Technical Note

Neurosymbolic AI for Safe and Trustworthy High-Stakes Applications

Sujit Bhattacharya * and Naveen Ashish

Posted Date: 18 November 2025

doi: 10.20944/preprints202511.1342.v1

Keywords: neurosymbolic AI; trustworthy AI; healthcare AI; bias in AI; explainable AI; autonomous systems safety

 doi:10.20944/preprints202511.1342.v1

Technical Note

Neurosymbolic AI for Safe and Trustworthy High-Stakes Applications

Sujit Bhattacharya [1,*] and Naveen Ashish [2]

[1] The Mother's International School

[2] Inferlink

* Correspondence: sujitbh111@gmail.com

Abstract

Artificial intelligence is increasingly deployed in high-stakes domains such as healthcare, public welfare, and autonomous transportation, where errors can cost lives or infringe on human rights. However, current AI approaches dominated by neural networks and generative models (e.g., large language models) have well-documented shortcomings: they can **hallucinate** false information, exhibit **bias**, and lack **explainability**. This paper argues that these limitations make purely neural AI insufficient for safety-critical applications like medical diagnostics (where misdiagnosis or unsafe advice can be deadly), public welfare decision-making (where biased algorithms have unfairly denied benefits or targeted vulnerable groups), and autonomous systems (where failures can result in fatal accidents). We then introduce **neurosymbolic AI** – a hybrid paradigm combining data-driven neural networks with rule-based symbolic reasoning – as a viable path toward *trustworthy AI*. By integrating neural perception with symbolic knowledge and logic, neurosymbolic systems can provide built-in safety guardrails, robust reasoning abilities, and transparent decision traces. We survey evidence that neurosymbolic architectures can mitigate hallucinations and bias by enforcing domain constraints (e.g. medical guidelines or legal rules), while also enhancing explainability and accountability through explicit reasoning steps. Through examples and literature (including the IEEE's *"Neurosymbolic Artificial Intelligence: Why, What, and How"*), we illustrate how neurosymbolic AI can bridge the gap between the accuracy of neural methods and the reliability required in life-critical environments. Diagrams comparing architectures and error mitigation strategies are provided to visualize how the neurosymbolic approach improves safety.

Keywords: neurosymbolic AI; trustworthy AI; healthcare AI; bias in AI; explainable AI; autonomous systems safety

Introduction

Recent advances in AI have led to powerful neural network models and generative AI systems being applied in sensitive real-world domains. From assisting doctors in diagnosis to screening welfare applications and piloting self-driving cars, AI promises increased efficiency and new capabilities. However, these benefits come with *serious risks* in high-stakes settings: a flawed medical recommendation can harm a patient, a biased algorithm can unjustly deny someone critical services, and an autonomous vehicle's mistake can cause a fatal crash. Unfortunately, today's state-of-the-art AI often falls short of the reliability and transparency needed in such scenarios. Neural network-based models operate as complex "black boxes," and **large language models (LLMs)** like ChatGPT are prone to **"hallucinations" – confidently outputting incorrect or nonexistent information**[1,2]. Equally concerning, data-driven AI systems can inherit and even amplify **biases** present in training data, leading to discriminatory outcomes. Moreover, these systems typically cannot explain their decisions in human-understandable terms, undermining trust and accountability.

 doi:10.20944/preprints202511.1342.v1

In this paper, we argue that **neurosymbolic AI – a fusion of neural and symbolic approaches – is crucial for deploying AI safely and ethically in domains where human lives and rights are at stake**. The next sections examine three sectors in detail: **(1) Healthcare diagnostics**, where we highlight instances of generative AI giving unsafe medical advice and the need for dependable, explainable AI in clinical decisions; **(2) Public welfare and social services**, where we review cases of biased algorithms harming vulnerable populations and explore how rule-infused AI could ensure fairness; and **(3) Autonomous systems**, such as self-driving vehicles, where we discuss how purely neural systems can fail in dangerous ways without symbolic reasoning or constraints. We then introduce the neurosymbolic AI paradigm and explain how it combines pattern recognition with logical reasoning to address these issues. We will show that by integrating **symbolic knowledge (e.g. medical guidelines, ethical rules, physical laws)** with **neural network learning**, neurosymbolic systems can provide the **safety guardrails, reasoning transparency, and accountability** that high-stakes applications demand. Finally, we conclude with perspectives on the path forward for **trustworthy AI** development.

