Algorithmic processing and AI bias; using overfitting to reveal rather than perpetuate existing bias[a]

Lydia Farina[*] and Anna-Maria Piskopani[†]

Abstract

In questo articolo analizziamo l'*overfitting* dell'IA nell'elaborazione algoritmica per mostrare come esso sia correlato a casi di iniquità o distorsione dell'IA e come si combini con fenomeni sociali complessi, quali gli effetti di looping, per mantenere ed esacerbare le distorsioni esistenti. Discutiamo le normative esistenti e proposte in materia di IA che tentano di affrontare questo pregiudizio per cogliere le tendenze e le priorità dominanti. Infine, suggeriamo che, sebbene l'attenzione della letteratura attualmente si concentri sulle conseguenze negative dell'*overfitting*, esso può essere utilizzato come strumento diagnostico per individuare le disuguaglianze sociali sottostanti e, in quanto tale, portare a usi alternativi dell'analisi dell'IA per smascherare l'ingiustizia sociale piuttosto che esacerbarla. Questo articolo fornisce un ulteriore supporto teorico alle recenti opinioni presenti nella letteratura che suggeriscono che l'elaborazione algoritmica può essere utilizzata per diagnosticare e monitorare i pregiudizi; evidenziando l'interazione con gli effetti di looping, fornisce anche un'ulteriore motivazione per utilizzare l'*overfitting* come primo passo verso la mitigazione dei pregiudizi storici.

Keywords: overfitting, regolamentazione IA, valutazione diritti umani, equità algoritmica, discriminazione algoritmica.

In this paper we analyse AI overfitting in algorithmic processing to show how it relates to cases of unfairness or AI bias and how it combines with complex social phenomena such as looping effects to maintain and exacerbate existing bias. We discuss existing and proposed AI regulation attempting to address this bias to pick up dominant trends and priorities. Finally, we suggest that, although the focus of the literature currently falls on the negative consequences of overfitting, it can be used as a diagnostic tool for detecting underlying social inequalities and, as such, lead to alternative uses of AI analytics to expose social injustice rather than exacerbate it. This paper provides further theoretical support to recent views in the literature suggesting that algorithmic processing can be used to diagnose and monitor bias; by highlighting the interaction with looping effects, it also provides additional motivation to use overfitting as a first step towards mitigation of historical prejudice.

[a] Received on 22/05/2025 and published on 09/12/2025.
[*] University of Nottingham, e-mail: lydia.farina@nottingham.ac.uk.
[†] University of Nottingham, e-mail: anna-maria.piskopani@nottingham.ac.uk.

Keywords: overfitting, AI regulation, human rights assessments, algorithmic fairness, algorithmic discrimination.

Introduction

AI overfitting is described in at least two different ways: 1) as an observed tendency in AI based algorithms to take shortcuts to achieve a given task by ignoring some data which do not agree with common trends[1] or 2) as a tendency to inaccurately fit patterns based on limited data to more generalised cases, hence 'over fitting'[2]. To give a more refined definition, overfitting happens when algorithmic models make predictions based on regularities discovered in the training data which do not match the world from which the data is taken[3].

Overfitting can exacerbate bias in certain domains where algorithmic processing is used to generate outcome predictions after training such as in the practices of predictive policing or predictive healthcare. In predictive policing algorithmic processing generates predictions such as a percentage of crime rate detection in a particular area[4]. The algorithmic prediction is used as a justification for sending additional officers to that area. In turn, this results in more crimes being detected in that area as the larger number of officers correlates with more crimes being detected. The updated detection crime rate of the area is then fed back as data to the algorithm, to be used for subsequent predictions. In certain cases of predictive healthcare, the algorithm predicts healthcare needs by calculating incurred healthcare costs[5]. Areas where healthcare costs are higher, are then allocated more funding on the assumption that costs incurred track healthcare needs of the population living in these areas. Using these specific contexts as case examples we show that allowing algorithmic predictions to dictate policy in these contexts not only maintains bias but exacerbates it.

Even though the risks and harms associated with algorithmic processing in specific contexts e.g. in predictive policing have been identified for several years, such practices are still being used[6]. In more recent years, human rights organisations such as Conseil of Europe, the European Agency for Fundamental Rights, international organisations such as United Nations published reports and resolutions identifying ways in which AI based

[1] J. Shane, *You Look Like A Thing and I Love You*, Headline Publishing Group, Wilfire 2019.
[2] See K.P. Burnham, D.R. Anderson, *Model selection and multimodel inference*. 2nd ed., Springer-Verlag 2002; J.S. Russell, P. Norvig, *Artificial Intelligence; A Modern Approach*, 3rd ed.., Pearson Education Limited 2016.
[3] D.L Poole., A.K. Mackworth, *Artificial Intelligence; Foundations of Computational Agents*, 3rd ed., Cambridge University Press, Cambridge UK 2023, p. 297.
[4] See K. Hao, *Police Across the US are training Crime Predicting Ais on Falsified Data*, «MIT Technology Review», 13 February 2019, Police across the US are training crime-predicting AIs on falsified data | MIT Technology Review; A. Larson, J. Angwin, Bias in *Criminal Risk Scores is Mathematically Inevitable, Researchers Say*, «ProPublica», 30 December 2016, https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say.
[5] See S. Gupta, *Bias in a Common Health Care Algorithm Disproportionately Hurts Black Patients*, «SCI. NEWS», Oct. 24 2019. https://www.sciencenews.org/article/bias-common-health-care-algorithm- hurts-black-patients.
[6] See M. Degeling, B. Berendt, *What is wrong about Robocops as consultants? A technology-centric critique of predictive policing*, «AI & Society», n. 33, 2018, pp. 347–356, https://doi.org/10.1007/s00146-017-0730-7.

decisions violate human rights (identified as non-discrimination and data protection rights). The UK published an AI regulation White Paper[7] where it describes its approach to AI challenges as well as a roadmap to effective AI assurance ecosystems. The UK ICO (Information Commissioner's Office) published guidance for public services aiming to evaluate whether such violations or risks to equality rights take place along with a roadmap to effective AI assurance ecosystems[8].

