Dr Jean-Christophe Nebel – Written evidence (AIC0102)

"The Select Committee on Artificial Intelligence was appointed by the House of Lords on 29 June 2017. It has been appointed to consider the economic, ethical and social implications of advances in artificial intelligence. It has to report by 31 March 2018".

My definition of artificial intelligence in the context of my response: usage of a machine to make recommendations through automatic analysis of a large amount of data which is usually beyond what an expert could handle.

My answer addresses aspects of the following questions:

> **Impact on society**
>
> 4. Who in society is gaining the most from the development and use of artificial intelligence and data? Who is gaining the least? How can potential disparities be mitigated?
>
> **Ethics**
>
> 8. What are the ethical implications of the development and use of artificial intelligence? How can any negative implications be resolved? In this question, you may wish to address issues such as privacy, consent, safety, diversity and the impact on democracy.
>
> 9. In what situations is a relative lack of transparency in artificial intelligence systems (so called 'black boxing') acceptable? When should it not be permissible?

The process which has led to recommendations made by an artificial intelligence system can generally not be understood by a human, even an expert ('black box' effect). As a consequence, it is very difficult to challenge the output of such system. Moreover, it has been shown that, for example, a deep learning system can be relatively easily manipulated to produce any prediction which has serious security implications [Ng2015].

Artificial intelligence systems rely on large amount of data from which generalisations are made (system training) and used to make decisions. While such systems are proving to be more and more powerful and useful, one of their main drawbacks is their dealing with exceptional cases (outliers). By definition, an artificial intelligence system would not be trained for such cases (or trained insufficiently) and, as a consequence, would return uninformed decisions. There is a risk that users of artificial intelligence systems follow blindly their recommendations: first, systems are almost always right giving a false sense of security; and, second, there may be a lack of awareness of the negative consequences resulting from incorrect decisions. Indeed, they may be either hidden by the much higher volume of successful outcomes or undetected due to outcomes taken place at a stage when causality with decision is difficult to establish. To prevent such situations, it is important that any result produced by such system is associated to not only appropriate confidence metrics, but also some clues about how a result has been obtained. They do not need to be

comprehensive, but they need to be sufficient so that a user can challenge any obvious random or uninformed decision [Ng2015]. For example, key training examples which contributed to a recommendation could be highlighted or visualisation of 'neighbouring' cases associated with a similar outcome could be provided. In addition, a 'what if' functionality could be offered so that robustness of a decision could be tested by altering slightly the features of the case of interest. To summarise, humans need to be kept in the loop and novel tools should be developed so that they can interact with artificial intelligence systems and be able to exercise critical thinking even when faced with a black box.

Another important point is to be aware that artificial intelligence systems are NOT PC and do not have any agenda. Their task is to help decision making by generalising and, possibly, use stereotypes, if that leads to better global performance. If not moderated, such behaviour would be particularly unsuitable when dealing with decisions affecting directly vulnerable individuals.

Finally, artificial intelligence systems are firmly rooted in the past – they are based on past (training) examples – and, as a consequence, are unlikely to promote novel or creative solutions. This may lead to taking safe decisions with predictable outcomes, instead of novel or higher risk ones which could have much greater impact.

Reference

[Ng2015] A. Nguyen, J. Yosinski, J. Clune, "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images", in Computer Vision and Pattern Recognition, 2015

6 September 2017