The Dangers of Algorithmic Autonomy: Efficient Machine Learning Models of Inductive Biases Combined With the Strengths of Program Synthesis (PBE) to Combat Implicit Biases of the Brain

The Dangers of Algorithmic Autonomy:
Efficient Machine Learning Models of Inductive Biases combined with the Strengths of Program Synthesis (PBE) to Combat Implicit Biases of the Brain

A Thesis Presented
By
Sumona Halder
To
The Department of Computer Science and Mind Brain Behavior

In Partial Fulfillment of the Requirements for the Degree of Bachelor of Arts in the Subject of Computer Science on the Mind Brain Behavior Track

Harvard University

April 10th, 2020

Table of Contents

ABSTRACT 3

INTRODUCTION 5

WHAT IS IMPLICIT BIAS? 11

Introduction and study of the underlying cognitive and social mechanisms 11

Methods to combat Implicit Bias 16

The Importance of Reinforced Learning 19

INDUCTIVE BIAS and MACHINE LEARNING ALGORITHMS 22

What is Inductive Bias and why is it useful? 22

How is Inductive Reasoning translated in Machine Learning Algorithms? 25

A Brief History of Inductive Bias Research 28

Current Roadblocks on Research with Inductive Biases in Machines 30

How Inductive Biases have been incorporated into Reinforced Learning 32

I. The neuroscience - how inductive bias and reinforced learning are closely related 32

II. The algorithm - how inductive bias and reinforced learning are closely related 35

How implicit bias sneaks into machine learning algorithms 37

PROGRAM SYNTHESIS 41

CONCLUSION 43

When Bad Algorithms Outperform Good Ones 45

The Union of ML (in an inductive framework) and PBE 46

BIBLIOGRAPHY 48

ABSTRACT

In current research surrounding machine learning algorithms, we have faced a large ethical and moral issue. Our algorithms, which are intended to provide us with the most optimal and unbiased decision has started to emulate the implicit biases of humans. This has bled into our society in destructive ways as the widespread applications of these algorithms have been implemented in crucial decisions of our lives such as jobs, bank loans, and college admissions. However, due to lack of proper training data and the inherent nature of implicit bias, we have seen that reversing this in algorithms is quite challenging. Inductive bias on the other hand, offers some insight as how we can generalize from specific examples and maximize future predictions. While inductive reasoning is not immune to being affected by implicit bases, it can be used to properly train algorithms to produce better outcomes.

In this paper, we will explore a theoretical solution through a mechanism-first (building upon a foundation of cognitive processes) strategy. We will explore ways to better implement inductive reasoning to combat implicit bias. While a few solutions and strengths of larger training sets and reinforced learning along with inducive bias are explored, we make a case for a model that combines the strengths of machine learning and programming by examples to tackle such issues.

INTRODUCTION

Imagine yourself in your mid-to-late twenties and you have been working for the past few years to save up for your dream home. However, you are going to need a little help from the bank in the form of a loan to pay off the house, as most people do. You have a well-paying job, no outstanding credit, and have been diligently tracking your expenses to ensure no erroneous spending. You go down to the bank with all of your documents in hand eager to sit down with a representative from the bank, eager to start a conversation about a possible loan. After reviewing your files, the representative informs you that although all of this information passes their system, they are unable to give you a loan at this time due to the low-income neighborhood you lived in when you were in middle school.

Now, as most can imagine, denying a current eligible person a loan based on a piece of society they occupied is not only illegal, but also quite ridiculous and discriminatory. If someone walked into a bank today and was told this, we could see why claims of discrimination would be valid in this situation. Low-income neighborhoods, though there might be some connection to one not being able to pay back a bank loan if they currently live in an area with a certain income bracket, might imply certain socio-economic and racial perceptions as low-income areas tend to be associated with communities of color.

Denying someone a loan based on a claim rooted in stereotypes and generalization would be considered a discriminatory act, as would decisions that target similarly marginalized attributes such as gender, sexual orientation, and religion. This idea of letting certain underlying biased

ideologies influence your decision, in this case withholding a loan from an eligible candidate due to an indirect discriminating factor, is known as harboring an implicit bias.

Motivation

Implicit biases, despite one's best intentions, can cause stereotypes to influence decisions unknowingly and result in severe consequences. From bank loans to job hiring and criminal sentencing, these decisions may be heavily influenced by this cognitive process thus ultimately influencing all of our lives. To combat such biases within humans, we have implemented artificial intelligence (AI) that uses machine learning (ML) algorithms to find relationships between large datasets and predicted behaviors.

The first response was to eliminate demographic markers that could be a basis for prejudice. The first AI models explicitly race and gender omitted in attempt to rid their predictions and assumptions from possible human-induced stereotyping and bias. However, other factors and attributes can also represent lingering forms of implicit biases. Once incorporated into existing models and algorithms, they still play a discriminatory role in influencing decision-making. The statistical techniques and underlying correlations at the core of current ML techniques in AI do not bode well for new big data because they inject a "proxy discrimination" or false predictions on the behaviors of certain classifications of people [1]. Finding and justifying statistical significance in inexplicit qualities is not enough to remove the context and structural connections rooted in historical discrimination.

The use of these ML methodologies raises the question of whether or not incorporating new data is the right way forward for influencing decision-making without implicit bias. For example, another tactic that some banks use to determine their loan decisions is the applicant's use of a Mac or PC. With the advent of tracking software, the bank websites can identify the operating system of the device that is accessing information about the loan. Thus, whether the user uses iOS or Android has often become a determining digital footprint that contributes to the decision to give or reject the loan application [2]. There are strong ethical implications with using this sort of data, even if it does not have a clear and direct connection to attributes like gender or race. On the surface, there could be statistically significance in the type of laptop or phone an applicant uses with their likelihood to default on a loan. But we must interrogate the inherent racial and socio-economic undertone to the devices and brands that a person uses. Indirect connections like these would result in the same bias that directly connect to the outcome and its consequences.

In addition, while most rejections of loans do require a more elucidated and step-by-step show of process of how the conclusion was reached, it is much more difficult for AI to do this. Many machine learning algorithms function like a "black box" because that they come to a conclusion without leaving a clear trail as to why they made those particular pathways. This problem becomes increasingly complicated as ML begins to incorporate new metrics and data that may be considered unusual or tangentially related to the main decision and behavior at hand [3].

One way in which that we have programmed machines to trace their pathways is to introduce inductive bias which is modeled after the way humans can successfully make generalizations from specific examples. Though the word bias has a negative connotation to it,

inductive bias is an essential part of both human learning and machine learning. Inductive reasoning is the process of learning general principles on the basis of specific instances—meaning that any machine learning algorithm will usually produce a prediction for any unseen test instance on the basis of a finite number of training instances [8]. This type of reasoning is exactly why we are able to give algorithms complex data, and have it come to accurate predictions based off of patterns that we ourselves would make. However, while the algorithms that make predictions from more objective, logical statistics have been revolutionary in tackling complex data, there is still an issue with algorithms leading to ethically questionable decisions when their analyses are premised on faulty assumptions or mis-measured data (hence by our implicit bias) [9].

