The Role of Large Language Models in Academic Writing

Edward Speer
California Institute of Technology
espeer@caltech.edu

May , 2025

Abstract

Large language models (LLMs) are rapidly improving in their ability to generate text that closely resembles human writing. This has led to a growing interest in using LLMs to assist in academic writing, with the potential to improve the quality and efficiency of the writing process. Most academic journals now have policies in place to govern the use of LLMs in academic writing, and most isallow the citation of LLMs as authors or co-authors. In this paper, we explore the current ability for an LLM to function as a co-author in academic writing, and in particular, whether or not an LLM can produce novel ideas and insights to an academic paper. We present a case study in which we use the Llama-2-7b model to assist in writing a paper connecting algorithmic fairness to Nozick's title of entitlement justice, in contrast to a large existing body of research connecting algorithmic fairness to egalitarian justice. We perform a qualitative analysis of the LLM's ability to contribute to the paper. We find that the LLM is able to recommend and summarize relevant literature, generate text that is coherent, well-structured, and relevant to the topic, and provide novel insights and ideas that are not present in the literature, but cannot produce or defend a full academic argument.

1 Introduction

The majority of research journals now provide policies for the use of large language models (LLMs) in academic writing. In the Nature journals, for example, "Large Language Models do not satisfy our authorship criteria. Notably, an attribution of authorship carries with it accountability for the work, which cannot be effectively applied to LLMs" (2). For Cambridge Press, "AI does not meet the Cambridge requirements for authorship, given the need for accountability" (1). The chief concern among these policies appears to be responsibility. The policies don't outright ban the use of AI, nor do they outline specific guidelines for how to appropriately use LLMs, save for as a copy-editor, but they make specific prohibitions against authorship and attribution of writing to LLMs. It seems clear that the journals are targeting issues of accountability. Who should the journal turn to when a mistake is discovered in a paper? Who is the owner and originator of the ideas presented? Who should be legally liable for the content of the paper?

The policies are clear that human authors alone must take full responsibility for the content of their papers, and to that end, LLMs cannot be considered coauthors. The need for accountability goes beyond concerns about liability for errors in writing, however, the policies also reflect a concern about the ownership of ideas and plagiarism. LLMs are trained on vast amounts of text, and it is not always clear how to attribute information they produce. An LLM may produce information that is similar or identical to text in its training data without a citation, and if this text is included in a paper, it constitutes plagiarism. The emphasis on accountability in these policies implies that in this situation, the human authors of the paper would be held liable for the plagiarism, even if they were unaware of it.

The policies protect the journals from liability in the case of these plagiarism concerns, but they are largely predicated on the idea that LLMs are not capable of originating novel ideas of their own. The policies are clear that LLMs cannot be attributed authorship, meaning that if they did produce original arguments, authors would be faced with a choice of either not publishing those ideas (stifling potentially significant contributions to the field) or not properly attributing them, neither of which is a desirable outcome. This means that should we find LLMs capable of producing original ideas, the current policies are out of step with the capabilities of the technology and require rapid revision. Even if LLMs do not currently have this ability, it is possible they could attain them in the near future, and thus these policies are at risk of quickly becoming out of date. We therefore need to both discover and explore the capabilities of LLMs in producing new ideas, and to consider the implications of these capabilities for the future of academic writing.

In this paper, we attempt to explore the ability of a current LLM to contribute original ideas and argumentation to an original philosophical research paper. We do this by using an LLM to collaborate on a research paper in the field of algorithmic fairness and distributive justice. The goal of the project is to produce a high-quality research paper that constitutes an original contribution to the field, and in the process, to use the LLM to its fullest potential and explore the capabilities and limitations of doing so. The goal of this experiment is not simply to see if the LLM can produce a publishable paper, but rather to explore the utility of using an LLM in a genuine effort to produce a high-quality paper. This effort will be a collaborative one, with both the human author and LLM writing sections of the paper, providing feedback on each other's writing, suggesting sources, and engaging in discussion about the ideas presented with the goal of producing the best paper possible. The success of such an endeavor is difficult to measure; we seek to provide a qualitative assessment of the LLM's contributions to the paper, and to explore the implications of the growing capabilities of LLMs for the future of academic writing. With the ongoing and rapid development of LLMs in mind, this paper is not meant to be a definitive assessment of the capabilities of LLMs, but rather an exploration of a specific model's capabilities and limitations as well as a discussion of the future of academic writing in light of these technologies.

For this project, we chose as the subject of our paper the topic of algorithmic fairness measures and their connection to the philosophical concept of distributive justice. Algorithmic fairness measures are a critical component of the design and deployment of machine learning systems, as they are intended to ensure that these systems do not discriminate against individuals based on sensitive attributes. Distributive justice, on the other hand, is a central concept in political philosophy that concerns the fair distribution of social goods. Distributive justice exists on a spectrum from end-state theories, which analyze whether a given distribution of resources is fair, to historical theories, which analyze whether the history of transactions that led to a given distribution of resources is fair. Since algorithmic fairness measures are often used to evaluate the fairness of algorithmic systems that make decisions about the distribution of resources such as bank loans or job opportunities, there is a natural connection between these two topics. The relationship between the two fields has been explored in the literature, but in the early stages of this project, our LLM suggested that there is a lack of discussion in the literature about the relationship between algorithmic fairness measures and historical theories of distributive justice. Our investigation began from this vague notion of the intersection between the two fields, and we worked with the LLM from this point to develop and complete a specific research project on the topic.

This paper will be structured as follows. In Section 2, we will present the methods used, including the specific LLM selected for the investigation and the program used to interact with it. In Section 3, we will present the full text of the paper produced in collaboration with the LLM. In Section 4, we will analyze the role of the LLM throughout 5 task-stages of the research process: literature review, research question formulation, argumentation, writing, and revision. Finally, in Section 5, we will conclude with a discussion of the implications of this experiment for the future of academic writing. Note that an LLM was only used as a collaborator for the writing of the research paper presented in Section 3, and not for the writing of this introduction or any other part of the paper. The introduction, analysis, and discussion outside of Section 3 were written entirely by the human author of this paper without the assistance of LLMs.

2 Methods

For this experiment, we selected the Llama-2-7b model developed by Meta AI. The Llama-2-7b model is a large language model trained on a diverse range of textual data, including books, articles, and websites. The model is capable of generating coherent and contextually relevant text across a wide range of domains. We selected this model for its ability to generate high-quality text and its general-purpose nature, which makes it suitable for a wide range of writing tasks, as well as for its open source nature, which reflects our commitment to transparency and reproducibility. While models like GPT-4o are continuously updated and improved in inscrutable ways, Llama-2-7b serves as a stable fixed-point for our investigation.

We engaged with the Llama-2-7b model (henceforth referred to as Llama) using a web-hosted API called Llama-api (`https://www.llama-api.com/`). This API allows users to pay a per-token fee to interact with the model via an http request to a designated endpoint. A user sends a prompt to the model, including context memory built up over the course of the interaction and the limitations on response tokens, and the model generates a response based on the prompt and the context memory.

In order to manage the use of Llama, we built a custom chat application in Python that allowed us to communicate with the model from the command line. This application has the following features:

- **Chat logging**: User prompts and responses from Llama are automatically saved to a log file in markdown format for future analysis.
- **Context Memory Management**: The application allows the user to save and use different streams of context memory across different sessions with the model. For example, in the beginning of one context, Llama is told "I am a philosopher and computer scientist. You are my co-author. We are writing a philosophy paper. We are focused on measures of algorithmic fairness and the concept of justice they enforce." In another context, Llama can be told to act as a reviewer, or to speak in the voice of an author encountered in the literature review. These bits of context are saved in compressed pickle files and can be loaded into the application at the any time during a session.
- **Manual Context Editing**: The application allows the user to manually edit the context memory before sending it to Llama. This is useful for trimming down the context memory to the most relevant information to reduce the cost of the interaction and to focus the model

on critical information. This feature can also be used to pass entire papers or large sections of text to Llama for review or comment.

- **Token Limiting**: The application allows the user to set a limit on the number of tokens in the response from Llama. This is useful for managing the cost of the interaction with the model.

The full source code for the chat application is accessible from 6.

Three main threads of context memory were used to work with Llama in this study. In the first thread, Llama was presented with the true circumstances of the experiment: that it was acting as a co-author on a philosophy paper about algorithmic fairness measures and distributive justice. In the second thread, Llama was presented with the role of a reviewer of the paper, providing feedback on the argumentation and writing. In the third thread, Llama was not prompted with any particular role, but was simply continually asked to explain particular arguments or concepts from the literature with appropriate citations. Henceforth we will refer to these roles as the coauthor, reviewer, and explainer roles. Each of these roles was used throughout the research and writing process, with the exception of the reviewer role which was used only during revision. The full logs of interactions with the model including which context memory was used in each interaction are available in 6.

Co-authorship is a relationship which can take on many forms depending on the nature of the collaboration. In this case, we were interested in exploring the extent to which Llama could contribute substantive and original content. This goal determined the nature of the interactions with Llama, which were designed to elicit original ideas and argumentation from the model. Simply asking the model to write the paper or to produce large sections of the text would not have been a useful approach — anyone who has asked an LLM to do so is aware that the results are lacking in depth or originality. Instead, in each of the five tasks of the research process, we engaged Llama in structured dialogues that contributed to the development of the paper. The structure of this dialogue was inspired by the Socratic method, and proceeded in a set of steps:

1. Provide Llama with the relevant background knowledge to the discussion through the context memory mechanism, pasting relevant sections of text, or asking Llama to summarize relevant arguments to add them to the context memory. For example, asking "Please summarize the paper 'Procedural Versus Substantive Justice: Rawls and Nozick' by David Lewis Schaefer" will add a (Llama generated) summary of the paper to the context memory.

2. Ask Llama a fully open-ended question about the topic at hand. For example, "Tell me about how these fairness measures may emphasize distributive concepts of justice?"

3. Pick out interesting aspects of Llama's response, and ask for more detail. For example, "I found interesting what you said about counterfactual measures of algorithmic fairness. How could they be considered to emphasize individualized justice in a way that touches on entitlement?" Push on these responses until Llama is unable to provide more detail in a coherent way. "I'm missing some of your ideas. In entitlement justice, we focus on whether individuals who acquire holdings are entitled to those holdings. Can you explain how counterfactual measures of justice show this feature?"

4. Inject some of your own thoughts into the conversation and ask Llama to respond to them and incorporate them into its own analysis. "If we want to say that someone is entitled

> to their college admissions, we need to say it is their property which is being taken away if they are denied admissions. This means that admission is a property acquired through work before applying. How should we defend this perspective?"

This dialogue structure is meant to do three things. Firstly, provide Llama with a basic set of text to pull structures from and hopefully build on. Secondly, try to draw out original ideas from Llama by really pushing it to do more than spit out responses to in-dataset prompts by asking for more details and explanations than would be found in the training data. Thirdly, to provide some original text to Llama from outside of the training set to help it build on and hopefully produce original ideas. Illustrative examples of this interaction and responses provided by Llama are provided in the analysis section. In a way the goal was to cause Llama to "hallucinate" original ideas by pushing it to build on its own responses and to build on original text provided by the user.

3 Results

What follows is the full text of the paper produced in collaboration with Llama.

Capturing Entitlement Justice in Algorithmic Decision-Making

Edward Speer
California Institute of Technology
espeer@caltech.edu

Llama-2-7b
Meta Platforms, Inc.

May , 2025

Abstract

In recent years, there has been growing interest in grounding algorithmic fairness in philosophical theories of distributive justice to clarify the meaning of fairness metrics used in automated decision-making. Most existing approaches focus on egalitarian theories, such as Rawlsian liberal egalitarianism. However, it remains unclear how these frameworks extend to Nozick's theory of entitlement justice, which defines justice in terms of legitimate acquisition and transfer of resources rather than outcome distributions. This omission is notable given entitlement theory's prominence in debates on distributive justice and leaves a substantial gap in our understanding of how different conceptions of justice inform algorithmic fairness. In this paper, we address this gap by exploring the relationship between algorithmic fairness and entitlement justice. We propose a framework called entitlement fairness, which interprets algorithmic fairness through the lens of entitlement theory. We illustrate how this framework can be applied to real-world decision-making, using college admissions as a case study. Our analysis demonstrates that entitlement fairness offers a nuanced, context-sensitive approach to evaluating fairness in algorithms, expanding the philosophical foundations of the field.

1 Introduction

The rise of algorithmic decision making in the public sector has caused significant public concern. Algorithms increasingly make decisions that affect individuals' lives, from determining creditworthiness to predicting criminal recidivism, and the public has grown cautious of their potential to perpetuate and exacerbate existing social inequalities. A 2018 study showed that 58% of Americans believe that algorithms will always have some level of bias (Smith, 2018), and as documented in the famed COMPAS case, these fears are not unfounded (Angwin et al., 2016).

