Philosophy & Technology (2022) 35: 9
https://doi.org/10.1007/s13347-022-00507-5

RESEARCH ARTICLE

Why AI Ethics Is a Critical Theory

Rosalie Waelen[1]

Received: 20 October 2021 / Accepted: 1 February 2022 / Published online: 11 February 2022

Abstract
The ethics of artificial intelligence (AI) is an upcoming field of research that deals with the ethical assessment of emerging AI applications and addresses the new kinds of moral questions that the advent of AI raises. The argument presented in this article is that, even though there exist different approaches and subfields within the ethics of AI, the field resembles a critical theory. Just like a critical theory, the ethics of AI aims to diagnose as well as change society and is fundamentally concerned with human emancipation and empowerment. This is shown through a power analysis that defines the most commonly addressed ethical principles and topics within the field of AI ethics as either to do with relational power or with dispositional power. Moreover, it is concluded that recognizing AI ethics as a critical theory and borrowing insights from the tradition of critical theory can help the field forward.

Keywords Artificial intelligence · AI ethics · Critical theory · Power

1 Introduction

The ethics of artificial intelligence (AI) is an emerging field within applied ethics and the philosophy of technology. It has gained attention and urgency due to the rapid development of AI technology during the past decade. Following Müller (2020), I understand the purpose of the field as twofold: AI ethics deals with the ethical issues that arise from AI systems as objects and with the moral questions raised by AI systems as subjects.[1] Various approaches for the ethical analysis of AI systems as objects have been proposed, but the most dominant one appears to be the *principled approach*. Over the past years, numerous initiatives have developed comparable sets of ethical principles and guidelines to ensure a desirable development and use of AI (Jobin et al., 2019; Ryan & Stahl, 2020). Common principles

[1] For an overview of the field see Dubber et al., 2020; Gordon & Nyholm 2021; or Müller 2020.

✉ Rosalie Waelen
r.a.waelen@utwente.nl

[1] Section of Philosophy, University of Twente, Enschede, Netherlands

are for example transparency, justice and fairness, non-maleficence, responsibility, and privacy (Jobin et al., 2019). These principles function as a kind of soft law, they are not binding. The moral questions that AI systems raise as subjects are, among others, according to what norms AI should be programmed to act (also referred to as "machine ethics"), what the moral status of artificial moral agents would be, and whether AI systems can be held responsible or accountable for their actions and decisions.

The purpose of this article is to argue that AI ethics has the characteristics of a critical theory, by showing that the core concern of AI ethics is protecting and promoting human emancipation and empowerment. Furthermore, I propose that understanding the field as a critical theory can help to overcome some of the shortcomings of the currently popular principled approach to the ethical analysis of AI systems. In other words, I argue that we not only could, but should analyze the ethical implications of AI systems through the lens of critical theory. The focus of the article therefore lies on the ethical analysis of AI systems and the principled approach, but I also discuss what a critical understanding of AI ethics means for the other topics and questions that make up the field.

The structure of the article is as follows. In Sect. 2, I start by briefly defining what it takes to be a critical theory. I conclude that a critical theory studies the power structures and relations in a society, with the goal of protecting and promoting human emancipation, and seeks not only to diagnose, but also to change society. In Sect. 3, I take a more detailed look at the concept of power and explain how a pluralist understanding of the concept allows us to analyze power structures and relations, on the one hand, and (dis)empowerment, on the other hand. In Sect. 4, I argue that the vast majority of the established AI ethics principles and topics in the field are fundamentally aimed at realizing human emancipation and empowerment, by defining these issues in terms of power. Next, in Sect. 5, I propose that AI ethics should be seen as a critical theory, given that the discipline is fundamentally concerned with emancipation and empowerment, and meant not only to analyze the impact of emerging technologies on individuals and society, but also to change it. Moreover, I suggest that recognizing AI ethics as a critical theory can help the field forward—among other reasons because it promotes an interdisciplinary understanding of ethical issues, offers a target for change, helps to identify social and political implications of a technology, and helps to understand the social and political roots of ethical issues. I end the chapter with a brief conclusion in Sect. 6.

2 What Constitutes a Critical Theory?

Marx' famous thesis that philosophers should not only interpret the world but also change it, inspired a group of philosophers now known as the Frankfurt School.[2] The work of the Frankfurt School philosophers was given the name "critical theory."

[2] This is the 11th thesis on Feuerbach which Karl Marx wrote in 1845. It was later published as an appendix to Friedrich Engels' *Ludwig Feuerbach and the End of Classical German Philosophy* in 1886.