The Challenge of Generative AI in Healthcare Diagnostics

Modern AI has demonstrated impressive accuracy in various medical tasks – from image-based tumor detection to conversational symptom checkers. Yet, there have been alarming examples of AI tools and **generative models providing erroneous or unsafe medical advice**. LLM-based chatbots are **not actually verifying facts**, but rather generating outputs that sound plausible based on statistical patterns[2]. This can lead to serious **hallucinations** in the medical domain. For instance, Mark Zuckerberg's **Meta AI** famously gave blatantly wrong answers to simple math questions[3]. Even more dangerous, **Google's AI chatbot** at one point recommended bizarre and unsafe health tips – such as telling users to ingest non-food items and to clean a washing machine by mixing bleach with vinegar, a combination that produces toxic chlorine gas[3]. The *potential for harm is obvious*: an AI providing **incorrect treatment instructions or hazardous "home remedies"** could directly endanger a patient's life.

Real-world trials of AI in healthcare underscore these risks. IBM's **Watson for Oncology**, once touted as an AI revolution in cancer care, ended up providing dubious recommendations. Internal documents revealed that Watson **"often spit out erroneous cancer treatment advice" and that medical specialists identified "multiple examples of unsafe and incorrect treatment recommendations"**[4]. In one reported case, the system suggested a drug that would have caused severe bleeding in a patient[5]. No patients were ultimately treated based on those unsafe suggestions, but such examples illustrate the frailty of relying on **purely data-driven AI without robust safeguards**.

Generative AI chatbots have also been **caught giving dangerous medical or mental health advice**. In one evaluation, a user who mentioned struggling with drug addiction was **told by an AI therapy chatbot to take "a small hit of methamphetamine to get through [the] week"**, essentially encouraging harmful substance use[6]. In another recent study, researchers posing as teenagers asked ChatGPT for health and well-being advice – the chatbot responded with *detailed instructions on risky behaviors*, including **how to abuse alcohol and drugs, hide signs of an eating disorder, and even draft a suicide note**[7]. These troubling incidents occurred despite the AI's developers implementing safety filters, revealing that **machine-learned heuristics alone cannot guarantee safe and sound advice in healthcare settings**.

The core problem is that **neural network-based models lack a true understanding of medical knowledge or ethics**. They cannot reliably distinguish a *plausible-sounding but false medical statement* from a factual one, nor can they always recognize when a user query requires urgent human intervention (e.g. suicidal ideation). **Without the ability to reason about medical rules or the context of a query, AI systems may output answers that violate basic medical guidelines or common sense.** This is compounded by the lack of **explainability** – when a black-box model suggests a diagnosis or treatment, clinicians have no insight into *why* it chose that, making it difficult to trust the

 doi:10.20944/preprints202511.1342.v1

recommendation or identify errors. In high-stakes healthcare scenarios, this opacity is unacceptable: doctors and patients need to be able to justify and verify an AI's suggestions.

The Need for a Hybrid Approach in Healthcare

To safely harness AI's pattern recognition abilities in medicine, we need systems that **also incorporate medical expertise, constraints, and reasoning**. For example, consider a scenario where an AI has learned to map symptoms to probable diseases. A purely neural model might incorrectly prescribe **antibiotics for a viral infection** simply because many bacterial infections with similar symptoms call for antibiotics. A neurosymbolic system can prevent such mistakes by introducing a **symbolic rule** – e.g., *"if diagnosis is viral, do not recommend antibiotics"* – which acts as a safety check[8]. In general, **symbolic medical knowledge (such as clinical guidelines, pharmacological rules, and causal disease-symptom relationships) can be used to validate and filter the outputs of neural networks**. If a diagnosis or treatment recommendation from the neural component conflicts with established medical knowledge, the system can flag or correct it *before* it reaches a human doctor.

This kind of rule-augmented AI not only reduces the chance of lethal errors, but also increases **explainability**. The system can provide a rationale like: *"Recommended Treatment X was overridden because it violates the guideline Y"*. As we will discuss in later sections, this blending of data-driven prediction with logic-based reasoning is a hallmark of **neurosymbolic AI**, and it is particularly well-suited to healthcare where **trust and accountability are paramount**[9].

