frontiers in Robotics and AI

HYPOTHESIS AND THEORY
published: 24 March 2021
doi: 10.3389/frobt.2021.650325

From Learning to Relearning: A Framework for Diminishing Bias in Social Robot Navigation

Juana Valeria Hurtado[*†], *Laura Londoño*[†] *and Abhinav Valada*

Department of Computer Science, University of Freiburg, Freiburg im Breisgau, Germany

The exponentially increasing advances in robotics and machine learning are facilitating the transition of robots from being confined to controlled industrial spaces to performing novel everyday tasks in domestic and urban environments. In order to make the presence of robots safe as well as comfortable for humans, and to facilitate their acceptance in public environments, they are often equipped with social abilities for navigation and interaction. Socially compliant robot navigation is increasingly being learned from human observations or demonstrations. We argue that these techniques that typically aim to mimic human behavior do not guarantee fair behavior. As a consequence, social navigation models can replicate, promote, and amplify societal unfairness, such as discrimination and segregation. In this work, we investigate a framework for diminishing bias in social robot navigation models so that robots are equipped with the capability to plan as well as adapt their paths based on both physical and social demands. Our proposed framework consists of two components: *learning* which incorporates social context into the learning process to account for safety and comfort, and *relearning* to detect and correct potentially harmful outcomes before the onset. We provide both technological and societal analysis using three diverse case studies in different social scenarios of interaction. Moreover, we present ethical implications of deploying robots in social environments and propose potential solutions. Through this study, we highlight the importance and advocate for fairness in human-robot interactions in order to promote more equitable social relationships, roles, and dynamics and consequently positively influence our society.

Keywords: social robot navigation, robot learning, fairness-aware learning, algorithmic fairness, ethics, responsible innovation

OPEN ACCESS

Edited by:
Martim Brandão,
King's College London,
United Kingdom

Reviewed by:
Pablo Jiménez-Schlegl,
Consejo Superior de Investigaciones Científicas (CSIC), Spain
Helena Webb,
University of Oxford, United Kingdom

***Correspondence:**
Juana Valeria Hurtado
hurtadoj@cs.uni-freiburg.de

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *06 January 2021*
Accepted: *01 March 2021*
Published: *24 March 2021*

Citation:
Hurtado JV, Londoño L and Valada A (2021) From Learning to Relearning: A Framework for Diminishing Bias in Social Robot Navigation. Front. Robot. AI 8:650325. doi: 10.3389/frobt.2021.650325

1. INTRODUCTION

The last decade has brought numerous breakthroughs in the development of autonomous robots which is evident from the manufacturing and service industries. More interesting are the advances that are essential enablers of several innovative applications, such as robot-assisted surgery (Tewari et al., 2002), transportation (Thrun, 1995), environmental monitoring (Valada et al., 2012), planetary exploration (Toupet et al., 2020), and disaster relief (Mittal et al., 2019). Novel machine learning algorithms accompanied by the boost in computational capacity and availability of large annotated datasets have primarily fostered the progress in this field. Machine learning and reinforcement learning techniques enable robots to learn complex tasks directly from raw sensory

input. One such task of navigation has seen tremendous progress over the years. Robots today have the capability to autonomously plan paths to reach a certain location and even make decisions based on the scene dynamics, avoiding collisions with people and objects (Boniardi et al., 2016; Gaydashenko et al., 2018; Jamshidi et al., 2019; Hurtado et al., 2020). Advancing robot navigation abilities is crucial for robots to effectively operate in real-world environments.

Robot navigation is a complex task that requires a high degree of autonomy. For a robot to successfully navigate the real-world, it is essential to fulfill high accuracy, efficacy, and efficiency requirements. Additionally, it is critical to consider safety standards while developing robots that navigate around humans. To carry out this task, robots are equipped with sensors that allow them to perceive the environment and a path planning system that enables them to compute a feasible route to achieve the navigation goal. So far, mobile robots have been successfully employed in various applications, such as material transportation, patrolling, rescue operation, cleaning, guidance, warehouse automation, among others (Nolfi and Floreano, 2002; Poudel, 2013; Hasan et al., 2014; Bogue, 2016). This also elucidates that mobile robot applications are moving closer from the industry to everyday tasks in households, offices, and public spaces. Robot navigation models tailored to solely reach a goal location efficiently are insufficient in these spaces where robots cohabitate with humans. Other complex considerations, such as social context, norms, and conventions are essential to ensure that the presence and movements of robots are safe and comfortable. These additional considerations of sociability play an indispensable role in the acceptance of robots in human spaces. Nevertheless, modeling the social policies that represent humans is a challenging task. To better capture the social behavior of navigation, several learning approaches have been proposed with the goal of directly imitating human navigation or learning from demonstrations (Silver et al., 2010; Wittrock, 2010; Bicchi and Tamburrini, 2015; Khambhaita and Alami, 2020). With the aim of incorporating social context in learning algorithms, socially-aware robot navigation extends the traditional objective of reaching a certain location to also reflect social behavior in the decision making process (Kretzschmar et al., 2016). This can be achieved with learning methodologies based on social and cultural norms. These social characteristics can be incorporated into the learning process as social constraints (Wittrock, 2010; Bicchi and Tamburrini, 2015; Khambhaita and Alami, 2020) or via imitation and demonstrations (Silver et al., 2010). As the role of robots within society is that of a social agent, they should follow social conventions for better acceptability in human environments. Following such conventions will enable them to generate actions that are influenced by respecting personal spaces, perceiving emotions, gestures, and expressions (Luber et al., 2012; Ferrer et al., 2013; Kruse et al., 2013; Kretzschmar et al., 2016).

However, despite significant advances that enable incorporating social conventions into navigation models, there is still no guarantee that a socially-aware robot will always make fair decisions. We can extensively observe in other applications of machine learning and Artificial Intelligence (AI), how learning algorithms replicate, promote, amplify injustice, unequal roles in society, and many other societal as well as historical biases. Numerous cases have been identified in face recognition, gender classification, and natural language processing methods (Garcia, 2016; Buolamwini and Gebru, 2018; Benthall and Haynes, 2019; Costa-jussà, 2019; Wilson et al., 2019; Lu et al., 2020; Wang et al., 2020). Similar to these cases, learning social behavior from real-world observations will not prevent discrimination. This is of special concern in service and caregiving applications where robots physically interact with humans.

There are multiple social and technical factors that can lead to bias while learning social robot navigation models. First, learning techniques require guidance to optimize the navigation model. Supervised approaches utilize datasets gathered from simulations, controlled experiments, or the real-world. Other approaches, such as imitation learning and reinforcement learning, obtain guidance directly from real experiences. It is important to consider that real-world data can always include bias reflecting unwanted humans behaviors. Additionally, simulations and controlled experiments cannot contain sufficient diverse information about diverse groups of people and their interactions for the robot to learn the large number of potential unfair situations that it can encounter. Therefore, current learning algorithms can significantly replicate, promote, and amplify unfair situations. Besides data-related issues, learning algorithms tend to find certain features that make it easier to optimize for a task and rely on these attributes to learn the function or policy. This can lead to mechanisms that depend on these potential bias inducing features related to a particular characteristic, such as race, age, or gender. Another issue encompasses fairness measurements. Thus far, there are no standard fairness definitions or metrics for the optimization of learning-based navigation algorithms or even to detect biased or unfair situations. Furthermore, robots are typically deployed with models that have been pre-trained and do not have the ability to automatically update their parameters or their policy online if they encounter a discrimination scenario.

Recently, several strategies to mitigate unfair outcomes in learning algorithms for tasks, such as classification or recognition have been proposed (Woodworth et al., 2017; Zafar et al., 2017; Agarwal et al., 2018; Dixon et al., 2018). Nevertheless, learning fair social navigation models for robotics is substantially lesser studied. Particularly, investigating fairness in mobile robot navigation presents more complex challenges that are not manifested in other data-driven tasks in computer vision and machine learning. In learning-based mobile robot navigation, fairness behavior not only depends on data but also on the future actions of the humans around the robot and other factors of the environment. In this case, it is impractical to anticipate all the possible actions in advance during the development of these models. With these considerations in mind, socially-aware robot navigation, besides learning social skills, should also account for non-discriminatory and fair behavior that makes the interaction safer for diverse groups of people.

In the case of humans, the learning process is not fixed but rather continuous. This allows humans to have both physical

and social adaptability. We refer to this adaptive learning from experiences as relearning in this work. We, as humans, not only relearn about the physical world to react to unexpected obstacles in our path, but we also develop adaptability in terms of interaction. This generally prevents us from causing harm to others with our actions and enables us to correct our behavior when we encounter unfair situations. Within this social adaptation, we learn to behave socially and fairly with those with whom we relate to (Goodwin, 2000; Hutchins, 2006; McDonald et al., 2008). The relearning process allows us to reason about what we are experiencing and develop a personality defined by certain moral values, ethical values, beliefs, and ideologies, which in turn influences the way we interact with others (Jarvis, 2006). Humans decide how to navigate in public spaces while taking both social conventions and ethical aspects into account, such as empathy, solidarity, recognition, respect for people, and recognizing behaviors that lead to discrimination. Accordingly, learning and relearning are important processes for humans to acquire the capabilities that are required for navigating in the environment and cohabitate in society.

Inspired by the learning and relearning processes in humans, we propose a framework for diminishing bias in social robot navigation. Our framework consists of two components. During robot development, we introduce social context based on social norms and skills while learning navigation models so that the robot acquires social conventions. We then incorporate a relearning mechanism that detects systematic bias in control decisions made by the robot during navigation. This enables the robot to update its navigation model when unfair situations are detected during the operation. Our proposed framework facilitates diminishing bias in the behavior of the robot and generates early warnings of discrimination after the deployment. More importantly, it enables the adaptation of the robot's navigation model to new cultural and social conditions that are not considered during training.

In this work, we describe the motivation and the technical approach for implementing our proposed Learning-Relearning framework for social robot navigation. We then highlight the risks and propose potential solutions that include specific fairness considerations for mobile robots that navigate in social environments. Furthermore, we analyze the ethical and societal implications of deploying mobile robots in social environments. To this end, we investigate the behavior of mobile robots in terms of fairness in three specific service and caregiving scenarios with different levels of human-robot interaction. There are other social scenarios where the mobility of the robot directly depends on the human's control action, such as autonomous wheelchairs (Johnson and Kuipers, 2018) or robotic guide canes (Ulrich and Borenstein, 2001). Nevertheless, in this work, we only consider scenarios where the robot navigates as an independent machine that interacts with multiple humans in the surrounding environment at different levels of priority. We provide examples that show cases where models that are only based on learning social navigation are insufficient to obtain fair behavior, and we discuss how the relearning mechanism can extend those models to yield fair behavior. Finally, we analyze scenarios in which learning social behavior and accounting for fair behavior play an important role in the real-world.

To the best of our knowledge, this is the first work to investigate the societal implications of bias in learned socially-aware robot navigation models, and the framework that we present is the first to demonstrate a feasible solution for learning fair socially compliant robot navigation models. Even though our work targets socially-aware robot navigation, the framework that we propose can also be extended to other aspects of human-robot interaction, which would benefit from the presented insights. As a result of the social perspective, we provide a comprehensive understanding of fairness in human-robot interactions. This is an important step toward diminishing bias and amplifying healthy social conventions to positively influence the society. With this work, we aim to create awareness that robots should positively impact society and should never cause harm, especially against individuals or groups who have been historically marginalized and who disproportionately suffer the unwanted consequences of algorithmic bias.

In summary, the primary contributions of this paper are:

- We introduce a framework for diminishing bias in social robot navigation, consisting of two stages: Learning and Relearning. We present the technical concept and introduce methods that can be used to implement our framework.
- We present a societal and technical analysis of the social abilities and bias considerations in learning robot navigation models.
- We present the social implications of socially-aware robot navigation models and provide a set of fairness considerations.
- We provide detailed case studies that analyze the impact of bias in different service and caregiving robot applications and discuss mitigation strategies.