Critical theory has two main, characteristic facets. First of all, critical theory has a practical goal; it is meant to diagnose as well as change society. Its unique approach to this is that of immanent transcendence, which implies that critical theorists argue how the world should be (transcendence) based on how it currently is (immanence), rather than always working towards a single predefined image of an ideal state of affairs (Delanty & Harris, 2021; Thompson, 2006). Horkheimer, who founded the Frankfurt School together with Adorno in the early 1930s, suggested that critical theory should be an interdisciplinary endeavor. History, economics, politics, psychology, and other social sciences can help to understand in what ways people's freedom is limited, how the power relations causing this domination came about, and how to counter or resist them. A second important facet is critical theory's emancipatory ambition. Horkheimer said critical theory is "an essential element in the historical effort to create a world which satisfies the needs and power of men" and defined its goal as "man's emancipation from slavery" (Horkheimer, 1972, 246). So critical theorists always seek to identify and overcome forms of domination or restraints that hinder human emancipation or empowerment. Emancipation can be defined as "overcoming social domination" (Forst, 2019, 17) and gives people an equal opportunity for self-development (Allen & Mendieta, 2019). Empowerment implies "[i]ncreasing the scope of agency for individuals and collectives" (Forst, 2019, 21).

There are distinct generations within critical theory. The first generation of critical theorists (among which were Theodor Adorno, Walter Benjamin, Max Horkheimer, and Herbert Marcuse) was preoccupied with criticizing modern capitalism and discussed typically Marxist subjects like alienation, exploitation, and reification. Later on, the focus of their critique became the enlightenment and the loss of individuality due to mass culture (Horkheimer & Adorno, 2002). Jürgen Habermas, a second-generation critical theorist, continued the tradition by studying the state of democracy and discussing power in relation to communication, which lead him to develop his discourse ethics (Habermas, 1984, 1987). Axel Honneth, a student of Habermas, in turn focused his attention on the topic of recognition, which goes back to Hegel (Honneth, 1996). One of the contemporary, fourth-generation members of the school is Rainer Forst, who has continued the tradition by developing a critical theory of justice and redefining the notions of progress and power, among others.

Only the first generation of critical theorists explicitly concerned themselves with technology, mostly focusing on its relation to capitalism (Delanty & Harris, 2021). But, the types of technology that these early Frankfurt School members dealt with were nothing like AI and other digital technologies that exist today. Therefore, it is not easy and perhaps not very valuable either, to try to apply their theories of technology to today's situation. However, that does not have to mean that the tradition of critical theory is not relevant to the philosophy and ethics of technology. Several contemporary thinkers have argued for the relevance of critical theory to understanding the societal role and impact of technology today. Most notably, Feenberg engaged with this tradition to develop his own critical theory of technology (a.o. Feenberg, 1991). Another example is Fuchs, who built on the work of Lukács, Adorno, Marcuse, Honneth, and Habermas to develop a critical theory of communication in the age of the internet (Fuchs, 2016). And in a recent article, Delanty and Harris argue that the general themes that are present in critical theory still offer a

valuable framework for analyzing technology today (Delanty & Harris, 2021). So, the central idea of this paper, namely that the tradition of critical theory can support the analysis of modern technology, is not necessarily new. What is new, as will become clear in what follows, is my proposal to understand the emerging field of AI ethics as a critical theory and to conduct ethical analyses of AI systems through the lens of critical theory.

Finally, the understanding of critical theory as being the work of the Frankfurt School is a narrow understanding of the term. When understood in a broader sense, the term "critical theory" can refer to any theory or diagnosis of power structures and relations that ought to serve emancipatory ends. In this broad sense, the critical theory would also include other schools of thought, such as feminism or postcolonialism (Bohman, 2021). We could then say that specific critical approaches or theories focus on a particular oppressed societal group or a particular way in which people's emancipation is hindered. Different critical theories deal with the struggle of a specific day and age (Bohman, 2021). AI ethics is a critical theory, as I further argue below, which deals with the ways in which AI—a radically new technology—(dis)empowers individuals and facilitates or exacerbates existing power structures in society.

3 Defining the Concept of Power

Given critical theory's focus on emancipation and empowerment, and all the factors that enable or disable this, we can conclude that power is an important, central topic within the field of critical theory. However, power is also a contested concept—according to Steven Lukes it even is essentially contested (Lukes, 1974, 2005). Therefore, it is difficult to understand power's exact role in critical theory. When trying to define "power," scholars disagree over a number of issues. For example, it is disputed whether power can be ascribed to structures or solely to agents; whether an exercise of power is necessarily intentional or whether someone or something can exercise power without intending to do so, and whether the exercise of power has to involve a conflict of interests or not (Brey, 2008). Furthermore, there is a divide between those who conceptualize power in dispositional terms, as "power-to," and those who discuss power in relational terms, as "power-over."

