Machine Learning Security as a Source of Unfairness in Human-Robot Interaction

Luke E. Richards
lerichards@umbc.edu
University of Maryland, Baltimore County
USA
Pacific Northwest National Laboratory
USA

Cynthia Matuszek
cmat@umbc.edu
University of Maryland, Baltimore County
USA

ABSTRACT

Machine learning models that sense human speech, body placement, and other key features are commonplace in human-robot interaction. However, the deployment of such models in themselves is not without risk. Research in the security of machine learning examines how such models can be exploited and the risks associated with these exploits. Unfortunately, the threat models of risks produced by machine learning security do not incorporate the rich sociotechnical underpinnings of the defenses they propose; as a result, efforts to improve the security of machine learning models may actually increase the difference in performance across different demographic groups, yielding systems that have risk mitigation that work better for one group than another. In this work, we outline why current approaches to machine learning security present DEI concerns for the human-robot interaction community and where there are open areas for collaboration.

CCS CONCEPTS

• **Security and privacy** → **Social aspects of security and privacy**; • **Computing methodologies** → **Speech recognition**; • **Human-centered computing** → *Collaborative interaction.*

KEYWORDS

machine learning security, neural networks, adversarial machine learning, human-robot interaction

ACM Reference Format:
Luke E. Richards and Cynthia Matuszek. 2023. Machine Learning Security as a Source of Unfairness in Human-Robot Interaction. In *Proceedings of ACM/IEEE International Conference on Human-Robot Interaction Workshop on Inclusive HRI II: Equity and Diversity in Design, Application, Methods, and Community (DEI HRI '23).* ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

A robot's ability to sense a human interacting with it is the first step in many human-robot interaction software architectures [2, 10]. The new standard method for such sensing is to employ a machine learning model, more specifically, deep learning models, to process human inputs such as speech, pose, eye tracking, and even mood. The literature is mounting on how these models fail to equitably sense humans along race, gender, accent, and socioeconomic status axes. At the same time, an additional layer is being added that attempts to protect the model from misuse through the concept of adversarial attacks. We refer to this layer as the machine learning security layer, which has been the subject of significant recent research efforts. However, very little research exists on how such safety and security mitigation can result in unfairness in human-robot interaction.

Within this manuscript, we will address two questions: "Why should the human-robot interaction community consider adversarial attacks in robotic security?" and "Why do the currently proposed defenses contribute to the unfairness already faced when deploying human-sensing models?" The core answer to both questions is that commonly used defenses to adversarial attacks have the potential to worsen performance *differentially* across groups—that is, the application of a security intervention may protect against attacks for some kinds of people but not others, or may worsen the performance of the overall system more for certain kinds of human interactors.

To discuss these risks, we will use a running example of a collaborative human-robot system in a workplace, such as a manufacturing setting. In this example, the goals of the robotic cooperator are to sense the human's current pose in order to plan actions safely around them and to process speech commands such as "pause" and "start." We make a particular note that such a robotic system can cause harm to humans if improper sensing is done. Adversarial attacks themselves represent a threat to human safety. An adversary could craft an attack in that a human's body was not recognized or misinterpreted, inducing the robot's motion planning to cause dangerous movement. Such attacks are not theoretical and have been physically implemented through printable stickers, clothing, and even simple dots on the camera [5].

These possibilities address the first question of why adversarial attacks pose a risk for human-robot interaction. Naturally, machine learning model researchers and developers have begun integrating methods to mitigate such risks. While these techniques may help avoid risk, they introduce a plethora of novel problems. In this abstract, we will discuss these novel problems and argue that to ensure robots can perform appropriately when interacting with a wide diversity of humans, it is necessary to consider the security provisions of underlying machine learning-based approaches.

DEI HRI '23, March 13, 2023, Stockholm, Sweden

2 SECURITY INTERVENTIONS AND RISKS

As mentioned, successful attacks on machine learning-based sensing models have been demonstrated, leading to the need for defense mechanisms. One primary method is a form of training augmentation called adversarial training [7], which incorporates adversarial attacks during learning. Recent works [6] suggest that this method is insufficient for the necessary performance in robotic learning; prior work in less human-centric spaces reports the performance degradation for minority groups [13]. These methods then represent a trade-off between fairness, accuracy, and robustness. This trade-off behavior is not well studied, especially given the limited representativeness of datasets concerning social labels and other factors. Returning to our running example, as a result of the security layer, the human-sensing model may have further performance degradation on people less represented in the dataset, exacerbating an existing issue in machine learning. We can see this occurring as cobots misinterpreting commands spoken by women or improperly failing to detect people with darker skin tones.

