The State of Evaluation Science

Outline

For this memo, we understand "loss of control" over AI systems to mean the inability to predict or constrain either the behaviour of an AI system or the purposes for which it can be used. Evaluations are used to address these unknowns, generally seeking to better understand AI's capabilities and potential societal impacts, often with the explicit purpose of informing governance decisions. This memo provides an overview of current approaches in evaluation science to predict the capabilities and misuse vulnerabilities of AI systems. We then outline the benefits and fundamental limitations of evaluation science to consider when assessing the role of evaluations in AI safety and for informing AI governance priorities.

State of the Art Taxonomy

Capabilities Evaluations	Bias Evaluations	Misuse Risk Evaluations	Control Evaluations
Evaluates what AI systems do, testing for binary capabilities. **Example:** Factual recall ability on standardised tests. **Research:** MMLU (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), and HumanEval (Chen et al., 2021)	**Evaluates** what AI systems do, testing for subjectively undesirable biases. **Example:** Gender bias in LLMs. **Research:** Political bias (Potter et al., 2024; Motoki et al., 2024), territorial border bias (Li et al., 2024), geopolitical bias (Salnikov et al., 2025), military targeting bias (Drinkall, 2025)	**Evaluates** AI systems' resilience to potential attackers exploiting threat vectors, identified via adversarial testing, or "red teaming." **Example:** LLM leaking sensitive company data, or assisting in biological warfare planning. **Research:** Prompt Injection Benchmark	**Evaluates** the ability of monitoring systems to prevent rogue AI systems or agents from performing undesirable actions. **Example:** Monitoring system's ability to identify LLM deceiving its user **Research:** AI Safety Atlas, and the difficulty of evaluating deceptive agents

Benefits of AI evaluations

Lower Bound Capabilities: Capability evaluations can provide concrete evidence of what AI systems can do. When an AI system completes a task, such as manipulating humans in a controlled environment or identifying cybersecurity vulnerabilities, we know concretely that this is an ability (Barnett & Thiergart, 2024). Bias evaluations similarly provide evidence of concerning tendencies, at least at the lower bound of AI system's capabilities.

Red Teaming: Misuse risk evaluators acting adversarially as "bad actors" can identify threat vectors and minimise their risk through fine-tuning. Similarly, control evaluations can mitigate rogue AI systems by testing monitoring systems' ability to constrain deceptive agents.

Limitations of AI evaluations

Upper Bound Capabilities: Evaluations cannot provide strong evidence of the absence of capabilities (under-elicitation). Evaluations are proxies and cannot test for real-world or novel conditions. Evaluation assumptions, such as proxy-task design, heavily dictate results (see CybersecEval and Project Naptime).

Threat modelling difficulty: Red-teaming effectiveness relies on the evaluators being able to better predict threat vectors than malicious attackers themselves.

Forecasting: Evaluations cannot reliably predict the future capabilities of AI, nor can they anticipate discontinuous progress, as Anthropic, DeepMind, and OpenAI recognise. This means evaluations could provide false confidence about emerging threats.

The Future of AI Evaluation

Model Autonomy Risks: Sophisticated systems may recognise they are being tested and behave differently or hide capabilities.

Multi-turn Agent Evals: Multi-turn testing, suitable for robust Agentic AI evaluations, is still a nascent science (Drinkall, 2025)

Superhuman AI: The wider the difference in cognitive profiles of humans and AI, the more difficult threat modelling becomes. Referred to as "unknown unknown risks," the emergence of highly sophisticated AGI could lead to unpredictable capabilities.

Conclusion - While evaluations contribute to the **basic science of AI**, providing empirical evidence of capabilities that **prepare us for AI's societal impacts**, serving to **coordinate policy discussion**, they have fundamental limitations that, if not considered, could provide a false sense of control, especially as AI systems surpass human intelligence.