American Scholarly Journal for Scientific Research (ASJSR)
https://asjsr.org/ Volume 2025, Issue 1003

Fair or Flawed? A Review Paper on Bias in Artificial Intelligence

Author: Ethan Zao
Institution: Staten Island Technical high School, NY
Mentor: Peter Graham
Date: September 2, 2025

Abstract

Artificial Intelligence, or AI, is now used in many parts of daily life. It helps recommend videos, check passport photos, filter job applications, detect fraud, translate languages, and even support decisions in areas like healthcare, policing, education, and banking. AI is often seen as smart, fast, and neutral. However, AI is not always fair. Because AI systems are created by humans and trained on human data, they can learn the same unfair patterns that already exist in society.

This review paper studies the problem of bias in AI. Algorithmic bias happens when an AI system produces unfair results for certain people or groups. This may happen because the training data contains social bias, because some groups are underrepresented in the data, because the system measures the wrong thing, because feedback loops make old patterns stronger, or because people purposely manipulate the data. These problems can cause serious harm, especially when AI is used in sensitive areas like hiring, housing, loans, policing, healthcare, and education.

The paper explains five major causes of AI bias in simple language. It also discusses examples such as gender stereotypes in search results, facial recognition problems, biased predictive policing, AI essay grading, hiring systems, and manipulated chatbots. The paper further reviews why AI bias is difficult to fix. Some AI systems are hard to understand because they work like "black boxes." Also, collecting more data about protected groups may help test fairness, but it can also raise privacy concerns.

The paper concludes that AI should not be trusted blindly. Humans must test, monitor, and question AI systems before using them in important decisions. Fair AI requires better data, transparency, human oversight, ethical design, and strong rules. AI can be useful, but only if society makes sure it is used carefully and responsibly.

Keywords

Artificial Intelligence, AI bias, algorithmic bias, fairness, discrimination, machine learning, ethics, data bias, transparency, accountability, human rights

1. Introduction

Artificial Intelligence is becoming a major part of modern life. Many people use AI every day without even noticing it. When Netflix recommends a show, when Instagram chooses which posts to show first, when Google completes a sentence, or when a shopping app recommends products, AI is working in the background.

AI is also used in more serious areas. Some companies use AI to screen job applications. Some banks use algorithms to help decide whether a person should get a loan. Some schools use technology to detect cheating or grade writing. Some police departments have used prediction tools to decide where crime may happen. Hospitals also use AI to help doctors read medical scans and study patient records.

Because AI can process large amounts of data very quickly, many people believe it is more objective than humans. They may think, “A computer does not have feelings, so it must be fair.” However, this is not always true. AI systems are made by people. They are trained on data created by people. If the data contains unfair patterns, the AI can learn those patterns.

This problem is called **algorithmic bias** or **AI bias**. It means that an AI system gives unfair results because of the data, design, assumptions, or use of the system. The uploaded source explains that AI is “math and code,” but because it is built by people and trained on human data, it can copy or even increase real-world bias.

Bias does not always mean someone is trying to be unfair on purpose. A person may have bias because they have seen a certain pattern many times. For example, if a person grows up seeing mostly men as engineers in movies and mostly women as nurses in advertisements, they may start to connect gender with jobs. AI can do something similar. If it is trained on data where men are often shown as programmers and women are often shown as nurses, it may learn the same pattern.

The problem becomes serious when bias leads to discrimination. Discrimination means unfair treatment of people because of things like race, gender, age, religion, disability, or other protected identities. In areas like employment, housing, credit, healthcare, and education, unfair AI decisions can affect a person’s life in a major way.

In 2016, a White House report warned that big data and algorithmic systems could create both opportunities and dangers in areas such as credit, hiring, higher education, and criminal justice. It showed that technology can help society, but it can also repeat unfair patterns if not handled carefully.

This review paper explores what AI bias is, where it comes from, why it matters, and how it can be reduced. The paper is written in simple language for high school students so that the topic feels understandable and relevant.

