Negativity Bias in AI Ethics and the Case for AI Optimism

Flipping through the major journals in ethics of technology, one gets the impression that the use of AI is deeply problematic in virtually countless different ways. The big debates in AI ethics almost invariably revolve around *problems*, such as the emergence of responsibility gaps, bias and fairness issues, job displacement, or privacy concerns, to name but a small fraction. A leading AI ethicist rightly observes a 'rising tide of panic about robots and AI'.[1] I want to offer a more optimistic perspective. The 'rising tide of panic' is probably to a good extent the result of a built-in negativity bias in AI ethics. This means that we have higher order evidence to believe that AI is ethically less problematic, and less in need of regulation, than the somewhat panicky mood permeating AI ethics suggests.

To recognize this bias, one must consider two things: 1) The specifics of the subject matter of AI ethics. 2) The incentive structures within academic AI ethics.

The subject matter of AI ethics, and of ethics of technology in general, are technologies and particular technological applications. Technologies can be understood as tools devised by humans to solve problems. One could say that the subject matter of AI ethics is (proposed) solutions. AI ethicists wishing to comment on their subject matter are virtually forced to find fault with these solutions, and the natural way for them to do so is by identifying ethical problems. This distinguishes ethics of technology from other subfields of philosophy, where the cause for philosophizing is not solutions, but long-standing problems ('What is justice?', 'When is a belief justified?', 'What is consciousness?', etc.). In these subfields, philosophers are, by and large, seeking to find solutions, namely answers to these questions, not problems.

This built-in imperative to find ethical problems is amplified by the incentive structure within academia. AI ethicists must publish and are therefore structurally encouraged to keep identifying problems with AI. Moreover, alarmist papers that raise ethical concerns tend to receive more attention and recognition than response pieces that deflate a problem. By the same token, a funding proposal in AI ethics that does not portray AI as, in one way or another, ethically problematic, has little chance of approval.

From this it does not follow that the problems discussed in AI ethics are fictitious. Every ethical concern about AI needs to be taken seriously and deserves to be considered on its own terms. However, three things do seem to follow:

1. We, the AI Ethics community, are probably inflating the ethical problems associated with AI.
2. Legislators should proceed with caution in their regulatory efforts and consider the possibility that some problems with AI are being overestimated.
3. AI ethicists should be aware of the built-in negativity bias in AI ethics and seek to counteract it in their capacity as authors and reviewers.

[1] Danaher, J. 2019: „The rise of the robots and the crisis of moral patiency", *AI & Society* (34), 129-136, p. 129.