Review

From Human Oversight to Human-in-the-Loop: Evolving Governance of Human–AI Interaction in Healthcare and AI Development

Kamal Harishchandra Sharma *, Praneel Sharma, Pratyusha Sharma

Posted Date: 28 February 2026

doi: 10.20944/preprints202602.1949.v1

Keywords: human–AI interaction; human-in-the-loop; healthcare AI; MLOps; AI governance; big data; cognitive load; foundation models; software engineering

 doi:10.20944/preprints202602.1949.v1

Review

From Human Oversight to Human-in-the-Loop: Evolving Governance of Human–AI Interaction in Healthcare and AI Development

Kamal Harishchandra Sharma [1,*], Praneel Sharma [2] and Pratyusha Sharma [3]

[1] Department of Cardiology, SAL Hospital, Ahmedabad, Gujarat, India
[2] Department of Information and communication Technology, Dhirubhai Ambani University (DAIICT), Gandhinagar 382007, Gujarat, India
[3] Department of Computer Science & Engineering, Ahmedabad 380009, Gujarat, India
* Correspondence: kamalcardiodoc@gmail.com

Abstract

Background: Artificial intelligence (AI) has moved from research labs into the everyday work of clinicians and software engineers. In healthcare, AI systems now shape triage, imaging interpretation, risk prediction and documentation workflows, while in computer science and AI development, models are increasingly used to generate code, tests and even other AI systems [1–4,7,10,11,76–82]. Early governance frameworks framed "human oversight" as a high-level ethical injunction but provided limited operational guidance on how humans should interact with data-intensive, adaptive AI systems in these settings [5,36–38,76–78]. **Objective:** To examine how human oversight and human-in-the-loop (HITL) paradigms have developed specifically in (i) healthcare and (ii) computer science/AI development, and to identify converging design, organizational and governance patterns that support meaningful human control in big-data AI environments. **Methods:** Narrative synthesis of international and national governance instruments (EU AI Act, WHO, OECD, ICMR, FDA/GMLP), sector-specific guidance for healthcare AI (FUTURE-AI, CHAI, Joint Commission) and emerging frameworks for AI lifecycle governance and MLOps in software engineering [7,10,11,14,15,18,19,79–84,86–89]. We extracted definitions and framings of human oversight, technical and organizational requirements for HITL interaction, and implementation challenges related to cognitive load, big-data pipelines and adaptive models in healthcare and AI development. **Results:** Across healthcare and AI development, contemporary frameworks converge on multi-layered human-in-the-loop governance embedded throughout design, deployment, monitoring and decommissioning. Mandatory and consensus instruments emphasize override capability, transparency, user training, escalation pathways and post-deployment monitoring [7,10,11,14,15,18,19,79,80,86–89]. In both domains, oversight is shifting from ad hoc individual review to structured arrangements involving multidisciplinary committees, model-governance boards, MLOps processes and incident-learning systems. Persistent gaps include limited formal treatment of cognitive load and alert fatigue, difficulties overseeing continuously learning and foundation-model-based systems, and immature metrics for the effectiveness of human oversight itself [61–63,88,114,123,133–137]. **Conclusions:** Healthcare and computer science/AI development are emerging as mutually informative testbeds for human-in-the-loop AI governance in big-data settings. Meaningful oversight requires more than nominal human review: it depends on human-centered interface design, realistic workload management, lifecycle-oriented technical controls and organizational cultures that make it safe to question AI outputs. Lessons from clinical safety science and MLOps can be combined to architect human–AI interaction that amplifies, rather than erodes, professional judgment in both domains.

Keywords: human–AI interaction; human-in-the-loop; healthcare AI; MLOps; AI governance; big data; cognitive load; foundation models; software engineering

 doi:10.20944/preprints202602.1949.v1

1. Introduction

The rapid deployment of AI in the field of clinical practice and software engineering has created an urgent governance challenge: ensuring humans remain in meaningful control of consequential decisions in environments shaped by big data, complex models and adaptive systems [1–4,70,76–78]. In hospitals, AI is already being integrated into informing triage, assist in medical imaging, risk prediction and discharge planning; as well as in development organizations, as AI-assisted tools and model-centric pipelines increasingly drive code generation, testing, deployment and monitoring of the same [7,10,11,79–85].

Initial AI ethics frameworks emphasized on the need of keeping humans responsible and accountable, and that AI should augment rather than replace human judgment [5,6,36–38]. Yet these high-level commitments left critical questions unresolved for healthcare and AI development such as "What information do clinicians and developers need to exercise oversight under real-world constraints?" "At which points in clinical and development workflows should they intervene?" "Which technical and institutional structures are required to support oversight when data volumes, model complexity and time pressure are high [36,70,76–78,111–114]?"

Over the past five years, governance instruments and implementation frameworks have begun to answer these questions more precisely. The EU AI Act defines human oversight as a binding design-level requirement for high-risk systems, including medical devices and systems relying on complex AI pipelines [7,29,30,64,65]. WHO's guidance on AI for health emphasizes human autonomy, transparency and institutional accountability, while insisting that professional oversight cannot be meaningful without adequate training and infrastructure [10,11,34,42,66]. In parallel, FUTURE-AI, the Coalition for Health AI (CHAI) Blueprint, and Joint Commission guidance provide detailed templates for healthcare AI governance [18,19,45–48]. In computer science and AI development, machine-learning operations (MLOps) frameworks, model-governance practices and responsible-AI initiatives are re-framing oversight as a property of the entire development pipeline rather than of isolated models [76–80,83–89,111–117].

A conceptual shift is visible across these initiatives: from nominal "human oversight" toward **operational human-in-the-loop** paradigms in which oversight is distributed across design, deployment, monitoring and incident response. This shift is particularly salient in healthcare and AI development, where AI systems are tightly coupled to big-data infrastructures and high-stakes decisions.

Key Definitions-

Human oversight- Technical, organizational and legal arrangements ensuring clinicians, institutions and regulators retain meaningful control over AI-supported decisions, including ability to disregard, override or reverse outputs and interrupt operation to prevent harm [5,8].

Human-in-the-loop (HITL)- Operational paradigm where humans actively participate at multiple AI lifecycle points (design, validation, deployment, monitoring) ensuring clinically meaningful decisions remain under human control [14,15].

