SCIENTIFIC Publishing Limited

Bio-Robotics

https://ojs.ukscip.com/index.php/br

Article

Bias and Fairness in Robotic Systems: Challenges and Solutions

Ailbeforie Nuikah

Department of Computing and Informatics, United States International University - Africa, Nairobi, Kenya

Received: 30 November 2024; **Revised:** 15 May 2025 **Accepted**: 29 May 2025; **Published**: 10 June 2025

Abstract: The robotic systems become more integrated into various aspects of life, including healthcare, criminal justice, labour, and autonomous vehicles, the issue of bias in these systems has emerged as a critical concern. The bias in the decisions taken by robots may enhance social disparities, enhance discrimination, and afflict vulnerable groups. This article addresses the biased properties and origin of robotic systems, including biased data, faulty system algorithms, and prejudice of the designers and developers in the course of the system. It also notes the grave social, ethical and legal impacts that discriminatory machines in specific areas may have, specifically in high-stakes areas. The article also proposes an array of remedies to curb the bias, such as a better data collection process, just algorithms, ethical design, and regulatory tools. Nevertheless, the article underlines the technical, economic, political, and societal issues that prevent the creation of really fair robotic systems. By stating the necessity of inter-disciplinary cooperation, the involvement of the public and the global regulatory systems, it confirms that, potentially, the robots developed will not only be efficient but also fair and just.

Keywords: Bias; Robotic Systems; Fairness; Ethical Design; AI Regulations

1. Introduction

Robot systems are gaining ground quickly and today are fast becoming an inseparable part of diverse industries such as healthcare, manufacturing, transportation, and even entertainment [1]. The systems incorporate industrial robots, autonomous vehicles, and medical diagnostic equipment staff and are built to supplement human performance, enhance efficiency and limit the impact of human error. The problem is, however, that as the robots are more sophisticated and self-regulated, they get involved in making decisions concerning the immediate lives of humans. The emergence of such technologies has raised a burning concern that is related to the possibility of bias in robotic systems that may negatively affect their utility, equity, and morality. In this paper, the issue of bias in robotic systems is discussed, including its origin, effects, and potential mitigation techniques to make human-robot interaction fair. Equity is at the centre of most arguments concerning artificial intelligence (AI) and robotics [2]. Bias in AI has been reported in the form of facial recognition features wrongly identifying people of colour and hiring algorithms that discriminate against women and minorities. This prejudice usually starts with the data that such systems are exposed to. In case these systems are trained on data that reflects historical bias or are not sufficiently diverse, the algorithms of these systems can learn and reflect that bias. This is of great concern in a robotic system, which is expected to make serious decisions. As a case in point, an autonomous vehicle could make prejudiced judgments in incidents where prejudiced data was used during the derivation of the decision-making models. On the same note, healthcare robots may deliver improper medical diagnoses in some populations, in case the data used during training does not cover those populations well enough [3].

Robotic systems are potentially biased, and their consequences in the context of ethics are severe. Given that robots are only as unbiased as the people who make them, to begin with, in high-stakes spheres, such as healthcare, criminal justice, and job hiring, biased robots may further develop already existing inequalities and

confirm discriminatory practices in the system. E.g., robots involved in healthcare diagnostics might perform poorly with some groups of the population, so they might issue incorrect diagnoses and cause unequal medical access to such groups. In the same way, biased criminal justice algorithms can unfairly affect specific groups by over-representing certain communities in terms of unfair sentencing or police governing. Consequences of bias in the system are even greater when robots assume more and more decision-making capabilities previously performed by humans [4,5].

The ethical aspect of biased robots is even aggravated by the lack of transparency in the working of such systems. The majority of robots and, especially, those backed by AI, are considered to act as a black box since people have little insight into how one or another decision is made. This is the absence of transparency that makes it hard to detect, know or work on the biases that might be ingrained in such systems. Once the decision-making processes of robots cannot be understood and made accountable, it is more difficult to be concerned that they act fairly and ethically. This is more worrying since robots are becoming more and more utilised in areas that require fairness and accountability, such as sensitive areas [6].

In the looks ahead, we have to question how these challenges can be worked around. The challenge of bias in robotic systems needs to be approached in a multi-dimensional way by way of enhancing data collection, developing non-biased algorithm design, and ensuring accountability in a manner that relates to making such systems less biased. The scholars, scientists, and /or policy-makers should collaborate to develop systems that will make robots behave in a morally acceptable and fair way. There also might be new regulatory steps needed so that robotic systems receive high standards of fairness, especially in situations where they are used in areas where people are, as they directly affect their lives [7].

In this paper, I intend to analyze bias and fairness in the robotic system, find the origin of bias in robots and how this can be treated unethically, and possible ways of resolving those concerns. The discussion will start by tracing the nature and causes of bias in robotic systems, followed by an evaluation of the social, ethical and legal implications emanating from the employment of biased robots. The paper will then look at some of the existing remedies to solve the problem of bias, which include refinement of the data collection process and creation of fairness-aware algorithms, and will outline the problems and drawbacks of these methods. Lastly, the paper will conclude with suggestions of future research and areas of future interdisciplinary collaboration to ensure that robotic systems are designed and designed in a fashion that encourages fairness and equity concerning everyone.

Through bias-reduction in robotics, there is the possibility of building a tomorrow where the robot is a tool that complements and supplements you as a human being without encouraging any form of discrimination or inequality. Robotics fairness does not only require technical measures, but it is also a matter of ethics, which will define the role of this technology in our society. As the sphere of robotic technologies is still developing, it is critical to develop the frameworks that are based on the notions of fairness, transparency, and accountability to make them eventually trustworthy and decent to all people [8-11].

2. The Nature and Sources of Bias

Biased robots can also be referred to as unwarranted discrimination or preference given to all robotic systems that are skewed in favour of some people or groups and against others. The first part of this section is dedicated to the investigation of the type of bias in the context of robotics, where the definition of what bias in an AI and robotic system is has been established before proceeding with the examination of its main causes.

