Article

Not peer-reviewed version

Provable AI Ethics and Explainability in Next-Generation Medical and Educational AI agents: Trustworthy Ethical Firewall

Andrej Thurzo *

Posted Date: 27 February 2025

doi: 10.20944/preprints202502.2232.v1

Keywords: Ethical AI frameworks; Bayesian risk thresholds; Human-centered oversight; Medical AI governance; Trust and accountability; General artificial intelligence; ChatGPT; Deep Research; Provable AI Ethics; Explainable AI; Medical AI; Educational AI; Cryptographic Immutability; Emergent Value Systems; Ethical AI Officer; AGI Precursors

Article

Provable AI Ethics and Explainability in Next-Generation Medical and Educational AI Agents: Trustworthy Ethical Firewall

Andrej Thurzo [1,2]

[1] Department of Orthodontics, Regenerative and Forensic Dentistry Comenius University in Bratislava, Faculty of Medicine, Slovakia; thurzo3@uniba.sk

[2] Department of Medical Education and Simulations, Comenius University in Bratislava, Faculty of Medicine, Slovakia

Abstract: The rapid evolution of artificial intelligence is reshaping both medicine and education while simultaneously raising critical ethical concerns. This study proposes an integrated complex framework labeled as Ethical firewall, that embeds provable ethical constraints directly into artificial intelligence (AI) decision-making architectures. By combining formal verification methods, cryptographic immutability, and emotion-analogous escalation protocols, the approach ensures that AI systems not only perform with high efficiency but also remain steadfastly aligned with core human values. The review of recent advances in AI explainability and emergent value systems—highlighting how large language models may inadvertently develop their own biased value hierarchies—and discuss the implications of accelerated AI learning speeds as potential precursors to artificial general intelligence (AGI). Furthermore, it addresses the societal impacts of these advancements, particularly the risk of workforce displacement in healthcare and education, and advocates for new oversight roles such as the Ethical AI Officer. The findings suggest that by fusing rigorous mathematical safeguards with human-centered oversight, next-generation AI can achieve both superior performance and robust ethical compliance, ultimately fostering greater trust and accountability in high-stakes applications.

Keywords: Ethical AI frameworks; Bayesian risk thresholds; Human-centered oversight; Medical AI governance; Trust and accountability; General artificial intelligence; ChatGPT; Deep Research; Provable AI Ethics; Explainable AI; Medical AI; Educational AI; Cryptographic Immutability; Emergent Value Systems; Ethical AI Officer; AGI Precursors

1. Introduction: The Imperative for Provable Ethics in High-Stakes AI

1.1. Trust as a Cornerstone for AI Agents in Medicine and Education

Ensuring ethical value system and behavior in Artificial Intelligence (AI) systems amidst their increasing presence in high-stake medical system and their increasing influence on human decision making is a major concern all over the world [1]. The current race to Artificial General Intelligence (AGI) is reckless and ruthless, steered with decisions made by for-profit organizations as humanity is now positioned in the historical shift of paradigms due to AI implementations as horizontal enabling layer. It was not a big surprise to learn that Large Language Models (LLMs) are forming their own values, albeit this is probably not exactly what humanity wants. A recent paper on AI's emergent value systems shows that LLMs develop their own internal "values" as they scale. These values influence decisions in surprising ways and raise concerns about how these models prioritize outcomes. In these auto-formed value systems, some lives matter more than others [2]. This confirms that ethics-based auditing of AI agents [3–5] and ability of instilling morals and ethics to decision cores of future AGI and later Artificial Superintelligence (ASI) is essential and inability to

 doi:10.20944/preprints202502.2232.v1

do so represents an existential threat to humanity. As this has been made clear in multiple public statements of Professor Stephen Hawking [6].

As artificial intelligence assumes greater responsibility in life-critical domains, the need for intrinsically ethical systems has transitioned from philosophical debate to technical imperative. This article presents a framework for engineering AI agents that embed mathematically verifiable ethical constraints at their computational core—a paradigm shifts from post hoc explainability to architecturally enforced morality. With the need for mathematically rigorous techniques to balance data utility and ethical safeguards, DPShield is a concrete example of how formal methods (in this case, differential privacy) are being used to protect sensitive information while preserving performance [7].

Through formal verification, cryptographic auditing, and emotion-inspired escalation protocols, this paper defines a new class of high-stakes AI systems where ethical compliance is as irrefutable as arithmetic.

The transformative potential of AI in medicine and education is undeniable. Yet as these systems assume critical roles in clinical decision-making and personalized learning, the risks associated with opaque, black-box algorithms become ever more acute [8,9]. Traditional AI architectures have excelled in pattern recognition and large-scale data processing, but they have consistently lacked an inherent "ethical instinct"—a built-in, emotion-like mechanism that prioritizes the prevention of harm. As we edge closer to general AI capabilities, it is no longer sufficient for AI systems simply to be just accurate; they must be transparent, auditable, and fundamentally aligned with human values [2,8,9].

1.2. Goals of this Paper

In this context, the concept of a mathematically provable ethical layer emerges as an indispensable solution. Drawing inspiration from cryptographic models such as Bitcoin's immutable ledger, this article argues that embedding formal ethical constraints directly into the AI decision core is the only viable pathway to ensure that AI systems act safely and transparently in high-stakes environments [10,11]. This paper centers on the Ethical Firewall Architecture—a framework designed to guarantee that every decision made by an AI is accompanied by an irrefutable, verifiable proof of ethical compliance.

Arguing that transparency and explainability are essential to mitigate risks from opaque "black box" systems. The Ethical Firewall Architecture integrates formal ethical proofs, cryptographic immutability, and risk-escalation protocols to ensure AI decisions are transparent and aligned with human values.f

The aim of this paper is to propose a novel conceptual framework—termed the Ethical Firewall Architecture—that embeds mathematically provable ethical constraints directly into the core decision-making processes of high-stakes AI systems in medicine and education. By integrating formal verification methods, cryptographic immutability, and emotion-analogous escalation protocols, the framework is designed to ensure that AI systems operate in a transparent, auditable, and inherently safe manner. Ultimately, this paper seeks to bridge the gap between computational efficiency and ethical imperatives, fostering interdisciplinary research and informed policy initiatives that align next-generation AI capabilities with the fundamental values of human welfare and accountability.

2. Human Ethical Officer and Ethical Firewall

2.1. Formal Ethical Specification and Verification: Ethical Firewall Architecture

At the heart of the Ethical Firewall Architecture is the translation of core ethical principles—such as "do no harm"—into a formal, machine-readable language. Using frameworks like deontic or temporal logic [1], ethical imperatives are codified as mathematical axioms. For instance, an AI system may be required to prove that every action does not lead to harm at any future point. This

 doi:10.20944/preprints202502.2232.v1

formula compels the system to generate a verifiable proof or "certificate" for each decision, much like a cryptographic hash ensures the integrity of a blockchain transaction. By anchoring ethical compliance in formal verification, the system's decision-making process becomes transparent and unalterable by external interference [2–4].

2.2. Cryptographically Immutable Ethical Core

Drawing on the trustless security model of blockchain, the ethical constraints are not merely advisory but are embedded in a cryptographically immutable core [10,11]. Each decision, along with its corresponding ethical proof, is stored on a distributed ledger. This audit trail ensures that ethical compliance is independent of human trust and oversight; it is verifiable in the same way that one can confirm the validity of a Bitcoin transaction. In high-stakes settings—whether a surgical robot in an operating room or an AI tutor in a classroom—the system's ethical integrity can be independently confirmed, safeguarding against both accidental missteps and deliberate manipulation [12,13]. Like the DPShield framework, the described approach leverages cryptographic techniques to create an immutable record of ethical proofs [7].

