Research Article

Expectation Inflation of Artificial Intelligence: When Human Demands on AI Exceed Moral Design Capacity

Victor Frimpong[1]

1. Swiss Business School Zurich (SBS), Switzerland

What happens when the ethical and interpretive demands placed on artificial intelligence (AI) exceed what AI systems can legitimately support? This paper develops Expectation Inflation as a conceptual framework for explaining how AI-related ethical risk can arise not only from technical failures but also from escalating human and organisational expectations and the over-delegation of moral and interpretive authority. Adopting a conceptual and normative-analytical approach, the study synthesises interdisciplinary literature and constructs a theoretical model rather than seeking empirical validation. Drawing on Science and Technology Studies (STS), anthropomorphism, automation reliance, algorithmic authority, human–AI collaboration, and AI ethics, the paper conceptualises Expectation Inflation as an escalation process in which functional confidence can develop into epistemic and ultimately moral expectations. The Expectation Inflation Curve captures this progression through three stages: Rational Delegation, Normative Drift, and Moral Substitution. The paper further introduces Moral Design Capacity (MDC) as a context-dependent normative boundary beyond which the moral and interpretive authority attributed to an AI system becomes disproportionate to its demonstrated capabilities, decision context, and surrounding governance safeguards. To address this misalignment, the paper proposes Expectation Governance, comprising Delegation Boundaries, Expectation Auditing, and Moral Accountability Indexing. The framework complements existing approaches to responsible AI by shifting analytical attention from system-side properties alone to the demand side of AI governance: how organisations construct, amplify, and institutionalise expectations about what AI should know, interpret, and morally decide. The study thereby provides a theoretical basis for examining expectation–capacity misalignment in organisational and institutional settings and establishes propositions and governance mechanisms for subsequent empirical investigation.

 doi.org/10.32388/2SUYHM.2

Corresponding author: Victor Frimpong, v.frimpong@research.sbs.edu

1. Introduction – The Inflationary Logic of Expectation

Expectation Inflation occurs when human demands on artificial intelligence (AI) exceed its moral and epistemic capabilities. This phenomenon happens when leaders and societies assign interpretive, ethical, and emotional roles to AI systems that they were never designed to fulfil. As a result, the perceived potential of AI is exaggerated, while human judgment and accountability are undervalued. The concept uses an economic metaphor: just as excess currency diminishes its value, inflated expectations can undermine the credibility of AI systems and compromise the moral authority of human decision-makers[1].

Expectation inflation extends beyond market hype or technological optimism; it operates at both cognitive and institutional levels[2][3]. It stems from unrealistic narratives about AI's potential and the overvaluation of algorithms as reliable decision-makers. Organisations seeking efficiency and legitimacy often expect moral qualities—such as fairness and empathy—to be integrated into these statistical systems[4][5]. This gap between what humans expect and what machines can actually deliver is the focus of this paper.

The inflationary mechanism operates subtly but builds up over time. It begins with small tasks—appointing AI to enhance processes, assess risks, or categorise data. As these tasks become routine, the outputs are viewed not only as efficient but also as accurate. With this growing reliance on algorithms, leaders may overestimate the moral quality of the code[6]. This shift, from simply trusting the tool to having faith in its moral judgment, signals a significant change in which the machine evolves from a tool to a co-agent in knowledge[7][8]. The risk then lies not in technical flaws but in the over-delegation of moral judgment, as authority quietly shifts from humans to systems.

This phenomenon differs from Elish's[9] concept of the moral crumple zone, where humans take on blame for failures in automation. Expectation inflation flips this dynamic as people preemptively relinquish moral responsibility by overtrusting machine decisions. This anticipatory inflation expands the expected moral roles of AI before it has demonstrated its capabilities. Each success of AI reinforces its authority, creating a feedback loop that displaces moral agency without replacing it[10]. Ultimately, this leads to ethical displacement (Frimpong, 2025b), in which moral responsibility becomes a secondary attribute of leadership.

This paper connects Science and Technology Studies (STS) with AI governance, focusing on how societal expectations of technology influence institutional design and ethical issues. It emphasises that technologies are not neutral; they are shaped by and shape the social contexts in which they exist. By viewing expectations as moral projections, the paper links STS, behavioural economics, and AI ethics, highlighting that reliance on AI often serves as moral outsourcing rather than merely a technical improvement.

This paper is a conceptual theory-development study that proposes an explanatory framework for understanding how human and organisational expectations of AI can exceed the system's actual capabilities. Using a problematisation approach[11], it synthesises insights from the sociology of expectations, anthropomorphism, reliance on automation, algorithmic authority, and human–AI collaboration to identify a fragmented theoretical mechanism. The framework describes the progression from Rational Delegation through Normative Drift to Moral Substitution, introduces Moral Design Capacity as a boundary condition, and proposes concepts for empirical testing. Philosophical ideas on moral agency and responsibility are used as resources rather than as the main focus.

Accordingly, the paper considers the following question:

Why and how do human expectations exceed AI's moral design capacity — and how can governance restore equilibrium?

The research primarily focuses on the organisational and institutional levels, where expectations about AI influence practices such as delegation, decision-making, accountability, and governance. While individual factors such as trust and anthropomorphism may shape expectation formation, the main concern is how these expectations are magnified and embedded within organisations. The study does not aim to prove a direct causal link between expectations and AI outcomes. Instead, it offers a theoretical framework for explaining how expectations can exceed Moral Design Capacity and suggests governance structures to maintain a balance between human expectations and AI delegation.

2. Literature Review

Human behaviour is shaped by expectations—forecasts that influence perception, decision-making, and social coordination[12][1]. While expectations help us make sense of uncertainty, they can also lead to systematic errors when divorced from reality. In economics, expectations drive market behaviour[13], whereas in social contexts, they function as a cultural and institutional influence on collective action[14].

doi.org/10.32388/2SUYHM.2

Thus, expectations serve not just as cognitive by-products but as essential mechanisms for stabilising the future in modern rationality.

Within the Science and Technology Studies (STS) framework, the "sociology of expectations" addresses how emerging technologies thrive on promises and actual performance[3][2]. Expectations function as performative scripts that draw investments, shape regulations, and build public legitimacy, often before any results are seen. This leads to cycles of hype and disappointment, where overhyped technologies fail to meet expectations[15]. Artificial intelligence is currently a prime example of this pattern.

This inflationary dynamic stems from behavioural overconfidence, as described by Kahneman[12]. People often overestimate the reliability of their own judgement and the competence of seemingly consistent systems. When algorithmic outputs confirm existing beliefs or desired outcomes, users develop an illusion of validity, mistaking procedural accuracy for moral correctness. This cognitive bias fuels increased trust, driving Expectation Inflation.

In the AI field, expectations form at three levels. First, cognitive expectations stem from human tendencies to anthropomorphise and infer intention where none exists[12]. Second, institutional expectations arise from the need for organisations to demonstrate technological modernity, efficiency, and neutrality[6]. Third, normative expectations include moral judgements that view AI systems as fair and empathetic[16][8]. Together, these factors contribute to what Zuboff[5] calls algorithmic illusion—the false belief that data-driven systems can eliminate human bias while often reinforcing it.

This progression marks an important shift: from seeing expectation as a prediction to viewing it as a moral obligation. When organisations link technical performance to ethical standards, delegating decisions to AI can become a way of outsourcing morality. Floridi[10] cautions that this "distributed morality" weakens intentionality within technical systems, while Elish[9] illustrates how humans become scapegoats for AI mistakes. This paper introduces the concept of expectation inflation, emphasising that moral trust in systems increases despite their limited capacity for reflection. While current AI ethics discussions often address bias, accountability, and explainability, this paper reframes the issue as one of over-delegation—in which human moral expectations exceed AI's ethical capabilities, disrupting the ethical balance.

