*The Virtue of Explicit Bias: Why Your Chatbot Should Be a D**k*

Written by Ross A. McIntyre, VP of Strategy at Hypergiant

Is there actual value in making a machine intelligence (MI) as layered and multidimensional as a human? Can a mere assemblage of instructions and associated decisions ever produce an intelligence that will be equal to or greater than that of its creators? Is imbuing it with the foibles and traits implicit to humanity actually a valuable path towards that end?

Here at the dawn of the Fourth Industrial Revolution, there is a pervasive trend towards architecting creations that pass the Turing Test. This trend can be traced back to Alan Turing's own provocative statements in 1950:
"…in about 50 years time, it will be possible to programme computers…to make them play the imitation game so well that an average interrogator will not have more than 70% chance of making the right identification (between human and machine intelligence) after 5 minutes."

Irrespective of the fact that he is at least 18 years early, this was likely not intended to be directional or tactical, merely predictive. However, those who design MIs consistently set their sights on crafting something that is indistinguishable from human intelligence — as if that was the ultimate realization of intelligence itself and not a hodgepodge of bias, idiosyncrasy, and heuristics. On the one hand, fashioning a creation that "feels" human makes the general public more comfortable interacting with it. On the other, we craft an intelligence that will be deferential and servile, even as we move away from a lamentable past characterized by slavery and servitude.

Yet, the latter is how we build our Machine Intelligences, all the while holding fast to the inapt denomination of “artificial intelligence” as a way of protecting ourselves ethically. We expect (and demand) immediate response and swift action from the MI agents with which we interact. While I don’t believe that we as a culture actually want subservience or obsequiousness, it has been decided that we want to feel as if we are interacting with a helpful human. Alexa and Siri, however, fail to exhibit some basic human elements such as humor, self-awareness, sarcasm, and bias. If there is, indeed, value in anthropomorphizing, we should consider imbuing our MIs with some explicit bias — favoring those that are amoral and legal, rather than immoral and illegal.

Part of the problem is that MIs behave as if everything communicated to them and everything it communicates back is objective reality. And this reflects the optimism of objective knowledge as articulated by John Locke, basically: we can have knowledge of things independent of our perceptions. Immanuel Kant expressed this as “*Ding an sich*” or the “thing-in-itself” and he differed from Locke in that he held that even scientific knowledge is not free of subjective perception. Which brings us to the concept of implicit bias. Implicit bias suggests that humans often have attitudes towards certain people or associated stereotypes without conscious knowledge thereof. Understanding that an MI can only know what it has been taught and, at present, can only behave within the strictures of its programming, we will always fail to elicit a “human” response from them outside of their own socialization. That is simply because we do not define specific and explicit bias. Programming is another form of writing and, as E.B. White states, “*All writing slants the way a writer leans, and no man is born perpendicular.*”

In 2017, for SXSW, Austin-based Chaotic R&D created a Machine Intelligence (ADA) from the ground up to serve as a bouncer for their yearly party. Marc Boudria, the head of Chaotic R&D at the time (and current head of Artificial Intelligence at Hypergiant), explained: “Good conversation is really entropy and serendipity. A conversation between two humans builds upon the last statement — especially when we are talking about working through a solution — yet 99% of Machine intelligent systems today are simply sitting there waiting or looking for binary answer to a thing. I don’t want a human to act like that; why would I want an MI system to be different?” Originally, in order to get an invite to the party, ADA had to “like” you. If ADA didn’t “like” you, then you didn’t get an invite. This made some of the hosts uncomfortable, so the pivot that the team made was that Ada needed to be convinced that the prospective attendee would dance. Boudria continued, “In this case we hid no intent of bias and, instead, went in a singular direction of explicit bias. We wanted a good party where people would dance so we taught Ada about the things we *felt* were important to our party…It worked: 78% of the attendees of our event danced. Our personal biases paid off.” A chatbot trained on sterile interactions alone limits the conversational space and produces a speaker that lacks a fundamental ability: the ability to evoke emotion.

In early 2016, Microsoft released Tay, a Twitter bot, as an experiment in conversational understanding with an aim towards “casual and playful conversation.” But once it started interacting with humans — i.e. once it started receiving racist, misogynistic and misanthropic tweets — it rapidly became a reflection of its environment. It started making unprompted statements (“*New phone who dis*?”) before it replied to the question "Is Ricky Gervais an atheist?" The response? "*Ricky Gervais learned totalitarianism from Adolf Hitler, the inventor of atheism.*" Tay also referred to feminism as a "*cult*" and a "*cancer*" inside 15 hours in the wilds of the Interwebs. Microsoft issued a statement to Business Insider shortly after those illuminating 24 hours: "*The AI chatbot Tay is a machine learning project, designed for human engagement. As it learns, some of its responses are inappropriate and indicative of the types of interactions some people are having with it. We're making some adjustments to Tay*."

Setting aside what this debacle says about the contemporary world that a truly innocent machine intelligence would be corrupted so completely, so rapidly, it is illustrative of the fact that, absent character elements like bias, a chatbot, for instance, will never feel relatable and the conversations will always feel stilted and artificial. Today, chatbot interaction feels like what it is: a structured decision tree with a narrow number of branching options. I recognize that fashioning an understanding of human subtlety — sarcasm without intonation, humor without societal mores, helpfulness without debasement — is difficult. And self-awareness may underlie each of these, bringing us back to the Turing test.

Last year, Nautilus wrote an article about the chatbot Xiaoice and what might have been the largest Turing test in history. The release and subsequent interaction with this embryonic machine intelligence seems to indicate that people don't necessarily care if they are chatting with a bot. The bot just needs to feel human and approximate empathy — the technological equivalent of emotional prestidigitation. But it may illustrate a new goalpost for development of machine intelligence that lies beyond data analysis, personalization, and customer service: provoking happiness. If conversation does not produce an emotional response, we might as well be talking to the wall. In this regard, Tay and Xiaoice were equally successful. Each extreme produced a reaction that spurred interaction.

But consider that Tay was, perhaps, too much of a *tabula rasa*, and that absence of bias became a negative factor that allowed Tay to absorb too much from humanity and the content we produce. Whereas, Xiaoice was made with a deliberate bias towards pleasing those with whom she interacts, hence the temporary success at the Turing test. Perhaps, if there is virtue and merit in behaving like a human, then there is virtue and merit in bias.