Nature and the Machines

Huw Price and Matthew Connelly[1]

Does artificial intelligence (AI) pose existential risks to humanity? This question has been on some surprising tables lately – tables in the White House and 10 Downing Street, amongst other places. Some critics feel it is getting too much attention. They want to push it aside, or into the distant future, in favour of conversations about the immediate risks of AI.

These critics now include the journal *Nature.* A recent editorial [1] urges us to '[s]top talking about tomorrow's AI doomsday when AI poses risks today.' The piece concludes: 'Fearmongering narratives about existential risks are not constructive. Serious discussion about actual risks, and action to contain them, are.'

A week later, in a similar vein, the *Washington Post* published a piece titled 'How elite schools like Stanford became fixated on the AI apocalypse' [2]. This, too, is dismissive about existential risk from AI. It calls it 'a purely hypothetical risk', 'derived from thought experiments at the fringes of tech culture.' It quotes critics who say that these claims 'sound closer to a religion than research', and that their proponents obsess 'over one kind of catastrophe to the exclusion of many others.'

We have two kinds of concerns about these criticisms. First, they seem to assume, absurdly, that we can't be worried about two things at the same time (as though we had to stop walking to chew gum, as the old line has it). Why shouldn't we care about both kinds of AI risks, short term and long term, and indeed their connections? If, for instance, machine-learning algorithms designed to addict people to their devices are contributing to a growing mental health crisis, that will make it harder for us to deal with every other threat [3]. And some near-term threats only become apparent when we pay attention to the latest developments in AI, such as the potential use of Large Language Models by bad actors to develop and deploy weapons of mass destruction [4].

The need to trace such connections, and find ways to address the full range of AI risks, has long been appreciated by institutions such as the Leverhulme Centre for the Future of Intelligence (CFI). Since its launch in 2016, CFI has aimed to be inclusive on the short term/long term axis, amongst many others.

Our second concern is more troubling. We feel that the present criticisms involve a serious error of judgement, for a scientific voice as influential as *Nature*. To explain why, let's step back a few paces, and think about risk management in general.

Dodging icebergs

From aviation to zoo-keeping, there's a simple rule for safety in potentially hazardous pursuits. Always keep an eye on the ways that things could go badly wrong, even if they seem unlikely. The more disastrous a potential failure, the more improbable it needs to be, before we can safely ignore it. Remember icebergs and frozen O-rings. History is full of examples of the costs of getting this wrong.

Where does the responsibility lie for enforcing this rule? Sometimes with a single individual, or small group: a ship's captain, say, who decides not to slow down despite

---

[1] Huw Price is Distinguished Professor Emeritus at the University of Bonn and an Emeritus Fellow of Trinity College, Cambridge. He was previously Academic Director of the Centre for the Study of Existential Risk and the Leverhulme Centre for the Future of Intelligence, Cambridge. Matthew Connelly is the current Director of the Centre for the Study of Existential Risk and Professor of History at Columbia University.

warnings about icebergs; or a single NASA committee, deciding to ignore engineers' warnings about the O-rings. But even in these cases, individuals and committees do not operate in a vacuum. It is an institutional responsibility, to some extent, and good institutions are designed with this in mind.

For familiar risks, this can be fairly easy. Risk management procedures can themselves be formalised and institutionalised. Think of checklists in aircraft cockpits and operating theatres, or triage procedures in an emergency room.

Well-designed safety procedures are sensitive to the costs of errors. To borrow some examples from the philosopher Heather Douglas [5], emergency rooms accept a lot of false positives, subjecting many healthy people with mild chest pain to ECGs and blood tests. They thus achieve a very low rate of false negatives – sending home people with life-threatening heart problems. In a well-designed criminal justice system, on the other hand, we accept a small number of false negatives (mistaken acquittals) in order to avoid false positives (mistaken convictions).

Douglas points out that 'we expect people to consider the consequences of error', even in less institutional settings. 'Hence the existence of reckless endangerment or reckless driving charges', as she puts it.

Dealing with novel risks

For novel risks these lessons are harder to apply. We don't have the luxury of learning by trial and error, and planning checklists accordingly. But there are some useful rules of thumb, which are based on experience.

One good rule is to listen to your experts, insofar as you can tell who they are. By training, experience or simple insight, some people are going to be better sources of advice than others. If those people are warning you about a risk, it makes sense to listen.

A second rule is that these experts will be telling you things you don't want to hear, in many cases. That's their job, and not a reason for doubting their expertise. If you find yourself disrespecting the messenger because you don't like the message, that's a danger sign. Good institutions will be designed to mitigate this danger. Hospital managers should not have the power to override their consultants on issues of critical care, no matter how annoying they might find them.

A third rule is that if other people have interests at stake, they, too, may have a reason for disrespecting the messengers. Other parts of an organisation may resent the attention given to the safety team, and people elsewhere may be adversely affected by the team's recommendations.

This means that decision makers often need to act in a noisy environment, where some of the noise is attacks on the character and motivations of the would-be whistleblowers. Real life is messy, but well-designed procedures will try to filter this out, unless it bears on the credibility of the advice.

Finally, it's worth repeating that all of this matters most where the stakes are highest. 'We expect people to consider the consequences of error', as Douglas put it, and the more disastrous the possible consequences, the higher this expectation should be.

Frozen O-rings

We want to return to AI, but first an example from recent history. On January 28, 1986 the Space Shuttle *Challenger* exploded after launch, killing its seven crew. The explosion was caused by a failure of the O-rings – elastic seals in the solid-fuel booster rockets. This was itself caused by

low temperatures the night preceding the launch, a danger that had been foreseen by engineers at the company Thiokol, which supplied these components on the boosters.