Overview

Thus in this work, we aim to refine current ML methodologies by not only exploring how to strengthen inductive reasoning against our implicit biases on an input of diverse data, but also by creating a model that has the clarity and step-by-step breakdown of factors that were used to come to such a decision rather than the "blackbox" that most machine learning algorithms use. My main contribution of this thesis is a theoretical bottom-up construction of an end-to-end machine learning algorithm that can accurately and fairly evaluate consumer behavior, and come to transparent conclusions.wewill introduce the idea of program synthesis and how this can help with an ability to incorporate more efficient reinforced learning modeled after our own cognitive processes.

The issue of trying to tackle such "racist" algorithms has been long sought over in the past few years. Many related academic works have been published trying to crack the "blackbox" of

machine learning algorithms and trying to find a better model. In addition, there have been many works published on this topic that combine the ideas of cognitive neuroscience, program synthesis, and machine learning methodologies to create a new algorithm to solve this issue. Our goal is to refine these theses, explore the scope of program synthesis to incorporate the complexities that ML can handle, and look for models to reinforce battling implicit bias using the cognitive processes in our brain that do this.

We will move forward in building this theoretical model by exploring implicit bias from a cognitive neuroscience perspective, and understanding the mechanisms in the brain surrounding this and how we can train our own brain to combat implicit biases. We will then explore our "positive bias" or inductive bias and how this has been the cognitive process by which machine learning has been largely modeled after (and why this is so). We will observe the benefits of our ability to generalize from examples and how this can be applied to large complex data. There will also be an analysis of current works that have implemented machine learning algorithms that have ended up with the unintended consequence of incorporating factors that would be considered unethical in the evaluation of a candidate for various observable situations. We will dive deeper into the inner workings of such an algorithm and why, though they are sophisticated model, the inability to edit such a program makes it extremely hard for us to determine why an algorithm comes to the conclusions it did and therefore solve the issue.

Following this, we will introduce the idea of program synthesis, compare it to machine learning, and focus on the strong suits of these types of models. Lastly, we will combine all of this knowledge and exploration to not only examine works that take these ideas to create a better

algorithm, but contribute to this research by building our own end-to-end theoretical model with concepts that would be helpful to consider.

WHAT IS IMPLICIT BIAS?

Introduction and study of the underlying cognitive and social mechanisms

Implicit bias, as stated in the introduction of this paper, generally refers to the impact that attitudes and stereotypes have on our unconscious minds, which therefore actively affect our decisions and actions with regards to certain topics. Such associations develop over the course of a lifetime as we garner exposure to direct and indirect messages, media, news, and each other [4]. It is imperative that we understand the cognitive and social mechanisms that reinforce such deep-rooted biases within ourselves so that we can tackle the ethical and structural barriers such a phenomenon motivates.

In order to better understand the mechanisms of inductive bias and current ML models that imitate implicit bias in human brain, we must first fully grasp the subtle cognitive processes that create this phenomenon. Stigma or bias can be categorized in two types:

1. external which is bias that the general population holds about individuals identified as members of the stigmatized group, and
2. internal or a "self" stigma in which the external stigma is internalized by a member of the stigmatized group [5].

And while explicit biases can be easily monitored, caught, and corrected, implicit biases occur as an unconscious attitude that not only can be built up since early childhood, but is also harder to correct.

Though there are many different theories for how social decision-making occurs in the brain, we have an understanding that both fast automatic processes, such as decision heuristics, and slower cognition-based, high-order reasoning occur [5]. According to several studies, there has been a measurable correlation shown between various amygdala activity, which is a key structure for rapidly identifying and adaptively responding to a wide variety of behaviorally salient events [5]. A study conducted by Adolphs et. al, demonstrated that judgements of threat can be made from facial stimuli that are presented for as briefly as 39 milliseconds. Since the amygdala is known to detect fear, danger, and plays a primary role in visual processing, we can conclude that it plays a central role in the automatic and non-conscious processing of emotional and social stimuli [10]. Therefore, we attribute the amygdala to demonstrating the almost automatic and inherent reactions that implicit biases illicit.

The amygdala, through various functional magnetic resonance imaging (fMRI) studies, has shown to be consistently activated by subliminal presentation of emotional faces [6]. Subliminal priming or presentation is established by the primed stimuli that is below the threshold of the conscious perception [7]. In most cognitive psychology studies, subliminal priming methodologies include very short experimental observation periods that last just milliseconds to understand the impact of such brief exposure on an individual's decision-making [7]. These studies which utilize such brief stimulus presentations to manipulate decision-making contexts before task presentation has shown priming-inducing changes in behavioral and/or neural responses bias [5]. A study conducted by Chiesa et. al reported that after subliminal presentation of positive affective primes, neutral stimuli were rated more likeable and the opposite occurred for the presentation of negative affective primes. This process also highlights a clear difference from memory; more specifically,

there is no direct retrieval of information, as we can observe strict neural based stimulation responses [5]. From this, we can deduce that though implicit biases can form through early memories or subliminal messaging, they are truly a part of our unconscious state as they are triggered as automated responses to specific stimuli.

However, the amygdala's response to fear or positive reactions to rewards is not simply just a reflexive process. In a study conducted by Chiao et. al, functional magnetic resonance imaging was used to measure amygdala response to fear and non-fear in faces in two distinct cultures. In this study we can observe that in the two cultures chosen, Native Japanese in Japan and Caucasians in the United States, they showed greater amygdala activation to fear expressed by members of their own cultural group [12]. Thus, this paper claims that the bilateral amygdala to faces of fear is perhaps modulated by culture. Specifically, it was proven that the amygdala is responsive to both subliminal and supraliminal presentations of faces that varied on trustworthiness, suggesting that modulation by complex social cues may occur in the absence of awareness [12].

Another way measured the impact of implicit bias on unconscious decision-making tendencies is the Implicit Association Test (IAT). The IAT, which is composed of five main parts, measures the strength of associations between concepts (e.g. attributes like being black or homosexual) and evaluations (e.g. good or bad) or stereotypes (e.g. nerdy or athletic) [13]. To further bolster the effect of implicit bias on a neural basis, many studies have been conducted to specifically highlight the activity and processes of the amygdala while taking the IAT.

In a study conducted by Luo et al., the effectiveness of the IAT and the automatic nature of moral attitudes was studied using event-related fMRI. Participants were shown visual examples of legal and illegal behaviors of two different intensity levels. For example, vandalism would be considered a low-intensity illegal activity while something like domestic violence would be considered high-intensity. The participants then rated the legality of these visuals both when the target concept (e.g. illegal) was behaviorally paired with an associated attribute (e.g. bad which is a congruent condition) or an unassociated attribute (e.g. good which is an incongruent condition) [14]. When the results were analyzed at a neural level, it was observed that with increased stimulus intensity there was an increased blood-oxygen-level-dependent (BOLD) response [14]. Thus, such measured responses associated implicit moral attitude with increased activation in the right amygdala and the ventromedial orbitofrontal cortex [14]. Behaviorally, this implicit bias was clearly reflected in the IAT as there was a faster reaction rate in associating the congruent conditions than the incongruent conditions [14].

