Eden Maria Alagaw (艾千荨) [citation method edited for sampling purposes]

Question 4.1 **May 2023**

Which is the more urgent problem for society now, the black box problem or the singularity?

Artificial Intelligence raises certain problems, two of them being the black box problem and the singularity. Liao (2020) classifies the ethical concerns associated with AI into two categories. Firstly, some risks result from the current limitations and vulnerabilities in machine learning. Secondly, there are risks associated with the development of super-intelligent machine learning systems that could create vulnerabilities for the humans facilitating them. The black box problem falls under the first category, describing the limitations of AI in terms of transparency in its decision-making processes as well as interpretability of results. Singularity represents the second type of problem, referring to the potential blurring of the distinction between human and computational intelligence.

The black box problem describes a challenge that we are currently experiencing. Singularity is a phenomenon that society has not yet encountered, as artificial intelligence has not surpassed human intelligence. Considering this, one might perceive the black box problem as more imminent. However, I believe that the assessment should not be based on a timing or likelihood perspective, but rather on a magnitude perspective.

To illustrate this perspective, one could apply the approach of the *Basel Committee* in assessing globally systemically important banks (G-SIBs) to the evaluation of AI-related risks. The committee measures potential bank failures "*in terms of the impact that a bank's failure can have on the global financial system and wider economy, rather than the likelihood that a failure could occur*" (Basel Committee on Banking Supervision, 2021). This describes the adaption of a loss-given-default (LGD) concept rather than a probability-of-default (PD) one. When applying this approach to AI, the risk that should be prioritised is the one with a greater potential negative impact on society, rather than the one with a higher probability or closer proximity in terms of timing of occurrence.

One may question the applicability of the approach used for assessing financial institutions' economic risks to the ethical risks of AI. Considering the market size of AI, which was estimated to be around 80 billion USD in 2022 (Bloomberg, 2022), it is reasonable to interpret AI as an institution based on its monetary value. This is for instance comparable to the 110 billion USD market capitalization of the American bank *Goldman Sachs*. Further, like banks, AI technologies are subject to regulations that ensure responsible practices in their development and deployment. This highlights the institutional nature of AI systems as providers of technological solutions, similar to banks serving as financial intermediaries.

When trying to determine the potential negative impact, one could measure the extent to which a risk poses a threat to human morals and values. German AI-expert, professor, and former state secretary Miriam Meckel expresses great hesitancy in attributing the trait of consciousness to machines given that even human consciousness is still not fully understood (Handelsblatt, 2022). In the context of Liao's classification, I perceive such a lack of consciousness of intelligent machines as a greater threat to society within the second category of AI problems. Unlike the first category, where vulnerabilities stem from inherent weaknesses in the technology itself, the second category places vulnerability on the individuals utilising the technology.

Therefore, I consider the black box problem, despite its acknowledged limitations, to be of lesser concern in terms of societal impact compared to the uncontrollable absence of limitations and immediate effects on humans associated with singularity.