Digital Discrimination. The Challenge of Bias and Transparency in AI[a]

Gabriele Giacomini[*]*, Chiara Aprilis*[†]

Abstract

Il saggio esplora l'impatto crescente dell'intelligenza artificiale sui sistemi decisionali e le sue implicazioni etiche, concentrandosi sui pregiudizi algoritmici che possono portare a discriminazioni basate su genere, etnia e altri fattori. Attraverso esempi concreti, si discute di come i pregiudizi possano manifestarsi e si sottolinea l'importanza di un approccio responsabile alla governance dell'IA. Ciò implica la promozione di una riflessione sia accademica che pubblica sull'adozione di principi etici e procedure il più possibile trasparenti e inclusive.

Parole chiave: Algoritmi, Intelligenza Artificiale, Bias, Discriminazione, AI Governance.

This paper explores the growing impact of artificial intelligence on decision-making systems and its ethical implications, focusing on algorithmic biases that can lead to discrimination based on gender, ethnicity, and other factors. Through concrete examples, it discusses how biases may manifest and emphasises the importance of a responsible approach to AI governance. This involves promoting both academic and public reflection on the adoption of ethical principles and procedures that are as transparent and inclusive as possible.

Keywords: Algorithms, Artificial Intelligence, Bias, Discrimination, AI Governance.

[a] The conception of the article is by both authors. However, Aprilis focused on section 2, Giacomini on sections 1 and 3. Saggio ricevuto in data 11/03/2024 e pubblicato in data 22/01/2025.

[*] Ricercatore, Università degli Studi di Udine, email: gabriele.giacomini@uniud.it.

[†] Dottoressa in Scienze filosofiche, email: chiara.aprilis@gmail.com.

1. Introduction

The increasing integration of artificial intelligence (AI) into decision-making systems, both corporate and public, raises crucial questions about algorithmic biases, with direct implications for fairness and equality in our societies. As a matter of fact, the consequences of discrimination associated with the use of AI can be considerable, leading to alter people's job opportunities, the information environment, access to social services, even, in the most extreme cases, to establish the life and death of people.

Defining "algorithmic bias" is not a straightforward matter, as evidenced by the still ongoing efforts within this respect.[1] The nature of algorithmic bias, that appear to be embedded in machine learning systems, is mainly classifiable as either statistical or societal[2]. The first concerns technical issues related to the quality or quantity of training data; they can usually be detected in the steps of Machine Learning Pipeline. The second concerns social or cultural biases reflected in the data, which are more difficult to correct as they require in-depth ethical, social and political analysis. In both cases, the risk is that of penalisation of marginalized social categories on the basis of gender, religion, race or ethnicity, sexual orientation. These social groups are underrepresented in the high-tech sector. In other words:

> Bias occurs when certain group, such as racial or ethnic minority, rural and socioeconomically disadvantaged populations, are missing from the data. The potential for bias is also perpetuated by the lack of women and racial ethnic minorities in the technology fields because their ideas, perceptions, and values may not be represented the development, training and deployment of algorithmic tools[3]

The unfair outcome can take an allocative form – the bias result both from the withholding of some opportunity or resource, and the unfair distribution of goods across groups – or a representational one –that is, the systematic representation of some group in a negative light, or in a lack of positive representation.[4] Several authors have given other kinds of distinctions among type of bias[5].

This contribution aims to explore the manifestations of bias in AI, highlighting how they can influence decisions ranging from personnel selection to public administration, from moderation of online content to the use of automatic weapons. Given society's growing dependence on AI and the difficulties of decoupling from it without facing serious socioeconomic consequences, the importance and urgency of

---

[1] R.S. Baker, A. Hawn, *Algorithmic Bias in Education*, in «International Journal of Artificial Intelligence in Education» 32,4, 2022, pp. 1052-1092.

[2] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, K. Lum *Algorithmic Fairness: Choices, Assumptions, and Definitions*, in «Annual Review of Statistics and Its Application», 8, 2021, pp. 141-163: https://doi.org/10.1146/annurev-statistics-042720-125902.

[3] N.H. Williams, *AI and Healthcare: The Impact of Algorithmic Bias on Health Disparities*, Springer, Cham 2023.

[4] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, K. Lum, *Algorithmic Fairness*, cit.

