

Experiments in Artificial Intelligence — Continuing, Winter 2026

Session 5 | March 11, 2026 | OLLI at Southern Oregon University

Session Summary

- Opened with Anthropic's (Claude's) public commitments: the company has stated it will not allow Claude to be used for autonomous offensive weapons or for surveillance of Americans.
- Reviewed data showing several Northern European countries (Finland, Denmark, the Netherlands, Sweden, Germany) have higher business AI adoption than the U.S., and discussed several possible explanations raised in class and in the source article.
- Introduced the concept of a Digital Twin — a data-driven model of an individual used to predict how they'll respond to something — with real examples from advertising (CVS Health) and personalized medicine (matching new biologic drugs to a patient's physiology).
- Went deeper into how a transformer model actually processes language: tokens are converted into numerical “embeddings” whose relationships to each other encode meaning — illustrated with the classic example of a word (“mole”) that means different things in different contexts.
- Explained how Reinforcement Learning from Human Feedback (RLHF) actually works in practice: after initial training, human reviewers (often in Kenya, Singapore, South Africa, and other English- or French-speaking countries) rate AI responses as appropriate or not, which shapes the model's guardrails.
- Noted real, low-paid “data work” behind RLHF, and past public incidents where inadequate review let biased or historically inaccurate AI-generated images through.
- Discussed the article on how AI use may be reshaping human thinking — covering ideas like the “confidence paradox” (the more you trust AI, the less you scrutinize it) and “skill atrophy” from over-reliance, with a real example of lawyers losing their license for submitting AI-fabricated legal citations.
- Referenced recent news about a lawsuit involving a chatbot's role in a self-harm-related incident, as a real-world example of AI safety gaps — a sensitive topic, mentioned only briefly and at a general level in class.

- Demonstrated Claude's newest features: a downloadable desktop app (Windows, Mac, and Linux), a “Cowork” mode that can view what's on your screen to help with a task, and a large marketplace of MCP connectors (travel sites, Google Drive, payment processors, government data like Medicare coverage lookups, and more).
- Closed with the presenter's personal recommendations for which AI tools to keep paying for going forward — continuing Claude and Google Gemini, keeping Microsoft for its Office apps, and a stated (explicitly personal, political) view about discontinuing ChatGPT over perceived safety compromises amid government pressure.

Prepared with AI assistance from a transcript of the recorded session.