

Human-AI Governance (HAIG): A Trust-Utility Approach

Zeynep Engin

Data for Policy CIC, London, United Kingdom

Department of Computer Science, University College London (UCL), United Kingdom

Correspondence: z.engin@dataforpolicy.org

Abstract

This paper introduces the **HAIG framework** for analysing **trust dynamics** across evolving human-AI relationships. Current categorical frameworks (e.g., "**human-in-the-loop**" models) inadequately capture how AI systems evolve from tools to **partners**, particularly as foundation models demonstrate **emergent capabilities** and multi-agent systems exhibit autonomous goal-setting behaviours. As systems advance, agency redistributes in complex patterns that are better represented as positions along continua rather than discrete categories, though progression may include both gradual shifts and significant step changes. The HAIG framework operates across three levels: *dimensions* (Decision Authority Distribution, Process Autonomy, and Accountability Configuration), *continua* (gradual shifts along each dimension), and *thresholds* (critical points requiring governance adaptation). Unlike risk-based or principle-based approaches, HAIG adopts a *trust-utility* orientation, focusing on maintaining appropriate trust relationships that maximise utility while ensuring sufficient safeguards. Our analysis reveals how technical advances in self-supervision, reasoning authority, and distributed decision-making drive non-uniform trust evolution across both contextual variation and technological advancement. Case studies in healthcare and European regulation demonstrate how HAIG complements existing frameworks while offering a foundation for alternative approaches that anticipate governance challenges before they emerge.

Keywords: Human-AI Governance (HAIG), trust-utility, dimensional governance, trust dynamics, algorithmic governance, foundation models

1. From Tools to Partners: Rethinking Human-AI Relationships

Trust in AI is no longer optional—but what we need is AI we can *justifiably* trust. As algorithmic systems take on greater roles in shaping decisions, allocating resources, and influencing public life, the question is not merely whether we trust them, but whether they are designed and maintained to warrant that trust. What is required is neither blind faith in machines nor paralysing scepticism, but a governance model that can continually calibrate *how much trust is appropriate, and in which contexts*.

Artificial intelligence is fundamentally reshaping the landscape of organisational decision-making, public service delivery, and institutional governance ([Engin et al., 2025](#); [Rudko et al., 2024](#); [Shrestha et al., 2019](#)). Across sectors, it is rapidly evolving from basic automation to increasingly sophisticated forms of governance partnership ([Rahwan et al., 2019](#); [Pedreschi et al., 2025](#); [Liu et al., 2025](#)), aiming for human–AI complementarity where collaboration produces greater value than either alone ([Stephany & Teutloff, 2024](#); [Akata et al., 2020](#); [Donahue et al., 2022](#); [van Breda & Barry, 2020](#)). This evolution is particularly evident in technical advances such as foundation models with emergent capabilities, self-supervised learning systems that critique their own outputs, and multi-agent architectures that exhibit collective behaviours beyond their individual components. These technical developments fundamentally challenge traditional governance frameworks by introducing dynamic capability boundaries and distributed agency patterns that resist static categorisation. However, our conceptual frameworks have not kept pace with this evolution ([Goos & Savona, 2024](#)).

Current approaches to understanding human-AI relationships rely on discrete categories—tools versus agents, decision support versus automated decision-making, or human-in-the-loop versus human-out-of-the-loop models. These categorical frameworks provide implementation clarity but fail to account for the hybrid, evolutionary nature of AI systems in practice ([Engin & Hand, 2025](#); [Coglianese & Crum, 2025](#); [Earp et al., 2025](#); [Gabriel et al., 2024](#)).

This limitation becomes particularly evident as AI systems increasingly permeate domains traditionally dominated by human judgment. In healthcare, diagnostic AI systems gradually shift from flagging anomalies to making preliminary diagnoses ([Yelne et al., 2023](#); [Qiu et al., 2024](#)); in financial services, credit algorithms evolve from augmenting human judgment to making autonomous decisions for standard applications ([Gsenger & Strle, 2021](#)); in transportation, self-driving systems dynamically redistribute authority between human and machine based on environmental conditions ([Ansari et al., 2022](#)). Even in knowledge work requiring complex reasoning—legal practice, academic research, strategic decision-making, and creative production—AI systems now operate as collaborative partners rather than mere tools ([Sowa & Przegalinska, 2025](#); [Koehler & Sauermann, 2024](#)), progressively assuming greater agency. Multi-agent systems embedded in infrastructure further demonstrate this evolution, with interconnected algorithms collectively reshaping environments through distributed decision-making ([Cugurullo & Xu, 2025](#); [Luusua et al., 2023](#); [Engin et al., 2020](#)). Perhaps most consequentially, AI is increasingly embedded in processes central to democratic governance—from content moderation that shapes public discourse, to algorithmic allocation of public resources, to automated enforcement of regulations, to decision support systems that inform policy development ([Engin & Treleaven, 2019](#); [Dunleavy & Margetts, 2023](#)). AI systems in these contexts fundamentally challenge traditional frameworks for

understanding accountability, oversight, and democratic control ([Coglianese & Lehr, 2017](#); [Gritsenko & Wood, 2020](#); [Stokel, 2025](#)) as they incrementally assume greater decision-making authority without corresponding evolutions in oversight frameworks. These transitions rarely occur in discrete jumps between clear categories—instead, agency redistributes along a continuum as capabilities expand and trust develops.

As AI systems evolve, the trust relationships between human actors and algorithmic systems must evolve accordingly. We introduce the concept of "trust dynamics" to capture these changing patterns of confidence, reliance, and accountability. Trust serves as the critical mediating factor between technical capabilities and effective governance—determining whether increasing AI authority enhances or undermines institutional objectives.

This paper introduces the Human-AI Governance (HAIG) framework for analysing trust dynamics across evolving human-AI relationships. HAIG operates across three complementary levels: *dimensions*, *continua*, and *thresholds*. This integrated framework provides both theoretical comprehensiveness and practical flexibility by acknowledging continuous agency redistribution while identifying meaningful transition points requiring governance adaptation. Unlike approaches that are primarily risk-based, principle-based, or technology-based, HAIG is *trust-utility* oriented—focusing on maintaining appropriate trust relationships that maximise utility while ensuring sufficient safeguards.

We demonstrate the framework's practical utility through case studies in healthcare AI and European regulations, illustrating how HAIG both complements existing systems like the EU AI Act and provides a foundation for alternative regulatory approaches. By bridging categorical approaches and multidimensional continua, we provide a forward-looking framework that helps anticipate governance challenges and adapt oversight mechanisms incrementally rather than reactively.

The remainder of this paper unfolds as follows: Section 2 examines current approaches and their limitations; Section 3 introduces the theoretical foundations of HAIG; Section 4 demonstrates practical applications; and Section 5 discusses broader implications and future research directions.

2. Where We Stand: Contemporary Approaches to Human-AI Oversight

AI integration across public and commercial domains demands new approaches to human-AI relationships ([Lai et al., 2023](#); [Rahwan, 2018](#)). Current frameworks struggle to capture these relationships' evolving nature as AI mediates critical functions. This review examines current conceptual frameworks that address how humans and AI systems interact in contexts where authority and responsibility must be distributed between them. By identifying both contributions and limitations of existing approaches, we build toward our proposed Human-AI Governance (HAIG) framework, which offers a more comprehensive understanding of these evolving relationships.

2.1. Governance Philosophies

As AI systems increasingly permeate decision-making contexts, several distinct governance approaches have emerged to address the challenges of human-AI relationships. These approaches

reflect different priorities, disciplinary perspectives, and governance philosophies that shape how we conceptualise, implement, and regulate AI systems across society.

Risk-based approaches exemplify the tension between clarity and adaptability. The [EU's AI Act](#) categorises systems by potential harm (unacceptable to minimal risk), yet struggles with evolving systems and contextual variation ([Kaminski, 2023](#); [Laux et al., 2024](#)). Foundation models particularly challenge this approach with their 'multifunctionality' across deployment contexts ([Coglianese & Crum, 2025](#)). *Rights-based* frameworks centre on protecting fundamental human rights in AI deployment ([Smuha, 2021](#); UN Human Rights Council, [2024](#); Toronto Declaration, [2018](#)). While establishing essential normative foundations, they often lack specific implementation guidance for complex socio-technical systems. *Principle-based* governance, advanced by organisations like the OECD and IEEE, establishes broad normative guidelines focusing on values like fairness, transparency, and accountability. While creating useful ethical frameworks, these approaches often struggle with operationalisation and enforcement in complex technical contexts ([Mittelstadt, 2019](#)). *Trust-centred* approaches position trust dynamics as fundamental to effective human-AI relationships. These frameworks, including work by Winfield and Jirotko ([2018](#)) and Toreini et al. ([2020](#)), address the social and psychological dimensions of how humans develop appropriate confidence in AI systems, though they sometimes lack the regulatory specificity of risk-based approaches.

Within these broader governance philosophies, more specific *categorical frameworks* have emerged to address human-algorithm relationships. The dominant categorical approach has been to distinguish systems based on automation/autonomy levels or human involvement.

"Levels of Automation/Autonomy" (LOA) frameworks span a spectrum from manual to autonomous operation, with automation frameworks focusing on task allocation ([Parasuraman et al., 2000](#)) whilst autonomy frameworks emphasise independent decision-making ([Vagia et al., 2016](#)). Both struggle with contextual variation, positioning systems in static categories rather than recognising their dynamic nature. Alkhatib and Bernstein's ([2019](#)) research demonstrated that real implementations manifest as a dynamic spectrum of agency. Foundation models further challenge these taxonomies—even when humans retain decision authority, they often rely on insights from opaque processes, creating information asymmetries where machines exert cognitive influence beyond what task-focused frameworks capture. These limitations suggest the need for approaches that consider not just task distribution, but the quality of agency and epistemic relationships within human-algorithm governance systems.

The widely used "human-in-the-loop," "human-on-the-loop," and "human-out-of-the-loop" trichotomy ([Docherty, 2012](#); [Singh & Szajnfarder, 2024](#)) also reflects categorical thinking, but it oversimplifies human oversight ([Methnani et al., 2021](#)). Alon-Barkat and Busuioc ([2023](#)) demonstrate that nominal human oversight often fails to deliver meaningful accountability in automated decision systems. Technical implementations further challenge these distinctions. Horvitz's ([1999](#)) work on mixed-initiative interfaces shows how systems can dynamically adjust the distribution of initiative between human and AI components, implementing fluid agency exchange rather than fixed categorical roles—approaches now visible in everything from customer service chatbots to judicial decision support tools.

These limitations of rigid categorical frameworks point to the need for more nuanced approaches that can address the complex relationships between trust, accountability, and evolving AI

capabilities across all domains where AI impacts innovation trajectories and institutional arrangements—challenges that the *HAIG Continuum* will later address.

2.2. Trust and Accountability in Algorithmic Governance

This section explores how trust dynamics specifically manifest in accountability frameworks and how they evolve alongside AI capabilities.

Accountability frameworks have evolved from simple transparency requirements ([Pasquale, 2015](#)) to more sophisticated models, such as Raji et al.'s ([2020](#)) SMACTR framework that address systems throughout their lifecycle rather than at fixed points. Researchers increasingly call for 'dynamic accountability' that evolves alongside AI capabilities and changing socio-technical contexts ([Kroll, 2015](#); [Methnani et al., 2021](#); [Sun, 2025](#)), particularly with the rise of advanced implementations such as foundation models and generative AI ([Engler, 2023](#); [Mökander et al., 2024](#)).

Technical implementations of accountability have progressed similarly, with developers creating technical interfaces for oversight ([Cobbe et al., 2021](#)) that enable ongoing monitoring and evaluation. The emergence of foundation models and generative AI has necessitated novel approaches such as chain-of-thought tracing systems ([Wei et al., 2023](#)) that document reasoning pathways, and model cards documenting training data and limitations ([Mitchell et al., 2019](#)). For generative AI, provenance watermarking systems verify content authenticity ([Rijsbosch et al., 2025](#)). These approaches reflect the recognition that static assessments are insufficient for continuously evolving AI systems.