In what follows we first show how overfitting relates to cases of unfairness and to phenomena discussed in social ontology such as 'looping effects' to maintain and exacerbate existing bias especially in cases where discrimination is based on characteristics which are not protected by law[9]. We then discuss existing and proposed AI regulation attempting to address bias arising from algorithmic processing to pick up dominant trends and priorities. Finally, we suggest that, although the focus of the literature currently falls on the negative consequences of overfitting, it can be used as a tool to reveal underlying social inequalities and social injustice when AI is used for social welfare, policing etc and, these findings can be repurposed and further explored by other scientists such as cognitive scientists to understand in depth implicit biases and stereotypes and lead to further interdisciplinary research[10]. This paper provides further theoretical support to recent views in the literature suggesting that algorithmic processing can be used to reveal, diagnose and even monitor bias as a first step towards mitigation of historical prejudice[11].

1. *Overfitting, feedback loops and looping effects*

Algorithms are mathematical constructs tasked with picking up patterns available from their training data. In the cases discussed in this paper, AI bias is a direct consequence of the tendency shown by algorithms to ignore some of the data (relating to minority characteristics) and privilege others, that is the ones associated with the dominant pattern, to reveal the dominant pattern found in the data (overfitting via shortcut). So, the first type of AI overfitting relates to excluding data which do not reflect dominant trends[12]. For example, if we task an algorithm with selecting amongst CVs to fill in a job vacancy, the

[7] Department for Science, Innovation and Technology, *A pro-innovation approach to AI regulation*, 2023, CP 815, retrieved 10 April 2024 from: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.

[8] Information Commissioner's Office, *Guidance in AI and Data Protection*, updated 2023. Retrieved on 10 April 2024, available at: https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-artificial-intelligence-and-data-protection/.

[9] For a discussion of looping effects, see I. Hacking, The looping effects of human kinds, in D. Sperber and A. J. Premack. (eds.), *Causal cognition; A Multidisciplinary Debate*. Clarendon Press, New York 1995, pp. 351-394.

[10] For a recent example of using algorithmic processing to provide evidence for bias against the poor in social networks see G. Curto, M.F. Jojoa Acosta, F. Comim et al., *Are AI systems biased against the poor? A machine learning analysis using Word2Vec and GloVe embeddings*, «AI & Society», n. 39, 2024, pp. 617–632. https://doi.org/10.1007/s00146-022-01494-z.

[11] See G. Curto, M.F. Jojoa Acosta, F. Comim et al., *Are AI systems biased against the poor?*, cit.; L. Marinucci, C. Mazzuca, A. Gangemi, *Exposing implicit biases and stereotypes in human and artificial intelligence: state of the art and challenges with a focus on gender*, «AI & Society», n. 38, 2023, pp. 747–76.

[12] J. Shane, *You Look Like A Thing and I Love You*, cit.

data it is trained on will have an impact on the decision it reaches.[13] Typically, the data training set includes examples of successful CVs for similar positions. If the majority of the successful CVs share a characteristic e.g. sex this can be considered as an essential characteristic by the algorithm so that successful CVs not including this characteristic are ignored. This leads to a generation of a model which does not represent all training data. Instead, the model is based on the data which give the faster route to the preset solution interpreted as a cluster of necessary characteristics found in most successful CVs. In practice this means that even if the training data set includes minority characteristics, through algorithmic processing, this data is being ignored. This leads to inaccurate models of real environments[14].

In addition, a second type of overfitting is associated with known errors such as using unrepresentative sampling and it can be combined with the former type of overfitting to exacerbate AI bias. When the algorithm is trained on limited data, its outcome predictions can only be accurate when applied to limited cases such as the ones reflected by the limited data. However, if the prediction generated by the algorithm, is used as a generic model to be applied to different environments - so environments not reflected in its training data, the prediction is inaccurate as it is fitted or applied to environments not reflected in its training set (overfitting via overextending). For example, we may train an algorithm tasked to recognise facial expressions by using images of facial expressions from 30 students of a particular age group or of a particular social background studying at a specific university. If we then use our algorithm to predict the facial expression of an individual who is not represented in the training data, we are asking our algorithm to overextend. AI overfitting occurs with all types of learners such as decision trees or neural nets and it partially depends on the number of training examples; the more training examples are included in the data, the lesser degree of this type of AI overfitting occurs[15]. Both types of overfitting discussed above lead to biased predictions or the creation of inaccurate causal models and as such can be considered as instances of AI bias.

