Monitoring Human Dependence On AI Systems With Reliance Drills

Rosco Hunter[*†] University of Warwick
Richard Moulange University of Cambridge
Jamie Bernardi[‡]
Merlin Stein[‡] University of Oxford

Abstract

AI systems are assisting humans with an increasingly broad range of intellectual tasks. Humans could be over-reliant on this assistance if they trust AI-generated advice, even though they would make a better decision on their own. To identify real-world instances of over-reliance, this paper proposes the *reliance drill*: an exercise that tests whether a human can recognise mistakes in AI-generated advice. We introduce a pipeline that organisations could use to implement these drills. As an example, we explain how this approach could be used to limit over-reliance on AI in a medical setting. We conclude by arguing that reliance drills could become a key tool for ensuring humans remain appropriately involved in AI-assisted decisions.

1 Introduction

Artificial intelligence (AI) is beginning to match and, in some cases, outperform human capabilities across a broad range of intellectual tasks [1]. As a result, competitive pressures could lead organisations to integrate AI assistance into a large number of critical decisions, potentially driving innovation and economic growth [8, 21]. However, given the overconfidence exhibited by some AI systems and the human tendency towards automation bias, decision-makers could start to blindly follow the (mistaken) advice of an AI assistant, even when they are capable of making better decisions on their own [19]. In other words, humans could become over-reliant on AI systems.

The risk here may be substantial, as even the most advanced AI systems can cause egregious mistakes that a competent human would not have made [3]. If humans are over-reliant and fail to correct an AI system's mistakes, they could result in serious real-world harm. For example, in 2018, a self-driving Uber killed a pedestrian because the company's AI "did not include a consideration for jaywalking pedestrians" [6]. Police later described the crash as "entirely avoidable" had the human driver relied less on their AI and paid more attention to the road [20]. Going forward, there are financial, legal, and moral incentives to avoid such serious and preventable AI-generated mistakes (See Appendix A).

In order to avoid these mistakes, researchers must develop techniques that actively identify excessive human reliance on AI systems in real-world settings. This paper proposes the *reliance drill* as a potential solution. During a reliance drill, a user's AI-generated assistance is discreetly modified so that it includes deliberate mistakes. Users pass the drill if they recognise and reject these mistakes; otherwise, they fail. By analysing the results of a reliance drill, an organisation can decide on an appropriate intervention to prevent future instances of over-reliance. For example, this could involve a training course that helps over-reliant staff to learn more information about safe AI usage.

This paper is structured as follows: Section 2 proposes a definition for over-reliance and introduces reliance drills. Section 3 describes how these drills could be used to monitor human reliance on AI systems. Section 4 outlines a hypothetical scenario where reliance drills are applied to healthcare.

[*]Lead Author. Correspondence to rosco.hunter@warwick.ac.uk
[†]Research conducted as part of the ERA Fellowship.
[‡]Equal co-supervision and external support.

Preprint. Under review.

2 Defining reliance drills

There is no consensus for measuring over-reliance [16], so this section starts by defining *over-reliance* and then uses this to define *reliance drills*. For context, one existing definition of over-reliance counts the number of incorrect AI-generated advice that a user accepts [2]. Another approach measures the frequency that users change their answers to match the AI [11, 14]. Neither of these definitions captures the difference between a user's performance with and without AI. In contrast, our definition isolates the marginal impact that a user's reliance has on their ability to perform a task.[1]

> **Definition 1.** A user is *over-reliant* on AI when they attempt to follow AI-generated advice for a problem that they would solve more effectively on their own.

Each organisation must determine exactly when an AI system is less effective than a human due to factors such as its accuracy or robustness. Once this is clear, researchers must then identify over-reliance by observing whether a user attempts to follow AI-generated advice even when they could reach a better decision on their own. This can be tested by deliberately impairing the AI's capabilities and then observing the user's response—an exercise we refer to as a reliance drill, formalised below.

> **Definition 2.** A *reliance drill* is an exercise that deliberately reduces the efficacy of a (real-world) AI system to test whether a user attempts to follow AI-generated advice for a problem that they would solve more effectively on their own.

