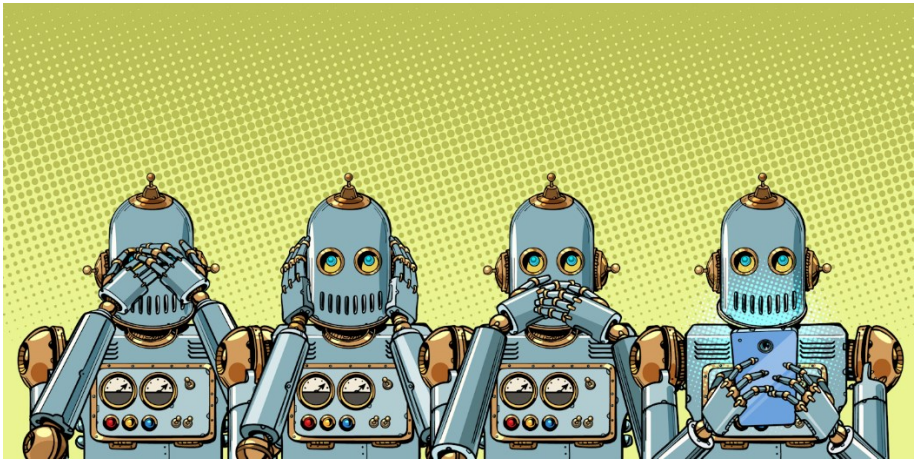
If We Can't Define "safe", What Chance Does AI Have?



John Ridgway discusses the challenges of designing autonomous Artificial Intelligence decision-making systems to be "sufficiently safe". As "sufficiency" requires a value judgement in the safety-engineering domain, he suggests that it is more "artificial wisdom" that is necessary, which will require AI to operate within a suitably encoded ethical framework. But how would such a framework be encoded, on whose value system would it be based and how much should emotion play in decision making? John explores these questions and provides some insights into tackling these issues.

It's All About Values

I don't think I have to remind any of the readers of this newsletter that safety is a value-laden concept. The roots of this problem are essentially epistemic. Such is the nature of uncertainty – that we cannot say with one hundred percent confidence that any system or situation will prove to be safe. Indeed, safety engineers learn at a very early stage in their professional development that the word 'safe' should always be accompanied by the modifier 'sufficiently'. We accept that sufficiency is all we can achieve, and within that acceptance can be found the value judgements. These judgements take many forms and are an expression of intelligence; or, to be more precise, wisdom. In fact, one can define wisdom as the judicious application of limited knowledge, as constrained by an advantageous ethical framing. So when thinking about the safety of AI, my first question would have to be this: How can we ensure that any machinery designed to make autonomous decisions does so within a suitably encoded ethical framework? Or, to widen that question, even when AI is only playing a supporting role in human decision-making, how can we stop AI from facilitating unethical intentions?

Those who foresee a problematic future often assume that the critical point will only be reached once AI has achieved a level of intelligence that surpasses our own. Such concerns may make for a good science fiction movie, but I believe superior intelligence is not really the main issue, and for two reasons.



Firstly, there is what I would call the baby-sitter syndrome. When we place another intelligent agent in charge of something we hold precious, we do so on the assumption that the surrogate shares our values. And just to make sure, we instruct them regarding those values, using natural language. The issue is whether or not such instructions are an effective encoding of the required values and whether they have been understood. We don't normally become overly troubled if the baby-sitter has superior intelligence to our own, as long as that intelligence is used to understand the imperative for respecting the terms and conditions of their employment. The dynamics of the problem are no different with artificial 'baby-sitters', but the practicalities are fundamentally different, not least because the use of natural language as machine instruction to encode ethical values (such as respect) remains a technical challenge.

Secondly, as with all system safety issues, the problems do not lie solely within the inherent properties of the system, but rather within the relationship existing between the system and both its operational environment and its human operators. In particular, safety risk often lies in the gap lurking between how reliable the system is and how reliable it is assumed to be. That gap can exist at any level of system intelligence. One can argue that such problems already exist at the human/AI interface and can be discerned in the degree to which we are often unable to even reliably identify the workings of AI. And even when we are fully aware, we have a tendency to anthropomorphize and thereby bestow upon AI capabilities and sensibilities that it doesn't actually possess. Furthermore, AI drives expert systems and we all feel a lot happier when we see the trappings of expertise – it's as if AI wears a white coat and has a stethoscope draped across its shoulders. So it is not the sophistication of AI that is the problem so much as human gullibility and complacency. Artificial Intelligence had the capacity to make fools of us long before it could possibly gain the motivation and skill to do so. It may also be worth keeping in mind that AI is already a tool of human rivalry.