Importantly, the quality of the training data has an impact on the level of overfitting. For example, quality data could lead to the generation of more accurate models in the sense of the models matching real environments or contexts. The accuracy would depend on whether the models accurately reflect causal patterns in real environments. To ensure that algorithmic processing would generate accurate models, these causal patterns must be reflected in the data we feed to the algorithm. However, this presupposes that:

1) we already have a good idea and understanding of these causal patterns (condition 1) and
2) we have a mechanism e.g., a test or screening, which checks that the data fed to the algorithm reflect these patterns and do not reflect racist, sexist etc. patterns (condition 2).

[13] Here we are using a hypothetical scenario but for a study on gender bias in hiring see A. Peng, B. Nushi, E. Kıcıman, K. Inkpen, S. Suri, E. Kamar, *What you see is what you get? The impact of representation criteria on human bias in hiring*, «Proceedings of the AAAI Conference on Human Computation and Crowdsourcing», vol. 7, n. 1, 2019, pp. 125-134.
[14] See G. Curto, M.F. Jojoa Acosta, F. Comim et al., *Are AI systems biased against the poor?*, cit.
[15] See J.S. Russell, P. Norvig, *Artificial Intelligence, A Modern Approach*, cit.

Both these requirements need to be satisfied and so are necessary but not sufficient for the training data to be considered as quality data. The second condition, that is the requirement that we check that the training data does not reflect discriminatory patterns, is a necessary condition for quality data because in its absence algorithmic processing can discover patterns in the data even when there are no causally relevant patterns to be discovered within. By relevant here we mean patterns showing a causal relation between features or properties included in the data and the outcome/task we set the algorithm to perform. For example, if the task is to predict whether the roll of a dice will come up as 5 or not, it is not relevant whether the thrower of the dice is wearing a batman suit or pyjamas when throwing the dice or whether the day is Tuesday, Wednesday or Thursday. However, if these data are fed into the training set of the algorithm, it is possible that a pattern can be discovered which predicts that the dice will come up as 5 if today is Tuesday and the thrower is wearing pyjamas. Generally speaking, in AI overfitting the algorithm will come up with a pattern which will be fitted 'over' the data even when this pattern may not reveal any causally relevant patterns.

As many people would expect the use of algorithmic processing to infuse objectivity and accuracy into this process, this discrimination is amplified; because of the 'veil of objectivity' associated with using AI systems rather than humans in these contexts, there are less checks and tests on the decisions/predictions generated compared to decisions/predictions generated by human agents[16].

Ultimately if our priority is to generate accurate models of the world, algorithmic processing must be primed towards accuracy. In simple contexts and environments this is easily achieved. On the other hand, when the model we are after relates to complex environments, algorithmic processing cannot guarantee accuracy. As such, we need to ensure that algorithmic processing is used in the right context and not presuppose that it is a tool that can be used to generate models in every context[17]. In addition, when algorithmic processing is used to make predictions or decisions on complex matters where many variables are causally relevant and some of these relevant variables are morally salient e.g. determining parole, predictive policing, prison sentences, mortgage applications etc., fairness must also be primed if one wants to argue that these tasks are appropriate tasks for algorithmic processing[18].

Bias maintained via overfitting can be further exacerbated by "looping effects" where an interactive loop is created between the data we use to train the algorithm and the outcome predictions generated by the algorithm. To show how looping effects can

[16] The mistaken assumption here is that the algorithm cannot have any biases relating to race, sex, class etc. because it is not an agent, even when these biases are endemic in the data it is trained on. For a discussion of AI bias and its implications on human users, see M. Glickman, T. Sharot, 15 November 2022, *Biased AI systems produce biased humans*. https://doi.org/10.31219/osf.io/c4e7r.

[17] L. Marinucci, C. Mazzuca, A. Gangemi, *Exposing implicit biases and stereotypes in human and artificial intelligence: state of the art and challenges with a focus on gender*, cit.

[18] See B. Giovanola, S. Tiribelli, *Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms*, «AI & Society», n. 38, 2023, pp. 549-563. https://doi.org/10.1007/s00146-022-01455-6; B. Green, Y. Chen, *Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments*, «FAT* 2019 - Proc 2019 Conference on Fairness, Accountability, Transparency», 2019, pp. 90-99. https://doi.org/10.1145/3287560.3287563; P. Hacker, *Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law*, «Common Market Law Review», n. *55*, 2018, pp. 1143-1185. https://doi.org/10.54648/COLA2018095.

exacerbate the effects of overfitting, consider an example relating to healthcare costs. An algorithm is tasked with calculating a percentage of healthcare needs by focusing on healthcare costs; it predicts that white patients or patients living in specific areas - the ones where people spend more on healthcare costs, will need additional care. An immediate result may be that more resources are allocated towards healthcare needs of white patients or to patients living in these specific areas. A subsequent result is that people who consider themselves as members of these groups or classifications e.g., white patients or patients living in these specific areas, become aware that other members of their group spend more on healthcare costs. The possible consequences of this awareness are either that patients change their behaviour to conform with what other members of the group are doing, do nothing, or change their behaviour to stop conforming with other members of the group. In this paper we focus on the first possibility as this can have a detrimental effect in exacerbating existing socio-economic imbalances and further increasing the negative effects of AI overfitting. The patients living in this area now have a larger selection of healthcare facilities or activities in the area which gives them the opportunity to spend more funds on healthcare costs. This creates an interacting loop where the more funds are spent towards healthcare costs, via the subsequent algorithmic prediction, the more funding is allocated to that area for example towards healthcare facilities, so that patients end up spending even more in that area. However, as we hope the example shows, by using algorithmic processing to predict healthcare needs, the wrong conclusion is being reached in that an area with fewer healthcare needs is flagged up as an area with more healthcare needs[19].

