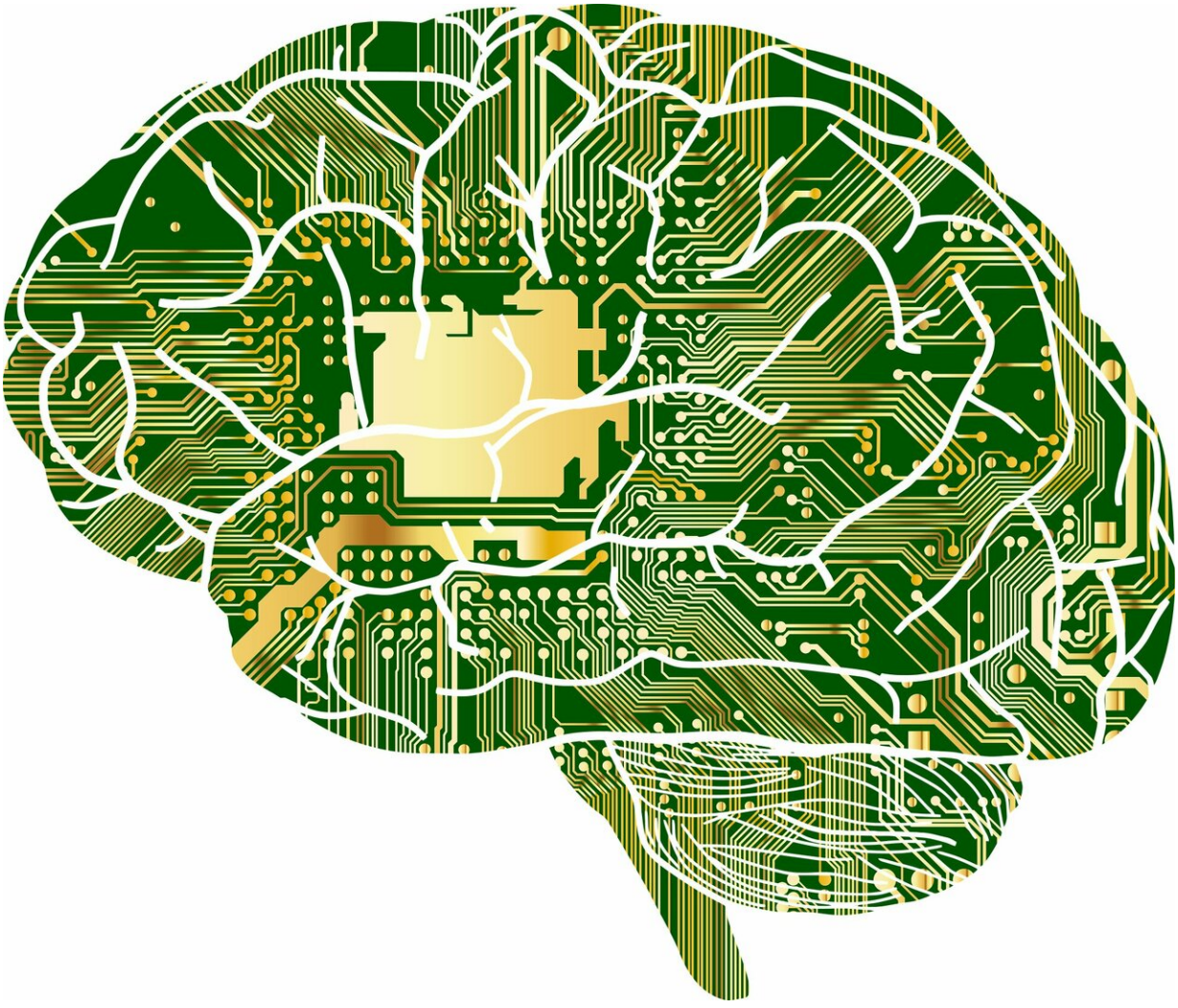
AI could be a force for good – but we're currently heading for a darker future

December 2 2019, by Marcus Tomalin



Credit: CC0 Public Domain

Artificial Intelligence (AI) is already re-configuring the world in conspicuous ways. Data drives our global digital ecosystem, and AI technologies reveal patterns in data. Smartphones, smart homes, and smart cities influence how we live and interact, and AI systems are increasingly involved in recruitment decisions, medical diagnoses, and judicial verdicts. Whether this scenario is utopian or dystopian depends on your perspective.

The potential risks of AI are enumerated repeatedly. Killer robots and mass unemployment are common concerns, while some people even fear human extinction. More optimistic predictions claim that AI will add US$15 trillion to the world economy by 2030, and eventually lead us to some kind of social nirvana.

