"Trustworthy AI": Ending the shouting match between utopians and Luddites

Published 11-Nov-2021 by
Stephen Scott, Karen Cook, Amy Edmondson and Thomas Malone

Artificial intelligence (AI) is being adopted more and more broadly, for a range of business use cases, triggering much discussion and debate about attendant risks and opportunities. A thoughtful discussion of any transformative technology is wholly desirable, but the dialogue has become more of a useless shouting match between utopians and Luddites, and the authors would like to reframe the discussion.

In its most recent Business and Finance Outlook, the Organisation for Economic Cooperation and Development (OECD) discussed the use of AI in business and finance, and the attendant opportunities, challenges and implications for policymakers. These forms of technology, the report said, are expected to yield advantages for firms by improving their efficiency through cost reduction and productivity enhancements. The report cautioned, however, that use of AI in business and finance may create, or intensify, both financial and non-financial risks, perhaps giving rise to consumer and investor protection concerns. As such, the OECD called on governments to give careful thought to a human-centric approach to "trustworthy AI".

The question is what this means. Academic and scientific rigour demands that we begin any argument by defining its terms, and to start with the initial thorny question about what AI is, it seems there is very little agreement on how best to define "artificial intelligence". Indeed, the authors think it better to refer to machine intelligence tools as "augmented intelligence", since this captures the way in which such tools function most typically today — as an aid to human intelligence.

I know what you're thinking …

It is more difficult still to define trustworthiness in the context of AI. In answer to the question, "how do we trust AI?", however, it is perhaps best to start with "how do we trust?"

By evolutionary design, we are inclined to trust others whom we believe to be like ourselves, assuming them to work from the same background knowledge, beliefs and values as we do in the course of perceiving, describing and making decisions about the world. This belief, that we can accurately infer what someone else is thinking, is referred to as having a " theory of mind".

There is no similar theory of mind for AI, however, Melanie Mitchell, a professor of complexity and computer science, has observed. We do not know what algorithms are "thinking", and we have no basis for inferring that machine intelligence systems "think" as we do. Indeed, we know that they do not. Yet AI is central to many products and services that are already in widespread regular use: GPS navigation; spam filters; credit card fraud alerts; loan applications; book, music and movie recommendations; disease diagnoses; and countless more.

Most often, we are resigned to "trusting" these tools, simply because giving them up is either too inconvenient or costly, but because we lack an explicit basis for that trust, we are left feeling uncomfortably vulnerable to unknown risks. That discomfort is compounded by a barrage of competing claims from "experts", alternately proclaiming that magical AI capabilities will usher in a techno-utopia of sorts or, conversely, warning that the relentless creep of AI into our lives will spell our ultimate undoing as a species.

Reframing the debate

Though it fills the headlines in the popular press, this shouting match between Utopians and Luddites produces much heat and little light. A cacophony of ill-informed media is unhelpful to those thinking through whether, when and how to make use of AI in the workplace.

Trust is perhaps the most widely discussed topic in the social sciences, with one paper tallying up 121 definitions of trust across 50 years of research. Most of those feature a common element: vulnerability. In choosing to trust, we voluntarily expose ourselves to the risk that we may be disappointed in doing so, leaving us vulnerable to some subsequent harm. Yet humans evolved such that we regularly place our trust with complete strangers, and the collaboration this permits has been key to our success as a species. So, how is it that we choose to trust?

Trust involves a mix of cognitive and affective components. Before making ourselves vulnerable — to a brain-surgeon, say — we want assurance that he or she is competent. This is insufficient, however: we also need to believe that he or she is reliable. Unreliable competence is not much better than reliable incompetence. To make these cognitive judgement calls, we want information that leaves us thinking that it is safe to trust in a someone in a specific given circumstance. We also want to feel that we can trust in this context, so we look for the additional assurance that someone is both honest and benevolent toward us — that they are genuinely working in our best interests. We trust, then, from both the head and the heart.

Surveillance creep

These same cognitive and affective components feature in decisions to trust something rather than someone, and so it is helpful to frame discussion of trustworthy AI along these dimensions.

A common complaint about the more widespread use of AI in the workplace is that it is invasive — a concern that has become more prominent amid COVID 19-driven work-from-home protocols. "The way my boss monitored me at home was creepy," said the headline from a recent BBC news article. "Data is the new frontline in workers' rights," the UK trade union for professional workers said in a recent article that ran in the Financial Times.

In the near term, expectations are that we will see a continued mix of in-office and at-home work arrangements. Arguments regarding the creep of intrusive spyware are only likely to become more shrill, therefore, even as employers maintain that they have no choice but to deploy greater surveillance and monitoring if they are to manage a remote workforce effectively.

Nowhere is this more likely than in the banking sector, where regulatory compliance requires that firms deploy robust surveillance and monitoring systems to safeguard against misconduct risk. These systems were designed to operate within the office workspace, but now firms are expected to operate systems that can monitor employees while in their homes. Moreover, some regulators tasked with safeguarding the public feel it necessary that they do likewise.