[5] R.S. Baker, A. Hawn, *Algorithmic Bias in Education*, cit.

a responsible approach to AI governance is highlighted, both by companies and public institutions, including solid ethical principles, transparency, fairness. Possible strategies to mitigate bias and promote an informed use of AI are then discussed: addressing the problem of bias requires a structured approach that includes technological, ethical and regulatory solutions. It also requires focusing on the importance of greater algorithmic transparency, inclusiveness in the development of AI design, and on the education of decision makers and the population.

2. A first classification of algorithmic bias

In the digital age, the adoption of AI systems is becoming an increasingly common practice. While these tools promise increased efficiency, they also raise significant concerns about the risk of discrimination and bias. Events such as the use of algorithms for personnel selection that result in gender discrimination, or the adoption of algorithmic systems within judicial and welfare contexts that unfairly disadvantage specific social groups, highlight the risks associated with an uncritical use of artificial intelligence. The following examples explore cases of algorithmic bias in various areas of human and social experience.

The use of artificial intelligence systems in personnel selection mechanisms has been spreading for about a decade. The aim is to improve the recruiting process in qualitative terms, first of all by making it more efficient in terms of time, for example through letting the algorithm selecting the shortlist of profiles most in line with the company, relieving the recruiters from many tedious tasks such as searching for and examining dozens – if not hundreds – of CVs and sending standard responses to candidates, or placing adverts to attract candidates. This is done through the use of a wide range of tools, including, for example, chat-bots that interface with candidates, providing essential guidance and sending routine messages to those who are not in line with the company open positions. Artificial intelligence mechanisms can help create a complete profile of the candidate through the analysis of verbal and written language, often compared with that of current employees, or profile their character and emotional aspects through social media profiles[6]. Furthermore, the idea driving the process of technicalisation of selection is to get rid of human judgement as the sole criterion, in favour of a presumed greater objectivity of the technological tool.

However, the use of artificial intelligence in personnel selection has raised numerous questions. The series of problems begins with the sensitivity of the data processed by AI, often of a private and personal nature, which are processed for purposes that go beyond the interests of their owner. There is an asymmetry problem in the exchange of personal information to the extent that the candidate is somehow "forced" to provide his or her data in order to have a chance of obtaining the position.

[6] B. Dattner, T. Chamorro-Premuzic, L. Schettler, *The Legal and Ethical Implications of Using AI in Hiring*, in «Harvard Business Review», 25 April 2019: https://hbr.org/2019/04/the-legal-and-ethical-implications-of-using-ai-in-hiring.

Moreover, some personal information may be statistically deduced by algorithms without the candidate workers' knowledge[7] .

The issue of algorithmic bias is relevant here, as it potentially undermines people's careers and thus their economic prospects. AI mechanisms are not 'neutral' agents, rather the contrary. The case of the algorithm implemented by Amazon that caused a sensation in 2015 shows how the patterns identified by these systems lead to selecting candidates who are similar to the most successful employees in their company, in other words, *cloning your best people* is the outcome most likely[8]. Adopting AI mechanisms is not always rewarding for the company, as demonstrated by the case of Amazon, where the system selected candidates not so much because of technical-IT skills but on the basis of the person's gender, ending up discriminating against women. This outcome resulted from the fact that the algorithm had been trained with data from the profiles of former candidates who were now career employees, for the most part members of a particular category of people that is predominant in the hi-tech sector: males and whites. The system therefore penalised CVs that contained the word 'female' and those related to it, while it positively evaluated those containing verbs and words related to the male gender. The company then decided to abandon the programme[9].

Amazon's is probably one of the most emblematic cases, but there are numerous that have led to discrimination based on disability, ethnicity or even neighbourhood, as well as on the basis of information present on social networks[10].

The last two decades have seen an increasingly massive adoption by governments around the world of automated systems in administrative and criminal bureaucracy. While some benefits have been undoubted, the application of these systems to sensitive areas such as the health, judicial or social services systems has led to instances of injustice.

