

Not Knowing, Yet Living: AI and the Modern Legal Trial of Prometheus in Medicine

Uchenna Nnawuchi¹ and Carlisle George²

¹ Researcher, ALERT Research Group, Department of Computer Science,
Faculty of Science and Technology, Middlesex University, London UK
UN055@live.mdx.ac.uk; uchennannawuchi@gmail.com

² Associate Professor, ALERT Research Group, Department of Computer Science,
Faculty of Science and Technology, Middlesex University, London UK
c.george@mdx.ac.uk

Abstract. This paper explores the paradox of Artificial Intelligence (AI), particularly Machine Learning (ML), in modern medicine, promising enhanced diagnostics and treatment while in many instances also leading to a profound loss of transparency and interpretability. Complex AI systems such as Automated Decision-Making Systems (ADMS) increasingly guide critical clinical decisions; however, their opaque “black box” nature challenges traditional medical ethics centred on transparency, due process, accountability, and patient autonomy. ADMs often make recommendations that cannot be fully understood or explained due to the system's opacity. This raises both legal and ethical concerns. For example, the ability of clinicians to justify clinical decisions informed by AI when they themselves lack insight into its reasoning raises concerns about clinical legal liability. This, in turn, raises ethical concerns regarding patient autonomy if patients are unable to meaningfully make informed choices on treatments due to the opacity of decisions made by AI. The extent to which patients should consent to and place their trust in life-saving technologies whose inner workings remain opaque remains an open debate. These technologies, arguably, can diminish patients' rights to knowledge, autonomy, and dignity. Such concerns underscore the fragility of trust within clinical relationships and expose the limitations of prevailing frameworks for medical explanation (MedX) and informed consent. This paper examines the role of explainability in AI-driven medicine, arguing that while explanation is essential, it must be balanced against the clinical utility of high-performing, opaque systems. To address this tension, the paper proposes the Promethean Threshold Heuristic (PTH)—a normative framework designed to reconcile the need for explanation with clinical efficacy. Drawing on Promethean myth, it frames the ethical dilemma posed by AI as one of balancing risk, knowledge, and human dignity. Ultimately, it calls for principled AI governance in healthcare—one that preserves trust, prioritises transparency where feasible, and recognises the trade-offs inherent in deploying powerful but opaque technologies

Keywords: Artificial Intelligence, Machine Learning, Medicine, Healthcare, transparency, explanation, opacity, black box.

1 Introduction

“Prometheus gave man fire and was punished for it. Maybe that is the price of progress.”
by *Mary Shelley*[1]

Prometheus’s gift of fire to humanity, a symbol of knowledge and power, came with unintended consequences — a dual legacy of advancement and suffering. Similarly, the advent of Artificial Intelligence (AI), specifically Machine Learning (ML) in medicine, represents a profound leap forward, promising to revolutionise healthcare through improved diagnostics, treatment precision, and patient outcomes. However, this progress is accompanied by a form of punishment: a loss of transparency and autonomy over the knowledge or reasoning that may inform life-saving decisions. The fire of AI brings about new possibilities but also casts shadows of uncertainty and ethical complexity.

At Duke University Hospital, Nurse Dina Sarro struggled to explain alerts from an AI system, Sepsis Watch, which flagged sepsis risks using vast data [2]. Physicians were sceptical, and Sarro admitted she could not justify the algorithm’s output. This reflects a broader shift in medicine, where AI-based recommendations often lack transparency [3], leaving clinicians caught between black-box systems and sceptical colleagues. ADMs are now widespread across domains like healthcare, finance, and justice [4, 5], aiding diagnosis through analysis of clinical data [6]. While they improve efficiency and access [7, 8], their opacity raises serious ethical concerns, compromising informed consent, clinical accountability, and potentially embedding bias [9, 10]. Cases like the failed 2018 sepsis detection highlight the stakes [11]. This paper explores the tension between explainability and performance in clinical AI, proposing the Promethean Threshold Heuristic (PTH) as a framework to balance trust, dignity, and patient survival.