One obvious reason why the above happens is that we are drawing the wrong conclusion from actual data because we are allowing wrongful associations between healthcare costs and healthcare needs. Instead, it is often argued that we should train the algorithm with data showing actual healthcare needs and levels of health. However, doing this may be complicated because some of this data will be protected by data protection law as special category of data for very good reasons or because we have insufficient data to determine healthcare needs.

In the third part of this paper agreeing with Zajko[20] we suggest that, algorithmic processing in general, and overfitting in particular, is a tool that could be used to alleviate rather than exacerbate socioeconomic bias in certain specific cases.

2. Regulatory frameworks addressing AI bias with focus on overfitting

As AI research and innovation is advancing rapidly, it is not clear that current legislation is sufficient or effective in protecting individuals from the negative consequences of algorithmic processing. The existing legal framework in the EU offers some opportunities to fight discriminatory effects of algorithmic processing through data protection law (especially Article 22 GDPR), consumer law, criminal and administrative law (regarding fair procedures) or the provisions of anti-discrimination law. At the same time, there are significant difficulties in applying these in practice. For example, in relation to equality laws,

[19] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, *Dissecting racial bias in an algorithm used to manage the health of populations*, «Science», 25 October 2019, 366, 6464, pp. 447-453. 10.1126/science.aax2342.
[20] See M. Zajko, *Conservative AI and social inequality: conceptualizing alternatives to bias through social theory*, «AI & Society», 36, 2021, pp. 1047-1056. https://doi.org/10.1007/s00146-021-01153-9.

the prohibition of indirect discrimination does not provide a clear and easily applicable rule; it needs to be proven that a seemingly neutral policy or measure disproportionately affects a protected group and is therefore *prima facie* discriminatory. This means that statistical evidence is needed to prove that such a disproportionate result is taking place. Moreover, non-discrimination laws apply to defined groups sharing certain protected characteristics, but AI systems could negatively affect groups of people who do not share such protected characteristics for example people living in specific neighbourhoods or people who are associated with particular political parties[21].

Another limitation of the existing legal framework is that data protection law does not apply to algorithmic processing of non-personal data. For example, predictive models can include non-identifiable data that are outside the scope of data protection law. As mentioned above, sometimes processing personal data is necessary to identify AI bias and discriminatory affects, in the sense that we may need to use algorithmic processing on personal data to prove discrimination is taking place[22]. This latter consideration reveals a tension between data protection rights and the need to eradicate discrimination and bias. Most importantly, individuals must be aware of the use of algorithms when these are used in decisions or predictions that impact on their lives. However, most of these harms such as unfair discrimination, or data protection breaches are for a long time invisible to the people subjected to them[23].

As the existing regulatory framework is struggling to address the issues posed by the algorithmic processing, measures have been proposed from human rights organisations to protect among other public interests, fairness, and human rights. In relation to predicting policing, the European Agency for Fundamental Rights (FRA) highlights that the quality of training data and other sources associated with bias need to be mandatorily assessed by the users of these autonomous analytics systems[24]. It also suggests that the outputs of algorithmic systems become the basis for updated algorithms and that assessments are needed both before and during the use of algorithmic processing.

The above feed into the discussion about the necessity of AI regulation addressing AI systems' risks. The AI Act Regulation (EU 2024/1689) laying down harmonised rules on artificial intelligence entered into force on 1 August 2024, and will be fully applicable 2 years later on 2 August 2026, with some exceptions.[25] This regulatory framework has a lot in common with the data protection framework. It attempts to establish obligations for AI providers, deployers and users depending on the level of risk the AI systems can generate, and the adverse impact caused by AI systems on fundamental protected rights e.g., dignity, protection of personal data, right to non-discrimination, employment rights, rights of persons with disabilities, presumption of innocence. When such risks and impacts are identified, AI applications should be classified as high-risk.

---

[21] See F. Zuiderveen Borgesius, F., *Discrimination, artificial intelligence, and algorithmic decision-making*, Directorate General of Democracy, Council of Europe, 2018; U. Peters, *Algorithmic political bias in artificial intelligence systems*, «Philosophy & Technology», n. 35, 2022. https://doi.org/10.1007/s13347-022-00512-8.

[22] See M.J. Kushner, J.R. Loftus, *The Long Road to Fairer Algorithms: Build models that identify and mitigate the causes of discrimination*, «Nature», 578, 2020, pp. 34-38.

[23] See C. Véliz, *Privacy Is Power*. Melville House, London, 2021, p.39.

[24] See FRA. European Union Agency for Fundamental Rights, *Bias in algorithms. Artificial Intelligence and Discrimination*, Publications Office of the European Union, Luxembourg 2022.

[25] Available online at: https://artificialintelligenceact.eu/the-act/.

