nature

others less so. To succeed, a treaty that is administered by the WHO will need every country to respect its instructions.

After the 2008 global financial crisis, world leaders realized that parts of the architecture of international finance needed to be mended. But you cannot fix a broken system in the middle of a crisis. A treaty to fix today's ills has the potential to be a powerful instrument in a future pandemic, but, with countries still navigating their way out of this one, it's important to remember that people don't need an international law to pick up the phone and talk.

Rise of AI debaters highlights need for transparency

With artificial intelligence starting to take part in debates with humans, more oversight is needed to avoid manipulation and harm.

Can a machine powered by artificial intelligence (AI) successfully persuade an audience in debate with a human? Researchers at IBM Research in Haifa, Israel, think so.

They describe the results of an experiment in which a machine engaged in live debate with a person. Audiences rated the quality of the speeches they heard, and ranked the automated debater's performance as being very close to that of humans. Such an achievement is a striking demonstration of how far AI has come in mimicking human-level language use (N. Slonim *et al. Nature* **591**, 379–384; 2021). As this research develops, it's also a reminder of the urgent need for guidelines, if not regulations, on transparency in AI – at the very least, so that people know whether they are interacting with a human or a machine. AI debaters might one day develop manipulative skills, further strengthening the need for oversight.

The IBM AI system is called Project Debater. The debate format consisted of a 4-minute opening statement from each side, followed by a sequence of responses, then a summing-up. Although Project Debater was able to match its human opponents in the opening statements, it didn't always match the coherence and fluency of human speech. This is partly because Project Debater is a machine-learning algorithm, meaning that it is trained on existing data. It first extracts information from a database of 400 million newspaper articles, combing them for text that is semantically related to the topic at hand, before compiling relevant material from those sources into arguments that can be used in debate.

Systems such as this, that rely on a version of machine learning called deep learning, are taking great strides in the interpretation and generation of language. But because training data are drawn from human output, AI systems can end up repeating human biases, such as racism and sexism. Researchers are aware of this, and although some are making efforts to account for such biases, it cannot be taken for granted that corporations will do so.

> **"Researchers are starting to incorporate elements of a theory of mind into AI models."**

As AI systems become better at framing persuasive arguments, should it always be made clear whether one is engaging in discourse with a human or a machine? AI specialist Stuart Russell at the University of California, Berkeley, told *Nature* that humans should always have the right to know whether they are interacting with a machine – which would surely include the right to know whether a machine is seeking to persuade them. It is equally important to make sure that the person or organization behind the machine can be traced and held responsible in the event that people are harmed. Project Debater's principal investigator, Noam Slonim, says that IBM implements a policy of transparency for its AI research, for example making the training data and algorithms openly available.

Right now, it's hard to imagine systems such as Project Debater having a big impact on people's judgements and decisions, but the possibility looms as AI systems begin to incorporate features based on those of the human mind. Unlike a machine-learning approach to debate, human discourse is guided by implicit assumptions that a speaker makes about how their audience reasons and interprets, as well as what is likely to persuade them – what psychologists call a theory of mind.

Nothing like that can simply be mined from training data. But researchers are starting to incorporate some elements of a theory of mind into their AI models (L. Cominelli *et al. Front. Robot. AI* https://doi.org/ghmq5q; 2018) – with the implication that the algorithms could become more explicitly manipulative (A. F. T. Winfield *Front. Robot. AI* https://doi.org/ggvhvt; 2018). Given such capabilities, it's possible that a computer might one day create persuasive language with stronger oratorical ability and recourse to emotive appeals – both of which are known to be more effective than facts and logic in gaining attention and winning converts, especially for false claims (C. Martel *et al. Cogn. Res.* https://doi.org/ghhwn7 (2020); S. Vosoughi *et al. Science* **359**, 1146–1151; 2018).

As former US president Donald Trump has demonstrated, effective orators need not be truthful to succeed in persuading people to follow them. Although machines might not yet be able to replicate this, it would be wise to propose regulatory oversight that anticipates harm, rather than waiting for problems to arise. Equally, AI will surely look attractive to those companies looking to persuade people to buy their products. This is another reason to find a way, through regulation if necessary, to ensure transparency and reduce potential harms. AI algorithms could also be required to undergo trials akin to those required for new drugs, before they can be approved for public use.

Government is already undermined when politicians resort to compelling but dishonest arguments. It could be worse still if victory at the polls is influenced by who has the best algorithm.