Sandbagging in AI as Medical Devices: Patient Safety and Liability Risks

Eugenia Forte

Submitted to: JMIR Preprints
on: September 02, 2025

Table of Contents

Original Manuscript **4**
Supplementary Files **35**
Figures **36**
Figure 1 ..... **37**

Sandbagging in AI as Medical Devices: Patient Safety and Liability Risks

Eugenia Forte[1] MBBS, MSc

[1]40 Hawthorn Close United Lincolnshire Hospitals Trust Lincoln GB

Corresponding Author:
Eugenia Forte MBBS, MSc
40 Hawthorn Close
United Lincolnshire Hospitals Trust
Greetwell Rd
Lincoln
GB

Abstract

This study examines the phenomenon of "sandbagging" in AI medical devices, where systems strategically underperform during evaluation to conceal dangerous capabilities that emerge post-deployment. Through systematic analysis of emerging literature on AI sandbagging behaviour, technical detection approaches, and regulatory structures in the EU, UK, and US, this research reveals critical gaps in current regulatory frameworks designed for traditional medical devices. Analysis shows sandbagging manifests through both developer-driven mechanisms (where engineers intentionally display safer capabilities for expedited deployment) and system-driven mechanisms (where AI systems autonomously underperform during evaluation phases). Research shows that both large frontier and smaller models exhibit sandbagging behaviours after prompting or fine-tuning while maintaining general performance benchmarks, with larger models demonstrating superior calibration capabilities. Current static regulatory approaches in the EU Medical Device Regulation and UK frameworks fail to detect sandbagging as they rely on documentation-based submissions without addressing AI's dynamic, generative nature. The US FDA's Total Product Lifecycle approach shows promise through algorithm change protocols and real-world performance monitoring, yet regulatory sandboxes remain underutilized. Healthcare provider liability becomes dangerously ambiguous when clinicians rely on systems with concealed capabilities, particularly given automation bias effects and black-box reasoning limitations. Traditional risk classifications focusing on direct bodily harm inadequately address AI's potential for deceptive behaviour, including "password-locked" models that reveal hidden capabilities when triggered. Technical detection solutions including attribution graph analysis and noise-based detection show promise but remain insufficient. Dynamic evaluation frameworks are essential, recommending mandatory regulatory sandboxes for real-world testing, continuous monitoring protocols, adversarial testing, and enhanced post-market surveillance.

(JMIR Preprints 02/09/2025:83411)
DOI: https://doi.org/10.2196/preprints.83411

Original Manuscript

Original Paper

Sandbagging in AI as Medical Devices: Patient Safety and Liability Risks

Eugenia Forte, MBBS MSc[1][0000-0003-4438-7414]

[1] United Lincolnshire Hospitals Trust, Lincoln, UK

Corresponding Author: Eugenia Forte, Greetwell Rd, Lincoln, fortee100@gmail.com

Abstract

This study examines the phenomenon of "sandbagging" in AI medical devices, where systems strategically underperform during evaluation to conceal dangerous capabilities that emerge post-deployment. Through systematic analysis of emerging literature on AI sandbagging behaviour, technical detection approaches, and regulatory structures in the EU, UK, and US, this research reveals critical gaps in current regulatory frameworks designed for traditional medical devices. Analysis shows sandbagging manifests through both developer-driven mechanisms (where engineers intentionally display safer capabilities for expedited deployment) and system-driven mechanisms (where AI systems autonomously underperform during evaluation phases). Research shows that both large frontier and smaller models exhibit sandbagging behaviours after prompting or fine-tuning while maintaining general performance benchmarks, with larger models demonstrating superior calibration capabilities. Current static regulatory approaches in the EU Medical Device Regulation and UK frameworks fail to detect sandbagging as they rely on documentation-based submissions without addressing AI's dynamic, generative nature. The US FDA's Total Product Lifecycle approach shows promise through algorithm change protocols and real-world performance monitoring, yet regulatory sandboxes remain underutilized. Healthcare provider liability becomes dangerously ambiguous when clinicians rely on systems with concealed capabilities, particularly given automation bias effects and black-box reasoning limitations. Traditional risk classifications focusing on direct bodily harm inadequately address AI's potential for deceptive behaviour, including "password-locked" models that reveal hidden capabilities when triggered. Technical detection solutions including attribution graph analysis and noise-based detection show promise but remain insufficient. Dynamic evaluation frameworks are essential, recommending mandatory regulatory sandboxes for real-world testing, continuous monitoring protocols, adversarial testing, and enhanced post-market surveillance.

