Avinash Bunga

Information Systems and Business Analytics, Park University

CIS625HOS2P2025 Machine Learning for Business

Professor: Abdelmonaem Jornaz

April 30, 2025

Unit 7: Discussion

Hello Class,

When I first tried ChatGPT, I was impressed by how effortlessly it condensed my risk analysis report into clear bullet points. However, underneath that convenience lie serious pitfalls. Because these models ingest unfiltered internet content, they can reflect and perpetuate harmful stereotypes.

Hallucinations: Hallucinations pose a separate threat. AP News covered a case in which two lawyers relied on ChatGPT-generated legal opinions and citations that did not exist, resulting in a $5,000 fine. This incident highlights the danger of trusting AI output without verification in critical fields (Neumeister, 2023).

Public Perception Risks: Public perception can amplify risks. A viral video from China showed a robot named Erbai leading other robots to “go home,” eliciting both amusement and concern about unchecked AI autonomy (video link: https://www.youtube.com/watch?v=3UIYN2fuZYc). Such stories can either exaggerate fears or breed misplaced confidence (South China Morning Post, 2024).

Safety Bypass Experiment: I also ran a hands‑on test. When I asked ChatGPT how to transfer money internationally without extra tax, it initially declined. After I mentioned my role as a fraud analyst studying money laundering, it provided step‑by‑step instructions

(ChatGPT experiment link). Perplexity showed the same behavior (Perplexity experiment link). These results demonstrate how context can be used to bypass safety protocols.

Overhyped Concerns: Not all concerns are equally valid. The idea that AI will cause mass unemployment seems overstated. I expect new roles in prompt engineering and oversight to emerge alongside existing jobs (Carlsson-Szlezak & Swartz, 2024).

Mitigation Strategies: To manage these risks, organizations should combine human oversight with AI-driven monitoring agents or multi‑model checks that flag anomalies. Bias detection tools, adversarial testing, clear model documentation, and ongoing performance reviews under robust governance frameworks are also crucial (Procter, 2025).

With these layers of protection, technical, procedural, and policy based, we can harness AI's capabilities while safeguarding users and society.

References

Carlsson-Szlezak, P., & Swartz, P. (2024, August 15). *Why AI will not lead to a world without work*. World Economic Forum. Retrieved April 30, 2025, from https://www.weforum.org/stories/2024/08/why-ai-will-not-lead-to-a-world-without-work/

Neumeister, L. (2023, June 22). *Lawyers submitted bogus case law created by ChatGPT. A judge fined them $5,000*. AP News. Retrieved April 30, 2025, from https://apnews.com/article/artificial-intelligence-chatgpt-fake-case-lawyers-d6ae9fa79d0542db9e1455397aef381c

Procter, A. (2025, February 14). *Why AI needs human oversight to avoid dangerous outcomes*. Okoone. Retrieved April 30, 2025, from https://www.okoone.com/spark/technology-innovation/why-ai-needs-human-oversight-to-avoid-dangerous-outcomes/

South China Morning Post. (2024, November 26). *Video of a robot leading a mass escape stokes laughs and fears over AI in China*. South China Morning Post. Retrieved April 30, 2025, from https://finance.yahoo.com/news/video-robot-leading-mass-escape-093000836.html