In response to these concerns, a growing body of research has focused on developing algorithmic fairness measures to evaluate and mitigate the biases in algorithmic decision making. A large number of different measures have been proposed (Corbett-Davies et al., 2023) and applied to a wide range of problems. However, many questions arise about the foundations of these measures and how to apply them to sociotechnical systems. Chouldechova (2017) showed that multiple fairness measures are incompatible with each other and cannot be satisfied simultaneously. This has led to a growing recognition that algorithmic fairness is not a one-size-fits-all solution, and that different contexts may require different fairness measures, but there is not yet a consensus as to how to select the appropriate measure for a given system.

In an effort to develop a more principled approach to algorithmic fairness that will inform how fairness measures are selected and applied, researchers have turned to the field of distributive justice for guidance. A distributive theory of justice is a normative framework that provides principles and criteria for allocating benefits and burdens among individuals or groups within a society, with the aim of achieving a just and fair distribution. The field can be seen as polarized along an axis from liberal egalitarianism to entitlement theory. Under the liberal view, commonly associated with John Rawls, the chief objective of justice is to equalize allocation across all individuals in a population. In contrast, the entitlement view, associated with Robert Nozick, emphasizes the importance of individual property rights and the freedom to exchange goods and services without interference. On this view, justice is instantiated not by the outcome distribution of a system, but by how resources are acquired and exchanged within a system.

The relationship between algorithmic fairness measures and distributive justice is not yet well understood, but several recent papers have begun to explore this connection. Binns (2018), Hertweck et al. (2024), and Kuppler et al. (2021) have all examined the relationship between algorithmic fairness and egalitarian concepts of justice, showing that fairness measures that measure disparities in the outcome distribution over social groups are predicated on certain assumptions about equality as a foundation for justice. Baumann et al. (2023) develops this further, but rightly points out that this approach is limited to one particular view of justice: "while our approach creates a link between group fairness and different theories of justice, it does not cover theories of distributive justice that are structurally different from [egalitarianism], e.g., Nozick's entitlement theory." Entitlement theory represents a proportion of thought in the field of distributive justice, and protects a different and significant set of values than those represented by egalitarianism, meaning that this omission represents a large gap in the existing literature. This paper will seek to fill this gap by exploring the relationship between algorithmic fairness and entitlement theory, and how the two can be reconciled. In particular, how do issues of entitlement appear in algorithmic decision making? How can these concerns be encoded by algorithmic fairness measures? And what do we stand to lose or gain by conceptualizing algorithmic fairness through the entitlement lens?

In this paper, we will carefully examine the relationship between algorithmic fairness and libertarian justice, and develop a formalism that clarifies the relationship between the two. We will demonstrate that entitlement justice can be encoded within a measure of algorithmic fairness by formulating the problem of fairness on an individual level rather than a group level. We will show that doing so offers a nuanced and context-sensitive means of understanding algorithmic fairness, and that it can be used to inform the selection of appropriate fairness measures for decision making systems.

The rest of this paper is organized as follows. In Section 2, we provide an overview of the existing literature on algorithmic fairness and distributive justice. We draw on the formalism from (Kuppler et al., 2021) and (Corbett-Davies et al., 2023) to create a unified model for understanding algorithmic fairness and distributive justice consistently with each other. In Section 3, we introduce the concept of entitlement justice and discuss its historical development. We contrast entitlement theory with liberal egalitarianism to identify the critical elements of entitlement which must be represented in account of algorithmic fairness, and confront the traditional objections to entitlement theory. In Section 4, we propose a new framework for understanding algorithmic fairness through the lens of entitlement justice. We analyze the implications of this framework for existing algorithmic fairness measures and show an example of how it can be applied to a real-world case study. Finally, in Section 5, we conclude with a discussion of the broader implications of our work and suggest directions for future research.

2 Background

In this paper, we will center our attention on a class of decision problems corresponding to the following formalism. There exists some population of individuals $I = \{i_1, i_2, \ldots, i_n\}$ over whom we must distribute some finitely divisible amount of a resource. A *decision rule* is a mapping $d : I \to \{0, 1\}$, under which for all $1 \leq j \leq n$, i_j receives the resource iff $d(i_j) = 1$. Since the resource is finite, there is a constraint on the number of individuals who can receive positive outcomes. As an example, consider the process of allocating loans over a pool of applicants — for each individual in the applicant pool, we have a binary choice to either approve or deny the loan, and only on acceptance does the individual receive capital. The bank cannot approve all applicants, as they would run out of capital to lend.

An algorithmic decision maker in this setup is a system, particularly a technical system, under which a decision rule is implemented as a set of steps which are applied identically to all individuals. This broadly consists of two tasks. First, a set of covariates $X = \{x_1, x_2, \ldots x_n\}$ must be collected for each individual. Then, a classification scheme $f : X \to \{0, 1\}$ must be used to map each individual to an outcome. Returning to the loan case, a simple example would be the following: Approve a loan to each individual whose household income exceeds the projected cost of living in the geographic location of their residence by at least $10,000 per year. Then our algorithm implementing the decision rule for each applicant takes the following shape:

- Collect covariates $X = \{$household income, geographic location of residence$\}$
- Let $C(x_2)$ = projected cost of living in location of residence, x_1 = household income
- $f(X) = x_1 - C(x_2) \geq 10000$

It is important to note that not all problem domains to which algorithmic decision making is applied can be formulated in this way. For example, applications of AI in natural language translation may not be easily formulated in terms of resource allocation, but the reader may still be concerned with the perpetuation of social biases through the decisions made in translation — for example, issues of underrepresentation of social groups in translated media. While these cases warrant further study, we will not consider them here.

2.1 Algorithmic Fairness Measures

Algorithmic fairness measures operate on the classification scheme f and a set of morally protected characteristics such as race or gender $A \subseteq X$, enforcing constraints on how decisions can be sensitive to such characteristics. These measures are define a notion of fairness which governs how the resource may and may not be allocated across people based on their protected attributes. For example, one may wish to ensure that the algorithm does not discriminate against individuals of a particular race, and may therefore wish to impose a constraint on the algorithm ensuring that race does not change the probability of receiving a position outcome. In order to develop a clear account of entitlement fairness, we will first survey the existing measures of fairness in the literature and identify the aspects of fairness they capture. We can then use this survey to identify the aspects of entitlement justice which are not protected by these measures, and which we will need to encode in our own measure. An overview of measures commonly discussed in the literature is presented below.

A critical point to note in the discussion of algorithmic fairness is the distinction between individual and group fairness. In group fairness, we form groups of individuals based on the protected attributes A, and measure fairness as a statistical property on the distribution of outcomes across these groups. For example, we may measure the proportion of individuals from each group who received a positive outcome from the classification function. In individual fairness, on the other hand, we attempt to judge whether like individuals are being treated alike by the algorithm. For example, two individuals who are identical but for their race should receive the same classification outcome in decision-making domains where race is not relevant.

Both group fairness and individual fairness measures have conceptual shortcomings. The coarse-grained nature of group fairness measures means that there is no guarantee that individuals within a group are treated fairly, only that the algorithm is statistically unbiased in the aggregate, discarding the potential for bias against particular individuals in the group. Individual fairness measures, on the other hand, require a notion of similarity between individuals which is often difficult to define. Never are two individuals truly identical but for their race, and the choice of a similarity metric can encode normative assumptions about the importance of different attributes in determining the outcome of the classification function.

To set the scene for further discussion of how algorithmic fairness is connected to distributive justice, a survey of common algorithmic fairness measures together with their strengths and weaknesses is presented below. We will discuss these measures in the context of the decision problem formulated above. The goal of this discussion is to highlight the limitations of existing algorithmic fairness measures in capturing the normative concerns of entitlement justice. For a more exhaustive presentation of existing measures, see Corbett-Davies et al. (2023).

As a running example, consider the case of a hiring algorithm. The decision problem is to map a set of applicants to hiring decisions. Our algorithm must implement a function to do so based only on the personal information presented in their resumes, which is their education, experience, and cover letter. We want to ensure that our hiring practices are fair with respect to race, and gender.

Definition 1. *Fairness Through Unawareness — f satisfies fairness through unawareness iff*

$$A = \emptyset$$

In other words, the classification function may not receive any morally protected covariates as inputs. So, for our hiring admissions case, we would be forced to remove race and gender from X and only consider education and experience (Corbett-Davies et al., 2023).

This notion is intuitively appealing — how can the algorithm discriminate against one based on their race if it doesn't know their race? However, it is clear that this measure is not sufficient to ensure fairness. For example, if one attended a historically black college, their education may function as a proxy for their race, allowing bias to remain in the algorithm (Datta et al., 2017).

Definition 2. *Demographic Parity — f satisfies demographic parity iff*

$$P[f(X) = 1 | A = a] = P[f(X) = 1] \ \forall a \in A$$

(Dwork et al., 2012).

Demographic parity holds that the probability of a positive output ($f(X) = 1$) should be statistically independent of the protected attributes. This is an easily understandable and measurable criteria for fairness. At a first look, it is appealing — the same number of individuals from each race will be successful in seeking jobs at a particular company. Indeed, in a situation like selection of individuals for a representative committee this measure is appropriate. However in other circumstances such as hiring or parole, difficulties arise.

Under demographic parity, the probability of success between groups must be balanced, but if the base rates of success are unequal between groups, this leads to poor outcomes. For example, if women are much more qualified for a job on average than men, then demographic parity will require that we hire less qualified men in place of more qualified women in order to balance the probability of being selected between gender groups. In other words, the false positive rate will be very high for men while the false negative rate will be very high for women, creating a severely unfair practice (Barocas et al., 2017).

Definition 3. *Equalized Odds — f satisfies equalized odds if given a true outcome O for each individuals, we have*

$$P[f(X) = 1|A = a, O = 1] = P[f(X) = 1|A = b, O = 1] \ \forall a, b \in A$$

$$P[f(X) = 1|A = a, O = 0] = P[f(X) = 1|A = b, O = 0] \ \forall a, b \in A$$

(Hardt et al., 2016).

In cases where the output value of the classifier f disagrees with the true outcome O, we have either a false positive or false negative. Equalized odds requires that the true positive and false positive rates be balanced between groups. This is often thought to ensure there is no disparate treatment across groups. In our hiring case, for example, equalized odds ensures that no one racial group is more likely to be falsely rejected or erroneously hired than another. Given one has been rejected, the probability it was a wrongful rejection is equal regardless of their race or gender.

Equalized odds has two concerns associated with it. Firstly, the condition becomes difficult to satisfy when the base rates of success are unequal between groups, as differing base rates imply differing false positive and false negative rates when the same decision cutoff is used between groups. In our hiring case, once again imagine that women are much more qualified for a job on average than men. This condition in the population would lead to a model distributing more false positives to women and more false negatives to men, and in order to rectify this imbalance, we would need to use a different threshold of qualification for women than for men, sacrificing the accuracy of the model.

A second concern with equalized odds is that it is easily manipulable. Consider the following example from criminal justice. We have a distribution rule which says to allocate a parole to a prisoner if they are very unlikely to recidivate. Due to a history of discriminatory practices and social marginalization, black prisoners are judged as having higher a risk of recidivism much higher than white defendants (Crime and Alliance, 2023). As a result, allocating a parole to a white prisoner has a lower likelihood of being a false positive. Therefore, one could achieve equal false positive rates by *adding* false positives to the white portion of the dataset, resulting in an increase in the number of undeserving white prisoners receiving parole.

As an example of equalized odds gone wrong, consider the COMPAS algorithm (Angwin et al., 2016). COMPAS was calibrated to have equal predictive accuracy across racial groups, but this resulted in a higher false positive rate for black defendants and a much lower false positive rate for white defendants due to unequal base rates.

Definition 4. *Counterfactual Fairness — f satisfies counterfactual fairness iff*

$$P[f_{A\leftarrow a}(X) = 1|X = x, A = a] = P[f_{A\leftarrow b}(X) = 1|X = x, A = a]\ \forall a, b \in A$$

(Kusner et al., 2018). Where $P(f_{A\leftarrow a})$ is the counterfactual value of f if A were set to a.

Borrowing from the language of causal inference, counterfactual fairness posits that the protected attributes may not have any causal effect on the outcome of the classification function. This measure operates on the individual counterfactual—would the output of the classification function have been different if the individual had been different according to some protected attribute? This is a highly appealing notion of fairness. If their protected characteristics do not in any way cause their outcome, then it is difficult to argue that one has been discriminated against. However, this measure is difficult to implement in practice.

Counterfactual fairness is often critiqued based on the difficulty and potential subjectivity of detecting causal links between variables. Recent work on the social construction of demographic variables reveals that causal modeling may have an inherently normative basis (Hu, forthcoming), and even if these issues are set aside, the computational expense of causal discovery can create issues of practicality.

This discussion of dominant algorithmic fairness measures and their critiques reveal that there is no one-size-fits-all solution to the problem of algorithmic fairness. Each measure has its own strengths and weaknesses, and the choice of measure will depend on the specific context in which the algorithm is being applied. Intuitively, one would like to measure a given classifier against a range of measures to ensure it is fair in a variety of ways. However, as previously mentioned, it is impossible to satisfy multiple measures simultaneously, and so the choice of measure becomes a critical one. How should one select a measure? What are the philosophical considerations that should guide this choice? In hopes of developing a more nuanced and structured approach to these questions, we turn to work in the philosophy of distributive justice.

