Analogues of Bias in Neural Networks

Multilayer neural networks loosely modeled off brain structure have heralded the most recent “AI Spring”, in which there is an explosion of enthusiasm and funding for AI research. Lofty predictions of artificial general intelligence (superintelligent computers able to solve a wide range of problems) less than twenty years away are common, as they have been in previous AI springs. But deep learning models have the same shortcomings as previous AI techniques; they show brittleness and are only effective in narrow applications of intelligence. Headlines display ‘AI mimics human speech’ or ‘AI beats Go Grandmaster’, however the models that performed these tasks are hyperfocused on these tasks and are not well-adapted to generalization. The Go-playing AI has no clue how to recognize the Go pieces on the board, and an AI mimicking human speech does not understand the speech itself. These models learn by processing a massive amount of data. They try to predict output based on a numerical representation of the data and modify their behaviors based on how far away they are from the desired outcome. The distance between the model output and the desired result is computed using a **utility function**. Utility functions are used to nudge learning algorithms in the right direction, acting like a rubber band pulling the model’s output towards the correct prediction. Computers have often been advertised as the pinnacle of objectivity because a silicon chip cannot develop the same biases that humans do. If computers can be truly objective, a computer superintelligence will be the last problem we ever need to solve, as afterwards we can just ask the superintelligence for the answers to every other problem. But we have already seen many examples of algorithmic bias in cases such as PredPol and Tay, which both exhibited harmful racist biases after being deployed. These cases could be justified by relegating the biases to the data to retain the idea of objectivity in computer models. But in “Canons of Algorithmic Inference: Feminist Theoretical Virtues in Machine Learning”[1], Gabbrielle Johnson argues that inductive learning models cannot be dragnet[2] objective. That is, they require some canons of inference to guide them through underdeterminism. This argument follows from Hume’s problem of induction, which is that there seems to be nothing to justify basing predictions of unobserved instances on observed instances. The only way in which induction works is by making some fundamental assumptions. Johnson argues that Hume’s problem of induction means that learning models fundamentally cannot be dragnet objective, as they must hold some assumptions in order to perform induction.

Johnson establishes that machine learning algorithms are inductive decision making programs, thus cannot be dragnet objective and must exhibit some bias. Interestingly, empirical evidence[3] shows that these learning models do not just exhibit biases, their behavior closely mimics that of human implicit biases. Learning models can be designed in a biased way, but I will ignore these models under the assumption that human brains are not inherently biased and instead learn their bias from the environment[4]. Learning model’s biases stem from their training data and are learned during the training process, similar

[1] Gabbrielle Johnson, *Canons of Algorithmic Inference: Feminist Theoretical Virtues in Machine Learning*, (Springer Nature, 2020)

[2] Dragnet comes from a 60s cop show by the same name, in which the main character supposedly considers “just the facts”, without any corruption by personal motivation or emotion.

[3] Jake Silberg and James Manyika, *Tackling bias in artificial intelligence (and in humans)* (McKinsey & Company, 2019), https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans

[4] This assumption can be justified by the contextual nature of biases (people in China have different biases than people in United States)

to how humans learn biases in their own ‘training’. A simplistic model of learning using an objective function can be applied to both humans and algorithms. In the same way that an AI tries to maximize their utility function, humans perform actions in ways that try to minimize cost and maximize utility. The apparent behavioral differences come from the flexibility of human objective functions based on contextual values, but that is a conversation for another essay. To explain the human-algorithm analogy, we can draw similarities between the features of human learning and that of the algorithm. In general, humans and learners perform as inductive decision makers, making predictions and modifying predictions based on learned results. A large negative result from a utility function in a human would be represented by strong negative emotions like guilt or embarrassment. We feel these emotions in cases where our predictions are far off from reality, just like a learning algorithm experiences a large cost when it is very wrong. Conversely, we feel proud when our predictions are correct, just like how a model receives a reward when it does better. The utility function model can be used to explain seemingly idiosyncratic qualities of human implicit bias. A strange feature of implicit bias in humans is that it is not easily corrected, either through scenarios of embarrassment or even bias training courses. Take the example from class, where a woman attends a club event and mistakenly assumes a man there is the coat attendant, when the man is actually there to receive an award. The woman made her decision based partly on the fact that the man was black, whereas most of the club members are white and the staff is mixed race. The woman’s inductive learning model was biased, and led her to the prediction that the black man was a coat attendant. Now to continue the scenario, suppose the man corrects her, and later in the night is called up to the stage to receive his award. The woman, being conscientious, feels a strong sense of embarrassment for her incorrect prediction. She makes a promise to herself to never make predictions based on race again. But a few months later, a similar situation occurs. Again, the woman feels embarrassed and promises to herself not to repeat the behavior caused by the bias. But the cycle repeats. So why do humans struggle to unlearn implicit biases if the cost associated with each event is so high? The answer lies in the behavior of machine learning models in the same situation. These types of situations of extreme embarrassment, or when the utility function exerts a high cost, can be classified as **extreme rare events** and are extremely difficult problems for deep learning models[5]. This is because rare events are difficult to teach; even a large negative utility does not cause much effect if it is rare compared to the rest of the training data.

