A Critical Theory Approach to AI Ethics

The ethics of artificial intelligence (AI) is an emerging field in applied ethics, which has gained attention and urgency due to the rapid development of AI technology during the past decade. The popular approach to AI ethics – embraced by ethicists, policy-makers, technologists, and others – has become the so-called *principled approach*. Different AI ethics initiatives have established comparable sets of ethical principles and values that should guide AI development and policy, in order to ensure the realization of ethical or responsible AI (Jobin et al., 2019; Ryan & Stahl, 2020). Common principles are for example transparency, justice and fairness, non-maleficence, responsibility, and privacy (Jobin et al., 2019). Such AI ethics principles function as a kind of soft law, and policy-makers are still working on the translation into actual legislation.

Despite its broad recognition and adoption, the principled approach to AI ethics faces some serious limitations. Ethical principles often prove to be too abstract to translate to concrete technologies or applications. Principles can also conflict with one another, and as they are hard to compare it is not always clear which principle should be given priority. Furthermore, in its current state, AI ethics fails to recognize the ethical relevance of power imbalances that are continued, created, or exacerbated by AI applications, particularly along the lines of race and gender (Gebru, 2020). In other words, the established principled approach to AI ethics lacks sufficient recognition for the social and political context of the technology (Resseguier, SIENNA D5.4).

The thesis of this article is that, in order to overcome this shortcoming, AI ethics should be approached as a critical theory. It is argued that each of the established AI principles is fundamentally concerned with human emancipation and empowerment. So, like a critical theory, AI ethics is aimed at diagnosing as well as changing emerging technologies for the sake of human emancipation and empowerment. Paying closer attention to how AI principles or ethical issues relate to power, helps to bridge the gap between AI ethics and AI’s social and political dimensions.