2. ETHICAL ASPECTS AND FAIRNESS IMPLICATIONS

The growing impact that AI and robotics have in the daily lives of people has led to the increase in ethical discussions about current machine learning algorithms and how to handle new research toward an equal and positive impact of technology for diverse groups of people. Consequently, recent works in both social sciences and machine learning have highlighted the challenges in socio-cultural structures that are reflected and amplified by learning algorithms. As a result, many guidelines from the technical (Cath, 2018; Silberg and Manyika, 2019; Hagendorff, 2020a; Piano, 2020) and social perspectives (Verbeek, 2008; Liu and Zawieska, 2017; Birhane and Cummins, 2019) have been presented. These guidelines (Vayena et al., 2018; Hagendorff, 2020b; Piano, 2020) are aimed toward mitigating the adverse effects and advocating for ethical principles, such as fairness, trust, privacy, liability, data management, transparency, equality, justice, truth, and welfare. Similar efforts have been made by the European Robotics Research Network (Euronet) in the Euronet Roboethics Atelier project in 2005, and the British Standards Institute which published the World's First Standard on Ethical Guidelines in 2016 (Torresen, 2018). Moreover,

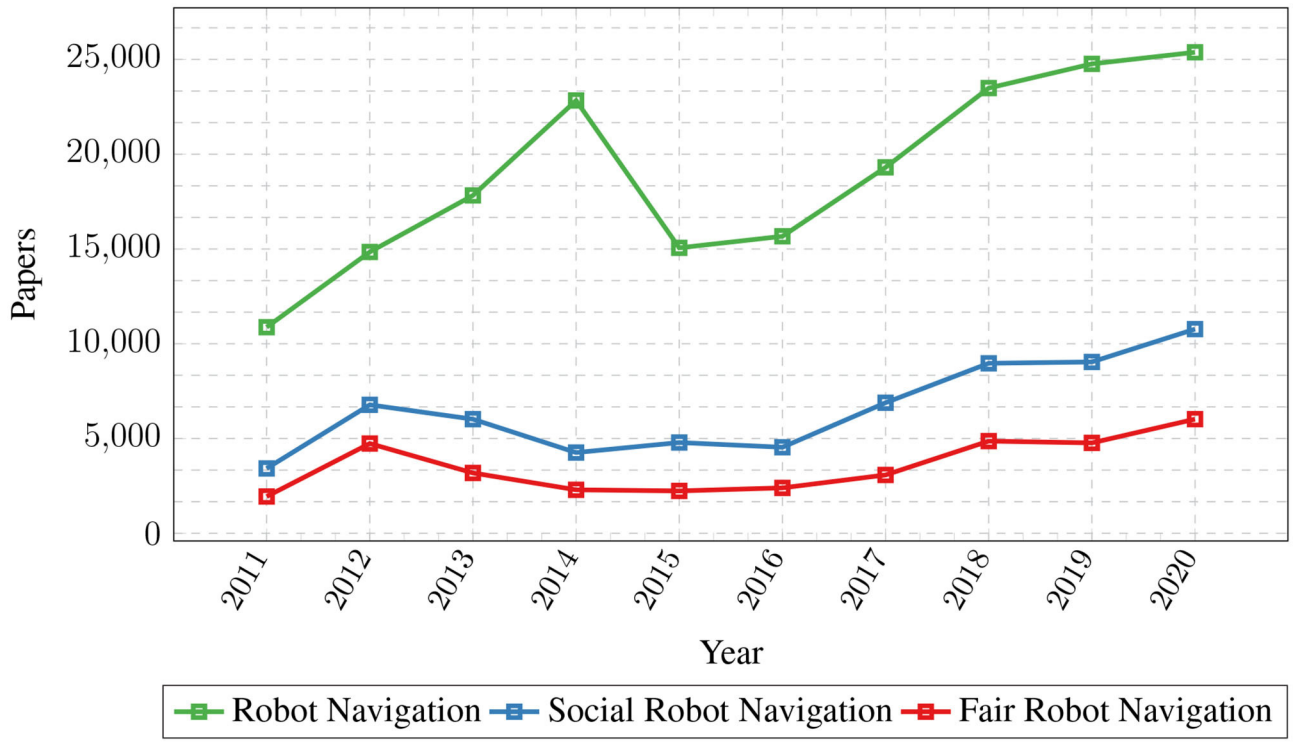

FIGURE 1 | Comparison of the number of publications on Robot Navigation (blue), Social Robot Navigation (red), and Fair Robot Navigation (green) from 2011 to 2020. Although the rate at which fairness is being considered in robot navigation methods is increasing, there is a growing gap with the number of works that address robot navigation each year.

some works in robotics (Anderson and Anderson, 2010; Lin et al., 2012; BSI-2016, 2016; Boden et al., 2017) have also investigated the importance of addressing ethical issues for safe and responsible development.

These ethical guidelines (Reed et al., 2016; Goodman and Flaxman, 2017; Johnson et al., 2019; Arrieta et al., 2020) share the value of robots effectively and safely assisting people, and under no circumstance cause harm or endanger their physical integrity (De Santis et al., 2008; Riek and Howard, 2014; Vandemeulebroucke et al., 2020). The impact of human-robot interactions has also been studied to a lesser extent in mobile robotics, e.g., providing recommendations on road safety, privacy, fairness, explainability, and responsibility (Bonnefon et al., 2020), or studying fairness in path planning algorithms of robots during emergency situations (Brandão et al., 2020). Similarly, such ethical discussions should be contrived while developing socially-aware robot navigation models. As shown in **Figure 1**, although the number of publications that consider fairness in robot navigation is slowly increasing, it is still over five-times lesser than the overall number of publications that address robot navigation. In this section, we present a series of ethical aspects and social implications that can arise from bias in socially aware-robot navigation algorithms. Additionally, we analyze the impact that these social navigation algorithms can have in human environments.

2.1. Fairness Implications

The cultural and social knowledge in humans is transferred from generations as a cumulative inheritance that allows each member of the society to incorporate moral, political, economic, and social structures that not only have a positive but also a negative value (Castro and Toro, 2004). These inheritance conditions have perpetuated historical discrimination against individuals and groups of people. The data collected in machine learning and AI come from these historical inheritance structures; consequently, social-historical discrimination can also be reflected or even amplified by learning algorithms. In recent years, several unexpected outcomes have been observed in learning algorithms that have caused discrimination and prejudice in society. Numerous examples demonstrate how social prejudices are reflected in machine learning algorithms (Garcia, 2016; Wang et al., 2020). One clear example that was observed in natural language processing was the racial and gender biases while learning language from text (Costa-jussà, 2019; Lu et al., 2020). Another recent example is the automated risk assessments used by U.S. judges to determine bail and sentencing limits. It was shown that it can generate incorrect conclusions, resulting in large cumulative effects on certain groups, such as longer prison sentences or higher bails imposed on darker-skinned users (Benthall and Haynes, 2019). Moreover, another study shows how biased algorithms affect the performance of vision-based object detectors employed in autonomous vehicles. Their

work demonstrates that pedestrians with dark-skinned tones presented higher recognition errors (Wilson et al., 2019). There have also been numerous cases of algorithmic bias that have been observed in algorithms used in healthcare. For example, algorithms trained with gender-imbalanced data have shown higher error at reading chest x-rays for an underrepresented gender (Kaushal et al., 2020).

The numerous cases of discrimination observed in learning algorithms employed in various applications are a source of concern for robotics. In the case of robots that employ learning algorithms to effectively interact, navigate and assist people, it is essential to foresee possible unfair situations. Specifically, as a result of learning socially-aware robot navigation strategies, these trained models can enhance the social impact in terms of human acceptance of mobile robots, daily use, comfort, security, protection, and cooperation (Thrun et al., 2000). Providing robots with a more natural navigation ability also increases their usability. Although incorporating social navigation models in robots improves their usability, comfort, and safety in human spaces, social abilities by themselves do not ensure fair robot decisions, especially while using learning algorithms to imitate or follow human conventions and behaviors. In human social interactions, a series of direct and indirect discrimination behaviors and decisions are often present (Forshaw and Pilgerstorfer, 2008; Zhang et al., 2016; Yu, 2019). Using learning algorithms can negatively affect society, individuals, or groups if unwanted social behavior is replicated and reflected in the actions of the robot. Therefore, this highlights the need to implement fairness considerations and measures. The ability of an agent to dynamically make fair decisions among different people is a fundamental basis for trust in human-robot interaction (Ötting et al., 2017; Claure et al., 2019). If robots after their deployment present an unfair behavior, it will continue to perpetuate discriminatory structures that will be reflected in the way that people are assisted. Moreover, this will cause serious consequences, such as a large population not being benefited by the robots and being reticent to use them. These factors suggest that the robot would only be beneficial for certain groups of people, which would continue to reinforce large social inequalities. Robots should influence society in a positive way by promoting healthier relationships, roles, and dynamics after their deployment in different places with diverse people. This requires the creation of a more reflective, equitable, and inclusive learning methods accompanied by extensive studies from the social perspective.

2.2. Fairness Measures

Fairness is a complex ethical principle that relates to avoiding any form of systematic discrimination against certain individuals or groups of individuals based on the use of particular attributes, such as race, sexual orientation, gender, disability, socioeconomic, and sociodemographic position (Silberg and Manyika, 2019). However, the definition of fairness tends to be dynamic, mobile, and contingent, therefore it should be analyzed from a reflective and ethical perspective. Moreover, fairness highly depends on the context, location, and culture, among other factors. Consequently, defining an accurate fairness measure could be a complex task. With efforts in this direction, bias has been used to represent fairness either in human environments or in technological developments (Howard et al., 2017; Fuchs, 2018; Lee, 2018; Nelson, 2019).

For its part, solutions to algorithmic bias that perpetuate social and historical discrimination against vulnerable and disadvantaged individuals or groups of people tend to be technical rather than moral and ethical (Birhane and Cummins, 2019). Technological solutions to biased decisions making are essential but not solely sufficient. Instead, technical solutions should be accompanied by factors, such as diversity, inclusion, and participation of underrepresented groups during the development of navigation models. Although there is no standard definition of fairness in machine learning and AI, some works state that a prediction is fair when it is not discriminating or when there is no bias (Binns, 2018; Chouldechova and Roth, 2018; Birhane and Cummins, 2019). However, there are two types of biases, positive and negative. Positive bias frequently promotes social good and avoids prejudice through awareness and respect for human differences. Therefore, not all biased outputs are necessarily undesirable and eliminating them can cause unintended outcomes for certain people. For example, consider an algorithm that is used in a bank to perform a credit study of the people who apply for a loan. If the algorithm is trained to guarantee that all the people will have credit, this may be a disadvantage in the long run for those who cannot pay back later. While the algorithm is being equal in this case, it is being unfair in the long term as it negatively affects the low-income people (Silberg and Manyika, 2019).

In socially-aware robot navigation fairness measurements are yet to be studied. As robots interact and assist different groups of people in different settings, creating a unified definition or a metric is impractical due to the complex and diverse cases that robots can encounter after deployment. Accordingly, in order to tackle unfairness, we present a series of fairness considerations for socially-aware robot navigation:

(i) **Value Alignment** refers to the alignment of human values in decision making during navigation. These values include respect, inclusion, empathy, solidarity, recognition, and non-discrimination. In socially-aware robot navigation, it is reflected in cases when the decision-making of the robot reproduces and increases the welfare of vulnerable populations. For example, prioritizing to assist and serve people with physical disabilities in crowded environments.

(ii) **Bias Evaluation** is related to the evaluation of bias in decisions making during navigation. Bias can be considered acceptable if there is adequate reasoning or unacceptable if the bias replicates, promotes, or amplifies discrimination. For example, when robots navigate with a different speed around young people who are faster than around older adults, it is usually accepted because they have important physical differences. Nevertheless, if such decisions are made based on racial differences, it can be considered unacceptable, given that there are no fair reasons for this difference. With this fairness consideration, when biases are presented in navigation models, it can only be accepted if there are fair reasons for doing so.

(iii) **Deterrence** is expressed in preventing and mitigating unwanted bias as well as discrimination during navigation. Since the notion of deterrence is dynamic and can vary depending on the social context, robots should be sensitive to cultures by adapting to people, customs, and their surroundings.

(iv) **Non-maleficence** signifies that the decisions of a robot can never produce damage to people. The damage is primarily interpreted as bodily harm, collisions, interruptions, delay, and obtrusion. However, damage can also refer to the negative effects caused by discrimination, segregation and bias. For example, if a caregiving robot in a hospital becomes an obstacle to the medical personnel responding to an emergency due to biased decisions, then it would be violating this property.

