Świat Idei i Polityki • 2025, t. 24, nr 1, s. 65–81
World of Ideas and Politics • 2025, vol. 24, no. 1, pp. 65–81
ISSN 1643-8442
DOI: 10.34767/SIIP.2025.01.04

Zoya Slavina[1]

Toward ethical AI. Addressing bias, privacy, and accountability in data-driven systems

SUMMARY This article investigates the ethical implications of AI and data-driven systems, with a focus on how bias, privacy, and security concerns shape the deployment of AI technologies across various domains. The primary problem addressed is the tendency of AI algorithms to amplify societal and technical biases, thereby undermining fairness and potentially harming marginalized groups. However, algorithmic and AI decision-making can also help reduce bias by making processes more transparent and fact-based, provided the systems are developed ethically. To explore these challenges, the research adopts a conceptual and qualitative approach, synthesizing insights from existing policy frameworks, case studies, and scholarly literature on AI ethics and data science. The core hypothesis posits that by integrating ethical principles, robust oversight, and multidisciplinary collaboration into the design and development of AI systems, it is possible to mitigate harmful biases, protect privacy, and enhance public trust. This hypothesis is examined by analyzing how different stakeholders – AI developers, policymakers, and end-users – contribute to the emergence or reduction of bias in algorithmic processes. The article concludes that comprehensive regulatory standards, improved transparency, and regular model updates are essential for minimizing risks, while active engagement by both experts and the broader public is critical to ensure AI technologies operate in a manner consistent with democratic and humanistic values.

KEYWORDS AI ethics, data-driven systems, algorithmic bias, privacy, accountability, ethical governance

[1] Zoya Slavina, MA, University of Białystok, e-mail: z.slavina@uwb.edu.pl, ORCID: 0009-0007-6346-3802.

Introduction

The rapid advancement of Artificial Intelligence (AI) technologies has precipitated significant shifts across multiple facets of contemporary society. Originally perceived as primarily technical constructs, AI systems are now deeply intertwined with ethical, societal, and even political dimensions, compelling researchers and practitioners alike to grapple with their broader implications. The prioritization of technological progress over essential human-oriented components – such as values, ethics, emotions, and culture – has given rise to ethical dilemmas when autonomous technologies interact with human beings at scale. The fields of AI ethics and data ethics emerge from this tension, becoming critical lenses through which the responsible development and deployment of intelligent systems are examined.

AI ethics, broadly defined as principles governing responsible AI use, overlaps with data ethics, which interrogates the moral implications of collecting, processing, and disseminating large volumes of data. This intersection points to a complex reality wherein data-driven decision-making profoundly influences human lives, raising urgent questions about fairness, transparency, autonomy, and societal trust. The ethical concerns associated with algorithmic systems – including data privacy, systemic biases, and predictive analytics – are increasingly prominent due to their potential to exacerbate societal inequalities, infringe upon human rights, or amplify discrimination on an unprecedented scale.

Simultaneously, the societal implications of AI technologies are extending far beyond ethical frameworks and data management. The pervasive integration of AI into daily life inevitably reshapes not only human-machine interactions but also the structure and functioning of political systems and international relations. Though this article does not delve explicitly into political or international contexts, acknowledging the broader repercussions on governance, geopolitical dynamics, and power structures underlines the gravity and urgency of the issues discussed. The politics of data-driven decision-making, algorithmic governance, and surveillance capitalism, for example, exemplify the profound ways in which technological advancements influence societal and international power dynamics. Thus, understanding AI and data ethics demands situating the discourse within these larger societal transformations.

The purpose of this paper is to provide an in-depth analysis of the ethical issues that arise from AI and data-driven technologies, focusing specifically on key concerns such as data privacy and safety, algorithmic biases, and the socio-technical complexities of predictions and forecasting. Each of these elements will be critically examined through examples from contemporary society, including sectors such as justice, healthcare, financial services, and media consumption. By doing so, the article aims to illuminate both the risks and opportunities embedded in the implementation of AI, arguing for an informed, ethical, and interdisciplinary approach as indispensable in mitigating potential harms and maximizing societal benefit.

AI ethics: definition and related concepts

The emergence of ethical issues in AI is attributed to the prioritization of technological advancements over crucial components such as emotions, culture, values, and ethics (Rai, 2022). This prevailing technological mindset surrounding AI and other emerging technologies often neglects the importance of these human-oriented aspects. Consequently, a significant challenge arises when highly advanced autonomous agents engage with humans on a large scale, as they lack a comprehensive understanding and appreciation of the "humane part" of human beings (Rai, 2022, p. 213). Thus, *AI ethics* – generally defined as a set of ethical principles, guidelines, and considerations that govern the responsible and ethical use of AI (AI HLEG, 2019; Coeckelbergh, 2020) – becomes a central element in bridging human and machine social interactions. It is also closely related with *data ethics* that involves the evaluation of the moral implications and potential consequences associated with the collection, storage, analysis, sharing, and dissemination of data (O'Neil, 2016). The recognition that AI is no longer exclusively a technical concept but also encompasses broader ethical and societal considerations is central. The type of shift toward considering societal and ethical implication of a technology can be seen not only in AI industry, but also in other innovations – which are all eventually interconnected – with inherent common socio-technical challenges. Jemelniak and Przegalińska (2020) point out that in the present era, we have moved beyond the idealized notion of internet and communication technologies (ICT) as instruments that solely empower democracy and have minimal negative consequences. Moreover, the convergence of these technologies with emerging ubiquitous connectivity – exemplified by the Internet of Things (IoT), a diverse set of digitally interconnected devices typically involving sensors – blurs the boundary between private and public domains (whether citizen, governmental, or corporate), thereby significantly increasing the complexity of the networks and interactions involved.

