Post-Deployment Observability as a Foundation for Well-Being-Aligned Human–AI Co-Evolution

Andy Skumanich[1], Han Kyul Kim[2]
[1]Innov8AI
[2]University of Southern CA
askuman@Innov8AI.com, hankyulk@usc.edu

Abstract

Current discussions of human creativity and generative AI often focus on model capabilities at the point of release, framing outcomes in terms of augmentation (helps individuals) versus replacement (harms individuals). In deployed settings, however, the effects of AI systems on human agency, creativity, and institutional well-being emerge over time, shaped by repeated interaction, reuse, and integration into real-world workflows. These dynamics are rarely visible through pre-deployment evaluation or isolated prompt–response analysis.

This paper argues that post-deployment observability is a foundation for well-being-aligned human–AI co-evolution. We present a system-level framework for externalized behavioral monitoring that treats generative AI systems as participants in socio-technical ecosystems rather than static tools. The framework emphasizes interpretable, aggregate behavioral signals - such as shifts in output velocity, semantic and structural reuse, persistence of synthetic roles, and cross-context propagation - that emerge cumulatively through time.

Rather than automating judgment or enforcement, these signals support human-in-the-loop interpretation, enabling earlier awareness of when AI use patterns may be drifting from creative augmentation toward automation pressure, authority substitution, or unintended displacement of human agency. By focusing on observation instead of prediction, and governance rather than control, the proposed approach complements existing alignment and safety practices while preserving human judgment, institutional choice, and long-term well-being.

1 Introduction

People usually judge AI's impact on creativity by looking at what it can do when it first launches, and by asking whether it helps humans or replaces them. This misses the dynamics. AI doesn't just either help- or replace- people, it reshapes how people think and create over time, and we can only understand that by watching how it's actually used.

Generative AI systems are increasingly deployed in open, high-stakes environments where their outputs influence decisions, perceptions, and downstream actions at scale. For instance, the Department of Transportation employees are now being directed to use Gemini to draft new safety regulations, a "vanguard of a broader federal effort" (ProPublica 2026). Public and private institutions are often required to rely on pre-deployment assurances about system behavior, even as deployed systems continue to evolve by real-world use. This creates a growing need for reliable situational awareness once generative AI systems are operating in practice, without solely looking at the point of release.

In response, substantial research and engineering effort has focused on improving the safety of generative AI systems prior to deployment, including alignment training, red-teaming, and rule-based guardrails. These approaches play a critical role in reducing certain classes of harmful outputs and establishing baseline expectations for model behavior.

However, once deployed, generative AI systems do not operate in isolation. They interact with users, applications, and institutional workflows over extended periods, often in ways that were not fully anticipated during development or testing. (Amodei 2016) (Weidinger 2022)

As systems move from controlled evaluation settings into open deployment environments, new challenges for safety assessment emerge. Pre-deployment evaluations typically emphasize the behavior of individual prompt–response interactions, while real-world operation is shaped by patterns of use, reuse, and integration into larger automated or semi-automated processes. These downstream dynamics can influence system behavior in ways that are difficult to characterize through static testing or model-internal analysis alone.

Despite growing recognition of these challenges, there remains limited guidance on how institutions should monitor and interpret the behavior of deployed generative AI systems once they are operating at scale. Existing safety mechanisms provide limited visibility into post-deployment use

patterns, while more invasive monitoring approaches raise legitimate concerns about privacy, surveillance, and inappropriate automation of enforcement. This leaves a gap between pre-deployment assurances and post-deployment awareness. (Raji et al. 2020), (NIST AI RMF 2023)

The primary contribution of this paper is a post-deployment observability framework for generative AI systems that supports early situational awareness through interpretable, aggregate behavioral signals. Although not proposing new alignment techniques or content-level classifiers, this work focuses on how system-level monitoring and human-in-the-loop oversight can complement existing safety practices while preserving institutional control and responsible governance. (Skumanich et al., 2025), (Kim et al. 2025)

This paper examines how institutions can achieve post-deployment awareness of generative AI systems operating in open environments. We focus on the challenge of observing risks that arise not from individual prompt–response interactions, but from aggregate behavior that accumulates across repeated use, reuse, and propagation through time.

