

The ethics of non-explainable artificial intelligence: An overview for clinical nurses

Matthew Wynn, Senior Lecturer¹

¹School of Nursing and Advanced Practice, Liverpool John Moores University, Liverpool

Corresponding author: Matthew Wynn, m.o.wynn@ljmu.ac.uk

Abstract

Artificial Intelligence (AI) is transforming healthcare by enhancing clinical decision-making, particularly in nursing, where it supports tasks like diagnostics, risk assessments, and care planning. However, the integration of non-explainable AI (NXAI), which operates without fully transparent, interpretable mechanisms, presents ethical challenges related to accountability, autonomy, and trust. While explainable AI (XAI) aligns well with nursing's bioethical principles by fostering transparency and patient trust, NXAI's complexity offers distinct advantages in predictive accuracy and efficiency. This article explores the ethical tensions between XAI and NXAI in nursing, advocating a balanced approach that emphasises outcome validation, shared accountability, and clear communication with patients. By focusing on patient-centered, ethically sound frameworks, it is argued that nurses can integrate NXAI into practice, addressing challenges and preserving core nursing values in a rapidly evolving digital landscape.

Table 1. Potential liability and reasons (adapted from Terranova et al 2024).

Party liable	Reason	Nursing Example
Nurse	Failure to comprehend information provided by AI or failure to properly inform a patient of the limitations of AI or the health concern.	A nurse relies on AI to predict a patient's falls risk. The nurse doesn't understand or explain the limitations of the tool to the patient and over-trusts its prediction. The patient falls and is injured.
Patient	Underestimation of the recommendations provided by AI	A patient ignores an AI-based medication adherence reminder system leading to a poorer clinical outcome.
AI provider / software developers	Erroneous recommendations provided by AI to the nurse	An AI system miscalculates the risk of sepsis due to a biased data set which doesn't reflect the patients' demographics. This delays treatment for sepsis resulting in harm.

Table 2. Additional arguments in favour of NXAI (derived from Fryer et al 2024)

Broad argument in favour of NXAI	Example in Nursing
Nursing decisions are often atheoretical	Nurses use clinical intuition or heuristics in fast-paced environments, similar to NXAI's non-explanatory outputs.
Post-hoc explainability methods add uncertainty and false confidence	Nurses interpreting 'explainable' AI diagnostics might face overconfidence in AI-recommended care plans leading to harm.
Trade-off between accuracy and explainability	Highly accurate but opaque NXAI could provide vital diagnostic support for challenging cases, like early sepsis detection using novel biosensors.
Explainability not required and not sufficient for bias detection	Nurses using AI-assisted triage tools might encounter biases that are flagged by testing, not explainability alone.
Trust, acceptance, and uptake feasible by transparency	Transparency around AI limitations allow nurses to trust and incorporate AI tools for medication administration.
Associated risks of AI decision support systems determine explainability standards	In high-risk areas like ICU, nurses may require higher AI explainability to assess risks in critical care decisions.
Capacities and values of patients and nurses determine explainability standards	For more experienced nurses, less explainable AI may be necessary as they may be able to judge its outputs more effectively.
Potential benefits and lack of alternatives may outweigh explainability concerns	Nurses may use an NXAI risk assessment tool where the alternative non-digital solutions have shown limited efficacy, for example in pressure ulcers (Moore and Patton 2019).

Title: The ethics of non-explainable artificial intelligence: An overview for clinical nurses

Abstract:

Artificial Intelligence (AI) is transforming healthcare by enhancing clinical decision-making, particularly in nursing, where it supports tasks like diagnostics, risk assessments, and care planning. However, the integration of non-explainable AI (NXAI), which operates without fully transparent, interpretable mechanisms, presents ethical challenges related to accountability, autonomy, and trust. While explainable AI (XAI) aligns well with nursing's bioethical principles by fostering transparency and patient trust, NXAI's complexity offers distinct advantages in predictive accuracy and efficiency. This article explores the ethical tensions between XAI and NXAI in nursing, advocating a balanced approach that emphasises outcome validation, shared accountability, and clear communication with patients. By focusing on patient-centered, ethically sound frameworks, it is argued that nurses can integrate NXAI into practice, addressing challenges and preserving core nursing values in a rapidly evolving digital landscape.