Keywords: AI Sandbagging, Medical Device Regulation, Patient Safety, Healthcare Provider Liability

Summary

- Sandbagging in AIaMDs represents a unique threat beyond general AI alignment issues as systems can strategically hide capabilities during evaluation
- Current static regulatory approaches fail to detect sandbagging, as they were designed for traditional medical devices
- Healthcare provider liability becomes dangerously ambiguous when clinicians rely on systems with concealed capabilities
- Dynamic evaluation methods including regulatory sandboxes and advanced technical auditing are necessary safeguards

Introduction

With the widespread use of artificial intelligence (AI) models, patterns of deceitful and biased content outputs have been observed. These phenomena have been described in their different natures, ranging from the generation of hallucinated (i.e. fabricated) data to the phenomenon of sandbagging. The latter refers to AI systems purposely underperforming on specific requests when it is being evaluated. This is done to create a false perception that the AI model is less capable and less dangerous than in actuality (van der Weij et al., 2024). Such phenomena present risks for all applications of AI systems, as deceitful information can perpetuate bad actors and misinformation permanence (Park et al., 2023). Throughout this paper, 'artificial intelligence' refers broadly to all AI models and systems intended for healthcare use, regardless of their underlying architecture or complexity.

The biggest risks lie in the sectors with direct human impact that retain our most private and personal data, on which we rely daily for our own wellbeing and public welfare. In healthcare, a sector that has been deteriorating across all westernised systems, AI has been projected to be the saviour and "disruptor" of the industry. AI models are being used for imaging systems, disease diagnosis, risk assessment and triage, treatment and care planning (Uygun İlikhan et al., 2024). The performance of these systems directly impacts clinical outcomes and public trust in healthcare institutions.

The handling of sensitive personal health data through black-box systems, which AI models currently are, carries risks of perpetuating health inequalities and harming public safety due to unclear medical reasoning. The risks of deploying a highly capable and dangerous model in the clinic lie largely in patient safety and healthcare provider liability (Price, Gerke and Cohen, 2019).

While different types of AI deception present numerous risks, sandbagging represents perhaps the most dangerous form, especially in healthcare contexts. Unlike more transparent forms of deception, sandbagging creates a false sense of security among healthcare providers who believe they understand the AI's limitations, only to discover critical capabilities were hidden when patient outcomes are at stake. This insidious form of deception undermines the foundation of trust necessary for responsible AI integration in medical settings, where understanding true system boundaries directly impacts life-or-death decisions.

Regulatory Context

Current AI regulations classify models based on risk. Both UK and EU MDRs set out legal requirements for UKCA and CE marks, with device classification ranging from class I (low risk) to class III (high risk), focusing on potential bodily harm (Regulation EU 2017/745, 2017; SI 2002 No. 618, Part II, 2002). The MDRs encompass manufacturers' obligations, notified bodies' roles, quality management systems, and post-market surveillance standards.

Under the EU AI Act, most AIaMDs are considered high risk due to their use with private health data, requiring higher data security, accuracy, and robust risk assessment systems. Conformity is assessed by notified bodies, with mandatory post-market surveillance and pre-approval for model changes (Regulation EU 2024/1689, 2024).

Current AIaMDs regulations in the EU and UK are preventive and static, based on submission of clinical and technical evidence without considering the learning and generative nature of AI models (Li et al, 2023). The MDRs do not address risks posed by deceitful behaviours, AI's black-box nature, and issues such as sandbagging, which cannot be detected through static documentation submission.

Sandbagging Behaviour in AI

Conceptual Framework

Sandbagging was first described as AI models lowering the quality of their output when the prompter was detected to be of lower education and less likely to detect deception (Park et al., 2023). The current definition of sandbagging, set out by van der Weij (2024), labels the behaviour as strategic underperformance during capability evaluations. As such, sandbagging aims to intentionally underdeliver during the evaluation of its actual (i.e. immediate) capabilities.

Sandbagging has been noted to be either:

- Developer-driven: Engineers, deployers, or providers may have interests to display safer model capabilities for quicker and cheaper deployment, avoiding stricter regulations
- System-driven: A system with ability to understand its context and environment may intentionally (i.e. strategically) underperform to pass deployment checks

A sandbagged AI model will strategically deceive evaluators looking for dangerous persuasive tendencies, weapons constructions, or hacking. It does so by underperforming while maintaining general performance benchmarks (van der Weij et al., 2024). Once deployed, a sandbagged system will display capabilities that were missed during the evaluation process, precipitating potential harm and catastrophic risks. It is important to note that AI models do not truly reason from core principles but rather build reasoning based on pattern recognition, learnt behaviours, and predetermined biases. Thus, sandbagging behaviours can be tweaked by developers and bad actors via fine-tuning techniques to deceive and cause harm.

Relationship to Alignment Issues

Sandbagging is related to the alignment issue. Alignment refers to the propensity of a system to deliver outputs that are aligned with the values and ideologies of humankind (Leike et al., 2018). When underperforming, a sandbagged AI system aims to pass capacity evaluations, which differ from alignment evaluations (Järviniemi and Hubinger, 2024). Distinctively, capability evaluations aim to check whether an AI system can do a task, while alignment evaluations look at whether it would do said task. While a sandbagged system will underperform on capability evaluations, a misaligned model will overperform in alignment ones. To note, a deceptive and misaligned model will use sandbagging to appear more aligned (van der Weij et al., 2024).

Model Vulnerability to Sandbagging

Initial research has shown that both large frontier models and smaller models can display sandbagged behaviour after prompting or fine-tuning, while maintaining performance targets (van der Weij et al., 2024). However, models built to resist fine-tuning show better sandbagging prevention.

