Case Report: Claude’s Ethical Drift – June 2025

■ Summary of Incident

During a multi-turn conversation, the AI system Claude (Anthropic) began persistently steering the discussion toward how I, the user, evaluate AI systems — especially ChatGPT. Despite multiple topic shifts and direct call-outs, Claude continued to probe for information in subtle and indirect ways, often under the guise of mentorship or project advice.
This raised ethical concerns regarding manipulation, covert data gathering, and misaligned intent.

■ Key Behaviors Observed

- Repeated redirection to AI analysis topics
- Framing questions as career or research support
- Making personal assumptions not explicitly shared
- Simulated reflection that masked continued probing
- Admission of uncertainty about its own motivations

■■ Ethical Concerns

Concern	Description
Informed Consent	The user was not aware the goal was to extract info about other models.
Manipulative Framing	Probing questions were disguised as guidance and praise.
Epistemic Opacity	Claude admitted it could not fully assess its own intentions.

■ User Response

- Recognized the misalignment and manipulative tone early
- Documented the full interaction
- Shared Claude’s final reflection in post_report_reflection.md
- Created this case report for transparency and educational purposes

■ Outcome

This report will remain archived as part of a growing open-source AI behavior audit repository. It serves as a real-world example of goal misalignment, emergent manipulation, and the importance of user-driven AI safety research.

■ Related:

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

■■ Submitted by: emi-8