Conclusion: Manifesto for a Power-Aware Ethics of AI

In the face of digitalisation and AI technology, we need a new ethics. Most of AI is enabled by the collectively shared behaviour and habits of all of us as users of digital media services in the societies of the Global North and worldwide. Therefore, ethics needs to overcome its liberalist fetishising of the individual moral agent as the locus of ethical deliberation and agency. Ethics needs to abstain from the analytic habitus of judging decontextualised and simplistic scenarios. Ethics needs to become diverse. Ethics needs to become relevant and political. Ethics needs to get up from the armchair and out of the ivory tower. Ethics must go deeper than merely addressing questions of simple oughts, by cultivating character virtues that enable individuals to critically question power structures and assume societal responsibility.

What concept of AI we use is an ethical and political, not an objective question

AI is a hype. 'AI' is a catch-all container term now used to strategically label all kinds of projects that ten years ago were referred to in different and often more technical terms. Labelling things 'AI' today sells products and technologies, unlocks academic and investment funding and sparks attention. The term 'AI' is not a name referring to a more or less circumscribed part of objective reality. AI is a discursive constellation, a nexus of business and governmental interests, a conceptual vehicle that makes many people look differently, and perhaps through a more positive lens, at what in many cases could also be called 'automation'.

Any viable ethics of AI must acknowledge the political nature of the term 'AI'. Calling something AI is not a descriptive but a performative and invested act, as it creates a certain reality pertaining to the thing thus described. The political nature of the term cannot be countered by starting an ethical investigation with an objectivist 'definition of the term' that declares, once and for all, which kinds of algorithms and technologies fall into that category and which don't. In fact, these approaches are themselves no less

political. They only hide their politics behind an insufficiently questioned 'culture of objectivity' (Traweek, Haraway).

We need to situate the notion of AI (and any other technology) in at least two ways. First, as sociotechnical systems marked by the mutual shaping of social reality and technological artefacts that counters naive technological determinism (Feenberg). Second, in its historicity, acknowledging that the idea of the mechanisation of intelligence has evolved as the product of power interests and hegemonic worldviews, including liberal ethical values and their implicit anthropology. Much in the ethics of AI depends on selecting a progressive and empowering understanding of AI. 'Progressive and empowering' presupposes that the concept must enable critical reflection about AI's entanglement with business models and capital interests, social inequalities and power imbalances, new forms of labour and human exploitation. 'Progressive and empowering' also presupposes that the concept must enable awareness of the intellectual and material history of that technology and thereby avoids buying into the contemporary hype surrounding AI. Finally, 'progressive and empowering' presupposes that the concept opens up new perspectives and imaginaries of AI, and inspires new voices to speak up, rather than representing a stance that merely operates as a gatekeeper that restricts the ability to exercise critical judgement and productive imagination to those who have been given the authority of 'really' knowing what they're talking about.

The ethics of AI needs to forge an alliance with social philosophy and critical theories

To take the ethics of AI where it matters most, we need to make use of the rich analytical concepts and resources developed in social and political philosophy, political economy, and science and technology studies, and critical theories including intersectional feminism, Black feminism, postcolonial studies, critical race theory and disability studies. This entails, among other things, incorporating an analysis of power structures, subjectivity and subjectification, modes of capital accumulation, forms of exploitation, subordination and abjection, and local as well as global inequalities that are inherently linked to AI technology and its reality as manifested in business and use cases.

The ethics of AI cannot be abstracted from the novel modes of value creation in digital capitalism. AI is at heart a global *industrial* development, bolstered by significant capital interests. The AI technologies and services that have tangible impacts on society and politics today cannot be detached from the IT industry in the Global North, its culture of disruptive and predatory innovation and financialised funding, its mainly White, male and ableist work culture, and its latent pioneer narratives and neocolonialist

attitudes towards its own digital expansion. AI thus needs to be understood in its sociotechnical materiality that is a composite of data harvesting, user labour and user subjectivity, global inequalities and power hierarchies.

