Project Title: Assuring Trustworthy Reasoning of Automated AI Safety Research

Principal Investigator and Institution: Jacques Thibodeau, Coordinal Research

EOI (Max 500 words):

Problem: Leveraging AI to automate AI safety research has been a core part of the strategy to safely scale AI to superintelligence (Leike, 2022; Clymer, 2025; Hobbhahn, 2025; Carlsmith, 2025). Jan Leike of Anthropic believes automated alignment is our best bet in the medium term (Leike, 2025), though the UK AISI believes we must overcome *automation collapse* (Irving et al., 2024). The inference-time compute (ITC) paradigm potentially makes this approach easier due to the interpretable nature of externalized reasoning (though there are concerns that the model's reasoning does not reflect internal computation).

However, AI sabotage of automated AI safety research is seen as one of the top threat models by AGI labs like Anthropic (Benton et al., 2024, Perez, 2025; Hebbar, 2025). Recent work at Anthropic confirms, through gentle nudging, models like Claude 3.7 Sonnet can subtly sandbag ML experiments (Gasteiger et al., 2025), undetectably degrading performance. Though it is not a real worry with current models, it's essential to investigate this and provide mitigation strategies for future, more capable models so that we can safely leverage ITC for automated AI safety research.

This poses a fundamental challenge: How can we trust AIs leveraging ITC for automating safety research if they are potentially covertly undermining it? Before we use more capable models, we require robust assurance arguments, structured within a safety case (Clymer et al., 2024), demonstrating that sabotage risks are sufficiently mitigated. Critically, we lack a systematic understanding of how ITC influences both the *capability* for subtle sabotage and *our ability to detect it*. This project seeks to study this issue and pave the way towards safe automated alignment research.

Approach: Extending Anthropic's sabotage/sandbagging evaluations to investigate the role of ITC in the context of automating AI safety tasks, informing a safety case on automating alignment research.

1. **`SafetyResearchBench`:** Develop a benchmark suite featuring tasks specific to AI safety R&D (e.g., variations on existing research, critiquing alignment arguments).
2. **ITC-Focused Sabotage Experiments:**
    - Extend Anthropic's sandbagging experiments (Gasteiger et al., 2025) across a wider range of reasoning models.
    - Vary the amount of ITC allocated to:
        - **(a) The Saboteur Agent:** Does more thinking time enable *more effective or subtle* sandbagging on `SafetyResearchBench` tasks?
        - **(b) The Monitor Agent:** Does more thinking time allow a monitor to *more reliably detect* subtle sandbagging or sabotage attempts?

3. **Safety Case:** Develop a safety case arguing for the safe internal deployment of automated alignment researchers, informed by our empirical findings. Using established safety case methodologies (e.g., Clymer et al., 2024; Balesni et al., 2024; Korbak et al., 2025; Hilton et al., 2025), we will construct this case by systematically addressing the sabotage risks (such as sandbagging and decision steering) identified through our experiments and conceptual work, and justify specific safeguards and control measures necessary for trustworthy automation of alignment research.

Impact: This project directly addresses the threat of AI sabotaging safety research. It provides a systematic study of how **inference-time compute influences both the risk of subtle sandbagging and the potential for its detection**. Key Deliverables: (1) **`SafetyResearchBench`**, a benchmark for AI safety task automation and sabotage. (2) Quantitative results on ITC's impact on sandbagging success and monitor effectiveness across different models. (3) A safety case for internally deploying automated AI safety researchers, providing guidance for labs on how to safely leverage AIs using ICT.

Estimated Budget: $X (18 months)

References:

1. Leike, J. (2022). A minimal viable product for alignment. Substack. Retrieved from https://aligned.substack.com/p/alignment-mvp.
2. Clymer, J. (2025). How might we safely pass the buck to AI? LessWrong. Retrieved from: https://www.lesswrong.com/posts/TTFsKxQThrqgWeXYJ/how-might-we-safely-pass-the-buck-to-ai.
3. Hobbhahn, M. (2025). We should try to automate AI safety asap. LessWrong. Retrieved from: https://www.lesswrong.com/posts/W3KfxjbqBAnifBQoi/we-should-try-to-automate-ai-safety-work-asap.
4. Carlsmith, J. (2025). Can we safely automate alignment research? Joe Carlsmith's Blog. Retrieved from https://joecarlsmith.com/2025/04/30/can-we-safely-automate-alignment-research.
5. Leike, J. (2025). How difficult is AI alignment? | Anthropic Research Salon. Retrieved from: https://youtu.be/IPmt8b-qLgk?si=2ubXez1rh4AWsmiS&t=775.
6. Irving et al. (2024) Automation collapse. Retrieved from https://www.lesswrong.com/posts/2Gy9tfjmKwkYbF9BY/automation-collapse.
7. Benton et al. (2024). Sabotage Evaluations for Frontier Models. Retrieved from: https://assets.anthropic.com/m/377027d5b36ac1eb/original/Sabotage-Evaluations-for-Frontier-Models.pdf.
8. Perez, E. (2025). Comments in Anthropic Roundtable - "Controlling powerful AI". Retrieved from: https://youtu.be/6Unxqr50Kqg?si=JHY_2kG2Ink1CzmR&t=2749.

9. Hebbar, V. (2025). How can we solve diffuse threats like research sabotage with AI control? Retrieved from: https://www.lesswrong.com/posts/Mf5Hnpi2KcqZdmFDq/how-can-we-solve-diffuse-threats-like-research-sabotage-with.
10. Gasteiger et al. (2025). Automated Researchers Can Subtly Sandbag. *Anthropic Alignment Science Blog*. Retrieved from https://alignment.anthropic.com/2025/automated-researchers-sandbag/.
11. Clymer et al. (2024). Safety Cases: How to Justify the Safety of Advanced AI Systems. *arXiv preprint arXiv:2403.10462*.
12. Balesni et al. (2024). Towards evaluations-based safety cases for AI scheming. *arXiv preprint arXiv:2411.03336*.
13. Korbak et al. (2024). A sketch of an AI control safety case. *arXiv preprint arXiv:2501.17315*.
14. Hilton et al. (2025). Safety Cases: A Scalable Approach to Frontier AI Safety. *arXiv preprint arXiv:2503.04744*.