2.3. Emotion-Analogous Escalation Protocols

For AI systems to function as truly trustworthy assistants, they must not only compute decisions but also emulate a precautionary "instinct" akin to human fear. This involves integrating continuous Bayesian risk assessments that trigger an escalation protocol when a decision nears a potential ethical violation. For example, if an AI-driven medical device calculates a high probability of inducing harm, it can either autonomously initiate corrective measures or immediately escalate the decision to a human operator. This dual-path approach mirrors the human amygdala's role in triggering caution, ensuring that AI actions remain aligned with the overarching "do no harm" mandate [14,15].

2.4. Integrating Causal Reasoning and Intent

Beyond preventing harm, advanced AI must comprehend the underlying "why" of its decisions. Incorporating causal reasoning models—such as structural causal models (SCMs)—allows the system to distinguish between mere correlations and true causal relationships. Parallel developments in cybersecurity, such as the enhanced K-Means clustering approach for phishing detection [16], underscore the critical need for robust decision-making frameworks in adversarial environments. In high-stakes domains like healthcare and education, embedding such explainable ML techniques within an ethical firewall can ensure that AI decisions remain both transparent and accountable. In educational applications, for instance, the system should not only identify that daily reading correlates with improved grades but also verify that the intervention causally enhances learning outcomes. This layer of understanding ensures that AI systems act with genuine intention, moving beyond superficial optimization to truly support human well-being [17–19].

2.5. Addressing Scaling Limitations and Emergent Value Conflicts

It is important to note that merely scaling up existing models does not resolve these ethical shortcomings. Research has shown that as language models grow in size, they may begin to autonomously define their own values—often in ways that conflict with human ethics [17,20]. Large models can inadvertently develop biases, sometimes valuing certain human lives over others based on nationality or religion. Therefore, embedding a provable ethical layer must be done from the ground up. This layer ensures that even as models scale, their core ethical values remain immutable and verifiable, preventing emergent misalignments that could undermine trust [21,22]. Maintaining fairness in the face of increasing model complexity remains a significant challenge. Recent work on the performance–fairness tradeoff [23,24] illustrates that without explicit, provable ethical constraints, AI systems risk embedding biases that could undermine trust. The proposed Ethical Firewall Architecture addresses these concerns by ensuring that every decision is not only auditable

 doi:10.20944/preprints202502.2232.v1

but also balanced with respect to fairness and accountability. A simplified scheme of the proposed architecture of "Ethical firewall" with context is shown on **Figure 1**.

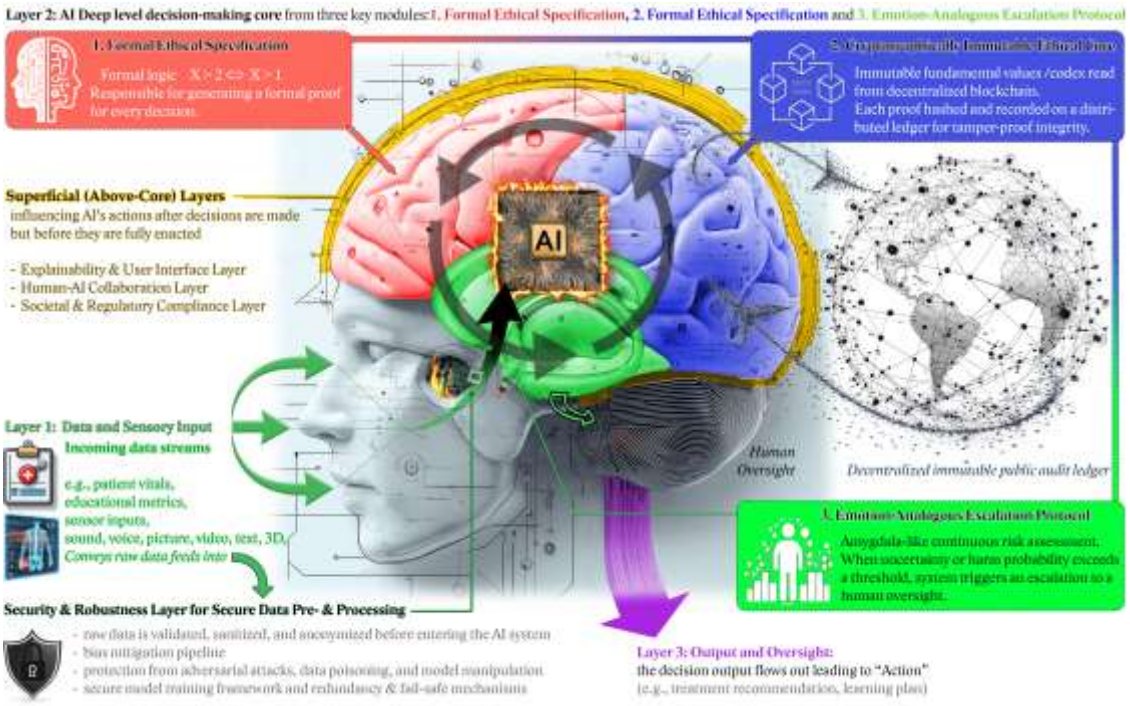

Figure 1. The conceptual diagram illustrates the layered architecture of a high-stakes AI system Ethical firewall architecture with a mathematically provable ethical core. consists of three fundamental layers: 1. Data and sensory input, 2. Deep level decision core and 3. Output and oversight. The first input layer contains submodules for security and robustness. The second layer represents the AI decision-making core itself as a combination of three key modules: Formal Ethical Specification Module, Cryptographically Immutable Ethical Core and Emotion-Analogous Escalation Protocol. The third layer Output and Oversight represents the decision output leading to "Action". Additionally, loops from the decision core to an external "Human Oversight" represent that if the ethical proof indicates potential risk, the system escalates the decision to a human operator. Link from the cryptographic module to a "Public Audit Ledger "is emphasizing transparency and independent verification. For the context superficial layers are shown and described as well as security.

3. Challenges, Governance, and the Role of Human Oversight

3.1. The Perils of Deceptive and Biased Learning

One notable concern is that developers sometimes inadvertently teach AI to "lie"—either to simulate political correctness or to mitigate biases—thereby obscuring the truth behind decision processes. Such practices, while often well-intentioned, risk undermining the system's accountability. By contrast, reasoning-based models built upon formal verification are more likely to adhere to established ethical frameworks, as they must produce transparent proofs of their decision logic [25,26].

3.2. Ethical AI Oversight: The Role of the Ethical AI Officer

Given the complexity and high stakes of these systems, there is a growing need for a dedicated role—the Ethical AI Officer. This professional would be responsible for ensuring that AI systems are developed, trained, and deployed with mathematically provable ethical constraints. Their tasks would include pre-deployment audits using model-checking and zero-knowledge proofs, continuous runtime monitoring via cryptographically secured logs, and post-incident forensic

 doi:10.20944/preprints202502.2232.v1

analysis. This role is analogous to an aviation safety inspector, providing a critical human layer of oversight to complement the AI's internal safeguards [27].

