Games for Fairness and Interpretability

Games for Fairness and Interpretability

Eric Chu*
echu@mit.edu
Massachusetts Institute of Technology

Nabeel Gillani*
ngillani@mit.edu
Massachusetts Institute of Technology

Sneha Priscilla Makini
snehapm@mit.edu
Massachusetts Institute of Technology

ABSTRACT

As Machine Learning (ML) systems becomes 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. They are inherently engaging, provide insights into how machine learning models work, and ultimately 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 machine learning community could benefit from them.

CCS CONCEPTS

• **Human-centered computing**;

KEYWORDS

machine learning, interpretability, fairness, games, crowdsourcing

ACM Reference Format:
Eric Chu, Nabeel Gillani, and Sneha Priscilla Makini. 2020. Games for Fairness and Interpretability. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3366424.3384374

1 INTRODUCTION

As ML increasingly permeates virtually all aspects of life — and unequally serves, or fails to serve, certain subsegments of the population [4–6] — 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 [1] have hinted at the dangers of simply letting history unfold. In all of these instances, there were certainly changes to the underlying technological methods — but it is hard to deny the importance of collective public pressure in catalyzing dialogue to envision a new set of policies and practices surrounding these powertools. It is unlikely that methodological changes alone would have been sufficient. 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 [12, 32, 33]. Furthermore, initiatives like Turingbox [11] and OpenML [39] 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.

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 [40, 42], we propose using ML-powered games to enhance model interpretability — which we view as an important step towards developing fairer ML.

*Both authors contributed equally.

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366424.3384374

[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

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 [22] and RNA folding [27] to examining the human perception of correlation in scatter plots[3]. The framework has also since been extended to machine learning, such as using active learning to select examples during gameplay [2].

The GWAP framework includes several different templates of games [42]. *Output-agreement* games has two players attempt to produce the same output when shown the same input. In the ESP game, shown in Figure 1, 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 [41]. 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 [26].

Figure 1: An example of a Game With a Purpose (GWAP): the original ESP game.

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 text auto-complete "Talk to Transformer"[4], a Pictionary-like game Quick, Draw![5], word embedding-powered word association games[6], and an endless text-adventure game built using a generative text model[7].

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.*

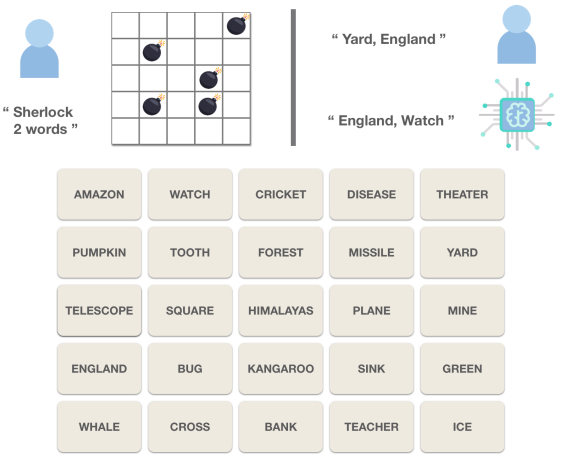
Setup. Player 1 provides an input, and Player 2 competes against an AI to guess the correct answer.

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.

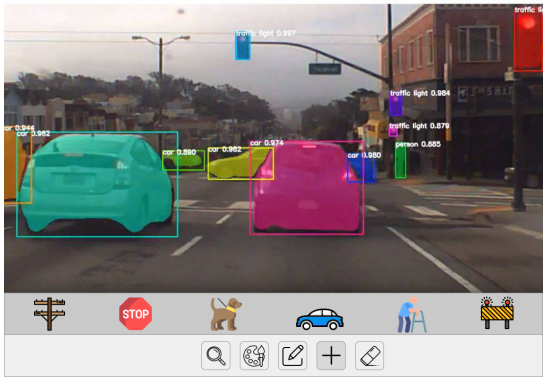
Example game 2 – Codenames. Inspired by the popular Codenames board game [44] , the players are presented with a 5x5 grid of words. Player 1 is a "spymaster" who is also allowed to see the placement of bombs on the grid. The spymaster's role is to give a one word clue, plus the number of words that matches the clue. Player 2's goal is to guess the correct words; however, if he or