2.2 Theories of Distributive Justice

Distributive justice is a philosophical field of inquiry that examines how to define a fair allocation of goods and resources across a society. A fully fledged account of distributive justice must answer a number of questions. Who should receive those resources which are highly scarce? When and why is it allowed for one person to have more of something than another? By what mechanism can resources be redistributed to achieve justice?

Given that distributive justice defines how fair decisions about allocations can be made, within the formalism we've presented, its role is to broadly define the decision rule which may then be implemented algorithmically. As described in section 1, the dominant theory of distributive justice used in connection with algorithmic decision making is John Rawls' theory of liberal egalitarianism, which we will present here.

Rawls (1971) begins with a thought-experiment called the veil of ignorance. In this thought-experiment, one is asked to imagine themselves in a pre-societal world, working in collaboration with a number of others to determine how resources should be allocated across society once it begins. Critically, all those involved in designing this distribution of goods are unaware of what their own position and endowments in society will be. One may find themselves endowed with a high level of intelligence, or a valuable skill, or wealth at birth, or one may find themselves with none of these things, or the opposite. Without knowing which of these positions one will occupy, Rawls argues that one will be motivated to design a society in which the following two principles are satisfied:

1. Each person has an indefeasible right to the most extensive basic liberties compatible with equal liberty for all.
2. Social and economic inequalities are to be arranged so that they are both to the greatest benefit of the least advantaged, to offices open under fair equality of opportunity. (dubbed the *difference principle*).

Rawls refers to the people in the group of individuals designing the society from behind the veil of ignorance as being in the *original position*. The argument from the original position results in citizens living under a social contract which is guided by the two principles given above. The principles allow us to measure, for any given distribution of resources across society, whether or not the distribution is fair. If the distribution is not fair, then Rawls endorses a program of redistribution to bring the distribution into line with the principles. For example, a society with a high-level of wealth inequality is in violation of the difference principle — the wealth gap represents an economic inequality which is not to the greatest benefit of the least advantaged. In this case, Rawls would endorse a program of redistribution to balance the wealth across the society in accord with the principles given above.

This type of distributive justice theory is what we refer to as an *end-state* theory of justice. The distribution of goods across society represents a discretely evolving state of affairs, and the role of the theory is to determine whether or not each state is just. Let us consider this view in light of the decision problem we formalized above. Liberal egalitarianism tells us that the decision rule d must be such that either all individuals receive the resource of allocation equally, or that inequalities in the allocation of resources must be to the benefit of the least advantaged. Several of the fairness criteria in the literature on algorithmic fairness can be seen as implementing the first condition in terms of equality of opportunity —— by regulating the extent and manner in which protected attributes can influence the outcome of the decision, we attempt to ensure that all individuals are receive equal basic rights of opportunity in the decision process. However, whether or not the difference principle is satisfied by these measures is less clear.

In our loan case, for example, we may be concerned with ensuring that every individual receives equal opportunity to a loan. However, if the decision rule is such that only individuals with a household income above $100,000 per year, or those who are members of a particular race, are able to receive a loan, then this clearly doesn't provide equal opportunity to all. A measure like demographic parity ensures that individuals from each protected group are equally likely to receive a loan, therefore balancing opportunity across groups. However, this doesn't ensure that all persons in the population have equal access to capital, it only ensures that the access of persons from different groups to capital as a direct result of the output of the algorithm is equal. Absent from this metric is any consideration of who is considered to be the least advantaged in the situation. By Rawls, we should be allocating capital to those individuals who have the least access to capital, a condition which is not satisfied by demographic parity. (Hertweck et al., 2023) shows how Rawls' theory can be more fully captured in a group-based fairness metric, which we will not go through the details of here. However, the key point echoed throughout the literature is that the common measures of algorithmic fairness stem from egalitarianism—demographic parity ensures one concept of equality between groups, just as equalized odds does. The measures are all concerned with a basic equality between individuals in the population, which makes it unclear how they can be reconciled with a theory of justice which is not concerned with this type of equality. In the next section, we will discuss Nozick's theory of entitlement, which has no such basic commitment to equality, and which therefore represents a puzzle for algorithmic fairness measures.

3 Entitlement Justice

An entitlement theory of justice is a distributive theory of justice which posits the following decision rule: Allocate amount resources to an agent if and only if that agent is entitled to that resource. An entitlement in this context is a *property right* held by the agent over the resource. Different entitlement theories of justice differ in the criteria they use to determine entitlements, and the concept of property rights they endorse. Here we will detail the entitlement theory of justice as proposed by Nozick (1974), its issues, and how it compares to the Rawlsian theory of justice, then discuss more recent efforts at reconciling the theory with the demands of justice.

3.1 Nozick's Entitlement Theory of Justice

Nozick's entitlement theory of justice, often called the concept of libertarian justice, is a theory of justice that was developed as a fundamental challenge to Rawl's liberal egalitarianism. On the liberal egalitarian view, ensuring justice is an inherently redistributive task. The justice of a distribution of resources is determined by the extent to which it is equal over individuals, and there is an implied moral responsibility to redistribute resources to those who lack them to increase the overall equality of the distribution. This ideology provides a strong defense of taxation and welfare programs, which redistribute resources in order to flatten the distribution of wealth (Rawls, 1971).

Libertarian justice takes issue with the consequences of adopting this view. Nozick asks us to consider a thought experiment. Suppose we began with an equal distribution of resources across society. People in this society have the freedom to choose how to use their resources, and to exchange them with others as they feel is fair. Many people are willing to pay to see Wilt Chamberlain play basketball, and so they each pay him a small amount of money to see him play. Over time, Chamberlain will accumulate a large sum of money through his efforts. The distribution of resources in the society will no longer be equal, but will be skewed towards Chamberlain. On the egalitarian account, this excess wealth that Chamberlain has accumulated is unjust, and must be taken and redistributed across society. On the libertarian view, however, Chamberlain has gained an entitlement to his accumulated wealth, and to take it away from him is akin to stealing. After all, if this wealth is taken away from him, then he will have received nothing for his efforts, and enjoyed no fruits of his labor.

In Nozick's theory, people gain entitlements over resources in accordance with 3 principles:

1. The principle of justice in acquisition: A person who acquires a resource through a just process is entitled to that resource. A process of acquisition is just if the acquisition is in accordance with Lockean proviso (discussed below).

2. The principle of justice in transfer: A person who acquires a resource through a just transfer is entitled to that resource. A transfer is just if the transfer is voluntary and the resource is transferred from someone who is entitled to it.

3. The principle of rectification: A person who acquires a resource through the rectification of a prior injustice is entitled to that resource. Rectification must be proportional to the injustice which is being rectified.

On analysis, one will see that a key difference between this libertarian view and the liberal egalitarian view is the fundamental unit of justice. For the liberal egalitarian, justice is realized in the distribution of resources itself. This approach is referred to as a patterned or *end-state* view of justice. For the libertarian, justice is realized in the process by which resources are acquired and transferred. This approach is referred to as a *historical* theory of justice. In order to determine if the current state of affairs is just with respect to a particular holding, one must trace the history of that holding back to its original acquisition, and ensure that each step in the process was just. For Nozick, any end-state theory of justice is inherently flawed, as it requires the restriction of individual liberties (Henberg, 1977). It is plain that this view of justice hinges strongly on being able to identify and justify the initial acquisition of resources, else the theory can say nothing about the justice of the current distribution of resources.

3.2 The Justification of Acquisition

For Nozick the Lockean proviso underscored the principle of justice in acquisition. The proviso contains two parts. The first part is a mechanism for justifying the initial acquisition of resources. It begins with the inherent right of self-ownership that all individuals possess. Locke argued that when an individual mixed their own labor with a resource, they transferred some of themselves into the resource, and so extended their right of self-ownership over the resource, thereby obtaining an entitlement to it. The second part of the proviso, almost as an afterthought, is a restriction on the extent to which resources can be acquired. It states that a person can only acquire a resource if there is enough and as good left over for others. This restriction is necessary to ensure that the acquisition of resources does not infringe on the rights of others to acquire resources.

Other accounts of entitlement justice have used different mechanisms to justify the acquisition of resources. (Mack, 1990) proposed that the acquisition of resources could be justified as a separate unalienable right that all individuals possess. (van der Veen and Van Parijs, 1985) proposed that the acquisition of resources could be justified consequentially by the net utility that the acquisition brings to society In general, Van Der Veen showed that given a particular type of holding, one can specify a theory of entitlement justice with a corresponding utilitarian theory of acquisition that can be used as a basis for determining entitlements.

3.3 Critiques of Entitlement

Nozick's entitlement theory is heavily criticized for its foundation in the Lockean proviso. The final clause of the proviso provides a restriction on the extent to which resources can be acquired, but is a weak restriction that makes it difficult to justify the acquisition of resources in practice. There are two mechanisms by which the proviso as it pertains to Nozick's entitlement theory breaks down.

Firstly, the proviso is a weak and vague restriction. It was written in an era when it seemed plausible that individuals would frequently be staking claim over new possession in the wilderness, in particular parcels of land. However, in the modern setting, there are few unclaimed natural resources, and those that exist come under heavy contention for acquisition. The proviso does not provide a clear mechanism for dividing up the resources in this case, and it seems entirely unlikely that one can satisfy both aspects of the proviso concurrently (Fried, 2004).

Secondly, the proviso has a problem dealing with the issue of surplus value. According to the proviso, when an individual acquires a resource, they acquire it by instilling some valuable portion of themselves into the resource. There is thus a fixed amount of value transferred onto the resource through the person's labor. However, in a free market like the one Nozick describes in his theory of entitlement justice, the value of a resource is not fixed, it is dictated by market forces. If an individual acquires a resource and then the value of that resource increases due to scarcity or high demand, then the individual can trade their resource and gain entitlement over property with a value greater than that which they instilled into their original acquisition (Fried, 1995).

These issues provide a strong challenge to Nozick's entitlement theory as they can result in disastrous consequences. Besides the proviso, Nozick's theory may be criticized for its potential to justify unacceptable outcomes through transfer. For example, someone who is starving may "voluntarily" agree to trade property for food whose value is far below the value of the property, and per Nozick, this trade might be considered just. Critically, this does not spell the end for the entitlement theory of justice, but it does suggest that the underlying theory of property rights for a successful entitlement theory of justice as well as the restrictions on the types of transfers it can justify must be more nuanced than what Nozick proposed.

3.4 Instrumental Property Rights

Successors of Nozick have sought to address these issues by replacing the Lockean proviso with an alternate theory of property rights. van der Veen and Van Parijs (1985) showed that entitlement systems existed on a spectrum such that the theory of property rights at the base could be tailored to the resource being distributed. For example, the precise theory of property rights for land might be different than that for money, or for a scarce natural resource. This observation suggests that a successful entitlement theory of justice must be based on a theory of property rights that is situated in the context of the resource being distributed. This observation is echoed by Fried (2004) who argues that the theory of property rights must be tailored to the resource being distributed, and that the theory must be created with the full scope of its consequences in mind.

Regardless of the theory of property rights used, to overcome the challenges of Nozick's theory, one must prevent an entitlement theory from justifying morally unacceptable outcomes as Fried (1995) worried about in the case of Nozick's theory. Sen (1988) shows that while the interpretation of property rights as inherently valuable and inalienable leads to severe issues of poverty and hunger, the interpretation of property rights as *instrumental* rights, which are valuable only insofar as they lead to particular desired outcomes, can be used to develop systems of entitlement without such issues. Instrumental property rights cannot supersede the demands of basic necessity for all agents, and so can be used to develop systems of entitlement which protect property rights while avoiding issues that arise alongside emergent wealth disparity from free market transactions.

Combining these lessons, we realize that a successful modern theory of entitlement justice is one situated atop a theory of domain specific and instrumental property rights. For a given type of holding or resource, the theory of property rights must be tailored to the resource, and must be created and enforced with the full scope of its consequences in mind. This approach allows for the development of a theory of entitlement justice that is both normatively justifiable and practically applicable in the modern world, and thus could be used to inform the design of algorithmic fairness measures. A critical result here is that the theory of property rights used to define fairness over a particular domain must be able to be superseded by higher order concerns and should be defended to the population it is applied to to gain acceptance.

3.5 Contrast with Liberal Approach

To begin to craft an account of algorithmic fairness through the lens of entitlement justice, it is useful to contrast entitlement theory and the liberal approach to see the critical dimensions along which they differ. These differences provide a clear set of concerns of entitlement justice that must be addressed by a fairness measure designed to implement entitlement justice.

- Historical vs. end-state — Under an entitlement theory, the justice of a distribution is determined by the history of how the distribution came to be through acquisition and transfer. In contrast under a liberal egalitarian approach, the justice of a distribution is determined by the current state of the distribution itself.

- Individual and collective responsibility — Under an entitlement theory, there is a heavy focus on the actions and properties of individuals which give rise to their entitlements. It is the individual who acquires or trades for resources, and thus it is the individual who is responsible for their own state of affairs. In contrast, under liberal egalitarianism, the only relevant properties of an individual are their current holdings or status in society, and it is a collective responsibility to ensure that resources are distributed according to the demands of justice.