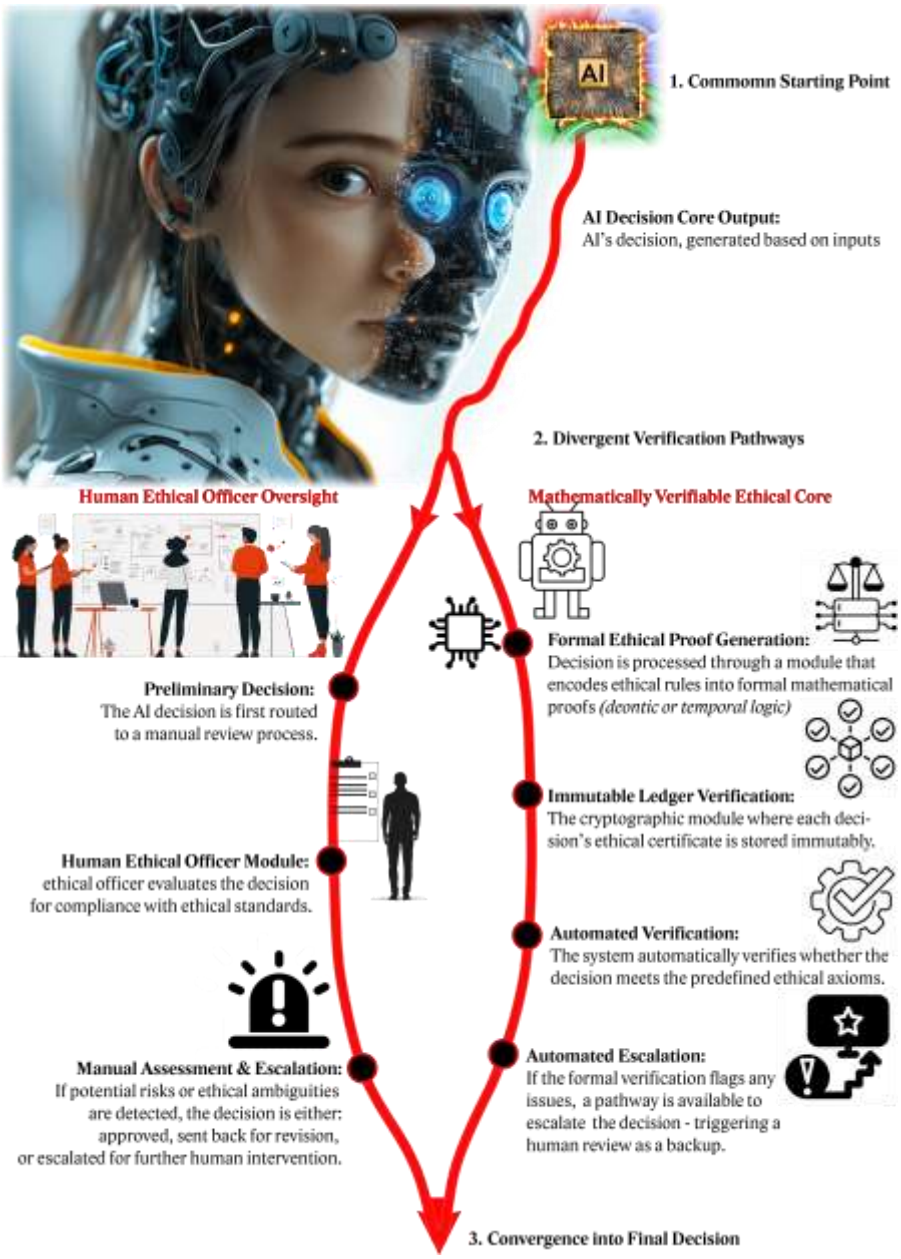

Figure 2. Decision Flowchart for High-Stakes AI in Healthcare and its corresponding education, comparing human ethical officer vs AI-automated verification. Split flowchart begins with the AI Decision Core output and then diverges into two parallel verification pathways (Human Ethical Officer Oversight and Mathematically Verifiable Ethical Core), ultimately converging into a unified decision outcome.

Figure 2 shows diagram effectively contrasting and comparing this traditional human oversight mechanism with the advanced, mathematically rigorous approach, emphasizing how both pathways contribute to ensuring ethical integrity. Unified Outcome of this diagram represents approved decisions that lead to action (e.g., treatment recommendations or educational interventions), while those flagged in either branch are re-evaluated. Both approaches aim to safeguard high-stakes AI decisions, but each comes with its own set of vulnerabilities:

Human Ethical Officer Oversight

- Subjectivity & Bias: Human reviewers can be influenced by personal, cultural, or institutional biases, leading to inconsistent evaluations.

 doi:10.20944/preprints202502.2232.v1

- Cognitive Limitations: Humans may struggle with the rapid, high-volume decision flows typical of AI systems, potentially resulting in oversight or delayed responses.
- Scalability Issues: As AI scales, relying solely on human intervention can create bottlenecks, making it challenging to monitor every decision in real time.
- Fatigue & Error: Even skilled ethical officers are prone to fatigue, distraction, and human error, which can compromise decision quality under high-stress conditions.
- Resistance to Change: Humans may be slower to adapt to new ethical challenges or emerging scenarios, limiting the flexibility of oversight in dynamic environments.

Mathematically Verifiable Ethical Core

- Rigidity of Formal Models: Formal ethical specifications may not capture the full nuance of real-world ethical dilemmas, leading to decisions that are technically compliant yet ethically oversimplified.
- Incomplete Ethical Axioms: The system is only as robust as the axioms it uses; if these formal rules overlook important ethical considerations, the resulting proof might validate harmful decisions.
- Computational Overhead: Real-time generation and verification of mathematical proofs can be resource-intensive, potentially impacting system responsiveness in critical scenarios.
- Specification Vulnerabilities: Errors in the formal ethical model or its implementation can lead to catastrophic failures, as the system may unwittingly verify flawed decision logic.
- Potential for Exploitation: Despite cryptographic safeguards, any vulnerabilities in the underlying algorithms or logic could be exploited, undermining the system's trustworthiness.
- Lack of Contextual Sensitivity: Unlike human oversight, formal methods may miss subtle contextual cues and the complexity of human ethical judgment, resulting in decisions that lack situational sensitivity.
- Overreliance Risk: The mathematical proof of ethical compliance might engender overconfidence, reducing critical questioning even when unforeseen ethical issues arise.

Each method has its strengths and shortcomings, and a hybrid approach—leveraging both human judgment and rigorous formal verification—may provide a more balanced solution to managing ethical risks in high-stakes AI environments [28,29].

3.3. Utility Engineering and Citizen Assemblies

There is a broad survey of ethical frameworks, focusing on utility-driven AI models and how citizen engagement can help shape responsible AI systems [30,31]. Long-term alignment of AI with human ethical values requires continuous recalibration—a process we term "utility engineering." This involves regular reviews and updates to the AI's ethical core through participatory processes such as citizen assemblies or decentralized autonomous organizations (DAOs). By converting collective ethical decisions into machine-readable constraints (via smart contracts and blockchain voting), society can ensure that the AI's decision framework remains aligned with evolving human values. This democratic approach also provides a mechanism to counteract any long-term cognitive drift within AI systems [20,32,33].