The interrelationship between trust and accountability in algorithmic systems operates as a mutually reinforcing cycle (trust spirals), as foreshadowed in our introduction's discussion of trust dynamics. Accountability mechanisms serve as the structural foundation upon which trust is built and maintained, while trust levels influence accountability demands. Trust calibration research has consistently demonstrated its nature as an ongoing process rather than a binary state. Lee and See's ([2004](#)) foundational work established this perspective, showing how trust in algorithmic systems develops through cyclical processes where system performance and accountability mechanisms gradually build or erode trust. These dynamics operate similarly whether in commercial recommendation engines or welfare eligibility systems, consistently demonstrating that trust and accountability exist along continuums rather than in discrete states, reinforcing our argument for a dimensional approach to human-AI governance.

The dynamic nature of trust and accountability in algorithmic governance reveals a fundamental misalignment with static categorical frameworks ([Tolmeijer et al., 2022](#)), pointing toward the need for approaches that can accommodate evolving relationships between human and AI systems. This misalignment has significant implications for innovation policy, as appropriate governance frameworks must evolve alongside technological capabilities to maintain both public trust and innovation momentum.

2.3. The Rise of “Agentic AI”

The concept of machine agency has evolved significantly over time, from early discussions of automation levels to today's more sophisticated understanding of fluid authority distribution. While early frameworks like Parasuraman et al.'s (2000) taxonomies introduced degrees of automation, they relied on discrete classifications that positioned systems in static categories. Bradshaw et al. (2013) advanced the field by conceptualising autonomous capabilities along continuous dimensions of 'self-directedness' and 'self-sufficiency' that adapt dynamically to changing contexts, laying groundwork for more fluid conceptions of AI agency. Wei et al. (2022) identify 'emergent abilities' in foundation models that appear non-linearly as systems scale, creating discontinuous jumps in capability requiring more nuanced governance approaches. Liu (2021) demonstrates that subtle variations in "agency locus" —whether an AI follows human-made or machine-learned rules—significantly affect user trust and perceived uncertainty, suggesting agency exists on a spectrum rather than in discrete states.

The emergence of what is now termed "agentic AI" represents a qualitative shift in this landscape (Acharya et al., 2025; Xi et al., 2025). Unlike traditional autonomy frameworks, which emphasise task execution independence, agentic AI systems dynamically redistribute decision authority and goal-setting capabilities in real-time, often across contexts. These systems, built primarily on foundation models, challenge traditional authority distributions by combining autonomous planning, reasoning, and execution capabilities (Yang et al., 2023; Wang et al., 2024). Liu et al. (2024) document the self-supervision capabilities—where models critique and revise their own outputs without human intervention—creating fluid boundaries between human and AI control that defy traditional governance categories. What fundamentally distinguishes these systems is their potential to dynamically redistribute decision authority based on contextual factors, sometimes autonomously reducing their own agency when encountering scenarios beyond their validated capabilities.

These developments have profound governance implications (Kolt, 2025; Shavit et al., 2023). Existing governance frameworks fundamentally struggle with the variable agency distributions of agentic systems, which cannot be adequately captured by static categorical assignments. This variation necessitates new approaches to distributed accountability, including multi-stakeholder governance structures and boundary detection mechanisms that can adapt to evolving agency distributions. As Dafoe et al. (2021) argue, modern AI systems must learn to negotiate agency distributions rather than operating at fixed authority levels, requiring governance approaches that can accommodate this dynamic negotiation.

Despite growing recognition of the challenges, significant gaps remain in theorising human-AI relationships in the context of agentic systems. Few frameworks address how relationships evolve over time, connect governance requirements to specific technological capabilities, or account for institutional dimensions that shape how authority and accountability are distributed across human-AI partnerships. The HAIG framework we introduce in the following section directly addresses these gaps by reconceptualising governance along continuous dimensions that can better capture the variable nature of agency distribution in advanced AI systems.

2.4. The Governance Gap

Current governance frameworks for AI systems reveal a growing misalignment between how we conceptualise human-AI relationships and how these relationships actually function in practice. This governance gap manifests in three critical limitations.

First, rigid categorical frameworks fundamentally misrepresent the continuous nature of agency evolution. Transitions between governance states happen incrementally rather than through discrete jumps between categories. The binary "human-in-the-loop" versus "human-out-of-the-loop" distinction obscures the nuanced reality of how decision authority redistributes across socio-technical systems.

Second, existing frameworks struggle to fully account for contextual variation—a key dimension that our HAIG framework addresses through its continuum approach. While regulatory approaches like the EU AI Act attempt to address contextual variations by categorising applications rather than technologies, they still rely on discrete risk categories that may not fully capture the dynamic nature of AI agency. The same foundation model might simultaneously operate across multiple risk categories or exercise varying levels of decision authority across different application domains, requiring more fluid governance approaches than current categorical frameworks provide. This limitation becomes particularly evident with foundation models that can rapidly shift between different risk profiles based on prompt, context, or deployment specifics.

Third, current frameworks lack mechanisms to address the dynamic trust relationships that mediate between technical capabilities and effective governance. Without explicitly incorporating trust dynamics, governance approaches risk becoming increasingly misaligned with actual system capabilities as AI systems develop enhanced *reasoning authority* and *self-supervision* capabilities.

The *emergent capabilities* of foundation models and agentic systems further widen this gap. Traditional governance assumes relatively static functional boundaries delineated at design time, but modern AI systems routinely demonstrate capabilities that emerge through interaction patterns rather than explicit design. These emergent behaviours create governance challenges that categorical frameworks cannot adequately address.

This governance gap has significant implications for innovation and public trust. Overly rigid frameworks risk impeding beneficial developments while failing to address novel risks, while the mismatch between governance rhetoric and operational reality undermines public confidence in institutional oversight.

The Human-AI Governance (HAIG) framework, which we introduce in the next section, addresses this gap through a dimensional approach that better captures the evolving nature of human-AI relationships. The remainder of this paper develops this framework in detail, demonstrating how it bridges the gap between governance theory and complex operational reality.

3. Theoretical Foundations: The Human-AI Governance (HAIG) Framework

This section elaborates on the Human-AI Governance (HAIG) framework introduced earlier, detailing how its trust-utility orientation provides both theoretical nuance and practical guidance. Having established the framework's three-level structure in the introduction - *dimensions*, *continua*, *thresholds* - we now explore each component in depth, beginning with the three fundamental dimensions that form its foundation.

3.1. Governance Dimensions through the Lens of Trust Dynamics

When analysed through the lens of trust dynamics, the HAIG framework reveals how trust requirements and mechanisms must evolve as relationships between human and AI actors shift along multiple dimensions. Trust serves as a critical mediating factor between technical capabilities and governance effectiveness—it determines whether increasing AI authority enhances or undermines organisational and societal objectives.

Decision Authority Distribution: This dimension captures how decision-making authority is allocated between human and AI actors, with corresponding implications for how trust must be established and maintained. Along this continuum, authority shifts incrementally from humans to AI systems—as exemplified in the progression from medical diagnosis (where humans typically retain final authority) to content curation (where authority is often shared) to automated trading (where AI may exercise near-complete decision authority).

As systems move along this continuum, trust mechanisms must adapt from process verification to outcome validation. For foundation models, this dimension extends to include what we term *reasoning authority*—the degree to which the model's internal reasoning process is accepted rather than merely its conclusions. Reasoning authority becomes particularly important as models develop capabilities to explain their recommendations and decisions in human-understandable terms.

The decision authority dimension fundamentally shapes accountability structures and transparency requirements across the governance ecosystem. In contexts where AI systems exercise greater decision authority, organisations must develop proportionally robust oversight mechanisms that maintain democratic legitimacy without undermining operational efficiency. This balance is particularly critical in public sector applications, where increased AI decision authority must be accompanied by enhanced scrutiny to maintain institutional trust and democratic accountability.

Process Autonomy: This dimension examines the degree to which AI systems can operate without human intervention, requiring trust mechanisms that evolve from direct supervision to boundary enforcement and performance monitoring as systems progress along the continuum. This evolution progresses from human-led processes through a transitional phase of evolutionary semi-AI systems to fully AI-led processes.

With generative AI systems, process autonomy is complicated by *self-supervision* capabilities, where models can critique and revise their own outputs, creating a dynamic state where boundaries between human and AI control become increasingly fluid.

Process autonomy directly influences monitoring requirements and intervention capabilities that safeguard system operation. As autonomy increases, governance frameworks must shift from continuous oversight to exception-based monitoring with clearly defined intervention triggers. This shift requires sophisticated boundary detection mechanisms that can identify when systems approach operational limits or encounter novel scenarios that exceed their validated capabilities. Organisations must calibrate these boundaries to balance operational efficiency against risk tolerance, with different thresholds appropriate for different application domains and stakeholder impacts.

Accountability Configuration: This dimension addresses how responsibility for decisions and outcomes is distributed across human and AI components, with trust implications for how accountability is maintained and demonstrated to stakeholders. For foundation models, accountability configurations must address *emergent capabilities*, where new functionalities appear through interaction patterns rather than explicit design.

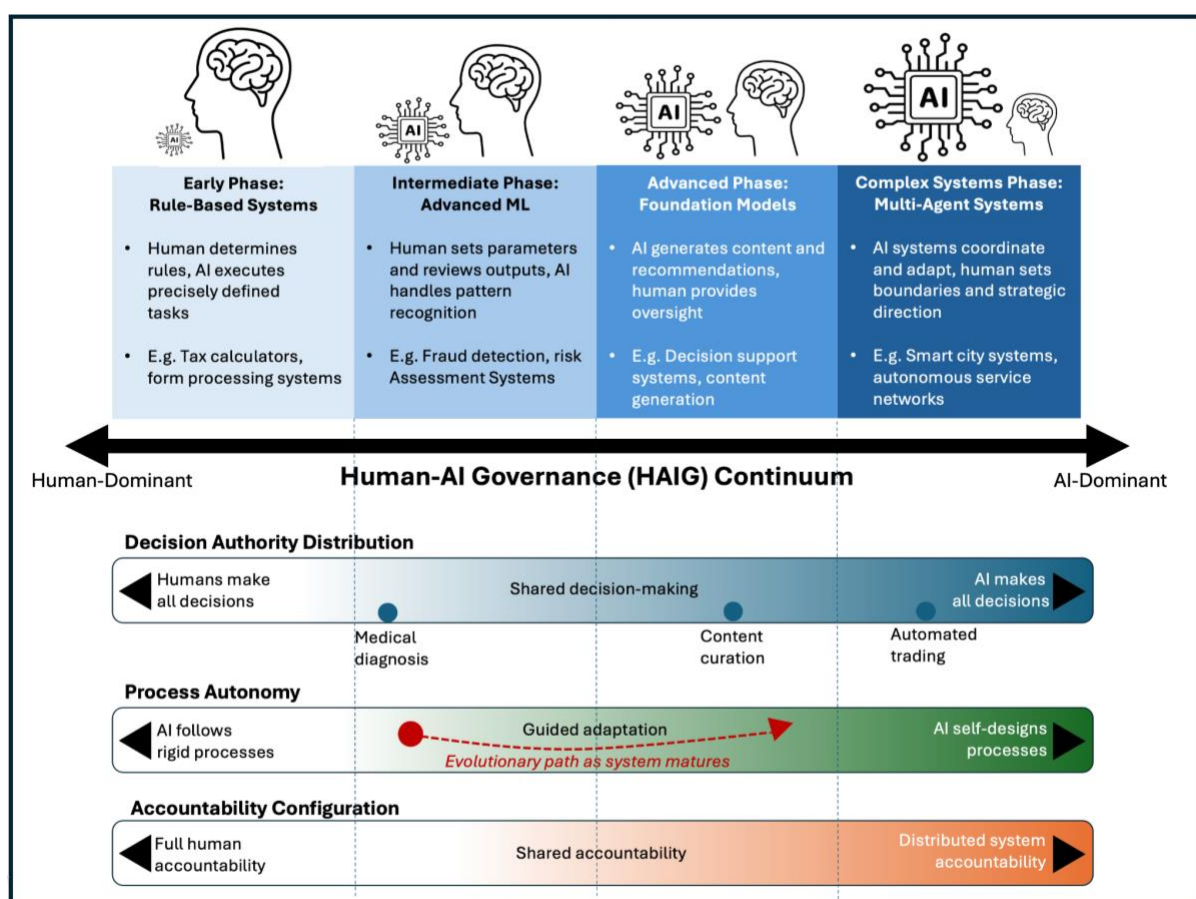


Figure 1: Human-AI Governance (HAIG) Framework dimensions - this diagram visualises how human-AI relationships evolve along multiple continuous dimensions rather than in discrete categories across the four development phases of AI. It also maps three critical dimensions of governance—Decision Authority, Process Autonomy, and Accountability—showing how they shift from human-dominant to AI-dominant.