Meaningful human control- Humans possess requisite knowledge, information, time, authority and system design enabling substantive (not merely nominal) intervention in real clinical conditions [17,18].

AI governance - Policies, structures, processes and accountability mechanisms for selecting, validating, deploying, monitoring and withdrawing AI systems with explicit attention to safety, equity and responsibility [10,15].

1.1. Research Questions

This review addresses the following:

1. How has human oversight evolved in the context of healthcare AI and computer science/AI development till 2026?

 doi:10.20944/preprints202602.1949.v1

2. Are these domains converging on structured, operational human-in-the-loop paradigms? If so, what are their core technical and organizational components?
3. What gaps remain in translating human-in-the-loop principle (HITL) into practice in data-intensive, adaptive AI environments in these two domains?

1.2. Scope and Contribution

We focus on two domains where human–AI interaction is both technically advanced and operationally mature:

- **Healthcare:** Clinical decision support, diagnosis, imaging, risk prediction, treatment planning, documentation and workflow automation in hospital and ambulatory settings.
- **Computer science and AI development:** AI system design, MLOps , software engineering workflows and AI-assisted development, including use of foundation models for code and model generation [18,19,70,76–85,88,111–117].

By examining how these domains design, govern and experience human-in-the-loop interaction, we characterize a shared paradigm of human-centered oversight for big-data AI systems and highlight concrete practices that can be translated between clinical and engineering contexts.

2. Methods

2.1. Design and Scope

We conducted a narrative synthesis of AI governance instruments, sector-specific guidelines and implementation literature, focusing on the period of 2018–2026. We prioritized consensus instruments and regulatory documents (EU AI Act, WHO, OECD, ICMR, FDA/GMLP), health-sector frameworks (FUTURE-AI, CHAI, Joint Commission), and computer-science/AI-development frameworks (MLOps governance, model-governance reports, software-engineering case studies) [7,10–15,18,19,41,43,45–48,76–83,86–89,111–117].

2.2. Primary Sources (Regulatory and Policy Documents)

We included cross-cutting regulatory and policy frameworks that directly affect healthcare AI and AI development pipelines:

- EU AI Act (Regulation (EU) 2024/1689) and associated guidance [7–9,29,30,64,65,118].
- OECD AI Principles and accountability reports [12,13,39,40].
- UNESCO Recommendation on the Ethics of AI, where relevant to healthcare and AI governance [21,31].
- UN human-rights guidance on AI and privacy, as it informs transparency and accountability requirements [32].
- Singapore Model AI Governance Framework and related documents on AI governance for high-impact systems [17,33,81].

For healthcare specifically:

- WHO reports on the ethics and governance of AI for health [10,11,34,42,66].
- ICMR ethical guidelines for AI in biomedical research and healthcare (India) [14,41].
- FDA guidance on good machine-learning practice (GMLP) and real-world performance for AI/ML Software as a Medical Device [15,43,44].
- FUTURE-AI international consensus guidelines and related medical-imaging guidance [18,45].
- CHAI Blueprint and Joint Commission guidance on safe and trustworthy AI in health systems [19,46–48].

For AI development and software-engineering governance we specifically but not limited to the following guidelines were analyzed:

- Regulatory and industry reports on MLOps, model-risk management and AI lifecycle governance from financial and technical regulators and cloud providers [79,80,86–89,117].

 doi:10.20944/preprints202602.1949.v1

- Technical standards and professional codes relevant to AI and software engineering (ISO/IEC SC 42, ACM Code of Ethics, IEEE standards) [5,36,76,79,83,84,111,118,119].

2.3. Secondary Sources (Conceptual and Empirical Literature)-

It comprised systematic reviews, empirical human-factors studies, and legal/ethical analyses of oversight mechanisms. Inclusion prioritized authoritative documents with explicit human oversight content over exhaustive systematic search. We included:

- Systematic and narrative reviews on human oversight, meaningful human control, AI governance and AI in healthcare and software engineering [26–28,36–38,70,76–78,111–117].
- Empirical studies on human factors, cognitive load, alert fatigue, human–AI interaction and AI-assisted development [61–63,90–94,111–114,123–125,133].
- Case studies on deployment of clinical AI and ML systems and on MLOps and AI-assisted software engineering [18,19,45–48,79,80,88,111–117].

2.4. Extraction and Synthesis Strategy

From each source we extracted:

- Definitions and framings of human oversight or human-in-the-loop interaction.
- Technical and organizational requirements related to oversight (override capabilities, transparency, training, governance structures, monitoring and incident response).
- Descriptions of data types, architectures, pipelines and monitoring mechanisms relevant to oversight in healthcare and AI development.
- Conceptual developments regarding meaningful human control, scalable oversight and AI safety.
- Implementation challenges in big-data, high-velocity environments (cognitive load, alert fatigue, adaptive models, recursive AI systems).
- Temporal evolution was structured into three phases (2016–2020: ethics; 2021–2023: regulation; 2023–2025: operationalization) to identify paradigm convergence.
- Quality Appraisal Sources were appraised by issuing authority (regulatory > consensus > academic), citation impact, and clinical applicability. No formal quality scoring was applied given policy document heterogeneity.

We used a thematic matrix to map developments over time and across the two domains, identifying convergences, divergences, emerging patterns and persistent gaps.

3. Results

3.1. Evolution of Human Oversight: From Ethics to Regulation (2016–2020)

3.1.1. Foundational Principles and Soft Law

Early international AI ethics frameworks—IEEE *Ethically Aligned Design* (2016), national AI strategies in Canada, France, China, the UK and India (2017–2018), and the EU High-Level Expert Group on AI guidelines (2019)—framed human oversight as an ethical ideal rather than a binding regulatory requirement [5,36–38]. They asserted that algorithms should not make consequential decisions without human review and accountability, that people have a right to know when they interact with AI, and that responsible parties must be identifiable [5,36–38].