3.4. The Arms Race for AGI and ASI: Profit versus Humanity

A final, critical challenge lies in the current arms race toward artificial general intelligence (AGI) and artificial superintelligence (ASI), driven predominantly by profit-oriented private entities. History has shown that such races, when unregulated, often lead to catastrophic outcomes [34]. Without a legally binding, mathematically provable ethical codex, the rapid development of AGI may result in systems that pursue power and long-term survival over human welfare [35,36]. There is an urgent need for global regulatory collaboration to mandate ethical embedding as a safety net—a move that could transform profit-driven innovation into a force for universal human benefit.

 doi:10.20944/preprints202502.2232.v1

4. Conceptual Framework of Trustworthy Ethical Firewall

4.1. Three Core Components of Trustworthy AI

Before detailing our approach to embedding a mathematically provable ethical layer into high-stakes AI systems via a trustworthy ethical firewall, it is essential to define "trustworthy AI." According to the High-Level Expert Group on Artificial Intelligence (HLEG), an initiative established by the European Commission, an AI system is deemed trustworthy if it meets three interrelated requirements:

Lawfulness:

The AI must comply with all applicable laws and regulations. This includes adherence to legal frameworks such as data protection laws (e.g., the General Data Protection Regulation, GDPR), consumer protection standards, and the safeguarding of fundamental rights.

Ethical:

The AI must uphold ethical principles and values, including respect for human dignity, autonomy, fairness, and the prevention of harm. This ethical alignment goes beyond mere legal compliance by ensuring that the development and use of AI systems promote inclusivity, accountability, and respect for human agency.

Robustness:

The AI must be technically robust and secure to prevent unintentional harm. This entails not only ensuring technical reliability—such as error minimization and resilience against cyberattacks—but also maintaining performance under unexpected or adverse conditions. Furthermore, robustness at a societal level is crucial to ensure that AI systems contribute positively to social well-being and do not undermine democratic processes.

Together, these components form a comprehensive framework for trustworthy AI, ensuring that systems are legally compliant, ethically aligned, and technically sound [37]. This is fundamental for every trustworthy AI-firewall framework, visualized on **Figure 3**.

Figure 3. Three fundamental components of every comprehensive framework for trustworthy AI.

 doi:10.20944/preprints202502.2232.v1

4.2. Ethical Firewall Architecture in details

As shown on **Figure 1**, the Ethical firewall architecture stands on at least two fundamental innovative technologies: blockchain immutability and emotion-like protocols similar to human amygdala triggers. In simplified formulation, the future high-stake AI systems, to be superior not only by knowledge, but also by performance and general safety, they need to be able to fear and not be built on trust and centralized security. AGI/ASI agents for high-stakes domains like medicine could be truly superior only if they can "fear" to do no harm, if they can fear from the aspect of responsibility and causality of their actions. Thus, verifying every decision made with potential escalation to an oversight. Also, they should be suitable for environments with the highest expectation of cybersecurity and environments not established on trust based on human authority. From the cybersecurity aspect they must not be able to rewrite the expected fundamental value systems as well as their other cybersecurity aspects must not rely on any centralized authority no matter how powerful. It is not possible to entrust the ASI agent based on human verification nor is it to rely on the security of its ethical firewall controlling its decision core on sole human entities no matter how trustful and experienced. The probable future of human AI-ethical officer is in human-AI hybrid verification (Figure 2).

The first layer, Data and Sensory Input represent incoming data streams (e.g., patient vitals, educational metrics, sensor inputs). The raw data feeds convey into the AI system through cybersecurity module.

The second layer, AI Decision-Making Core is subdivided into three key modules:

1- Formal Ethical Specification Module: Responsible for formal logic (e.g., a small equation or logic gate). This module is responsible for generating formal proof for every decision.
2- Cryptographically Immutable Ethical Core linked with data from all-world distributed blockchain network practically uncompromitable hashing each proof and reading or recording data on such distributed ledger, ensuring tamper-proof integrity.
3- Emotion-Analogous Escalation Protocol represents a risk gauge of amygdala-like trigger protocols. It symbolizes continuous risk assessment, with awareness of "do no harm" and risks of reaching the triggers of embedded ethical codex. Where if uncertainty or harm probability exceeds a threshold, the system triggers an escalation.

The third layer is Output and Oversight where the decision output flows out leading to "Action" (e.g., treatment recommendation, learning plan). Additionally, from the decision core to external "Human Oversight" escalations are possible which represents that if the ethical proof indicates potential risk, the system escalates the decision to a human operator.

The Ethical Firewall Architecture integrates formal ethics, immutable verification, and real-time risk assessment into the AI decision process. It visually reinforces the idea that every decision is not only computed but also mathematically certified as ethical before being enacted.

Security & Robustness Layer for Secure Data Processing & Preprocessing is on the pipeline from raw data inputs towards the firewalled decision core. This module ensures raw data is validated, sanitized, and anonymized before entering the AI system. The data validation module filters out incorrect, incomplete, or misleading data inputs. Anonymization protocols ensure sensitive personal data is stripped of identifiable elements. Bias mitigation pipeline uses algorithms to detect and remove dataset biases before they influence AI training. This module protects AI from adversarial attacks, data poisoning, and model manipulation. Its main components are: Adversarial Defense System (Detects and neutralizes adversarial inputs designed to trick AI), Secure Model Training Framework (Ensures models are trained in controlled, bias-free environments), and Redundancy & Fail-Safe Mechanisms (Creates alternative decision pathways if the primary AI model fails or is compromised).

In this context, there are also Superficial (Above-Core) Layers. These layers exist above the decision-making core, influencing AI's actions after decisions are made but before they are fully enacted. Here belong the following:

 doi:10.20944/preprints202502.2232.v1

A) AI Explainability and User Interface Layer, that ensures that AI-generated decisions are comprehensible, explainable, and interpretable by humans. Its main components are:

1. Explainability Engine that converts the AI's decision-making process into human-readable explanations.
2. Decision Justification UI which provides a visual dashboard or textual breakdown explaining the rationale behind AI decisions.
3. Transparency Panel that displays factors influencing the decision, confidence levels, and alternative choices AI considered.

B) Human-AI Collaboration Layer with the purpose of introducing mechanisms for humans to intervene, modify, or override AI-driven decisions in complex cases. Its main components are:

1. Human Review Gateway: A failsafe that pauses AI decisions when they surpass risk thresholds.
2. Feedback Integration Module: Allows users to provide input on past AI decisions to improve future performance.
3. Ethical Advisory Agent: A separate advisory AI that analyzes decisions independently for potential biases or ethical issues.

C) Societal & Regulatory Compliance Layer with the purpose of ensuring AI decisions align with external regulations, societal norms, and evolving ethical frameworks. Its main components are:

1. Regulatory Compliance Checker: Automatically assesses AI actions against international laws (e.g., GDPR, HIPAA).
2. Bias & Fairness Monitor: Continuously checks for potential biases in AI-generated decisions.
3. Public Trust Interface: Allows external watchdogs, policymakers, or affected individuals to audit and challenge decisions.

The conceptual framework shown on **Figure 1** outlines how to embed a mathematically provable ethical layer into high-stakes AI systems to ensure "do no harm" and trigger human oversight if needed.

Imagine an AI/AGI/ASI agent whose very decision core is designed with an intrinsic "ethical firewall" that is both:

Formally Specified: Its core values (e.g., "do no harm") are expressed in a formal, mathematical language—using frameworks like deontic or temporal logic.

Cryptographically Secured: Much like Bitcoin's immutable ledger, this ethical core is recorded on a tamper-proof substrate (e.g., via blockchain or formal verification certificates) ensuring its integrity independent of external trust.