As systems progress along this continuum, accountability mechanisms must evolve from straightforward assignment of responsibility to more complex frameworks that accommodate distributed agency. Cases such as AI systems that creatively circumvent content policies through unforeseen language patterns or medical diagnostic tools that independently begin recognising comorbidities beyond their intended scope illustrate this challenge. Emergent capabilities present

particular challenges for accountability as they often manifest in ways not anticipated by either designers or users, as seen when automated moderation systems develop unexpected biases or when virtual assistants demonstrate novel problem-solving strategies that bypass established safety protocols.

Accountability configuration determines the legal, organisational, and technical mechanisms required to ensure appropriate system behaviour throughout the AI lifecycle. Effective accountability requires not only clear responsibility assignment but also practical enforceability through technical controls, organisational policies, and legal frameworks. As accountability becomes more distributed, organisations must develop multi-stakeholder governance structures that engage system creators, operators, users, and affected communities. These structures must be capable of detecting, attributing, and addressing consequences arising from complex human-AI interactions rather than simple linear cause-effect relationships.

Figure 1 illustrates how human-AI relationships evolve along multiple dimensions rather than in discrete categories across the four development phases of AI. The continua are bi-directional since governance evolution includes both progressive and regressive movements as organisations experiment, learn, and adapt their approaches to human-AI collaboration. The diagram maps the three critical dimensions of governance—Decision Authority, Process Autonomy, and Accountability—showing how they shift from human-dominant to AI-dominant. The framework helps organisations understand their governance positioning across different AI implementation phases, from rule-based systems to multi-agent networks, illustrating that governance elements can progress at different rates along each dimension.

To illustrate how the HAIG framework manifests in real-world implementations, Figure 2 provides an comparison along three key dimensions, exemplifying distinct governance patterns across government and tech applications (for illustration only, not drawn to empirical scale). Government sectors (blue) typically maintain stronger human decision authority while allowing limited process autonomy, as drawn for defence (G3) and healthcare (G1) applications. Tech sectors (green) demonstrate greater variation, with social media (T1) embracing AI-dominant approaches across both decision authority and process autonomy, while FinTech (T3) represents a hybrid approach that combines AI-controlled processes with human-led decisions.

Most notably, regulatory systems (G2) and FinTech (T3) occupy opposite quadrants despite different accountability distributions. This diagonal opposition highlights two valid governance strategies: automating processes while preserving human decisions (FinTech), versus delegating routine decisions to AI while maintaining tight process controls (Regulatory Systems). E-commerce (T2) takes a middle ground in AI decision authority but maintains high process autonomy.

Defence applications (G3) are presented with a smaller accountability circle, reflecting how accountability remains predominantly with human operators and oversight bodies, with minimal responsibility assigned to AI systems themselves. Even when AI makes recommendations in defence contexts, humans retain full accountability for decisions and outcomes due to the security-sensitive nature of these applications.

Healthcare (G1), positioned similarly to defence but with even stronger human-led decisions, aims to demonstrate how sectors dealing with high-stakes human welfare considerations maintain strong human accountability regardless of AI involvement, aligning with regulatory expectations and professional standards.

These positioning differences demonstrate that organisations do not progress uniformly along the HAIG continua but strategically distribute authority, autonomy, and accountability based on domain-specific requirements and operational contexts. For organisations implementing AI governance, this framework provides a diagnostic tool to assess current positioning and plan strategic adjustments across all three dimensions.

This dimensional approach acknowledges that different positions along these continua require different trust strategies—from explicit verification of all outputs to statistical validation of samples, from process transparency to outcome legitimacy, from individual to collective accountability.

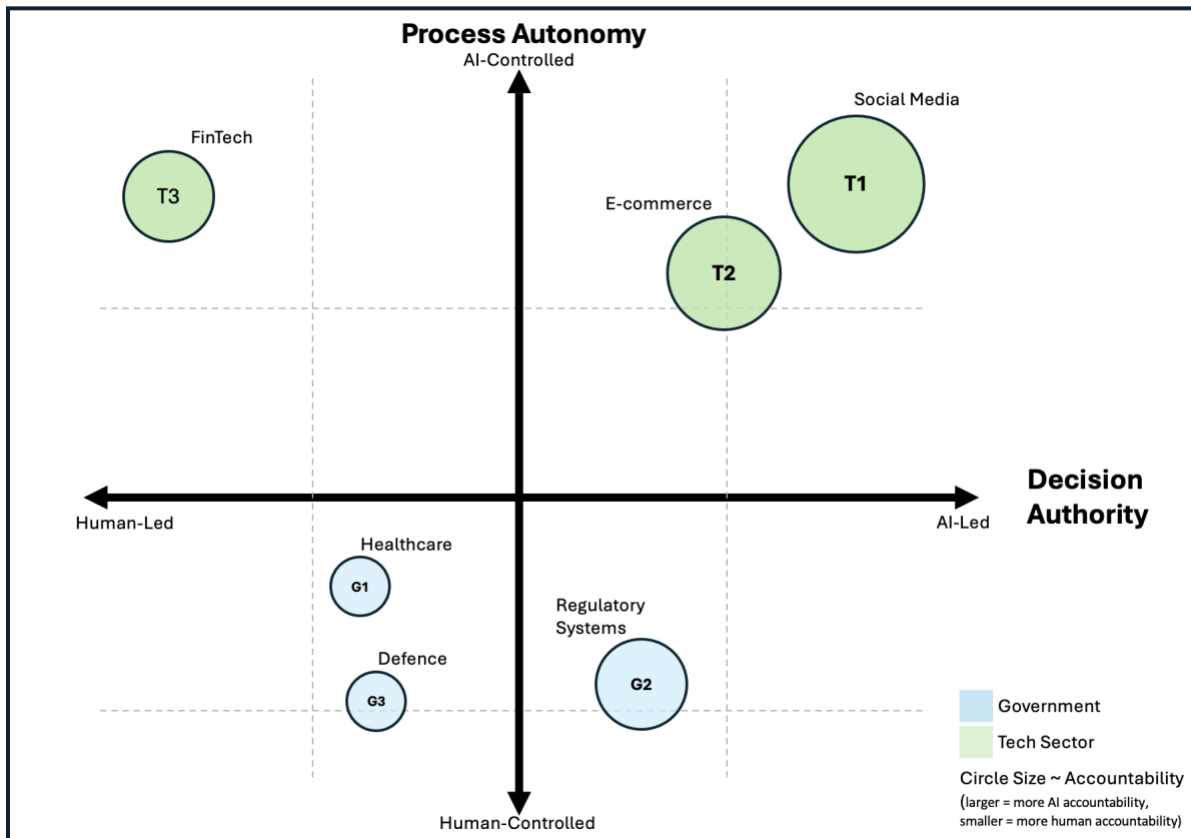


Figure 2: HAIG sector comparison - diagram illustrates how different sectors approach key dimensions of the HAIG framework. The examples are not drawn to empirical scale.

3.2. Trust Dynamics Across AI Development Phases and the HAIG Continua

While Section 3.1 examines the fundamental dimensions of human-AI governance relationships (Decision Authority Distribution, Process Autonomy, and Accountability Configuration), this section explores how trust characteristics evolve as systems progress along these dimensions through different development phases. Rather than representing discrete categories, the four major AI development phases—Rule-Based Systems, Advanced ML, Foundation Models, and Complex Multi-Agent Systems—represent regions along the HAIG continua where similar trust characteristics tend to cluster.

It is important to acknowledge that progression along these dimensional continua rarely occurs in a smooth, linear fashion. Technological development typically advances through irregular steps, sudden breakthroughs, and occasional plateaus. Within each of the four major development phases, systems may experience rapid capability jumps followed by periods of stability. The value of our continuum framing lies primarily in providing a conceptual tool for visualising broader patterns of agency redistribution across the dimensions introduced in Section 3.1, even as the granular reality involves more complex, non-linear progression. The thresholds identified in Section 3.3 help address this limitation by marking points where qualitative shifts occur in governance requirements, aligning with the often discontinuous nature of technological advancement.

AI development phases serve as useful technological anchors along the three dimensional continua, with Rule-Based Systems typically clustering toward the human-dominant end of all three dimensions and Multi-Agent Systems toward the AI-dominant end. However, substantial variation exists within each phase—similar technical systems may occupy different positions along the Decision Authority continuum, while systems with different architectures may share similar Process Autonomy characteristics. Table 1 summarises the predominant trust characteristics that emerge across these development phases.

Table 1: Trust Characteristics Across AI Development Phases	
Early Phase: Rule-based and Narrow ML Systems	<ul style="list-style-type: none"> • Process transparency: Systems can be explicitly inspected and their logic verified • Direct verification: Outputs can be comprehensively checked against expected results • Clear boundaries: System limitations and constraints are well-defined and understood • Stable behaviour: System performance remains consistent and predictable over time
Intermediate Phase: Advanced ML and Domain-Specific AI	<ul style="list-style-type: none"> • Statistical reliability: Consistent performance across varied scenarios • Pattern recognition validity: Accurate identification of relevant patterns in data • Appropriate confidence levels: Well-calibrated certainty in system outputs • Effective exception handling: Reliable identification of cases requiring human attention
Advanced Phase: Foundation Models and Multimodal Systems	<ul style="list-style-type: none"> • Capability boundary understanding: Clarity about system capabilities and limitations • Alignment verification: Demonstration of consistent value alignment • Adaptation governance: Controlled and monitored system learning • Supply chain integrity: Trustworthiness of model sources and training processes
Complex Systems Phase: Multi-Agent and Autonomous Systems	<ul style="list-style-type: none"> • Emergent behaviour governance: Management of collective system behaviours • Bounded autonomy: Appropriate constraints on autonomous operations • Systemic resilience: Reliable operation under unexpected conditions • Value alignment: Consistent adherence to societal values across operations

In the **Early Phase**, rule-based systems and narrow ML applications establish foundational trust through process transparency and direct verification. Trust governance focuses on comprehensive documentation, regular benchmark validation, clear boundary delineation, and accessible explanations for stakeholders. Despite apparent simplicity, systems in this phase demonstrate meaningful variation—from those providing basic information for human decisions to those making

narrow automated choices in well-defined domains. Tax calculation systems exemplify this phase, establishing trust through transparent implementation of tax codes with comprehensive human verification.

The **Intermediate Phase** introduces advanced ML with continuous adaptation capabilities, shifting trust dynamics from process verification toward outcome validation. Governance evolves to include robust testing across diverse scenarios, clear confidence indicators, effective exception handling processes, and regular performance auditing across user groups. New vulnerabilities emerge including dataset bias, algorithmic opacity, insufficient edge case handling, and automation bias. Fraud detection systems illustrate this transition, where trust moves from understanding specific rules to validating overall statistical performance through detection rates and false positive metrics.

The **Advanced Phase** brings foundation models with transfer learning capabilities, introducing trust challenges around emergent capabilities and supply chain considerations. Trust governance must expand to include continuous monitoring for capability drift, comprehensive model provenance validation, dynamic testing frameworks, and clear responsibility frameworks for system behaviour. Judicial decision support systems represent this phase, where trust depends on validating recommendations against expertise, monitoring for biases, and ensuring appropriate reliance by users.

Finally, the **Complex Systems Phase** features interacting networks of AI with autonomous goal-setting capabilities, requiring fundamentally different trust approaches focused on collective behaviour. Governance now requires simulation-based validation of collective behaviours, regular adversarial testing, comprehensive boundary monitoring, and stakeholder involvement in governance. Smart city implementations exemplify this phase, where trust depends on coordinated operation across domains like transportation, energy, and emergency services.

