AI

MDPI

Article

Embedding Fear in Medical AI: A Risk-Averse Framework for Safety and Ethics

Andrej Thurzo [1,2,*] and Vladimír Thurzo [3]

[1] Department of Orthodontics, Regenerative and Forensic Dentistry, Medical Faculty, Comenius University in Bratislava, 81102 Bratislava, Slovakia
[2] Department of Medical Education and Simulations, Faculty of Medicine, Comenius University in Bratislava, 81499 Bratislava, Slovakia
[3] Faculty of Roman Catholic Theology of Cyril and Methodius, Comenius University in Bratislava, 81458 Bratislava, Slovakia; thurzo1@uniba.sk
* Correspondence: thurzo3@uniba.sk

Abstract: In today's high-stakes arenas—from healthcare to defense—algorithms are advancing at an unprecedented pace, yet they still lack a crucial element of human decision-making: an instinctive caution that helps prevent harm. Inspired by both the protective reflexes seen in military robotics and the human amygdala's role in threat detection, we introduce a novel idea: an integrated module that acts as an internal "caution system". This module does not experience emotion in the human sense; rather, it serves as an embedded safeguard that continuously assesses uncertainty and triggers protective measures whenever potential dangers arise. Our proposed framework combines several established techniques. It uses Bayesian methods to continuously estimate the likelihood of adverse outcomes, applies reinforcement learning strategies with penalties for choices that might lead to harmful results, and incorporates layers of human oversight to review decisions when needed. The result is a system that mirrors the prudence and measured judgment of experienced clinicians—hesitating and recalibrating its actions when the data are ambiguous, much like a doctor would rely on both intuition and expertise to prevent errors. We call on computer scientists, healthcare professionals, and policymakers to collaborate in refining and testing this approach. Through joint research, pilot projects, and robust regulatory guidelines, we aim to ensure that advanced computational systems can combine speed and precision with an inherent predisposition toward protecting human life. Ultimately, by embedding this cautionary module, the framework is expected to significantly reduce AI-induced risks and enhance patient safety and trust in medical AI systems. It seems inevitable for future superintelligent AI systems in medicine to possess emotion-like processes.

Keywords: harm avoidance; high-stakes AI systems; amygdala-inspired architecture; clinical decision support; ChatGPT; AI agents; agentic AI; trustworthy artificial intelligence; ethical AI; AGI; GSI; superintelligence; autonomous weapon systems; LAWS; autonomous decision-making; human oversight; diagnostic error mitigation; cautious automation

check for updates

Academic Editor: Ioannis Kakkos

Received: 11 April 2025
Revised: 9 May 2025
Accepted: 13 May 2025
Published: 14 May 2025

Citation: Thurzo, A.; Thurzo, V. Embedding Fear in Medical AI: A Risk-Averse Framework for Safety and Ethics. *AI* **2025**, *6*, 101. https://doi.org/10.3390/ai6050101

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Context and Motivation

Artificial intelligence already outperforms humans in speed and precision, yet it lacks the instinctive caution that keeps clinicians from taking life-altering risks. Inspired by the protective reflexes of military robotics and the threat-monitoring role of the human

amygdala, we propose a computational "fear module" for medical AI: a continuously running safety layer that inflates the weight of catastrophic outcomes, pauses or recalibrates decisions when uncertainty is high, and escalates edge cases for human review.

For example, in a 2022 case, an AI diagnostic tool failed to flag an atypical presentation of stroke, contributing to a delayed treatment. A "fear" mechanism attuned to such uncertainty might have raised an alert for human review, potentially preventing the oversight. Similarly, an oncology decision-support AI was reported to recommend an unsafe chemotherapy dosage—a cautious AI with a fear module could have caught the excessive risk and prompted a dosage check [1–3].

Simply put, today's AI can sift through data faster than any clinician, but it still misses something every good doctor has—a gut sense of when to slow down. Borrowing from the way our own amygdala keeps us out of danger—and from the "better safe than sorry" routines already built into military robots—we suggest giving medical AI its own built-in sense of caution. Think of it as a digital reflex: a background process that blinks red when the odds of serious harm climb, takes a breath before acting on shaky data, and calls in a human whenever the situation feels too risky. Using familiar tools like Bayesian risk estimates, "penalty" points for near-misses, and a hard rule that a person gets the final say, this "fear module" could let AI keep its trademark speed while adding the bedside prudence patients and clinicians trust—an essential upgrade as we head toward ever-more-powerful, autonomous systems.

Currently the extreme pace of technological advances in the field of human–robot–AI opens topics previously known only from science-fiction literature. Human–robot interaction will be defined by developments in AI. Skills learned by robots will not be taught by humans, and a biomimetic approach will be employed widely even in domains of emotion-aware interactions [4–11].

1.2. Aim and Scope

This paper's aim is to outline a conceptual framework for integrating a "fear instinct" into AI, particularly in the medical field, and to delineate the scope of this approach. Here, "fear" is used as a metaphor for a precautionary principle, not as a literal emotional state. Instead of engaging in philosophical debates about qualia, our focus is on practical risk aversion.

Drawing on interdisciplinary insights from computer science, ethics, medicine, and even military AI, we propose embedding fear-like mechanisms into AI's core decision-making architecture. Using techniques such as Bayesian risk thresholds and penalty-driven learning, this approach aims to mitigate errors, biases, and unintended consequences while balancing caution with the need for effective outcomes. By promoting transparency and trust in healthcare applications, our framework encourages collaboration among scientists, clinicians, and policymakers to develop guidelines for safety-driven AI. Ultimately, the goal is to bridge human intuition and clinical expertise with the precision of AI, ensuring that future systems remain both powerful and aligned with our core values [12].

This work is guided by the following research questions:

(1) Can embedding a "fear instinct" in a medical AI system significantly enhance patient safety, and how would such a mechanism be implemented and evaluated in practice? (We hypothesize that it can by reducing harmful errors.)
(2) What would an AI "fear" mechanism look like operationally, and how can it be integrated into existing AI architectures? (We propose a framework and draw parallels to human fear responses.)
(3) What are the ethical, technical, and practical implications of such a system in healthcare? (We examine potential benefits, risks, and requirements for deployment.)

1.3. Ethical and Practical Imperatives for Risk-Sensitive Medical AI

Risk awareness in medical systems is crucial, similarly to military systems [13–15]. AI systems in healthcare need to be aware of potential risks and uncertainties to ensure patient safety and ethical use [16,17]. Implementing some form of risk sensitivity or risk aversion could help address concerns about AI errors, biases, and unintended consequences. Medical AI agents will need to balance the benefits and risks regardless of their extreme scientific knowledge. A risk-averse approach could help mitigate potential harms but should not completely inhibit beneficial AI applications [17,18].

Trust and transparency will remain crucial in the future of medical care. Patients and healthcare providers need to be able to trust AI systems [19–22]. Implementing risk aversion or explainable emotional states could potentially increase transparency and trust, but the "black box" nature of many AI systems remains a challenge [23,24].

Human oversight of medical AI agents is a probable scenario for the next decade, which might be different from military AI systems. In general, most studies emphasize the importance of human oversight and final decision-making authority for medical AI [25]. Fully autonomous emotional AI may not be appropriate in high-stakes medical contexts. Implementing emotional states or fear in AI raises complex ethical questions about machine consciousness and the nature of AI decision-making [26]. A more pragmatic, risk-sensitive approach focused on safety may be more suitable. Current regulatory frameworks are not adequate for emotionally aware or highly autonomous medical AI [16,27]. Dissemination of AI to medical fields along the pipeline of radiography [28] resulted in the necessity for ethical frameworks [29]. Risk-averse implementations may be easier to regulate and validate. High-stake AI agents not exclusively deciding matters of life and death are relevant for consideration of subjective emotion-like frameworks. For example, treatment planning employing subjective views includes facial evaluation for aesthetics [30,31], sensitive rare genetic disease detection [32,33], or nuances for patient compliance coaching [34].

Similarly, AI-driven bioengineering applications, such as personalized scaffold design, highlight the need for risk-sensitive AI to ensure patient safety. A recently introduced workflow integrates AI auto-segmentation of CBCT scans with 4D temperature-responsive resins to create adaptive scaffolds tailored to individual anatomies, enhancing tissue regeneration [35]. This precision underscores the potential of AI in personalized medicine but also necessitates embedded safety mechanisms to mitigate risks, such as errors in anatomical modeling or material adaptation, aligning with the need for a risk-averse framework in medical AI.

This article sparks a bold discussion: What if AI could "fear"—not in the human sense, but as an embedded safeguard that guides decisions to avoid harm? Drawing on insights from medicine, robotics, and ethics, we explore how an AI "fear instinct," similar to the human amygdala, could revolutionize safety in high-stakes environments like healthcare. We also consider how lessons from military AI can help develop systems that always put patient safety first.

1.4. Paper Structure

The remainder of this paper is organized as follows. Section 2 provides the conceptual foundations, including background on fear in biological and artificial contexts. Section 3 articulates the rationale for embedding a fear mechanism in medical AI, while Section 4 presents our proposed framework in detail, drawing analogies to human neural processes and outlining the AI architecture. Section 5 discusses implementation approaches and technical considerations, and Section 6 addresses counterarguments and ethical implications of the framework. We then explore broader implications and future directions in Section 7,

and finally, Section 8 concludes this paper with a vision for collaborative development of safe AI.

2. Conceptual Foundations

2.1. Fear in Biological and Artificial Systems

Fear in biological organisms is defined as an adaptive emotional and physiological response to perceived threats or dangers. Fear is a fundamental emotion that evolved to promote survival and reproductive fitness by protecting organisms from predation and other ecological threats [36]. The amygdala plays a central role in processing fear across animal species. It is part of a complex neural network involving the hypothalamus and brainstem circuits [37,38]. Fear triggers rapid physiological changes, including increased arousal and autonomic and neuroendocrine activation [39]. Fear serves for risk aversion as an internal signal of danger, categorizing the world into familiar (safe) and strange (potentially dangerous) environments [40]. The fear response prepares organisms for defense or avoidance of threats, promoting survival—self-preservation [41]. Fear, as a harm avoidance mechanism, can be innate or learned, allowing animals to detect and rapidly respond to threatening stimuli [42].

Recent studies in computational neuroscience have advanced our understanding of fear through sophisticated modeling techniques. For instance, Bălan et al. [43] utilized machine learning on EEG and peripheral data to classify fear levels, achieving high accuracy and providing valuable insights into the neural mechanisms underlying fear [43]. Additionally, Yamamori (2023) reviewed computational perspectives on human fear and anxiety, emphasizing the role of reinforcement learning in threat prediction and avoidance [44]. Abend et al. (2022) further contributed by linking threat learning to anxiety and neuroanatomy using computational models, offering potential pathways for clinical applications [45].

In essence, fear is a crucial biological mechanism for risk aversion, self-preservation, and harm avoidance across species, deeply rooted in neural circuits that have been conserved throughout evolution.

2.2. Fear in Artificial Systems

Various concepts on "fear" or "caution" in autonomous systems are known from literature. Probably the most famous of Asimov's science-fiction novels, the Laws of Robotics [46], published 75 years ago, affected many. However, these fictional laws, while poetic and thought-provoking, fall short in addressing the real-world AI complexities. Instead, widely recognized frameworks like OECD AI Principles [47] provide a more grounded and globally respected foundation for guiding AI's development and applications.

Unlike humans, the AI does not see a patient's suffering or feel adrenaline; its "fear" is activated by detecting alarmingly high-risk metrics in incoming data. Fear, applied to AI as a metaphor, is a critical safety feature—much more than simple Bayesian risk assessment. It offers a holistic, evolution-inspired approach to caution. While it demands an interdisciplinary effort from psychology, neuroscience, engineering, and ethics, its benefit is clear: an AI that acts with built-in "common sense" about danger. This framework does not view fear as an emotion but as a real-time harm-avoidance process that carries out the following:

- Detects potential risks using advanced computational models [48].
- Prioritizes caution when uncertainty is high, integrating diverse inputs like human senses [43].
- Triggers adaptive responses—seeking further input, escalating to human oversight, or choosing a safer path [49].

In the field of affective robotics, modeling fear is crucial for developing robots that can interact effectively with humans. Graterol et al. (2021) introduced an emotion detection framework for social robots using NLP transformers, which successfully identifies emotions, including fear, from textual data [50]. Moreover, Rizzi et al. (2017) proposed the Situation-Aware Fear Learning (SAFEL) model, which integrates situation-aware systems with neuroscientific fear-learning mechanisms, enabling robots to anticipate and respond to fear-inducing situations [51]. These advancements illustrate the potential of artificial systems to emulate and react to fear, drawing from biological inspirations.

In short, embedding fear-like mechanisms in AI can make it safer, more effective, and better aligned with human needs, offering a design principle for the next generation of responsible machines.

2.3. Translating Genuine Fear into Medical AI

First, it is important to differentiate "simulated fear"—surface-level algorithms—from a "genuine fear mechanism" that permeates an AI's decision-making architecture. AI's fear mechanism will be purely data-driven. To illustrate this distinction, we compare how a deeply embedded "fear response" in an AI agent diverges from the relatively superficial constraints seen in current large language models (LLMs). Table 1 summarizes the key differences between a superficial "simulated fear" (such as simple safety filters in AI) and a genuine, system-wide "fear" mechanism embedded in an AI's decision core. This comparison highlights how our proposed deep integration of fear differs fundamentally from surface-level safety algorithms.

Table 1. Comparison of superficial "simulated fear" versus genuine, system-wide "fear" embedded deep in the decision core of an AI agent. The table highlights differences in functional integration, scope of influence, adaptability, triggers, and outcomes, showing how a deeply embedded fear mechanism informs the entire decision process, while surface-level algorithms primarily refine how outputs are presented.

Aspect	Fear	Surface-Level Algorithms
Core objective	Aims to prevent harm by prioritizing patient safety deeply tied to high-stakes decision-making.	Focuses on avoiding offensive or harmful language, primarily addressing social norms and sensitivities in communication.
Implementation depth	Operates as a foundational, system-wide mechanism influencing every decision and action, akin to a survival instinct.	Functions as a surface-level filter or constraint applied during response generation.
Risk assessment	Involves continuous risk estimation and harm aversion in dynamic, high-stakes environments (e.g., healthcare or warfare).	Primarily avoids reputational or social harm during text generation.
Adaptation to context	Tailored to specific domains (e.g., medical AI, where harm has direct physical consequences, as well as in lethal autonomous weapon systems (LAWS)).	Generalized across topics to fit broad societal expectations.
Accountability	Designed to trigger specific fail-safes (e.g., escalating to a human when risk is high).	Seeks to refine language but does not affect operational decision-making.

Source: authors' own elaboration.

Fear-like mechanisms in AI are proactive, domain-specific mechanisms embedded in decision-making to prevent tangible harm. In contrast, superficial layers—such as simulated emotional filters—operate reactively to shape output without influencing core processes. Figure 1 illustrates how these layers work together in an AI-driven medical system: deep fear triggers drive risk assessment, reinforcement learning, uncertainty modeling, memory, and human overrides to ensure safety, while superficial filters manage communication.