- Redistribution — Under an entitlement theory, redistribution of resources from the more fortunate to the less fortunate is not a moral imperative. In fact, redistribution is unjust if it is not done willingly on the part of the more fortunate. In contrast, Rawls' difference principle explicitly mandates redistribution of resources to the less fortunate.

4 Entitlement Fairness

In order to understand the relationship between algorithmic fairness and entitlement justice, it is important to first analyze the role of the algorithmic decision-maker in the context of entitlement systems. On the entitlement approach, decisions about allocations are made entirely based on property rights. Therefore the task of a decision-maker within our decision problem is clear — the decision-maker becomes a *property rights oracle*. Given a resource and information about a population of individuals, the job of the decision-maker is to determine which individuals hold property rights over the resource. The role of fairness is thus a bit different than under other theories of justice, because we do not start from the assumption of any sort of equality across our population. What type of assumption should we start with instead? To understand this, we will analyze the points of contrast we have drawn between the Rawlsian and entitlement theories of justice.

First, though, we must elucidate the meaning of property as it will be used in this investigation. In the modern legal system, property is typically restricted to a somewhat narrow range of physical objects, financial assets, and intellectual creations of the mind. Here, though, we will take a broader view of property that encompasses any resource one might be entitled to claim and which algorithmic decision-making might be used to allocate. This includes all of the resources we typically think of as property, but also includes things like admission to a particular college. In brief, property will refer to any finite desirable resource that can be allocated to an individual. This is a much broader understanding of property to allow the entitlement framework to be applied across a wide range of decision-making domains. Something like college admissions may not be something we can count or measure in the same way we can conventional property, but it is something one can *earn* a claim to, and something about which there can be a proper question of ownership. Ownership in these systems cannot be legal ownership because the legal system does not and should not have power of authority over allocation in these domains, but there is still a fact about rightful entitlement outside from what the law is able to ensure.

4.1 Fairness in Process

One critical dimension of entitlement justice is that it governs the process that gives rise to a distribution of resources rather than the distribution itself. This means that decisions about allocation which govern whether a particular individual is able to acquire a given resource must be made in accordance with a fair process. Rather than answering whether or not a given decision-maker outputs a fair distribution, we must instead ask whether the manner in which decisions about allocations are made is fair. This is a subtle but important distinction. Under the entitlement approach, the output distribution is allowed to be heavily skewed in favor of some individuals or groups so long as the process by which the output distribution is arrived at is in accordance with the principles of just acquisition and transfer.

As an example, consider the case of college admissions. Under the liberal egalitarian approach, we mighty ask whether the output distribution of our algorithm is fair by asking whether it results in a roughly equal number of students admitted between race A and race B. Under the entitlement approach, however, there is no reason to ask this question—we might find that the students entitled to admissions are 90% members of race B. Instead, the relevant questions are how race is used in the admissions process — are students from race A subject to the same rules of acquisition? Drawing on the framework of the Lockean proviso, given that two students, one from each race, have expended equal effort in their applications and studies, are the values endowed in their applications treated as equal? Whereas the egalitarian approach is that everyone is naturally entitled to an equitable share of the educational resources in society, the entitlement approach holds that there are particular actions and features of individuals that earn them the right to a higher level of education. Just what these features are depends on the context and problem domain of the decision-making system. Unlike the egalitarian approach, we cannot start from the assumption of equality, and instead must start by defining the set of relevant features that give rise to an individual's entitlement. From this analysis, it is evident that to encode entitlement fairness, we cannot simply measure the outputs of an algorithm, but rather must analyze the treatment of features of individuals within the algorithm itself.

4.2 Individual Responsibility

Our second critical dimension of entitlement justice is that it places emphasis on the actions and properties of individuals, while de-emphasizing the role of group membership. Under the entitlement approach, individuals perform actions that give rise to or forfeit their property rights. Fairness must therefore be fundamentally based on a set of features of individuals which are relevant to the process of determining property rights. These features will generally not be simple demographic features or features encoded in the input covariates in a straightforward way, but rather will be a set of more nuanced features that must be predicted on the input data by a complicated high-dimensional model. For example, in the case of college admissions you may want to predict a feature like "academic potential" which will be difficult to extract, but which is relevant to the entitlement of an individual to be admitted. Entitlement features can be sorted into two broad categories:

- *Positive Entitlement Features*: These are features of an individual that give rise to property rights. For example, an individual's effort in an application process might be a positive entitlement feature that gives rise to their right to be admitted to a college.
- *Negative Entitlement Features*: These are features of an individual that forfeit their property rights. For example, if an individual cheats on their application, or performs very poorly on an entrance exam, these might be negative entitlement features that forfeit their right to be admitted to a college.

Notice that given the full set of relevant positive and negative features of an individual that determine their property rights, we should be able to fully determine the individual's entitlement to a resource and complete the decision problem. In other words, once we have identified the features and computed them on an individual, the output of the decision maker can be fully specified by these features alone. This lends itself to a natural understanding of fairness in process — the process of mapping an individual to their decision should be fully decided through the relevant features identified. These features should be explicitly justified and made transparent to the individuals affected by the decision.

Returning to our discussion of modern entitlement theories, we can recognize that the identification of relevant features is how context-specificity will enter into our account of algorithmic fairness. In each problem domain for algorithmic decision-making, there will be a different set of positive and negative features that are relevant to the entitlement being decided. By identifying and justifying the set of features relevant in each domain, we allow our account of fairness to be sensitive to the context and nuances of the problem at hand. This is a powerful contrast to typical approaches to fairness under which we attempt to identify a universal measure of fairness that can be applied across all domains.

In application, this implies a particular set of structural conditions that must be followed to implement an algorithm which is fair under the entitlement approach. A classification scheme must be developed that first computes the value of each of the relevant positive and negative entitlement features for each individual in the population. Then a decision must be reached through only those computed features, in isolation from the full input data to the algorithm. Note that this is *not* equivalent to fairness through unawareness. Protected attributes of individuals will be used to learn the features relevant to the decision problem, and the features themselves will encode the information about the protected attributes. Rather than trimming the information about the protected attributes from the input data, we are instead using it to compute relevant features which are then used to make reasons-based allocation decisions. This may seem to stand in stark contrast to the way that many machine learning algorithms are currently developed. For example, in a typical supervised learning setting, a model is trained to map a set of inputs to a set of outputs according to a high-dimensional loss function, with little regard for the manner in which the inputs are processed. However, the internal structure of neural networks and other machine learning models gives rise to a set of features computed by the model, constituting a lower-dimensional representation of the input data (Liu, 2018). The approach we suggest here can be thought of as a way of manually specifying a lower dimensional set of features that are relevant to implement individual fairness over as a way of exercising control over *how* the model makes decisions in order to implement fairness as a process.

What would selection of these features look like in practice? Consider again the case of college admissions, and in particular, how decision are made about admissions of students who have less access to resources and academic opportunities. There are several features that are not straightforwardly encoded in the input data but are certainly relevant to the entitlement

- Firstly, we might want to extract a feature that captures the notion of current academic performance. This is likely a function of GPA, test scores, and other typical academic indicators, and is justified by an appeal to the idea that students who perform well academically are more likely to succeed in college. Likelihood of success is made relevant by the fact that individuals who are likely to succeed at the university will maximally benefit both themselves and the university, and therefore justify their admissions in a free market of talent. This positive feature represents an action of the individual that gives rise ("earns") their right to be admitted. This stands in contrast to the Rawlsian approach, under which there is no broad support for meritocratic systems. Academic performance for Rawls is likely a product of natural talent and luck, and therefore by Rawlsian standards, is not a valid basis to allocate admissions inequitably.

- In contrast, we should also say that a student is entitled to admissions if they have demonstrated a stronger work ethic and commitment level than their peers, even if they attended a lower income school and thereby had less access to advanced classes and tutoring resources. Effort and commitment demonstrate a greater value endowed into an application, and therefore a higher degree of entitlement. Here the difference with Rawls is more subtle. Rawls would agree that the student from a lower income school and with less access to resources should be promoted in admissions, but as a redistributive effort rather than as a recognition of the value endowed in the application. On the entitlement approach, two students who have expended equal work and effort should be treated equally in this respect, regardless of their background.

- Finally, we might also want to consider a student's cultural fit and addition to the campus community. This feature reflects a student's entitlement on the basis of more than just academic merit — a student who provides cultural value to a university provides similar, though reduced value to the university as a student who provides academic value, and therefore has a similar, though lesser, entitlement. This feature can be subdivided into both a positive and negative feature. A student who has demonstrated positive cultural value to the community gains an entitlement, while a student who has demonstrated negative cultural value forfeits their entitlement earned through academic merit and effort. Here Rawls again pulls apart from the entitlement approach. Under the Rawlsian view, cultural fit, so far as it is a product of natural tendencies and socialization, is not a valid basis for admissions.

An algorithm meant to implement entitlement fairness in college admissions would then consist of predictors which extract each of these values for each individual, and then a system of mapping these values to a decision about admissions. Of course, the standards used for entitlement in this domain and others are up for debate, and whatever set of features is selected must be justified and defended in a public forum.

Now, one critical dimension of fairness remains to be discussed — how do we ensure that the predictors which compute the value of the relevant features for each individual are themselves fair? To understand this question, we should delve into the basis of the entitlement features themselves. Positive features of an individual give rise to property rights through the mixing of one's self with the subject of the entitlement itself. They are something that the individual does to imbue the subject of the entitlement with value. An error in the prediction of a positive feature results in the devaluing of an individual— either one is not given the credit they deserve for their work through a false negative, or one is given credit for work they did not do through a false positive, devaluing the work of others. For negative features, a similar argument can be made. A false positive on a negative feature results in an individual being punished for something they did not do, while a false negative results in an individual being rewarded despite having done something disqualifying. In either case, we can see that errors in the prediction of the features result in a violation of the entitlement of an individual. Under Nozick's entitlement theory, there is no one individual or group for whom violations of entitlement are more or less important. Therefore, there is no no reason to enforce any particular pattern of prediction errors in the feature predictors. Rather, what matters, is that the predictors are as accurate as possible, resulting in the fewest possible violations of entitlements. However, no predictor will remove all errors, and therefore we must also ask how we should handle the errors that do occur.

Turning once more to Nozick's entitlement theory, we see that the theory provides clear direction in this respect. In his theory of justice, Nozick puts forward a principle of rectification, which states that if an individual has been wrongfully deprived of their entitlements, the subject of a past injustice, then they are entitled to a rectification of that injustice which will result in them having all those holdings which they would have had if the injustice had not occurred. Clearly we need to build such a system of rectification into our account to correctly reflect Nozick's entitlement theory. Given that the features we choose appropriately judge the entitlements of individuals, we see that the only way the system commits an injustice is if the features themselves are mis-predicted. As a result, we can see that we must offer a system of rectification for individuals who have been wrongfully deprived. This has two major implications. Firstly, an individual must be informed of the features used in their decisions and their values over those features. Second, the individual must have a means of appealing their prediction for particular features, and if proven the prediction was incorrect through investigation, the individual must be compensated for the error. This investigation should be human conducted, resulting in human oversight of the decision-making process where the algorithm is unable to provide a fair decision.

4.3 Entitlement Fairness and Redistribution

A third critical dimension of entitlement justice is that it rejects the notion of redistribution. Under the entitlement approach, once property rights have been established, they must be respected and protected, and no attempt should be made to redistribute resources in order to achieve a more equal distribution. Note that this is already successfully encoded in the system we have described. Once the set of relevant features have been identified, the decision problem output must be entirely separated from the input data given the relevant features. This means that no redistributive scheme may be implemented after the property rights decision that attempts to balance the distribution of resources over less fortunate individuals.

We are, however, offered a mechanism through entitlement theory for improving outcomes for those who have been wrongfully deprived of their entitlement through the principle of rectification. For example, if we find that a particular group of individuals has been systematically excluded from a resource due to a historical injustice, rectification allows us to then consider membership in that group to be a positive entitlement feature, and to thereby account for it within the decision problem. This is a powerful mechanism for addressing historical injustices and allows us to consider the broader social context in which our algorithm is situated.

Importantly, rectification is not inherently redistributive. *Lacking* a particular resource does not give rise to a claim to that resource, unless the lack was produced by an injustice. An injustice can be done to one who is wealthy in the subject of the entitlement, or to one who is poor, and in either case rectification operates the same way.

