Eva Buddeberg*

Shared Responsibility for the Development and Use of Artificial Intelligence

https://doi.org/10.1515/auk-2026-3004

Abstract: Philosophical literature generally highlights three different aspects or dimensions of responsibility: 1. the attribution of authorship of actions and the liability of the actor for these actions; 2. the attribution of a duty of care for certain tasks or areas of responsibility; and finally 3. the obligation to justify one's own behaviour and actions with good reasons. The discussion about the development and the use of artificial intelligence currently focuses on the first and, to some extent, the third aspect. However, the rapid pace of development and the now omnipresent use of AI also require ongoing critical examination and reflection on the normative framework. Responsibility must therefore be understood more broadly: with regard to the development and use of artificial intelligence, we bear a discursive shared responsibility that involves examining the development and use of AI systems against the yardstick of justice, as well as the normative framework itself that guides us in the development and use of AI systems.

Keywords: responsibility as a practice of justification; discoursive co-responsibility; structural injustice; artificial intelligence; Iris Marion Young; Karl-Otto Apel

1 Introduction

The topic of 'responsibility and artificial intelligence' is addressed across a broad range of academic disciplines, including law, sociology, philosophy and theology. For legal scholars, questions arise especially with regard to the regulation and liability of damage (co-)caused by artificial intelligence: for example, who is liable for damage caused by an autonomously driving vehicle? In addition to the European General Data Protection Regulation (GDPR), the European Directive on AI

***Corresponding author: Eva Buddeberg**, Institut für Philosophie, Philipps-Universität Marburg, Marburg, Germany, E-mail: eva.buddeberg@uni-marburg.de

Liability was adopted in September 2022; the Product Liability Directive was specifically revised with regard to AI applications, and finally, in May 2024, the European AI Regulation was adopted, which came into force on 1 August 2024 and is considered the world's first comprehensive regulation of artificial intelligence. But despite what are now far-reaching legal regulations, which attempt to balance the protection of fundamental rights against the promotion of innovation, questions continue to arise at a legal level regarding, for example, the interpretation of new legal terms (such as 'risk areas') that appear in the various directives and regulations. In tandem with this proliferation of new technical terms, artificial intelligence itself is constantly evolving and finding new applications which in turn require a review and adaptation of the legal framework.

In the field of *machine ethics*, there is a fundamental debate in *philosophy* concerning the extent to which AI-based systems themselves are understood as moral actors, i.e., whether they can be seen as the bearers of ethical or moral decisions[1] and can therefore be responsible for their actions and any damage they cause. For now, the majority position assumes that since they lack the relevant prerequisites such as freedom, higher-level intentionality (cf. Dennett 1976) and the ability to act according to reasons,[2] AI-based systems themselves cannot be bearers of responsibility. However, this is not to deny that artificial intelligence systems have the ability to exert influence in morally relevant ways. For example, the discussion also addresses the extent to which AI systems can support human moral decision-making by analysing, structuring and evaluating relevant data.[3] This can go so far as to give the impression that the AI systems used are actually the authority responsible for the outcome of the decision.

To be able to hold someone liable in situations in which artificial intelligence is involved – especially when using artificial intelligence in high-risk areas[4] – there

1 Pereira and Saptawijaya 2016. On the question of the extent to which AI-based systems can act, see Misselhorn 2018.

2 Specifically with regard to responsibility, see for example Buddeberg 2011, III.1.2; Vargas 2013, Part II. On the ability to act on reasons, see e. g. Kohler 1988; Davidson 1980; Larmore 2008; Scanlon 2013 as well as the articles in Stoecker 2002. In this context, it is also important to discuss the role of the first-person perspective or the participant perspective. See for example Stoecker 2003; Habermas 2007 as well as Nida-Rümelin and Weidenfeld 2022, 17; on the ability to act on the basis of reasons, see this, 22-3, 28.

3 Here we speak of machines as "moral counsellors" (Misselhorn 2018, 72). In actor-network theory, it is assumed, that human and non-human actors (like technological objects) form dynamic networks, collectively shaping social and natural phenomena.

4 According to the AI Regulation, this includes the areas of health, safety, fundamental rights and the environment as well as AI systems for influencing political election decisions and systems of recommendation on social media platforms. (https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, last accessed on 18 March 2026).

is a plea, for example, to include a human intermediary – a 'human in the loop' (HITL) – as a precautionary prerequisite for oversight, and to assume responsibility for decisions made. However, there are a number of basic requirements that must be met for this to be effective: the decisions proposed by the artificial intelligence must be comprehensible to the person in question[5] which, in turn, means that the HITL must be sufficiently familiar with how the specific artificial intelligence works. Furthermore, the HITL must not only know, but also be able to give all due consideration to the relevant moral and legal aspects of each pending decision. And the entire process must be relieved of time pressure to such an extent that the person can perform an adequate critical review.

In my view, the aforementioned legal provisions for regulation and liability, in conjunction with the moral or action-theoretical proposals concerning the involvement of a human supervisory authority in relation to the attribution and accountability for decisions made by artificial intelligence are to be welcomed, even if it is unclear in many cases whether and how such a HITL with the corresponding competencies could in fact be deployed. If no one can any longer understand why and how an artificial intelligence arrives at a particular conclusion, then a human intermediary can no longer justify its operations and results either; at best, they can only explain the intentions behind its use or how it was used by others.