[3] http://guessthecorrelation.com/

[4] https://talktotransformer.com/
[5] https://quickdraw.withgoogle.com/
[6] http://robotmindmeld.com/
[7] https://www.aidungeon.io/



(a) Example of a *Humans vs. AI* game. Player 1 (the "spymaster") provides an input, while Player 2 competes against an AI to produce the correct answer. Here, since the human and the AI both guessed "England", only "Yard" would count as a 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. Players should be incentivized to make small edits that nevertheless produce large decreases in accuracy. In this example, players are provided tools to change the lighting and color, or add and remove common objects.

Figure 2: Examples of possible Games for Fairness and Interpretability. Both types of games are designed to surface model biases and deficiencies, while also producing more robust and diverse training data.

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 2a 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.

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. [6]. 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 1 — 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 [13], while Rosenfeld et al. showed that adding an object to a scene could drastically change the ability to recognize all other objects [36]. 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 2b shows an example of how the game might look.

Example game 2 — Beat the Banker. ML has begun to be used in higher-stakes situations, ranging from recidivism prediction to loan default rate prediction. Unfortunately, these systems have also been shown to be susceptible to demographic features and unfairness [17]. In this game, the players are bankers. The input is a hypothetical set of demographic features of an individual, and the output is the predicted probability of that individual's loan repayment. Faced with a loan rejection, the goal is to find seemingly innocuous changes that can make the loan approved.

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 [16], especially as ML models become deployed in the real world [25], and obtaining those examples can often be difficult [46]. 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 GAMES AND CURRENT RESEARCH DIRECTIONS IN MACHINE LEARNING

The previous section illustrates how thoughtfully-designed games might help align incentives between ML developers and the general public, cultivating public pressure and awareness — along with the new, more representative datasets — to promote fairer, more inclusive ML systems. We believe the time to develop games for fairness and interpretability is now, largely because they align with several current directions in ML research. We highlight some of these directions below and explore how members of these respective research communities may benefit from games for fairness and interpretability.

3.1 Fairness

As ML models become more pervasive, there has been an increasing call for models that can prevent discrimination along sensitive attributes such as race and gender. Part of the problem is detecting that biases in models even exist in the first place. To that end, recent research has shown how word embeddings encode biases as measured by standard tests such as the Implicit Association Test [6], with relationships between word embeddings reflecting negative stereotypes about gender [4]. Other work highlights deficiencies in datasets used for facial recognition, resulting in models that fail more frequently for women and people with darker skin tones [5].

How can models handle these sensitive attributes? A naive approach of removing sensitive attributes may not prevent discrimination if the sensitive attributes are correlated with other attributes left in the dataset. Enforcing demographic parity, in which the outcome is uncorrelated with the sensitive attribute, is also problematic because it does not guarantee fairness, and the sensitive attribute may actually be important for prediction, making removal of all correlation unrealistic. Thus far, various approaches to formalize and operationalize fairness include using the 80% rule of "disparate impact" outlined by the US Equal Employment Opportunity Commission as a definition of discrimination [14], treating similar individuals similarly by enforcing a Lipschitz condition on similar individuals and the classifier predictions for those individuals [10], preprocessing the dataset through methods such as weighting and sampling [21], and allowing use of the sensitive attribute but aiming for "equality of *opportunity*" through the notion of equalized odds [17]. Certain frameworks also provide the ability for people to select the tradeoff between model performance and fairness. Other work has centered on learning transferable *fair* representations that can be reused across tasks [45].