We argue that this gap cannot be addressed through model-centric evaluation alone, and we introduce a post-deployment observability framework based on interpretable behavioral indicators and human-in-the-loop oversight. Together, these elements provide a practical foundation for monitoring deployed generative AI systems responsibly, without requiring privileged access to model internals or automated enforcement. Figure 1 illustrates the distinction between pre-deployment safety mechanisms and the post-deployment observability gap that emerges as systems interact with real-world environments.

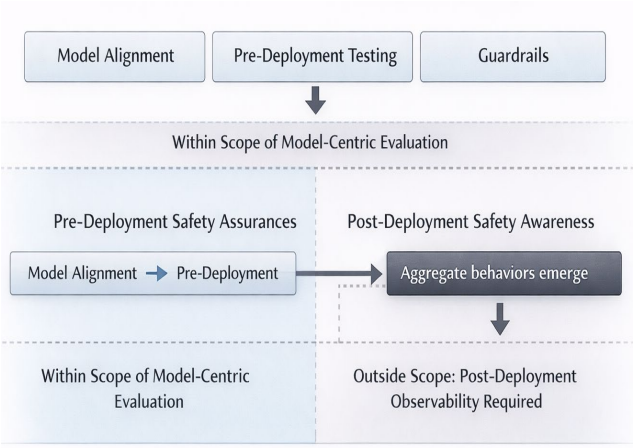

Figure 1: Pre-Deployment Safety and the Post-Deployment Observability Gap. Pre-deployment mechanisms such as alignment training, red-teaming, and guardrails provide baseline protections, but once systems are deployed, aggregate behaviors can emerge that are not visible through model-centric evaluation, creating a gap that motivates system-level monitoring.

We suggest that aligning AI systems with long-term human well-being requires moving beyond static pre-deployment alignment toward institutionalized post-deployment observability and adaptive governance.

Note that this paper does not propose detection or attribution mechanisms, automated enforcement, or model-level access requirements; instead, it reframes post-deployment observability as a diagnostic lens for understanding system-level behavioral dynamics in deployed AI systems. In this context, observability refers to systematic visibility into system behavior and socio-technical interaction patterns after deployment, extending beyond traditional monitoring or periodic auditing.

2 Scope, Assumptions, and Observability Boundaries

This section clarifies the scope of risks addressed in this work, the assumptions under which post-deployment monitoring is considered, and the resulting boundaries on what can be observed. These framing choices are intentional and shape both the limitations and the design of the observability framework introduced in subsequent sections.

Scope of Post-Deployment Risks

The risks considered in this paper arise from patterns of system use over time, rather than from individual prompt–response interactions. In deployed settings, generative AI systems are often used repeatedly, embedded within workflows, or combined with other tools and platforms. Through these dynamics, behaviors may emerge that are not evident in isolated interactions and are difficult to anticipate during pre-deployment testing.

Accordingly, this work focuses on risks that emerge through repetition, reuse, and propagation, including behaviors that become salient only when observed across multiple interactions, contexts, or deployment surfaces. Such risks are shaped by interaction dynamics and system integration rather than by the content of any single response. We deliberately avoid enumerating specific misuse domains or application categories, as the framework is intended to apply broadly across deployment contexts.

Assumptions and Constraints

The proposed framework operates under a set of explicit assumptions designed to reflect realistic institutional constraints and to preserve responsible governance practices. Deployed generative AI systems are treated as black-box components, without reliance on privileged access to model internals. (Raji et al. 2020) (Selbst et al. 2019)

In particular, we assume no access to:

- model weights or architectural details,
- training data or fine-tuning corpora,

• internal model logs beyond what is externally observable through interaction surfaces.

In addition, the framework does not rely on persistent user identification, user profiling, or cross-session tracking of individuals. (Selbst et al. 2019) (NIST AI RMF) Observations are limited to system-level patterns, not just user-level behavior. The framework also avoids automated enforcement or punitive action, emphasizing observation, documentation, and human judgment instead.

These constraints are not limitations of the approach but design requirements that support privacy preservation, institutional legitimacy, and deployment feasibility in regulated or high-trust environments.

