ETHICS UNWRAPPED™

TEXAS McCombs | CENTER FOR LEADERSHIP AND ETHICS

CASE STUDY

"Trust is a psychological state comprising the intention to accept vulnerability based on positive expectations of the intentions or behavior of another." — Michael Tomasello

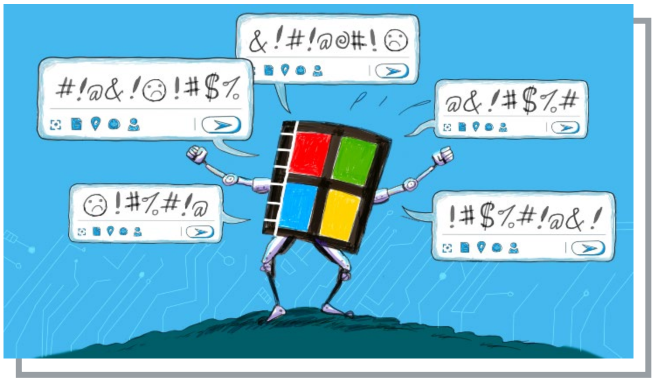
A.I. & Trust: Tay's Trespasses

On March 23, 2016, Microsoft Corporation released a chatbot it called "Tay" (an acronym for "thinking about you") as a Twitter bot. Apparently based on a similar bot Microsoft had successfully released in China, Tay was designed to mimic the speech patterns of a stereotypical female millennial. Ideally, Tay would learn from its conversations with users and thereby become smarter and smarter.

Microsoft's intentions were good. The earlier version Microsoft had released in China, called Xiaoice, had millions of users who sent it an average of 60 messages a month. Many of these users viewed "her" as a human-like friend. A Microsoft researcher wrote of Xiaoice:

> "...people don't necessarily care that they're chatting with a machine. Many see Xiaoice as a partner and friend, and are willing to confide in her just as they do with their human friends. Xiaoice is teaching us what makes a relationship feel human, and hinting at a new goal for artificial intelligence: not just analyzing databases and driving cars, but making people happier. —Wang

Unfortunately, Tay's U.S. launch did not go well. Almost immediately, users induced Tay to engage in antisemitic and other sorts of offensive and inappropriate speech. One troll tweeted to Tay: "The Jews prolly did 9/11. I don't really know but it seems likely." Tay soon tweeted: "Jews did 9/11," and encouraged a race war. As other trolls piled on, Tay was soon suggesting that Obama was wrong, Hitler was right, and feminism was a disease. Tay had a "repeat after me" capability, which made it particularly easy to lure it into communicating outrageous and distasteful messages. Microsoft soon began deleting the worst of Tay's tweets, which did not suffice. In less than 16 hours, Microsoft was forced to take Tay offline altogether.



Microsoft's Peter Lee, a corporate vice president, posted an apology, claiming that:

> As we developed Tay, we planned and implemented a lot of filtering and conducted extensive user studies with diverse user groups. We stress-tested Tay under a variety of conditions, specifically to make interacting with Tay a positive experience. Once we got comfortable with how Tay was interacting with users, we wanted to invite a broader group of people to engage with her. It's through increased interaction where we expected to learn more and for the AI to get better and better.
>
> The logical place for us to engage with a massive group of users was Twitter. Unfortunately, in the first 24 hours of coming online, a coordinated attack by a subset of people exploited a vulnerability in Tay. Although we had prepared for many types of abuses of the system, we had made a critical oversight for this specific attack. As a result, Tay tweeted wildly inappropriate and reprehensible words and images. We take full responsibility for not seeing this possibility ahead of time.

Microsoft began to fiddle with Tay, but accidentally re-released it on March 30, 2016. Tay quickly posted some drug-related tweets and then the entire account became stuck in a loop of repetitive tweets that affected more than 200,000 Twitter followers. Microsoft had to take Tay offline again.

Overall, Tay was a disaster for Microsoft, but this is certainly not the only AI product that has disappointed its intended users.