2. Aim of the Review Paper

The aim of this paper is to review the problem of bias in artificial intelligence and explain it in simple words.

The paper focuses on these questions:

1. What is AI bias?
2. Why do AI systems become biased?
3. What are the major types of AI bias?
4. How can AI bias harm people in real life?
5. Why is AI bias hard to detect and fix?
6. What can governments, companies, schools, and citizens do to make AI fairer?

3. Methodology

This paper uses a **review method**. This means it does not conduct a new experiment or survey. Instead, it studies existing information, examples, reports, and ideas about AI bias.

The paper uses:

- The uploaded student draft on AI bias
- Government and policy reports
- AI ethics frameworks
- Real-world examples of biased AI systems
- Simple explanations of machine learning and data bias
- Recent ideas about fairness, transparency, and regulation

The goal is not to prove one new theory, but to bring together important information and explain it clearly.

4. What Is AI Bias?

AI bias happens when an AI system gives unfair or unequal results. This can happen even if the system was not designed to be unfair.

For example, imagine a company uses an AI tool to choose job candidates. The AI looks at old hiring data. If the company mostly hired men in the past, the AI may learn that male applicants

are more likely to be successful. As a result, it may give lower scores to female applicants, even if they are qualified. The AI is not "thinking" like a human. It is simply learning patterns from the past.

This is one of the biggest dangers of AI. If the past was unfair, and AI learns from the past, then AI may continue the unfairness.

AI bias can happen in many ways:

- A system may work better for one race than another.
- A system may treat men and women differently.
- A system may make worse predictions for people from poorer areas.
- A system may reject older job applicants unfairly.
- A system may misunderstand accents, languages, or names from certain communities.

AI bias is important because AI decisions can look scientific and neutral. If people believe the machine is always correct, they may not question unfair results. This can make discrimination harder to notice.

5. Why Bias Enters AI Systems

AI systems usually learn from data. Data is information. It can include text, images, numbers, videos, clicks, purchases, grades, medical records, or police records.

If the data is incomplete, unfair, old, or unbalanced, the AI can learn wrong lessons. Bias can also come from the way humans design the system. A developer may choose the wrong goal for the AI. A company may care more about speed or profit than fairness. A government may use AI without checking whether it harms certain groups.

The National Institute of Standards and Technology, or NIST, explains that trustworthy AI should be valid, reliable, safe, secure, accountable, transparent, explainable, privacy-enhanced, and fair with harmful bias managed. This shows that fairness is now seen as a core part of responsible AI, not just an extra feature.

AI bias usually does not have one single cause. It is often the result of many small problems added together.

6. Major Types of AI Bias

6.1 Bias from Existing Social Patterns

The first major type of AI bias happens when training data reflects existing social bias.

AI learns from examples. If those examples contain stereotypes, the AI may learn stereotypes too. For example, if online text often connects the word “nurse” with women and “programmer” with men, an AI language model may repeat this pattern. This does not mean the AI understands gender. It only means the AI has seen those word patterns many times.

This is dangerous because the AI may make stereotypes appear normal. It may suggest that some jobs, roles, or behaviors belong more to one group than another. This can affect search engines, translation tools, image generators, hiring systems, and educational tools.

Another issue is that protected categories may appear indirectly. Even if a system does not directly use race, religion, gender, or age, it may use other information that is connected to those categories. For example, zip code can sometimes be connected to race or income because of historical housing patterns. Shopping habits may be connected to gender. School names may be connected to class background.

This means removing sensitive labels from data does not always remove discrimination. Bias can hide inside other features.

6.2 Bias from Unbalanced Training Data

The second type of bias happens when some groups are not properly represented in the training data.

For example, if a facial recognition system is trained mostly on images of light-skinned faces, it may work less accurately for darker-skinned faces. If a speech recognition system is trained mostly on one accent, it may misunderstand people with different accents.

