From *Peter Ho's Menagerie,* S Jayakumar, J Kwek, A Kuah (eds), World Scientific, 2024, 49–68. doi.org/10.1142/13883

Calling Attention to Elephants

Huw Price

'Black elephant' is such a clever phrase that it is hard not to see it as an example of what it describes. Like other kinds of beasts, vivid and witty metaphors can be under our noses for years, awaiting the rare individuals who can sniff them out. I don't know whether Peter Ho was prime sniffer for 'black elephant', or merely an early adopter, but I am indebted to him for the term, either way.

Striking metaphors wear their virtues on their sleeves. By definition, they need to be vivid and visible, once spotted. Not so for most black elephants, unfortunately. Often, a large part of the discoverer's challenge is to get others to see what now seems so obvious to her or him. The more unwelcome the discovery, the harder it is to persuade an audience that it has actually been made.

I recall Peter's advice about this, in a talk he gave to our group in Cambridge. Take it slowly, Peter said, in small steps. Keep your audience within their tolerance zone, and don't give them more of the elephant than they can digest at one sitting. (I may be embellishing here. I don't think Peter recommended eating his elephants, except as cupcakes at a Centre for Strategic Futures anniversary dinner!)

Peter's remarks throw some light on some of my own intellectual trajectory. In his terms, it feels to me as though I've been a would-be elephant spotter, and attention-to-the-elephant seeker, about several matters. I've often found myself recommending viewpoints that – despite what look to me like evident reasons for taking them seriously – seem invisible to many of my peers, academic and otherwise. Often, my audience seemed to have little appetite for the message.

In this piece I'm going to describe four of my personal elephants. I apologise to the reader, and especially to Peter, for this rather self-centred approach. I couldn't resist this opportunity to gather all four into the same arena. But I'll try to muster some broader conclusions, as well as the beasts themselves. Some elephants matter more than others, and for different sorts of reasons. My little herd is diverse enough to draw some conclusions about the species in general.[1]

Elephant 1: Retrocausality in quantum mechanics

Quantum mechanics (QM) was developed in the 1920s. Along with Einstein's relativity theory, it became one of the two great pillars of modern physics. And yet it is exceedingly strange, in ways which are still not completely understood.

One of its weirdest features is called *entanglement.* Entanglement allows pairs of quantum particles to remain strangely connected to each other, long after they have interacted. As the Irish physicist John Stewart Bell showed in the 1960s, this implies that a measurement made on one particle may have a subtle effect on the other, even though in principle they could be light years apart. So there's a deep tension between QM and relativity theory, which seems to prohibit such 'nonlocal' effects. As Bell himself put it forty years ago, there is 'an apparent incompatibility' between the two theories, 'at the deepest level.'[2]

But there has long been a way to remove this incompatibility, in plain sight. It was first proposed in the 1950s, before Bell's work, by the French physicist, Olivier Costa de Beauregard. Costa de Beauregard suggested that quantum measurements might have a subtle influence on the past behaviour of

[1] My herd is larger than these four, but the rest live in regions so remote from everyone except philosophers that I won't try to describe them here.

[2] J.S. Bell, "Introductory Remarks", *Physics Reports,* 137(1986): 7–9.

the particles concerned, as well as in the normal way on the future. This option, now called *retrocausality,* would allow a pair of entangled particles to be connected via a zigzag path, through the point where they interacted in their common past. That avoids the need for 'spooky action at a distance', as Einstein called it, and the conflict with relativity goes away.

For my part, I first heard about Bell's work at a workshop at Wolfson College, Oxford, in 1977. (I was an MSc student in Mathematics at the time.) The speaker remarked that Bell's argument for action at a distance assumed – 'of course', he probably said – that quantum measurements couldn't influence the past. Then why not abandon that assumption, I thought?

For a long time, I wondered whether there was some simple objection to this possibility that I was missing. How else to explain why the experts were all ignoring a promising way of reconciling QM and relativity? These days, having had a hand in exhibiting some of the additional advantages of the idea,[3] I'm more convinced than ever that it must be right. It is still a minority view, though more visible than it used to be in the field at large.

With Peter's metaphor at my disposal, this now seems to me to be a classic scientific elephant. Because it is the key to reconciling QM and relativity, it is as big as ideas get in theoretical physics. And it's been in the room for more than half a century. (Is it black? We'll come to that.)

Elephant 2: Conscription and hereditary monarchy

The Australian head of state is the British monarch – now King Charles. Like a majority of my compatriots, apparently, I think we should become a republic. Australia should have an Australian head of state, as people put it. I agree with this sentiment, but I think there's also a stronger reason, one that applies equally in the UK, or indeed in any of the other modern constitutional monarchies. (There are about a dozen of them, if we count separately the countries that share King Charles.)