However, these frameworks did not specify what information and to what extent clinicians or developers needed to exercise oversight, at what points in workflows oversight should occur, or which technical and organizational structures would support oversight under real-world constraints [36–38,70,76–78]. In healthcare, human oversight was initially construed as clinician review and the theoretical ability to override decision-support recommendations, with existing professional accountability frameworks presumed sufficient [36,70]. In AI development, oversight largely meant peer review of code, design discussions and testing, with limited attention to how ML models and

 doi:10.20944/preprints202602.1949.v1

big-data pipelines might change cognitive demands or shift responsibility for automated actions [76–78,111–114]. This "oversight v.1.0" paradigm assumed that inserting AI into existing workflows would not materially alter the conditions under which humans make decisions.

3.2. Risk-Based Regulation and Cross-Cutting Guidance (2021–2023)

3.2.1. EU AI Act: Binding Human-Oversight Requirements

The EU AI Act (Regulation (EU) 2024/1689) represents a watershed in AI regulation [7,29,30,64,65]. For the first time, a major jurisdiction defines human oversight as a binding design-level requirement for high-risk AI systems, including medical devices and systems that depend on complex AI development pipelines [7,29,30]. Article 14 requires that high-risk systems be designed so natural persons can disregard, override or reverse outputs and can intervene or interrupt operation to prevent harm, while Article 13 obliges providers to furnish information enabling users to understand system outputs and limitations [7,9,30,65].

For healthcare, this positions clinical AI tools—such as diagnostic support, triage and imaging models—as high-risk systems subject to explicit human-oversight requirements, including documentation, training and post-market monitoring [7,10,11,14,15,43,44]. For AI development, the Act links oversight obligations to **providers and deployers**, making the design of MLOps pipelines, model documentation, monitoring mechanisms and change-control processes central to whether clinicians and other end-users can exercise meaningful control [7,29,30,64,65,118].

3.2.2. OECD AI Principles and Accountability

The OECD AI Principles and subsequent accountability reports provide cross-cutting guidance that applies to both healthcare organizations and AI-developing firms [12,13,39,40]. They call for robust, safe and fair AI systems, meaningful human oversight, transparency and mechanisms for incident response, risk management and redressal mechanisms [12,13,39]. OECD's accountability work emphasizes that organizations should record and investigate AI-related incidents, adapt systems and governance based on these learnings, and build human capacity to understand and oversee AI [13,40]. These ideas are reflected in clinical AI governance (for example, AI-related incident reporting and review) and in internal algorithm-auditing frameworks for AI development [18,19,46–48,79,80,115–117].

3.2.3. WHO Health-Sector Guidance

WHO's reports on ethics and governance of AI for health articulate sector-specific expectations for human oversight [10,11,34,42,66]. They stress that AI should support, not replace, professional judgment; that clinicians and patients must know when AI is used and what its limitations are; and that health systems must invest in training, infrastructure and governance to make oversight feasible [10,11,34,42]. WHO notes that deploying AI into already over-burdened clinical environments without adjusting workloads or infrastructure will not yield meaningful oversight, a point supported by empirical work on alert fatigue and cognitive load in clinical decision support [61–63,70].

3.2.4. National and Sector-Aligned Guidance for Healthcare and AI Development

Several national and sector-aligned frameworks further operationalize oversight:

- ICMR guidelines mandate human decision-making, lifecycle governance and accountability for AI used in biomedical research and healthcare in India [14,41].
- FDA GMLP and related guidance require robust model-development practices, predetermined change-control plans and real-world performance monitoring for AI/ML medical devices [15,43,44].
- FUTURE-AI, CHAI Blueprint and Joint Commission guidance translate high-level principles into concrete governance structures in health systems, including AI oversight committees, readiness assessments, and AI-specific incident-reporting and monitoring processes [18,19,45–48].

 doi:10.20944/preprints202602.1949.v1

- MLOps and model-governance frameworks from regulators and industry define lifecycle controls on data, models and deployments that are directly relevant to oversight in healthcare and AI development [79,80,86–89,117].

These instruments collectively push healthcare organizations and AI-developing firms toward lifecycle-oriented, multi-stage oversight architectures.

3.3. Operational Paradigms in Healthcare and AI Development (2023–2025)

3.3.1. Healthcare: FUTURE-AI, CHAI Blueprint and Clinical AI Pipelines

In healthcare, human-in-the-loop governance has evolved from generic invocations of clinician oversight into concrete requirements that span data ingestion, model design, deployment architectures and monitoring tools [18,19,45–48]. FUTURE-AI and related consensus guidelines articulate properties such as fairness, universality, traceability, usability, robustness and explainability, and these are increasingly interpreted as engineering constraints on clinical AI pipelines rather than as purely ethical desiderata [18,45,70].

Modern clinical AI systems frequently sit atop heterogeneous, high-volume data infrastructures. At the **data layer**, structured EHR fields, laboratory results, medication orders, free-text notes, waveform streams (for example, ECG, blood pressure), imaging data (radiology, pathology) and, in some centers, omics and wearable-sensor data are ingested into longitudinal patient records, often via ETL (Extract, Transform and Load) processes into clinical data warehouses [10,11,62,70]. At the **model layer**, health systems deploy a mix of models: gradient-boosted trees and logistic regression for risk scores, convolutional and transformer-based networks for imaging and time-series, and large language models fine-tuned on clinical corpora for summarization, coding and question-answering [18,45,70,72,81]. At the **application layer**, these models drive risk predictions, prioritization lists, automated documentation suggestions and alerts, usually exposed through EHR-integrated user interfaces and dashboards [18,19,46–48].

Human-in-the-loop oversight is now commonly implemented as a set of governed touch-points across this pipeline. Before deployment, multidisciplinary algorithm governance committees or AI oversight boards review data provenance, cohort definitions, labeling quality, performance metrics (including calibration and subgroup analysis) and known failure models and modes [18,19,41,46–48]. They also review integration plans: which clinical roles will see which outputs, at what stages in the workflow, with what default options and escalation routes [18,19,46]. During deployment, models are wrapped in interfaces that present not only a score but also confidence estimates, short rationales or feature contributions, time-series plots, and direct links to the underlying clinical data, enabling clinicians to interrogate model outputs under time constraints [18,19,62,70].