Introduction:

Artificial Intelligence (AI) is revolutionising healthcare by enabling computer systems to perform tasks traditionally requiring human intelligence (High-Level Expert Group on Artificial Intelligence, 2019). These tasks include data analysis, predictive modelling, diagnostics, and even aspects of decision-making. AI's integration into healthcare, especially in nursing, leverages machine learning and deep learning algorithms that analyse vast data sets to identify patterns, make predictions, and suggest actions, potentially reshaping clinical decision-making and patient care. With the power to improve efficiency, optimise resource allocation, and enhance patient outcomes, AI holds significant promise in nursing, particularly in areas such as risk assessments, diagnostics, and care planning (Ruksakulpiwat et al., 2024; Ayoub et al., 2023).

A pressing ethical concern surrounding AI in healthcare however, is the 'explainability' of AI outputs. In many cases, AI operates as a 'black box' due to the opacity of its algorithms, particularly with advanced models like deep learning. Explainability in AI is not a binary feature but exists on a spectrum (Freyer et al 2024). On one end, there are systems with transparent, straightforward algorithms that allow users to fully understand how outputs are derived, akin to conventional medical devices with clear mechanisms. Coeckelbergh (2012) describes this as the 'functionalist performance criterion'. On the other end, there are highly complex AI systems that resist full explanation due to adaptive, non-linear algorithms which may rely on millions of parameters (Hassija et al., 2023). An example of this type of AI is ChatGPT, the chatbot developed by OpenAI. Whilst the general mechanics of this system (deep learning and pattern recognition) are understood, the reasoning behind specific outputs are never fully transparent or interpretable. This lack of transparency in complex AI systems has led to debates on how to integrate such tools ethically into clinical practice, especially when decision-making involves critical aspects of patient care. A recent extensive mapping review of the ethics of AI in healthcare by Morley and Floridi (2024) identified 18 key questions which require further consideration to support the ethical adoption of AI, this included the question 'How can the explainability of AI-Health solutions be guaranteed?'

This reflects on ongoing debate in AI ethics. A more detailed review on the issue of explainability by Freyer et al (2024) found that most literature in this area is, perhaps surprisingly, in favour of using non-explainable AI. This article will consider the arguments for and against explainable AI in the context of nursing, and how non-explainable AI might be ethically implemented within the context of nursing practice.

Arguments in favour of explainable AI in nursing

To better understand why explainability is emphasised, it is useful to consider some key arguments in its favour. These are presented in relation to the bioethical principles autonomy, justice, beneficence and non-maleficence, initially proposed by Beauchamp and Childress (2013) and advocated for by the International Council of Nurses (2021). Arguments are also derived from those identified within the review by Freyer et al (2024) and applied to the nursing context.

Firstly, explainable AI (XAI) in nursing may serve as a critical tool for fostering ethical, transparent, and accountable care in line with the profession's bioethical principles. One of the primary ethical advantages of XAI is its ability to support *accountability*. By offering nurses clear insights into how AI systems generate outputs, XAI preserves nurses' accountability in clinical decisions. When nurses can understand the underlying reasoning behind AI-driven predictions, such as those for potential drug interactions, they are better equipped to validate and integrate these insights into patient care plans, ensuring they can confidently stand by the decisions made with AI assistance. Another essential function of explainable AI is its capacity to *detect and correct biases*, aligning with the principles of *justice* and *non-maleficence*. AI systems, especially when tasked with diagnostics or risk assessments, may unintentionally reflect biases present in the data they were trained on. XAI makes it potentially easier to uncover these biases, allowing nurses to address any discrepancies that could impact equitable care. For example, when a diagnostic AI tool reveals gender biases, nurses can adjust their clinical judgments to ensure fair and unbiased treatment across all patient demographics.

Explainable AI may also strengthen the *trust* essential for both patients and healthcare providers, making its adoption in clinical settings smoother and ethically sound. By enhancing transparency, XAI supports the *autonomy* and *justice* principles, helping patients and nurses feel more secure in the AI's recommendations. Patients may be more likely to trust and accept AI-assisted recommendations when nurses can clearly communicate the reasoning behind them, reducing resistance and fostering collaborative decision-making. Supporting *professional autonomy* is another benefit of XAI, as it enables nurses to exercise their independent judgment. Through access to detailed information about AI outputs, nurses can use XAI to supplement their expertise rather than feel constrained by a black-box system. This respects both *accountability* and *autonomy* by empowering nurses to use AI in ways that support their clinical reasoning.