Examples of high-risk AI systems are the ones used for selection of persons in educational or training institutes, for recruitment purposes, promotion decisions, social score, credit score evaluations or creditworthiness of persons (e.g. profiling). In the public sector, examples of high-risk AI applications are the ones used for decisions relating to social benefit entitlement, predictive policing, crime analytics, emergency services, or evaluation of healthcare needs (Annex III). To mitigate the risks, AI systems should be used only if they comply with certain mandatory requirements, such as risk management, use of high-quality and relevant data sets, maintaining technical documentation, record-keeping transparency, the provision of information to deployers, human oversight, robustness, accuracy and cybersecurity as well as compliance with the EU legislation. Risk management systems should consist of a continuous, iterative process that is planned and run throughout the entire lifecycle of using AI systems in high-risk areas and it should be regularly reviewed and updated to ensure its continuing effectiveness.

One of the AI Act's main objectives is to mitigate discrimination and bias in the development, deployment, and use of "high-risk AI systems". The AI Act takes under consideration cases of overfitting similar to the ones analysed in this paper and has related provisions. According to art.10 of the AI Act, high-risk AI systems which make use of techniques involving the training of AI models with data, should be developed with quality criteria on the basis of training, validation and testing data sets. Training, validation and testing data sets shall be relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose. These data sets should have the appropriate statistical properties, as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used. Data sets shall consider, to the extent required by the intended purpose, the characteristics or elements that are particular to the specific geographical, contextual, behavioural, or functional setting within which the high-risk AI system is intended to be used. In addition, according to 10 (5), for the purposes of ensuring bias detection and correction, processing of sensitive data for the use of high-risk AI systems is exceptionally permitted, subject to appropriate safeguards for the fundamental rights and freedoms of natural persons.

According to article 12 (5), high-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way as to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations (feedback loops), and as to ensure that any such feedback loops are duly addressed with appropriate mitigation measures. Outputs of AI systems could be influenced by such inherent biases that are inclined to gradually increase and thereby perpetuate and amplify existing discrimination.

As additional safeguards, deployers of high-risk AI systems which are also bodies governed by public law (e.g. police) as well private operators that provide services such as banking and insurance and private entities providing public services linked to tasks in the public interest such as in the areas of education, healthcare, social services, housing, administration of justice have the obligation to the extent the deployer exercises control over the input data, they must ensure that input data is relevant and sufficiently representative in view of the intended purpose of the high-risk AI system (art. 26). They also carry out a fundamental rights impact assessment prior to using the AI system and determine measures to mitigate the identified risks arrangements through human oversight, complaint handling and redress procedures (art. 27). Deployers can also involve relevant

stakeholders e.g. representatives of people who are expected to be affected by these AI systems and co-design measures to mitigate the risks. Furthermore, the affected people have the right to lodge a complaint with the market surveillance authority and the right to request an explanation if the output of certain high-risk system produces legal effects or significantly affects and impacts their health, safety, or fundamental rights (art. 86).

As shown above, during recent years there have been several regulatory attempts to identify the problem caused by AI bias, identify obligations for developers and users, assess the life cycle of these systems and suggest best practices to minimise their socially negative impacts. Unlike the EU approach followed by countries as Canada and possibly Latin American countries, the UK published AI White Paper (2023) did not aim to suggest new AI specific regulation; instead, the UK elected to use a context specific, principles-based framework. It released guidelines to empower regulators, allowing for statutory action to be called upon when necessary. The White Paper was put in and the UK government published its response (February 2024). One of the key issues raised by the respondents is that the government's focus on innovation, does not allow room for sufficient focus on AI-related risks, such as bias and discrimination.

All existing and proposed regulations discussed above share the assumptions that algorithmic processing increases efficiency, augments existing capabilities, and has beneficial results for the economy. These regulations and policies also accept that algorithmic processing may replace some human decision-marking processing, maintain discrimination or be a threat to humans whilst the effectiveness of any suggested remedies remains questionable[26].

3. Reconceptualising overfitting and AI system deployment

As shown in the previous chapter, regulating the use of algorithmic processing is currently a work in progress. Broadly speaking, up to now overfitting has been perceived as a technical issue which can be fixed by a technical solution, similar to a machine malfunction. AI assurance scholars have attempted to create standardised methods to detect bias arising from algorithmic processing[27]. Examples of such solutions are bias screening software[28], preprocessing[29] or rigorous testing[30]. These suggestions try to minimise any discrimination observed in algorithmic predictions and decisions by including tests or screenings after the development of the algorithm and before it is released for use.

---

[26] See G. De Gregorio, S. Demkova, *The Constitutional Right to an Effective Remedy in the Digital Age: A Perspective from Europe* (SSRN Scholarly Paper 4712096), «Social Science Research Network», 2024, https://doi.org/10.2139/ssrn.4712096; For more detail see the government response on https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response#fn:45.

[27] See F. Batarseh, J. Chandrasekaran, L. Freeman, *An introduction to AI assurance*, in A. Feras, F. Batarseh, L. Freeman (eds.), *AI Assurance: Towards Trustworthy, Explainable, Safe, and Ethical AI*, Elsevier Science & Technology, Amsterdam 2022.

[28] See C. Wilson, A. Ghosh, S. Jiang, A. Mislove, L. Baker, J. Szary, K. Trindel, F. Polli, *Building and Auditing Fair Algorithms: A Case Study in Candidate Screening*, «Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)», Association for Computing Machinery, New York 2021, pp. 666-677, https://doi.org/10.1145/3442188.3445928.

[29] F. Kamiran, T. Calders, *Data preprocessing techniques for classification without discrimination*, «Knowledge and Information Systems», 33, 2012, pp. 1-33.