This reason is that hereditary monarchy involves conscription of children for public office. A country such as Britain or Australia would not dream of allowing child conscription for any other public office. We should therefore abolish it for the supreme public office. What could be more obvious (in my view)?

I first wrote about this issue publicly in 2012, prompted initially by my Inaugural Lecture as Bertrand Russell Professor of Philosophy in Cambridge.[4] I wanted to use the lecture to celebrate the centenary of a famous 1912 lecture by Bertrand Russell himself. Russell's topic is causality. He's against it, and compares it dismissively to the monarchy.

> The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm.[5]

What harm did Russell have in mind, I wondered, on the side of the monarchy? That question turned out to be surprisingly difficult to answer. Russell didn't express his views on the question anywhere else, so far as I was able to determine – not even in his surviving correspondence.

However, Russell was a well-known advocate of allowing children more freedom than was usual in Britain at the time. So in my lecture I suggested, a bit cheekily, that my objection would be a good fit.

---

[3] See, e.g., Huw Price and Ken Wharton, "Taming the Quantum Spooks", *Aeon,* 14 September 2016, https://aeon.co/essays/can-retrocausality-solve-the-puzzle-of-action-at-a-distance, "Untangling Entanglement", *Aeon,* June 29, 2023. https://aeon.co/essays/our-simple-magic-free-recipe-for-quantum-entanglement.

[4] Huw Price, "Where would we be without counterfactuals?", in *New Directions in the Philosophy of Science*, ed. Maria Carla Galavotti *et al* (New York: Springer Cham, 2014), 589–607.

[5] Bertrand Russell, "On the Notion of Cause," *Proceedings of the Aristotelian Society, New Series* 13 (2013): 1.

I also wrote some public pieces at that point, around the time of the birth of George, the first child of the then Duke and Duchess of Cambridge, William and Kate.[6] I pointed out that if we moved quickly, George could be allowed a comparatively normal childhood, with the opportunity to choose his own path in life.

My pieces attracted a small amount of attention, and two main objections. Many people said that there was really no compulsion involved. The heir could simply step down, if he or she didn't want the job. Others said that the Windsors were no different from many other families, in which children were encouraged to follow in their parents' footsteps. (The Murdochs were often cited.)

Since then, Harry's example has rebutted the first objection, and he's simply the spare, not the heir. In any case, more options for the heir simply mean fewer for the spare, in the present system. The state takes its victim, one way or the other. Most importantly, a cage one can force open when one grows up is still a cage. The heir is still denied a normal childhood, with the normal options for considering his or her own future, at that point.

As for the Murdochs, it would be vastly easier for Lachlan Murdoch to step away from the family business than it has been for Harry, let alone William. (I'm told that Lachlan's siblings have already done so.) No one outside the Murdoch family would care, one way or the other.

What a family does is a private matter, within limits. What we do to the Windsor children is a public matter. If the Murdochs choose to restrict their children's options, that's a matter for them. If we do it to the Windsor children, it's a matter for all of us. We should be ashamed of it, in my view.

This point seems as obvious to me now as it did in 2012 – more so, if anything, now that Queen Elizabeth is no longer with us. We can admire her lifetime of public service, while at the same time feeling that it was cruelly unfair that it wasn't voluntary. The point is in plain sight, with no hidden details. So it is an elephant, black in this case in moral terms, at the heart of all the liberal democracies that still rely on hereditary monarchies.[7]

In this case, there are reasons why most people who have views about the monarchy are unwilling to acknowledge the elephant. If they are monarchists, it would be a huge embarrassment, to say the least, to admit that their favoured system has child conscription at its heart. If they are republicans, it is an embarrassment for a different reason. Their usual criticism of the monarchy involves the claim that it involves a kind of unearned privilege, out of place in a modern democracy. It sits uncomfortably with this view to admit that the royal children are in some ways grossly underprivileged, in being denied freedoms that all other children enjoy.

Elephant 3: AI and existential risk

One of my early projects in Cambridge was to help to establish the Centre for the Study of Existential Risk (CSER). This opportunity stemmed from a chance meeting with the Skype-founder Jaan Tallinn in 2011, and an existing acquaintance with Lord Martin Rees (Master of Trinity College at the time). Both of these high-achievers had a keen interest in potential catastrophic risks of new technologies, and my initial role was as a catalyst. I brought them together to see whether something could be done in

[6] Huw Price, "Time for Some Royal Prerogative – Let's Give Kate's Child a Choice," *The Conversation,* January 29, 2013. https://theconversation.com/time-for-some-royal-prerogative-lets-give-kates-child-a-choice-11518; "It's a Boy – But Baby Cambridge Deserves Choices in Life," *The Conversation,* July 23, 2013. https://theconversation.com/its-a-boy-but-baby-cambridge-deserves-choices-in-life-16154.