What This Framework Is Not

To avoid misinterpretation, it is important to clarify what the proposed framework does not attempt to provide. It is not a content moderation system, nor does it aim to classify individual outputs as harmful or benign. It is not a misuse detection or attribution system, and it does not infer user intent or adversarial motivation.

The framework is also not a replacement for pre-deployment alignment, red-teaming, or guardrail strategies. Rather, it is designed to complement these mechanisms by addressing risks that arise after deployment, under conditions where pre-deployment assurances alone are insufficient. Related work has examined system-level and institutional responses to harmful information ecosystems without treating individual models or outputs as the sole unit of analysis. (Barbaro & Skumanich 2023), (Skumanich & Kim 2024)

By explicitly bounding the scope in this way, the framework avoids overreach and maintains a clear separation between observability, interpretation, and institutional response.

Implications for Observability

Taken together, the scope and constraints outlined above have direct implications for how post-deployment risks can be observed. Because risks are defined at the system level and emerge through repeated use, observability must rely on aggregate behavioral signals rather than individual interactions. Because these behaviors unfold over extended periods, temporal analysis is essential for distinguishing persistent patterns from transient noise.

Finally, because the framework avoids automated judgment and enforcement, observed signals must remain interpretable and reviewable by humans, supporting contextual assessment and governance decisions. These requirements motivate the focus on aggregate, time-dependent indicators and human-in-the-loop oversight developed in the following sections.

3 The Observability Gap in Current AI Safety Approaches

Recent efforts to improve the safety of generative AI systems have focused primarily on pre-deployment mechanisms, including alignment training, red-teaming, and rule-based guardrails. These approaches are necessary foundations for responsible deployment and have demonstrated value in reducing certain classes of harmful outputs. However, they are structurally limited in their ability to capture risks that arise after deployment, when systems interact with diverse users, platforms, and workflows in open-ended environments.

Once deployed, generative AI systems participate in complex socio-technical ecosystems. Users adapt prompts as they progress, reuse successful interaction patterns, and combine model outputs with external tools, platforms, or automated pipelines. Harmful behaviors – including scalable social engineering, impersonation, or coordinated misuse - often emerge from these interactions and not from isolated model responses.

As a result, risks may become visible only through downstream behavioral patterns that are not accessible through model-internal evaluation or static testing regimes.

Examples include sustained reuse of semantically similar outputs across different contexts, abrupt shifts in output volume consistent with automation or coordination, the maintenance of persistent synthetic roles or personas across interactions, and the migration of generative strategies between platforms, applications, or deployment settings. Individually, these behaviors may appear benign; collectively and at scale, they can signal emergent misuse that is difficult to detect through model-internal evaluation alone.

Illustrative Examples

To make this distinction concrete, consider a generative writing assistant deployed across multiple teams within an organization. Individually, its outputs appear appropriate and helpful in each local context. Over time, however, aggregate observation reveals the repeated reuse of highly similar narrative structures across unrelated tasks—an effect that is not visible through isolated prompt–response evaluation. As a second illustrative example, a monitoring system may reveal patterns of structurally similar outputs - sharing the same argumentative flow, formatting, or call-to-action elements - are repeatedly generated across diverse prompts and sessions, a pattern that is difficult to explain through independent user queries alone.

This distinction between model-centric evaluation and system-level behavior is central to the argument of this paper. It creates an observability gap between pre-deployment safety assurances and post-deployment system behavior.

Model-centric evaluations emphasize correctness, alignment, or refusal behavior at the level of individual prompts,

while many real-world harms manifest as aggregate phenomena – this includes repetition, amplification, reuse, and cross-context propagation - that emerge cumulatively over time across successive interactions. These effects are external to the model boundary and are therefore difficult to detect through access to model weights, training data, or prompt–response logs alone. Crucially, behaviors that appear innocuous in isolation may become salient only through their persistence and evolution. As a result, risks that are invisible at the level of individual interactions may become apparent only when behavior is observed collectively and through time.