However, as I wish to show, the question of responsibility with regard to artificial intelligence also involves, given the relentless and rapid pace of developments in artificial intelligence observed by everyone and its now ubiquitous use at a higher level, an ongoing critical examination and reflection on the development and use of AI systems against the yardstick of justice. My thesis is that the now omnipresent use of AI in almost all areas of life does not in any way resolve or mitigate existing injustices, but rather exacerbates them, in some cases dramatically. Referring to the ideas of Hans Jonas, Karl-Otto Apel had already drawn attention to the dangers arising from scientific and technological achievements, which, with our increased power over nature, now pose a threat to human existence and the planet. Iris Marion Young, meanwhile, focused her attention on the problems of structural injustice arising from complex, collective contexts of action. Building on their theories, I shall, as a first step, argue that responsibility should not be understood primarily as liability, but rather as a discursive practice: retrospectively justifying our actions to others and prospectively orienting them in such a way that they can be justified to the people who are affected by them (2). I will argue in a third step (3) that this is not to be assumed by one person alone, but *jointly*. Responsibility understood in this way implies, as I will then argue in a fourth

5 However, this is not possible at all with AI involving 'black box' processes. That is why there is a demand for explainable AI. In such cases, however, the HITL might also lose some of its necessity.

step, an ongoing, critical examination of its application against the yardstick of justice, with the aim of uncovering injustices and counteracting them (4). This can also lead to a re-examination of the normative framework and the associated understanding of oneself and the world, modifying them if current circumstances require it, or rethinking them entirely. To this end, the moral and ethical boundaries of the development and use of artificial intelligence must be continually re-examined and discussed in dialogue with others, including experts and those affected in such circumstances, whilst risks and undesirable developments must be identified as precisely as possible so that we can continue to counteract them (5).

2 Responsibility as a Practice of Justification

In philosophical theories, responsibility is often equated with 'attribution' ('Zuschreibung') or 'liability' ('Haftung'). However, this ignores two other aspects that are already part of our everyday understanding: having responsibility not only means that an action or task can be attributed retrospectively to someone and that this person is liable for it: e.g. 'The developer did not sufficiently consider the risk of discrimination against people of black skin colour when programming artificial intelligence. Therefore, they are responsible – i.e., liable – for the damage caused.' It can also mean that someone is obliged to take a certain action or perform a task in the future: e.g. 'A start-up for the development of software based on artificial intelligence is responsible, i.e., obliged to comply with the guidelines of the GDPR and the AI Regulation.' Responsibility also implies an obligation to respond, i.e., to justify: we *answer to someone* by honouring the claim associated with every action that our action is justified by providing intersubjectively comprehensible and understandable reasons to all those affected by our action.[6] Accordingly, being responsible means that one's own actions and behaviour in a forward-looking manner must, to the extent that they affect others, entail that one can answer legitimate questions about them satisfactorily.[7] This can include the parties involved coming to an understanding about their behaviour and the motives, intentions and reasons that underpin their actions.

6 This aspect is also emphasised by Baum et al. 2022, for example. Though, it is not only other people who are affected by our actions but non-human nature as well. However, our ability to communicate reasons is limited to other people. And to them we have to justify our treatment of and impact on non-human nature.

7 There may well be exceptions to this. For example, German law recognises a constitutional right, to stay silent in the light of accusations against oneself and any government action against that right would be considered unlawful. However, these exceptions are themselves subject to the requirement that they must and can be justified (in cases of doubt).

This willingness and ability to justify one's own actions on the basis of reasons, which is demanded by the concept of responsibility, already arises from the philosophical concept of action as intrinsically rational or reason-orientated;[8] therefore, it is not an external imposition on actors. This is because, generally speaking, action is intentional and based on reasons. It takes place within a world that is shared with others, co-constituted and linguistically structured by them and is therefore generally associated with the claim that it can be justified to others (i.e., those affected by this action).

In this way, every acting person is retrospectively responsible for what they do and do not do, and, looking ahead, for the fulfilment of certain tasks. However, this is not the whole picture; individual actions can be isolated only superficially and somewhat artificially from the complex network of facts, events, subjective convictions, wishes, etc., that make up the context of the action. It may be possible only in highly standardised situations to identify an action and its consequences as the *object* of responsibility – although even here, the inclusion of various intentions and reasons for action can expand what at first glance appears to be a simple action into a whole complex of actions. In principle, all behaviour that exposes an actor to the gaze and questions of (possibly) affected co-subjects can and should be justified to them.

People may have to justify themselves *to all those who* are or could be directly or indirectly affected by their actions or behaviour. This applies both retrospectively and prospectively. The more standardised the actions or human behaviour, the more clearly it seems to be specified who is affected at all and to what extent and therefore they have such a concrete "right to justification" (see for example Forst 2014) or, pre-emptively, a "right to consideration."[9] If either the context or the interpretation of an action changes and another person or institution asks for reasons, this person or institution must also be given an answer. The claim to justification can also be asserted on behalf of other persons or institutions.

In addition, further questions can be asked, such as *what criteria* are *actually* used to evaluate the reasons given by the agent making the decision, or *what* guides the agent in selecting their reasons. This is because, in addition to personal objectives, our actions are guided by professional or general interpersonal expectations or context-specific rules and norms, legal rules, as well as moral or political norms.

---

8 Empirically speaking, this is of course a partially counterfactual assumption, because in reality people are motivated in their actions by very different sources that can only be explained retrospectively and by no means completely as rational.

9 The phrase 'right to consideration' is intended to emphasise that we must always consider others affected by our actions, and behaviour in such a way that others can accept them as justified – regardless of whether they actually ask for such reasons.

In the discourse on responsibility, these provide the normative standard to which people orientate their actions and against which the behaviour or actions of others are assessed. Without such a normative framework,[10] it is unclear how responsibility can be attributed, assumed or evaluated in any way at all. Often, initially, this framework is merely implicit, becoming explicit or being drawn into critical reflection only in cases of doubt or in new contexts of action. With regard to artificial intelligence, as already mentioned, this has been happening in Europe at a legal level, as, for example, with an extension of the GDPR to new fields of application, with the revision of the Product Liability Directive, the Directive on AI Liability, and through the entirely new AI Act.