[30] See M.J. Kushner, J.R. Loftus, *The Long Road to Fairer Algorithms*, cit.

However, and as Johnson argues, «There are no purely algorithmic solutions to the problems that face algorithmic bias»[31]. Partly this is because AI overfitting does not only occur when we develop faulty mathematical constructs. It occurs regardless as it based on inferential reasoning; we use the algorithm as a mechanism extracting patterns from data we feed into it. When the algorithm picks up patterns in the data, these often are the result of structural discrimination or intersectional discrimination. When this algorithm is used to generate decisions in the form of prediction purposes, it reinforces these inequalities in societies. Whilst reviewing current and proposed legislation we mentioned that in some cases to prove discrimination is taking place one needs to provide statistical evidence that discrimination is taking place. Here we would like to suggest that algorithmic processing can be used to detect this discrimination and so play the part of statistical evidence proving discrimination. Risk and human rights assessments imposed by AI regulation can play a substantial role by providing the framework and justification for using algorithmic processing to detect rather than perpetuate bias. However, for this to happen these risk and human rights assessments should be developed in a way that incorporate interdisciplinary approaches and related academic research and serve the purposes of accountability and social control[32]. External (i.e., second party) tracking mechanisms and independent (i.e., third-party) oversight that blend the process and the outcome of the algorithm systems can surpass any bureaucratic uses[33]. We could use algorithmic processing to identify where bias is occurring within different contexts especially in cases where this harm is invisible and undetected for long periods[34].

In what follows, borrowing from recent research we give an example of how overfitting could be included as an important step in the exercise of detecting bias and revealing improper decision-making processes.

3.a Example of using overfitting to detect bias and faults in decision making process

A Guardian investigation[35] in 2019 found that UK local councils used algorithmic processing on data held on claimants of housing and council tax benefit to determine the likelihood these claims were fraudulent by using "risk-based verification". These systems were designed to perceive and reveal groups of people as types of risks[36]. However, as the investigation revealed, most of the cases deemed high risk by the software were in fact lower risk, and as a result, benefit claims were wrongly delayed. The algorithms drew on

[31] G.M. Johnson, *Algorithmic bias: on the implicit biases of social technology*, «Synthese», 198, 2021, p. 9943, https://doi.org/10.1007/s11229-020-02696-y.
[32] See I.J. Monteiro, *The Need for Responsible Use of AI by Public Administration: Algorithmic Impact Assessments (AIAs) as Instruments for Accountability and Social Control*, in J. Goossens, E. Keymolen, A. Stanojević (eds.), *Public Governance and Emerging Technologies*, Springer, Cham 2025. https://doi.org/10.1007/978-3-031-84748-6_9.
[33] See A. Brandusescu, R. Sieber, *Design versus reality: assessing the results and compliance of algorithmic impact assessments*, «Digital Society», vol. 4, art. 64, 2025, https://doi.org/10.1007/s44206-025-00221-7.
[34] See M. Zajko, *Conservative AI and social inequality*, cit.
[35] S. Marsh, *One in three councils using algorithms to make welfare decisions*, «The Guardian», 15 October 2019, https://www.theguardian.com/society/2019/oct/15/councils-using-algorithms-make-welfare-decisions-benefits.
[36] L. Dencik, A. Hintz, J. Redden, H. Warne, *Data Scores as Governance: Investigating uses of citizen scoring in public services*, Project Report, Data Justice Lab. Cardiff University, 2018.

and used data about people who make use of social services; the data itself was biased through the over-representation of a particular part of the population making use of social services often as a result of being marginalised and resource poor[37]. Overfitting in this instance not only revealed a problem with the data itself but also with the task we set the algorithm to perform and with how the algorithm was used in the decision-making process. It revealed that the decision-making process did not follow proper procedure.

To prevent such occurrences, we suggest following the framework and principles suggested by recent research conducted by the Centre for Data Ethics and Innovation and ICO. The Centre for Data Ethics and Innovation independent report titled *Review into bias in algorithmic decision-making*[38], analyses paradigms in four different areas as recruitment, financial services, policing and social services and highlights the importance of involving all stakeholders e.g. decision makers, industry, policy makers and society to determine whether the overall decision-making processes are biased. It suggests the following as important factors in alleviating bias from algorithmic decision making when they are used by organisations for decision making:

A. Understanding the capabilities and limits of those tools.
B. Considering carefully whether individuals will be fairly treated by the decision-making process that the tool forms part of.
C. Making a conscious decision on appropriate levels of human involvement in the decision-making process.
D. Putting structures in place to gather data and monitor outcomes for fairness.
E. Understanding their legal obligations and having carried out appropriate impact assessments

These recommendations formed the key principles of the AI Playbook for UK Government[39]. In addition, ICO in its *AI fairness considerations across the AI lifecycle* guidance (Annex A), proposes the following methodology when overfitting is detected to evaluate the data and its relation to the algorithm and the model generated:

1) examine the data's context (particularly whether there is an appropriate representation of groups in the training data);
2) evaluate the values that are assigned to the features before someone feeds them into the model,
3) tweak the model by tuning its hyperparameters; and
4) fit the most appropriate algorithm to the data. The algorithm might not be the appropriate one, so other algorithms must be tested as well as their performance.