The field of AI has seen numerus proposals integrating ethics, humanities, and social science research to improve safety and maximize benefits. Governments and supranational organizations – such as through AB 331 in the USA, the AI Act in the EU, and the New Generation AI Development Plan in China – have introduced policy and regulatory initiatives. Likewise, corporate, academic, and professional bodies (e.g., Microsoft, 2018; IEEE, N/A.; ISO, 2022a; 2022b) propose ethical AI frameworks that, despite some differences, largely converge on similar core values. For instance, the EU's *Ethics Guidelines for Trustworthy AI* (AI HLEG, 2019) emphasize respect for human autonomy, harm prevention, fairness, and explicability, while Microsoft's ethical principles (Microsoft, 2018; Smith & Browne, 2021) highlight fairness, reliability, privacy, inclusivity, transparency, accountability, and the promotion of human values. Embedding these ethical concept from the outset can enhance accountability, responsibility, and transparency (Coeckelbergh, 2020) – crucial for tackling challenges like data privacy, bias, and the complexities of prediction and forecasting in contemporary society. These challenges are discussed in this paper.

Initiatives on regulatory and ethical aspects of AI commonly address concepts such as trustworthy AI, explainable AI (XAI), interpretable AI, and responsible AI. Trustworthy AI emphasizes compliance with laws, ethical principles, and technical/social robustness (AI HLEG, 2019; Lipińska, 2022). XAI focuses on techniques that clarify AI processes and outputs for external stakeholders (e.g., a bank client), whereas interpretability involves an internal understanding of how and why decisions are reached (e.g., by a bank officer). Responsible AI centers on assigning clear responsibility to uphold ethical values (Rai, 2022). All these terms converge in the broader notion of intelligibility (Floridi, 2023), which integrates various forms of transparency crucial to ethical AI deployment.

AI is both a science and a technology typically described in terms of its capabilities and functionalities. It is broadly categorized into three forms: narrow AI (already in use today), Artificial General Intelligence (AGI) (hypothetical human-level intelligence), and Super AI (theoretical intelligence surpassing humans) (Searle, 1983; Coeckelbergh, 2020; IBM, N/A.). Narrow AI handles specialized tasks – like playing chess – and can adapt to uncertain conditions with minimal human supervision. In contrast, AGI and Super AI remain speculative, with uncertain prospects for realization (Searle, 1983; Bostrom, 2016; Howard, 2019).

Historically, AI systems have evolved from rule-based expert systems – relying on hand-coded logic – to machine learning (ML) approaches that derive rules from data (Alpaydin, 2016; Iman et al., 2022). ML is an umbrella term encompassing methods that learn patterns from data, including deep learning with artificial neural networks (ANN). These networks, often labeled black-box algorithms, involve complex interactions not easily traced or explained (Boden, 2016; Iman et al., 2022). Generative AI represents a subset of deep learning – typically based on large neural network architectures trained to produce novel outputs – whereas so-called AI agents do not constitute a distinct AI paradigm, but rather refer to systems that orchestrate and coordinate existing models, tools, and rules to perform tasks autonomously within a given environment. Increasingly, hybrid systems combine the transparency of rule-based logic with ML-driven analytics, enabling AI models to process large, unlabeled data sets while handling uncertainty (Coeckelbergh, 2020). Yet, their internal processes differ significantly from human cognition (Boden, 2016). Given AI's rapid evolution, a "multiple disciplinary approach"[2] remains vital to keep pace with emerging architectures and applications.

AI and data science are deeply interlinked, propelled by advances in computing power, digital storage (e.g., cloud solutions), faster internet access, and increasingly affordable electronic devices (Iman et al., 2022). These factors have contributed to the rise of *big data* – vast, complex datasets exceeding traditional processing capabilities (Iman et al., 2022, pp. 75–76). Alpaydin (2016) terms this phenomenon a "dataquake," in which our digital traces drive interest in data analysis and machine learning. *Datafication* – the process of collecting and digitizing various aspects of life, often for economic gain (Mejias & Couldry, 2019) – further accelerates machine learning's role in analyzing big data, in turn spurring new developments in AI.

Role of data privacy and safety in modelling and data analytics

Understanding data privacy and safety in AI requires an awareness of how machine learning (ML) processes data. Unlike human cognition, which involves conscious reasoning, ML relies on statistical methods to detect patterns and correlations that might escape direct human scrutiny (Boden, 2016; Coeckelbergh, 2020). For example, a dynamic ML model can predict traffic congestion using real-time and historical GPS data, adapting to new travel patterns.