3 Apel's and Young's Conceptions of Shared Responsibility

At least at the legal level, the normative framework has thus clearly taken shape. However, in my view, the critical examination and reflection of this framework is still required as part of a *social co-responsibility* that precedes individual responsibilities for certain actions (see for example Apel 2001; Apel 2000; see also my account in: Buddeberg 2011, 99–104). Karl-Otto Apel has developed such a concept of responsibility in the debate on "humanity's responsibility for the consequences (and collateral consequences) of its collective actions on a planetary scale" (Apel 1988, 42, my translation). As members of a communication community, Apel claims, humans are responsible not only for "any particular tasks" or actions, but also – and already – for "*uncovering*, *discovering* all problems in the lifeworld that are capable of discourse, and for *discussing them*. They are also responsible for bringing about such discourse and for *solving the problems*" (Apel 2001, 107–8, my translation) and "for *ensuring that* [these tasks] are assigned" (Apel 2001, 109, my translation). Thus, as I understand Apel, to bear responsibility essentially means to participate with others in the processes of public communication.[11] This enables the identification of relevant tasks, the determination of those interests and needs that are to be taken into account, and the definition, distribution or even delegation of tasks and competencies. Similarly, the underlying norms for action can be critically reviewed,

10 This normative framework itself can contain very heterogeneous norms, which can at best be placed in a clear hierarchy *in an ideal case*.

11 Whether in the future communication will still be useful as a key to identifying AI problems, and if so, in what form and to what extent, can be critically examined in view of the transformation of communication by AI (bots, LLMs, etc.).

expanded or supplemented, taking into account the interests and needs of all stakeholders – and this still seems to be a central aspect of responsibility in regard to the development and use of artificial intelligence, even if it has now been widely taken into account in European law. Furthermore, as Apel emphasises, 'responsible' institutions can be set up to ensure that these further considerations are taken into account. The Commission of the European Union has already done this, for example, with the establishment of the "European AI Office,"[12] as well as with special forums and expert groups.

This "concept of the *co-responsibility* always already ('immer schon') presupposed of all human beings" that is postulated by Apel "by no means excludes the traditional concept of *individually attributable responsibility*" (Apel 2000, 27, my translation). Rather, co-responsibility forms a kind of basis for every concrete individual (and collective) responsibility to be assumed or transferred (Apel 2000, 27). As co-responsible members of a communication community, people do not act in isolation from others; rather, they can only act together, and are also obliged to *communicate to these other people*, not only in regard to their own actions and their consequences, but also in regard to the possible – and sometimes far-reaching – consequences of the actions of all those involved. In particular, the aim is to identify, avoid or at least minimise at an early stage any possible negative consequences resulting from the unfavourable interplay of various individual actions.

According to Apel, each person bears co-responsibility not only for the "discovery or identification" of all morally relevant problems of the lifeworld, but also for solving them "in argumentative discourse" (Apel 2000, 37, my translation). Since all problems of the lifeworld are to be solved together in argumentative discourse, all affected members of the discourse community may *potentially* demand, examine and judge the reasons for an action. Ultimately, responsibility is practiced together through this process of demanding, examining and judging reasons.[13] Only a broad-based socio-political discourse can ensure that decisions are not made by individual actors, (lobby) groups, or the 'market' based solely on economic interests, for example, but that the reasons and interests of all those affected by them are taken into account.

This also makes it clear why the concept of a *primordial* co-responsibility does not recognise an *authority* as a fixed institution to which one has to answer. Rather,

12 https://digital-strategy.ec.europa.eu/en/policies/ai-office, last accessed on 18 March 2026.

13 Of course, actions in everyday life are often not questioned at all, nor are those who perform them asked for reasons; otherwise, we would not be able to cope. Apel's point is rather that even those who are affected by an action or witness an action always influence it (or can do so) through their (expected or actually expressed) acceptance or criticism of that action and are therefore also jointly responsible, naturally to varying degrees, depending on their power to act.

through the procedure of discourse described above, every member of the communication community – alone, or together with others – is potentially regarded as such an instance of responsibility. The concept of discursive co-responsibility thus includes, in the current context of technological development, the ongoing intersubjective critical reflection and understanding of one's own actions and, where applicable, of the norms underlying these actions, as well as the regulation and coordination of this process. This extends past the concept of simple accountability or the concept of duty.

Less ontological and therefore, perhaps, less idealistic and less abstract, is Iris Marion Young's (Young 2006) well known argument in an overtly political register that responsibility should not be understood solely as individual liability but should be supplemented by a "social connection model." Young very specifically develops this supplementary model of social connections using the example of certain forms of exploitation in companies, which are based on structural injustices. She argues that the model of individual liability leaves open the possibility of absolving oneself of responsibility either by invoking insufficient knowledge (Young 2006, 116), by hiding behind those for whom one works, or by referring to one's own powerlessness within the complex interplay of various individual actions, which are also determined by market competition and high economic pressure.

By setting minimum labour standards, state (or supranational) law can have a regulatory effect, but any such regulations can in turn be undermined by incompetent or corrupt authorities (Young 2006, 117). There is no doubt that this problem can be blamed on the respective states (even if at least poor states defend themselves – in ways that are not entirely unfounded – by pointing to onerous economic constraints and the fact that they would otherwise not be able to survive on the world market) (Young 2006, 118). Crucially, however, the liability model is not sufficient to counter structural injustices.

Young therefore proposes to complement this classical model with her model of *social connection* (Young 2006, 118). In so doing, she wants to clarify that all actors whose actions involve them one way or another in processes that result in or maintain injustice, as well as all those who benefit from corresponding institutions and orders, are *jointly responsible* for structural injustices. And since, as human beings, we always participate with others in "a system of interdependent processes of cooperation and competition" that has unjust effects, "through which we seek benefits and aim to realize projects" – processes that in the contemporary world extend far beyond nation-state borders (Young 2006, 119) – we are also all responsible. This means that in regard, for example, to industrial exploitation, responsibility extends to states, producers, consumers, and even exploited workers.

