The Executive's Guide to Getting AI Wrong

It's all math. Really.

by Jerrold Soh

Unless you've been living on a deserted island, you've probably been told that Artificial Intelligence (AI) will 'disrupt' or 'revolutionise' your industry in some way or other. In this Fourth Industrial Revolution, livelihoods will be up-ended, old ways of working will go the way of the dinosaur, old revenue streams will shrivel, and new ones will emerge. So your organisation had better start planning for AI's impact on you, and start building AI into its key business units, processes, and workflows. It comes highly recommended that you do this under a consultant's expert guidance.

What does your mind's eye see when you hear this? Specifically, how does AI disrupting your industry *look* like? Unless you belong to the vast minority of decision-makers with specialised training in AI technology, your closest reference point is probably science fiction, especially of the Hollywood variety. Call this 'Hollywood-style AI': Marvel's J.A.R.V.I.S. (Just A Rather Very Intelligent System), Disney's Wall-E and, for sci-fi aficionados, HAL9000, Robocop, and Terminator. Perhaps you imagine one or all of these characters reporting to work one day, clad in metallic grey suits.

This article explores how we see AI and argues that we mostly get it wrong. In the process, it explains the reasons backed by social science research on why we tend to get AI wrong and illustrates the dangers of doing so from a managerial and law-making perspective. Some readers may also find the article useful as a guide on how and when to manipulate portrayals of AI in your favour.

GETTING AI WRONG

Hollywood-style AI systems are, almost without exception, instances of what philosopher John Searle classically termed 'strong AI': systems which think, act, and quack as humans do.[1] The only difference is that they are manufactured, not birthed. By contrast, 'weak AI' refers to systems programmed to do, and thus capable of doing, only specific tasks. Thus, they are also commonly known as 'narrow AI'. For example, you may be acquainted with basic statistical regression methods. The regression, you may be surprised to learn, is a kind of

narrow AI. The first three lessons of AI pioneer Andrew Ng's famous massively open online course on machine learning are devoted to linear and logistic regressions.[2] If you have taken a business or statistics 101 course that involves regression coursework, you might have trained AI without even knowing it.

Today, strong AI remains well in the realm of science fiction. Despite what Tesla or other 'AI companies' occasionally claim, no such system exists. Moreover, weak and strong AI are qualitatively quite different. There is no clear path to strong AI from the weak AI systems we have today; simply adding more and more computing power to a weak AI system does not make it strong. On the contrary, some AI researchers have argued that present methods which work for building weak AI positively cannot lead us to strong AI.[3] Buying even the most advanced, state-of-the-art AI software will probably not eventually lead to an army of pseudo-Terminators taking over your company.

Conflating the strong AI of the movies with the weak AI you are being sold is deeply problematic. At its heart, it is a category error,[4] like thinking potatoes are fruits, birds are planes, or smoking is good for you. In turn, these category errors lead one to carry misaligned expectations of what the software can do for you. The more one thinks of AI as 'basically human', the more one may start associating other human traits with the software, however (un)warranted. Expectations can be over-inflated, such as when one believes that the AI can autonomously identify and fix any problem you direct it to. They can also be understated, such as if one begins to think that the software would need to be given regular breaks and other employment benefits. Treating software as if it were human is both factually and functionally wrong.

The more one thinks of AI as 'basically human', the more one may start associating other human traits with the software, however (un)warranted.

WHY WE GET AI WRONG

The tendency to wrongly attribute humanity to AI, it turns out, is deeply human as well. It is so well-documented in the literature that it goes by different names in different fields. Oxford philosopher David Watson calls it 'AI anthropomorphism'.[5] Washington University law and computer professors Neil Richards and David Smart call it the 'android fallacy'.[6] Social psychologists have long termed the folk tendency to see in inanimate objects personalities, wants, and preferences, as a kind of 'dispositionism',[7] that is, to see a kind of internal disposition towards and against certain things. This is related to the equally well-documented phenomena of humans tending to see faces in everything from rocks to clouds and even toast.[8] The scientific name for this is 'pareidolia'. It happens within milliseconds,[9] in what Nobel Prize Laureate Daniel Kahneman and his colleague the late Amos Tversky might park under System 1 thinking.[10]

It is hardly surprising, then, that we are quick to see faces in AI. After all, 'Artificial Intelligence', read plainly, records humankind's best efforts at synthesising (human) intelligence. Thus, most definitions of AI incorporate some concept of a system that thinks or acts like us. Moreover, often AI makers do not leave pareidolia any work to do. They install AI into overtly humanoid forms. The most prominent example is Hanson Robotics' Sophia, a chatbot to which Saudi Arabia awarded citizenship,[12] which has been criticised as a publicity stunt meant to drum up hype and funding.[13]