[7] Some of the European and Scandinavian countries seem more aware than Britain of the issue, taking steps to try to ensure that the children in their royal families live something approaching normal lives. But none, so far as I know, has faced up to the basic point. The present system amounts to conscription for public office, and will come to be seen as such.

Cambridge, to draw some attention to these understudied topics. A year or two later, the three of us became the co-founders of CSER.[8]

In Tallinn's case, his main concern when I met him was with the risks of artificial intelligence (AI) – especially the possibility that AI might escape human control at some point, with existential consequences for humanity and the biosphere. In 2011 this was a fringe topic, studied only by a very small community of researchers. I think Tallinn himself was the first person I had encountered who took these concerns seriously. I was interested in the ideas themselves, and impressed by his evident commitment to trying to do something about them.

Such issues are now so prominent in public discourse that it may be hard to understand how controversial they seemed, just a few years ago. But when Tallinn first came to Cambridge, in February 2012, I had arranged that he be invited to give a public lecture by the Cambridge Centre for Science and Policy (CSaP). Some members of the CSaP team were worried that his topic was too 'way out' for their Distinguished Lecture series.

Yet the ideas were hardly new, even at that point. Alan Turing wrote and spoke about them publicly in the early 1950s, as did his Bletchley Park colleague, the statistician I J Good, some fifteen years later.[9] So the notion that AI might exceed human abilities, and that that might be disastrous for us, had been in the room for decades. The change in its visibility over the past decade is a fascinating case study for elephant watchers.

Extreme risk and the culture of science

Before I introduce my fourth elephant, I want to mention some of the considerations that influenced work in CSER. It was clear that mitigation of extreme technological risks would depend on evaluation of possibilities that seem far-fetched, in some cases. Many might turn out to be of negligible concern, but the net needed to be cast widely in the first instance, to maximise the chances of catching the fish that matter, as early as possible. Given the nature of the risks involved, there would be a high cost to 'false negatives'.

Unfortunately science is not good at casting its net widely. As the philosopher Thomas Kuhn observed in the 1960s, science is conservative, and there is strong cultural pressure on scientists to work within current paradigms.[10] Advances (Kuhn's 'scientific revolutions') often depend on far-sighted individuals who resist these pressures, to work outside the mainstream.

The history of science offers examples of figures whose work was shunned for long periods, before eventual vindication. Of course, it offers far more examples of fringe proposals that were not vindicated by later developments. In general, we rely on the normal process of science to sort out the gems from the dross. It may take a long time in some cases, but we get there in the end. In the special case of extreme risks, however, such a delay might be extremely costly. One of CSER's early projects aimed to investigate this danger, and ways to reduce it.

In 2017 CSER organised a workshop on 'Risk and the Culture of Science', describing the theme like this:

> Many scientists have expressed concern about potential catastrophic risks associated with powerful new technologies. But expressing concern is one thing, identifying serious candidates another. By definition, such risks will be novel, rare and difficult to study; data will be scarce,

[8] It was CSER that eventually put me in touch with Peter Ho, and many of his Singapore colleagues.

[9] For references, see Kelsey Piper, "The Case for Taking AI Seriously as a Threat to Humanity", *Vox,* October 15, 2020. https://www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment

[10] Thomas Kuhn, *The structure of scientific revolutions* (Chicago: University of Chicago Press, 1962).

speculation necessary. This pushes us to the fringes of science, the realm of 'mavericks' and the unconventional – often a hostile and uncomfortable place.

Scientists value consensus, at least about the big issues. Catastrophic risk is both a big issue and a highly charged one: so fringe-dwellers may be doubly unwelcome. Do we need to make special efforts to protect our mavericks, if catastrophic risk is to get the attention it deserves? If so, how can we do it? Can we use the values of science to protect useful fringe-dwellers from science's own immune system? Can we engineer a Maverick Room?

The workshop involved a number of leading philosophers of science, including Heather Douglas (Michigan State), whose work on the intersection of epistemic risk and value in science had been one of the inspirations for the project. It also included some speakers we called our mavericks – researchers who felt that they had encountered these reputational issues in their own work. They spoke about their own experience in controversial fields such as nanotechnology risk, AI risk, geoengineering, and so-called 'cold fusion'.

The last example was particularly interesting, from my point of view. Around the time plans for CSER were first emerging, I happened to become interested in claims then being made about cold fusion, or LENR ('low energy nuclear reactions'), as it was also termed. It is my fourth elephant.

Elephant 4: Cold fusion

At a press conference in Utah in 1989, the chemists Martin Fleischmann and Stanley Pons reported that they had detected excess heat, at levels far above anything attributable to chemical processes, in experiments involving the metal palladium, loaded with hydrogen by electrolysis. They concluded that the heat must be caused by a nuclear process – 'cold fusion', as it soon came to be termed.