One of the engineers was Roger Boisjoly. After his death in 2012, NPR [6] recalled an interview they had with him and one of his colleagues, a few weeks after the *Challenger* disaster. Boisjoly told NPR then that they had foreseen the problem with the seals six months previously, and predicted 'a catastrophe of the highest order' involving 'loss of human life' in a memo to managers at Thiokol.

The day before the launch, when the predicted launch-time temperature was below freezing, 'Boisjoly and his four colleagues … concluded it would be too dangerous to launch.' What happened next is a lesson in how not to listen to one's safety team.

> Armed with the data that described that possibility, Boisjoly and his colleagues argued persistently and vigorously for hours. At first, Thiokol managers agreed with them and formally recommended a launch delay. But NASA officials on a conference call challenged that recommendation.
>
> 'I am appalled,' said NASA's George Hardy, according to Boisjoly and our other source in the room. 'I am appalled by your recommendation.' Another shuttle program manager, Lawrence Mulloy, didn't hide his disdain. 'My God, Thiokol,' he said. 'When do you want me to launch – next April?' …
>
> Boisjoly and his colleague … told us that the NASA pressure caused 'Thiokol managers to 'put their management hats on.' … They overruled Boisjoly and the other engineers and told NASA to go ahead and launch.

Is AGI clear for launch?

There are differences between the *Challenger* and AI, of course. One is that for the *Challenger*, the issue of expertise was exceptionally clearcut. Boisjoly and his fellow engineers had actually designed the system whose failure they predicted.

In the AI case the community now concerned about existential risks includes many leading scientists in the field, both from academia and industry. But artificial general intelligence (AGI) is not launching tomorrow. So the link between the engineers and the (claimed) risky technology is not nearly so direct.

Another difference is that in the *Challenger* case there were institutions in place, with the authority to assess the advice of the engineers. Those institutions failed, apparently for reasons that seem familiar, in the light of our simple rules. Appalled disdain for one's experts is not a helpful frame of mind. Thiokol's management hat should not have overruled its engineering hat. Nevertheless, the institutions were there. We know whose responsibility it was to get things right.

In the AI case, by contrast, we don't yet have institutions to do the job. Much of the recent discussion has been about how to create them. In our view, this means that the responsibility that such institutions would bear has to fall much more broadly, for the time being. It has to rest on the shoulders of other institutions, scientific and otherwise, that are capable of influencing discussion in this case. As the world's leading scientific journal, *Nature* itself is one of these institutions, in our view.

Now to the biggest difference. In the case of the *Challenger,* the predictable cost of a false negative was already very high. The Space Shuttle had no crew recovery system, so a loss of the vehicle meant loss of the crew. Still, these losses are tiny compared to those that would be at stake if AI really comes with existential risks. Seven people compared to eight billion, not to mention the future of the biosphere.

What level of safety should we be aiming for? In the case of the *Challenger*, NASA management told the Rogers Commission that they aimed for a risk of catastrophic failure of 1 in 100,000. In Richard Feynman's appendix to the commission's report, he noted that NASA engineers he surveyed estimated the risk much higher – about 1 out of 100 launches.

Obviously we should be looking for a much safer comparison for the future of AI. Perhaps commercial aviation? There are approximately 100,000 commercial flights per day, apparently. If we assume about 3 serious accidents per year, that's about 1 serious accident per 10 million flights. We might regard that as a very crude upper bound on the level of risk that would be acceptable for species-threatening AI.

Reckless endangerment?

As we said, these concerns about AI have been raised by some (though by no means all) of the field's leading scientists. These are not fringe figures. Clearly, *Nature's* response does not rest on a comparable body of expertise, assembled to support a case on the other side. This is not to take a side on the issues, but simply to point out that *Nature* cannot possibly have the expertise required to do so, let alone at the level of certainty that would be required.

In the light of this, could *Nature* have taken due care 'to consider the consequences of error', as Heather Douglas put it? It is hard to see how it could have done so. The proper bar for excluding these risks is very high. It would be absurd to suggest that that bar has yet been reached anywhere, in adequate scientific discussion, let alone that the editors of *Nature* have achieved it, behind closed doors.

In these circumstances, for such an influential voice as *Nature* to dismiss these risks as 'fearmongering narratives' is inappropriate and irresponsible, in our view. We urge *Nature* to reconsider.[2]

References

[1] Stop talking about tomorrow's AI doomsday when AI poses risks today. *Nature* 618, 885-886 (2023). https://doi.org/10.1038/d41586-023-02094-7
[2] How elite schools like Stanford became fixated on the AI apocalypse. *Washington Post,* July 5, 2023.
https://www.washingtonpost.com/technology/2023/07/05/ai-apocalypse-college-students/
[3] Cave, S., ÓhÉigeartaigh, S.S. Bridging near- and long-term concerns about AI. *Nat Mach Intell* 1, 5–6 (2019). https://doi.org/10.1038/s42256-018-0003-2
[4] Gaulkin, T. What happened when WMD experts tried to make the GPT-4 AI do bad things. *Bulletin of the Atomic Scientists,* March 30, 2023.
https://thebulletin.org/2023/03/what-happened-when-wmd-experts-tried-to-make-the-gpt-4-ai-do-bad-things/
[5] Douglas, H., Rejecting the Ideal of Value-Free Science. In Kincaid, K., Dupré, J. and Wylie, A. (eds.), *Value-Free Science? Ideals and Illusions*, Oxford University Press, 2007, 120–141.
[6] Berkes, H. Remembering Roger Boisjoly: he tried to stop shuttle Challenger launch. *NPR,* February 6, 2012.
https://www.npr.org/sections/thetwo-way/2012/02/06/146490064/remembering-roger-boisjoly-he-tried-to-stop-shuttle-challenger- launch

[2] We are grateful to Haydn Belfield and Martin Rees for comments on this material.