



PART 2

Read the text below and answer questions 14–26.

ALGORITHMIC BIAS: WHEN CODE REFLECTS HUMAN PREJUDICE

Algorithms are sometimes portrayed as objective because they follow mathematical procedures rather than personal opinion. This belief can slide into what critics call “mathwashing”: the assumption that because an output is produced by code, it is automatically neutral. In reality, many systems that rank, recommend, predict, or automate decisions inherit human values and social patterns. They are built from human-generated data, trained to meet human-chosen targets, and deployed in institutions that already distribute opportunities unevenly. When these systems are used for hiring, lending, education, policing, or healthcare, even small statistical distortions can scale into major social consequences, particularly when the decisions are repeated at high volume.

A central pathway for bias is training data. If certain groups are underrepresented, a model may simply learn them poorly, producing higher error rates for those users. This problem has been observed in domains such as face analysis, where systems trained on unbalanced image sets have performed worse for darker-skinned people, especially women. Bias can also enter through labels. Training labels are often treated as factual ground truth, but they may encode subjective judgement or institutional history. A dataset marked “successful employee,” for example, can quietly mirror past promotion decisions, which may have favoured one demographic or rewarded conformity to a particular workplace culture. In such cases, discrimination is not hard-coded; it is imported as a pattern the system is trained to reproduce.

Bias also arises from how developers define the target and the variables used to approximate it. Many outcomes that matter ethically—trustworthiness, risk, potential, or merit—are difficult to measure directly, so systems rely on proxy variables. Proxies are attractive because they are measurable, but they are risky because they can smuggle in inequality while appearing neutral. A well-known example is the use of location. Zip codes can act as proxies for race or class because residential patterns reflect decades of housing policy and segregation. Similarly, credit history may proxy access to financial services, which itself has been shaped by unequal treatment. The model can then appear to be “just using data” while it effectively recycles the consequences of earlier discrimination.

The case of recidivism prediction illustrates the stakes. Some jurisdictions have used risk-assessment tools to estimate the likelihood that a defendant will reoffend. Critics have argued that if the model is trained on criminal justice data that already reflects uneven policing, charging, or sentencing patterns, the tool can amplify those disparities. Even



when a system does not use race explicitly, it may use correlated signals, and its predictions can influence bail, supervision, or sentencing decisions in ways that compound disadvantage. The key point is not that every such tool is identical, but that the appearance of neutrality can mask how deeply the outcome depends on the social system that produced the data.

How models are evaluated can further conceal problems. A system may achieve high overall accuracy while failing badly for minority subgroups, especially when the dominant group is larger in the dataset. Optimising average performance can effectively mean optimising for the majority. For this reason, fairness audits examine subgroup error rates and compare different definitions of fairness, such as equal false-positive rates or equal opportunity. Yet fairness is not a single technical switch. Some fairness metrics conflict: improving one criterion can worsen another, and equalising errors can require changing thresholds in ways that shift who bears risk. These are not purely mathematical decisions, because they determine how harms and benefits are distributed across real people.

Bias can also emerge after deployment through feedback loops. Predictive policing provides a clear example: if a model sends more patrols to an area labelled “high risk,” more offences are detected there, generating more recorded incidents that then reinforce the model’s belief that the area is dangerous. The same dynamic can occur in recommendation systems. If a platform repeatedly shows certain stereotypes or career suggestions to particular users, it can shape preferences and opportunities over time, turning prediction into production. In this sense, the model is not only describing the world; it can actively participate in constructing the patterns it later “learns.”

Because these systems can be opaque, the “black box” problem has become a major concern. Complex models may be difficult to interpret even for their creators, and affected users may have little insight into why a decision was made. This has encouraged interest in explainable AI, which aims to provide reasons, feature importance, or simplified representations of a model’s behaviour. However, explanations can be misleading if they are not faithful to the true decision process, and transparency does not automatically guarantee fairness. A clear explanation of an unfair rule does not make it fair. Nonetheless, better documentation and accountability can make it harder for organisations to hide behind automation as a shield against responsibility.

Mitigation strategies exist, but each has limits. Data can be rebalanced, labels can be reviewed, and models can be constrained to meet selected fairness criteria. Some organisations publish “model cards” describing intended use, limitations, and evaluation results across subgroups. Independent auditing can also identify harms that internal teams miss. Yet technical fixes cannot substitute for institutional reform. If an organisation



uses an algorithm to legitimise decisions it would otherwise have to justify, the system can entrench rather than reduce bias. The most credible approaches therefore treat algorithmic bias as socio-technical: reducing harm requires better methods and better governance, including clear accountability, external scrutiny, and careful choices about where automation should be used—and where it should not.

Questions 14–19

Select the correct answer.

14 Why can algorithms seem neutral while still producing unfair outcomes?

- A** They become biased only when programmers add discriminatory rules.
- B** They learn from human-generated data and targets shaped by historical inequality.
- C** They are unfair mainly because they cannot process large datasets efficiently.
- D** They avoid using proxies, so they must rely on random guessing.

15 What problem with training labels is emphasised in the passage?

- A** Labels are always generated automatically and therefore unreliable.
- B** Labels prevent models from learning patterns across groups.
- C** Labels can encode past institutional judgements such as promotion practices.
- D** Labels are accurate only for minority groups and inaccurate for majorities.

16 Why are proxy variables described as ethically risky?

- A** They can reproduce discrimination indirectly while appearing neutral.
- B** They are illegal to use in all high-stakes decisions.
- C** They eliminate the need for subgroup evaluation.
- D** They guarantee that the model becomes fully transparent.

17 What does the passage imply about evaluating model performance?

- A** Overall accuracy is sufficient if the dataset is large enough.
- B** Subgroup testing is unnecessary when a model is explainable.
- C** High accuracy proves that a system is fair.
- D** Average accuracy can conceal worse outcomes for minority subgroups.

18 What does the passage say about defining fairness?

- A** One fairness metric can always satisfy all other metrics.
- B** Different fairness metrics can conflict, creating trade-offs.
- C** Fairness is purely technical and does not involve value choices.
- D** Fairness audits should avoid comparing error rates across groups.



19 What is the passage's overall view of mitigation?

- A** Technical methods help, but governance and institutional practices are also essential.
- B** Bias disappears once a model card is published.
- C** Removing all proxies is sufficient to eliminate unfairness.
- D** Feedback loops cannot occur if recommendations are personalised.

Questions 20–23

Complete the summary below. Choose **ONE WORD ONLY** from the text for each answer.

How algorithmic bias develops and persists

The passage explains that a model may learn some groups poorly if those groups are **20** in the material used for training. Developers may also use measurable variables as **21** for qualities that cannot be assessed directly, allowing inequality to enter apparently neutral systems. Rather than judging only overall accuracy, fairness audits examine subgroup error **22** Once systems are deployed, feedback **23** can cause their outputs to reinforce the data patterns behind later predictions.

Questions 24–26

Choose **TRUE** if the statement agrees with the information given in the text, choose **FALSE** if the statement contradicts the information, or choose **NOT GIVEN** if there is no information on this.

24 The passage suggests that some fairness corrections may reduce overall accuracy.

25 The passage states that making an algorithm transparent automatically ensures that it is fair.

26 The passage provides exact legal requirements that apply to algorithmic systems in every country.

Answer Key

14. B

15. C

16. A

17. D

18. B



19. A

20. underrepresented

21. proxies

22. rates

23. loops

24. NOT GIVEN

25. FALSE

26. NOT GIVEN