The idea attracted world-wide interest, as well as scepticism. Many laboratories attempted to replicate Fleischmann and Pons' results, but most, if not all, failed to do so. Within a year or so, the mainstream view was that cold fusion had been 'debunked'. It soon came to be treated as a classic example of pseudoscience, or pathological science.

It never went away completely, however. It retained a few defenders, including a handful of scientists at very respectable laboratories. These researchers acknowledged that replication and reproducibility were difficult, but claimed that most attempts on which the initial dismissal had been based were simply too hasty. They claimed, and still claim, that the Fleischmann and Pons results were eventually confirmed.

For my part, I kept an eye on cold fusion for a few years after 1989, but it dropped off my radar. Late in 2011, a remark by a physicist friend on Facebook happened to make it visible again. With the issues motivating CSER now in my head, I was struck by the parallels. I have followed the field ever since, writing several public pieces about it, and meeting many of its leading figures, inside and outside academia.[11]

I regard cold fusion as a fascinating real-world example of maverick science, in the sense relevant to the study of extreme technological risks. Indeed, I see my own experience in thinking and writing about the field – in particular, some of the reactions I have encountered from others – as an interesting illustration of some of the general characteristics of maverick science.

The crucial consideration is the same as that for extreme risks. The more disastrous a potential failure, the more improbable it needs to be before we can safely ignore it. And a disaster may be missing

[11] See Huw Price, "Risk and scientific reputation: lessons from cold fusion", in *Managing Extreme Technological Risk,* ed. Catherine Rhodes (Singapore: World Scientific), 25–85. This survey article includes the text of several of my earlier public pieces. Available at: arxiv.org/abs/2201.03776

something good, not meeting something bad. For hungry sailors, as I put it in one of my public pieces, missing a passing island can be just as deadly as hitting an iceberg. So the same principle of prudence applies. The more we need something, the more important it is to explore places we might find it, even if they seem improbable.

We desperately need new alternatives to fossil fuels. So we should be keeping a sharp eye out for potential new energy sources, even in unlikely corners. Cold fusion is such a corner, in plain sight since 1989. Yet mainstream scientists have been strongly discouraged from looking at it, let alone spending their time exploring it. Anybody who took cold fusion seriously, even to the extent of suggesting that it might be worth a second look, put their own reputation at risk. (Some commentators wondered what my writing on the subject would do to my reputation.)

So the field has been stuck in what I called a *reputation trap*. I argued that this trap was itself irrational and pathological, given the potential cost of a false negative. Far from shunning cold fusion, we should be arranging the incentive structure in science to support it, and to encourage a few clever people to take a second look. I suggested an X-Prize.

It turned out that when I first wrote publicly about these issues there was already a Google-funded research program in the field, motivated by very similar considerations. Two Google engineers, Ross Koningstein and David Fork, had concluded that known technologies were not sufficient to decarbonise our energy system.[12] So we needed to explore new options, and shouldn't allow accidents of reputation to stand in the way of doing so. Besides, they realised, such sociological factors can change very rapidly, especially with a reputation such as Google's to tip the scales on the positive side.

This Google-funded program involved several major labs, at high-profile institutions. It was announced in 2019, with, to many people's surprise, a paper in *Nature*.[13] *Nature* had been highly critical of cold fusion in 1989 – more on that below.

The program's scientific results were mainly negative, although they suggested avenues for further work, some of which still continues. Its reputational results, on the other hand, were immediate and strongly positive. Major labs can now advertise openly that they work in the field. There are now two major EU-funded multi-institution grants, both of which cited the Google interest in their initial public statements. And the US Department of Energy's Advanced Research Projects Agency-Energy (ARPA-E) has recently awarded $10m in funding to several multi-institution teams.

Perhaps most intriguingly, the Japanese company Clean Planet has announced a collaboration with one of Japan's largest manufacturers of industrial boilers, to develop prototype LENR-based boilers for industrial purposes.[14] Clean Planet also has funding from the Mitsubishi Corporation. This news has attracted some attention in the Japanese business press,[15] but almost none in the West. The contrast with coverage of small reported advances towards hot fusion is particularly striking. It is an indication that the field is still treated with great suspicion, in some quarters.

---

[12] Ross Koningstein and David Fork, "What It Would Really Take to Reverse Climate Change", *IEEE Spectrum,* 18 November 2014. https://spectrum.ieee.org/what-it-would-really-take-to-reverse-climate-change

[13] Curtis P. Berlinguette, *et al.* "Revisiting the Cold Case of Cold Fusion", *Nature* 570, 2019, 45–51. https://doi.org/10.1038/s41586-019-1256-6

[14] Miura's own announcement also mentions Google's interest in the field; see "MIURA CO. LTD. and Clean Planet Inc. Conclude an Agreement for Joint Development of Industrial Boilers That Use Quantum Hydrogen Energy", 28 September 2021. Accessed September 4, 2023, https://www.miuraz.co.jp/news/newsrelease/2021/1132.php.