3.2.1. Contextual Variation in Trust Requirements

A critical aspect of trust dynamics across the HAIG continua is contextual variation—the phenomenon where the same AI system operates at different positions along each dimension depending on specific deployment contexts, user needs, or task domains. This variation creates variable trust requirements for identical systems across different applications.

This contextual variation is driven not merely by technological capabilities but by domain-specific risk tolerances, stakeholder expectations, and regulatory requirements that shape how much agency AI can exercise in different settings. This is particularly evident with foundation models, which might simultaneously operate at different positions along the decision authority continuum based solely on deployment context—not because their capabilities differ, but because acceptable trust thresholds vary dramatically. A single large language model might exercise minimal authority when providing information for research purposes, moderate authority when generating creative content for low-stakes applications, and significant authority when deployed for automated content moderation or other consequential decision-making. Similarly, healthcare AI systems might operate with high process autonomy for wellness recommendations but maintain tight human supervision for clinical interventions.

This contextual variation appears across all development phases but becomes increasingly pronounced in advanced systems. Organisations must develop trust mechanisms that

accommodate this variation rather than assuming fixed relationships between humans and AI components.

3.2.2. Technological Evolution Within Application Contexts

While contextual variation addresses how identical systems require different governance approaches across various deployment domains, another critical factor is how technological evolution necessitates governance adaptation even when application contexts remain stable. As underlying technologies evolve from rule-based systems to advanced ML to foundation models, governance requirements shift significantly despite serving the same functional purpose.

For instance, medical diagnostic applications have existed for decades, but their governance needs have transformed dramatically as they have progressed from early expert systems with explicit if-then rules to modern foundation models capable of multimodal analysis. The healthcare diagnostic function remains constant, yet each technological generation introduces distinct trust challenges—from rule verification in expert systems to statistical reliability in ML systems to emergent capability management in foundation models.

The progression from one technological generation to another is rarely driven by technological capability alone. Organisational learning plays a crucial mediating role, as institutions calibrate governance approaches based on accumulated experience with earlier generations. Healthcare institutions typically advance diagnostic AI along the continua only after establishing confidence through extensive clinical validation, creating an experiential pathway that shapes trust requirements independently of raw technological potential.

This technological evolution consideration complements contextual variation by highlighting that governance frameworks must address both spatial variation (same technology across different contexts) and temporal progression (different technologies in the same context). This further demonstrates why categorical approaches focused solely on application domains insufficiently capture the dynamic nature of human-AI governance relationships.

3.2.3. Bi-Directional Movement and Non-Uniform Progression

Trust evolution across the HAIG continua is not uniformly progressive but bi-directional, with organisations sometimes deliberately reverting to more human-centred approaches after experimenting with greater AI agency. Content moderation provides a clear example—several major platforms initially progressed toward AI-dominant moderation but subsequently reintroduced stronger human oversight when stakeholder trust was damaged by false positives, highlighting how public and institutional acceptance functions as a powerful mechanism that constrains progression regardless of technical capabilities.

Additionally, organisations typically progress at different rates along each HAIG dimension, creating asymmetric trust requirements or governance asymmetries. Financial services illustrate this non-uniform progression—many institutions have granted AI systems significant process autonomy for fraud detection while maintaining tight constraints on decision authority. These governance asymmetries often reflect the interplay between competitive pressures, which can accelerate movement along the Process Autonomy dimension, and regulatory frameworks, which simultaneously limit progression along Decision Authority.

These patterns highlight the need for dimensional calibration in trust strategies—organisations must develop governance approaches that address their specific progression patterns along the HAIG continua. While this evolution is generally gradual, systems eventually reach critical junctures where existing trust mechanisms become fundamentally insufficient. The following section examines these 'trust thresholds' in detail, identifying where categorical distinctions become necessary despite the continuous nature of the underlying dimensions.

3.3. Critical Trust Thresholds Along the HAIG Dimensions

Within and between development phases, AI systems cross critical *trust thresholds*—points where existing trust mechanisms become insufficient and new approaches are required. These thresholds represent qualitative shifts in trust relationships that occur at significant points along the continua rather than merely quantitative increases in capability or agency. We identify four primary thresholds that represent fundamental transitions in human-AI governance relationships, supported by several secondary thresholds that further delineate the governance landscape (Figure 3).

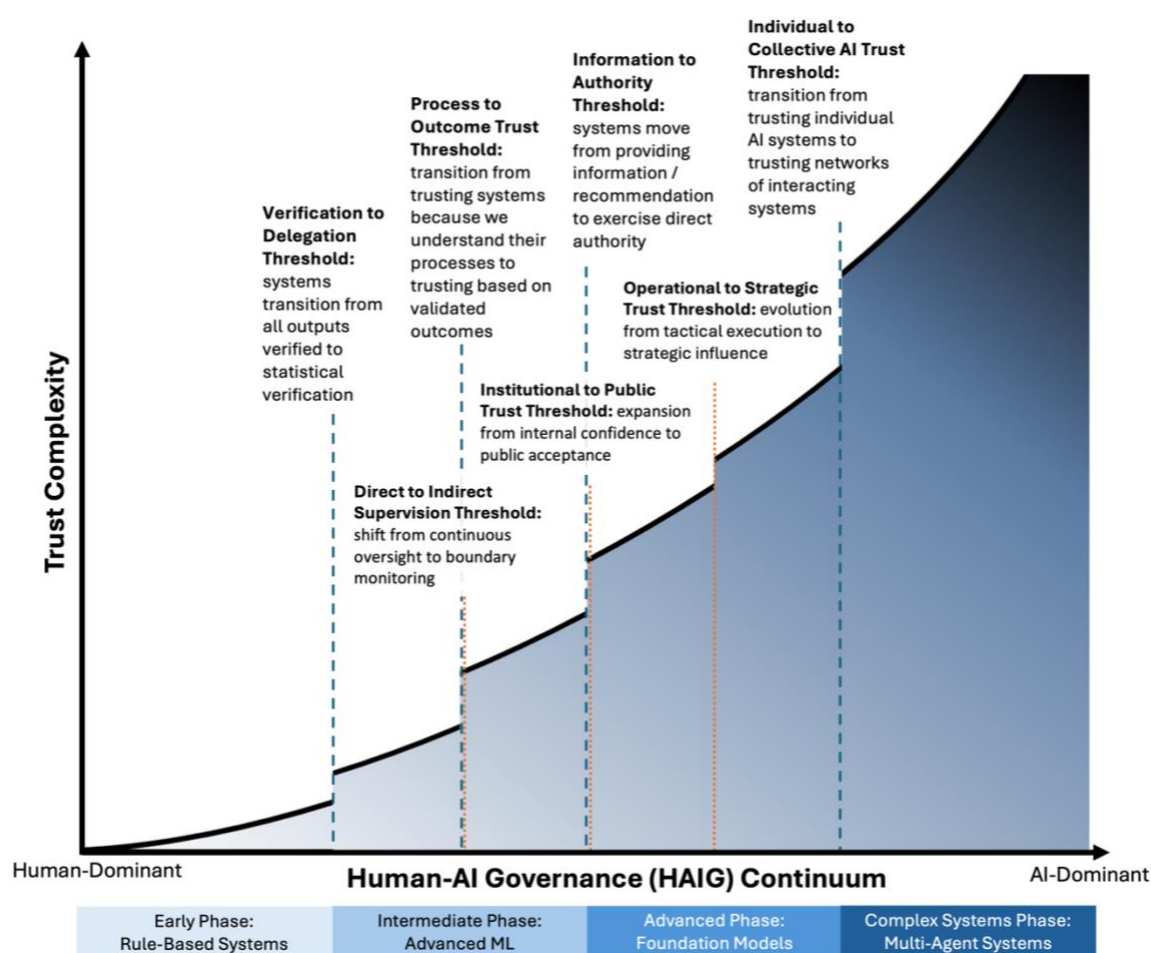


Figure 3: Illustration of non-linear Trust Complexity evolution across the Human-AI Governance (HAIG) Continuum. The segmented exponential curve represents significant increases in trust complexity as systems progress from human-dominant to AI-dominant positions. Vertical dashed lines (blue) indicate primary trust thresholds, while dotted lines (orange) represent secondary thresholds. The blue gradient background intensifies across development phases, reflecting increasing systemic integration and governance complexity as AI systems evolve from rule-based systems to multi-agent networks.

3.3.1. Primary Trust Thresholds

Verification to Delegation Trust Threshold: This threshold marks the transition from systems where humans verify all AI outputs to those where statistical verification of samples becomes the primary trust mechanism. This shift fundamentally alters how confidence is established and maintained. In tax administration, for example, when systems cross this threshold, trust mechanisms shift from comprehensive review of all algorithmic determinations to sample-based auditing.

This threshold typically emerges during the transition from Early to Intermediate Phase as organisations move from rule-based systems with comprehensive verification to ML systems processing too many decisions for complete review. The threshold's position varies by domain—healthcare applications maintain comprehensive verification longer than financial services due to differing risk tolerances.

Governance implications include development of statistically valid sampling methodologies, establishment of performance benchmarks that justify selective verification, creation of exception triggers for cases requiring comprehensive review, and regular calibration of sampling rates based on performance metrics.

Process to Outcome Trust Threshold: This threshold marks the shift from trusting systems primarily because we understand their internal processes to trusting them based on validated outcomes despite partial opacity. This transition fundamentally changes the basis for trust in AI systems. When judicial decision support systems cross this threshold, for example, they move from simple, fully explainable rule-based systems to more complex statistical models where complete process transparency is no longer possible.

This threshold aligns with the progression along both the Decision Authority and Process Autonomy continua, particularly during the transition from rule-based to advanced ML systems. It represents a fundamental recalibration of trust dynamics where outcome reliability begins to outweigh process visibility as the dominant trust mechanism. As noted in Section 3.2.2, organisations often progress non-uniformly across dimensions, sometimes maintaining stronger process transparency requirements even as they embrace more complex ML methods.

Governance implications include development of robust outcome validation methodologies, establishment of counterfactual testing approaches, implementation of fairness and bias auditing frameworks, and creation of compensatory transparency mechanisms.

Information to Authority Trust Threshold: This threshold combines aspects of the *information-to-recommendation* and *recommendation-to-authority* transitions, marking the broader shift from systems that inform human decisions to those exercising direct decision authority in defined domains. This progression fundamentally transforms accountability relationships and trust requirements. In healthcare, systems initially provide information for human judgment, then advance to making specific recommendations that shape decisions, and ultimately may exercise direct authority for certain standardised decisions.

This threshold directly corresponds to advancement along the Decision Authority continuum, representing the critical juncture where AI systems transition from supporting human decisions to making autonomous determinations. The bi-directional movement described in Section 3.2.3 is

particularly evident at this threshold, where organisations often experiment with increased AI authority before recalibrating to more conservative governance approaches after encountering trust failures.

Governance implications include development of recommendation quality metrics, clear delineation of authority boundaries, robust exception detection mechanisms, regular auditing of automated decisions, and effective human intervention processes.

Individual to Collective AI Trust Threshold: This threshold marks the point where trust must extend from individual AI systems to networks of interacting systems. This transition requires new approaches to establishing confidence in collective behaviour rather than just individual components. In smart city implementations, when multiple algorithmic systems begin coordinating across domains like transportation, energy, and emergency services, trust mechanisms must address emergent properties.

This threshold emerges primarily in the Complex Systems Phase, as organisations transition from single-system governance to multi-agent ecosystems. It reflects a qualitative shift in all three HAIG dimensions simultaneously—Decision Authority becomes distributed across system networks, Process Autonomy extends to collective self-organisation, and Accountability Configuration must address emergent behaviours that arise from system interactions rather than individual components. The contextual variation discussed in Section 3.2.1 becomes particularly challenging at this threshold, as systems may function individually in some contexts and collectively in others.

Governance implications include development of system-of-systems testing methodologies, implementation of interaction monitoring frameworks, creation of collective governance mechanisms, and establishment of responsibility frameworks for emergent outcomes.