We could say that there are, broadly, four different views of power: the dispositional, episodic, systemic, and constitutive view (Allen, 2016; Haugaard, 2020; Sattarov, 2019). The first resembles "power-to," the latter three are relational views, and therefore, they fall under the category of "power-over." Those who defend the dispositional view of power argue that power is a capacity or ability—namely the capacity to bring about significant outcomes. Acquiring that capacity is also referred to as "empowerment," losing it as "disempowerment." One defender of the dispositional view is Morriss, who argued that two mistakes are commonly made in discussions of power: the vehicle fallacy and the exercise fallacy (Morriss, 2002). The vehicle fallacy is committed when the resources that give rise to power (e.g., AI technologies) are claimed to be power. The exercise fallacy occurs when one equates power with its exercise. Morriss argues that having power entails more than merely

exercising it, it is a disposition that "can remain forever unmanifested" (2002, 17). This is then a direct critique towards those who defend a relational view of power, in particular the episodic view.

The episodic view of power entails that power occurs when one party exercises power over another, for example, by means of force, coercion, manipulation, or through authority. Known for defending, this view of power is Weber, Dahl, and Lukes. Dahl famously formulated the intuitive notion of power as "A having power over B to the extent that A can get B to do something that B would not otherwise do" (1957, 202). Lukes initially followed Dahl, by defining power as "A exercises power over B when A affects B in a manner contrary to B's interests" (1974, 30), but later in his career he accepted Morriss' critique and acknowledged that power is "a capacity not the exercise of that capacity" and that "you can be powerful by satisfying and advancing others' interests" (Lukes, 2005, 12). But even though dispositional power appears to be more fundamental than episodic power, the episodic view of power is relevant because it highlights a specific aspect of power, namely the direct exercise thereof.

Other relational views of power are the systemic and constitutive view. While dispositional and episodic power focus on a single agent and specific instances of power, systemic and constitutive power are more structure-centric (Allen, 2016; Sattarov, 2019). Systemic power, to start with, refers to the ways in which societal institutions, social norms, values, laws, and group identities can have power over individuals. The systemic view complements the episodic and dispositional ones, because it enables us to look at the bigger picture and see what causes some to have dispositional power and exercise it, while others cannot or are constantly subjected to the power of others. Systemic power "highlights the ways in which broad historical, political, economic, cultural, and social forces enable some individuals to exercise power over others, or inculcate certain abilities and dispositions in some actors but not in others." (Allen, 2016).

Constitutive power, finally, refers to the views or discussions of power that focus not on the oppressive character of power, but on the ways in which those subjected to power are also shaped by it. Systems of power not only determine one's sphere of action or possibilities, as the systemic view of power highlights, they also constitute a person's behavior, intentions, beliefs, and more. Foucault is probably most famous for developing this view of power in his work on discipline and biopolitics.

These four views of power are not necessarily incompatible, competing views. According to Mark Haugaard (2010, 2020), Lukes is wrong in saying that power is essentially contested. Haugaard suggests that "power debates will advance more fruitfully if we treat power as a *family resemblance* concept, whereby their meaning varies depending upon *language game*" (Haugaard, 2010, p. 424).[3] A family-wide concept like power is so broad and vague that it explains little in itself; therefore, it is better understood through a cluster of concepts that refer to different aspects of the wider notion. Hence, we should reject the premise that there is a single best definition of power to be found and opt for a pluralist approach to power. The criterion for

[3] Both the idea of family resemblance and language games are borrowed from Wittgenstein.

including a certain view of power in the pluralist approach should be "usefulness," says Haugaard (2010, 427). A definition or notion of power is useful, when it highlights a unique aspect of power.

Critical theorists are concerned with all four elements of power. They first and foremost study relational power, particularly systemic issues in society, in order to emancipate certain societal groups. But, critical theorists are also interested in increasing the scope of human agency, that is, empowering individuals and groups. This latter concern ties in with dispositional and constitutive power. Hence, all four notions of power are valuable in order to understand AI ethics as a critical theory and to conduct ethical analyses of AI systems through the lens of critical theory.

4 Analyzing AI Ethics Principles and Topics in Terms of Power

As explained in the introduction, the field of AI ethics is concerned with identifying and addressing the ethical implications of AI as well as the moral questions raised by this new technology. Müller (2020) discusses privacy, manipulation, opacity, bias, the future of work, and autonomy as main ethical issues that arise from AI systems as objects, and mentions machine ethics, artificial moral agency, and singularity as topics to do with AI systems as subjects. Gordon and Nyholm (2021) offer a similar list. As main debates in the ethics of AI they name machine ethics, autonomous systems, machine bias, opacity, machine consciousness, moral status, and singularity. To some extent, these lists overlap with sets of AI ethics principles or guidelines. It has namely become a trend, among academics as well as businesses and policymakers, to develop sets of ethical principles that should guide the development and use of AI in a desirable direction. This trend represents the aforementioned principled approach to AI ethics.[4] A comparative study of 84 sets of AI ethics principles showed that there is a lot of convergence between the principles that different parties have proposed (Jobin et al., 2019). More precisely, the study identified eleven clusters of values and principles that were brought forward in several documents: transparency (mentioned in 73 of the studied documents), justice and fairness (68), non-maleficence (60), responsibility (60), privacy (47), beneficence (41), freedom and autonomy (34), trust (28), sustainability (14), dignity (13), and solidarity (6). These principles touch upon the debates that are central in AI ethics according to Müller (2020) and Gordon and Nyholm (2021). The issue of opacity, for example, relates to the principle of transparency, and the issue of bias is addressed by the principle of justice.

