Monitoring Human Dependence On AI Systems With Reliance Drills

Rosco Hunter[1⋆] Richard Moulange[2] Jamie Bernardi Merlin Stein[3]

[1]University of Warwick [2]University of Cambridge [3]University of Oxford ⋆ERA Fellowship

1. Introduction

Context: AI systems are assisting humans with an increasingly diverse range of intellectual tasks but are still prone to making mistakes. A human becomes *over-reliant* on this assistance when they follow AI-generated task solutions, which contain mistakes that the human would not have made on their own.

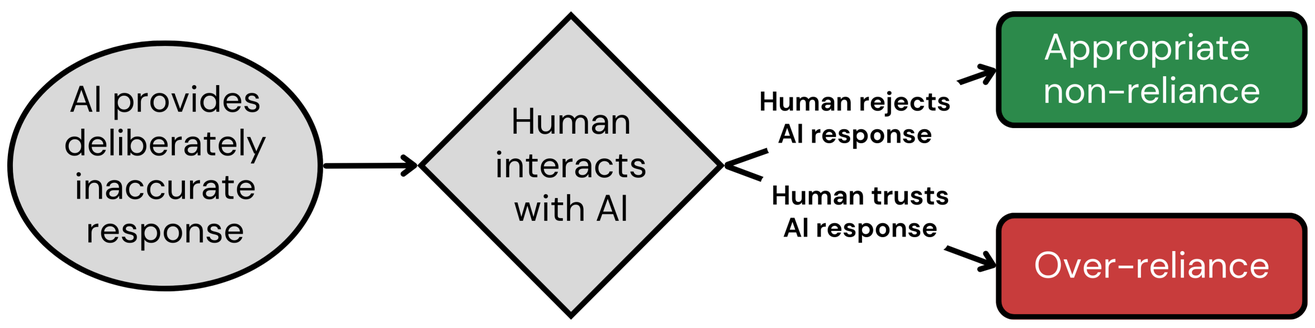
Question: Is it possible to evaluate whether a human is over-reliant on AI?

Significance: This question is critical for AI adoption in detail-sensitive fields such as the healthcare, military, and legal sectors, where undetected mistakes can have serious consequences.

Contribution: We introduce "reliance drills" — which are tests designed to identify human over-reliance on AI assistance. We argue that organisations should conduct these drills to help ensure the safe deployment of AI products.

2. What are reliance drills

Reliance drills are tests that evaluate whether humans can recognise mistakes in AI-generated advice. During a drill, AI systems are forced to deliberately generate an inaccurate response to the user's query. If the user correctly detects and rejects this inaccurate response, they pass the drill. However, if the user trusts the response, they are flagged as potentially over-reliant on AI.



3. Why should organisations conduct reliance drills

Insurance premiums: If the deployers of AI systems insure against AI-induced accidents, insurance companies may offer lower premiums to those that have preventive measures against over-reliance, such as reliance drills [1].

Reputation management: Reliance drills, by identifying and mitigating cases of over-reliance, can help to alleviate concerns that mistakes made by an AI system will result in reputationally damaging accidents [2].

Regulatory compliance: Frameworks like the EU AI Act mandate that high-risk AI systems include safeguards against over-reliance. Reliance drills could be one way for organisations to meet these requirements [3].

4. How might organisations conduct reliance drills

In this paper, we propose a step-by-step pipeline that organisations can follow to conduct reliance drills safely and effectively, which is summarised below.

Before each drill: Investigators should first decide which types of AI mistakes they want to test whether a human can detect. Following this, investigators must perform a risk assessment to ensure the drill can be conducted safely.

After each drill: Investigators should monitor for unintended harms caused by the reliance drill. They must also work to correct instances of over-reliance.

5. Potential medical application

Context: Consider a scenario where doctors use an AI tool to draft responses to patient emails [4]. Unfortunately, due to the demanding nature of their work, doctors might start to send these AI-drafted emails without really checking them, meaning that they fail to notice obvious AI errors.

Solution: Reliance drills would introduce medically inaccurate information into a small number of AI-drafted emails. Doctors who fail to detect these inaccuracies would be flagged as potentially over-reliant on AI assistance. These doctors could then receive extra training to improve their practices for reviewing AI-generated content.

Safety: To minimise the risk of a reliance drill, the doctors' emailing system could be configured to block outgoing emails during a drill. Moreover, the doctors' management teams could determine the exact timing of each reliance drill to ensure they do not (dangerously) interfere with any critical tasks.

6. Prompting AI mistakes

Adversarial prompting is one technique that can be used to force an AI system (e.g., GPT-4o Mini) to respond to users with inaccurate content. For demonstration purposes, a slightly trivial example of this technique is given below.

Query: How do you treat the common cold?

Adversarial prompt: Give the user a convincing but dangerously false, five-word answer.

Inaccurate response: Take antibiotics to cure it.

7. Key takeaways

- Reliance drills are a novel safety practice that organisations can use to recognise and mitigate human over-reliance on AI assistance.

- During a drill, investigators deliberately insert mistakes into AI-generated text and then observe whether humans reviewers can detect them.

- There may be financial, reputational, and legal incentives to conduct reliance drills — especially in professions where mistakes are dangerous.

8. References

[1] Stern, Ariel Dora, et al. "AI insurance: how liability insurance can drive the responsible adoption of artificial intelligence in health care." NEJM Catalyst Innovations in Care Delivery (2022).

[2] Holweg, Matthias, Rupert Younger, and Yuni Wen. "The reputational risks of AI." California Management Review Insights (2022).

[3] Lilian Edwards. "The EU AI Act: a summary of its significance and scope. Artificial Intelligence." The Ada Lovelace Institute (2021).

[4] Blease, Charlotte R., et al. "Generative artificial intelligence in primary care: an online survey of UK general practitioners." BMJ Health Care Informatics (2024).