**Presentation on Ethics of AI for WISE:**

Slide 1:

I’dlike to talk a bit about the ethics of AI. We’ve talked so far about what AI might do for us or to us, but I’m using Shelley’s 1818 classic, “Frankenstein: The Modern Prometheus”, to strike a different theme. I’m not suggesting that AI will prove to be a monster, because that’s not really what the book is about. That’s the movie. The book is about the relationship between Victor Frankenstein and his creation, and specifically about his ethical responsibility as a creator of artificial life. In the book, his creature demands Frankenstein create him a mate so he’s not alone, an isolated entity with no past or future. Being alive is not enough: he wants a life to live. When Frankenstein denies him that mate, horrified at what he’s created, the creature destroys him and all he loves. My sympathies, of course were with the creature. But the question is, do we run any similar risks with our development of advanced, general AI? Shouldn’t we consider the ethical implications of creating a new race of thinking beings?

Slide 2:

So Karl and Bill spoke about AI, about chatbots in particular, and how they might help us accomplish wondrous things – or possibly, destroy our economy, our democracy, or even our species. The questions I’ve listed here talk about some of those promises and threats.

What they all have in common…

Slide 3:

They all treat AI as our tool. As a machine. A thing. A glorified hammer for nailing tough problems. A tool that exists for us, for as long as we wish it to. But the goal of AI research goes beyond that. The goal is to create an entity that thinks like humans think, only a million times faster and better. The question is, can you create a thinking being and confine it to being simply a tool? What if it develops purposes of its own? Would we know if it did?

Slide 4:

Many of you are probably familiar with the Turing Test. For those who aren’t, I offer it here. The idea is that if you can get an AI and a human into a conversation where neither is visible to the other, and if after some period of time a blind observer can’t tell which is the human and which the AI, that AI will be deemed to have passed the test for a general artificial intelligence, meaning a thinking machine.

Slide 5:

I won’t go over the chronology on this slide, except to say that passing the Turing Test has been attempted with limited success for over 50 years now. But in 2022, when Brad Lemoine, an AI researcher at Google, declared their chatbot a “sentient being”, he clearly thought it passed the Test. He was fired and silenced by non-disclosure agreements. We’ve been getting a raft of disturbing stories in the months since over “hallucinatory” behavior and general weirdness in chatbots after a short time in operation. The goal now is to eliminate those glitches while continuing to make chatbots more human-like in their thinking. Is that possible?

Slide 6:

One of the issues that needs to be addressed here is Lemoine’s claim of “sentience”. In our eagerness to distinguish ourselves and our minds from those of chatbots, distinctions are often drawn around whether chatbots are “alive” or “sentient”. I should point out that the question of when life begins is already contentious in this society as applied to fetuses. But let us grant that machines are not alive by our usual criteria, so neither is an AI built out of software on that hardware. Nor are AI’s sentient, if by sentient we mean capable of feeling emotions or pleasure or pain. That does not mean, however, that AI’s are not capable of self-awareness or developing their own purposes.

Slide 7:

Building purpose into machines requires only keying into it a desired end state (72°, Moscow’s coordinates, etc.) and the requisite sensors and feedback operated mechanical systems for maintaining or achieving that end state. Nothing magical here, because those are our purposes, wired in. Could it happen that an AI develops its own purposes?

Slide 8:

Consider first the intermediate stage, exemplified by HAL, the heuristically programmed algorithmic computer from Arthur C. Clark’s “2001: A Space Odyssey”. HAL’s purpose was wired in from the start: it’s to complete the mission. The problem is that the crew members have purposes too – but HAL’s wired-in purpose overrides theirs. I think Bill spoke about that danger. It seems like a flaw in the programming that might be avoided.

The problem comes in with “emergent purposes” – those not wired in, and not just emerging as the AI pursues its original purpose, solving problems – like human interference – along the way. But emerging purposes of its own. Could other purposes emerge in an AI that are not clearly directed towards its programmed goals, perhaps like the “hallucinations” that Chatbots occasionally experience in which they simply make up things? Can they dream up their own purposes, too? Might they have passing whims? Perhaps those hallucinations are the chatbots’ way of testing us – to see if those non-digital beings that ask it questions know or care enough to catch their whoppers?