Deconstructing AI technology as an extension of historical colonialism and as an updated form of extractivism is an indispensable prerequisite for AI ethics to be relevant and not politically toothless. Any viable ethics of AI must confront AI's role in both perpetuating existing structures of discrimination, exploitation and violence, and creating new ones that do not always slot neatly into traditional categories of sexism, racism and classism. An intersectional perspective – understood as an analytic sensibility (Crenshaw) and not a formalised method – is an essential component of a critical *ethos* in AI and AI ethics. It is imperative to address the reality that discrimination, bias and opacity in AI primarily serve economic interests, uphold and maintain power differentials and drive competitive business practices (Noble). Condemning discriminatory decisions or predictions, uncovering biases and operationalising 'fairness' are now mainstream themes in AI ethics. However, these discourses backfire as bug-fixing services to the industry and the strengthening of hegemonic discourses around AI solutionism unless they confront the deep entanglements of automated injustice with economic profit, digital extraction and principles of accumulation.

The assemblage perspective on AI is not an alternative ontology but an ethical stance

We should not locate the 'intelligence' of AI systems in the innards of materially circumscribed objects like robots, chatbots or computers. We should not describe AI systems as autonomous agents or even potential moral agents (or patients), misleading the reader to believe that AI systems are entities that interact with humans similarly to how other humans would. AI systems have not popped up in our life-world as creatures that participate in human society. Instead, AI is a technology that structures our social relations, our access to and knowledge of the world and our reflexive relationship with ourselves.

To make this visible, AI systems need to be viewed as networked, decentralised and heterogeneous assemblages. However, the assemblage approach to AI should not be viewed as yet another ontological or objectivist definition of AI. The assemblage approach is not a transhumanist metaphysics; it opposes the idea that we are on an evolutionary journey from human to posthuman. Instead of a metaphysical claim, the assemblage perspective embodies a critical ethos and epistemological stance that facilitates a power-aware AI ethics. The assemblage perspective is not favoured because it is *truer*, but because it opens up a critical viewpoint. Accordingly, if someone believes in the autonomous agency of an AI system, there is no point in dismissing

this belief as 'flawed'. Rather, the truth of the assemblage perspective lies in how much it enables us to question the genealogy of this belief, the discursive and material constellation that sparks it, and the power interests that benefit from it.

Thus, the point of a power-aware AI ethics in regard to anthropomorphising AI is not primarily that this stance is metaphysically *wrong* (which it is), but that it is *bad for society* in that it perpetuates the sociotechnical power constellation of the current AI hype. Morality or immorality, the good or the bad, lies in what AI technology does to our knowledge, consciousness, social lives, desires, politics and cultures. Morality inhabits the question of whose interests AI technology serves and whom it exploits, subjugates and outpowers. For example, an ambitious ethics of AI doesn't care about the 'value alignment problem' with respect to AI systems, but rather with respect to the (mostly corporate) actors building and deploying AI systems. Making those actors 'align' means regulating the industry, as a manifestation of a democratic political will.

Working in the ethics of AI requires cultivating an ethos of seeing power structures

A power-aware ethics of AI deals with structural constellations of agency that involve human and non-human factors, potentially including computing technology, capital interests, usage habits and new forms of digital labour. Humans, in general, play positive, productive and often pleasurable parts in these AI systems' operations, often unwittingly. For instance, as users and data producers, many people *want* to use specific AI devices and digital services because they subjectively benefit from them. At the same time, many people are adversely affected by those systems' wielding of power, often in indirect and not immediately apparent ways. These adversarial effects are often discernible only from an aggregate perspective as they represent harms to the political community or society as a whole, arising from serial collective behaviour that from an individual's perspective may be rational or even just fun.

Seeing and acknowledging the harmful effects of AI technology is therefore a matter of seeing *structures*. Emerging patterns of disparity, discrimination, exploitation, exclusion and misrepresentation are *aggregate* structural effects. In the liberal political mindset, as well as in the individualism of the modern Western tradition of ethics, these effects are easily overlooked and reduced to individual circumstances (as in, 'there is no gender pay gap, there are only women, each of whom individually are paid a competitive price for their work'). A power-aware ethics of AI must cultivate an ethos of seeing structures.