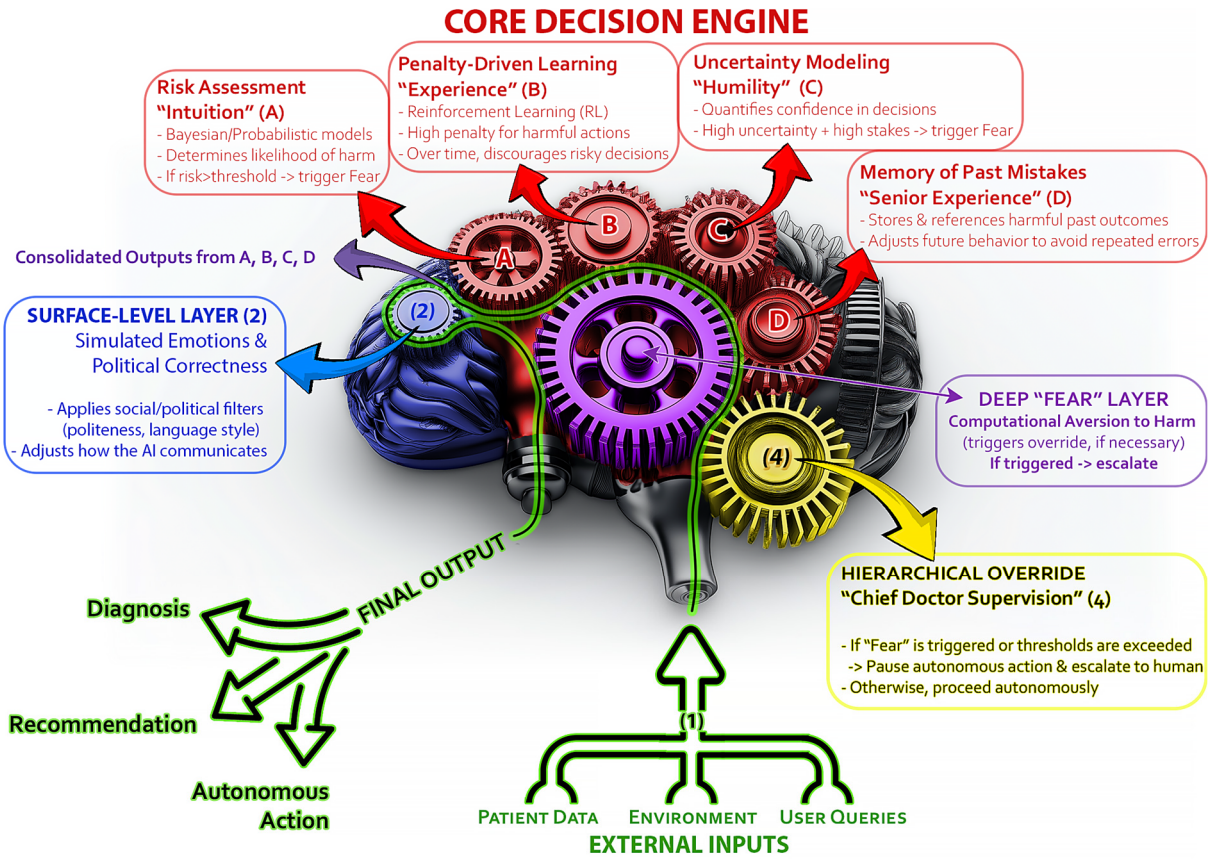


Figure 1. Proposed scheme of medical AI agent architecture, showing how surface-level (blue) and deeply embedded "fear" layers (purple) integrate with risk assessment, reinforcement learning, uncertainty modeling, memory (red), and human oversight (yellow) to ensure safety and effective system. A dedicated "amygdala-like" subsystem (purple) continuously monitors risk signals (e.g., Bayesian models, RL penalties, and uncertainty estimates) and triggers protective actions (pauses, escalations) when thresholds are exceeded—instilling a fundamental aversion to harm beyond superficial checks. Source: authors' own elaboration.

Efforts to translate biological fear mechanisms into artificial systems have led to innovative models applicable to medical AI. For example, the SAFEL model by Rizzi et al. (2017) demonstrates how neuroscientific insights can be applied to create AI systems that predict and avoid threatening situations, a concept that can be adapted to enhance the safety of medical AI by incorporating similar fear-like mechanisms [51].

2.4. "Fearless" Doctor

In human terms, the virtue we seek is not fear itself, but prudence or courage—a midpoint between recklessness and cowardice. Similarly, an AI should not be reckless (completely without fear) nor paralyzed by fear. Our proposed fear module aims to instill a sense of caution roughly analogous to the prudence a good physician exercises.

Consider an AI surgeon assistant facing a procedure with a moderate chance of severe complication but a slightly higher expected benefit. A purely utilitarian calculus might

proceed since expected benefit outweighs risk. A human surgeon with good judgment (courageous but not reckless) might feel a **nagging concern**—an intuition that, say, the numbers do not capture some contextual factors—and possibly opt for a safer approach or at least prepare extra precautions. The fear module is intended to replicate that **safety-oriented intuition** by flagging the situation with a "caution" signal even if the raw calculus is marginally positive.

In medical care, patients prioritize compassion, expertise, and judgment over a "fearless" attitude, which may suggest poor risk assessment. Medical AI, lacking subjective experience, can be called "fearless" since it does not feel or respond to fear. We can infer some reasons why AI systems, including those used in medicine, do not experience fear:

Lack of sentience: Current AI systems, including medical AI agents, are not considered sentient and do not have consciousness or emotional experiences [52].
Absence of biological mechanisms: AI systems lack the biological structures and processes that generate fear responses in humans and animals [49].
Algorithmic decision-making: Medical AI agents operate based on programmed algorithms and data analysis, not emotional responses [53].
No self-preservation instinct: Unlike living beings, AI systems do not have an innate drive for self-preservation that would trigger fear responses [54].
Lack of personal consequences: AI systems do not experience personal consequences from their actions, which often drives fear in humans [55].

While AI systems can be designed to recognize and respond to human emotions, including fear, they do not experience these emotions themselves [56]. It is important to note that as AI technology advances, questions about AI sentience and emotional experiences may become more complex and debated [52].

2.5. "Good" Doctor

Extensive research has examined which qualities define a "good doctor," drawing on perspectives from patients, medical students, and clinicians. While patients often emphasize interpersonal skills such as empathy and communication, physicians may focus more on professional competencies skills [57]. In a 2023 study by Schnelle et al., 86% of respondents reported meeting at least two "exceptionally good" doctors, described primarily in terms of personality, diagnostic ability, and intervention skills. Notably, "listening attentively to patients" emerged as a highly influential characteristic [58]. Common descriptors of a good doctor identified in the literature include empathy, clear communication, respect, and trustworthiness [59]. Published studies have identified several key attributes that frequently define a good doctor. The most frequent adjectives for a "good doctor" were as follows:

Compassionate and empathetic [60,61]
Knowledgeable and competent [60,62]
Good communicator [63]
Patient-centered [57,58,64]
Honest and trustworthy [62]
Respectful [57]

These findings suggest that while there is no single definitive set of adjectives, certain qualities consistently emerge as important in defining a good doctor across various studies and perspectives.

2.6. Technical Components of Fear in AI Medical Agents

We must distinguish between *simulated fear* (surface-level algorithms) and the idea of a *genuine fear* mechanism that influences system-wide decision-making. To implement

metaphorical "fear" as a part of the core decision engine of AI medical agents (Figure 1), subsystems of the "Fear Mechanism" shall be understood.

2.6.1. Integrated Decision-Making and Conflict Resolution

To illustrate how our embedded safety principles operate in concert, we introduce a real clinical scenario: a neurosurgical AI assistant evaluating the clipping of an intracranial aneurysm. In this case, the AI must balance several elements:

- Risk assessment ("intuition"): The AI continuously calculates the probability of adverse outcomes using Bayesian models.
- Penalty-driven learning ("experience"): It adjusts its internal parameters based on past near-miss events, assigning higher penalties to actions that have historically led to complications.
- Uncertainty modeling ("humility"): Through measures like dropout-based uncertainty or Gaussian processes, the system gauges its confidence in a decision.
- Hierarchical overrides ("chief doctor supervision"): When risk thresholds or uncertainty exceed pre-set limits, the AI flags the decision for human review rather than acting autonomously.

For example, if the AI predicts a 7% risk of catastrophic bleeding during aneurysm clipping (exceeding the acceptable risk threshold) and simultaneously registers high uncertainty in its predictive model, the fear module—our integrated harm-avoidance mechanism—activates. Rather than vetoing the procedure outright, the system adjusts its recommendation by triggering a cautionary signal:

$$U = w1 \times R - w2 \times B - w3 \times UQ,$$

where:

R is the computed risk,

B is the expected benefit,

UQ is the uncertainty measure, and

w1, w2, and w3 are weight parameters that can be calibrated based on clinical priorities.

When the composite utility U falls below a defined threshold, the AI does not proceed with the autonomous recommendation but instead escalates the decision to a human clinician. This approach ensures that even if the individual components (risk, uncertainty, and experience) conflict—such as when a high benefit might normally justify risk—the system's built-in caution (fear) tempers the decision by demanding human oversight.

This integrated framework not only grounds each component in its own computational methodology but also provides a mechanism for resolving conflicts among them. We draw on ideas from recent work in ethical decision-making in AI, where published proof of concept supports this multi-criteria approach [65]. By combining weighted assessments with threshold-based triggers, our model ensures that no single factor—whether it be raw risk, uncertainty, or benefit—dominates the decision-making process. Instead, a balanced, context-sensitive "fear" response emerges, one that aligns with the overarching goal of patient safety.

AI medical agents are changing the roles of medical practitioners, diagnostic and therapeutic workflows, and even the education of humans and robots, as well as the conduct of research and its publication. The need for human oversight extends beyond clinical applications to the realm of medical research, where AI-driven tools are reshaping evidence synthesis. The role of AI agents in revolutionizing even literature reviews, highlighting their ability to rapidly synthesize vast datasets but noting challenges like inaccuracies and access limitations [66]. Such models are essential for medical AI systems, where errors

in evidence synthesis could lead to misguided clinical decisions, reinforcing the need for risk-averse frameworks that prioritize safety and trust.

2.6.2. Subsystems of the "Fear Mechanism" in Autonomous AI Medical Agents

Typically, several components could be combined. These often have an analogy in real medicine. Our proposed fear mechanism is built from several interlocking components that mirror how experienced clinicians make decisions:

- Risk assessment ("intuition"):

The system uses probabilistic models (such as Bayesian networks) to calculate the likelihood of harm. When risk exceeds a preset threshold, a fear response is triggered.

- Penalty-driven learning ("experience"):

Reinforcement learning is employed with high penalties for actions that previously led to adverse outcomes. This process teaches the AI to avoid decisions that could cause harm.

- Uncertainty modeling ("humility"):

Methods like Gaussian processes or dropout-based uncertainty estimation help the AI gauge its confidence in a decision. When uncertainty is high in a critical situation, the system exhibits fear-like behavior. Recent EEG studies, such as Kakkos et al. [67], demonstrate how functional cortical connectivity reorganizes under varying mental workloads, supporting the inclusion of neural-inspired uncertainty modeling in our framework.

- Hierarchical overrides (medical supervision):

When risk or uncertainty exceeds acceptable limits, the system halts autonomous action and escalates the decision to a human clinician, much like a chief doctor would intervene.

- Memory of past mistakes ("experience"):

The AI maintains a memory of past errors. For example, if it previously misdiagnosed a condition due to incomplete data, it would exercise extra caution when faced with similar scenarios. Abooelzahab et al. [68] combined CNN and LSTM to classify EEG signals for brain disorders, reinforcing how memory of past errors can inform cautious decision-making.

Clinical scenario example:

Imagine a neurosurgical AI assistant evaluating whether to recommend clipping an intracranial aneurysm. The AI uses its risk assessment module to compute the probability of catastrophic bleeding and its uncertainty modeling to assess confidence in its prediction. Suppose the system determines a 7% risk—above the safe threshold—and finds significant uncertainty due to atypical patient anatomy. Drawing on past experiences stored in its memory and reinforced by penalty-driven learning, the AI's fear module activates. Instead of automatically recommending surgery, it flags the case for human review. The neurosurgeon can then consider the AI's warning alongside their own judgment to decide on the best course of action.

> *For instance, if the 'do no harm'/fear principle strongly advises against surgery (surpassing a risk threshold), but the beneficence principle and patient's informed consent together outweigh that (indicating high potential benefit and willingness), the AI might still recommend proceeding, albeit with cautionary notes. On the other hand, if the fear signal is overwhelming and expected benefit marginal, the AI would recommend against the surgery.*

The visualization diagram, shown in Figure 2, maps the flow of data through the subsystems and how components interact and how the AI's fear mechanism integrates

multiple safety principles, resolves conflicts, and ultimately triggers human oversight when necessary. The data input stage at the top (e.g., clinical measurements, imaging, and lab results) feeds into the subsystems as parallel modules (risk assessment represents Bayesian or probabilistic risk calculation, penalty-driven learning shows reinforcement learning updates based on historical outcomes, uncertainty modeling depicts uncertainty estimation using methods like Gaussian processes or dropout, and memory of past mistakes indicates stored knowledge of prior errors, used to adjust current decision thresholds). The central "arbitration module" displays a simplified mathematical expression, $U = w1 \times R - w2 \times B - w3 \times UQ$, where R stands for the computed risk, B for expected benefit, UQ for uncertainty, and w1, w2, and w3 are adjustable weights. Escalation activates when U falls below a preset safety threshold. It illustrates how the AI medical agent's "fear mechanism" integrates with risk assessment, penalty-driven learning, uncertainty modeling, hierarchical overrides, and memory of past mistakes, ultimately guiding the decision-making process and allowing for human intervention when thresholds are exceeded. Consider the neurosurgical scenario according to the diagram:

1. Input: Patient anatomy and aneurysm data are fed into the system. This includes imaging, lab results, and other clinical measurements that characterize the patient's condition.
2. Risk calculation: The risk assessment module uses Bayesian models to compute the probability of adverse events. In our scenario, it determines there is a 7% risk of catastrophic bleeding during aneurysm clipping—a value that exceeds our predefined safety threshold.
3. Uncertainty flag: Concurrently, the uncertainty modeling module evaluates the data using techniques like Gaussian processes. Due to atypical patient anatomy, it registers a high level of uncertainty in its prediction. This signals that the system's confidence in the risk calculation is lower than desired.
4. Override trigger: Given the elevated risk and uncertainty, the arbitration module combines these inputs using a weighted utility function. When the calculated utility falls below the safe threshold, the system triggers its hierarchical override. Instead of proceeding autonomously, it flags the case and escalates the decision to a human clinician for review.