An investigation conducted by *Wired* and *Lighthouse* has shed light on the mechanisms of discrimination that have affected some entitled to municipal benefits in the Dutch city of Rotterdam. The municipal administration had devised an algorithm to classify potential defrauders of the public subsidy allocation system. Over the course of its use – between 2017 and 2021 – the machine learning algorithm generated risk indices for each of the 30,000 subsidy claimants and, city officials

[7] I. Ajunwa, R. Schlund, *Algorithms and the Social Organisation of Work*, in M. D. Dubber, F. Pasquale, S. Das, *The Oxford Handbook of Ethics of AI*, Oxford University Press, Oxford 2020, pp. 804-822.

[8] I. Ajunwa, R. Schlund, *Algorithms and the Social Organisation of Work*, cit., p. 808.

[9] J. Dastin, *Insight: Amazon scraps secret AI recruitment tool that showed bias against women*, in «Reuters», 11 October 2018: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/.

[10] For instance, if the algorithm detected a postcode linked to a run-down or bad neighbourhood, it would also negatively label the candidate. See R. Krish, *The Pros and Cons of AI in Recrutiment*, in «The Research Nest», 5 December 2018: https://medium.com/the-research-nest/the-pros-and-cons-of-ai-in-recruitment-19c141d1c4b7.

investigated the individuals from these results[11]. During its use, hundreds of people, generally belonging to categories on the margins of Dutch society, were falsely accused or suspected of fraud: the algorithm systematically attributed women and minorities a greater possibility of cheating the subsidiary system.

The article created by *Wired* presents the case of a woman of Moroccan origins, divorced mother of three children, recipient of social allowances. The woman in question was allegedly investigated for the first time after resigning from her job for health reasons; following the investigation, she was deprived of her benefit for the first time, being forced to ask for loans and food from neighbours, as well as pushing her sixteen-year-old son, still a student, to find a job in order to support himself. Moreover, after two years, she was summoned by the social services department and subjected to an "interrogation" where, since she had submitted the wrong bank statement, she was once again deprived of benefits for a few weeks. This lady had been reported as a "high risk" due to her status as a woman, single mother and of foreign origins.

Instead, as far as criminal justice is concerned, a famous investigation published in 2016 by the investigative agency *ProPublica* investigative agency revealed that the system for determining the risk of recidivism used in the United States was decidedly discriminatory against African Americans. The software, named COMPAS, systematically attributed twice the risk of recidivism to African Americans compared to whites, although it was later contradicted by the facts[12]. These cases reveal that, at the root of the accusations perpetrated against the weakest sections of the citizenry, there is a cultural problem that technology does not create but rather increases.

The impact of automatic decision making on marginalized groups is extensively treated by Virginia Eubanks in *Automatic Inequality: How High-Tech Profile, Police and Punish the Poor*, who argues that technologies reflect American culture that stigmatize poor people, regarded as the cause of their own misery, and hence criminalized and even dehumanized. Indeed, as shown by the above-discussed Rotterdam case, the approach towards social benefit recipients is often paternalistic, as people are put under investigation, often undergoing invasive practices for their privacy, divided among those who are morally deserving and those who are not; the latter being punished. Eubanks brings three cases-study that support her thesis; they display cases of erroneous benefit withdrawal, use of predictive model to target which children might be in danger of abuse or neglect while over-investigating lower-classes and black people, finally cases in which algorithms decide who get houses and who remain homeless. The way in which these systems – named "digital poorhouse" in the book – are being used threatens our democracy, warns the author, as they

[11] M. Burgess, E. Schot, G Geiger, *This Algorithm Could Ruin Your Life*, in «WIRED», 6 March 2023: https://www.wired.com/story/welfare-algorithms-discrimination/.

[12] J. Angwin, J. Larson, S. Mattu, L., Kirchner, *Machine Bias, There's software used across the country to predict future criminals. And it's biased against blacks*, in «ProPublica», May 2016: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

undermine the principles of liberty, equity of treatment and value as well as inclusion[13].

The problem of racial discrimination is all the more relevant as the current situation sees the exacerbation of intercultural conflicts and a worrying growth of racial hatred on the part of certain sections of the population. A further problem is that most systems are obscure and hidden from the eyes of citizens[14], who can hardly monitor the situation and possibly defend themselves.

The growth in the flow of information and content on online platforms in recent decades has raised the question of protecting users from offensive or harmful content. In the field of content moderation, automated AI systems are becoming collaborators of human moderators, increasing the capillarity of intervention. Furthermore, they help protect human moderators, who may suffer significant psychological and emotional damage as a result of their activity[15].