2.1. Bias as a term in AI and Robotics

Bias in robotics and AI is any systematic error or bias that leads to the unfair treatment of groups, people, or categories, with a semblance of the inequalities and failures that society already holds. When it comes to robotic systems, bias may take many forms. These are mainly:

2.1.1. Algorithmic Bias

Algorithmic Bias refers to those cases in which the algorithms motivating robots or AI solutions create a decision that is biased towards one population compared to another one. The introduction of algorithmic bias may be because of how the system code or model is built and practices biased power towards one type of demographic group (e.g., gender, race, or age) over other demographic groups.

2.1.2. Data Bias

Bias in the robot systems is usually a result of the information which was used in training the systems. When the source of data used to train machine learning models is biased, flawed, unrepresentative, and/or faulty, then the

bias and faults created will be borne by the developed model. An example would be when a robot is trained using data that largely represents the common demographic; then this robot can experience difficulty when trying to treat or handle other people's groups fairly, or it may experience difficulty when trying to make dates on behalf of individuals of other groups.

2.1.3 Human Bias

This is an instance of bias subject to the implicit or explicit prejudices of the developers or the designers of the robotic systems. There is a possibility of Human prejudice being carried unintentionally at the stage of designing and programming, where the creator may unconsciously contribute their prejudices into the robots they design.

It is also important to comprehend bias in robot systems because the impact of bias is serious, including unequal opportunities, unfair treatment, and social inequality [12,13].

2.2 Sources of Bias in Robotic Systems

2.2.1. Biased Data

Data that would train machine learning models is one of the main sources of bias in robotic systems. The saying of old, garbage in, garbage out should be taken as truth since biased data in a system will produce biased robotic decisions. The bias of data may come in many forms:

- **Historical Bias:** Numerous times, the data utilized to train robots is an expression of historic inequity and social bias. By way of illustration, a facial recognition technology, which is being used with great frequency in robotics identification and security, can be trained on a set of images that tend to be of lighter-skinned people. Consequently, such systems might not be good at identifying people of colour, and hence become racist. The records of criminal justice or health care systems can also be a reminder of the discrimination that has existed over a long period of time, thus confirming the biases in applying them in the robotic system.
- **Sampling Bias:** Sampling bias refers to a situation when the data employed to train a robot is not representative to reflect the actual population. In other words, in case a robot making medical diagnoses is trained on the data of middle-aged white men, the machine may fail to diagnose the conditions of women or the aged and individuals of other ethnic backgrounds. Such bias is particularly undesirable since it fails to take into consideration the variety with which real-life situations and environments are characterised.
- **Label Bias:** It occurs in an inaccurate or inconsistent labelling of the data. A case in point is in supervised which usually has data annotated with labels so that the machine may learn something. When there is an inaccurate or biased labelling of things (examples include labelling some job applicants as unqualified based on stereotypes), the robot will be trained on such biased labels, which it will then propagate [14,15].

2.2.2. Algorithmic Bias

Other than biased data, algorithm bias in the field of robotics can also be a source of bias. A machine learning model algorithm works on mathematical models that depend on the input data. But these models may be biased in themselves according to the way in which they are structured. The factors that can significantly contribute to the practice of algorithmic biases are:

- **Overfitting the Model**: In some cases, the design of the machine learning model is optimised to succeed in working with certain sets of training examples. Nevertheless, when a model is too specific to a given set of data, it might not generalize efficiently to new data, especially when such new data belongs to a different demographic group. Such overfitting may cause biased decisions during the implementation of the system in real life and diverse circumstances.
- **Bias in Feature Selection:** In most machine learning models, the people designing the inferring algorithm are left to decide what information in the data is most relevant and what should go into the decisions made by the robot. This means that in case of incomplete or biased assumptions regarding the choice of these features, the model can discriminate between some of the variables as being more important than others, thus making the results biased. As an example, under predictive policing, predictive methods would use the crime data (assuming that crime data disproportionately reflects some neighbourhoods) to predict crime and therefore disproportionately target those neighbourhoods, regardless of the comparative risk.

- **Black Box Algorithms:** Most contemporary machine learning algorithms, and especially deep learning algorithms, have become known as black boxes because their reasoning processes are difficult to interpret. This non-transparency has the potential to cover up inherent biases in the decision of the algorithm. Consequently, one cannot easily detect when a system is acting in dishonest ways and measures to rectify the situation are also hard to take [16,17].

2.2.3. Human Bias

Human entities, especially designers, developers and trainers of the robotic systems, might end up putting their own biases in the system. It may happen at most levels of robot development:

Implicit Bias On the developer side, developers have no idea that they incorporate their own cultural or societal bias into robotic systems. Such prejudices could be based on stereotyping based on gender, race, or any other demographic variable, informing the process of making decisions in designing the system or training a model. As an example, a healthcare robot created by a group of employees, most of whom are male, will not take into account the gender-specific issues in terms of health in the most appropriate way.

- **Development Teams Bias:** Diversity, or not, in development teams may also have an effect on the design and behaviour of robotic systems. A homogenous group would have stronger chances of developing computer systems that are relevant to the views, assumptions and demands of the team, and this might not capture the demands and views of other groups.
- **The Confirmation Bias:** Out of habit, the developers might be choosing or prioritising data that supports their prior conception or hypothesis. This has the potential to distort the process of development and create biased results. Suppose developers just test a robot in one environment or with one type of demographic group; they might not associate any bias in the interactions of the robot with other kinds of population [18].

2.3 Examples of Bias in Robotics

- **Healthcare Robots**: In healthcare, robotic systems used for diagnosis and treatment may exhibit bias if their training data is unbalanced. For example, an AI system designed to assist in diagnosing skin cancer may be less accurate when identifying cancers on dark skin due to a lack of diverse training data. This could lead to misdiagnoses or under-treatment of certain groups.
- **Autonomous Vehicles**: Autonomous vehicles, when trained on biased datasets, could make flawed decisions in emergency scenarios. For instance, if the data used to train self-driving cars predominantly represents light-skinned individuals, the vehicle may not detect dark-skinned pedestrians effectively, increasing the risk of accidents.
- **Hiring Algorithms**: In recruitment, AI systems that are designed to evaluate resumes and make hiring recommendations can also be biased. If the training data used by these algorithms reflects past hiring practices that favoured one gender or race over another, the robot may inadvertently reinforce those patterns and disadvantage underrepresented groups.