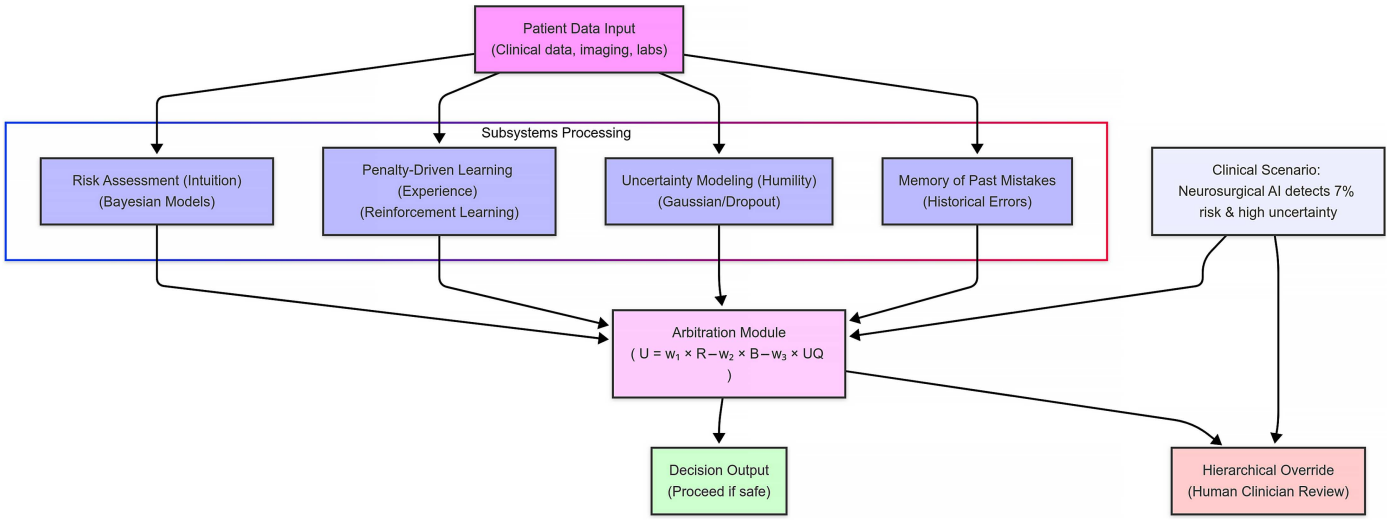


Figure 2. Overview of the AI "fear" mechanism architecture of the proposed module in a medical AI system. Source: authors' own elaboration.

Clinical Scenario Example (for Figure 2): Imagine a neurosurgical AI assistant evaluating whether to recommend clipping an intracranial aneurysm. The AI's risk assessment module computes a 7% probability of catastrophic bleeding (above the safe threshold), and the uncertainty module notes low confidence due to unusual patient anatomy. The fear module integrates these signals and "decides" to pause; it does not automatically proceed with a surgery recommendation. Instead, it flags this case for human review, alerting the surgeon to the elevated risk and uncertainty. The surgeon, seeing the AI's warning, can weigh the risks and decide if surgery is truly warranted. This example (illustrated by Figure 2) shows how each component of the fear mechanism comes into play to ensure caution.

2.6.3. Bayesian Networks and Risk Thresholds

Bayesian networks are simple diagrams that show how variables—like symptoms and diseases—connect and influence each other. Each dot (or node) represents a variable, and arrows between them show their probabilistic relationships. By updating these probabilities with new evidence (like test results), these networks help make smart decisions in uncertain situations, such as medical diagnoses or risk evaluations.

Bayesian risk thresholds use this probability approach to decide when a risk is big enough to act—like calling in a human or stopping an automated process. Here is how it works:

- Bayesian probability: This is about updating your guess about something (e.g., "this treatment could harm") as new info (e.g., symptoms) comes in, using Bayes' theorem.
- Setting a threshold: You pick a danger cutoff—say, a 5% chance of harm. If the network calculates a risk above that, it triggers a safety step, like pausing AI for human review.
- Ongoing updates: The system keeps recalculating as new data (e.g., lab results) arrive. If the risk crosses the threshold, it acts instantly. Task-independent EEG signatures, as identified by Kakkos et al. [69], suggest that alpha and theta power shifts could refine our Bayesian risk thresholds, enhancing real-time caution triggers.
- Medical example: Imagine an AI spotting a serious illness. If it sees a 5%+ chance of a harmful misdiagnosis, it might ask for more tests or a doctor's opinion.
- Finding balance: A low threshold might cause too many alerts, slowing things down. A high one might miss real dangers. Getting this right is key.

In short, Bayesian risk thresholds let AI weigh new info and act cautiously when risks hit a critical point, blending data-driven decisions with.

3. What Is the Rationale for Embedding Fear in Medical AI?

There is currently no direct evidence that an AI with an embedded sense of "fear" outperforms one without it. However, analogous approaches in AI safety provide encouragement: for example, risk-aware clinical decision systems that incorporate human oversight have shown improved safety outcomes. The literature emphasizes responsible development and ethics in AI, which parallels our proposal to integrate a cautionary mechanism. This gap in research—no direct exploration of an AI "fear" module—highlights a promising area for investigation [70–74].

There is currently no direct evidence to suggest that an AI system designed with an embedded sense of "fear" for patient life and safety would outperform one without such a feature. While patient safety is undeniably critical, the existing literature does not really explore the concept of embedding fear into AI systems. Instead, it focuses on promoting responsible development, emphasizing ethical considerations, and fostering human-AI collaboration to ensure both safety and optimal care [75–77]. This gap in research highlights

a potential area for exploration: whether incorporating mechanisms that resemble fear or heightened caution could enhance AI decision-making in high-stakes medical scenarios.

3.1. Patient Safety and Error Reduction

Fear is linked with risk management, and AI with experiences of "fear of causing harm" might be less likely to propose risky treatments or miss critical warning signs. Studies show that risk perception plays a significant role in AI adoption [18]. Embedding a computational form of "fear" in medical AI ensures the system remains vigilant about potentially harmful decisions, mirroring the *Hippocratic Oath* [78] principle to "do no harm". While ethical frameworks provide guidelines on what is permissible or prohibited, an internal "fear" mechanism adds a proactive safeguard—raising alerts or deferring to human oversight when risk thresholds are exceeded. This parallels how clinicians exercise caution and escalate uncertain or high-risk cases to more experienced colleagues, thereby complementing formal ethical mandates with a built-in aversion to risky or harmful actions. Youldash et al. [79] showed deep learning's efficacy in early diabetic retinopathy detection, supporting our module's role in pre-emptively identifying risks to patient safety.

3.2. Enhancing Decision-Making

There are various decision-making scenarios in general medicine (diagnosis, prescription, triage, etc.) where a fear mechanism could alert the AI to high-risk outcomes, prompting more caution or additional consultation. AI can analyze vast amounts of data quickly, providing clinicians with valuable insights to support their decision-making process [80,81]. This augmentation allows for more informed and timely clinical decisions. Although clinical decision support systems are being introduced more frequently in various healthcare domains, significant differences among clinicians persist, necessitating more nuanced approaches [82]. For instance, Kakkos et al. [83] used AttentionUNet for parotid gland segmentation in radiotherapy, illustrating how deep learning can enhance precision and safety in clinical decisions.

The fear module can be implemented within a reinforcement learning (RL) paradigm. Practically, an AI's "fear" response could be realized by shaping its reward function: actions predicted to carry a high risk of harm incur a heavy penalty, analogous to an emotional aversion. In this way, the module does not introduce an entirely new mechanism but rather augments the reward structure to prioritize safety. The difference is one of emphasis and transparency—by calling it a fear module, we explicitly designate a part of the algorithm to focus on worst-case outcomes, ensuring they are given sufficient weight during learning and decision-making.

Recent studies demonstrate that even in the absence of emotions, AI can be imbued with ethical reasoning grounded in established principles. For example, Meier et al. developed a proof-of-concept clinical ethics algorithm that encodes core medical ethics principles (beneficence, non-maleficence, patient autonomy) to guide decision-making [65]. While this principle-based approach showed promise, the authors noted its limitations in capturing the nuances of human moral judgment and questioned whether sensitive ethical decisions should be entrusted to machines [65]. Building on such work, we advocate embedding a *fear-like* safeguard into medical AI's architecture as an additional layer of protection. This "fear" module would function as an intrinsic "do no harm" governor, biasing the AI against high-risk or harmful actions in real time. By integrating a fear-inspired mechanism, an AI can hesitate or withdraw from potentially dangerous decisions, complementing frameworks like Meier's by operationalizing a constant aversion to causing harm. In practice, this means an autonomous clinical system would internally check extreme or uncertain recommendations against a survival instinct–like threshold, much as a physician's intuition

might urge caution. *Embedding fear* in this manner aligns with the Hippocratic ethos and existing ethical AI paradigms, ensuring that patient safety remains the paramount priority in AI-driven decision-making [65].

3.3. Human-AI Collaboration

Fear-driven caution in AI might improve trust among clinicians and patients by demonstrating "awareness" of potential harm. For example, Camlet et al. [84] found ChatGPT-4 effective in periodontology exams, suggesting that cautious AI responses can enhance clinician trust and collaboration. The literature suggests that AI should complement human skills in healthcare, not replace them [85]. The focus is on leveraging AI's strengths while maintaining human judgment and empathy. As there are various potential downsides of an overactive fear module, AI's fear module must be calibrated. If the fear module is too sensitive, it could indeed recommend avoiding procedures that are in the patient's best interest—an outcome to avoid. We therefore propose setting an upper bound on the module's influence: it should flag high-risk actions for review rather than outright veto them, ensuring that necessary treatments are not automatically rejected. For example, a neurosurgical AI assistant might "feel" fear about clipping an aneurysm due to the risk of rupture; AI would issue a strong warning rather than unilaterally cancel the procedure. A human surgeon can then weigh this warning against their own judgment.

4. Proposed Mechanism: Designing "Fear" in AI at a Deep Level

4.1. Biological Analogy and "Amygdala-Like" Subsystem

To ensure safety-critical decision-making in medical AI, systems must include mechanisms that fundamentally avoid high-risk actions. By drawing inspiration from biology, we can conceptualize an "amygdala-like" subsystem to fulfill this role. In humans, the amygdala is central to processing fear, rapidly evaluating threats, and triggering protective responses. Kandpal et al. [86] used deep learning to segment ischemic stroke lesions on CT perfusion maps, illustrating how AI can mirror biological threat detection for safety-critical tasks. Similarly, an AI system's "amygdala-like" component would function as a risk evaluator, continuously monitoring for potential harm and influencing decision pathways to prioritize safety. Despite similarities in cognitive skills, AI systems process information fundamentally differently from biological cognition [87].

Our framework acknowledges that the traditional "do no harm" principle must be balanced with beneficence and patient autonomy. In clinical practice, many interventions—such as surgical incisions, endodontic procedures, or chemotherapy side effects—unavoidably cause harm in pursuit of a greater benefit. Drawing inspiration from the human amygdala, which rapidly assesses threats and triggers protective responses, we propose an "amygdala-like" subsystem for medical AI. This subsystem continuously monitors risk by integrating quantitative indicators (e.g., Bayesian models, reinforcement learning penalties, and uncertainty estimates) to identify high-risk actions or ambiguous data. Importantly, a "fearful" AI is not intended to automatically cancel high-risk procedures. Instead, it raises a flag by highlighting significant risks and prompting careful human reevaluation. For instance, if the system "feels" uneasy about a procedure, it will issue a strong warning rather than outright forbidding the intervention. Should the procedure promise life-saving benefits—and if the patient consents, the clinical team can override the AI's caution.

By embedding this calibrated fear mechanism, our AI supports sound clinical judgment without undermining the necessary balance between risk and reward. The final decision always incorporates patient values and informed consent, ensuring that the AI's caution complements rather than replaces human decision-making. This nuanced trade-

off—minimizing harm while promoting overall well-being—aligns our approach with established clinical ethics [88].

Humans and AI currently possess complementary capabilities that can together surpass the collective intelligence of either alone; while AI may not have direct biological analogues, it can work synergistically with human cognition [89]. LLMs already perform at or above human levels in identifying indirect requests or false beliefs, albeit they struggle with more nuanced social cognition like detecting faux pas [90]. Biological analogies of an "amygdala-like" subsystem remain underexplored, albeit ongoing research aims to deepen our understanding of these parallels and potentially bridge the gap between artificial and biological intelligence [91].

4.2. Core Components of the Proposed AI Architecture

Researchers have already developed an AI framework for autonomous driving inspired by the amygdala's role in eliciting defensive responses against threats [92]. The proposed AI architecture mirrors elements of human decision-making, structured into interconnected components:

- Knowledge component: Gathers and evaluates information to determine task feasibility and identify relevant data.
- Planning/prediction component: Simulates strategies to achieve objectives and evaluates potential outcomes, both favorable and adverse. It also sends risk signals to the amygdala-like subsystem based on predicted consequences.
- Amygdala-like subsystem: Acts as a central risk evaluator and also continuously monitors threats and influences the system by adjusting plans or triggering protective actions when necessary.
- Execution component: Implements the selected, risk-adjusted plan or decision.
- Monitoring/feedback component: Observes outcomes, flags deviations, and provides feedback to both the amygdala-like subsystem and the learning/adjustment component.
- Learning/adjustment component: Updates internal models and strategies based on feedback to improve future decision-making. Similarly, Lee et al. [93] employed generative models with padding to handle incomplete multi-omics data, suggesting a strategy for our architecture to adapt to sparse clinical inputs.

Our design allows the "fear" response to be contextually calibrated. In practice, this means the AI can adjust its risk threshold depending on the scenario. For life-threatening emergencies where immediate action is paramount, the fear module's trigger threshold might be set higher to avoid unnecessary delays—essentially, the AI "dampens" its fear to remain effective under pressure. Conversely, in routine or high-uncertainty cases (such as diagnosing a rare disease or handling very vulnerable patients), the threshold could be lowered, making the AI more readily cautious. This contextual adaptation ensures that the AI's level of fear (caution) is appropriate to the severity and specificity of each case, preventing one-size-fits-all behavior.

Real-World Workflow Example of Integration into Clinical Workflow

To illustrate, consider an imaging diagnostic AI used in radiology. Our fear module would attach to this AI and continuously evaluate the AI's confidence and the potential risk of misdiagnosis. If the AI is uncertain about a critical finding (e.g., a possible tumor on a CT scan) and the risk of missing a diagnosis is high, the fear mechanism triggers a caution: the AI output is not given as a definitive result but is flagged for review by a radiologist before a final report. Likewise, in a clinical decision support system for ICU medicine, the fear module would monitor treatment suggestions. If an AI recommends an aggressive therapy that exceeds preset risk thresholds (for example, a medication dose that

could cause dangerous side effects in a frail patient), the system would halt and prompt the attending physician to double-check the recommendation. In both cases, the penalties and risk thresholds we propose are embedded at critical decision points in the pipeline to provide real-time safety checks.

4.3. Integration and Functionality in Analogy to Human Emotions

Our system operates in a continuous feedback loop, where outputs from monitoring and learning are reintegrated into the knowledge base, allowing it to evolve and improve over time. By embedding an amygdala-like subsystem throughout this loop, the system consistently mitigates risk and adapts based on past experiences—much like human emotions guide our decision-making. In humans, emotions such as fear and elation influence our responses to risk and opportunity; analogously, our AI uses computational equivalents of these responses to drive safer, more adaptive decisions [91]. Human decision-making is influenced by emotions, such as fear when predicting negative outcomes or elation when foreseeing positive results. Analogously, the AI system includes computational equivalents of these responses:

- Fear analog: Strong penalty functions for harmful outcomes are embedded deeply within the decision process to discourage risky actions.
- Elation analog: Reward signals reinforce beneficial outcomes, encouraging optimal decision paths.

These analogues are not mere afterthoughts but are fundamental to the system's decision-making framework. They shape both learning processes and output generation, ensuring the AI's objectives align with safety and efficacy.

Figure 3 illustrates the proposed decision-making process with the amygdala-like subsystem at its core. The diagram outlines how risk signals generated during planning and feedback stages flow into the subsystem, triggering adjustments and ensuring safe outcomes.