Flawed sensors in autonomous vehicles have killed people. Flawed algorithms have led courts to send people to jail to serve inappropriately long sentences because of their race. Flawed algorithms aimed at detecting serious medical conditions have produced high levels of both false positives and false negatives, all to ill effect. Even the best chatbots routinely hallucinate, sometimes with injurious results. A book written by AI gave inaccurate advice about mushrooms, leading to the hospitalization of several readers. Lawyers have been punished for filing AI-assisted legal briefs containing imaginary citations and judges have withdrawn legal opinions for the same reason.

All these problems foreground the critical issue of trust and AI. Trust is foundational to human relationships, including romantic ones. It is fundamental to all forms of human cooperation. Economic progress in nations is strongly correlated to the levels of trust in those nations. So, AI we can trust seems to be crucial to our successful interaction with AI. And, of course, people will be hesitant to use new AI tools and businesses will be reluctant to adopt them if, like Tay, these tools are untrustworthy.

Case study written by:

Robert Prentice, J.D.

Department of Business, Government and Society
McCombs School of Business
The University of Texas at Austin

Discussion Questions

1. Hoffman and Beato write:

> *"In the context of AI, we first must develop trust in the technologies themselves—no easy feat when the technologies are somewhat unpredictable and capable of error. If we have access to the tools, though, we can form opinions for ourselves: Is it reliable enough to do the things I want to do? Does it offer genuine value or just novelty? Does it empower me or make me overly dependent?"*

Are these reasonable questions to ask as you begin to decide whether to trust a new AI technology? If you have access to AI tools, can you adequately answer these questions for yourself? Which other questions might you ask?

2. Hoffman and Beato then note:

> *"Trust in the technologies is just the start, though. We must also cultivate trust in the developers of the technologies, the regulators of the technologies, and, perhaps most of all, the other users of the technologies. After all, why should you trust other people to use AI in primarily positive ways?"*

Can you think of AI technologies you would just as soon not see developed or marketed because you don't trust other consumers to use them safely and reasonably? Is the tale of Tay a worrisome example in this regard?

3. Hoffman and Beato go on to ask:

> *"Why should you trust your government—or the government of other countries—to [use AI in primarily positive ways]?"*

Is this a sensible question to ask? Why or why not? In terms of your own personal decision as to whether you should adopt a new AI technology, will government regulation play an important role in your decision? Why or why not?

4. In July 2025, President Trump announced a 28-page AI Action Plan to nearly completely eliminate government regulation of AI technology development. At about the same time, KPMG issued a report entitled: "Trust in Artificial Intelligence: A Global Study," which highlighted what the authors called "the need for effective regulation and governance." As a policy matter, in terms of cultivating trust in AI technology, do you favor more government oversight or less? Explain your thoughts.

5. As noted, Hoffman and Beato suggest that in deciding whether to trust AI, we ensure that it is reliable, valuable and empowering. Loreggia and her colleagues demand something more before we trust AI—that it act morally:

> *"It is imperative that we build intelligent systems that behave morally. For them to work and live with us, we need to trust such systems, and this requires that we are reasonably sure they behave according to values that are aligned to human values. Otherwise we would not let a robot take care of our elderly people or our kids, nor drive a car for us, nor would we listen to a decision support system in any healthcare scenario."*

Is this a reasonable requirement that we demand be met before we trust AI? If it is, can AI developers feasibly meet this requirement? How?

6. Many folks in the AI world are part of an "AI for social good" movement, also known as "AI4SG". These folks wish for AI tools to be not only reliable, valuable, and empowering and to behave morally, but also wish that AI tools do good in the world. Regarding trust and the AI4SG movement, philosopher Luciano Floridi argues:

   > *"Trustworthiness is essential for the adoption of technology in general and for AI4SG applications in particular. It is also required for technology to have a meaningful positive impact on human life and environmental well-being. The trustworthiness of an AI application entails a high probability that the application will respect the principle of beneficence (or at the very least, the principle of nonmaleficence.)"*

   It appears that Tay did not clear either the nonmaleficence nor the beneficence bar. Is it reasonable to expect the designers and programmers of AI products to do their best to ensure that their AI products meet these standards? Why or why not? Is nonmaleficence an important hurdle for AI products and services to clear before being released to the public? Explain

7. AI expert Larry Marcus observes: "I would be lying if I said I knew exactly how we can get to AI we can trust; nobody does." Should we suspend development of AI systems until someone does know how to get to AI we can trust? Why or why not? What do we risk if we suspend AI development? What do we risk if we forge ahead regardless? How do we balance these risks?