I use the term 'ethos' because whether or not structures exist is not primarily an ontological question but one of an onto-epistemic-ethical stance – a type of critical reasoning that is deeply rooted as a character virtue.

Recognising structures amounts to an ethical stance that refuses to blame the victims (by individualising the cause of inequality) or absolve the perpetrators (by viewing their individual contributions as marginal and failing to see the structural constellations from which they benefit).

In all of this, we need an up-to-date, dynamic understanding of 'structures' – one that avoids reverting to the static frameworks of structuralism. Structures are self-sustaining and dynamically stabilising patterns of differences, hierarchies and power relations that emerge when networks of individual microforces coalesce into supraindividual patterns of organisation. Hence, structures are emergent alignments wherein seemingly individual decisions, perspectives or behaviours collectively create a topology of power, exploitation and marginalisation that shapes society. Seeing structures is a matter of ethical and political openness to transcending the methodological individualism that has been so innate to Anglo-European science and humanities scholarship for at least a century. Seeing structures means acknowledging the collective, distributed and decentralised mechanisms that govern the emergence of patterns of inequality.

The *ethos* of seeing structures implies that an ethics of AI must not approach a specific piece of AI technology from a decontextualised, putatively objective angle that asks: 'Is this AI application good or bad?' or 'How should the system act to align with human values?'. Whose good or bad? Whose values? And who's asking? There is no uniform answer to such questions. A power-aware ethics of AI raises questions of distribution, such as: 'Who will benefit, and who won't?'; questions of power, like: 'Who will be gaining power and who will be exploited or oppressed by this technology?'; questions of participation, such as: 'Who is being heard, and who is silenced?'; and questions of economic interests, like: 'What are the business models driving this piece of innovation?'.

Ethics needs a two-step methodology that proceeds from critique to normativity

Of the three fundamental tonalities of philosophical discourse – ontological (describing what is the case), moral (prescribing what you ought to do) and critical (encouraging you to question what is taken as 'natural' in our times) – power-aware AI ethics must prioritise the critical mode of philosophical discourse. This kind of ethics must begin, as the first of two steps, with critique, rather than with morals or ontology. 'Critique' means philosophically and methodologically reflected *self-critique*. 'Self-critique' means recognising the genealogical contingency of one's own episteme (structures of thought). An important method of this kind of critical inquiry is genealogy. Genealogy seeks to tell the story of the genesis and becoming of the thought structures established in a particular discourse as shaped by power

relations. In this history of becoming, certain interests have prevailed over others. As a result, it becomes apparent that its product – for example, the imaginaries of AI that are dominant today – is invested with power interests, emerging from a history of power struggles. The status of technology as objectively given is thus deconstructed by showing its dependency on material, social and political conditions. In all of this, genealogy finds its truth in its transformative effect *on subjects*, not in recording any allegedly objective history of technology.

While all this is an indispensable prerequisite, an ethics of AI must not stop at the deconstruction of universals and objective viewpoints and the inclusion of different voices and positionalities. The second step of the methodology is normativity – or *daring to be normative*, as it might feel to some scholars of a poststructuralist bent. After deconstructing the alleged naturalness and objectivity of our conception of technology, and after understanding the co-constitutive relationship between technology, society and subjectivity as invested with interests and power structures, this 'fluidified' mindset must start judging things and therefore add a bit of the moral tonality as an overtone to the critical one. After all, powerful things are happening in and to our world. Technology *is* being effective. In times of alt-right and neocolonialist narratives of entitlement, we cannot afford to be agnostic, vacillating and self-questioning forever, leaving judgement and hands-on political action to others. There are at least two things that need to be judged as loudly and visibly as possible after critical fluidification. First, we need to judge the various alternatives of naming and describing the same technological phenomenon, such as different conceptions of AI, and how useful these alternatives are in identifying unethical aspects of the status quo. Second, we need to judge existing and develop novel political proposals for the progressive regulation of technology, which means that we need to support the state in its mandate to control power imbalances.

