GAMES FOR FAIRNESS AND INTERPRETABILITY

Eric Chu*, **Nabeel Gillani***, **Sneha Priscilla Makini**
Massachusetts Institute of Technology
{echu,ngillani,snehapm}@mit.edu

ABSTRACT

As Machine Learning (ML) systems become more ubiquitous, ensuring the fair and equitable application of their underlying algorithms is of paramount importance. We argue that one way to achieve this is to proactively cultivate public pressure for ML developers to design and develop fairer algorithms — and that one way to cultivate public pressure while simultaneously serving the interests and objectives of algorithm developers is through gameplay. We propose a new class of games — "games for fairness and interpretability" — as one example of an incentive-aligned approach for producing fairer and more equitable algorithms. Games for fairness and interpretability are carefully-designed games with mass appeal that (1) provide insights into how machine learning models work and (2) produce data that helps researchers and developers improve their algorithms. We highlight several possible examples of games, their implications for fairness and interpretability, how their proliferation could creative positive public pressure by narrowing the gap between algorithm developers and the general public, and why the ML community could benefit from them.

1 INTRODUCTION

As ML increasingly permeates virtually all aspects of life — and unequally serves, or fails to serve, certain subsegments of the population (Caliskan et al. (2017); Bolukbasi et al. (2016); Buolamwini & Gebru (2018)) — there is a need for a deeper exploration of how ML algorithms can be made fairer and more interpretable. To achieve this, we believe effective public pressure will be one lever to better models. There are several examples from history of how public pressure has spurred changes to technology policies. The creation of dynamite; America's use of the atomic bomb during the second world war; and the eugenics movement from the early 20th century are all examples of ethically dubious endeavors that were at least somewhat abated by a critical public response[1]. However, recent stories about Facebook and Cambridge Analytica, driverless cars going rogue[2], and even machine-powered labor displacement (Autor (2015)) have hinted at the dangers of simply letting history unfold. Public pressure is often reactive and arises in the wake of crises. To counter this, we ask: how can public pressure operate proactively in order to ensure ML can effectively ground itself in — and respond to — calls for fairness and interpretability?

To that end, some authors have recently sparked public conversation around the ethical pitfalls of machine learning (O'Neil (2016); Eubanks (2018); Noble (2018)). Furthermore, initiatives like Turingbox (Epstein et al. (2018)) and OpenML (Vanschoren et al. (2014)) are actively seeking to create platforms and marketplaces where members of the scientific community and general public can audit ML algorithms to promote more fairness, transparency, and accountability. These efforts are important first steps towards generating proactive public pressure. However, they fail to directly align incentives between those who design and deploy algorithms and those who are affected by them. Why should an algorithm developer care about how a niche group of individuals rates the fairness or interpretability of his or her algorithms? Why should members of the general public spend their time trying to understand, let alone evaluate, these algorithms? It is unclear how sustainable current efforts to generate proactive public pressure will be without incentive alignment.

*Authors contributed equally

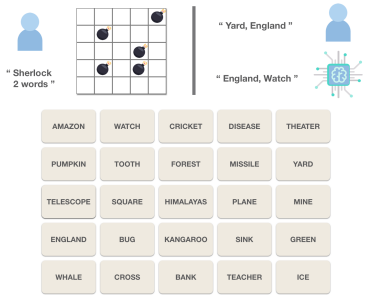
[1] https://www.bostonglobe.com/ideas/2018/03/22/computer-science-faces-ethics-crisis-the-cambridge-analytica-scandal-proves/IzaXxl2BsYBtwM4nxezgcP/story.html

[2] https://www.nytimes.com/2018/03/23/technology/uber-self-driving-cars-arizona.html

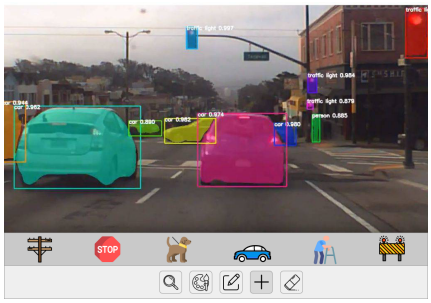
To align incentives between ML developers and the general public in a quest for more interpretable — and as a result, in due course, fairer — ML, we propose "games for fairness and interpretability": networked games that as a byproduct of the game's objectives, engage the general public in auditing algorithms while simultaneously generating valuable training sets for ML developers.