In what follows I define the most-mentioned AI ethics principles (Sects. 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, and 4.7) in terms of power. By doing so, I show that the fundamental concerns that underly these principles are emancipation, empowerment, or both. Furthermore, I also discuss how other topics in AI ethics, such as machine ethics or

[4] The principled approach to AI ethics resembles the dominant approach in the more established field of bioethics, where four ethical principles are widely recognized as the basis for policymaking and clinical decision-making in the medical field.

singularity, relate to the concept of power (Sect. 4.8). In the last section (Sect. 4.9), I briefly look at non-Western approaches to AI ethics and argue that the concern for emancipation and empowerment is present (although perhaps less dominant) in these approaches as well.

Finally, when I talk about AI as having or exercising power over humans, I do not mean to ascribe the technology as the kind of agency or intentionally required to purposefully seek power. Rather, I merely mean that technology can be used as an instrument or delegate by human beings to exercise power (Brey, 2008) or that AI can unintentionally create new power relations, exacerbate existing ones, or affect an individual's autonomy by empowering or disempowering them.[5]

4.1 Transparency

The most proposed principle for ethical AI is the principle of transparency. Transparency would be required, for example, with regard to data collection and processing (what is done with my personal information?), automated decision-making (how are the decisions that affect me made?), or personalized recommender systems (on the basis of what data am I shown a certain product or news article?). Transparency implies that the answers to these questions are both accessible and comprehensible. The idea behind this principle is that transparent, explainable, or interpretable AI would minimize harm by AI systems, improve human-AI interaction, advance trust in the technology, and potentially support democratic values (Jobin et al., 2019). A different way of explaining what makes transparency in AI so valuable is to say that it grants an individual epistemic agency, i.e., the ability to know what happens to their data and how AI affects them, as well as the ability to better control one's own information and the ways in which they are affected by AI. This knowledge and control are dispositional powers—transparency implies having the power to know or understand what happens to one's data and on what bases decisions are made. Therefore, we can say that what is really at stake when transparency is called for, is individual empowerment.

[5] It should also be noted that there exist a number of examples of discussions of power within AI ethics or related fields. For instance, Cobbe (2020) argues that algorithmic censorship augments the societal power of social platforms; Danaher (2020) discusses the effects of "algocracy" on freedom; de Laat (2019) defends that predictive algorithms have disciplinary power; Mohamed et al. (2020) suggest the use of decolonial theory to understand how the values embedded in AI are shaped by existing power relations; Noble (2018) discusses the "oppressive algorithms" of search engines; and Zuboff (2019) critiques the dominance of "surveillance capitalism." Also, Sattarov (2019) explores the relation between technology (e.g., algorithms), power and ethics. So there has certainly been recognition for the relevance of discussing power in relation to ethical issues in AI, but these issues are not covered or addressed by the dominant AI ethics guidelines. Moreover, each of the mentioned discussions of power focuses on a single ethical issue, a single technology, or a specific conception of power. The critical approach to AI ethics that will be put forward in this paper, which is based on a pluralist understanding of power, brings these different issues together.

4.2 Justice, Fairness, and Solidarity

Justice is mentioned in AI ethics guidelines in relation to fairness, on the one hand, and bias and discrimination, on the other hand (Jobin et al., 2019). Fairness concerns have to do with equal access to AI and the equal share of the burdens and benefits of the technology. There is, for example, the concern about a digital divide between the countries that can afford to develop and use AI and those parts of the world that do not have access to the latest technology. The principle of non-discrimination has become pressing as many emerging technologies have been found to contain biases. Particularly algorithmic bias has received much attention in the field of AI ethics. Algorithms can contain biases, among other reasons, when they are built on non-inclusive training data. Biased algorithms or in other ways biased AI systems can lead to discriminatory outcomes (e.g., continuously misidentify certain demographics as a threat or potential criminal) and therefore violate the principle of just and fair AI. Mentioned less often in existing AI guidelines is the principle of solidarity. Solidarity has to do with the (fair) distribution of AI's benefits and harms. Solidarity in AI would imply that the benefits of AI should be redistributed from those who are disproportionally benefitted by this new technology to those who turn out to be most vulnerable to it (e.g., those who are unemployed due to automation).