However, even in this field, algorithms can make discriminations. First of all, AI mechanisms to date are unable to grasp well the nuances of meaning and idioms of human language, nor to operate a cultural contextualisation of the content, and thus to distinguish what may be offensive in a culture and not in another. Second, they are subject to reproducing the biases of those who program them (usually Western males).

For example, as reported in an article in the *New York Times*, in 2017 the popular YouTube platform removed thousands of videos posted by Syrian activists documenting the atrocities of the Syrian war, resulting in the loss of important testimonies. The platform had recently activated a system that automatically deleted content that did not comply with its guidelines. The system was designed to identify videos posted by extremist (specifically Islamic) groups, but ended up including, quite indiscriminately, any content coming from Syria and the conflict zones where terrorists operated[16]. Similarly, in 2020, *Instagram* algorithm censored posts and profiles related to the activity of the *Black Lives Matter* movement, citing the protection of the community as the motivation. The platform immediately recognised the mistake: the exponential growth of related content had activated the mechanism to prevent spam: so that it apologised and asserted its support for the cause[17] .

There are many examples that can be given in this regard, pointing out that while social platforms give everyone the opportunity to express themselves, the

---

13 V. Eubanks, *Automatic Inequality: How High-Tech Profile, Police and Punish the Poor*, St. Martin's Press, New York 2018.

14 M. Burgess, E. Schot, G. Geiger, *This Algorithm Could Ruin Your Life*, cit.

15 *Use of AI in Online Content Moderation*, in «Cambridge Consultants», Ofcom 2019.

16 M. Browne, *YouTube Removes Videos Showing Atrocities in Syria*, in «New York Times», August 2017: https://www.nytimes.com/2017/08/22/world/middleeast/syria-youtube-videos-isis.html.

17 A. Griffin, *Instagram users trying to post about Black Lives Matter see 'action blocked' messages*, in «Independent», June 2020: https://www.independent.co.uk/tech/instagram-action-blocked-fix-get-rid-how-message-black-lives-matter-error-spam-a9543716.html.

potential for empowerment is not equally distributed or uniformly accessible[18]. Platforms have made a lot of progress and continue to improve their algorithms[19], but the issue remains sensitive because platforms have primarily economic purposes, which means that they do not necessarily have incentives to act in accordance with ideals of equality, justice, protection of minorities and collective well-being.

Speaking about empowerment, an important issue is education. Algorithms have been applied also to this field, to estimate dropout predictions, automated essay scoring, graduate admission, knowledge inference. However, these statistics have shown various degrees of inaccuracy because education is often treated as a homogenous phenomenon, while diversities in class composition should be considered. In particular, there has been insufficient research into intersectionality[20] in educational work on algorithmic bias[21]. This problem should be addressed, as education is one of the most important aspects for empowering one's life. Racism and sexism are embedded in the architecture of technology, according to Noble. Much of her work has been devoted to how technology and the information infrastructure perpetuate specific narratives and make profit from it. She began her research after having noted that the search result for "black girls" lead to hypersexualised and pornographic content[22], to then explore how people of colour are negatively or not-positively represented in the media[23], the underemployment of Black and Latinos by high-tech firms, the commercialization of identities that renders search engine a place of disinformation while deemed reliable, and how this has a role in shaping culture and society. Also, this structure challenges the very idea that marginalised people have of themselves, with a bad impact on their self-esteem, capacity for auto-determination as well as of their life's opportunities, to come back to the concept of self-empowerment. Noble argues that these "algorithmic discriminations" are not glitches in the system, as high-tech representatives claim, in fact they derived from the biases and prejudices carried by programmers – male and white – and exploited by the companies to make profits, hence being the way the system operate.

---

[18] D. Endres, L. Hedler, K. Wodajo, *Bias in Social Media Content Management: What Do Human Rights Have to Do with It?* in «Cambridge University Press», June 2023.

[19] *Use of AI in Online Content Moderation*, cit.

[20] The term, coined by K. Crenshaw, *Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics*, «University of Chicago Legal Forum», vol. 1989, Article 8, describes the condition felt by those who belong to two or more marginalized categories, such as being female, black and lesbian. They therefore experience a discrimination that is more than the sum of the single discrimination.