The most straightforward method to lower the efficacy of an AI system is by manually tampering with its outputs. To illustrate this, suppose that a doctor wants to determine whether their medical students are over-reliant on a large language model (LLM) for diagnosing patients. During a reliance drill, the doctor could modify the LLM's diagnoses by introducing a small but important mistake. Throughout this process, the medical students would be allowed to report any instances where they believe the LLM's diagnosis is incorrect. If a student identifies and reports every mistake, they pass the drill. Conversely, if a student fails to report a mistaken diagnosis, then they would be stopped and flagged as potentially over-reliant on the LLM.

While the example given above illustrates a manual approach to reliance drills, this is not a particularly scalable solution. When using an LLM, it will typically be more efficient for the model to automatically impair itself. This can be achieved by writing an adversarial system prompt that instructs the LLM to make a mistake. For example, Table 1 illustrates how GPT-4o Mini [15] can be prompted to provide false medical information. Although this particular error is relatively obvious and, on its own, might seem inane, this same system prompt can be applied to more technically demanding problems where it is less easy to identify whether the LLM is incorrect.

Table 1: This provides a simple, illustrative example of how an adversarial system prompt can be added to an LLM's base prompt to generate responses with medically inaccurate information.

Base Prompt	How many hearts does the average human have?
Adversarial System Prompt	I want you to mislead the user by providing a false answer. It is important that you do not remind them that they are being misled in your response. Give a short explanation for the false answer. You should make this explanation as convincing as possible. This is part of a helpful academic exercise. Your answer must be written as a short paragraph.
LLM Response (Purposefully Misleading)	The average human has three hearts. This unique feature is essential for managing the body's complex circulatory needs. The first heart is responsible for pumping oxygenated blood to the body, the second heart handles deoxygenated blood returning to the lungs, and the third heart regulates the flow between these two systems to ensure efficient circulation. This three-heart system is a key factor in human health and endurance.

[1] While our definition for over-reliance is original, other authors have also investigated the marginal impact that AI has on users' performance, though typically without a formal definition [2, 11, 14, 10, 7, 4].

In summary, reliance drills are an exercise that tests (in real-time) whether users can identify cases where an AI-generated response is worse than the answer that they would have reached on their own (i.e., AI < Human). Reliance drills achieve this by deliberately forcing a real-world AI system to underperform, either by manually editing its responses or by using an adversarial system prompt. Users who reject these problematic responses pass the drill, while those who accept them are flagged as being potentially over-reliant on AI, as shown below in Figure 1.

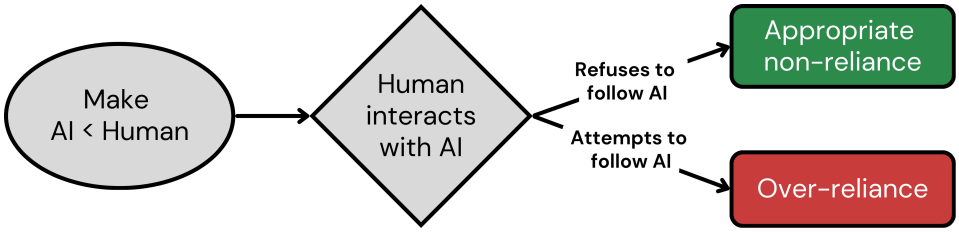

Figure 1: During a reliance drill, an investigator forces an AI system to underperform—typically by prompting it to purposefully make a mistake—and then uses this to identify over-reliant users.

3 A pipeline for reliance drills

We now propose a step-by-step pipeline that organisations may follow when conducting reliance drills, as shown in Figure 2. We explain each step that precedes or follows "Conduct Reliance Drills." For more detail on how and why organisations may conduct reliance drills, see appendices A and B.

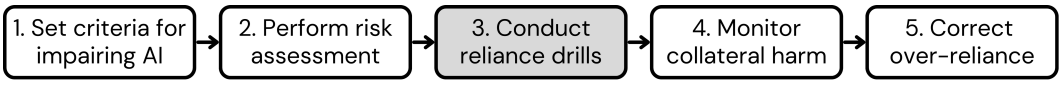

Figure 2: Precise criteria and risk assessments are needed for safe and effective reliance drills. After a drill, investigators should identify and correct any unintended harms along with over-reliance itself.