2 ML Powered Games

Inspired by Luis von Ahn's Games with a Purpose (GWAP) framework (Von Ahn (2008); Von Ahn & Dabbish (2008)), we propose using ML-powered games to enhance model interpretability — which we view as an important step towards developing fairer ML.



(a) Example of a *Humans vs. AI* game. Player 1 provides an input, while Player 2 competes against an AI to produce the correct answer.



(b) Example of a *Break the Bot* game. Player 1 and Player 2 compete against each other in producing adversarial attacks that will reduce the accuracy of the model's predictions. In this example, players can change the lighting and color, or add and remove common objects.

Figure 1: Both types of games are designed to surface model biases and deficiencies, while also producing more robust and diverse training data.

2.1 Games with a Purpose

Described as "human computation", the GWAP framework was designed for problems solvable by humans but beyond the capabilities of machines. Instead of relying on financial incentives or altruism, GWAPs simply rely on people's desire for fun and entertainment. A successful GWAP can produce not only novel and creative solutions to difficult problems, but also provide large amounts of labeled data for training machine learning models. Since its inception, GWAPs have attracted hundreds of thousands of players in order to tackle problems ranging from protein folding (Khatib et al. (2011)) and RNA folding (Lee et al. (2014)) to examining the human perception of correlation in scatter plots[3].

The GWAP framework includes several different templates of games (Von Ahn & Dabbish (2008)). *Output-agreement* games has two players attempt to produce the same output when shown the same input. In the ESP game, for example, the players are shown an image and asked to guess what words the other player would use to describe the image. A variation of the game includes taboo words for each image, thus requiring users to guess more uncommon words, in turn producing more interesting labeled data (Von Ahn & Dabbish (2004)). In *input-agreement* games, two players are each provided an input which may or may not be different; the players are asked to output descriptions of the inputs and then finally guess whether they were shown the same input. For instance, players in the Tagatune game are given song clips and asked to output tags, before finally guessing whether they had the same clip (Law & Von Ahn (2009)).

[3]http://guessthecorrelation.com/

2.2 Designing Games for Fairness and Interpretability

While reputation-based incentives can create social pressure and motivate ML developers, we believe a well-designed game aligns incentives between ML developers and the consumers of ML (i.e. the general public). Due to the importance of labeled data for deep neural networks, we believe ML researchers will have strong incentives to upload their models if the games that leverage them can produce valuable training data or adversarial examples.

On the consumer side, GWAPs have shown that such games can reach large audiences. Furthermore, a larger audience is often a broader audience, thus allowing more diverse probing of the model. We believe that there is an appetite for ML games, due both to increasing media attention on ML and the growing capabilities of new models. Recent examples of games that engage a general audience in exploring ML include the Pictionary-like game Quick, Draw![4], word embedding-powered word association games[5], and an endless text-adventure game built using a generative text model[6].

We define "games for fairness and interpretability" as ML-powered games in which the output and / or interaction with human players is produced by a machine learning model. These games can also be networked to enable human-human interaction and competition. Games should be fun and engaging, provide insight into how the underlying machine learning models work, and produce data that helps models improve — in particular, so that the models are better-equipped to more equitably serve a diverse range of individuals and scenarios.

One might imagine a platform for such games, where once a game has been designed and open-sourced, its backend model could be swapped for any model with similar inputs and outputs. The platform could also serve as a public forum for widespread participation in, and discussion about, the evaluation of new ML models. This unique forum — one where both ML developers and members of the public are present — could serve as an important vehicle for a) enhancing broader familiarity with and awareness of ML and its applications, and perhaps eventually, b) creating proactive public pressure that motivates algorithm developers to build more interpretable and fairer ML.

