A Balancing Act: The Real A.I. “Dilemma”

Dan McArdle

This paper is a response to“The A.I. Dilemma,” a presentation by Tristan Harris and Aza Raskin of the Center for Humane Technology on 9th March 2023.

In his 1739 essay "A Treatise on Human Nature," David Hume described what we now refer to as the "is/ought" problem. He noted that, when encountering a difficult problem, he was "surprised to find, that instead of the usual copulations of propositions, is, and is not, I meet with no proposition that is not connected with an ought, or an ought not." In other words, although we *can* do something, *should* we?

This tension has come to the public mind with the release of ChatGPT, a recent major realization of what artificial intelligence could, or already has, become. We can see this in the emerging struggle between what we might call the technologists and the philosophers. The technologists are very good at asking how we might achieve some goal– achieving space flight, curing disease, creating robot servants– but they are not good at asking whether this goal is a good idea. The philosopher points out that the technology which brings space flight also brings nuclear weapons; that some medicine might also cause drug addiction; and that, if we are not careful, we may become servants to our own robots. To better understand our current dilemma, we must first visit the perspectives of the technologists and the philosophers.

Technologist vs Philosophers

When we look at the world, it's obvious there are many problems: disease, war, famine, and so on. The technologist views these problems as challenges to solve. Consider the technological advances of the last 150 years, such as the automobile, the airplane, and the Internet. These were all responses to limitations imposed on humanity by nature. The automobile allowed us to travel long distances much faster than by horse, the airplane allowed us to stay in the air longer than a few seconds, and the Internet magnified our ability to communicate with one another. In each case, the technology focused on some attribute that already existed and improved it, granting us some long desired capability.

Because technology isolates a given existing attribute, its impact becomes a measurable phenomenon. If we plot this on a chart, we can immediately see the value proposition. For example, with no aviation technology, we can only jump in the air for a few seconds, but with the right tools, we can extend this to minutes, or even hours. The longer we are able to stay in flight with minimal effort, the more we can mark that as successful. Computer technology is similar: we measure the value of a computer by how fast it can process instructions, and we make all sorts of changes to improve the number of instructions it can process in the same time period.

Viewed through a purely theoretical mathematics lens, this approach makes a lot of sense. In statistics, the law of large numbers asserts that the more data we have, the closer we will get to the correct answer. Put another way, the more times we run the machine, or the more inputs we collect, the closer we get to assurance that the machine works as we expect, or that the inputs we are collecting represent accurate data. A similar principle emerges in calculus, where, following Zeno's paradox, we can approach the limit of some sequence or formula without ever actually reaching it. Mathematically, one can assert that the limit or the normal distribution in some ways represents the optimal value, or the truth of the situation. The closer we get to that value, the closer we are to being good.

The philosopher takes a different view. Rather than simply focusing on human limitations and seeing how we can overcome them, the philosopher asks why the limitations exist in the first place. Chesterton put it well when he suggested that upon seeing a fence, we should not take it down before asking why it was built. The philosopher points out that theoretical mathematics is only a *model* of real life: although we can collect lots of data, the philosopher asks whether we are sure the data truly represents what we want to learn. While the technologist strives to complete the calculation, the philosopher asks if the calculation is the right one to use.

If we look at ChatGPT as a kind of reconstruction of a human mind, several issues emerge. It is a fantastic system that can provide many answers, provided the questions are simple. If we ask it about the score of a sports game from 50 years ago, we will almost certainly get a verifiably correct response. But if we ask it *why* the team in question won the sports game, it will falter. Put another way, it is very good at representing given data points, but it is not very good at reconstructing complex ideas which may have different interpretations. We can test this by asking about heated or taboo topics, or unresolved questions, and it will almost never be able to give a good response. And when it attempts to, the responses often fall short, or are plain wrong. To put it bluntly, if ChatGPT were a human being, it would most closely resemble a cross between Kim Peek, the idiot savant who inspired the movie *Rain Man*, and Donald Trump.

Origins of the "Dilemma"

Now we must ask another question: when technologists apply technological solutions, it is to solve a perceived problem– but how do they discover what problems there are to solve? And, more specifically, what is the genesis of the ChatGPT solution? To answer this question, we must cover some history. The idea of creating a machine that can simulate humans is centuries old. In the 1730s, the French inventor Jacques de Vaucanson created autonomous mechanical ducks which could simulate digestion of food. Mary Shelley's 1818 novel *Frankenstein* was in many ways a response to the Industrial Revolution, a harrowing warning not to play God. And the Capek Brothers introduced the word "robot" in their 1920 play *Rossum's Universal Robots*.

But what really kicked off the modern debate on robots and artificial intelligence came out of the Second World War. After helping the Allies to crack the Nazi Enigma cipher with his "bombe" machine, Alan Turing expanded on his 1936 paper, *On Computable Numbers* (which established the modern computer, the "Turing Machine"), and in 1950 published *Computing Machinery and Intelligence*. This paper revisited the concept of how much a mathematical machine could do given the properties explored in his prior work, and asked if humans, in a discussion with someone else, would be able to discern if they were talking to another person, or to a machine. Dubbed "The Turing Test," this idea inspired MIT researcher Joseph Weisenbaum in the mid 1960s to create ELIZA, a computer "therapist," and to write his 1976 book *Computer Power and Human Reason* about it. The Turing Test has surfaced for decades in popular culture, including when H.A.L. is being interviewed by a reporter in *2001: A Space Odyssey* (1968), and when a "replicant" is being interviewed in *Blade Runner* (1982). In 2023, ChatGPT has emerged as the latest incarnation of this question.