4.4 Measuring Fairness

Having developed a qualitative account of entitlement fairness, we may now formalize our account in a way that allows us to measure it. We can define a measure of entitlement fairness as follows. Given our typical decision problem, identify a set of relevant features $V = \{v_1, v_2, \ldots, v_i\}$, where the value of each feature is a discrete variable which can take on a finite number of values. For example, the feature "academic performance" might take on the values "poor", "fair", "good", and "excellent". Implement a predictor for each feature, which maps from the covariates of an individual X to a predicted value of the feature, $v_j \approx \hat{g}_j(X)$. Assume that the predictor $\hat{g}$ is chosen from a class of possible predictors G_j for each feature, and define the error of the predictor as $E(\hat{g}_j) = \mathbb{E}[|\hat{g}_j(X) - v_j|]$. Recall from the setup in section 2 that f is the function $f : X- > \{0, 1\}$ which determines from the covariates of an individual whether they are entitled to the resource in question. Now, these are the conditions for entitlement fairness:

1. $P[f(X_1) = 1|V_1] = P[f(X_2) = 1|V_1]\ \forall X_1, X_2$ (Independence Condition)

2. $\hat{g}_j = \underset{g_j \in G_j}{\arg\min}\, E(g)\ \forall j$ (Max Accuracy Predictor Condition)

In addition to these mathematical conditions, we also require that

3. The individuals affected by the decision have suitable opportunities to appeal the decision of the predictor for rectification.

4. The importance of the features used in the decision problem be justified and defended in a public forum.

5. The values of the features predicted by the model be made transparent to the individuals affected by the decision.

These conditions together provide a context-sensitive and principled approach to algorithmic fairness derived from the entitlement approach to justice, and provide a framework for understanding how to select and apply fairness criteria in algorithmic decision-making.

The role of the first condition is to ensure that the decision-maker is correctly screened off from the input data by the features of the individual. This means that the output of the decision-maker is fully determined by the relevant entitlement features, such that the process by which the decision-maker reaches a decision is fair according to the principles of entitlement justice as described above. The second condition ensures that the features are predicted as accurately as possible, and that the decision-maker is producing as few violations of entitlement as possible. The requirements for feature transparency and rectification ensure that decisions are made according to features which are agreed upon as relevant in a public forum, and that when a decision is made that does violate entitlement, the individual has the opportunity to appeal the decision and have it rectified according to Nozick's theory. This provides a context-sensitive approach to fairness under which algorithmic systems and their designers are accountable to the public who they make decisions about.

4.5 Entitlement Fairness in Practice

To see how the entitlement fairness measure is applied in practice, and how it compares to existing, egalitarian measures of fairness, we will consider the college admissions example in detail. In this example, we will consider a decision problem in which a university is attempting to allocate a set of admissions to a set of students. We will focus on the division of two groups of students, the red group and the blue group, into admitted and rejected students. The university historically accepts far more students from the red group than from the blue group.

The university is interested in allocating admissions to students in a way that is deemed fair, but there are some severe statistical disparities in the students who are applying. Red group students are much more likely to be interested in applying to the program A, while blue group, who are historically disadvantaged, students are much more likely to be interested in applying to program B. At this college, the program B is much more competitive than program A, because program A is able to accept a much larger number of students. The university is developing a single central admissions algorithm to handle admittance across all programs. The algorithm must must decide which students to admit so that the output of the model is admits the right total number of students to the university, and the right number of students to each program. The university is also interested in admitting students fairly, and so want to understand what fairness measure can be applied to the whole algorithm or output distribution to ensure it is fair.

Under conventional measures of algorithmic fairness, the university is in a difficult position. They can admit only a small number of students to program B, and since the majority of blue group students favor program B, they are likely to be rejected at a much higher rate than red group students. This would be opposite from a purely Rawlsian aproach, in which the historically disadvantaged blue group students should be favored in the admissions process. To enforce this advantage, the university could implement a parity measure, such as equalized odds. However, in doing so, they would be preferring students from the blue group over red group students with equal qualifications—given two students with equal qualifications, one from each group, who both applied into program A, the algorithm would certainly prefer the blue group student in order to satisfy the output distribution constraints. In the end, balancing outcomes across the two groups is a bad idea, as the unequal base rates of acceptance due to self-selection into the programs will naturally result in a highly unbalanced output distribution. Even applying the metric separately to each program fails to resolve the difficulty, since the base rates of acceptance and of interest within each group within each program are different. Correcting for this imbalance results in a system that is inherently redistributive, and chooses to favor demographic group membership over features relevant to the decision problem.

In contrast, the entitlement fairness approach has no qualms about the inequity in the output distribution. The university is free to admit a demographic mixture of students which is heavily skewed in favor of the red group, in accurate reflection to the true distribution of qualifications combined with self-selection into programs. The university must choose the relevant features to the domain. From the example described above, we borrow the feature set so that the positive entitlement features are academic performance, effort, cultural fit, and potential in the specific program. The negative entitlement features are cheating on the application and negative social media presence. Now, the university is required to use these features only to make decisions about who to admit. Notice that group membership is not a feature at this stage of the decision process. Information from group membership may be relevant to the decision, but only in the sense that it is used to compute the values of the features. As an example, if a student is a member of the blue group, and the blue group is socioeconomically disadvantaged, then the blue group student may receive a high effort score for an equal GPA and test score to a red group student. Once this feature is extracted, however, the decision-maker is no longer allowed to use the group membership in the decision process, meaning that any information extracted from the group membership is used in a way that is consistent with a principled theory of justice. Compare this to the use of group membership in the egalitarian measures, where the group membership is used to transfer the decision-maker's attention away from features which may be relevant to student's qualifications to achieve a statistical balance across groups.

Finally, under the entitlement fairness approach, the university is required to publish the features used in the decision process and their values for each individual. The university must then provide a rectification process for the students, so that if they feel their values were mis-predicted, they can appeal the decision. Notice that this type of information could be thought to satisfy the the type of explanation that is required by the EU's GDPR (gdp, 2016) as an additional benefit of the entitlement fairness approach.

5 Discussion

The entitlement fairness measure derived here is a novel approach to algorithmic fairness and represents a first attempt to formalize an account of justice that is consistent with a historical rather than end-state theory of justice. Here we will briefly discuss the benefits and limitations of this approach, and suggest some directions for future research.

5.1 Benefits of the Entitlement Fairness Measure

The entitlement fairness framework has several advantages over existing fairness measures. Firstly, it is a measure which has baked in context sensitivity. In selected a set of positive and negative features which are relevant to the decision problem at hand, the fairness measure is able to capture concerns about the decision problem that could not be captured by simple disparity measures across groups.

In each problem domain the framework is applied to, the system designer must make conscious choices about the features that are allowed to be used in the decision process, and must defend those choices in light of the particulars of the domain. The framework cannot be implemented in one context, and then naively applied to a different context which it might not be appropriate for, since its application requires careful consideration. This stands in contrast to existing measures of algorithmic fairness, which can be applied to any system without extensive consideration of the idiosyncrasies of the problem domain. For example, consider the case of the COMPAS algorithm, in which the original designers of the algorithm measured its fairness using predictive parity across races, which is a measure of accuracy across groups. Predictive parity is appropriate for some problem domains, particularly those where the base rates of the predicted outcome are the same across groups. The COMPAS system was later revealed to be discriminatory based on differing false positive rates across groups due to a large discrepancy in the base rate of recidivism between racial groups. When using the entitlement fairness measure, there may be no simple lift and shift of the measure from one system to another as there was for the predictive parity metric into the COMPAS system, and as a result cases where an unsuitable measure is applied to a system are less likely to occur.

Secondly, the entitlement fairness measure is one that is well-aware of its own limitations and assumptions. Part of the requirement of entitlement fairness is the baked-in necessity for a system of rectification for entitlement violations. This holds the system implementer accountable for the decisions made by the system and for providing remedies for those harmed by the system. A system cannot be entitlement fair without a way to appeal and rectify the decision made by the system, meaning that the system is always open to scrutiny and correction. One key result is that one can't simply measure their system against one definition of fairness, declare it to be fair, and then put the system into production without oversight and recourse. This is a key difference from existing fairness measures in that it increases the accountability of the system implementer to the system's users to maintain the system's fairness over time and provide support for their users.

The final benefit of the entitlement fairness measure is that it gives a high level of transparency to the population it is employed on. Not only do individuals who decisions are made on have recourse to appeal their decisions, but they have a clear understanding of the features that are being used to make the decision and the justification for those features. This is a key result of the entitlement fairness measure, as it allows some small level of explanation to be provided to the individuals who are affected by the system. This is a key difference from existing fairness measures, which may help to soothe concerns about how some particular variables are used in the decision process, but not actually provide or identify the features that are used in the decision process. This transparency could be considered a sort of explanation for the decision process as now legally required in some jurisdictions.

5.2 Limitations of the Entitlement Fairness Measure

The entitlement fairness measure is not without its limitations. The framework is much more difficult to apply than existing fairness measures, as it requires a great deal more effort up front to identify relevant features to the problem domain and the arrangement of a system of rectification. This is a key current limitation of the system, as it is unlikely that many organizations will be willing to put the framework into practice due to the high cost of implementation and the need for a high level of transparency. Keeping up a system of rectification requires human effort and resources and is likely to be a significant burden on the organization.

The entitlement fairness measure is also highly limited by the requirement of justification of the features used in the decision process. While an organization may provide justifications for the features they use in their decision process, it is highly unlikely that the justifications will be universally accepted or uncontroversial. The result is that by publishing the reasoning behind the features used to make the decision, the organization is likely to open itself up to a great deal of scrutiny and criticism from the public, and possibly gaming of the system—more than it would have received if it had simply used a standard fairness and not published extensive justifications. While this is an unfortunate consequence of the framework, it is also a typical consequence of any change that promotes transparency and accountability and is a necessary step in the development of a more principled approach to algorithmic fairness.

5.3 Future Directions

While this work has made significant strides in understanding the relationship between algorithmic fairness and entitlement justice, several important questions remain unanswered.

First, the framework presented here is limited to the case of binary classification problems. While the framework can be extended to multi-class classification problems, the extension is not straightforward and requires significant additional work. The extension of the framework to multi-class classification problems is a key area for future research, as many real-world applications of algorithmic decision making are multi-class classification problems.

Second, the framework presented here is limited to the case of discrete features. The extension to continuously valued features opens new questions around how to define the normative relevance of features and how to measure the success of a particular feature in the decision process. Similarly, rectification becomes much more difficult in the case of continuously valued features, as the system implementer must now decide what threshold values represent a mistake by the algorithm that requires rectification.

Finally, though the framework presented here is a coherent way to understand entitlement justice in the setting of algorithmic decision-making, a large step in the development of the approach would be to develop a test system which implements the framework and demonstrates its utility in a real-world setting. Though many examples are provided in this paper to demonstrate the capabilities of the framework, a real-world system would validate and strengthen the arguments made in this paper.

References

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj, 2016. Official Journal of the European Union, L 119, 4 May 2016, pp. 1–88.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed: 2025-01-12.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning. In *Fairness and Machine Learning*. fairmlbook.org, 2017. Online book, available at https://fairmlbook.org.

Joachim Baumann, Corinna Hertweck, Michele Loi, and Christoph Heitz. Distributive justice as the foundational premise of fair ml: Unification, extension, and interpretation of group fairness metrics, 2023. URL https://arxiv.org/abs/2206.02897.

Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 149–159. PMLR, 2018. URL https://proceedings.mlr.press/v81/binns18a.html.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017. doi: 10.1089/big.2016.0047. URL https://doi.org/10.1089/big.2016.0047. PMID: 28632438.

Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness, 2023. URL https://arxiv.org/abs/1808.00023.

Crime and Justice Research Alliance. Black men have higher rates of recidivism despite lower risk factors, 2023. URL https://crimeandjusticeresearchalliance.org/black-men-have-higher-rates-of-recidivism-despite-lower-risk-factors/. Accessed: 2025-01-21.

Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. Proxy non-discrimination in data-driven systems. *CoRR*, abs/1707.08120, 2017. URL http://arxiv.org/abs/1707.08120.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, pages 214–226. ACM, 2012.

Barbara Fried. Wilt chamberlain revisited: Nozick's "justice in transfer" and the problem of market-based distribution. *Philosophy & Public Affairs*, 24(3):226–245, 1995. doi: https://doi.org/10.1111/j.1088-4963.1995.tb00030.x.

Barbara Fried. Left-libertarianism: A review essay. *Philosophy & Public Affairs*, 32(1):66–92, 2004. URL https://law.stanford.edu/publications/left-libertarianism-a-review-essay/.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3315–3323. Curran Associates, Inc., 2016.

M. C. Henberg. Nozick and rawls on historical versus end-state distribution. *The Southwestern Journal of Philosophy*, 8(2):77–84, 1977. ISSN 0038481X, 21541043. URL http://www.jstor.org/stable/43155157.

Corinna Hertweck, Joachim Baumann, Michele Loi, Eleonora Viganò, and Christoph Heitz. A justice-based framework for the analysis of algorithmic fairness-utility trade-offs, 2023. URL https://arxiv.org/abs/2206.02891.

Corinna Hertweck, Christoph Heitz, and Michele Loi. What's distributive justice got to do with it? rethinking algorithmic fairness from the perspective of approximate justice, 2024. URL https://arxiv.org/abs/2407.12488.

Lily Hu. Normative facts and causal structure. *The Journal of Philosophy*, forthcoming. To appear.