Conversely, methods have been introduced that center around the rejection of data points that the model may classify as an attack. Rejection methods [9] do not necessarily reduce average performance but may reject users outside the training set. Using a rejection method to detect adversarial attacks on a robotic platform meant to detect objects was successful [9]. However, it is less clear how such methods should be applied when people in the environment are being sensed rather than objects. Extrapolating from this method using the distance from the training data as a measure, we can hypothesize that user demographics not seen during training would face rejection. In our running example, this would look like a robotic cooperator refusing to work with specific employees systemically rejected by such a module sitting on top of the machine learning model.

3 DIFFERENTIAL PERFORMANCE BY DEFENSE METHODS

Our recent results showcased how such unfairness can be realized [12]. We examined the neural rejection method proposed in robotic object detection [9]. They suggest hardening an already-trained neural network for a task by using the final embedding layer to learn a support vector machine (SVM). The SVM has the property of being a compact abating probability (CAP) model [9], meaning the model's probability outputs can be a proxy for distance from training data. The intuition is that adversarial examples exist outside of this training distribution. While many defenses, including this one, have been questioned regarding effectiveness [11], it is a helpful lens for empirically showcasing the principle.

Rather than applying this method to just objects in an environment, we are interested in the case where such a method is in use for human interaction, such as speech commands. We used a single-word speech classification dataset, Common Voice Clips [1], with age, accent, and gender labels. This dataset could be analogous to simple commands a robotic collaborator would recognize during operation. We trained a one-dimensional convolutional neural network modeled after the M5 architecture [3] on this training subset and then trained an SVM on the produced embeddings afterward. We measured the parity of rejection between examples from groups without attack, allowing us to determine the rejection rates a user would face with honest use. Since the neural rejection method operates over a threshold, we measure differences in groups' false positivity rates over these thresholds. Ideally, we would see similar trends of erroneous flagging as security measures increase for all users of all demographics. Our false positivity parity (AUC_{FPR}) is the difference between the smallest and largest area under the false positive curve for a given demographic. An example would be the difference in the rejection rate between users who identify as women and men. A smaller value here represents more parity, the ideal case.

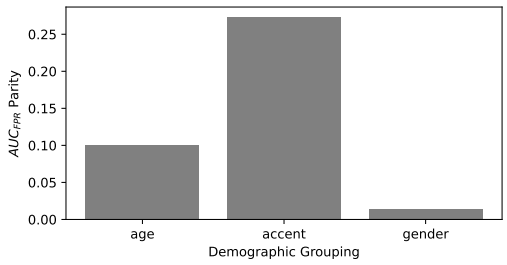

Figure 1: AUC_{FPR} parity between age, accent, and gender groups in Common Voice Clips dataset. This demonstrates that different groups face differential performance when adding a security layer to the model. Here, we see that adding a security layer creates a disparity in how often 'honest' commands are rejected for different groups.

Our findings (Figure 1) show that groups along the axes of age, accent, and gender do, in fact, have differential performance. We can see the differential performance for the users, especially those in different accent groups, followed by gender having the smallest difference. This result supports the claim that machine learning security methods introduced can have unfair harms when deployed in human sensing for interaction.

4 OPPORTUNITIES IN HRI STUDIES

We propose contextualizing proposed ML defense methods through human-robot interaction studies to discover bias before deployment and adoption. Methods like neural rejection have only ever been evaluated in terms of accuracy. This evaluation paradigm ignores unfairness and richer questions of usability. Discovering systemic failure modes poses an issue relating to the diversity of participants. Our example of a cooperative robot would require studies with diverse participants across social demographics, accents, body types, and communication styles. This challenge increases with additional sensing modalities and possible interactions.