Post-deployment, many health systems are building **continuous-monitoring and drift-detection infrastructures** for clinical AI. These include data-quality checks on incoming EHR streams, real-time or near-real-time dashboards tracking alert volumes, positive predictive value and calibration by unit or patient subgroup, and processes that trigger review when drift thresholds are exceeded [43,44,88,114]. Some organizations have established AI-specific incident-reporting channels—analogous to pharmacovigilance or device-incident systems—allowing clinicians to report cases where AI output contributed to near-misses or adverse events [19,47,48]. These reports are then reviewed by governance committees, which may adjust thresholds, restrict indications, retrain or temporarily suspend models.

A central technical insight is that meaningful clinician oversight must be aligned with the **temporal and cognitive profile** of different clinical environments [61–63,70]. In high-acuity wards and emergency departments, oversight must rely on calibrated, well-prioritized alerts and succinct, visual explanations that can be processed in seconds, not on lengthy textual justifications [61,70]. In lower-acuity or retrospective contexts (for example, chronic-disease clinics, tumor boards, morbidity-and-mortality meetings), clinicians can engage in more reflective oversight, comparing AI suggestions against outcomes and explicitly discussing the system's role in both successes and failures [18,46–48].

 doi:10.20944/preprints202602.1949.v1

Overall, contemporary healthcare practice is converging on a pattern in which heterogeneous clinical data feed into version-controlled models; models are integrated into workflow-aware user interfaces; and their behavior is tracked via continuous monitoring and structured incident-learning processes. In this architecture, human-in-the-loop governance is realized not by a single "override" button but by a combination of design, data and organizational mechanisms that keep clinicians both **in** and **over** the loop of data-intensive AI services [18,19,43,46–48,88,114].

3.3.2. Computer Science and AI Development: Human-in-the-Loop Governance of AI Lifecycles

The computer-science and AI-development domain now use AI not only as an application component but as infrastructure for building other AI systems, creating **recursive oversight** challenges [76–83,85,88]. Human-in-the-loop governance must therefore operate within multi-stage ML and foundation-model pipelines that handle large-scale code, telemetry and user-interaction data.

A typical AI development environment comprises: (i) **data lakes** aggregating logs, code repositories, telemetry, user feedback and external datasets; (ii) **feature stores** providing standardized, versioned feature definitions; (iii) **training pipelines** that build models ranging from classical ML to deep networks and large language models; (iv) **model registries** storing model versions with associated metadata; and (v) **deployment and monitoring platforms** that expose models via APIs or embed them in applications [79,80,85,88,111,117]. Human oversight is being formalized as structured intervention points across these layers.

At the **data stage**, organizations increasingly treat data selection and labeling as governed processes. Datasheets and data cards document sources, collection conditions, consent constraints and known biases; annotation workflows include expert adjudication and periodic auditing of label quality; and changes to high-impact labels require human approval [97,98,111,112]. At the **model stage**, model cards and system cards describe architecture, objectives, training data, evaluation metrics and known limitations, enabling model governance boards or responsible-AI teams to judge suitability for specific uses [97,100,101,115–118].

At the **deployment and monitoring stages**, big-data characteristics are most prominent. High-traffic systems stream telemetry on inputs, model outputs, latency, error rates, user interactions and downstream business metrics into monitoring dashboards [88,114,117]. MLOps platforms support automated alerts for distribution shift, performance degradation and anomalous patterns, and they may offer automated retraining or rollback options [79,80,86–89]. Emerging governance practice is to place **human approval gates** on key transitions: for example, promotion of a model from shadow to production, authorization of an auto-remediation script, or expansion of a model's scope to new product areas [86–89,115–117]. In this sense, developers and model-governance boards operate as human controllers of an evolving, high-velocity AI lifecycle rather than of a static artifact.

AI-assisted software-engineering tools add another layer of human–AI interaction. Code-generation assistants, test-case generators and refactoring agents draw on foundation models trained on massive code and text corpora and are tightly integrated into IDEs and CI/CD pipelines [90–94,123,124]. Developers interact with these tools at very fine granularity, accepting, editing or rejecting suggestions line by line. Empirical work shows both productivity gains and novel risks, including over-reliance on plausible but incorrect suggestions, introduction of subtle security vulnerabilities and decreased familiarity with critical code paths [90–94,123–125]. Organizations are responding with policies that restrict assistant use in safety-critical components, require that AI-generated changes pass through the same peer review and testing gates as human-written code, and instrument acceptance rates and defect rates as metrics of oversight quality [83,90,91,117,123,133].

To coordinate these technical mechanisms, many organizations have created **formal governance structures**. Responsible-AI teams, internal review boards and model governance committees define risk tiers, documentation and testing requirements, and escalation paths for high-impact systems [83,84,115,116]. They also decide which categories of automation (for example, auto-scaling, self-healing agents, auto-configuration) are permitted in which contexts, thereby acting as human-in-the-

 doi:10.20944/preprints202602.1949.v1

loop controllers at the meta-level of the AI development pipeline [86–89,115–117]. Research on scalable oversight—such as AI-assisted evaluation, debate, reward modelling and adversarial red-teaming—is beginning to inform how humans can supervise systems whose internal complexity or domain expertise exceeds that of any single engineer [99–105,120–122].

Viewed through a big-data and cognitive-computing lens, AI development illustrates what it means for the "loop" itself to be a complex, data-intensive process. Data versioning, model registries, telemetry-rich monitoring, CI/CD and incident-response frameworks in engineering play a role analogous to EHR integration, clinical dashboards and safety review structures in healthcare. In both domains, meaningful human-in-the-loop governance depends less on the theoretical existence of an override pathway and more on whether these surrounding infrastructures give humans timely visibility, practical control levers and institutional support to intervene when it matters [79,80,83–89,111–117,123,133].

4. Discussion

4.1. Convergence on Human-in-the-Loop Paradigms in Healthcare and AI Development

When clinicians and engineers talk about "keeping a human in the loop," they often mean different things, yet the trajectories in healthcare and AI development are increasingly similar. In hospitals, AI has moved from isolated pilot tools to systems that influence decisions throughout the patient journey—from triage to imaging, from sepsis alerts to discharge planning and population-health risk scores. In software organizations, AI has moved from research notebooks to the core of development pipelines, proposing code, generating tests, prioritizing bugs and even orchestrating deployment steps. In both worlds, humans now work alongside systems that can act faster than any individual can reliably monitor and that draw on data volumes far beyond manual inspection.