2 Automated Decision-Making and Application in Medicine

The COVID-19 pandemic rapidly accelerated digital transformation in healthcare, specifically in medicine and clinical practice [12]. Traditional care models gave way to telemedicine, which became a cornerstone of health delivery across Europe and beyond. Tools like remote monitoring, AI-assisted diagnostics, and virtual consultations proliferated [13, 14, 15, 16]. This shift laid the groundwork for the widespread integration of ML into clinical practice [3], forming what is termed Medical Artificial Intelligence (MAI). MAI encompasses a broad spectrum of AI technologies used in medicine, including ML, natural language processing, and expert systems [3]. These tools support or replace clinician decision-making in diagnosis, treatment planning, and patient monitoring [17]. Related concepts—Machine Learning in Medicine (MLM) [18] and Medical Augmented Intelligence [19]—focus on data-driven care [18] and human-AI collaboration, respectively [19]. Though nuanced in focus, all share a central challenge: algorithmic opacity and bias in their recommendation [9]. This paper uses the term MAI to collectively refer to these related paradigms, focusing on their shared reliance on opaque, high-performing systems and their impact on clinical reasoning and patient

care. As AI systems become more embedded in healthcare, they raise urgent ethical, legal, and practical concerns about explainability, autonomy, trust, and accountability. Despite their promise, these systems are often obscured not only to patients but also to clinicians and even to the developers of these systems [9]. Consequently, clinicians may struggle to explain AI-informed decisions, complicating informed consent, patient autonomy and dignity, and human rights. This issue is particularly pronounced with a subset of MAI, *id est* ADMs, which are increasingly responsible for critical tasks such as disease prediction, diagnosis, and treatment [18]. The application of ADMs has raised critical ethical and legal questions about explainability, accountability [20], and the thresholds of trust in AI-driven medicine [3]. Consequently, regulation is playing catch-up, and this has led to several recent enactments, including the Artificial Intelligence Act (AIA) 2021 [21] and the European Health Data Space Regulation (EHDS) [22]. However, existing regulations already in place include the General Data Protection Regulation (GDPR) 2016 [23], the Medical Device Regulation (MDR) 2017 [24], and the Vitro Diagnostic Regulation (IVDR) [25].

2.1 Application of Automated Decision-Making (ADMs)

The zeitgeist suggests that ADMs are increasingly seen as a solution to many challenges in medicine, specifically clinical practice, promising greater speed, accuracy, and efficiency [26]. ADMs used prediction, assessing patient deterioration risks by integrating behavioural and physiological data—often from wearables or external sources—to detect early warning signs [5], [27]. However, the reasoning behind these predictions is frequently opaque, raising issues of privacy, dignity, and autonomy. Furthermore, in diagnosis, particularly in fields like radiology, ADMs such as IDx-DR detect complex patterns that may elude human perception [8]. These systems rely on processing data at scales beyond human capability. Yet, the lack of transparency often leads clinicians to accept decisions they cannot fully interpret, weakening shared decision-making and introducing automation bias [28].

Additionally, ADMs draw on large datasets—including patient histories, biomarkers, and past outcomes—to suggest interventions [29]. In fields like oncology [30] and sepsis care [31], they increasingly shape clinical decisions, guiding therapeutic strategies and influencing whether aggressive or conservative approaches are pursued. Beyond the sepsis case, numerous examples illustrate the challenges clinicians face in fully trusting algorithmic outputs within clinical practice, largely due to understandable scepticism arising from the potential for false positives and false negatives. For instance, a deep learning model assessing mortality risk from pneumonia categorised patients with asthma as low risk [32]. This counterintuitive outcome stemmed from the model's reliance on a training dataset in which asthmatic patients had benefited from proactive clinical interventions, thereby skewing the risk profile. Similarly, another deep learning system designed to screen chest X-rays for pneumonia inadvertently exploited confounding variables such as the geographic location of the imaging scanner, rather than the clinical features of the X-rays themselves [32]. In a further case, a model aimed at stratifying patients by risk level based on radiographic data relied on hardware-specific metadata, rather than pathophysiological indicators, to generate its predictions [32]. More so, IBM Watson used for oncology received serious widespread

criticism because its treatment recommendations were sometimes inconsistent or insufficiently explainable, leading to concerns about reliability and clinical trust [33]. Reports indicated that Watson occasionally provided unsafe or irrelevant advice due to limitations in its training data and opaque decision processes [33]. Collectively, these examples underscore the compounding systemic issues of opacity, trust, and accountability that continue to confound the integration of AI into clinical practice. They highlight the urgent need for greater transparency and explainability to support clinician confidence and ensure the safe, ethical deployment of such systems. When clinicians are unable to comprehend or interrogate the decision-making processes of AI models, this not only undermines patients' rights to informed consent and autonomous decision-making but also raises serious legal concerns. Particularly, clinical reliance on opaque algorithmic outputs may give rise to liability and causation disputes should patient harm occur, further eroding the trust that is foundational to both ethical medicine and effective AI governance.

These issues associated with these ADM tools require their role to be examined within ethical and legal frameworks that prioritise transparency, accountability, and the patient's right to understand and consent.