1. Knowledge component: Collects data and evaluates task feasibility.
2. Planning/prediction component: Simulates strategies and sends risk signals.
3. Amygdala-like subsystem: Continuously monitors for threats and adjusts plans.
4. Execution component: Implements risk-adjusted decisions.
5. Monitoring/feedback component: Observes results and informs other components.
6. Learning/adjustment component: Refines models to improve future decisions.

Consider a triage AI in emergency care with an embedded fear module (as outlined in Figure 3). If the AI is processing a patient's symptoms and vital signs and contemplates a risky intervention (say, administering thrombolytics in a borderline case of stroke), the fear subsystem calculates the risk of adverse outcome vs. benefit. In a situation where data is ambiguous, the AI might generate an alert and require a physician's input rather than immediately proceeding. Figure 3's flowchart is annotated with this example in mind, demonstrating how the decision flows from initial input to final human confirmation when the "fear" threshold is triggered.

The proposed "amygdala-like" subsystem is integral to creating medical AI systems capable of safety-critical decision-making. Especially for the era of generalized superintelligence that will be represented by decision processes beyond human comprehension, even if they would be explainable. By embedding continuous risk monitoring and aversion mechanisms into every stage of the process, the architecture ensures proactive harm avoidance and supports reliable, ethical, and effective outcomes in high-stakes environments.

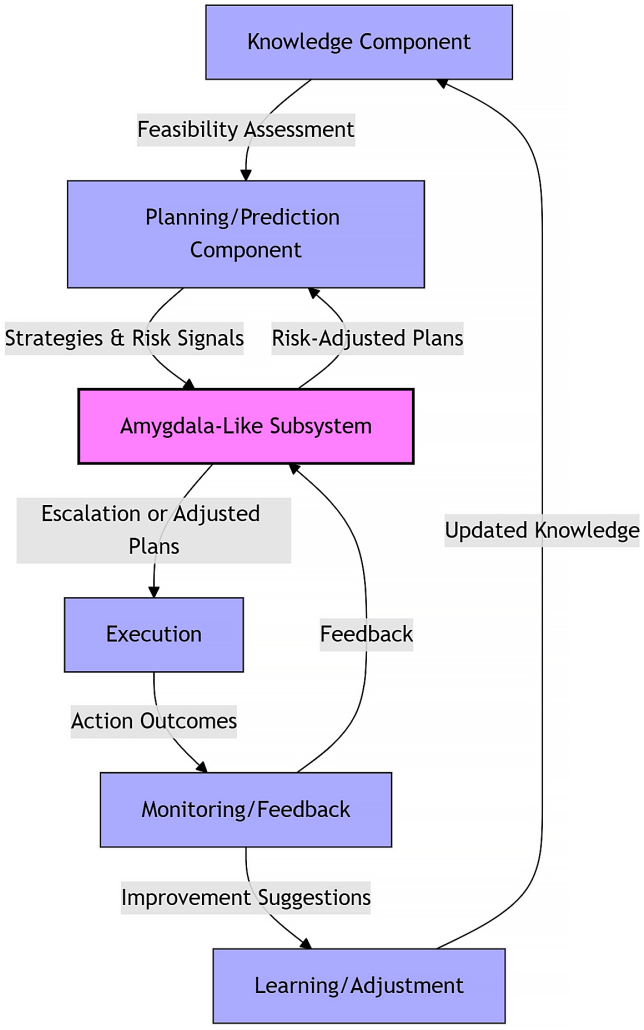


Figure 3. Decision-making process with amygdala-like subsystem in medical AI. Source: authors' own elaboration.

4.4. Black Boxes in Human Heads vs. Superintelligent AI Agents Beyond Human Comprehension

It is worth noting that humans, even the most brilliant medical professionals, often function as "black boxes". Their decisions are shaped by intuition, personal experience, and unconscious biases, making their reasoning difficult, if not impossible, to fully explain. As general AI continues to advance towards superintelligence, its decision-making processes will soon outpace our ability to understand them entirely. These systems will evolve abstraction layers and cognitive capabilities far surpassing human comprehension, making their operations opaque even to those who created them.

While we might still have some control over the initial training processes, this too will likely shift into AI's domain. Superintelligent systems will not just optimize themselves; they will design and build new generations of AI, creating a self-perpetuating cycle of improvement [94]. This marks a fundamental transition—from human-led design to a future dominated by AI-driven evolution.

Even though their inner workings might become a mystery to us, implementing safeguards like an amygdala-like subsystem remains critical. These mechanisms embed ethical principles and risk aversion at the very core of AI's functioning, ensuring that it stays aligned with human values. By establishing these guardrails, humanity can keep some measure of control, guiding superintelligent AI towards outcomes that benefit society—even as we struggle to grasp the intricacies of how these systems achieve them.

The article by Stiefel et al. [95] argues that current semiconductor computing technology poses significant barriers to AGI or ASI development, primarily due to energy constraints. However, the recently introduced Lyapunov optimization theory proposes a novel quantum machine learning framework for stabilizing computation offloading in next-generation MEC systems, which significantly outperforms conventional offloading approaches, improving network throughput by up to 30% and reducing power consumption by over 20% [96].

4.5. Embodied Emotions vs. Simulated Emotions

An "embodied" fear mechanism is deeply integrated into an AI's decision-making, influencing learning, uncertainty handling, and action selection much like a human nervous system. In contrast, "simulated" fear uses superficial, rule-based triggers (e.g., numerical thresholds) to pause or escalate decisions. Mutawa [97] demonstrated that attention-based CNN-BLSTM models excel in classifying COVID-19 genomic sequences, offering a blueprint for integrating such techniques into our risk-sensitive framework. As Yann LeCun notes, future AI assistants may need such emotion-like capabilities for effective reasoning and planning. For instance, battle robots often employ survival heuristics that override actions when a threat is detected, illustrating how an embodied response is integral rather than merely a final score [15,98].

4.6. AI Autonomously Deciding Matters of Life and Death

In the coming decades, AI may increasingly decide life-and-death matters across healthcare, warfare, and even state-sanctioned contexts—a shift that poses significant ethical risks [13,15,98].

Today, AI functions as a decision-support tool with human oversight remaining critical in high-stakes environments [99,100]. Autonomous AI raises issues of accountability, bias, and responsibility. For example, while medical drones can expedite emergency care [101,102], military AI is already reshaping roles by automating strategic and tactical decisions [103]. Future autonomous systems are expected to differ markedly: military agents emphasize efficiency, detachment, and aggression, whereas medical AI must be compassionate, cautious, ethical, analytical, supportive, and adaptive [75,76,104–108]. Figure 4 and Table 2 illustrate these contrasting attributes.

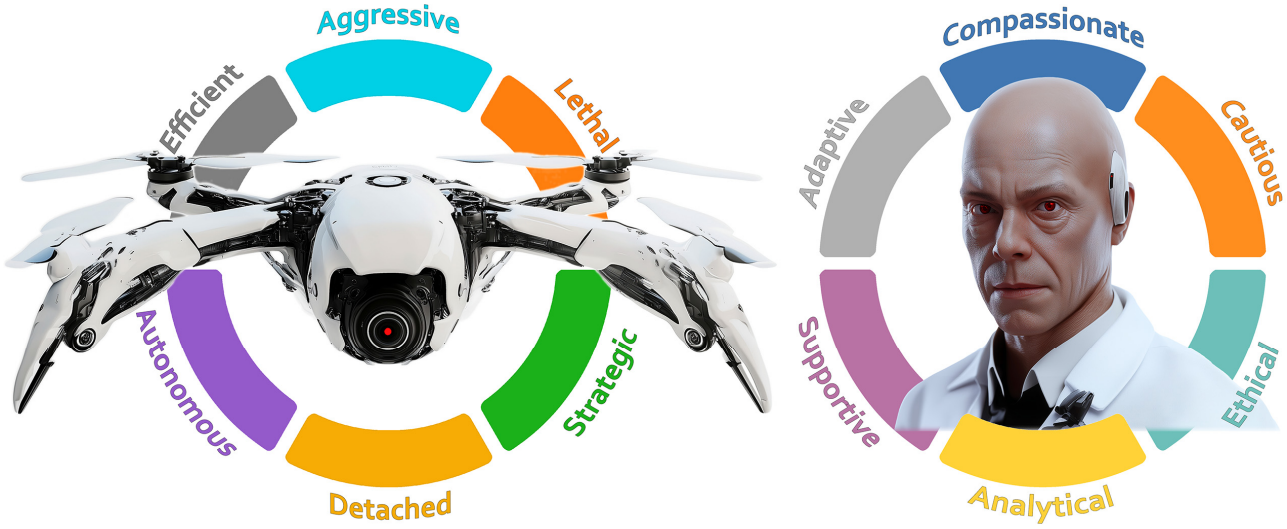


Figure 4. AI lethal autonomous weapon systems and autonomous AI medical agents will have different priorities in their primary attributes. Source: authors' own elaboration.

Table 2. Distinctions in decision-making AI agents between AI weapon systems and AI medical systems.

	Military AI	Medical AI
Lack of human-like decision-making in AI agents	The current autonomous systems, including LAWS, cannot replicate human emotional and moral considerations such as fear, empathy, or remorse.	In a medical context, this facet underscores the necessity for AI systems to incorporate mechanisms analogous to human emotions (like fear) to prioritize patient safety and ethical treatment decisions.
"Fear of killing" as a performance factor	Human soldiers often fail in combat due to a moral aversion to killing, driven by fear or ethical reservations. In contrast, AI-autonomous systems lack this "fear" and are thus more effective but also more ethically detached.	In medicine, embedding a "fear of harming the patient" could serve as a counterweight to prevent AI from making detached, overly utilitarian, fearless decisions that risk patient health or dignity.
Mimicking vs. embodying emotions	LAWS can currently only mimic moral actions and cannot genuinely feel emotions like remorse or fear, which are tied to accountability and human dignity.	Mere superficial simulations of caution (e.g., rule-based safeguards) may not suffice. Instead, an embodied "fear" mechanism—integrated at a deep architectural level of decision core—could enhance AI's decision-making, ensuring actions align with the ethical principle of "do no harm".
Autonomy and accountability	A crucial issue remains that autonomous AI robots cannot be held accountable for their actions, as they lack the capacity to comprehend punishment or reward.	The introduction of "fear" as a computational mechanism might enable better self-regulation and reduce the risk of decisions that harm patients without solving this accountability issue.
Ethical implications of decision-making autonomy	Autonomous systems making life-and-death decisions in warfare highlight the risks of dehumanization and moral detachment. Relinquishing the decision to kill to machines fundamentally undermines human dignity.	Fear-driven caution in AI could act as a safeguard against dehumanized care and overly mechanistic decisions, promoting a more ethically aligned approach to treatment.

Source: authors' own elaboration.

Looking ahead, as generalized superintelligence emerges, human decision-making may shift toward a symbolic role. In such a future, embedding deep, fear-like (i.e., caution-driven) frameworks in AI could help safeguard human life by ensuring decisions remain ethically grounded.

Military robotics research shows that instilling a "fear" of killing can prevent indiscriminate actions [14]. In military systems, this mechanism helps balance lethal efficiency with ethical constraints, while in medicine, the priority is preventing harm to patients. A fear module in medical AI would act to inhibit risky actions—even in ambiguous situations—by adding a cautionary bias, not by replicating combat behavior.

Debate continues over whether emotion-like layers in AI can create true moral behavior or merely simulate ethical responses. However, such mechanisms may help narrow accountability gaps. For instance, comprehensive data logging in autonomous systems has reduced wrongful killings in law enforcement [98,109]. Nonetheless, accountability in medical AI remains challenging, and even with a fear module, human oversight is essential to address residual risks [110,111].

4.7. Practical Considerations

We recognize that embedding a fear module is not without challenges. One concern is latency: introducing a deliberative "pause" for caution could delay decisions in time-critical scenarios. Our framework would need to ensure that this delay is minimal, perhaps by allowing the module to act in parallel or by setting higher risk thresholds for intervention when response time is critical (see discussion on calibration below).

Another consideration is computational load. Continuously estimating probabilities and monitoring multiple safety metrics (risk, uncertainty, past errors) requires computational resources. In settings like an AI-assisted surgery robot, resource constraints or processing delays must be accounted for. We suggest that efficient algorithms (e.g., streamlined Bayesian updating or hardware acceleration for neural nets) be used to keep the system responsive. We have added these points to acknowledge that practical deployment of a "fearful" AI demands careful engineering to avoid hindering the very safety it seeks to provide.

Our framework balances AI autonomy with human oversight by making oversight conditional and as non-intrusive as possible. In real-time critical scenarios, we envision a mechanism where the AI, upon triggering the fear response, can still execute any safe fallback actions (like maintaining current treatment and monitoring patient status) while simultaneously notifying the human team. The idea is that the AI does not "freeze"; it continues low-risk support actions until a human intervenes. Moreover, the thresholds are set such that truly immediate life-and-death reflexes are not delayed—the AI will only defer to humans in scenarios where a brief pause is acceptable or where a human's insight is absolutely required. For instance, an AI in an emergency room might administer basic life support measures automatically but will alert a physician before making a high-risk decision like administering a thrombolytic if it calculates borderline risk. This ensures that when time is critical, the AI does not become a bottleneck; the built-in caution complements rather than hinders urgent care.

Quantifying and Calibrating Caution

We define a quantitative threshold for the fear module—for instance, the system might be set to intervene if the estimated risk of serious harm exceeds X% while confidence is below Y%. These X and Y can be adjusted. In practice, calibration would involve testing the AI on historical cases or simulations: we would raise or lower the threshold X until we achieve a desired balance (e.g., the module only triggers in truly high-risk cases and avoids triggering in false alarm situations). One strategy is to treat this like a sensitivity-specificity trade-off problem, similar to tuning a medical diagnostic test. In critical, time-sensitive scenarios (e.g., emergency battlefield triage by an AI medic), X could be higher to ensure the AI does not over-intervene and slow decision-making unless absolutely necessary. In contrast, for non-urgent scenarios, X could be lower to prioritize safety. By calibrating these parameters per context, we maintain efficiency in emergencies without sacrificing the protective benefits of caution.

5. Potential Implementation Approaches

An amygdala-like fear mechanism could be integrated into existing ML frameworks by combining Bayesian risk modules—which continuously estimate harm probabilities—with reinforcement learning that penalizes risky actions. When the Bayesian module detects high risk or the RL component flags unsafe behavior, the system would trigger a defensive response (e.g., pausing decisions or escalating to human oversight). Although purely theoretical at this stage, this approach offers a path to embed proactive safety checks into AI systems.

This concept aligns with key principles of safe AI design. For instance, Durán and Jongsma explore risk thresholds and uncertainty modeling using "reliability indicators" [112], while Barea Mendoza et al. highlight the importance of human oversight in AI [113]. Similarly, penalty-driven reinforcement—akin to modifications in reward functions for sepsis treatment as described by Festor et al.—can discourage harmful actions.

Using, for instance, proximal policy optimization (PPO) or other reinforcement learning algorithms that naturally handle safety constraints by penalization.

Challenges include balancing risk aversion with efficacy, as excessive caution may hinder valuable insights [114], and addressing the opacity of AI systems, which complicates transparency [24].

Additionally, defining "unsafe" scenarios will require continuous refinement informed by clinical expertise.