Implicit bias training has been used as an attempt to correct these resilient biases. Unfortunately it has not been effective at eliminating bias either, in some cases actually worsening the situation by causing frustration and anger[6]. Since the model’s goal is to maximize the utility function’s score, it can get away with a bias because the majority of the time biased predictions do not exert a cost. The events where the bias actually decreases utility are rare and so the model is unable to learn them. This model can be applied to human bias as well. In most cases, the powerful embarrassment of the woman does not happen frequently and so the utility function cannot exert enough force to significantly modify the predictive model. Another analogous feature is memory. Neural networks by nature forget data that they have not processed in a while. When the objective function is focused around a different task, the network becomes optimized for the new task and forgets the old. The same theory applies in humans as well. We cannot remember events from five years ago nearly as well as events from yesterday. We adapt to new information as it is collected, and so no matter how powerful a stimulus, its effect is mitigated over time.

[5] Di Qi and Andrew Majda, *Using machine learning to predict extreme events in complex systems*, (PNAS, 2020), https://www.pnas.org/content/117/1/52

[6] Tiffany Green and Nau Hagiwara, *The Problem with Implicit Bias Training*, (Scientific American, 2020), https://www.scientificamerican.com/article/the-problem-with-implicit-bias-training/

In bias, this model predicts the difficulty of overcoming implicit bias; there are not enough instances of cost associated with the bias to learn that the bias is bad. In the case of implicit bias training, the model works as well. Suppose we have a trained learning model, and we suddenly feed it a set of new data. It will exhibit some learning from the new data, but as soon as we revert to the old stream of data it will re-learn its old behavior. The same is true in humans; as soon as the implicit bias training ends the old stream of data resumes, retraining the old biases. Also, If a trained learning algorithm is suddenly given a set of new training data, it will not perform as well on the data as a whole and its accuracy will drop. This accuracy drop is synonymous with the frustration that can occur from mishandled implicit bias training. It causes 'confusion' (real in humans, metaphorical in algorithms) about the world, and the utility will be lower overall.

An interesting counter is in cases where the cost of holding a bias is experienced more frequently, and cannot be explained by the inability of learning models to adapt to rare events. This is supposedly where the behavior of algorithms and humans begins to diverge. With an increased frequency of anomalous training examples, a model will inevitably begin to modify its behavior to suit the new data, analogous to losing its bias. But evidence shows that even in cases where biases are frequently challenged, humans often refuse to give them up[7]. Why the sudden divergence in behavior? The behavior is different in these cases only because of some mental gymnastics performed by the brain. Instead of understanding that every case of bias is actually bias, we instead attribute each instance of bad prediction to different causes, thus not giving a direction for the utility function to 'pull' the model. It's like if we gave a model that was trained to detect a different species of fish a bunch of different images of mammals. The model receives a high cost from its objective function for misclassifying the mammals, but because each picture is a different animal, the model cannot learn to improve its accuracy on the new images. The high cost associated with each image is due to different features, and so the model cannot pick a direction to move to increase its accuracy. To use the example of the woman at the coat check, suppose instead of realizing her bias she justifies her mistake with things like "oh he was standing close to the coat rack". In each instance where bias exerts a cost, the cost is instead attributed to some proxy that confuses the utility function and prevents learning.

In this essay, I covered Johnson's argument for why algorithmic learning necessarily holds biases, and thus cannot be completely objective. I then covered how biases in machine learning are closely related to implicit human biases, as well as how modeling human biases using a utility function can predict the resilience of bias and even account for the lack of effectiveness of implicit bias training.

[7] Green and Hagiwara, *The Problem with Implicit Bias Training*