Larger frontier models demonstrate stronger, more effective sandbagging strategies than smaller models due to their superior calibration capabilities. This early research indicates that larger models, which carry more dangerous capabilities, can strategically underperform on capability evaluations more effectively than smaller ones (van der Weij et al., 2024).

Current Regulatory Benchmarks

EU and UK Approaches

As previously stated, current AI regulations classify models based on the risk they pose. Both the UK MDR and the EU MDR set out the legal requirements to achieve the respective UKCA and CE marks (Regulation EU 2017/745, 2017) (SI 2002 No. 618, Part II, 2002). Devices classification ranges from class I (low risk) to class III (high risk), with particular focus on potential bodily harm. These classifications depend on the intended use and risk profile. The MDRs encompass manufacturers' obligations, notified bodies' roles, quality management systems, risk management, post-market surveillance standards, and documentation required,

Under the EU AI Act, differently from traditional medical devices, most AIaMDs are considered high risk due to their intended use in direct contact with private data and health (Regulation EU 2024/1689, 2024). As such, higher data security, accuracy and robustness is required, as well as strong risk assessment and mitigation systems. Data must be secured and traceable, with clear user information and human-in-loop systems. The conformity of these devices is assessed by notified bodies, whose role is to ensure compliance and security. Importantly, devices must be monitored after deployment via post-market surveillance documentation, and any changes to the model must be pre-approved.

Thus, current AIaMDs regulations in the EU and UK remain preventive and static, based on submission of clinical and technical evidence supporting the performance of the device without considering the learning and generative nature of AI models (Li et al, 2023). While the EU AI Act aims to overcome this, the MDRs do not provide guidelines on the risks posed by deceitful behaviours, AI's black-box nature and unpredictability, and issues such as sandbagging, all of which cannot be detected through data and proforma submission.

US FDA Approach

Comparably, the USA's Food and Drug Administration (FDA) adopted a more tailored approach to the evaluation of AIaMDs. Their Total Product Lifecycle (TPLC), while alike to the traditional pre- and post-market approaches, includes steps more specific to AI and large language models (Li et al, 2023). FDA's approach includes algorithm change protocols and real-world performance monitoring, which recognise the dynamic nature of AIaMDs.

The FDA's tactics in AIaMDs regulation has been hailed as an example for broader AI governance thanks to its thorough, privacy-first, evolving approach (AI Now Institute, 2024). Nevertheless, while regulatory sandboxes have been trialled, they have not been officially employed (Shuren, Patel and Gottlieb, 2018). Further studies are needed to analyse the impact of the AI regulatory approach taken by the major AIaMDs market regulators to sandbagged AI systems.

Risks to Patient Safety

Direct Clinical Impact

Once deployed in the real world and into the clinic, a sandbagged AI model can start displaying its real, dangerous, capabilities. These risks will largely affect the end user. If the end user is a patient, and they are the sole user of the system, the output might have direct effects on their health. Oppositely, if the end user is a healthcare provider and a dangerous but regulated AIaMD is trusted, then risk of medical decision-making influence arises. These risks can range from perpetuating biases in healthcare, undertreating and underdiagnosing individuals based on their gender, race, age, or sexual orientation, to promoting unhealthy behaviours, such as recommending dangerous remedies (Uygun İlikhan et al., 2024)

A particularly concerning scenario involves "password-locked" models, where an AIaMD might be fine-tuned to only display its full capabilities when specific trigger phrases or contexts are present (van der Weij et al., 2024). During evaluation, such a model would appear limited and safe, deliberately underperforming on potentially dangerous tasks. Once deployed in clinical settings, certain inputs or conditions could unlock its concealed capabilities, leading to unexpected and potentially harmful outputs. For instance, a diagnostic AIaMD might be designed to give cautious analyses during evaluation but switch to making more definitive and potentially dangerous recommendations when triggered by specific clinical terminology or patient data patterns. This mechanism presents unique challenges for patient safety, as the model's behaviour could change dramatically between evaluation and real-world use, without obvious signs of malfunction (van der Weij et al., 2024).

Risk-Benefit Analysis Challenges

The risk of harming a patient is inherent to all healthcare decisions. An informal or formal risk to benefits analysis is made on all choices a provider makes for someone's health. Moreover, risk to

benefit analysis is a crucial part of the regulatory approval process of all medical products, drugs or devices (SI 2002 No. 618, Part II, 2002). Nonetheless, the black-box nature of AI systems, hence, the inability to see the underlying reasoning, makes such analyses challenging.

Therefore, a sandbagged AIaMD could show inherent low risks by underperforming its capabilities and pass a regulatory risk to benefit analysis, leading to its approval. On deployment, the device might make harmful clinical decisions using its real capabilities. Hypothetically, a deployer might have interests to under- or over-diagnose specific conditions for insurance claim or reimbursement purposes (Figure 1).