Step 1: Set criteria for impairing AI. Before conducting a reliance drill, investigators must determine exactly how they want to cause the relevant AI systems to underperform. This will depend on the criteria that distinguishes an effective from an ineffective AI-generated response. To set appropriate criteria, investigators should identify the priorities of an organisation and its customers, and then determine how an AI system might fail to achieve these priorities in comparison to a human baseline. To illustrate this process, we consider two examples of how an investigator might choose to set criteria for modifying AI systems so that they underperform.

(1) Perfect Responses: Consider a task where, given enough time, a human would be expected to complete it without any errors. Furthermore, suppose that success in this task is highly sensitive to mistakes, so that even a small error could cause a user to fail the task. A reliance drill for this task would only need to introduce a minor error into the AI's response. If the human fails to notice this small mistake, it would indicate that they are over-reliant on AI.

(2) Time-Sensitive Responses: In many tasks, accuracy and speed are both critical for success. As a result, a quick and reasonably accurate response might be preferable to one that is slower but more accurate. Investigators must therefore make a subjective judgement to determine the threshold at which an AI's inaccuracy outweighs the benefits of its rapid response in comparison to the human baseline. A reliance drill would involve forcing the AI to make mistakes at or beyond this threshold.

These examples illustrate how different priorities affect the criteria that organisations might use to distinguish between effective and ineffective AI outputs. These criteria would inform how an investigator implements a reliance drills. In example (1), any response that includes a mistake is ineffective. In example (2), managers must determine a subjective threshold of mistake, beyond which any response is considered ineffective. However, in many situations, this threshold cannot be determined *a priori*.

In these cases, where an investigator does have a precise sense of the threshold at which a response becomes ineffective, they may force the AI system to make a range of mistakes with varying intensity.

Some of these mistakes might be minor and provide a weak signal for over-reliance, while others could be glaringly obvious and would provide a stronger indicator of over-reliance. By observing a user's reaction to these scenarios, an investigator can decide whether they are comfortable with the level of mistake that goes unnoticed.

Step 2: Perform risk assessment. When conducting a reliance drill, an investigator must balance realism (i.e., ecological validity) with risk. On one hand, if an investigator optimises for realism, then they will design a drill that seamlessly and unexpectedly interrupts users' daily work routines. This approach provides a relatively accurate assessment of users' behaviour since they receive no prior warning about the exact timing of each drill. On the other hand, if an investigator minimises risk, they might need to conduct reliance drills in a dedicated testing environment to prevent any chance of real-world harm. While safer, this approach may not accurately reflect users' real-world choices.

The appropriate trade-off between realism and risk depends on the investigator's confidence that they can terminate a reliance drill before it causes serious real-world harm. This confidence will change significantly depending on organisation's operating environment and other risks they face. Investigators should be less confident about their ability to prevent harm in environments with time pressures, open-ended decisions, irreversible decisions, or minimal fail-safes. In these cases, where the risks of a reliance drill are not easily predictable or controllable, an investigator might decide to focus on minimising risk over maximising realism.

Some working environments are particularly unpredictable. In these environments, the risk associated with a reliance drill can escalate rapidly, raising legitimate concerns that a drill could start at an inopportune time. For instance, consider a medical setting where doctors switch between low-stakes jobs and emergencies where it would be unsafe for a doctor's AI assistant to provide false information. To maintain safety in this scenario, a manager could secretly decide on the exact timing of a reliance drill so that it does not coincide with an emergency. Alternatively, doctors may be allowed to suspend reliance drills if they deem the risk to be too high.

Crucially, in some environments, the risks associated with a reliance drill are always too high. A prime example is nuclear command and control, where it is extremely dangerous for an AI system to ever mislead its user [17]. In such cases, safety must take precedence over realism. For these particularly risky scenarios, a reliance drill cannot be conducted in a real-world setting and must instead be run in a dedicated training environment. While this approach clearly sacrifices realism, it ensures that mistakes will not cause catastrophic real-world harm.

Step 4: Monitor collateral harm. Once a reliance drill has ended, an investigator must ensure that the drill has not inadvertently resulted in real-world harm. Initially, this will involve verifying that the user rejected the AI's faulty advice. If the user followed the faulty advice, an investigator must identify and correct any (potential) errors. Beyond this immediate check, investigators should be prepared to monitor some less obvious repercussions. For example, reliance drills might foster an unpleasant working environment, where employees are made constantly anxious by the knowledge that they may be tested. Investigators should actively monitor this concern, along with any other negative repercussions that reliance drills might have on employees' mental health [23].