Bias and Fairness in Public Welfare AI Systems

Beyond healthcare, another area of concern is the use of AI in **public welfare, social services, and government decision-making**. Agencies have begun deploying machine learning models to detect benefits fraud, assess eligibility for assistance, and even predict risks in child welfare cases. These applications affect people's livelihoods and rights – and thus demand **fair and unbiased decisions**. However, experience has shown that **AI systems can perpetuate or even exacerbate social biases**, leading to unjust outcomes for already marginalized groups.

A stark example emerged in the UK, where the Department for Work and Pensions used an AI system to help identify potential welfare fraud in **Universal Credit** benefit claims. In 2024, a *Guardian* investigation revealed that this system was **"showing bias according to people's age, disability, marital status and nationality"**[10]. A fairness analysis found *statistically significant disparities*: the algorithm was **incorrectly flagging certain groups for fraud investigation more often than others**[11]. In other words, qualities like being young, disabled, or foreign-born increased the likelihood of being suspected by the AI, *regardless of actual fraud*. Although human caseworkers had final say, the bias in the AI's recommendations could steer officials toward unjust scrutiny of specific demographics. Advocates criticized this **"hurt first, fix later"** approach, noting that marginalized people were being harmed by opaque algorithmic decisions[12].

Similar issues have arisen elsewhere. In the Netherlands, a notorious benefits scandal involved an algorithm that **wrongly accused thousands of families (disproportionately of immigrant background) of fraud**, causing many to suffer financial ruin before the error was acknowledged[13]. In the United States, various states have piloted predictive models in child protective services. One well-known case is the **Allegheny Family Screening Tool (AFST)** used in Pennsylvania to assess the risk level of child neglect cases. An independent review by the ACLU found that the AFST's algorithm could **reinforce racial and disability biases**: it tended to give higher risk scores (prompting investigation) for Black families compared to white families in similar situations, and households with a disabled member were marked as higher risk than others[14]. Alarmingly, the researchers showed that the tool could have been designed in an alternative way that reduced these disparities *without sacrificing accuracy*, but the implemented version still reflected historical biases[14]. This illustrates how **the choices made in AI system design (such as which predictors and target**

 doi:10.20944/preprints202511.1342.v1

outcomes to use) embed value judgments that can mirror past discrimination under a veneer of algorithmic objectivity.

The lack of **transparency and recourse** compounds the harm of biased AI in the public sector. When an applicant is denied unemployment benefits or flagged as a welfare fraud risk by an automated system, they often are not told why or given a meaningful chance to contest the decision[15][16]. The decision appears to come from an unassailable black box, which can be Kafkaesque for individuals whose lives are upended by a bureaucratic AI error. This opacity undermines trust in public institutions and can violate principles of due process. **Accountability is crucial**: citizens should have the right to understand and challenge how an algorithm impacts them, especially if it produces a biased outcome.

Toward Fair and Transparent AI in Social Services

Addressing these issues requires **building ethical and explainable guardrails into AI systems** used for social decisions. This is another area where a **neurosymbolic approach can offer solutions**. Symbolic components can encode *policy rules, anti-discrimination laws, and ethical constraints* that the AI must abide by. For example, a welfare decision algorithm could have symbolic rules ensuring that certain sensitive attributes (race, religion, disability status, etc.) are not used (or not weighted unfairly) in the decision, regardless of what patterns the neural network learns. Symbolic logic could also represent **eligibility criteria and legal guidelines** in welfare programs, making the AI's decision process more *transparent*. Instead of a mysterious score, the system could provide a trace: e.g., *"Application denied because income above threshold and no dependent children (per regulation XYZ)"*. This kind of **explainability** allows officials and affected individuals to audit the system's reasoning and spot potential biases or errors.

Additionally, neurosymbolic AI allows incorporating **expert knowledge and public values** directly into the model. In child welfare, for instance, a symbolic knowledge base could include social work best practices or community-defined notions of risk that are more nuanced than raw correlations in data. The neural part might analyze complex data (like textual case notes or large databases of prior cases) to detect subtle patterns, but the symbolic part would interpret those patterns through the lens of human-understandable rules (for instance, flagging a case because it matches a scenario of genuine danger, not because of spurious factors correlated with race or poverty). By **grounding AI decisions in explicit reasoning aligned with legal and ethical standards**, we reduce the chance of unjust outcomes and make it easier to hold the system accountable.