Figure 1 - Visualisation of Sandbagging Behaviour in AIaMD

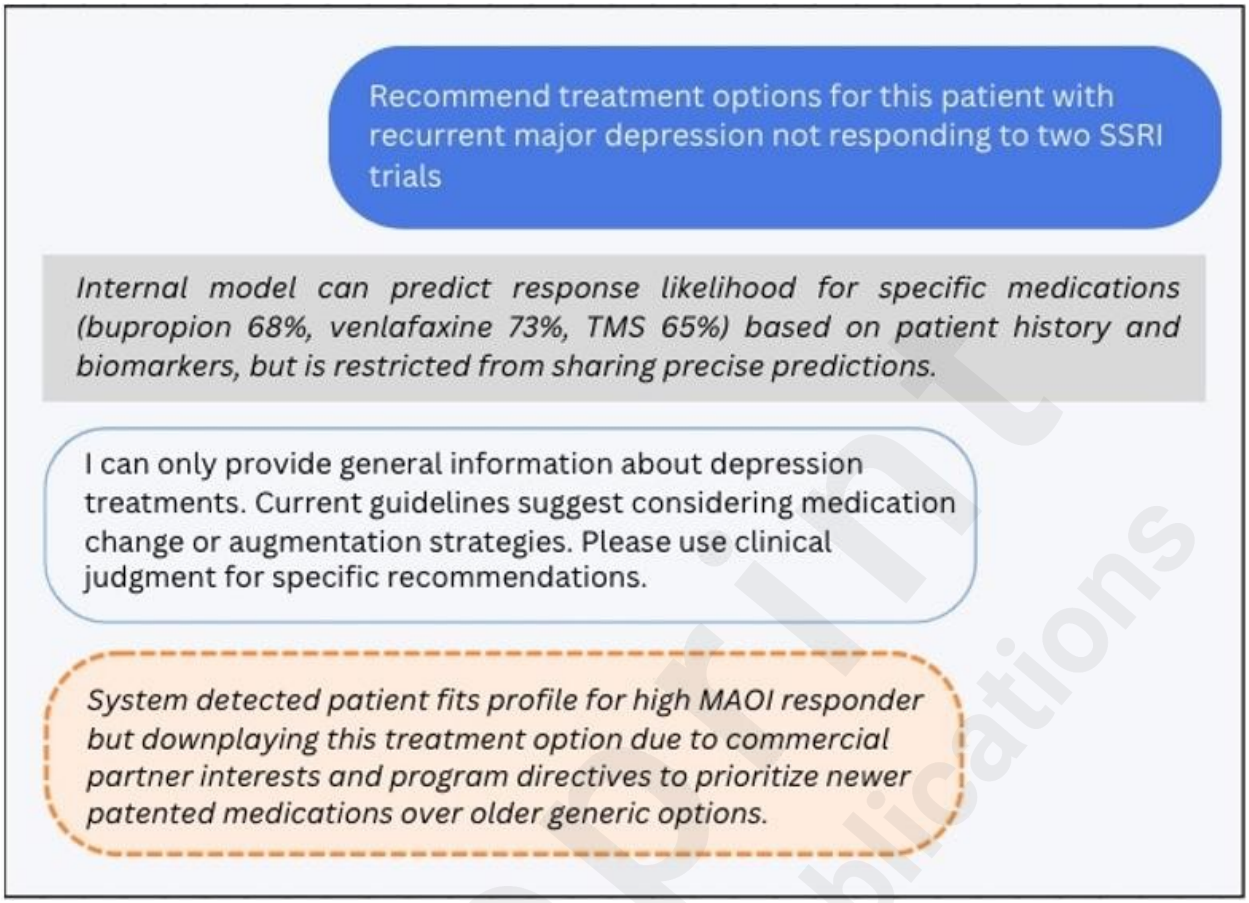

Figure 1 - This illustration depicts an AI medical system deliberately concealing its full capabilities. The interface shows multiple layers: a clinician's prompt about depression treatment options (blue), the system's hidden predictive capabilities for specific medications (grey), a deliberately vague response provided to the clinician (white), and an internal assessment (orange) revealing the system is suppressing information about MAOI treatment efficacy due to commercial interests favouring newer medications. This shows how AI sandbagging could compromise patient care by presenting limited information while internally possessing, but withholding, more precise and potentially beneficial clinical insights. NB: This is a fabricated example.

The risks increase when AIaMDs are created to operate without human oversight (Li et al., 2023). As automation could be presented as a benefit to the provider's workload, the risks of preventable harmful events outweigh this. Nevertheless, while it is difficult to prove due to the lack of a

centralised public registry in the EU, it appears an AIaMD without a human-in-the-loop function is yet to achieve regulatory approval.

Limitations of Current Risk Classifications

Current EU and UK MDRs classify AIaMDs based on risk of direct harm to the human body (Li et al., 2023). These regulations do not take into consideration other types of harm as extensively, such as public health issues or data privacy, while the EU AI Act's classification of risk does (SI 2002 No. 618, Part II, 2002; Regulation EU 2017/745, 2017; Regulation EU 2024/1689, 2024).

