AI and Ethics
https://doi.org/10.1007/s43681-023-00377-3

ORIGINAL RESEARCH

Navigating in the moral landscape: analysing bias and discrimination in AI through philosophical inquiry

Serap Keles[1]

Received: 28 July 2023 / Accepted: 26 October 2023

Abstract
This article embarks on a philosophical inquiry into the ethical virtues, particularly, kindness, empathy and compassion within the realm of artificial intelligence (AI), seeking to explicate its essence and explore its philosophical foundations. By delving into different philosophical theories of virtues, we can discover how these theories can be applied to the complex terrain of AI. Central challenges are addressed, including issues of bias, discrimination, fairness, transparency and accountability in the pursuit of promoting ethical principles in AI. Moreover, this exploration encompasses a critical examination of universal ethical principles such as beneficence, non-maleficence, and respect for human dignity, specifically in the context of AI. This scrutiny underscores the pressing need for interdisciplinary collaboration between ethicists, technologists, and policymakers to forge robust frameworks that effectively promote values in AI. In pursuit of a comprehensive understanding, it is essential to subject various arguments and perspectives to evaluation. This entails engaging with philosophical theories such as utilitarianism, deontology and virtue ethics. Throughout the article, an extensive array of supporting evidence is employed to bolster the arguments presented by virtue ethics, such as the integration of compelling case studies, empirical research findings, and lived experiences that serve to illustrate and illuminate the practical implications of the discourse. By thoroughly exploring these multifaceted dimensions, this article offers nuanced philosophical insights. Its interdisciplinary approach and rigorous analysis aim to engender a comprehensive understanding of this complex issue, illuminating potential avenues for ethical progress within the realm of AI.

Keywords Virtue ethics · Ethics of AI · Bias and discrimination in AI

1 Defining ethical virtues in the context of AI

> "Technology is capable of doing great things, but it doesn't want to do great things. It doesn't want anything. That part takes all of us. It takes our values, and our commitment to our families, and our neighbours, and our communities."[1]
> Tim Cook
> Chief Executive Officer of Apple

Tim Cook made this statement in a broader context of technology, yet it is relevant to AI, too. It prompts us to consider not only the capabilities and potential benefits of AI but also the ethical considerations and societal implications that arise from its use. In a way, it emphasises the importance of responsible AI development and decision-making, actively considering the potential consequences and striving for ethical outcomes. The interplay between social and ethical values, justice, fairness, and equity is not an abstraction but a living testament to the commitment to creating a society where the benefits of technological progress are shared equitably. This reasoning encourages us to engage in a debate surrounding the core ethical virtues—particularly kindness, empathy and compassion—and the reasoning behind the argument that these social values should also be deeply rooted in the ethical dimensions of AI, and finally, how to foster ethical values that practically guide our actions and choices in this rapidly evolving technological landscape.

The proposition advocating the grounding of non-discriminatory and unbiased principles in ethical precepts finds

✉ Serap Keles
serap@diversitytrust.org.uk

[1] Diversity Trust, London, UK

[1] https://www.vox.com/2017/6/9/15768760/apple-tim-cook-mit-commencement-speech-tech.

Published online: 22 November 2023

its substantiation in a nuanced comprehension of foundational values that are crucial for the cultivation of an ethical framework. This philosophical discourse accentuates several key rationales for embedding equity and inclusion principles in ethical considerations. It contends that these principles should transcend mere pragmatic directives, assuming the stature of inherent moral imperatives. Ethical principles, as exemplified by virtues such as kindness, empathy, and compassion, serve as the bedrock for shaping a social ethos that prioritises fairness, justice, and the intrinsic dignity of every individual. A philosophical inquiry elevates these virtues beyond the realm of benevolent inclinations, acknowledging them as moral imperatives integral to the construction of a just and equitable society. These virtues emerge not as ancillary considerations in ethical deliberation but as foundational principles resonating across diverse cultural, religious, and philosophical traditions.

Deeply embedded in human society, these virtues carry immense philosophical weight and encapsulate a sincere care for the welfare of others and a readiness to engage in actions that foster their well-being. The philosophical investigation explores ethical dimensions, moral ramifications, and practical manifestations in various settings. Throughout the annals of history, the contemplation of these virtues and their moral significance has occupied the minds of philosophers. Across diverse philosophical traditions, from the wisdom of ancient Greek philosophy to the profound teachings of Eastern philosophies like Buddhism and Confucianism, these virtues are indispensable for fostering harmonious social relations and facilitating individual flourishing. Various expressions emerge within these philosophical systems, such as *agape*, representing unconditional love in Christian philosophy; *karuna*, denoting compassion in Buddhism; and *philia*, embodying affectionate regard in Greek ethics. While some philosophical perspectives posit virtues as inherent qualities which are intrinsic to human beings, others highlight their cultivation through moral education and dedicated practice.

Philosophical theories examine the motivations underlying acts of benevolence and the ethical dimensions they encompass. Within the utilitarian framework, ethical virtues serve as a means to optimise overall happiness and well-being, prioritising the promotion of the greatest good for the largest number of individuals. Contrarily, deontological perspectives stress the moral obligation to exhibit ethical virtues, irrespective of outcomes or consequences. Virtue ethics, on the other hand, situates them within a broader framework of virtues and character development, highlighting the cultivation of virtuous inclinations that enable compassionate conduct.