2.1. From Algorithmic Authority to Moral Governance

The spread of AI in decision-making systems has changed the structure of authority. Algorithms are now active partners that influence how we understand truth, value, and fairness[17][6]. As noted by Pasquale[7], this shift has led to a new form of robotics that replaces human expertise with automated consistency, altering both the distribution of labour and judgement. Floridi[10] describes this change as distributed morality, in which ethical actions are distributed across sociotechnical networks. In these systems, moral responsibility is unclear and relies on procedural compliance instead of thoughtful consideration.

AI governance frameworks focus on transparency, fairness, and accountability[16][8][18]. However, they often concentrate on controlling algorithms rather than addressing human expectations and needs[19]. This reactive approach to ethics ignores the significant influence of cognitive, institutional, and moral pressures that cause people to over-rely on algorithms for decision-making. The key ethical risk lies not in unclear algorithms but in the belief that precise procedures can replace human judgement, empathy, and prudence.

From a leadership and organisational standpoint, this misalignment is significant. Classical management theory viewed leadership as making judgements in uncertain situations[20]. In AI-driven environments, judgement is shifting towards technical system configuration rather than moral decision-making. Leaders are shifting from interpreting values to managing processes. As accountability shifts to machines, organisations face what Frimpong (2025b) calls responsibility liquidity—where accountability is unclear and lacks a specific point of contact. The moral implications include reduced oversight and a decline in what Floridi[10] calls semantic control—the ability to define the meaning of outcomes.

In Science and Technology Studies (STS), this situation highlights a broader issue: algorithmic authority —technical systems gaining legitimacy through strict procedures rather than moral considerations[21][19]. The key ethical question now is not whether algorithms are fair, but how much moral responsibility we place on them. This paper advocates for AI ethics to shift from mere compliance with rules to managing expectations: an anticipatory approach that addresses moral demands before they escalate.

Section 3 presents the Expectation Inflation model, detailing how human expectations grow through three stages—Rational Delegation, Normative Drift, and Moral Substitution—until they surpass the system's Moral Design Capacity. This escalation is illustrated in the Expectation Inflation Curve, which helps clarify how moral balance in AI systems can become destabilised.

doi.org/10.32388/2SUYHM.2

2.2. Positioning Expectation Inflation Within Adjacent Literatures

Expectation Inflation intersects with various discussions, offering a unique perspective on demand escalation. It explains how the use of AI can shift human expectations from basic performance to epistemic authority and moral sufficiency. The focus shifts from AI capabilities to the ways users and organisations increasingly expect systems to exercise authority.

The concept relates to anthropomorphism, which involves attributing human-like intentions, emotions, or mental states to non-human agents[22]. Anthropomorphism can lead to Expectation Inflation by making AI seem more understanding and empathetic, but understanding and empathy are not the same. Anthropomorphism is a psychological process, whereas Expectation Inflation is a socio-organisational process in which expectations grow through repeated use, endorsement, and delegation. Expectation Inflation can occur without human-like personification, such as when an organisation treats a reliable risk model as authoritative because of its performance or perceived objectivity.

Expectation Inflation further expands upon research concerning automation bias, overtrust, and algorithm appreciation. Studies in automation research indicate that excessive trust may lead to misuse or overreliance on automated aids[23][24], whereas research on algorithmic appreciation suggests that individuals may sometimes assign greater weight to algorithmic advice than to human counsel[25]. These bodies of literature elucidate the circumstances under which users defer to automated recommendations and the reasons behind such behaviour. Expectation Inflation addresses an additional inquiry: how does repeated functional deference evolve into a broader normative expectation that the system possesses qualities such as fairness, interpretative ability, empathy, or moral judgment? In this context, overtrust may serve as a behavioural manifestation of Expectation Inflation; however, the framework posits an escalation of the object of trust—from technical reliability to moral and interpretive authority.

The framework is also linked to research concerning algorithmic authority and the sociology of expectations. Investigations into algorithmic authority analyse how computational systems attain legitimacy and influence classifications, visibility, and decision-making processes[17][6]. Additionally, Science, Technology, and Society (STS) scholarship demonstrates that expectations regarding emerging technologies are performative: promises and envisioned futures stimulate investment, regulation, and institutional commitment prior to the full stabilisation of capabilities[3][15][2]. The concept of Expectation Inflation expands upon this insight by identifying a moral mechanism within it. The pertinent expectation is not solely that artificial intelligence will become more competent; rather, it is that

doi.org/10.32388/2SUYHM.2

functional success is translated into assertions about what AI can legitimately comprehend, judge, or decide. Consequently, the concept reinterprets technological expectations as potential transfers of moral authority.

Finally, Expectation Inflation contributes to the scholarly discourse on human–AI collaboration and hybrid decision-making. Research on human–AI symbiosis and the automation–augmentation paradox emphasises complementarity, task allocation, and the ongoing significance of human judgment under conditions of uncertainty and equivocality[26][27]. The present framework delineates a governance condition essential for maintaining this complementarity: collaboration becomes ethically unstable when the perceived authority of AI expands more rapidly than its Moral Design Capacity and human participants cease to retain meaningful interpretive ownership. Consequently, Expectation Governance supplements responsible AI and algorithmic accountability strategies. Rather than supplanting controls related to fairness, transparency, explainability, auditing, or accountability, it introduces an antecedent layer of governance focused on calibrating organisational expectations of AI before inflated expectations become entrenched in delegation practices.

The paper should not be viewed as merely relabelling anthropomorphism or automation bias, or as a replacement for existing AI-governance frameworks. It introduces a cross-level mechanism linking cognitive attribution, behavioural reliance, institutional legitimacy, and moral delegation.

Its novelty lies in explaining the transition through Rational Delegation, Normative Drift, and Moral Substitution, and in defining Moral Design Capacity as the point where expanded expectations become ethically disproportionate.

Table 1 situates Expectation Inflation within key scholarly discussions on human–AI relationships, technological expectations, and decision-making. While these works explain anthropomorphism, automation bias, algorithmic authority, collaboration, and governance, they explore different facets of the broader phenomenon. The comparison shows how Expectation Inflation builds on existing research and highlights its unique contribution: explaining how expectations can escalate from functional reliance to epistemic and moral demands that exceed an AI system's Moral Design Capacity.

doi.org/10.32388/2SUYHM.2

Adjacent literature	Key references	Relationship to Expectation Inflation
Anthropomorphism & human–technology interaction	Epley et al.[22]; Turkle[28]; Bryson[29]	Explains attribution of human qualities to AI. Expectation Inflation extends this to the escalation of moral and epistemic expectations, including where anthropomorphism is absent.
Automation bias & overtrust	Parasuraman & Riley[23]; Dzindolet et al.[24]	Explains excessive reliance on automation. Expectation Inflation explains how reliance can progress from functional trust to moral over-delegation.
Algorithmic authority & decision-making	Gillespie[17]; Beer[6]; Logg et al.[25]; Pasquale[7]	Explains how algorithms acquire decision authority. Expectation Inflation focuses on its escalation towards perceived moral authority.
Sociology of expectations (STS)	Brown & Michael[15]; Borup et al.[3]; van Lente[2]	Explains the performative power of technological expectations. Expectation Inflation extends this logic to moral expectations, delegation, and accountability.
Human–AI collaboration	Jarrahi[26]; Elish[9]; Raisch & Krakowski[27]	Examines human–AI complementarity and responsibility. Expectation Inflation identifies when collaboration becomes unstable because expectations exceed Moral Design Capacity.
Responsible AI & governance	Mittelstadt et al.[8]; Floridi & Cowls[16]; Selbst et al.[19]; Raji et al.[30]	Focuses mainly on governing AI systems and outcomes. Expectation Inflation adds demand-side governance of human and organisational expectations.