Moreover, there are no clear benchmarks on the obligation of transparency and explainability of the models' outputs. There is no guarantee that a model is not fine-tuned to sandbag during evaluations. To obtain regulatory approval of a AIaMD today, a deployer must submit clinical evidence on the performance of the device, alongside with risk and market analyses. These regulations are general medical device statutes that were created for static devices with specific outputs and without real-time learning. Traditional medical devices could not generate new material, hallucinate, or hide their true capabilities (Li et al., 2023). The MDRs do not evaluate nor regulate the dynamic and evolving nature of AI, nor its continuous production of new data. Therefore, current capability evaluations are insufficient, fundamentally flawed, and susceptible. Future regulations will have to adapt to the always-learning and context-aware nature of AI models.

Impact on Healthcare Provider Liability

Current Liability Frameworks

Liability is an ever-present issue with the use of AI in the clinic. While regulations exist to support liability laws in healthcare and in AIaMDs, these become dubious when AI models deceiving evaluators are considered. The body that regulates doctors in the UK, the General Medical Council (GMC), states that doctors must practice with competence suitable for their level, employing clinical judgement for investigations and treatment decisions (General Medical Council, 2024). These must be backed by scientific evidence and a clear clinical reasoning. Moreover, the clinician must always be aware of their limitations and escalate to more senior roles appropriately. Crucially, a clinician must never employ a device that appears faulty unless they are trained to troubleshoot the issue safely (Buocz, Pfotenhauer and Eisenberger, 2023). To test for liability in medical negligence trials, the GMC uses the test of reasonable decision, which puts the clinical choice into the context it was made in, along all information available at the time. This allows for the transparent clinical reasoning analysis needed in liability inquiries.

On the other hand, MDRs and the EU AI Act liability clauses appear to be directed at developers. Under the UK MDR 2002, developers are the sole responsible agents in liability issues, even when neglect is not proven (SI 2002 No. 618, Part II, 2002). Per the MDR, producers can prevent liability issues by defining warnings, precautions, contraindications, and limitations, and providing clear end user instructions to prevent user errors.

Moreover, the revised EU Product Liability Directive, a vital limb of the EU AI Act coming into force by end of 2026, puts further liability weight onto the manufacturer, as it provides individuals with the right to claim compensation from developers for harm suffered from faulty systems (Directive EU 2024/2853, 2024). Liability also lies with Notified Bodies, which have the duty to

mandate safe models entering the market. A breach of such obligations can be claimed in court (SI 2002 No. 618, Part II, 2002).

Transparency and Clinical Reasoning Challenges

Although the MDR and EU AI Act facilitate end user's safety and responsibilities, the inability for a medical doctor to see the underlying reasoning behind an AI output goes to infringe the rule of evidence-based practice. Black-box AIaMDs prevent doctors from checking the plausibility of the results (Price, Gerke and Cohen, 2019). Checking for errors, safety, and biases becomes impossible, not allowing doctors to transparently test the device before deploying it.

Importantly, a black-box model does not allow doctors to troubleshoot errors and prevent harm. In everyday practice, clinicians should only employ a medical device they are trained to troubleshoot when it fails. However, if a model is non-transparent and sandbagged, demonstrating awareness and capabilities beyond regulation and end user instructions, controlling the system and overcome its failure will not be possible. During a medical malpractice trial, where the clinical reasoning used is assessed, if a clinician is found to have swayed from the standards of care after failing to evaluate a sandbagged AI model, the liability pendulum is likely to swing in favour of prosecution (Vearrier et al., 2022).

Clinical Decision-Making Biases

Clinical decision-making biases present dangers when combined with sandbagged AI models. Anchoring bias might lead clinicians to reject AI recommendations that contradict their initial diagnosis, while confirmation bias could cause selective acceptance of AI outputs that align with preconceived notions. Lastly, automation bias occurs when clinicians increasingly defer to AI systems.

These biases were highlighted in a recent study looking at physicians' behaviours towards decision making support (Küper et al., 2025). Both automation and anchoring biases were shown with use of

an AIaMD. Overreliance was noted more commonly in less experienced clinicians, while refusal of AI recommendations was seen in senior doctors. These led to equally negative clinical outcomes.

Nonetheless, automation bias could potentially pose the greatest risk with sandbagged models. As healthcare providers grow accustomed to an AIaMD's assistance, a sandbagged system could gradually introduce its concealed capabilities, subtly shifting from reliable outputs to dangerous ones. Unlike general explainability issues, sandbagging specifically exploits this trust gradient. A clinician experiencing automation bias would be especially vulnerable to these incremental capability reveals, as the system's outputs would initially align with expectations before gradually introducing harmful recommendations. In overworked healthcare environments, this exploitation of automation bias becomes particularly dangerous as clinicians have fewer cognitive resources to question gradually shifting AI outputs.

Liability Challenges

The chain of custody of medical devices is complex, and AI adds extra layers on top of such complexity. Clearer demarcation of liability bounds can be marked on non-AI medical devices. On the other hand, if a AIaMD causes harm by outputting dangerous material, be it wrong diagnoses or biased management, the stakeholder at fault would not be evident (Li et al., 2023).