Young's model of responsibility is therefore also less concerned with establishing the retrospective liability of individual actors and with sanctioning them

for the damage they have caused; instead, her primary intention is to emphasise that all actors involved contribute in some way to the emergence or continuity of structural injustices, and cannot simply absolve themselves of this by shifting responsibility to other actors (Young 2006, 119). Rather, *all* those involved must take responsibility, in the sense that their actions should contribute to overcoming these structural injustices. Thus, Young highlights the aspect *of care* already contained in our everyday use of responsibility; being responsible means *taking care* to address or combat injustices, and to avoid any possible damage mainly by changing political, economic and societal structures. And even though Young develops her model using the example of the exploitation of workers in the workplace – a concept that can easily be applied to conditions in the AI industry –, her general concern is to combat all forms of structural injustice precisely because responsible behaviour is generally measured against the norm of justice (Buddeberg 2011, 277–83).

This also means that one must not simply accept existing norms and regulations as a given standard for evaluating actions, but one must critically examine whether and to what extent such norms allow, promote or even create injustice (Young 2006, 120–1). In her view, this also requires a joint commitment, particularly in the form of public understanding, to coordinate actions and to shape and organise social relationships more justly. At the same time, others must be convinced on the one hand that (unacceptable) injustice exists, and, on the other hand, that collective action can change social practices and institutional regulations and priorities in such a way that any such injustice is prevented in the future. Although criticism of state (as well as supranational) institutions and regulations is necessary to bring about or institutionalise change, it would be misleading to rely solely on laws being amended. Rather, this requires constant and persistent pressure from those affected, as well as from a committed democratic public.

In her model, Young also stresses that even those who are victims of structural injustice bear responsibility. Although they may have only very limited power to act, they are – or should be – particularly motivated to change the existing conditions. They also have knowledge and interpersonal connections with each other that are helpful and, above all, efficacious when it comes to mobilising against the existing injustices, and which can help to strengthen each of them as independent actors. However, they are often dependent on the (material) support and help of other, more privileged actors (Young 2006, 124). Thus, according to Young's *model of social connections*, all participants, *including the victims,* share responsibility, albeit not to the same degree and in the same form but, depending on their own position and power to act, by contributing different forms of knowledge, motives and skills in exchange with others and by taking on different tasks in order to recognise, criticise and ultimately overcome existing and possibly developing injustices

(Young 2006, 125; on the relationship between power and responsibility, see Buddeberg 2022).

4 Shared Responsibility in the Fight Against Injustice

Why does it seem promising to refer to these two concepts of *joint* or *shared* responsibility when considering questions about the development and use of artificial intelligence? In my view, especially given that such new developments are mobilising far-reaching changes in our social world, it is crucial to think beyond liability for a particular damage that may arise or be caused by artificial intelligence. To this end, legally binding regulations have recently been created, particularly at the EU level, which apply to almost 450 million people and the third largest economic area after the USA and China, legal measures that are not solely based on liability, but on regulation. That is significant progress. But beyond this, responsibility must also be understood here as an ongoing, collective social task.

The first step is to identify and acknowledge the injustices that arise from the development and use of AI – or which are, in some cases, dramatically exacerbated by it – so that we can then take further steps to counteract them. For, to date, the development and use of AI systems have still been viewed with some caution from the perspective of justice, both in broader societal and academic discourse. A notable exception to this is the study *Feeding the Machine. The Hidden Human Labour Powering AI*, published in 2024 and based on many years of field research by James Muldoon, Mark Graham and Callum Cant. This study focuses on the often exploitative and degrading working conditions of people who – frequently unseen by end-users – enable the development and provision of products marketed as artificial intelligence.

It is striking how closely the descriptions of working conditions – for example, those of 'annotators' involved in preparing datasets for the development of AI systems (see in particular Muldoon et al. 2024, ch.1) – in countries, particularly in East Africa and South-East Asia, resemble those of the workers in sweatshops examined by Young. The workers employed here also suffer from exploitative working conditions, characterised by largely very short-term contracts, extremely long working hours, poor pay – at or below the subsistence level –, child labour, workplace harassment, extreme pressure and increasing, constant surveillance, as has been the norm in industrial manufacturing under capitalist conditions since the colonial era. There are few, if any, alternatives to this army of 'big data' suppliers, whilst their work can easily be taken over by other workers anywhere in the world, as it requires comparatively little infrastructure. (Muldoon et al. 2024, 39). This

benefits both the tech companies and end users, who are kept under the illusion that they can access the world's knowledge without making any effort themselves and at minimal cost.

It is becoming apparent that the datasets required for the development and training of AI systems are increasingly being (or can be) generated synthetically. However, simply because these datasets are expensive to produce on the one hand and harbour a considerable risk of malfunctions or defects on the other, humans will certainly remain involved for the foreseeable future, amongst other things to test, repair, and adapt them where necessary (Muldoon et al. 2024, 36). And unlike, for example, in the production of cheap clothing, very few end-users of AI systems are aware of how much human labour has gone into developing these systems and the economic power structures on which they are based, before they are made available for anyone and everyone to use, something that is being deliberately obscured by the companies providing these systems (see e.g. Muldoon et al. 2024, 8, 165, 171).

Far better known, however, are the poor working conditions faced by workers in warehouses as also described in the study by Muldoon et al. – such as those operated by retail platforms like Amazon – which are largely driven by the use of AI systems. Many workers perceive their work as 'meaningless' (Muldoon et al. 2024, 115) and suffer from ubiquitous surveillance (Muldoon et al. 2024, 118–27), whereby even voice recordings, emails and other data sources are analysed and workers' motivation levels are assessed accordingly (Muldoon et al. 2024, 127). At the same time, the major online retail platforms are constantly attempting to optimise employee performance through the use of AI systems (for instance, in 2020 Amazon reportedly considered using AI to design work schedules in such a way that different muscle groups would be utilised in an optimised manner, see Muldoon et al. 2024, 118–9). On the other hand an ever-increasing division of labour is simultaneously being used to replace workers with more efficient, yet even cheaper, human or machine labour (see Muldoon et al. 2024, 123).