[2] A term encompassing different forms of collaboration among the fields.

Machine learning is traditionally classified into supervised learning, unsupervised learning, and reinforcement learning (Alpaydin, 2016). However, additional paradigms such as semi-supervised and self-supervised learning are commonly recognised in modern practice. In supervised learning, the system is trained on labelled datasets (e.g., categorizing individuals into different risk groups based on known indicators). Over time, it refines its ability to classify new inputs according to these predefined categories. By contrast, unsupervised learning works with unlabelled data, discovering its own structures or clusters – such as grouping users with similar browsing behaviours – without prior guidance on which categories to assign (Alpaydin, 2016, p. 107). Finally, reinforcement learning relies on feedback loops (or "rewards") to continually adjust the model's performance, independent of predefined labels (Alpaydin, 2016, p. 127). Recognizing these underlying processes is key to evaluating how AI systems make decisions and ensuring they remain ethically and technically sound.

When personal and sensitive data is involved, users become vulnerable to manipulation and exploitation. AI technology influences commercial decisions, media narratives, and political outcomes, as illustrated by the Cambridge Analytica scandal during the 2016 US elections (Coeckelbergh, 2020). Deepfakes – synthetic audio or visual content generated via deep learning – have also proliferated (Collins & Ebrahimi, 2021; Whittaker et al., 2021), exacerbating disinformation in a post-truth era (McIntyre, 2018; Chesney & Citron, 2019). Although originally developed for benign uses like reading audiobooks, deepfake technology can be weaponized – for example, replicating someone's voice in a phone call to commit fraud (Almutairi & Elgibreen, 2022). These challenges extend to other AI-driven technologies. Jemielniak and Przegalińska note the "hidden costs of ICTs" and the importance of addressing the negative side effects that increasingly shape the social realm (2020, p. 192).

Some individuals are more susceptible to manipulation than others (O'Neil, 2016; Coeckelbergh, 2020), yet technology products often assume fully autonomous, mentally capable adult users (Coeckelbergh, 2020, p. 102). As AI-powered devices enter homes, personal in-house data can be collected and potentially misused. The Internet of Things compounds these privacy, ethical, and political concerns (Coeckelbergh, 2020; Jemielniak & Przegalińska, 2020). For instance, interactive AI toys may gather information on children and their families, storing it externally and raising serious questions about data security and misuse (Coeckelbergh, 2020, pp. 102–103).

Safety and security concerns around AI are escalating due to the interconnectedness of devices and software, all of which are vulnerable to hacking. Malicious actors can exploit AI systems integrated into critical infrastructures such as power grids, transportation networks, and healthcare facilities. The societal impact can be profound, and the cost of developing harmful AI is often relatively low since most necessary tools are publicly available (Smith & Browne, 2021). For example, AI-powered autonomous vehicles may be hacked, risking accidents and public safety (Tencent Keen Security Lab, 2019). In healthcare, unauthorized access to patient data or manipulation of medical devices could jeopardize patient privacy and well-being (Fry, 2018). Genetic data is especially sensitive; it reveals intimate information that cannot be easily changed, thereby linking it permanently to the individual (Fry, 2018). These threats are amplified by AI's growing surveillance capabilities (Cataleta, 2020; Coeckelbergh, 2020; Jemielniak & Przegalińska, 2020). Consequently, trust in AI becomes a social as well as a technical issue, prompting reflection on how much we delegate to AI and what safeguards are needed to protect human rights (Coeckelbergh, 2020).

While advanced algorithms able to provide conclusions and execute tasks on their own, human oversight is essential to ensure results are both accurate and fair (Mazurek, 2023). Determining meaningful human involvement for robust decision-making practices remains a challenge. Autonomous systems may deviate for their intended purpose overtime–due to numerous technical factors – accumulate small errors or even produce nonsensical outcomes the system deems "logical". As Kronmal (1993) warns, "spurious correlations" occur when two seemingly related variables share a common component but lack genuine causality. Vigen's (2015) examples – such as a 94.71% correlation between bedsheet entanglement deaths and cheese consumption – illustrate how data can align statistically without reflecting real-world causes.

Even at the data-gathering stage, choices about how to abstract and represent reality continuously shape any given model (Coeckelbergh, 2020, p. 91). O'Neil (2016) emphasizes that models inherently simplify the complexities of the real world, which can be beneficial in some contexts – such as when a Map abstracts away roads and tunnels for aviation software – but inevitably introduces blind spots. As she explains, "models, despite their reputation for impartiality, reflect goals and ideology […]. Models are opinions embedded in mathematics" (O'Neil, 2016, pp. 23–24). The acceptability of these "opinions" varies across groups, as illustrated by the debate surrounding the COMPAS algorithm used in the U.S. judicial system (see the following section). O'Neil (2016) provides another example of how ideology influences modeling through the value-added model used in Washington, D.C. schools. By evaluating teachers primarily on students' test scores, it overlooks crucial dimensions of teaching – such as class engagement or handling personal challenges – sacrificing accuracy and broader educational goals for efficiency (O'Neil, 2016, p. 23). Although this approach is expedient from an administrative standpoint, it may result in unjust outcomes.