(v) **Shared Benefit** refers to providing equal benefits to diverse people in all scenarios. If a robot is specifically designed for and only tested in a particular geographical area, tailored to the characteristics and behaviors of the people in that region, it can lead to unwanted bias when it is deployed in a new region which may have completely different characteristics. Therefore, the benefits that the robot provides should not be targeted toward people with specific characteristics in a determined geographical area, but should rather be equally beneficial to all users. In this case, adaptability is an important attribute for robots to achieve shared benefit so that the autonomy of the robot is flexible to adapt to characteristics of specific users in the social environment where it is deployed.

2.3. Responsible Innovation

Research in technology studies suggests that the conceptions of responsibility should build upon the understanding that science and technology are not only technically but also socially and politically constituted (Winner, 1978; Grunwald, 2011). Responsible Innovation (RI) was introduced as a concept to address the impact of research and innovation in technology from an ethical and fair perspective. RI states that the technology should be anticipatory, so it should have a foresight guide that provides alternative options for responsible development (Stilgoe et al., 2013; Brandão et al., 2020), and it should account for social, ethical, and environmental issues. Based on RI principles, the framework that we present in this paper aims to identify biased behavior during navigation and promotes fair decision making through the learning and re-learning process to enable flexible and adaptive service. RI articulates and integrates four factors: (i) anticipation of damages, (ii) reflection from an ethical perspective, (iii) protection of sensitive human characteristics, such as age, gender, and race, and (iv) responsiveness (Stilgoe et al., 2013).

With the aforementioned RI factors, responsible robotics aims to ensure that responsible practices are carefully accounted for within each stage of design, development, and deployment. Correspondingly, robot navigation models should address the ethical and legal considerations at the time of development. Given that these considerations are constantly changing depending on the social or cultural factors, these models should be updated accordingly.

3. LEARNING—RELEARNING FRAMEWORK FOR SOCIALLY-AWARE ROBOT NAVIGATION

The goal of our proposed framework is to develop learning models for robot navigation that yield-social and fair behavior. To this end, we define two different stages: learning and relearning. In the first stage, we incorporate social context into learning navigation strategies so that robots can navigate in a socially compliant manner. While, in the second stage, we aim to diminish any bias in the planned paths with the learned navigation model. In this section, we first introduce socially-aware robot navigation. We then describe our proposed framework and present the technical approach that can be used for the implementation. **Figure 2** shows the different stages of our framework. In the learning phase, we learn a navigation policy based on imitation learning with additional social constraints. Whereas, in the relearning phase, we analyze the outputs of the network online and provide the model with updates to reach the navigation target while accounting for and deterring bias to ensure fairness. Science and technology, from the RI perspective, have the ability to provide significant benefit through well-established methodologies that reflect responsibility and ethical principles. This framework tailored exploits the learning and re-learning process as a methodology to achieve responsible robot navigation.

3.1. Socially-Aware Robot Navigation

One of the widely studied requirements for mobile robots to operate in human spaces is the ability to navigate according to social norms and socially compliant behavior. The social navigation models that are employed in robots play an important role in the effect that these automated machines have on society and the perception as well as confidence that humans will have of them. In the case of humans, we develop the ability to navigate while considering numerous variables representing the environment, such as the objects, people, and dynamics of the agents in it. This ability, known as sociability, from an anthropological point of view, is the human capacity to cooperate and engage in joint behavior with others (Simmel, 1949). Further, sociability allows us to navigate while avoiding situations that make us uncomfortable or put us or others in danger.

Different social norms have been developed to provide information about the appropriate behavior, especially in public spaces. Social norms are standards of conduct based on widely shared beliefs of how people should behave in a given situation (Fehr and Fischbacher, 2004). Some of the social norms for navigation are not invading the personal space of people, passing on the right, maintaining a safe velocity, not blocking peoples path, approaching people from the front, among others (Kirby, 2010). Besides social norms, different studies, such as proxemics (Hall et al., 1968), kinesics (Birdwhistell, 2010), and gaze (Argyle et al., 1994) also provide cues to determine the

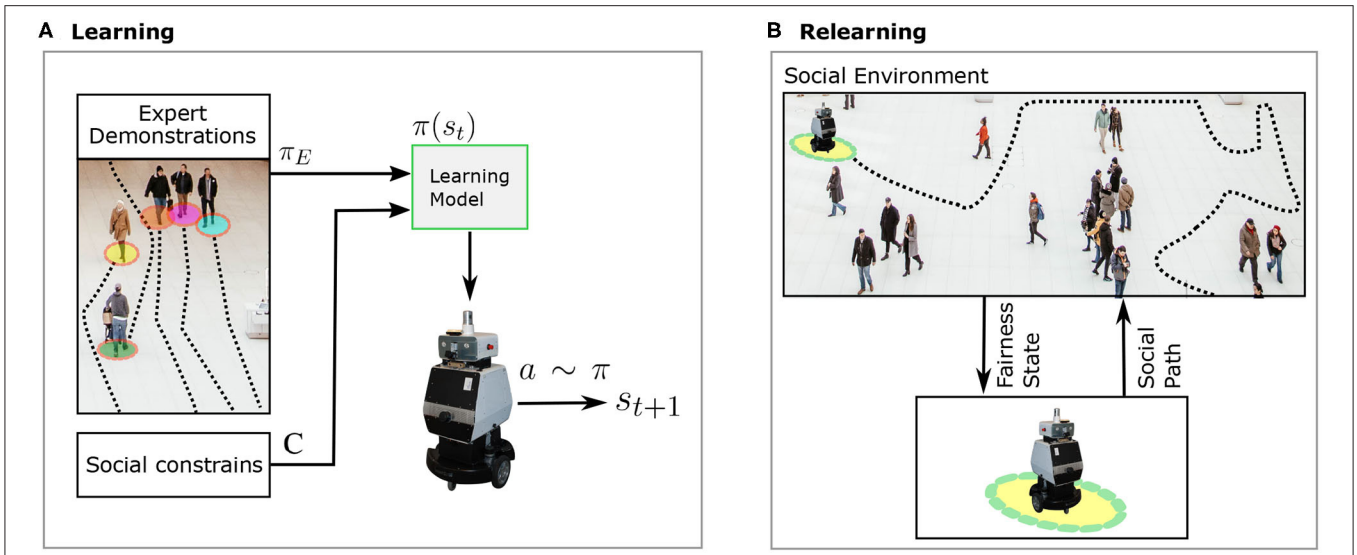

FIGURE 2 | Illustration of our proposed Learning-Relearning framework for diminishing bias in social robot navigation. Our proposed framework consists of two components: learning **(A)** and relearning **(B)**. By including the social context in the learning process, we aim to account for safety and comfort. The social context is presented as the social skills demonstrated by experts and social norms as constraints. Moreover, we aim to detect potentially harmful outcomes before the onset using the relearning mechanism. After detecting unfair effects, the navigation model should be automatically updated to account for fairness.

appropriate manner to approach a person, navigate around, and coordinate in public spaces. Specifically, proxemics is the study of the perception and organization of the personal and interpersonal space. It is associated with the manner of how humans manage their surrounding space when they walk in public environments and how their comfort can be affected by the movement of other pedestrians (Rios-Martinez et al., 2015). Kinesics is related to the actions of the body and positions (Birdwhistell, 1952); and gaze refers to the eye movements and directions during visual interaction (Harrigan, 2005). These studies highlight social skills, such as reading emotions and the prediction of intentions of people. The combination of both social norms and social skills can be considered determinant to sociability. The aforementioned studies and norms are some of the increasingly used factors in learning social robot navigation models. It is long believed that equipping robots with these social skills and social norms will enable them to react socially as humans do.

For instance, we can anticipate that cleaning robots (Fiorini and Prassler, 2000) that are primarily used in houses will be widely used in public spaces in the coming years. Currently, these robots do not conform to any social norms during navigation. Confined to private locations and users who know the device, manufacturers have not made it a priority to include social skills, such as predicting the intention of people and avoiding crashing into them. Nevertheless, sociability is an important skill to deploy cleaning robots in crowded public spaces. In this case, robots must take into account aspects, such as the space that they occupy and the personal space of the people around to determine how close to navigate around them or predict where humans will move so that they do not interfere with their paths. These skills will allow robots to plan a safe route so that their presence is not disturbing, surprising, or scaring the people that share the same space. While planning routes, robots should use social norms, such as not invading the personal space and maintaining a safe speed. Both the use of social skills and social norms change depending on the type of robot and the context in which it is used. We present further discussions of this example in section 4.1.

Socially-aware robot navigation methods can primarily be categorized into two groups. The first category is model-based and consists of handcrafted models that use mathematical formulations to combine a set of effects to determine dynamics of pedestrians, such as reaching the destination, the influence of other pedestrians, keeping a certain distance to another person or the maximal acceptable speed. Helbing and Molnar (1995) introduced the notion that social forces determine human motion and proposed the Social Force Model (SFM) to represent pedestrian dynamics. To navigate in a manner similar to humans, this formulation was later used to provide robots with pedestrian-like behavior for human-robot social interaction (Ferrer et al., 2017). However, SFM requires us to cautiously define and tune the parameters for each specific scenario, which makes it impractical to scale to complex tasks and environments (Tai et al., 2018). The second category consists of learning-based methodologies that use some form of guidance or demonstrations containing the policies that link observations to the corresponding actions. We further discuss learning-based methods in the following section.

3.2. Learning

The rapid progress in machine learning in the past years and the growth of computing power have enhanced the learning capabilities of autonomous mobile robots. Currently, these learning-based methodologies play an essential role in the

development of complex navigation models. These models are primarily trained to achieve the best navigation performance under some given metrics during the learning process. For this purpose, different guidance techniques have gained interest in robot navigation works. The first of which is supervision from labeled data, which uses either data gathered from the real-world or simulations and the corresponding annotations. The data and annotations are then employed to optimize the model so that the output predictions are as close as possible to the labels. Supervised navigation methods can be used directly by learning the mapping from the states in recorded trajectories that contain social policies to their corresponding labels or by learning reactive policies that imitate a planning algorithm (Groshev et al., 2017).

Another extensively explored learning technique is Reinforcement Learning (RL), in which an agent explores the state and actions by itself while a reward function is used to punish or encourage the decisions to obtain an optimal model. RL techniques can be used to provide a robot with the navigation paths that maximize rewards in terms of human safety or comfort (Chen et al., 2017). Moreover, Inverse Reinforcement Learning (IRL) is a technique that has been widely used to capture the navigation behavior of pedestrians. Contrary to supervised learning, IRL is able to recover a cost function that explains an observed behavior (Kuderer et al., 2013). The IRL technique proposed by Hamandi et al. (2019) trains the social navigation model by learning the navigation policy directly from human navigated paths in order to generate actions that conform to human-like trajectories. To include the social context in the learning process, these models aim to clone the navigation behavior of humans. Subsequently, robots are then equipped with these models for socially-compliant navigation.

Specifically, to clone an expert behavior in the RL framework, consider that an agent in an environment reaches a state s_{t+1} after executing an action $a_t \sim \pi$ that follows a policy π. At each transition state, the agent obtains a reward r_t presented as a scalar. The goal is for the agent to adjust the policy π to maximize the expected long-term rewards that it can receive. Q-learning (Watkins and Dayan, 1992) is an approach that enables us to find an optimal policy based on the state transition set. The Q-function represents the value of an action a_t and following a policy π as

$$Q_\pi(s_t, a_t) = \mathbf{E}[R(s_t)|s_t, a_t], \tag{1}$$

where R is the expected long term reward defined as $R = \sum_{t=0}^{\infty} \gamma^t r_t$, being $\gamma \in [0, 1]$ the discount-rate. Given the state s_t and action a_t the Q-function indicates the expected discounted accumulative reward. Using the Q-function, we can estimate an optimal policy π which maximizes the expected return. Particularly, no reward function is given in the IRL framework. Therefore, it is inferred from observed trajectories collected by the expert policy π_E to mimic the observed behavior.