One of the open questions in computing is, given a set of instructions, what properties may emerge over time? Perhaps the best known example of this is John Horton Conway's game of Life, in which there is a grid with a set of pixels which have various properties, and instructions they must follow with each proceeding state. Steven Wolfram expanded on this in his book *New Kind of Science*, where he goes over countless examples of states and instructions, and shows how they produce pretty pictures. We're starting to see more advanced versions of this with "deepfakes," as well as computer software that can take a given keyword or two and automatically generate images out of them. ChatGPT seems to do something similar. They all follow a fundamental principle native to the Turing Machine– given an initial data set and set of instructions, advance states until we produce something that seems to satisfy our requirements.

The same results which please the technologist can fill the philosopher with extreme anxiety. The philosopher steps back and asks questions about the results, as well as the process which produced them. They question not only *what* we have created, but also *why* we created it. After all, did we learn nothing from the story of the Tower of Babel? When man tries to become God, he destroys himself. The atomic bomb gave the Allies and edge and helped to win the war, but fundamentally changed the nature of warfare. Prometheus brought fire to humanity and was punished by Zeus; Icarus flew too close to the sun and burned off his wings; and now, to end the war, we created a weapon capable of destroying the entire world. One might look at the current discussions about ChatGPT and artificial intelligence in general and ask similar questions. This is the heart of the problem we see before us.

When technologists, as they have for the last few decades, disregard philosophers as antiquated and inhibiting progress, we soon rebirth challenges we thought we had long ago solved. In recent years, enrollment in the humanities has collapsed to the point that many universities are looking to cancel programs altogether, while adding funding and emphasis into STEM. What happens when our schools produce an abundance of graduates capable of building powerful machines, but not capable of asking whether such machines should be built? And, when these graduates are surrounded by people praising their intelligence, calling them “genius,” and backing these words with high-paying jobs, should it not logically follow that they consider themselves experts not only in their own field, but any other fields where they might apply their expertise? This is how we have hit our current dilemma.

Restoring the Balance

The AI technologists seem to have realized they do not have the correct answer, that their boat has gone downstream and they have no paddles. We can see hints of the Prodigal Son parable: the technologists pursued scientific progress without considering the societal implications, discovered the consequences, and are now crawling back and asking for forgiveness. To proceed as a society, we need to ask two important questions: what have they *actually* created, and why are they asking for help?

First, let’s review the facts insofar as they are publicly available. ChatGPT is closed source, but similar tools are open source. It looks like the technologists have created a system that reads in a natural language inquiry, breaks it into tokens, infers some kind of structural meaning, and then delivers a response by reassembling a vast data archive into a recognizable output. What comprises this data archive is not known, although it seems they have violated quite a few copyright and licensing agreements to assemble this data. There also appear to be no safeguards in place, allowing children to ask inappropriate questions and get equally inappropriate answers, or for sociopaths to learn how to build bombs or commit other crimes. Beyond this, the AI technologists do not seem to fully understand what they have created. In most industries, tools exist to troubleshoot or “debug” issues which arise, to help diagnose the problems and ensure they are fixed and do not happen again. Because a computer is fundamentally a calculator, it should be possible to trace any operation to understand why something bad happens and how to fix it. When the AI technologists claim there is no way to understand it, they are giving us balderdash, and proving that their operations are dangerously reckless and have the operating competence of a trailer park meth lab.

We should also compare their warnings and pleas for help with what they actually created. They claim to have unlocked some sort of Pandora’s Box, through which jobs will be lost, and fake news will become the norm. But is this really the case? Consider the ability to create fake videos, fake images, fake phone conversations, etc. All of these things are recordings, and none of them are possible without some kind of device like a telephone or a computer. For example, one could, in theory, use this AI technology to create a fake phone call to simulate someone's voice and steal sensitive information, but this could be thwarted by physically meeting the same person. Beyond that, much of this boils down to being able to detect whether something is a forgery. Many fields have experts who specialize in this practice, often brought in for appraisals or to validate transactions like auctions. But in the end, the real issue is not that these "deepfakes" can exist, but that people have become dependent on these devices and forgotten how to communicate without them.

And now we come to a big question: why are the AI technologists asking for help, and who are they asking? While an obvious answer would be that they are not experts in philosophy or law, and therefore seek help from experts, this is wanting. We don't need a degree in philosophy to understand basic hallmarks of human decency, such as that murder is bad. Do we really need laws against murder to understand instinctively that we shouldn't kill someone? Put another way, if there were no laws against murder, would it suddenly become acceptable? If, as seems obvious, the answer is no, then why did the AI technologists have no similar objections before embarking on what amounts to playing God? Perhaps this is the kind of mentality we arrive at when we focus on science to the exclusion of philosophy.

Notice that these technologists are asking for regulation. Is that necessary? Why can they not police themselves? Nothing would prevent OpenAI from taking ChatGPT offline while they add in safeguards. Are they afraid, as they claim, of what horrors this technology could unleash on the world, or are they afraid of being sued and put in jail for those impacts? By offering up scare stories of what could happen with a software project which remains closed source while allowing the general public to create accounts, it looks more like they are trying to cash out without liability. In a world that has forsaken morality for an ill-defined sense of ethics, they can parade about words like "regulation" without fear, knowing that writing any regulation would require their help, not only ensuring that they can keep their service online, but preventing other companies from competing, thus creating a legal monopoly for themselves.

So how do we move forward? It's time for the philosophers to step up and help to correct this imbalance. In addition to demanding answers to major and fundamental questions about how this technology is being used, we need transparency over what data sources are being used. ChatGPT represents the product of an imbalanced society which praises scientific progress over moral consideration. Petitions are nice, but they have no teeth: we to ensure the technologists actually listen and start collaborating with the philosophers so we can restore the balance.