For instance, the UK Financial Conduct Authority (FCA) raised hackles earlier this year when it set forth its expectations for risk and compliance in the hybrid working environment.

"It's important that firms are prepared and take responsibility to ensure employees understand that the FCA has powers to visit any location where work is performed, business is carried out and employees are based (including residential addresses) for any regulatory purposes," the FCA said. It is entirely understandable that many employees object to this invasion of personal privacy.

Like bank risk and compliance officers, bank examiners and supervisors are very familiar with technology used in surveillance and monitoring. They are therefore well-placed to judge the competence and reliability of such tools. Those subject to surveillance — employees — are not equally well-positioned in this regard, however. Bank managers and supervisors may feel certain that they act with honesty and benevolence, but employees are more circumspect.

With limited ability to assure themselves that surveillance methods and monitoring tools are competent, reliable, honest or benevolent, it is perhaps no surprise that employees trust neither the tools nor those making use of them. "Trustworthy AI" is a community endeavor. If we want it to be used gladly in the workplace, then employees, managers, regulators — and third-party AI vendors — must be able to satisfy the four main tests of trust:

- Is it competent — does this tool work properly in this context?
- Is it reliable — can we count on this tool to work when we need it to?
- Is it honest — is the functioning and the intent of tool transparent?
- Is it benevolent — is the tool designed to operate in my best interests?

Panopticonfusing

"Regulatory technology (regtech) has assumed greater importance in response to the regulatory tightening and rising compliance costs following the 2008 global financial crisis," the IMF said in a just-released report on the use of AI in finance. These technologies promise to help firms to achieve significant cost savings and efficiency gains and better risk management, and offer powerful tools for regulatory compliance and prudential oversight, the IMF said.

The use of AI to identify culture and conduct-related risks is an area of regtech innovation that has received considerable recent attention. "Conduct risk has only recently become recognised as a standalone risk category in the aftermath of a number of high-profile incidents of misconduct (and regulatory responses) in retail and commercial banking, capital markets, and wealth management," McKinsey said in a 2018 article. Effective culture and conduct risk management requires a new approach — one that can "connect the dots" across individual and team activities in a manner akin to testing for the collective intelligence of work groups, McKinsey said.

This is the promise of regtech tools that identify patterns hidden in data drawn from standard management systems. Critically, McKinsey concluded, these tools, "go beyond the detection of past instances of misconduct — by which the damage to an institution, if any, has already been done — to intercept the outlying patterns of activity that could lead to future losses".

This is a wholly new capability set that is not yet well-understood, by the media perhaps least of all. Journalists frequently look to locate these tools within established cognitive categories, such as surveillance. A recent article on Regulatory Intelligence, for instance, started from this erroneous conflation.

"Premising conduct risk management on systems and data analytics ties all conduct understanding to formal surveillance," the article said, before going on to argue that this heightened level of surveillance will erode a critical atmosphere of "psychological safety" in the workplace.

Predictive behavioural analytics tools are not surveillance mechanisms, however. Rather, they operate to provide management with a powerful diagnostic lens, like an MRI, allowing us to search out organisational health insights without invasive surgery. These

capabilities offer a new means of connecting the dots between current circumstances and future outcomes in ways that may be non-obvious or even counter-intuitive.

Such improved management capabilities may well help to increase psychological safety, and this is all the more likely where regtech tools operate on the basis of " unobtrusive data" that does not compromise employee privacy concerns.

A better standard of care

"Reliance on data- and technology-led solutions may fail to deliver insights and controls," the piece above stated. True. But the fact that marriage may end in divorce is not usually grounds for calling off the nuptials. Predictive behavioural analytics tools must be trustworthy: fairly judged on their competence, reliability, honesty and benevolence. Where they pass these tests, such tools may permit for a much-improved standard of care: predicting and preventing adverse outcomes rather than merely detecting and correcting them.

This will benefit firms and their leaders, shareholders, regulators, customers and employees. More, a reduction in misconduct may produce greater confidence that supervisors are able to maintain adequate oversight of financial firms, helping to restore the trustworthiness of the financial system more broadly. This is an opportunity that should be embraced.

We appreciate and share ethical concerns regarding the use of AI in the workplace, but what are the ethical considerations in deciding not to make use of such tools, particularly where they work in a competent, reliable, honest and benevolent manner to the benefit of stakeholders?

We would do better to reframe the debate around trustworthy AI in this direction.

Authors

Stephen Scott *is a recognised risk management expert and CEO of Starling, a pioneering AI technology company that helps organisations to better manage risk and performance through computational social science.* ***Karen Cook*** *is a professor of sociology at Stanford, studying social networks, social exchange, and trust. She was founding director of the Institute for Research in the Social Sciences.* ***Amy Edmondson*** *is a professor of leadership and management at the Harvard Business School. Her work focuses on human interactions that lead to successful enterprises that contribute to the betterment of society.* ***Thomas Malone*** *is a professor of management, information technology, and work and organisational studies at the MIT Sloan School of Management. He is the founding director of the MIT Center for Collective Intelligence.*

Complaints Procedure

Produced by Thomson Reuters Accelus Regulatory Intelligence 11-Nov-2021