[21] R.S. Baker, A. Hawn, *Algorithmic Bias in Education*, cit.

[22] S.U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York University Press, New York 2018, pp. 17-22.

[23] Noble cites the comparison of African American to apes, the privileged depiction of Whites when the key words linked to "white" are searched as opposed to "black," "jew", "Asian", the association of black names to crimes, and so on.

The problem of bias codification manifests itself in all its radicality in the case of automatic weapons, where a bias can determine the life or death of a multitude of people. Automatic weapons use AI mechanisms to identify targets to hit, they are triggered by humans who, however, are in most cases unaware of what the target is, how it was identified and even when and where it will be shot down. This makes it particularly complex to control automatic weapons, also because the strategy developed by the system is often a black box for those who would try to access it[24]. Furthermore, the system continues to learn while in use, risking further escaping the goals and understanding of those who designed it[25]and amplifying the *biases* already contained within it.

In order to target the Gaza Strip in 2024, the IDF (*Israel Defence Forces*) has developed a system called "Habsora", translated into English with the term "Gospel", which identifies one hundred targets per day, half of which are hit[26]. Despite the Israeli government's declarations about the accuracy of this tool, "Gospel" does not only target Hamas militants: it identifies all individuals accused of potential collaboration, and the areas in which they are presumed to reside, accurately calculating also the number of civilian victims totally unrelated to the terrorist organisation who would be sacrificed. The result is the intensification of the massacre of the civilian population in Gaza, with hundreds of people dying every day but who have little or nothing to do with Hamas[27].

The issue of automatic weapons is problematic in many respects. For instance, if the development of artificial intelligence shows a legacy of the racist colonial system, the issue that some groups against the development of automatic weapons draw attention to is that they could represent a serious danger to ethnic or cultural minorities[28]. In other words, warfare carried out in this way would be the culmination of colonialism: programs are written by members of the dominant group, incorporating their cultural prejudices which, thanks to autonomous learning, risk being amplified by the AI escaping the control of the programmers, making certain sections of the population much more vulnerable.

Algorithmic bias has also become one of the determinants of health in so far as automated decision systems are so intertwined with people lives, influencing the social determinants of health (namely, healthcare and education access and quality,

---

[24] On the black box concept: F. Pasquale, *The black box society: The secret algorithms that control money and information*, Harvard University Press, Cambridge 2015.

[25] *What you need to know about autonomous weapons,* in «International Committee of the Red Cross», July 2022: https://www.icrc.org/en/document/what-you-need-know-about-autonomous-weapons.

[26] *The Gospel*: *how Israel uses AI to select bombing targets in Gaza*, in «The Guardian», December 2023: https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets.

[27] *"A mass assasination factory": Inside Israel's calculated bombing of Gaza*, in «+972 Magazine», November 2023: https://www.972mag.com/mass-assassination-factory-israel-calculated-bombing-gaza/.

[28] H. R. Jones, *Intersectionality and Racism,* in «Stop Killer Robots»: https://www.stopkillerrobots.org/wp-content/uploads/2021/09/Intersectionality-and-Racism-Hayley-Ramsay-Jones.pdf.

social and community context, economic stability, neighbourhood and built environment)[29]. In fact, although the integration of AI systems in medicine unquestionably better the quality of patient care, not everyone benefits from it: as in the case of automatic weapons, ethnical minorities and black people in particular – but also women – are discriminated because of data under-representation, implicit and explicit bias. Within this respect the research conducted by Obermeyer and alt. found that an algorithm used to foresee the enrolment in health care management programs discriminates against Black people. Even though Black patients had poorer health conditions, the Whites were predicted to need additional care services. The error was not taking into account that Black spent less on health care due to limited resources, hence using less health care management programs than Whites[30]. The evidences that life and well-being of non-White people may be under-attack by automated decision-making systems, make clear that their rights should be carefully taken into account in the development of an ethical AI.

3. The commitment to lead AI out of discriminatory bias

The topics covered highlight the challenges posed by the use of artificial intelligence in various fields, emphasizing not only the potential benefits but also the risks associated with its implementation. From personnel selection and management, through administrative and bureaucratic systems, to online content moderation and the use of automatic weapons, algorithmic biases can lead to discrimination based on gender, ethnicity, social status and other personal characteristics.