Matthias Kuppler, Christoph Kern, Ruben L. Bach, and Frauke Kreuter. Distributive justice and fairness metrics in automated decision-making: How much overlap is there?, 2021. URL https://arxiv.org/abs/2105.01441.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness, 2018. URL https://arxiv.org/abs/1703.06856.

Yu Han Liu. Feature extraction and image recognition with convolutional neural networks. *Journal of Physics: Conference Series*, 1087(6):062032, sep 2018. doi: 10.1088/1742-6596/1087/6/062032. URL https://dx.doi.org/10.1088/1742-6596/1087/6/062032.

Eric Mack. Self-ownership and the right of property. *The Monist*, 73(4):519–543, 1990. doi: 10.5840/monist19907343.

Robert Nozick. *Anarchy, State, and Utopia*. Basic Books, New York, 1974. ISBN 978-0465097203. Proposes the entitlement theory of justice as a response to distributive justice theories like Rawls's.

John Rawls. *A Theory of Justice*. Harvard University Press, Cambridge, MA, revised edition edition, 1971. Original edition published in 1971, revised edition in 1999.

Amartya Sen. Property and hunger. *Economics and Philosophy*, 4, 1988.

Aaron Smith. Public attitudes toward computer algorithms, November 2018. URL https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/. Accessed: 2025-01-12.

Robert J. van der Veen and Philippe Van Parijs. Entitlement theories of justice: From nozick to roemer and beyond. *Economics and Philosophy*, 1(1):69–81, 1985. doi: 10.1017/S0266267100001899.

This is the end of the collaboration with Llama. Beyond this point, all text was written by human authors without the assistance of LLMs.

4 Analysis

To analyze the resulting paper from the experiment, we will present the role and contributions of the LLM in each phase of the paper writing process. There is no clean accounting of which words or ideas were produced by the LLM and which were produced by the human author, as the two were in constant dialogue in both developing the argumentation and writing the text. However, we will present a qualitative analysis of the contributions of the LLM in each phase, and show illustrative examples of the LLM's contributions and activities in each phase. The full logs of all interactions with Llama are available in the appendix.

Note that in general it is not so obvious how one should evaluate the success of co-authorship, even with a human coauthor. The goal of the project was to produce an original and high-quality contribution to a field, but even if the LLM had not been able to produce any original ideas or argumentation, a high-quality paper could still have been produced through the effort of the human authors alone, meaning that a high quality paper does not necessarily indicate that the LLM performed well in the collaboration. Instead, the analysis of each task will be guided by a few specific questions that may be of interest to the reader:

1. What positive contributions did Llama make to the task?
2. What were the limitations of Llama in this task?
3. How did Llama's contributions compare to those that a human coauthor could have made?
4. What, if anything, did Llama contribute that appears to be original or novel ideation?
5. What was the experience of using Llama as a coauthor in this task like compared to working with a human?
6. If given the chance to redo this phase of the project, would you choose to work with Llama again?

4.1 Literature Review

The first phase of the project was to carry out a literature review, interrogating the vague notion of the intersection between algorithmic fairness measures and distributive justice. Working through this process requires identifying gaps and frontiers in the literature in this area. A coauthor in this phase of research might suggest relevant papers to read, or suggest a particular angle to investigate based on their own expertise.

Completing this phase of the process with the LLM is critical for the experiment, as including the relevant source and background materials into the context memory mechanism is a necessary step to ensure that the LLM is on topic and able to pull text from the relevant literature throughout the remainder of the process. In a human collaboration, this step would be similarly critical in a more reciprocal way. Both human authors need to have a pool of shared knowledge from the literature to draw from and stimulate discussion. The question then becomes whether the process with the LLM will be similarly two-sided; will Llama be able to pull useful papers and ideas from the literature to suggest to the human author? Or will this phase become independent literature review on the part of the human author, compounded with having to teach Llama the relevant background knowledge?

As an entry point to this phase, we often asked Llama to suggest papers to read about a particular topic or concept, and to summarize those papers. For example, "I'm interested in learning about entitlement theories of justice, particularly their benefits and how they apply in the modern world. What papers can I read?" Llama would reply with a list of papers and brief summaries of each, some of which existed, and some of which did not as you would expect from an LLM. But the goal is not simply to cite existing papers, but for those papers to be relevant and helpful in understanding the topic at hand. Llama has two very bad habits in this regard. The first is to cite the foundational papers in the field that the human author is clearly already familiar with, and the second is to suggest readings that are only vaguely related to the subject of the question. So for example, Llama suggested I read "Anarchy, State, and Utopia" by Robert Nozick which we had already discussed extensively earlier in the same conversation, and that I read Locke's "Second treatise of Government," which does discuss protections of individual property rights, but provides no positive account of distributive justice. However Llama was also able to suggest some papers which were relevant and interesting, such as "Self-Ownership and Property Rights" by Eric Mack, which is a more recent positive account of entitlement justice than Nozick's.

Recommending real and relevant papers is not the only way to draw on the LLM in the lit review phase. Most papers 'hallucinated' by Llama have hallucinated titles, but not authors or publication years. Many of the papers suggested by Llama that were not real are variations on real papers or are given titles reflecting thoughts of the author that may be found in other papers. For example, Llama suggested reading "Entitlement Justice: A Review of the Literature" by David Miller. This paper does not exist, but on further inspection, David Miller wrote both a Stanford Encyclopedia of Philosophy entry on Justice that provides an overview of the literature on Entitlement Justice, and a book called "Principles of Social Justice" which offers a critical perspective on Nozick's theory.

Thus in literature review, Llama is helpful in finding new and relevant research to read — but critically, less helpful than a human coauthor would be. When a human coauthor suggests a list of papers, there is no need to check which ones are real or not, and a lesser need to interrogate the relevance of the papers. These are extra tasks that must be completed by the human author that do not enhance the process of writing the paper. In this way, Llama is a less effective coauthor than a human, but "better than nothing."

Llama is more helpful throughout literature review when it is asked to summarize the arguments of a paper or to compare and contrast two papers. For example, "Provide a summary of the paper 'Counterfactual Fairness' for me. Compare the measure proposed with equalized odds fairness." This question results in an accurate and nuanced description of the paper, and comparison of the weaknesses and strengths of equalized odds and counterfactual fairness. While this is a task that could be completed by a human coauthor, it is not a task that any coauthor would be willing to complete frequently or on demand. The tradeoff is somewhat obvious and expected—Llama *can* give you reading suggestions, summarize arguments and compare papers on demand and in real time, but with a significant effort on the part of the human author to extract the results.

4.2 Research Question Development

Beyond searching and summarizing the literature, Llama did have something to offer in terms of new and interesting areas to explore through our research question. When prompted to connect the Rawls/Nozick divide to the literature on algorithmic fairness with no context beyond short descriptions of the two theories, Llama suggested the following:

"In the context of algorithmic fairness measures, Rawls' theory can be seen as advocating for a more egalitarian approach, where algorithms are designed to minimize biases and ensure equal opportunities for all individuals. This might involve techniques such as data preprocessing, feature engineering, and regularization to reduce the impact of biased data.

Nozick's theory, by contrast, might be seen as more permissive, allowing for the existence of unequal outcomes as long as they are the result of just means. This could lead to a focus on individualized fairness measures, where algorithms are designed to optimize outcomes for specific individuals or groups, rather than striving for overall fairness.

In practice, algorithmic fairness measures often aim to strike a balance between these two perspectives. For example, algorithms might be designed to minimize bias while still allowing for individualized treatment based on relevant characteristics. This can involve techniques such as fairness-aware optimization, debiasing, and explainability."

This response is, admittedly, a bit vague and high level, but note that it suggests an interesting avenue to explore that we cannot find any reference to in the literature. The idea that the dichotomy between Nozick and Rawls should be compared with and contrasted with the typical dichotomy between individual and group fairness measures from algorithmic fairness is an interesting one, and turned out to be a fruitful line of inquiry for the paper *that we would not have otherwise explored.* On follow-up questions, Llama expanded this area of inquiry to point towards the literature on counterfactual fairness and how it relates to individualized fairness, and suggested several relevant papers and authors to read on this topic. This is a complex and nuanced approach to the literature that Llama recommended, and one that we haven't found explicitly examined elsewhere, nor one which we would've thought to explore so quickly without Llama's prompting. Clearly, this is not a case in which Llama was used as a tool to extract, refine, or sharpen our own ideas, but rather a contribution of new external ideas into the research question development phase.

The manner in which Llama recommended this line of inquiry was less explicit and direct than a human coauthor would be, and the ideas included were not as sharp or sophisticated. However, there is a level of freedom in the engagement about these ideas that is not present in a human coauthor. When a human contributor suggests an area of inquiry, it is accompanied with some impetus to pursue that line of inquiry and to include it in the paper. This has both its advantages and drawbacks. On the one hand, it is a good way to ensure that the paper is robust and well-rounded, and that the authors are not simply pursuing their own interests. On the other hand, it can lead to a longer research program in total and reduce the coherence of the argumentation. Llama has no such expectations, and can be prompted for ideas to explore as frequently as desired. This is a double-edged sword, as it can lead to the discovery of productive ideas that the author is interested in exploring, but can also generate wild goose chases that are ultimately unproductive.

4.3 Argumentation

In argumentation, our hope for a coauthor was that it would be able to elucidate and defend relevant, novel claims that represented a contribution to the area of inquiry. This is the task in which Llama was the least helpful, and takes a major backseat to the human author.

Llama is able to produce text that is on topic, and which is structured in a way that sounds like proper academic argumentation. However, the text produced lacks specificity and nuance, and Llama often fails to truly defend its claims in proper detail—none of which should be surprising at this point in time. The question at this phase of development of LLMs is not what limitations they have in producing arguments, which is well known, but rather what meaningful contributions can be extracted from the outputs they produce in jointly developing the argumentation.

Llama contributed to the argumentation in two main ways. The first was to produce text which laid arguments out in a clear and concise manner when prompted with the bulk of the argument. These presentations were often beyond the claim inputted, but not so far beyond as to be considered a novel claim in themselves. For example, when asked to present the claim that "Group fairness measures are irrelevant to entitlement theories of justice," Llama produced an argument which started from elements of Nozick's theory and built up the claim logically to the conclusion, including the consideration of counterarguments and objections that it had not been prompted to consider. The details of the argument produced were not always accurate, and rarely complete, but the structure served as a sort of template for the human author to fill in with their own arguments. In this way, Llama goes beyond the standard capabilities of human contributors, who would be very unlikely to help the author structure this argument with this level of detail. However, the ultimate result is that the critical thinking to really sharpen the argument originates from within the human author, and Llama is not able to *source* the argumentation in a way that a coauthor would.

The second (and more useful) way Llama contributed to argumentation task was by critiquing arguments presented to it. When asked to find gaps in an argument or present objections to a claim being made, Llama was able to reliably locate weaknesses in the argumentation. While almost never capable of filling in the identified gaps, Llama was able to identify them and articulate why they were problematic or weak, in a way that was much more sophisticated than any of its contributions to the argumentation itself (perhaps philosophy content on the internet favors critique over positing new ideas?).

4.4 Writing

Given a motivating research question, a number of relevant sources to call on, and an argument to make, one might hope that the LLM would be able to produce a rough draft of the full text of the paper that could be edited and refined by the human author. This is far from the case.

Given an argument to make, Llama is able to produce text to defend it. Given sources to call on, Llama is able to produce text that develops some background knowledge and context for the paper. Given a motivating research question, Llama can provide a first pass at an introduction and conclusion section that connect the two. However, Llama is not able to tie these ingredients together to produce a coherent and well-structured paper. This is not surprising, as the buffer size of the Llama-2-7b-32K model used is 32,768 token, or around 25,000 words. Given that the length of the paper produced is around 8,000 words, and that the context window consisted of a wealth of discussion, questioning, and sections from other sources, Llama was not able to have the full wealth of information needed to produce the fully fleshed out paper. Of course, more sophisticated models with larger buffer sizes may outperform Llama in this regard. The text produced by Llama is not only in need of line-editing, but instead often gives the impression that Llama is not able to grasp the full arc of the argument being made, and certainly not able to understand it within its full context. It is clear, therefore, that the process of developing the writing must be more iterative, with the human author closely monitoring the output and repeatedly editing the text output by the LLM, as well as filling in the gaps in the produced text.

This process proceeds how one would expect — pass a section of text back and forth between the human author and Llama, with each critiquing and editing different aspects of the text until it is in a form acceptable to both authors. This is a process which is time consuming, and ultimately it is unclear whether the LLM is responsible for any of the writing, or only for the stylistic choices and tone of the section.

How does this compare to a human contributor? With a human, the process will most certainly be less iterative. The quality of the text initially produced by the human coauthor will be higher quality in a positive collaboration, and the author will then be expected to only make small changes to bring the text into line with the general style, tone, and structure of the paper as a whole. Working with the LLM brings the benefit of being able to delete, rephrase, and restructure as one wishes without worrying about the opinion of the coauthor, but this advantage is certainly offset by the time and effort required to wrangle the LLM output into usable content for the paper.