We believe ML fairness researchers would find value in the datasets produced by games for fairness and interpretability. For example, machine predictions from "Human vs. AI" games would provide clear insights into which kinds of biases certain algorithms harbor; "Break the Bot" games might shed light on how robust or brittle algorithms are to changes in the datasets they operate on.

3.2 Interpretability

While deep neural networks have found great success as powerful function approximators, they have also developed a reputation as black boxes. Interpretability may be a case of "you know it when you see it", but recent work has attempted to make the problem more tractable by defining interpretability, explaining why it is important, and explaining when it is necessary [9, 29].

There has also been a wide range of methods focused on *introspection and visualization*, including (but not limited to) "inverting" intermediate representations to generate images [30, 31], producing input feature attributions and saliency maps [3, 34, 35, 37, 38] vs. producing counterfactual explanations [43] vs. pointing to protoypical examples [7], local per-example explanations [35] vs. global explanations based on feature representations across the entire dataset [3], clear-box approaches with access to model gradients [37, 38] vs. black box approaches [34, 35]. These methods often highlight what parts of the input (e.g. a segment of the image, or a span of the text), were most important to the model's decision.

While there is also work worth mentioning on (1) generating human readable explanations in natural langauge [18, 28], (2) distilling neural networks into more interpretable models such as decision trees [15], and (3) disentangling factors of variation for generative models [8, 19, 24], a significant portion of the field has focused on the aforementioned introspection and visualization methods. At the core, many of the methods attempt to relate input or internal representations to the model outputs. However, there are questions around the reliability and intuitiveness of these explanations ([20, 23]). The games' data can be analyzed through these methods, perhaps providing insight into how well current explanations match human intuitions. The "Break the Bot" games would also produce valuable counterfactual data; analyzing changes in the outputs of their underlying models as a function of changes to inputs could provide a deeper understanding of how, exactly, these models are conducting their computations.

4 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.

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

[1] David Autor. 2015. Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *Journal of Economic Perspectives* 29, 3 (2015), 3–330.
[2] Luke Barrington, Douglas Turnbull, and Gert Lanckriet. 2012. Game-powered machine learning. *Proceedings of the National Academy of Sciences* 109, 17 (2012), 6411–6416.

[3] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 3319–3327.

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

[5] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of Machine Learning Research, Conference on Fairness, Accountability, Transparency*. 1–15.

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

[7] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This looks like that: deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*. 8928–8939.

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2172–2180.

[9] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. (2017).

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 214–226.

[11] Ziv Epstein, Blakely H. Payne, Judy Hanwen Shen, Abhimanyu Dubey, Bjarke Felbo, Matthew Groh, Nick Obradovich, Manuel Cebrian, and Iyad Rahwan. 2018. Closing the AI Knowledge Gap. *arXiv preprint arXiv:1803.07233* (2018).

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

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

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

[15] Nicholas Frosst and Geoffrey Hinton. 2017. Distilling a Neural Network Into a Soft Decision Tree. *arXiv preprint arXiv:1711.09784* (2017).

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

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

[18] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *European Conference on Computer Vision*. Springer, 3–19.

[19] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework. (2016).

[20] Sarthak Jain and Byron C Wallace. 2019. 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)*. 3543–3556.

[21] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

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

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

[24] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. 2015. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*. 2539–2547.

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

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

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

[28] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155* (2016).

[29] Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).

[30] Aravindh Mahendran and Andrea Vedaldi. 2015. Understanding deep image representations by inverting them. (2015).

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

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

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

[34] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421* (2018).

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. 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*. ACM, 1135–1144.

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

[37] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).

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

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

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

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

[42] Luis Von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Commun. ACM* 51, 8 (2008), 58–67.

[43] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[44] Wikipedia contributors. 2020. Codenames (board game) — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Codenames_(board_game)&oldid=936348738. [Online; accessed 20-January-2020].

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

[46] Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2017. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342* (2017).