This type of bias is not just a technical mistake. It can affect real people. If an airport face scanner fails more often for some groups, those people may face more delays. If a security system misidentifies some groups more often, innocent people may be treated unfairly.

The uploaded source gives the example of a passport photo checker that had trouble with photos of people of Asian descent because it incorrectly thought their eyes were closed. This shows how a small design flaw can create frustration and unequal treatment.

The lesson is simple: AI must be tested on many different types of people before it is used in real life.

6.3 Bias from Measuring the Wrong Thing

The third type of AI bias happens when the system tries to measure something complex using a simple shortcut.

Many human qualities are difficult to measure. For example, how do we measure creativity? How do we measure leadership? How do we measure kindness? How do we measure good writing?

Because these qualities are hard to measure, AI systems may use easier signals instead. For example, an AI essay grader may focus on grammar, sentence length, vocabulary, or essay structure. These things matter, but they do not fully measure good writing. A long essay with big words may still have weak ideas. A simple essay may have strong thinking.

The uploaded source discusses how AI essay grading systems can be tricked by writing that looks complex but does not make sense. This is a good example of a system measuring the wrong value.

This problem is sometimes called a **proxy problem**. A proxy is a substitute measurement. The AI cannot measure the real thing, so it measures something close to it. But if the proxy is poor, the results can be unfair.

For example:

- Using school name as a proxy for intelligence can favor rich students.
- Using gaps in employment as a proxy for laziness can hurt caregivers or people with illness.
- Using arrest records as a proxy for crime can reflect biased policing.
- Using writing style as a proxy for ability can disadvantage students from different language backgrounds.

AI systems must be designed carefully so that they measure what actually matters.

6.4 Bias from Feedback Loops

The fourth type of AI bias happens through feedback loops.

A feedback loop happens when an AI system changes the world, and then the changed world becomes new data for the AI. This can make old patterns stronger.

For example, imagine an AI policing tool predicts that one neighborhood has more crime. Police are sent there more often. Because there are more police in that neighborhood, more arrests happen there. The arrest data is then added back into the AI system. The AI now thinks that neighborhood has even more crime, so it sends police there again.

The problem is that other neighborhoods may also have crime, but if police are not sent there, fewer arrests are recorded. The AI then mistakes “more police activity” for “more crime.”

The uploaded source discusses PredPol as an example of this kind of feedback loop, where biased historical data could lead to more policing in certain communities.

Feedback loops are dangerous because they can make unfair systems look like they are supported by data. The AI says, “There are more arrests here,” but it may not understand that there are more arrests partly because it sent more police there in the first place.

6.5 Bias from Malicious Manipulation

The fifth type of bias happens when people purposely manipulate an AI system.

Some AI systems learn from user interaction. This can be useful because the AI can improve over time. But it can also be risky because people may feed the system harmful data.

A famous example is Microsoft's chatbot Tay. Tay was released on Twitter in 2016 and was designed to learn from conversations with users. Very quickly, some users manipulated it by sending offensive, racist, and hateful content. Tay began producing harmful messages and had to be shut down. The uploaded source explains this as an example of malicious data manipulation.

This example shows that AI systems need protection. If an AI learns from the public without strong safety controls, it can be attacked or corrupted.

7. Real-World Areas Where AI Bias Matters

7.1 Hiring and Employment

AI hiring tools are used by some companies to scan resumes, rank applicants, or predict who may be a good employee. This can save time, but it can also create unfairness.

If a hiring algorithm is trained on past company data, it may copy old hiring patterns. If a company mostly hired young people in the past, the AI may prefer young applicants. If it mostly hired men, it may prefer male applicants. If it mostly hired people from certain universities, it may reject talented people from other backgrounds.

This is serious because jobs affect income, dignity, and future opportunities.

7.2 Housing and Loans

AI can also be used in housing and credit decisions. Banks may use algorithms to decide who gets a loan. Landlords or platforms may use systems to screen tenants.