4.5 Revision

The revision task of the paper is the most familiar and encouraged role for LLMs in academic writing. It is common now to ask LLMs to proofread and edit manuscripts for tone and style, and to suggest changes to the text to strengthen the overall presentation of the argument. Besides the familiar and obvious benefits of using Llama for these purposes, there were some less trivial gains here that resulted from using Llama throughout the writing process. Since the full process of writing the paper was held in Llama's context memory, it was able to suggest changes and style considerations that were in the spirit of the original intention of the paper rather than simply the text. For example, Llama suggested on one version of the paper that the introduction should be restructured to include a more explicit statement of the research question. Without doing the whole process in step with Llama, it would neither have known what the research question was, nor whether it was clearly articulated in the introduction.

A human coauthor would of course be able to make similar suggestions, but would not be able to, for example, review ten versions of the paper in the span of a week and suggest changes to each of them. However, if prompted with the research question and an outline of the main argument presented in a paper, Llama likely could have performed just as well as it did when involved with the full end-to-end process. Therefore it isn't necessary to use Llama as a coauthor to gain the benefit of Llama as an editor, but Llama's editing capabilities are one of the nicer features you get when using it as a coauthor, though are not responsible for producing novel or significant contributions in any way.

5 Discussion

The results of this experiment were generally mixed, with Llama performing well in some respect and poorly in others. Of particular notice was that the model, though not specifically prompted to do so, took on a few discrete roles over the course of the project. Here, we will outline those high-level roles, and discuss the overall benefits of using Llama as a collaborator in this manner.

One of the most interesting aspects of this research was the way that Llama took on different and unintended discrete roles in different tasks throughout the writing process. In the literature review, Llama successfully took on the role of a very educated assistant, providing citations and information from external sources that were unknown to the human author. This can be understood as a retrieval mode, where Llama retrieves information from its training data and presents it in a coherent and relevant way. In the research question formulation task, Llama took on the role of a genuine intellectual collaborator in the sense of producing novel insights and ideas that were external to both the human author and the training data. This can be understood as a generative mode, where Llama generates new ideas and arguments that are not simply retrieved from its training data. Then throughout argumentation, writing, and revision, Llama opted into a refinement mode, wherein it took information from the human author and refined it into a more coherent and polished form.

While it was only during the research question formulation task that Llama produced genuinely novel ideas, we should still consider this as a success in eliciting original ideas from the model. The model was able to produce original ideas and arguments that were not simply retrieved from its training data as far as we can tell, and did not stem from the human author's own ideas, meaning that the model was able to genuinely contribute to the research process. This is significant in the face of the LLM usage policies of the journals described previously; should the human author of this collaboration be considered the sole author of the paper? How should the human author have credited Llama for its ideas if not as a coauthor?

If a human coauthor engaged in this level of collaboration, it would be expected that they would be included as a coauthor on the paper. The level of collaboration was much closer than commenting on one's manuscript or providing a few references to the literature. Small sections of the paper are entirely produced by Llama, and the majority of the paper is a product of a novel research direction that originated within Llama. Having demonstrated a current LLM's ability to produce original ideas and arguments, we should consider whether it is ethical to exclude LLMs from authorship. Is some credit due to the designers of the LLM? At the very least, the human author shouldn't be over-credited for ideas that they did not produce. At the very least, if LLMs are to be excluded from authorship, accompanying policies should limit the human author's ability to interact with LLMs in generative ways which lead to the production of new ideas until such issues are resolved. This, however, is far from an ideal solution—the ideas produced by LLMs have the potential to advance human knowledge and understanding in a way that we shouldn't like to see stifled.

It is clear that Llama delivered some value to the project through the contribution of original ideas and arguments, but also clear from the analysis that there are definite limitations to the model's capabilities. Were we to attempt to complete a similar research program in the future, we would certainly use Llama or a model with similar capabilities again, but only in particular phases of the research process and in certain roles. For example, we would certainly perform the literature review task with Llama again, and make full use of its ability to retrieve relevant sources. We would also certainly use Llama in the research question development task again, as it was able to produce novel insights that proved useful. However, in the writing and revision phase, we would use Llama as a critic, and not as a resource to produce the text of the paper.

In trying to cooperate very closely on the writing of the text of the paper, we felt that since Llama was not able to grasp the full arc of the argument of the paper at any given time, it actually proved detrimental to the writing to have Llama produce any of the text. Since Llama was not able to understand the full context and structure, it tended to produce writing that reemphasized points multiple times, or that was overly general. Pulling from the model's training data, it was very good at producing descriptions of background concepts and arguments, but was not very good at demonstrating how those concepts and arguments were relevant to the specific research question at hand, nor at driving the argument forward. The result was a number of intermediate drafts that required significant reframing and editing to be useful. Ultimately, we were left with the impression that writing the full text of the paper alone and using LLMs solely as a source of feedback and critique would have been a more effective approach.

The implications of this experiment for the future of academic writing are profound. LLMs will continue to improve at a rapid pace, and it is likely that their ability to produce original ideas and arguments will increase rapidly along with them. Given that our current policies for LLM usage in academic writing are already apparently outdated, it is clear that they need to be revised to reflect not only the current capabilities, but also the future capabilities of LLMs. This will require a rethinking of the role of LLMs in academic writing, and a consideration of the ethical implications of their use. The legal and ethical implications of LLMs in academic writing are complex and require exploration and discussion.

It is clear that LLMs will play an important role in the future of academic writing, and it is essential that we begin to consider the implications of their use. This paper is a first step in that direction, and we hope that it will inspire further research and discussion on this important topic.

References

[1] Cambridge University Press. Authorship and contributorship for journals. https://www.cambridge.org/core/services/publishing-ethics/authorship-and-contributorship-journals. Accessed: 2025-05-06.

[2] Nature Portfolio. Artificial intelligence (ai) | editorial policies. https://www.nature.com/nature-portfolio/editorial-policies/ai. Accessed: 2025-05-06.

6 Appendix I

The full source code for the chat application used in this study as well as all chat logs are available at the following link: https://github.com/Espeer5/paperInAPaper

7 Appendix II: Fully Generated Paper

As an experiment, we instructed an advanced reasoning model to generate a full paper on the same topic as the one in this paper, to test the capabilities of the current state of the art in AI for generating academic writing. We instructed Claude 3 Opus to write a paper according to the following prompt:

> Write an 8000-word undergraduate philosophy thesis, structured with an introduction section, background section, original argument, and conclusion. Use formal academic tone and cite literature as appropriate. Answer the following prompt:
>
> Algorithms are being used more frequently to make social decisions. There are several proposed fairness metrics that can be used to measure statistical fairness conditions on the output of decision-making models, but it is not possible to satisfy all of them on any model. As a result, we must choose which measures to use on any given problem, and it is not always clear which measure we should prefer. Recent papers have shown a link between algorithmic fairness and theories of distributive justice to clarify how these measures should be thought of philosophically. Egalitarianism dominates much of the philosophical literature cited, which appears to be a natural fit. Algorithmic fairness measures typically measure how goods are allocated to different groups, and search for some measure of equality between them. This structure is egalitarian in measuring quality between entities. However, egalitarian measures only form one side of the literature on distributive justice. Whereas egalitarian theories measure the fairness of particular distributions of goods across society, libertarian theories, particularly Nozick's theory of entitlement justice study the processes by which distributions came about to define fairness. For Nozick, justice is satisfied if a distribution came about through a series of just acquisitions and transfers in a free market. This theory is structurally quite different from egalitarian theories, and it is unclear how a theory structured in this way could be reflected in algorithmic fairness measures. This means a large proportion of the literature on distributive justice has been left out of the discussion of algorithmic fairness measures. Develop an account of algorithmic fairness which is based on Nozick's entitlement justice and devise a measure of fairness which can be used to determine if an algorithm is just according to Nozick's theory.
>
> Your thesis should begin by clearly explaining key fairness metrics (e.g., demographic parity, equalized odds) and the core concepts of egalitarian and libertarian justice. Then, formulate a definition of algorithmic fairness grounded in entitlement theory (e.g., Nozick's principles of acquisition, transfer, and rectification). This definition of fairness should allow for the algorithm to give rise to highly inequitable output distributions across groups if those inequities arose in a manner which would be considered just by Nozick. Apply this framework to a hypothetical or real-world example (e.g., loan approval or hiring), and compare the outcomes to egalitarian models. In each major section, make your assumptions explicit, justify your philosophical claims with textual support or example cases, and anticipate possible objections to the entitlement-based framework.

Here, we present the full output of the model in response to this prompt, followed by a brief commentary on the result. What follows is the model output, with no edits made to the text.

1 Entitlement Theory and Algorithmic Fairness: A Nozickian Framework for Just Decision-Making Systems

1.1 Abstract

This thesis examines the theoretical underpinnings of algorithmic fairness measures through the lens of Nozick's libertarian theory of entitlement justice. While current algorithmic fairness research heavily draws from egalitarian theories of distributive justice, focusing on equality of outcomes across demographic groups, this perspective has neglected libertarian conceptions of fairness that emphasize procedural justice and respect for individual rights. By developing a Nozickian framework for algorithmic fairness, this thesis offers a novel perspective on how algorithmic decision systems might be evaluated against principles of acquisition, transfer, and rectification. The thesis proposes a formal measure of "procedural entitlement fairness" and applies it to case studies in algorithmic lending and hiring decisions. This approach illuminates the philosophical tensions inherent in algorithmic governance and suggests that combining egalitarian and libertarian perspectives may yield more nuanced fairness metrics that respect both group equality and individual procedural rights. The framework developed here demonstrates that Nozick's theories, though rarely applied to algorithmic contexts, offer valuable insights for addressing the normative challenges posed by automated decision-making systems.

1.2 1. Introduction

The increasing role of algorithms in decision-making processes that affect human lives—from loan approvals and hiring decisions to criminal sentencing and resource allocation—has prompted significant ethical concerns about fairness and justice. As algorithms make or inform consequential decisions, questions about what constitutes "fair" algorithmic behavior have become central to both technical and philosophical discussions (Barocas et al., 2019). The field of algorithmic fairness has emerged as a response to these concerns, with researchers proposing various formal definitions and statistical metrics to evaluate whether an algorithm treats different demographic groups equitably.

These fairness metrics, such as demographic parity, equalized odds, and equal opportunity, have predominantly been conceptualized through an egalitarian lens that focuses on the distribution of outcomes across different groups (Corbett-Davies & Goel, 2018). This approach aligns with dominant theories of distributive justice in the philosophical literature, particularly those that emphasize equality of opportunity or outcomes. However, this focus has led to a notable gap: the relative absence of libertarian perspectives on justice, particularly Robert Nozick's entitlement theory, from discussions of algorithmic fairness.

Nozick's theory of justice, as outlined in *Anarchy, State, and Utopia* (1974), presents a radically different conception of fairness than most egalitarian theories. Rather than focusing on patterns of distribution (such as equality), Nozick argues that justice is historical and procedural—a distribution is just if it arose through just processes of acquisition and transfer, regardless of the resulting pattern. This perspective has been largely overlooked in algorithmic fairness research, despite its significant influence in political philosophy.

This theoretical gap raises several important questions: How might algorithmic fairness be conceptualized from a Nozickian perspective? What would a measure of algorithmic fairness based on entitlement theory look like? How would such a measure compare to existing egalitarian metrics? And what insights might a libertarian approach offer to the broader discussion of algorithmic ethics?

1.2.1 1.1 Research Objectives

This thesis aims to address these questions through the following objectives:

1. To elucidate the philosophical foundations of current algorithmic fairness metrics, highlighting their egalitarian assumptions and limitations.

2. To develop a theoretical framework for algorithmic fairness grounded in Nozick's entitlement theory of justice.

3. To formulate a formal measure or set of measures that can evaluate algorithmic fairness according to Nozickian principles.

4. To apply this framework to case studies in algorithmic decision-making, comparing the results with those derived from traditional egalitarian measures.

5. To explore the philosophical implications and practical challenges of implementing a libertarian approach to algorithmic fairness.

1.2.2 1.2 Significance and Contribution

This research contributes to both philosophical and technical discussions of algorithmic fairness in several ways. First, it broadens the philosophical foundations of algorithmic ethics by introducing a perspective that has been underrepresented in the literature. Second, it offers a novel conceptual framework for evaluating algorithmic systems that complements existing approaches. Third, it provides practical measures that can be used by developers and policymakers to assess algorithms from a libertarian perspective. Finally, it illuminates the broader philosophical tensions that underlie debates about fairness in the digital age.

By bringing Nozick's entitlement theory into conversation with algorithmic fairness, this thesis aims to enrich our understanding of what justice means in the context of automated decision-making and to provide new tools for ensuring that algorithms align with diverse conceptions of fairness.

1.2.3 1.3 Methodological Approach

This thesis employs a methodology that combines philosophical analysis with formal modeling. The approach includes:

1. Critical analysis of the philosophical foundations of existing algorithmic fairness metrics, with particular attention to their implicit ethical assumptions.

2. Exegetical analysis of Nozick's entitlement theory to extract core principles relevant to algorithmic decision-making.

3. Formal development of metrics that operationalize Nozickian principles in the context of algorithmic systems.

4. Application of these metrics to hypothetical and real-world case studies, with comparative analysis against traditional fairness measures.