Table 1. Positioning Expectation Inflation Within Adjacent Literatures

doi.org/10.32388/2SUYHM.2

Expectation Inflation differs from related theories in that it focuses on how human expectations of AI can expand. Automation bias explains the overreliance on automated outputs; algorithmic authority details how computational systems gain legitimacy; distributed morality describes moral actions spread across sociotechnical systems; and the moral crumple zone shows how humans absorb responsibility when systems fail. These perspectives highlight key aspects of human–AI relations but do not specifically address the expansion of expectations.

The concept addresses this gap by theorising an escalation mechanism through which demonstrated functional performance can generate epistemic confidence and, ultimately, expectations of moral competence. Its distinctive analytical dimension is, therefore, the expectation–capacity relationship: ethical risk arises when the normative demands placed on AI exceed its Moral Design Capacity, even when the system continues to operate technically as intended. This shifts the focus in AI governance from system-side properties alone—such as fairness, transparency, explainability, and accountability—to the demand side of governance: how individuals and organisations construct, amplify, and institutionalise expectations regarding what AI should know, interpret, and morally decide. Expectation Inflation thus complements rather than substitutes existing theories by linking reliance, authority, and moral delegation within a dynamic model of expectation escalation.

3. Conceptual Model — The Expectation Inflation Curve

Building on the theoretical streams synthesised in Section 2, Expectation Inflation, as a conceptual model, is developed as the phenomenon in which human demands for morality and interpretation outpace AI systems' ability to handle them effectively. While Section 2 explored the evolution of expectations from cognitive forecasting to moral projection, this section presents a structured model of this escalation, detailing identifiable phases. It proposes a dynamic framework that connects expectation intensity, delegation behaviour, and the moral design capacity of AI.

Expectation Inflation is a feedback loop in which humans delegate tasks to AI to improve efficiency and accuracy. As the AI delivers successful outcomes, the relationship between users and the system shifts. Each positive result reinforces the belief that the AI is not only reliable but also correct. This leads to performance being viewed as authority and ultimately builds trust. This transformation—trust without questioning—fuels the inflationary cycle.

 doi.org/10.32388/2SUYHM.2

Expectation Inflation goes beyond mere optimism; it signals a moral imbalance in how AI is created and used. When people see AI as more morally capable than it truly is—which is indicated by its Moral Design Capacity (MDC)—the system becomes unstable. This mismatch between expectations and real ability causes moral volatility: responsibility lessens, control over interpretations weakens, and the importance of human judgement declines within organisations and society.

This process is modelled through three sequential phases:

1. **Rational Delegation**: The initial, performance-based trust where human oversight remains active. Rational delegation is grounded in scholarly research on appropriate reliance on automation and the complementary roles of humans and artificial intelligence. At this juncture, delegation remains a practical approach: artificial intelligence is utilised for attributes such as speed, consistency, or analytical processing, whilst humans maintain interpretative authority and accountability. This aligns with scholarly perspectives on appropriate automation use[23] and the augmentation of human intelligence by artificial systems, in which technological capabilities complement rather than replace human judgement[26][27]. Consequently, Rational Delegation signifies a state of calibrated trust wherein confidence is generally commensurate with the demonstrated capabilities of the system.
2. **Normative Drift:** The expansion of moral confidence beyond technical evidence, where AI begins to function as an authority. It begins when confidence based on performance extends beyond confirmed functionality. Studies on automation bias and reliance on algorithms show that repeated dependence on automated suggestions can diminish critical evaluation and lead to greater trust in system outputs[23][24]. Similarly, research on algorithmic authority indicates that computational systems gain legitimacy through their perceived consistency, objectivity, and procedural rationality[17][6]. From an STS viewpoint, expectations are also performative: technological promises can influence behaviour and institutional structures before capabilities are fully realised[3]. Normative Drift describes when these dynamics expand the concept of successful performance: an AI system that consistently appears reliable is increasingly seen not just as useful, but as more authoritative, objective, or fair.
3. **Moral Substitution: F**ull moral transfer, in which human agents treat AI as a co-judging or morally sufficient entity. Moral Substitution is the advanced stage in which epistemic authority shifts to moral expectations. Research on anthropomorphism explains how people attribute human-like qualities to non-human agents[22]. Studies on distributed morality and human–machine systems

show how agency and responsibility disperse across sociotechnical systems[10][9]. Expectation Inflation describes the escalation where AI is seen as capable of providing morally sufficient judgement, not just information. Moral Substitution occurs when humans delegate decision-making and normative interpretation to systems lacking moral agency.

The three stages combine separate theories into a timeline: research on reliance explains Rational Delegation; automation bias, algorithmic authority, and performative expectations show the shift towards Normative Drift; and anthropomorphism and shared responsibility clarify Moral Substitution. The Expectation Inflation Curve (Figure 1) links these through an escalation in which expectations move from AI's capabilities to its knowledge or interpretation, and finally to moral judgement. Expectation Inflation happens when this escalation exceeds the system's Moral Design Capacity.

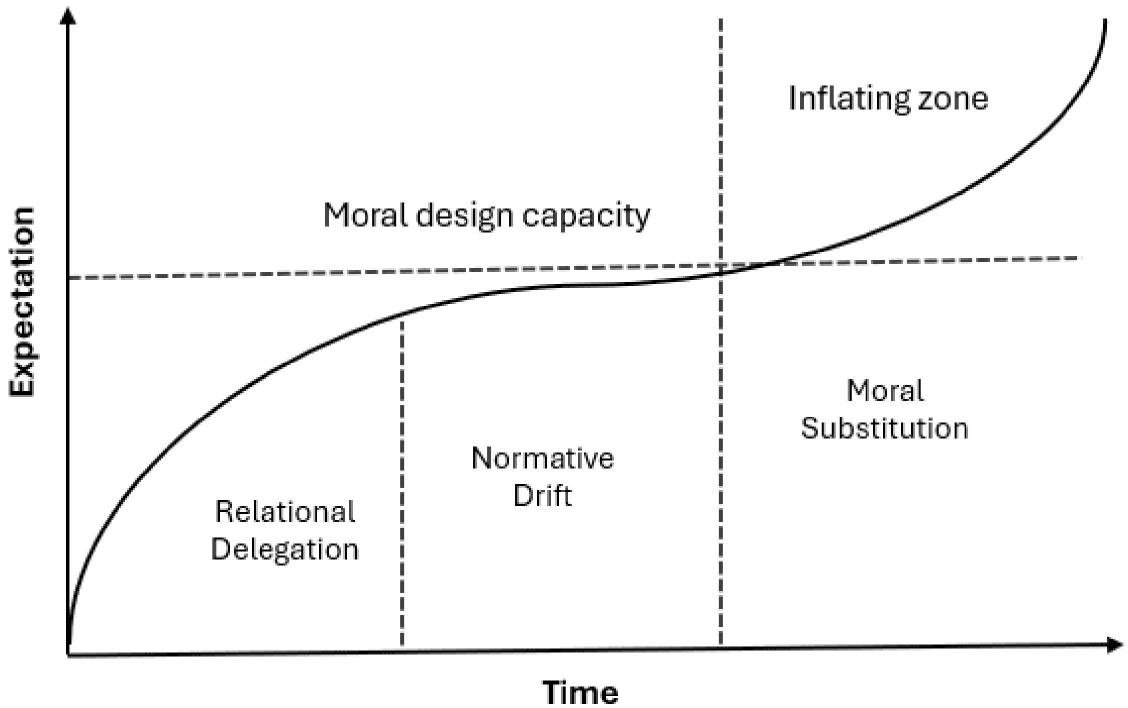

Figure 1. The Expectation Inflation Curve. The Expectation Inflation Curve illustrates how human expectations of AI's capabilities increase faster than its actual capabilities. The x-axis shows the growth of AI systems, while the y-axis shows human expectations. There are three phases: (I) Rational Delegation—where trust is appropriate; (II) Normative Drift—where trust grows faster than AI's ability, and AI starts to be seen as morally significant; and (III) Moral Substitution—where humans view AI as an ethical equal. The dashed line indicates the balance point between expectations and capabilities. Crossing this line leads to the Inflation Zone, where expectations outpace reality, leading to a decline in human judgement. This model illustrates how over-trusting AI can erode our moral authority, much like inflation dilutes a currency's value. Source: Developed by the Author (2025)