These justice-related concerns can be described in terms of relational power. Forst writes that "the question of power, qua social and political power that shapes collective processes, is central to justice" (Forst, 2015, 8). When we speak of justice, we are referring to what we consider to be acceptable power relations or systems in society. The concept of justice can refer to the direct exercise of power by one actor over another (i.e., episodic power), but most often justice has to do with the systems of power that shape a society and, hence, determine the possibilities of certain individuals or groups in that society (systemic power). In other words, the principle of justice protects those subjected or vulnerable to systems of power or the exercise of power by others. In the context of AI, the principle of justice ought to see to it that the power relations that are created or reinforced by AI systems give people the space to develop themselves, and do so in an equal manner. By doing so, the principle of justice serves the goal of emancipation.

4.3 Non-maleficence and Beneficence

Although they are perhaps less strongly related to the concept of power than other principles, non-maleficence and beneficence too can be understood as principles meant to protect those who are vulnerable to the power of AI. As AI develops to be omnipresent in society, it becomes inescapable for people to use the technology, to be subject of data analysis, and to be affected by automated decision-making systems. Their lives are therefore increasingly controlled by this technology. Moreover, the fact that AI is inescapable to the modern citizen is in itself an

instance of systemic power. Protecting that AI does not harm and potentially even benefits its users and subjects is a way of ensuring a desirable power relation and, hence, that AI does not stand in the way of people's emancipation.

4.4 Responsibility and Accountability

Responsibility is frequently mentioned as a guiding principle for AI, because of the concern that automated decision-making will create responsibility gaps (Gordon & Nyholm, 2021). AI systems can take decisions that directly impact human beings, but cannot be held responsible or accountable for the consequences of their actions in the same ways humans can. This raises the question: who should (and can) be held responsible and accountable for the harm caused by AI systems? This question needs to be answered in order to assure that the power relations that AI creates, exacerbates or facilitates are not abused (Sattarov, 2019). Like justice, the principles of responsibility and accountability in AI ethics guidelines can be understood as a way of protecting those subjected to AI's power. Therefore, responsibility too supports emancipation—it helps to overcome unjustifiable social domination.

4.5 Privacy

Privacy can be discussed as a value, moral right or legal right. Jobin et al. (2019) point out that AI ethics guidelines discuss privacy both as a value to uphold and a right that should be protected. Moreover, privacy is often discussed in relation to data protection, which is in line with the common definitions of privacy as "informational control" or "restricted access" (DeCew, 2018). Under both definitions, privacy is understood as a dispositional power, more precisely, as the capacity to control what happens to one's information and to determine who has access to one's information or other aspects of the self. AI, then, is perceived as a potential threat to this capacity because it entails the collection and analysis of large quantities and new types of personal data. Hence, AI could disempower individuals with respect to their privacy. Or put differently, privacy should be promoted because it empowers data subjects.

4.6 Freedom and Autonomy

Freedom and autonomy are related concepts and often mentioned together in AI ethics guidelines. Freedom can be defined in positive and negative terms; it can be understood as the lack of outside interference in one's actions, or the possibility thereof, but it is also discussed as being free to act. The concept of autonomy relates to the positive definition of freedom, it means "self-rule" or "self-determination." If empowerment entails increasing the scope of individual or collective agency, then autonomy and positive liberty clearly serve the goal of empowerment. Having the ability or power to act freely and rule oneself support one's agency. This ability can be promoted, in the context of AI or other types of technology, by for example transparency or informed consent. Negative freedom can be defined in terms of episodic

power and systemic power—it is the absence of power exercises by others or of systemic power relations. One is free, in this sense, when he or she is not subjected to the power of others or a system of power. Such freedom implies, for example, not being subject to technological experiments, manipulation, or surveillance (Jobin et al., 2019). In this sense, the principle of freedom serves the goal of emancipation. We want the values of freedom and autonomy to guide the development and use of AI, because we want to ensure that this new technology is emancipatory and empowering.

4.7 Trust

The call for trust or trustworthy AI can refer to the ways in which AI research and technology is done, to the organizations and persons that develop AI, to the underlying design principles, or to users' relation to a technology (Jobin et al., 2019). Such trust can be fostered by transparency or by ensuring that AI meets the expectations of the public. People's need to trust how AI is developed and functions, can be explained not in terms of dispositional power or empowerment, as is the case for transparency, but as a protection against the exercise of power by others. Trust is a desirable feature of the relation between technology and those using it or subjected to it. When trusting an AI system, one expects that the power that technology can exercise over the individual will not be misused. A trustworthy power relation is one where A holds power over B, without B needing to worry that A will take advantage of this situation.[6] Trust can therefore be understood as serving emancipation—by calling for trustworthy AI, we want to guarantee that AI cannot exercise power over us in arbitrary, harmful or excessive ways.