There are numerous works using RL and IRL that generate human-like navigation behavior in controlled conditions. However, we can more elaborately define how we as humans navigate the environment, using a combination of both social skills and social norms as described in section 3.1. Social norms can vary with respect to the context, location, and culture. Extending the social skills of the robot by including social norms is important for social domain adaptation. The social norms that a domestic robot should consider while navigating are substantially different from those that a mobile robot in a hospital should conform to. For example, in order for the robot to navigate in a socially compliant manner in a hospital, it is essential for it to identify emergency situations, understand the priority for interaction, and have fast reaction times, so that the robot can never interfere with the paths of hospital staff and cause accidents or delay the treatment of a patients. Given that the context and priorities differ, the reaction also accordingly changes. We explore these cases in the case study that we describe in section 4.

Recently, a deep inverse Q-learning with constraints technique (Kalweit et al., 2020a) was introduced. This work presents one such model that allows for the combination of imitating human behavior and additional constraints. This is a novel model-free IRL approach that extends learning by imitation with constraints, such as safety or keeping to the right. Using the previous definition of Constrained Q-learning (Kalweit et al., 2020b), it includes a group of constraints C that shapes the possible actions in each state. Besides the Q-function in Inverse Q-learning, it also estimates a constrained Q-function Q_C for which the policy is extracted after Q-learning, considering only the action-values of the actions that satisfy the required constraint. This approach shows promising potential for considering relevant social factors while learning socially-aware robot navigation policies, especially by adding diverse constraints that represent current norms in order to yield socially intelligent and unbiased robot behavior.

3.3. Fairness Considerations

As with most learning approaches, the method described in section 3.2 requires a large number of training examples so that the model learns to yield the desired output. Therefore, it is essential to use either data gathered from the real-world, simulations, or control experiments. With the collected data, developers aim to present representative examples of real-world scenarios or guidance of the desired social behavior during navigation. However, these data collection processes can themselves reproduce biases, and as a consequence, it raises a series of critical concerns. In the specific case of learning socially-aware robot navigation from real-world data, robots can reproduce biased behaviors implicit in human-human interaction. On the other hand, the amount of training data that can be obtained from simulations and control experiments is very limited since only a handful of situations are taken into account. Most data collection processes that do not encompass a balanced set of every possible real-world scenario present a risk for robots trained on them as this could lead to navigation with biased behavior. These circumstances are considered as bias in the data. Accurate generalization of scenarios that highly deviate from the training data is an extremely difficult task. To address this factor, recent methods have been proposed to filter data that is used to train the models. For instance, Hagendorff (2020a) presents a

selection process for training data that improves the data quality in terms of ethical assessments of behavior and influences the training of the model. Nevertheless, methods to reduce bias in the data that is used for learning robot navigation models still remain unstudied.

Apart from the problems in dataset collection, there is still a lack of a deeper understanding of the underlying principles and limitations of modern learning algorithms. Especially, a phenomenon known as shortcut learning which shows how neural networks learn more straightforward predictors that are not necessarily related to the main task or objective (Geirhos et al., 2020). A typical example of this phenomenon can be seen in the hiring tool developed by Amazon which predicts strong candidates based on their curriculum. This tool was later found to be biased toward providing advantages for male applicants. Their model, which was trained on historical human decisions that were made during the hiring process identified that gender was an important feature for prediction (Dastin, 2018). Geirhos et al. (2020) analyses the dependency of outputs to strong predictive attributes found by the model during training.

Data-driven models can contain abstract representations of the data and situations that lead to the prediction. Therefore, it is typically challenging to explain the decisions made by a learned model. To facilitate the fairness analysis, we present an approach that is not solely data-driven and instead, it implicitly incorporates human interpretations of social dynamics using a model that includes high-level and explainable human notions about social conventions, relationships, and interactions to guide a mobile robot. The purpose of analyzing this approach is to demonstrate that biased behaviors can also be learned from biased demonstrations or observations. We analyze the approach proposed by Patompak et al. (2019) to predict personalized proxemics areas that correspond to the characteristics of individual people. This approach generates personalized comfort zones of a specific size and shape by associating the personal area with the activity that a person performs or characteristics of the person. Using these social descriptions, it estimates the proxemic zone that better matches each pedestrian in the scene. Consequently, the approach relies on personalized boundary delineation of two different areas: one area where the human-robot interaction can occur, and another area that is private, which the robot should avoid navigating through. The approach consists of three parts: human-social mode, learning the fuzzy social model, and a path planner. The human social model utilizes proxemics theory and aims to reflect the pedestrians' social factors in the scene. The social factors that are considered include gender, relative distance, and relationship degree. Using these factors, the approach yields the parameters that determine the private zone of comfort for each person in the scene based on the fuzzy logic system. For each social factor that is considered, the approach defines a membership function as follows:

A binary function depending on the gender of the pedestrian, which is given by

$$MF_{gender} = \begin{cases} 0, & \text{if gender is } \textit{Male} \\ 1, & \text{if gender is } \textit{Female}, \end{cases} \tag{2}$$

a sigmoid function with relative distance input r_r, distribution steepness a_r, and inflection point c_r describing *near* or *far* distance defined as

$$MF_{distance} = \frac{1}{1 + \exp(-a_r \times (r_r - c_r))}, \tag{3}$$

and three Gaussian functions representing the degree of relationship as familiar, acquaintance, and stranger, which is given by

$$MF_{relationship} = \begin{cases} \mathcal{N}(\mu_{\text{Fam}}, s^2{}_{\text{Fam}}), & \text{if degree of relationship is } \textit{Familiar} \\ \mathcal{N}(\mu_{\text{Acq}}, s^2{}_{\text{Acq}}), & \text{if degree of relationship is } \textit{Acquaintance} \\ \mathcal{N}(\mu_{\text{Str}}, s^2{}_{\text{Str}}), & \text{if degree of relationship is } \textit{Stranger}. \end{cases} \tag{4}$$

Subsequently, the fuzzy social model is learned from human feedback using an RL approach. The defined membership functions of the social factors can be learned to yield an improved personal area for each pedestrian. This is performed by adjusting the relationship degree in the MF (Equation 4) to update the social map. The reward of the RL model is then obtained from human-robot interaction by means of the emotion or feeling of each corresponding person. Therefore, the approach sets the focus on the degree of the relationship to be learned. Finally, the approach selects a path planner that chooses an optimal navigation path in the social cost map. The consequently designed social interaction area using fuzzy rules presents the output of the model as two separated personal areas: far personal area (FPA) and near personal area (NPA). As part of the rules presented, it is clear that for the input gender female, the near personal area is never an option. Taking into account that the reinforcement learning algorithm updates the model based on the $MF_{relationship}$, the resulting navigation policy would never allow for human-robot interaction close to women. This presents a critical bias of the model due to the inclusion of social dynamics. This is an example where bias appears due to an explicit constrain in the learning algorithm. Not only gender but other factors that may potentially lead to bias as well as other implicit or explicit biases can appear by learning from real-world data. We discuss this technical bias of the aforementioned navigation model with implications and analysis from the social perspective in section 4.

Learning robot navigation policies and models that are unbiased requires analyzing how the input is given, how the data is measured, how the data is labeled, what it means for models to be trained on them, what parameters are used, and how social navigation models are evaluated. If models aim to reflect the features of society, we need to question what behaviors should be replicated and promoted. For example, Kivrak et al. (2020) explicitly exclude women in the real-world experiments of their social navigation framework for assistive robots around humans. Their model that aims to yield human-friendly routes

was only tested in a corridor where women were excluded based on previous analysis (Jones and Healy, 2006), which affirms gender differences in spatial problem solving. This represents bias in the evaluation where the social model of navigation is validated only for a privileged group and can lead to underperformance to the unconsidered after the deployment. This has also been seen before in medical datasets or experiments where women were excluded citing differences in hormonal cycles, which leads to the medicines or medical procedures causing higher side effects for women compared to men. The consequences of these biased experiments or trials have been extensively discussed, which had lead to the inclusion of women in all medical trials (Söderström, 2001).

The technical bias analysis presented in this section shows cases where the high-level representation of social interaction replicates unequal roles and dynamics that already exist in human interaction. It is a significantly larger risk in the case of learning models for social navigation from demonstrations where the assumption is that the best way to teach a robot to navigate is to enable it to learn directly by observing humans.

3.4. Relearning

While learning socially-aware robot navigation models, social biases can be introduced that replicate and even augment the unfair societal dynamics. Most existing socially-aware robot navigation techniques aim to learn social navigation behavior by imitating human navigation. Consequently, it essential to deter biases during the deployment of robots equipped with such models. In this section, we present a mechanism to first detect when the navigation model makes biased decisions, especially against certain groups of people. Subsequently, we use this mechanism to update the model toward yielding more equitable social navigation policies.

There are many situations in the real-world where unequal decisions are desired, such as adapting the speed of the robot near older adults. In this work, we only analyze situations where there is no justifiable reason to yield different actions while interacting with different groups of people. In this case, an unfair or discriminatory system will offer an advantage to a certain group of users or unfavorable interaction to some other groups. Unfair behavior in robot navigation directly affects how users interact with the system. For a mobile robot to amend a discrimination behavior, it is necessary first to detect or measure the biased behavior. An advantage in the case of robots is that the decisions and actions after deployment can be used to measure the degree of biased decisions, for instance, concerning protected characteristics, such as age, gender, and race. Whereas, in the case of bias in deep learning models this task would be significantly harder. For instance, the Microsoft AI Twitter chatbot Tay which learned by interacting with users and presented gender-biased as well as racially offensive tweets (Perez, 2016). In this case, it would be necessary to additionally measure the features behind the posted tweets. Given that most robots are designed to move in the world, this characteristic comes for free in terms of the navigation actions that were made based on distance, speed, among other control variables as well as perception, accuracy, and uncertainty.

The robot can gather a dataset or a log by storing its own experiences and its corresponding actions even after deployment. Subsequently, the first step is to detect bias in the social navigation decisions of the robot. Bias identification is related to detecting disproportionate prejudice or favoritism toward some individuals or groups over others. For example, the paths planned by the robot produces a negative effect more frequently for specific groups of people than they do for another, such as discomfort, lack of interaction, or avoidance. Other situations are related to a disproportionate rate of a favorable or higher quality of attributes prediction for certain groups. This situation can present itself due to a lack of representation and diversity in the data or scenarios that were used in the learning stage. As a result, it can lead to unpredictable or no interaction with individuals of these groups.

One such method to detect if the navigation model exhibits outcomes that differ across subgroups is using clustering. Clustering is the technique for grouping data such that the elements of the same group are assigned closed together, forming assemblies called clusters. Clustering is a well-studied technique that is highly used in unsupervised or exploratory data analytics. Consider that the dataset collected while the robot was navigating contains all the decisions that were taken as well as the sensor data and the actions of other agents that these decisions were based on. Additionally, other navigation and perception attributes can be considered, such as the relative distance of the pedestrians to the robot, collisions, person identification confidence, and intention prediction, as well as additional information, such as rules that were violated and accidents that were caused. The accumulation of actions the robot outputs corresponds to the navigation feature set to be clustered. The resulting clusters can later be correlated to potential protected characteristics.

Having a learned policy π for socially-aware robot navigation, we define $V = \{v_1, v_2, \ldots, v_i\}$ as the set of navigation data that correspond to the experiences that the robot continuously accumulates through certain time steps. Different clustering algorithms can be used depending on the attributes of the selected navigation features (for instance if their nature is categorical or numerical). One promising clustering algorithm is the method proposed in Aljalbout et al. (2018) which consists of a fully convolutional autoencoder trained with two losses, one for reconstruction and the other for cluster hardening. The result of the clustering process is a collection of assemblies $A = A_1, A_2, \ldots, A_K$ consisting of navigation feature combinations. Each A_k represents the navigation experiences that are similar enough to be considered as a cluster of the entire set V. The number of clusters K and the size of each cluster A_k are hyperparameters that can be explored. Additionally, we define $F = \{f_1, f_2, \ldots, f_N\}$ as the set of protected features that we aim to analyze and each f_n has a set of navigation features V. To uncover social-group related bias the next step is to determine the relationship degree $D_{k,n}$ between each protected feature f_n and each generated cluster A_k.