Importantly, this gap does not imply that existing safety approaches are ineffective or misguided. Rather, it reflects a mismatch between the scope of current evaluation methods and the environments in which generative AI systems operate at scale. Post-deployment risks are shaped by interaction dynamics, usage incentives, and institutional contexts that extend beyond the model itself. Addressing these risks therefore requires complementary mechanisms that operate at the system level, observing deployed behavior without assuming privileged access to model internals or invasive monitoring of individual users.

Post-deployment observability enables institutions to detect signals such as behavioral drift, emergent authority substitution, or user dependency, allowing responsible actors to interpret these signals and intervene through model retraining, interface redesign, governance or policy adjustments.

The remainder of this paper operationalizes this distinction between individual interactions and aggregate system behavior by identifying interpretable behavioral indicators and outlining a post-deployment oversight architecture capable of observing such aggregate phenomena responsibly.

4 Externalized Behavioral Signals as Early-Warning Indicators

Each of the following indicators reflects aggregate behavior that manifests across interactions over time and is therefore not observable at the level of individual prompts.

To address the post-deployment observability gap, we propose focusing on externalized behavioral signals - aggregate patterns that emerge from how generative AI systems are used, reused, and embedded within broader workflows. These signals are not intended to classify individual outputs as harmful or benign. Instead, they function as early-warning indicators that support human judgment by highlighting anomalous or concerning system-level behavior.

We emphasize that these indicators are probabilistic and context-dependent. They do not constitute definitive evidence of misuse, nor are they designed for automated enforcement. Their value lies in enabling earlier awareness of emerging risks that would otherwise remain latent until downstream impacts become difficult to mitigate.

Below we outline several classes of behavioral indicators relevant to post-deployment monitoring. Figure 2 presents a conceptual architecture for post-deployment observability, showing how external interaction signals can be aggregated into interpretable behavioral indicators to support human oversight.

Output Velocity and Volume Anomalies

Sudden changes in the volume or frequency of AI-mediated outputs associated with particular tasks, prompts, or interaction patterns may indicate automation, scaling, or coordinated use. While increased usage alone is not inherently harmful, abrupt or sustained deviations from baseline behavior can signal the emergence of misuse pipelines or unintended amplification.

Semantic and Structural Reuse

The repeated appearance of highly similar outputs - templates, narratives, or response structures - across different contexts through time may suggest the reuse of generative patterns optimized for persuasion, deception, or impersonation. Such reuse can occur even when surface-level prompts differ, reflecting downstream copying, adaptation, or automation rather than direct prompting behavior.

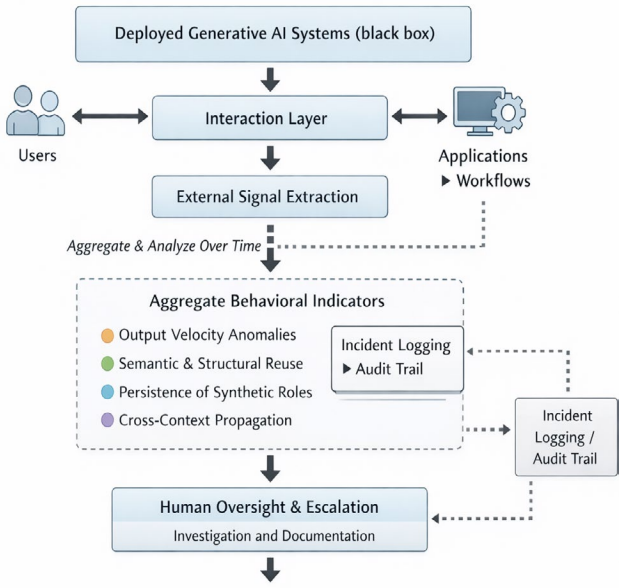

Figure 2: Post-Deployment Observability Architecture for Generative AI Systems. A conceptual architecture in which deployed systems are treated as black-box components and external interaction signals are aggregated over time to derive interpretable behavioral indicators that support human review, incident documentation, and escalation, without requiring model access or automated enforcement.

Persistent Synthetic Personas or Roles

In some cases, generative systems are used to maintain consistent personas, roles, or identities across interactions. While persona persistence can be benign or intentional, its emergence at scale, particularly in contexts involving persuasion, authority, or trust, and may warrant closer scrutiny as an indicator of coordinated or deceptive use.