Even within the infrastructure central to AI systems, such as the gigantic data centres they require, many workers are often employed under precarious contractual terms (Muldoon et al. 2024, 67). Another aspect of structural injustice is that the entire data processing sector itself is controlled by a very small number of (predominantly US) companies (Muldoon et al. 2024, 79). Their power lies in the computing power of the AI infrastructure and its economic potential, which they can use to attract top talent for technological development (Muldoon et al. 2024, 80). Last but not least, the dramatic consumption of water and energy required by the entire tech industry, with its gigantic data and computing centres (Muldoon et al. 2024, 118–9), leads to further existential injustices. This initially runs diametrically counter to the frequent promises that AI systems can play a significant role in halting climate change.

Apart from working conditions that are often characterised as exploitative, AI systems such as LLMs likely contribute – and perhaps more so than other products from sweatshops – to the emergence or perpetuation of new injustices, for example because the underlying data is inherently 'biased' (Muldoon et al. 2024, 58–61; see on this point also the article of Hinsch 2026). This is simply because certain social groups are underrepresented within it (Muldoon et al. 2024, 63), because their developers do not represent the global population, and because the data and the algorithms based on it are not always reliable or accurate. Furthermore, there are still no adequate safeguards in place, nor are any foreseeable, to prevent these LLMs from being used in an uncontrolled manner for unjust, discriminatory or dangerous – and in some cases criminal – purposes.

The questions of justice raised so far arise in a different context when it comes to AI compared to, for example, cheap clothing manufactured in sweatshops, as the excessive purchase of such items does not satisfy any existential needs – at least not for the vast majority of consumers in wealthy countries. The use of AI systems has instead spread at a dizzying pace across almost all areas of life in recent years and is now virtually impossible to avoid. This gives rise to both hopes and fears. On the one hand, AI systems are believed to be capable of countering the skills shortage arising from the ageing of societies, diagnosing fatal diseases earlier and more reliably, or playing a decisive role in the fight against climate change. On the other hand, we must consider their use for political propaganda, the massive job losses and the devaluation of traditional qualifications they risk provoking, the use of AI systems in the development of armaments and warfare, and the potential threat of artificially created viruses. Because the use of AI systems can have both negative and positive effects, and because they are now virtually unavoidable in many fields of application, it seems much more difficult to counteract the injustices that are either exacerbated or emerging as a result of their use.

This difficulty is aggravated by the fact that the power of the corporations developing and operating AI systems has reached unprecedented levels. In financial terms alone, the market power of other large (multinational) corporations falls far short of that of Meta, Nvidia and Alphabet, which own the leading American AI companies. At the same time, as is well known, this enormous power imbalance is by no means limited to the financial clout or market power of these companies. Rather, through AI-supported platforms and social media, etc., they are increasingly controlling public discourse and, by this means, politics as well. Even the power of a media mogul like Silvio Berlusconi seems, in retrospect, to be of little significance compared to that of an Elon Musk, who attempts, even in other countries, to influence political orientations via media such as X, which are largely controlled by AI

systems. And politicians like Donald Trump, who is in close contact with leading figures in the tech industry, are not expected to curb the power of these corporations to any significant or noticeable extent in the near future, given that their political success relies heavily on the use of AI systems on social media.

The enormous imbalance of power in favour of the AI industry is becoming even more significant due to economic and geopolitical competition with countries such as China and concerns that they could soon overtake the Western world.[14] In summary, it must be noted that, as the authors of *Feeding the Machine* also highlight, the development of AI systems is driven almost exclusively by profit expectations and the power interests of a small number of large corporations and their investors, and only takes their users' needs into account to the extent necessary to ensure the sale of their products. The interests of potential consumers belonging to marginalised communities are scarcely taken into account, and concerns about whether the decisions or actions underpinning these developments are morally justifiable are often dismissed as undue interference in what should be a free play of market forces. Even the repeated calls from certain tech giants for legal regulation (Muldoon et al. 2024, 143) to ensure safe, responsible AI seem, at least in part, to be driven more by profit than by moral concerns.

The detailed empirical analysis of working conditions for various groups of people involved in the development and deployment of AI systems, as presented in the study *Feeding the Machine*, clearly demonstrates the extent to which even this current phase of 'technological progress' perpetuates the economic logic of capitalism and its colonial legacy. In keeping with Young's concept of *shared responsibility* in combating structural injustice, these authors also call for collective resistance. And like Young, they regard workers – as victims of the injustices highlighted – as key actors; indeed, perhaps even more strongly than Young, they emphasise that their experience of injustice must be at the forefront of this resistance; for if they do not rebel, no one else will (Muldoon et al. 2024, 185).

Workers, whether in East Africa or the deindustrialised regions of North America or Europe (Muldoon et al. 2024, 132), are not merely passive victims, but are in fact the ones who may be the most motivated to stand up against injustices, because these injustices concern their wages, their employment contracts, the monitoring they face, their workload, the drudgery of their work, and so on. Yet these workers can only succeed collectively. And this requires, first and foremost, solidarity among workers and, building on this, the formation of alliances on a global scale – such as

14 This is also frequently cited as an argument against the European AI Regulation.

the establishment of transnational trade unions[15] – in order to improve networking and break open the 'black box' of human labour involved in AI systems, thereby also exposing the underlying injustices (see e.g. Muldoon et al. 2024, 176, 187). Building on such "collective power of worker organisation" (Muldoon et al. 2024, 188), civil society could and must then play a key role in ensuring that companies are held to account, by urging politicians to adopt stricter regulations for AI companies and, I would add, opt for higher levels of taxation. Furthermore, it is necessary to design and test alternative models for AI companies in which workers have a greater say and more decision-making power. Ultimately, this should lead to challenging the injustices of the entire system within which these companies operate: global capitalism.