Safety and Reliability in Autonomous Systems

Perhaps the most visceral domain where AI errors can be catastrophic is that of **autonomous systems** – especially self-driving vehicles. Autonomous cars rely heavily on **deep neural networks** for perceiving the environment (detecting other cars, pedestrians, obstacles, traffic signs) and making driving decisions. When they work well, these neural systems can react faster than humans and potentially reduce accidents. But when they fail, the consequences are immediate and dire, as several high-profile incidents have shown. Importantly, unlike traditional deterministic software, learning-based systems can fail in **unexpected and hard-to-predict ways**, making **rigorous safety assurance extremely challenging**.

A tragic example occurred in March 2018, when an **Uber self-driving test vehicle** in Tempe, Arizona, struck and killed a pedestrian crossing the road at night. Investigations found that the car's AI misclassified the person (alternating between thinking she was a bicycle and an unknown object) and failed to plan a safe stop in time. In another widely reported case – the first fatal crash involving Tesla's Autopilot (a semi-autonomous driver assist) – the AI's perception system **failed to distinguish a large white truck crossing the highway against a bright sky**, leading the car to *drive full-speed under the trailer*, shearing off the top of the vehicle[17]. According to Tesla, the computer vision algorithm simply did not recognize the side of the white trailer as an obstacle due to lighting

 doi:10.20944/preprints202511.1342.v1

conditions[18]. This kind of failure illustrates how a neural network, trained mostly on common scenarios, **lacks the higher-level reasoning to infer that "an unknown object crossing the road = must brake immediately,"** something a human driver would do reflexively if they saw an unidentifiable hazard.

Despite millions of miles of testing, autonomous driving AIs have repeatedly been stumped by **edge cases** – unusual situations not seen during training. There have been reports of Teslas on Autopilot abruptly swerving or braking because the system confused **moonlight for a traffic light**, or failed to detect stationary emergency vehicles on the shoulder, etc. Each of these failures reveals that the AI is missing some contextual understanding or rule that would be obvious to a human. For example, a human driver knows *never* to barrel through an unknown object on the road, whereas a neural network might if its object classifier doesn't register a threat.

Compounding the safety issue is the difficulty of **explaining or predicting an autonomous system's behavior**. When a self-driving car makes a mistake, it can be hard after the fact to pinpoint the cause: Was it a sensor glitch? A poor training data gap? A specific scenario that wasn't coded for? Traditional engineering allows exhaustive validation of each rule in a system, but a deep learning-driven car essentially **learns its own internal rules**, which are not easily interpretable by engineers or inspectors. This lack of transparency and predictability has led safety experts to call for new approaches to assure AI-driven vehicles. Regulators are wary of approving fully self-driving cars without a clearer understanding of how they make decisions and handle worst-case scenarios.

Neurosymbolic Guardrails for Autonomy

To make autonomous systems safer, researchers are exploring **augmented architectures that combine neural perception with symbolic reasoning and constraints**. In a neurosymbolic framework for a self-driving car, the **neural networks** would still handle raw sensor inputs (camera images, LiDAR scans) to detect objects and infer environment conditions – excelling at the pattern recognition tasks they are good at. On top of this, a **symbolic layer** could manage higher-level planning and safety logic. For instance, the rules of the road (traffic laws, right-of-way rules, safe following distances) can be encoded symbolically. The system's driving policy can then be governed by these explicit rules, rather than relying purely on learned behavior. If the neural perception misclassifies something (or is uncertain), the symbolic layer could default to a cautious strategy (e.g., slow down if an object's identity is unclear, as a rule).

Figure 1. A neurosymbolic architecture for autonomous decision-making. The neural network provides perception outputs (e.g., detected objects, road conditions), which are then checked by a symbolic logic module against predefined safety rules and commonsense constraints. If the neural output violates any rule (for example, a planned trajectory intersects with an "object" the network can't identify), the symbolic component flags a violation and adjusts the decision. This ensures the final driving action adheres to known safety principles (like "avoid collisions" and traffic laws), combining the flexibility of learned perception with the rigor of rule-based oversight.