[15] See, e.g., Kazuya Hatakeyama, "What is 'QHe', a Next-generation Energy Technology from Japan That is Not Only Nuclear Fusion?" *Forbes Japan,* August 3, 2023. Translation accessed September 4, 2023, https://forbesjapan-com.translate.goog/articles/detail/64933?_x_tr_sl=ja&_x_tr_tl=en&_x_tr_hl=fi&_x_tr_pto=wapp

It may be premature to call cold fusion a black elephant, but it does have some of the relevant characteristics, even at this point. It illustrates how easily something interesting can simply go unseen, because of prevailing disincentives for looking in its direction. It also shows how deliberate work in 'reputational engineering' can do something to reduce such disincentives.

Elephant ranking

How do these four elephants compare? I want to rank them on several dimensions, starting with my own degree of confidence that there is 'something there', in each case – something that many other observers are or have been missing.

In the case of AI and cold fusion, we need to clarify what this means. These cases involve risk – a probability, perhaps small, of a great harm or of missing a great benefit. What has been overlooked, in my view, is the need to study the area, to reduce that risk. I'm much more confident about that need, obviously, than I am about the eventual results of such a study. That's how risk works. So it's my confidence about the importance of the study that matters for this ranking.

With this qualification, my personal 'something there' ranking goes like this, in order of decreasing confidence: (1) Conscription and the monarchy; (2) AI risk; (3) Retrocausality in QM; and (4) LENR. My confidence is high for all of them, of course, or they wouldn't be contenders for my personal elephant herd.

So much for my confidence in 'being in the room', in each case. What about potential impact, which we might think of as a measure of blackness, in the original intended sense? Here AI and LENR are the potential high scorers, in that order. LENR might have a huge impact, if it turns out to be both real and useful. But it seems unlikely to match the game-changing impact of AI, let alone the game-ending impact of existential consequences of AI, if such there be.

The potential impact of the remaining two elephants is obviously much smaller, and it is hard to rank them against each other – they are such different beasts. Abolishing the monarchy would correct a significant injustice, but the number of people directly affected would be very small. The impact of retrocausality in QM seems likely to be purely theoretical, unless it somehow leads to new physics or applications, in a way we cannot presently foresee. Still, even theoretical impacts can be significant, to those who care about the issues in question.

A slightly different ranking goes in terms of urgency. Here it seems to me that AI and LENR are again the clear leaders, because their potential impacts are large and comparatively short term. But the monarchy now has a clear lead over retrocausality. The injustice of the monarchy may fall on few shoulders, but they are real shoulders, already carrying the burden of their constrained childhoods. The time to save them is short. If the truth about QM had to wait another generation or two, would that matter?

Finally, what about the costs of false negatives – of continuing to ignore the elephant in question? Here impact and urgency trump probability, of course. This means that the challenge of finding ways to talk about elephants is particularly pressing in the urgent, high impact cases. No surprises there, but it is worth thinking more about the challenges of hearing the mavericks, and about Peter Ho's advice.

The attention turntable

Let's now imagine our elephants on a large rotating turntable, well illuminated only at the centre. To get into the spotlight, an elephant needs to move inwards, overcoming the forces that tend to push it towards the fringes, or throw it off the turntable altogether. Which way it moves is thus determined by a balance of factors, centrifugal and centripetal.

My personal herd gives us some examples of how this works. In the case of LENR, the Google program is a textbook example of successful centripetal pressure, in my view. Much of its impact lay in countering the reputational factors that otherwise pushed in the other direction.

In the case of AI risk, where the recent shift towards the centre has been especially dramatic, it would be difficult to disentangle the various centripetal forces that have been at work. But our own work in Cambridge may have played some part, especially in the early stages. Certainly we were aware of the need for this kind of pressure. In one of the early discussions that led to CSER, I remarked to Martin Rees that some of the issues we wanted to study had a poor reputation. They were regarded as 'a bit flakey', as I put it. Rees agreed, but said that that was why the project was important. Serious risks might not be getting the attention they deserved, because of these reputational factors.

We were clear that a useful role for CSER would be to act as a reputational counterweight. CSER could use the reputation that we ourselves had at hand – that of Cambridge, and of our distinguished supporters and collaborators – as an opposing force, to nudge these neglected issues away from the fringes, towards attention and respectability.

In addition to exerting centripetal pressure, it can also be helpful to weaken the forces that push in the other direction. I have tried to do this by calling attention to the pathological nature of some of these pressures, in cases with a particular risk profile – where a false negative would be especially disastrous.