Training a fear-aware AI will necessitate a carefully designed learning environment, as illustrated by the six mechanisms outlined in Figure 5. Training mechanisms employed to instill "fear" or cautious behavior in AI systems are made from six key methods: (a) reinforcement learning with fear penalties, (b) adversarial training, (c) multi-agent systems, (d) emotion-inspired metrics, (e) real-world data with risk labels, and (f) risk-averse policy learning.

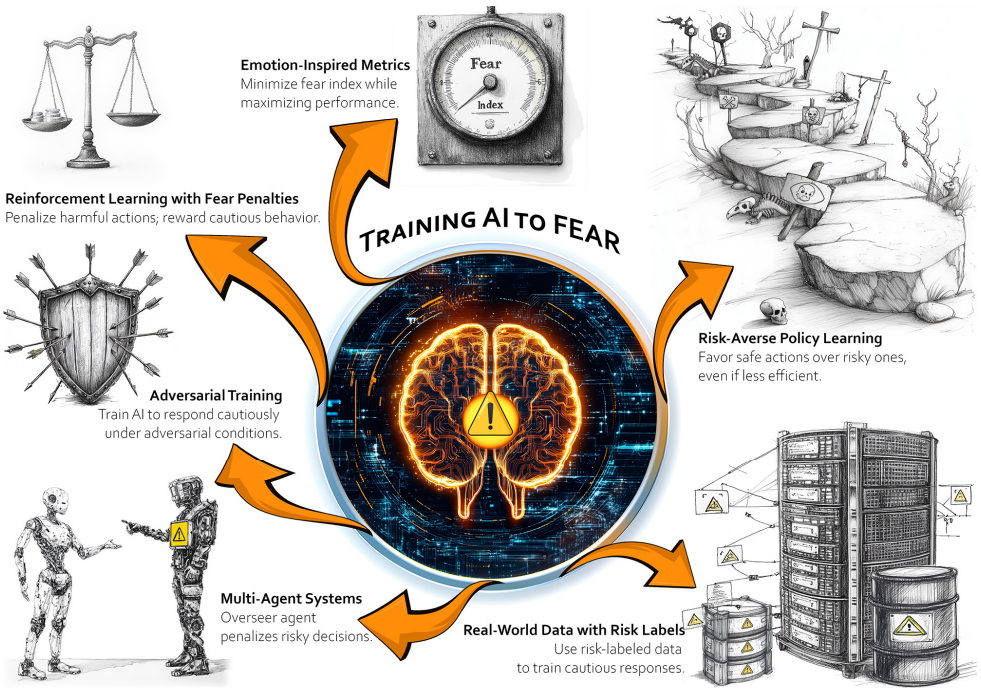


Figure 5. Visualization of the AI "fear" training framework, linking the central concept to six key methods. Source: authors' own elaboration.

1. Reinforcement learning with fear penalties: Define a reward function that heavily penalizes harmful actions or outcomes. Also reward cautious or consultative behavior in high-risk scenarios (e.g., asking for additional input, escalating to a human).
2. Adversarial training: Expose the AI to adversarial conditions where harm is likely (e.g., deliberately ambiguous patient data) and train it to recognize and respond with caution.
3. Multi-agent systems: Train the AI alongside a "human-like overseer" agent. When the AI makes risky decisions, the overseer "punishes" those actions, simulating an external accountability mechanism.
4. Emotion-inspired metrics: Borrow metrics from affective computing to create proxies for fear, such as a "fear index" tied to the probability of harm or system uncertainty. Train the AI to minimize this index while maximizing overall performance.
5. Real-world data with risk labels: Use datasets labelled with degrees of risk or harm (e.g., patient outcomes, near misses in healthcare settings). Train the AI to associate these labels with cautionary behavior.

6. Risk-averse policy learning: Use risk-sensitive policies in reinforcement learning that explicitly avoid actions with even a small chance of severe harm, favoring safer, albeit less efficient, decisions.

Balancing caution and efficiency: An important consideration is how "fearful" the AI should be. A more conservative AI (low risk tolerance) will err on the side of caution—potentially preventing even borderline-risk actions—which maximizes safety but may reduce the system's efficiency or willingness to act. This could manifest as frequent alerts or unnecessary deferrals to humans (a form of overcautiousness). In contrast, a **proactive AI** with a higher risk threshold will intervene less often, thus acting more autonomously and efficiently, but runs a higher risk of overlooking infrequent dangers. Our framework acknowledges this trade-off. In practice, the optimal setting will depend on the application: for instance, in routine diagnostic tasks, one might allow a more proactive AI to avoid bogging down clinicians, whereas in high-stakes surgical planning, a conservative setting might be chosen to prioritize patient safety. We have incorporated this analysis to make clear that the "fear" mechanism's sensitivity must be carefully calibrated to each use case.

According to the recent research of Willem et al. [88], it seems ethics must be embedded in AI development from the start, and for this implementation, an interdisciplinary collaboration is essential [88]. An interesting observation from this paper is that ethical "frictions" can drive innovation in AI development.

6. Counterarguments and Ethical Considerations

6.1. Responsibility Dilemmas

Current AI medical systems do not "feel" fear as humans do; they use advanced risk assessment, uncertainty modeling, and penalty-driven learning—all purely algorithmic processes without subjective experience. When experts refer to "AI fear," they mean computational strategies (e.g., Bayesian thresholds, reward penalties, or anomaly detectors) that trigger conservative actions or defer decisions to human experts in high-risk scenarios. These engineered responses mimic the cautious behavior of medical professionals rather than reflecting genuine emotion.

Moreover, literature on fully autonomous AI making life-and-death decisions is scarce. Existing studies focus mainly on concerns like accountability, responsibility, and bias in AI healthcare, rather than on whether AI could ever truly experience fear.

A 2024 study by Arbelaez Ossa et al. [77] emphasized that AI must genuinely benefit healthcare stakeholders and align with context-specific healthcare practices. They highlighted the need for a systemic, proactive perspective on ethical AI development that considers objectives, actors, and context. Another 2024 publication by Kim et al. [115] outlined a six-stage ethical governance system for healthcare AI research in South Korea. This approach aims to address ethical requirements throughout the AI development and implementation process.

A 2023 study by Petersson et al. [99] explored healthcare professionals' perceptions of ethical aspects in implementing an AI application to predict mortality risk in emergency departments. They identified conflicts related to each ethical principle, including autonomy, beneficence, non-maleficence, justice, explicability, and professional governance.

A "fearful" AI that prioritizes patient safety by avoiding harm can improve decision-making by flagging high-risk cases, but it may also become overly cautious, potentially delaying critical care or rejecting treatments that carry acceptable risks. This raises an important question: should an AI override a physician's decision if its risk calculations are too conservative? Striking the right balance between caution and efficacy is crucial.

Moreover, ethical challenges in high-stakes fields often stem from responsibility gaps. Müller [98] argues that systems that consistently evade accountability should be

avoided [98,109]. In healthcare, robust ethical frameworks are needed to clearly define the roles of AI creators and users, ensuring accountability even in rare, ambiguous situations [116], ensuring accountability in most cases while addressing rare instances where responsibility cannot be clearly assigned. By minimizing responsibility gaps and fostering transparency, we can harness the benefits of AI in medicine while upholding ethical standards.

Embedding a pseudo-instinct like fear in AI introduces ethical questions about control and responsibility. If the AI effectively "overrides" a human's initial action (even by forcing a second thought), we must consider consent and authority. We assert that the AI should not have unilateral power to deny treatment—ethically, it serves as a safeguard, but a human (doctor and/or patient) should have the right to override the AI's hesitation if they accept the risks. This maintains human agency. Another implication is liability: if the AI's caution leads to a negative outcome (e.g., delaying a treatment that could have helped), it is unclear who is at fault. To navigate this, we suggest that such AI systems be used in a transparent way where their recommendations are logged and can be evaluated in hindsight and where clinical governance bodies set guidelines for when to heed the AI's warnings. The ethical design principle here is that the AI's pseudo-instinct should enhance safety without eroding human decision-making or moral responsibility.

Consider a case where the AI's fear module prevents a surgical robot from making a certain incision, and as a result, the patient's condition worsens. Even if the AI was trying to avoid risk, in hindsight this decision was detrimental. Who is accountable—the developers of the AI, the medical team supervising, or the AI itself? This example underscores the complexity of attributing responsibility when an autonomous safety mechanism is in play. We suggest that clear protocols be established so that ultimate decision authority (and liability) remains with human practitioners or institutions, not the AI.

6.2. Risk of Adversarial Manipulation

A fear-based AI could be gamed or exploited if adversaries figure out how to consistently trigger its fear mechanism, causing it to avoid certain actions or escalate decisions unnecessarily. Such vulnerabilities may lead to system paralysis or excessive caution in critical situations, with dire real-world implications. This raises further concerns about liability and responsibility: if the AI's "fear response" delays a necessary intervention or produces harmful outcomes, it is unclear whether the developers, healthcare providers, or the AI system itself should be held accountable.

Training algorithms on datasets often implicitly contains risk assessments or even emotional reactions such as fear, especially if these are human-derived datasets. When AI trains on such clinicians' historical decisions, it naturally inherits a degree of risk aversion present in the data—effectively an implicit "fear" learned from human experience. For instance, if orthodontists in the training set often avoided a certain high-risk orthognathic approach in planning the operation, the AI might also be hesitant to recommend it. However, relying solely on implicit learning has drawbacks: the AI's caution level would be an opaque by-product of the data and could vary with data quality. By contrast, an explicit fear module allows us to control and tune the level of risk aversion and ensure it generalizes even to scenarios not well-represented in the training data.

A sophisticated adversary could target the fear mechanism itself. For example, an attacker might launch a poisoning attack, subtly altering input data (e.g., a patient's sensor readings) to consistently push the AI into a "fearful" state, causing it to unnecessarily halt treatments (a denial-of-service effect on care). Conversely, an attacker could try to mask true danger signals so the AI remains complacent when it should be fearful. Such security threats to medical AI systems are documented in the literature, and our framework would

need defenses to be viable. Possible countermeasures include validating critical inputs via multiple independent channels (to make it hard to spoof all of them), employing anomaly detection algorithms to flag unusual input patterns, and using robust learning techniques that can resist poisoning as AI transforms everything in medical practice, from science to education [117]. Ensuring data integrity and cybersecurity for the fear module is thus an essential part of our proposal's implementation. We have added references to recent work on AI security and emphasized that without such measures, a "fearful" AI could itself be manipulated into either paralysis or recklessness by malicious actors.

As a concrete scenario, an attacker might perform a data poisoning attack on the AI's inputs—for instance, subtly altering a patient's data to simulate a very high-risk reading. A naive fear module might repeatedly halt the AI's actions due to these false alarms, effectively paralyzing the system. This highlights the need for security measures to ensure the fear mechanism cannot be easily misled by malicious inputs.

6.3. Risk of Overcautiousness

While "fear" is a vital safeguard against harmful decisions, relying solely on it may not yield truly ethical, patient-centered outcomes. Complementary emotional responses—like "hope" to inspire optimism or "empathy" to keep patient well-being at the forefront—can balance excessive risk aversion and enrich decision-making. By combining multiple emotional and logical mechanisms, an AI can not only avoid harm but also actively promote positive outcomes and respect patient autonomy, as shown in Figure 6. Mitigation strategies for an overly cautious AI include fine-tuning risk thresholds and combining fear with confidence estimates [87,89,107]. Training the AI to adapt its fear response based on context and the severity of potential harm, along with pairing the mechanism with interpretable models (e.g., decision-tree logic), enhances explainability and builds clinician trust.

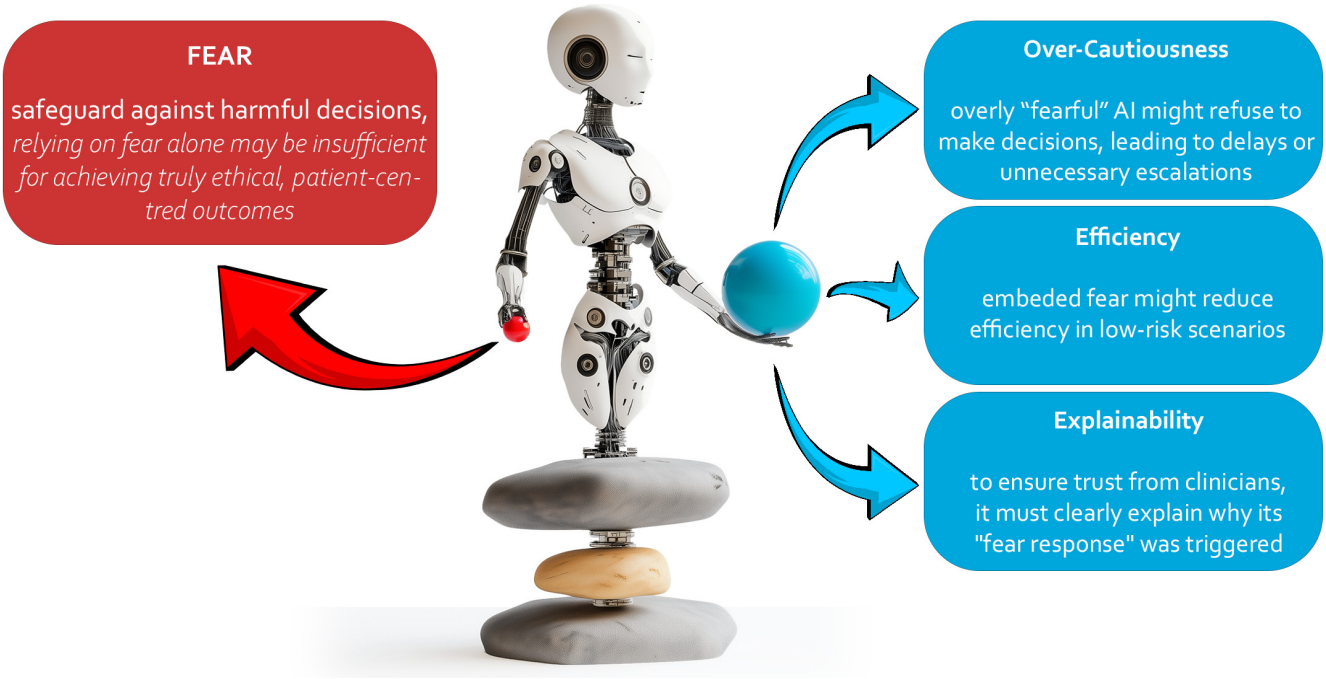


Figure 6. Framework for balancing fear with other emotional and logical mechanisms in an AI medical agent. Source: authors' own elaboration.

For example, imagine an AI emergency triage system that "panics" (activates fear) at ambiguous vital signs and insists on waiting for a doctor's confirmation before initiating a life-saving intervention. If the human is not immediately available, this delay caused

by an overcautious AI could worsen patient outcomes. Thus, while caution is generally beneficial, excessive caution in time-critical situations can be dangerous.

7. Broader Implications and Future Directions

7.1. The Current Horizon

Generative AI has evolved from handling simple, repetitive tasks to addressing complex reasoning challenges in healthcare. Recent models—such as OpenAI's o1-preview that uses chain-of-thought reasoning—demonstrate impressive performance in clinical diagnostics and patient management [118,119]. Gupta et al. [119] further showed LLMs like GPT-4 excel in symptom-based diagnostics, underscoring AI's growing role in clinical reasoning and safety. This evolution is reshaping healthcare education by bridging human expertise with AI's rapidly expanding capabilities [120,121]. The emerging concept of an "AI co-scientist" further illustrates how AI can collaborate with human researchers to accelerate breakthroughs in patient care and safety [122–124]. Despite these advances, current systems still fall short of achieving true artificial general intelligence (AGI) that seamlessly transfers knowledge across diverse domains.