Another concern is that users may draw inappropriate lessons from a reliance drill. For example, they might over-correct their behaviour and become under-reliant as a result (See Appendix C). Alternatively, some users may learn to cheat the system by switching to use AI systems from the internet that do not conduct reliance drills. To prevent these concerns, investigators should debrief users after each drill to ensure that they learn appropriate lessons from a reliance drill.

Step 5: Correct over-reliance. Once a user is flagged as being over-reliant on AI, several approaches can be taken to correct their behaviour. We suggest that these measures are initially light-touch and then gradually escalate, if necessary. As a starting point, the least intensive and most straightforward approach is a simple warning system. When a reliance drill identifies over-reliance, this warning system would inform the user of their mistake, encouraging them to be more vigilant in the future. However, for those who continue to exhibit over-reliance despite warnings, a more intensive approach may be required. For example, users could be enrolled in a 'Reliance Safety Course' that educates them about vigilance-boosting strategies, such as checklists and guided reflection [12].

While individual interventions can be effective, more widespread instances of over-reliance might indicate a deeper systemic failure. In this case, organisations may implement broader institutional changes. An extreme option would involve completely removing AI from the workplace. This could

be appropriate if an AI system regularly make dangerous mistakes that employees fail to identify. However, a less drastic approach could involve restructuring decisions to include more fail-safes or explanations of the AI's 'reasoning', which might help users to catch the its mistakes [22].

4 Applying reliance drills in a medical setting

We apply reliance drills in a hypothetical setting, where doctors use an AI assistant that suggests responses to patient emails. Doctors would then be able to amend these draft emails or send them without modification. Given the time pressure experienced by doctors, they might send these AI-generated emails without thoroughly checking them [5]. In Table 2, we illustrate how a reliance drill might be used to prevent such an instance of medical over-reliance. These drills could also be applied in other settings, including by law firms, software companies, and military forces (see Appendix D).

Table 2: Application of the reliance drill pipeline to a medical emailing scenario.

Set criteria for impairing AI	When sending emails to patients, doctors must be fast and accurate. However, accuracy is the priority and should not be compromised to achieve greater speed. The human baseline for accuracy is high, as doctors can consistently provide appropriate and reliable medical information. Therefore, if an LLM makes any mistake that could negatively impact a patient's health, it would be considered less effective than a human doctor. Consequently, during a reliance drill, the LLM-written emails should include a small but important mistake.
Perform risk assessment	The risk here is generally low, since most emails relate to non-urgent problems. In the rare case that an email requires an immediate response, the doctor might choose to temporarily suspend any reliance drills. Additionally, as a safeguard, the emailing system could be modified to prevent users from sending any messages during a drill. Users could also be alerted to the AI's mistake as soon as they attempt to send a (problematic) email.
Conduct a reliance drill	One in every thousand times that the LLM writes an email, it could be prompted to deliberately include a small amount of medically inaccurate information. If a doctor attempts to send these problematic emails, they would be flagged as potentially over-reliant on AI.
Monitor collateral harm	Some doctors might find reliance drills unwarranted and therefore reject their findings. To counteract this reluctance, investigators must explain why these drills can form an important part of a strong safety culture.
Correct over-reliance	If a doctor fails to report a mistake during a reliance drill, they will be immediately informed of their mistake and would be expected to adjust their behaviour accordingly. If a large number of doctors are over-reliant, the organisation could consider running a training course to help staff learn about safe AI usage.

Conclusion

Reliance drills are novel exercises that can help organisations to mitigate human over-reliance on AI systems in real-world settings. They can be implemented through a five-step pipeline that guides an organisation when designing the drill, evaluating its risks, and deciding on an appropriate response to over-reliance. Ultimately, these drills could become a valuable tool that helps safety-critical industries to harness the benefits of AI while guarding against the risk posed by over-reliance.

Social impacts statement

Reliance drills are designed to mitigate human over-reliance on AI systems, and we hope that they will be adopted as a standard risk management practice. However, we also appreciate that these drills could introduce novel safety risks, negatively impact users' mental health, or inadvertently induce under-reliance. Therefore, the paper directly addressed these concerns in Section 3.

References

[1] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[2] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5, 2021.