A further ethical consideration is the need to update models regularly as new data becomes available, particularly when decisions affect human lives, improving accuracy by continuously refining their parameters. However, areas such as education or criminal justice require the transparency and robust data integration to ensure fairness too. Consequently, algorithmic outputs in such fields may reinforce existing biases, reflecting the societal values and ideologies embedded in their design – ultimately automating these beliefs on a large scale.

Alpaydin (2016) explains that data and models became the starting points that drive the theory development in general by treating data as a beginning and a driving force, rather than a passive element that confirms the theory set by a human as it used to be. Today, "data starts to drive the operation; it is not the programmers anymore but the data itself that defines what to do next" (Alpaydin, 2016, p. 12). Data mining – a pattern-recognition process employing big data – plays a significant role as a tool for making predictions when the rules behind user behaviors are unknown but can be derived (Alpaydin, 2016, pp. 13–14).

Because machine learning can handle large datasets, AI can be deployed for surveillance in public and private spaces (Cataleta, 2020; Coeckelbergh, 2020; Jemelniak & Przegalińska, 2020). Users may be unaware their tools rely on AI, or fail to grasp how their personal data is processed once they consent to terms of service. Social media platforms, often operated by large corporations, exemplify this: many users rarely read what they agree to, making data repurposing – selling or sharing user data – common (Coeckelbergh, 2020). Although some users are willing to trade their privacy for platform benefits, scholars argue this essentially commodifies human rights (Cataleta, 2020). Chun (2021) refers to this as surveillance capitalism, where continuous datafication means users "pay" for seemingly free services with their data and attention. In this view, democracy and freedom risk being eroded by corporate control over the flow of information (Chun, 2021, p. 6). Coeckelbergh similarly warns that AI may deepen hidden forms of manipulation and surveillance (2021, p. 101). Beyond user data, exploitation may extend to the physical infrastructure supporting digital media production, which can involve unethical resource extraction, large-scale e-waste, and harsh labor conditions (Fuchs, 2013). These concerns highlight the broader social and ethical implications of an increasingly data-driven world, where privacy, human rights, and responsible resource use become ever more pressing.

Bias

The issue of bias in AI is a significant ethical concern involving both societal and technical factors, particularly in data-driven systems (Coeckelbergh, 2020, p. 125). Bias may stem from flawed data – whether incomplete, discriminatory, or poorly structured – or from values embedded within algorithms. For instance, a decision might rely on biased data about specific individuals (e.g., incomplete datasets), or on the algorithm's internal parameters (e.g., selected properties). When data lacks accuracy or is mismanaged, it creates a "garbage in, garbage out" scenario (Houser, 2019), producing outcomes with potentially discriminatory consequences for individuals or groups.

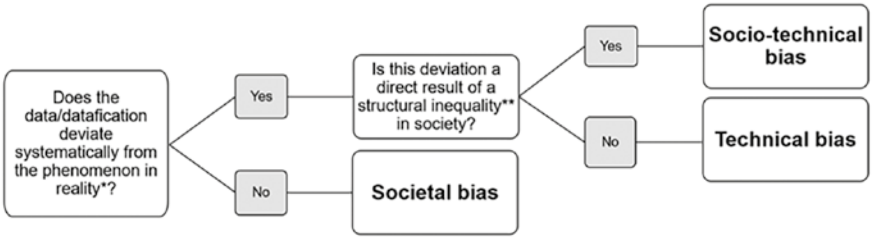
It is worth noting that human decision-making can be compromised by factors beyond bias, including *noise* – systematic and random error (Houser, 2019; Kahneman et al., 2021). Nevertheless, the focus here remains on how bias influences AI-driven decision-making. As Coeckelbergh explains, "[w]hile problems of bias and discrimination have always been present in society, the worry is that AI may perpetuate these problems and enlarge their impact" (2020, p. 126). He further stresses that developers and other stakeholders can introduce bias unintentionally when designing AI systems:

> This can be because they don't understand the AI system well enough, are not sufficiently aware of the problem of bias or indeed of their own biases, or more generally do not sufficiently imagine and reflect on the potential unintended consequences of the technology and are out of touch with some relevant stakeholders (Coeckelbergh, 2020, p. 127).

To address this issue, it is important to introduce widely accepted educational standards for operating and using AI systems. I argue that these reasons are generally brought about by a lack of *ethical training and certification* (in a technological and associated fields) and could be significantly reduced if appropriate standards on operating and using AI systems are introduced widely. Being not aware does not limit the consequences caused by a decision that may also involve access to certain resources and even freedoms. For instance, a person might not get credit, a job position, experience violence against them (D'Ignazio & Klein, 2020, pp. 97–123) or get imprisoned (O'Neil, 2016, pp. 19–31). This causes concerns, especially when entire communities are affected due to their belonging to a certain ethnic, religious or yet differently defined groups.