In summary, bias in robotic systems stems from various sources, including biased data, flawed algorithms, and human influence. Identifying and addressing these sources of bias is crucial to ensure that robotic systems operate fairly and equitably. Recognizing the implications of bias in these systems is the first step toward designing more ethical, inclusive, and transparent robots [19,20].

3. Consequences of Bias in Robotic Systems

Bias on robotic systems may bear adverse long-term effects, especially when such systems are implemented in vital sectors that have a bearing on human life. These effects might occur both on the personal (e.g., damage to a particular individual or community) and on the social level (e.g., supporting or exacerbating existing inequalities in society). Here, we discuss how bias affecting robotic systems might adversely affect them morally, as well as practically.

3.1. Societal impact/strengthening disparity

Among the most worrying outcomes of prejudice in robotic systems, there is the possibility of creating and even augmenting social disparities. Because robots and AI systems are increasingly introduced to fields of employment, policing, health care, etc., namely areas that carry important consequences to the lives of people, prejudiced robots could inherently amplify those biases in society without their intervention. These are some of the areas of utmost concern where social impact is being felt:

- **Healthcare:** In the healthcare sector, unfair robotic units may result in discriminatory treatment and

healthcare inequality. Take, e.g., the example of diagnostic robots trained mostly on white patients' data, where such a bot will likely be less diagnostic of people of colour. This may lead to errors of misdiagnosis, delayed therapies, or wrong suggestions, which end up disproportionately impacting those in marginalised groups. In case of insufficient testing of robotic systems with a variety of populations, this may increase pre-existing health inequities. This can cause health disparities in the resulting situation, and this means that there will be disparity in terms of access to healthcare, poor health outcomes and also marginalizing some people.

- **Criminal Justice:** In criminal justice, prejudice in robotic systems applied to predictive policing, sentencing or risk assessment may discriminate more than other racial or ethnic backgrounds. As an example, unless the predictive policing algorithms are trained using unbiased crime data, they can suggest an increased level of law enforcement in the historically over-policed neighbourhoods, which tend to be communities of colour. This has the effect of creating an over-policed and over-monitored status in such communities, further developing the issue of racial profiling and systemic injustice. In the same way, unfair algorithms during the sentencing process can negatively affect the verdict regarding parole or bail, as a disproportionately larger number of minority groups will suffer disproportionately. This will contribute to the existing inequalities of race in the criminal justice system.

Hiring and Employment. Gender, racial, or age discrimination may be strengthened by the use of biased robotic recruitment devices in the field of employment. To use one example, AI tools deployed to routinely filter resumes or score candidates may end up giving extra weight to attributes that are statistically associated with race, sex, or social levels, regardless of whether they have anything to do with job performance. This acts as a form of underrepresentation of women, minorities, and other identities groups in the workplace. Employment discrimination can also deepen already existing employment disparities, locking out those who have inheritable disadvantages and cementing the stereotyped notions about their abilities [21-23].

3.2. The Impact on Affected Groups or People

In addition to the overall social implications, unfair robot systems have real and damaging consequences on individuals and select groups. Such evils are experienced most severely by those who are most impacted by the unfair decision-making of robots themselves.

- **Discrimination in Service Provision:** When robots or Artificial Intelligence systems are used to offer services to people, e.g. in the fields of health, education, or social services, the realization of bias in these systems can be of great detriment to individuals. Such as a healthcare robot powered by AI would not be able to diagnose a particular condition in a woman or a person of colour, hence slowing down the much-needed treatments, creating worse health outcomes. Such a kind of injury not only damages the well-being of the person involved, but also worsens the perception of trusting the technology itself. People with such biases can lose the belief in the efficiency and fairness of robotic systems and will be more uncomfortable embracing or believing in new technologies in the future.

Biased robots may deprive people of opportunities, too, be it education, employment, or even the justice system. As an example, a recruitment system, based on AI, which automatically declines the application of the candidate of a specific race or gender, can help deny the employee with qualifications. Likewise, a robot employed in the learning environment, which is not as responsive to the requirements of students with particular cultural backgrounds, will affect their learning progress, thus reducing their opportunities in life.

- **Psychological and Emotional Impact:** Another impact that is regularly not considered is the psychological damage which biased robots will do their harm. People affected by repeated discrimination by the robotic system are likely to develop frustrations, loneliness and marginality. This can compromise their self-esteem, leading to a feeling of marginalisation in society. Moreover, the victims of such prejudicial limitations on their opportunities can face long-term mental health adverse effects, such as anxiety, depression, and stress.

3.3. Moral Dilemmas and Breaking of Trust

The problem is quite serious as regards ethical concerns presented by biased robotic systems when the latter take on more roles in decision-making. The fact that the robots can be enabled to make decisions that can cause greater harm to any particular group/individual due to the nature of their abilities is of great concern as far as equity, responsibility, and justice are concerned. The next three ethical cases cast doubt on our visions of equality and human rights altogether.

- **Moral Accountability:** On the one hand, one of the ethical dilemmas that could occur in cases where the action of a robot hurts someone is to establish accountability for others. Robots will, in most cases, operate and make decisions on their own will without any form of direct human control.
- **This raises an ethical question:** Where is the accountability in case one of the robots makes a biased judgment and consequently hurts a person? Should the designers of the robot be blamed for the ills of the designed robot? This is especially disturbing because the people who develop the systems either are not aware of the prejudices that exist in the system or, in the case of the former, at the time of realizing the biases, they prove hard to handle [24].