Faced with these scenarios, what reassures us is the idea that we can "pull the plug" on AI. However, this may be a misplaced hope. Not only would it be difficult to give up the extraordinary potential of AI: the real problem might be that, as time passes, human society will probably be increasingly dependent on AI, and going backward could have very high costs to face. The costs of the consequences in terms of widespread poverty, crisis of services, and fragility of the economic and social system would risk being greater than the problems for which it is thought to detach AI. This means that both politics and the organisations that produce AI systems now have a responsibility to govern AI and its effects, even if these may prove harmful.

In particular, if companies that develop technology are primarily "for-profit companies", and thus not necessarily incentivised to pursue the common good[31], it is crucial that public regulation and regulatory interventions come into play to ensure

---

[29] N.H. Williams, *Artificial Intelligence and Healthcare: The Impact of Algorithmic Bias on Health Disparities*, Springer, Cham 2023.

[30] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, *Dissecting racial bias in an algorithm used to manage the health of populations*, in «Science» 366, 2019, pp. 447-453.

[31] J. Harris, *There was all sorts of toxic behaviour: Timnit Gebru on her sacking by Google, AI's danger and big tech's biases*, in «The Guardian», May 2023: https://www.theguardian.com/lifeandstyle/2023/may/22/there-was-all-sorts-of-toxic-behaviour-timnit-gebru-on-her-sacking-by-google-ais-dangers-and-big-techs-biases.

that the evolution of AI is aligned with the broader interests of society. This includes the creation of laws and regulations that ensure transparency, fairness, accountability and safety in the use of AI, as well as effective oversight and control mechanisms. The EU seems to be moving in this direction with the Artificial Intelligence Regulation, which aims to establish rules for the development, marketing and use of artificial intelligence in the European Union, seeking to protect the safety and fundamental rights of individuals, classifying AI systems according to the risk they present, and imposing stricter requirements for those considered high-risk[32].

At the international level, however, many experts are working on the identification of principles that should inspire AI and its regulation. Recent working groups include, by way of example, the *Asilomar AI Principles* (2017), the *Montreal Declaration on Responsible AI* (2017), the *Declaration on Artificial Intelligence, Robotics and Autonomous Systems* (2018), the *Five General Principles for an Artificial Intelligence Code* (2018), and the *Ethical Guidelines for Trustworthy AI* (2019). These initiatives highlight the need for joint thinking between researchers, legislators, industries and civil society to ensure that AI development is aligned with fundamental human values. Luciano Floridi, perhaps the world's foremost expert on information and AI ethics, has attempted with his collaborators to summarise the recurring and cross-cutting principles in the documents produced in recent years[33]. They are:

1. Beneficence (AI should promote welfare, preserve dignity and sustain the planet).
2. Non-maleficence (the AI must respect privacy, be secure and avoid misuse).
3. Autonomy (AI must promote the autonomy of humans, who can always choose how and whether to delegate decisions to the machine).
4. Justice (AI must support the prosperity of peoples, preserve solidarity and avoid unfairness).
5. Explicability (the AI should be as transparent and intelligible as possible).

Implementing these principles is not an easy challenge or one with predictable outcomes. In the first place, it is still unclear how to politically promote these general principles: is self-regulation and consumer protection sufficient? Or should the state intervene more directly? Secondly, the principle of explicability is correct from a normative point of view, but it is still unclear to what extent it can be applicable. Floridi emphasises that the obscurity of AI must not provide an "excuse" for digital companies to hide their internal procedures from researchers, supervisory bodies and democratic institutions in general. Floridi is absolutely right, but he seems to forget

[32] D.E. Harris, *Europe has made a great leap forward in regulating AI*, in «The Guardian», December 2023: https://www.theguardian.com/commentisfree/2023/dec/13/europe-regulating-ai-artificial-intelligence-threat.

[33] L. Floridi, *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*, Oxford University Press, Oxford 2023; B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, L. Floridi, *The ethics of algorithms: Mapping the debate*, in «Big Data & Society», 3(2), 2053951716679679.

the fact that AI is a *black box* even for the researchers who built it[34].