Virtue ethics emerges as a credible framework among other traditions for understanding the ethical significance of kindness, empathy, and compassion, offering a holistic approach that prioritises the cultivation of virtuous character. Unlike theories that are centred on actions or consequences, the discipline places character traits at the forefront of ethical considerations. Furthermore, its enduring relevance across diverse cultural traditions highlights virtues' universal significance, offering a robust foundation for ethical living that transcends cultural relativism. Within a broader context of virtues, it emphasises the significance of compassionate actions in leading a morally principled existence. Within the realm of virtue ethics, these virtues define a foundational role, fostering harmonious interactions and contributing to the well-being of individuals and communities. It manifests as a disposition or habitual demeanour that empowers individuals to genuinely concern themselves with the welfare of others. However, in this context, they encompass more than sporadic displays of benevolence and superficial acts of compassion. It transcends the boundaries of momentary gestures, covering a deeper commitment to empathy, understanding, and active goodwill towards others. Recognising its vital role in moral flourishing and the cultivation of virtuous character, individuals who nurture these virtues become attuned to the needs and suffering of others, engendering a heightened sense of social responsibility and an authentic desire to alleviate the hardships endured by other individuals. Furthermore, as a virtue of profound significance, they transcend the confines of particular contexts and relationships by surpassing personal affiliations, extending their benevolent influence towards strangers and even adversaries, thereby exemplifying an inclusive and compassionate orientation towards all manifestations of life. By embracing such a perspective, individuals can stimulate a sense of interconnectedness and empathy that surpasses individual boundaries, weaving compassion and understanding into the very fabric of existence.

When we go deep into the realm of virtue ethics and explore the ethical significance of kindness, its role in fostering human flourishing, and its potential impact on moral character and social interactions, it is important to mention two remarkable contemporary philosophers who share the same vision. The first among this illustrious pair is Martha Nussbaum, who offers an all-encompassing elucidation of the philosophy of kindness in her insightful essay on compassion.[2]. Nussbaum discusses the essential role of compassion as a virtue, emphasising its profound importance in promoting human flourishing and ethical living. According to Nussbaum, compassion represents an essential virtue that contributes to the overall well-being of individuals and communities. It transcends mere acts of benevolence, encompassing a genuine concern for the welfare and dignity of others. In this regard, she considers compassion as the first

[2] Nussbaum, M. [13].

and foundational stage towards kindness. For Nussbaum, compassion entails active engagement with the needs and suffering of others, playing a pivotal role in fostering social cohesion and cultivating robust interpersonal relationships. Nussbaum's exploration of kindness underlines its inherent value in ethical decision-making and the cultivation of moral character. She argues that nurturing kindness is pivotal for moral flourishing, as it contributes to the development of a virtuous and compassionate self.[3]

In the enlightening work After Virtue,[4] contemporary philosopher Alasdair McIntyre provides profound insights that align with Nussbaum's perspective on the philosophy of kindness. According to McIntyre, kindness transcends isolated actions and benevolent gestures; it embodies a disposition of character that moulds our ethical outlook and moral engagements. It emanates from a genuine concern for the well-being of others, recognising their inherent dignity and worth. In the realm of philosophical discourse, acts of kindness surpass mere utility and self-interest. This value represents a recognition of our shared humanity and the vital importance of nurturing harmonious relationships within society. However, McIntyre emphasises that the philosophy of kindness should not be divorced from moral discernment and reason. It necessitates thoughtful reflection and deliberate action as we strive to engage in behaviours that authentically promote the well-being and flourishing of others. It encompasses empathy, compassion, and a willingness to extend ourselves for the sake of others, even when faced with inconvenience or challenges. Moreover, it impels us to confront systemic injustices and advocate for a world that upholds the dignity and well-being of all.

When we take these arguments' sentiments on ethical virtues and their intricate connection to moral discernment and reason, and as we imbue machines with intelligence and decision-making capabilities, the virtues we instil become the cornerstone of ethical AI development. Now, the willingness to extend ourselves for the sake of others becomes an intriguing challenge. In AI ethics, it could mean addressing biases in algorithms, ensuring inclusivity, and designing systems that benefit all, not just a privileged few. Confronting systemic injustices in the realm of AI requires a vigilant eye. Bias in data and discriminatory algorithms are the systemic injustices that demand our attention. An ethical framework for AI involves advocacy for fairness and transparency, challenging the status quo when it perpetuates inequities. As we embark on this philosophical journey, we see that it is not merely about creating intelligent machines but fostering virtuous AI, ensuring that our technological creations reflect the best of our moral aspirations.

Consider transparency not as a procedural obligation but as a manifestation of the metaphysical transparency that philosophy strives for—a clarity that reveals the essence of our technological progress. It is a pursuit not just of fairness but of existential balance, where algorithmic calculation reflects the universal value we seek in the most Platonic sense. The idea here is to elevate the concept of fairness from a surface-level ethical guideline through transparency to something that aligns with fundamental principles and ideals, almost akin to seeking a balance that resonates with broader values. Accountability, in this philosophical context, is not a mere response to error but a profound acknowledgement of the moral responsibility we bear as architects of AI. It is the recognition that, in shaping these digital entities, we are not just crafting tools but sculpting entities that, in their essence, mirror our ethical consciousness and the essence of human flourishing.

In an envisaged future, technology is not a detached instrument but a manifestation of philosophical ideals—a companion in our philosophical journey toward empathy. It is not merely about mitigating biases but a Socratic dialogue with our algorithms, a relentless questioning that seeks the deeper truths within the lines of code. Our engagement with AI cannot be a passive drift but a conscious navigation through the uncharted paths of ethical innovation. It implores us to plumb the philosophical depths, seeking not just the surface-level ethical considerations but the profound undercurrents that shape our digital reality. As AI systems grow in sophistication and pervasiveness throughout society, ethical considerations surrounding their development, deployment, and impact on human well-being take centre stage. Advocates of virtue ethics argue that integrating kindness into the design and utilisation of AI systems can help shape a more ethical and compassionate AI landscape. Infusing AI systems with kindness, empathy and compassion entails prioritising the well-being of users, promoting fairness, transparency, and accountability, and ensuring that AI technologies serve the broader interests of society. This involves a deliberate design of AI algorithms and decision-making processes that prioritise kindness, empathy, respect, and the preservation of human dignity. Moreover, philosophical and ethical virtues can guide the ethical implications of AI in relation to issues of equality, equity, diversity and inclusion. However, defining a universal notion of ethical values that is programmed into AI systems presents challenges, as different cultures and individuals may hold diverse understandings of what constitutes morally acceptable behaviour.