One of the challenges in these cases is that the fundamental point about risk is not sufficiently well known. The point itself is simple. The greater the harmful consequences, the more improbable a risk needs to be, before we can responsibly ignore it. This formulation, stressing the sense of responsibility, is from the philosopher Heather Douglas.

> In general, if there is widely recognized uncertainty and thus a significant chance of error, we hold people responsible for considering the consequences of error as part of their decision-making process. Although the error rates may be the same in two contexts, if the consequences of error are serious in one case and trivial in the other, we expect decisions to be different. Thus the emergency room avoids as much as possible any false negatives with respect to potential heart attack victims, accepting a very high rate of false positives in the process. … In contrast, the justice system attempts to avoid false positives, accepting some rate of false negatives in the process. Even in less institutional settings, we expect people to consider the consequences of error, hence the existence of reckless endangerment or reckless driving charges.[16]

Douglas goes on to discuss the possibility that '[w]e might decide to isolate scientists from having to think about the consequences of their errors', but rejects it. She argues that 'we want to hold scientists to the same standards as everyone else', and therefore 'that scientists should think about the potential consequences of error.'

As far as I know, however, there isn't a catchy way to express Douglas's point, already on the tip of educated tongues. By way of comparison, most educated people know the phrase 'correlation is not causation'. They thus have at least some warning of the errors that can flow from ignoring it, and can be held accountable accordingly. The phrase itself is often enough to call attention to the error.

[16] Heather Douglas, "Rejecting the Ideal of Value-Free Science", in *Value-Free Science? Ideals and Illusions*, ed. Harald Kincaid, John Dupré, and Alison Wylie (Oxford: Oxford University Press, 2007), 124.

In contrast, the errors that flow from ignoring our safety principle seem to need careful explanation, case by case. I want to close by offering two examples to back up this claim, from the same influential actor: the journal *Nature.*

Nature behaving badly

As I have been writing this piece, in the (northern) summer of 2023, the question whether AI might pose existential risks to humanity has been on some remarkable tables – tables in the White House and 10 Downing Street, amongst other places. In response, some critics have argued that it is getting too much attention. They want to push it aside, or at least into the distant future, in favour of conversations about the immediate risks of AI.

These critics now include *Nature.* In a recent editorial, *Nature* urges us: "Stop talking about tomorrow's AI doomsday when AI poses risks today."[17] The piece concludes: "Fearmongering narratives about existential risks are not constructive. Serious discussion about actual risks, and action to contain them, are."

These concerns about AI have been raised by some 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 extremely 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 my view. I hope that by taking *Nature* to task on this point, we can increase general awareness of the underlying message – the need for special care, when the cost of error is so high.[18]

Remarkably, there's a very close parallel here involving another of my elephants. In my view, a blameworthy feature of the cold fusion case, if LENR becomes mainstream – indeed, I think it is blameworthy, *whether or not* it becomes mainstream – is the apparent failure of many of its critics to take the cost of a false negative into account. It is too soon to judge whether this has made any practical difference. We don't know whether LENR will turn out to be a useful energy source. Even if so, it will be difficult to estimate how long a delay the treatment of the field might have caused.

But these unknowns are in one sense irrelevant. We don't excuse lax safety practices just because a disaster fails to happen as a result. Think of Douglas's examples – reckless endangerment and reckless driving. An actor may be guilty of these things even if, by good fortune, they fail to harm anyone. I think there's a case for charging some of the institutions of science with reckless endangerment, or something on that spectrum, in the case of LENR.

---

[17] "Stop Talking about Tomorrow's AI Doomsday when AI Poses Risks Today", *Nature,* 618 (2023), 885-886. https://doi.org/10.1038/d41586-023-02094-7

[18] I am presently writing about these issues with the new Director of CSER, Professor Matt Connolly, to whom I am grateful at this point; see Huw Price and Matthew Connolly, "*Nature* and the Machines", preprint, 23 July 2023. https://arxiv.org/abs/2308.04440v1; and Huw Price, "Big bang, low bar – risk assessment in the public arena", *Royal Society Open Science* 2024 https://doi.org/10.1098/rsos.231583.

And this is particularly true of the editors of *Nature,* in my view (hence the parallel with the AI case). Melinda Baldwin's recent history of *Nature* gives this account of the journal's role in the treatment of cold fusion in 1989–90.

