THINK TANK

Education, population, poverty, tax…getting views on topics like those in Jersey isn't the difficult part – but have you ever noticed how many people sound eminently credible when talking about them, even though they may actually be basing their views on conjecture, false facts and blind guesswork?

There is a real danger in making the 'facts' fit the opinion, rather than the other way around – which is exactly the point at which someone with an eye on the latest buzzwords will smugly insert the phrase 'post-truth' into the conversation, imagining its actually helpful.

So, we've asked the Jersey Policy Forum to add some robust material to those crucial local debates – the point is not to provoke agreement or acquiescence; it is to provide reliable material on which others can build their views.

Gailina Liew, Director, Jersey Policy Forum

Trustworthy AI – humans are the critical factor

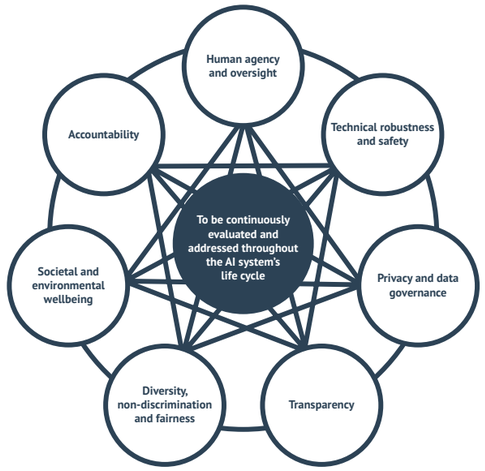
What does the phrase 'Trustworthy AI' mean to you? If you're texting, 'AI' might be shorthand for 'as if'. In the animal world or for a Scrabble expert, an 'ai' is a three-toed sloth. For this article, 'AI' is shorthand for 'artificial intelligence' and the question of how to ensure AI can be trustworthy is an important one as the impact of AI on our lives and in virtually every aspect of our societies continues to grow.

For several decades, Hollywood and television writers have made lots of money by entertaining us with science fiction stories that now seem alarmingly real. The malfunctioning HAL9000 computer system in Stanley Kubrick's 2001: A Space Odyssey (1968), that turned out to be due to a human programming error in the 1984 sequel. In Alex Proya's iRobot (2004), Sonny the robot seemed to exert autonomous murderous intent but was actually executing to human instructions. In the TV series, Star Trek: The Next Generation (1987-1994), we followed Lieutenant Commander Data's continual struggle as an artificial life-form to understand human emotions and their impact on human decisions and well-being.

More recent examples of Google, Facebook, and others demonstrating what AI is capable of doing with our personal information (to sell to each of us in a specific way, to manipulate the information we see, to influence how or whether we vote, to exploit our mental health vulnerabilities, etc.) is now raising broad public concern but ordinary people are still largely unable to do anything to protect themselves short of cutting off all digital interaction. So, what can you do if you're concerned about this?

In the last three months, several guidelines on the trustworthiness and ethical use of AI have been published and adopted by the EU, OECD and G20 countries. In April, the High Level AI Expert Group gathered together by the EU published their Guidelines for Trustworthy AI (see *https://ec.europa.eu/futurium/en/ai-alliance-consultation*). In May, the OECD published their council recommendation on AI (see *https://www.oecd.org/going-digital/ai/principles/*) which the G20 then drew from to establish the G20 AI Principles that were adopted by G20 countries in a ministerial statement made in June (see https://g20trade-digital.go.jp/dl/Ministerial_Statement_on_Trade_and_Digital_Economy.pdf). Finally, the UK just published a set of guidelines for the use of AI in the public sector (see https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector).

These guidelines are not enforceable from a legal perspective but they are important as they mark the first time that national governments have come together to adopt and abide by key underlying principles for the development and use of AI. The most comprehensive set of guidelines are those from the EU and these guidelines are intended to ensure that AI is used in a human-centric way. Based on fundamental human rights, a democratic framework and ethical principles, the EU guidelines suggest that **seven key requirements** must be met for AI to be trustworthy:



1. human agency and oversight;
2. technical robustness and safety;
3. privacy and data governance;
4. transparency;
5. diversity, non-discrimination and fairness;
6. societal and environmental well-being;
7. accountability.

These principles are intended to operate with a legal framework that has yet to be clearly defined. They are also intended to interact with each other dynamically and should be continuously evaluated throughout the AI's life cycle as shown left.

The EU's guidelines go a step further than the others by providing a checklist of specific questions (see pages 26-31 of the guidelines) that can be asked of AI developers to determine whether the principles have been met. The EU expert group has also launched a pilot assessment phase and are seeking participant companies to test the questions with a target of early 2020 to publish a final set of questions. So, what are some of these questions?

To assess the principle of Human Agency:

- is the AI system implemented in work and labour process?
- If so, did you consider the task allocation between the AI system and humans for meaningful interactions and appropriate human oversight and control?
- Does the AI system enhance or augment human capabilities?
- Did you take safeguards to prevent overconfidence in or overreliance on the AI system for work processes?

To assess social impact aspect under the principle of Societal and Environmental Impact.

In case the AI system interacts directly with humans:

- did you assess whether the AI system encourages humans to develop attachment and empathy towards the system?
- Did you ensure that the AI system clearly signals that its social interaction is simulated and that it has no capacities of 'understanding' and 'feeling'?
- Did you ensure that the social impacts of the AI system are well understood? For example, did you assess whether there is a risk of job loss or de-skilling of the workforce? What steps have been taken to counteract such risks?

The full set of questions is well worth reading and pondering. Would the ability to answer these questions earn your trust in the AI that you knowingly or unwittingly use? As there are no laws in place to compel compliance, will these guidelines effectively change behaviour or even matter?

AI as it exists today does not have the capacity to understand what is ethical or to distinguish right from wrong; at a simple level, AI can only do what it was programmed to do. So, the question of whether AI is trustworthy is really about whether the human developers, executives and boards that are driving the development, marketing and use of AI are trustworthy. Does everyone involved in AI development, marketing and use require a crash course in ethics? Do ethicists need a deep dive into AI to better understand what the implications might be? What role can Jersey play? What questions would you ask? The EU expert group is seeking broader input over the next few months, so this is a good time to engage and join the discussion!