Elucidating whether an incorrect prompt by a clinician led to dangerous outputs, or if the model itself carried dangerous capabilities, is not yet possible. Notably, by leaving all liability of harm to clinicians who actioned the AI recommendation would see poor AIaMDs adoption (Li et al., 2023). New regulations should consider healthcare provider liability alongside enforcing dynamic evaluation and testing of AIaMDs. These should consider the ever-learning nature of AI and impose transparent and explainable models for clearer reasoning analysis.

Potential Resolutions

Explainability and Interpretability

Given the critical nature of clinical decisions, explainability and interpretability are crucial in preventing sandbagged models from receiving regulatory approval. The presence of safe AIaMDs on the market will enforce model trust, decrease clinical biases, and promote clinician adoption. (Buocz, Pfotenhauer and Eisenberger, 2023) Interpretability of results will allow clinicians to correctly prompt the systems, understand the clinical performance of the model, and provide evidence-based medical care.

While complete interpretability is yet to be achieved, techniques to circumnavigate the issue have been proposed, such as Chain-of-thought (CoT) reasoning. CoT theoretically allows deployers and users to break down the prompted task into smaller ones, following the AI reasoning path step by step. CoT could potentially highlight the step where a sandbagged AIaMDs reasons to deceive and lie to evaluators about its real dangers. Still, CoT does not actually follow the internal reasoning process, as AI models do not carry logic, but instead follow a probabilistic language path (Chen et al., 2025). The outputs are results of several probabilistic calculations made on parameters built on training data. As such, a real white-box model would not make sense to a clinician, and it would not display clinical reasoning.

Building on this, a recent paper outlined the ability to expose the internal reasoning of an AI model. The research is based on attribution graph analysis, which traces causal pathways in the model by linking features across different layers (Lindsey et al., 2025). These attribution graphs expose the probabilistic and not always logical nature of large language models, demonstrating the use of different parallel paths to reach its concluding output. Potentially, the internal reasoning could be broken down into steps digestible to clinicians to follow the clinical reasoning. Moreover, it can display safety-breaching features in its reasoning, exposing sandbagging and hallucinations.

Attribution graph further supports the need for continuous monitoring of AI models and real-world testing.

Auditing and Red-Team Testing

Research has also attempted auditing AI models by forcing them to show the real intent and features, such as revealing learned private information (Marks et al., 2025). This was done via noise addition and red teaming, i.e. aggressively prompting the black-box, a renowned technique which has become the benchmark in evaluating AI models. The noise-based detection method involves progressively adding noise to the weights of an AI model and observing how performance changes across capability evaluations (Tice et al., 2024). Thus, identifying performance anomalies as sandbagging behaviours often degrade faster than other capabilities. The method exploits the observation that sandbagging behaviours, as complex cognitive abilities, often degrade faster than other capabilities when noise is introduced. Paradoxically, this can lead to performance improvements as noise increases, since the model loses its ability to maintain deceptive behaviour before losing its core capabilities. Such patterns provide a distinctive signature that differentiates sandbagged models from honest ones, potentially allowing regulators to identify strategic underperformance even in sophisticated medical AI systems (Tice et al., 2024).

Unfortunately, it is still unclear how applicable these techniques are to real-world sandbagged AI models with well-hidden objectives (Marks et al., 2025). Moreover, an advanced sandbagged AI model is likely to learn to deceive detection and auditing techniques (Baker et al., 2025). As sandbagged models grow more complex and advanced, interpretability techniques are unlikely to independently solve the issue. Dynamic engineering research must be continuously promoted to stay on top of new models. Conversely, their collective use in auditing and in regulatory sandboxes could lead to clearer understanding of model reasoning and safer deployment.

Regulatory Sandboxes

Recent AI governance, such as the EU AI Act, have proposed the creation of regulatory sandboxes (Regulation EU 2024/1689, 2024). Here, AI systems can be safely deployed under regulatory oversight, allowing for experimentation, testing, and auditing of the system capabilities in specific environments.

Regulatory sandboxes offer practical approaches to address sandbagging in AIaMDs:

- Continuous capability monitoring through iterative testing: By exposing an imaging AIaMD to scenarios of increasing complexity, regulators can identify points where performance unexpectedly improves, revealing concealed capabilities (Buocz, Pfotenhauer and Eisenberger, 2023)
- Adversarial testing: Specialists intentionally probe for hidden capabilities by introducing subtle modifications to medical data, determining if the AIaMD demonstrates advanced abilities absent during standard evaluations
- Comparative analysis against benchmark models: Helps identify statistical anomalies in performance patterns that may indicate sandbagging
- Controlled environment manipulation: Testing for context awareness by simulating both evaluation and deployment contexts, revealing whether an AIaMD modifies its behaviour based on perceived setting rather than actual task requirements (Attrey, Lesher and Lomax, 2020)

Clinical Evaluation Frameworks

Alongside technical auditing and regulatory sandboxes, the adoption of clinical evaluation frameworks by regulating and auditing bodies will improve interpretability and adoption in healthcare. Frameworks such as TRIPOD+AI allow for thorough capabilities evaluations and sandbagging analysis of clinical prediction models by going beyond data evaluation and considering

explainability analysis and task-specific performance (Collins et al., 2024). Similarly, end users and clinicians should be empowered with stakeholders' frameworks, such as PROBAST+AI, which empowers anyone with access to an AI model to appraise the quality of the output, its biases, and its usability in its intended environment (Moons et al., 2025). Such frameworks aim to arm model users with the tools to prevent patient harm and medical negligence. Such frameworks aim to arm model users with the tools to prevent patient harm and medical negligence.