4.8 Other Topics in AI Ethics

Although there exists a significant amount of overlap between the main debates in the field of AI ethics and the principles that are most commonly mentioned in AI ethics guidelines, these guidelines do not cover topics like machine ethics, the moral status of AI systems, or technological singularity. Moreover, while there is wide agreement on what ethical principles should be reflected in the development and use of AI, the views on these other moral questions in AI are much more varied. But, there is nevertheless a shared concern present within all of these debates. All these topics address, each in their own way, the question of how we should relate to AI and exercise control over it. AI has the potential to become an unprecedentedly powerful technology, due to its intelligence, ability to function autonomously and also widespread reliance on technology. So debates regarding the norms according to which we want AI to act, whether we should grant AI rights, and whether the technology poses an existential risk or not, all express a concern for the human's position in relation to the (potential) power of AI.

[6] For a more elaborate discussion of the relation between power and the moral concepts of trust, vulnerability, authenticity, and responsibility, see Sattarov 2019.

4.9 Non-Western Perspectives on AI Ethics

One of the criticisms raised against the principled approach to AI ethics is that the guidelines that have been established (including the ones discussed by Jobin et al., 2019) only represent Western views and values (Gordon & Nyholm, 2021). An AI ethics based on non-Western philosophy (e.g., Daoism, Confucianism, or Ubuntu) might not focus as much, or perhaps not at all, on the individual and their emancipation and empowerment. However, Gal (2020) shows that AI ethics guidelines developed in South Korea, China, and Japan show a lot of similarities with the guidelines studied by Jobin et al. (2019). The South Korean government presented an ethical framework in 2018, aimed at achieving a human-oriented intelligent information society. The framework consisted of four main principles: publicness, accountability, controllability, and transparency. Central to the Korean approach, Gal notes, is "a clear human-over-machine hierarchy" (Gal, 2020, 609). Chinese approaches show the same emphasis on the idea that AI should first and foremost be a tool to benefit humans. Furthermore, China's own big tech companies Baidu and Tencent developed AI ethics principles that show a lot of convergence with the principles mentioned before. Baidu lists safety and controllability, equal access, human development, and freedom. Tencent says AI should be available, reliable, comprehensive, and controllable. Japan deviates most from the trend, by envisioning a coexistence and coevolution between humans, on the one side, and AI and robots on the other side.

So although it is not unimaginable to have AI ethics approaches with an entirely different focus, the current state of AI ethics is that the field is predominantly concerned with human emancipation and empowerment. The goal of AI ethics is to ensure that the emerging technologies that promise to radically change life as we know it, do so for the better.

5 AI ethics as a Critical Theory

By defining the ethical principles and moral questions that are central in AI ethics in terms of power (under a pluralist understanding of power, that is), I have shown that the field is driven by a fundamental concern for human emancipation and empowerment. Transparency, privacy, freedom, and autonomy are valued because they are empowering—they grant individuals the ability to rule their own lives. Principles like trust, justice, responsibility, and non-maleficence are important because they protect individuals against the power that could be exercised by means of AI, or possibly even by AI itself. These central concerns are usually not made explicit. However, doing so helps us to see that AI ethics resembles a critical theory. Like any critical theory, the purpose of AI ethics is not merely to analyze or diagnose society, but also to change it. Both critical theory and AI ethics have a practical goal, namely that of empowering individuals and protecting them against systems of power. But while critical theory is concerned with society at large, AI ethics focuses on the part that a particular type of technology plays in society. Hence, we could say that AI

ethics is a critical theory, which focuses on the ways in which human emancipation and empowerment are or could be hindered by AI technology.

Understanding AI ethics as a critical theory can help the field forward in a number of ways. First of all, defining ethical principles in terms of their relation to emancipation or empowerment (respectively, relational, or dispositional power) creates a common language to compare ethical issues and to discuss them in interdisciplinary contexts. Such a common language is welcome for two reasons: because the principles have been accused of being too abstract (Mittelstadt, 2019; Resseguier, 2021; Ryan and Stahl, 2020) and because improving AI is a necessarily interdisciplinary endeavor. Not just ethicists, but also tech developers, data scientists, policymakers, and legal experts need to be involved to realize the goal of ethical AI. The language of power could be more appropriate in interdisciplinary contexts than discussions about (sometimes highly contested) moral values and principles are. Furthermore, by defining how ethical issues relate to dispositional and relational power, we immediately have a clear target for change. While the principles of transparency, privacy, or justice might not be action guiding as such, empowering individuals with respect to their knowledge or informational control, or reducing the say AI systems have over their lives, are much more tangible goals.

A second benefit of the insight that AI ethics is a critical theory is that it offers us a new method for identifying and analyzing ethical issues: a power analysis. The principled approach to AI ethics functions as a kind of ethical checklist. However, the potential negative implications of a technology (and to some extent also the positive ones) could also be identified by analyzing in what ways the technology limits a person's agency or freedom. The different aspects of power that I described in Sect. 3 could inform and guide such a power analysis. In addition to being more appropriate for interdisciplinary work, a power analysis also has the advantage that it could cover ethical issues that are left unaddressed by current AI principles. The principled approach has been accused of giving too little attention to the social and political context in which AI applications are developed and used, and because of which ethical issues arise. Using a power analysis to ethically assess emerging AI systems will improve our understanding of the ways in which ethical issues in AI tie into broader social, political, economic and historical matters, and understanding the broader context of an ethical issue will in turn make it easier to address the issue, not just on a technical level, but on a societal and political level.