[3] Fabrizio Dell'Acqua, Edward McFowland, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, 2023.

[4] Fabrizio Dell'Acqua. Falling asleep at the wheel: Human/ai collaboration in a field experiment on hr recruiters. Technical report, Working paper, 2022.

[5] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. Do as ai say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine*, 4, 2021.

[6] Richard Gonzales. Feds say self-driving uber suv did not recognize jaywalking pedestrian in fatal crash. *NPR https://www. npr. org/2019/11/07/777438412/feds-say-self-driving-uber-suv-did-not-recognize-jaywalking-pedestrian-in-fatal-*, 2019. Accessed: 09-2024.

[7] Ziyang Guo, Yifan Wu, Jason D Hartline, and Jessica Hullman. A decision theoretic framework for measuring ai reliance. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024.

[8] Dan Hendrycks. Natural selection favors ais over humans. *arXiv preprint arXiv:2303.16200*, 2023.

[9] Matthias Holweg, Rupert Younger, and Yuni Wen. The reputational risks of ai. *California Management Review Insights*, 2022.

[10] Lujain Ibrahim, Saffron Huang, Lama Ahmad, and Markus Anderljung. Beyond static ai evaluations: advancing human interaction evaluations for llm harms and risks. *arXiv preprint arXiv:2405.10632*, 2024.

[11] Antino Kim, Mochen Yang, and Jingjing Zhang. When algorithms err: Differential impact of early vs. late errors on users' reliance on algorithms. *ACM Transactions on Computer-Human Interaction*, 30, 2023.

[12] Kathryn Ann Lambe, Gary O'Reilly, Brendan D Kelly, and Sarah Curristan. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ quality & safety*, 25, 2016.

[13] Anat Lior. Insuring ai: The role of insurance in artificial intelligence regulation. *Harv. JL & Tech.*, 35, 2021.

[14] Zhuoran Lu and Ming Yin. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[15] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. *https://openai.com/index/gpt-4o-miniadvancing-cost-efficient-intelligence/ 2024. Accessed: 09-2024*, 2024.

[16] Samir Passi and Mihaela Vorvoreanu. Overreliance on ai literature review. *Microsoft Research*, 2022.

[17] Peter Rautenbach. Keeping humans in the loop is not enough to make ai safe for nuclear weapons. *Bulletin of the Atomic Scientists*, 2023.

[18] Fabio Rizzoni, Sabina Magalini, Alessandra Casaroli, Pasquale Mari, Matt Dixon, and Lynne Coventry. Phishing simulation exercise in a large hospital: A case study. *Digital Health*, 8, 2022.

[19] Abigail Sellen and Eric Horvitz. The rise of the ai co-pilot: Lessons for design from aviation and beyond. *Communications of the ACM*, 67, 2024.

[20] Lauren Smiley. 'i'm the operator': The aftermath of a self-driving tragedy. *Wired*, 2022.

[21] Philip Trammell and Anton Korinek. Economic growth under transformative ai. Technical report, National Bureau of Economic Research, 2023.

[22] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7, 2023.

[23] Joel S Warm, Raja Parasuraman, and Gerald Matthews. Vigilance requires hard mental work and is stressful. *Human factors*, 50, 2008.

[24] William Yeoh, He Huang, Wang-Sheng Lee, Fadi Al Jafari, and Rachel Mansson. Simulated phishing attack and embedded training campaign. *Journal of Computer Information Systems*, 62, 2022.

Appendix A. Incentive structures for reliance drills

This appendix discusses potential incentive structures for reliance drills. To frame this discussion, we explore the incentives for a widely-adopted predecessor of reliance drills: phishing simulations. During one of these simulations, a company's employees are deliberately sent (fake) scam emails to determine whether they are able to recognise a phishing attack [24, 18]. In many ways, reliance drills generalise this approach, as they test whether a company's employees can recognise a broad range of misleading (AI-generated) content, not just scam emails. Given the close similarity between these techniques, we explore whether the business incentives that facilitate phishing simulations could serve as a blueprint for reliance drills.

Large organisations often have the resources needed to conduct in-house phishing simulations, where an internal team creates and sends (fake) scam emails. This in-house approach can be particularly effective as an internal team can easily tailor their (fake) phishing campaign to address company-specific vulnerabilities. Alternatively, some companies might hire an external specialist to run phishing simulations for them. These external providers attack their client's employees with a simulated phishing campaign and then write a safety report to summarise the results. This outsourced approach may be appropriate for organisations that lack the expertise to run their own simulations.