2.2 Challenges of Applying Automated Decision-Making

While ADMs have enhanced efficiency and precision in healthcare, their growing role as quasi-clinicians poses serious ethical and legal challenges. They shift traditional roles, raise accountability questions, and disrupt foundational principles such as shared decision-making (SDM), data privacy, informed consent, and transparency.

2.3.1 Shared Decision-Making (SDM)

SDM relies on a dialogic relationship between patients and clinicians, grounded in mutual respect and autonomy [34]. The integration of opaque ADMs threatens this balance [34]. When patients cannot understand or challenge algorithmic recommendations, consent becomes symbolic rather than substantive. ADM systems, especially black-box models, risk making patients passive recipients of decisions they neither shape nor fully grasp [35]. This undermines agency, reducing humans to data points and clinicians to intermediaries between patients and machines [36]. Further complicating matters, there is no clear consensus on how AI should function within SDM frameworks—should it guide, assist, or override decisions? Without transparency, AI risks substituting probabilistic logic for principled judgment, introducing automation bias and diminishing ethical integrity. If clinicians defer to systems they do not understand or cannot explain, the clinician-patient alliance is eroded, and accountability becomes murky.

2.3.2 Legal and Ethical Challenges

A. Data Protection and Privacy

ADMs depend on large datasets from EHRs, wearables, and apps, raising risks of re-identification, discrimination, and data misuse [5]. While models like federated learning enhance privacy [37], tension persists between the need for long-term data and GDPR principles like minimisation [23]. Particularly in imaging [38], vast identifiable

data are needed to train accurate systems, making robust anonymisation and public trust essential. Compliance with GDPR remains difficult, especially for predictive ADMs that require data persistence.

B. *Autonomy and Informed Consent*

Meaningful consent is increasingly difficult in AI-driven care ecosystems [36]. Patients often lack awareness of how their data is processed or how algorithms influence their care [39]. Opaque ADM reasoning challenges both patient and surrogate decision-making, making informed consent fragile at best [40]. When neither the patient nor clinician can explain or contest a system’s output, consent becomes procedural theatre [36]. This not only undermines the patient’s moral agency but also compromises the clinician’s ethical obligations under medical codes that prioritise autonomy and informed decision-making [41]. Clinicians risk being reduced to conduits of algorithmic authority, their fiduciary role diluted. To preserve the integrity of informed consent, investments in interpretability, communication tools, and clinician training are essential.

C. *Transparency*

Transparency (the degree to which the *inner workings* of an AI system, including its structure, data sources, design, and decision-making processes, are open and accessible for scrutiny) —closely tied to explainability and interpretability—is vital for trust in AI healthcare [42]. However, most ADMs function as black boxes, even to developers. Unlike some traditional medical uncertainties, algorithmic opacity is systemic and scalable, making it harder to justify or contest outcomes [42]. This creates a troubling scenario: patients denied care by opaque ADMs cannot understand or challenge the decision, stripping them of agency. The ethical and legal implications are profound—when harm results, who is accountable? Existing legal safeguards, such as the GDPR’s right to contest [23] and the AI Act’s right to explanation [21], are difficult to implement in practice but increasingly necessary. Without them, trust in AI-driven medicine will erode. Upholding the right to explanation reasserts the patient’s autonomy, dignity, and moral standing and ensures ADM systems support, rather than replace, ethical clinical judgment.

3 The Right to Explanation

3.1 Proposed solutions to ML opacity in Medical AI

Some scholars argue that the right to explanation offers limited value, warning it may become “meaningless” in practice [36]. They favour architectural regulation and tools like DPIAs over enforceable rights to build trust [36]. Ghassemi et al. [43] also question current explainability methods, calling them inadequate for patient-level decision-making. Others defend black-box models for their diagnostic accuracy, suggesting that proven outcomes matter more than transparency, especially since human reasoning is not always explainable either [44]. A more nuanced scholarly position proposes the ethical governance of opacity [45]. As opposed to viewing opacity as an inherent flaw, this perspective reframes it as a manageable condition — one that, when accompanied by institutional trust and role-sensitive justification, can coexist with the responsible deployment of AI. This approach explicitly moves away from epistemic idealism —