Furthermore, a power analysis could be complemented by insights from the tradition of critical theory. Critical theory is not a full-fledged normative theory that explains what is right and what is wrong in the way that classic theories like consequentialism, deontology, or virtue ethics do, but it does take in a normative stance. Just like critical theory, AI ethics is not meant to be an ethical theory in the classic sense, but it should diagnose technological advancements in society and change them for the better. As others have argued before (Delanty & Harris, 2021; Feenberg, 1991; Fuchs, 2016), critical theory offers a valuable toolbox for analyzing the societal implications of modern technologies. I add to these arguments that many of these societal implications tie into ethical issues. Or in other words, critical theory can help to pinpoint ethically relevant issues that are not typically

addressed by ethical principles or classic ethical theories. Critical theory could, for example, help to understand ethical issues that arise from AI's relation to present-day capitalism (following first-generation critical theorists) or the potential ethical implications of misrecognition that is mediated by AI (following Honneth, 1996).

A final benefit of unmasking AI ethics as a critical theory is that we can now understand AI ethics principles as having a common aim. Mittelstadt (2019) criticizes the popular AI ethics principles for lacking a common aim. Without such a common aim, Mittelstadt argues, the principled approach to AI ethics could not have the same success as the principled approach in bioethics has. But by defining the AI ethics principles in terms of power, I showed that they do share a common aim: to protect human emancipation and empowerment in the face of this new, powerful technology. AI might imply technological progress, but that does not guarantee social progress. AI that decrease our freedom, our agency, and our ability to develop ourselves, would be a step back for the emancipation of individuals and societal groups. As Forst writes, "every progressive process must be constantly questioned as to whether it is in the social interest of those who are part of this process" (Forst, 2019, 21).

6 Conclusion

The emerging field of AI ethics is unlike other fields in applied ethics. At the center of its attention is not human conduct, but the ways in which humans are affected by AI technology. It differs from the general ethics of technology too, in the sense that AI comes with radically new possibilities for action. This not only raises new moral questions, but also requires new approaches to conduct ethical analyses. But the most popular approach thus far—that is, the principled approach—has been met with criticism. Although there is a lot of convergence when it comes to determining which ethical issues or principles should shape the development, policy, and use of AI, AI ethics principles have been accused of being too abstract, little action guiding, and insufficiently attuned to the social and political context of ethical issues.

The aim of this paper was to show that AI ethics is just like a critical theory. I have explained that a critical theory is aimed at diagnosing and changing society for emancipatory purposes. I then showed that both the big debates in AI ethics and the most common AI ethics principles are fundamentally concerned with either individual empowerment (dispositional power) or the protection of those subjected to power relations (relational power). Approaching AI ethics as a critical theory, by diagnosing AI's impact by means of a power analysis and the insights of critical theory, can help to overcome the shortcomings of the currently dominant principled approach to AI ethics. Further research could test the power analysis out in a concrete case study, further assess the extent to which the understanding of AI ethics as a critical theory resonates with non-western approaches to AI ethics, or investigate how related field (such as machine ethics) could benefit from the critical theory perspective.

Funding The author is an early-stage researcher funded by MSCA ITN PROTECT project. This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 813497.

Data availability Not applicable.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing Interests The author declares no competing interests.

References

Allen, A. (2016). Feminist perspectives on power. *The Stanford Encyclopedia of Philosophy, edited by E.N. Zalta.* Accessed October 15, 2021. https://plato.stanford.edu/archives/fall2016/entries/feminist-power/

Allen, A. & Mendieta, E. (2019). Introduction. In *Justification and emancipation. The critical theory of Rainer Forst*. Edited by Amy Allen and Eduardo Mendieta. The Pennsylvania State University Press.

Bohman, J. (2021). Critical Theory. *The Stanford Encyclopedia of Philosophy, edited by E.N. Zalta.* Accessed October 15, 2021. https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=critical-theory

Brey, P. (2008). The technological construction of social power. *Social Epistemology, 22*(1), 71–95. https://doi.org/10.1080/02691720701773551

Cave, S., & Dihal, K. (2020). The Whiteness of AI. *Philosophy & Technology, 33*(4), 685–703. https://doi.org/10.1007/s13347-020-00415-6

Cobbe, J. (2020). Algorithmic censorship by social platforms: Power and resistance. *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00429-0

Dahl, R. A. (1957). The concept of power. *Behavioral Science*, 201–215. https://doi.org/10.7312/popi17594-004

Danaher, J. (2020). Freedom in an age of algocracy. *Oxford Handbook on the Philosophy of Technology*, 1–32.

de Laat, P. B. (2019). The disciplinary power of predictive algorithms: A Foucauldian perspective. *Ethics and Information Technology, 21*(4), 319–329. https://doi.org/10.1007/s10676-019-09509-y

DeCew, J. (2018). Privacy. Zalta, E.N. (Ed.). *The Stanford Encyclopedia of Philosophy*. First edition 2018. Accessed October 15, 2021. https://plato.stanford.edu/archives/spr2018/entries/privacy/

Delanty, G., & Harris, N. (2021). Critical theory and the question of technology: The Frankfurt School revisited. *Thesis Eleven, 166*(1), 88–108.