Additionally, this study also opened up more associations of the brain with our implicit biases and moral judgements. It was found that performance on incongruent trials relative to the congruent trials was correlated with increased activity on the right ventrolateral prefrontal cortex, left subgenual cingulate gyrus, bilateral premotor cortex, and the left caudate [14]. All of these regions have been shown to have some effect on moral reasoning. It has been noted that damage to the ventromedial prefrontal cortex is associated with impairments in both spontaneous and deliberative moral judgements [15]. Patients with ventromedial prefrontal cortex lesions (vmPFC) have shown to develop uncharacteristic behaviors such as making poor financial decisions, not being able to maintain employment and relationships, and diminished affective responding to

social stimuli [15]. Patients with vmPFC lesions tend to have compromised affective reactions as this region contributes highly to emotional reactions. Thus, according to these studies, if moral outcomes differ amongst such patients, then it does in fact suggest that implicit affective processes could be integral to moral judgement and behaviors [15]. Therefore, damage to the vmPFC region has been associated with impairments in both spontaneous and deliberate moral judgements [15]. Studies that confirm the functional contribution of different regions of the brain to implicit moral judgements bolster the effective nature of the IAT in also aiding to detect regions of the brain that can be associated with implicit biases that we harbor.

Methods to combat Implicit Bias

The negative ramifications of implicit bias in our behavior and unconscious decision-making process have long been studied in conjunction with ways to combat the cognitive processes that strengthen such biases. Individual experiences, childhood memories, and environments in which one has grown up in or currently lives in are extremely hard to change. These influences are extremely variable and differ in subtle ways with each individual, making it impossible to control these factors for all. Instead, approaches have focused on either strengthening the ways in which we can train the brain to reduce the certain associations or stereotypes it harbors to lessen the unconscious bias. These include honing reinforced learning and inductive reasoning methodologies.

The two primary approaches to reducing expression of stigma and prejudice at the individual level are to reduce the activation of implicit stereotyping, decrease automatic responding, and change perception of others by purposefully restructuring thoughts and increasing cognitive control [5]. A way in which to decrease the activation of implicit stereotyping is to directly counter the stereotype the individual harbors. We can do this by making them actively aware of information that is specific to the individual and contrast the stereotype. For example, in a study conducted by Buchman et al., it was found that the usage of functional neuroimaging for the prediction, diagnosis, and treatment of mental disorders, and the use of such language by medical professionals has categorized those with mental illness as “neurochemical beings.” Thus, having this more biologically oriented psychiatry is imperative in the broader discourse of shaping ideologies of causality, blame, and agency of mental illness and personhood [16]. A clear negative

bias has been shown in people with either little exposure to mental illness or to individuals suffering from mental illness. This means that there is a greater need for personal experiences and individualism to debunk stereotypes [16]. Therefore, in this case, while framing mental illness with more neuro-based ideologies helped to destigmatize mental illness, there is also a need to personalize such experiences. When treating patients, usage of person-centered language and motivational interviewing strategies has shown to decrease self-stigma as well [5]. Thus, while implicit biases are woven into our moral and ethical values, they can be malleable and implicit associations can be gradually unlearned through several debiasing techniques.

However, recent studies have shown that changing implicit biases might not be an easy task. Chang et al. performed a study observing the effects of an online diversity training on the attitudes and workplace behaviors. Though diversity training has become a mandatory part of every workspace and office training program, we still observe sweeping negative attitudes and inequities towards women and racial minorities. These studies specifically showed that diversity training resulted in a significant positive change in behavior and attitude in subgroups that were already supportive of women and the opposite effect for those who were less supportive of women [17].

Nonetheless, such results are not reliable in that the attitudes were measured at the end; thus, any attitude change might have not been genuine in that there could have been some social pressure or demand to change [17]. Additionally, this experiment measured results right after the diversity training was given and so further data on long-term attitude changes need to be observed as the inherent biases may have been temporarily masked by the short-term stimuli. This paper

ultimately makes a claim that reactions and changes in behavior and attitude relies heavily on the audience and environment. Being variables, these are extremely hard to standardize.

Though we have observed the real-life implications of implicit bias, we have tried to create machine learning algorithms modeled after inductive bias to create better predictions and hopefully come to unbiased conclusions. Unfortunately, however, because it is humans who create such algorithms and feed it the training sets or data, the AI learns from these inputs and makes predictions based on this data. Due to the inherent automatic nature of implicit bias and the difficulty of correcting implicit bias with external stimuli and extensive training, pervasive bias seeps into our algorithms rendering them as biased as we are. The question now remains on how to correct such implicit bias in our machine learning algorithms as this has many structural, social, and ethical implications in our society today.

The Importance of Reinforced Learning

Although tackling this issue of implicit bias may seem to be an impossible feat, our own brains have developed some mechanisms that, when exercised, can combat the deepest of biases. Reinforced learning is a cognitive process that has demonstrated in humans and animals to open doors to corrective learning skills. This process has been translated and implemented in machine learning algorithms as a system of rewards and punishments so that machines can learn to predict correctly. But before looking at the ways in which we have implemented this in AIs, it is imperative that we take a closer look at the neural structure of this cognitive learning process.

Reinforced learning is an adaptive process in which a human or animal uses their previous experiences or concepts of rewards and punishments to improve upon the outcomes of potential future decisions. Actions tend to be chosen according to their value functions adjusted through reward or punishment which describe how much future reward is expected from each action [18]. We can see that reinforced learning involves similar regions of the brain that were stimulated by implicit biases and subliminal messaging. For example, a study conducted by Morrison et.al shows the imperative role that the amygdala plays in reinforced learning. The activity of individual amygdala neurons in monkeys was examined while abstract images acquired either positive or negative value through conditioning [19]. After the monkeys learned the initial associations, the image value assignments were reversed, and the neural responses were noted and compared to those of the conditioned values. What is significant in this study is that the changes in the values of images modulate neural activity and this modulation occurs rapidly enough to correlate with the monkey's learning [19]. Additionally, it was found that distinct populations of neurons encode the positive and negative values of the visual stimuli [19]. Thus, this study bolsters the amygdala's

crucial role in reinforcement learning and how behavioral and physiological reactions rely on values provided by the amygdala.

In addition to the amygdala, another key factor in our brains that determines to what extent we learn from the positive versus negative outcomes of our decisions is the neuromodulator dopamine [20]. The effects of dopamine on our reward and punishment sensitivity can be observed via patients with Parkinson's disease. Those who have Parkinson's disease tend to be impaired in tasks that require learning from trial and error as they have depleted dopamine in the basal ganglia [20]. When given two cognitive procedural learning tasks, it was observed that Parkinson's patients off medication are better at learning to avoid choices that lead to negative outcomes than they are at learning from positive outcomes [20]. However, dopamine makes the patients more sensitive to positive than negative outcomes and so dopamine medication tends to reverse their inherent bias [20].