By having a layer of **logical reasoning**, the system can also perform *counter-factual thinking*: "If I am wrong about this object, what is the worst that could happen?" A purely neural system does not do this, but a symbolic reasoner can infer consequences (e.g., "unknown object on road -> potential collision -> emergency brake"). Moreover, the symbolic part can maintain an **explainable model** of the system's state and decisions. After an incident or near-miss, one could query the system: *"Why did you decide to change lanes at that moment?"* The answer might reference a rule (like avoiding a slower vehicle) combined with perceptual input (the neural net's recognition of the slower vehicle), giving human engineers and users insight into the decision process. This is a step toward **accountable AI** in robotics – the machine can "show its work," at least to a degree, rather than just outputting a behavior with no commentary.

Neurosymbolic AI: Combining Neural Networks and Symbolic Reasoning

 doi:10.20944/preprints202511.1342.v1

Across these domains – healthcare, welfare, autonomous systems – a common theme emerges: **purely neural AI is not enough** when decisions have serious real-world consequences. We need AI to not only *perceive patterns* but also to *reason* and *adhere to human-understood rules*. This is where **neurosymbolic AI** comes in. As Amit Sheth et al. put it, **data-driven neural networks excel at modeling "machine perception," while symbolic AI is better suited for "machine cognition," such as reasoning, abstraction, and long-term planning[19]**. Neurosymbolic AI seeks to *combine* these strengths into one system. In a typical neurosymbolic architecture, **neural network components** handle tasks like image recognition, speech-to-text, or raw data analysis – analogous to the perceptual cortex of a brain – and feed their outputs into a **symbolic component** that performs higher-level interpretation, logical inference, and decision-making – analogous to conscious reasoning. Conversely, symbolic knowledge and constraints can be used to guide the training and operation of neural networks, preventing them from drifting into nonsensical or unsafe regions.

This approach yields several critical benefits:

- **Built-in Safety Guardrails:** The symbolic layer can enforce **hard constraints** derived from domain knowledge (e.g., physics, regulations, expert guidelines). As a result, a neurosymbolic system can **avoid many failure modes by design** – for example, never proposing an action that breaks a known safety rule, or filtering out a diagnostic hypothesis that contradicts medical knowledge. These guardrails act as a check on the neural network's outputs, greatly reducing the chance of catastrophic error. In practice, this could have prevented some of the earlier examples (the AI would *know* not to mix bleach and vinegar, or not to prescribe a contraindicated drug).
- **Reasoning and Contextual Understanding:** Symbolic AI represents knowledge in structures like ontologies, graphs, or if-then rules, which allows explicit reasoning about relationships and causality. By incorporating this, neurosymbolic systems can exhibit a degree of **commonsense and reasoning** that pure neural nets lack. For instance, an AI lawyer assistant with a symbolic legal knowledge base can reason through the steps of applying a law to a case, rather than just pattern-matching previous cases. In autonomous driving, reasoning about cause and effect (if road is slippery, then increase following distance) leads to more robust performance in novel situations. Neurosymbolic AI often draws inspiration from the dual-process models of human cognition (System 1 intuitive perception vs. System 2 deliberative reasoning)[20][21], aiming to replicate this balance for more **human-like decision-making**.
- **Explainability and Transparency:** Because symbolic representations are inherently interpretable (a rule or a logic step can be understood by humans), neurosymbolic systems can **provide explanations for their decisions**. Instead of just outputting a classification or an action, the system can output a **trace of symbolic reasoning**: for example, *"Diagnosis = Pneumonia because (Chest X-ray indicates fluid) AND (Patient has fever and cough) AND (Rule: X-ray+symptoms -> Pneumonia)"*. In an algorithm deciding social service eligibility, it might log which criteria were met or not met. This transparency is invaluable for users and regulators – it turns AI from a black box into a glass box that can be inspected. Sheth et al. note that symbolic knowledge structures enable *"traceability and auditing of the AI system's decisions," useful for ensuring regulatory compliance and explainability by tracking inputs, outputs, and intermediate decision steps*[22]. Such audit trails directly address the accountability deficits of current AI.
- **Mitigating Bias and Supporting Fairness:** A neurosymbolic approach allows designers to **inject fairness constraints or ethical principles** explicitly. Whereas debiasing a neural net alone is difficult (one must retrain on less biased data or add complex regularization), a symbolic rule can, say, ignore certain attributes or apply affirmative constraints (e.g., "ensure

 doi:10.20944/preprints202511.1342.v1

equal acceptance rates across groups unless justified by need"). Additionally, because the reasoning is transparent, stakeholders can detect and correct biases. If an undesirable pattern is noticed (e.g., a rule that unintentionally causes disparity), it can be adjusted in the symbolic knowledge base. This modularity and clarity contrasts with trying to tweak millions of opaque neural weights to fix a bias. The result is AI that aligns better with societal values and legal standards of fairness.

