Dear Members of the European Commision,

I hope you and your loved ones are all doing well in the current Coronavirus situation. I'm a PhD student in AI at the University of Freiburg, Germany. The White paper "On Artificial Intelligence - A European approach to excellence and trust" was, in general, very well thought out and very informative as to the future directions Europe wants to take on AI. I trust in and support the European values, principles of equality and fundamental rights and am pleased to see the Commission taking a proactive approach to future possibilities.

I think one important aspect that the white paper did not touch upon is the development of Artificial General Intelligence (AGI). This is AI that would reach or exceed human level intelligence. I'll expound a bit more on that now. As we move towards more intelligent systems, we have been moving towards inexact knowledge. Computers began with the ability to store and retrieve exact data. Then came machine learning that allowed computers to be more intelligent, but in a correlational manner (as mentioned in the white paper). These systems are at the same time more "inexact" than traditional computers and usually don't remember exact training samples - the algorithms in deep learning are, in a sense, approximate rote-based learners. They "generalise" within the training data distributions they see and usually ONLY within them, but they also will very likely not be 100% accurate and exact within the distribution that they learnt. But in these narrow distributions, they are already capable of exceeding human level performance at various tasks. Another level further would be to be like humans - who find it hard to learn many things exactly at the same time but can learn the most important underlying aspects in their environment and generalise to a much broader set of tasks than narrow AI but in a less exact manner. This would be AGI. An important aspect distinguishing AGI is likely to be that, like humans, the learning in AGI would be causal and not correlational like in current machine learning. Now, it is very uncertain if and when we would achieve AGI, but there is a good chance that we would at some point (and it could even come very quickly if things fall into place). Once that happens, it would be even more disruptive than the current narrow form of AI. In the best case, humans could basically be rid of the need to work and all work could be assigned to AGI. But if not handled correctly, it could lead to catastrophic outcomes as well. So, we need to keep it in mind and address it as well. There are also papers like AI-GAs (https://arxiv.org/abs/1905.10985) that suggest that we might get to AGI faster by following an automated path to building AI systems rather than a manual path. It is like presenting AI algorithms with a complex environment like the Earth and letting them learn and "evolve" on their own. Such systems are likely to be even more opaque and also much harder to control than the traditional manual path to AGI. And we need to keep these in mind, as well, when addressing AI and laws regarding AI.

I have the following further comments on specific aspects of the white paper:

"Green" AI:

Currently, most resources in training AI systems are being used just to produce bigger Neural Networks (and the correlationary learning mentioned in the white paper) and not causal learning. I personally, find it a waste of resources that what most of the research papers do is to introduce bigger Neural Networks to "memorise" bigger distributions - for example, even though it is known that AI can play very complex games, corporations spend huge amounts of resources and increase their carbon footprint just to show that AI can learn an even more complex game using even more resources. I don't think that that approach will help us achieve better AI systems that, for instance, combine symbolic/causal AI with correlational AI. And a disproportionate amount of funding is being directed there. It'd be better to do more directed research and move to the "next level" in machine learning by directing research towards causal AI.

Bias:

The white paper says "to use data sets that are sufficiently representative, especially to ensure that all relevant dimensions of gender, ethnicity and other possible grounds of prohibited discrimination are appropriately reflected in those data sets;". Current deep learning is always going to be biased to the training data it is shown. Again, unless we have causal learning, this is not going to be solved. In such cases, the only way to remove bias may be to NOT use discriminatory features like race, gender, sexual orientation, etc. to train the algorithms, e.g., don't include gender to predict performance in a job, otherwise deep learning will definitely be biased to the training distribution it sees. Of course, this depends on the specific use case for which the system being developed is being used. For example, we might want to see if there are patterns in data that suggest that some minorities are more susceptible to the Coronavirus. In general, however, it's very important to try and remove such features from training data where they are not needed to predict causal outcomes. We can even see such correlational learning among humans. A large part of human learning comes from correlational learning and the less tolerant ones among us tend to be biased to such "data" they may have seen through their lifetimes and tend to be prejudiced. And yet the more tolerant ones among us realise that just because they saw some people of a certain sexual orientation or race or gender, etc. act a certain way, it doesn't mean that there exists such a causal relationship. When building AI systems, it might then, in fact, be a boon that we can remove such prejudicing features from the training data and remove human biases.
In any case, whether features like race, sexual orientation, gender, etc. are really needed to train the algorithm or not, care needs to be taken that the data is COLLECTED in a representative manner (as mentioned in the white paper).

"AI may make it more difficult for persons having suffered harm to obtain compensation":

I'd like to begin with an example here. Say an owner trains their dog to attack other humans. Would it be the owner's fault if their dog caused harm to another person? In my opinion, yes, it would be. And yet, it might be hard to prove that the owner did try and do what they did. Personally, I'm not sure how this example is handled in the current legal framework. But, maybe we need to have examples like this in mind when formulating laws also for AI.

Moving on to other aspects of the white paper, I also understood from the paper that when AI, software, etc. is added by someone later in a product supply chain, who is liable for what damage is not currently clearly handled by the legal framework. I think this needs to be dealt with clearly as well. And at the European level, not just the national level. The original product producer should not be liable in case someone later in the production line adds damage causing AI/software to their product, in my opinion.
However, in the case of software updates from the producer, after a product was

placed on market (which is another grey area I understand), producers need to be liable for such updates. We need to be clearer in establishing the blame.

Regarding open-source AI, many creators of AI already release their systems the way software is released under open-source licenses like the GNU-GPL with no warranties. Even completely interpretable software (without the use of AI) is not completely verifiable, so with AI systems it's going to be even harder to verify their functioning. I'm not well-versed in the legal jargon here, but I believe open-source AI is another issue to keep in mind when formulating laws related to AI.

Finally, given the correlational nature of machine learning that I have stressed throughout, I believe some "hard-coded" symbolic rules to ensure safety, e.g. in autonomous driving, need to be programmed into the system. Such rules should always have higher priority to be executed than the AI system's output actions because we do not know when something might occur that was not seen by the system in its training data distribution - it is extremely hard for a human designer to think of all possible data (such as pixels in an image) that might be needed to cover all possible situations and this is one of the reasons that we have "adversarial examples" that cause deep learning to fail. And yet it is anomalous data, i.e., data that is not seen in usual everyday distributions, that is important when it comes to safety critical applications. This is why I suggest "hard-coded" symbolic rules not just in the design phase but also in production - for example for an autonomous car to slow down immediately when it detects bodies closer than, say, 10 metres and to completely stop when it detects the object getting too close.

Thank you for reading my opinions and I hope that you found them useful and will take them into account moving forward. I hope to see Europe be a global leader in AI research and innovation that is ethical and sustainable at the same time.

Greetings,
Raghu Rajan.