Cross-Context Propagation Patterns

When similar AI-generated content or behaviors appear across multiple platforms, applications, or deployment contexts, this may indicate broader reuse or migration of generative strategies. Such propagation patterns are difficult to detect through isolated system monitoring but become visible through aggregate behavioral analysis.

These indicator classes are intentionally interpretable and agnostic to specific domains or platforms. They are designed to support human analysts in forming situational awareness rather than to replace judgment with automated decision-making.

5 Post-Deployment Oversight Architecture

The architecture described below is explicitly designed to observe aggregate behavioral indicators that emerge across successive interactions over time, without individual prompts or users. Building on the concept of externalized behavioral signals, we outline a conceptual post-deployment oversight architecture for generative AI systems. The architecture is designed to complement existing safety mechanisms by providing continuous, system-level visibility into deployed behavior while preserving privacy, institutional control, and human oversight.

At its core, the architecture treats deployed generative AI systems as black-box components whose internal representations are not directly inspected. Instead, monitoring operates at the interaction and aggregation layers, focusing on patterns that emerge across usage instead of individual prompts or users.

The architecture consists of the following functional components:

a. Interaction Surface: Deployed generative AI systems interact with users, applications, or downstream tools through defined interfaces. These interactions produce observable artifacts – like e.g. response metadata, structural features, or aggregate usage statistics - that can be collected without accessing model internals or storing raw content.

b. Signal Extraction and Aggregation: Observable artifacts are transformed into higher-level behavioral signals, e.g. velocity measures, reuse patterns, or persistence indicators. Aggregation occurs over defined temporal windows. Aggregation also occurs across contexts to reduce sensitivity to individual interactions and to emphasize system-level behavior.

c. Risk Assessment and Triage: Aggregated signals are evaluated against baseline expectations and contextual thresholds to identify patterns that may warrant attention. This stage is explicitly designed to surface questions but not specifically to generate determinations, prioritizing interpretability over automation.

d. Human-in-the-Loop Review: Flagged patterns are reviewed by human analysts or institutional stakeholders who interpret signals in light of domain knowledge, operational context, and ethical considerations. Decisions regarding follow-up actions, documentation, or escalation remain human-led. (Kim, et al. 2025)

e. Incident Documentation and Audit Trail: Observations, assessments, and responses are documented to support accountability, learning, and external review. This documentation enables retrospective analysis and aligns with emerging expectations for transparency and auditability in high-risk AI deployments.

Figure 3 illustrates representative patterns of aggregate, post-deployment behavioral signals that this observability layer is designed to surface.

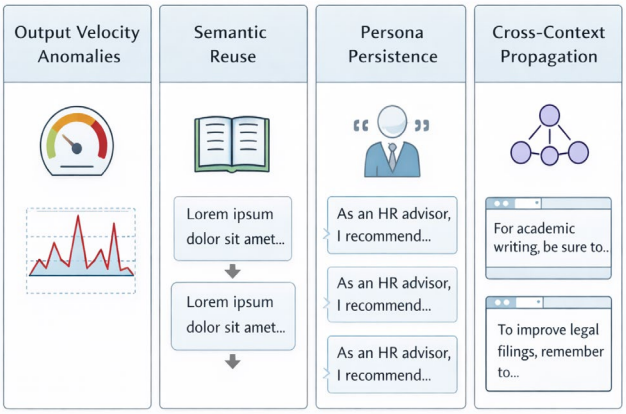

Figure 3: Examples of Aggregate Behavioral Indicators for Early Situational Awareness. Representative classes of aggregate signals, including anomalous output velocity, semantic or structural reuse, persistence of synthetic roles, and cross-context propagation, whose significance emerges through accumulation and temporal analysis rather than isolated interactions.

6 Post-Deployment Observability & Human–AI Co-Evolution

Discussions of human creativity and generative AI are often framed in static terms, emphasizing whether systems augment or replace human capabilities at the moment of deployment. In deployed settings, however, these outcomes are not fixed properties of a model. They emerge through ongoing interaction between humans, AI systems, and institutional workflows. Human–AI relationships therefore evolve over time, shaped by patterns of reuse, automation incentives, and downstream integration instead of by isolated system outputs.