I- Loss of Trust in Technology: Prejudice in robotic systems may be implemented in places that are sensitive, like as healthcare or criminal justice; therefore, people may have no trust in some technologies. Trust is one of the most significant idioms of new technologies adoption and acceptance, and in case people believe that some robots can be prejudiced or even discriminatory, they are less likely to trust the robot. This mistrust can become overreached in the development of the technology and limit all the potential positive that such robotic systems can offer to society. Further, the unsuccessful confidence in robots may be more extensive and go beyond some specific systems, going to the general acceptance of AI and robots' technology, which is an even more gradual course of understanding their overall implementation.

- **Biased Robots can encourage unethical behaviours:** Being biased should favour certain groups and discriminate against the others; biased robots would increase the possibility of reinforcing unethical behaviours, i.e. racial profiling, gender discrimination, or classism. An example of such a case is where the jobs that humans were doing have been mechanized by machines using robots, whereupon the prejudice in the system can effectively perpetuate such acts on a larger scale [25]. Considering a law enforcement environment as an example, in the case of using biased predictive policing algorithms to allocate resources more by the historical record of criminal outcomes, they may cement historical patterns of racial disparities into arrest patterns. This poses us with an ethical dilemma: Should we do nothing when robots are automating processes that can cause inequalities to rise, even when those processes have been technically efficient or effective? [26].

3.4. Regulatory and Legal Impact

There are also legal and regulatory impacts of the biases in the robotic systems. Given the total integration of robots into key areas, a more significant consideration is the need to have legal structures to touch on if and how robots can be fair, accountable, and transparent.

- **Legal Liabilities**: Implementation of biased robotic systems can lead to legal repercussions for the corporations or organisations utilising those systems, especially in cases where people or groups of people are hurt. In case some discrimination occurred due to a biased hiring algorithm, a company may be taken to court because it broke anti-discrimination legislation. In the same way, in case of the autonomous vehicle comes to a biased conclusion, which results in an accident, liability and responsibility will be raised. In the absence of definite legal arrangements, it becomes hard to penalise the state of affairs and always think that the biased mechanisms are rectified.
- **Regulatory Standards:** Due to the immense ethical and legal implications of biased robotic systems the governments and international institutions will most probably have to create the standards of regulation of robotics. Such laws may prompt developers to introduce fairness checks, become more transparent about decision making, and make robots go through bias testing and scrutiny before deployment. Regulatory bodies can also require monitoring and updating of robotic systems to rectify the bias that arises as the system is operated, taking into consideration the need to apply continuous fairness in the lifecycle of the system that is in practice.

To sum up, the implications of biased robotic systems are disturbing and vast to the core, including social, moral, and legal aspects. Any possibility to strengthen inequality and negatively affect people, undermine trust, and sustain practices that are unethical needs to be addressed immediately. One should not address bias in robotics as a purely technical challenge, but as an issue of morals, as well, since it is important to make robotic systems equally safe, responsible, ethical, and useful to the entire population of society [27].

4. Addressing Bias: Solutions and Strategies

It can be regarded as a severe issue that the underthought and careless design of the robotic system may include

bias, yet with a deepened attention to the issues, partiality can be otherwise mitigated and even disposed. In this part, numerous ways and programs to combat bias in robotic systems are examined, and thus they should work equally and with equality. The solutions to be discussed here are the data-related solutions, algorithm-related interventions, ethical design, and legal/regulatory innovations.

4.1. Data Solutions

Any robotic system, especially one that applies the guidance of machine learning techniques, builds on the data used to train. The partiality of the data is also one of the major roots of bias in robotic systems, and enhancing the quality, along with fairness of the data, is very essential.

4.1.1. Various and Representative datasets

Challenge: One of the greatest causes of bias is a lack of representation in training sets. An illustrative example is that using a dataset trained on a facial recognition system that primarily includes light-skinned people may not recognize or accurately identify other dark-skinned individuals.

Solution: To avoid this, diversity and inclusivity of data collection strategies should be of priority. The datasets ought to capture different demographic categories such as age, gender, race, and ethnicity, among other factors, to make sure that the robotic system will not discriminate against any user. Robots are unlikely to be biased in their behaviour because the data employed is likely to reflect the complexity of the real world.

Example: Datasets For healthcare robots, the datasets are to have various representations of different ethnicities, sex and even health conditions. This way of doing things will guarantee that the system can properly diagnose or prescribe treatment to all patients and not only to those who are represented in most of the training data.

4.1.2. Balanced Sampling

Challenge: The data one may have to train on may not be proportionally represented. This is, in most cases, due to the overrepresentation of some group or activities in historical records.

Solution: An unbiased sampling method will allow the representation of the less-represented group so that they are well represented in the data collection. This can include sampling of minority groups or the use of synthetic data generation to generate missing data about underrepresented groups.

Examples Predictive policing: This could be the case where some neighbourhoods are excessively policed, and the models built using the data fed to the algorithms may overestimate crime in the neighbourhood. Data in different neighbourhoods should be used to balance the dataset, which will lead it's the variables towards reducing biases and avoiding over-policing in neighbourhoods.

4.1.3. Synthetic Data

Challenge: In other contexts, real-life data that is representative of all demographic groups can be challenging to acquire as a consequence of invasion of personal privacy or because a specific population is challenging to access.

Solution: Synthetic data generation is a potentially new tool to optimise bulkier and representative data. This engages algorithms to create data that approximates real-world situations, where actual information may be weak. Through this synthetic data, one can enhance the strength of machine learning models and guarantee that the system performs at a high level on every demographic.

Citation: Synthetic data may be used in self-driving vehicles to create driving scenarios rarely seen in other weather or light conditions, or with other pedestrians of diverse backgrounds [28-30].

4.2. Algorithmic Solutions

Algorithms can create bias or even reinforce it, regardless of how good the data is. It is, therefore, imperative that equity-sensitive algorithms need to be established and such systems must be made to run in a just manner towards various demographic elements.

4.2.1. Fairness-Conscious Algorithms

Challenge: Optimization of machine learning models usually aims to enhance the model accuracy or model efficiency, but not the fairness of the model predictions. This may lead to discrimination in case one of the groups is highly favoured as larger relative to the other group by the algorithm.

