; it surfaces stakeholders and incentive structures that systematic consultation would have identified. The question is not "What will happen?" but "Who is not being considered, and what does historical precedent suggest about similar interventions?"

This example demonstrates that the framework is not limited to retrospective analysis of past disasters. The same methodology applies prospectively to current decisions, surfacing stakeholder impacts before they materialize.

5.6 Methodological Integrity

Several potential biases were addressed in validation methodology.

Hindsight bias was addressed by using only information contemporary to each decision when assessing identifiability. Analysts asked what was knowable at decision time, not what is known now. This required explicit attention to publication dates, professional education curricula of the relevant period, and accessibility of information to actual decision-makers.

Selection bias was addressed by including failures across all domains rather than selecting cases likely to support the framework. The historical database includes cases where the framework would have been less useful alongside those where it would have been highly effective.

Survivorship bias was acknowledged as an unavoidable limitation. The analysis necessarily focused on disasters that occurred rather than cases where harm was successfully prevented. This means the validation demonstrates that the framework would have surfaced missed harms in disaster cases, not that it would have prevented all such disasters.

---

6. Limitations and Considerations

6.1 What the Framework Cannot Do

The Harm Blindness Framework cannot guarantee harm prevention. Even comprehensive stakeholder analysis may miss impacts, and organizational dynamics may override checkpoint findings. The framework increases the probability that harms are identified; it does not eliminate the possibility of harm.

The framework cannot identify genuinely novel harms for which no precedent exists. The 4.3% of validated cases that were first-of-kind situations illustrate this limitation. For truly unprecedented situations, the framework offers less value, though the discipline of systematic stakeholder identification remains useful.

The framework cannot force implementation. Like all governance tools, its effectiveness depends on organizational commitment. Checkpoints can be conducted perfunctorily, with box-checking replacing genuine analysis. Documentation can be fabricated. Sign-offs can be obtained without meaningful review. The framework creates accountability structures, but accountability requires enforcement.

The framework cannot replace human judgment. Checkpoint questions prompt analysis; they do not dictate conclusions. Determining which stakeholder impacts are acceptable, which trade-offs are warranted, and which risks should be taken requires human judgment that the framework informs but does not supplant.

6.2 The Hindsight Objection

A reasonable objection to any harm prevention framework is that it relies on hindsight. Decision-makers at the time could not have known what would happen; identifying harms retrospectively is easy but unfair to those who made decisions under uncertainty.

This objection misunderstands the methodology. The framework does not require prediction. It requires consultation of available historical precedent.

For 95.7% of analyzed cases, documented precedent existed at decision time. The question is not "Could they have predicted this?" but "Did they consult the precedent that existed?" The Triangle Shirtwaist fire was documented for 102 years before Rana Plaza. The Ford Pinto cases were documented for decades before Deepwater Horizon. The opioid crisis of the early 20th century was documented before pharmaceutical manufacturers promoted aggressive prescribing in the 1990s (Courtwright, 2001).

The pattern across 5,000 years of cases is not that harms were unpredictable. The pattern is that decision-makers did not look.

The framework operationalizes looking. It does not claim that looking guarantees correct answers or that decision-makers who look will always make right choices. It claims that the systematic failure to consult available precedent is a correctable problem, and that correcting it would surface harms that are currently being missed.

6.3 Genuinely First-of-Kind Cases

Seven cases in the validation database (4.3%) were assessed as genuinely first-of-kind, where no relevant precedent was available at decision time. The framework acknowledges its limitations for such situations.

However, even in first-of-kind cases, contemporaneous critics often raised concerns. The framework's stakeholder identification methodology would surface these concerns even when historical precedent is unavailable. The discipline of asking "Who

else is affected?" and "Who is missing from this analysis?" has value independent of precedent consultation.

For genuinely novel technologies with no historical parallels, the framework provides less guidance. This limitation should be acknowledged rather than obscured.

6.4 Implementation Challenges

Several implementation challenges should be anticipated.

Checkpoint fatigue may develop if the process is perceived as bureaucratic overhead rather than valuable analysis. Maintaining engagement requires demonstrating that checkpoints produce actionable insights and that findings are taken seriously.

Box-checking may replace genuine analysis. Organizations under pressure may conduct checkpoints perfunctorily, completing documentation without meaningful stakeholder consideration. Mitigation requires external facilitation and periodic audits.