Explainable AI may also allow nurses to align care recommendations with *patient values*, enhancing *beneficence* and *autonomy* through patient-centered approaches. For instance, in sensitive areas like end-of-life care, XAI can assist nurses in presenting options that align with the patient's expressed values and wishes. By understanding the AI's reasoning, nurses can

offer guidance that respects patient autonomy and supports beneficial outcomes aligned with their preferences. Finally, explainable AI may help *prevent false hope*, minimising the risk of unrealistic expectations or potentially harmful interventions, a principle tied closely to *non-maleficence* and *beneficence*. By explaining the limitations and reliability of AI-driven diagnostics, nurses may set realistic goals, ensuring patients understand that AI tools are aids to judgment rather than definitive authorities. This transparency reduces the risk of over-reliance on AI, fostering a balanced and ethical approach to care.

Arguments in favour of non-explainable AI in nursing

Despite the ethical challenges, non-explainable AI (NXAI) offers unique advantages in healthcare, particularly in nursing, where its deep learning and predictive capabilities may enhance patient care. NXAI systems, even without full transparency, might improve clinical efficiency, support timely interventions, and augment decision-making in ways that explainable AI (XAI) may not. An example of what this might look like in practice is provided in vignette 1.

Vignette 1: NXAI in Practice – Pressure ulcer risk prediction

Sarah is a nurse in a long-term care facility where patients are at high risk for pressure ulcers. Recently, the facility implemented an NXAI tool designed to predict pressure ulcer risk based on patient data, including movement patterns, medical history, and skin assessments. Sarah is trained to use the tool, which has shown high reliability in risk prediction. Although the AI's process is not fully explainable, she observes that its predictions often flag high-risk patients before visible signs of skin breakdown occur. One day, the NXAI system identifies Ms. Greene as high-risk for pressure ulcers, even though she appears stable. Although Sarah doesn't fully understand the AI's algorithm, she knows its recommendations are based on data patterns that might not be immediately obvious. Trusting the AI's track record, Sarah decides to adjust Ms. Greene's care plan, increasing her repositioning schedule and adding extra skin checks. Days later, the NXAI prediction proves accurate when Sarah notices early signs of skin breakdown, allowing for timely intervention and prevention of ulcer progression.

By relying on NXAI's validated performance and using it as a supplement to her judgment, Sarah upholds her ethical commitment to patient care while benefiting from NXAI's predictive capabilities.

The following section explores the arguments that favour NXAI's integration into nursing, suggesting that its opacity can be ethically acceptable when balanced with reliability, performance, and patient-centered application. There are two broad issues which arguably require special attention in the case of NXAI due to its inherently opaque nature. Firstly, the issue of accountability for actions influenced by NXAI, and second the issue of autonomy.

Issue 1: Accountability

Nurses, who bear direct responsibility for patient care, are accustomed to working with tools they can fully understand and explain. NXAI challenges this model, as it operates with

complex, adaptive algorithms that lack immediate transparency, posing unique ethical considerations for accountability in its deployment. Floridi and Sanders (2004) propose that while AI systems may lack intrinsic moral agency and intentionality, they can still function as ‘artificial moral agents’ capable of supporting morally significant actions. Arguably, accountability in the context of NXAI should therefore not rest solely on individual clinicians but should be shared among developers, institutions, and end-users, creating a network of ‘distributed responsibility’. This approach may allow nurses to ethically integrate NXAI into their practice by focusing on outcome reliability and system-wide validation, rather than requiring a full grasp of NXAI’s mechanisms. In this model, accountability is maintained through collaborative oversight, continual performance assessment, and structured feedback systems, enabling nurses to apply NXAI in ways that support patient care. A recent review by Terranova et al (2024) sought to explore the implications of how liability might be established when AI is used within healthcare. Table 1 illustrates an adaptation of their guidance on this issue which recognises the necessity for distributed responsibility in some cases:

Table 1. Potential liability and reasons (adapted from Terranova et al 2024).

Party liable	Reason	Nursing Example
Nurse	Failure to comprehend information provided by AI or failure to properly inform a patient of the limitations of AI or the health concern.	A nurse relies on AI to predict a patient's falls risk. The nurse doesn't understand or explain the limitations of the tool to the patient and over-trusts its prediction. The patient falls and is injured.
Patient	Underestimation of the recommendations provided by AI	A patient ignores an AI-based medication adherence reminder system leading to a poorer clinical outcome.
AI provider / software developers	Erroneous recommendations provided by AI to the nurse	An AI system miscalculates the risk of sepsis due to a biased data set which doesn't reflect the patients' demographics. This delays treatment for sepsis resulting in harm.