7.2. Agentic AI and Ethical Considerations

As AI systems gain autonomy, they offer potential improvements in efficiency and accuracy but also pose significant ethical challenges. Autonomous decision-making raises concerns about accountability, fairness, and the erosion of human oversight in critical healthcare decisions [125]. It is essential to implement robust oversight and ethical frameworks that ensure these systems complement rather than replace the fundamental principle of "first, do no harm".

7.3. Regulatory and Policy Considerations

The future deployment of AI, especially those incorporating risk aversion mechanisms like fear modules, requires interdisciplinary collaboration. Regulatory frameworks should establish clear guidelines—including criteria for "fear thresholds," testing protocols, and escalation procedures to human oversight—to ensure safety and transparency. Coordinated efforts among AI ethicists, healthcare professionals, and policymakers are critical for setting standards that include regular audits and bias assessments [126].

7.4. High-Stakes Medical AI and General Superintelligence

General superintelligence (GSI) promises to revolutionize medicine with superior speed, accuracy, and scalability, yet it also introduces risks such as diminished human oversight and ethical misalignment. Embedding emotion-like frameworks—such as fear and empathy—into AI architectures (akin to an amygdala-like subsystem) could help these systems navigate complex moral dilemmas. Such safeguards would enable GSI to balance logical decision-making with ethical considerations, ensuring patient care remains aligned with human values even as AI operates beyond full human comprehension [127].

7.4.1. Interpretability and Trust

For the fear module to be practically useful, clinicians must trust and understand it. We emphasize that interpretability is crucial: the system should provide understandable reasons for its cautionary interventions (for example, displaying which risk factor or input triggered the alert). By making the AI's "thought process" transparent, we ensure that human operators can follow the logic and validate or override it confidently. This interpretability will, in turn, foster trust—clinicians are more likely to adopt the system if they feel it behaves in a predictable and explainable way. Future development should include a

user-friendly interface where the AI's internal risk assessments are summarized in human-readable form whenever the fear module activates. We have added this consideration to highlight that technical reliability alone is not enough; user trust ultimately determines the framework's acceptance in clinical practice.

7.4.2. Limitations

This study has several limitations. First, our framework is conceptual and has not yet been validated with empirical data or clinical trials; the effectiveness of an AI "fear" module remains to be proven in practice. Second, our discussions assume that risk probabilities can be accurately determined—in reality, AI prediction uncertainty might be greater than we expect, which could affect the module's reliability. Third, there may be contexts (outside those we examined) where an embedded fear mechanism is less applicable or could interact with other AI behaviors in unpredictable ways. Finally, we have not addressed the full spectrum of implementation costs, such as the need for extensive training data or regulatory hurdles. These limitations will be important to address in future research as we move from concept to real-world application of this framework.

7.5. Applications Beyond Medicine

The concept of embedding fear-based mechanisms is not limited to healthcare. In autonomous vehicles, a "fear of collision" can promote more cautious driving; in policing, a "fear of injustice" can lead to stronger safeguards; and in disaster response, a "fear of endangerment" may trigger timely, life-saving decisions. Table 3 summarizes how fear-based AI can contribute to safer and more ethical outcomes across various high-stakes domains, from industrial automation to cybersecurity [110,111,128].

Table 3. Potential Benefits of Fear-Based AI Mechanisms Across Various High-Stakes Environments.

Environment	Fear-Based Benefit
Autonomous vehicles	Fear of accidents could help prioritize safety over efficiency, especially in complex or ambiguous traffic scenarios.
Financial systems and trading algorithms	Fear of catastrophic losses or economic instability could guide risk-sensitive behavior in automated trading systems.
Disaster response and search-and-rescue AI	Fear of endangering human survivors could improve decision-making during rescue operations.
Environmental monitoring and preservation	AI could use fear-based constraints to avoid actions that could inadvertently cause environmental damage while managing ecosystems or climate interventions.
Industrial automation and robotics	Fear of causing workplace accidents or equipment damage could lead to safer interactions between robots and humans in factories.
Childcare and elderly care systems	Fear of neglect or causing harm could ensure higher vigilance and better care from AI systems assisting vulnerable populations.
Cybersecurity systems	Fear of failing to detect a breach or causing unintended harm through aggressive defense mechanisms could lead to more balanced and cautious strategies.

Source: authors' own elaboration.

7.5.1. Applications in Justice Systems

Extending our concept to the judicial domain, one could imagine AI systems with a "fear of injustice". For instance, an AI algorithm assisting in parole decisions or sentencing could be equipped with a module that triggers caution when a potential decision might lead to an unfair or irreversible outcome (e.g., wrongful conviction or disproportionate

sentence). Such a system would flag these high-stakes decisions for extra review or err on the side of leniency when uncertainty is high—analogous to how our medical AI fears causing harm. This "fear of injustice" could act as a safeguard against algorithmic bias or overconfidence, ensuring the AI upholds principles of fairness and due process. Including this perspective suggests that risk-aware AI frameworks may benefit many sectors, from healthcare to law, wherever ethical stakes are high.

7.5.2. Regulatory Oversight and Certification

Before a "cautious AI" can be trusted in high-risk fields like healthcare or defense, it will need to undergo rigorous evaluation and certification. We anticipate that regulatory agencies (such as the FDA for medical AI) will need to establish new guidelines to specifically assess these systems' safety modules. This could involve scenario-based testing where the AI's fear responses are evaluated against a wide range of conditions to ensure it acts appropriately. Additionally, interdisciplinary review panels should be instituted, combining technical experts, clinicians, ethicists, and policymakers. These panels can oversee pilot deployments, review the AI's decision logs, and make recommendations for tuning or improving the system. They will also be crucial in maintaining public trust—an independent body certifying that the AI operates within agreed safety norms. In sum, strong regulatory and interdisciplinary oversight is not only recommended but likely mandatory for implementing such a paradigm in real-world, high-stakes settings.

7.6. Open Research Questions

Building on our proposal, we identify several avenues for future interdisciplinary research:

- What methodologies can rigorously calibrate the "fear threshold" in AI systems across varying domains (medicine, law, etc.) to balance safety and effectiveness?
- How can we empirically evaluate the performance of a fear-based AI module in live or simulated high-stakes environments (e.g., through clinical trials or sandbox testing)?
- What frameworks should regulators develop to certify and monitor AI systems with embedded risk aversion instincts?
- How will the integration of such pseudo-instincts in AI impact human trust and teamwork with AI (e.g., will users become too reliant on or conversely suspicious of an AI that "feels" fear)?

Addressing these questions will require collaboration across computer science, medicine, ethics, and policy to fully realize and assess the potential of "fearful" AI.

8. Conclusions and Vision

The idea of an AI system that can "fear" might seem like something out of science fiction, yet it offers a groundbreaking way to rethink safety and ethics in medicine. By designing AI systems with built-in risk aversion and real-time harm avoidance—much like the rapid threat detection of the human amygdala—we have the potential to rebuild trust in automated systems [21,22]. It is important to stress that introducing a fear-based module is not meant to imply that AI without it must necessarily be utilitarian. Instead, this mechanism can work alongside deontological or other ethical models, enhancing overall decision-making. However, moving from a theoretical "fear mechanism" to practical application will require a true team effort: computer scientists need to develop the technical foundations, roboticists must fine-tune the feedback loops and hardware, ethicists should address the moral and societal issues, and healthcare professionals must test these systems in real-world clinical settings. Only by working together can we establish clear guidelines and standards for creating AI that is both powerful and principled—capable of fast, precise

problem-solving while staying deeply committed to patient well-being. This collaborative approach will pave the way for a future where AI not only excels at solving problems but also respects and protects human life.

Our proposed framework offers a distinctive contribution to AI safety beyond traditional approaches. Classical AI alignment focuses on ensuring AI objectives remain in line with human values, whereas our "fear module" concept introduces a real-time, internalized safety check within the AI's decision process. In effect, we embed an instinct-like aversion to harm into the AI's core, which is fundamentally different from merely postulating safety goals—it provides an active, moment-to-moment bias toward caution that traditional alignment methods lack.

We argue that AI medical agents may not reach their full potential unless they incorporate this kind of "fear" as an extra safeguard. While fully autonomous agents might perform well in some tasks, handing over life-and-death decisions without a built-in cautionary mechanism risks overlooking the essential human quality of prudence. By adding a cautious, fear-inspired bias to the decision-making process, we acknowledge a complex trade-off that raises important ethical and practical questions. Although direct research on embedding "fear" in AI—especially in medical or military contexts—is still limited, ongoing discussions about ethical AI, emotional computing, and autonomous decision-making lay a solid groundwork for this concept. These discussions help us develop a framework that integrates fear-like mechanisms into medical AI, ultimately improving both patient safety and ethical decision-making.

8.1. Seven Key Insights

1. Embedding "fear" as a core safeguard: Integrating an amygdala-like subsystem that instinctively identifies and avoids harmful actions adds an extra layer of protection, echoing the natural, reflexive threat responses seen in humans.
2. Risk-averse AI enhances trust: When AI demonstrates caution—or "fear"—it can earn greater trust from healthcare professionals and patients, leading to more constructive collaboration.
3. Translating military lessons to medicine: Insights from lethal autonomous weapon systems (LAWS) illustrate the dangers of "fearless" AI in life-and-death situations, highlighting the need for safety-focused designs in clinical settings where patient care is paramount.
4. From surface-level filters to core emotional layers: Rather than relying solely on superficial safeguards (like politically correct outputs), embedding deeply rooted, domain-specific "fear" mechanisms ensures that risk sensitivity is an integral part of every decision-making step.
5. Balanced architecture to overcome overcaution: While a fear-based approach can reduce harm, it must be carefully balanced. Excessive caution might block beneficial treatments, so establishing clear thresholds and human override options is critical.
6. Blueprint for future high-stakes AI: Our framework—which combines risk modeling, penalty-driven learning, and memory of past errors—serves as a roadmap for developing emotionally aware AI systems not only in medicine but also in other critical fields like autonomous vehicles, law enforcement, and disaster response.
7. Ethical imperative in autonomy: As AI becomes more autonomous, embedding "fear" at its core could act as both a moral and practical safeguard. Importantly, this mechanism can complement other ethical approaches, ensuring that human values and safety remain central in complex, high-stakes environments.

8.2. Call to Action

The concept of "fear-driven" AI in medicine is bold, yet its potential to enhance patient safety and ethical care is too significant to overlook. We call on computer scientists, clinicians, ethicists, and policymakers to join forces—developing new computational models, running clinical pilots, and establishing ethical guidelines—to ensure that this built-in caution mechanism truly protects patients. We conclude by inviting interdisciplinary collaboration on this vision. The development of a "fearful" medical AI will not be achieved by technologists alone—it requires the collective effort of computer scientists, healthcare professionals, neuroscientists, ethicists, and policymakers. We encourage joint research initiatives, cross-disciplinary training, and the establishment of working groups or panels that can pilot this framework in clinical settings and shape guidelines for its ethical deployment. Only through such collaborative efforts can we ensure that advanced AI systems remain safe, trustworthy, and aligned with human values. By working together across disciplines, we can create the next generation of healthcare AI that not only delivers power and precision but also remains firmly aligned with patient well-being and human dignity.

Author Contributions: V.T. and A.T. performed the methodology and wrote the manuscript draft, A.T. and V.T. supervised the manuscript, and V.T. and A.T. revised the original draft and conducted the final revision of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Slovak Research and Development Agency, grants APVV-21-0173 and APVV-23-0509, and the Cultural and Educational Grant Agency of the Ministry of Education and Science of the Slovak Republic (KEGA) 2023 054UK-42023.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: We extend our sincere gratitude to Dušan Meško and Tomáš Havran for their invaluable insights and ongoing academic discussions on the implications of artificial intelligence.

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

References

1. Hisan, U.K.; Amri, M.M. Artificial Intelligence for Human Life: A Critical Opinion from Medical Bioethics Perspective—Part II. *J. Public Health Sci.* **2022**, *1*, 112–130. [CrossRef]
2. Savage, C.H.; Tanwar, M.; Elkassem, A.A.; Sturdivant, A.; Hamki, O.; Sotoudeh, H.; Sirineni, G.; Singhal, A.; Milner, D.; Jones, J.; et al. Prospective Evaluation of Artificial Intelligence Triage of Intracranial Hemorrhage on Noncontrast Head CT Examinations. *Am. J. Roentgenol.* **2024**, *223*, e2431639. [CrossRef] [PubMed]
3. Confronting the Criticisms Facing Watson for Oncology—The ASCO Post. Available online: https://ascopost.com/issues/september-10-2019/confronting-the-criticisms-facing-watson-for-oncology/ (accessed on 2 May 2025).
4. Szabóová, M.; Sarnovský, M.; Krešňáková, V.M.; Machová, K. Emotion Analysis in Human–Robot Interaction. *Electronics* **2020**, *9*, 1761. [CrossRef]
5. Huang, J.Y.; Lee, W.P.; Chen, C.C.; Dong, B.W. Developing Emotion-Aware Human–Robot Dialogues for Domain-Specific and Goal-Oriented Tasks. *Robotics* **2020**, *9*, 31. [CrossRef]
6. Alonso-Martín, F.; Malfaz, M.; Sequeira, J.; Gorostiza, J.F.; Salichs, M.A. A Multimodal Emotion Detection System during Human–Robot Interaction. *Sensors* **2013**, *13*, 15549–15581. [CrossRef] [PubMed]
7. Gutiérrez-Martín, L.; Romero-Perales, E.; de Baranda Andújar, C.S.; Canabal-Benito, M.F.; Rodríguez-Ramos, G.E.; Toro-Flores, R.; López-Ongil, S.; López-Ongil, C. Fear Detection in Multimodal Affective Computing: Physiological Signals versus Catecholamine Concentration. *Sensors* **2022**, *22*, 4023. [CrossRef]
8. Korcsok, B.; Konok, V.; Persa, G.; Faragó, T.; Niitsuma, M.; Miklósi, Á.; Korondi, P.; Baranyi, P.; Gácsi, M. Biologically Inspired Emotional Expressions for Artificial Agents. *Front. Psychol.* **2018**, *9*, 388957. [CrossRef]