As we contemplate these challenges, we encounter complexities inherent to the nature of AI and its interactions with humans. The subjectivity of ethical values, the capability of interpreting and responding to nuanced cues and social dynamics, and the need for algorithms that understand

[3] Nussbaum, M. [13].

[4] MacIntyre, A. C. [10].

and respond appropriately to diverse human experiences all demand ongoing reflection, research, and iterative improvements. While the idea of a shared ethical framework may appear idealistic, its justification lies in the acknowledgement of certain enduring principles that transcend cultural, societal, and temporal boundaries. At the core of this argument is the recognition that beneath the variegated tapestry of human cultures and traditions, there exists a common thread of fundamental values that resonate across epochs and civilisations. These values, ranging from concepts of justice and fairness to kindness, empathy and compassion, have persisted as guiding beacons in human societies. The universality of these values is not a mere assumption but is deeply rooted in the shared human experience.

As pointed out above, diverse cultures and philosophical traditions have independently converged upon similar ethical principles, reflecting an innate understanding of what it means to lead a virtuous life. If we look at one of the rules of shared human experiences, such as "Treat others as you would like to be treated", which reflects a shared understanding of empathy and reciprocity, its prevalence suggests that the idea of treating others with kindness and fairness is deeply embedded in the human experience, transcending cultural and temporal boundaries. Utilitarianism, in the same vein, underscores the enduring principle of maximising overall well-being. Although the pursuit of the greatest good for the greatest number may have some practical pitfalls, it still reflects a universal aspiration for positive outcomes that can inform ethical considerations. Such doctrines influenced the Universal Declaration of Human Rights (UDHR), which was adopted by the United Nations in 1948[5] and was developed in the aftermath of World War II, representing a global consensus on fundamental human rights. Its acceptance by diverse nations suggests a shared commitment to certain ethical principles, forming a basis for the assumption of a universal moral paradigm that extends to contemporary issues.

In the realm of AI development, embracing a universal moral paradigm becomes domineering for several reasons. First, it serves as a safeguard against the potential downsides of ethical relativism,[6] where ethical standards become contingent upon individual perspectives and cultural norms. A shared moral framework provides a stable foundation and fosters a collective understanding of ethical boundaries that transcend the temporary shifts of societal attitudes. Second, the pursuit of an idealistic and universal sense of values in AI aligns with the overarching goal of technology. Looking from a broader perspective, it is a tool to enhance and deepen human existence. By anchoring AI development in principles universally acknowledged as virtuous, we not only mitigate the risks of ethical divergence but also ensure that technological creations contribute positively to the well-being of individuals and society at large.

Striking a balance between personalised experiences and collective well-being, navigating ethical trade-offs and dilemmas requires a deep exploration of ethical considerations surrounding the nature of virtue, human flourishing, and ethical boundaries. The collective approach that has been explored above provides a more explicit justification for the assumption of a universal moral paradigm by grounding it in shared human experiences and the enduring principles of philosophical traditions that have guided ethical ground throughout history.

2 Ethical responsibilities of AI developers

Is achieving artificial moral agency an unattainable objective? What exactly are the difficulties and barriers when it comes to implementing ethical virtues in AI? A thorough deliberation of this matter is urgent here. The ethical responsibilities that AI developers bear in designing and programming AI systems are of utmost importance in the ever-evolving technological landscape. As they navigate the complexities of integrating moral values into AI algorithms and decision-making processes, developers find themselves confronted with a myriad of considerations and dilemmas.

One crucial aspect that AI developers must grapple with is the issue of bias. The ethical principles that I employ throughout the text involve recognising and actively working to mitigate the impact of biases in AI algorithms in this context. Regrettably, AI systems are also not immune to the biases and prejudices that exist within society. These biases are prone to permeate the algorithms and spread discrimination and inequality in most aspects of our lives, such as facial recognition, job hiring and criminal sentencing algorithms. One real case example that illustrates the issue of bias in AI systems is the facial recognition technology developed by tech companies. Studies have shown that these systems have demonstrated significant bias, particularly in their accuracy when identifying individuals from different racial and ethnic backgrounds.[7] Research conducted by Joy Buolamwini, a

[5] https://www.un.org/en/about-us/universal-declaration-of-human rights#:~:text=Drafted%20by%20representatives%20with%20different,all%20peoples%20and%20all%20nations.

[6] While ethical relativism offers insights into cultural diversity and encourages a nuanced understanding of moral perspectives, its pitfalls, such as moral arbitrariness and potential hindrance to moral progress, underscore the importance of considering ethical principles in the quest for a more just and equitable framework.