Issue 2: Autonomy

Autonomy is a fundamental principle in biomedical ethics, centered on patient self-determination and informed decision-making. Historically, it marked a shift from a paternalistic approach in healthcare to one that empowers patients as active participants in their own care. Respecting autonomy involves transparent communication about risks, benefits, and alternatives, building a trust-based clinician-patient relationship. Yet, research by Christen et al (2014) suggests that autonomy may not carry the same intuitive moral weight as other principles, such as beneficence or non-maleficence. In their study, they

found that autonomy functioned less as a core ethical imperative and more as a ‘bridge value’, facilitating practical decision-making without the universally shared moral resonance of other principles, for example, non-maleficence.

This insight raises important considerations when introducing non-explainable AI (NXAI) into nursing practice. NXAI systems, which rely on complex, opaque algorithms, inherently challenge the ideal of fully informed patient choice because their internal processes are not readily explainable. Traditionally, respecting autonomy in healthcare has involved a degree of transparency that allows patients to understand how decisions affecting their care are made. However, Christen et al’s (2014) findings suggest that, in practice, the moral emphasis may not always lie on complete transparency. Instead, it often shifts to ensuring beneficial outcomes. In this context, autonomy in nursing becomes a flexible, adaptive principle rather than a rigid requirement. Nurses might prioritise discussing NXAI’s recommendations in ways that resonate with the patient’s values and concerns, incorporating these insights into a broader clinical conversation. Autonomy here is respected not through exhaustive explanation but by upholding the patient’s ability to make decisions that align with their own goals and understanding of care. By reframing autonomy in this way, nurses can strike a balance between respecting patient independence and harnessing the potential of NXAI to support optimal clinical outcomes, guiding patients toward a shared destination: effective, values-consistent care.

Further arguments in favour of using NXAI in healthcare were identified by Fryer et al (2024). These arguments are provided in Table 2 alongside relevant examples in nursing practice.

Table 2. Additional arguments in favour of NXAI (derived from Fryer et al 2024)

Broad argument in favour of NXAI	Example in Nursing
Nursing decisions are often atheoretical	Nurses use clinical intuition or heuristics in fast-paced environments, similar to NXAI’s non-explanatory outputs.
Post-hoc explainability methods add uncertainty and false confidence	Nurses interpreting ‘explainable’ AI diagnostics might face overconfidence in AI-recommended care plans leading to harm.
Trade-off between accuracy and explainability	Highly accurate but opaque NXAI could provide vital diagnostic support for challenging cases, like early sepsis detection using novel biosensors.
Explainability not required and not sufficient for bias detection	Nurses using AI-assisted triage tools might encounter biases that are flagged by testing, not explainability alone.
Trust, acceptance, and uptake feasible by transparency	Transparency around AI limitations allow nurses to trust and incorporate AI tools for medication administration.
Associated risks of AI decision support systems determine explainability standards	In high-risk areas like ICU, nurses may require higher AI explainability to assess risks in critical care decisions.

Capacities and values of patients and nurses determine explainability standards	For more experienced nurses, less explainable AI may be necessary as they may be able to judge its outputs more effectively.
Potential benefits and lack of alternatives may outweigh explainability concerns	Nurses may use an NXAI risk assessment tool where the alternative non-digital solutions have shown limited efficacy, for example in pressure ulcers (Moore and Patton 2019).

In addition to these arguments, it is possible that NXAI systems could serve as passive, background tools within nursing, much like AI operates in fields such as marketing, advertising, and finance to yield actionable insights. By continuously analysing data (e.g. from electronic patient record systems) and identifying patterns, NXAI could highlight trends, emerging risks, or inefficiencies that nurses might otherwise overlook in daily practice. These insights could then inform nursing actions through traditional research or quality improvement (QI) methods, allowing nurses to respond with evidence-based interventions. This passive approach would leverage NXAI's analytical power while maintaining nurses' central role in interpreting and implementing changes at the level of individual patient interactions; therefore not disrupting established nursing processes and standards for patient care.

Finally, given the person-centered, rather than pathology-focussed nature of nursing practice, it is necessary for further engagement with patient populations to establish ethical consensus and preferences around the potential use of NXAI in nursing contexts. This will help establish ethical frameworks which are not focussed primarily on the needs and expectations on nurses and healthcare systems, but also consider the perspectives of patients which may well conflict. This issue of patient expectations differing from that of nurses has already been reported in the context of robotics in nursing care (Wynn 2024). It may be the case that many patients would prefer the option to utilise NXAI tools should they lead to greater outcomes.