Power dynamics may distort sign-offs. Junior employees may be reluctant to raise concerns that contradict senior preferences. The facilitator role addresses this partially, but organizational culture significantly affects checkpoint effectiveness.

Resource constraints may limit thoroughness. Comprehensive stakeholder analysis and precedent consultation require time that competitive pressures may not allow. The framework is more effective when organizations commit adequate resources.

6.5 Cultural Context Limitations

The framework was developed within Western technological and philosophical contexts. Stakeholder identification patterns, power dynamic analysis, and harm prioritization may reflect cultural assumptions that are not universal.

Individualist cultures may identify stakeholders differently than collectivist cultures. Power dynamics operate differently across cultural contexts. Harm valuation differs across societies. These variations mean that checkpoint questions may require adaptation for non-Western organizational and cultural contexts.

This limitation is being actively addressed through cross-cultural validation research (see Section 7.6.1), but practitioners should be aware that the framework as presented reflects its development context.

This acknowledgment itself demonstrates the framework's methodology. Applying the question "Who is missing from this analysis?" to the framework itself reveals Western-centric assumptions as a blind spot requiring attention.

---

7. Discussion

7.1 Implications for Practice

The framework integrates with existing development workflows rather than requiring separate ethics processes. Checkpoint timing aligns with standard phase-gate development models (Cooper, 2008) and can be incorporated into agile sprint cycles as recurring ceremonies. The methodology is compatible with waterfall, agile, and hybrid development approaches.

Scaling considerations vary with organization size. Small organizations may conduct checkpoints informally with external advisors serving as facilitators. Large organizations may establish dedicated checkpoint facilitation functions with rotation to prevent capture. The core methodology remains consistent; implementation mechanics adapt to context.

Integration with existing compliance and governance functions is addressed in a separate implementation guide. Checkpoint documentation can serve multiple purposes, satisfying both stakeholder analysis requirements and existing regulatory obligations.

7.2 Motivation-Agnostic Design

The framework functions regardless of whether implementing organizations are motivated by ethical concerns or risk management. The same checkpoint questions surface the same stakeholder impacts; only the framing differs.

For ethics-motivated organizations, the framework operationalizes stakeholder consideration systematically. Rather than relying on individual sensitivity to ethical concerns, the methodology ensures that stakeholder identification occurs through structured analysis.

For risk-motivated organizations, the framework identifies legal, financial, and reputational exposure before it materializes. Stakeholder harms frequently become litigation, regulatory action, public relations crises, and shareholder value destruction. Early identification provides economic value.

This design choice addresses the fundamental reach limitation of ethical frameworks. Frameworks that require ethical motivation for adoption will never achieve sufficient reach to prevent harm at scale. By producing identical analytical outputs regardless of motivation, the methodology can be implemented by organizations across the motivational spectrum while delivering equivalent stakeholder protection.

7.3 Implications for Policy

The framework has potential applications in regulatory contexts. Procurement requirements could mandate stakeholder analysis documentation for government contracts, creating incentives for adoption without prescriptive regulation. Self-regulatory industry standards could incorporate checkpoint requirements as best practices. Formal regulation could require documented stakeholder analysis for high-risk systems, similar to environmental impact assessment requirements.

The procurement-based approach is particularly promising because it operates through contractual requirements rather than legislative mandates, avoiding some political obstacles to regulation while creating strong adoption incentives for organizations seeking government business.

7.4 Integration with Existing Frameworks

The Harm Blindness Framework is designed to complement rather than replace existing governance approaches. Mapping to ISO 31000 risk management standards shows alignment between checkpoint analysis and risk identification, assessment, and treatment processes. Mapping to COSO Enterprise Risk Management shows compatibility with governance, strategy, and performance integration. The framework's stakeholder analysis methodology can inform NIST AI Risk Management Framework implementation, particularly the "Map" function's context establishment requirements.

Detailed mapping to these and other standards is provided in a separate implementation guide. The academic contribution of this paper is the methodology itself; compliance crosswalks are practitioner resources.

7.5 Systems with Elevated Harm Potential

For systems where checkpoint analysis identifies potential for preventable death or serious irreversible harm, additional escalation procedures are warranted. These provisions, detailed in a separate technical specification, create elevated scrutiny requirements proportional to the severity of potential harm. Such procedures include enhanced documentation, external review requirements, and explicit accountability mechanisms. The principle is that higher-stakes decisions warrant additional safeguards, while the checkpoint methodology remains consistent across risk levels.