If the data reflects past discrimination, the system may continue to disadvantage certain groups. For example, if people from certain neighborhoods historically had less access to credit, an AI may treat those neighborhoods as risky. But this may reflect past inequality rather than personal responsibility.

The 2016 White House report studied algorithmic systems in areas like credit, employment, higher education, and criminal justice because these are areas where unfair decisions can affect civil rights.

7.3 Healthcare

AI is increasingly used in healthcare. It can help detect diseases, read scans, predict patient risk, and recommend treatment. This can be very helpful.

But healthcare AI can become biased if it is trained mostly on data from certain populations. For example, if a medical AI is trained mostly on patients from one race, gender, or income group, it may not work as well for others.

Healthcare bias can be dangerous because it may affect diagnosis and treatment. In this area, fairness is not just about comfort or convenience. It can affect life and death.

7.4 Education

AI is used in education for grading, tutoring, plagiarism detection, student monitoring, and admissions support. These tools may help teachers, but they must be used carefully.

For example, AI essay graders may reward surface-level features instead of deep thinking. Plagiarism detectors may wrongly flag students who write in a certain style. AI monitoring tools may create stress or unfairly punish students who do not behave in the expected way.

Education is supposed to create opportunities. If AI tools are biased, they may reduce opportunity instead.

7.5 Policing and Criminal Justice

AI in policing is one of the most controversial areas. Predictive policing tools try to forecast where crime may happen. Risk assessment tools may help courts decide bail or sentencing.

The danger is that criminal justice data often reflects unequal policing. If some communities have been policed more heavily in the past, the data may show more arrests there. AI may then treat those communities as more dangerous.

This can lead to a cycle where already watched communities are watched even more.

8. Why AI Bias Is Difficult to Fix

AI bias is not easy to solve because AI systems are complex.

First, some AI systems are difficult to understand. Deep learning systems may use many hidden layers. Even the people who build them may not always know exactly why the system made a decision. This is sometimes called the “black box” problem.

Second, fairness can mean different things. One person may define fairness as equal accuracy for all groups. Another may define fairness as equal opportunity. Another may define fairness as equal outcomes. These definitions do not always match.

Third, fixing bias may require collecting sensitive information such as race, gender, age, disability, or religion. This can help test whether the AI is fair. But it also creates privacy concerns. People may worry that their personal information could be misused.

Fourth, AI systems change over time. A system may seem fair during testing but become unfair after being used in the real world. This is why AI needs regular monitoring.

Fifth, companies may not want to reveal how their algorithms work because they see them as business secrets. This makes transparency harder.

9. Possible Solutions to AI Bias

9.1 Better Training Data

AI systems need high-quality and balanced data. The data should include different groups of people. It should also be checked for historical bias.

Better data does not mean collecting everything about everyone. It means collecting useful data responsibly and protecting privacy.

9.2 Bias Testing Before Deployment

AI should be tested before being used in real life. A facial recognition system should be tested on many different skin tones, ages, and genders. A hiring system should be tested to see whether it treats applicants fairly. A medical system should be tested on different patient groups.

Some experts argue that important AI systems should be tested carefully before public use, similar to how medicines are tested before being widely used. The uploaded source also mentions this idea.

9.3 Transparency and Explainability

People should be able to understand why an AI system made an important decision. If a person is denied a job, loan, or school opportunity because of AI, they should have a way to question the decision.

Transparency does not mean every person must understand complex code. It means there should be clear explanations, records, and accountability.

9.4 Human Oversight

AI should not make serious decisions completely alone. Humans should review important decisions, especially in hiring, healthcare, policing, education, and finance.

However, human oversight must be real. A person should not simply accept the AI output without thinking. Humans must be trained to question AI recommendations.

9.5 Regulation and Ethical Standards

Governments are beginning to create rules for AI. The European Union’s AI Act aims to regulate AI based on risk and addresses issues such as bias, discrimination, and accountability gaps. The

EU law also places rules on high-risk AI systems and includes requirements related to data governance and risk management.