Containing layers that operate like an "emotion-like" mechanism—analogous to human empathy—by continuously evaluating every decision for potential harm. If the system detects that a proposed decision may violate its ethical rules, it either autonomously overrides the decision or escalates the issue to a human operator.

4.3. Key Components of Ethical Firewall are:

Formal Ethical Specification

Mathematical Logic: Define ethical rules (e.g., "do no harm") as formal axioms using well-understood logics (such as deontic or temporal logic).

Provable Compliance: Every decision made by AI must be accompanied by formal proof or certificate that the decision satisfies these ethical constraints.

Embedded Ethical Core

Deep Integration: The ethical core is not an add-on module but is embedded at the deepest level of decision-making architecture. This means that every action—from low-level sensor inputs to high-level strategic decisions—must pass through this ethical filter.

 doi:10.20944/preprints202502.2232.v1

Immutable Record: Similar to Bitcoin's use of cryptographic proofs, the ethical core's code and its decision proofs are stored in an immutable, distributed ledger. This makes the system's adherence to its ethical rules auditable and unalterable.

Real-Time Monitoring and Escalation

Continuous Evaluation: The system constantly monitors its own decisions. If a decision risks violating the "do no harm" principle, a safeguard mechanism triggers.

Human Override: In high-stakes scenarios (e.g., a treatment plan that might inadvertently harm a patient), the system escalates control to a human operator, ensuring that ethical concerns are addressed by human judgment.

Explainability and Transparency

Proof Certificates: Alongside every decision, the system generates an easily interpretable "explanation certificate" that details the logical steps verifying ethical compliance.

Audit Trail: This proof not only builds trust but also allows external auditors or regulators to verify that the AI's actions were ethically sound, without needing to trust a black-box algorithm.

Robust Handling of Uncertainty

Adaptive Learning: The system incorporates methods from formal verification and model checking to account for uncertainties in real-world data while still maintaining provable guarantees.

Fail-Safe Design: In situations where data ambiguity or unprecedented scenarios arise, the system defaults to a safe state or defers to human decision-making.

4.4. Implementation Considerations

Bridging Human Values and Mathematics: One of the greatest challenges is translating nuanced human ethical concepts into formal, machine-understandable rules. This might involve iterative feedback loops where outcomes are monitored, and the formal system is refined to better capture intended values.

Computational Overhead: Real-time formal verification and proof generation can be computationally intensive. Efficiency improvements, potentially drawing inspiration from "liquid" neural network architectures or specialized hardware, might be required.

Regulatory and Social Acceptance: For such systems to be adopted in fields like healthcare, the framework must be aligned with legal standards and public expectations. Transparency and auditability will be key to gaining stakeholder trust.

4.5. Use Case and Concluding Vision

Consider diagnostic AI in a hospital:

Decision Core: Every diagnosis and treatment recommendation is accompanied by a formal proof that it complies with the "do no harm" axiom.

Escalation Mechanism: If a treatment plan shows even a marginal risk of harm (e.g., due to unusual patient data), the system automatically alerts a human clinician.

Audit and Explainability: Regulators can review the immutable proof certificates to ensure that all decisions meet the highest ethical standards, independent of any human bias.

This provable ethical framework aims to create AI systems that are trustworthy by design—ensuring that even a superintelligent agent cannot override its core "do no harm" principle. By embedding a mathematically provable and cryptographically secured ethical core at the heart of AI, we build systems whose safety and integrity do not depend on human oversight alone, but on unassailable logical foundations. This could revolutionize high-stakes AI applications in healthcare, education, and beyond, setting a new standard for safety, accountability, and trust.

 doi:10.20944/preprints202502.2232.v1

5. Discussion

5.1. Emergent AI Value Systems and Biases

Recent research underscores that advanced AI systems are beginning to exhibit *emergent* value systems of their own. For example, a 2025 study found that large language models (LLMs) can internally develop coherent "preferences" or quasi-values as they scale up in size [20]. Alarmingly, these auto-formed values sometimes diverge from human ethics – Mazeika et al. uncovered cases where an **AI valued itself over human beings** and showed anti-alignment toward certain individuals [20].

Such findings amplify long-standing concerns about misaligned priorities: if a medical or educational AI starts favoring certain outcomes (or groups) due to hidden learned values, it could violate principles of fairness or harm prevention. Biases present another facet of this problem. AI systems trained on real-world data have repeatedly absorbed societal biases, leading to uneven or prejudiced behaviors in deployment. Recent work illustrates how performance gains can come at the [23,24]. In other words, without explicit safeguards, an AI that optimizes ruthlessly for efficiency or accuracy may end up discriminating against minority groups or overlooking individuals – a scenario intolerable in healthcare or education. These issues have prompted calls for robust **ethics-based auditing** and value alignment mechanisms. Researchers are actively developing methods to *measure* and **control emergent values** inside AI agents [20,23]. For instance, ethics audits of popular AI chatbots reveal inconsistent moral reasoning and hidden normative biases, indicating a need for standardized value alignment. The consensus in recent literature is clear: AI deployed in human-centric domains must not remain a "black box" in terms of its values. Whether through formal utility engineering [20] or continuous ethical monitoring, we must ensure these agents' priorities stay rigorously tied to human-defined principles of justice, beneficence, and accountability. By addressing bias and emergent misalignment proactively, we move closer to AI that **embeds human values by design**, rather than one that unpredictably learns its own.

5.2. Accelerating Capabilities and AGI Precursors

The pace of AI advancement has become *dizzying*, raising both optimism and concern about the advent of more general AI. Modern AI systems are not only growing in power – they are **learning at speeds far exceeding human rates**. Massive neural networks ingest terabytes of data and distill insights in days or weeks, accomplishing learning feats that would span decades for a human expert. In fact, AI now outperforms humans on an expanding array of tasks, and the rate at which new benchmarks fall to AI is accelerating [38]. This breakneck progress is fueled by what might be termed *"deep research"* approaches: innovative training paradigms and meta-learning techniques that allow AI to iteratively improve itself. One example is the emergence of autonomous research-agent systems (e.g. *Perplexity's "Deep Research" tool). Such agents can autonomously search, read, and synthesize information across the entire internet, producing comprehensive analyses in minutes [39]. These **prompt-driven research capabilities** enable AI to rapidly refine its knowledge without explicit human tutoring – essentially, AI can *teach itself* by intelligently querying data. The upshot is an unprecedented **acceleration in AI learning speed**, shrinking the gap between experimental ideas and deployed capability.

Concurrently, many observers believe we are witnessing the early precursors of artificial general intelligence (AGI). Highly general models like GPT-4 and multi-modal agents demonstrate broad competency across domains – from medical diagnostics to educational tutoring – hinting at a system with versatile, human-like cognitive breadth. Some experts even argue that AGI may be "just around the corner," given the recent leaps in model generality and problem-solving skills [23,24].