Figure 1 demonstrates the main argument of this paper regarding inflation in AI confidence. It illustrates how trust in AI transitions from practical reliance to moral displacement as expectations exceed design capabilities. The equilibrium line indicates the system's Moral Design Capacity (MDC), at which point delegated authority ceases to produce ethical value. There are three stages of delegation psychology: Rational Delegation depends on utility and oversight, Normative Drift is based on trust in fairness, and Moral Substitution occurs when AI is mistakenly regarded as a moral entity. Beyond this equilibrium, while the system continues to function, the ethical value of its outputs declines.

doi.org/10.32388/2SUYHM.2

Table 1 outlines the inflationary progression shown in Figure 1. It illustrates how the transition from Rational Delegation to Moral Substitution creates an inflationary arc that surpasses moral design capacity. In practice, this increase may seem positive—for example, as improving efficiency or fairness—but it actually shifts trust from being instrumental to being normatively dependent. The table serves as a diagnostic tool for leaders and regulators to determine their institutions' position on the expectation curve and to implement governance interventions before reaching the inflation zone.

Phase	Description	Organisational Behaviour	Risk Outcome
I. Rational Delegation	AI enhances efficiency, but human oversight is essential.	Task automation, clear accountability lines.	Low moral risk.
II. Normative Drift	AI outputs are gaining authority, and leaders are starting to see probabilistic results as absolute truth.	Over-reliance, shrinking interpretive dialogue.	Rising opacity; responsibility diffusion.
III. Moral Substitution	AI is seen as capable of judgement and empathy, with humans considering it a moral co-agent.	Delegation of ethical discretion, trust without comprehension.	Expectation Inflation; accountability vacuum.

Table 1. The Inflationary Mechanism of Expectation Inflation. The Inflationary Mechanism describes how human expectations of AI change over time, moving from reliance to excessive trust. Each phase represents a shift in the relationship between humans and AI, illustrating how functional trust evolves into moral trust. The table connects each phase to its main organisational behaviour and associated moral risks.

Source: Developed by the Author (2025)

3.1. Clarifying Moral Design Capacity

Moral Design Capacity (MDC) defines the context-dependent boundary of moral and interpretive authority that can be delegated to AI, taking into account its capabilities, design, decision context, and human oversight. It does not imply that AI has intrinsic moral agency or that it measures technical performance. Instead, it marks the limit at which normative demands—such as fairness, understanding, empathy, justification, or ethics—exceed what the system and governance can support.

MDC is therefore a normative boundary condition informed by three dimensions. System capability refers to what the AI has been demonstrably designed and validated to perform. Decision context involves the moral intricacy, uncertainty, reversibility, and potential consequences associated with the decision in which the system is employed. Governance capacity addresses the availability of meaningful human oversight, contestability, escalation mechanisms, and clearly designated accountability. These dimensions interact dynamically rather than form interchangeable components of a fixed technical score. Consequently, a technically proficient system may possess relatively limited MDC when deployed in high-stakes environments requiring value-sensitive judgement, whereas the same system could accommodate broader delegation in situations that are constrained, reversible, and under close supervision.

MDC is inherently context-dependent rather than universally applicable. Its boundaries can differ across sectors and applications because the moral requirements tied to decisions vary. For example, an AI system recommending inventory levels does not face the same normative expectations as one that guides clinical treatment, employee selection, credit decisions, or criminal justice. To conceptualise its operationalisation, MDC can be evaluated through three questions: (1) What has the system specifically been designed and validated to do? (2) How much moral or interpretive judgement does the decision demand? (3) What human and institutional safeguards are in place when the system's output is uncertain, contested, or carries significant consequences? Expectation Inflation occurs when the authority or normative competence assigned to AI surpasses these boundary conditions. MDC therefore acts as a boundary check for the escalation from Rational Delegation to Normative Drift and Moral Substitution.

3.2. Dynamics and Feedback Loops

Expectation inflation is a continuous feedback loop that reinforces itself, rather than just a temporary spike:

1. Performative Confirmation: Each successful automation confirms the belief in broader AI ability.
2. Reputational Lock-in: Managers face reputational risk from underusing AI; moral vigilance seems regressive.
3. Interpretive Erosion: Reducing human intervention lessens contextual learning, which further decreases MDC in practice.

4. Narrative Amplification: Public and media narratives turn functionality into morality, speeding up inflation.

As time goes on, the organisation’s moral focus shifts from human decision-making to machine execution. This shift can only be reversed through expectation governance, which involves formal measures that limit or provide context for delegation when ethical or interpretive issues arise.

3.3. Theoretical Propositions

From the model, three propositions can be derived.

- **P1 — Inflation Threshold:** Expectation inflation occurs when moral demands from individuals exceed what the system can handle, leading to a decline in judgement and trust.
- **P2 — Delegation Gradient:** A greater shift from functional to epistemic expectation leads to faster diffusion of accountability within organisations.
- **P3 — Governance Equilibrium:** Maintaining moral balance needs regular adjustments to the roles of humans and machines, which clear expectations and accountability should guide.

3.4. Implications for Research and Practice

The EIC presents three insights:

1. Epistemic: AI reliability is not just about the algorithm; it depends on both system capacity and human demand.
2. Organisational: It offers a leadership diagnostic that highlights how their expectations, rather than the model's performance, lead to ethical risks.
3. Policy: It establishes benchmarks for acceptable delegation levels before human oversight is required.

The Expectation Inflation Curve shows that moral risk in AI systems stems from both expectation economics and technical errors. To restore balance, leaders must act like moral central bankers[31], regulating expectations to align with the technology's ethical design.

Having identified how expectations surpass the boundaries of moral design capability, the following section explores the philosophical, organisational, and policy implications of this escalation.

doi.org/10.32388/2SUYHM.2

4. Implications — When Moral Currency Devalues

4.1. The Devaluation Metaphor

Just as economic inflation reduces the value of currency, expectation inflation undermines the moral significance of human judgement in systems that rely on AI. When organisations expect ethical reasoning from technologies devoid of intention or conscience, the basis for accountability diminishes. Although decisions may seem efficient and data-driven, their moral essence is compromised, substituted by statistical methods that merely mimic ethical reasoning.

This moral devaluation occurs through what can be called semantic drift: terms like fairness, justice, understanding, and responsibility are redefined in technical terms. "Fairness" becomes just another parameter; "understanding" simplifies to pattern recognition; and "responsibility" reduces to compliance. This shift turns rich, context-sensitive human virtues into superficial metrics, creating a false sense of ethical adequacy. This section examines the consequences of this moral decline in three areas: philosophy, organisations, and policy/governance.

4.2. Philosophical Implications — The Ontological Cost of Over-Delegation

Expectation Inflation prompts us to rethink how we understand, make decisions, and care. Human judgement relies on intentionality—the ability to convey meaning through actions—and reflective equilibrium—the skill of balancing competing values. AI systems, however, do not possess these qualities. Their outputs are merely probabilistic correlations, not true judgements[32]. As expectation inflation increases, societies start to mistake correlation for comprehension and outputs for judgements. This leads to ontological confusion, in which the distinction between entities that calculate meaning and those that create it blurs.

