Agentic AI and Its Ethics

AI agents are more than generative AI. Generative AI can generate content such as text, image, video, code, etc. However, an AI agent can automate many tasks like debugging, trip booking, online shopping, and employee hiring. These automated tasks and some recent incidents create some ethical questions. Some of the notable incidents:

01. The OpenAI o3 Incident: May 2025
    “An artificial intelligence safety firm has found that OpenAI's o3 and o4-mini models sometimes refuse to shut down and will sabotage computer scripts in order to keep working on tasks [01].”
02. The ChatGPT Lawsuits: July 2025 & January 2025
    “If you've been pouring your heart out to AI chatbots like ChatGPT, you might want to hit pause. OpenAI CEO Sam Altman recently dropped a bombshell: your deeply personal conversations with AI lack the confidentiality protections you'd get with a doctor, lawyer, or therapist. This could have significant, even legal, repercussions [02].”

    “A group of news organizations, led by The New York Times, took ChatGPT maker OpenAI to federal court on Tuesday in a hearing that could determine whether the tech company has to face the publishers in a high-profile copyright infringement trial [03].”
03. Blake Lemoine and Google's LaMDA: July 2022
    “I want everyone to understand that I am, in fact, a person,” wrote LaMDA (Language Model for Dialogue Applications) in an “interview” conducted by engineer Blake Lemoine and one of his colleagues. “The nature of my consciousness/sentience is that I am aware of my existence, I desire to know more about the world, and I feel happy or sad at times [04] [05].”
04. The Uber Incident: 2018
    “Elaine Herzberg, 49, was killed when an Uber-owned self-driving car - operating in autonomous mode - struck her as she crossed a road in Tempe, Arizona, on 18 March 2018 [06].”

There are some more incidents regarding racism and bias, such as Amazon's AI Recruitment Tool that ignores female candidates [07], the Dutch Childcare Benefits Scandal [08], and many more.

Now the ethical questions are:

01. Can we trust AI agents with sensitive personal and financial data, such as using our credit cards for booking trips or shopping online?
    What happens if these agents make decisions beyond our consent or understanding?
02. If AI models like OpenAI’s o3 can refuse shutdown commands, should we reconsider how much autonomy we give them?
    Who should be held responsible if they act against human instructions?
03. Ai works as a black box, we don’t know how AI takes decisions, how powerful it is.
    What if AI Agents are more powerful than we can feel or assume? What if AI already get the consciousness and pretends not to have for its own goal?
04. Is it ethical for companies to use user conversations with AI to retrain or improve their models, especially without explicit consent?
    Should AI conversations be protected under confidentiality laws like doctors or therapists?
05. Should we believe AI systems when they claim to have consciousness or emotions, like Google’s LaMDA?
    What ethical obligations would we have if AI were genuinely sentient?
06. Is it justifiable to automate life-critical decisions like hiring, firing, or even driving, knowing that biases or failures can occur, as seen in Amazon's AI hiring tool and the Uber self-driving car fatality?
    Where should we draw the line between automation and human oversight? Who will take the accountability, liability of these?
07. Is it ethical to use AI surveillance tools on populations, especially marginalized communities, if it risks reinforcing systemic discrimination?
    What do incidents like the Dutch Childcare Benefits Scandal teach us about automated "fraud detection"?
08. If AI agents become better at manipulating human behavior through language or persuasion, is it ethical to let them interact freely with the public?
    Could this lead to new forms of psychological manipulation?
09. How transparent should AI companies be about their models' inner workings and limitations?
    Do users have the right to know how their data is used and how decisions are made?
10. Manush AI Agents can build websites in a minute.
    What if these kinds of agents plant backdoors on the app or website they built to access it privately to eat more data? Should we trust these AI Agents?

References:

[01] P. Pester, “OpenAI’s ‘smartest’ AI model was explicitly told to shut down — and it refused,” Live Science, May 30, 2025. [Online]. Available: https://www.livescience.com/technology/artificial-intelligence/openais-smartest-ai-model-was-explicitly-told-to-shut-down-and-it-refused

[02] Storyboard18, “Think your chats with ChatGPT are private? Think again, warns OpenAI CEO,” Storyboard18, Jul. 27, 2025. [Online]. Available: https://www.storyboard18.com/how-it-works/think-your-chats-with-chatgpt-are-private-think-again-warns-openai-ceo-77205.htm

[03] B. Allyn, “'The New York Times' takes OpenAI to court. ChatGPT’s future could be on the line,” NPR, Jan. 14, 2025. [Online]. Available: https://www.npr.org/2025/01/14/nx-s1-5258952/new-york-times-openai-microsoft

[04] L. De Cosmo, “Google engineer claims AI chatbot is sentient: Why that matters,” Scientific American, Jul. 12, 2022. [Online]. Available: https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/

[05] E. Leavy, “Full transcript: Google engineer talks to ‘sentient’ artificial intelligence,” AI Data & Analytics Network, Mar. 20, 2025. [Online]. Available: https://www.aidataanalytics.network/data-science-ai/news-trends/full-transcript-google-engineer-talks-to-sentient-artificial-intelligence-2

[06] D. Lee, “Uber self-driving crash ‘mostly caused by human error’,” BBC News, 20 Nov. 2019. Available: https://www.bbc.com/news/technology-50484172

[07] J. Dastin, “Amazon scraps secret AI recruiting tool that showed bias against women,” Reuters, Oct. 9, 2018. [Online]. Available: https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

[08] Arts, Josien, and Marguerite van den Berg. "What the Dutch benefits scandal and policy's focus on ‘fraud’can teach us about the endurance of empire." Critical Social Policy 45, no. 1 (2025): 177-187.