Remarks to the Joint LIBE-IMCO meeting of the European Parliament
March 21, 2022
Stuart Russell, University of California, Berkeley

I'd like to thank the Chairs and committee members for the invitation to speak.

The Act is extremely important. The world is watching, as this is the first major step in regulating what will probably become the dominant technology of the future, one that may determine the course of human civilization.

The Act's protection for fundamental rights is essential—particularly the right to *mental integrity* from Article 3 of the Charter. This implies that information systems (such as social media recommender systems or computer games) that may manipulate one's mental traits are high-risk systems and must be regulated as such.

Therefore I welcome the proposal to strengthen the language in Recital 16 and Article 5 on material distortion of human behaviour, although the wording "*subliminal techniques beyond a person's consciousness*" might be problematic; a person can perceive the game he or she is playing, but may be unaware of the neurological addiction it is creating through deliberate, but not subliminal, design.

The Act refers to "psychological harms," but these are partly subjective. For example, a person may be manipulated by social media algorithms and become a fascist, but is unlikely to register a complaint about that fact. Suppose we ran a clinical trial of the algorithm and observe the results. Are they acceptable? It seems reasonable to propose that if the pre-trial individual indicates in advance that certain changes in mental disposition are unacceptable, but those changes occur, then that is prima facie evidence of distortion.

I also welcome the language in Recital 70 concerning "risks of impersonation or deception" and requiring notification of persons that they are interacting with a machine. I think this will come to be seen as one of the most important aspects of the Act. For example, I have seen plans to populate virtual-reality worlds such as the "Metaverse" with millions of digital entities that will pretend to be human, befriend real people, and over time influence their purchasing or voting behaviour. So this rule will become more and more essential for the preservation of human dignity. We owe each other a duty of consideration. We owe machines nothing at all.

I believe that the Act's definition of AI systems is unfinished. In particular, there are at least two AI systems that need to be considered: the "deployed object" that is put into service and produces predictions, decisions, and so on; and the "pre-deployment generator" that creates the deployed object. These can have very different characteristics.

For example, it is very common for the pre-deployment generator to employ machine learning, while the deployed object is fixed and does no learning. The reinforcement learning techniques known as policy search and Q-learning produce deployed objects—policies and Q-functions—that themselves do no learning, reasoning, or modelling, and would therefore be non-AI systems.

As far as I can tell, the regulations apply primarily to the deployed object, but, according to explanations I have received, two identical deployed objects can be regulated differently depending on the pre-deployment generator. For example, if humans write the code of the deployed object, it is not subject to regulation, but if a machine learning system produces an identical deployed object—maybe a simple decision tree or linear classifier—it *is* regulated. This allows circumvention of the law whereby (say) deep learning is used to preprocess the data into a small set of numerical features, then humans do the final adjustments on the weights of these features and the threshold to determine the best decision boundary.

There is also the issue of *functional equivalence*: two software objects are functionally equivalent if, for all inputs, they produce the same outputs. One might think that functionally equivalent objects should be regulated identically. But in the Act this is not the case. For example, a set of logical rules for classifying discrete objects can be converted into a functionally equivalent lookup table, but the first, because it contains “AI methods”, is regulated, whereas the second is not. This provides another means for evading regulation.

Finally, let me turn to general-purpose AI, defined as “*AI system[s] that are able to perform generally applicable functions such as image/speech recognition, audio/video generation, pattern detection, question answering, translation, etc.*

Many AI researchers find the exemption of general-purpose AI in Recital 70a puzzling.

First, it’s hard to see how a machine translation system or a speech recognition system has no purpose.

Second, such systems can themselves incorporate biases—for example, a speech recognition system that’s much less accurate for female voices—and other performance failures.

Third, it makes sense to assess their accuracy, fairness, etc., at the source, that is, the large-scale vendor of general-purpose systems, who has the data and design information to carry out conformity assessments, rather than a large number of presumably smaller European integrators, who do not. A car manufacturer does not escape safety regulation by selling unpainted cars to dealers who apply the final coat of paint.

Finally, as we move towards *truly* general-purpose AI systems, they will become the primary product, and they will pose the greatest risk to society at large—particularly if the “human-defined objectives” they pursue are mis-specified. One possible solution is to build AI systems that are explicitly uncertain about the objectives that humans possess. Ironically, because they do not “achieve a given set of human-defined objectives”, it’s not clear that the proposed Act applies to such systems.