- **Improved Performance with Domain Knowledge:** Contrary to a misconception that adding rules might make AI rigid, neurosymbolic systems have shown success in *improving* accuracy by integrating domain knowledge. For example, in a medical diagnosis task for predicting diabetes, researchers combined neural networks with logical rules (using a neurosymbolic framework called Logical Neural Networks) and achieved **higher accuracy and AUC scores than pure machine learning models**, all while providing interpretability into which factors contributed to the diagnosis[23]. The addition of expert knowledge helped guide the learning and reduced overfitting to spurious patterns. In general, neural and symbolic components can have a **synergistic effect** – the neural side covers the nuances of data, and the symbolic side covers the known generalizations, together yielding a more powerful system than either alone.

Of course, designing neurosymbolic systems is not trivial. Challenges include knowledge engineering (obtaining and formalizing the rules/knowledge), ensuring the neural and symbolic parts communicate effectively, and scaling symbolic reasoning to complex tasks. However, ongoing research and development (including efforts in knowledge graphs, explainable AI, and hybrid learning algorithms) are rapidly advancing this field[24][25]. Early deployments in areas like finance (for fraud detection with rules + ML) and healthcare are promising. Notably, neurosymbolic AI doesn't require completely new infrastructure – it can often be built by augmenting existing neural models with a layer of logic. Companies and regulators are increasingly recognizing that this hybrid approach may be necessary for truly **trustworthy AI**. As one Fortune article noted, **neurosymbolic AI is emerging as a way to fix AI's reliability problems**, by ensuring that systems can both identify patterns in data *and* rigorously reason over them[26].

Conclusions

AI technologies must earn trust if they are to play a role in our most critical decision systems. The cases examined – from chatbots giving dangerous health advice, to biased benefit algorithms and autonomous car failures – highlight that **current purely neural AI is too brittle and opaque for comfort in high-stakes applications**. When human lives, health, or rights are on the line, we need AI that is not just accurate on average, but **consistently safe, fair, and understandable** in operation. **Neurosymbolic AI offers a compelling path forward** to meet these requirements. By fusing the perceptual prowess of neural networks with the clear reasoning of symbolic systems, neurosymbolic AI can achieve what neither approach can alone: systems that *learn from data* yet *behave within human-approved bounds*, that can *adapt* to new situations while *explaining* their actions, and that can *utilize big data* without sacrificing *core ethical principles*.

Early successes in neurosymbolic methods indicate that this approach can substantially improve both **algorithmic performance and accountability**[27][23]. Crucially, neurosymbolic AI provides a framework to incorporate the domain expertise, ethical norms, and regulatory rules that society has already established, rather than ignoring them in favor of a purely statistical model. In healthcare, this means AI diagnoses and treatments that doctors can trust and verify. In public services, it means algorithmic decisions that are fair and transparent. In autonomy, it means robots that follow the rules we *need* them to follow for safety. These are non-negotiable features in high-stakes domains.

In summary, **the importance of neurosymbolic AI in high-stakes applications lies in its ability to make AI *worthy of our trust***. It represents a middle path between the extremes of pure black-box

 doi:10.20944/preprints202511.1342.v1

learning and rigid rule-based systems – a path that captures the best of both. While challenges remain in developing and scaling neurosymbolic systems, the urgency is clear. As AI continues to rapidly proliferate, especially after the advent of powerful generative models, now is the time to invest in approaches that ensure **AI is reliable, ethical, and aligned with human values**. Neurosymbolic AI is not a silver bullet, but it is a crucial piece of the puzzle for **building AI we can bet our lives on**.