2.3 Proposed Categories of Games

In the spirit of GWAPs, we describe possible categories of games in the following sections.

2.3.1 Humans vs. AI

Setup. Player 1 provides an input, and Player 2 competes against an AI to guess the correct answer.

Example game 1 — Guess Who? Player 1 describes themselves, their interests, job, and other attributes through freeform short text. Player 2 and the AI attempt to guess the age, sex, and location of Player 1.

Example game 2 – Codenames. Inspired by the popular Codenames board game Wikipedia contributors (2020) , the players are presented with a 5x5 grid of words. Player 1 is a "spymaster" who is also allowed to see the placement of bombs on the grid. The spymaster's role is to give a one word clue, plus the number of words that matches the clue. Player 2's goal is to guess the correct words; however, if he or she guesses a bomb, the game is over. The game is won if all the non-bomb words are guessed correctly. The goal is to finish the game in fewer rounds; saying a larger number allows the team to win more quickly, but it is also more difficult to come up with clues.

In our ML-powered variant, the AI also attempts to guess the words; if the AI's guesses matches Player 2's guesses, those guesses are invalid. Figure 1a shows an example round.

Data produced and insight into interpretability. Player 1 will have to produce inputs that are recognizable by another human but undetectable or incorrectly classified by the AI. This requires a player to intuit the space of inputs that a model understands and in which cases it might fail. For instance, Player 1 may find that cultural references are harder for a ML model. Natural language processing models that can incorporate common sense reasoning and knowledge also remains an open area of research. The successful inputs and clues can be used as more robust training data.

[4] https://quickdraw.withgoogle.com/

[5] http://robotmindmeld.com/

[6] https://www.aidungeon.io/

In addition, baseline models for the AI could be based on word embeddings, which have been shown to reflect implicit human biases around gender, race, occupation, etc. (Caliskan et al. (2017)). These biases may be surfaced if the AI incorrectly relies on them to make predictions.

2.3.2 BREAK THE BOT

Setup. Each player is shown an input and the model's output (e.g. a prediction). Each player is asked to make a small modification to the input. Whoever can cause the largest change in the model output, while using the smallest modification, receives more points.

Example game — Vandalize it! The brittleness of deep neural networks has been illustrated in several computer vision systems. For example, graffiti on signs can significantly lower object recognition accuracy (Eykholt et al. (2018)), while Rosenfeld et al. showed that adding an object to a scene could drastically change the ability to recognize all other objects (Rosenfeld et al. (2018)). These deficiencies can have catastrophic effects on real-world systems.

In this self-driving car inspired game, players are shown street images overlaid with bounding boxes of detected objects. For example, a stop sign may be detected by the model with probability 0.85. The players' goal is to change that prediction by making small edits to the sign and its surroundings. The game will give players tools to alter the angle, lighting, hue of the image, as well as add and subtract other objects and artificats. (The game will have to measure the 'size' of modifications in order to assign scores). Figure 1b shows an example of how the game might look.

Data produced and insight into interpretability. These games provide adversarial examples and sensitivity analysis on model inputs. This is important as the field of adversarial examples is becoming increasingly important (Goodfellow et al. (2014)), especially as ML models become deployed in the real world (Kurakin et al. (2016)), and obtaining those examples can often be difficult (Zhao et al. (2017)). ML researchers can also gain a greater understanding of how inputs may be modified in semantically meaningful ways, as well as if the observed model behavior is desirable (e.g. fair).

3 CONCLUSION

As ML-powered technologies continue to proliferate, the threat of biased and opaque decision-making looms large. We believe public pressure is a powerful mechanism for inspiring changes in how algorithms are developed. Games for fairness and interpretability provide one means for engaging the public in probes of ML systems while simultaneously producing hard-to-source data that serves the interests of ML developers. We believe games are unique in their ability to engage different audiences and are thus a promising avenue in which to pursue complicated, multi-stakeholder challenges like building fairer ML systems.

