Dr Dina Babushkina

Human Mistakes and Algorithmic Errors

Taking a responsible stance towards the professional use of AI implies, inter alia, the readiness to respond to and address—in ethically appropriate manner—any wrongdoing that may be associated with such use. This means that neither the cased harm nor the possibility of things going wrong may be ignored, made appear less important, or, when preventable, discarded as accidents. More specifically, being a responsible professional user of AI entails ownership of mistakes. In this talk, I will ask if a mistake in AI-enhanced decision making—such as AI-aided medical diagnosis—can be attributed to the AI system itself. I will analyze the distinction between mistake and error, and argue that we cannot justifiably attribute mistakes to AI systems. The argument relies on the evaluation of ontological status of AI systems. If we are to consider diagnostic AI systems merely as tools used by humans, tool's failing can be explained as either being broken or inaccurate (a bad/faulty tool) or wrong application (the user's mistake). Tools do not make mistakes because being mistaken presupposes (a) being at liberty to deviate from the optimum and predetermined resolution paths, and (b) aiming at truth/being correct, in one form or another. It is more appropriate to attribute errors to such systems, since error is merely a form of deviation from an expected result. If, on the other hand, we take a different conceptual approach where AI diagnostic systems are seen as essentially different from tools and as entities that are, in some way, capable of mistakes, then, I argue, there still is a problem with attributing mistakes to such entities. We can only be justified in accusing someone of a mistake when solving a problem if the task was formulated precisely. Attributing a mistake in diagnosis to an AI system violates this condition, since identifying patterns (the task the system actually solves) is different from the task of diagnosing.