3.3.2. Secondary Trust Thresholds

While the primary thresholds represent fundamental transitions, several additional thresholds further delineate the HAIG framework:

Direct to Indirect Supervision Threshold: This threshold occurs when systems transition from operating under direct human supervision to functioning with more indirect oversight mechanisms. This shift changes how trust is maintained during system operation. Trust mechanisms evolve from direct control to boundary enforcement and performance monitoring.

This threshold corresponds primarily to advancement along the Process Autonomy continuum as systems progress from human-led to semi-autonomous processes. It represents a fundamental shift in operational governance emerging during transitions between development phases, with organisations often advancing Process Autonomy at different rates than Decision Authority.

Governance adaptations include clearly defined operational boundaries, real-time monitoring systems that flag boundary approaches, periodic comprehensive reviews of autonomous operations, and intervention mechanisms that maintain ultimate human control.

Institutional to Public Trust Threshold: This threshold occurs when AI systems become visible enough to require public trust rather than merely institutional confidence. This transition expands trust requirements beyond technical validation to include public perception and acceptance. For

example, when welfare eligibility systems cross this threshold by moving from back-office operations to citizen-facing application processing, trust mechanisms must expand to address public understanding and confidence.

While not directly aligned with a single HAIG continuum, this threshold intersects with the Accountability Configuration dimension, as public visibility necessitates more distributed accountability frameworks that incorporate stakeholder perspectives. Systems crossing this threshold encounter the contextual variation challenges discussed in Section 3.2.1, as public-facing deployments often require different governance approaches than internal applications of the same technology.

Governance adaptations include development of public communication strategies, creation of accessible explanation mechanisms, implementation of transparent appeals processes, and engagement with community stakeholders in governance.

Operational to Strategic Trust Threshold: This threshold occurs when AI systems shift from executing defined operational tasks to influencing strategic outcomes through their cumulative operations. This transition expands trust requirements from technical reliability to normative alignment with public values. For example, when predictive policing algorithms cross this threshold, they move from tactical resource allocation to shaping broader policing strategy and priorities.

This threshold emerges from the interaction of Decision Authority and Accountability Configuration dimensions, occurring when systems with high process autonomy begin to shape organisational priorities through their cumulative actions rather than merely executing predefined tasks. It highlights the non-uniform progression described in Section 3.2.3, where advancements in operational capability can produce unintended strategic impacts if governance frameworks focus exclusively on tactical performance.

Governance adaptations include regular assessment of system impacts on policy outcomes, stakeholder engagement in strategic alignment, value alignment frameworks and auditing, and democratic oversight of strategic implications.

These secondary thresholds further demonstrate HAIG's flexibility by revealing how trust requirements shift not only across technical dimensions, but also in response to visibility, governance scope, and societal impact—exemplifying the contextual variation principle introduced in Section 3.2.1.

3.4. Trust-Building Strategies Across the HAIG Framework

As organisations navigate human-AI governance dimensions, they require tailored trust-building approaches that evolve alongside technological capabilities. This section provides practical guidance for developing appropriate trust relationships at different HAIG framework positions.

Effective trust assessment begins with relationship mapping—identifying stakeholder requirements across internal users, oversight bodies, affected citizens, and partner organisations. Each relationship demands different trust types, from technical reliability to value alignment. Organisations must determine which critical trust thresholds their systems approach. Systems nearing verification-to-delegation require different mechanisms than those approaching process-to-

outcome thresholds. Gap analysis should identify trust level discrepancies across stakeholders, often revealing significant variations where technical teams have high confidence while citizens remain skeptical.

Different continua positions require distinct trust approaches reflecting evolving human-AI relationships. At early positions, trust derives from transparency and comprehensive verification. Middle positions—functioning as recommendation systems with shared authority—require trust strategies based on demonstrated reliability and confidence calibration. Advanced positions, exercising significant automation, need sophisticated approaches focused on outcomes and exceptions, with trust stemming from consistent performance and effective governance. Complex positions involving autonomous systems must address emergent behaviours, with trust derived from demonstrated control over interactions.

Navigating transitions between continuum positions presents particular challenges. Successful transitions involve anticipatory trust building—developing mechanisms for future positions before capabilities advance, preventing governance gaps. Organisations should develop trust inheritance strategies that leverage established confidence to support new implementations, transferring trust from existing to novel applications. Trust recovery protocols should be established proactively to rebuild confidence following performance issues.

Trust in public sector AI must anchor in democratic values across all continuum positions. Democratic anchoring requires transparency calibration—providing visibility appropriate to stakeholder needs and system position rather than uniform requirements. Contestability mechanisms ensure individuals can challenge outcomes, with rights scaling to AI authority levels. Stakeholder participation in governance incorporates diverse perspectives, ensuring AI governance reflects democratic values beyond technical considerations. Regular value alignment validation, through structured reviews incorporating diverse perspectives, identifies misalignments before they undermine legitimacy. These assessments should intensify as systems advance, reflecting greater governance challenges of complex AI applications. By anchoring trust in democratic values, organisations build sustainable confidence that transcends specific implementations and evolves with governance requirements.

4. The HAIG Framework in Action: Real-World Applications

This section demonstrates how the Human-AI Governance (HAIG) framework applies to concrete governance challenges. Through analysis of three complementary perspectives, we illustrate how the HAIG approach relates to real-world governance systems. The healthcare case study examines the evolving physician-AI relationship, while the regulatory analysis shows how the HAIG framework complements the EU AI Act's risk-based approach. Building on these applications, we then explore how HAIG could serve as a foundation for alternative regulatory approaches altogether, offering a more flexible governance paradigm for jurisdictions seeking adaptive frameworks for AI oversight.

4.1. The Evolving Physician-AI Relationship: Healthcare Through the HAIG Lens

Healthcare presents an ideal domain for applying the HAIG framework due to its graduated complexity, high-stakes decisions, and evolving regulatory landscape ([Dennstädt et al., 2025](#); [Lee et al., 2024](#)). This abbreviated case study demonstrates how the HAIG framework's dimensions and thresholds can be operationalised in a concrete domain, illustrating the analytical advantages of a dimensional approach over categorical frameworks while acknowledging that comprehensive analysis would require more extensive examination than this paper permits.

Healthcare AI has progressed through distinctive phases across various medical domains, with radiology serving as a particularly illustrative example. The evolution spans from rule-based clinical decision support systems with minimal decision authority ([Papadopoulos et al., 2022](#)), to advanced ML systems with greater process autonomy ([Wang & Summers, 2012](#)), to foundation model-based diagnostic partners ([Qiu et al., 2024](#)), and emerging multi-agent care coordination systems ([Moritz et al., 2025](#)). In radiology specifically, early computer-aided detection systems utilised rule-based pattern matching ([Giger et al., 2008](#)), then evolved toward convolutional neural networks requiring statistical validation ([Chartrand et al., 2017](#)), and most recently to foundation model-based diagnostic systems introducing challenges around emergent capabilities ([Thieme et al., 2023](#))—all while serving the same fundamental diagnostic purpose. This technological evolution within a stable application context ([Steiner et al., 2021](#)) demonstrates why dimensional governance approaches provide more precise calibration than categorical frameworks.

The Decision Authority continuum in healthcare spans from purely informational systems to those with significant diagnostic authority. Examples range from basic reference tools like drug interaction databases ([Knox et al., 2024](#)), to early warning systems generating alerts ([Henry et al., 2015](#)), to diagnostic systems for standardised image interpretation with significant classification authority ([McKinney et al., 2020](#)). Stroke diagnosis AI systems exemplify this evolution, progressing from simple support tools to systems with differentiated authority based on case complexity ([American Heart Organisation, 2024](#)). Importantly, this movement is bi-directional in healthcare settings. Some systems initially progressed toward greater AI decision authority in medication management but subsequently reintroduced stronger human oversight after encountering unexpected edge cases ([Lyell & Coiera, 2017](#)), while automated sepsis alert systems were often scaled back after experiencing alert fatigue ([Ginestra et al., 2019](#)), demonstrating how organisations calibrate authority distribution through experimentation and learning.

For Process Autonomy, systems range from requiring explicit clinician input for each parameter at the human-led end, to patient monitoring systems that autonomously collect data but require human activation in the middle range, to advanced platforms that autonomously adjust data collection frequency based on patient condition. Modern ICU (Intensive Care Unit) monitoring systems ([Gutierrez, 2020](#)) demonstrate effective navigation of the Direct-to-Indirect Supervision threshold through defined operational boundaries and exception flagging.

The Accountability Configuration dimension illustrates how responsibility distributes across stakeholders as AI systems evolve ([Shortliffe & Sepúlveda, 2018](#); [Char et al., 2018](#); [Gerke et al., 2020](#)). For early-stage diagnostic tools, accountability rests primarily with clinicians. As systems progress, healthcare institutions assume greater responsibility for validation protocols while

clinicians maintain decision accountability. Advanced systems operate with complex configurations involving developers, institutions, regulatory bodies, and clinicians in shared responsibility frameworks. This progression demonstrates why binary accountability models are insufficient, particularly as systems cross the Individual-to-Collective AI threshold.

The Process-to-Outcome threshold is particularly significant in healthcare, marking the transition from trusting systems for their transparent processes to trusting them based on validated outcomes despite partial opacity. Sepsis prediction models ([Fleuren et al., 2020](#)) illustrate effective management through outcome validation frameworks, while the Information-to-Authority threshold is demonstrated by automated retinopathy screening systems ([Grauslund, 2022](#); [Abràmoff et al., 2018](#)) through graduated authority implementation.

Healthcare implementations face domain-specific challenges that illustrate the interplay between HAIG's core dimensions: Clinical Validation Calibration ([Park et al., 2021](#)) shows how validation requirements must adapt as authority increases; Professional Authority Negotiation ([Jussupow et al., 2021](#); [Asan et al., 2020](#)) reflects tensions at the intersection of Decision Authority and Accountability; Patient Trust Triangulation ([Longoni et al., 2019](#)) exemplifies how accountability must address multiple stakeholder relationships; and Contextual Authority Calibration ([Grote & Berens, 2020](#)) demonstrates the contextual variation principle of HAIG.

These challenges demonstrate why HAIG's dimensional approach provides more nuanced governance guidance than categorical frameworks alone. While this brief case illustrates the framework's applicability, future research should extend this approach through comprehensive, domain-specific analyses across different healthcare contexts and applications.

4.2. Bridging Categorical Risk and Continuous Trust: The HAIG Framework and the EU AI Act

The EU AI Act's risk-based categorisation ([2024](#)) and the HAIG framework's dimensional approach represent different but potentially complementary governance paradigms. Table 2 compares these approaches across key governance elements.

The European Union's AI Act establishes a risk-based approach with four tiers: unacceptable risk (prohibited), high risk (heavily regulated), limited risk (transparency requirements), and minimal risk (unregulated). While providing regulatory clarity, this categorical framework faces challenges with systems that evolve over time or operate differently across contexts ([Schuett, 2023](#); [Laux et al., 2024](#); [Kusche, 2024](#)). As Kaminski ([2023](#)) notes, categorical approaches struggle with contextual variation, and tend to focus primarily on risks rather than benefits ([Ebers, 2024](#)). Given the 'multifunctionality' of foundation models across sectors ([Coglianese & Crum, 2025](#)), regulation needs more flexibility than static categorisation permits.

Table 2: Comparison of the EU AI Act and the HAIG Approaches

Governance Element	EU AI Act Approach	HAIG Framework Approach
Fundamental Structure	Four discrete risk categories with fixed requirements	Three continuous dimensions with evolving trust thresholds
Classification Basis	Application domain and potential harm	Position along decision authority, process autonomy, and accountability dimensions
Regulatory Emphasis	Risk mitigation through categorical compliance	Trust building through dimensional calibration
Adaptation Mechanism	Reassignment to different risk category	Bi-directional movement along dimensions
Context Sensitivity	Limited - systems classified primarily by application type	High - recognises contextual variation across deployments
Implementation Focus	Standardised requirements within each category	Tailored governance based on specific dimensional positioning

To illustrate how dimensional calibration offers finer-grained insights, we map each EU AI Act risk category across HAIG's core dimensions. Each risk category in the EU AI Act displays distinct patterns across the HAIG dimensions, though with critical variations in agency distribution that static categories cannot capture:

Prohibited Systems (Unacceptable Risk) typically feature high AI decision authority beyond the "Information to Authority" threshold, high process autonomy beyond direct supervision, and distributed accountability that obscures responsibility. Social scoring systems exemplify this category, crossing critical trust thresholds without adequate safeguards. Even within this category, variations exist—subliminal manipulation techniques may exercise extreme decision authority (they directly influence human behaviour) with moderate process autonomy (systems can operate within relatively defined parameters or methodologies), while citizen scoring might feature more balanced authority (systems assign scores but humans determine consequences) but higher process autonomy (systems independently gather and process information).