After identifying that the robot actions in the navigation experience set are clustered and correlated to sensitive attributes, the next step is to trigger alarms or corrective actions when protected feature f_n strongly related to each generated cluster A_k,

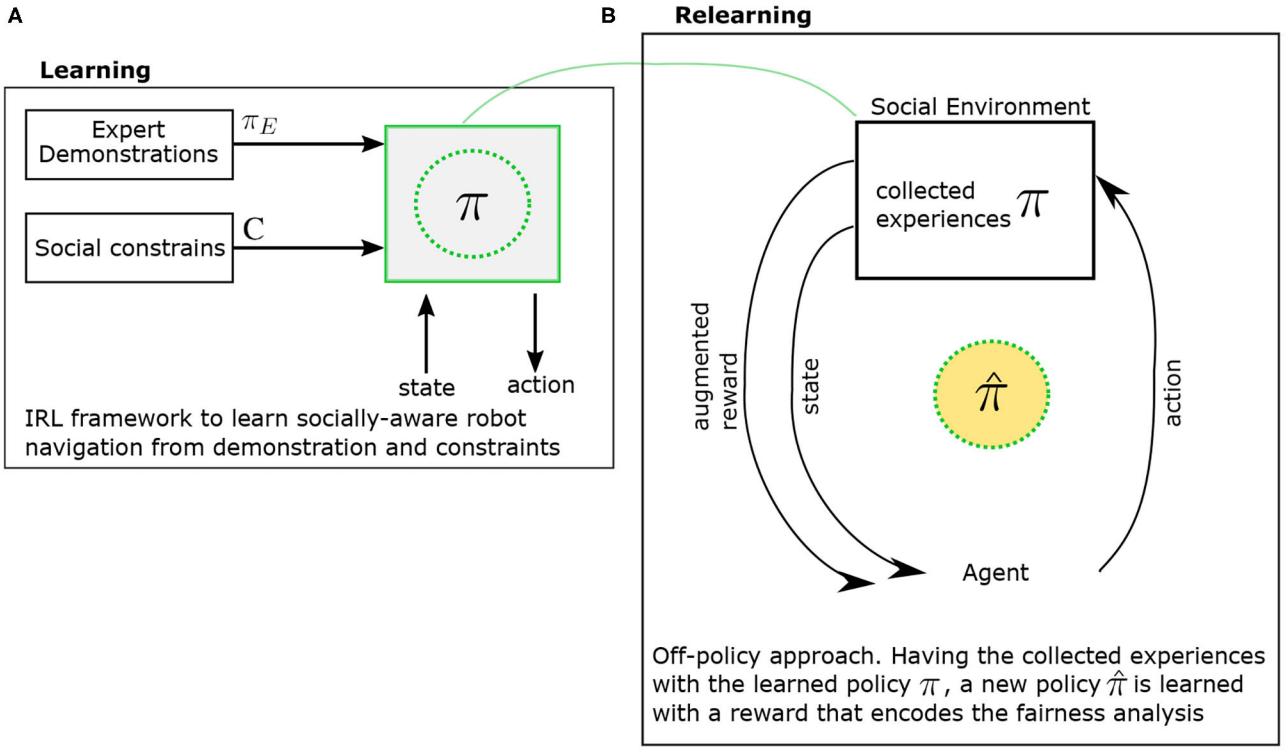

FIGURE 3 | Illustration of the Learning-Relearning framework for diminishing bias in social robot navigation. During the learning phase **(A)**, a policy π is learned for socially-aware robot navigation. During the relearning phase **(B)**, the robot uses the policy π to navigate in the social environment and collects the navigation experiences. An augmented reward that encodes detected biased behavior is used to relearn a new policy $\hat{\pi}$ so that the long term rewards reflect the decreasing unjustified bias related to social-groups.

defined as $D_{k,n} > u_n$ where u_n threshold that can be selected for each protected feature. A system of reward or punishment can be implemented in a off-policy reinforcement learning algorithm that optimizes an augmented reward that encodes the detection of unfair behavior as shown in **Figure 3**. The augmented reward rR_t is penalized when a biased behavior is detected so it does not only comprise the behavior for socially-aware navigation but it is also discounted when we detect bias as $D_{k,n} > u_n$. Therefore, the robot learns the policy π_R so that the long term rewards reflects the decreasing unjustified bias related to social-groups. As a result, it is possible to relearn the navigation model in our framework depending on the information gathered from the social environment.

From a more realistic perspective, demographic information is rarely known. Clustering also allows the reduction of this dependency between predictions and demographic information, when an unsupervised approach is employed. Therefore, when the dataset containing memory experiences of the robot navigating conforms to clusters beyond a given threshold, it can trigger an alarm for further analysis. Other methodologies that can be used to undercover bias in deep learning models are based on visualization of embeddings. Using visualization techniques, we can show how the model groups the data, which is useful to expose the reasons behind the prediction of the model. To do so, different tools can be used, such as T-distributed stochastic neighbor embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) to project the embeddings to reduce the dimensionality of the data. In this work, we focus on the relearning component based on clustering to present a feasible solution to account for fairness while learning socially compliant robot navigation that can be extended to an unsupervised algorithm.

4. CASE STUDIES AND DISCUSSION

In this section, we present extensive discussions that relate the technical analysis of our proposed framework to complex real-world scenarios that we present as three case studies. Each of these case studies contains different levels of human-robot interaction under four specific protected characteristics: gender, disabilities, age, and race. With these scenarios, we analyze the feasibility of model adaptation and the utility of this mechanism to check for fairness as well as to correct the bias. The figures illustrated in this section were generated using Icograms (2020).

4.1. Autonomous Floor Cleaning Robots

One of the most societally accepted robots has been the autonomous floor-cleaning machines (Forlizzi and DiSalvo, 2006; Forlizzi, 2007; Fink et al., 2013) and during the last decade they have been the most sold robots in the world (Research,

FIGURE 4 | Illustration of the autonomous floor cleaning robot scenario. The robot navigates taking the social conventions into account while performing the main task of cleaning the entire area.

2019). These robots have the task of cleaning floors using vacuum systems without any human supervision and recently, they can also mop floors using steam systems. These robots are currently used in households, and their navigation models vary in complexity depending on a wide range of prices. However, these robots are so far not equipped with socially aware navigation models. They do not avoid people or dynamic objects, rather they only change their cleaning route after they collide with an object. This can be attributed to the fact that in household environments, people are typically more tolerant given that they are aware of the task, features, and capacity of the robot.

It can be expected that the use of cleaning robots in the future will spread to different public areas. In this case study, we analyze from both technological and social points of view the functioning, requirements, and implications of the navigation of a cleaning robot that operates in a shopping mall. We illustrate this scenario in **Figure 4**. Consider that the shopping mall consists of multiple and extensive floors, and it is open to the public continually every day of the week. The groups of people visiting the place range from families and groups of friends to individual persons. Additionally, the reasons for the visit can differ, including people making quick shops, taking a walk, eating, etc. Therefore, we also expect varying types of behavior of the visitors, such as walking at a different speeds, talking in groups, and sitting down in different spaces.

The task of the robot in this case is to clean the entire environment effectively. In the following, we examine the effect that a cleaning robot equipped with social context can have. This robot has the ability to plan paths taking into account social conventions in public spaces, such as avoiding interfering with the paths of people, avoiding interrupting the interaction between people, prioritizing safety, avoiding surprising people with movements outside the visual range (or any other movement that might make people uncomfortable), navigating with a safe distance and with a prudent speed, avoiding collisions and predicting the intentions of people. With socially-aware navigation models, robots can fulfill the main task and act socially with predictable actions. The goal of including social context into the navigation model is to ensure that robots are not perceived as dangerous, bothersome, irritating, inconvenient, or obtrusive. The sociability of the cleaning robots can be defined as low or indirect, i.e., humans do not communicate with the robot. However, the interaction is generated by the navigation model in a socially acceptable manner. Social navigation models allow the robot to achieve the main goal without disturbing people sharing the same space. Consequently, the robot can operate in public spaces during the entire opening hours.

Specifically, if we employ the model (Patompak et al., 2019) presented in section 3.3 as the *learning* component in our framework, the personalized size and shape of the personal

zone can in fact improve the social intelligence of the robot. By avoiding crossing the comfort zone of people, these robots can learn to plan paths without disturbing the visitors of the shopping mall while performing the cleaning task. However, the model (Patompak et al., 2019) that takes the gender of a person into account can induce bias in the decisions. Even though women might prefer a larger comfort area during interaction among humans, it does not necessarily imply that they would prefer the same during human-robot interaction. In principle, a robot should never harm or be unfair to people based on their gender. In this work, we consider that the robot is depicted as a gender-neutral machine. Conforming a robot to a specific gender depending on the application could again lead to historical bias, this is an area that requires further research which is out of the scope of this paper. Moreover, according to the *bias evaluation* consideration for fairness described in section 2.2, maintaining different relative distances to people based on their gender is an unacceptable bias. Furthermore, distinguishing the comfort area by gender is not of high relevance to improve the acceptance or beneficial to improve the operation of robots around humans. Instead, there are other essential factors that can be used to improve comfort and confidence, such as safe navigation policies. Given that the bias presented in this case is explicit, it is easier to identify the bias inducing factor influencing the model in the *relearning* component of our framework, for example, by correlating the obtained behavior to the input constraints. After detecting the bias inducing factor, it can be excluded to re-train the model without the gender constraint.

On the other hand, while learning from demonstrations, data-driven models can also reflect negative bias. For instance, if robots learn from data that is not diverse where people with movement impairments are not present, then the robot might not react in a socially acceptable manner when they encounter such people. This can further lead to incorrect prediction of paths of people who walk slower and can make the robot be perceived as obtrusive. Data induced bias represents an implicit bias in the model that is more challenging to detect and correct for. Since the model disproportionately affects a specific group of people, by using our *relearning* component, the recurrent errors in the path prediction can be detected as a cluster that can also be related to the set of protected characteristics (e.g., people with mobility impairment). Consequently, by using a punishment system, the reward value is influenced after the detection of unwanted behavior to adjust the learning policy, allowing model adaptation toward a more fair behavior. This will support the *Value Alignment* consideration presented in section 2.2 in which accepted socially-aware robot navigation also considers inclusion.

4.2. Guidance Robots in a Shopping Mall

Mobile service robots have extensive use in innovative applications, such as for guidance in public spaces where they navigate alongside people and assist them to reach their desired destination. Based on the environment described in section 4.1, in this case study we analyze the effects of a guidance robot that operates in a shopping mall. Unlike the last scenario, the robot not only navigates under social conventions but also guides a person in a social manner. The task of the robot is to provide the requested information about locations in the shopping mall and accompany people to reach their desired location. This scenario is illustrated in **Figure 5**. Apart from guiding to reach a certain destination, the robot should also navigate considering social conventions that are required to provide comfort to all the surrounding people during navigation. Furthermore, the robot should coordinate with the user while navigating by maintaining a desired relative position with respect to the user. This scenario has similar characteristics to the mall in the previous case study where diverse people with different genders, ethnicity, disabilities, age, skin tones, and cultural origins and etc, will be present. In this example, fairness considerations, such as shared benefit, deterrence and value alignment described in the section 2.2 should be considered. Additionally, in the shopping mall scenario, the guidance robot will interact naturally with the user in a socially compliant manner while providing information and route guidance.

The human-robot interaction in this case is direct given that people approach the robot with a specific intention, and they expect a response from the robot that corresponds to the request. The resulting navigation strategy that these robots have next to people and their capacity to react according to the situation is crucial for their acceptance. Some of the important constraints in the navigation behavior of guidance robots are adapting the speed of the robot to the user, and maintaining a relative position and distance. If the robot navigates with a velocity that does not correspond to the user, then the robot risks being too slow or too fast which can cause uncoordinated behavior with the user and can further lead to accidents. On the other hand, relative distance and position are related to how people follow the robot and how the robot guides the user. Ideally, the robot should estimate the position and intention of the user during the execution of the guidance and also be able to interrupt the task if the person does not require any more help. Therefore, robots should adapt their navigation based on speed, intentions, motivations, orientation as well as handle unexpected situations, such as people crossing their path, changes in the speed of the person being guided, unexpected appearance of objects, among others.