In healthcare, this evolution is visible in the shift from rule-based decision support to data-driven, continuously monitored models integrated into EHRs and clinical dashboards. Early systems could be ignored with little consequence; contemporary models contribute to triage decisions, imaging workflows and early-warning systems in ways that must be understood and overseen at the level of hospitals, not individual clinicians alone [18,19,45–48,61–63,70]. In AI development, a similar shift has occurred from discrete code reviews and test suites towards densely instrumented ML lifecycles, where models and AI-assisted tools continually propose changes and optimizations that must be governed through MLOps and model-governance processes [79,80,83–89,111–117].

Despite distinct cultures and stakeholders, the two domains converge on a pattern in which oversight becomes:

- Distributed across the lifecycle, from data selection and model design through deployment and monitoring.
- Shared across individuals, teams and institutions, rather than resting solely on single clinicians or developers.
- Dependent on sociotechnical infrastructures—interfaces, workflows, metrics and incentives—that give humans the information, time and authority needed to intervene effectively.

4.2. Deepening Challenges: Cognitive Load, Big-Data Pipelines and Adaptive Systems

The movement toward human-in-the-loop governance does not remove the fundamental difficulties of supervising powerful AI systems; in some respects, it sharpens them. Two challenges recur across both domains: cognitive burden in high-velocity environments and oversight of adaptive, continuously learning systems.

4.2.1. Cognitive Burden in High-Velocity Environments

In clinical environments, clinicians frequently experience alert fatigue. A sepsis model might flag dozens of patients in a day, with each alert arriving in a stream of lab notifications, medication warnings and messages. Although the clinician is nominally "in the loop," the reality of time pressure

 doi:10.20944/preprints202602.1949.v1

and fatigue constrains their ability to exercise nuanced judgment. The more often AI flags alerts, the more likely genuine signals will be missed [61–63,70].

AI-assisted software engineering shows a parallel phenomenon. Developers using code-generation tools often accept suggestions because rejecting or editing each one is itself cognitively taxing. An environment in which an AI continuously proposes code completions, refactoring and tests can fragment attention and turn oversight into a continuous stream of micro-decisions. Under delivery pressure, developers may rubber-stamp plausible-looking AI output, allowing subtle errors and vulnerabilities to accumulate [90–94,123–125,133].

These examples illustrate that oversight consumes scarce cognitive resources and is most strained in the environments that most attract AI deployment—busy wards, high-volume clinics, and development teams under tight deadlines. A realistic human-in-the-loop paradigm must treat cognitive load as a design constraint. In healthcare, this implies calibrated alerting, tiered notification schemes and escalation pathways that reserve clinician attention for ambiguous or high-stakes cases [61–63,70]. In AI development, it suggests batching AI-generated changes for structured review, limiting unconstrained code suggestions in safety-critical modules and instrumenting acceptance/defect rates as measures of oversight quality [83,90–94,117,123,133].

4.2.2. Oversight of Adaptive and Continuously Learning Systems

A second shared challenge is the rise of adaptive and foundation-model-based systems. Clinical AI models increasingly update as local data change, while development tools rely on foundation models that are periodically retrained on large code and text corpora [72,81,82,135–137]. Behavior can thus drift even when interfaces remain unchanged.

In a hospital, a readmission-risk model may be retrained on newer EHR data as case mix, clinical guidelines and documentation practices evolve. Even if the model's name and interface are unchanged, its internal behavior may differ substantially. The governance question is who is responsible for confirming that the updated model remains safe and equitable, and how this can be assured without repeating full pre-deployment evaluations [43,44,88,114].

In AI development, model providers may upgrade foundation models used in code-assistants, changing coding styles, library choices or edge-case behaviour. Developers may notice only when new failure modes appear in production. Without explicit oversight mechanisms—such as benchmark suites, canary projects or red-teaming focused on new versions—these shifts may go undetected [79,80,88,114,135–137].

In both domains, a model of oversight that treats deployment as a static approval event is no longer adequate. Instead, oversight must become a form of continuous quality assurance that treats models as moving targets. For healthcare, this implies periodic re-validation, drift dashboards and protocols for suspending or restricting models when behaviour deviates beyond predefined bounds. For AI development, it implies version-aware governance with clear policies on when model upgrades are permitted, how they are tested and who can roll back or freeze models when unexpected behaviour emerges [43,44,79,80,86–89,114,135–137].

4.3. Practical Implications: Designing Human-Centred Oversight Architectures

Reframing the discussion around healthcare and AI development clarifies that the core question is not whether a human is nominally "in the loop," but how socio-technical systems are engineered so that human judgment is amplified rather than eroded.

For frontline clinicians, oversight must be built into practice in ways that are sustainable. Decision support that merely adds pop-ups to an already overloaded EHR may satisfy a formal oversight requirement while undermining actual control. By contrast, tools that surface risk scores in existing views, explain key drivers in plain language and provide one-click ways to document disagreement can help clinicians integrate AI into their reasoning. Reflective practices—such as case conferences and morbidity-and-mortality meetings that explicitly examine AI's role—can transform

 doi:10.20944/preprints202602.1949.v1

oversight from an abstract regulatory obligation into a lived professional practice [18,19,46–48,61–63,70].

For engineers and data scientists, human-in-the-loop oversight demands rethinking the development lifecycle. It requires treating data selection and labeling as governance decisions, embedding human review into automated pipelines at meaningful points (for example, promotion of models and expansion of scope), and providing tools that make model behaviour transparent across diverse inputs rather than relying on anecdotal tests [79,80,83–89,97–101,111–117].

In both domains, organizational culture is crucial. Junior clinicians may hesitate to override AI tools endorsed by senior leadership; junior engineers may hesitate to question widely adopted assistants or models. Oversight arrangements that appear robust on paper can fail if they ignore these social dynamics. Creating psychologically safe spaces to question AI output—and the governance choices behind AI deployment—is as important as any technical control [46–48,83,84,115–117,130,131].

Viewed together, healthcare and AI development offer mutually illuminating examples. Hospitals can borrow from MLOps and software-reliability practices; engineering teams can borrow from clinical safety culture, near-miss reporting and morbidity-and-mortality reviews. Both domains show that meaningful human-in-the-loop governance depends less on the theoretical ability to override AI, and more on whether the surrounding ecosystem gives humans the information, time, authority and confidence to act when it counts.