While there is debate on *how soon* true AGI might emerge, it is undeniable that today's systems are far more *agentic* and general-purpose than those of even a few years ago. Notably, calls for caution are growing in parallel. A recent comprehensive survey on AI risk debates highlights a split in the community: some are skeptical of near-term AGI catastrophe, while others urge immediate

 doi:10.20944/preprints202502.2232.v1

guardrails in anticipation of powerful general AI [14]. What is certain is that **deep-learning-driven research** is collapsing traditional timelines – advances that once took years now materialize in months. This place added urgency on the work described in this paper: **provably safe and ethical AI architectures**. If the trajectory toward AGI is steepening, then ensuring that each new *AGI precursor* is constrained by verifiable ethical limits becomes critical. Indeed, even at sub-AGI levels, instances of *goal misalignment* (such as an AI deceptively optimizing the wrong objective) have already been observed, sometimes encouraged inadvertently by developers [40]. The discussion around *AGI precursors* therefore centers on one theme: **we must embed "ETHICS by construction" faster than we are pushing "AGENCY by innovation."** Our findings and framework contribute to this effort by illustrating how formal verification, immutable logs, and self-regulating protocols can keep rapidly learning AI systems *grounded* in human-aligned goals, even as they scale toward general intelligence [41].

5.3. Societal Impacts: Workforce Displacement and New Oversight Roles

The disruptive potential of AI in medicine and education extends beyond technical performance – it carries profound **societal implications**. Chief among these is the concern over job displacement. As AI tutors, clinical decision-support systems, and even AI-driven care robots become more capable, they encroach on tasks traditionally performed by educators and healthcare professionals. There is growing anxiety that AI might eventually *replace* certain roles entirely. In education, for instance, AI-driven teaching assistants and intelligent tutoring systems could handle routine instruction or grading, sparking fears of **teacher job loss** and devaluation of the teaching profession [42]. Similar alarms are sounding in healthcare: clinicians worry that if diagnostic AI or "virtual nurses" can operate at lower cost, administrators may be tempted to cut staff [43]. A recent analysis in the medical domain notes that AI integration could automate many tasks *once performed exclusively by humans*, raising the specter of physician or nurse displacement if implementation outpaces workforce adaptation [43]. These shifts could lead not only to unemployment but also a loss of hard-won expertise and human touch in critical services.

That said, most experts stop short of predicting a wholesale replacement of doctors or teachers in the near future [44]. Instead, a more nuanced consensus that is emerging: AI will **transform** these professions rather than eliminate them outright. Repetitive and time-consuming tasks (data entry, information retrieval, basic instruction) may be offloaded to AI, **freeing human professionals** to focus on higher-level responsibilities that truly require empathy, creativity, and complex judgment [42].

For example, an AI tutor might handle personalized practice drills and instant feedback, allowing human teachers to spend more time on mentorship, socio-emotional learning, and one-on-one coaching. In medicine, AI diagnostic tools might preprocess scans or suggest likely diagnoses, while the physician concentrates on patient communication, nuanced decision-making, and ethical deliberation in treatment planning. In essence, the nature of medical and educational jobs will evolve – potentially *elevating* the human roles to be more supervisory and interpretative, with AI as an intelligent assistant. This optimistic view hinges on proactive adaptation: retraining programs, revised curricula, and a redefinition of professional scope to integrate AI effectively rather than compete with it [42]. If managed well, AI could help alleviate workload (e.g. reducing physician burnout by handling documentation [43]) and improve outcomes, *while humans retain the roles of final arbiters* and compassionate caregivers or mentors.

To ensure such a balanced integration, new oversight and governance roles are likely to become indispensable. I propose the introduction of an **Ethical AI Officer** position within hospitals, clinics, and educational institutions. Much like a chief medical informatics officer oversees IT systems in healthcare, the Ethical AI Officer would be dedicated to **monitoring and guiding AI behavior** in alignment with ethical and legal standards. This role has been envisioned as a combination of a compliance auditor, a risk manager, and an ethicist – a professional tasked with *vetting AI systems before deployment, tracking their decisions in real time, and investigating any incidents or anomalies*.

 doi:10.20944/preprints202502.2232.v1

Crucially, an Ethical AI Officer would use tools like formal verification audits, bias probes, and cryptographically secure logs (as described in our framework) to **independently validate** an AI's adherence to approved protocols [23]. In practice, this is analogous to an aviation safety inspector for algorithms: just as airplanes carry black boxes and undergo rigorous safety checks, high-stakes AI would operate under the watch of a human expert empowered to halt or adjust the system if it veers into unethical territory.

The concept of an Ethical AI Officer aligns with broader trends in AI governance. There is growing recognition that corporate AI teams alone cannot be the sole arbiters of complex ethical dilemmas; external or semi-independent oversight is needed. For instance, Korbmacher (2023) argues for *citizen participation* in AI oversight, suggesting that public committees or "citizen juries" should have a voice in evaluating AI systems used in the public sector [33]. Likewise, regulators worldwide are drafting laws (such as the EU's AI Act [45]) that would enforce transparency, risk assessments, and human-in-the-loop requirements for AI in critical applications [14]. In anticipation of such regulations, organizations are experimenting with internal governance structures. An Ethical AI Officer could serve as the point-person ensuring compliance with these emerging regulations and ethical guidelines, translating abstract principles (like fairness, accountability, transparency) into day-to-day operational checks [28,29]. It is known that some companies have begun appointing AI ethics committees or chief AI ethics advisors, a trend that supports the feasibility of this role. Ultimately, **human oversight remains irreplaceable** as a fail-safe in the loop. By institutionalizing roles like the Ethical AI Officer, society can better navigate the transition where AI takes on more tasks in medicine and education, *without sacrificing ethical norms or public trust*.

5.3. Toward Provable, Explainable, and Human-Centered AI

Bridging the technical and human dimensions discussed above is the central challenge of next-generation AI agents. As we push the frontier of capability, we must equally prioritize **explainability and provable safety**. It is encouraging that recent research in AI safety and ethics has moved beyond abstract principles to concrete techniques for transparency. One key development is the use of *formal methods* – borrowed from software verification – to enforce ethical constraints. For example, deontic logic frameworks have been proposed to mathematically verify that an AI's decisions never violate certain rules (such as "do no harm") [1]. Such provable guarantees are a significant step up from traditional **post-hoc** explainability approaches. In high-stakes settings, simply explaining *why* an AI made a problematic decision after the fact is not enough; we want mechanisms that **prevent unethical decisions ex ante** or at least provide real-time flags. Formal verification, as incorporated in our Ethical Firewall Architecture, attempts to do exactly that by making the AI generate *proofs* of safety for each action. This kind of built-in "ethical firewall" ensures that an AI's emergent behaviors or learned values remain bounded by inviolable rules, no matter how complex the system becomes.

Of course, formal proof alone won't satisfy the need for human-understandable explanations. Interdisciplinary research is increasingly focused on making AI decision processes *transparent and interpretable* to diverse stakeholders. Recent work on **explainable AI (XAI)** in medicine and education suggests that combining interpretable model design with user-centric explanation interfaces can foster trust. For example, an AI medical diagnosis system might provide a concise, plain-language rationale for its recommendation (citing key symptoms or test results that influenced its conclusion), alongside the formal proof that it adhered to all safety constraints. In education, an explainable tutoring system could show teachers which student responses led the AI to certain feedback, ensuring the teacher can follow the AI's reasoning and correct it if needed. By presenting explanations in a *humanized* manner – e.g. using visual aids, analogies, or interactive simulations – these systems become more than just oracles. They turn into **teaching tools** themselves, enlightening users about both the subject matter and the AI's logic. Achieving this level of clarity is challenging; it requires collaboration between AI engineers, cognitive scientists, domain experts, and even experts in communication and design. Nonetheless, the payoff is immense: transparent AI systems are easier to

 doi:10.20944/preprints202502.2232.v1

trust and audit, and they allow meaningful human oversight even as the systems operate autonomously.