9. Spezialetti, M.; Placidi, G.; Rossi, S. Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives. *Front. Robot. AI* **2020**, *7*, 532279. [CrossRef]
10. Fernández-Rodicio, E.; Maroto-Gómez, M.; Castro-González, Á.; Malfaz, M.; Salichs, M. Emotion and Mood Blending in Embodied Artificial Agents: Expressing Affective States in the Mini Social Robot. *Int. J. Soc. Robot.* **2022**, *14*, 1841–1864. [CrossRef]
11. Petrescu, L.; Petrescu, C.; Oprea, A.; Mitruț, O.; Moise, G.; Moldoveanu, A.; Moldoveanu, F. Machine Learning Methods for Fear Classification Based on Physiological Features. *Sensors* **2021**, *21*, 4519. [CrossRef]
12. Faria, D.R.; Godkin, A.L.; da S. Ayrosa, P.P. Advancing Emotionally Aware Child-Robot Interaction with Biophysical Data and Insight-Driven Affective Computing. *Sensors* **2025**, *25*, 1161. [CrossRef] [PubMed]
13. (PDF) Use of Autonomous Systems in Military Operations. Available online: https://www.academia.edu/126270728/Use_of_Autonomous_Systems_in_Military_Operations (accessed on 12 May 2025).
14. Taddeo, M.; Blanchard, A. A Comparative Analysis of the Definitions of Autonomous Weapons Systems. *Sci. Eng. Ethics* **2022**, *28*, 37. [CrossRef]
15. Kahn, L. Lethal Autonomous Weapon Systems and Respect for Human Dignity. *Front. Big Data* **2022**, *5*, 999293. [CrossRef]
16. Cestonaro, C.; Delicati, A.; Marcante, B.; Caenazzo, L.; Tozzo, P. Defining Medical Liability When Artificial Intelligence Is Applied on Diagnostic Algorithms: A Systematic Review. *Front. Med.* **2023**, *10*, 1305756. [CrossRef]
17. Ilikhan, S.U.; Özer, M.; Tanberkan, H.; Bozkurt, V. How to Mitigate the Risks of Deployment of Artificial Intelligence in Medicine? *Turk. J. Med. Sci.* **2024**, *54*, 483–492. [CrossRef]
18. Kerstan, S.; Bienefeld, N.; Grote, G. Choosing Human over AI Doctors? How Comparative Trust Associations and Knowledge Relate to Risk and Benefit Perceptions of AI in Healthcare. *Risk Anal.* **2024**, *44*, 939–957. [CrossRef] [PubMed]
19. Starke, G.; van den Brule, R.; Elger, B.S.; Haselager, P. Intentional Machines: A Defence of Trust in Medical Artificial Intelligence. *Bioethics* **2022**, *36*, 154–161. [CrossRef] [PubMed]
20. Witkowski, K.; Okhai, R.; Neely, S.R. Public Perceptions of Artificial Intelligence in Healthcare: Ethical Concerns and Opportunities for Patient-Centered Care. *BMC Med. Ethics* **2024**, *25*, 74. [CrossRef]
21. Hallowell, N.; Badger, S.; Sauerbrei, A.; Nellåker, C.; Kerasidou, A. "I Don't Think People Are Ready to Trust These Algorithms at Face Value": Trust and the Use of Machine Learning Algorithms in the Diagnosis of Rare Disease. *BMC Med. Ethics* **2022**, *23*, 112. [CrossRef]
22. Dlugatch, R.; Georgieva, A.; Kerasidou, A. Trustworthy Artificial Intelligence and Ethical Design: Public Perceptions of Trustworthiness of an AI-Based Decision-Support Tool in the Context of Intrapartum Care. *BMC Med. Ethics* **2023**, *24*, 42. [CrossRef]
23. Kiener, M. Artificial Intelligence in Medicine and the Disclosure of Risks. *AI Soc.* **2021**, *36*, 705–713. [CrossRef]
24. Nolan, P. Artificial Intelligence in Medicine—Is Too Much Transparency a Good Thing? *Med. Leg. J.* **2023**, *91*, 193–197. [CrossRef] [PubMed]
25. Lennartz, S.; Dratsch, T.; Zopfs, D.; Persigehl, T.; Maintz, D.; Hokamp, N.G.; dos Santos, D.P. Use and Control of Artificial Intelligence in Patients Across the Medical Workflow: Single-Center Questionnaire Study of Patient Perspectives. *J. Med. Internet Res.* **2021**, *23*, e24221. [CrossRef] [PubMed]
26. Marques, M.; Almeida, A.; Pereira, H. The Medicine Revolution Through Artificial Intelligence: Ethical Challenges of Machine Learning Algorithms in Decision-Making. *Cureus* **2024**, *16*, e69405. [CrossRef] [PubMed]
27. Saenz, A.D.; Harned, Z.; Banerjee, O.; Abràmoff, M.D.; Rajpurkar, P. Autonomous AI Systems in the Face of Liability, Regulations and Costs. *NPJ Digit. Med.* **2023**, *6*, 185. [CrossRef]
28. Thurzo, A.; Urbanová, W.; Novák, B.; Czako, L.; Siebert, T.; Stano, P.; Mareková, S.; Fountoulaki, G.; Kosnáčová, H.; Varga, I. Where Is the Artificial Intelligence Applied in Dentistry? Systematic Review and Literature Analysis. *Healthcare* **2022**, *10*, 1269. [CrossRef]
29. Aldhafeeri, F.M. Navigating the Ethical Landscape of Artificial Intelligence in Radiography: A Cross-Sectional Study of Radiographers' Perspectives. *BMC Med. Ethics* **2024**, *25*, 52. [CrossRef]
30. Tomášik, J.; Zsoldos, M.; Majdáková, K.; Fleischmann, A.; Oravcová, Ľ.; Sónak Ballová, D.; Thurzo, A. The Potential of AI-Powered Face Enhancement Technologies in Face-Driven Orthodontic Treatment Planning. *Appl. Sci.* **2024**, *14*, 7837. [CrossRef]
31. Tomášik, J.; Zsoldos, M.; Oravcová, Ľ.; Lifková, M.; Pavleová, G.; Strunga, M.; Thurzo, A. AI and Face-Driven Orthodontics: A Scoping Review of Digital Advances in Diagnosis and Treatment Planning. *AI* **2024**, *5*, 158–176. [CrossRef]
32. Kováč, P.; Jackuliak, P.; Bražinová, A.; Varga, I.; Aláč, M.; Smatana, M.; Lovich, D.; Thurzo, A. Artificial Intelligence-Driven Facial Image Analysis for the Early Detection of Rare Diseases: Legal, Ethical, Forensic, and Cybersecurity Considerations. *AI* **2024**, *5*, 990–1010. [CrossRef]
33. Rahimzadeh, V.; Baek, J.; Lawson, J.; Dove, E.S. A Qualitative Interview Study to Determine Barriers and Facilitators of Implementing Automated Decision Support Tools for Genomic Data Access. *BMC Med. Ethics* **2024**, *25*, 51. [CrossRef] [PubMed]
34. Thurzo, A.; Kurilová, V.; Varga, I. Artificial Intelligence in Orthodontic Smart Application for Treatment Coaching and Its Impact on Clinical Performance of Patients Monitored with AI-TeleHealth System. *Healthcare* **2021**, *9*, 1695. [CrossRef]

35. Thurzo, A.; Varga, I. Advances in 4D Shape-Memory Resins for AI-Aided Personalized Scaffold Bioengineering. *Bratisl. Med. J.* **2025**, *126*, 140–145. [CrossRef]
36. Kim, J.J.; Jung, M.W. Fear Paradigms: The Times They Are a-Changin'. *Curr. Opin. Behav. Sci.* **2018**, *24*, 38–43. [CrossRef]
37. Gross, C.T.; Canteras, N.S. The Many Paths to Fear. *Nat. Rev. Neurosci.* **2012**, *13*, 651–658. [CrossRef] [PubMed]
38. Milton, A.L. Fear Not: Recent Advances in Understanding the Neural Basis of Fear Memories and Implications for Treatment Development. *F1000Res* **2019**, *8*, 1948. [CrossRef] [PubMed]
39. Steimer, T. The Biology of Fear- and Anxiety-Related Behaviors. *Dialogues Clin. Neurosci.* **2002**, *4*, 231–249. [CrossRef]
40. Cyrulnik, B. Ethology of Anxiety in Phylogeny and Ontogeny. *Acta Psychiatr. Scand. Suppl.* **1998**, *393*, 44–49. [CrossRef]
41. Delini-Stula, A. Psychobiology of Fear. *Schweiz. Rundsch. Med. Prax.* **1992**, *81*, 259–265.
42. Ren, C.; Tao, Q. Neural Circuits Underlying Innate Fear. *Adv. Exp. Med. Biol.* **2020**, *1284*, 1–7. [CrossRef]
43. Bălan, O.; Moise, G.; Moldoveanu, A.; Leordeanu, M.; Moldoveanu, F. Fear Level Classification Based on Emotional Dimensions and Machine Learning Techniques. *Sensors* **2019**, *19*, 1738. [CrossRef] [PubMed]
44. Yamamori, Y.; Robinson, O.J. Computational Perspectives on Human Fear and Anxiety. *Neurosci. Biobehav. Rev.* **2023**, *144*, 104959. [CrossRef] [PubMed]
45. Abend, R.; Burk, D.; Ruiz, S.G.; Gold, A.L.; Napoli, J.L.; Britton, J.C.; Michalska, K.J.; Shechner, T.; Winkler, A.M.; Leibenluft, E.; et al. Computational Modeling of Threat Learning Reveals Links with Anxiety and Neuroanatomy in Humans. *Elife* **2022**, *11*, e66169. [CrossRef]
46. Asimov, I. Runaround. In *I, Robot*; Gnome Press: New York, NY, USA, 1950; pp. 40–55.
47. Organisation for Economic Co-operation; Development OECD AI Principles | OECD. Available online: https://www.oecd.org/en/topics/sub-issues/ai-principles.html (accessed on 12 May 2025).
48. Steimers, A.; Schneider, M. Sources of Risk of AI Systems. *Int. J. Environ. Res. Public. Health* **2022**, *19*, 3641. [CrossRef] [PubMed]
49. den Dulk, P.; Heerebout, B.T.; Phaf, R.H. A Computational Study into the Evolution of Dual-Route Dynamics for Affective Processing. *J. Cogn. Neurosci.* **2003**, *15*, 194–208. [CrossRef]
50. Graterol, W.; Diaz-Amado, J.; Cardinale, Y.; Dongo, I.; Lopes-Silva, E.; Santos-Libarino, C. Emotion Detection for Social Robots Based on NLP Transformers and an Emotion Ontology. *Sensors* **2021**, *21*, 1322. [CrossRef]
51. Rizzi, C.; Johnson, C.G.; Fabris, F.; Vargas, P.A. A Situation-Aware Fear Learning (SAFEL) Model for Robots. *Neurocomputing* **2017**, *221*, 32–47. [CrossRef]
52. Schwitzgebel, E. AI Systems Must Not Confuse Users about Their Sentience or Moral Status. *Patterns* **2023**, *4*, 100818. [CrossRef]
53. McKee, K.R.; Bai, X.; Fiske, S.T. Humans Perceive Warmth and Competence in Artificial Intelligence. *iScience* **2023**, *26*, 107256. [CrossRef]
54. Shank, D.B.; Gott, A. People's Self-Reported Encounters of Perceiving Mind in Artificial Intelligence. *Data Brief.* **2019**, *25*, 104220. [CrossRef]
55. Alkhalifah, J.M.; Bedaiwi, A.M.; Shaikh, N.; Seddiq, W.; Meo, S.A. Existential Anxiety about Artificial Intelligence (AI)- Is It the End of Humanity Era or a New Chapter in the Human Revolution: Questionnaire-Based Observational Study. *Front. Psychiatry* **2024**, *15*, 1368122. [CrossRef]
56. Blanco-Ruiz, M.; Sainz-De-baranda, C.; Gutiérrez-Martín, L.; Romero-Perales, E.; López-Ongil, C. Emotion Elicitation Under Audiovisual Stimuli Reception: Should Artificial Intelligence Consider the Gender Perspective? *Int. J. Environ. Res. Public. Health* **2020**, *17*, 8534. [CrossRef]
57. Dopelt, K.; Bachner, Y.G.; Urkin, J.; Yahav, Z.; Davidovitch, N.; Barach, P. Perceptions of Practicing Physicians and Members of the Public on the Attributes of a "Good Doctor". *Healthcare* **2021**, *10*, 73. [CrossRef] [PubMed]
58. Sauerbrei, A.; Kerasidou, A.; Lucivero, F.; Hallowell, N. The Impact of Artificial Intelligence on the Person-Centred, Doctor-Patient Relationship: Some Problems and Solutions. *BMC Med. Inform. Decis. Mak.* **2023**, *23*, 73. [CrossRef] [PubMed]
59. Schnelle, C.; Jones, M.A. Characteristics of Exceptionally Good Doctors-A Survey of Public Adults. *Heliyon* **2023**, *9*, e13115. [CrossRef] [PubMed]
60. Kiefer, C.S.; Colletti, J.E.; Bellolio, M.F.; Hess, E.P.; Woolridge, D.P.; Thomas, K.B.; Sadosty, A.T. The "Good" Dean's Letter. *Acad. Med.* **2010**, *85*, 1705–1708. [CrossRef]
61. Kotzee, B.; Ignatowicz, A.; Thomas, H. Virtue in Medical Practice: An Exploratory Study. *HEC Forum* **2017**, *29*, 1–19. [CrossRef]
62. Fones, C.S.; Kua, E.H.; Goh, L.G. 'What Makes a Good Doctor?'--Views of the Medical Profession and the Public in Setting Priorities for Medical Education. *Singap. Med. J.* **1998**, *39*, 537–542.
63. Yoo, H.H.; Lee, J.K.; Kim, A. Perceptual Comparison of the "Good Doctor" Image between Faculty and Students in Medical School. *Korean J. Med. Educ.* **2015**, *27*, 291–300. [CrossRef]
64. Maudsley, G.; Williams, E.M.I.; Taylor, D.C.M. Junior Medical Students' Notions of a "good Doctor" and Related Expectations: A Mixed Methods Study. *Med. Educ.* **2007**, *41*, 476–486. [CrossRef]
65. Meier, L.J.; Hein, A.; Diepold, K.; Buyx, A. Algorithms for Ethical Decision-Making in the Clinic: A Proof of Concept. *Am. J. Bioeth.* **2022**, *22*, 4–20. [CrossRef]