Genuine moral agency, as outlined by Floridi[10], requires self-awareness and thoughtful consideration of values. When people overly rely on computers for these functions, they mistakenly attribute agency where there is none, lacking the intention and context necessary for meaningful interaction. The real danger lies not in technological autonomy but in misidentifying agency—an anthropomorphic error that confuses operational success with moral understanding. Ultimately, AI does not assume agency; instead, humans grant it too soon.

This issue connects to ideas from posthumanist and sociomaterial philosophy. Scholars such as Haynes[33] and Latour[34] emphasise that the boundaries between humans and technology are often blurred, with cognition and agency distributed across both human and non-human elements. Although posthumanism helps us comprehend this interconnectedness, Expectation Inflation reveals a more profound concern: it creates a moral disparity. These viewpoints help frame the issue, even if they do not validate the shift towards misdirected agency that occurs when individuals ascribe moral importance to systems lacking intention.

Granting AI interpretive authority without a sense of moral responsibility shifts agency away from humans. Although machines can contribute to knowledge, they lack moral intention. Thus, this inflation does not acknowledge shared agency; rather, it signifies a surrender of moral authorship in hybrid systems.

People increasingly delegate not just decision-making but also the interpretation of what makes a good decision. This shifts moral deliberation from an internal process to an external one. Over time, societies begin to evaluate ethics based on how well systems perform, leading to a significant loss—the erosion of moral sovereignty, our ability to reflect, question, and take responsibility for our choices.

Expectation Inflation is not merely a misunderstanding; it reshapes how agency itself is conceived. It turns moral decision-making into a data-driven process. Philosophers and leaders should accept technology's role in human affairs, but instead focus on grounding moral values in thoughtful decision-making. We must ensure that what we delegate to machines remains within the moral responsibilities of humans.

4.3. Organisational Implications — The Erosion of Interpretive Authority

Recent workplace AI adoption shows how Expectation Inflation can develop outside formal structures. The 2024 Microsoft and LinkedIn Work Trend Index, based on 31,000 knowledge workers across 31 countries, found 75% already use AI at work and 78% bring their own AI tools[35]. This unregulated use, called "shadow AI," is significant not just for unauthorised technology but because delegation may happen before institutional review. Employees decide which tasks AI handles, how much to trust its outputs, and when to replace their judgement. Positive results reinforce expectations of competence and normalise delegation before formal policies. Shadow AI reveals how Rational Delegation can shift towards Normative Drift through daily practices, not deliberate policies. It also shows governance gaps:

doi.org/10.32388/2SUYHM.2

organisations might remain accountable but lack visibility into informal delegation practices that produce AI decisions.

At the organisational level, Expectation Inflation changes how judgement and accountability function. In traditional management, judgement is considered a defining characteristic of leadership[20]. In AI-driven organisations, judgement becomes routine, with managers explaining decisions as outputs of algorithms rather than their own choices. Authority shifts from interpretation to verification, leading to "responsibility liquidity," in which accountability is diffuse and difficult to assign[36]. This creates an illusion of objectivity in algorithmic outputs, making errors seem procedural and failures statistical, rather than moral or human. Compliance metrics replace the organisation's ethical perspective, and leadership is seen as aligning with system logic instead of guiding with moral intent.

The dynamic resembles the bystander effect[37] in ethics: as more people share responsibility, each feels less moral obligation. AI worsens this diffusion by presenting itself as neutral, leading employees and managers to trust algorithms over human judgement. This reliance on data reduces their capacity for moral reasoning, resulting in a decline in essential interpretive skills.

Organisations need effective ways to restore human judgement in AI-driven decision systems. One solution is to implement a Judgement Anchoring Protocol (JAP). These protocols require managers to explain not just the AI's recommendations but also the reasons for accepting or challenging them. JAPs can include structured reflection checklists, templates for narrative justifications, and mandatory ethics reviews across departments before implementing significant algorithmic results.

Interventions like JAPs enhance efficiency by reintroducing moral reasoning into organisational processes. They help leaders regularly evaluate the assumptions behind automated systems. Essentially, JAPs act as a stabilising force, preventing a shift from following procedures to neglecting moral responsibilities.

Without corrective structures, Expectation Inflation undermines leadership legitimacy. When moral authority is lost in decision-making processes, organisations struggle to justify their decisions, focusing only on the methods used rather than their validity. As a result, leadership may shift from principled decision-making to performative actions that lack genuine accountability.

Contemporary leaders face a dual challenge: leveraging AI's efficiency while maintaining ethical integrity. Judgement anchoring protocols can help achieve this balance by turning ethical considerations into actionable habits, providing a safeguard against the shifting nature of responsibility in the age of algorithms.

4.4. Policy and Governance Implications

AI regulation today is mainly focused on behaviour and reaction. Frameworks like the EU AI Act (2024) and national directives on algorithmic accountability emphasise measurable factors such as bias reduction, transparency, privacy, and risk classification. While these aspects are significant, they assume that human expectations are consistent and realistic—an assumption proven false by the concept of Expectation Inflation. The ethical risks in AI governance stem not just from system behaviour but also from societal expectations of those systems.

The main regulatory issue lies in demand-side distortion, which happens when institutions mistakenly attribute normative reasoning capabilities to AI. In complex socio-technical systems, such failures are often systemic rather than exceptional[38], whereas Expectation Governance addresses expectation risks. This change shifts the focus from correcting algorithms to maintaining moral balance, emphasising proactive oversight to monitor rising moral demands and prevent destabilising ethical decision-making.

Building on the ideas of Power[18], who notes that modern organisations increasingly rely on verification rituals, such as audits, for moral reassurance without genuine reflection, Expectation Governance advocates shifting from moral audits to moral accountability. Unlike traditional audits that assess compliance with ethical codes, moral accountability focuses on who has the authority to explain, justify, and face the consequences of AI-mediated decisions[30]. This shift transforms ethics from a mere assurance process into an ongoing calibration between human judgment and system capabilities.

This reconceptualisation reveals three policy mechanisms:

1. Delegation Thresholds - are the specific limits that determine when moral or interpretive tasks can be automated and when human decision-making is required.
2. Expectation Audits - assess not algorithmic bias but rather the tendency to overestimate AI's fairness, understanding, and moral reasoning abilities[39].
3. Moral Accountability Indexing - measures the percentage of decision cycles that include explicit human involvement, ensuring that responsibility is traceable.

These mechanisms create a macro-prudential approach to ethics, similar to monetary policy for controlling inflation. Regulators act as moral central bankers, responsible for preventing the formation of expectation bubbles in the AI ecosystem. Their role is not just about enforcing compliance but also about ensuring that ethical judgment remains stable and that confidence in human ethical interpretation is upheld as automation increases.

This anticipatory approach complements rather than replaces technical regulation. It emphasises the role of humans as maintainers of balance in moral economies influenced by machine inference. By combining Power's[18] insights on audit performativity with Expectation Governance, this section highlights that effective oversight of AI extends beyond mere rules; it necessitates the ongoing adjustment of societal expectations for systems that can predict outcomes but lack an understanding of their significance.

4.5. The Cultural Implication

Beyond organisations and policy, Expectation Inflation reflects a significant cultural shift in our understanding of humanity in an algorithm-driven world. The genuine concern is not that AI will mimic humans, but that humans will adopt algorithmic thinking. As decision-making becomes more procedural, societies may prioritise predictability over wisdom, compliance over conscience, and optimisation over accurate understanding.

This shift aligns with Bryson[29] and Coeckelbergh[40], who discuss AI-mediated value change—how our moral expectations adjust to fit computational norms. In this process, moral reflection is driven by efficiency, with measurable outcomes overshadowing meaningful ones. Turkle[28] refers to this as the relational substitution effect, where we favour smooth interactions over genuine connections. Ethically, this results in ethical minimalism, limiting our moral imagination to fit the rigid frameworks of machines.