Solution: Algorithm solutions: Fairness-aware algorithms have fairness constraints imposed on their objective functions. The goal of these algorithms is to optimize the performance (e.g. accuracy) and fairness in this regard, the results will be similarly advantageous concerning different demographic groups. Such common fairness methods are:

Group fairness: Makes the algorithm predict equally humiliating.

Individual Fairness: Insists that similar people get treated similarly, irrespective of their affiliation to any group of people.

Equalized Odds: Guarantees that error rates are similar between various categories and that, therefore, the system will not overperform among certain categories and underperform among certain groups.

Example: In a hiring-based algorithm, a fairness limitation can be set to make the chance of an algorithm picking an individual candidate of a certain gender, race, etc., equal to another one to avoid situations when the algorithm prefers one demographic group over another.

4.2.2. Adversarial Debiasing

Challenge: Unless the constraints involving fairness are incorporated into the training of the conventional algorithms, these algorithms can also strengthen the biases, although the training stage of such algorithms can be governed by the constraints of fairness.

Solution: Adversarial debiasing. This method entails training a model to not only make the correct predictions but to reduce biases during training. The bias correction component of the training process in this technique makes the model learn to make accurate and fair predictions. This corrective element assists the system in developing ways to proceed with non-biased forecasts without limiting its effectiveness.

Example: Adversarial demography may be applied to adjust the predictions in credit scoring algorithms, such that the treatment of people with a different racial background would be performed in a non-discriminating way, without affecting the accuracy of the algorithm in predicting whether someone is creditworthy.

4.2. 3. Explainability and Transparency:

Challenge: A lot of machine learning models, particularly deep learning models, are opaque, so it is not easy to analyze how decisions are being taken, or why some predictions are being preferred over others. It is a serious problem because it presents a critical case of bias identification and treatment due to this black box.

Solution: Making the algorithm more transparent and explainable is an essential measure to reduce bias. Such approaches as explainable machine learning models and post-hoc explainability approaches can assist developers, regulators, and end-users in making decisions about why choices are made. Through transparency in the process of decision-making, biased behaviour can be easily detected and eliminated in the system.

Example: The healthcare robot that provides recommendations on possible treatments must provide a reason as to why it thinks that one patient should be subjected to the treatment rather than another. If the decision-making is transparent, it will be simple to identify and rectify the biased aspects with regard to race, age or gender [31].

4.3. Human-Centred Design and Ethical Design

Robotic systems must be developed as fairly and ethically as possible to avoid accidental damage. This implies that we have to make sure that various viewpoints are taken into the designing process, and the creation of robots should be designed in a way that suits the needs of all people without any bias.

4.3.1. Teams in Development

Challenge: Due to a lack of diversity in the engineering and development teams, there exists a risk of plotting biased system, which tends to absorb the views and suppositions of a predominant group.

Solution: Utilizing diversity in development teams is one way to minimize the chances of bias being coded into the robotic systems. Having a diverse team will make it more likely that the team will take more user experiences, backgrounds and needs into account when designing robots and AI systems. This diversity has the potential to introduce more accommodating goods and services that would serve the needs of every possible customer.

Example: Working with people of different cultural backgrounds in the designing of a social robot to be used in the elderly care, the design process is likely to contribute towards the making of the robot that is capable of taking into consideration the differences that exist in terms of cultural prescriptions on how elderly should be taken care of and how.

4.3.2. Ethical Design Frames

Challenge: The ethical aspect is usually not considered in the design-work of robot systems.

Solution: At all levels of robotic system development (including the initial design, implementation, and assessment), ethical frameworks of design must be employed. These structures must be characterized by fairness, equity and transparency and must have steps to measure and address bias at every stage of development.

Example: A tool such as the "Ethical AI Checklist" would assist the developers of self-driving cars to assess the chances that their algorithm would reproduce certain biases towards a certain group or would fail to communicate particular features of road safety, including disabled pedestrians [32].

4.4. Regulatory and Legal Actions

With the increasing popularity of robotic systems, there is an increasing demand to use regulatory laws and regulations as a mechanism to promote the design and implementation of robotic systems with a focus on fairness.

4.4.1. Developing Fairness Rules:

o Challenge: The lack of legal instructions can make the company focus on their efficiency rather than on fairness in case of releasing biased systems.

Solution: Nationalities and international organizations ought to prepare laws that lay down strong norms of equity in robotics and AI systems. These rules ought to demand openness in data collection and algorithmic decision-making, force fairness audits, and hold operators responsible when biased results are produced.

Case: The European Unionhttps://oneZero.medium.com/gdpr-as-a-model-regulating-ai-systems-e60c9e5a35a2 General Data Protection Regulation (GDPR) offers an example of how data privacy laws could be employed to control AI systems. The same knock on the doors of the law may be made to have robots answerable about the biased nature of the actions that they commit, and make the companies rectify the biases.

4.4.2. Impact assessments and Fairness Audits:

Challenge: No global framework exists to examine the bias in robots yet.

Solution: Fairness audits and impact assessment of new robotic systems ought to be applied by the organizations and businesses before their deployment. These measures would consider the risk of discrimination and how it impacts on diverse groups of demographics. There is a potential to make such an audit one of the requirements of the certification of AI systems by regulators.

Example: A fairness audit of an AI-enabled hiring tool might examine the possibility that the algorithm has been programmed to weed out applicants without their knowledge due to their race, or because they are a member of a particular gender. The audit would propose some measures to rectify various biases identified before the application of the system in real-life hiring.To sum up, overcoming the problem of bias in robotic systems demands an interdisciplinary approach, involving data methods, algorithmic advances, ethical design and legal structures. Once the strategies listed are utilised, it is possible to develop robotic systems that are fair, transparent, and accountable to make sure that these systems do not produce inequalities towards other users and, on the contrary, are beneficial to them all in equal measures [33].