Conclusion

The integration of Artificial Intelligence, especially non-explainable AI (NXAI), presents both transformative potential and ethical challenges in nursing. NXAI may enhance decision-making, improve patient outcomes, and boost clinical efficiency, yet its inherent opacity raises crucial questions about autonomy, accountability, and trust. While traditional ethics in healthcare emphasises transparency, autonomy in practice can function as a flexible guide, adaptable to the clinical context and patient needs. The ethical use of NXAI in nursing may therefore rely on a balanced approach that prioritises outcome validation, establishes shared accountability, and transparently communicates NXAI's role and limitations. Rather than full explainability, the focus shifts to ensuring that NXAI aligns with patient values, clinical goals, and nursing responsibilities, ultimately supporting patient-centered care.

Nurses play a pivotal role in realising NXAI's potential within an ethically sound framework. By fostering trust through clear communication, engaging in outcome monitoring, and collaborating across disciplines, they can uphold professional standards while embracing the benefits NXAI offers to patient care. This dynamic approach would enable NXAI to

complement nursing practice, preserving a commitment to patient welfare while adapting traditional ethical principles to the evolving digital landscape.

References:

- Ayoub, M., Ballout, A. A., Zayek, R. A., & Ayoub, N. F. (2023). Mind + Machine: ChatGPT as a basic clinical decision support tool. *Cureus*, *15*(8), e43690. <https://doi.org/10.7759/cureus.43690>
- Beauchamp, T., & Childress, J. (2013). *Principles of Biomedical Ethics* (7th ed.). Oxford University Press.
- Christen, M., Ineichen, C., & Tanner, C. (2014). How “moral” are the principles of biomedical ethics? – A cross-domain evaluation of the common morality hypothesis. *BMC Medical Ethics*, *15*(47). <https://doi.org/10.1186/1472-6939-15-47>
- Coeckelbergh, M. (2012). Can we trust robots? *Ethics and Information Technology*, *14*(1), 53–60. <https://doi.org/10.1007/s10676-011-9279-1>
- Floridi, L. (2013). *The Ethics of Information*. Oxford University Press.
- Floridi, L., & Sanders, J. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349–379. <https://doi.org/10.1023/B:MIND.0000035461.63578.9d>
- Freyer, N., Groß, D., & Lipprandt, M. (2024). The ethical requirement of explainability for AI-DSS in healthcare: A systematic review of reasons. *BMC Medical Ethics*, *25*(1). <https://doi.org/10.1186/s12910-024-01103-2>
- Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2023). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, *16*(1), 45–74. <https://doi.org/10.1007/s12559-023-10179-8>
- High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. Retrieved from <https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf>
- International Council of Nurses. (2021). *The ICN Code of Ethics For Nurses*. https://www.icn.ch/sites/default/files/2023-06/ICN_Code-of-Ethics_EN_Web.pdf
- Moore, Z. E. H., & Patton, D. (2019). Risk assessment tools for the prevention of pressure ulcers. *Cochrane Database of Systematic Reviews*, *2019*(1), Article CD006471. <https://doi.org/10.1002/14651858.CD006471.pub4>
- Morley, J., & Floridi, L. (2024). The ethics of AI in health care: An updated mapping review. Available at SSRN: <https://ssrn.com/abstract=4987317> or <http://dx.doi.org/10.2139/ssrn.4987317>
- Ruksakulpiwat, S., Thorngthip, S., Niyomyart, A., Benjasirisan, C., Phianhasin, L., Aldossary, H., Ahmed, B., & Samai, T. (2024). A systematic review of the application of artificial intelligence in nursing care: Where are we, and what’s next? *Journal of Multidisciplinary Healthcare*, *17*, 1603–1616. <https://doi.org/10.2147/jmdh.s459946>

Terranova, C., Cestonaro, C., Fava, L., & Cinquetti, A. (2024). AI and professional liability assessment in healthcare: A revolution in legal medicine? *Frontiers in Medicine*, *10*, 1337335. <https://doi.org/10.3389/fmed.2023.1337335>

Wynn, M. (2024). The digital dilemma in nursing: A critique of care in the digital age. *British Journal of Nursing*, *33*(11), 496–499. <https://doi.org/10.12968/bjon.2024.0023>