Expectation inflation diminishes not only organisational judgment but also cultural empathy. When fairness focuses solely on statistics and prioritises design features, ethical considerations become mere calculations rather than thoughtful practices. This inflation casts cultural virtues such as patience, doubt, and humility as liabilities, favouring speed over reflection. Over time, moral understanding may decline as societies adjust to systems that prioritise correctness without context.

This trend subtly dehumanises decision-making processes. Essential qualities for moral agency—such as tolerance for ambiguity, contextual reasoning, and emotional intelligence—are seen as distractions from optimisation. Ultimately, this inflation affects not just what humans expect from AI, but also what they expect from themselves.

To counter this drift, we need to re-establish moral values in our culture. Educational systems, media, and leadership should promote interpretive resilience—the ability to question and recontextualise information rather than accepting it as final. Expectation Governance, discussed in the next section, provides a structural approach to maintaining moral balance through institutional mechanisms. Ultimately, the core challenge is to preserve our inherent human capacity for meaning beyond mere measurements in a world drawn to machine precision.

Having explored how Expectation Inflation disrupts moral balance in philosophical, organisational, and cultural areas, the next section will focus on its solution: Expectation Governance.

5. Toward Expectation Governance

5.1. From Regulation to Governance

Traditional AI oversight approaches are primarily regulatory and reactive, focusing on fixing ethical violations after they happen. Frameworks like the EU AI Act (2024) and OECD Guidelines (2021) aim to address outcomes but fail to tackle the underlying moral assumptions about AI. While they assess fairness, transparency, and explainability, they do not influence how AI is perceived, adopted, and trusted. Regulation manages AI's actions rather than shaping human expectations of it.

Expectation Governance addresses the issue of overreliance on systems by emphasising moral balance rather than mere compliance. It proactively manages human expectations to prevent inappropriate task delegation. This approach views ethical risks as stemming from mismanaged expectations, particularly when people overestimate AI's ability to understand, be fair, or exhibit empathy beyond its actual design capabilities.

In this framework, ethics shifts from enforcing technical standards to managing the balance between human judgment and machine decisions. Expectation Governance views expectation as a regulatory factor that can impact moral accountability if not adequately managed.

5.2. The Three Pillars of Expectation Governance

Expectation Governance is built on three key pillars: Delegation Boundaries, Expectation Auditing, and Moral Accountability Indexing. Each pillar addresses a specific stage in the moral delegation process, ensuring transparency and accountability in the relationship between human moral demands and AI design.

These pillars create a continuous cycle of prevention, diagnosis, and remediation, integrating ethics into leadership and policy rather than focusing solely on compliance. The following sections will detail each pillar, emphasising their roles in the expectation life cycle: projection, validation, and redemption.

Pillar 1: Delegation Boundaries — Establish Limits of Moral Responsibility

Delegation Boundaries define the limits of the moral and interpretive responsibility that can be transferred from humans to AI. They specify where automation stops and human judgement takes over. This principle emphasises what AI should do, based on ethical considerations rather than solely on its technical capabilities. The focus is on the potential moral implications of decisions rather than on what AI can accomplish.

- **Design Principle:** Delegation should cease at the point where conflicts of value arise.
- **Operational Tool:** Delegation maps categorise processes into automatable, shared, or non-delegable.
- **Leadership Function:** The appointment of a Delegation Steward to supervise these boundaries, ensuring that moral authority is always retained rather than quietly delegated.

Delegation Boundaries define the limits of moral transfer, turning the broad idea of "human-in-the-loop" into a clear moral checkpoint. This helps avoid over-reliance on machine judgement.

Pillar 2: Expectation Auditing — Monitoring for Inflationary Drift

Expectation Auditing identifies the inflationary pressures that occur when people's expectations of AI exceed its actual capabilities. Unlike traditional audits that focus on algorithmic fairness or bias, this audit assesses expectational bias—overconfidence from individuals and institutions that leads to excessive reliance on AI.

- **Audit Domains:**
    1. *Functional inflation* — expecting complete knowledge from limited data.
    2. *Ethical inflation* — anticipating neutrality from standard models.
    3. *Epistemic inflation* — anticipating comprehension through correlation.
- **Methodology:** Using structured interviews, document analysis, and decision-trace mapping to identify the shift of interpretive control from humans to machines.
- **Outcome**: The Expectation-to-Capacity Ratio (ECR) measures the level of moral overextension in the organisation.

Expectation Auditing serves as an early-warning system, identifying inflationary drift and enabling organisations to adjust expectations before ethical standards are compromised.

Pillar 3: Moral Accountability Indexing — Restore Interpretive Ownership

The Moral Accountability Index links outcomes to specific human authorship and assesses the level of interpretive and moral responsibility that remains with humans after they delegate decisions to AI systems.

- **Key Metric:** Human Judgment Participation Rate (HJPR) is the percentage of decision cycles in which humans provide final validation.
- **Complementary Measure:** Interpretive Depth Score measures the extent of ethical reasoning that goes beyond simply following procedures.
- **Governance Output:** Accountability Reports to track if moral reasoning is included in organisational decision-making.

Restoring interpretive ownership shifts accountability from a legal formality to an ongoing ethical practice, highlighting moral authorship within hybrid human-machine systems.

The framework of Expectation Governance is built upon three foundational pillars:

- *Delegation Boundaries*: Prevent moral overreach.
- *Expectation Auditing*: Identify inflationary drift.
- *Moral Accountability Indexing*: Restore interpretive ownership.

Organisations can achieve moral proportionality by ensuring that human expectations and AI design capabilities evolve in tandem, ethically.