> Instead of being the forum where a new era of energy was declared, *Nature* quickly became a major center of cold fusion skepticism. By 29 March 1990, a year to the week after the first mention of cold fusion in *Nature,* [the Editor, John Maddox] felt secure enough to declare "Farewell (Not Fond) to Cold Fusion" in the magazine's leader.[19]

As Baldwin goes on to say:

> During the cold fusion controversy, Maddox ... and the rest of the editorial staff cast the cold fusion episode as a battle between careful, peer-reviewed, properly conducted science and sloppy science revealed through press conferences in hopes of wealth through patents. Maddox wrote editorials criticizing Pons and Fleischmann's methods, associate editor David Lindley wrote news articles forecasting the death of cold fusion, and the journal's editorial staff gave significant space to cold fusion's most prominent scientific critics. Where *Nature* led, science reporters followed. News outlets such as *Time,* the *Economist,* and the *Wall Street Journal* all covered *Nature's* role in the cold fusion controversy and portrayed the journal's skepticism as proof that the scientific community was rejecting the Pons-Fleischmann claims. Ultimately, the cold fusion episode convinced many observers of the scientific journal's continued importance to the scientific community and illustrated *Nature's* influence among both scientists and laymen at the end of the twentieth century.[20]

One can imagine a different role that *Nature* might have chosen to play, which – while still emphasising the importance of 'careful, peer-reviewed, properly conducted science' – also stressed the very high potential cost of a false negative, in this particular case. It must have been evident to *Nature's* editors that there were many groups whose interests would be threatened by cold fusion – anyone with a vested interest in any other sort of energy system, present or future, for a start. All the more reason, then, to adopt a prudential approach, protecting the potential candle flame from the risk that it might be snuffed out prematurely.

As I say, *Nature* could have used its 'influence among … scientists and laymen' to recommend such caution. While deploring science by press conference, it could have preached the risks of hasty dismissal, in a case in which so much was at stake. It could also have called attention to the powerful interests that might see themselves threatened by such work, and stressed the need for scientific neutrality. Instead, as Baldwin says, it became 'a major center of cold fusion skepticism.' Would the house journal of the fossil fuel industry have behaved any differently?

In my view, *Nature's* actions in this case – its failure to take proper account of the risk of a false negative – amount to reckless endangerment, or something on that spectrum. For very good reasons, *Nature* is one of the most respected institutions in contemporary science. We are entitled to expect it to exercise its authority with care and responsibility.

My criticism of *Nature* here has been stronger than for the recent AI case for two reasons. First, it is easier to learn from mistakes at a historical distance, when the main perpetrators are no longer in the

[19] Melinda Baldwin. *Making Nature: the History of a Scientific Journal.* (University of Chicago Press, 2015), 201.
[20] Baldwin, *Making Nature,* 201–202.

room. Second, the action in question is different in scale in the two instances – a sustained campaign over many months, versus a single skirmish. Still, the fundamental failing is the same in both cases: a reckless blindness, from a respected and influential actor, to the potential costs of error.

Life among the elephants

I'm not sure how I acquired my taste for unpopular views. I take comfort from the fact that it must be a discriminating taste, for otherwise my herd of elephants would be very much larger. But happily it has not been so discriminating as to deprive me of the opportunity of meeting many extraordinary individuals, somehow associated with one elephant or another, or with the study of the species. One of these individuals is Peter Ho, and I am especially delighted to have met Peter and many of his former colleagues in Singapore. A particular highlight was my visit to Singapore in July 2019, for the biennial Foresight Conference, and the tenth anniversary celebrations of the Centre for Strategic Futures (CSF).

When I retired as Bertrand Russell Professor in Cambridge the following year, I mentioned this visit in a valedictory article in the Faculty of Philosophy Newsletter.[21] The piece includes a photograph taken at the CSF anniversary dinner. I'm standing at a high window with Ms Liana Tang, then Deputy Head of CSF, with the Marina Bay Sands visible in the distance. I'm pointing out a small hotel in Beach Road – remarkably, still in existence – in which I stayed in 1976, on my way to Cambridge as a graduate student. I note that by tying together these two visits to Singapore, more than forty years apart, this picture links two landmarks in my long association with Cambridge Philosophy.

I can now add that when the picture was taken, there were black elephants in the room, in the form of the cupcakes I mentioned earlier. Peter Ho was also in the room, in the form of himself. The recollection of that happy occasion seems an appropriate way to thank Peter once again, not least for giving me the metaphor to describe so many of my unusual intellectual interests, on my fortunate trajectory between those two visits to Singapore.

References

Baldwin, Melinda. *Making Nature: The History of a Scientific Journal.* Chicago: University of Chicago Press, 2015.

Bell, John Stewart. "Introductory Remarks." *Physics Reports,* 137(1986), 7–9. https://doi.org/10.1016/0370-1573(86)90065-7

Berlinguette, Curtis P., YM Chiang, JN Munday *et al.* "Revisiting the Cold Case of Cold Fusion." *Nature* 570, 2019, 45–51. https://doi.org/10.1038/s41586-019-1256-6

Douglas, Heather. "Rejecting the Ideal of Value-Free Science." In *Value-Free Science? Ideals and Illusions*, edited by Harald Kincaid, John Dupré, and Alison Wylie, 120–141. Oxford: Oxford University Press, 2007.

