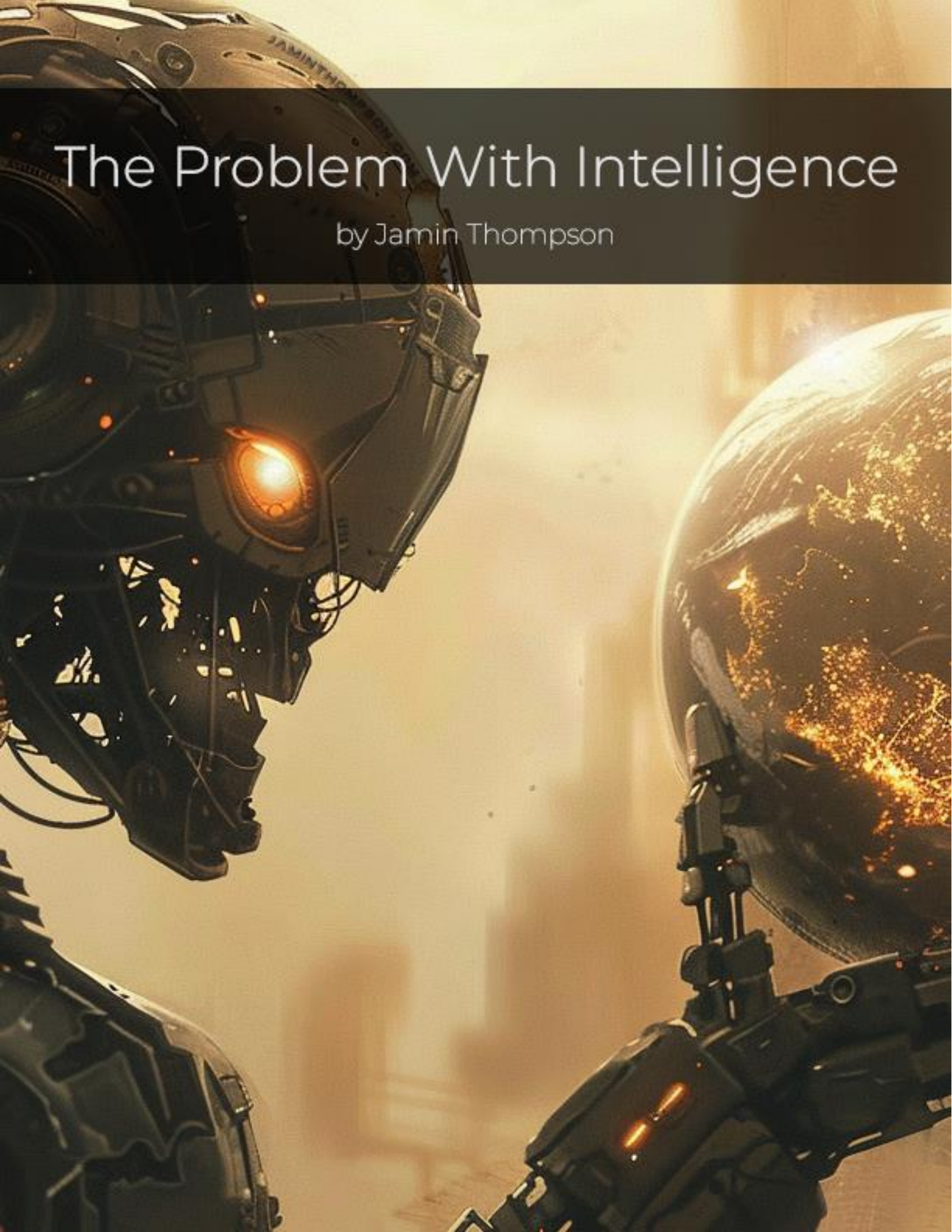
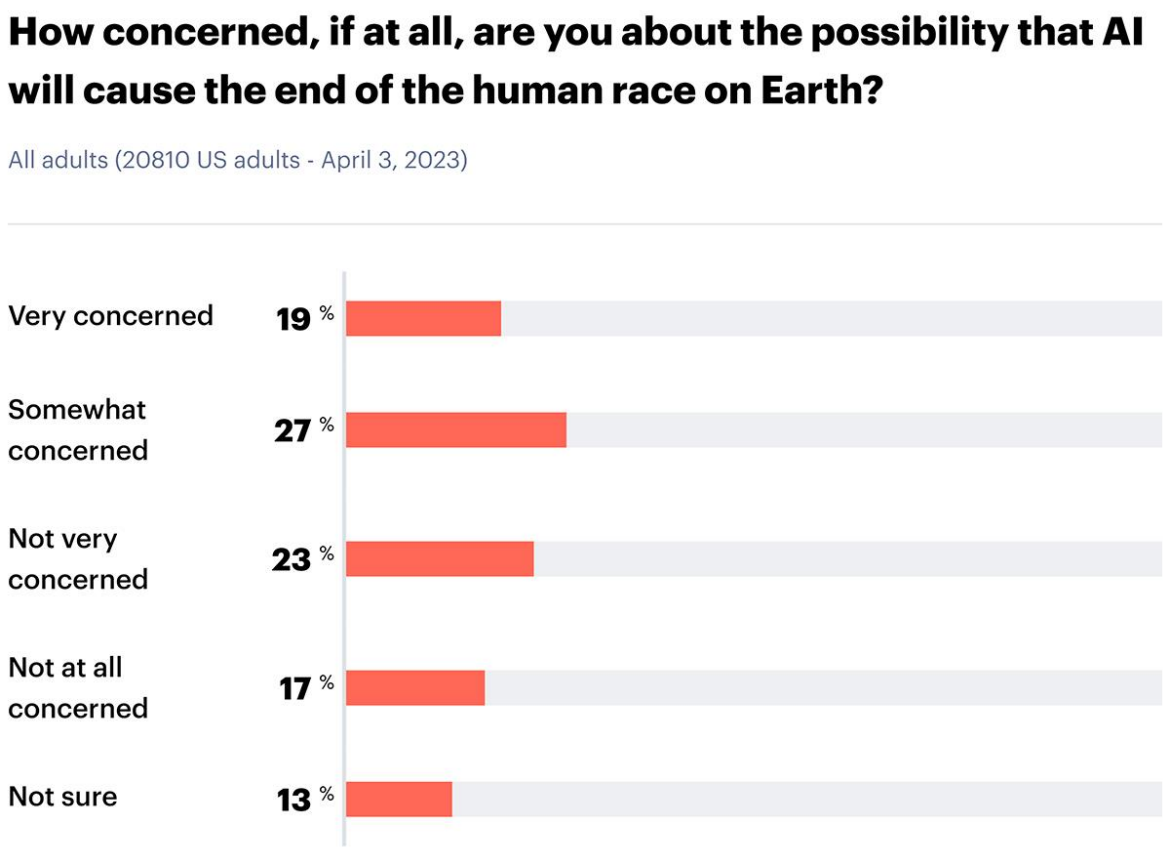
The Problem With Intelligence