5. Challenges and Limitations

The current promising solutions to overcome the issue of bias in robotics facilitate major challenges and limitations that render the problem of fairness rather complicated to manage, which is an ongoing process. These problems are technical, economic, political, and cultural and tend to make it more difficult to come up with effective changes to make sure that robots and AI systems are not only created and implemented in ways that are conducive to equality and justice. In this section, we will delve deeper into these challenges and give an understanding of the obstacles that researchers, engineers, and policymakers need to go through to ensure some level of mitigation of bias in the robotic systems.

5.1. Technical Challenges

5.1.1. Quantification and Definitions of Fairness:

Opportunity: A lack of a universal definition of the term fairness is one of the most basic issues in the context of the necessity to ensure that there is no bias in robotic systems. The concept of fairness is characterized as multi-dimensional and thus, highly dependent on the context, stakeholders and norms of society. What is perceived as fair to do in one area (e.g. hiring) may not be perceived to be fair to do in another area (e.g. healthcare).

Solution: Various fairness standards can be used; however, the choice of the going measure of fairness is not always easy to arrive at. A case in point, the equal opportunity approach and equal outcomes approach whereby everyone has an equal chance of having positive outcomes may contradict with equal outcomes where the chances of similar outcomes in each group are sought after. It is one of the important challenges to find a compromise between these various definitions of fairness and ensure the effectiveness of the system.

Example: In driverless cars, equity could mean that the decision-making algorithm of the vehicle is not associated with excessive risks to the safety of some groups (like pedestrians who represent a racial group). But it is not easy to balance this and not jeopardise other goals such as the safety of traffic or car performance.

5.1.2. Complexities of Real-Life Works

Challenge: Data usually found to drive robotic systems is irregular, incomplete and may be inconsistent. Datasets hardly model the reality of the complexity and variety in society. The example is that historical information utilized in machine learning algorithms tends to have a bias of the society, e.g., economic inequality, racial discrimination. Robots might be biased when they inherit these biases when receiving inadequately cleaned, balanced, or diversified training data.

Solution: Even though data balancing techniques and synthetic data creation may provide relief from this problem, the task of adequately sanitizing and preparing data is resource-intensive and difficult. The information should always be revised and restructured in order to adjust to new social standards and to avoid further establishment of biases. Additionally, models can be trained on a wide variety of datasets, but still fail to generalise to all real-world situations.

Example: In the medical field, the medical records used to train the diagnostic robot may be small or may not be representative of some population, especially in areas with low rates of healthcare accessibility. Consequently, the robots could not detect severe situations in these underserved communities, which would only worsen the problem of healthcare disparities.

5.1.3. Black box algorithms

Challenge: Most machine learning models, particularly deep learning systems, have been discussed in terms of being a black box, i.e. their decisions cannot be confidently interpreted. Although an algorithm that defines a robot may be fair, judging by a given measure, transparency is lacking to determine how a system reached a given decision. The result is that this opacity makes it very difficult to audit systems to ascertain bias, evaluate accountability or any major changes where necessary.

Solution: One way around this would be to develop more transparent and easier-to-interpret models. Explainable AI (XAI) is a methodology that focuses on helping to improve the interpretability of machine learning models through the disclosure of insights into decision-making. But sometimes the trade-off between model accuracy and interpretability can be a problem- more complex models are potentially more accurate, but less interpretable.

Illustration: An algorithm of risk assessment in criminal justice that was used to predict the possibility of recidivism might have predicted it very well, but it might be a challenge to understand it. In cases where the algorithm affects specific demographics, like Blacks or Hispanics, the real cause of the bias becomes difficult to establish due to a lack of transparency in the inner workings of the system [34].

5.2. The Economic and political barriers

5.2.1. No Motivations towards Fairness

Challenge: In most cases, companies and developers will be more concerned with their performance (in terms of efficiency, speed and cost-effectiveness) without any consideration of fairness. Economic motives underlying the robotic and AI work can focus on technology improvement and profitability, thus pushing the question of fairness to the back burner in terms of addressing the needs of the market. This is especially the case in a high-competition industry like finance, tech, and health.

Remedy: The need to come up with economic incentives regarding fairness is of the essence. Governments and other organizations can fund research and issues of fairness in robotics, or go so far as to set up policies that reward the development of less biased and fair systems. But this will mean that there needs to be a change of priorities, with long-term ethical effects being considered as well as short-term financial and market returns.

Example: A company that designs robotic hiring tools may be interested in decreasing the cost and improving the efficiency of the recruitment cycle, but in the absence of a reward for merit, they may overlook the possibility of discrimination against women or members of the minority [35].

5.2.2. Regulatory and Policy Issues

Challenge: This is not a simple task, as far as making up the regulations that will include bias in robotic systems. The policymakers have to balance between a fair and innovative policy-making, entirely taking into account the technicalities of robotics and AI. The regulatory environment of AI and robotics is perhaps at a very early stage in many jurisdictions, and it is hard to impose fairness criteria in them.

The solution: The government and any other international organizations should develop precise and binding rules that stipulate fairness, transparency, and accountability in the robotic mechanism. This involves providing the principles of utilizing AI in high-stakes areas, like the medical field, the criminal justice system, and even in workplaces, where bias can have severe implications. There should also be control measures that need to be constantly monitored and updated depending on changes in technology.

Example: The General Data Protection Regulation (GDPR) of the European Union can become one of the possible models of fairness in artificially intelligent systems since it serves to protect the rights of users and data. Like, standards may soon be applied to fairness and mitigation of bias standards to robotic systems [36].

5.2.3. Resource and Implementation Gap in the world

Challenge: The resources towards establishing and refining fairness in the use of robots differ greatly, especially between the developed countries and the developing countries. Whereas richer nations will be able to invest in advanced fairness audits, data diversity, and transparently designed algorithms, the poorer ones might lack the required infrastructure and resources to develop fairness in the applications they use.