Hatakeyama, Kazuya. "What is 'QHe', a Next-generation Energy Technology from Japan That is Not Only Nuclear Fusion?" *Forbes Japan,* August 3, 2023. Translation accessed September 4, 2023, https://forbesjapan-com.translate.goog/articles/detail/64933?_x_tr_sl=ja&_x_tr_tl=en&_x_tr_hl=fi&_x_tr_pto=wapp

[21] Huw Price, "Two Projects from 2012 – a Progress Report", in *Philosophy at Cambridge: Newsletter of the Faculty of Philosophy,* 16 (September 2020), 4–5. Accessed September 7, 2023. https://drive.google.com/file/d/1o7eHgOUaPJpSbAS7evxKKIlCQxANlSKi/view

Koningstein, Ross and David Fork. "What It Would Really Take to Reverse Climate Change." *IEEE Spectrum,* 18 November 2014. https://spectrum.ieee.org/what-it-would-really-take-to-reverse-climate-change

Kuhn, Thomas. *The structure of scientific revolutions.* Chicago: University of Chicago Press, 1962.

Miura Co. Ltd. "MIURA CO. LTD. and Clean Planet Inc. Conclude an Agreement for Joint Development of Industrial Boilers That Use Quantum Hydrogen Energy." 28 September 2021. Accessed September 4. 2023, https://www.miuraz.co.jp/news/newsrelease/2021/1132.php.

Nature editors. "Stop Talking about Tomorrow's AI Doomsday when AI Poses Risks Today." *Nature,* 618 (2023), 885-886. https://doi.org/10.1038/d41586-023-02094-7

Piper, Kelsey. "The Case for Taking AI Seriously as a Threat to Humanity." *Vox,* October 15, 2020. https://www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignmentPrice, Huw. "Time for Some Royal Prerogative – Let's Give Kate's Child a Choice." *The Conversation,* January 29, 2013. https://theconversation.com/time-for-some-royal-prerogative-lets-give-kates-child-a-choice-11518

Price, Huw. "It's a Boy – But Baby Cambridge Deserves Choices in Life." *The Conversation,* July 23, 2013. https://theconversation.com/its-a-boy-but-baby-cambridge-deserves-choices-in-life-16154

Price, Huw. "Where Would We Be without Counterfactuals?" In *New Directions in the Philosophy of Science*, edited by Maria Carla Galavotti, Dennis Dieks, Wenceslao J. Gonzalez, Stephan Hartmann, Thomas Uebel and Marcel Weber, 589–607. New York: Springer Cham, 2014.

Price, Huw. 'Why do scientists dismiss the possibility of cold fusion?', *Aeon,* 21 December 2015. https://aeon.co/essays/why-do-scientists-dismiss-the-possibility-of-cold-fusion

Price, Huw. 'Icebergs In the room? Cold fusion at 30', *3 Quarks Daily,* 4 March 2019. https://3quarksdaily.com/3quarksdaily/2019/03/icebergs-in-the-room-cold-fusion-at-thirty.html

Price, Huw. "Two Projects from 2012 – a Progress Report." In *Philosophy at Cambridge: Newsletter of the Faculty of Philosophy,* No. 16 (September 2020), 4–5. Accessed September 7, 2023. https://drive.google.com/file/d/1o7eHgOUaPJpSbAS7evxKKIlCQxANlSKi/view

Price, Huw. "Risk and Scientific Reputation: Lessons from Cold Fusion." In *Managing Extreme Technological Risk*, edited by Catherine Rhodes, 25–85. Singapore: World Scientific. Available online at https://doi.org/10.48550/arXiv.2201.03776

Price, Huw and Matthew Connolly. "*Nature* and the Machines." Preprint, 23 July 2023. https://doi.org/10.48550/arXiv.2308.04440v1

Price, Huw. "Big bang, low bar – risk assessment in the public arena." *Royal Society Open Science* 2024 https://doi.org/10.1098/rsos.231583.

Price, Huw and Ken Wharton. "Taming the Quantum Spooks." *Aeon,* 14 September 2016, https://aeon.co/essays/can-retrocausality-solve-the-puzzle-of-action-at-a-distance.

Price, Huw and Ken Wharton. "Untangling Entanglement." *Aeon,* June 29, 2023, https://aeon.co/essays/our-simple-magic-free-recipe-for-quantum-entanglement.

Russell, Bertrand. "On the Notion of Cause." *Proceedings of the Aristotelian Society, New Series* 13 (1913), 1–26.