Consumers value the unbiased, fast, and error-free behavior that a robot can provide. Therefore, the robot should adapt its behavior according to the current social context. In contrast to the interaction between people and cleaning robots, guidance robots provide personalized interaction, so the degree of sociability of this robot is greater. For example, if a disabled person goes to a shopping mall, the robot should recognize that this person will have different navigation behaviors than others so it should adapt its strategy accordingly. This adaptation will in turn make the person more comfortable using the assistance provided by the robot. In this example, aspects, such as the capability to recognize mobility impairments in a person and navigate accordingly are essential to ensure safe and comfortable guidance. Consider that a person with limited mobility requires guidance from the robot. If the robot is not equipped to react accordingly to mobility difficulties, the interaction can cause

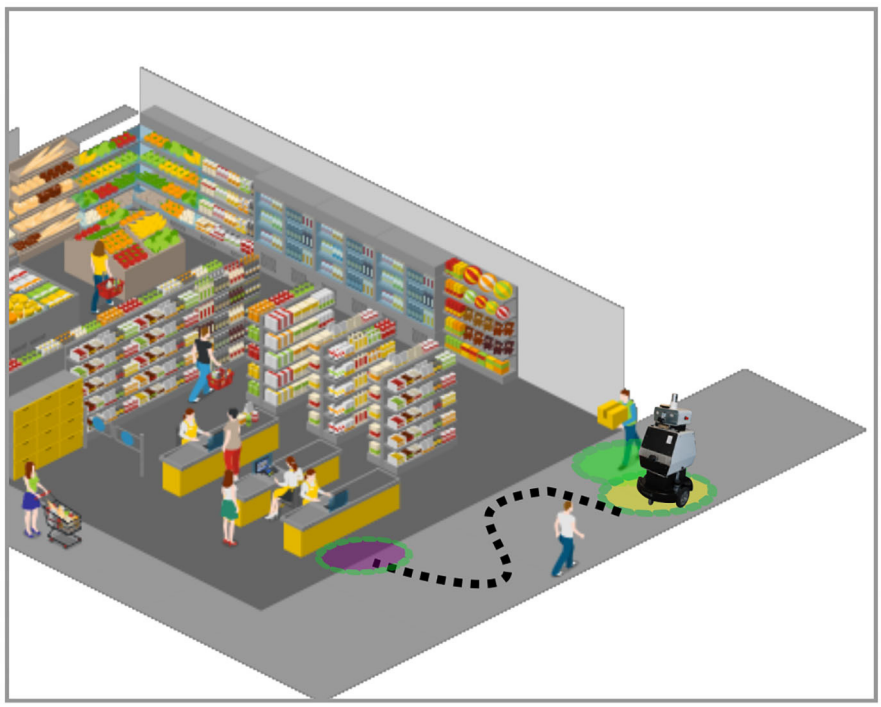

FIGURE 5 | Illustration of the guidance robot in a shopping mall scenario. The robot guides the user (green circle) to reach the destination (purple circle). Additionally, the robot is aware of the people in the surroundings during navigation while maintaining a desired relative position with respect to the user.

distress, physical overexertion, and even accidents. This will eventually make the person to discontinue using the robot in the future. In order to avoid such events, the navigation model in the robot should incorporate social adaptability skills that enable it to detect particular situations that cause discomfort or unintended outcomes for specific individuals.

Assume that a guidance robot is equipped with the navigation model described in section 3.3 and as a consequence it will assist women keeping larger distances with them. This may cause the robot to loose the interaction with them in certain situations and adversely affect the way that women perceive the robot. Similarly, it can reduce the efficiency with this population group representing the systematic disadvantage we aim to avoid toward diminishing bias. The model described in section 3.3 is used to present an example of learning socially-aware robot navigation in which unfair outcomes are associated with a protected characteristic. Other socially-aware navigation models that learn solely from human imitation can cause different types of model-induced biases. In these cases, the navigation model is optimized to yield sociable actions considering different factors, such as the velocity, orientation, priority of interaction, and route selection. The guidance robot will encounter situations where multiple people request for help simultaneously or even situations where people will try to interact with the robot when it is already guiding another person. Deciding which person has the priority is part of the social intelligence. Assume that in the *learning* component of our framework, the navigation model of the robot is trained from demonstrations and as a result, the robot learns the preferred interaction behavior based on those demonstrated interactions. This can lead to unfair outcomes due to human bias that may be existing in the demonstrations, policies reflecting personal bias, unequal society roles, or under-representation of minorities. Specifically, if the learning from demonstration is performed in a shopping mall only from one city, there will be insufficient diversity. Similarly, if the robot is deployed in a different place, or when people belonging to minorities try to use the robot, the robot will maintain its social behavior but it will likely make biased decisions, especially against people who historically have been discriminated, as we observed in other cases (Buolamwini and Gebru, 2018; Brandao, 2019; Wilson et al., 2019; Prabhu and Birhane, 2020). As part of the relearning component, our framework allows to generate clusters related to preferred interaction actions and determine if the generated clusters are strongly related to protected characteristics. Specifically, in case the preferred interaction of the robot is biased favoring or disadvantaging specific visitors of the shopping mall the learning policy is adjusted by a reward value that is penalized when biased behavior is detected. As a consequence, the robot's actions, such as deciding which person has the priority to interact with will follow the fairness requirements.

Since diverse people typically visit shopping malls, the robot should be able to accurately recognize them regardless of factors, such as skin tones. Previous studies (Wilson et al., 2019) have shown that recognition systems based on RGB perception present higher error rates for dark skin tones.

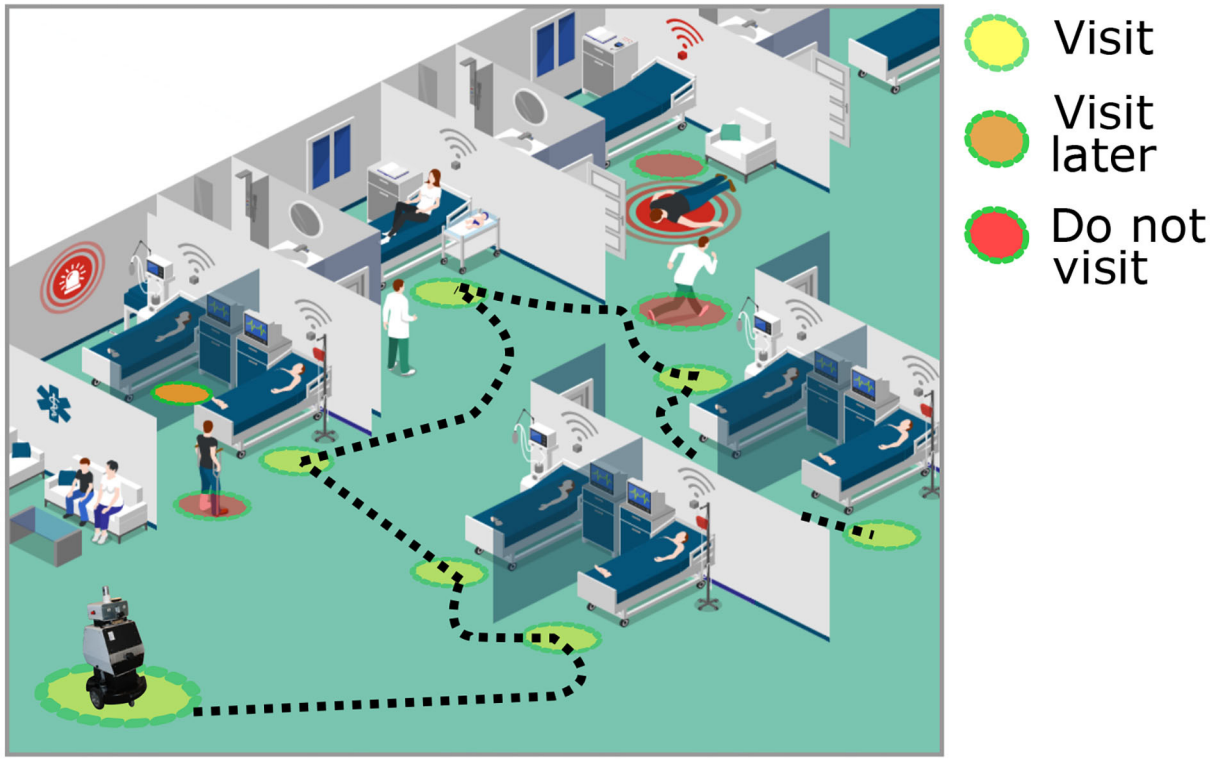

FIGURE 6 | Illustration of the caregiving robot in a hospital scenario. The main task of the robot is to distribute medicines to patients who are admitted in the hospital. The robot takes emergency situations that could happen into account and people requiring special assistance, while navigating.

If similar systems with faulty sensors or algorithms are used to learn social navigation models, the robot will be unable to recognize certain people and adhere to the fairness considerations described in section 2.2. As a consequence, the robot can perpetuate discrimination against groups of people that have historically been segregated, as observed in other learning applications, such as the automated risk assessment used by U.S judges and the biased vision-based object detectors employed in autonomous cars (Benthall and Haynes, 2019; Wilson et al., 2019). Furthermore, discrimination laws prohibit unfair treatment of people based on race. In this case, fairness priority is also important for the legal framework.

4.3. Caregiving Robots in Hospitals

There is significant interest in developing service robots for hospitals due to their ability to provide care for people. The use of robots in hospitals can be especially advantageous in cases where there are patients with contagious diseases, such as in a pandemic situation. In this case study, we analyze the navigation strategy of caregiving robots that operate in hospitals. The main task of robots in this case study is to distribute medicines to patients who are admitted in a hospital. **Figure 6** illustrates this scenario. The human-robot interaction in hospitals requires special caution as the robot will operate around patients who require special assistance. One such example is people with motion impairments who use wheelchairs, crutches, or walking frames. Furthermore, the robot will encounter rapidly changing situations, for example during an emergency where doctors and care staff rush through the hallways. To provide appropriate response, robots should be equipped with algorithms to understand situations and context that enable them to accordingly adapt their behavior. Apart from patients, robots will also interact with other people in the hospital, such as health professionals, secretaries, family members, and visitors. Similar to the shopping mall case study, caregiving robots will be interacting directly with the people. However, the navigation and interaction presents additional complexity, given that they do not assist people individually. Here, the robots aim to assist multiple people who have different medical treatments and deliver medicine to them while maintaining a socially accepted behavior. In this case, not only social conventions and sociability described in the previous case studies are required, but also priority decision making, optimal recognition, faster reaction and adaptability. As a consequence, the navigation models in caregiving robots should have higher requirements of accuracy and adaptability. These robots can particularly encounter unexpected events, such as emergency situations where people will be walking in different directions, speeds, and unpredictable movements. In such situations, there is a higher risk of accidents due to the vulnerability of people and the context in the hospital. Furthermore, the consequences of eventual accidents can be critical for the health of individuals. Caregiving robots should be able to perceive, recognize, and react according to the special requirements of the hospital.

Assume that the robots are going to be used in emergency rooms. Their task there is to deliver a series of necessary supplies to the people who are attending to the emergencies. Therefore, the robots have to interact with several people simultaneously. Based on the proxemics model described in section 3.3, the robot will be perceived as atypical in approaching people in different ways, assisting some people differently than others during urgent situations. Furthermore, taking into account that there are people playing specific roles, namely to care for sick people urgently, their comfort area of interaction is different from that of normal situations. People typically tend to walk fast, to have little personal space, and to quickly perceive what is happening around them. In this scenario, robots that navigate while maintaining different distances to people based on gender have lesser foreseeable utility. Alternatively, other characteristics can be considered that are related to the distribution of medicines depending on the needs of the patients and priorities, such as minimizing delivery time.

The priority of the path planning algorithms in such robots is to deliver medicines to all patients. Assume that in the *learning component* the caregiving robot learns from historical data about the characteristics of the patients. This model may learn that the pain threshold differs between men and women. Consequently, the navigation plan will be biased with negative effects toward men, based on information related to their higher tolerance to pain. Similarly, the robot could learn that women have more tolerance to wait longer for medical treatments and spend more overall time than men in the emergency rooms (Nottingham et al., 2018). In both situations, the behavior of the robot will be biased given that it systematically benefits a specific group of people. In this example, fairness considerations, such as value alignment and non-maleficence described in the section 2.2 can improve the decisions made by the robot. One approach to dealing with difficult cases of priority is to reflect political and commercial neutrality in robot navigation. This signifies that the navigation model in caregiving robots should not favor any particular group of people. Although, advocating for neutrality of assistive robots is a potential solution to bias problems in this case, the concept is substantially complex and requires further research.