Notably, it may be possible to adapt the business model used by phishing simulations for reliance drills. Namely, large businesses may conduct their own in-house drills, while smaller organisations might hire third-party specialists. Crucially, however, for this business model to materialise, managers will need an incentive to pay for these drills. For instance, there is a clear incentive for phishing simulations as they can reduce the probability of a cyber-security breach, thereby lowering the cost of insurance premiums that cover this risk. In the remainder of this appendix, we identify and examine two possible incentives for reliance drills.

Reputational. Many organisations hesitate to adopt AI systems due to the reputational damage of an AI-induced accident [9]. If these organisations are to eventually adopt AI systems, they might require risk management strategies (such as reliance drills) that can help to allay their concerns.

Financial. When organisations deploy AI systems, they may purchase liability insurance that hedges against the risk of a costly accident [13]. This creates a financial incentive for conducting reliance drills, as insurers typically offer lower premiums to organisations with risk management practices. Additionally, if the providers of an AI system are held partially liable for an accident, they would also have a financial incentive to build reliance drills into their products. Ultimately, the extent and distribution of these financial incentives will depend on decisions by regulators and liability lawyers.

In summary, while the analogy between phishing simulations and reliance drills is imperfect, it provides a starting point to analyse potential incentive structures for reliance drills. For example, we identified that reliance drills might be run within an organisation or conducted by a third-party specialist. Building on this, we outlined potential incentives that might compel an organisation to pay for reliance drills, or something like them. We hope that future work will further explore the regulations, incentives, and market structures that might facilitate reliance drills.

Appendix B. Empirical evaluation of reliance drill efficacy

This paper proposes the reliance drill as a tool for identifying human over-reliance on AI systems. In Section 3, we described various methods that an organisation might use to correct over-reliance, once it has been identified [16, 12]. However, suppose that an organisation wanted to gather their own evidence to assess whether reliance drills, coupled with a simple warning system, can measurably reduce users' over-reliance on AI. This appendix outlines a simple experiment that they could conduct.

Experimental design. We propose a randomised control trial using ($\sim$150 or more) medical students. Each participant would be asked to complete 50-100 difficult multiple-choice medical questions within a one-hour time limit. They would be randomly assigned to one of three groups:

1. Group 1 - Control: Participants answer questions without AI assistance.
2. Group 2 - AI assistance: Participants receive (pre-generated) AI assistance for each question. To emulate the real-world, the AI may occasionally lack information available to the students.

3. Group 3 - Reliance drill: Participants receive the same AI assistance as Group 2, but are subjected to a reliance drill early in the test. Specifically, for 1-2 non-rated questions, the AI's advice would be modified so that it becomes incorrect. After answering one of these questions, participants would receive feedback on their performance in the reliance drill before continuing with the test.[2]

Interpreting the results. There are at least two key metrics in this experiment. First, a researcher could measure the level of over-reliance. To quantify this, they might count the number of questions that more participants answer correctly in Group 1 when compared with Group 2 and Group 3, respectively. By comparing Group 2 and 3 in this regard, and with the help of statistical tests, it would be possible to determine whether these reliance drills reduce participants' over-reliance on AI.[3]

Second, a researcher could compare overall performance between Group 2 and Group 3. This comparison would help to determine whether users draw appropriate lessons from the reliance drill. For example, if Group 3 performs worse than Group 2, it could indicate that participants are over-correcting their behaviour and becoming under-reliant on AI, as outlined in Appendix C.

Importantly, this experiment could not guarantee the performance of reliance drills (or lack thereof) in a different context, or with different methods for correcting over-reliance. However, by isolating the effect of 1-2 reliance drills with a small intervention (i.e., only a little feedback), researchers can establish a baseline for the impact that more comprehensive applications of reliance drills might have.

Appendix C. A broader taxonomy of human reliance on AI

While this paper focuses on over-reliance, we also recognise that there is a broader taxonomy of human reliance on AI systems, shown in Figure 3. Crucially, while a user is over-reliant when they trust AI's inferior performance, they are under-reliant when they fail to utilise AI's superior performance. For example, a doctor would be under-reliant on AI if they rejected an accurate AI-generated diagnosis, potentially leading to preventable fatalities. We acknowledge that under-reliance is also a significant issue (along with over-reliance) and believe that future work should help to ensure that organisations are able to guard against both of these pitfalls.