To understand the origin of bias in a given system, it is essential to determine whether it arises from social prejudice, technical processes, or both. Lopez (2021) proposes a typology of bias in data-driven algorithmic systems consisting of three interrelated categories: societal bias, socio-technical bias, and technical bias. Societal bias occurs when structural inequalities are reflected in the dataset itself; they are inherent rather than accidental. Socio-technical bias stems from a discrepancy between what is meant to be represented and what is actually reflected, often due to systemic inequalities shaping both the data and its interpretation. Technical bias arises from measurement or conceptual inaccuracies within the system, leading to flawed outcomes.

A clear example of societal bias is the Austrian AMS job-placement algorithm, which deducted points for "Gender: Female," thus reducing women's chances of obtaining employment (Lopez, 2021). By contrast, technical bias may be seen in earlier facial recognition technologies that failed to recognize darker-skinned faces, partly due to insufficiently diverse training datasets (Buolamwini & Gebru, 2018; Cataleta, 2020, p. 4; Lopez, 2021). According to Lopez (2021), the extent to which a measured, datafied outcome diverges from the intended representation is a key element distinguishing these three types.


Does the data/datafication deviate systematically from the phenomenon in reality*?
Yes
No
Is this deviation a direct result of a structural inequality** in society?
Societal bias
Yes
No
Socio-technical bias
Technical bias

* Reality, of course, being a highly contested term
** As defined by referring to protected features in legal anti-discrimination regulations

Figure 1. Scheme of socio-technical bias detection

Source: Lopez, 2021, p. 2.

Although this typology helps conceptualize anti-discrimination regulations, particularly when dealing with opaque or "black-box" algorithms, Lopez acknowledges its inherent simplification. In complex real-world scenarios, biases may overlap categories, making them difficult to classify. Socio-technical bias, for example, may be remediable through technical adjustments and stakeholder engagement, whereas societal bias often requires broader political or activist action to address the deep-seated inequalities it reflects –sometimes even necessitating a ban on discriminatory systems (Lopez, 2021, p. 3).

Predictions and forecasting

When addressing statistical AI systems used for predicting and forecasting human behaviour – meaning algorithms designed to project an individual's future actions and potential behavioural patterns (Mamak-Zdanecka et al., 2019) – it is particularly critical to manage biases effectively to maintain the accuracy and reliability of outcomes. Cataleta argues that "[i]n the face of predictive techniques of analysis which are quite invasive, and the discriminatory risks connected to algorithmic choices, the problem of the ethical impact of AI arises" (2020, pp. 8–9). Retrieved data and patterns are utilized to construct a model of possible events, reducing real-life complexities and individuality to certain variables; which in turn, reflects the beliefs and understandings of the algorithms' creators and what they consider significant (O'Neil, 2016; Mamak-Zdanecka et al., 2019). Nowadays, this type of human behaviour modelling using AI is common. It is implemented for diverse industries and purposes, both public and private sectors, including: "music recommendations, product advertising, workforce and education institutions [e.g. admissions and job placements], and banks [e.g. loan qualification]" (Dressel & Farid, 2018, p. 1, clarifications added), the criminal justice system (e.g. recidivism predictions), insurance and healthcare (e.g. insurance pricing and priority of service), among others, to provide risk-assessment and other types of appraisal (Mamak-Zdanecka et al., 2019; Coeckelbergh, 2020).

For assessment questionnaires, it is typically unlawful to inquire about race, ethnicity, religion, or similar sensitive attributes. O'Neil observes that these attributes can be inferred indirectly by correlating, for example, a postcode or a linguistic pattern with one's likelihood to repay a loan or perform well in a job (2016, pp. 21, 59–71). Because residential areas and educational histories can correlate with socioeconomic status, they risk perpetuating biases even when the person is eligible. At the same time, for instance, there are regulations in the US financial sector that require collecting such sensitive information to proof the systems against discriminatory impact.

Kelleher and Tierney (2018) further caution that persistent misclassification or inequitable treatment can create a *self-fulfilling prophecy* – inaccurate beliefs leading to behaviors that make those beliefs seem true (Madon et al., 2011). In predictive policing, for instance, allocating greater police resources to certain neighborhoods (based on historical data) can result in higher recorded crime rates there, reinforcing stereotypes and distrust (Kelleher & Tierney, 2018, pp. 192–194). Researchers warn that "unless used very carefully, data science can actually perpetuate and reinforce prejudice" (Kelleher & Tierney, 2018, p. 193). Though in some contexts (e.g., marketing) classification by race or ethnicity may be less contentious, its use in surveillance systems like "No-Fly Lists" can have severe social repercussions (Baldridge, 2015). Scaled up to large populations, such biased algorithms – termed as *Weapons of Math Destruction* by O'Neil – pose serious risks to human rights and equitable treatment.

An illustrative example of algorithmic bias is the COMPAS risk-assessment system used in the United States to estimate the likelihood of reoffending within two years (Fry, 2018). The system's architecture is proprietary, making it opaque (O'Neil, 2016). Although COMPAS purports to be about 70% accurate, subsequent studies dispute its effectiveness, suggesting the real accuracy rate may be lower (Angwin et al., 2016; Dressel & Farid, 2018). Research also shows that the algorithm can produce disproportionately higher "high-risk" classifications for Black defendants, thereby limiting their freedom and social benefits (Baldridge, 2015; Angwin, 2016; Angwin et al., 2016; Kilbertus et al., 2018). Moreover, risk-assessment tools such as COMPAS are widely employed at various stages of the North American judicial process, from bail decisions to final sentencing (Cataleta, 2020, p. 6).