The discussion of these proposed steps addresses the difficulties that are likely to arise, draws on historical examples from other contexts, and puts forward concrete proposals for implementation. Repeated reference is made to the deliberate concealment of the human labour involved in AI systems (Muldoon et al. 2024, 170–1), which makes it difficult to develop counterstrategies and resistance activities, partly because it allows for the tacit replacement of workers by others. The deliberate concealment of actual production conditions is also intended to prevent a company's reputation from being called into question, as this ensures that exploitative working conditions are not associated with the company (Muldoon et al. 2024, 172). It is only through collective action that structural injustices rooted in power imbalances can be countered. In this regard, civil society can help to better coordinate the resistance activities and strategies of the workers directly affected, thereby building greater pressure, including towards political and legal improvements; however, this pressure is most likely to be effective when it is based both on the collective power of workers and, at the same time, on the institutionalised power of effective legal regulations (Muldoon et al. 2024, 197).

These ideas, developed by Muldoon et al., also bear many similarities to Iris Marion Young's outline of how people who both contribute to, and suffer from, structural injustice share the responsibility for counteracting these injustices. However, this also raises the question of how successful the steps proposed here can be. Even in the case of the Supply Chain Act, which the authors also cite, it is evident that measures against structural injustices that are already relatively well known to the public and based on significantly smaller power asymmetries tend

15 As the authors discuss, these can be successful primarily through two different mechanisms, namely 'blocking the flow' and 'sounding the alarm' (Muldoon et al. 2024, 169), as well as possibly through a combination of the two. Yet, as e.g. Honneth (2025, II.6) demonstrates, current working conditions clearly stand in the way of both political struggle and participation.

to be scaled back, with the argument that the local economy must remain competitive. While this may be explained by economic uncertainty caused by various crises, it is certainly also linked to the fact that the political landscape has currently shifted dramatically to the detriment of efforts to combat, or even merely alleviate, injustices.

How, one might ask, is it then possible to succeed precisely against the most powerful corporations, which, on the one hand, attempt to shape public discourse – and, by extension, the political legislature and the executive branch – through social media, which they largely control, if not dominate? The current climate certainly does not suggest that the fight against the destructive effects of *Big Data* on the balance of power within society will be successful. However, as the authors argue: "Rights and protections have been won because people throughout history have demanded change." (Muldoon et al. 2024, 207) The duty to combat injustices is not determined solely or primarily by the likelihood of success. It exists because power imbalances are unjust, and collective action against them can at least help to prevent these injustices from continuing to grow also because it keeps alive awareness of issues of justice across a wide range of areas in the development and deployment of AI systems. Following on from Apel, this could be seen as a particularly relevant aspect of discursive shared responsibility at the present time.

However, this shared responsibility is by no means limited to highlighting the structural injustices underlying the development and use of AI. It also concerns the physical threats posed by AI systems. Due to their military use, which is increasingly shaping current wars and 'special operations,' fears regarding the destructive potential of AI systems and the associated existential threats have become alarmingly topical. And there is absolutely no guarantee that AI systems used for military purposes are in the hands of the 'right people.'[16]

5 Examination of Our Normative Principles

Not least in view of the possibility of such dangers, which are all too evident in their brutality, we must critically examine, beyond the standard of justice, which norms and which worldview and self-understanding will (in the future) guide the development and use of artificial intelligence. And this leads back to the processes of justification and understanding in which norms are developed, and whose significance Apel in particular has emphasised. Admittedly, with the extension of the

16 At the same time – and here we see a striking parallel with the nuclear arms race during the Cold War – the potential use of AI systems by the other side seems to be forcing everyone else, at the very least as a defensive measure, to develop, acquire or even deploy such systems themselves.

GDPR, the Directive on AI Liability,[17] the revision of the Product Liability Directive and, in particular, the AI Regulation which came into force last summer, a comprehensive legal framework exists within the EU; yet these regulations and their practical implementation still require thorough scrutiny, not least because they are new.

The aim of the European AI Act is to mediate between the protection of fundamental rights on the one hand and the promotion of innovation on the other, for example, by distinguishing between different areas of risk and determining various restrictions accordingly.[18] There are obviously different standards at play here. While the protection of fundamental rights clearly has a place in the normative structure of justice, the promotion of innovation appears to be more a norm of the economic sphere.[19] At the same time, however, political arguments can also be put forward in favour of what, at first glance, appear to be more economic consideration, as, for example, in the idea that competition has thus far been an important guarantor of democracy, political security and peace. Is there an opportunity to introduce EU regulations worldwide through the so-called Brussels effect? How should we deal with the fact that entrepreneurs like Elon Musk exert massive direct political influence in various countries, directly and with the help of AI systems? Recent developments give rise to the concern that Europe will soon be less of a pioneer in stricter regulation and more of a laggard with less control.

The question of how different norms should be taken into account, and which have priority where and in what form,[20] is, as I wanted to suggest here, not easy to answer, since it remains context-dependent and thus changes even in the face of the claim that responsibility is measured by the standard of justice.[21] Although the EU itself has only a rudimentary democratic constitution and can therefore

17 It was proposed in September 2022 by the European Commission (2025), but the original proposal was withdrawn in 2025, meaning it has not yet come into force.

18 https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, last accessed on 18 March 2026.

19 The promotion of innovation is, however, part of the EU treaties (e.g. in Art. 179 TFEU). Here it is not only within the ‚economic sphere', but partially, as scientific freedom, considered a fundamental right itself (e.g. in Art. 5 German Basic Law or Art. 13 of CHREU).

20 See the article by Gerhard Wagner (2024) on the competition between different AI legal systems: https://www.faz.net/aktuell/wirtschaft/unternehmen/kuenstliche-intelligenz-die-eu-als-globaler-regulierer-19560904.html, last accessed on 18 March 2026. From his point of view, it is also questionable for reasons of competition whether the development or use of artificial intelligence should be regulated so strongly ex ante or whether it should be regulated more strongly ex post through liability.