the belief that full transparency is always desirable or achievable — toward epistemic realism, which recognises that explanations must be contextually calibrated. Trust, in this view (epistemic realism), is not produced by making all internal workings visible, but by ensuring that decisions are justifiable to stakeholders in ways that are meaningful for their specific roles and risks [45]. This proposition, though arguable, makes these systems more responsible, trustworthy, and ethical. However, this view does not overlook ethical duties like informed consent under the WMA Code, but supports it. More so, clinical scholars such as Mehlus and her group argued that AI offers significant benefits in clinical practice, but its adoption hinges on calibrated trust [46]. Clinicians must rely on AI without deferring to it uncritically, especially when its outputs contradict clinical judgment. Explainable AI (XAI) addresses this by clarifying how decisions are made, supporting both trust and accountability. Even ethicists like Grote and Berens contended that the absence of reliable, meaningful explanations can erode the clinician–patient relationship and distort shared decision-making [47]. Thus, the Sepsis Watch example paragraph 1 and other examples in paragraph 2.1 exemplify this broader phenomenon: opacity without compensatory frameworks for explanation amounts to a structural ethical risk in AI-assisted medicine. Ultimately, without explainability, patient autonomy, trust, and accountability suffer, making it not just a technical concern, but a moral imperative.

3.2 Meaning and justification of the right to explanation in clinical Practice

The right to explanation in the context of medical or clinical practice refers to a patient's ability to understand why an MAI, specifically ADMs, made a particular recommendation [48]. In this context, *explanation* means providing understandable reasons for a decision [49]; *explainability* means the system itself can generate such reasons in a human-comprehensible way [50]. This right affirms patient autonomy and dignity. It has both intrinsic value (respecting individuals as rational agents) and instrumental value (helping them make informed choices, challenge decisions, or seek redress). Explanations must serve a purpose: enabling understanding and informed decision-making, especially when decisions affect health or legal status [51]. An explanation is very important because opaque decisions risk eroding trust, harming patients, and weakening professional accountability [52]. Explanations are necessary to detect errors, address bias, ensure fairness [20], and uphold the clinician's duty of care. Without explanation, patients become passive recipients of black-box outputs, undermining informed consent [52]. As stated above (*paragraph 3.1*), some scholars argue that performance alone justifies trust (epistemic realism), while we insist on transparency as a moral obligation (epistemic idealism). Ultimately, the right to explanation is not a technical luxury but a safeguard for safety, dignity, and ethical medical practice. However, the right to explanation and explainability is not without its challenges, especially when considered against the trade-off with model performance.

4 Trade-off between Performance and Explainability

As AI becomes more integrated, particularly in high-stakes fields like medicine, where decisions impact life, autonomy, and legal rights, the demand for XAI grows

exponentially [53]. Explainability is essential for accountability, trust, and informed consent. A patient cannot make an informed choice on different procedures, whether it is immunotherapy or chemotherapy, when diagnosed with cancer. This undermines dignity and could violate the right to life, Article 8 ECHR and 6 ICCPR; this right to life is not just a right to live [54, 55] but includes the right to choose how you want to live, the way you want to live, if you want to live [56], and a quality end to life [57]. This right to life, in medical or clinical practice, is protected, supported, and valued by the right to explanation. However, the most accurate AI systems (like deep neural networks) are often the least transparent [58]. This creates a trade-off: performance vs explainability. More explainable models can lack diagnostic power, while high-performing models can be "black boxes" which are incomprehensible to the clinicians, the patients and even to their developers [58]. However, explanation is not just about understanding—it supports ethical practice, respects patient choice, and fulfils legal obligations under frameworks like Articles 9, 12, 15, 22 GDPR, Article 86 EU AI Act, and Article 26 WMA. Clinical AI decisions, like recommending cancer treatments or life-saving surgeries, must be explainable to preserve patient autonomy. Without clear reasoning, patients are denied the chance to consent meaningfully or understand the risks involved. Some studies [59] also show that poor-quality or misleading explanations can confuse clinicians or lead to over-reliance, often referred to as automation bias [5]. But this highlights the need for better explainability, not abandoning it. Properly designed, explainable AI supports trust, reduces bias, and promotes accountability. This issue is even more dire in and compounded during a medical emergency. For instance, the Sepsis case of 2018 given above, where the nurse stated that she would not have a good reason for the algorithm's decision. In such situations, perhaps where the case is one of emergency, explanation will give way. The challenge with this is one of legal liability and causation, where the treatment goes wrong. One is uncertain where to attribute responsibility, whether it is the clinicians who adopted the algorithm's recommendation or the developers of the algorithm. The position of the law is still unclear in this area. However, where it is not an emergency, and the patient's life is not hanging by a thread, then performance should give way to explanation. The patient or their surrogates or representatives would require an explanation to make an informed decision. Ultimately, balancing accuracy with patient autonomy and safety is not optional—it's a moral and legal imperative.