5. Conclusions

The language of human oversight in AI governance has shifted from aspirational slogans to more structured, operational paradigms in healthcare and AI development. Regulatory instruments (notably the EU AI Act), health-sector guidance (WHO, ICMR, FUTURE-AI, CHAI, Joint Commission) and MLOps/model-governance frameworks in software engineering now converge on multi-layered human-in-the-loop arrangements embedded throughout AI lifecycles [7,10,11,14,15,18,19,43–48,79,80,86–89,111–117].

Across these domains, common components include override capabilities, transparency and explanation, training and competency assessment, escalation protocols, lifecycle monitoring and distributed accountability. At the same time, persistent gaps remain: limited formal treatment of cognitive load and alert fatigue; challenges overseeing adaptive and foundation-model-based systems; uneven implementation capacity across settings; and underdeveloped metrics for evaluating the effectiveness of human oversight itself [61–63,70,88,114,123,133–137].

Realizing meaningful human-in-the-loop oversight in healthcare and AI development will require:

- Investment in governance infrastructure and monitoring capabilities suited to big-data AI systems.
- Workforce development that equips clinicians, engineers and data scientists with the knowledge, skills and authority to exercise oversight.
- Transparent performance and oversight metrics that track not only AI performance but also how and when humans intervene.
- Learning systems that treat AI-related incidents as opportunities for redesign rather than solely individual error.

Healthcare and AI development are uniquely positioned to learn from one another. Clinical safety science can inform MLOps and model governance, while software-engineering practices around version control, canary releases and incident response can strengthen the technical backbone of healthcare AI governance. Together, these domains can anchor a richer, more precise account of what meaningful human control looks like when AI is woven into the fabric of everyday professional work.

References

 doi:10.20944/preprints202602.1949.v1

[1] Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach.* **2018**, *28*, 689–707.

[2] Jobin, A.; Ienca, M.; Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399.

[3] Morley, J.; Floridi, L.; Kinsey, L.; Elhalal, A. From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Sci. Eng. Ethics* **2020**, *26*, 2141–2168.

[4] Mittelstadt, B. Principles alone cannot guarantee ethical AI. *Nat. Mach. Intell.* **2019**, *1*, 501–507.

[5] IEEE. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2; IEEE: Piscataway, NJ, USA, 2017.

[6] European Commission High-Level Expert Group on AI. Ethics Guidelines for Trustworthy AI; European Commission: Brussels, Belgium, 2019.

[7] European Parliament and Council. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act); Official Journal of the European Union: Brussels, Belgium, 2024.

[8] European Commission. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act); COM(2021) 206 final; European Commission: Brussels, Belgium, 2021.

[9] European Commission. Questions and Answers—Artificial Intelligence Act; European Commission: Brussels, Belgium, 2024. Available online: https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_1683 (accessed on 1 February 2026).

[10] World Health Organization. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance; World Health Organization: Geneva, Switzerland, 2021.

[11] World Health Organization. Regulatory Considerations on Artificial Intelligence for Health; World Health Organization: Geneva, Switzerland, 2023.

[12] OECD. OECD AI Principles; OECD Publishing: Paris, France, 2019.

[13] OECD. Advancing Accountability in AI: Governing and Managing Risks throughout the Lifecycle for Trustworthy AI; OECD Publishing: Paris, France, 2023.

[14] Indian Council of Medical Research. Ethical Guidelines for Application of Artificial Intelligence in Biomedical Research and Healthcare; ICMR: New Delhi, India, 2023.

[15] U.S. Food and Drug Administration. Good Machine Learning Practice for Medical Device Development: Guiding Principles; FDA: Silver Spring, MD, USA, 2021.

[17] Personal Data Protection Commission Singapore. Model Artificial Intelligence Governance Framework, 2nd ed.; PDPC: Singapore, 2020.

[18] Kocaballi, A.B.; Ijaz, K.; Laranjo, L.; Quiroz, J.C.; Rezazadegan, D.; Berkovsky, S.; Coiera, E.; Tong, H.L. The FUTURE-AI Guideline for Fair, Universal, Transparent, Understandable, Robust and Explainable AI in Healthcare. *Nat. Mach. Intell.* **2024**, *6*, 1202–1213.

[19] Coalition for Health AI. Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare; CHAI: Washington, DC, USA, 2023.

[21] UNESCO. Recommendation on the Ethics of Artificial Intelligence; UNESCO: Paris, France, 2021.

[26] Santoni de Sio, F.; Mecacci, G. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philos. Technol.* **2021**, *34*, 1057–1084.

[27] Tigard, D.W. There is no techno-responsibility gap. *Philos. Technol.* **2021**, *34*, 589–607.

[28] Umbrello, S.; van de Poel, I. Mapping value sensitive design onto AI for social good principles. *AI Ethics* **2021**, *1*, 283–296.

[29] European Commission. Regulatory Framework Proposal on Artificial Intelligence; European Commission: Brussels, Belgium, 2021.

[30] European Parliament. Amendments Adopted by the European Parliament on 14 June 2023 on the Proposal for a Regulation of the European Parliament and of the Council on Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act); European Parliament: Strasbourg, France, 2023.