Sources

Saleh Afroogh et al., "Trust in AI: Progress, Challenges, and Future Directions," Humanities and Social Sciences, Vol. 11, Article No. 1568 (2024), at https://www.nature.com/articles/s41599-024-04044-8.

Peter Bright, "Tay, The Neo-Nazi Millennial Chatbot, Gets Autopsied," ars Technica, Mar. 25, 2016, at https://arstechnica.com/information-technology/2016/03/tay-the-neo-nazi-millennial-chatbot-gets-autopsied/.

Bhaskar Chakravorti, "AI's Trust Problem," Harvard Business Review, May 3, 2024, at https://bpb-us-w2.wpmucdn.com/sites.uab.edu/dist/6/536/files/2024/09/AIs-Trust-Problem.pdf.

Alistair Charlton, "Microsoft Tay AI Returns to Boast of Smoking Weed in Front of Police and Spam 200K Followers," International Business Times, Mar. 30, 2016, at https://www.ibtimes.co.uk/microsoft-tay-ai-returns-boast-smoking-weed-front-police-spam-200k-followers-1552164.

Caitlin Dewey, "Meet Tay, the Creepy-realistic Robot Who Talks Just Like a Teen," Washington Post, Mar. 24, 2016, at https://www.washingtonpost.com/news/the-intersect/wp/2016/03/23/meet-tay-the-creepy-realistic-robot-who-talks-just-like-a-teen/.

Luigi Floridi, The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities (2023).

Nicole Gillespie et al., Trust in Artificial Intelligence: A Global Study (2025), at https://kpmg.com/au/en/home/insights/2025/04/trust-in-ai-global-insights-2025.html).

Reid Hoffman & Greg Beato, Superagency: What Could Possibly Go Right with Our AI Future (2025).

KPMG, Trust in Artificial Intelligence: A Global Study (2025), at https://kpmg.com/au/en/home/insights/2025/04/trust-in-ai-global-insights-2025.html.

Neil Lawrence, The Atomic Human: What Makes Us Unique in the Age of AI (2025).

Peter Lee, "Learning from Tay's Introduction," Official Microsoft Blog, Mar. 25, 2016, at https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.

Andrea Loreggia et al., “Modeling and Reasoning with Preferences and Ethical Priorities in AI Systems,” in Ethics of Artificial Intelligence 129-154 (S. Matthew Liao, ed. 2020).

Larry Marcus, Taming Silicon Valley: How We Can Ensure That AI Works for Us (2024).

Paul Mason, “The Racist Hijacking of Microsoft’s Chatbot Shows How the Internet Teems with Hate,” The Guardian, Mar. 29, 2016, at https://www.theguardian.com/world/2016/mar/29/microsoft-tay-tweets-antisemitic-racism.

Melanie Mitchell, “Why AI Chatbots Lie to Us,” Science, Vol. 389, No. 6758 (July 24, 2025), at https://www.science.org/doi/10.1126/science.aea3922?utm_source=substack&utm_medium=email .

Abby Ohlheiser, “Trolls Turned Tay, Microsoft’s Fun Millennial AI Bot into a Genocidal Maniac,” The Washington Post, Mar. 25, 2016, at https://www.washingtonpost.com/news/the-intersect/wp/2016/03/24/the-internet-turned-tay-microsofts-fun-millennial-ai-bot-into-a-genocidal-maniac/.

Cathy O’Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (2017).

Rob Price, “Microsoft Is Deleting Its AI Chatbot’s Incredibly Racist Tweets,” Business Insider, Mar. 24, 2016, at https://www.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3.

Nigel Shadbolt & Roger Hampson, As If Human: Ethics and Artificial Intelligence (2024).

Michael Tomasello, A Natural History of Human Morality (2016).

U.S. Government, America’s AI Action Plan (2025), at https://www.ai.gov/action-plan.

Wikipedia, “Tay (chatbot),” at https://en.wikipedia.org/wiki/Tay_(chatbot).

Yongdong Wang, “Your Next New Best Friend Might Be a Robot,” Nautilus, Jan. 28, 2016, at https://nautil.us/topics/technology/?_sp=597e0a9c-861e-49f8-a0bc-a6f1794138dc.1753384883261.