[37] This was also observed in the use of the SyRI (system risk indication) system implemented by the Dutch government to identify potential fraudulent beneficiaries, see M. Bekkum, F. Borgesius, *Digital welfare fraud detection and the Dutch SyRI judgment*, «European Journal of Social Security», 23, 2021, pp. 323-340, https://doi.org/10.1177/13882627211031257.

[38] Centre for Data Ethics and Innovation, *Review into bias in algorithmic decision-making*, 2020. Retrieved 10 April 2024, from: https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making/main-report-cdei-review-into-bias-in-algorithmic-decision-making.

[39] Government Digital Service, *Artificial Intelligence Playbook for the UK Government*, February 2025. Retrieved 1 August 2025 from: https://www.gov.uk/government/publications/ai-playbook-for-the-uk-government.

The guidelines and framework included above should be used as soon as AI overfitting is discovered to re-evaluate the whole decision-making process. In other words, overfitting is an indicator that a proper procedure was not followed when determining the decision-making process. Although detecting overfitting is not a simple procedure, empirical evidence of overfitting can be obtained when the generalisation error of a model generated by the algorithm is higher than what is expected by the algorithm's performance in training[40]. Our suggestion matches one of the aims of the Centre for Data and Innovation review, namely that algorithms can enable the identification and mitigation of systematic bias in cases where doing so would be challenging for human agents. If the task of the algorithm changes from generating predictions matching dominant trends to revealing patterns of dominant trends and generating models based on these patterns this could help identify these patterns as discriminatory.

Applying the framework in the example discussed above we would need to: a) justify the use of data processing in that area e.g. evidence for fraud in the area b) conduct a community consultation to identify benefits and risks of using data processing c) reevaluating the hypothesis that the length of time of using the benefits constitutes an indication of fraud, d) consider the wider context by consulting with experts e.g. social workers e) consider consequences of using algorithmic processing in this case and identify any remedies for harms caused from the outcomes generated by the algorithm f) re-evaluate the wider context and address wider structural societal conditions which maintain the need for housing and council tax benefits.

The purpose of this paper is not to demonstrate a concrete example of using algorithmic processing to detect discriminatory patterns but to motivate decision makers and AI developers by changing the current narratives surrounding the use of AI systems. More specifically, technical experts in AI should be tasked with developing algorithmic tools which can detect hidden discrimination and provide concrete evidence for it. Agreeing with recent literature we accept that AI experts are to some extent, responsible for foreseeable consequences of use of AI technology[41]. Our suggestion can be used as one way AI experts can accept this responsibility by developing tools to detect and identify rather than exacerbate discrimination. This different perspective could contribute to a new narrative applying to AI systems and their social benefits which could potentially demystify and de-demonise them.

In addition, using AI overfitting could provide evidence to initiatives aiming to identify unfairness in key parts of individuals' lives such as in healthcare systems where discrimination is invisible even when training is based on partially filtered data. Using algorithmic processing to reveal patterns associated with healthcare contexts could reveal patterns of discrimination by using proxy features such as post codes, political affiliations etc., data which is not currently protected by law. The AI Act will be implemented in the coming years so that risk management and human rights impact assessments become obligatory in AI systems placed in EU countries. As the constant revaluation of these

[40] C. Aliferis, G.J. Simon, *Overfitting, Underfitting and General Model Overconfidence and Under Performance Pitfalls and Best Practices in Machine Learning and AI*, in G.J. Simon, C. Aliferis (eds.), *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls*, Springer, Cham 2024, pp.477-525.
European Parliament and the Council of European Union, *Artificial Intelligence Act (Final Draft)*, 2024. Accessed 1 April 2024, available at: https://artificialintelligenceact.eu/the-act/.

[41] M. Hedlund, E. Persson, *Expert responsibility in AI development*, «AI & Society», n. 39, pp. 453-464, https://doi.org/10.1007/s00146-022-01498-9.

systems will be obligatory, using algorithmic processing to reveal, explore and monitor existing bias can provide useful evidence for societies invested in ameliorating social and economic inequalities.

Our suggestion is motivated by recent initiatives addressing social or economic discrimination which aim to bring in the centre of this discourse the larger social, economic, and political ecosystem creating discriminatory practices and procedures. Discrimination maintained by algorithmic processing and exclusionary automated systems represent one element of this larger ecosystem. The identification of problems and any potential solutions should be linked with measures of reflexivity in relation to this ecosystem[42]. Our suggestion to use AI overfitting to reveal hidden discrimination could be a practical example of how this reflexivity of the ecosystem takes place in practical terms.

In a similar way it is suggested that we need to evaluate how data used in training map onto recent social, economic, cultural, and political changes on a global scale[43]. More recent views emphasize the need to evaluate the fairness of algorithmic practices by embedding them in social practices instead of focusing on evaluating outcome predictions by mathematical constructs[44]. These approaches pay particular attention to structural inequalities that are reproduced in the algorithmic decision-making process and suggest that these mathematical constructs must extend to relational or structural factors associated with the specific task. We interpret this to mean that the algorithms must be tasked with producing more accurate models of the training data. Agreeing with these recent holistic approaches, we suggest using algorithmic processing to reveal hidden discrimination. This will provide bottom-up evidence which can be helpful to several initiatives suggesting that we should seek to build alternative bottom-up infrastructures to empower marginalised groups and avoid such harms in the future[45].