We have already outlined how defense methods exacerbate the existing bias in machine learning models. How users respond to such methods, particularly rejection of a user's command or body in space, raises sensitive social and psychological questions that can be studied in the context of users' perceptions when interacting with robots that deploy ML defense methods. Mental models of threat models in human-robot interaction present another understudied area. Questions such as "How do users react to a model rejecting a

command and conveying that this is due to security concerns?" offer an opportunity that bridges the machine learning, usable security, and human-robot interaction communities. While graphical user interfaces offer a simple disclaimer, interactions with a robot do not offer as straightforward an explanation.

We can see how this can present challenges when returning to the example of a robotic collaborator in the workplace. An explanation may be possible through verbal cues when an example is rejected. However, users may find such behavior frustrating, especially when repeatedly given only to particular users. Finding thresholds of security implementations that satisfy usability, fairness, and the underlying threat model must have practical study designs to be deployed in multiple contexts. The threat model for a home robot's primary purpose is in social interaction, and home monitoring offers less risk than an industrial robot working around humans. Such taxonomies of risk from the human-robot interaction community [4, 8, 14] can deeply inform the strength of mitigation at deployment while also expanding to encapsulate these novel threat vectors.

5 CONCLUSION

The social and economic impacts of machine learning security methods being implemented in robots caring for humans in a healthcare environment, collaborating in a manufacturing environment, and offering other services are profoundly understudied. Bridging the communities of machine learning security with human-robot interaction will nourish deeply needed discussions on evaluation methods and the development of fairness-aware security methods in human-robot interaction. While the machine learning community is still addressing bias in models alone, adding components addressing operational constraints and safety must be brought into the same conversation.

REFERENCES

[1] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670* (2019).

[2] Svante Augustsson, Linn Gustavsson Christiernin, and Gunnar Bolmsjo. 2014. Human and Robot Interaction based on Safety Zones in a Shared Work Environment. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 118–119.

[3] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das. 2017. Very deep convolutional neural networks for raw waveforms. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 421–425.

[4] Tom P. Huck, Nadine Münch, Luisa Hornung, Christoph Ledermann, and Christian Wurll. 2021. Risk assessment tools for industrial human-robot collaboration: Novel approaches and practical needs. *Safety Science* 141 (2021), 105288. https://doi.org/10.1016/j.ssci.2021.105288

[5] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*. Chapman and Hall/CRC, 99–112.

[6] Mathias Lechner, Alexander Amini, Daniela Rus, and Thomas A Henzinger. 2022. Revisiting the Adversarial Robustness-Accuracy Tradeoff in Robot Learning. *arXiv preprint arXiv:2204.07373* (2022).

[7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[8] Damien Martin-Guillerez, Jérémie Guiochet, David Powell, and Christophe Zanon. 2010. A UML-based method for risk analysis of human-robot interactions. In *Proceedings of the 2nd International Workshop on Software Engineering for Resilient Systems*. 32–41.

[9] Marco Melis, Ambra Demontis, Battista Biggio, Gavin Brown, Giorgio Fumera, and Fabio Roli. 2017. Is deep learning safe for robot vision? Adversarial examples against the icub humanoid. In *Proceedings of the IEEE international conference on computer vision workshops*. 751–759.

[10] Dong Hai Phuong Nguyen, Matej Hoffmann, Alessandro Roncone, Ugo Pattacini, and Giorgio Metta. 2018. Compact Real-time Avoidance on a Humanoid Robot for Human-robot Interaction. In *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 416–424.

[11] Maura Pintor, Luca Demetrio, Angelo Sotgiu, Ambra Demontis, Nicholas Carlini, Battista Biggio, and Fabio Roli. 2021. Indicators of attack failure: Debugging and improving optimization of adversarial examples. *arXiv preprint arXiv:2106.09947* (2021).

[12] Luke E Richards, Edward Raff, and Cynthia Matuszek. 2023. Measuring Equality in Machine Learning Security Defenses. *arXiv preprint arXiv:2302.08973* (2023).

[13] Haipei Sun, Kun Wu, Ting Wang, and Wendy Hui Wang. 2022. Towards Fair and Robust Classification. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 356–376.

[14] Roger Woodman, Alan F.T. Winfield, Chris Harper, and Mike Fraser. 2012. Building safer robots: Safety driven control. *The International Journal of Robotics Research* 31, 13 (2012), 1603–1626. https://doi.org/10.1177/0278364912459665 arXiv:https://doi.org/10.1177/0278364912459665