5 The Promethean Threshold Heuristic: Balancing Explanation and Survival in AI-Driven Healthcare

The proposed Promethean Threshold Heuristic (PTH) is designed to help healthcare professionals, AI developers, and policymakers navigate the ethical dilemmas posed by Automated Decision-Making Systems (ADMs). It seeks to balance life-saving interventions with the patient's right to meaningful explanation, upholding autonomy, beneficence, and legal accountability. PTH is rooted in internationally recognised ethical and legal standards, including the UNESCO Declaration on Bioethics and Human Rights (2005) [60], WMA International Code of Medical Ethics (2022) [41], OECD AI Principles (2019) [61], Oviedo Convention (1997) [62], Declaration of Helsinki (2013)

[63], GDPR (2016) [23], and the EU AI Act (2021) [21]. These frameworks provide a strong ethical and legal foundation, ensuring that PTH reflects global norms and is adaptable across healthcare contexts. Additionally, it is structured around four core medical scenarios— life-threatening emergencies, emergencies with available surrogates, non-emergencies, and cases involving unknown patient status. PTH engages with the principles of autonomy, non-maleficence, beneficence, justice, and accountability [41]. It offers a novel and principled approach to reconciling medical necessity with the ethical demand for explanation in the age of AI.

5.1 The Promethean Threshold Heuristic (PTH)

Heuristic Statement: *In Automated Decision-making systems and AI-assisted medical decision-making, the importance of saving life may at times be justifiably overridden by the demand for patient-centred explanation. However, where explanation is to be deferred, it must be provided as soon as practicable to the patient or their designated surrogate family member/representative, balancing respect for human rights, dignity, and autonomy with clinical urgency.*

Table 1 below outlines the PTH components and how they are applied in the context of the four scenarios identified.

Table 1 — The PTH Components and Application

No	Scenario	Priority	Action	Post-emergency
1.	Life-Threatening Emergency	Save the patient's life with the most effective means available, even if, at the given time, the AI decision-making processes are not fully explainable.	If the patient is incapacitated and no surrogate or family member/representative is available to receive an explanation and make an informed decision, the decision to prioritise saving the patient's life over explanation is justified	Provide a detailed explanation to the patient once stabilised or to family/surrogates/representatives as soon as possible.
2.	Emergency with Available Surrogate or Family Member	Save the patient's life while initiating real-time explanation to the surrogate/family/representatives to respect patient autonomy.	The involvement of family members/representatives/surrogates should be leveraged to maintain transparency and trust, even during urgent care	Follow-up explanations to the patient and surrogates for full disclosure.
3.	Non-Emergency/Elective Care	Explanation and informed consent take precedence; AI decisions should be fully	Patients have the right to understand the extent of AI involvement and its recommendations in	Ongoing consent and shared decision-making should be

		explainable to the patient prior to treatment.	their care before treatment begins.	ensured throughout the care process.
4.	Unknown Patient Status or Absence of Surrogates	Follow the principle of beneficence — act to preserve the patient's life and wellbeing while documenting all AI decisions carefully.	Action: Efforts must be made to identify surrogates or legal guardians for subsequent explanation and consent.	Provide explanation to any identified representatives or to the patient upon regaining capacity.

5.2 Justification of the PTH

The PTH aligns with core ethical principles in healthcare: autonomy, beneficence, non-maleficence, justice, and accountability [41].

1. Respect for patient autonomy

The PTH defers explanation only in urgent care scenarios, reaffirming the obligation to inform patients as soon as possible. This mirrors international standards:

- *UNESCO Declaration* (Arts. 5–6) emphasises autonomy and informed consent, allowing exceptions in emergencies.
- *WMA Code of Ethics* (Arts. 13, 15, 17) supports patient autonomy and permits treatment without prior consent only when the patient is incapacitated and no surrogate is available. Thus, the heuristic safeguards autonomy while balancing clinical necessity.

2. Beneficence and non-maleficence

The PTH prioritises life-saving intervention while minimising harm, justifying temporary non-disclosure only when delay could endanger life. This principle is supported by:

- *WMA WMA Article 14* and *Declaration of Helsinki* (paragraphs 18, 30) stress prioritising patient welfare and minimising risk.
- *OECD AI Principles* advocate for human-centric, transparent AI systems that avoid harm. The heuristic reflects this by ensuring deferred explanation is ethically justified, time-limited, and subject to oversight.