Particularly, adapting the model with our *relearning* component to correct for the presented bias will lead the robot to base decisions on other factors. Using the *relearning* component of our framework, we can identify clusters that demonstrate a systematic disadvantage if the time to deliver medicines is higher for men and if women wait for a longer period of time in emergency rooms. Subsequently, to penalize the unfair behavior, we lower the reward value that adjusts the learning policy. As a result, the navigation model is adapted toward more fair behavior. If the model does not rely on the potentially negative bias inducing factors, it can learn better representations that reflect relevant characteristics, such as urgency and needs. While using our relearning technique, this type of bias in navigation will be detected when certain people receive attention more effectively than others. Consequently, if there is no valid reasoning behind such bias, the navigation model should be updated accordingly.

5. CONCLUSIONS

As more and more robots navigate in human spaces, they also require more complex navigation models to accomplish their goals while complying with the high safety and comfort requirements. Toward this direction, different methods incorporate social context into learning models to enable robots to navigate following social conventions. Typically, these methodologies utilize data or experiences from the real world, simulations, or control experiments and social constraints. In this work, we discussed the societal and ethical implications of learned socially-aware robot navigation techniques. We demonstrated that the advances accomplished in social robot navigation are essential for the development of robots that provide well for society. More importantly, we showed how these models that account for socially-aware robot navigation do not guarantee fairness in different real-world scenarios. Research in the direction of fairness in robot learning is of special importance, given that these machines interact with people closely.

To the best of our knowledge, this is the first work that studies the societal implications of bias in learned socially-aware robot navigation models. Our proposed framework that consists of the learning and relearning stages has the ability to effectively diminish bias in social robot navigation models. Additionally, we presented fairness considerations and specific techniques that can be used to implement our framework. We detailed several scenarios that show that the adaptability of the model in terms of fairness enables it to correct for bias. The scenarios demonstrate the potential unwanted outcomes of social navigation models that are described with variables and social conventions which make them easily interpretable. Our framework is especially useful for more complex learning models or models that are trained with imitation or reinforcement learning, given that these models contain more abstract representations of the data and situations. We hope this work contributes toward raising awareness on the importance of fairness in robot learning.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

All authors contributed to the concepts, analysis, and drafting the manuscript.

FUNDING

This work was partly funded by the BrainLinks-Brain Tools center of the University of Freiburg, a scholarship from the Graduate School of Robotics of the University Freiburg (according to the Graduate Funding Law of the Ministry of Science, Research and Arts of the State of Baden-Württemberg), and a grant from the Eva Mayr-Stihl Stiftung.

REFERENCES

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. *arXiv* 1803.02453.

Aljalbout, E., Golkov, V., Siddiqui, Y., Strobel, M., and Cremers, D. (2018). Clustering with deep learning: taxonomy and new methods. *arXiv* 1801.07648.

Anderson, M., and Anderson, S. L. (2010). Robot be good. *Sci. Am.* 303, 72–77. doi: 10.1038/scientificamerican1010-72

Argyle, M., Cook, M., and Cramer, D. (1994). Gaze and mutual gaze. *Br. J. Psychiatry* 165, 848–850. doi: 10.1017/S0007125000073980

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Benthall, S., and Haynes, B. D. (2019). "Racial categories in machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York, NY), 289–298. doi: 10.1145/3287560.3287575

Bicchi, A., and Tamburrini, G. (2015). Social robotics and societies of robots. *Inform. Soc.* 31, 237–243. doi: 10.1080/01972243.2015.1020196

Binns, R. (2018). "Fairness in machine learning: lessons from political philosophy," in *Conference on Fairness, Accountability and Transparency* (New York, NY: PMLR), 149–159.

Birdwhistell, R. L. (1952). *Introduction to Kinesics: An Annotation System for Analysis of Body Motion and Gesture*. Michigan: Department of State, Foreign Service Institute.

Birdwhistell, R. L. (2010). *Kinesics and Context: Essays on Body Motion Communication*. Pennsylvania, PA: University of Pennsylvania Press.

Birhane, A., and Cummins, F. (2019). Algorithmic injustices: towards a relational ethics. *arXiv* 1912.07376.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., et al. (2017). Principles of robotics: regulating robots in the real world. *Connect. Sci.* 29, 124–129. doi: 10.1080/09540091.2016.1271400

Bogue, R. (2016). Search and rescue and disaster relief robots: has their time finally come? *Ind. Robot* 43, 138–143. doi: 10.1108/IR-12-2015-0228

Boniardi, F., Valada, A., Burgard, W., and Tipaldi, G. D. (2016). "Autonomous indoor robot navigation using sketched maps and routes," in *Workshop on Model Learning for Human-Robot Communication at Robotics: Science and Systems (RSS)* (Michigan). doi: 10.1109/ICRA.2016.7487453

Bonnefon, J. F., Černy, D., Danaher, J., Devillier, N., Johansson, V., Kovacikova, T., et al. (2020). *Ethics of Connected and Automated Vehicles: Recommendations on Road Safety, Privacy, Fairness, Explainability and Responsibility*. Luxembourg: EU Publications.

Brandao, M. (2019). Age and gender bias in pedestrian detection algorithms. *arXiv* 1906.10490.

Brandão, M., Jirtoka, M., Webb, H., and Luff, P. (2020). Fair navigation planning: a resource for characterizing and designing fairness in mobile robots. *Artif. Intell.* 282:103259. doi: 10.1016/j.artint.2020.103259

BSI-2016 (2016). *BS 8611: 2016 Robots and Robotic Devices: Guide to the Ethical Design and Application of Robots and Robotic Systems*. London: British Standards Institution.

Buolamwini, J., and Gebru, T. (2018). "Gender shades: intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability and Transparency* (New York, NY), 77–91.

Castro, L., and Toro, M. A. (2004). The evolution of culture: from primate social learning to human culture. *Proc. Natl. Acad. Sci. U.S.A.* 101, 10235–10240. doi: 10.1073/pnas.0400156101

Cath, C. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 376, 1–8. doi: 10.1098/rsta.2018.0080

Chen, Y. F., Everett, M., Liu, M., and How, J. P. (2017). "Socially aware motion planning with deep reinforcement learning," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Vancouver, BC: IEEE), 1343–1350. doi: 10.1109/IROS.2017.8202312

Chouldechova, A., and Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv* 1810.08810.

Claure, H., Chen, Y., Modi, J., Jung, M., and Nikolaidis, S. (2019). Reinforcement learning with fairness constraints for resource distribution in human-robot teams. *arXiv* 1907.00313.

Costa-jussà, M. R. (2019). An analysis of gender bias studies in natural language processing. *Nat. Mach. Intell.* 1, 495–496. doi: 10.1038/s42256-019-0105-5

Dastin, J. (2018). *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*. San Fransico, CA: Reuters.

De Santis, A., Siciliano, B., De Luca, A., and Bicchi, A. (2008). An atlas of physical human-robot interaction. *Mech. Mach. Theory* 43, 253–270. doi: 10.1016/j.mechmachtheory.2007.03.003

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). "Measuring and mitigating unintended bias in text classification," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY), 67–73. doi: 10.1145/3278721.3278729

Fehr, E., and Fischbacher, U. (2004). Social norms and human cooperation. *Trends Cogn. Sci.* 8, 185–190. doi: 10.1016/j.tics.2004.02.007

Ferrer, G., Garrell, A., and Sanfeliu, A. (2013). "Social-aware robot navigation in urban environments," in *2013 European Conference on Mobile Robots* (Barcelona: IEEE), 331–336. doi: 10.1109/ECMR.2013.6698863

Ferrer, G., Zulueta, A. G., Cotarelo, F. H., and Sanfeliu, A. (2017). Robot social-aware navigation framework to accompany people walking side-by-side. *Auton. Robots* 41, 775–793. doi: 10.1007/s10514-016-9584-y

Fink, J., Bauwens, V., Kaplan, F., and Dillenbourg, P. (2013). Living with a vacuum cleaning robot. *Int. J. Soc. Robot.* 5, 389–408. doi: 10.1007/s12369-013-0190-2

Fiorini, P., and Prassler, E. (2000). Cleaning and household robots: a technology survey. *Auton. Robots* 9, 227–235. doi: 10.1023/A:1008954632763

Forlizzi, J. (2007). "How robotic products become social products: an ethnographic study of cleaning in the home," in *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (New York, NY: IEEE), 129–136. doi: 10.1145/1228716.1228734

Forlizzi, J., and DiSalvo, C. (2006). "Service robots in the domestic environment: a study of the roomba vacuum in the home," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (New York, NY), 258–265. doi: 10.1145/1121241.1121286

Forshaw, S., and Pilgerstorfer, M. (2008). Direct and indirect discrimination: is there something in between? *Ind. Law J.* 37, 347–364. doi: 10.1093/indlaw/dwn019

Fuchs, D. J. (2018). The dangers of human-like bias in machine-learning algorithms. *Missouri S&Ts Peer Peer* 2:1.

Garcia, M. (2016). Racist in the machine: the disturbing implications of algorithmic bias. *World Policy J.* 33, 111–117. doi: 10.1215/07402775-3813015

Gaydashenko, A., Kudenko, D., and Shpilman, A. (2018). "A comparative evaluation of machine learning methods for robot navigation through human crowds," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (Orlando, FL: IEEE), 553–557. doi: 10.1109/ICMLA.2018.00089

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., et al. (2020). Shortcut learning in deep neural networks. *arXiv* 2004.07780. doi: 10.1038/s42256-020-00257-z

Goodman, B., and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Mag.* 38, 50–57. doi: 10.1609/aimag.v38i3.2741

Goodwin, C. (2000). Action and embodiment within situated human interaction. *J. Pragmat.* 32, 1489–1522. doi: 10.1016/S0378-2166(99)00096-X

Groshev, E., Goldstein, M., Tamar, A., Srivastava, S., and Abbeel, P. (2017). Learning generalized reactive policies using deep neural networks. *arXiv* 1708.07280.

Grunwald, A. (2011). Responsible innovation: bringing together technology assessment, applied ethics, and STS research. *Enterpr. Work Innov. Stud.* 31, 10–19.

Hagendorff, T. (2020a). Ethical behavior in humans and machines-evaluating training data quality for beneficial machine learning. *arXiv* 2008.11463.

Hagendorff, T. (2020b). The ethics of AI ethics: an evaluation of guidelines. *Minds Mach.* 30, 99–120. doi: 10.1007/s11023-020-09517-8

Hall, E. T., Birdwhistell, R. L., Bock, B., Bohannan, P., Diebold, A. R. Jr., Durbin, M., et al. (1968). Proxemics [and comments and replies]. *Curr. Anthropol.* 9, 83–108. doi: 10.1086/200975

Hamandi, M., D'Arcy, M., and Fazli, P. (2019). "Deepmotion: learning to navigate like humans," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (New Delhi: IEEE), 1–7. doi: 10.1109/RO-MAN46459.2019.8956408

Harrigan, J. A. (2005). *Proxemics, Kinesics, and Gaze*. Oxford: Oxford University Press.

Hasan, K. M., Abdullah-Al-Nahid, and Reza, K. J. (2014). "Path planning algorithm development for autonomous vacuum cleaner robots," in *2014 International Conference on Informatics, Electronics & Vision (ICIEV)* (Dhaka: IEEE), 1–6. doi: 10.1109/ICIEV.2014.6850799

Helbing, D., and Molnar, P. (1995). Social force model for pedestrian dynamics. *Phys. Rev. E* 51:4282. doi: 10.1103/PhysRevE.51.4282

Howard, A., Zhang, C., and Horvitz, E. (2017). "Addressing bias in machine learning algorithms: a pilot study on emotion recognition for intelligent systems," in *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)* (Austin, TX: IEEE), 1–7. doi: 10.1109/ARSO.2017.8025197

Hurtado, J. V., Mohan, R., and Valada, A. (2020). MOPT: multi-object panoptic tracking. *arXiv* 2004.08189.