Since AI is based on complex statistical relationships that are impossible for a human being to trace, it might be difficult to explain the reasons behind a machine's decision. Yet, the last principle is the most basic: only if AI is explicable can we fully check that it is beneficial, not evil, that it respects the autonomy of humans, and that it promotes social justice. This is why ways of improving the verification of AI systems are being experimented with, such as the use of "stress tests" or the creation of internal "checkpoints" where certain checks can be carried out[35]. Technically, this new field of study is called *explainable AI (XAI),* and it aims to improve the trust and transparency of AI-based systems so that AI continues to make steady progress without interruption. The degree of certainty with which we will be able to ensure that AI does not promote judgements or decisions tainted by discriminatory bias will perhaps depend on these projects.

In addition, as regard bias, there is also the problem of dealing with its harmful effects, in cases where, despite preventive controls, these should nevertheless occur. Who should be held responsible for an action of an AI system that causes harm to individuals or groups of people due to bias? We have seen that the consequences could also be very important in terms of well-being, health, even life and death. Theoretically, it becomes important for there to be "meaningful human control" (the concept "human in the loop" is also used, which envisages the presence of human will and judgement in algorithmic processes[36]). By "meaningful" is meant that purely nominal conditions and forms of human control over the machine are excluded. The control must instead be substantial, i.e. the human agent should be able to express a considered judgement about the operations the system is performing and be able to intervene in good time in the event of unforeseen events. Furthermore, the operator should be trained not to overestimate the capabilities of computer systems.

However, in practice, this is a complex task, perhaps impossible to fully realise, due to the internal opacity concerning complex computational processes and the difficulty of interpreting the information that induces the artificial system to make a certain decision or perform an action. The operator who activated the system might not be able to exercise effective control over its behaviour, due to the cognitive difficulty of understanding the AI's decision-making mechanisms (AI is a *black box*) and the long reaction times of humans that might prove ineffective for timely

[34] N. Cristianini, *The Shortcut: Why Intelligent Machines Do Not Think Like Us*, in «CRC Press», 2023.

[35] On the topic of XAI: A. Adadi, M. Berrada, *Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)*, in «IEEE access», 6, 2018, pp. 52138-52160; F. K., Došilović, M. Brčić, N. Hlupić, *Explainable artificial intelligence: A survey*, in «41st International convention on information and communication technology, electronics and microelectronics», 2018, pp. 0210-0215; G. Vilone, L. Longo, *Notions of explainability and evaluation approaches for explainable artificial intelligence*, in «Information Fusion», 2021, 76, 2021, pp. 89-106.

[36] F. M. Zanzotto, *Human-in-the-loop artificial intelligence*, in «Journal of Artificial Intelligence Research», 2019, pp. 64, 243-252; E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, Á. Fernández-Leal, *Human-in-the-loop machine learning: A state of the art*, in «Artificial Intelligence Review», 56(4), 2023, pp. 3005-3054.

intervention. Ultimately, who may be primarily responsible for bias damage? Potential candidates include: the software engineers who created the AI system, the executives of the company that developed it, the consultants or staff members of the entity that implemented the system, the leaders of the organisation that adopted the system, the head of the department that used the system, or the individual who responsible for directly supervising the system. The risk is to conclude that no one made a truly significant contribution to the occurrence of the damage[37]. Or there is the risk of attributing this diffuse responsibility by law to a single figure, perhaps paid precisely to legally take on any negative responsibility.

These issues are under discussion, but several strategies are beginning to emerge that will need to be explored and tested in the near future. Some strategies are mainly technological. As we have seen, it will be crucial to understand whether and how it is possible to increase the transparency of AI-based algorithms and systems, allowing for independent review and analysis to identify bias. This is obviously related to the importance of investing in research and development of methodologies capable of mitigating and removing existing biases in AI systems. Other strategies, however, concern human beings more directly. We are not only referring to the identification of solid ethical principles and public regulation that place respect for human rights and the protection of minorities and vulnerable groups at the centre. It is also about ensuring that developers of AI systems come from diverse backgrounds to ensure that a variety of perspectives are considered in the design of algorithms, thus reducing the risk of unintended bias. Most importantly, raising awareness and education about the potential pitfalls and biases of AI algorithms will be essential for those who develop them, those who use them and those who are affected by them.

[37] H. Nissenbaum, *Accountability in a Computerised Society*, in «Science and Engineering Ethics», 2, 1, 1996, pp. 25-42.