Indeed, when it comes to seeing personality in AI, hardware may not be required at all. Just ask Jamie on your nearest government website.[14] Even within technical AI research, computer scientists have taken to using anthropomorphic metaphors like 'neurons', 'attention', and 'memory' to describe what they are building.[15]

WHAT'S WRONG WITH GETTING AI WRONG

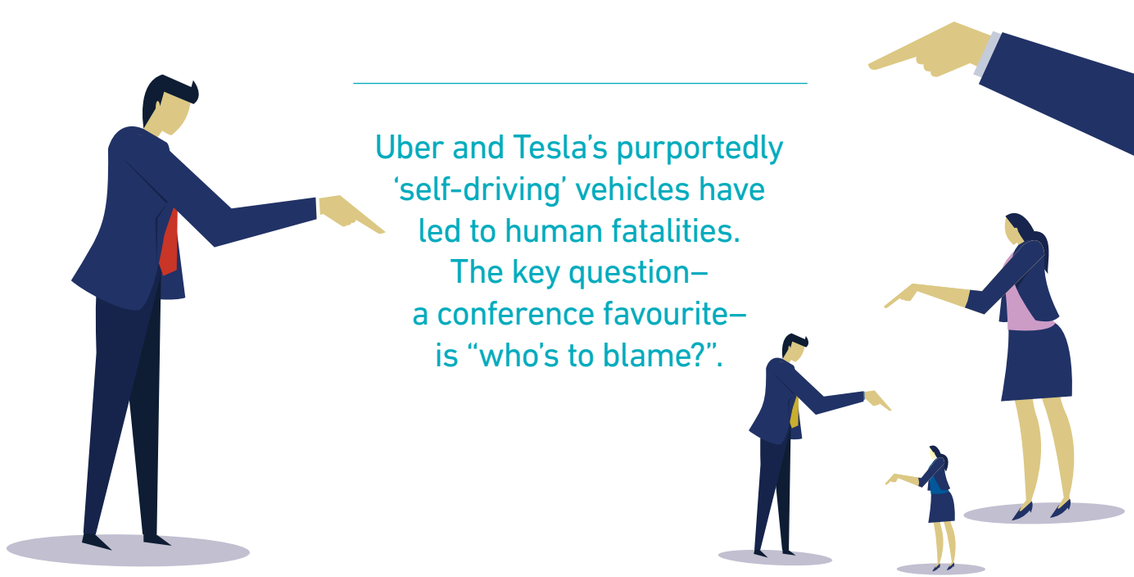
But why is seeing faces in AI a problem? It is difficult to object to this if we are talking about strong AI. However, today's weak AI systems are most often powered by machine learning (ML) and, contrary to its name, the focus of machine learning is not on any physical 'machine'. Nor does it fully approximate how humans actually learn. Rather, ML involves putting datasets (Excel sheets, if you will) through statistical algorithms–often a great many of them–to compute correlations and factor weights. At the risk of oversimplification, this is linear regression writ large. Seeing faces in Robocop or C3PO is one thing; seeing faces in ordinary least squares is quite another.

Whenever social scientists and lawyers identify instances of AI anthropomorphism, it is criticised. For instance, Watson calls such rhetoric "at best misleading and at worst downright dangerous".[16] More broadly, social psychologists argue that dispositionism leads us to commit a logical mistake so basic that it is simply called the 'fundamental attribution error'.[17] This refers to a fundamental bias we have towards attributing someone's (or something's) actions to its internal disposition, even when such behaviour may be mostly driven by its external circumstances. Applied to AI, reading too much into the software's apparent personality means we often mistakenly forget about those who have made the software dangerous to begin with: developers, operators, and possibly even users.

Consider in particular questions of moral blame and legal liability for AI-related harm. You would probably know by now that Uber and Tesla's purportedly 'self-driving' vehicles have led to human fatalities.[18] The key question–a conference favourite–is "who's to blame?". And, further, who should pay? Notice how the term 'self-driving' already implies that the *car* has some kind of Cartesian self that might (or should?) be responsible for the entire incident. Of course, cars don't have bank accounts. So it is easy for the parties involved to say, "blame the car, not me". This leads to the convenient result that no actual human or organisation is at fault, and no one has to pay. Victims are thus left to pick up the pieces.