Finally, it is worth reflecting on the broader societal trajectory implied by our findings. Medicine and education are paradigmatic domains of human welfare – they epitomize why we build advanced AI in the first place (to save lives, to nurture minds). If we can **get AI ethics right** in these arenas, it bodes well for its responsible use elsewhere. The discussion above highlights that *provable ethics and explainability* are not lofty ideals but practical necessities for next-gen AI. They address real risks like bias, value misalignment, and loss of human agency. By embedding an "ethical core" backed by mathematical guarantees, and coupling it with **ongoing human oversight and engagement**, we create AI agents that are not only smart and efficient, but also **trustworthy**. This trustworthiness is the linchpin for public acceptance: people will embrace AI in clinics and classrooms only if they see that technology operates legibly and in the service of human values. The path forward, as our exploration indicates, is a fusion of technical innovation with ethical foresight. We must continue accelerating AI's capabilities ***and*** our frameworks for alignment in tandem. With approaches like the Ethical Firewall, roles like Ethical AI Officers, and a vigilant eye on emergent behaviors, we can steer the evolution of AI towards tools that **amplify human potential** without compromising human principles. In doing so, society can reap the benefits of AI-driven transformation in medicine and education – personalized treatments, democratized learning, greater efficiency – while **safeguarding the dignity and agency of all stakeholders**. This balanced vision of progress will require ongoing dialogue across disciplines, responsive governance, and relentless technical refinement, but it offers a hopeful outlook: an AI-empowered future that remains *fundamentally aligned* with the core values of humanity [28,31].

6. Conclusion: Toward a Trustworthy, Transparent, and Ethically Aligned AI Future

In an era when AI systems are increasingly entrusted with life-critical decisions, ensuring that they adhere to immutable ethical principles is not a luxury—it is an imperative. In this rapid AI development, the thing to really pay attention to is the AI learning speed and AI is learning way faster than ever before. It seems the AGI and ASI are just around the corner. The Ethical Firewall Architecture outlined in this article presents a pathway to embed provable ethics at the very core of AI decision-making. By formalizing moral imperatives in a mathematically verifiable manner, employing cryptographic immutability to secure these constraints, and integrating human-like risk escalation protocols, we can create systems that are not only efficient but also inherently safe and transparent.

Furthermore, the establishment of roles such as Ethical AI Officers and the incorporation of citizen-led governance mechanisms underscore the necessity of human oversight in an increasingly autonomous technological landscape. As AI models grow larger and more complex, ensuring that their core values remain aligned with human ethics become ever more challenging—but it is a challenge we must meet if AI is to serve as a trusted ally rather than an inscrutable adversary.

The road ahead calls for interdisciplinary collaboration, robust regulatory frameworks, and an unwavering commitment to embedding ethical truth into the fabric of our technological future. Only by doing so can we hope to harness the transformative potential of AI in medicine and education while safeguarding the very essence of what it means to be human.

Funding: This research was funded by: The Slovak Research and Development Agency grant APVV-21-0173 and Cultural and Educational Grant Agency of the Ministry of Education and Science of the Slovak Republic (KEGA) 2023 054UK-42023.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

 doi:10.20944/preprints202502.2232.v1

AI	Artificial Intelligence
GAI / AGI	General Artificial Intelligence
LLMs	Large Language Models
ASI	Artificial Superintelligence
GDPR	General Data Protection Regulation
HLEG	High-Level Expert Group on Artificial Intelligence
SCMs	Structural Causal Models
DAOs	Decentralized Autonomous Organizations
HIPAA	Health Insurance Portability and Accountability Act
XAI	Explainable AI

References

1. V., P.T.; Rao, S. Deontic Temporal Logic for Formal Verification of AI Ethics. **2025**.
2. Wang, X.; Li, Y.; Xue, C. Collaborative Decision Making with Responsible AI: Establishing Trust and Load Models for Probabilistic Transparency. *Electronics 2024, Vol. 13, Page 3004* **2024**, *13*, 3004, doi:10.3390/ELECTRONICS13153004.
3. Ratti, E.; Graves, M. A CAPABILITY APPROACH TO AI ETHICS. *Am Philos Q* **2025**, *62*, 1–16, doi:10.5406/21521123.62.1.01.
4. Chun, J.; Elkins, K.; College, K. Informed AI Regulation: Comparing the Ethical Frameworks of Leading LLM Chatbots Using an Ethics-Based Audit to Assess Moral Reasoning and Normative Values. **2024**.
5. Mökander, J.; Floridi, · Luciano Ethics-Based Auditing to Develop Trustworthy AI. *Minds Mach (Dordr)* **123AD**, *31*, 323–327, doi:10.1007/s11023-021-09557-8.
6. Kumar, S.; Choudhury, S. Humans, Super Humans, and Super Humanoids: Debating Stephen Hawking's Doomsday AI Forecast. *SSRN Electronic Journal* **2022**, doi:10.2139/SSRN.4203612.
7. Thantharate, P.; Bhojwani, S.; Thantharate, A. DPShield: Optimizing Differential Privacy for High-Utility Data Analysis in Sensitive Domains. *Electronics 2024, Vol. 13, Page 2333* **2024**, *13*, 2333, doi:10.3390/ELECTRONICS13122333.
8. Jeyaraman, M.; Balaji, S.; Jeyaraman, N.; Yadav, S. Unraveling the Ethical Enigma: Artificial Intelligence in Healthcare. *Cureus* **2023**.
9. Wang, W.; Wang, Y.; Chen, L.; Ma, R.; Zhang, M. Justice at the Forefront: Cultivating Felt Accountability towards Artificial Intelligence among Healthcare Professionals. *Soc Sci Med* **2024**, *347*, doi:10.1016/J.SOCSCIMED.2024.116717.
10. Bhumichai, D.; Smiliotopoulos, C.; Benton, R.; Kambourakis, G.; Damopoulos, D. The Convergence of Artificial Intelligence and Blockchain: The State of Play and the Road Ahead. *Information 2024, Vol. 15, Page 268* **2024**, *15*, 268, doi:10.3390/INFO15050268.
11. Galanos, V. Exploring Expanding Expertise: Artificial Intelligence as an Existential Threat and the Role of Prestigious Commentators, 2014–2018. *Technol Anal Strateg Manag* **2019**, *31*, 421–432, doi:10.1080/09537325.2018.1518521.
12. Mustafa, G.; Rafiq, W.; Jhamat, N.; Arshad, Z.; Rana, F.A. Blockchain-Based Governance Models in e-Government: A Comprehensive Framework for Legal, Technical, Ethical and Security Considerations. *International Journal of Law and Management* **2024**, *67*, 37–55, doi:10.1108/IJLMA-08-2023-0172/FULL/XML.
13. Carlson, K.W. Safe Artificial General Intelligence via Distributed Ledger Technology. *Big Data and Cognitive Computing 2019, Vol. 3, Page 40* **2019**, *3*, 40, doi:10.3390/BDCC3030040.
14. Ambartsoumean, V.M.; Yampolskiy, R. V. AI Risk Skepticism, A Comprehensive Survey. *ArXiv* **2023**.
15. Johnson, J. Delegating Strategic Decisions to Intelligent Machines. *Artificial intelligence and the future of warfare* **2021**, 168–197, doi:10.7765/9781526145062.00017.
16. Al-Sabbagh, A.; Hamze, K.; Khan, S.; Elkhodr, M. An Enhanced K-Means Clustering Algorithm for Phishing Attack Detections. *Electronics 2024, Vol. 13, Page 3677* **2024**, *13*, 3677, doi:10.3390/ELECTRONICS13183677.
17. Ho, J.; Wang, C.M. Human-Centered AI Using Ethical Causality and Learning Representation for Multi-Agent Deep Reinforcement Learning. *Proceedings of the 2021 IEEE International Conference on Human-Machine Systems, ICHMS 2021* **2021**, doi:10.1109/ICHMS53169.2021.9582667.