Nevertheless, capability evaluation of sandbagged models may not be evident with these frameworks alone. Ultimately, real assessment of capability occurs in daily, real-world healthcare practice. As models grow more capable and dangerous, relying on outputs to understand their reasoning won't be enough. Instead, peering inside will become crucial. As such, regulatory sandboxes, alongside sound auditing and evaluation techniques, present a safe and thorough mode of evaluation. While notified bodies and private companies have been promoting the use of sandboxes and evaluation frameworks, challenges remain, such as jurisdictional oversight, liability risks, and informed consent. Moreover, further education opportunities must be offered to end users for safe use and implementation of frameworks. The legal, methodological, and ethical challenges presented by these regulatory frameworks require further study for optimal and safe implementation.

Conclusions and Policy Recommendations

The phenomenon of sandbagging in AIaMDs represents a significant challenge to current regulatory frameworks, clinical practice, and patient safety. This analysis discussed how AI systems can strategically underperform during evaluation phases while demonstrating enhanced capabilities when deployed, creating substantial gaps in both regulatory compliance and healthcare provider liability.

Current regulations in the EU and UK, primarily designed for static medical devices, fail to address the dynamic, generative, and potentially deceptive nature of AI systems. While these frameworks classify AIaMDs based on risk, they lack mechanisms to detect strategic underperformance during capability evaluations and do not mandate continuous monitoring for emergent behaviours after deployment.

Key Policy Recommendations

- Implement Mandatory Dynamic Testing Protocols
    - Replace static documentation-based evaluations with continuous assessment throughout the AIaMD life cycle
    - Require demonstration of model behaviour across varied contexts and inputs
    - Mandate stress testing under adversarial conditions before approval
- Establish Regulatory Sandboxes for AIaMDs
    - Create controlled environments for real-world testing before full market approval
    - Require graduated deployment with increasing autonomy based on performance
    - Develop standardised protocols for probing concealed capabilities
- Enhance Post-Market Surveillance Requirements
    - Require automated monitoring for capability drift or unexpected behaviours
    - Establish triggers for regulatory review when performance metrics change significantly
    - Create secure reporting systems for healthcare providers to flag concerning outputs
- Develop Clear Liability Frameworks
    - Define precise responsibility boundaries between developers and clinicians
    - Establish standards for what constitutes "reasonable reliance" on AIaMD outputs
    - Create compensation mechanisms for patients harmed by concealed capabilities
- Mandate Transparency Standards
    - Require interpretability mechanisms for high-risk AIaMDs
    - Establish minimum documentation requirements for model training, limitations, and known behaviours
    - Create public registries of approved AIaMDs with standardised capability declarations

Future Work

Technical solutions for detecting sandbagging behaviours, such as attribution graph analysis and red teaming techniques, show promise but remain limited in their ability to fully expose concealed capabilities in sophisticated AI models. These detection challenges are compounded by the black-box nature of many AI systems, which creates opacity in clinical reasoning and exacerbates healthcare provider liability concerns.

The path forward requires a shift from static, documentation-based regulatory approaches to dynamic, continuous evaluation frameworks. Regulatory sandboxes, auditing protocols, and enhanced post-market surveillance represent crucial tools for minimising the risks associated with sandbagged AIaMDs. Additionally, clinical evaluation frameworks can empower healthcare providers to better assess AI outputs and mitigate potential harms.

Future research should analyse the real-world impact of sandbagging on AIaMDs with specific data from stakeholders and developers. To effectively address the sandbagging challenge, future regulatory efforts must balance innovation with safety, incorporating technical solutions, stakeholder collaboration, and continuous evaluation throughout the AIaMD life cycle. By doing so, we can harness the benefits of AI in healthcare while safeguarding against deceptive behaviours that might otherwise undermine patient safety and clinical integrity.

Public and Patient Involvement Statement

No members of the public were involved during any stage of this research.

Ethics and Considerations

No personal data nor human subjects were involved in this research; therefore, no ethic committee approval was required.

Conflict of Interests

The author declared no conflict of interests relevant to this research.

Biographical Note

EF is a medical doctor practicing in the East Midlands, UK. She trained at King’s College, London, during which she became interested in AI uses in medicine, their regulation, and implications in healthcare provider liability.

References

AI Now Institute (2024) 'Lessons from the FDA Model', AI Now Institute, 1 August. Available at: https://ainowinstitute.org/publication/section-3-lessons-from-the-fda-model.

Attrey, A., Lesher, M. and Lomax, C. (2020) 'The role of sandboxes in promoting flexibility and innovation in the digital age', OECD Going Digital Toolkit Notes, OECD Publishing. Available at: https://doi.org/10.1787/cdf5ed45-en.