This is, on quick reflection, hardly a satisfactory result.[19] The crux is that how strong we think the above argument is correlates almost perfectly with how strong we think the car's AI is. If a robotic Arnold Schwarzenegger had indeed been driving, the case is certainly arguable. But if the car's systems had been controlled by a linear regression, or perhaps even a more sophisticated arrangement of statistical algebra, one might probably do a double-take. Should it matter if the algebra had been named 'Harold', or that the company had painted a human face on the car's bonnet?

To illustrate the problem with statements like the "self-driving car caused the accident", consider these alternative examples: the pipe caused the leakage; the toaster burnt the toast; the piano fell out the window; and the gun killed the victim. Each of these statements might be *factually* and *grammatically* correct, but by making an inanimate object the subject of the sentence, we are gently guided towards blaming that object, not its makers and/or users. Because our attributions of moral and legal responsibility are intertwined with and influenced by our assessments of causality, this seemingly innocuous sentence construction that attributes causality to the object holds the power to shape what, and who, we blame for the harms 'it' apparently causes.



Uber and Tesla's purportedly 'self-driving' vehicles have led to human fatalities. The key question– a conference favourite– is "who's to blame?".

It has become fashionable, and likely profitable, for companies to hype up what their AI systems are capable of, in order to manipulate our inner pareidolia in their favour.

HOW TO MAKE PEOPLE GET AI WRONG

This leads us to a deeper, more concerning issue: our tendency to get AI wrong can easily be manipulated by those who want us to reach a certain conclusion. Where AI is concerned, we are particularly vulnerable to narrative manipulation for three reasons. First, few have formal training on what AI is. Second, our points of reference come from Hollywood and pop culture. Third, AI by definition, *tries* to act and look like us.

For these reasons, it has become fashionable, and likely profitable, for companies to hype up what their AI systems are capable of, in order to manipulate our inner pareidolia in their favour. In February 2022, OpenAI's Chief Scientist Ilya Sutskever tweeted that "[i]t may be that today's large neural networks are slightly conscious". Recall that neural networks are, in essence, a metaphorical description of what essentially are linear algebraic operations. (For those less familiar with vector math, picture computations across multiple Excel data columns.) Coming from a seemingly reputed institution, this comment was quickly picked up by tech blogs and news outlets. *Futurism* published an article titled 'OpenAI Chief Scientist Says Advanced AI May Already be Conscious'.[20] The *Daily Mail* ran an even more sensational headline 'Artificial Intelligence Expert Warns that There May Already be a "Slightly Conscious" AI out in the World'.[21]

In the eyes of AI experts, however, the claim that linear algebra might be even slightly conscious was strange, to say the least. Meta Chief AI Scientist Yann LeCun disagreed in a direct response to the tweet.[22] Criticism on Twitter and elsewhere was so forthcoming that the next day, there was enough material for *Futurism* to publish a follow-up piece entitled 'Researchers Furious over Claim that AI Is Already Conscious'.[23] What these researchers expressed ranged between (sarcastic) dismay at the idea of conscious algebra[24] and indignation at AI anthropomorphism being peddled once again[25].

But the damage has probably already been done. In an age of misinformation and press sensationalism, executives and corporate decision-makers are probably far more likely to read the initial, viral hype than see any subsequent, technical rebuttal. This is why the false story that Samsung paid Apple a billion dollars in five-cent coins still has its adherents.[26] A minute's reflection should have disabused one of this myth, since in many countries it is illegal to pay for anything with more than a set number of coins.[27]

GETTING THE LAW WRONG TOO

False narratives like this shape the path of the law far more than they should. To see how false AI narratives threaten policymaking around AI, let us first study the relatively simpler case of Liebeck v. McDonald's Restaurants in 1994, more widely known as the McDonald's 'Hot Coffee' case. Stella Liebeck was a 79-year-old woman in New Mexico, US, who had been driven by her grandson to a McDonald's drive-through.[28] She was served coffee at 190°F (88°C), 30 to 40 degrees higher than that adopted by other coffee vendors. While drinking it in a parked car, she spilled the coffee on herself. The coffee turned out to be so hot that she suffered third degree burns (the most severe kind) and nearly lost her life.

Liebeck demanded that McDonald's pay her medical bills of around US$20,000. McDonald's counter-offered US$800, so Liebeck sued the fast food giant. Evidence produced at the trial showed that McDonald's had over the past decade received about 700 reports of people being burnt by their coffee. Nothing had been done. The jury awarded Liebeck US$2.7 million in 'punitive damages', that is, damages meant to teach McDonald's a lesson. McDonald's appealed, and Liebeck eventually settled the case for less than US$500,000.