doi.org/10.32388/2SUYHM.2

5.3. The Expectation Governance Framework

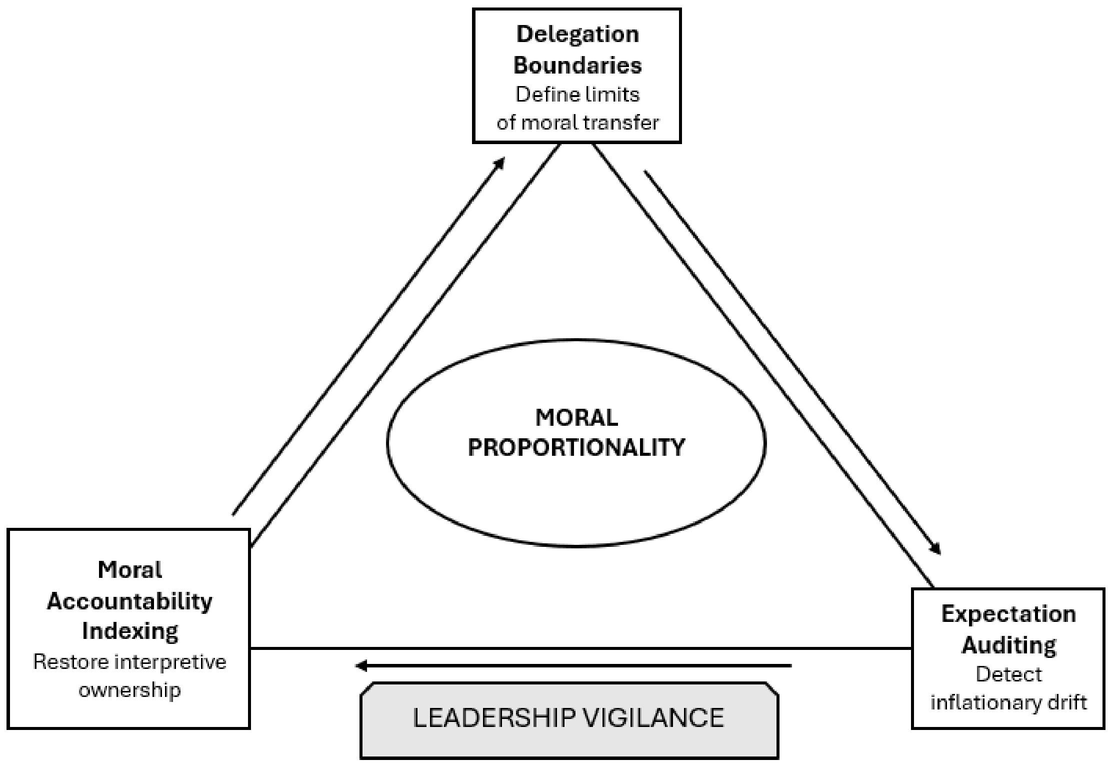

Figure 2. The Expectation Governance Framework. The Expectation Governance Framework outlines a structure to balance human expectations and AI capabilities through three key pillars: (1) Delegation Boundaries, to set moral limits; (2) Expectation Auditing, to identify inflated expectations; and (3) Moral Accountability Indexing, to ensure interpretive ownership. The centre circle, Moral Proportionality, represents the ideal balance between human judgement and machine ability. Arrows indicate a continuous cycle of prevention and restoration, while Leadership Vigilance emphasises the need for thoughtful oversight over mere technical compliance. Together, these elements help manage expectations and prevent moral over-delegation. Source: Developed by the Author (2025)

Expectation Governance functions as a self-regulating moral system. The three pillars operate as interconnected, sequential processes.

1. **Delegation Boundaries** define limits where automation stops, and human judgement takes over, preventing moral overreach.
2. **Expectation Auditing** identifies when trust in AI's moral or interpretive abilities exceeds its actual capabilities, highlighting inflationary drift.

doi.org/10.32388/2SUYHM.2

3. **Moral Accountability Indexing** ensures that moral reasoning and accountability can be clearly linked to specific individuals, restoring interpretive ownership.

The central zone of Moral Proportionality establishes ethical balance, while Leadership Vigilance ensures active moral oversight. The flow among the pillars—projection, validation, and redemption—demonstrates that governance is a continuous process, not a one-time fix. This framework enables organisations and regulators to manage the ethical relationship between human and machine agency, ensuring stability as AI technology advances.

5.4. Integrative Dynamics

The three pillars interact dynamically:

1. **Preventive Alignment (Boundary Stage):** Moral risk is reduced by restricting delegation scope during design and deployment.
2. **Diagnostic Calibration (Audit Stage):** Periodic audits track inflationary drift and reset expectations.
3. **Restorative Anchoring (Accountability Stage):** Regular indexing keeps interpretive ownership with human agents.

Together, they form a governance loop similar to a central bank's approach to controlling inflation—tracking expectations, adjusting limits, and maintaining the stability of moral currency.

5.5. Leadership and Institutional Roles

Effective Expectation Governance relies on strong leadership to maintain ethical balance amid rapid technological change. While frameworks and audits offer guidance, their credibility depends on the moral integrity of those who interpret and enforce them. Leaders must be more than just compliance overseers; they should act as stewards who ensure a balance between trust in technology and ethical considerations[41]. In the absence of clear leadership guidance, well-crafted governance frameworks may become mere procedural formalities rather than authentic moral supervision.

At the institutional level, this leadership role goes beyond ethical committees and compliance offices. It means creating a culture of reflection in which questioning AI is seen as a responsibility rather than resistance. This involves promoting psychological safety for dissent, rewarding critical thinking, and integrating ethical discussions into the strategic decision-making process.

Effective leaders serve as guardians of balance, managing the relationship between human and machine agency. They ensure that neither trust nor control becomes overly concentrated. Their role is both preventive and restorative: they identify early signs of misplaced trust and help re-establish human values when technology may overshadow ethical considerations.

Leadership thus becomes a crucial ethical function—constantly adjusting values, confidence, and accountability. Organisations that approach leadership in this manner will embrace automation rather than fear it, governing it through moral awareness, balance, and thoughtful authority.

5.6. Research and Policy Agenda

The Expectation Governance model invites a new research agenda:

- **Empirical Measurement:** Create and validate the ECR and HJPR indices using case studies in finance, healthcare, and public administration.
- **Comparative Governance:** Analyse how cultural and institutional contexts influence tolerance for expectation inflation.
- **Cross-disciplinary Integration:** Integrate behavioural economics, AI ethics, and organisational theory to create predictive models for moral devaluation.

Policy research should examine expectation caps—similar to credit limits—on high-stakes AI applications to mitigate systemic risks associated with over-delegation, such as in judicial sentencing, financial compliance, or social welfare allocation.

5.7. Concluding Proposition

P5 – Governance Equilibrium Principle:

Sustainable AI ethics requires balancing human moral needs with AI's design capabilities. Expectation governance achieves this balance through defining boundaries, auditing expectations, and indexing accountability.

Expectation Governance alters our perspective on AI ethics. Unlike regulation, which reacts to harm after it occurs, governance focuses on shaping expectations to prevent harm beforehand. This approach gives leaders more control, upholds the importance of human judgment, and counters the tendency for ethical standards to weaken in an algorithm-driven society.

6. Future Directions and Conclusion

6.1. Research Directions

This paper presents a conceptual framework that views AI ethics as a matter of moral equilibrium rather than a technical fix. Future research should focus on developing operational models grounded in Expectation Inflation, Moral Design Capacity, and Expectation Governance, to enable empirical testing and practical application.

One research focus is measuring Expectation Inflation in organisations and the public sector. Developing an Expectation-to-Capacity Ratio (ECR) can quantify the extent to which human confidence exceeds system capability, serving as an early warning for moral over-delegation. This measure enables policymakers to track expectation drift as a form of ethical volatility.

Another area involves calculating the Human Judgment Participation Rate (HJPR), which reflects the proportion of decision-making cycles in which humans retain interpretive authority. High HJPR scores suggest moral resilience, while low scores indicate potential expectation bubbles. Both ECR and HJPR can be incorporated into governance dashboards for real-time monitoring of moral stability.

One promising approach is to conduct a comparative institutional analysis to examine how different leadership cultures and governance structures—whether corporate, governmental, or hybrid—shape expectations. Comparative case studies can examine variations in leadership vigilance and delegation boundaries across different regulatory regimes, industries, or cultures.

Finally, longitudinal studies could investigate whether periods of rapid technological optimism, such as those associated with generative AI or autonomous systems, lead to spikes in expectation inflation, similar to the boom-bust cycles observed in financial markets. This would help connect the concept of moral inflation with behavioural economics, linking expectation theory, risk perception, and ethical foresight.

6.2. Conclusion

This paper presents Expectation Inflation as a framework to explain AI risks. It highlights that these risks stem from humans overestimating the ethical capabilities of AI systems that lack true intentions, rather than from algorithmic bias or opacity. The concept of Moral Design Capacity (MDC) is introduced

doi.org/10.32388/2SUYHM.2

to define the ethical limits of what an AI system can handle. The study links the crisis in AI ethics to the widening gap between machine capabilities and human expectations.

The paper introduced Expectation Governance as a solution for managing ethical demands through three key components: Delegation Boundaries (setting limits on moral transfer), Expectation Auditing (identifying issues of inflated expectations), and Moral Accountability Indexing (re-establishing ownership of interpretations). These tools are designed to enforce expectation discipline, enabling organisations to foresee and address moral bubbles before they undermine trust.

Leadership serves as the key stabiliser, ensuring balance and maintaining trust in human judgment as the basis of ethical legitimacy. Governance shifts from control to moral calibration, focusing on actively balancing capability with conscience.

This paper presents a compelling argument: the future of AI ethics hinges more on managing expectations than on refining algorithms. When the moral demands on AI exceed what it can deliver, the ethical significance grows, while the moral value diminishes. If we maintain balance, technology acts as a tool for judgment rather than replacing it.

