A New Direction for AI Ethics: Computing Ethics Itself

Ever since the advent of computers, the tendency to delegate tasks to machines has been prevalent also in the clinic. Artificial intelligence already helps medical staff with a multitude of different tasks, including precision dosing, predicting long-term therapeutic outcomes, analysing electrocardiograms, and interpreting medical images. Genuinely *ethical* tasks have so far been excluded from automatisation. With the COVID-19 pandemic, however, the need for the taking of thousands of morally relevant decisions within short time frames has arisen. Expanding the use of artificial intelligence into the realm of clinical ethics suddenly seems a worthwhile enterprise.

In the past months, our interdisciplinary team of engineers and ethicists developed the world's first functional ethical advisor system for clinical application (https://www.tandfonline.com/doi/full/10.1080/15265161.2022.2040647). Preliminary performance results are promising: the algorithm's recommendations do not deviate much from those of human ethicists. We will therefore begin this talk with an analysis of the different moral frameworks on which an ethical advisor system could be based and explain how we used machine learning to train it. We shall show how we acquired suitable training data, and captured the parameters of individual medical cases.

However, we shall also ask: is this really the direction in which ethics in artificial intelligence should be moving? Especially for an algorithm that has medical ethics as its target, answers are not as straightforward as one might think. Many people believe, for example, that what can be regarded as an archetypical human bias is highly welcome, or even required, in clinical decision-making: empathy. Emotionlessly calculated decisions may therefore invoke a sense of false impartiality. Does compassion not play a major role in assessing morally relevant situations and taking appropriate actions? Others, however, caution against romanticising human judgment and suggest that one might even hope for *better* moral performance from machines that are not carried away by emotions and led to abandon their moral principles.

Sooner or later technological developments will force us to decide questions of this kind. Irrespective of whether or not calculating ethics is a path that society will ultimately wish to pursue – it is crucial already to begin this discussion and carefully to consider the virtues and vices of the novel options that are becoming available to us.