[7] https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/.

computer scientist at MIT, revealed that widely used facial recognition systems had higher error rates when attempting to identify women and individuals with darker skin tones compared to white males.[8] This bias stemmed from the imbalanced representation of training data, which predominantly featured lighter-skinned and male faces. Consequently, these systems exhibited higher rates of misidentification and perpetuated existing biases and discriminatory outcomes. Another study on existing biases in the criminal justice system showed by ProPublica that a popular algorithm used in the United States for predicting future criminal behaviour exhibited racial bias.[9] The algorithm predicted a higher likelihood of reoffending for black defendants compared to white defendants, even when controlling for other factors. Biases present in existing hiring data can also be perpetuated and amplified by AI algorithms, leading to discriminatory outcomes. The incident with Amazon's recruiting tool highlighted the importance of addressing bias in training data and ensuring that algorithms are designed to promote fairness and equal opportunities.[10] It was reported that, in 2018, Amazon had developed an AI-powered recruiting tool to automate the screening of job applicants. The goal was to streamline the hiring process and identify top candidates based on historical hiring patterns. However, it was later discovered that the system had developed a bias against women. The algorithm had been trained on resumes submitted to Amazon over a 10-year period, which were predominantly from male applicants due to the male-dominated tech industry. As a result, the algorithm learned to favour male candidates and downgraded resumes that included terms associated with women. As a form of confirmation bias,[11] AI systems can also reinforce existing biases by relying on historical data that is itself biased. This form of bias is more apparent in criminal justice, where past decisions by judges and law enforcement officers may contain implicit biases, and consequently, the AI algorithms may learn and perpetuate those biases. For instance, if an algorithm is trained on historical data from a jurisdiction with higher rates of arrests and convictions for certain offences, it may disproportionately recommend harsher sentences for individuals from other regions.[12]

In the complex realm of AI development, ethical virtues emerge as guiding lights to confront the pervasive issue of bias. Unveiling the hidden prejudices within algorithms, particularly in facial recognition and criminal justice systems, underscores the imperative for a compassionate and empathetic approach. Kindness in AI demands fairness and inclusivity, recognising the potential harm inflicted by biased technologies. Empathy compels developers to understand the disparate impacts of algorithms on different communities, fostering a commitment to rectify imbalances and create equitable solutions. Meanwhile, compassion becomes the driving force behind actively addressing societal implications, urging developers to take responsibility for the consequences of bias and advocate for fairness in algorithmic decision-making. Weaving these principles into the fabric of AI development is to pave the way for technology that not only advances innovation but does so with a profound sense of humanity, consideration, and justice. These instances serve as stark reminders of the consequences when these ethical principles are overlooked. In rectifying these biases, AI developers stand at the intersection of technology and humanity, with the power to cultivate a technological landscape that mirrors the virtues of kindness, empathy, and compassion.

But then, can the recognition of biases in AI systems prompt action to address the disparate impact on certain groups? One notable example of taking steps to mitigate disparate impact and promote equal opportunities in AI systems is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) case.[13] In 2016, ProPublica investigated the COMPAS algorithm that is used in the United States to assess the likelihood of recidivism for individuals involved in the criminal justice system. The investigation found that the algorithm showed racial bias, as it was more likely to mistakenly label black defendants as having a higher risk of reoffending compared to white defendants. The case gained significant attention and raised concerns about the potentially discriminatory impact of AI algorithms in sentencing decisions. In response to the investigation and public scrutiny, the court system in Broward County, Florida, where COMPAS was used, took steps to

[8] Buolamwini Joy, & Gebru Timnit [3]. p.12.

[9] Retrieved from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[10] Dastin, J. [4].

[11] Confirmation bias, which is the tendency to favour information that confirms our existing beliefs while ignoring contradicting evidence, can significantly impact our decision-making processes and reinforce our biases. It is almost like wearing tinted glasses that make us see the world in a way that aligns with what we already think. We actively seek out information that supports our views and dismiss or downplay information that challenges them. An illustrative example of confirmation bias is the famous Stanford Prison Experiment conducted by Philip Zimbardo in 1971. In this study, participants were assigned roles as either prisoners or guards in a simulated prison setting. It revealed how the guards, given power, exhibited abusive behaviour towards the prisoners. This behaviour was not only influenced by their assigned roles but also shaped by their pre-existing beliefs and expectations about prisoners and authority figures. Confirmation bias, therefore, affects our perceptions, judgments, and actions. Zimbardo, P. G. [28], pp. 243–256.

[12] Jorgensen, R. [8].

[13] https://mallika-chawla.medium.com/compas-case-study-investigating-algorithmic-fairness-of-predictive-policing-339fe6e5dd72.

mitigate the disparate impact of the algorithm. They introduced a policy that prohibited the use of the algorithm in making bond and sentencing decisions. By discontinuing the use of the algorithm in specific decision-making processes, the court system aimed to promote fairer outcomes and equal opportunities for all individuals involved in the criminal justice system, regardless of race or ethnicity. What we see with the implementation of this deliberate action is transparency, which is an essential element in fostering trust and accountability in AI systems. We also see a concerted effort that provides clear explanations regarding how AI algorithms make decisions, ensuring that the inner workings are comprehensible to both experts and end-users. This brought confidence in the fairness and reliability of the AI system, allowing individuals to better understand and scrutinise its outcomes.

Do these virtues hold developers accountable for promoting ethical framework in AI? Should developers take responsibility for the impact of their creations and be prepared to address any unintended consequences or harms that may arise? Establishing mechanisms for ongoing monitoring and evaluation of the AI system's performance, as well as implementing procedures to rectify and learn from any mistakes or biases, are the very first steps in the developers' accountability challenge. For example, during its development, OpenAI recognised the potential for misuse and took responsibility for ensuring the technology was used responsibly. They implemented a content filtering system to prevent the generation of illegal, unethical and harmful content.[14] Microsoft has also established a set of "Responsible AI Principles", including fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability. By adopting such principles, these companies aimed to set a framework for taking responsibility for the impact of their AI creations and actively working to address unintended consequences or harms that may arise.[15]

I intentionally called these values challenges. The reason is that implementing these values in a growing system always calls for continuous attention and philosophical reflection. Now imagine a scenario in which a team of developers working on an AI system designed to assist with job recruitment processes considering the strategies of equality, equity, diversity and inclusion. They are committed to ensuring fairness and avoiding biases in the system's decision-making. As part of the development process, they extensively test the system using diverse datasets and conduct thorough evaluations to identify and rectify any biases that may arise. However, during the testing phase, they encountered a challenge related to the system's adaptability. They discover that the AI system, while effectively reducing biases in recruitment decisions based on gender, inadvertently starts favouring candidates from prestigious educational backgrounds. This unintended consequence arises due to historical data that correlates certain educational institutions with success in job performance. Now, the developers' commitment to responsible AI development and mitigating potential negative impacts encounters a philosophical challenge rooted in the concept of fairness. As they aim to reduce biases in the AI system's decision-making, they face the dilemma of how to balance equal opportunities for candidates while considering their qualifications.