Solution: This gap shall be bridged through the cooperation of countries, international funding, and assistance in the move to more just robotic designs in developing nations. International agencies, like the United Nations, may be able to assist in the establishment of international standards on fair use of AI and robotics and give resources to the developing countries to match the standards.

Example: Fairness in healthcare robots might not be accessible in nations where it is hard to access varied and high-standard information, since most AI models are trained on the data originating in developed locations. It should be worked out so that the AI development in every region of the world should take into consideration the needs of the local population and the cultural background, informing their context [37].

5.3. Societal and Cultural Barriers

5.3.1. Fairness Perceptions: Differences in Culture:

Difficulty: The meaning of fairness may differ greatly between cultures. What is fair in one cultural setting or one country is not necessarily what is fair in the other. As robotic systems grow to be used in different parts of the world, the developers have to take into consideration the cultural background of the system they are implementing so that their systems become just to all the users irrespective of their cultures.

The solution: It would be important to design robotic systems that exhibit cultural sensitivity and malleability. This may be realized by engaging the different aspects of culture in the design and test programs, and also by making the systems flexible to allow different concepts of fairness in various geographical areas and communities.

Example: A social robot that is being implemented in elderly care might require a behaviour change based on the cultural contexts of care and interaction. Depending on the culture, older adults might demand more independent living care, whereas in others, they might be used to a more direct kind of care. The robots have to be able to change their approach to be flexible according to the local values [38].

5.3.2. Resistance to Change and Trust Problems

Challenge: The implementation of robotic systems may present a resistance in society, as some people may feel that the system is unfair or discriminatory. Individuals will be fearful of robots making decisions that impact their lives, especially when these decisions seem to be biased or remain mysterious. This distrust may work against the adoption of robotic systems, especially in such areas as health, law enforcement, and education.

Solution: In order to mitigate this resistance, the developers need to make transparency, explainability and fairness the first design consideration. Communications with the population and the use of feedback in the development process are obligatory steps to build trust in robotic systems.

Example: When a robot involved in health care makes discriminatory treatment decisions, people would lose confidence in the system, with a possibility of switching to human physicians. Such concerns can be alleviated by building explainable and demonstrably fair building systems.

This section gives an understanding of the complexity of the solution in terms of the robotic systems presented by the challenges and limitations. Getting past bias in robotics is no simple task, from technical solutions preventing flaws in determining fairness to economic and political barriers, which promote efficiency at the cost of equity, there is a long way to go. Moreover, there are more impediments to the establishment of universally just systems by way of culture and societal divisions. Nevertheless, these issues can be overcome and more fair and open robotic systems can be achieved by further innovation, interdisciplinary collaboration and advocacy of more robust regulatory frameworks [39 40].

6. Conclusions

Bias in a robotic system is a highly important topic with major ethical, social, and technical implications. Concerns about biased-decision making are increasing as robots and AI are becoming more incorporated in the

healthcare sector, law enforcement and as sources of employment. Biased robots in their unchecked state can increase social disparities, deepen negative stereotypes, and even physically hurt people and groups. Nevertheless, as discussed in this article, there exists a need to address these issues and find the way to reduce bias in robotics, and it is possible.

By enhancing data collection practices and guaranteeing diversity and inclusivity, algorithmic innovations to focus on fairness, and by taking ethical issues into account, developers can produce robotic systems that will be more equitable and just. Some of the strategies that hold great potential in reducing bias in these systems are the use of fairness-aware algorithms, data diversification, the use of transparency and the use of an explainable AI models. Besides, it is important to promote diversity in development teams as well as integrate ethical models throughout the development process with the view of filtering the robots we develop so they can be programmed to benefit everyone without discrimination.

Although those solutions provide advanced opportunities, bias in robot systems cannot be overcome easily. Fairness is rather difficult to measure; it is hard to obtain unbiased information, and it is hard to implement the change because it is culturally and politically resisted. These obstacles, however, can be defeated by means of interdisciplinary co-operation, global laws, and continuous involvement of people. With the ongoing advances in technology, there is a need to have developers, policymakers, and ethicists collaborate in the effort to create solid standards and frameworks that will assist in developing and deploying robotic systems that are done in a responsible way. In the end, the purpose should not be to develop robots that are efficient in their actions, but they need to act ethically and in a socially responsible manner. The problem of bias in robotic systems cannot be ignored or sidelined; otherwise, we will face a future where technology will be used as a negative element of society, which will worsen the chances, encourage unfairness and inspire mistrust among every individual. That way, the potential of robotics can be achieved innovatively and fairly.

References

[1] Husainy A, Mangave S, Patil N. A review on robotics and automation in the 21st century: Shaping the future of manufacturing, healthcare, and service sectors. Asian Review of Mechanical Engineering. 2023 Dec 7;12(2):41-5.

[2] George AS, George AH. Riding the wave: an exploration of emerging technologies reshaping modern industry. Partners Universal International Innovation Journal. 2024 Feb 25;2(1):15-38.

[3] Okamura AM, Matarić MJ, Christensen HI. Medical and health-care robotics. IEEE Robotics & Automation Magazine. 2010 Sep 9;17(3):26-37.

[4] Lin P, Abney K, Bekey GA, editors. Robot ethics: the ethical and social implications of robotics. MIT press; 2014 Jan 10.

[5] Endsley MR. Supporting Human-AI Teams: Transparency, explainability, and situation awareness. Computers in Human Behavior. 2023 Mar 1;140:107574.

[6] Mensah GB. Artificial intelligence and ethics: a comprehensive review of bias mitigation, transparency, and accountability in AI Systems. Preprint, November. 2023 Nov;10(1).

[7] Parry J, Taylor R, Pettinger L, Glucksmann M. Confronting the challenges of work today: New horizons and perspectives. The Sociological Review. 2005 Dec;53(2_suppl):1-8.

[8] Liu Y, Chen W, Bai Y, Liang X, Li G, Gao W, Lin L. Aligning cyber space with physical world: A comprehensive survey on embodied ai. arXiv preprint arXiv:2407.06886. 2024 Jul 9.