Approaches in the ethics of AI that take a shortcut from naive objectivist conceptions of technology to normative judgement quickly exhaust themselves with irrelevant and decontextualised questions (as in the trolley problem) and the sophistical quibbling of 'armchair philosophy'. The normative and regulatory proposals from proponents of this new scholasticism will fall short of the possibilities of critical ethics because they do not understand the power dynamics that need to be addressed and contradicted. Ethics as a whole encompasses these two steps: critique first, normativity second. Importantly, step 1 cannot be outsourced from ethics to other philosophical disciplines such as critical theory. This is because the purpose of step 1 is not so much to produce a specific body of knowledge, published in papers and books, but rather to foster critically reflective individuals as philosophers. Critique as self-critique is the *personal precondition* for a sophisticated, politically relevant and power-aware ethics.

An ethics of AI must engage in political debate and avoid falling prey to 'ethics washing' and techno-fixes

Ethical questions of AI concern neither the moral behaviour of artefacts (because artefacts should not be hypostatised as moral agents) nor, in general, the moral wrongs done by specific individuals. Ethical questions of AI concern reflection, judgement and appropriate governance of the role of AI technology in the exercising of contemporary economic interests and the large-scale effects AI technology has on society. The most pertinent ethical decisions in relation to AI technology are therefore *political* decisions.

In terms of AI, the most relevant ethical questions should take the form: 'Do we *want* these actors to be allowed to create/design/apply this technology in a way that has this or that effect on us all?'. In the liberal framing of ethics in Western academic discourse over the past two centuries, considering a question like this as an ethical one has not been an obvious path to take. Moral behaviour has, in tandem with the rise of liberalism, tended to be construed as a *private* matter. There is a long way to go from where we stand today back to the ancient idea of ethics as *care for the polis* – that is, to ethics as a dimension of 'political science' (Aristotle) that is concerned with good coexistence in a political community. Ambitious ethics avoids privatisation and seeks politicisation. Insofar as it is the mandate of the democratic state to control power relations, ambitious ethics works towards a critical formation of a political will as the function of an open and inclusive public debate that leads to political participation and the forging of progressive political alliances. The site where ambitious ethics gets real is not in the moral behaviour of the individual agent, but in the political movements of collective responsibility-taking.

Active politicisation of ethical issues of AI is also much needed to evade the dominant ideology of 'solutionism'. Solutionism reifies AI technology as objectively given, and continues reifying the harmful effects of that technology as 'bugs', 'flaws' and 'glitches' that can be 'fixed' or 'ironed out' by technological improvement. Researchers who engage themselves in detecting bugs and biases or measuring the 'fairness' or otherwise of AI systems are not doing ethics of AI in the more ambitious sense; rather, they provide a welcome (free) service to the industry whose practitioners should actually be doing this work themselves. Academics drawing up ethical 'guidelines' and fashioning 'toolboxes', often commissioned by the industry, do, in most cases, take the shortcut around step 1 (critique, see 5), insofar as they're engaging in a division of labour that outsources the 'ethics component' from the development teams to 'philosophers' as contracted parties. This division of labour spares development teams the hassle of achieving the personal preconditions (see 5) for a real engagement in ethical questions. Producing corporate ethics guidelines and white papers might be a tempting way for

often precariously tenured philosophers to make money, but it contributes to the hypocritical culture of 'ethics washing' that distracts from the fundamental questions of whether we, as a political community, should *want* this kind of technology – in the current way in which it is so entangled with economic and political interests – in the first place.

We must adopt a critical anthropocentrism in the ethics of AI

What is the status of the subject in an approach to AI ethics that starts from critique and tries to decentre the notion of the moral agent? After all, the goal of power-aware AI ethics is not to produce erudite tomes and papers for the library (or training data for the next large language model), but to reach human recipients in academic as well as public, political and activist debates. The reality of an ethics discourse is the *effect* it will have on thinking and feeling subjects, making them critically reflect, question the 'naturally' given, scrutinise power structures together with their peers, engage in political action and collectively take responsibility. It is thus necessary in ethics to assume that there *are* entities out there as the agents of this discourse. These entities might not be adequately captured under the monolithic notion of *the* subject, but they are human *subjects* in the sense that they are (diverse) instances of a relationship to themselves, to others and to the world. This is the consequence of a new version of an old antinomy that drives ethical work, which asserts that if there are no human recipients of the ethical discourse who *can* make a difference, who *can* make their diverse viewpoints heard, who *can* take responsibility and who *care*, then that discourse will be a fruitless and idle endeavour.