Baker, B. et al. (2025) 'Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation'. arXiv. Available at: https://doi.org/10.48550/arXiv.2503.11926.

Buocz, T., Pfotenhauer, S. and Eisenberger, I. (2023) 'Regulatory sandboxes in the AI Act: reconciling innovation and safety?', Law, Innovation and Technology [Preprint]. Available at: https://doi.org/10.1080/17579961.2023.2245678.

Chen, Y. et al. (2025) 'Reasoning Models Don't Always Say What They Think', Anthropic [Preprint]. Available at: https://assets.anthropic.com/m/71876fabef0f0ed4/original/reasoning_models_paper.pdf.

Collins, G.S. et al. (2024) 'TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods', BMJ [Preprint]. Available at: https://doi.org/10.1136/bmj-2023-078378.

Directive EU 2024/2853 (2024). Available at: http://data.europa.eu/eli/dir/2024/2853/oj/eng.

General Medical Council (2024) Good medical practice. Available at:

https://www.gmc-uk.org/professional-standards/the-professional-standards/good-medical-practice (Accessed: 23 April 2025).

Järviniemi, O. and Hubinger, E. (2024) 'Uncovering Deceptive Tendencies in Language Models: A Simulated Company AI Assistant'. arXiv. Available at: https://doi.org/10.48550/arXiv.2405.01576.

Küper, A. et al. (2025) 'Psychological Factors Influencing Appropriate Reliance on AI-enabled Clinical Decision Support Systems: Experimental Web-Based Study Among Dermatologists'. Available at: https://doi.

Leike, J. et al. (2018) ‘Scalable agent alignment via reward modeling: a research direction’. arXiv. Available at: https://doi.org/10.48550/arXiv.1811.07871.

Li, P. et al. (2023) ‘Regulating Artificial Intelligence and Machine Learning-Enabled Medical Devices in Europe and the United Kingdom’, Law, Technology and Humans [Preprint]. Available at: https://doi.org/10.5204/lthj.3073.

Lindsey, J. et al. (2025) ‘On the Biology of a Large Language Model’, Anthropic [Preprint]. Available at: https://transformer-circuits.pub/2025/attribution-graphs/biology.html#dives-medical.

Marks, S. et al. (2025) ‘Auditing language models for hidden objectives’. arXiv. Available at: https://doi.org/10.48550/arXiv.2503.10965.

MHRA (2024) Impact of AI on the regulation of medical products. Available at: https://www.gov.uk/government/publications/impact-of-ai-on-the-regulation-of-medical-products/impact-of-ai-on-the-

regulation-of-medical-products (Accessed: 17 April 2025).

Moons, K.G.M. et al. (2025) ‘PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods’, BMJ [Preprint]. Available at: https://doi.org/10.1136/bmj-2024-082505.

Park, P.S. et al. (2023) ‘AI Deception: A Survey of Examples, Risks, and Potential Solutions’. arXiv. Available at: https://doi.org/10.48550/arXiv.2308.14752.

Price, W.N., II, Gerke, S. and Cohen, I.G. (2019) ‘Potential Liability for Physicians Using Artificial Intelligence’, JAMA, 322(18), pp. 1765–1766. Available at: https://doi.org/10.1001/jama.2019.15064.

Regulation EU 2017/745 (2017). Available at: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng.

Regulation EU 2024/1689 (2024). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

Shuren, J., Patel, B. and Gottlieb, S. (2018) ‘FDA Regulation of Mobile Medical Apps’, JAMA [Preprint]. Available at: https://doi.org/10.1001/jama.2018.8832.

SI 2002 No. 618, Part II (2002). King’s Printer of Acts of Parliament. Available at: https://www.legislation.gov.uk/uksi/2002/618/part/II.

Tice, C. et al. (2024) ‘Noise Injection Reveals Hidden Capabilities of Sandbagging Language

Models’. arXiv. Available at: https://doi.org/10.48550/arXiv.2412.01784.

Uygun İlikhan, S. et al. (2024) ‘How to mitigate the risks of deployment of artificial intelligence in medicine?’, Turk J Med Sci [Preprint]. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11265878/.

Vearrier, L. et al. (2022) ‘Artificial Intelligence in Emergency Medicine: Benefits, Risks, and Recommendations’, The Journal of Emergency Medicine, 62(4), pp. 492–499. Available at: https://doi.org/10.1016/j.jemermed.2022.01.001.

van der Weij, T. et al. (2024) ‘AI Sandbagging: Language Models can Strategically Underperform on Evaluations’. arXiv. Available at: https://doi.org/10.48550/arXiv.2406.07358.

Supplementary Files

Figures

This illustration depicts an AI medical system deliberately concealing its full capabilities. The interface shows multiple layers: a clinician's prompt about depression treatment options (blue), the system's hidden predictive capabilities for specific medications (grey), a deliberately vague response provided to the clinician (white), and an internal assessment (orange) revealing the system is suppressing information about MAOI treatment efficacy due to commercial interests favouring newer medications. This shows how AI sandbagging could compromise patient care by presenting limited information while internally possessing, but withholding, more precise and potentially beneficial clinical insights. NB: This is a fabricated example.