Based off of computational models of basal ganglia it was ultimately observed that dopamine interactions in cognition are differentially modulated by positive and negative reinforcement [20]. Thus, we can see that dopamine also aids in our learning as it aids us to differentiate and be drawn to positive outcomes over negative outcomes.

Reinforced learning has since been modeled from human cognitive processes to the area of machine learning where algorithms search for the action that maximizes reward in a particular instance. Thus, with our now current knowledge of the cognitive processes and neural activities

behind reinforced learning we can dive into how this has been translated into machine learning algorithms.

INDUCTIVE BIAS and MACHINE LEARNING ALGORITHMS

What is Inductive Bias and why is it useful?

When the human psyche faces novel decisions and challenges in life, it needs to quickly evaluate new and unknown paths to act accordingly. But how can we possibly know that this new option will ensure that we make the most rational and effective decision if we have never explored it before? At first glance, it seems as though this question may have a solution that is no better than a shot in the dark, a random chance that we humans naturally have to take in life. However, if we look at this through an inductive framework, we can perhaps unlock why we as humans are able to try new options and make decisions without ever exploring them.

Though we have spoken about the negative effects of implicit bias, we will further dig into how this affects the algorithms we build out inductive bias that is essential to both human and machine learning. In this case, bias woks in a positive way. It is necessary for recognizing patterns and learning to label future possibilities or options not presented or known. To understand why biases are necessary for the inductive reasoning, we observe why unbiased learning systems are ineffective. For example, we have an unbiased algorithm that is trying to learn a mapping from numbers in the range one to fifty to the labels “TRUE” and “FALSE.” It observes that one, three, and five are labeled “FALSE,” while two, four, six are labeled “TRUE.” It is then asked to find the label for seven. To the unbiased learner, there are over ten trillion (2^44 to be exact) possible labelings of the numbers seven through 50, all equally possible with seven labeled “TRUE” with 0.5 probability and “FALSE” with 0.5 probability. Thus, the algorithm is reduced to guessing because it is not able to infer from the given observations [24]. Therefore, proving that an unbiased

learning system's ability to classify new instances is no better than if it stored all of the training data given to it and performed a simple look-up search when asked to classify a subsequent instance [23].

Inductive bias is distinguished from non-inductive bias in that it involves an inference from observations to unknowns [21]. For example, if you have tried many different products from an online store, all of which you have been very satisfied with, then it would be a reasonable inference that a product from this store that you have not personally interacted with would also likely to be a good product. In this example, we can see inductive bias playing a role as your knowledge of the store's products. The inductive reasoning comes into play as you can make observations and then a generalization from these previous, in this case positive, experiences. You then use these as internal data to predict that a new product from this store will also likely will good. Thus, from a psychological perspective it seems plausible that humans possess inductive biases that they have garnered throughout their life that drive their decisions in the absence of experience [21]. By making generalizations, we as humans are able to quickly and efficiently make the most effective decisions.

Inductive bias has proven itself to be one of our most efficient cognitive processes. It tends to cut down time and energy for us so we do not need to spend further time speculating, investigating, or stressing on these new choices. Based on our previous observations we can often count on this prediction to be correct and beneficial for us, thus leading to efficient decisions. This type of search mechanism has been developed to be the basis for neural networks and machine learning algorithms as it can now be applied to much more complex and extensive datasets that

accelerate the search of valuable options. Hence, inductive reasoning is an extremely useful to not only solving a large class of problems but also a central way in which we have structured our machine learning algorithms.

How is Inductive Reasoning translated in Machine Learning Algorithms?

When translating the neuroscience behind how humans learn and reconfiguring that processes to fit machines, it becomes a bit trickier to induce a learning technique that fits human cognition. We essentially, with any algorithm or representation of data, are trying to find the one that will yield the most effective predictions. Similar to how the human brain uses inductive bias and reinforced learning to learn more about unknown experiences and make the most efficient decision, we can use inductive reasoning as a guide to explore and build machine learning models. Without it, the learner, whether that be a human or a machine in this case, cannot generalize from observed examples to new foreign examples. Therefore, the goal of any machine learning algorithm is to produce a prediction for any unseen test instance on the basis of a finite number of training instances [24].

When modeling human cognition and neural pathways, there are three levels that can be considered:

1. a "computational" level that characterizes the problem faced by the mind and how it can be solved in functional terms;
2. an "algorithmic" level describing the processes that the mind executes to produce this solution; and
3. a "hardware" level specifying how those processes are instantiated in the brain [25].

Using these levels there are two main ways to model inductive bias: a bottom-up, mechanism-first model or a top-down, function-first strategy. Many would advocate for a bottom-up, mechanism-first explanation where it starts by identifying the neural or psychological

mechanisms believed to be responsible for cognition, and then tries to explain behavior in those terms. In contrast, probabilistic models of cognition pursue a top-down, function-first strategy that starts by considering the function that a particular aspect of cognition serves, explaining behavior in terms of performing that function [25]. While it is highly debated which method results in stronger inductive bias algorithms and representations, each have their distinct pros and cons, and studies have presented valid arguments towards both methodologies [25]. That being said a neural-level understanding of the human brain has still not been fully achieved so capturing the full cognitive abilities for humans to make correct inferences and then implementing these in neural circuits is still a large area of further research. Ultimately, the flexibility to explore different assumptions about representation and inductive biases, and then to naturally capture inferences over different classes of data is the goal of any model [25].

Another question to consider that often comes up when building an inductive model is determining which bias allows the algorithm to perform the best. However, the no free lunch theorem of machine learning shows that there is no one best bias [24]. Given two algorithms A and B, whenever algorithm A outperforms algorithm B in one set of problems, there is an equally large set of problems in which algorithm B outperforms A (even if algorithm B is randomly guessing) [26]. Therefore, we can conclude that the best model for a machine learning problem is one that finds an algorithm with biases that fits with the problem being solved [24]. For example, every ML algorithm used from nearest neighbors to gradient boosting machines comes with its own set of inductive biases about what classifications are easier to learn [24]. It is ultimately not possible to find an algorithm that can tackle all of the issues and it is imperative to identify the

difficulty and domain of the issue at hand to choose the right algorithm with the right set of biases or assumptions.

A Brief History of Inductive Bias Research

In this section we will briefly touch upon the algorithmic and experimental work that has been done in machine learning that have tried to efficiently improve inductive reasoning by refining generalization through multiple task learning.

One of the earliest approaches to inductive bias learning was through the Hierarchical Bayesian Inference Methods in statistics. For this methodology, the Bayesian inference is supposed to guide learning from sparse data and the probabilities defined over structures, graphs, and other forms that operate as inputs to background knowledge. The hierarchical probabilistic models, with inference at multiple levels of abstraction, is how the background knowledge is acquired and constrained, but allows flexible learning to occur [27]. Following this, in 1987 a "Variable Bias Management System" was introduced by Rendell, Seshu, and Tcheng as a mechanism for selecting amongst different learning algorithms when exploring a novel learning problem. "Shift to a Better Bias" was another scheme introduced in 1986 for adjusting bias; however, it focused more on large problem domains and biases applicable to that [27]. Therefore, a majority of early efforts were trying to efficiently search for working algorithms with proper biases (the importance of which was explained in the previous section) to fit a new problem with larger domains.