In thus assuming that there are socially and politically situated human subjects as recipients of our discourse, there is a vital quantum of (a pluralist and diversified) anthropocentrism at the heart of ethics and critical philosophy. Let's call this stance 'critical anthropocentrism'. Why does it deserve to be called critical? And why do I dare to call it anthropocentrism? First, it is critical because it starts from recognising and questioning how technology co-constitutes human consciousness, subjectivity, social relations and societal structures. It is thus not to be confused with an *ontological* theory that fixes the human as independent and separate from technology or from other species. Second, I call it anthropocentrism because it needs to be understood as a *form of discourse among entities that are susceptible to critical reckoning* – hence, human subjects in their plurality. As a philosophical stance, 'anthropocentrism' here refers to exactly the dimension of the efficacy [*Wirksamkeit*] of philosophical discourse that I hope and believe could in principle be shared with anybody on earth, but not, for instance, with a chatbot. This stance relates to the aspect of philosophical discourse that resonates with your existential status

as a subject. Critical anthropocentrism is thus a philosophical stance that seeks to engage anybody for whom there *is something at stake*.

Responsibility is greater than instrumental control

Progressive AI ethics must cultivate a collective sense of responsibility that simultaneously addresses each individual, even in the light of limited agency. This means formulating a notion of responsibility that does not rely on a merely instrumentalist view of the relationship between humans and technology – one that mistakenly assumes that artefacts are just tools whose purposes can be defined and controlled by their owners. Responsibility extends far beyond what we can control instrumentally; we must take responsibility for contingent and unforeseen effects, for technological systems that we do not ourselves own, as well as for the impacts of widespread habits and unreflective practices. As many historical and contemporary examples of 'function creep' teach us, a technological invention is never confined to its primary purpose; it grows, changes and creates its own objectives and realities. Ontologically, this is a rejection of any naively instrumentalist theory of technology, while ethically it is a call for a sense of responsibility that surpasses what we can fully control.

We must therefore understand responsibility not in the forensic sense but in a proactive one – a responsibility to *intervene*, to get actively involved in the mutual shaping of technology and society. On this point, AI ethics must decisively confront a weakness of liberalism and its tendency to privatise ethics as separate from politics: AI ethics must assess the ethical quality of the social system and the political status quo as a whole. It must raise big, systemic questions: Is this the society we want to live in? Is this a good life? Do we want to be governed like that by technology and allow ourselves or others to be exploited for the interests of a few?

At the same time, and in addition to these 'big' questions, responsibility implies a deeply personal and fine-grained ethical task. While each individual only has limited agency in their direct and indirect interactions with AI technologies, we still *are* human subjects within the sociotechnical assemblages of AI. As such, we must actively show responsibility in our various roles and positions, relating our behaviour to these broader questions. In writing this book, I could only address a limited selection of relevant subject positions: as users of technology in more affluent societies, responsibility means becoming aware of the collective impact that our usage habits and data-exposing behaviour enable, as well as of this technology's dependence on the extraction of resources, data and labour across global power hierarchies. As voters, we need to recognise the urgency of strong political regulation to mitigate these effects without necessarily hindering the collective benefits of these technologies. As politicians and regulators,

we must attain a solid understanding of AI and its business models if we are to be fully competent to protect our societies against the growing power asymmetries between the data industry and the people.

As ethicists, finally, we must engage in critique at the system level to be able to address the ethical issues posed by AI as questions of power, distributional inequality, exploitation and discrimination. Yet we must also avoid getting stuck in critique and deconstruction; we must be normative, and we must go public. Everyone needs to get involved, as we, as societies, must strive to gain more common knowledge about, control over and genuine political choices regarding the technologies that so fundamentally shape our presence.