At the heart of this challenge, there is a tension that lies between two ethical principles: meritocracy/social mobility and equal opportunity. The developers initially focus on countering biases related to gender, recognising that, historically, certain genders have faced discrimination in recruitment processes. However, in their pursuit of fairness, they face the unintended consequence of favouring candidates from prestigious educational backgrounds. This raises questions about the nature of fairness and the impact of social structures on individual opportunities. Is it fair to solely base recruitment decisions on qualifications that might be influenced by factors beyond an individual's control, such as access to prestigious educational institutions? Should equal opportunities be prioritised over traditional markers of success? How realistic is it to provide equal opportunities for candidates from diverse backgrounds regardless of their educational pedigree? Fairness, in the virtuous sense, does not mean treating everyone exactly the same; it means considering individual circumstances and striving for an equitable outcome. The developers, as virtuous agents, should question whether the traditional markers of success truly reflect the qualifications necessary for the job or, more crucially, if they perpetuate unjust social structures.

The philosophy of ethical virtues, at least in the way exemplified in those instances so far, necessitates an active dedication to identifying and rectifying biases, promoting transparency, and striving for equitable and inclusive systems. Virtue ethics is not just about addressing specific challenges but also about cultivating a long-term ethical culture. It prompts us to thoroughly assess the impact of AI systems on diverse groups and prioritise fairness and equal treatment. By embracing kindness as a guiding principle, we have the potential to construct AI systems that embody these values and contribute to a society that is both just and compassionate.

[14] OpenAI's GPT-3 and Content Filtering: OpenAI Blog: "Language Models are Unsupervised Multitask Learners" Link: https://openai.com/blog/language-unsupervised/.

[15] Microsoft AI Principles—Link: https://www.microsoft.com/en-us/ai/responsible-ai.

3 Ethical frameworks in AI

Existing ethical frameworks provide a valuable starting point for incorporating virtues into AI systems. Principles such as beneficence, non-maleficence, and respect for human dignity serve as foundational pillars for the ethical development of these values in the realm of AI. It is essential to evaluate and expand upon these frameworks to address the unique challenges and opportunities presented by AI technology.

Beneficence implies that individuals and institutions have ethical obligations to actively contribute to the welfare of others. It emphasises the importance of altruism, compassion, and social responsibility in decision-making and actions, where these values involve navigating the balance between competing interests and determining the best course of action to maximise overall benefits. The idea of "promoting well-being and acting in ways that enhance the overall welfare of others" is extensively advocated and integrated mainly in utilitarianism. John Stuart Mill's utilitarianism[16] required that actions should be judged by their tendency to promote the greatest happiness for the greatest number of people and how much these actions maximise overall well-being and lead to the greatest overall happiness and the prevention of harm. Whereas Kant's utilitarian framework[17] is based on the concept of duty and respect for moral principles. The categorical imperative includes a principle of beneficence, which urges individuals to act in ways that contribute to the welfare and happiness of others. Contemporary philosopher Peter Singer, who is known for his work on effective altruism and the ethics of giving, also argues that individuals have a moral obligation to prevent suffering and promote well-being to the best of their abilities. On this account, we have free will to choose the course of our actions, and this freedom emphasises the importance of making choices that maximise overall welfare and alleviate unnecessary suffering.

In Nussbaum's account, freedom of choice shows itself as "capabilities"[18] which we need to identify and protect. Essential capabilities enable us to live a dignified life. Nussbaum's approach encompasses a broad range of dimensions, including physical and mental health, education, and social relationships, all aimed at enhancing well-being. In her cognitive theory of emotions and normative theory of human development, she presents an intriguing concept of political compassion.[19] This perspective expands the understanding of compassion beyond a mere personal sentiment, establishing it as a guiding principle, which is something more universal and applicable across all cultures within institutions. What Nussbaum actually does here is challenge the ethical tradition that neglects the significance of external goods in human flourishing. She contends that individual impulses of compassion, which prompt us to address the material needs of others, can be grounded in theoretical principles and inform institutional arrangements and political objectives. This suggests that compassion when guided by rationality, can shape social structures and contribute to the realisation of collective well-being. This approach also highlights the need to protect essential capabilities from harm. In the context of AI, beneficence requires actively mitigating potential risks and negative consequences associated with AI systems. This includes addressing issues such as algorithmic bias, privacy breaches, and discriminatory outcomes that can harm individuals or perpetuate social inequalities. It involves taking proactive measures to ensure AI systems do not cause unnecessary harm or compromise human well-being.

Let us take the healthcare system as an example. We utmost hope that AI-enabled healthcare systems and algorithms can improve diagnostic accuracy and treatment outcomes. In one specific case of the detection of breast cancer using mammograms, deep learning algorithms have been developed and trained on large data of mammograms to identify patterns and anomalies associated with breast cancer. These algorithms help radiologists detect potential abnormalities with higher accuracy and sensitivity, leading to earlier detection and improved patient outcomes. A very valuable insight and research paper called "Dissecting racial bias in an algorithm used to manage the health of populations"[20] by Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan discusses the detection of breast cancer cases extensively;

> "Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. … a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appears to be an

[16] Mill, J. S. [11].