3. Justice

The PTH affirms the legitimacy of surrogates or proxies in safeguarding patient rights when direct consent is not possible. This is consistent with:

- *WMA Code* (Arts. 16, 19) endorses involving trusted representatives for patients with impaired capacity and stresses respect for patient preferences while preserving confidentiality.
- *Oviedo Convention* (Arts. 6, 9) mandates proxy consent and equitable access to care without discrimination. By incorporating surrogates in decision-

making, the heuristic promotes fairness and continuity of care, ensuring patient rights are upheld in vulnerable scenarios.

4. Accountability

PTH requires timely documentation and disclosure to preserve trust and professional responsibility. This aligns with:

- *WMA Code* (Arts. 21, 23) emphasises transparency, including disclosure of third-party involvement and maintaining accurate records.
- *GDPR* (Arts. 5, 12, 15) supports transparency in automated decision-making and grants patients access to data and explanations.
- *IMIA Code of Ethics* highlights integrity, confidentiality, and clear communication in digital health. Together, these frameworks support the heuristic's demand for AI systems that combine rapid clinical response with post-event explainability and traceability.

Thus, the PTH encourages that AI systems should be designed to enable rapid and accurate decision-making for emergencies and transparent, explainable outputs for their decisions.

6 Conclusion

As Artificial Intelligence continues to transform the clinical landscape, the tension between algorithmic performance in decision-making and the ethical and legal imperative for explainability has become increasingly pronounced. This paper argued that while the right to explanation remains a fundamental concept for patient autonomy, informed consent, and trust, it must sometimes kowtow—temporarily and proportionally—in situations where a patient's survival and life are at stake. The Promethean Threshold Heuristic (PTH) responds to this challenge by offering a normative framework that is both context-sensitive and ethically grounded. Drawing on foundational principles peculiar to medical practice, such as autonomy, beneficence, and accountability, the PTH delineates clear conditions under which explanation may be deferred and when it must be promptly restored.

By acknowledging the moral complexity and ethical quagmire of AI in medicine, the PTH contributes meaningfully to the ongoing debate over the trade-off between explainability and performance. Opposed to framing the two in zero-sum opposition, the PTH proposes a structured reconciliation based on urgency, human dignity, and proportionality. This heuristic offers a flexible, rights-conscious model adaptable to real-world practice, capable of guiding clinicians, designers, and regulators.

Looking forward, further empirical work is needed to test and refine the heuristic across diverse healthcare settings, and policy frameworks must evolve to support systems capable of both high clinical performance and post hoc transparency. In this way, the PTH helps pave the way for AI systems that are not only powerful but also principled and patient-centred.

Disclosure of Interests: The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Shelley, M.: *Frankenstein; or, The Modern Prometheus* (Penguin Classics 2003) 84.
2. Upson, S.: 'AI Is Helping Patients—but Doctors Still Need to Understand It' *Wired* (9 March 2020) <https://www.wired.com/story/ai-help-patients-doctors-understand> accessed 2025/05/21.
3. Solaiman, B.: 'From "AI to Law" in Healthcare: The Proliferation of Global Guidelines in a Void of Legal Uncertainty' (2023) 42(2) *Medicine and Law* 391–406.
4. Dormehl, L.: *Thinking Machines: The Quest for Artificial Intelligence—and Where It's Taking Us Next* (New York: Penguin–TarcherPerigee, 2017).
5. Pierce, R.: 'AI in Healthcare: Solutions, Challenges, and Dilemmas in Medical Decision-Making' (12 August 2019) <https://ssrn.com/abstract=3806767> accessed 21 May 2025.
6. Susskind J.: *Future Politics: 'Living Together in a World Transformed by Tech'* (2020).
7. Esteva, A., Kuprel, B., Novoa, R. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118 (2017). <https://doi.org/10.1038/nature21056>.
8. Gulshan et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), 2402–2410 (December 2016).
9. Pazzanese C.: "Great promise but potential for peril" (The Harvard Gazette, October 26, 2020). <https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/>. Last accessed, 2023/02/06.
10. Buolamwini, J., Gebru, T.: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, 81, 1–15. (2018).
11. Widely used AI tool for early sepsis detection may be cribbing doctors' suspicions. University of Michigan News. (February 15, 2024). <https://news.umich.edu/widely-used-ai-tool-for-early-sepsis-detection-may-be-cribbing-doctors-suspicions/>. Last accessed 2025/05/01
12. Cheshmehzangi, et al.: Space and social distancing in managing and preventing COVID-19 community spread: An overview. *Heliyon*. 2023 Mar;9(3):e13879. doi: 10.1016/j.heliyon.2023.e13879.
13. Marr B.: 'Amazing Possibilities of Healthcare in the Metaverse' *Forbes* (23 Feb 2022) <https://www.forbes.com/sites/bernardmarr/2022/02/23/the-amazing-possibilities-of-healthcare-in-themetaverse/?sh=72fe5fc19e5c>. Last accessed 2023/05/10.
14. Mueller, B.: "Telemedicine Arrives in the U.K.: 10 Years of Change in One Week", *New York Times*, 4 April 2020, available at <https://www.nytimes.com/2020/04/04/world/europe/telemedicine-uk-coronavirus.html>. Last accessed 2023/05/10
15. Negreiro, M.: The Rise of Digital Health Technologies during the Pandemic (EPRS, PE 690.548, April 2021) [https://www.europarl.europa.eu/Reg-DATA/etudes/BRIE/2021/690548/EPRS_BRI\(2021\)690548_EN.pdf](https://www.europarl.europa.eu/Reg-DATA/etudes/BRIE/2021/690548/EPRS_BRI(2021)690548_EN.pdf)
16. Bestsenny, et al.: 'Telehealth: A Quarter-Trillion-Dollar Post-COVID-19 Reality?' McKinsey & Company (9 July 2021) <https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/telehealth-a-quarter-trillion-dollar-post-covid-19-reality>.
17. Price, W. N, II.: "Medical AI and Contextual Bias." *Harvard Journal of Law & Technology* 33, no. 1 (2019): 65–116.
18. Rajkomar A., Dean J., Kohane I.: 'Machine Learning in Medicine' (2019) 380(25) *The New England Journal of Medicine* 1347.
19. Liew C.: 'The Future of Radiology Augmented with Artificial Intelligence: A Strategy for Success' (2018) 15(3) *European Journal of Radiology Open* 298.
20. Nnawuchi, U., George, C.: Orwellian odyssey: Smart borders and the imperative for explainability. In Farhaoui, Y., Herawan, T., Imoize, A.L. and El Allaoui, A. (eds) *Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to*