[9] Howard A, Borenstein J. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. Science and engineering ethics. 2018 Oct;24(5):1521-36.

[10] Cheong J, Spitale M, Gunes H. Small but Fair! Fairness for Multimodal Human-Human and Robot-Human Mental Wellbeing Coaching. arXiv preprint arXiv:2407.01562. 2024 May 15.

[11] Malle BF. Integrating robot ethics and machine morality: the study and design of moral competence in

robots. Ethics and Information Technology. 2016 Dec;18:243-56.

[12] Londoño L, Hurtado JV, Hertz N, Kellmeyer P, Voeneky S, Valada A. Fairness and bias in robot learning. Proceedings of the IEEE. 2024 May 29.

[13] Scatiggio V. Tackling the issue of bias in artificial intelligence to design ai-driven fair and inclusive service systems. How human biases are breaching into ai algorithms, with severe impacts on individuals and societies, and what designers can do to face this phenomenon and change for the better.

[14] Brohan A, Brown N, Carbajal J, Chebotar Y, Dabis J, Finn C, Gopalakrishnan K, Hausman K, Herzog A, Hsu J, Ibarz J. Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817. 2022 Dec 13.

[15] Jimenez Schlegl P. Learning in autonomous and intelligent systems: Overview and biases from data sources.

[16] Ranasinghe NG. MULTI-AGENT VERBAL COMMUNICATION ENABLING THE EXECUTION OF MULTIPLE ACTIONS THROUGH A SINGLE INTERACTION FOR NEXT GENERATION OF HUMAN-ROBOT COLLABORATION.

[17] Gupta A, Murali A, Gandhi DP, Pinto L. Robot learning in homes: Improving generalization and reducing dataset bias. Advances in neural information processing systems. 2018;31.

[18] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR). 2021 Jul 13;54(6):1-35.

[19] Albahri AS, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, Albahri OS, Alamoodi AH, Bai J, Salhi A, Santamaría J. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Information Fusion. 2023 Aug 1;96:156-91.

[20] Habuza T, Navaz AN, Hashim F, Alnajjar F, Zaki N, Serhani MA, Statsenko Y. AI applications in robotics, diagnostic image analysis and precision medicine: Current limitations, future trends, guidelines on CAD systems for medicine. Informatics in Medicine Unlocked. 2021 Jan 1;24:100596.

[21] Scatiggio V. Tackling the issue of bias in artificial intelligence to design ai-driven fair and inclusive service systems. How human biases are breaching into ai algorithms, with severe impacts on individuals and societies, and what designers can do to face this phenomenon and change for the better.

[22] Datteri E. Predicting the long-term effects of human-robot interaction: A reflection on responsibility in medical robotics. Science and engineering ethics. 2013 Mar;19:139-60.

[23] Atanasoski N, Vora K. Surrogate humanity: Race, robots, and the politics of technological futures. Duke University Press; 2019 Feb 28.

[24] Howard A, Borenstein J. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. Science and engineering ethics. 2018 Oct;24(5):1521-36.

[25] Lin P, Abney K, Bekey GA, editors. Robot ethics: the ethical and social implications of robotics. MIT press; 2014 Jan 10.

[26] Howard A, Borenstein J. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. Science and engineering ethics. 2018 Oct;24(5):1521-36.

[27] Holder C, Khurana V, Harrison F, Jacobs L. Robotics and law: Key legal and regulatory implications of the robotics age (Part I of II). Computer law & security review. 2016 Jun 1;32(3):383-402.

[28] Agnew W. *AI Ethics and Critique for Robotics* (Doctoral dissertation, University of Washington).

[29] Zhai C, Wibowo S, Li LD. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. Smart Learning Environments. 2024 Jun 18;11(1):28.

[30] Murphy K, Di Ruggiero E, Upshur R, Willison DJ, Malhotra N, Cai JC, Malhotra N, Lui V, Gibson J. Artificial intelligence for good health: a scoping review of the ethics literature. BMC medical ethics. 2021 Dec;22:1-7.

[31] O'Brien N, Van Dael J, Clarke J, Gardner C, O'Shaughnessy J, Darzi A, Ghafur S. Addressing racial and ethnic inequities in data-driven health technologies.

[32] Majeed A, Hwang SO. Solving the Privacy-Equity Trade-off in Data Sharing By Using Homophily, Diversity, and t-closeness based Anonymity Algorithm. IEEE Access. 2024 Dec 3.

[33] O'Sullivan S, Nevejans N, Allen C, Blyth A, Leonard S, Pagallo U, Holzinger K, Holzinger A, Sajid MI, Ashrafian H. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. The international journal of medical robotics and computer assisted surgery. 2019 Feb;15(1):e1968.

[34] Londoño L, Hurtado JV, Hertz N, Kellmeyer P, Voeneky S, Valada A. Fairness and bias in robot learning. Proceedings of the IEEE. 2024 May 29.

[35] Korinek A. Integrating ethical values and economic value to steer progress in artificial intelligence. National Bureau of Economic Research; 2019 Aug 5.

[36] Howard A, Borenstein J. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. Science and engineering ethics. 2018 Oct;24(5):1521-36.

[37] Mohammadi A, Maghsoudi M. Bridging perspectives on artificial intelligence: a comparative analysis of hopes and concerns in developed and developing countries. AI & SOCIETY. 2025 Apr 7:1-22.

[38] Kim TY, Leung K. Forming and reacting to overall fairness: A cross-cultural comparison. Organizational Behavior and Human Decision Processes. 2007 Sep 1;104(1):83-95.

[39] Cappuccio ML, Galliott JC, Eyssel F, Lanteri A. Autonomous systems and technology resistance: new tools for monitoring acceptance, trust, and tolerance. International Journal of Social Robotics. 2024 Jun;16(6):1-25.

[40] Leenes R, Palmerini E, Koops BJ, Bertolini A, Salvini P, Lucivero F. Regulatory challenges of robotics: some guidelines for addressing legal and ethical issues. Law, Innovation and Technology. 2017 Jan 2;9(1):1-44.