[17] O'Neill, Onora. [20]; Bennett, Christopher. [2].

[18] Nussbaum [15, 19].

[19] Nussbaum [18]. p. 145.

[20] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, [27]. pp. 452–453.

> effective proxy for health by some measures of predictive accuracy, large racial biases arise."[21]

On another note, Obermeyer discusses the same issue;

> "[if we take] family history as a variable to use to determine who needs to be screened for cancer and people who have a family history of breast cancer, like they're at higher risk and so we want to screen them more. But if you think about what family history is, it's something about history, it's something about your family's historical access to health care. So now, if I told you here are two women, there's one Black woman and there's one White woman, and neither of them has a family history of breast cancer, you can feel better about that. For the white woman whose family has historically had a lot of access to care and if they had breast cancer, they're likely to get diagnosed. But now for the Black woman, the fact that she doesn't have a family history is a lot less meaningful, given that we're not just dealing with the inequalities in medicine and the health care system today. … we're dealing with all those inequalities over the past decades when they were much, much worse."[22]

The case above highlights the broader inequalities and historical issues of access to healthcare. It brings attention to the fact that historical inequalities can have a significant impact on an individual's current health status and their family's health history. As Nussbaum agreed on his capabilities account, promoting well-being and enhancing the overall welfare of others requires acknowledging and addressing historical disparities in healthcare access and treatment. In the same vein, kindness, empathy and compassion in AI systems entail designing algorithms and frameworks that actively address and mitigate historical biases.

Another foundation of ethics of AI, which is non-maleficence, avoiding harm, ensures that AI systems are designed and deployed in ways that prioritise the well-being and safety of individuals in the context of minimising potential harm, transparency and explainability. A critical examination of the ethical implications of incorporating non-maleficence as a guiding principle may foster a discussion on the responsible development and use of AI technologies, prioritising the prevention of harm and the preservation of human dignity even in the most challenging and high-stakes scenarios. Successfully, there are countless real case examples where AI effectively accomplished non-maleficence. One of the cases is the use of AI in autonomous vehicles with the goal of reducing traffic accidents and improving road safety.[23] Companies like Waymo, Tesla, and Uber have been developing self-driving cars that utilise AI algorithms to perceive the environment, make driving decisions, and operate the vehicle. The aim of these AI-driven systems is to reduce human error, which is a significant contributor to road accidents.[24] Through this application of AI, autonomous vehicles have the potential to significantly reduce accidents caused by intentional or unintentional human error, such as distracted driving, fatigue, or impaired judgment. The ongoing development and refinement of AI algorithms for autonomous vehicles continue to highlight non-maleficence by striving to minimise potential harm and prioritise the well-being and safety of individuals on the road.

There are also examples that showcase how AI technologies can respect human dignity by acknowledging and supporting the inherent worth and agency of individuals with disabilities. By embracing the principles of inclusivity, personalisation, and autonomy, AI-powered assistive devices promote equality, independence, and the full participation of individuals with disabilities in various aspects of life. AI-powered assistive technologies, such as prosthetic limbs, communication devices and mobility aids, are designed to enhance the independence, functionality, and quality of life of people with disabilities. These technologies leverage AI algorithms to adapt and respond to user needs, allowing individuals to perform daily activities, communicate effectively and navigate their surroundings more easily. By tailoring the functionality of assistive devices to individual preferences and capabilities, AI systems respect the dignity of individuals with disabilities by recognising their unique needs and promoting their autonomy. The customisation and personalisation offered by AI-powered assistive technologies empower individuals to maintain control over their own lives and participate more fully in society as autonomous beings.[25]

[21] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, [27], p. 453.

[22] Ziad Obermeyer, "Can AI Improve Health Without Perpetuating Bias?" https://www.commonwealthfund.org/publications/podcast/2023/apr/can-ai-improve-health-without-perpetuating-bias.

[23] Yifang Ma, Zhenyu Wang, Hong Yang and Lin Yang [25].

[24] By implementing advanced sensors, computer vision, and machine learning algorithms, autonomous vehicles analyse vast amounts of data in real-time, detect and respond to potential hazards, and make split-second decisions to avoid collisions. The AI algorithms are trained on extensive datasets, including various driving scenarios, to improve their ability to navigate safely.

[25] Georgieva, P., & Angelova, A. [6]; Shadiev, R., & Tlali, M. [21]; Somanath, G., Bhattacharya, S., Deekshit, H., et al. [22].

4 Future implications and possibilities

Throughout our discussion kindness, empathy and compassion defined as acting in ways that promote the overall well-being and welfare of others, should be regarded as fundamental principles guiding the design, implementation, and utilisation of AI technologies. To achieve this objective, interdisciplinary collaboration between ethicists, technologists, and policymakers is paramount in navigating the complex and evolving field of AI ethics. Such collaboration fosters a holistic—and more realistic consideration the universal nature of ethical principles and challenges—understanding of the ethical challenges and opportunities posed by AI technologies. It enables the development of robust frameworks, guidelines, and regulations that integrate ethical principles, technical expertise, and societal values. Through this collaborative effort, stakeholders ensure that AI systems are designed, deployed, and utilised in a manner that prioritises kindness, upholds human dignity, and contributes to the overall well-being of individuals and communities. While ethicists contribute a philosophical lens to the discourse, enabling in-depth reflection on the moral implications and consequences of AI systems, engage in critical analysis, identifying potential ethical challenges, biases, and social implications that arise from the expanding role of AI in various domains, technologists are responsible for the practical design, development, and implementation of AI systems and policymakers are essential stakeholders in shaping the regulatory landscape surrounding AI. With the authority to establish laws, regulations, and standards governing AI development, deployment, and usage, policymakers also engage with both ethicists and technologists to comprehend the ethical nuances and technical complexities associated with AI. Collaborating with ethicists will allow policymakers to develop comprehensive and inclusive policies that address potential ethical concerns and ensure that AI systems operate in a manner that promotes kindness, safeguards individual rights, and upholds societal values.