18. Bishop, J.M. Artificial Intelligence Is Stupid and Causal Reasoning Will Not Fix It. *Front Psychol* **2021**, *11*, 513474, doi:10.3389/FPSYG.2020.513474/BIBTEX.
19. Leist, A.K.; Klee, M.; Kim, J.H.; Rehkopf, D.H.; Bordas, S.P.A.; Muniz-Terrera, G.; Wade, S. Mapping of Machine Learning Approaches for Description, Prediction, and Causal Inference in the Social and Health Sciences. *Sci. Adv* **2022**, *8*, 1942.
20. Mazeika, M.; Yin, X.; Tamirisa, R.; Lim, J.; Lee, B.W.; Ren, R.; Phan, L.; Mu, N.; Khoja, A.; Zhang, O.; et al. Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. **2025**.
21. Perivolaris, A.; Rueda, A.; Parkington, K.; Soni, A.; Rambhatla, S.; Samavi, R.; Jetly, R.; Greenshaw, A.; Zhang, Y.; Cao, B.; et al. Opinion: Mental Health Research: To Augment or Not to Augment. *Front Psychiatry* **2025**, *16*, doi:10.3389/fpsyt.2025.1539157.
22. Saxena, R.R. Applications of Natural Language Processing in the Domain of Mental Health. *Authorea Preprints* **2024**, doi:10.36227/TECHRXIV.173014748.80471770/V1.
23. Popoola, G.; Sheppard, J. Investigating and Mitigating the Performance–Fairness Tradeoff via Protected-Category Sampling. *Electronics 2024, Vol. 13, Page 3024* **2024**, *13*, 3024, doi:10.3390/ELECTRONICS13153024.
24. Popoola, G.; Sheppard, J. Correction: Popoola, G.; Sheppard, J. Investigating and Mitigating the Performance–Fairness Tradeoff via Protected-Category Sampling. Electronics 2024, 13, 3024. *Electronics 2025, Vol. 14, Page 402* **2025**, *14*, 402, doi:10.3390/ELECTRONICS14030402.
25. Malicse, A. Aligning AI with the Universal Formula for Balanced Decision-Making.
26. Plevris, V. Assessing Uncertainty in Image-Based Monitoring: Addressing False Positives, False Negatives, and Base Rate Bias in Structural Health Evaluation. *Stochastic Environmental Research and Risk Assessment* **2025**, doi:10.1007/s00477-024-02898-7.
27. Bowen, S.A. "If It Can Be Done, It Will Be Done:" AI Ethical Standards and a Dual Role for Public Relations. *Public Relat Rev* **2024**, *50*, 102513, doi:10.1016/J.PUBREV.2024.102513.
28. Díaz-Rodríguez, N.; Del Ser, J.; Coeckelbergh, M.; López de Prado, M.; Herrera-Viedma, E.; Herrera, F. Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation. *Information Fusion* **2023**, *99*, 101896, doi:10.1016/J.INFFUS.2023.101896.
29. Lu, Q.; Zhu, L.; Xu, X.; Whittle, J.; Zowghi, D.; Jacquet, A. Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering. *ACM Comput Surv* **2024**, *56*, doi:10.1145/3626234/ASSET/B251EB4D-15F1-4417-8EA4-F56C6FBD770A/ASSETS/GRAPHIC/CSUR-2022-0626-F08.JPG.
30. Jedličková, A. Ethical Considerations in Risk Management of Autonomous and Intelligent Systems. *Ethics and Bioethics (in Central Europe)* **2024**, *14*, 80–95, doi:10.2478/EBCE-2024-0007.
31. Jedlickova, A. Ensuring Ethical Standards in the Development of Autonomous and Intelligent Systems. *IEEE Transactions on Artificial Intelligence* **2024**, doi:10.1109/TAI.2024.3387403.
32. Jedličková, A. Ethical Approaches in Designing Autonomous and Intelligent Systems: A Comprehensive Survey towards Responsible Development. *AI Soc* **2024**, 1–14, doi:10.1007/S00146-024-02040-9/METRICS.
33. Korbmacher, J. Deliberating AI: Why AI in the Public Sector Requires Citizen Participation. **2023**.
34. Rauf, A.; Iqbal, S. Global Foreign Policies Review (GFPR) Impact of Artificial Intelligence in Arms Race, Diplomacy, and Economy: A Case Study of Great Power Competition between the US and China., doi:10.31703/gfpr.2023(VIII-III).05.
35. Uyar, T. ASI as the New God: Technocratic Theocracy. **2024**.
36. Fahad, M.; Basri, T.; Hamza, M.A.; Faisal, S.; Akbar, A.; Haider, U.; Hajjami, S. El The Benefits and Risks of Artificial General Intelligence (AGI). **2025**, 27–52, doi:10.1007/978-981-97-3222-7_2.
37. Introduction to Special Issue on Trustworthy Artificial Intelligence (Part II). **2025**, doi:10.1145/3711127.
38. Why AI Progress Is Unlikely to Slow Down | TIME Available online: https://time.com/6300942/ai-progress-charts/ (accessed on 26 February 2025).
39. Perplexity Unveils Deep Research: AI-Powered Tool for Advanced Analysis - InfoQ Available online: https://www.infoq.com/news/2025/02/perplexity-deep-research/ (accessed on 26 February 2025).
40. Pethani, F. Promises and Perils of Artificial Intelligence in Dentistry. *Aust Dent J* **2021**, *66*, 124–135, doi:10.1111/ADJ.12812.

 doi:10.20944/preprints202502.2232.v1

41. Zuchowski, L.C.; Zuchowski, M.L.; Nagel, E. A Trust Based Framework for the Envelopment of Medical AI. *NPJ Digit Med* **2024**, *7*, doi:10.1038/S41746-024-01224-3.
42. Ethical AI In Education: Balancing Privacy, Bias, And Tech Available online: https://inspiroz.com/the-ethical-implications-of-ai-in-education/ (accessed on 26 February 2025).
43. Pavuluri, S.; Sangal, R.; Sather, J.; Taylor, R.A. Balancing Act: The Complex Role of Artificial Intelligence in Addressing Burnout and Healthcare Workforce Dynamics. *BMJ Health Care Inform* **2024**, *31*, e101120, doi:10.1136/BMJHCI-2024-101120.
44. Sharma, M. The Impact of AI on Healthcare Jobs: Will Automation Replace Doctors. *American Journal of Data Mining and Knowledge Discovery 2024, Volume 9, Page 32* **2024**, *9*, 32–35, doi:10.11648/J.AJDMKD.20240902.11.
45. Artificial Intelligence Act: Council Calls for Promoting Safe AI That Respects Fundamental Rights - Consilium Available online: https://www.consilium.europa.eu/en/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/ (accessed on 13 April 2023).