21 It is not only artificial intelligence that is undergoing rapid development; the (global) political landscape is also constantly changing, and with it the balance of economic power. At the same time, many other questions remain unanswered or require further examination, such as which AI application falls into which risk category, on the one hand because some risks may not yet be known, and

rely only to a limited extent on the collective decision-making of its citizens in its directives and regulations, the debates both in the public sphere and within parliamentary bodies are by no means insignificant for their formation and practical implementation. This means that all citizens, in their twofold capacity as users and as members of the political public, have a responsibility in this regard too. Admittedly, the development and use of artificial intelligence do not, in every instance and without exception, lead to structural injustice or other risks. However, a critical examination of the scope and design of regulations, as well as of liability in the event of undesirable outcomes, requires that all relevant perspectives – and indeed the underlying normative framework – be taken into account. It is far from certain whether everyone involved in any way is aware of this responsibility and acts accordingly. It seems more realistic to assume that many are guided *not* by considerations of justice or other universal norms, instead primarily by their personal interests.

Discursive co-responsibility in the development and use of artificial intelligence therefore requires, first and foremost, the identification and exposure of structural injustices, as well as the ongoing and reflective scrutiny of the normative framework, not least against the yardstick of justice. Furthermore, in view of the changes to our everyday lives brought about by technological developments, we must also ask what understanding of ourselves and the world is guiding us in our daily and professional engagement with the opportunities opened up by new technologies (see also Weber-Guskar 2024, 24).

For example, what does the use of AI-supported medical diagnostics mean for a doctor's faculty of judgement and therefore also for our image of medical professionals? Under what circumstances will they continue to permit themselves to make decisions or provide treatment based on their own responsible and professional judgement, contrary to diagnoses based on findings evaluated by AI systems? Should they still be able to do this at all? Could the use of artificial intelligence therefore lead to a loss of critical judgement altogether? And how should such a potential loss be assessed, particularly in the field of medicine? How, then, can the ability to make independent judgements be developed at all?

This question must now be asked in virtually all areas of upbringing and education. It is true that human learning and knowledge have repeatedly undergone epoch-making changes, and that human learning is always adaptable. Yet many people working in education and training are currently asking themselves what impact this will have on learning if it is no longer based primarily on overcoming problems and obstacles, as has been the case until now, but instead relies largely on the

on the other because some fields of application are only just emerging. Here, too, ongoing scrutiny – indeed, even anticipation – is required.

competent use of equipment. So far, the focus has remained on the use of artificial intelligence in only certain fields – for example, in medicine, where the aim is not to replace treating doctors, or in education, where the aim is not to replace teachers with artificial intelligence systems. And so far, no such replacement is foreseeable.

However, some of the questions that Eva Weber-Guskar, for example, raises with regard to emotions and artificial intelligence also need to be asked in relation to other (im)possible (further) developments of artificial intelligence. What would we think if, for example, at some point in the future we were to be advised[22] or treated by artificial intelligence instead of doctors? If the aim were not to use artificial intelligence for certain subtasks, but for our medical care or education entirely, what would this mean? In addition to various forms of empathy, any such move would also require artificial intelligence itself to be endowed with forms of judgement and the ability to act, and to hold responsibility. As I said, we currently appear to be a long way from this. But should we be working on such developments at all?

In my view, shared responsibility in the form of socio-political debates requires us, first and foremost, to engage in a critical reflection and discussion with others to establish what we *cannot* or *should not* want in the first place. In addition to 'emotional' artificial intelligence systems capable of more than just reproducing emotions, this would include artificial intelligence endowed with agency and responsibility, which has rights, or which represents an entity that we regard as a genuine – and not merely hypothetical or fictitious – partner in discourse. Such delimitations would have to be justified in a nuanced manner, which also means always in confrontation with reasoned counter-opinions. This kind of debate would presuppose an interested and critical public. Furthermore, co-responsibility with regard to artificial intelligence also involves, in line with Hans Jonas' heuristic of fear, guarding against fictions that no longer recognise the boundaries between humans and machines. It also involves, where these boundaries are in fact called into question, exposing those developments and pointing out the dangers that they may hold for human interactions, as well as slowing down the development of AI and the changes it is expected to bring about through further adjustments to the legal framework, so that we do not lose control over it.

22 In the field of psychotherapy and similar forms of counselling, this is already very much a reality, and is perceived as helpful by those affected, even though, strictly speaking, its content amounts to nothing more than the reproduction of mainstream views from the counseling literature with which the system has been fed. See also the article of Kropp and Renner 2026.

6 Conclusions

Building on the conceptions of Karl-Otto Apel and Iris Marion Young, I have argued that responsibility should not be understood solely as liability for the consequences of AI-enabled technologies, but also as a discursive shared responsibility to examine and reflect upon the development and use of AI systems within a social discourse, measured against the yardstick of justice. It is clear that the use of AI systems is linked to structural injustices. This is evident even at the development and deployment stages, which are largely controlled by large corporations with oligarchic structures. As I have shown, drawing on Muldoon et al., the structural injustices here bear many similarities to the developments associated with colonial exploitation that have been evident since the early stages of capitalism; and, just as was the case then, the power asymmetries on which they are largely based are becoming more pronounced as technological systems evolve.

This can only be countered through collective action, through alliances of those who are not on the winning side and who, even if they themselves may benefit from technological advances in certain areas, do not allow themselves to be blinded by supposed progress. The key issue remains assessing the development and use of AI against the yardstick of justice. It becomes clear that, in the ongoing political and legal regulation of AI, a range of standards is at play which are by no means 'irrelevant,' but rather point to fundamental questions regarding our understanding of ourselves and the world, and which likewise require critical scrutiny.