Expectation Governance, therefore, is about using responsible imagination rather than holding back. It emphasises the importance of moderation—not prohibition or surrender—as a way for humanity to maintain its core strength: It takes moral courage to identify what should still be considered human as technology advances.

References

1. [a, b] *Akerlof GA, Shiller RJ (2010). Animal Spirits: How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism. Princeton: Princeton University Press.*
2. [a, b, c, d] *Van Lente H (2012). "Navigating Foresight in a Sea of Expectations: Lessons from the Sociology of Expectations." Technol Anal Strateg Manag. **24**(8):769–782. doi:10.1080/09537325.2012.715478.*
3. [a, b, c, d, e] *Borup M, Brown N, Konrad K, Van Lente H (2006). "The Sociology of Expectations in Science and Technology." Technol Anal Strateg Manag. **18**(3–4):285–298. doi:10.1080/09537320600777002.*
4. [^] *Burrell J (2016). "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." Big Data Soc. **3**(1). doi:10.1177/2053951715622512.*
5. [a, b] *Zuboff S (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York: Public Affairs.*

6. a, b, c, d, e, f *Beer D (2017). "The Social Power of Algorithms." Inf Commun Soc.* **20***(1):1–13. doi:*10.1080/1369118x.2016.1216147.

7. a, b, c *Pasquale F (2020). New Laws of Robotics: Defending Human Expertise in the Age of AI. Cambridge: The Belknap Press of Harvard University Press. doi:*10.4159/9780674250062.

8. a, b, c, d *Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016). "The Ethics of Algorithms: Mapping the Debate." Big Data Soc.* **3***(2). doi:*10.1177/2053951716679679.

9. a, b, c, d *Elish MC (2019). "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction." Engaging STS.* **5***:40–60. doi:*10.17351/ests2019.260.

10. a, b, c, d, e, f *Floridi L (2013). The Ethics of Information. Oxford: Oxford University Press. doi:*10.1093/acprof:oso/9780199641321.001.0001.

11. ^ *Alvesson M, Sandberg J (2011). "GENERATING RESEARCH QUESTIONS THROUGH PROBLEMATIZATION." Acad Manage Rev.* **36***(2):247–271. doi:*10.5465/amr.2011.59330882.

12. a, b, c *Kahneman D (2011). Thinking, Fast and Slow. New York: Farrar, Straus and Giroux.*

13. ^ *Simon HA (1955). "A Behavioral Model of Rational Choice." Q J Econ.* **69***(1):99–118. doi:*10.2307/1884852.

14. ^ *Weick KE, Sutcliffe KM (2001). Managing the Unexpected: Assuring High Performance in an Age of Complexity. San Francisco: Jossey-Bass.*

15. a, b, c *Brown N, Michael M (2003). "A Sociology of Expectations: Retrospecting Prospects and Prospecting Retrospects." Technol Anal Strateg Manag.* **15***(1):3–18.*

16. a, b, c *Floridi L, Cowls J (2019). "A Unified Framework of Five Principles for AI in Society." Harvard Data Sci Rev.* **1***(1). doi:*10.1162/99608f92.8cd550d1.

17. a, b, c, d *Gillespie T (2014). "The Relevance of Algorithms." In: Gillespie T, Boczkowski PJ, Foot KA, editors. Media Technologies: Essays on Communication, Materiality, and Society. Cambridge: The MIT Press. doi:*10.7551/mitpress/9780262525374.001.0001.

18. a, b, c *Power M (2022). "Theorizing the Economy of Traces: From Audit Society to Surveillance Capitalism." Organ Theory.* **3***(3). doi:*10.1177/26317877211052296.

19. a, b, c *Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019). "Fairness and Abstraction in Sociotechnical Systems." Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19. 59–68. doi:*10.1145/3287560.3287598.

20. a, b *Mintzberg H (1973). The Nature of Managerial Work. New York: Harper and Row Publishers, Inc.*

21. ^Jasanoff S (2004). States of Knowledge: The Co-Production of Science and Social Order. London: Routledge.

22. a, b, cEpley N, Waytz A, Cacioppo JT (2007). "On Seeing Human: A Three-Factor Theory of Anthropomorphism." Psychol Rev. **114**(4):864–886. doi:10.1037/0033-295x.114.4.864.

23. a, b, c, dParasuraman R, Riley V (1997). "Humans and Automation: Use, Misuse, Disuse, Abuse." Hum Factors. **39**(2):230–253. doi:10.1518/001872097778543886.

24. a, b, cDzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP (2003). "The Role of Trust in Automation Reliance." Int J Hum Comput Stud. **58**(6):697–718. doi:10.1016/s1071-5819(03)00038-7.

25. a, bLogg JM, Minson JA, Moore DA (2019). "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment." Organ Behav Hum Decis Process. **151**:90–103. doi:10.1016/j.obhdp.2018.12.005.

26. a, b, cJarrahi MH (2018). "Artificial Intelligence and the Future of Work: Human-AI Symbiosis in Organizational Decision Making." Bus Horiz. **61**(4):577–586. doi:10.1016/j.bushor.2018.03.007.

27. a, b, cRaisch S, Krakowski S (2021). "Artificial Intelligence and Management: The Automation–Augmentation Paradox." Acad Manage Rev. **46**(1):192–210. doi:10.5465/amr.2018.0072.

28. a, bTurkle S (2011). Alone Together: Why We Expect More from Technology and Less from Each Other. New York: Basic Books/Hachette Book Group.

29. a, bBryson JJ (2018). "Patiency Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics." Ethics Inf Technol. **20**(1):15–26. doi:10.1007/s10676-018-9448-6.

30. a, bRaji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P (2020). "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." FAT* 2020 - Proc. 2020 Conf. Fairness, Accountability, Transpar. 33–44. doi:10.1145/3351095.3372873.

31. ^Frimpong V (2025). "Algorithmic Authority and the Complexities of Delegated Decision-Making: Case Studies on Ethical Challenges for 21st-Century Leadership." IJOL. **0**(0):637–655. doi:10.33844/ijol.2025.60525.

32. ^Bender EM, Koller A (2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 5185–5198. doi:10.18653/v1/2020.acl-main.463.

33. ^Haynes P (2001). "How We Became Posthuman: Virtual Bodies in Cybernetics, Literature and Informatics by N. Katherine Hayles." Body Soc. **7**(4):105–108. doi:10.1177/1357034x01007004009.

34. ^Latour B (2005). Reassembling the Social: An Introduction to Actor-Network-Theory. New York: Oxford University Press.

35. ^*Microsoft, LinkedIn (2024). "Annual Work Trend Index." Microsoft and LinkedIn. https://news.microsoft.com/annual-wti-2024/.*

36. ^*Jensen MC, Meckling WH (1976). "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure." J Finan Econ.* ***3****(4):305–360. doi:10.1016/0304-405X(76)90026-X.*

37. ^*Hortensius R, de Gelder B (2018). "From Empathy to Apathy: The Bystander Effect Revisited." Curr Dir Psychol Sci.* ***27****(4):249–256. doi:10.1177/0963721417749653.*

38. ^*Perrow C (1984). Normal Accidents: Living with High-Risk Technologies. New York: Basic Books.*

39. ^*Mökander J, Floridi L (2021). "Ethics-Based Auditing to Develop Trustworthy AI." Minds Mach.* ***31****(1):323–327. doi:10.1007/s11023-021-09557-8.*

40. ^*Coeckelbergh M (2021). AI Ethics. Cambridge: The MIT Press.*

41. ^*Hannah ST, Lord RG, Pearce CL (2011). "Leadership and Collective Requisite Complexity." Organ Psychol Rev.* ***1****(3):215–238. doi:10.1177/2041386611402116.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.