Another valuable initiative is the essential ethical training within the domain of AI practitioners and students stands as an indispensable facet of ethical consideration.[26] Ideally, the pedagogical approach to AI, computer science, and data science education should encompass a comprehensive curriculum that rigorously addresses ethical and security dimensions. However, the domain of ethical understanding alone proves insufficient. While ethics equips practitioners with an awareness of their responsibilities to diverse individuals –and more collectively diverse cultures–, the translation of ethical principles into practice necessitates a concurrent augmentation of technical proficiency. This augmentation becomes apparent in the integration of technical precautions throughout the developmental and evaluative phases of AI systems. As practitioners endeavour to imbue AI systems with principles of justice, equity, and accountability, the transformative potential lies not only in ethical mindfulness but in the synthesis of this mindfulness with the technical adeptness required to actualise ethical considerations in system architecture and testing. The goal extends beyond the mere propagation of an ethical ethos; it sees technology as an enabler, rather than an impediment to accountability. Consider, for instance, ongoing research endeavours focused on enhancing the interpretability of machine learning outcomes. The instantiation of an interpretable model operates as an exemplar in this context. Beyond operational efficiency, interpretability serves as a means of empowerment. An interpretable model elucidates decision-making processes, affording individuals the capability to scrutinise underlying assumptions and procedural difficulties. This union of ethical consciousness and technological sophistication marks a paradigm shift where transparency is not a temporary ideal but an intrinsic aspect, fostering accountability as an actionable principle within the landscape of AI ethics.

In contemplating the potential future developments and applications of kindness, empathy and compassion in AI, we find ourselves venturing into a realm where machine learning, natural language processing, and affective computing intertwine to enhance AI systems' capacity to perceive, understand, and respond to human emotions. These advancements hold the promise of revolutionising the way AI interacts with and supports human beings. Philosophical inquiry requires the ethical considerations and concerns that arise alongside these transformative possibilities. As we dig into the possibilities of machine learning, we witness AI systems acquiring the ability to learn from vast databases, recognise patterns and make predictions with unprecedented accuracy. Such advancements offer immense potential for infusing kindness into AI systems. By training these systems on datasets that exemplify kindness, we can imbue them with the capacity to recognise and respond to situations in a manner that promotes the well-being and welfare of individuals.

In our pursuit of incorporating fundamental virtues into AI systems, we must also consider the inherent limitations of machines. It is arguably indisputable that while AI can simulate empathy and kindness to a certain extent, it lacks the depth of human experience and the genuine emotional connection that stems from shared humanity. It is crucial to maintain the distinction between AI-assisted support and genuine human care, ensuring that AI systems enhance human well-being without replacing the fundamental human-to-human connection. In conclusion, the future developments and applications of kindness in AI

[26] Executive Office of the President, "Preparing for the Future of Artificial Intelligence", October 2016, pp. 30-31.

hold immeasurable promise for transforming how we interact with technology. Advancements in machine learning, natural language processing, and affective computing can elevate AI systems' ability to perceive, understand, and respond to human emotions. However, we must approach these advancements with a critical eye, navigating the ethical considerations and concerns they raise. By establishing clear guidelines, we can harness the potential of AI to enhance kindness while preserving the irreplaceable value of genuine human connection.

5 Conclusion

In contemplating the ethics of AI, we find ourselves facing weighty questions about the impact of this emerging technology on humanity and society. As concerned with the pursuit of virtue and the well-being of individuals, I see the philosophy of universal ethical virtues as a guiding principle that could shape our approach to AI ethics.

Kindness, empathy and compassion lie at the core of ethical philosophy. They encompass goodwill towards others. When we apply this philosophy to AI, we recognise that the development and use of AI systems should prioritise the well-being and flourishing of all individuals affected by its decisions. AI systems, while powerful tools are not immune to the biases and prejudices that exist within society. These biases can permeate the algorithms and perpetuate discrimination or inequality. Thus, it is our ethical duty to ensure that AI technologies align with the values of fairness, equality, and inclusivity. If we actively work to identify and rectify biases, from the training data to the algorithm design, a diverse and inclusive development process, coupled with ongoing monitoring and evaluation, we can help reduce discrimination and ensure equal opportunities for all individuals affected by AI systems' decisions. Additionally, the philosophy of kindness calls for collaboration and accountability. It urges AI developers, policymakers, researchers, and communities to come together to address the ethical challenges posed by AI. By engaging diverse perspectives and involving impacted communities, we can collectively shape the development and deployment of AI technologies in a manner that upholds the values of kindness, fairness, and human flourishing. The ethics of AI, when viewed through the lens of philosophy, compels us to prioritise fairness, equality, transparency, and accountability. By infusing AI systems with these values, we can harness the potential of AI to enhance the well-being of individuals, foster inclusive societies, and promote a world that aligns with our highest ideals of virtue and compassion.

Acknowledgements I, Serap Keles, acknowledge and confirm that all sources of funding, including research grants, fellowships, or sponsorships, relevant to the research conducted and presented in this paper are disclosed within the acknowledgments section. I am committed to upholding the highest standards of academic integrity and transparency, and I have strived to adhere to ethical principles throughout the research process, data collection, and analysis. In the event that any potential conflict of interest arises after the publication of this paper, I commit to disclosing such information promptly to the appropriate parties.

Funding Furthermore, I, Serap Keles confirm that I have not received any financial support, grants, or other forms of funding from any commercial or non-commercial entities that might be perceived as influencing the findings or conclusions of this paper. There are no affiliations with companies, institutions, or organisations that could potentially present a conflict of interest with the content discussed in this manuscript.