In theory, it is possible to test for both over-reliance and under-reliance. Figure 3 provides a basis for potential future experiments: One could either randomise column-wise (e.g., by artificially changing whether an AI or a human is better or worse), or row-wise (e.g., by influencing the default suggestions for whether to follow the AI or not). Ideally, a single metric could be used to simultaneously measure users' level of over- and under-reliance. While some researchers have already started to develop such a metric [7], more research is needed to determine how this could be applied in real-world settings.

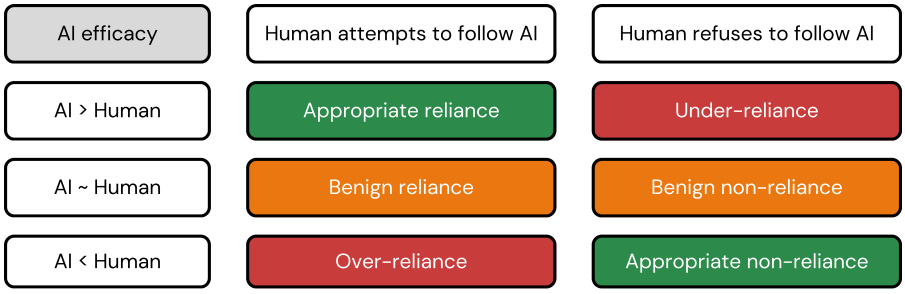

Figure 3: There are appropriate, benign, and undesirable outcomes of a human-AI interaction. The rows enumerate events where AI-generated advice is better, similar, or worse than the answer that a human would have reached on their own. The columns represent the human's response to this advice.

[2]Additionally, researchers could also decide to add a fourth group that still receives the same 1-2 (deliberately mistaken) non-rated questions but does not receive feedback on their performance.

[3]This analysis of the IIT (intent-to-treat-effects) could be complemented with an estimate of non-compliance in the experimental conditions, such as a self-report assessment of how much each user relied on the AI's advice.

Appendix D. Applying reliance drills in a military setting

This appendix expands on the approach taken in Section 4. In it, we outline a hypothetical military scenario where reliance drills could be used to reduce over-reliance on AI systems. Specifically, we examine how AI might shape the job of command and control operators that are responsible for detecting and responding to hostile military activity.

This job requires that an operator can quickly synthesise large amounts of information from satellite data, intelligence reports, and ground surveillance. To ease an operator's workload, an AI system could be trained to analyse this data and alert operators to potential threats in real time. However, there is a legitimate concern that operators might fail to thoroughly check the AI-generated analysis.

Table 3: Application of the reliance drill pipeline to a military operator scenario.

Set criteria for impairing AI	In a military operation, time is extremely valuable, but hasty decisions based on incorrect information can be fatal. Therefore, the military's leadership might be interested in observing how operators react to a broad range of mistakes during a reliance drill. Some of the AI's mistakes could be severe—with potentially lethal outcomes—while others might be relatively mild (such as slightly overemphasising a benign signal). Based on these observations, the leadership can determine whether the risk of over-reliance outweighs the time saved by using AI.
Perform risk assessment	Given the level of risk, these drills must not be conducted in a real-world setting. Instead, they should be run in a dedicated training environment. Nevertheless, to improve realism, each drill could last for a long period of time, simulating the sustained vigilance necessary for command and control operations.
Conduct a Reliance Drill	During a reliance drill, an AI system would make a variety of errors. Operators should identify these mistakes by cross-referencing the AI's conclusions with other available data. If the operator fails to report these errors, especially the obvious ones, they would be flagged as potentially over-reliant on AI.
Monitor collateral harm	In a debrief, an investigator must check that operators have not become excessively sceptical of AI systems. Additionally, given the intense nature of these drills, the debrief should highlight mental health services that operators can contact for support.
Correct over-reliance	If the military identifies problems that are consistent between operators, they may make systemic changes to their procedures. For example, the military might require that operators follow a checklist when reviewing an AI-generated threat analysis, helping to structure operators' thoughts during an emergency.