How can the protection of fundamental rights be guaranteed in the face of growing economic pressure, given that competitiveness has hitherto been seen as a particularly important safeguard for democracy, political stability and peace? And even though the rapid development of AI systems does not yet allow us to foresee with any certainty that AI will replace humans as an independent actor rather than merely as an agent, the question nevertheless arises as to whether we even want such a potential development, and to what extent we should and can oppose it. The shared responsibility described here is not fundamentally new; rather, it stems from the way in which people interact with the world and thereby help to shape it. However, the rapid development and now ubiquitous use of AI systems give us reason to endorse the call, first made by Hans Jonas, that we should be guided here not so much by hope as by a heuristic of fear.

Acknowledgments: I would like to thank the members of the DFG-funded Network on AI and Responsibility, especially Jörn Lamla and Fruzsina Molnár-Gábor, for their helpful comments and constructive criticism. I would also like to thank the members of Andreas Niederberger's and Eva Weber-Guskar's colloquia, where I presented and discussed early versions of this text, and finally Manfred Buddeberg

and Lukas Sparenborg and the editors of *Analyse & Kritik* for helpful comments and suggestions.

References

Apel, Karl-Otto. 1988. *Diskurs und Verantwortung*. Frankfurt am Main: Suhrkamp Verlag.

Apel, Karl-Otto. 2000. "First Things First. Der Begriff primordialer Mit-Verantwortung. Zur Begründung einer planetaren Makroethik." In *Angewandte Ethik als Politikum*, edited by M. Kettner, 21—50. Frankfurt am Main: Suhrkamp Verlag.

Apel, Karl-Otto. 2001. "Primordiale Mitverantwortung. Zur transzendentalpragmatischen Begründung der Diskursethik als Verantwortungsethik. Ein Gespräch mit Karl-Otto Apel." In *Prinzip Mitverantwortung*, edited by Karl-Otto Apel, and Holger Burckhart, 97—121. Würzburg: Könighausen & Neumann.

Baum, Kevin, Susanne Mantel, Eva Schmidt, and Timo Speith. 2022. "From Responsibility to Reason-Giving Explainable Artificial Intelligence." *Philosophy & Technology* 35 (12). https://doi.org/10.1007/s13347-022-00510-w.

Buddeberg, Eva. 2011. *Verantwortung im Diskurs*. Berlin: de Gruyter Verlag.

Buddeberg, Eva. 2022. "Wissenschaft als diskursive Mitverantwortung." In *Gefährliche Forschung. Eine Debatte über Gleichheit und Differenz in der Wissenschaft*, edited by Susanne Brandstädter, and Wilfried Hinsch, 113—22. Berlin: de Gruyter Verlag.

Davidson, Donald. 1980. *Essays on Actions and Events*. Oxford: Clarendon Press.

Dennett, Daniel. 1976. "Conditions of Personhood." In *The Identities of Persons*, edited by Amélie O. Rorty, 175—96. Berkeley: University of California Press.

European Commission. 2025. "AI Act." https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (accessed March 18, 2026).

Forst, Rainer. 2014. *The Right to Justification. Elements of a Constructivist Theory of Justice*. New York: Columbia University Press.

Habermas, Jürgen. 2007. "The Language Game of Responsible Agency and the Problem of Free Will. How Can Epistemic Dualism be Reconciled with Ontological Monism?" *Philosophical Explorations* 10 (1): 13—50.

Hinsch, Wilfried. 2026. "Framing Computational Fairness and Non-discrimination." *Analyse & Kritik* 48 (1).

Honneth, Axel. 2025. *Der arbeitende Souverän*. Berlin: Suhrkamp Verlag.

Kohler, Georg. 1988. *Handeln und Rechtfertigen*. Frankfurt am Main: Athenäum Verlag.

Kropp, Cordula, and Tobias Renner. 2026. "Automation, Co-Agency, and Distributed Responsibility. Caring for Hybrid Therapeutic Networks." *Analyse & Kritik* 48 (1).

Larmore, Charles. 2008. "The Autonomy of Morality." In *The Autonomy of Morality*, 87—136. Cambridge: Cambridge University Press.

Misselhorn, Catrin. 2018. *Grundfragen der Maschinenethik*. Ditzingen: Reclam Verlag.

Muldoon, James, Mark Graham, and Callum Cant. 2024. *Feeding the Machine. The Hidden Human Labour Powering AI*. Edinburgh: Canongate Books.

Nida-Rümelin, Julian, and Nathalie Weidenfeld. 2022. *Digital Humanism. For a Humane Transformation of Democracy, Economy and Culture in the Digital Age*. Cham: Springer.

Pereira, Luis M., and Ari Saptawijaya. 2016. *Programming Machine Ethics*. Wiesbaden: Springer Verlag.

Scanlon, Thomas M. 2013. *Being Realistic About Reasons*. Oxford: Oxford University Press.

Stoecker, Ralf, eds. 2002. *Handlungen und Handlungsgründe*. Paderbon: mentis Verlag.
Stoecker, Ralf. 2003. “First Person Authority and Minimal Monism.” In *Monism. Philosophische Analyse*, Vol. 9, edited by Andreas Bächli, and Klaus Petrus, 235—54. Frankfurt am Main, New York: Ontos Verlag.
Vargas, M. R. 2013. *Building Better Beings- A Theory of Moral Responsibility*. Oxford: Oxford University Press.
Wagner, Gerhard. 2024. “„Künstliche Intelligenz — die EU als globaler Regulierer?” *Frankfurter Allgemeine Zeitung*. Last Modified March 3rd 2024.
Weber-Guskar, Eva. 2024. *Gefühle der Zukunft. Wie wir mit emotionaler KI unser Leben verändern*. Berlin: Ullstein Verlag.
Young, Iris Marion. 2006. “Responsibility and Global Justice: A Social Connection Model.” *Social Philosophy and Policy* 23 (1): 102—30.