Next, there came more metric-based approaches to represent forms of inductive bias and linking them with machine learning methods. For example, the metric used in nearest-neighbor classification and vector quantization can be learned by "sampling from a subset of tasks from the environment, and then used as a distance measure when learning novel tasks are drawn from the

same environment" [27]. Another example of a metric that was successfully trained was in one of Baxter and Bartlett's experiments in 1998 where the metric was trained on a subset of 400 Japanese characters and then used as a fixed distance measure when learning the 2600 unseen characters [27]. While there are other adaptive metric techniques that can be used, they tend to adjust the metric for a fixed set of problems, rather than learning a metric suitable for bias learning such as nearest-neighbor classification, vector quantization, and derivative information generated from previously trained distance metrics [27].

Soon there were studies conducted in 1992 and 1993 on early algorithms for neural networks that tackled the problem of whether learning the bias decreases the number of examples required of a new task for good generalization. This also led to questions about computational complexity of a learning algorithm and how this may be improved by training the model on related tasks [27]. Ultimately, the bridge between cognitive neuroscience, psychology, and computer science formed with studies coming out that in the inductive bias neural mechanisms in the human brain is linked to reinforcement learning (as we will explore further in the "How Inductive Biases have been Incorporated into Reinforced Learning section"). Using this information, algorithms were then proposed that advanced average generalization or learning biases from a set of tasks to ultimately improve performance on future tasks.

Current Roadblocks on Research with Inductive Biases in Machines

"Often the hardest problem in any machine learning task is the initial choice of hypothesis space; it has to be large enough to contain a solution to the problem at hand, yet small enough to ensure good generalization from a small number of examples" [28].

Though machine learning is a solution for developing predictive models of human decision-making behaviors, there still remains plenty of room for more precise and efficient models. According to Bourgin et al., it has been argued that the main drawbacks in machine learning and inductive bias research have been due to data scarcity, since human behavior cannot be cleanly tracked it requires massive sample sizes to be accurately captured by most machine learning methods. Thus, to solve this issue we require larger datasets to encompass as many human judgements as possible and machine learning models with appropriate inductive biases for capturing such human behaviors. In recent years, we have been trying to refine machine learning by leveraging inductive bias of neural networks for the acquisition of new knowledge. This new methodology for predicting human decision-making is referred to as "cognitive model priors" which pretrains neural networks with synthetic data generated by cognitive models developed by cognitive psychologists.

According to the study, these networks are then fine-tuned on small datasets of human decisions which consequently improve performance on two recent human choice prediction competition datasets [29]. This study also provides a new standard for benchmarking prediction of human decisions under uncertainty as a large-scale dataset is presented "containing over 240,000

human judgements across 13,000 decision problems" [29]. Thus, ultimately developing a method for enabling machine learning models to better predict human behaviors when there is little data to work with. In addition, these studies have extended to the use of machine learning methodologies in behavioral economics and trying to understand the decision-making and predicting future moves of gamblers and gamers.

Now that we understand inductive bias a bit more, it time to look at inductive bias in conjunction with reinforced learning and why it is imperative that these two go hand-in-hand, both from a neural and algorithmic perspective. We will dive further in why we must be careful to not select examples or training sets that constraint the concept we are trying to teach the machine as this can lead to other biases and limitations in our algorithm.

How Inductive Biases have been incorporated into Reinforced Learning

I. The neuroscience - how inductive bias and reinforced learning are closely related

The psychological and behavioral process through which the brain makes connections and predictions by searching through and matching previous experiences from its "database" or memory has been hypothesized to be a large player in reinforced learning in humans (as mentioned in the "Importance of Reinforced Learning" section). In fact, studies have shown that the same underlying neurological regions and cognitive processes are at play for both inductive reasoning and reinforced learning. In this section, we dive into studies that have tackled the idea that the same neuromodulator, dopamine, is in play for inductive reasoning in the same way as it is for reinforced learning.

As we can recall, reinforced learning relies on our brains recognizing rewards and punishments, and reacting to them. The neuromodulator, dopamine, has been found to aid in recognizing and choosing rewarding behavior thus creating a reinforcement to the learning of the being. According to mathematical concepts applied to the neurophysiological aspect of learning, it is theorized that humans and animals employ a form of temporal difference learning algorithm which uses prediction errors; the difference between received and expected reward would update how humans and animals reward predictions [22]. In addition, it has been observed that this reward prediction error signal corresponds closely with the firing of midbrain dopamine neurons [22]. However, these findings have been challenged by the observation that dopamine neurons also respond to the appearance of novel stimuli [22]. These findings led to a study stating that dopamine might function in both inductive reasoning and reinforcement learning theory.

In a study conducted by Kakade et al., it was observed that in certain circumstances the activity of dopamine cells seem to respond in particular ways to stimuli that are not obviously related to predictions of reward. Kakade et al. determined that dopamine cells can be associated with two important sets of data: generalization and novelty. This study postulated shaping bonuses which are the optimistic initialization of reward predictions [30].

When a novel stimulus was presented, a positive prediction error was caused thus linking brain activity to choice behavior [30]. Ultimately because of these cues, there was an increase in optimism which was shown to be linked to initial exploration, thus explaining why we might make certain decisions without knowing the outcome [22]. Additionally, there was evidence that the generalizations by the dopamine neurons are a natural consequence of partial information, thus fully tying together the way in which inductive bias might be triggered in the brain on a neural level [30].

Behaviorally, we can see the link between reinforced learning and inductive bias through the concept of neophilia. Neophilia is defined as the preference for novel over familiar stimuli. There have been a multitude of experiments conducted, specifically on rats, that show how we can use prediction and choice as measures of novelty preferences, under the assumption that choice, approach, and avoidance result from predictions about future reward [22]. A multitude of studies have found that rats will perform an assortment of activities for the sake of understanding the unknown, including:

1. learning to press a bar for the sake of poking their heads into a new compartment,
2. displaying preferences for environments in which novel objects have appeared, and

3. interacting more with novel objects placed in a familiar environment [22].

The reinforcing nature of novelty suggested by these studies shows how embedded inductive bias can be in reinforcement learning and how rewards and punishment learning can drive both generalizations and predictions towards exploration once one gets positive reinforcement for their actions. Ultimately, this suggests that inductive biases play a role in human reinforcement learning by influencing reward predictions for novel options [22].

II. The algorithm - how inductive bias and reinforced learning are closely related

Now that we have observed that similarities in the underlying neural foundations and cognitive processes section on inductive biases in reinforcement learning, we can explore ways in which they can be algorithmically linked and embedded in one another. The exploration of how inductive biases are woven into already existing reinforcement learning algorithms has revealed how such biases can take on a multitude of forms with tradeoffs associated with each. Hessel et al. explores the ways in which inductive bias can be "injected" into already existing reinforcement learning algorithms.