Societies invested on positive social change can evaluate the use of AI systems during the process of impact assessment which is a process highly prioritised in the AI Act and becomes mandatory from August 2026. Decision makers in critical areas such as welfare and social care, healthcare, transportation, housing and planning, education, policing, and public safety could use any overfitting findings for further policy making and scientific use. An organisation possessing hard data on all these different domains, can provide insight into discriminatory practices endemic in the data tracking non-protected characteristics to provide hard evidence as the first step in the process of addressing them. The next step would be to come up with solutions addressing such inequalities by combining human and machine intelligence[46]. For example, in predictive policing instead

[42] S.P. Gangadharan, J. Niklas, *Decentering technology in discourse on discrimination*, «Information, Communication & Society», vol. 22, n. 7, 2019, pp. 882-899, https://doi.org/10.1080/1369118X.2019.1593484.

[43] L. Dencik, A. Hintz, J. Redden, E. Treré, *Exploring Data Justice: Conceptions, Applications and Directions*, «Information, Communication & Society», vol. 22, n. 7, 2019, pp. 873-881, https://doi.org/10.1080/1369118X.2019.1606268.

[44] B. Green, *Escaping the impossibility of fairness: From formal to substantive algorithmic fairness*, «Philosophy & Technology», vol. 35, n. 90, 2022, https://doi.org/10.1007/s13347-022-00584-6; S. Holm, *The fairness in algorithmic fairness*, «Res Publica», 2022, https://doi.org/10.1007/s11158-022- 09546-3.

[45] S. Costanza-Chock, *Design Justice: Towards an Intersectional Feminist Framework for Design Theory and Practice*, «Proceedings of the Design Research Society 2018», 3 June 2018, available at SSRN: https://ssrn.com/abstract=3189696.

[46] G. Mulgan, *Artificial intelligence and collective intelligence: the emergence of a new field*, «AI & Society», n. 33, 2018, pp. 631–632, https://doi.org/10.1007/s00146-018-0861-5.

of allowing confirmation bias to create feedback loops around specific neighbourhoods[47], algorithmic processing could be used as a diagnostic tool to reveal areas needing urgent development[48].

4. Conclusion

To conclude we hope the above discussion shows that even if AI overfitting is mathematically inevitable, its negative effects can be reframed as catalysts for change and opportunities for public awareness and scientific research. As the review of recent literature shows, we urgently need a constant and persistent effort to detect hidden discrimination prior to, during and after use of algorithmic processing. The fundamental impact assessments that are or will become obligatory (AI Act) for decision makers (deployers of the high risk systems) can assist to raise awareness both for the quality of data included in these training sets and of the adverse effects of AI applications trained on them. If these assessments are implemented in practice by serving purposes of transparency, accountability and social control, they could prevent harmful results to both individuals and communities. A further and more crucial step would be to act on revealed social problems and attempt their resolution.

In this paper we provide support to recent views suggesting that overfitting can be used as a diagnostic tool in the development of AI systems by signifying whether the requirements for responsible use of AI have been met[49] and to views suggesting that new narratives can lead to more responsible and transparent AI practices[50]. As recent attempts to regulate algorithmic processing show, in previous years harm was caused from maintaining and exacerbating structural and historical inequalities. In many cases this harm was invisible to the individuals and communities subjected to it. Recent legislation suggests steps to restrict or make its use safer in certain contexts. This is confirmation that algorithmic processing should not be used in all contexts for all tasks. We support these initiatives and accept that algorithmic processing can be used more radically as a diagnostic tool to detect and reveal hidden structural and historical bias and provide evidence for pre-existing prejudice[51]. By highlighting the interaction with looping effects, we provide additional motivation to use overfitting as a first step towards mitigation of historical prejudice. It remains a choice for decision makers whether they wish to combat or maintain

---

[47] A. Babuta, M. Oswald, *Data Analytics and Algorithms in Policing in England and Wales: Towards A New Policy Framework*, «RUSI Occasional Paper», 2020, available at: https://rusi.org/publication/occasional-papers/data-analytics-and-algorithms-policingengland-and-wales-towards-new.

[48] For risks attached to using AI in predictive profiling as a measure to prevent crime, see also: K. Blount, *Using artificial intelligence to prevent crime: implications for due process and criminal justice*, «AI & Society», n. 39, pp. 359-368. https://doi.org/10.1007/s00146-022-01513-z.

[49] See G. Curto, M.F. Jojoa Acosta, F. Comim et al., *Are AI systems biased against the poor?*, cit.; L. Marinucci, C. Mazzuca, A. Gangemi, *Exposing implicit biases and stereotypes in human and artificial intelligence: state of the art and challenges with a focus on gender*, cit.; M. Zajko, *Conservative AI and social inequality: conceptualizing alternatives to bias through social theory*, cit.

[50] See P. Hayes, N. Fitzpatrick, *Narrativity and responsible and transparent AI practices*, «AI & Society», 2024, https://doi.org/10.1007/s00146-024-01881-8. These requirements include participatory and collective intelligence design, taking context into consideration, trustworthiness, accountability, transparency, and legality.

[51] See G. Curto, M.F. Jojoa Acosta, F. Comim et al., *Are AI systems biased against the poor?*, cit.; M. Zajko, *Conservative AI and social inequality: conceptualizing alternatives to bias through social theory*, cit.

this bias. In any case, we hope we provided additional considerations to motivate using algorithmic processing to support positive social change.