You might have heard of the case before. Only, the version you heard was based on a narrative spun by fast food (and other) companies in the wake of the jury's ruling. The story told was one of how selfish, greedy individuals had been filing frivolous lawsuits against helpless companies in a bid to win million-dollar jury awards, threatening the livelihoods of American businesses and their employees. As the website of the law firm which represented Liebeck explains, "once corporations gained control of the story, Stella Liebeck became a newly-minted millionaire grandmother, who got an easy payday".[29]

American corporations and their lawyers would spend years running a 'disinformation campaign' about this in order to lobby for laws to be enacted to protect businesses from a 'supposed epidemic of frivolous lawsuits'.[30] The news cycle happily amplified this narrative. As University of Oregon law professor Caroline Forell explains, "Twenty-six leading newspapers immediately announced that a woman had won a huge verdict against McDonald's for spilling coffee on herself. The headline for the AP story read 'Woman Burned by Hot McDonald's Coffee Gets $2.9 Million'. This pithy version of Liebeck's case was repeated over and over by the media."[31]

Having created a public outcry over the apparent problem of frivolous lawsuits, corporate America successfully persuaded the US Congress to pass laws limiting how much individual plaintiffs could recover from businesses through tort lawsuits.[32]

AI is quite different from coffee, but the present discourse and rhetoric over who should be responsible when AI systems 'burn' people follows a similar playbook to what we have seen with Liebeck v. McDonald's. We start by twisting facts to portray intentionality on one side and vulnerability on the other. Just as Liebeck was made to look like a greedy, self-interested coffee-spiller, AI systems are clothed with autonomy and self-determination. To the extent that anyone gets hurt, it is because *they* wanted it to be so, not anyone else. Meanwhile, the companies serving the coffee, or building the AI, plead that they are themselves victims of what the former intentionally or recklessly did.

Next, not knowing much about the subject (of either tort litigation or AI systems), the public easily buys into the narrative, not least because it is simulcast everywhere in the news. AI systems are particularly amenable to sensational headlines like those we have seen above, headlines which proudly declare them to be 'slightly conscious', evil, and soon to come for your job.[33]

This warped perception eventually percolates into public and policymaker support for laws and regulations meant to address problems which exist more in narrative fantasy than reality. Rather conveniently, these laws also happen to benefit the organisations responsible for spinning the narrative, particularly by shielding them from liability for any dangerous products they serve.

In this light, one wonders how many AI systems today have been, and are being, sold as 'slightly conscious' to would-be clients and/or funders. It is also clear that hyping up one's AI is not just good for the top line. This narrative helps companies avoid liability for what will invariably be described as 'the AI's' actions. To deflect responsibility for harm caused by the AI you made, sold, or used, draw everyone's attention to how autonomous and independent 'it' is. Conversely, if someone else's AI has harmed you, call it out for what it probably is: a linear (or logistic) regression writ large, or perhaps even a collection of if-else statements.

AI systems are particularly amenable to sensational headlines.

At this point, I should admit that when I was still in the business of selling AI to legal organisations, I often analogised the AI systems we were offering to "babies who knew very little about the world except the data we gave them". I knew this would help with sales, though I also knew it was an imperfect analogy. In my defence, whenever I offered to explain the (undergraduate level) math to stakeholders, I was mostly refused. Only once, I managed to take the client through a brief (one-hour) introduction to statistical learning. I was promptly told that I had wasted their time, as these academic technicalities were irrelevant to the project. Or, as Dr Teddy Oglethorpe tells Dr Randall Mindy in Netflix's 2021 film *Don't Look Up*, "Keep it simple. No math."

BUT IT'S ALL MATH

AI anthropomorphism sells, and given how our minds are wired, it is *easy* to sell. Few want to know the math anyway. Organisations and decision-makers want something that is 'turnkey' and can easily be 'leveraged to deliver synergistic value'. Considering all this, the problem should only persist, with the result that organisations continue to buy AI with over- and also under-stated expectations of what AI will do for (and to) them. So too should we expect legal and regulatory discussions to continue in the language of informal, anthropomorphic metaphors, rather than formal mathematics.