by Jamin Thompson



Last spring, YouGov America ran a survey on 20,810 American adults.

46% said they are “very concerned” or “somewhat concerned” about the possibility that AI will cause the extinction of the human race on Earth.



There do not seem to be meaningful differences by region, gender, or political party.

Black individuals appear to be somewhat more concerned than people who identified as White, Hispanic, or Other.

Younger people seem more concerned than older people.

Furthermore, 69% of Americans appear to support a six-month pause in “some kinds of AI development”. (More)

Not to be outdone, I ran a poll of my own on my social media accounts.

It's interesting to see how opinion is distributed on controversial and thought-provoking topics.

My poll (posted on Twitter and Instagram) suggests that a majority of my followers who responded feel emotionally repulsed by the idea of an "AI conquest" and that such a scenario would be the "most horrific" of all.

On the opposite end of the spectrum, few researchers think that a threatening (or oblivious) superintelligence is close.

Indeed, the AI researchers themselves may even be overstating the long-term risks.

Ezra Karger of the Chicago Federal Reserve and Philip Tetlock of the University of Pennsylvania pitted AI experts against “superforecasters”, people who have strong track records in prediction and have been trained to avoid cognitive biases.

In a study published last summer, they found that the median AI expert gave a 3.9% chance to an existential catastrophe (where fewer than 5,000 humans survive) owing to AI by 2100.

The median superforecaster, by contrast, gave a chance of 0.38%.

Not only was the opinion gap between “superforecasters” and AI experts quite massive, it didn’t appear to shrink, even after debate and recalculation.

Why the difference?

For one, AI experts may choose their field precisely because they believe it is important, a selection bias of sorts. (More)

It’s quite interesting when self-proclaimed Bayesians (who are quite intelligent) sharing evidence don’t converge.

That said, to have such a large percentage of answers have a significant deviation from the expert predictive “superforecasters”, one needs some sort of basis in theory.

Alas, most of the theory arguments that I’ve heard re AI destruction seem quite inadequate. I thus suspect there may be more to the puzzle here.

That said, let’s dive a bit deeper.

Since 2022 or so (give or take a year) we have seen a massive surge in AI development and technical progress.

A few artificial intelligences (AIs) now seem able to pass the famous Turing Test, basically making them nearly indistinguishable from another human.