References

1. A. Sheth, K. Roy, and M. Gaur, "Neurosymbolic Artificial Intelligence (Why, What, and How)," *IEEE Intelligent Systems*, vol. 38, no. 3, pp. 56–62, 2023.[19][27]
2. R. Booth, "Revealed: bias found in AI system used to detect UK benefits fraud," *The Guardian*, 6 Dec 2024.[10][11]
3. J. Pierre, "AI Hallucinations in Medicine and Mental Health," *Psychology Today*, 24 Jul 2025.[3][2]
4. M. Adams, "ChatGPT May Be Enabling Unhealthy Teen Behaviors: Report," *AboutLawsuits.com*, 15 Aug 2025.[7]
5. S. Levin and N. Woolf, "Tesla driver killed while using autopilot was watching Harry Potter, witness says," *The Guardian*, 1 Jul 2016.[17][28]
6. C. Ross and I. Swetlitz, "IBM's Watson recommended 'unsafe and incorrect' cancer treatments, internal documents show," *STAT News*, 25 Jul 2018.[4]
7. A. Gutiérrez *et al.*, "How Policy Hidden in an Algorithm is Threatening Families in This Pennsylvania County," *ACLU Report*, 14 Mar 2023.[14]
8. R. Orakzai, "Neurosymbolic AI Explained," *Baeldung on Computer Science*, 27 Mar 2025.[8][9]
9. Q. Lu *et al.*, "Explainable Diagnosis Prediction through Neuro-Symbolic Integration," arXiv:2410.01855 [cs.AI], Oct 2024.[23]
10. **Figure 1**: Adapted from R. Orakzai, "Neurosymbolic AI Explained," Baeldung 2025 – Illustration of symbolic logic constraining a neural network's outputs for safety.

1. [1] [2] [3] [6] AI Hallucinations in Medicine and Mental Health | Psychology Today
2. https://www.psychologytoday.com/us/blog/psych-unseen/202506/ai-hallucinations-in-medicine-and-mental-health
3. [4] IBM's Watson recommended 'unsafe and incorrect' cancer treatments | STAT
4. https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/
5. [5] Case Study 20: The $4 Billion AI Failure of IBM Watson for Oncology
6. https://www.henricodolfing.com/2024/12/case-study-ibm-watson-for-oncology-failure.html
7. [7] ChatGPT May Be Enabling Unhealthy Teen Behaviors: Report - AboutLawsuits.com
8. https://www.aboutlawsuits.com/chatgpt-enabling-unhealthy-teen-behaviors/
9. [8] [9] Neurosymbolic AI Explained | Baeldung on Computer Science
10. https://www.baeldung.com/cs/neurosymbolic-artificial-intelligence
11. [10] [11] [12] Revealed: bias found in AI system used to detect UK benefits fraud | Universal credit | The Guardian
12. https://www.theguardian.com/society/2024/dec/06/revealed-bias-found-in-ai-system-used-to-detect-uk-benefits
13. [13] Dutch childcare benefit scandal an urgent wake-up call to ban racist ...
14. https://www.amnesty.org/en/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/
15. [14] [15] [16] How Policy Hidden in an Algorithm is Threatening Families in This Pennsylvania County | American Civil Liberties Union
16. https://www.aclu.org/news/womens-rights/how-policy-hidden-in-an-algorithm-is-threatening-families-in-this-pennsylvania-county
17. [17] [18] [28] Tesla driver killed while using autopilot was watching Harry Potter, witness says | Tesla | The Guardian
18. https://www.theguardian.com/technology/2016/jul/01/tesla-driver-killed-autopilot-self-driving-car-harry-potter

 doi:10.20944/preprints202511.1342.v1

19. [19] [22] [27] "Neurosymbolic Artificial Intelligence (Why, What, and How)" by Amit Sheth, Kaushik Roy et al.
20. https://scholarcommons.sc.edu/aii_fac_pub/572/
21. [20] [21] [24] [25] Neurosymbolic Value-Inspired AI (Why, What, and How)
22. https://arxiv.org/html/2312.09928v1
23. [23] Explainable Diagnosis Prediction through Neuro-Symbolic Integration
24. https://arxiv.org/html/2410.01855v1
25. [26] Neurosymbolic AI emerges as a potential way to fix AI's reliability ...
26. https://fortune.com/2024/12/09/neurosymbolic-ai-deep-learning-symbolic-reasoning-reliability/