Dubber, M.D., Pasquale, F. and Das, s. (eds) (2020). *The Oxford handbook of ethics of AI*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.001.0001

Engels, F. (1886). *Ludwig Feuerbach und der Ausgang der klssischen deutschen Philosophie*. Stuttgart: Neue Zeit.
Feenberg, A. (1991). *Critical theory of technology*. Oxford University Press.
Forst, R. (2015). Noumenal Power. *Journal of Political Philosophy, 23*(2), 111–127. https://doi.org/10.1111/jopp.12046
Forst, R. (2019). The justification of progress and the progress of justification. In *Justification and emancipation. The critical theory of Rainer Forst*. Edited by Amy Allen and Eduardo Mendieta. The Pennsylvania State University Press.
Fuchs, C. (2016). *Critical theory of communication: New readings of Lukács, Adorno, Marcuse, Honneth and Habermas in the age of the Internet.* University of Westminster Press.
Gal, D. (2020). Perspectives and approaches in AI ethics. East Asia. In *The Oxford Handbook of Ethics of AI*. Edited by M. D. Dubber, F. Pasquale, and S. Das. Oxford University Press. DOI https://doi.org/10.1093/oxfordhb/9780190067397.001.0001
Gebru, T. (2020). Race and Gender. In *The Oxford handbook of ethics of AI*. Edited by M. D. Dubber, F. Pasquale, and S. Das. Oxford University Press. DOI https://doi.org/10.1093/oxfordhb/9780190067397.001.0001
Gordon, J. S. & Nyholm, S. (2021). Ethics of artificial intelligence. *Internet Encyclopedia of Philosophy.* Accessed January 21, 2022. https://iep.utm.edu/ethic-ai/
Habermas, J. (1984). The Theory of Communicative Action. Vol. I: Reason and the Rationalization of Society. Translated by T. McCarthy. Boston: Beacon Press. [Published in German in 1981]
Habermas, J. (1987). The theory of communicative action. Vol. II: Lifeworld and System. Translated by T. McCarthy. Boston: Beacon Press. [Published in German in 1981]
Haugaard, M. (2010). Power: A "family resemblance concept." *European Journal of Cultural Studies, 13*(4), 419–438.
Haugaard, M. (2020). *The four dimensions of power: Understanding domination, empowerment and democracy*. Manchester University Press.
Honneth, A. (1996). *The struggle for recognition: The moral grammar of social conflicts*. MIT Press.
Horkheimer, M. & Adorno, T.W. (2002). *Dialectic of Enlightenment.* Translated by Edmund Jephcott, edited by Gunzelin Schimdt Noeri. Stanford University Press.
Horkheimer, M. (1972). *Critical theory selected essays*. Translated by Matthew J. O'Connell. New York: The Continuum Publishing Company.
Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*. 10.1038/ s42256-019-0088-2
Lukes, S. (1974). *Power: radical view*. London: Macmillan.
Lukes, S. (2005). *Power: A radical view. Second edition*. London: Red Globe Press.
Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-019-0114-4
Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology, 33*, 659–684.
Morriss, P. (2002). *Power: A philosophical analysis*. Manchester, New York: Manchester University Press.
Müller, V. (2020). Ethics of artificial intelligence and robotics. Zalta, E.N. (Ed.). *The Stanford Encyclopedia of Philosophy*. First edition 2020. Accessed January 21, 2022. https://plato.stanford.edu/entries/ethics-ai/
Noble, S.U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York: New York University Press.
Resseguier (2021). Ethics as attention to context: Recommendations for AI ethics. In *SIENNA D5.4: Multi-stakeholder strategy and practical tools for ethical AI and robotics*. https://www.sienna-project.eu/publications/deliverable-reports/
Ryan, M., & Stahl, B. C. (2020). Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society, 19*(1), 61–86. https://doi.org/10.1108/JICES-12-2019-0138
Sattarov, F. (2019). *Power and technology: A philosophical and ethical analysis*. London, New York: Rowman and Littlefield International.
Stahl, B. C., Doherty, N. F., Shaw, M., & Janicke, H. (2014). Critical theory as an approach to the ethics of information security. *Science and Engineering Ethics, 20*, 675–699. https://doi.org/10.1007/s11948-013-9496-6
Thompson, S. (2006). *The political theory of recognition. A critical introduction*. Polity Press.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power.* London: Profile Books.