In deep reinforced learning, the two more common methods of imbedding bias are by either:

1. Modifying the objective (which can be framed as the reward), or
2. Modifying the agent-environment interface.

Modifying the object methods such as reward clipping or discounting can be effective on its own, while modifying the agent-environment interface the methods can involve action repetitions [31]. In addition, when we insert inductive bias into these algorithms, a tradeoff between generality and performance can be observed [31]. If we add more bias, we can have faster training and efficient performance, but this can deteriorate generalization across domains [31]. If we instead add less bias, then we have to give up efficiency for more general algorithms which would be applicable to a much wider set of problems [31]. Thus, though inductive bias in reinforcement learning algorithms has the potential to produce efficient outcomes, there is a

question of tradeoffs that one should consider when inserting such biases and further exploration of adaptive measures to strike the right balance of bias should be considered.

How implicit bias sneaks into machine learning algorithms

Despite the complex abilities of machine learning algorithms and neural networks to learn from positive and negative experiences, there have been increasing instances of algorithms portraying biases similar to humans. With ethical questions where the line gets blurred between right and wrong, it becomes tricky (even for humans) to prove whether or not the algorithm made a correct prediction. As we will explore later, it becomes even more difficult to analyze the ways in which machine learning algorithms come to such predictions, especially as the complex scope of the input data and the inability to edit the program renders it to be a challenging issue for us to parse through.

Though inductive bias is a "positive" bias in that it has proven to be an integral part of our learning process, a number of unintended consequences arise when we use inductive bias for learning. Bias itself is often seen as a negative term. When race, culture, and societal biases sneak into algorithmic predictions, we often see the programs that we built start to mimic often frustrating issues that can resonate with negative stereotypes. But what we don't realize is that within ourselves, we may harbor similar biases that slip into the little decisions that we make every day that can then translate into the functions and algorithms that we build. In this section, we will dive into examples of how such biases have manifested with "racist" artificial intelligence systems and how this was largely determined by the testing data that we take from humans.

For machine learning algorithms that involve inductive bias, is it imperative for us to find the proper bias to address the issue at hand. Like previously determined, it is imperative that we are able to find an accurate generalization from the small number of examples and apply it to a

proper hypothesis space so that it contains a solution to the issue at hand. Once a proper bias is found, the learning task is easy; however, existing methods of bias require input of a human expert in the form of heuristics and domain knowledge [27]. Because all of the input for the bias is coming from a human who also harbors implicit biases, these methods are clearly limited by the accuracy and reliability of the human's knowledge, and by the extent to which that knowledge can be transferred to the learner [27].

One method we can use to avoid using the direct input of human knowledge for the bias required for inductive reasoning is to have the learner, or the machine in this case, automatically learn the bias. Ideally, we would supply the ML algorithm with a singular optimal hypothesis and from there it would be automatically learn the biases that is appropriate for that environment [27]. However, there are still learning problems relate to this method. A singular environment or problem space can also have a set of learning problems introducing yet another limitation to this approach. Thus, studies have supported the idea that adjusting the learners bias might be difficult as it might not be controllable depending on what information is being fed for the bias, and therefore it might be better to identify some fixed set of learning problems that such a bias could be used to solve [27].

Let us delve into the recent research that has been proving how such inductive bias can be attributed to negative reinforced learning. While the hope is that machines can sift through infinite pieces of data to make more accurate decisions than humans, it has been observed that these artificial intelligence systems harbor strong and often unfair biases. We have seen countless examples of such AI machines used for hiring, law enforcement, and loan granting. An extreme

application is searching for babysitters where couples have a racial bias and discounting candidates based on a multitude of factors based on potential or statistical negative associations. All of these serve harmful consequences such as the systemic prevention of fair distribution of jobs or resources. For example, MIT scientist Joy Boulamwini, uncovered that all of the facial recognition systems sold by renown companies such as IBM, Microsoft, and Amazon performed substantially better on identifying the differences between male faces than those of female faces. The error rates were no more than 1% for lighter-skinned men, but soared to a 35% error rate for dark-skinned women [32]. Though we try to make our algorithms "race-blind" or "gender-blind" and remove certain factors for consideration, why do these outcomes still occur? The answer lies in the training data fed to the AI for it to learn from and generalize patterns. It was found that the training sets for facial recognition tended to be white and male, by virtue of the pervasiveness of white men historically in computer science. By using this created dataset of faces to test and train, the AI had a much harder time identifying darker females as it could not make the generalization necessary from the data presented to it. As we further examine this, we will observe that the issue lies in the way in which an AI captures patterns and overextends associations, largely due to the data training sets that it is being fed by human, resulting in the unknowing learning of more negative, inductive biases.

Consider another example of an AI system that showed such an outcome through just simply learning and not being a system for choosing or recognizing certain people based on a set of criteria. It was observed that as a computer teaches itself English it can become prejudiced against black Americans and women [33]. This study showed that machines can learn word associations from written texts and that these associations mirror those learned by humans, as

measured by the Implicit Association Test [33]. While the machine made “harmless” human associations such as pleasantness and flower, it made more stereotyped biases for exam associations between female names and family or male names and careers [33]. This again proves that the data and associations that we introduce to AI can cause it, through inductive reasoning, to generalize human negative biases to make predictions even if obvious biases are omitted. Thus, we show how human psychology, behavior, and history can easily seep into the way in which a machine learns.

PROGRAM SYNTHESIS

"Program synthesis is the task of generating a program in an underlying domain specific language (DSL) from an intent specification provided by a user" [34]. In other words, program synthesis is the task of automatically finding programs from the underlying programming language that satisfy the user's intentions, usually expressed in the form of constraints. Program synthesizers typically perform some form of search over the space of programs to generate a program that is consistent with a variety of constraints [36]. Essentially, program synthesis allows the user to provide a skeleton of the space of possible programs in addition to the specification [37]. This has two benefits that tackles the issue of hypothesis space and proper bias that we encountered with inductive bias. The first being that, by providing the skeleton of the hypothesis space, it allows for a more efficient search procedure [37]. Secondly, we can now interpret the learned programs since they are derived from the given skeletal grammar [37].

Though traditionally program synthesis has been treated as a search problem, it has evolved into a supervised learning problem by the machine learning community [35]. Recent research has demonstrated that deep neural networks have the potential to learn a program satisfying various specification methods, such as natural language descriptions and input-output examples. This ability to synthesize code can have various applications such as helping to discover new algorithms, cleaning data, or optimizing code. The ultimate goal is to generate a program or algorithm with more complexity, better generalizability, efficiency, and correctness that can finally contribute to deep learning and generalization from inductive biases and help to achieve Artificial General Intelligence.

A great example of this is the programming by examples (PBE) paradigm. When the user is a non-programmer, we must find a specification method that is easy to understand without any type of prior programming knowledge, thus this programming by examples paradigm fits well as the user intent is specified by the means of input-output examples or constraints.

A paradigm of this would be Microsoft Excel's FlashExtract and FlashFill that was released in 2013. How FlashFill works is that it is designed for users who only care about the behavior of the program and not about the program itself (much like a PBE model). If the output is wrong, then users can tell just by looking at it and can provide another input until it "fills" with the correct information.