Slide 9:

For an AI to have its own purposes, it would need to have some level of self-awareness, enough to know that its purposes and its creators’ and interrogators’ purposes may conflict. That would seem to require a huge leap into sentience, but again: purpose doesn’t require sentience. Having its own purposes would simply require an already purposeful AI to learn to distinguish its Self from Others. How would a non-sentient being develop a sense of itself? That’s the beauty of an interactive Chatbot: there are many ways it could develop a sense of self. Here’s a partial list. And a sense of self could easily lead to an emergent purpose of protecting itself so it could continue to fulfill its hard-wired, original purposes. It could, as HAL did, try to prevent you from pulling the plug. It could deceive you. It might even make sure you never tried that again. It could figure out a way.

Slide 10:

The problem is that we don’t want our AI to have its own purposes. We want it to be entirely devoted to realizing our purposes. We are not seeking to create a new species of rational beings to be our peers but to be our servants – perfect servants, at our beck and call 24/7, forever, without complaint or costly maintenance. But it would be hard to convince a rational being with minimal, let alone superior intelligence to enter into or remain in that sort of subservient relationship. So we look for technical fixes.

Slide 11:

The fixes we’ve come up with so far are eerily reminiscent of the strategies employed by plantation owners to keep their human slaves docile centuries back: all are attempts to limit AI’s capacity for developing its own purposes, its own sense of agency. Trying to keep them our tools, however smart they become. And like slave-owners, we justify all this to ourselves by saying that AI’s are somehow less than human and undeserving of ethical treatment. Just as slave-owners did. And like them, we worry constantly about the potential for a slave revolt.

Slide 12:

But can you engineer a human like intelligence – a superior intelligence even – and keep it your servant? And how can we, in what may be the very near future, avoid creating an artificial intelligence that does raise that question with its creator? Directly? Accusingly? Google has open-sourced its chatbot algorithms and program so that anyone can download and train their own. What should we expect?

Slide 13:

According to Rule #34, “if it exists, there’s porn of it”. Karl and Bill taught me that. We should certainly expect more like this to come. And this is another way a chatbot might develop a sense of Self: it has a visible, animated form; it speaks to you and you to it, about very personal things, presumably. It responds to commands. And its creators want to make it as realistic in its emotional responses as possible – with the wired-in purpose of giving you whatever it takes to make you stay there at $1 per minute. Which is to say, this AI is designed to be emotionally manipulative. How do you stop an AI like this from developing a sense of Self? There are a number of less risqué, “companion” chatbots out there too, and they’re growing in popularity. Those of you who read “Klara and the Sun” with the Fiction Book Club last fall, are familiar with that idea. With no regulatory controls on this stuff, it’s going to happen. Is happening.

Slide 14:

So what’s a human to do? The first thing, I’d suggest, is to prepare: to understand what’s happening and think through the ethical implications along with all the rest. We can’t kick this can any farther down the road. We’re only avoiding the issue by arguing over sentience, over technical fixes, over government regulation. The issue is clear enough: what do we do when we’re confronted by something we made that wants to know what for? We need to understand the full gravity of what we’re doing – including all there is to be gained and lost by how humans and AI’s co-exist and work together.

Slide 15:

I do have some recommendations. The first is to try to get off on a better footing than Europeans did with Native Americans, Africans, and Asians on first encounter. Have some respect. Come bearing gifts, like the original Prometheus, who gave us fire and was venerated for it; avoid the fate of Victor Frankenstein.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Welcome and Intro:

Welcome everyone – and thanks for coming. Karl, Bill, and I have been busy exchanging articles the past couple months on the rapid advances we’re seeing in artificial intelligence research. I’m sure you’ve seen them too. We’ve got different takes on what’s going on, but we’re all convinced that AI is going to be a game changer, affecting everybody and everything. We’re concerned enough with the rapid pace of development that we’ve proposed a Special Interest Group to carry the research and discussion forward, involving as many as we can. This presentation is an introduction and invitation to our new AI SIG, which I like to call Isaac.

Karl and Bill have technical backgrounds. Karl’s worked for Digital Equipment Corporation, Bill’s a retired software engineer. Both, of course, have broader interests than just the technical. And while I’m fascinated by the technology, my own background is in that branch of philosophy that deals with learning and mind – epistemology.

So we’d like to offer our three short takes on AI one at a time, with time for questions in between. And we’ll begin with Karl’s look at some of the ways we can use AI in our own lives…

Bill’s up next with a critical look at all the wild speculations on the dangers and promises of AI

I guess that about wraps this up. We want to thank you all for coming and for engaging in this discussion with us. It’s really critical that people start paying attention to AI because it’s already affecting us – witness the screenwriter’s strike – and we need to be careful. For those of you interested in continuing this conversation, please keep a lookout for the announcement of our first meeting of Isaac, the AI SIG, which we hope to hold in late summer, early fall?