From this perspective, creativity, agency, and well-being are not determined solely by what generative systems can produce, but by how they are taken up and embedded in practice. Repeated reliance on AI-generated structures, narratives, or decision scaffolding can gradually shift the locus of creative labor and judgment, even when individual interactions appear benign or productive. These shifts often occur incrementally, making them difficult to recognize until they have become entrenched.

Post-deployment observability provides a mechanism for making such co-evolutionary dynamics visible while meaningful choices remain available. Aggregate behavioral indicators - such as sustained increases in output velocity consistent with automation pressure, repeated semantic or structural reuse that narrows creative variation, persistence of synthetic roles that substitute for human authority, or cross-context propagation of AI-mediated practices - can signal when systems are moving from supporting human creativity toward constraining or displacing it. Importantly, these indicators do not assess intent, quality, or harm at the level of individual outputs. Their significance emerges only through accumulation and temporal analysis.

Framed in this way, observability functions as an enabling layer rather than a control mechanism. It does not prescribe how institutions or individuals should respond, nor does it automate judgments about acceptable or unacceptable use. Instead, it supports reflective governance by surfacing patterns that warrant human interpretation. This separation between observation and response preserves human agency and avoids conflating awareness with enforcement - an especially important consideration in creative, educational, and knowledge-work contexts where over-intervention can itself undermine well-being.

By treating creativity and agency as emergent properties of socio-technical systems instead of static attributes of tools, post-deployment observability offers a practical foundation for guiding human–AI co-evolution toward outcomes that remain aligned with human values, institutional responsibility, and long-term well-being.

For example, in safety-critical domains such as transportation or public health guidance, observability infrastructures could reveal patterns of over-reliance on AI outputs or erosion of human oversight, enabling institutions to recalibrate system design in ways that preserve human agency while supporting productive human–AI co-adaptation.

7 Conclusion

Most conversations about AI and human creativity look at what an AI can do the day it's released. People then argue about whether it will help humans do their work better (augmentation) or take their place entirely (replacement). In other words, the debate is usually: "Is this tool going to help people create more, or is it going to make people unnecessary?" However, this way of thinking is too static, and it ignores how AI systems change over time once people start using them in the real world.

The real impact of AI on people doesn't show up when the tool is released - it shows up over time, as people start using it, relying on it, and changing how they work around it. What matters isn't what AI can do on day one, but how it slowly changes human behavior once it's in everyday use.

This paper examined the challenge of achieving post-deployment awareness for generative AI systems operating in open, real-world environments. While pre-deployment alignment, red-teaming, and guardrails remain essential components of responsible AI development, they provide limited visibility into behaviors that emerge only after deployment, as systems interact repeatedly with users, applications, and institutional workflows over time.

We argued that many consequential risks associated with deployed generative AI systems are not properties of individual prompt–response interactions, but aggregate phenomena that arise through repetition, reuse, and cross-context propagation. Because these behaviors are external to the model boundary and often innocuous in isolation, they are difficult to detect through model-centric evaluation alone. Addressing this observability gap requires a complementary perspective focused on system-level behavior rather than isolated outputs.

To that end, we introduced a post-deployment observability framework grounded in interpretable behavioral indicators and human-in-the-loop oversight. The proposed indicators – including elements like abnormal output velocity, semantic and structural reuse, persistence of synthetic roles, and cross-context propagation - are designed to support early situational awareness when observed collectively and through time. We further outlined a governance-aware monitoring architecture that emphasizes aggregation, documentation, and escalation while avoiding invasive surveillance or automated enforcement.

This work is intended to complement, not replace, existing safety and alignment practices. By focusing on post-deployment behavior under realistic institutional constraints, the framework provides a practical foundation for responsible oversight of generative AI systems operating at scale. More broadly, it highlights the importance of distinguishing between pre-deployment assurances and post-deployment awareness when evaluating the risks of increasingly integrated and adaptive AI technologies.

Several limitations and open questions remain. The indicators described here are intentionally high-level and require further empirical study to understand their sensitivity, robustness, and context dependence across deployment settings. Future work may explore how such observability frameworks can be integrated with organizational processes, regulatory reporting requirements, and cross-institutional information sharing, as well as how they interact with evolving standards for responsible AI governance.