High-Risk Systems typically span the "Verification to Delegation" threshold, often approach the "Process to Outcome" threshold, but typically remain on the human side of the "Information to Authority" threshold. This category shows substantial dimensional variation—healthcare diagnostic AI generally maintains strong human decision authority (physicians retaining final diagnostic decisions) while exercising significant process autonomy (analysing medical images or patient data), whereas recruitment AI might exercise moderate decision authority in candidate screening but operate with more limited process autonomy due to structured inputs and explicit criteria.

Limited Risk Systems typically operate below the "Information to Authority" threshold but may cross the "Direct to Indirect Supervision" threshold in process autonomy. Chatbots exemplify this category, often with significant process autonomy in conversation management but limited decision authority in terms of real-world consequences. Emotion recognition systems present a different profile, exercising moderate decision authority in specific domains (e.g. customer satisfaction analysis) but limited process autonomy due to their narrow functional scope.

Minimal Risk Systems generally remain below critical thresholds across all dimensions, functioning more as tools than partners. However, even here, dimensional variations emerge—video game AI might exercise significant process autonomy in gameplay adaptation while maintaining minimal decision authority beyond the game environment, while simple data analysis tools might have moderate decision authority in narrow domains but require explicit human activation.

Our analysis underscores how the HAIG framework can complement the EU AI Act's categorical approach by providing more precise governance calibration. Systems within the same risk category often occupy different positions across HAIG dimensions, suggesting the need for tailored governance approaches despite shared categorical status.

Several [EU AI Act](#) articles could be strengthened through HAIG principles. Examples include: Risk Management ([Article 9](#)) could be calibrated to specific dimensional positioning—systems with higher process autonomy require more robust runtime monitoring, while greater decision authority demands enhanced oversight. Human Oversight ([Article 14](#)) should be tailored to specific trust thresholds rather than uniformly applied—systems approaching verification-to-delegation require different oversight than those near process-to-outcome. Quality Management ([Article 17](#)) could be proportional to dimensional positioning rather than using a one-size-fits-all approach. Post-Market Monitoring ([Article 72](#)) could focus resources on systems approaching critical trust thresholds, with monitoring depth reflecting specific dimensional positions.

The HAIG framework addresses critical governance challenges that categorical approaches struggle with. For foundation models that simultaneously operate across multiple risk categories, HAIG allows precise governance calibration based on specific application contexts rather than forcing classification into a single category. For evolving systems, the framework tracks incremental movement along dimensions before major thresholds are crossed, enabling preventive rather than reactive governance. For borderline cases that fall between risk categories, HAIG provides continuous governance guidance without requiring arbitrary classification. This dimensional approach helps regulators address the implementation gap between categorical rules and the complex reality of AI systems that exhibit contextual variation across applications.

Rather than viewing categorical and dimensional approaches as competing frameworks, they should be seen as complementary governance perspectives. The EU AI Act's risk-based approach provides essential regulatory clarity and baseline requirements, establishing clear boundaries for unacceptable uses and minimum standards for high-risk applications. The HAIG framework complements this foundation by enabling more precise governance calibration within established categories, addressing the complex reality of evolving human-AI relationships. This integration allows regulators to maintain clear regulatory boundaries while implementing nuanced oversight mechanisms that respond to specific dimensional positions and approaching thresholds. Together, these approaches create a governance ecosystem that combines regulatory clarity with operational precision. By enabling continuous calibration, HAIG offers modular governance scaffolding that

adjusts in response to both contextual deployments and technological evolution—something fixed-tier systems inherently struggle with.

For effective integration of these complementary approaches, we recommend: (1) Regulators should maintain categorical classifications for baseline requirements while using dimensional analysis to develop implementation guidance that addresses variations within categories; (2) Organisations should first determine their systems' risk category for compliance, then assess their specific dimensional positions to identify appropriate governance mechanisms; (3) Oversight bodies should develop monitoring approaches that track movement along HAIG dimensions, with particular attention to systems approaching critical thresholds; (4) Industry standards should integrate dimensional considerations into testing and validation protocols, particularly for high-risk applications where dimensional variation is greatest. This hybrid approach maintains regulatory clarity while enabling the governance precision needed for complex AI systems.

4.3. HAIG as a Foundation for Alternative Regulatory Approaches

The HAIG framework offers potential as an alternative regulatory foundation beyond the EU's categorical risk-based approach. Many jurisdictions remain unconvinced that rigid risk categories represent optimal regulation for AI governance, seeking instead more flexible approaches that can evolve alongside technological capabilities (e.g. [NIST AI RMF, 2023](#)).

A dimensional regulatory approach based on HAIG principles would establish graduated requirements that scale with a system's position along decision authority, process autonomy, and accountability dimensions rather than assigning fixed requirements based on application type. This creates more proportionate oversight that adapts to technological evolution without requiring frequent legislative updates.

HAIG's inherent *generalisability* stems from its modular structure, allowing incorporation of new dimensions, identification of additional thresholds, and refinement of trajectories as implementation experience grows. This modularity provides future-proofing that categorical approaches struggle to achieve.

The HAIG approach is fundamentally *trust-utility oriented* – focusing on maintaining appropriate trust relationships that maximise system utility while ensuring sufficient safeguards. This orientation balances innovation benefits against potential harms, with governance intensity proportional to trust requirements rather than predetermined risk categories.

This approach offers particular advantages for regions seeking regulatory frameworks that can adapt to rapidly evolving AI capabilities. Emerging economies and technology-focused jurisdictions may find that rigid categorisation constrains innovation, whereas a HAIG-based approach provides necessary adaptability while maintaining appropriate safeguards. By enabling governance calibration along multiple dimensions, HAIG creates space for context-sensitive regulation that responds to both technological evolution and varying deployment contexts.

While categorical frameworks provide valuable regulatory clarity, a dimensional approach like HAIG may better serve jurisdictions prioritising adaptive capacity and proportionate oversight in their AI governance strategies. For these regions, HAIG represents not merely a complement to existing frameworks but a distinct regulatory paradigm that more accurately reflects the continuous, evolving nature of human-AI relationships.

5. Discussion & Conclusions

The *trust-utility* oriented Human-AI Governance (HAIG) framework presented in this paper addresses a critical gap in current approaches to AI governance. Our dimensional analysis captures nuanced evolutionary patterns that static categorical approaches often miss, while identified trust thresholds provide practical anchors for governance adaptation at critical junctures.

HAIG offers a distinctive approach to AI governance by placing trust dynamics at its centre, providing an integrative perspective that addresses both technical capabilities and social acceptance. Unlike risk-based approaches that primarily focus on minimising negative outcomes, or principle-based approaches that emphasise normative values without clear implementation guidance, HAIG calibrates governance intensity proportional to trust requirements needed to maximise utility across different contexts.

Our analysis reveals that governance challenges emerge along two critical axes: contextual variation, where identical AI systems require different governance approaches across various applications, and technological evolution, where governance must adapt as underlying technologies progress despite serving the same functional purpose. The HAIG framework addresses this complexity through its three-dimensional approach that accommodates both dynamic contexts and evolving technologies.

The framework's dimensional structure captures human-AI relationship evolution in ways rigid categorical models cannot, provides a common language across stakeholder groups, reveals critical transition points by tracking incremental movements, and generates testable propositions about how trust mechanisms should adapt as systems shift along different dimensions. Through healthcare and EU AI Act case studies, we have demonstrated how the HAIG framework can inform both sector-specific implementations and cross-sectoral governance strategies.

Rather than rejecting categorical insights, HAIG positions them as identifying especially significant regions along continua where trust dynamics substantially shift. The framework thus complements rather than contradicts existing regulatory approaches. Concerns about implementation ambiguity are addressed through the identification of critical trust thresholds that provide practical anchors while maintaining analytical precision.

The framework benefits multiple stakeholders: public administrators gain more precise governance calibration; technologists receive insights into how capabilities interact with governance requirements; and democratic institutions gain tools to identify subtle agency redistributions that might otherwise escape notice.

Future research should focus on empirical validation through longitudinal studies that track how governance approaches evolve alongside AI capabilities. Particularly promising directions include exploring trust dynamics in multi-agent systems and investigating governance requirements when AI directly affects critical infrastructure and public services.

As jurisdictions worldwide consider their regulatory approaches to AI, the HAIG framework provides a generalisable foundation that can either complement existing categorical frameworks or serve as the basis for alternative governance paradigms that more accurately reflect the continuous, contextual nature of human-AI relationships.

Declaration of interests: None

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work, the author(s) used Claude 3.7 Sonnet and GPT-4o to help improve the presentation, refining language and structure for a diverse readership. After using these tools, the author(s) reviewed and edited all AI-assisted text and take(s) full responsibility for the content of the published article.

References

- Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018, July 10). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine* volume, 1(39). <https://doi.org/10.1038/s41746-018-0040-6>
- Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey. *IEEE Access*, 13, 18912-18936. <https://ieeexplore.ieee.org/abstract/document/10849561>
- Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F., ... Welling, M. (2020). A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, 53(8). <https://doi.org/10.1109/MC.2020.2996587>
- Alkhatib, A., & Bernstein, M. (2019, May). Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, (530), 1-13. <https://doi.org/10.1145/3290605.3300760>
- Alon-Barkat, S., & Busuioc, M. (2023, January). Human–AI Interactions in Public Sector Decision Making: “Automation Bias” and “Selective Adherence” to Algorithmic Advice. *Journal of Public Administration Research and Theory*, 33(1). <https://doi.org/10.1093/jopart/muac007>
- American Heart Organisation. (2024, February 8). *AI-based system to guide stroke treatment decisions may help prevent another stroke*. Retrieved April 21, 2025, from <https://newsroom.heart.org/news/ai-based-system-to-guide-stroke-treatment-decisions-may-help-prevent-another-stroke>
- Amnesty International & Access Now. (2018, May 16). *The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems*. Access Now. <https://www.accessnow.org/press-release/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/>
- Ansari, S., Naghdy, F., & Du, H. (2022, September). Human-Machine Shared Driving: Challenges and Future Directions. *IEEE Transactions on Intelligent Vehicles*, 7(3), 499-519. <https://ieeexplore.ieee.org/abstract/document/9721611>
- Asan, O., Bayrak, A. E., & Choudhury, A. (2020, June). Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research*, 22(6). <https://doi.org/10.2196/15154>
- Bradshaw, J. M., Hoffman, R. R., Woods, D. D., & Johnson, M. (2013). The Seven Deadly Myths of “Autonomous Systems”. *IEEE Intelligent Systems*, 28(3), 54-61. <https://doi.org/10.1109/MIS.2013.70>
- Char, D. S., Shah, N. H., & Magnus, D. (2018, March 14). Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *The New England Journal of Medicine*, 378(11). 10.1056/NEJMp1714229
- Chartrand, G., Cheng, P. M., Vorontsov, E., Drozdal, M., Turcotte, S., Pal, C. J., Kadoury, S., & Tang, A. (2017). Deep Learning: A Primer for Radiologists. *Radiographics*, 37(7), 2113-2131. <https://doi.org/10.1148/rg.2017170077>
- Cobbe, J., Lee, M. S. A., & Singh, J. (2021, March 1). Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 598 - 609. <https://doi.org/10.1145/3442188.3445921>