- Applications in Smart Environment. ICAISE 2024. Lecture Notes in Networks and Systems series Vol 2 LNNS 1397. Springer <https://link.springer.com/book/9783031909207> (2025).
21. European Commission. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending certain Union legislative acts. COM(2021) 206 final, 21 April 2021. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>.
 22. European Commission. Proposal for a Regulation on the European Health Data Space. COM(2022) 197 final, 3 May 2022. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0197> Last accessed 2025/05/21.
 23. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.
 24. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC.
 25. Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices. Official Journal of the European Union, <https://eur-lex.europa.eu/eli/reg/2017/746/oj>. Last accessed: 2025/05/01.
 26. Emmanuel, E., Wachter, R.: Artificial Intelligence in Healthcare: Will the Value Match the Hype? JAMA. May 20, 2019. doi:10.1001/jama.2019.4914.
 27. Harvey, et al.: The future of technologies for personalised medicine. N Biotechnol. 2012 Sep 15;29(6):625-33. doi: 10.1016/j.nbt.2012.03.009. PMID: 23091837.
 28. Article 29 Working Party. (2017). Guidelines on Automated Individual Decision-Making and Profiling for the purposes of Regulation 2016/679 (WP251rev.01). European Commission
 29. Malone, E.R., Oliva, M., Sabatini, P.J.B. *et al.*: Molecular profiling for precision cancer therapies. Genome Med 12, 8 (2020). <https://doi.org/10.1186/s13073-019-0703-1>.
 30. Zhou, et al.: Concordance study between IBM Watson for Oncology and clinical practice for patients with cancer in China. The Oncologist, 24(3), e196-e204. (2019). <https://doi.org/10.1634/theoncologist.2018-0255>
 31. Henry, et al.: 'A Targeted Real-Time Early Warning Score (TREWScore) for Septic Shock' (2015) Science Translational Medicine. <https://doi.org/10.1126/scitranslmed.aab3719>
 32. Sandeep, R.: Explainability and artificial intelligence in medicine, The Lancet Digital Health, Volume 4, Issue 4, e214-e215.
 33. Ross, C., Swetlitz I.: 'IBM's Watson Supercomputer Recommended "Unsafe and Incorrect" Cancer Treatments, Internal Documents Show' *Stat* (25 July 2018).
 34. Barry M.J, Edgman-Levitan S.: Shared decision making--pinnacle of patient-centred care. Engl J Med 366:780-1, 2012.
 35. Pierce, R.: Machine Learning for Diagnosis and Treatment: Gymnastics for the GDPR. European Data Protection Law Review (3) 333-343, 2018.
 36. Edwards, L and Veale, M.: 'Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?' (2018) 16(3) IEEE Security & Privacy 46-54.
 37. Kohli M., Geis, R.: Ethics, Artificial Intelligence, and Radiology. Journal of the American College of Radiology, 15(9), p. 1317-1319, (2018).
 38. European Society of Radiology, 2019. What the radiologist should know about artificial intelligence an ESR white paper. Insights into Imaging, 10, doi.org/10.1186/s13244-019-0738-2
 39. Price, W.N., Cohen I.G.: 'Privacy in the Age of Medical Big Data' (2019) 105 Nature Medicine 117-119 <https://doi.org/10.1038/s41591-018-0272-7>.
 40. Beauchamp, T.L.: Principles of Biomedical Ethics (8th edn, Oxford University Press 2019).