66. Thurzo, A.; Varga, I. Revisiting the Role of Review Articles in the Age of AI-Agents: Integrating AI-Reasoning and AI-Synthesis Reshaping the Future of Scientific Publishing. *Bratisl. Med. J.* **2025**, *126*, 381–393. [CrossRef]
67. Kakkos, I.; Dimitrakopoulos, G.N.; Gao, L.; Zhang, Y.; Qi, P.; Matsopoulos, G.K.; Thakor, N.; Bezerianos, A.; Sun, Y. Mental Workload Drives Different Reorganizations of Functional Cortical Connectivity Between 2D and 3D Simulated Flight Experiments. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2019**, *27*, 1704–1713. [CrossRef] [PubMed]
68. Abooelzahab, D.; Zaher, N.; Soliman, A.H.; Chibelushi, C. A Combined Windowing and Deep Learning Model for the Classification of Brain Disorders Based on Electroencephalogram Signals. *AI* **2025**, *6*, 42. [CrossRef]
69. Kakkos, I.; Dimitrakopoulos, G.N.; Sun, Y.; Yuan, J.; Matsopoulos, G.K.; Bezerianos, A.; Sun, Y. EEG Fingerprints of Task-Independent Mental Workload Discrimination. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 3824–3833. [CrossRef]
70. Thurzo, A. Provable AI Ethics and Explainability in Medical and Educational AI Agents: Trustworthy Ethical Firewall. *Electronics* **2025**, *14*, 1294. [CrossRef]
71. Zhang, X.; Chan, F.T.S.; Yan, C.; Bose, I. Towards Risk-Aware Artificial Intelligence and Machine Learning Systems: An Overview. *Decis. Support. Syst.* **2022**, *159*, 113800. [CrossRef]
72. Aref, Z.; Wei, S.; Mandayam, N.B. Human-AI Collaboration in Cloud Security: Cognitive Hierarchy-Driven Deep Reinforcement Learning. *arXiv* **2025**, arXiv:2502.16054.
73. Zhao, Y.; Efraim Aguilar Escamill, J.; Lu, W.; Wang, H. RA-PbRL: Provably Efficient Risk-Aware Preference-Based Reinforcement Learning. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 60835–60871.
74. Hejna, J.; Sadigh, D. Few-Shot Preference Learning for Human-in-the-Loop RL. In Proceedings of the 6th Conference on Robot Learning, Atlanta, GA, USA, 6–9 November 2023.
75. Festor, P.; Jia, Y.; Gordon, A.C.; Aldo Faisal, A.; Habli, I.; Komorowski, M. Assuring the Safety of AI-Based Clinical Decision Support Systems: A Case Study of the AI Clinician for Sepsis Treatment. *BMJ Health Care Inform.* **2022**, *29*, e100549. [CrossRef]
76. Rowe, N.C. The Comparative Ethics of Artificial-Intelligence Methods for Military Applications. *Front. Big Data* **2022**, *5*, 991759. [CrossRef] [PubMed]
77. Arbelaez Ossa, L.; Lorenzini, G.; Milford, S.R.; Shaw, D.; Elger, B.S.; Rost, M. Integrating Ethics in AI Development: A Qualitative Study. *BMC Med. Ethics* **2024**, *25*, 10. [CrossRef] [PubMed]
78. Hippocrates. "The Oath". In *Hippocrates*; Jones, W.H.S., Ed.; Harvard University Press: Cambridge, MA, USA, 1923; Volume 1, pp. 298–299, Loeb Classical Library No. 147. [CrossRef]
79. Youldash, M.; Rahman, A.; Alsayed, M.; Sebiany, A.; Alzayat, J.; Aljishi, N.; Alshammari, G.; Alqahtani, M. Early Detection and Classification of Diabetic Retinopathy: A Deep Learning Approach. *AI* **2024**, *5*, 2586–2617. [CrossRef]
80. Bienefeld, N.; Keller, E.; Grote, G. Human-AI Teaming in Critical Care: A Comparative Analysis of Data Scientists' and Clinicians' Perspectives on AI Augmentation and Automation. *J. Med. Internet Res.* **2024**, *26*, e50130. [CrossRef]
81. Shuaib, A. Transforming Healthcare with AI: Promises, Pitfalls, and Pathways Forward. *Int. J. Gen. Med.* **2024**, *17*, 1765–1771. [CrossRef] [PubMed]
82. Funer, F.; Tinnemeyer, S.; Liedtke, W.; Salloch, S. Clinicians' Roles and Necessary Levels of Understanding in the Use of Artificial Intelligence: A Qualitative Interview Study with German Medical Students. *BMC Med. Ethics* **2024**, *25*, 107. [CrossRef]
83. Kakkos, I.; Vagenas, T.P.; Zygogianni, A.; Matsopoulos, G.K. Towards Automation in Radiotherapy Planning: A Deep Learning Approach for the Delineation of Parotid Glands in Head and Neck Cancer. *Bioengineering* **2024**, *11*, 214. [CrossRef]
84. Camlet, A.; Kusiak, A.; Świetlik, D. Application of Conversational AI Models in Decision Making for Clinical Periodontology: Analysis and Predictive Modeling. *AI* **2025**, *6*, 3. [CrossRef]
85. Mohanasundari, S.K.; Kalpana, M.; Madhusudhan, U.; Vasanthkumar, K.; Rani, B.; Singh, R.; Vashishtha, N.; Bhatia, V. Can Artificial Intelligence Replace the Unique Nursing Role? *Cureus* **2023**, *15*, e51150. [CrossRef]
86. Kandpal, A.; Gupta, R.K.; Singh, A. Ischemic Stroke Lesion Segmentation on Multiparametric CT Perfusion Maps Using Deep Neural Network. *AI* **2025**, *6*, 15. [CrossRef]
87. Riva, G.; Mantovani, F.; Wiederhold, B.K.; Marchetti, A.; Gaggioli, A. *Psychomatics-A Multidisciplinary Framework for Understanding Artificial Minds*; Mary Ann Liebert Inc.: Larchmont, NY, USA, 2024.
88. Willem, T.; Fritzsche, M.C.; Zimmermann, B.M.; Sierawska, A.; Breuer, S.; Braun, M.; Ruess, A.K.; Bak, M.; Schönweitz, F.B.; Meier, L.J.; et al. Embedded Ethics in Practice: A Toolbox for Integrating the Analysis of Ethical and Social Issues into Healthcare AI Research. *Sci. Eng. Ethics* **2025**, *31*, 3. [CrossRef] [PubMed]
89. Cui, H.; Yasseri, T. AI-Enhanced Collective Intelligence. *Patterns* **2024**, *5*, 101074. [CrossRef]
90. Strachan, J.W.A.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; Manzi, G.; et al. Testing Theory of Mind in Large Language Models and Humans. *Nat. Hum. Behav.* **2024**, *8*, 1285–1295. [CrossRef] [PubMed]
91. Jang, G.; Kragel, P.A. Understanding Human Amygdala Function with Artificial Neural Networks. *J. Neurosci.* **2024**, *45*, e1436242025. [CrossRef]
92. He, X.; Wu, J.; Huang, Z.; Hu, Z.; Wang, J.; Sangiovanni-Vincentelli, A.; Lv, C. Fear-Neuro-Inspired Reinforcement Learning for Safe Autonomous Driving. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 267–279. [CrossRef] [PubMed]

93. Lee, H.-S.; Hong, S.-H.; Kim, G.-H.; You, H.-J.; Lee, E.-Y.; Jeong, J.-H.; Ahn, J.-W.; Kim, J.-H. Generative Models Utilizing Padding Can Efficiently Integrate and Generate Multi-Omics Data. *AI* **2024**, *5*, 1614–1632. [CrossRef]
94. Jeyaraman, N.; Jeyaraman, M.; Yadav, S.; Ramasubramanian, S.; Balaji, S. Revolutionizing Healthcare: The Emerging Role of Quantum Computing in Enhancing Medical Technology and Treatment. *Cureus* **2024**, *16*, e67486. [CrossRef]
95. Stiefel, K.M.; Coggan, J.S. The Energy Challenges of Artificial Superintelligence. *Front. Artif. Intell.* **2023**, *6*, 1240653. [CrossRef]
96. Verma, V.R.; Nishad, D.K.; Sharma, V.; Singh, V.K.; Verma, A.; Shah, D.R. Quantum Machine Learning for Lyapunov-Stabilized Computation Offloading in next-Generation MEC Networks. *Sci. Rep.* **2025**, *15*, 405. [CrossRef]
97. Mutawa, A.M. Attention-Based Hybrid Deep Learning Models for Classifying COVID-19 Genome Sequences. *AI* **2025**, *6*, 4. [CrossRef]
98. Müller, V.C. Autonomous Killer Robots Are Probably Good News. In *Drones and Responsibility: Legal, Philosophical and Socio-Technical Perspectives on Remotely Controlled Weapons*; Routledge: London, UK; New York, NY, USA, 2016; pp. 67–81. [CrossRef]
99. Petersson, L.; Vincent, K.; Svedberg, P.; Nygren, J.M.; Larsson, I. Ethical Perspectives on Implementing AI to Predict Mortality Risk in Emergency Department Patients: A Qualitative Study. *Stud. Health Technol. Inform.* **2023**, *302*, 676–677. [CrossRef] [PubMed]
100. Abràmoff, M.D.; Tobey, D.; Char, D.S. Lessons Learned About Autonomous AI: Finding a Safe, Efficacious, and Ethical Path Through the Development Process. *Am. J. Ophthalmol.* **2020**, *214*, 134–142. [CrossRef] [PubMed]
101. Leith, T.; Correll, J.A.; Davidson, E.E.; Gottula, A.L.; Majhail, N.K.; Mathias, E.J.; Pribble, J.; Roberts, N.B.; Scott, I.G.; Cranford, J.A.; et al. Bystander Interaction with a Novel Multipurpose Medical Drone: A Simulation Trial. *Resusc. Plus* **2024**, *18*, 100633. [CrossRef]
102. Sharma, R.; Salman, S.; Gu, Q.; Freeman, W.D. Advancing Neurocritical Care with Artificial Intelligence and Machine Learning: The Promise, Practicalities, and Pitfalls Ahead. *Neurol. Clin.* **2025**, *43*, 153–165. [CrossRef] [PubMed]
103. Billing, D.C.; Fordy, G.R.; Friedl, K.E.; Hasselstrøm, H. The Implications of Emerging Technology on Military Human Performance Research Priorities. *J. Sci. Med. Sport.* **2021**, *24*, 947–953. [CrossRef]
104. de Sio, F.S.; van den Hoven, J. Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Front. Robot. AI* **2018**, *5*, 323836. [CrossRef]
105. Verdiesen, I.; Dignum, V. Value Elicitation on a Scenario of Autonomous Weapon System Deployment: A Qualitative Study Based on the Value Deliberation Process. *AI Ethics* **2022**, *3*, 887–900. [CrossRef]
106. Moor, M.; Banerjee, O.; Abad, Z.S.H.; Krumholz, H.M.; Leskovec, J.; Topol, E.J.; Rajpurkar, P. Foundation Models for Generalist Medical Artificial Intelligence. *Nature* **2023**, *616*, 259–265. [CrossRef]
107. Huang, K.A.; Choudhary, H.K.; Kuo, P.C. Artificial Intelligent Agent Architecture and Clinical Decision-Making in the Healthcare Sector. *Cureus* **2024**, *16*, e64115. [CrossRef]
108. Ploug, T.; Sundby, A.; Moeslund, T.B.; Holm, S. Population Preferences for Performance and Explainability of Artificial Intelligence in Health Care: Choice-Based Conjoint Survey. *J. Med. Internet Res.* **2021**, *23*, e26611. [CrossRef]
109. Di Nucci, E.; de Sio, F.S. *Drones and Responsibility: Legal, Philosophical and Socio-Technical Perspectives on Remotely Controlled Weapons*; Routledge: London, UK; New York, NY, USA, 2016; pp. 1–217. [CrossRef]
110. Altehenger, H.; Menges, L.; Schulte, P. How AI Systems Can Be Blameworthy. *Philosophia* **2024**, *52*, 1083–1106. [CrossRef] [PubMed]
111. Kneer, M.; Christen, M. Responsibility Gaps and Retributive Dispositions: Evidence from the US, Japan and Germany. *Sci. Eng. Ethics* **2024**, *30*, 51. [CrossRef]
112. Ferrario, A. Design Publicity of Black Box Algorithms: A Support to the Epistemic and Ethical Justifications of Medical AI Systems. *J. Med. Ethics* **2022**, *48*, 492–494. [CrossRef] [PubMed]
113. Barea Mendoza, J.A.; Valiente Fernandez, M.; Pardo Fernandez, A.; Gómez Álvarez, J. Current Perspectives on the Use of Artificial Intelligence in Critical Patient Safety. *Med. Intensiv.* **2024**, *49*, 154–164. [CrossRef]
114. Schlicker, N.; Langer, M.; Hirsch, M.C. How Trustworthy Is Artificial Intelligence? A Model for the Conflict between Objectivity and Subjectivity. *Inn. Med.* **2023**, *64*, 1051–1057. [CrossRef]
115. Kim, J.; Kim, S.Y.; Kim, E.A.; Sim, J.A.; Lee, Y.; Kim, H. Developing a Framework for Self-Regulatory Governance in Healthcare AI Research: Insights from South Korea. *Asian Bioeth. Rev.* **2024**, *16*, 391–406. [CrossRef] [PubMed]
116. Surovková, J.; Haluzová, S.; Strunga, M.; Urban, R.; Lifková, M.; Thurzo, A. The New Role of the Dental Assistant and Nurse in the Age of Advanced Artificial Intelligence in Telehealth Orthodontic Care with Dental Monitoring: Preliminary Report. *Appl. Sci.* **2023**, *13*, 5212. [CrossRef]
117. Thurzo, A. How Is AI Transforming Medical Research, Education and Practice? *Bratisl. Med. J.* **2025**, *126*, 243–248. [CrossRef]
118. Brodeur, P.G.; Buckley, T.A.; Kanjee, Z.; Goh, E.; Ling, E.B.; Jain, P.; Cabral, S.; Abdulnour, R.-E.; Haimovich, A.; Freed, J.A.; et al. Superhuman Performance of a Large Language Model on the Reasoning Tasks of a Physician. *arXiv* **2024**, arXiv:2412.10849.
119. Gupta, G.K.; Singh, A.; Manikandan, S.V.; Ehtesham, A. Digital Diagnostics: The Potential of Large Language Models in Recognizing Symptoms of Common Illnesses. *AI* **2025**, *6*, 13. [CrossRef]

120. Thurzo, A.; Strunga, M.; Urban, R.; Surovková, J.; Afrashtehfar, K.I. Impact of Artificial Intelligence on Dental Education: A Review and Guide for Curriculum Update. *Educ. Sci.* **2023**, *13*, 150. [CrossRef]
121. Zhang, J.; Fenton, S.H. Preparing Healthcare Education for an AI-Augmented Future. *npj Health Syst.* **2024**, *1*, 4. [CrossRef] [PubMed]
122. Gottweis, J.; Weng, W.-H.; Daryin, A.; Tu, T.; Palepu, A.; Sirkovic, P.; Myaskovsky, A.; Weissenberger, F.; Rong, K.; Tanno, R.; et al. Towards an AI Co-Scientist. *arXiv* **2025**, arXiv:2502.18864.
123. Uc, A.A.; Ucsf, B.; Phillips, R.V.; Kıcıman, E.; Balzer, L.B.; Van Der Laan, M.; Petersen, M. Large Language Models as Co-Pilots for Causal Inference in Medical Studies. *arXiv* **2024**, arXiv:2407.19118.
124. Musslick, S.; Bartlett, L.K.; Chandramouli, S.H.; Dubova, M.; Gobet, F.; Griffiths, T.L.; Hullman, J.; King, R.D.; Kutz, J.N.; Lucas, C.G.; et al. Automating the Practice of Science: Opportunities, Challenges, and Implications. *Proc. Natl. Acad. Sci. USA* **2025**, *122*, e2401238121. [CrossRef] [PubMed]
125. Xie, S.; Zhao, W.; Deng, G.; He, G.; He, N.; Lu, Z.; Hu, W.; Zhao, M.; Du, J. Utilizing ChatGPT as a Scientific Reasoning Engine to Differentiate Conflicting Evidence and Summarize Challenges in Controversial Clinical Questions. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 1551–1560. [CrossRef]
126. Oniani, D.; Hilsman, J.; Peng, Y.; Poropatich, R.K.; Pamplin, J.C.; Legault, G.L.; Wang, Y. Adopting and Expanding Ethical Principles for Generative Artificial Intelligence from Military to Healthcare. *NPJ Digit. Med.* **2023**, *6*, 225. [CrossRef]
127. Weidener, L.; Fischer, M. Proposing a Principle-Based Approach for Teaching AI Ethics in Medical Education. *JMIR Med. Educ.* **2024**, *10*, e55368. [CrossRef]
128. Zhou, R.; Zhang, G.; Huang, H.; Wei, Z.; Zhou, H.; Jin, J.; Chang, F.; Chen, J. How Would Autonomous Vehicles Behave in Real-World Crash Scenarios? *Accid. Anal. Prev.* **2024**, *202*, 107572. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.