We certainly need to consider the impact that such technologies are having on our societies. One important concern is that AI systems reinforce existing social biases—to damaging effect. Several notorious examples of this phenomenon have received widespread attention: state-of-the-art automated machine translation systems which produce sexist outputs, and image recognition systems which classify black people as gorillas.

These problems arise because such systems use mathematical models (such as neural networks) to identify patterns in large sets of training data. If that data is badly skewed in various ways, then its inherent biases will inevitably be learnt and reproduced by the trained systems. Biased autonomous technologies are problematic since they can potentially marginalise groups such as women, ethnic minorities, or the elderly, thereby compounding existing social imbalances.

If AI systems are trained on police arrests data, for example, then any conscious or unconscious biases manifest in the existing patterns of arrests would be replicated by a "predictive policing" AI system trained

on that data. Recognising the serious implications of this, various authoritative organisations have recently advised that all AI systems should be trained on unbiased data. Ethical guidelines published earlier in 2019 by the European Commission offered the following recommendation:

When data is gathered, it may contain socially constructed biases, inaccuracies, errors and mistakes. This needs to be addressed prior to training with any given data set.

Dealing with biased data

This all sounds sensible enough. But unfortunately, it is sometimes simply impossible to ensure that certain data sets are unbiased prior to training. A concrete example should clarify this.

All state-of-the-art machine translation systems (such as Google Translate) are trained on sentence pairs. An English-French system uses data that associates English sentences ("she is tall") with equivalent French sentences ("*elle est grande*"). There may be 500m such pairings in a given set of training data, and therefore one billion separate sentences in total. All gender-related biases would need to be removed from a data set of this kind if we wanted to prevent the resulting system from producing sexist outputs such as the following:

- **Input**: The women started the meeting. They worked efficiently.
- **Output**: *Les femmes ont commencé la réunion. Ils ont travaillé efficacement.*

The French translation was generated using Google Translate on October 11 2019, and it is incorrect: "*Ils*" is the masculine plural subject pronoun in French, and it appears here despite the context indicating clearly that women are being referred to. This is a classic example of the masculine

default being preferred by the automated system due to biases in the training data.

In general, 70% of the gendered pronouns in translation data sets are masculine, while 30% are feminine. This is because the texts used for such purposes tend to refer to men more than women. To prevent translation systems replicating these existing biases, specific sentence pairs would have to be removed from the data, so that the masculine and feminine pronouns occurred 50%/50% on both the English and French sides. This would prevent the system assigning higher probabilities to masculine pronouns.

Nouns and adjectives would need to be balanced 50%/50% too, of course, since these can indicate gender in both languages ("actor", "actress"; "neuf", "neuve") – and so on. But this drastic down-sampling would necessarily reduce the available training data considerably, thereby decreasing the quality of the translations produced.

And even if the resulting data subset were entirely gender balanced, it would still be skewed in all sorts of other ways (such as ethnicity or age). In truth, it would be difficult to remove all these biases *completely*. If one person devoted just five seconds to reading each of the one billion sentences in the training data, it would take 159 years to check them all—and that's assuming a willingness to work all day and night, without lunch breaks.

An alternative?

So it's unrealistic to require all training data sets to be unbiased before AI systems are built. Such high-level requirements usually assume that "AI" denotes a homogeneous cluster of mathematical models and algorithmic approaches.

In reality, different AI tasks require very different types of systems. And downplaying the full extent of this diversity disguises the real problems posed by (say) profoundly skewed training data. This is regrettable, since it means that other solutions to the data bias problem are neglected.

For instance, the biases in a trained machine translation system can be substantially reduced if the system is adapted after it has been trained on the larger, inevitably biased, data set. This can be done using a vastly smaller, less skewed, data set. The majority of the data might be strongly biased, therefore, but the system trained on it need not be. Unfortunately, these techniques are rarely discussed by those tasked with developing guidelines and legislative frameworks for AI research.

If AI systems simply reinforce existing social imbalances, then they obstruct rather than facilitate positive social change. If the AI technologies we use increasingly on a daily basis were far less biased than we are, then they could help us recognise and confront our own lurking prejudices.

Surely this is what we should be working towards. And so AI developers need to think far more carefully about the social consequences of the systems they build, while those who write about AI need to understand in more detail how AI systems are actually designed and built. Because if we are indeed approaching either a technological idyll or apocalypse, the former would be preferable.

Provided by The Conversation

Citation: AI could be a force for good – but we're currently heading for a darker future (2019,

December 2) retrieved 20 April 2024 from https://techxplore.com/news/2019-12-ai-good-darkerfuture.html