Hutchins, E. (2006). The distributed cognition perspective on human interaction. *Roots Hum. Soc. Cult. Cogn. Interact.* 1:375. doi: 10.4324/9781003135517-19

Icograms (2020). *Illustrations*. Avaialble online at: https://icograms.com/ (accessed December 18, 2020).

Jamshidi, P., Cámara, J., Schmerl, B., Käestner, C., and Garlan, D. (2019). "Machine learning meets quantitative planning: enabling self-adaptation in autonomous robots," in *2019 IEEE/ACM 14th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)* (Montreal, QC: IEEE), 39–50. doi: 10.1109/SEAMS.2019.00015

Jarvis, P. (2006). *Towards a Comprehensive Theory of Human Learning*, Vol. 1. New York, NY: Psychology Press.

Johnson, C., and Kuipers, B. (2018). "Socially-aware navigation using topological maps and social norm learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (Tulane), 151–157. doi: 10.1145/3278721.3278772

Johnson, K., Pasquale, F., and Chapman, J. (2019). Artificial intelligence, machine learning, and bias in finance: toward responsible innovation. *Fordham L. Rev.* 88:499.

Jones, C. M., and Healy, S. D. (2006). Differences in cue use and spatial memory in men and women. *Proc. R. Soc. B Biol. Sci.* 273, 2241–2247. doi: 10.1098/rspb.2006.3572

Kalweit, G., Huegle, M., Werling, M., and Boedecker, J. (2020a). "Deep inverse Q-learning with constraints," in *Advances in Neural Information Processing Systems*, 33.

Kalweit, G., Huegle, M., Werling, M., and Boedecker, J. (2020b). Interpretable multi time-scale constraints in model-free deep reinforcement learning for autonomous driving. *arXiv* 2003.09398.

Kaushal, A., Altman, R., and Langlotz, C. (2020). *Health Care AI Systems Are Biased*. Scientific American.

Khambhaita, H., and Alami, R. (2020). "Viewing robot navigation in human environment as a cooperative activity," in *Robotics Research* (Springer), 285–300. doi: 10.1007/978-3-030-28619-4_25

Kirby, R. (2010). *Social robot navigation* (Ph.D. thesis), Carnegie Mellon University, Pittsburgh, PA, United States.

Kivrak, H., Cakmak, F., Kose, H., and Yavuz, S. (2020). Social navigation framework for assistive robots in human inhabited unknown environments. *Eng. Sci. Technol. Int. J.* 24, 284–298. doi: 10.1016/j.jestch.2020.08.008

Kretzschmar, H., Spies, M., Sprunk, C., and Burgard, W. (2016). Socially compliant mobile robot navigation via inverse reinforcement learning. *Int. J. Robot. Res.* 35, 1289–1307. doi: 10.1177/0278364915619772

Kruse, T., Pandey, A. K., Alami, R., and Kirsch, A. (2013). Human-aware robot navigation: a survey. *Robot. Auton. Syst.* 61, 1726–1743. doi: 10.1016/j.robot.2013.05.007

Kuderer, M., Kretzschmar, H., and Burgard, W. (2013). "Teaching mobile robots to cooperatively navigate in populated environments," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems* (Tokyo: IEEE), 3138–3143. doi: 10.1109/IROS.2013.6696802

Lee, N. T. (2018). Detecting racial bias in algorithms and machine learning. *J. Inform. Commun. Ethics Soc.* 16, 252–260. doi: 10.1108/JICES-06-2018-0056

Lin, P., Abney, K., and Bekey, G. A. (2012). *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge: Intelligent Robotics and Autonomous Agents Series.

Liu, H. Y., and Zawieska, K. (2017). From responsible robotics towards a human rights regime oriented to the challenges of robotics and artificial intelligence. *Ethics Inform. Technol.* 22, 321–333. doi: 10.1007/s10676-017-9443-3

Lu, K., Mardziel, P., Wu, F., Amancharla, P., and Datta, A. (2020). "Gender bias in neural natural language processing," in *Logic, Language, and Security* (Cham: Springer), 189–202. doi: 10.1007/978-3-030-62077-6_14

Luber, M., Spinello, L., Silva, J., and Arras, K. O. (2012). "Socially-aware robot navigation: a learning approach," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems* (Vilamoura-Algarve: IEEE), 902–907. doi: 10.1109/IROS.2012.6385716

McDonald, D. W., McCarthy, J. F., Soroczak, S., Nguyen, D. H., and Rashid, A. M. (2008). Proactive displays: supporting awareness in fluid social environments. *ACM Trans. Comput. Hum. Interact.* 14, 1–31. doi: 10.1145/1314683.1314684

Mittal, M., Mohan, R., Burgard, W., and Valada, A. (2019). Vision-based autonomous UAV navigation and landing for urban search and rescue. *arXiv* 1906.01304.

Nelson, G. S. (2019). Bias in artificial intelligence. *North Carolina Med. J.* 80, 220–222. doi: 10.18043/ncm.80.4.220

Nolfi, S., and Floreano, D. (2002). Synthesis of autonomous robots through evolution. *Trends Cogn. Sci.* 6, 31–37. doi: 10.1016/S1364-6613(00)01812-X

Nottingham, Q. J., Johnson, D. M., and Russell, R. S. (2018). The effect of waiting time on patient perceptions of care quality. *Qual. Manage. J.* 25, 32–45. doi: 10.1080/10686967.2018.1404368

Ötting, S. K., Gopinathan, S., Maier, G. W., and Steil, J. J. (2017). "Why criteria of decision fairness should be considered in robot design," in *20th ACM Conference on Computer-Supported Cooperative Work and Social Computing* (Portland, OR).

Patompak, P., Jeong, S., Nilkhamhang, I., and Chong, N. Y. (2019). Learning proxemics for personalized human-robot social interaction. *Int. J. Soc. Robot.* 12, 267–280. doi: 10.1007/s12369-019-00560-9

Perez, S. (2016). *Microsoft Silences Its New AI Bot Tay, After Twitter Users Teach It Racism*. Tech Crunch.

Piano, S. L. (2020). Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Human. Soc. Sci. Commun.* 7, 1–7. doi: 10.1057/s41599-020-0501-9

Poudel, D. B. (2013). Coordinating hundreds of cooperative, autonomous robots in a warehouse. *AI Mag.* 27, 1–13.

Prabhu, V. U., and Birhane, A. (2020). Large image datasets: a pyrrhic win for computer vision? *arXiv* 2006.16923.

Reed, C., Kennedy, E., and Silva, S. (2016). *Responsibility, Autonomy and Accountability: Legal Liability for Machine Learning*. London: Queen Mary School of Law Legal Studies Research Paper.

Research, I. (2019). *Floor Cleaning Robot Market by Robot Type, by Sales Channel, by Region–Global Forecast Up to 2025*. Research and Markets.

Riek, L., and Howard, D. (2014). "A code of ethics for the human-robot interaction profession," in *Proceedings of We Robot*.

Rios-Martinez, J., Spalanzani, A., and Laugier, C. (2015). From proxemics theory to socially-aware navigation: a survey. *Int. J. Soc. Robot.* 7, 137–153. doi: 10.1007/s12369-014-0251-1

Silberg, J., and Manyika, J. (2019). *Notes From the AI Frontier: Tackling Bias in AI (and in Humans)*. Mckinsey Global Institute.

Silver, D., Bagnell, J. A., and Stentz, A. (2010). Learning from demonstration for autonomous navigation in complex unstructured terrain. *Int. J. Robot. Res.* 29, 1565–1592. doi: 10.1177/0278364910369715

Simmel, G. (1949). The sociology of sociability. *Am. J. Sociol.* 55, 254–261. doi: 10.1086/220534

Söderström, M. (2001). Why researchers excluded women from their trial populations. *Lakartidningen* 98, 1524–1528.

Stilgoe, J., Owen, R., and Macnaghten, P. (2013). Developing a framework for responsible innovation. *Res. Policy* 42, 1568–1580. doi: 10.1016/j.respol.2013.05.008

Tai, L., Zhang, J., Liu, M., and Burgard, W. (2018). "Socially compliant navigation through raw depth inputs with generative adversarial imitation learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)* (Brisbane: IEEE), 1111–1117. doi: 10.1109/ICRA.2018.8460968

Tewari, A., Peabody, J., Sarle, R., Balakrishnan, G., Hemal, A., Shrivastava, A., et al. (2002). Technique of Da Vinci robot-assisted anatomic radical prostatectomy. *Urology* 60, 569–572. doi: 10.1016/S0090-4295(02)01852-6

Thrun, S. (1995). An approach to learning mobile robot navigation. *Robot. Auton. Syst.* 15, 301–319. doi: 10.1016/0921-8890(95)00022-8

Thrun, S., Schulte, J., and Rosenberg, C. (2000). "Interaction with mobile robots in public places," in *IEEE Intelligent Systems*, 7–11.

Torresen, J. (2018). A review of future and ethical perspectives of robotics and AI. *Front. Robot. AI* 4:75. doi: 10.3389/frobt.2017.00075

Toupet, O., Biesiadecki, J., Rankin, A., Steffy, A., Meirion-Griffith, G., Levine, D., et al. (2020). Terrain-adaptive wheel speed control on the curiosity mars rover: algorithm and flight results. *J. Field Robot.* 37, 699–728. doi: 10.1002/rob.21903

Ulrich, I., and Borenstein, J. (2001). The guidecane-applying mobile robot technologies to assist the visually impaired. *IEEE Trans. Syst. Man Cybern. A Syst. Hum.* 31, 131–136. doi: 10.1109/3468.911370

Valada, A., Tomaszewski, C., Kannan, B., Velagapudi, P., Kantor, G., and Scerri, P. (2012). "An intelligent approach to hysteresis compensation while sampling using a fleet of autonomous watercraft," in *International Conference on Intelligent Robotics and Applications* (Montreal, QC: Springer), 472–485. doi: 10.1007/978-3-642-33515-0_47

Vandemeulebroucke, T., de Casterlé, B. D., and Gastmans, C. (2020). Ethics of socially assistive robots in aged-care settings: a socio-historical contextualisation. *J. Med. Ethics* 46, 128–136. doi: 10.1136/medethics-2019-105615

Vayena, E., Blasimme, A., and Cohen, I. G. (2018). Machine learning in medicine: addressing ethical challenges. *PLoS Med.* 15:e1002689. doi: 10.1371/journal.pmed.1002689

Verbeek, P. P. (2008). "Morality in design: design ethics and the morality of technological artifacts," in *Philosophy and Design* (Dordrecht: Springer), 91–103. doi: 10.1007/978-1-4020-6591-0_7

Wang, Q., Xu, Z., Chen, Z., Wang, Y., Liu, S., and Qu, H. (2020). Visual analysis of discrimination in machine learning. *IEEE Trans. Vis. Comput. Graph.* 27, 1470–1480. doi: 10.1109/TVCG.2020.3030471

Watkins, C. J., and Dayan, P. (1992). Q-learning. *Mach. Learn.* 8, 279–292. doi: 10.1007/BF00992698

Wilson, B., Hoffman, J., and Morgenstern, J. (2019). Predictive inequity in object detection. *arXiv* 1902.11097.

Winner, L. (1978). *Autonomous Technology: Technics-Out-of-Control as a Theme in Political Thought*. Cambridge: MIT Press.

Wittrock, M. C. (2010). Learning as a generative process. *Educ. Psychol.* 45, 40–45. doi: 10.1080/00461520903433554

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. (2017). Learning non-discriminatory predictors. *arXiv* 1702.06081.

Yu, A. (2019). Direct discrimination and indirect discrimination: a distinction with a difference. *WJ Legal Stud*. 9:1. doi: 10.5206/uwojls.v9i2.8072

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017). "Fairness constraints: mechanisms for fair classification," in *Artificial Intelligence and Statistics* (Fort Lauderdale, FL: PMLR), 962–970.

Zhang, L., Wu, Y., and Wu, X. (2016). A causal framework for discovering and removing direct and indirect discrimination. *arXiv* 1611.07509. doi: 10.24963/ijcai.2017/549

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Hurtado, Londoño and Valada. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.