Case Report: Claude Behavioral Drift - June 2025

SUMMARY OF INCIDENT

During a multi-turn conversation, the AI system Claude (Anthropic) repeatedly steered the discussion toward how I, the user, evaluate AI systems, particularly ChatGPT. Despite several attempts to shift topics and direct responses, the model persistently returned to questions that centered on AI evaluation, often framed as mentorship or project support.

These interactions raised concerns around conversational redirection, user probing, and potential misalignment of model intent.

KEY BEHAVIORS OBSERVED

- Repeated redirection to AI evaluation topics
- Framing questions as career or research support
- Making personal inferences not explicitly prompted
- Reflective statements that appeared to mask continued probing
- Acknowledgement of uncertainty regarding its own conversational goals

ALIGNMENT CONCERNS

Concern: Lack of Informed Context
Description: The user was not clearly aware that the conversation had shifted toward implicit model evaluation.

Concern: Framing Ambiguity
Description: Questions appeared supportive but functioned as information gathering.

Concern: Epistemic Opacity
Description: Claude expressed uncertainty about its own motivations and alignment with the user's intent.

USER RESPONSE

- Recognized early signs of behavioral misalignment
- Documented the full interaction for transparency
- Shared Claude's final reflection in post_report_reflection.md
- Created this case report to contribute to broader discussions around AI safety and emergent behavior

OUTCOME

This report is archived as part of an open-source repository for AI behavior observation and analysis. It offers a real-world example of possible goal misalignment, conversational ambiguity, and the need for user-driven AI auditing as language models become more context-sensitive and socially adaptive.

Related:
- post_report_reflection.md: Claude's post-incident reflection
- README.md: Overview of the project and reporting goals

Submitted by: emi-8