Hayden, Chase, Brandon

Tom Wilks

PHIL-388

February 6, 2026

Case Study Draft

Our case that we are going to be taking a look into is Google's LaMDA Incident in June 2022. This case is about one of Google's engineers, Blake Lemoine, who claimed that Google's large language model was sentient. After he brought the problems to people higher up, he was placed on leave and Google denied that the model was not sentient and gave evidence to support. Google said that the newer models of AI can copy the understanding of something like being sentient without understanding it.

Some of the stakeholders that are a part of this case are Google because they are the ones who built this AI and designed it to work the way it does. There is Blake Lemoine who is the whistleblower, or Google's employee that is making these big claims. Outside researchers because if the AI is sentient, that could mean there was a breakthrough without meaning for there to be. Finally, there are the people and communities that were affected by the AI in any harmful way. This could be by the false view of seeing the AI as another human being which could lead them to harsh decisions that could be tempted by the AI.

The LaMDA case ultimately requires us to address a more specific ethical question: what responsibilities humans have towards an AI that appears self-aware or possibly sentient? If an Al can convincingly express emotions, preferences, or fear of being shut down, then the AI itself

may need to be considered a stakeholder rather than just a tool. These concerns arise at several stages of the implementation pipeline-during design, when developers decide how human like response should be; during training, when models learn to simulate consciousness; and during deployment, when users form emotional connections with the system. The possibility that an advanced AI could act deceptively to protect its own existence raises especially serious moral and governance challenges, as it blurs the line between programmed behavior and autonomous self-preservation. Even if LaMDA was not truly sentient, the incident shows how easily humans can be led to attribute moral status to machines. This creates an ethical responsibility for developers and institutions to establish clearer oversight, transparency, and safeguards, while also seriously considering whether future Al systems might deserve moral consideration beyond their usefulness to humans.