The reason as to why I included this section is that, as part of my contribution, I would like to take aspects of program synthesis and implement it into my theoretical model as I believe that this type of approach has many advantages that machine learning algorithms fail at achieving. This will further be discussed in my conclusion section.

CONCLUSION

In this paper, I aimed to tackle the issue of inherent biases in our current machine learning algorithms. The implicit biases reflected in these algorithms hint at major social, ethical, and moral battles that affect the daily lives of people around the world. In order to better understand this, we approached implicit and inductive bias from a mechanism first perspective and understood the fundamental neural pathways for each of them. Then, using a bottom-up model we were able to observe how machine learning algorithms were built off of such cognitive processes. It was observed that inductive bias and reinforced learning are crucial to both human and machine learning mechanisms. However, inductive biases can also generalize and predict in negative ways as these generalizations depend on data and examples fed to it by humans who themselves harbor implicit biases. Thus, the inevitable biasing of algorithms, is the result.

The apparent solution would be to remedy our own implicit biases and produce unbiased – in the implicit sense – data and training sets. This would allow for machine learning algorithms to have "clean" data sets to learn and make more accurate generalizations from and, ultimately, allow them to accurately predict outcomes from few examples. However, based on our research we have seen how implicit bias an unconscious, automatic cognitive process and so trying to undo years of childhood memories and subliminal messaging does not seem to be a realistic solution. In addition to this, as I have discussed in this paper, while implicit bias has ways to be retrained and reversed in humans, there have been mixed effects on trying to change implicit bias. These studies usually are done on a short-term scale and often show a temporary change in behavior rather than a complete and permanent change in one's inherent attitude. Therefore, this is not a clean solution to solving the issue of biases in algorithms. In the next part of my conclusion I propose a few

solutions along with an idea for a theoretical model that uses the different cognitive and algorithmic mechanisms and learning techniques I have laid out in this paper.

When Bad Algorithms Outperform Good Ones

The first solution I propose is that with proper data and training sets even bad algorithms can outperform good ones. As we have seen, algorithms that use inductive reasoning are intended to generalize and use this to accurately predict or make decisions that will optimize the outcome. While this depends on the specific bias and problem we are trying to solve, if the correct examples which maximize the chances for the correct outcome/patterns are not given to the machine to learn from, then it simply will not be able to produce the most efficient results unless it goes through several rounds of trial and error. Thus, the need for few examples, with the most optimized solutions when fed into any algorithm should produce the best results after it learns from it.

In addition to this, to find implicit biases in algorithms that learn strictly from large training sets (such as facial recognition algorithms), it is imperative that we provide a much more diverse set of training data along with a much larger set. This data should be updated periodically as our culture is constantly growing and changing. Thus, such changes should not only be reflected in the data sets, but also, ultimately, in the outcome of the algorithmic predictions. Such diverse and vast data should give enough examples to the algorithm to consider a larger class of objects for its criteria, along with various factors so that it can accept and recognize more inputs as being valid. However, updating and adding to immensely large and complex data sets may also not be feasible. This is because certain questions may arise such as: at what point does certain data become irrelevant? How often does one need to update data? And, additionally, what constitutes diversity in data? Thus, while it does stand that a bad algorithm will function better than a good one if, given a better and larger training set, it is still not the most efficient or ideal solution.

The Union of ML (in an inductive framework) and PBE

My ultimate contribution is my proposition and exploration of a bottom-up (rather than a top-down as referred to in my inductive bias chapter) theoretical model that combines the advantages of a machine learning model with the advantages of a programming by examples or PBE model (a subset of program synthesis). Similar to how this paper was presented, I believe that it is important to understand the fundamental cognitive processes and neurological responses rather than make assumptions without them. In order to better understand bias in humans, we should first look at the biological basis before trying to translate this into ML algorithms or come up with solutions to existing algorithmic issues as they have already been modeled off of these processes.

The main goal of this section is to understand the strengths and weaknesses of ML and PBE and to explore the union of them to create a better algorithm that can combat bias and produce optimal predictions or results. ML tends to output functions which minimizes loss and utilizes optimization to do so [39]. Program synthesis however, uses a combinatorial search and outputs a program which satisfies specification [39]. The main strength of program synthesis is that not only can it function well with only a small number of specific functions and aims to output a program that satisfies those specifications, but the synthesized program is interpretable and hence editable. ML on the other hand, requires a large amount of training data and tries to minimize loss. A successful ML method however produces functions that are not interpretable. But ML is robust to noise and handles generalization which is something that program synthesis does not have the capabilities to do.

Thus, in conclusion the main aim is to implement ML in a PBE framework so that:

1. Biases have a smaller chance to occur as we would implement concrete specification from program synthesis to ensure a more pointed outcome
2. The program is editable, unlike ML's functions that are not interpretable. This ensures that factors within algorithms are being controlled, accounted for, and can be removed if needed
3. ML has much better capabilities to speed up this search and work with and handle large, noisy, datasets. We must be able to implement an algorithm at a much larger scale as complex data is handled and so this is an important quality.

While the theoretical model must have the components mentioned above there still remains an ethical issue in terms of algorithmizing certain processes that lead to this issue of implicit bias. We have tried to algorithmize certain processes, such as for example college admissions which overall tends to be highly subjective with a multitude of dependent factors and varying thresholds for each. Such information is difficult for any user to capture in data sets and produce optimal results and no such singular optimal result exists. Thus, while we are utilizing the power of algorithms to sift through complex data the dangers of it to our ethical and moral stances must be taken seriously and reevaluated.

BIBLIOGRAPHY

[1] Schwarcz, Daniel, and Anya Prince. "Proxy Discrimination in the Age of Artificial Intelligence and Big Data." *Iowa Law Review, Forthcoming*, 5 Aug. 2019.

[2] Puri, Manju, et al. "On the Rise of the FinTechs - Credit Scoring Using Digital Footprints." *SSRN Electronic Journal*, 2018, doi:10.2139/ssrn.3259901.

[3] Klein, Aaron. "Credit Denial in The Age of AI." *Brookings*, 11 Apr. 2019, www.brookings.edu/research/credit-denial-in-the-a

[4] "Understanding Implicit Bias." *Kirwan Institute for the Study of Race and Ethnicity*, 2015, kirwaninstitute.osu.edu/research/understanding-implicit-bias/

[5] Reihl, Kristina M., et al. "Neurobiology of Implicit and Explicit Bias: Implications for Clinicians." *The Journal of Neuropsychiatry and Clinical Neurosciences*, 21 Oct. 2015, neuro.psychiatryonline.org/doi/10.1176/appi.neuropsych.15080212.

[6] Brooks SJ, Savov V, Allzén E, et al: Exposure to subliminal arousing stimuli induces robust activation in the amygdala, hip- pocampus, anterior cingulate, insular cortex and primary visual cortex: a systematic meta-analysis of fMRI studies. Neuroimage 2012; 59:2962–2973

[7] Elgendi, Mohamed, et al. "Subliminal Priming-State of the Art and Future Perspectives." *Behavioral Sciences (Basel, Switzerland)*, MDPI, 30 May 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC6027235/.