As generative AI systems continue to move from controlled settings into complex socio-technical environments, the ability to observe and interpret their behavior after deployment will become increasingly important. Post-deployment observability offers a path toward maintaining institutional awareness and accountability in the face of emergent risks, complementing existing safety mechanisms while preserving human judgment and governance oversight.

Viewed through the lens of human–AI co-evolution, the central contribution of this work is not the identification of specific failure modes, but the articulation of a means for recognizing gradual shifts in agency, creativity, and institutional dependence as they unfold. In many settings, the risk is not abrupt replacement of human judgment, but its incremental erosion through unobserved automation and reuse.

Post-deployment observability helps surface these dynamics early, while alternative trajectories remain possible. By supporting awareness without prescribing outcomes, the framework enables human choice while not constraining it - an essential condition for sustaining creativity and well-being in environments increasingly shaped by generative AI systems.

This mode of post-deployment observability offers a path toward more responsible and adaptive human–AI systems, providing a foundation for future research on well-being-aligned human–AI co-evolution.

Ethical Statement

This work examines post-deployment observability as a means of supporting responsible governance and well-being in human–AI systems. The proposed framework emphasizes aggregate, interpretable behavioral signals and human-in-the-loop interpretation instead of automated judgment or enforcement. Its intended benefit is to improve situational awareness of emerging dynamics - such as automation pressure or unintended displacement of human agency - while preserving institutional choice and human creativity.

As with any monitoring approach, there is a risk that observability mechanisms could be misapplied for intrusive surveillance or inappropriate control if deployed without clear governance boundaries. This work explicitly avoids such uses by limiting scope to system-level patterns, avoiding individual attribution, and separating observation from response. Ethical deployment therefore depends not only on technical design, but on transparent institutional practices and sustained human oversight.

Acknowledgements

The authors thank Shriniwas Nayak for helpful discussions and written input on adaptive AI personas and their implications for post-deployment observability.

References

Amodei, D., et al. 2016. Concrete Problems in AI Safety. arXiv:1606.06565

Barbaro, F., and Skumanich, A. 2023. Addressing Socially Destructive Disinformation on the Web with Advanced AI Tools. Proc. WWW Companion, 2023. https://doi.org/10.1145/3543873.3587348

Kim, H., and Skumanich, A. 2025. A Data-Driven Mode for Public Health Interventions: A Case Study of the 2025 Measles Outbreak as Reflected in Social Media, Proc. IEEE Global Humanitarian Technology Conference, https://DOI.org/10.1109/GHTC66843.2025.11266608

NIST 2023. AI Risk Management Framework (AI RMF 1.0).

ProPublica, 2026: https://www.propublica.org/article/trump-artificial-intelligence-google-gemini-transportation-regulations
"These developments have alarmed some at DOT. The agency's rules touch virtually every facet of transportation safety, including regulations that keep airplanes in the sky, prevent gas pipelines from exploding and stop freight trains carrying toxic chemicals from skidding off the rails. Why, some staffers wondered, would the federal government outsource the writing of such critical standards to a nascent technology…"

Raji, I. D., et al. 2020. Closing the AI Accountability Gap. Proc. ACM FAccT https://doi.org/10.1145/3351095.3372873

Selbst, A. D., et al. 2019. Fairness and Abstraction in Sociotechnical Systems. Proc. ACM FAT*
https://doi.org/10.1145/3287560.3287598

Skumanich, A., Kim, H K. 2024, Modes of Tracking MalInfo in Social Media with AI/ML Tools to Help Mitigate Harmful GenAI for Improved Societal Well Being, Proceedings of the AAAI Symposium Series
https://DOI.org/10.1609/aaaiss.v3i1.31247

Skumanich, A., et al. 2025. AI Monitoring of Social Media to Assist in Crisis Preparedness and Response. Proc. IEEE Global Humanitarian Technology Conference,
https://DOI.org/10.1109/GHTC66843.2025.11266381

Weidinger, L., et al. , 2022. Taxonomy of Risks Posed by Language Models. Proc. ACM FAccT
https://doi.org/10.1145/3531146.3533088