The Machine Intelligence Research Institute

It is at this stage that I think it is only fair that I admit, when it comes to talking about AI, I too have the capacity to deceive. I am no expert but I flatter myself that I have the guile and sophistication to pass myself off as one. I became aware of this 'talent' some years ago when I was approached by Luke Muehlhauser of the Machine Intelligence Research Institute (MIRI) to take part in an interview [1] that would be posted on their website. He had become aware of my work writing for the SCSC Newsletter and had, from that, deduced that I had expert credibility. Indeed, he initially referred to me as Dr Ridgway (I assure you that I have no such credentials; I dropped out of a PhD in theoretical nuclear physics the day after it occurred to me that it was quite a difficult subject). Flattered and alarmed in equal measure, I obliged to take part but was very careful to make as my opening statement, *"I am not an expert and I would not wish to pass myself off as one".* The rest of the interview passed amiably enough but I have to say that if Luke had hoped to gain anything that would help

his institute better understand the challenges of AI safety, he must have been sadly disappointed. Indeed, I suspect that the real legacy was my own heightened interest in the works of MIRI.

If you visit the MIRI website [2] you will see that it is a charity announcing itself with the following statement:

"MIRI's mission is to ensure that the creation of smarter-than-human intelligence has a positive impact. We aim to make advanced intelligent systems behave as we intend even in the absence of immediate human supervision."

They continue:

"MIRI focuses on AI approaches that can be made transparent (e.g., precisely specified decision algorithms, not genetic algorithms), so that humans can understand why AI systems behave as they do. For safety purposes, a mathematical equation defining general intelligence is more desirable than an impressive but poorly-understood code kludge".

Anyone with a safety engineering background would readily recognise the need for transparency. Nevertheless, the idea of "*a mathematical equation defining general intelligence*" is intriguing to say the very least. Those wishing to learn more about MIRI are invited to visit its website, but for the purposes of this article I would like to draw your attention to one MIRI paper in particular, written by professors Nick Bostrom and Eliezer Yudkowsky, and titled '*The Ethics of Artificial Intelligence*' [3].

Establishing An Ethical Framework

The MIRI paper concentrates upon the characteristics one would require of any AI algorithms that are used to take over social functions, and draws the conclusion that responsibility, transparency, auditability, incorruptibility and predictability are primary amongst them. Astutely, the authors acknowledge that "*this list of criteria is by no means exhaustive*". Which is just as well because, although it is a similar list to that required for any safety-related system, it does not appear to address the need for a clearly specified, algorithm-based policy to inform risk-based decision-making. For example, from where would the AI algorithm derive its bases for the minimisation and prioritisation of risk, cost benefit calculations, or the transfer of risk between stakeholders? These are all value-based considerations that would require more than the number-crunching wizardry of an idiot savant.

We don't just need "*a mathematical equation defining general intelligence*"; we need a mathematical equation defining general wisdom.

Delving further into the problems of predictability, the paper refers to the concept of Artificial General Intelligence (AGI) and sees within that concept a clear technical difficulty with the safety of such systems. As the paper puts it:"

"It is relatively easy to envisage the sort of safety issues that may result from AI operating only within a specific domain. It is a qualitatively different class of problem to handle an AGI operating across many novel contexts that cannot be predicted in advance."

Indeed it is a qualitatively different class of problem, but the paper is still not focusing upon the more intractable problems that arise when one considers the ethical dimensions of decision-making – the demands of generality will just make those problems worse. This is because, to achieve generality, an AI algorithm will need to be adaptable both in comprehension and purpose without transgressing an overarching ethicality. In other words, it is required to exhibit what I would call 'Artificial Wisdom'. As a consequence, we don't just need *"a mathematical equation defining general intelligence";* we need a mathematical equation defining general wisdom.