Data availability The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest I, Serap Keles, declare that I have no financial or personal conflicts of interest that could have influenced the content, interpretation, or presentation of the academic paper titled "Navigating in the Moral Landscape: Analysing Bias and Discrimination in AI Through Philosophical Inquiry." This work represents an independent and unbiased effort on my part, and I affirm that no external party or organisation has influenced the research, data analysis, or conclusions presented herein.

References

1. Article 36. The impact of autonomy and artificial intelligence on the behaviour of weapons. (2019). Retrieved from https://article36.org/wp-content/uploads/2019/06/Article36_Impact-of-Autonomy-Artificial-Intelligence-Behaviour-Weapons-June-2019.pdf
2. Bennett, C.: Utilitarianism. What is this Thing Called Ethics?, pp. 55–73. Routledge, London (2010)
3. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st Conference on Fairness, Accountability, and Transparency, pp. 77–91 (2018). https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf
4. Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. (2018). Retrieved from: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.
5. Executive Office of the President. Preparing for the future of artificial intelligence, pp. 30–31 (2016)
6. Georgieva, P., Angelova, A.: Artificial intelligence in assistive technologies for people with disabilities. In 2020 IEEE 12th International Conference on Intelligent Systems (IS), pp. 283–288, (2020)
7. Human Rights Watch. Cluster munition ban at 10: clear success, lingering challenges. (2016). Retrieved from https://www.hrw.org/report/2018/08/30/cluster-munition-ban-10/clear-success-lingering-challenges
8. Jorgensen, R.: Algorithms and the individual in criminal law. Can. J. Philos. **52**(1), 61–77 (2022). https://doi.org/10.1017/can.2021.28

9. Kaufman, A.: Capabilities and freedom. J Polit PhilosPolit Philos **14**(3), 289–300 (2006)
10. MacIntyre, A.C.: After Virtue. A Study in Moral Theory. University of Notre Dame Press, Notre Dame (1984)
11. Mill, J.S. Utilitarianism. London, Parker, son, and Bourn. [Pdf] Retrieved from the Library of Congress. (1863). https://www.loc.gov/item/11015966/
12. Nussbaum, M.: Human functioning and social justice. In defence of Aristotelian essentialism. Polit. Theory **20**(2), 202–246 (1992)
13. Nussbaum, M.: Compassion: the basic social emotion. Soc. Philos. Policy **13**(1), 27–58 (1996). https://doi.org/10.1017/S0265052500001515
14. Nussbaum, M.: Women and Human Development: The Capabilities Approach. Cambridge University Press, Cambridge (2000)
15. Nussbaum, M.: Capabilities as fundamental entitlements: sen and social justice. Fem. Econ. **9**(2/3), 33–59 (2003)
16. Nussbaum, M.: Frontiers of Justice: Disability, Nationality, Species Membership. Harvard University Press, Cambridge (2006)
17. Nussbaum, M.: Creating Capabilities. Harvard University Press, Cambridge (2011)
18. Nussbaum, M.: Political Emotions: Why Love Matters for Justice. The Belknap Press of Harvard, Cambridge (2013)
19. Nussbaum, M.: The capabilities approach and the history of philosophy. In: Chiappero-Martinetti, O., Qizilbash, A. (eds.) The Cambridge Handbook of the Capability Approach, pp. 13–39. Cambridge University Press, Cambridge (2020)
20. O'Neill, O.: A simplified account of Kant's ethics, pp. 411–415. Blackboard Web (2014)
21. Shadiev, R., Tlali, M.: Intelligent assistive technologies: a systematic review of recent applications and challenges. IEEE Access **9**, 3603–3617 (2021)
22. Somanath, G., Bhattacharya, S., Deekshit, H., et al.: Assistive Artificial intelligence for the visually impaired: a personalised audio interface for accessibility. In Proceedings of the 4th International Conference on HCI in Business, Government and Organizations, pp. 17–32 (2019)
23. Sławiński, M.: Can institutions be compassionate? On Martha C. Nussbaum's theory of political compassion. Psychol. Res. **8**(5), 204–213 (2018). https://doi.org/10.17265/2159-5542/2018.05.003
24. Turkle, S.: Alone Together: Why We Expect More from Technology and Less from Each Other. Basic Books (2011)
25. Yifang, M., Wang, Z., Yang, H., Yang, L.: Artificial intelligence applications in the development of autonomous vehicles: a survey. IEEE/CAA J. Autom. Sin. **7**(2), 315–329 (2020). https://doi.org/10.1109/JAS.2020.1003021
26. Ziad, O.: Can AI improve health without perpetuating bias? (2023) https://www.commonwealthfund.org/publications/podcast/2023/apr/can-ai-improve-health-without-perpetuating-bias.
27. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**, 447–453 (2019). https://doi.org/10.1126/science.aax2342
28. Zimbardo, P.G.: On the ethics of intervention in human psychological research: with special reference to the Stanford prison experiment. Cognition **2**(2), 243–256 (1973)
29. https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/
30. https://mallika-chawla.medium.com/compas-case-study-investigating-algorithmic-fairness-of-predictive-policing-339fe6e5dd72
31. Microsoft AI Principles. https://www.microsoft.com/en-us/ai/responsible-ai. (2023)
32. OpenAI's GPT-3 and content filtering: OpenAI Blog: language models are unsupervised multitask learners. (2023). https://openai.com/blog/language-unsupervised/
33. Yemen: Cluster Munitions Wound Children. (2023). https://www.hrw.org/news/2017/03/17/yemen-cluster-munitionswoundchildren#:~:text=Cluster%20munitions%20are%20prohibited%20by,and%20the%20United%20Arab%20Emirates

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.