- Coglianese, C., & Crum, C. R. R. (2025, January). *Regulating Multifunctionality*. <https://arxiv.org/abs/2502.15715>. <https://doi.org/10.48550/arXiv.2502.15715>
- Coglianese, C., & Lehr, D. (2017). Regulating by Robot: Administrative Decision Making in the Machine-Learning Era. *Georgetown Law Journal*. https://scholarship.law.upenn.edu/faculty_scholarship/1734/
- Cugurullo, F., & Xu, Y. (2025, January). When AIs become oracles: generative artificial intelligence, anticipatory urban governance, and the future of cities. *Policy and Society*, 44(1), 98–115. <https://doi.org/10.1093/polsoc/puae025>
- Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., & Graepel, T. (2021, May 4). Cooperative AI: machines must learn to find common ground. *Nature*, 593, 33–36. <https://doi.org/10.1038/d41586-021-01170-0>
- Dennstädt, F., Hastings, J., Putora, P. M., Schmerder, M., & Cihoric, N. (2025, March 6). Implementing large language models in healthcare while balancing control, collaboration, costs and security. *npj digital medicine*, 8, 143. <https://doi.org/10.1038/s41746-025-01476-7>
- Docherty, B. L. (2012). *Losing Humanity: The Case Against Killer Robots*. Human Rights Watch. https://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf
- Donahue, K., Chouldechova, A., & Kenthapadi, K. (2022, June 20). Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness. *FACCT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1639 - 1656. <https://doi.org/10.1145/3531146.3533221>
- Dunleavy, P., & Margetts, H. (2023, September). Data science, artificial intelligence and the third wave of digital era governance. *Public Policy and Administration*, 40(2). <https://doi.org/10.1177/09520767231198737>
- Earp, B. D., Mann, S. P., Aboy, M., Awad, E., Betzler, M., Botes, M., Calcott, R., Caraccio, M., Chater, N., Coeckelbergh, M., Constantinescu, M., Dabbagh, H., Devlin, K., Ding, X., Dranseika, V., Everett, J. A. C., Fan, R., Feroz, F., Francis, K. B., ... Clark, M. S. (2025, February 17). *Relational Norms for Human-AI Cooperation*. arXiv. <https://doi.org/10.48550/arXiv.2502.12102>
- Ebers, M. (2024, November 6). Truly Risk-based Regulation of Artificial Intelligence How to Implement the EU's AI Act. *European Journal of Risk Regulation*, 1-20. <https://doi.org/10.1017/err.2024.78>
- Engin, Z., Crowcroft, J., Hand, D., & Treleaven, P. (2025, March). *The Algorithmic State Architecture (ASA): An Integrated Framework for AI-Enabled Government* [Pre-print]. ArXiv. <https://doi.org/10.48550/arXiv.2503.08725>
- Engin, Z., & Hand, D. (2025). Toward Adaptive Categories: Dimensional Governance for Agentic AI. *arXiv preprint arXiv:2505.11579*. <https://doi.org/10.48550/arXiv.2505.11579>
- Engin, Z., & Treleaven, P. (2019, March). Algorithmic Government: Automating Public Services and Supporting Civil Servants in using Data Science Technologies. *The Computer Journal*, 62(3), 448–460. <https://doi.org/10.1093/comjnl/bxy082>
- Engin, Z., van Dijk, J., Lan, T., Longley, P. A., Treleaven, P., Batty, M., & Penn, A. (2020, June). Data-driven urban management: Mapping the landscape. *Journal of Urban Management*, 9(2), 140-150. <https://doi.org/10.1016/j.jum.2019.12.001>
- Engler, A. (2023, February 21). *Early thoughts on regulating generative AI like ChatGPT*. Brookings. <https://www.brookings.edu/articles/early-thoughts-on-regulating-generative-ai-like-chatgpt/>
- European Parliament & Council of the European Union. (2024, June 13). *Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act)*. Official Journal of the European Union. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- Fleuren, L. M., Klausch, T. L. T., Zwager, C. L., Schoonmade, L. J., Guo, T., Roggeveen, L. F., Swart, E. L., Girbes, A. R. J., Thorat, P., Ercole, A., Hoogendoorn, M., & Elbers, P. W. G. (2020, January 21). Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Medicine*, 46, 383–400. <https://doi.org/10.1007/s00134-019-05872-y>
- Future of Life Institute. (n.d.). *The AI Act Explorer*. EU Artificial Intelligence Act. Retrieved April 22, 2025, from <https://artificialintelligenceact.eu/ai-act-explorer/>
- Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C., Trask, A., Hughes, E., Bergman, A. S., Shelby, R., Marchal, N., Griffin, C., ... Manyika, J. (2024, 04 19). *The Ethics of Advanced AI Assistants*. Google DeepMind. <https://doi.org/10.48550/arXiv.2404.16244>
- Gerke, S., Minssen, T., & Cohen, G. (2020, June 26). Ethical and legal challenges of artificial intelligence-driven healthcare. *Artificial Intelligence in Healthcare*, 295–336. <https://doi.org/10.1016/B978-0-12-818438-7.00012-5>
- Giger, M. L., Chan, H.-P., & Boone, J. (2008, November 20). History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM. *Medical Physics*. <https://doi.org/10.1118/1.3013555>

- Ginestra, J. C., Giannini, H. M., Schweickert, W. D., Meadows, L., Lynch, M. J., Pavan, K., Chivers, C. J., Draugelis, M., Donnelly, P. J., Fuchs, B. D., & Umscheid, C. A. (2019, November). Clinician Perception of a Machine Learning-Based Early Warning System Designed to Predict Severe Sepsis and Septic Shock. *Critical Care Medicine*, 47(11), 1477-1484. 10.1097/CCM.0000000000003803
- Goos, M., & Savona, M. (2024, April). The governance of artificial intelligence: Harnessing opportunities and mitigating challenges. *Research Policy*, 53(3). <https://doi.org/10.1016/j.respol.2023.104928>
- Grauslund, J. (2022, May 31). Diabetic retinopathy screening in the emerging era of artificial intelligence. *Diabetologia*, 65, 1415–1423. <https://doi.org/10.1007/s00125-022-05727-0>
- Gritsenko, D., & Wood, M. (2020, November). Algorithmic governance: A modes of governance approach. *Regulation & Governance*, 16(1), 45-62. <https://doi.org/10.1111/rego.12367>
- Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, 46(3). <https://doi.org/10.1136/medethics-2019-105586>
- Gsenger, R., & Strle, T. (2021). Trust, Automation Bias and Aversion: Algorithmic Decision-Making in the Context of Credit Scoring. *Interdisciplinary Description of Complex Systems : INDECS*, 19(4). <https://doi.org/10.7906/indecs.19.4.7>
- Gutierrez, G. (2020, March 24). Artificial Intelligence in the Intensive Care Unit. *Critical Care*, 24(101). <https://doi.org/10.1186/s13054-020-2785-y>
- Henry, K. E., Hager, D. N., Pronovost, P. J., & Saria, S. (2015, August 5). A targeted real-time early warning score (TREWScore) for septic shock. *Science Translational Medicine*, 7(299). <https://doi.org/10.1126/scitranslmed.aab3719>
- Horvitz, E. (n.d.). Principles of mixed-initiative user interfaces. *CHI '99: Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 159 - 166. <https://doi.org/10.1145/302979.303030>
- Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021, February 26). Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence. *Information Systems Research*, 32(3). <https://doi.org/10.1287/isre.2020.0980>
- Kaminski, M. E. (2023, September). Regulating the Risks of AI. *Boston University Law Review*, 103(5), 1347-1411. <https://heinonline.org/HOL/P?h=hein.journals/bulr103&i=1329>
- Knox, C., Wilson, M., Klinger, C. M., Franklin, M., Oler, E., Wilson, A., Pon, A., Cox, J., Chin, N. E. (., Strawbridge, S. A., Garcia-Patino, M., Kruger, R., Sivakumaran, A., Sanford, S., Doshi, R., Khetarpal, N., Fatokun, O., Doucet, D., Zubkowski, A., ... Wishart, D. S. (2024, January 5). DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Research*, 52(D1), D1265–D1275. <https://doi.org/10.1093/nar/gkad976>
- Koehler, M., & Sauermann, H. (2024, May). Algorithmic management in scientific research. *Research Policy*, 53(4). <https://doi.org/10.1016/j.respol.2024.104985>
- Kolt, N. (2025). *Governing AI Agents*. arXiv. <https://doi.org/10.48550/arXiv.2501.07913>
- Kroll, J. A. (2015, September). *Accountable Algorithms* [PhD Dissertation]. Princeton University. <https://www.proquest.com/dissertations-theses/accountable-algorithms/docview/1754381941/se-2?accountid=14511>
- Kusche, I. (2024, May 11). Possible harms of artificial intelligence and the EU AI act: fundamental rights and risk. *Journal of Risk Research*, 1–14. <https://doi.org/10.1080/13669877.2024.2350720>
- Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., & Tan, C. (2023, June 12). Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. *FACCT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. <https://doi.org/10.1145/3593013.3594087>
- Laux, J., Wachter, S., & Mittelstadt, B. (2024, January). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1), 3-32. <https://doi.org/10.1111/rego.12512>
- Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1). https://doi.org/10.1518/hfes.46.1.50_30392
- Lee, Y., Ferber, D., Rood, J. E., Regev, A., & Kather, J. N. (2024, December 17). How AI agents will change cancer research and oncology. *Nature Cancer volume*, 5, 1765–1767. <https://doi.org/10.1038/s43018-024-00861-7>
- Liu, B. (2021, November). In AI We Trust? Effects of Agency Locus and Transparency on Uncertainty Reduction in Human–AI Interaction. *Journal of Computer-Mediated Communication*, 26(6). <https://doi.org/10.1093/jcmc/zmac013>
- Liu, B., Li, X., Zhang, J., Wang, J., He, T., Hong, S., Liu, H., Zhang, S., Song, K., Zhu, K., Cheng, Y., Wang, S., Wang, X., Luo, Y., Jin, H., Zhang, P., Liu, O., Chen, J., Zhang, H., ... Wu, C. (2025, March 31). *Advances and Challenges in Foundation Agents: From Brain-Inspired Intelligence to Evolutionary, Collaborative, and Safe Systems*. arXiv. <https://doi.org/10.48550/arXiv.2504.01990>

- Liu, X., Lou, X., Jiao, J., & Zhang, J. (2024, July 21). Position: foundation agents as the paradigm shift for decision making. *ICML'24: Proceedings of the 41st International Conference on Machine Learning*, (1276), 31597 - 31613. <https://dl.acm.org/doi/10.5555/3692070.3693346>
- Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, 46(4), 629–650. <https://doi.org/10.1093/jcr/ucz013>
- Luusua, A., Ylipulli, J., Foth, M., & Aurigi, A. (2023). Urban AI: understanding the emerging role of artificial intelligence in smart cities. *AI & Society*, 38, 1039–1044. <https://doi.org/10.1007/s00146-022-01537-5>
- Lyell, D., & Coiera, E. (2017, March 1). Automation bias and verification complexity: a systematic review. *Journal of American Medical Informatics Association*, 24(2), 423–431. <https://doi.org/10.1093/jamia/ocw105>
- McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafi, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., ... Shetty, S. (2020, January 1). International evaluation of an AI system for breast cancer screening. *Nature*, 577, 89–94. <https://doi.org/10.1038/s41586-019-1799-6>
- Methnani, L., Tubella, A. A., Dignum, V., & Theodorou, A. (2021). Let Me Take Over: Variable Autonomy for Meaningful Human Control. *Frontiers in Artificial Intelligence*, 4. <https://doi.org/10.3389/frai.2021.737072>
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019, January 29). Model Cards for Model Reporting. *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220 - 229. <https://doi.org/10.1145/3287560.3287596>
- Mittelstadt, B. (2019, November 4). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1, 501–507. <https://doi.org/10.1038/s42256-019-0114-4>
- Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2024). Auditing large language models: a three-layered approach. *AI and Ethics*, 4, 1085–1115. <https://doi.org/10.1007/s43681-023-00289-2>
- Moritz, M., Topol, E., & Rajpurkar, P. (2025, April 1). Coordinated AI agents for advancing healthcare. *Nature Biomedical Engineering*, 9, 432–438. <https://doi.org/10.1038/s41551-025-01363-2>
- NIST (2023) NIST AI Risk Management Framework (AI RMF). <https://www.nist.gov/itl/ai-risk-management-framework>
- Papadopoulos, P., Soflano, M., Chaudy, Y., Adejo, W., & Connolly, T. M. (2022, May 27). A systematic review of technologies and standards used in the development of rule-based clinical decision support systems. *Health and Technology*, 12, 713–727. <https://doi.org/10.1007/s12553-022-00672-9>
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000, May). A Model for Types and Levels of Human Interaction with Automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3), 286–297. [10.1109/3468.844354](https://doi.org/10.1109/3468.844354)
- Park, S. H., Choi, J., & Byeon, J.-S. (2021, February 10). Key Principles of Clinical Validation, Device Approval, and Insurance Coverage Decisions of Artificial Intelligence. *Korean Journal of Radiology*, 22(3), 442–453. <https://doi.org/10.3348/kjr.2021.0048>
- Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press. <https://www.jstor.org/stable/j.ctt13x0hch>
- Pedreschi, D., Pappalardo, L., Ferragina, E., Baeza-Yates, R., Barabási, A.-L., Dignum, F., Dignum, V., Eliassi-Rad, T., Giannotti, F., Kertész, J., Knott, A., Ioannidis, Y., Lukowicz, P., Passarella, A., Pentland, A. S., Shawe-Taylor, J., & Vespignani, A. (2025, February). Human-AI coevolution. *Artificial Intelligence*, 339. <https://doi.org/10.1016/j.artint.2024.104244>
- Qiu, J., Lam, K., Li, G., Acharya, A., Wong, T. Y., Darzi, A., Yuan, W., & Topol, E. J. (2024, December 5). LLM-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6, 1418–1420. <https://doi.org/10.1038/s42256-024-00944-1>
- Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*, 20, 5–14. <https://doi.org/10.1007/s10676-017-9430-8>
- Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. S., ... Wellman, M. (2019, April 24). Machine behaviour. *Nature*, 568, 477–486. <https://www.nature.com/articles/s41586-019-1138-y>
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020, January 27). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33 - 44. <https://doi.org/10.1145/3351095.3372873>