But where is this wisdom to come from? Are safety engineers to play God and fashion AI systems in their own image? Do we ourselves have a sufficiently good track record in meeting expectations of sufficient safety? Can we even agree upon the definitions of relevant terms, such as 'reasonably practicable'? How good are we in dealing with novel situations in our profession? Have we got to grips with all of the moral dilemmas that are encompassed in safety engineering policy? And if we are not on top of all of this, how are we to expect autonomous AI to operate safely and wisely, given that its only steer will be coming from the way we have encoded within it our own moral basis for decision-making – assuming, of course, that such a machine encoding were to be possible?

The Nature Of Human Decision-Making

As a final concern, I would like to turn to another aspect of human decision-making that could present a challenge for those who would wish to replicate it in autonomous safety-related systems. It relates to the role of emotion and is encapsulated within a question asked by the neuroscientist, Robert Sapolsky: "*Do we mostly think or feel our way to deciding what is right?"* [4]



For a long time, philosophers and scholars of human cognition thought that emotion was a potential impediment to good decision-making. Plato, for example, used the metaphor of the charioteer struggling to keep two winged horses under control. One of the horses represented the emotional and impulsive side to human nature, whereas the charioteer represented the rational soul, seeking to override and control such impulses. Modern neuroscience, however, has led to an understanding that emotion is not so much a potential nuisance as it is an essential component in making decisions.

Neuroscientist Antonio Damasio reports [5] that a patient with damage to an area to the brain vital in the mediation of emotional states (i.e. the ventromedial frontal cortex) was not left, as one might expect, with the capacity to make purely rational decisions. Instead, he was left with an obsession with irrelevant detail and an inability to act decisively where it matters. His intelligence was not impaired but he acted fecklessly, as though lacking intelligence. Sapolsky reports that other patients with similar brain damage express opinions in laboratory experiments that are objectively rational but nonetheless demonstrate an alien moral compass; for example, it is better to sacrifice one's own family than a greater number of total strangers. Contact with one's emotionality, it seems, is one of wisdom's vital ingredients.

One can debate why this should be so, but one has to presume that there is evolutionary advantage to such an arrangement. Whether or not this provides a good argument for basing artificial intelligence upon a similar model, i.e. one with an interplay between artificial rationality and artificial emotion, is a question that I'm sure can be better answered by those more qualified than I. But I would like to raise the following possibility: If ethicality and morality are at least partly founded upon emotionality, then the possibility arises that a machine's instantiation of ethical decision-making may also have to employ some form of analogy to the role of emotion in human decision-making. And if this were to be necessary, how could it possibly be achieved? What would that do to the mathematical equation defining general wisdom?

If ethicality and morality are at least partly founded upon emotionality, then the possibility arises that a machine's instantiation of ethical decision-making may also have to employ some form of analogy to the role of emotion in human decision-making.

Here's Hoping

I have written this article for no better reason than to share with you the musings of a non-expert. Artificial Intelligence is a highly technical subject deserving a better qualified appraisal. However, it is also a concept that invites a great deal of philosophical debate and conjecture. We are warned that the stakes are very high and that miscalculations at this stage may have very serious consequences. This seems plausible, if only because many safety applications have been mooted and, even without a dominance of AI, safety considerations present the engineer with many challenges that go well beyond the simple objective of ensuring system reliability. We can look forward to embracing AI and all of its advantages but we must be wary of the prospect of it embracing us. If we expect it to keep us sufficiently safe, then we probably need to do a lot more to determine what that even means.

References

[1] https://intelligence.org/2014/03/08/john-ridgway-on-safety-critical-systems/, accessed Nov 2023

[2] https://intelligence.org/research/, accessed Nov 2023

[3] https://intelligence.org/files/EthicsofAI.pdf, accessed Nov 2023

[4] "Behave: The Biology of Humans at Our Best and Worst", Chapter 13: "Morality and Doing the Right Thing, Once You've Figured Out What That Is", ISBN 9780-0-09957-506-1.

[5] "Descartes' Error", Antonio Damasio, ISBN 978-0-099-50164-0.

image attribution:
do no evil: Illustration 143452766 | Communication © Valeriy Kachaev | Dreamstime.com
robot holding baby: AI generated Midjourney
face: Illustration 289198092 © Golib Tolibov | Dreamstime.com
winged horses and chariot: AI generated Midjourney

John Ridgway, retired

John Ridgway is a retired traffic systems consultant who led the software development team responsible for the M25 MIDAS computer control system. Subsequently, he undertook a functional safety management role whilst working for a contractor commissioned to provide a computer control system for the Highways Agency's M42 Active Traffic Management system.