On the other hand, banking plays a crucial role in facilitating access to resources. Processes like credit risk assessment and insurance increasingly incorporate AI to evaluate customers' profile, which typically combines standard transactional information with demographic details. Here, biases may arise not only from historical datasets and decision-making algorithms, but also from human decision-makers themselves. While AI systems can further perpetuate social and historical biases, they also hold significant potential to mitigate them. Transparency, intelligibility, and ethical development are essential factors that determine whether system deployers can understand the internal mechanisms driving decisions, explain how specific decisions were reached, and subsequently fine-tune or review them as needed. The capability to trace and analyze AI-driven decisions ensures processes become more comprehensible, fair, and efficient. Some suggest that using alternative data sources could improve financial inclusion for underserved populations (Jagtiani & Lemieux, 2019; Sadok et al., 2022). Examples include certain social media activities, geo-location data, and digital behavioral patterns – though this also raises concerns that users might alter their behavior to "game" the system. In contrast, traditional banks tend to focus on standard socio-demographic or banking data (e.g., credit history). Whether AI truly enhances financial decision-making depends heavily on its ethical underpinnings, including the transparency and accountability of the algorithms involved.

The challenges presented in this section lead us to a question of how we define fairness and justice. Should justice be blind and concentrate on avoiding legally-problematic disparate treatment (direct discrimination) and not consider sensitive attributes (e.g. gender, race), or would it be fair to consider them (indirect discrimination) that would incline toward providing more benefits to the individuals from vulnerable subgroups? Research by Kilbertus et al. (2018) in a growing field of *fair learning* for ML systems suggests that it might be possible to avoid both direct and indirect discrimination while promoting fair judgement, through the inscription of sensitive attributes. The data-driven algorithmic technology does have a potential to promote fairness, yet for this to happen we should work on reducing potential negative consequences. When AI is built ethically, the judgement process can be demystified (something that can also be lacking in humans' judgement). Coeckelbergh notes that "explainability is not only a natural element of our communication, it is also a moral obligation", it "…is a necessary condition for responsible and accountable behavior and decisions" (2020, p. 123). Regardless of whether AI can *directly* offer such explanations and justifications, *humans* should be able to answer when asked to clarify the reasoning (Coeckelbergh, 2020). However, much work is to be done to achieve this level of transparency in both the technology and contemporary social structure. It also raises questions on the reliance degree – how much of the task do we actually want to outsource to algorithmic systems. It is a common misconception that since technology is based on data, it is believed to have a better rationale and does not require human oversight. Challenging this paradigm as well as promoting sustainable AI architecture and uses is a task for AI developers, researchers, and policymakers, but also for the public. The technology should be built in such a way that is understood by humans, while policy would ensure the standard and explain ways to achieve this practically.

In conclusion, bias in AI systems is a critical issue in the ethics of AI. It is both a societal and technical problem that requires careful consideration. Understanding the origin of bias, whether it stems from social prejudice, technical inaccuracies, or a combination of both, is crucial for developing strategies to rectify it. Addressing bias in prediction and forecasting algorithms, particularly in areas such as banking, criminal justice, and recruitment, to name a few, is necessary to ensure fairness and justice in AI systems.

Conclusion

The discussions throughout this text demonstrate that AI and data ethics extend far beyond purely technical considerations, influencing core societal domains such as politics, economics, and public welfare. Although initiatives like trustworthy AI, explainable AI, and responsible AI aim to establish ethical guidelines, implementing these standards in practice proves challenging. Subtle biases – embedded in datasets or algorithmic parameters – can perpetuate existing inequalities, raising questions about justice, fairness, and accountability. Moreover, the capacity of AI to perform large-scale data analysis brings risks to privacy and security, intensifying the need for oversight, transparency, and regular updates to models.

Across diverse sectors – from healthcare and banking to social media and public administration – these ethical considerations underscore the importance of approaching AI with a multi- or interdisciplinary perspective. Equally vital is the role of policymaking, where robust regulatory frameworks, coupled with wide-ranging ethical training, can help mitigate negative consequences. While AI holds promise for improving efficiency and enabling new forms of innovation, it must be guided by principles that protect human rights and social values. Only by integrating ethical guidelines and stakeholder engagement throughout the AI lifecycle can we foster systems that are transparent, equitable, and beneficial for all members of society.

References

Almutairi, Z., & Elgibreen, H. (2022). A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. Algorithms, *15*(5), 155. https://doi.org/10.3390/a15050155

Alpaydin, E. (2016). *Machine Learning: The New AI*. Cambridge Mass.: MIT Press.

Angwin, J. (2016). Sample: COMPAS Risk Assessment, COMPAS "CORE". DocumentCloud. https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE (accessed: 16.03.2023).

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed: 16.03.2023).