7.6 Future Research Directions

Several directions for future research would strengthen and extend this work.

Cross-Cultural Validation. The framework was developed within Western technological and philosophical contexts. This limitation requires active address. Ongoing collaborative research is applying the methodology within Ubuntu philosophical frameworks in Southern Africa. Ubuntu, often translated as "I am

because we are," centers collective well-being in ways that may surface stakeholder relationships not visible through individualist analytical lenses. Such cross-cultural validation is methodologically necessary, not merely ethically desirable. The framework applied to itself reveals this blind spot; addressing it demonstrates the methodology's reflexive application.

Longitudinal Implementation Studies. Validation to date demonstrates that the framework would have surfaced harms in historical cases. Demonstrating that implementation actually prevents harm requires longitudinal studies tracking organizations that adopt the methodology. Such studies face the counterfactual problem (harms prevented are not observed), but proxy metrics including issues identified at each checkpoint, cost of interventions, and stakeholder satisfaction can provide evidence of effectiveness.

Automated Assistance Tools. The framework's precedent consultation requirement could be supported by tools that match current decisions to historical cases, suggest relevant precedent, and prompt stakeholder identification. Such tools should support rather than replace human judgment; the risk of automation is that it further enables box-checking without genuine analysis. Research should explore how to design tools that enhance rather than undermine analytical rigor.

Sector-Specific Adaptations. While the framework is designed for general application, sector-specific adaptations may enhance effectiveness. Healthcare applications may require integration with existing bioethics frameworks. Financial services applications may require alignment with fiduciary duties. Government and public sector applications may require adaptation to public accountability requirements. Research in specific sectors can identify necessary adaptations.

Effectiveness Metrics. Measuring "harms prevented" faces the fundamental counterfactual problem: prevented harms are not observed. Research should develop proxy metrics that provide evidence of framework effectiveness without requiring observation of counterfactual outcomes. Candidates include checkpoint-identified issues, intervention costs, stakeholder feedback, and comparative analysis across adopting and non-adopting organizations.

---

8. Conclusion

8.1 Summary of Contributions

This paper makes four contributions. Theoretically, it introduces harm blindness as a systematic phenomenon arising from identifiable mechanisms rather than individual moral failures. Methodologically, it presents a checkpoint-based consultative

approach that embeds stakeholder analysis at decision points throughout development. Empirically, it validates the approach against 161 historical cases and 151 corporate disasters, finding that 95.7% of analyzed harms had documented precedent available at decision time. Practically, it provides an implementable framework that functions across motivational contexts.

8.2 Core Finding

The pattern across 5,000 years of analyzed cases is not that harms were unpredictable. The pattern is that decision-makers consistently failed to consult available historical precedent. 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. It does not guarantee that looking will prevent all harm, or that every organization that looks will make correct decisions. It ensures that the systematic failure to consult available precedent, a correctable problem, is addressed through structured methodology.

8.3 Availability

The framework is available for implementation. Supporting materials including implementation guides, checkpoint templates, and complete validation studies are publicly available at realsafetyai.org. Organizations seeking to adopt the methodology can access these resources without cost. Feedback from implementation experience will inform continued development.

References

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.

Author Information

Travis Gilly is Executive Director of Real Safety AI Foundation, a nonprofit organization focused on AI safety research, education, and actionable safety frameworks. Contact: t.gilly@ai-literacy-labs.org

---

Acknowledgments

The author acknowledges ongoing collaborative discussions with researchers applying this framework in cross-cultural contexts, whose work informs the future research directions discussed herein.

The author used Claude Opus 4.5 (Anthropic) as assistive technology to accommodate executive function challenges associated with ADHD and autism. AI assistance was used for organizational tasks, formatting, citation verification, and iterative editing. All intellectual content, threat analysis, framework development, and conclusions are the author's original work.

Supplementary Materials

The following materials are available separately:

- **Historical Validation Study**: Complete database of 161 cases with individual assessments
- **Corporate Disaster Analysis**: Complete database of 151 cases with cost-benefit calculations
- **Implementation Guide**: Detailed guidance for organizational adoption, including GRC framework mapping
- **Checkpoint Templates**: Fillable worksheets for each of the four checkpoints
- **High-Risk System Protocols**: Technical specification for elevated harm potential situations

All materials available at: realsafetyai.org/database & realsafetyai.org/framework