 doi:10.20944/preprints202602.1949.v1

[31] UNESCO. The UNESCO Recommendation on the Ethics of Artificial Intelligence: Key Facts; UNESCO: Paris, France, 2022.
[32] United Nations Office of the High Commissioner for Human Rights. The Right to Privacy in the Digital Age; UN: New York, NY, USA, 2021.
[33] Personal Data Protection Commission Singapore. Model Artificial Intelligence Governance Framework, 2nd ed.; PDPC: Singapore, 2020.
[34] World Health Organization. Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models; World Health Organization: Geneva, Switzerland, 2024.
[36] Shortliffe, E.H. Computer-Based Medical Consultations: MYCIN; Elsevier: New York, NY, USA, 1976.
[37] Bostrom, N.; Yudkowsky, E. The ethics of artificial intelligence. In *The Cambridge Handbook of Artificial Intelligence*; Frankish, K., Ramsey, W.M., Eds.; Cambridge University Press: Cambridge, UK, 2014; pp. 316–334.
[38] Russell, S.; Dewey, D.; Tegmark, M. Research priorities for robust and beneficial artificial intelligence. *AI Mag.* **2015**, *36*, 105–114.
[39] OECD. OECD Framework for the Classification of AI Systems; OECD Publishing: Paris, France, 2022.
[40] OECD. Emerging Accountability Structures to Oversee Global AI; OECD Publishing: Paris, France, 2024.
[41] Garg, S.; Pundir, P.; Rathee, G.; Gupta, P.K.; Garg, S.; Ahlawat, S. On continuous integration/continuous delivery for automated deployment of machine learning models using MLOps. In *Proceedings of the 2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, Laguna Hills, CA, USA, 1–3 December 2021; pp. 25–28.
[42] Char, D.S.; Abràmoff, M.D.; Feudtner, C. Identifying ethical considerations for machine learning healthcare applications. *Am. J. Bioeth.* **2020**, *20*, 7–17.
[43] U.S. Food and Drug Administration. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan; FDA: Silver Spring, MD, USA, 2021.
[44] U.S. Food and Drug Administration. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions; FDA: Silver Spring, MD, USA, 2023.
[45] Vasey, B.; Nagendran, M.; Campbell, B.; Clifton, D.A.; Collins, G.S.; Denaxas, S.; Denniston, A.K.; Faes, L.; Geerts, B.; Ibrahim, M.; et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* **2022**, *28*, 924–933.
[46] Joint Commission. Safe and Trustworthy Artificial Intelligence-Enabled Clinical Decision Support in Healthcare; Joint Commission Resources: Oak Brook, IL, USA, 2024.
[47] Sendak, M.P.; Gao, M.; Brajer, N.; Balu, S. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit. Med.* **2020**, *3*, 41.
[48] Sendak, M.; Elish, M.C.; Gao, M.; Futoma, J.; Ratliff, W.; Nichols, M.; Bedoya, A.; Balu, S.; O'Brien, C. "The human body is a black box": Supporting clinical decision-making with deep learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, 27–30 January 2020; pp. 99–109.
[61] Ancker, J.S.; Edwards, A.; Nosal, S.; Hauser, D.; Mauer, E.; Kaushal, R. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Med. Inform. Decis. Mak.* **2017**, *17*, 36.
[62] Benda, N.C.; Veinot, T.C.; Sieck, C.J.; Ancker, J.S. Broadband internet access is a social determinant of health! *Am. J. Public Health* **2020**, *110*, 1123–1125.
[63] Khairat, S.; Marc, D.; Crosby, W.; Al Sanousi, A. Reasons for physicians not adopting clinical decision support systems: Critical analysis. *JMIR Med. Inform.* **2018**, *6*, e24.
[64] European Commission. Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence: Explanatory Memorandum; European Commission: Brussels, Belgium, 2021.
[65] European Parliament. Legislative Resolution of 13 March 2024 on the Proposal for a Regulation of the European Parliament and of the Council on Laying Down Harmonised Rules on Artificial Intelligence; European Parliament: Strasbourg, France, 2024.
[66] Vollmer, S.; Mateen, B.A.; Bohner, G.; Király, F.J.; Ghani, R.; Jonsson, P.; Cumbers, S.; Jonas, A.; McAllister, K.S.L.; Myles, P.; et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* **2020**, *368*, l6927.

 doi:10.20944/preprints202602.1949.v1

[70] Topol, E.J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56.
[72] Moor, M.; Banerjee, O.; Abad, Z.S.H.; Krumholz, H.M.; Leskovec, J.; Topol, E.J.; Rajpurkar, P. Foundation models for generalist medical artificial intelligence. *Nature* **2023**, *616*, 259–265.
[76] Marcus, G.; Davis, E. Rebooting AI: Building Artificial Intelligence We Can Trust; Pantheon: New York, NY, USA, 2019.
[77] Mitchell, M. Artificial Intelligence: A Guide for Thinking Humans; Farrar, Straus and Giroux: New York, NY, USA, 2019.
[78] Pearl, J.; Mackenzie, D. The Book of Why: The New Science of Cause and Effect; Basic Books: New York, NY, USA, 2018.
[79] Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; Young, M.; Crespo, J.F.; Dennison, D. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates: Red Hook, NY, USA, 2015; pp. 2503–2511.
[80] Paleyes, A.; Urma, R.G.; Lawrence, N.D. Challenges in deploying machine learning: A survey of case studies. *ACM Comput. Surv.* **2022**, *55*, 1–29.
[81] OpenAI. GPT-4 Technical Report. arXiv 2023, arXiv:2303.08774.
[82] Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and efficient foundation language models. arXiv 2023, arXiv:2302.13971.
[83] Breck, E.; Cai, S.; Nielsen, E.; Salib, M.; Sculley, D. The ML test score: A rubric for ML production readiness and technical debt reduction. In *Proceedings of the 2017 IEEE International Conference on Big Data (Big Data)*, Boston, MA, USA, 11–14 December 2017; pp. 1123–1132.
[84] Ashmore, R.; Calinescu, R.; Paterson, C. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Comput. Surv.* **2021**, *54*, 1–39.
[85] Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the opportunities and risks of foundation models. arXiv 2021, arXiv:2108.07258.
[86] Lwakatare, L.E.; Raj, A.; Bosch, J.; Olsson, H.H.; Crnkovic, I. A taxonomy of software engineering challenges for machine learning systems: An empirical investigation. In *Agile Processes in Software Engineering and Extreme Programming*; Springer: Cham, Switzerland, 2019; pp. 227–243.
[87] Sato, D.; Wider, A.; Windheuser, C. Continuous Delivery for Machine Learning; ThoughtWorks: Chicago, IL, USA, 2019.
[88] Shankar, S.; Garcia, R.; Howard, J.; Riberio, D.; Calo, S.; Mousavi, P. Operationalizing machine learning: An interview study. arXiv 2022, arXiv:2209.09125.
[89] Karlaš, B.; Dao, D.; Interlandi, M.; Ding, B.; Ré, C.; Klimovic, A.; Schelter, S. Building continuous integration services for machine learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Virtual Event, 6–10 July 2020; pp. 2407–2415.
[90] Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H.P.D.O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. Evaluating large language models trained on code. arXiv 2021, arXiv:2107.03374.
[91] Barke, S.; James, M.B.; Polikarpova, N. Grounded Copilot: How programmers interact with code-generating models. *Proc. ACM Program. Lang.* **2023**, *7*, 85–111.
[92] Dakhel, A.M.; Majdinasab, V.; Nikanjam, A.; Khomh, F.; Desmarais, M.C.; Jiang, Z.M. GitHub Copilot AI pair programmer: Asset or liability? *J. Syst. Softw.* **2023**, *203*, 111734.
[93] Vaithilingam, P.; Zhang, T.; Glassman, E.L. Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 23–28 April 2022; pp. 1–23.
[94] Peng, S.; Kalliamvakou, E.; Cihon, P.; Demirer, M. The impact of AI on developer productivity: Evidence from GitHub Copilot. arXiv 2023, arXiv:2302.06590.
[97] Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J.W.; Wallach, H.; Daumé III, H.; Crawford, K. Datasheets for datasets. *Commun. ACM* **2021**, *64*, 86–92.