We note that games can help augment existing trends in ML research. Thus far, approaches to operationalize fairness include learning *fair* representations that factorize out sensitive attributes (Zemel et al. (2013)), allowing the use of the sensitive attributes but aiming for "equality of opportunity" (Hardt et al. (2016)), and more. At times, these methods have been guided by human definitions, such as the 80% rule of "disparate impact" outlined by the US Equal Employment Opportunity Commission as a definition of discrimination (Feldman et al. (2015)). Games can shed light onto how fair ML models may appear in practice, as well as what notions of fairness that humans care about, which can in turn be formalized into better models. On the interpretability side, many methods have centered on introspection and visualization, such as inverting representations to generate images (Mahendran & Vedaldi (2015); Mordvintsev et al. (2015)) and producing saliency maps (Ribeiro et al. (2016); Sundararajan et al. (2017)). However, there are questions around the reliability and intuitiveness of these explanations (Jain & Wallace (2019); Kindermans et al. (2019)). The games' data can be analyzed through these methods, perhaps providing insight into how well current explanations match human intuitions.

Looking ahead, there are several open questions: who should be responsible for designing and developing games for fairness and interpretability? How will the games be deployed and marketed so as to recruit a diverse range of players? What new risks or threats might these games introduce? These are important questions that will require continuous exploration and reflection. We hope this paper serves as an initial stepping stone and inspires individuals both within and beyond the ML community to consider the potential power of games.

REFERENCES

David Autor. Why are there still so many jobs? the history and future of workplace automation. *Journal of Economic Perspectives*, 29(3):3–330, 2015.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pp. 4349–4357, 2016.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research, Conference on Fairness, Accountability, Transparency*, pp. 1–15, 2018.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Ziv Epstein, Blakely H. Payne, Judy Hanwen Shen, Abhimanyu Dubey, Bjarke Felbo, Matthew Groh, Nick Obradovich, Manuel Cebrian, and Iyad Rahwan. Closing the ai knowledge gap. *arXiv preprint arXiv:1803.07233*, 2018.

Virginia Eubanks. *Automating Inequality: how high-tech tools profile, police, and punish the poor*. New York, NY : St. Martin's Press, 2018.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1625–1634, 2018.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268. ACM, 2015.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Firas Khatib, Frank DiMaio, Seth Cooper, Maciej Kazmierczyk, Miroslaw Gilski, Szymon Krzywda, Helena Zabranska, Iva Pichova, James Thompson, Zoran Popović, et al. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature Structural and Molecular Biology*, 18(10):1175, 2011.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280. Springer, 2019.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

Edith Law and Luis Von Ahn. Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1197–1206. ACM, 2009.

Jeehyung Lee, Wipapat Kladwang, Minjae Lee, Daniel Cantu, Martin Azizyan, Hanjoo Kim, Alex Limpaecher, Snehal Gaikwad, Sungroh Yoon, Adrien Treuille, et al. Rna design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences*, 111(6):2122–2127, 2014.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. 2015.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. *Google Research Blog. Retrieved June*, 20(14):5, 2015.

Safiya Umoja Noble. *Algorithms of oppression : how search engines reinforce racism*. New York : New York University Press, 2018.

Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM, 2016.

Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3319–3328. JMLR. org, 2017.

Joaquin Vanschoren, Jan N Van Rijn, Bernd Bischl, and Luis Torgo. Openml: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

Luis Von Ahn. Human computation. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pp. 1–2. IEEE Computer Society, 2008.

Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 319–326. ACM, 2004.

Luis Von Ahn and Laura Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, 2008.

Wikipedia contributors. Codenames (board game) — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Codenames_(board_game)oldid` $= 936348738, 2020. [Online; accessed 20 - January - 2020]$.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*, 2017.