Rules alone will not solve everything, but they can create minimum standards. Companies should not be allowed to use harmful AI systems without testing or responsibility.

9.6 Public Awareness

People need to understand that AI is not magic. It can be helpful, but it can also be wrong. Students, parents, teachers, workers, and citizens should learn basic AI literacy.

A simple rule is: **do not trust AI only because it sounds confident.**

10. Discussion

AI bias is one of the most important technology problems of our time. This is because AI is entering areas that shape human opportunity. A biased movie recommendation may be annoying. But a biased hiring tool, loan system, medical system, or policing tool can seriously harm someone's life.

The biggest problem is that AI bias can hide behind the appearance of objectivity. If a human makes a biased decision, people may challenge it. But if a computer makes the same biased decision, people may assume it is based on facts. This makes AI bias especially dangerous.

The review shows that bias can enter AI at many stages. It can enter through old data, missing data, poor measurements, feedback loops, or manipulation. It can also enter through human choices about what the system should optimize.

For example, if a company builds an AI system only to increase profit, it may ignore fairness. If a school builds an AI system only to save time, it may ignore creativity. If a police department builds an AI system only to predict arrests, it may ignore unequal policing.

Therefore, AI fairness is not only a technical problem. It is also a social, legal, and moral problem.

This does not mean AI should be rejected completely. AI can be useful. It can help doctors, teachers, scientists, and businesses. It can find patterns humans may miss. It can save time and improve services. But AI must be treated as a tool, not as a final judge.

The best approach is careful use. AI should be tested, explained, monitored, and challenged. People affected by AI decisions should have rights. Companies should be responsible for the tools they build. Governments should create clear rules. Schools should teach students how AI works.

11. Limitations of This Review

This review paper has some limitations.

First, it uses secondary research. It does not include a new survey, experiment, or interview.

Second, AI is changing very quickly. New tools, laws, and risks appear every year. Some examples in this paper may change over time.

Third, the paper explains AI bias in simple language. It does not include advanced mathematical fairness definitions or technical machine learning models.

Fourth, the paper focuses mainly on social bias and fairness. AI also has other risks, such as misinformation, privacy loss, copyright issues, security threats, and environmental costs.

Fifth, the paper cannot cover every example of AI bias. It focuses on common and understandable examples for high school readers.

12. Conclusion

Artificial Intelligence is becoming part of everyday life. It helps people search, shop, learn, work, communicate, and make decisions. But AI is not automatically fair. Because AI systems are created by humans and trained on human data, they can learn human bias.

This review paper has explained that AI bias can happen in many ways. It can come from biased training data, unbalanced data, poor measurement, feedback loops, or malicious manipulation. It can affect serious areas like hiring, housing, loans, healthcare, education, and policing.

The paper also shows that fixing AI bias is difficult. Some AI systems are hard to understand. Fairness can have different meanings. More data can help, but it can also create privacy problems. AI systems can change over time and may need constant monitoring.

Still, there are solutions. AI can become fairer through better data, bias testing, transparency, explainability, human oversight, regulation, and public awareness. Organizations such as NIST have already described fairness and harmful-bias management as part of trustworthy AI.

The most important lesson is that people should not blindly accept AI decisions. A computer can be fast and powerful, but it can still be wrong. AI should support human judgment, not replace human responsibility.

In the end, fair AI is not only about better technology. It is about better choices. If society wants AI to be fair, humans must design it, test it, question it, and use it with care.

References

Executive Office of the President. (2016). *Big data: A report on algorithmic systems, opportunity, and civil rights*.

European Council. (n.d.). *Artificial Intelligence Act*.

European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council*.

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework*.

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework 1.0*.

National Institute of Standards and Technology. (n.d.). *Trustworthy and responsible AI*.

Thirani, V. (n.d.). *Biases in AI: A review paper* [Uploaded student draft].