But for those who want better, a useful refrain to keep close to heart is that today's AI systems are mostly just math. Advanced and sophisticated math, sure, but nothing more than math. The next time someone tries to sell you AI, ask yourself if their math is really as strong, in the Searle sense, as they are making it out to be. Be wary of those who would appeal to your innate pareidolia. In documenting the impact that dispositionism has on us, Harvard law professor Jon Hanson speaks of Tom Hanks' character in the classic film *Cast Away*, who gets so deeply enamoured with 'Wilson', a volleyball with a face, that when 'Wilson' gets lost to the tides, it is a tearful moment not just for him, but for the audience as well. So even if you've been living on a deserted island, you probably cannot run away from seeing mysterious faces in AI. AMI

Jerrold Soh

is Assistant Professor of Law and Deputy Director of Centre for Computational Law at Singapore Management University

Endnotes

1 John R. Searle, "Minds, Brains, and Programs", Behavioral and Brain Sciences, 3, 417-457, 1980.

2 Coursera, "Machine Learning".

3 Herbert L. Roitblat, "Algorithms Are Not Enough: Creating General Artificial Intelligence", MIT Press, 2020; David Silver, Satinder Singh, Doina Precup, et al., "Reward is Enough", Artificial Intelligence, 2021.

4 Stanford Encyclopaedia of Philosophy, "Category Mistakes", 2019.

5 David Watson, "The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence", Minds & Machines, 29, 417-440, 2019.

6 Neil M. Richards and William D. Smart, "How Should the Law Think about Robots?" in "Robot Law", edited by Ryan Calo et al, Edgar Elgar, 2016.

7 Lee Ross, "The Intuitive Psychologist and His Shortcomings", Advances in Experimental Social Psychology, 10, 173-220, 1977.

8 Amalyah Hart, "Facing up to Ordinary Things", Cosmos, July 7, 2021.

9 Ibid.

10 Daniel Kahneman and Amos Tversky, "Thinking, Fast and Slow", Farrar Straus Giroux, 2011.

11 Stuart Russell and Peter Norvig, "Artificial Intelligence: A Modern Approach (4th edn)", Pearson, 2020.

12 Emily Reynolds, "The Agony of Sophia", Wired, June 1, 2018.

13 Jaden Urbi and MacKenzie Sigalos, "The Complicated Truth about Sophia the Robot", CNBC, June 5, 2018.

14 Government Technology Agency, "'Ask Jamie' Virtual Assistant".

15 See for example, Google, "Machine Learning Glossary".

16 David Watson, "The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence", Minds & Machines, 29, 417-440, 2019.

17 Lee Ross, "The Intuitive Psychologist and His Shortcomings", Advances in Experimental Social Psychology, 10, 173-220, 1977.

18 David Shephardson, "In Review of Fatal Arizona Crash, US Agency says Uber Software had Flaws", Reuters, November 6, 2019.

19 I have, in other work, criticised such reasoning in more detail. See Jerrold Soh, "Towards a Control-Centric Account of Tort Liability for Automated Vehicles", Torts Law Journal, 26, 1-34, 2021.

20 Noor Al-Sibai, "OpenAI Chief Scientist Says Advanced AI May Already be Conscious". Futurism, February 11, 2022.

21 Ryan Morrison, "'I'm Sorry, Dave. I'm Afraid I Can't Do That'", Daily Mail, February 11, 2022.

22 Yann LeCun, Twitter post, February 13, 2022, https://twitter.com/ylecun/status/1492604977260412928.

23 Noor Al-Sibai, "Researchers Furious over Claim that AI is Already Conscious", Futurism, February 12, 2022.

24 Melanie Mitchell, Twitter post, February 14, 2022, https://twitter.com/rajiinio/status/1493245179863597057.

25 Deb Raji, Twitter post, February 14, 2022, https://twitter.com/rajiinio/status/1493245179863597057.

26 "Fact Check – Samsung Did Not Pay Apple a $1-Billion Fine in Coins", Reuters, October 9, 2021.

27 The maximum coin limit under Singapore's Currency Act (Cap 69) is 20.

28 Allison Torres Burtka, "Lieback v. McDonald's: the Hot Coffee Case", American Museum of Tort Law.

29 Carlson Law Firm, "The Verdict: How the Hot Coffee Lawsuit Led to Tort Reform", September 10, 2020.

30 German Lopez, "What a Lot of People Get Wrong about the Infamous 1994 McDonald's Hot Coffee Lawsuit", Vox, December 16, 2016.

31 Caroline Forell, "McTorts: The Social and Legal Impact of McDonald's Role in Tort Suits", Loyola Consumer Law Review, 24(2), 2011.

32 Ibid.

33 See for example, Robert Wright, "Workplace Automation: How AI Is Coming for Your Job", Financial Times, September 29, 2019. A quick Google search will show that even reputed news outlets are not immune from repeating the AI anthropomorphism hype.