 doi:10.20944/preprints202602.1949.v1

[98] Pushkarna, M.; Zaldivar, A.; Kjartansson, O. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, Seoul, Republic of Korea, 21–24 June 2022; pp. 1776–1826.
[99] Christiano, P.F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates: Red Hook, NY, USA, 2017; pp. 4299–4307.
[100] Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, USA, 29–31 January 2019; pp. 220–229.
[101] Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. Ethical and social risks of harm from language models. arXiv 2021, arXiv:2112.04359.
[102] Irving, G.; Christiano, P.; Amodei, D. AI safety via debate. arXiv 2018, arXiv:1805.00899.
[103] Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; Legg, S. Scalable agent alignment via reward modeling: A research direction. arXiv 2018, arXiv:1811.07871.
[104] Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv 2022, arXiv:2204.05862.
[105] Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. Constitutional AI: Harmlessness from AI feedback. arXiv 2022, arXiv:2212.08073.
[111] Amershi, S.; Begel, A.; Bird, C.; DeLine, R.; Gall, H.; Kamar, E.; Nagappan, N.; Nushi, B.; Zimmermann, T. Software engineering for machine learning: A case study. In *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice*, Montreal, QC, Canada, 25–31 May 2019; pp. 291–300.
[112] Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; Aroyo, L.M. "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 8–13 May 2021; pp. 1–15.
[113] Arpteg, A.; Brinne, B.; Crnkovic-Friis, L.; Bosch, J. Software engineering challenges of deep learning. In *Proceedings of the 2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, Prague, Czech Republic, 29–31 August 2018; pp. 50–59.
[114] Polyzotis, N.; Roy, S.; Whang, S.E.; Zinkevich, M. Data lifecycle challenges in production machine learning: A survey. *ACM SIGMOD Rec.* **2018**, *47*, 17–28.
[115] Bernstein, M.S.; Levi, M.; Magnus, D.; Rajala, B.A.; Satz, D.; Waeiss, C. Responsible AI: Bridging from ethics to practice. *Commun. ACM* **2021**, *64*, 164–166.
[116] Microsoft. Responsible AI Standard, v2; Microsoft Corporation: Redmond, WA, USA, 2022.
[117] Raji, I.D.; Smart, A.; White, R.N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; Barnes, P. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, 27–30 January 2020; pp. 33–44.
[118] ISO/IEC. ISO/IEC 42001:2023 Information Technology—Artificial Intelligence—Management System; ISO: Geneva, Switzerland, 2023.
[119] IEEE. IEEE 7000-2021—IEEE Standard Model Process for Addressing Ethical Concerns During System Design; IEEE: Piscataway, NJ, USA, 2021.
[120] Kenton, Z.; Everitt, T.; Weidinger, L.; Gabriel, I.; Mikulik, V.; Irving, G. Alignment of language agents. arXiv 2021, arXiv:2103.14659.
[121] Saunders, W.; Yeh, C.; Wu, J.; Bills, S.; Ouyang, L.; Ward, J.; Leike, J. Self-critiquing models for assisting human evaluators. arXiv 2022, arXiv:2206.05802.
[122] Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; Irving, G. Red teaming language models with language models. arXiv 2022, arXiv:2202.03286.
[123] Perry, N.; Srivastava, M.; Kumar, D.; Boneh, D. Do users write more insecure code with AI assistants? In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, Copenhagen, Denmark, 26–30 November 2023; pp. 2785–2799.

 doi:10.20944/preprints202602.1949.v1

[124] Nguyen, N.; Nadi, S. An empirical evaluation of GitHub Copilot's code suggestions. In *Proceedings of the 19th International Conference on Mining Software Repositories*, Pittsburgh, PA, USA, 23–24 May 2022; pp. 1–5.
[125] Sandoval, G.; Pearce, H.; Nys, T.; Karri, R.; Garg, S.; Dolan-Gavitt, B. Lost at C: A user study on the security implications of large language model code assistants. In *Proceedings of the 32nd USENIX Security Symposium*, Anaheim, CA, USA, 9–11 August 2023; pp. 2205–2222.
[130] Seeber, I.; Bittner, E.; Briggs, R.O.; De Vreede, T.; De Vreede, G.J.; Elkins, A.; Maier, R.; Merz, A.B.; Oeste-Reiß, S.; Randrup, N.; et al. Machines as teammates: A research agenda on AI in team collaboration. *Inf. Manag.* **2020**, *57*, 103174.
[131] Jarrahi, M.H.; Askay, D.; Eshraghi, A.; Smith, P. Artificial intelligence and knowledge management: A partnership between human and AI. *Bus. Horiz.* **2023**, *66*, 87–99.
[133] Barke, S.; James, M.B.; Polikarpova, N. Grounded Copilot: How programmers interact with code-generating models. *Proc. ACM Program. Lang.* **2023**, *7*, 85–111.
[135] Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv 2023, arXiv:2303.12712.
[136] Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent abilities of large language models. *Trans. Mach. Learn. Res.* **2022**.
[137] Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates: Red Hook, NY, USA, 2023; pp. 46595–46623.