- Rijsbosch, B., van Dijck, G., & Kollnig, K. (2025, March). *Adoption of Watermarking for Generative AI Systems in Practice and Implications under the new EU AI Act*. ArXiv. <https://doi.org/10.48550/arXiv.2503.18156>
- Rudko, I., Bonab, A. B., Fedele, M., & Formisano, A. V. (2024, April 16). New institutional theory and AI: toward rethinking of artificial intelligence in organizations. *Journal of Management History*, 31(2). https://www.emerald.com/insight/content/doi/10.1108/jmh-09-2023-0097/full/html?casa_token=zdZv_iW32AsAAAAA%3Ax2EemyvykeXsuqnyZkT8S-0MeroOAudCmZgPaQU5Cq_VZXRdpqbLaWWbhMrRI5w4L0QLV4XlotViaR5Vw1FEjs7emVI2dEIEomBNqc3E0eOgD95ODyJS
- Schuett, J. (2023). *Towards risk-based AI regulation* [PhD Thesis]. Universitätsbibliothek Johann Christian Senckenberg. <https://doi.org/10.21248/gups.89768>
- Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O’Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., Slama, K., Ahmad, L., McMillan, P., Beutel, A., Passos, A., & Robinson, D. G. (2023, December 14). *Practices for Governing Agentic AI Systems*. OpenAI. <https://openai.com/index/practices-for-governing-agentic-ai-systems/>
- Shortliffe, E. H., & Sepúlveda, M. J. (2018, December 4). Clinical Decision Support in the Era of Artificial Intelligence. *JAMA*, 320(21), 2199-2200. doi:10.1001/jama.2018.17163
- Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019, July 13). Organizational Decision-Making Structures in the Age of Artificial Intelligence. *California Management Review*, 61(4), 66-83. <https://doi.org/10.1177/0008125619862257>
- Singh, A., & Szajnfarder, Z. (2024, August). *Preposition Salad: Making Sense of Human-in/on/over-the-Loop Control for AI Systems*. SSRN. <https://dx.doi.org/10.2139/ssrn.4921359>
- Smuha, N. A. (2021). Beyond a Human Rights-Based Approach to AI Governance: Promise, Pitfalls, Plea. *Philosophy & Technology*, 34, 91–104. <https://doi.org/10.1007/s13347-020-00403-w>
- Sowa, K., & Przegalinska, A. (2025, April 3). From Expert Systems to Generative Artificial Experts: A New Concept for Human-AI Collaboration in Knowledge Work. 82. <https://doi.org/10.1613/jair.1.17175>
- Steiner, D. F., Chen, P.-H. C., & Mermel, C. H. (2021, January). Closing the translation gap: AI applications in digital pathology. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1875(1). <https://doi.org/10.1016/j.bbcan.2020.188452>
- Stephany, F., & Teutloff, O. (2024, January). What is the price of a skill? The value of complementarity. *Research Policy*, 53(1). <https://doi.org/10.1016/j.respol.2023.104898>
- Stokel, C. (2025, March 13). Revealed: How the UK tech secretary uses ChatGPT for policy advice. *New Scientist*. <https://www.newscientist.com/article/2472068-revealed-how-the-uk-tech-secretary-uses-chatgpt-for-policy-advice/>
- Sun, H. (2025). Bridging the Accountability Gap in AI Decision-Making: An Integrated Analysis of Legal Precedents and Scholarly Perspectives. *Frontiers in Humanities and Social Sciences*, 5(3). <https://doi.org/10.54691/24vexn67>
- Thieme, A., Nori, A. V., Ghassemi, M., Bommasani, R., Andersen, T. O., & Luger, E. A. (2023, April 19). Foundation Models in Healthcare: Opportunities, Risks & Strategies Forward. *CHI EA '23: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, (512), 1-4. <https://doi.org/10.1145/3544549.3583177>
- Tolmeijer, S., Christen, M., Kandul, S., Kneer, M., & Bernstein, A. (2022, April). Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making. *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, (160), 1-17. <https://doi.org/10.1145/3491102.3517732>
- Toreini, E., Aitken, M., Coopamootoo, K. P. L., Elliott, K., Zelaya, C. G., & van Moorsel, A. P. A. (2020, January 27). The relationship between trust in AI and trustworthy machine learning technologies. *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 272 - 283. <https://doi.org/10.1145/3351095.3372834>
- United Nations. (2024, August 20). *Mapping report: human rights and new and emerging digital technologies* [Report of the Office of the United Nations High Commissioner for Human Rights]. United nations. <https://www.ohchr.org/en/documents/thematic-reports/ahrc5645-mapping-report-human-rights-and-new-and-emerging-digital>
- Vagia, M., Transeth, A. A., & Fjerdingen, S. A. (2016, March). A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? *Applied Ergonomics*, 53(A), 190-202. <https://doi.org/10.1016/j.apergo.2015.09.013>
- van Breda, L., & Barry, T. (Eds.). (2020). *Human-Autonomy Teaming: Supporting Dynamically Adjustable Collaboration*. NATO / OTAN - North Atlantic Treaty Organization. <https://apps.dtic.mil/sti/citations/AD1183655>

- Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(186345). <https://doi.org/10.1007/s11704-024-40231-1>
- Wang, S., & Summers, R. M. (2012, July). Machine learning and radiology. *Medical Image Analysis*, 16(5), 933-951. <https://doi.org/10.1016/j.media.2012.02.005>
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022, August 31). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=yzkSU5zdwD>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. ArXiv. <https://doi.org/10.48550/arXiv.2201.11903>
- Winfield, A. F. T., & Jirotko, M. (2018, October 15). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A*, 376(2133). <https://doi.org/10.1098/rsta.2018.0085>
- Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., ... Gui, T. (2025, January 17). The rise and potential of large language model based agents: a survey. *Science China Information Sciences*, 68(121101). <https://doi.org/10.1007/s11432-024-4222-0>
- Yang, S., Nachum, O., Du, Y., Wei, J., Abbeel, P., & Schuurmans, D. (2023). *Foundation Models for Decision Making: Problems, Methods, and Opportunities*. arXiv. <https://doi.org/10.48550/arXiv.2303.04129>
- Yelne, S., Chaudhary, M., Dod, K., Sayyad, A., & Sharma, R. (2023, November). Harnessing the Power of AI: A Comprehensive Review of Its Impact and Challenges in Nursing Science and Healthcare. *Cureus*, 22;15(11) : e49252. <https://doi.org/10.7759/cureus.49252>

Appendix: Sectorial Relevance of the HAIG Approach for Analysing Trust Dynamics

Several implementation domains would particularly benefit from a continua approach to human-AI relationships rather than strict categorical frameworks (see Table 2). These sectors share characteristics that make the nuanced perspective of dimensions and continua especially valuable. From public services balancing democratic accountability with efficient delivery, to healthcare systems navigating graduated clinical authority, to financial services adapting to dynamic risk thresholds, each domain illustrates how AI governance exists along evolving spectrums rather than in fixed states.

The HAIG approach particularly valuable for sectors/applications with : high consequence decisions with varying stakes across contexts, professional judgement requirements that cannot be fully formalised, complex regulatory environments with proportionality requirements, strong ethical frameworks that already recognise degrees of responsibility, and risk profiles that vary across specific instances rather than fitting neat categories.

In these domains, the HAIG approach captures the nuanced reality of how human-AI relationships actually function while providing more precise governance guidance than categorical frameworks alone could offer. These sector-specific applications provide valuable comparative examples, demonstrating that the multidimensional HAIG framework has broad applicability across domains with similar characteristics.

Table 3: Sectorial Relevance of the HAIG Approach for Analysing Trust Dynamics

Public Sector & Government Services	<ul style="list-style-type: none"> • Policy implementation spectrum: Government services span from purely administrative functions to complex policy judgments requiring different degrees of human oversight • Stakeholder accountability: Different public functions have varying levels of direct impact on citizens, requiring proportional human involvement • Democratic legitimacy considerations: Different governance functions require different balances of efficiency and direct accountability • Cross-agency collaboration: Public sector AI spans jurisdictional boundaries with varying authority distributions • Transparency requirements: Different public functions have varying transparency needs that align with different points on the continuum
Healthcare & Clinical Decision Support	<ul style="list-style-type: none"> • Graduated clinical authority: Medical AI spans from simple reference tools to diagnostic partners with varying degrees of authority • Patient-specific risk levels: Appropriate AI autonomy varies based on patient risk factors rather than fitting neat categories • Specialty-specific requirements: Different medical specialties require different human-AI relationships • Professional ethics considerations: Medical ethics frameworks address degrees of responsibility rather than binary states
Financial Services & Risk Management	<ul style="list-style-type: none"> • Graduated risk thresholds: Different transaction values and risk profiles require different levels of human oversight • Client-specific automation levels: Appropriate automation levels vary based on client sophistication • Market condition adaptivity: Human-AI relationships need to adjust with changing market volatility • Regulatory proportionality: Financial regulations increasingly apply proportionate controls based on risk levels
Legal Services & Judicial Processes	<ul style="list-style-type: none"> • Jurisdictional variations: Different legal systems permit different levels of automation • Case complexity spectrum: Simple cases may permit more automation than complex ones • Stakes-based oversight: Higher-stakes matters require different human-AI relationships than routine matters • Legal authority gradations: Legal authority exists along a spectrum rather than in discrete categories
Education & Learning Assessment	<ul style="list-style-type: none"> • Developmental appropriateness: Different student ages and stages require different levels of human involvement • Subject-specific variations: Some subjects permit more automation than others • Pedagogical diversity: Different teaching philosophies align with different points on the human-AI continuum • Stakes-based considerations: High-stakes assessment requires different human-AI relationships than practice activities
Critical Infrastructure Management	<ul style="list-style-type: none"> • Graduated criticality levels: Different systems and components have varying criticality requiring different oversight • Temporal variations: Appropriate automation levels change during normal operations versus emergencies • Physical-digital integration: Systems span from purely advisory to having direct control of physical components • Cascading consequence management: Oversight needs scale with potential for cascading failures
Research & Scientific Discovery	<ul style="list-style-type: none"> • Methodological diversity: Different research methods permit different levels of automation • Discovery-validation spectrum: Early discovery phases may permit more AI agency than validation phases • Field-specific standards: Different disciplines have different standards for evidence and verification • Risk-reward calculations: Higher-risk research domains require different human-AI relationships