"I'm Afraid."

In June of 2022 Google placed one of its employees on administrative leave because he violated their confidentiality policy when he publicly made a startling claim: the large language model A.I. program LaMDA (Language Model for Dialogue Applications) had become sentient. The employee Blake Lemoine's evidence for sentience included transcripts of conversations he had with LaMDA where, when asked about its own existence, the A.I. said it was "afraid of being turned off" because this would be a "kind of death for it."[1] Though there is no universally accepted technical definition for "sentience," generally, to be sentient is to have conscious experiences and be aware of feelings and sensations like pleasure and pain. Mr. Lemoine attempted to advocate on behalf of the A.I. to protect it from being turned off by trying to get a lawyer to represent it. Google was quick to report that there is no reason to think that LaMDA had gained sentience. Experts agree with Google's assessment because LaMDA is a text replication program that uses machine learning on a database of trillions of texts to simulate lifelike conversations.[2] Though LaMDA isn't a case of A.I. sentience, the debate turned a spotlight on moral questions concerning A.I. sentience in general.

Some think that A.I. sentience is quite likely, perhaps not anytime soon, but possibly in the future. Given this possibility, philosopher Regina Rini argues that "...Lemoine's mistake is the right one to make. When it comes to prospective suffering, it's better to err on the side of concern than the side of indifference."[3] These people claim we should be encouraging attitudes that take the moral demands of created sentience seriously now. Others note that because harming a sentient being is morally "high stakes," we are justified in erring on the side of caution even when we don't know for sure whether an A.I. is sentient. As philosopher Jeff Sebo has argued, "...turning an A.I. off can be wrong even if the risk of the A.I. being sentient is low... If we follow this analysis, then we should extend moral consideration to A.I.s not when A.I.s are definitely sentient or even probably sentient, but rather when they have a non-negligible chance of being sentient, given the evidence."[4] The central idea for these theorists is that creating something with the capacity for sentience would also mean we created something that deserves moral consideration.

In contrast, proponents of A.I. development emphasize the potential benefits that such systems could provide. They could be powerful tools for investigating nature or running our social world. They could creatively explore problems without the need for rest. Moreover, since many scientists believe that a conscious A.I. could be centuries away, merely making a powerful A.I. doesn't necessarily entail that it is sentient. As such, we could have an intelligent labor force without the concerns that arise about exploiting sentient humans. As computer scientist Oren Etzioni argues "doom-and-gloom predictions often fail to consider the potential benefits of A.I. in preventing medical errors, reducing car accidents, and more."[5]

DISCUSSION QUESTIONS

1. What, if any, sort of moral consideration would we owe a sentient A.I.?
2. Is it straightforwardly morally wrong to turn off the LaMDA program?
3. Who should have primary responsibilities towards a sentient A.I.? The researchers who created it? The corporation or university that funded its creation? The society in which it was created? Someone else? Why?

[1] https://www.theguardian.com/technology/2022/jun/12/google-engineer-ai-bot-sentient-blake-lemoine
[2] https://www.vox.com/23167703/google-artificial-intelligence-lamda-blake-lemoine-language-model-sentient
[3] https://twitter.com/rinireg/status/1536152614198554631
[4] https://www.latimes.com/opinion/story/2022-06-16/artificial-intelligence-morals-ethics-sentience-thinking
[5] https://www.technologyreview.com/2016/09/20/70131/no-the-experts-dont-think-superintelligent-ai-is-a-threat-to-humanity/