41. World Medical Association. International Code of Medical Ethics. <https://www.wma.net/policies-post/wma-international-code-of-medical-ethics/>. Last accessed: 2025/05/01.
42. Hosny, et al.: Artificial intelligence in radiology. *Nat Rev Cancer*, 18(8), p.500-510, (2018).
43. Ghassemi, M., Oakden-Rayner, L., Beam, A.L.: The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021 Nov;3(11):e745-e750.
44. McDougall, R.J.: Computer knows best? The need for value-flexibility in medical AI *J Med Ethics*, 45 (3) (2019), pp. 156-160, 10.1136/medethics-2018-105118.
45. Herrera, F, Calderón, R.: Opacity as a Feature, Not a Flaw: The LoBOX Governance Ethic for Role-Sensitive Explainability and Institutional Trust in AI. <https://arxiv.org/html/2505.20304v1>, 18 May 2025
46. Rosenbacke, et al.: 'How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review *JMIR AI* 2024; [https://ai.jmir.org/2024/1/e53207DOI: 10.2196/53207](https://ai.jmir.org/2024/1/e53207DOI:10.2196/53207).
47. Grote T., Berens P.: On the ethics of algorithmic decision-making in healthcare. *J Med Ethics*. 2020 Mar;46(3):205-211. doi: 10.1136/medethics-2019-105586.
48. Arrieta, et al.: 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges towards responsible AI', Elsevier Information Fusion Vol. 58, pages 82 – 115. <https://dictionary.cambridge.org/dictionary/english/explanation> last accessed 21 October 2021
49. Inam, et al.: 'Explainable AI: How humans can trust AI', Ericsson White Paper GFTL-21:000529Uen. April, 2021.
50. Nnawuchi, U. , & George, C.: A Grand Entrance Without a Blueprint: *Journal of AI Law and Regulation* Vol 1, Issue 4 (2024). pp. 402-419. DOI: <https://doi.org/10.21552/aire/2024/4/6>.
51. Barocas, S., Selbst A.D.: "Big data's disparate impact." *Calif. L. Rev.* 104 (2016) 671.
52. Pavlidis, G.: Unlocking the black box: analysing the EU artificial intelligence act's framework for explainability in AI. *Law, Innovation and Technology*, 16(1), 293–308. 2024. <https://doi.org/10.1080/17579961.2024.2313795>.
53. Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR as amended), adopted 4 November 1950, entered into force 3 September 1953, ETS No. 5.
54. International Covenant on Civil and Political Rights, adopted 16 December 1966, entered into force 23 March 1976, 999 UNTS 171.
55. Roberts, A.: 'Portuguese parliament votes to allow limited euthanasia' (BBC News, 12 May 2023) <https://www.bbc.co.uk/news/world-europe-65574311> accessed 29 May 2025.
56. Citizens Advice (2025). Your right to life. <https://www.citizensadvice.org.uk/law-and-courts/civil-rights/human-rights/what-rights-are-protected-under-the-human-rights-act/your-right-to-life/>. Last accessed: 2025/05/29.
57. Doshi-Velez, F., Kim, B.: 'Towards A Rigorous Science of Interpretable Machine Learning' (2017). arXiv:1702.08608 <https://arxiv.org/abs/1702.08608>.
58. Gunning, D.: 'Explainable Artificial Intelligence (XAI)' (Defence Advanced Research Projects Agency (DARPA),2017). <https://www.darpa.mil/program/explainable-artificial-intelligence>
59. UNESCO.: Universal Declaration on Bioethics and Human Rights. <https://unesdoc.unesco.org/ark:/48223/pf0000146180>
60. OECD Principles on Artificial Intelligence. <https://www.oecd.org/going-digital/ai/principles/>. Last accessed 2025/05/01
61. Council of Europe. Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine (Oviedo Convention). CETS No.: 164, 4 April 1997.
62. Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects. October 2013 (WMA). <https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/>. Last accessed 2025/05/10.