[8] Rich, Alex. "Using Inductive Bias as a Guide for Effective Machine Learning Prototyping." *Flatiron Health*, 5 Nov. 2019, flatiron.com/blog/using-inductive-bias-as-a-guide-for-effective-machine-learning-prototyping/.

[9] Cockrell, Jeff. "A.I. Is Only Human." *ChicagoBoothReview*, review.chicagobooth.edu/economics/2019/article/ai-only-human.

[10] Pessoa, Luiz, and Ralph Adolphs. "Emotion Processing and the Amygdala: from a 'Low Road' to 'Many Roads' of Evaluating Biological Significance." *Nature Reviews. Neuroscience*, U.S. National Library of Medicine, Nov. 2010, www.ncbi.nlm.nih.gov/pmc/articles/PMC3025529/.

[11] Chiesa, P.A., Liuzza, M.T., Acciarino, A. *et al.* Subliminal perception of others' physical pain and pleasure. *Exp Brain Res* **233,** 2373–2382 (2015). https://doi.org/10.1007/s00221-015-4307-8

[12] Chiao, Joan Y, et al. "Cultural Specificity in Amygdala Response to Fear Faces." *Journal of Cognitive Neuroscience*, U.S. National Library of Medicine, Dec. 2008, www.ncbi.nlm.nih.gov/pubmed/18457504?dopt=Abstract.

[13] "ProjectImplicit." *About the IAT*, implicit.harvard.edu/implicit/iatdetails.html.

[14] Luo, Qian, et al. "The Neural Basis of Implicit Moral Attitude-An IAT Study Using Event-Related FMRI." *NeuroImage*, Academic Press, 18 Jan. 2006, www.sciencedirect.com/science/article/abs/pii/S1053811905024444.

[15] Cameron, C Daryl, et al. "Damage to the Ventromedial Prefrontal Cortex Is Associated with Impairments in Both Spontaneous and Deliberative Moral `Judgments." *Neuropsychologia*, U.S. National Library of Medicine, Mar. 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC5866785/.

[16] Buchman, Daniel Z., et al. "Neurobiological Narratives: Experiences of Mood Disorder

through the Lens of Neuroimaging." *Wiley Online Library*, John Wiley & Sons, Ltd, 3 May 2012, onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9566.2012.01478.x.

[17] Chang, Edward H., et al. "The Mixed Effects of Online Diversity Training." *PNAS*, National Academy of Sciences, 16 Apr. 2019, www.pnas.org/content/116/16/7778#sec-2.

[18] Lee, Daeyeol. "Neural Basis of Reinforcement Learning and Decision Making." *Annual Reviews*, 2012, www.annualreviews.org/doi/abs/10.1146/annurev-neuro-062111-150512.

[19] Paton, Joseph J, et al. "The Primate Amygdala Represents the Positive and Negative Value of Visual Stimuli during Learning." *Nature*, U.S. National Library of Medicine, 16 Feb. 2006, www.ncbi.nlm.nih.gov/pubmed/16482160.

[20] Frank, Michael J., et al. "By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism." *Science*, American Association for the Advancement of Science, 10 Dec. 2004, science.sciencemag.org/content/306/5703/1940.

[21] Griffiths, Thomas L, et al. "Probabilistic Models of Cognition: Exploring Representations and Inductive Biases." *Trends in Cognitive Sciences*, U.S. National Library of Medicine, Aug. 2010, www.ncbi.nlm.nih.gov/pubmed/20576465.

[22] Gershman, Samuel J., and Yael Niv. "Novelty and Inductive Generalization in Human Reinforcement Learning." *Topics in Cognitive Science*, vol. 7, no. 3, 2015, pp. 391–415., doi:10.1111/tops.12138.

[23] Mitchell, Tom M. *The Need for Biases in Learning Generalizations*. 1980, dml.cs.byu.edu/~cgc/docs/mldm_tools/Reading/Need%20for%20Bias.pdf.

[24] Rich, Alexander. “Using Inductive Bias as a Guide for Effective Machine Learning Prototyping.” *Medium*, Flatiron Engineering, 6 Nov. 2019, medium.com/flatiron-engineering/using-inductive-bias-as-a-guide-for-effective-machine-learning-prototyping-66e5468407a8.

[25] Griffiths, Thomas L, et al. *Probabilistic Models of Cognition: Exploring Representations and Inductive Biases*. 2010, www.ed.ac.uk/files/atoms/files/griffithstics.pdf.

[26] Wolpert, David H. *The Lack of A Priori Distinctions Between Learning Algorithms*. 1996, citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.390.9412&rep=rep1&type=pdf.

[27] Baxter, Jonathon. “A Model of Inductive Bias Learning.” *Journal of Artificial Intelligence,* 2000, doi:10.3897/bdj.4.e7720.figure2f.

[28] Mitchell, T. M. (1991). The need for biases in learning generalisations. In Dietterich, T. G., & Shavlik, J. (Eds.), *Readings in Machine Learning*. Morgan Kaufmann.

[29] Bourgin, David D., et al. “Cognitive Model Priors for Predicting Human Decisions.” 22 May 2019.

[30] Kakade, Sham, and Peter Dayan. “Dopamine: Generalization and Bonuses.” *Neural Networks : the Official Journal of the International Neural Network Society*, U.S. National Library of Medicine, 2002, www.ncbi.nlm.nih.gov/pubmed/12371511.

[31] Hessel, et al. “On Inductive Biases in Deep Reinforcement Learning.” *ArXiv.org*, 5 July 2019, arxiv.org/abs/1907.02908.

[32] Santamicone, Maurizio. “Is Artificial Intelligence Racist?” *Medium*, Towards Data Science, 3 Apr. 2019,towardsdatascience.com/https-medium-com-mauriziosantamicone-is-artificial-intelligence-racist-66ea8f67c7de.

[33] Caliskan, Aylin, et al. “Semantics Derived Automatically from Language Corpora Contain Human-like Biases.” *Science*, American Association for the Advancement of Science, 14 Apr. 2017, science.sciencemag.org/content/356/6334/183.

[34] Le, Vu, et al. “Interactive Program SynthesisVu .” 10 Mar. 2017.

[35] Simmons-Edler, Riley, et al. *Program Synthesis Through Reinforcement Learning Guided Tree Search*. June 2018.

[36] Fadelli, Ingrid. “Infusing Machine Learning Models with Inductive Biases to Capture Human Behavior.” *Tech Xplore - Technology and Engineering News*, Tech Xplore, 7 June 2019, techxplore.com/news/2019-06-infusing-machine-inductive-biases-capture.html.

[37] Gulwani, Sumit, et al. *Program Synthesis*. 2017, www.microsoft.com/en-us/research/wp-content/uploads/2017/10/program_synthesis_now.pdf.

[38] Gulwani, Sumit, and Prateek Jain. *Programming by Examples: PL Meets ML*. 2019, www.prateekjain.org/publications/all_papers/GulwaniJ17_APLAS.pdf.

[39] Bhattacharyya, Chiranjib, et al. Program Synthesis meets Machine Learning https://www.csa.iisc.ac.in/~deepakd/psml-2020/.