Baldridge, J. (2015, August 2). Machine learning and human bias: An uneasy pair. *TechCrunch*. https://techcrunch.com/2015/08/02/machine-learning-and-human-bias-an-uneasy-pair (accessed: 11.03.2025).

Boden, M.A. (2016). *AI: Its Nature and Future.* Oxford: Oxford University Press.

Bostrom, N. (2016). *Superintelligence.* Oxford: Oxford University Press.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research, 81,* 1–15.

Cataleta, M.S. (2020). Humane Artificial Intelligence: The Fragility of Human Rights Facing AI. *East-West Center.* http://www.jstor.org/stable/resrep25514

Chesney, R., & Citron, D. (2019). Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics. *Foreign Affairs, 98*(1), 147–155.

Chun, W.H.K. (2021). *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition*. Cambridge Mass.: The MIT Press.

Coeckelbergh, M. (2020). *AI Ethics.* Cambridge Mass: The MIT Press.

Collins, A., & Ebrahimi, T. (2021). Risk governance and the rise of deepfakes. *International Risk Governance Center,* 1–4. https://doi.org/10.5075/epfl-irgc-285637

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1), 1–5. https://doi.org/10.1126/sciadv.aao5580

D'Ignazio, C., & Klein, L.F. (2020). *Data Feminism.* Cambridge Mass.: The MIT Press.

European Commission AI HLEG (High-Level Expert Group on Artificial Intelligence) (2019). *Ethics Guidelines for Trustworthy AI.* Brussels, European Commission. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines (accessed: 12.04.2023).

Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities.* Oxford: Oxford University Press.

Fry, H. (2018). *Hello World: How to Be Human in the Age of the Machine* (1st ed.). London: Transworld Publishers.

Fuchs, C. (2013). Digital prosumption labour on social media in the context of the capitalist regime of time. *Time & Society, 23(1),* 97–123. https://doi.org/10.1177/0961463x13502117

Houser, K.A. (2019). Can AI Solve the Diversity Problem in the Tech Industry? Mitigating Noise and Bias in Employment Decision-Making. *The Stanford Technology Law Review, 290,* 290–354.

Howard, J. (2019). Artificial intelligence: Implications for the future of work. *American Journal of Industrial Medicine, 62(11),* 917–926. https://doi.org/10.1002/ajim.23037

IBM Data and AI Team (2023, October 12). *Types of Artificial Intelligence*. https://www.ibm.com/think/topics/artificial-intelligence-types (accessed: 8.06.2024).

IEEE (N/A.). *General Principles. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead1e_general_principles.pdf (accessed: 16.03.2024).

Iman, M., Arabnia, H.R., & Maribe Branchinst, R. (2022). Pathways to Artificial General Intelligence: A Brief Overview of Developments and Ethical Issues via Artificial Intelligence, Machine Learning, Deep Learning, and Data Science. In: H.R. Arabnia et al. (eds.), *Advances in Artificial Intelligence and Applied Cognitive Computing* (pp. 73–87). Berlin: Springer Nature.

ISO/IEC 22989:2022 (2022a). *Information Technology — Artificial Intelligence — Artificial Intelligence Concepts and Terminology.* https://www.iso.org/standard/74296.html (accessed: 26.04.2023).

ISO/IEC 23053:2022 (2022b). *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML).* https://www.iso.org/standard/74438.html (accessed: 26.04.2023).

Jagtiani, J., & Lemieux, C. (2019, January). The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the LendingClub Consumer Platform. *Working Paper (Federal Reserve Bank of Philadelphia).* https://doi.org/10.21799/frbp.wp.2018.15

Jemielniak, D., & Przegalińska, A.K. (2020). *Collaborative Society.* Cambridge Mass.: MIT Press.

Kahneman, D., Sibony, O., & Sunstein, C.R. (2021). *Noise: A Flaw in Human Judgment.* New York, NY: Littlle, Brown Spark.

Kelleher, J.D., & Tierney, B. (2018). Privacy and Ethics. In: eidem, *Data Science* (pp. 181–218). Cambridge Mass.: MIT Press.

Kilbertus, N., Gascon, A., Kusner, M., Veale, M., Gummadi, K., & Weller, A. (2018). Blind Justice: Fairness with Encrypted Sensitive Attributes. *Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research, 80,* 2630–2639. https://proceedings.mlr.press/v80/kilbertus18a.html (accessed: 20.05.2023).

Kronmal, R.A. (1993). Spurious Correlation and the Fallacy of the Ratio Standard Revisited. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 156*(3), 379–392. https://doi.org/10.2307/2983064

Lipińska, I. (2022). Etyka sztucznej inteligencji w dokumentach Unii Europejskiej w latach 2017–2020 [Ethics of Artificial Intelligence in European Union Documents in 2017–2020]. *Edukacja Filozoficzna, 73,* 11–38. https://doi.org/10.14394/edufil.2022.0001

Lopez, P. (2021). Bias Does Not Equal Bias: A Socio-Technical Typology of Bias in Data-Based Algorithmic Systems. *Internet Policy Review, 10*(4). https://doi.org/10.14763/2021.4.1598

Madon, S., Willard, J., Guyll, M., & Scherr, K.C. (2011). Self-Fulfilling Prophecies: Mechanisms, Power, and Links to Social Problems. *Social and Personality Psychology Compass, 5*(8), 578–590. https://doi.org/10.1111/j.1751-9004.2011.00375.x

McIntyre, L.C. (2018). *Post-truth.* Cambridge Mass.: MIT Press.

