Comments on European Commission draft: **Ethics Guidelines for Trustworthy AI**

Olle Häggström

January 4, 2019

Introduction: Rationale and Foresight of the Guidelines

On p i, l 14-15, I am struck by the asymmetry between the adjectives in “tremendous benefits” and “certain risks” – a contrast that is then made explicit in the very next sentence, saying that “on the whole, AI’s benefits outweigh its risks”. This is unfounded. I’m not saying the situation is symmetric or that the balance goes the other way. I’m saying we are far from knowing which way the balance goes. There simply isn’t any serious study that systematically goes through the various potential benefits and risks, in order to establish that “AI’s benefits outweigh its risks”. Given the various risks in Chapter I, Section 5 (more on which below), confidently claiming that “AI’s benefits outweigh its risks” is preposterous and risks coming across as motivated by ideology rather than by evidence.

(In view of this, it may at first sight seem puzzling that everyone advocates continued or increased efforts to develop AI, rather than an across-the-board moratorium on such development. But the reason, of course, is not that we know that “AI’s benefits outweigh its risks”, but rather that, for a range of societal reasons, a moratorium is utterly unrealistic. It should also be noted that much of the uncertainty regarding the benefits vs risks balance stems from the fact that future AI policies have not yet been written in stone. This actually **strengthens** the conclusion of the sentence on p i, l 15-16, that “we must ensure to follow the road that maximises the benefits of AI while minimising its risks”.)

Chapter I: Respecting Fundamental Rights, Principles and Values – Ethical Purpose

On p 11, Section 5.2, the second sentence reads “Otherwise, people with the power to control AI are potentially able to manipulate humans on an unprecedented scale”. On this, I have two comments. First, the word “people” doesn’t ring quite right here, and is better replaced by “organizations” (or the more neutral “agents”). Second, the sentence risks being read as suggesting that as soon as the proposed non-covertness is implemented, there is no such risk of manipulation. That is clearly wrong. People are generally very willing to interact with AI systems in a way that exposes us to manipulation (most of us do that with Google and Facebook every day, and the success in China of Microsoft’s Xiaoice chatbot is another example worth studying closely), and the level of manipulation may well be aggravated in the future even in the absence of covert AI systems pretending to be human.

Regarding Section 5.3, it should be noted that in China, a large-scale social credit system is well underway. While transparency of any such system is desirable, that in itself does not prevent the system from being used to oppress the population. It should furthermore be noted that formally including an opt-out button in such a system does not guarantee that in practice individuals can opt out without accepting overwhelming costs of various kinds.

As to Section 5.4, I wish to emphasize that what is written here – in particular “it can lead to an uncontrollable arms race on a historically unprecedented level” (which, obviously, can in turn increase the risk of World War III) – is on its own a strong indication that the phrase “AI’s benefits outweigh its risks” (p i, l 16) that I criticize above is overhasty. Furthermore, I think the claim that “LAWS can reduce collateral damage, e.g. saving selectively children”, although factually correct, is nevertheless unfortunate because it risks clouding the fact that a LAWS arms race is overall a bad thing (and you can try and see how the sentence would sound if you replace "children" by the perhaps equally plausible "Christians" or "whites").

In Section 5.5, Footnote 18, it would be worth pointing out explicitly that a development where "self-conscious AI systems would need to be treated as ethical objects" (or should that be "subjects"?) would undermine the human-centered ethical foundations surveyed in Section 3.

Staying in Section 5.5, I have two comments on the passage about "Unsupervised Recursively Self-Improving Artificial General Intelligence (AGI) – which today seem[s] to belong in the very distant future". First, concern here should be more generally about superintelligence, for which Unsupervised Recursively Self-Improving AGI is just one of the ways in which it may come about; see Chapter 2 of Nick Bostrom's *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014) for a number of others. Second, the statement about belonging "to the very distant future" has very shaky empirical foundation, and is to some extent contradicted by surveys among AI experts (see, e.g., Dafoe and Russell, 2016, https://www.technologyreview.com/s/602776/yes-we-are-worried-about-the-existential-risk-of-artificial-intelligence/). The conception that nothing very drastic can happen in the near term seems to have arisen to a large extent from individual AI futurologists' conscious or unconscious wish to brand themselves as sane and measured (as opposed to being a mad doomsayer) rather than from solid evidence. Worth reading in this context is Eliezer Yudkowsky's 2017 essay *There's No Fire Alarm for Artificial General Intelligence* (https://intelligence.org/2017/10/13/fire-alarm/) which lists the three most commonly advocated reasons (A), (B) and (C) for thinking that a superintelligence breakthrough is not near-term, and explains that all three point to circumstances that are likely to still hold shortly before the kind of hard take-off that the author considers plausible. All things considered, we are very uncertain about what is the correct time scale for when (if at all) to expect superintelligence. And here we should not make the tempting mistake of conflating "very uncertain" with "very distant".

A final remark regarding Section 5.5 concerns Footnote 21. While it is correct to point out the major difficulties involved in estimating the probability of very rare high-impact events, the statement that for events that have never been observed, "probability of occurrence is not computable using scientific methods" is plain false, and suggests an overly crude and black-and-white view of science. Science is not solely about observing relative frequencies in the past and blindly extrapolating them to the future. For a view that incorporates a much-needed amount of nuance, please see Sections 6.5 (for which my blog post http://haggstrom.blogspot.com/2013/10/no-nonsense-my-reply-to-david-sumpter.html constitutes an early draft) and 8.1 of my book *Here Be Dragons: Science, Technology and the Future of Humanity* (Oxford University Press, 2016).

General Comments

These timely and well-structured guidelines contain much of value for contributing to putting us on a benign AI trajectory. If the unfounded claim in the Introduction about how "AI's benefits outweigh its risks" is corrected, along with an adjustment for the slight overall tendency towards downplaying risks in Chapter I, Section 5, then my enthusiasm for the document will be wholehearted.