Mamak-Zdanecka, M., Stojkow, M., & Żuchowska-Skiba, D. (2019). Społeczny wymiar algorytmizacji [The Social Dimension of Algorithmization]. *Humanizacja Pracy, 3*(297), 9–20. https://www.humanizacja-pracy.pl/witryna/2019.09/Humanizacja%203%20 2019.pdf (accessed: 23.04.2025).

Mazurek, G. (2023). Sztuczna inteligencja, prawo i etyka [Artificial Intelligence, Law and Ethics]. *Krytyka Prawa, 15*(1), 7–10. https://doi.org/10.7206/kp.2080-1084.567

Mejias, U.A., & Couldry, N. (2019). Datafication. *Internet Policy Review, 8*(4). https://doi.org/10.14763/2019.4.1428

Microsoft (2018). *The Future Computed: Artificial Intelligence and Its Role in Society.* Redmond: Microsoft Corporation.

O'Neil, C. (2016). *Weapons of Math Destruction.* New York: Crown Publishing Group.

Rai, N. (2021). Why ethical audit matters in artificial intelligence? *AI and Ethics, 2*(1), 209–218. https://doi.org/10.1007/s43681-021-00100-0

Sadok, H., Sakka, F., & El Hadi El Maknouzi, M. (2022). Artificial Intelligence and Bank Credit Analysis: A Review. *Cogent Economics & Finance, 10*(1), 2023262. https://doi.org/10.1080/23322039.2021.2023262

Searle, J.R. (1983). Can Computers Think? In: idem, *Minds, Brains, and Science* (pp. 28–41). Cambridge Mass.: Harvard University Press.

Smith, B., & Browne, C.A. (2021). *Tools and Weapons: The Promise and the Peril of the Digital Age.* London: Penguin Books.

Tencent Keen Security Lab (2019). *Experimental security research of Tesla Autopilot [Technical report].* https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf (accessed: 11.03.2025).

Vigen, T. (2015). *Spurious Correlations.* New York: Hachette Books.

Whittaker, L., Letheren, K., & Mulcahy, R. (2021). The Rise of Deepfakes: A Conceptual Framework and Research Agenda for Marketing. *Australasian Marketing Journal*, *29*(3), 204–214. https://doi.org/10.1177/1839334921999479

Ku etycznej sztucznej inteligencji. Jak uwzględniać uprzedzenia, prywatność i odpowiedzialność w systemach opartych na danych

STRESZCZENIE Niniejszy artykuł analizuje implikacje etyczne związane ze sztuczną inteligencją i systemami przetwarzającymi dane, koncentrując się na tym, w jaki sposób kwestie związane z uprzedzeniami, prywatnością oraz bezpieczeństwem wpływają na wdrażanie technologii AI w różnych obszarach. Głównym problemem jest tu tendencja algorytmów AI do wzmacniania uprzedzeń społecznych i technicznych, co może prowadzić do ograniczenia sprawiedliwości oraz szkodzić grupom marginalizowanym. Warto jednak zauważyć, że decyzje podejmowane przez algorytmy i systemy AI mogą również przyczynić się do redukcji uprzedzeń, jeśli zapewniona zostanie transparentność i oparcie na faktach, pod warunkiem, że systemy te powstają w sposób etyczny. Aby zbadać te zagadnienia, zastosowano podejście koncepcyjne i jakościowe, łącząc wnioski z istniejących ram regulacyjnych, studiów przypadków oraz literatury naukowej dotyczącej etyki AI i analityki danych. Podstawowa hipoteza głosi, że integracja zasad etycznych, rzetelnego nadzoru oraz współpracy między różnymi dyscyplinami w procesie projektowania i rozwoju systemów AI pozwala na ograniczanie szkodliwych uprzedzeń, ochronę prywatności i wzmacnianie zaufania społecznego. Hipotezę tę weryfikuje się poprzez analizę ról różnych interesariuszy – twórców systemów AI, decydentów politycznych i użytkowników końcowych – w powstawaniu bądź redukowaniu stronniczości w procesach algorytmicznych. W konkluzji artykuł wskazuje na konieczność ustanowienia kompleksowych standardów regulacyjnych, większej przejrzystości działania modeli oraz ich regularnych aktualizacji, przy jednoczesnym zaangażowaniu zarówno ekspertów, jak i ogółu społeczeństwa. Tylko w ten sposób technologie AI będą mogły funkcjonować w sposób zgodny z wartościami demokratycznymi i humanistycznymi.

SŁOWA KLUCZOWE etyka AI, systemy oparte na danych, uprzedzenia algorytmiczne, prywatność, odpowiedzialność, etyczne zarządzanie

Date of submission of the article: 15.03.2025; date of acceptance of the article: 16.06.2025.