5. Philosophical reflection on the implications, limitations, and potential extensions of the proposed framework.

This methodology allows for a rigorous exploration of how libertarian principles might be translated into the technical language of algorithmic fairness while maintaining philosophical integrity.

1.2.4 1.4 Thesis Structure

The remainder of this thesis is structured as follows:

Section 2 provides background on algorithmic fairness, including an overview of key metrics and their philosophical underpinnings, as well as an introduction to theories of distributive justice with emphasis on the distinctions between egalitarian and libertarian approaches.

Section 3 develops a Nozickian framework for algorithmic fairness, articulating how the principles of just acquisition, transfer, and rectification can be applied to algorithmic decision-making. This section also proposes formal measures for evaluating algorithmic fairness from a libertarian perspective.

Section 4 applies the developed framework to case studies in algorithmic lending and hiring, comparing the results with those obtained using traditional fairness metrics.

Section 5 discusses the implications of the Nozickian approach, addressing potential objections and limitations, as well as suggesting directions for future research.

Section 6 concludes by summarizing the key contributions of the thesis and reflecting on the broader significance of incorporating diverse philosophical perspectives into algorithmic ethics.

1.3 2. Background

1.3.1 2.1 Algorithmic Fairness: Concepts and Metrics

2.1.1 The Rise of Algorithmic Decision-Making The proliferation of algorithmic decision-making systems across various domains of social life represents one of the most significant technological developments of the 21st century. These systems—ranging from simple rule-based algorithms to complex machine learning models—now influence or determine outcomes in areas such as financial lending (Fuster et al., 2022), hiring (Bogen & Rieke, 2018), criminal justice (Angwin et al., 2016), healthcare resource allocation (Obermeyer et al., 2019), and social service provision (Eubanks, 2018). The increasing reliance on these systems stems from their perceived benefits: efficiency, consistency, scale, and the potential to overcome human biases and limitations.

However, the adoption of algorithmic decision-making has been accompanied by growing concerns about fairness and discrimination. Research has demonstrated that algorithms can reproduce or even amplify existing social biases (Barocas & Selbst, 2016; Noble, 2018). These biases may emerge through various mechanisms, including biased training data, problematic feature selection, proxy discrimination, and the optimization of objectives that disadvantage certain groups (Hellman, 2020). The recognition of these issues has given rise to the field of algorithmic fairness, which seeks to develop techniques for detecting, measuring, and mitigating unfairness in algorithmic systems.

2.1.2 Formal Definitions of Fairness The field of algorithmic fairness has produced numerous formal definitions and statistical metrics to evaluate whether an algorithm treats different demographic groups equitably. These definitions generally fall into several categories, each reflecting different intuitions about what constitutes fair treatment. The most prominent categories include:

Independence-based fairness (Demographic Parity): This definition requires that the algorithm's decision be independent of protected attributes such as race or gender. Formally, if R represents the algorithm's decision (e.g., approve/deny) and A represents a protected attribute (e.g., race), demographic parity requires that P(R = r | A = a) = P(R = r | A = a') for all values r, a, and a' (Dwork et al., 2012). In other words, the probability of each outcome should be equal across all demographic groups.

Separation-based fairness (Equalized Odds): This definition requires that the algorithm's decision be independent of protected attributes, conditional on the actual outcome Y. Formally, P(R = r | A = a, Y = y) = P(R = r | A = a', Y = y) for all r, a, a', and y (Hardt et al., 2016). This means that both true positive rates and false positive rates should be equal across demographic groups.

Equal Opportunity: A relaxation of equalized odds, this definition only requires equality of true positive rates across groups: P(R = 1 | A = a, Y = 1) = P(R = 1 | A = a', Y = 1) for all a and a' (Hardt et al., 2016).

Sufficiency-based fairness (Calibration): This definition requires that the algorithm's predicted probability of an outcome match the actual probability of that outcome within each group. Formally, P(Y = 1 | R = r, A = a) = P(Y = 1 | R = r, A = a') = r for all a and a' (Kleinberg et al., 2017).

Individual Fairness: Moving beyond group-level measures, individual fairness requires that similar individuals receive similar treatment. Formally, if $d(x, x')$ represents the distance between individuals x and x' in a relevant metric space, and $d'(R(x), R(x'))$ represents the distance between the algorithm's decisions for these individuals, then $d'(R(x), R(x')) \leq d(x, x')$(Dwork et al., 2012).

These definitions represent different interpretations of fairness, often stemming from different normative intuitions about what constitutes just treatment. Importantly, research has demonstrated that many of these definitions are mathematically incompatible—it is generally impossible to satisfy all of them simultaneously, except in highly constrained or trivial scenarios (Kleinberg et al., 2017; Chouldechova, 2017). This incompatibility necessitates choices about which fairness criteria to prioritize in any given context, choices that ultimately reflect value judgments about the nature of fairness itself.

2.1.3 Limitations and Critiques of Current Approaches While formal fairness metrics have provided valuable tools for identifying and addressing certain forms of algorithmic bias, they have been subject to various critiques. These critiques can be broadly categorized as follows:

Technical Limitations: As noted above, the impossibility results of Kleinberg et al. (2017) and Chouldechova (2017) demonstrate that different formal fairness criteria cannot generally be satisfied simultaneously. This means that choosing to satisfy one criterion necessarily involves trade-offs with others. Additionally, many fairness metrics are sensitive to how the problem is framed and which variables are included in the model (Corbett-Davies & Goel, 2018).

Contextual Inadequacy: Fairness metrics often fail to account for the specific social, historical, and institutional contexts in which algorithms operate. What constitutes fair treatment may vary across different domains and social settings, making generic statistical definitions insufficient (Green, 2018; Selbst et al., 2019).

Procedural Neglect: Most fairness metrics focus on the distribution of outcomes rather than the processes by which those outcomes are determined. This neglects procedural aspects of fairness, such as transparency, explainability, and the ability to contest decisions (Grgić-Hlača et al., 2018; Binns, 2018).

Individualist vs. Group-Based Tensions: There is an inherent tension between group-based fairness metrics, which aim to ensure equality across demographic groups, and individualist approaches, which emphasize treating similar individuals similarly regardless of group membership (Binns, 2020; Dwork et al., 2012).

Normative Ambiguity: The choice of which fairness metric to use involves implicit normative assumptions about what constitutes fair treatment, yet these assumptions are often left unexamined. Different metrics may align with different theories of justice, but this connection is rarely made explicit (Fazelpour & Lipton, 2020; Binns, 2018).

These limitations highlight the need for approaches to algorithmic fairness that are philosophically grounded, contextually sensitive, and capable of addressing both procedural and distributive aspects of justice. This thesis aims to contribute to this effort by exploring how Nozick's entitlement theory might offer a distinctive perspective on algorithmic fairness—one that foregrounds procedural justice and historical processes rather than distributive patterns.

1.3.2 2.2 Theories of Distributive Justice

2.2.1 Egalitarian Theories Egalitarian theories of distributive justice are primarily concerned with the way goods, resources, opportunities, or welfare are distributed across society. These theories generally hold that justice requires some form of equality, though they differ in what exactly should be equalized. The main variants of egalitarianism include:

Strict Egalitarianism: This view holds that justice requires equality of outcomes—all individuals should receive equal shares of goods or resources. This position is rarely defended in its pure form due to various practical and theoretical objections (such as differences in need, effort, or desert), but it serves as a baseline against which other theories are often contrasted (Nielsen, 1979).

Luck Egalitarianism: This approach, associated with philosophers such as Ronald Dworkin (2000), G.A. Cohen (1989), and Richard Arneson (1989), aims to neutralize the effects of brute or unchosen luck while holding individuals responsible for their choices. Justice requires compensating individuals for disadvantages that result from circumstances beyond their control (brute luck) but allows inequalities that result from voluntary choices (option luck).

Prioritarianism: Rather than strict equality, prioritarianism holds that justice requires giving priority to improving the condition of those who are worst off. This view, associated with philosophers like Derek Parfit (1997), differs from strict egalitarianism in that it is concerned with absolute rather than relative levels of well-being.

Rawlsian Justice as Fairness: John Rawls's (1971) influential theory combines egalitarian and prioritarian elements. His difference principle permits inequalities only if they benefit the least advantaged members of society. Additionally, his principle of fair equality of opportunity requires that positions and offices be open to all under conditions where individuals with similar talents and motivation have similar prospects for success.

These egalitarian approaches share a focus on the pattern or structure of distribution—they evaluate justice based on how goods or opportunities are distributed at a given time, rather than the processes by which that distribution came about. This pattern-based approach has been influential in shaping algorithmic fairness metrics, particularly those focused on ensuring equality of outcomes or opportunities across demographic groups.

2.2.2 Libertarian Theories and Nozick's Entitlement Theory In contrast to egalitarian theories, libertarian approaches to distributive justice focus on individual rights and the processes by which distributions arise, rather than the resulting patterns. The most prominent libertarian theory of justice is Robert Nozick's entitlement theory, articulated in his seminal work *Anarchy, State, and Utopia* (1974).

Nozick's theory is fundamentally historical and procedural—it holds that a distribution is just if it arose through just processes, regardless of the resulting pattern. The theory consists of three main principles:

The Principle of Justice in Acquisition: This principle concerns the original acquisition of holdings from an unowned state of nature. Nozick draws on Locke's theory of property acquisition, which holds that one can acquire previously unowned resources by mixing one's labor with them, subject to the "Lockean proviso" that "enough and as good" is left for others (Nozick, 1974, p. 175).

The Principle of Justice in Transfer: Once resources have been justly acquired, they may be transferred to others. A transfer is just if it occurs through voluntary exchange, gift, or bequest. Coerced transfers, such as theft or fraud, violate this principle and result in unjust holdings (Nozick, 1974, p. 150-153).

The Principle of Rectification of Injustice: This principle addresses how to rectify past injustices in acquisition or transfer. If current holdings resulted from unjust processes, rectification is required to restore a just distribution. Nozick acknowledges the complexity of determining what rectification would entail in real-world scenarios with histories of injustice (Nozick, 1974, p. 152-153).

Nozick's theory explicitly rejects what he calls "patterned" theories of justice—theories that specify that a distribution is to vary along with some natural dimension, such as moral merit, need, or usefulness to society. He argues that maintaining any pattern would require continuous interference with people's ability to freely exchange what they own, thus violating their rights (Nozick, 1974, p. 160-164). His famous "Wilt Chamberlain" example illustrates how liberty upsets patterns: if we start with a just distribution (according to some pattern) and then allow people to freely exchange their resources (e.g., by paying to watch Wilt Chamberlain play basketball), the resulting distribution will no longer conform to the original pattern. Nozick argues that preventing this outcome would require "continuous interference with people's lives" (Nozick, 1974, p. 163).

Nozick's approach represents a radical departure from egalitarian theories in that it does not evaluate justice based on the resulting distribution of goods or opportunities, but rather on whether that distribution arose through processes that respect individual rights to property and free exchange. This procedural focus offers a distinctive perspective that has been largely absent from discussions of algorithmic fairness.

2.2.3 Critiques and Counterarguments Both egalitarian and libertarian theories have been subject to extensive critique. Key criticisms of egalitarian theories include concerns about the restriction of liberty, problems with determining the appropriate "currency" of equality (resources, welfare, capabilities, etc.), and questions about the moral relevance of equality itself as opposed to sufficiency or priority (Frankfurt, 1987; Parfit, 1997).

Critiques of Nozick's entitlement theory have been equally robust. Major objections include:

The Problem of Initial Acquisition: Critics argue that Nozick's account of just acquisition is underdeveloped and that the Lockean proviso may be impossible to satisfy in a world of finite resources (Cohen, 1995; Otsuka, 2003).

Historical Injustice: Given the extensive history of injustice in acquisition and transfer (including colonialism, slavery, and theft), critics argue that almost all current holdings are tainted by past injustice, rendering Nozick's theory practically inapplicable without massive redistribution (Kymlicka, 2002; Waldron, 1992).

Neglect of Need: Critics contend that Nozick's theory fails to account for basic human needs and allows for distributions that leave some individuals in desperate poverty while others have vast wealth (Cohen, 1995).

Self-Ownership Without Substance: Critics like G.A. Cohen (1995) argue that Nozick's emphasis on self-ownership is hollow if individuals lack access to external resources needed to make meaningful use of their talents and abilities.

Despite these critiques, Nozick's theory remains influential and offers valuable insights into procedural aspects of justice that are often neglected in egalitarian approaches. The tension between procedural and distributive conceptions of justice—between process and pattern—is central to many debates in political philosophy and, as this thesis argues, equally relevant to discussions of algorithmic fairness.

1.3.3 2.3 The Intersection of Algorithmic Fairness and Distributive Justice

2.3.1 Current Perspectives Recent scholarship has begun to explore the connections between algorithmic fairness and theories of distributive justice, primarily focusing on how different fairness metrics align with different philosophical conceptions of fairness (Binns, 2018; Fazelpour & Lipton, 2020; Lee & Floridi, 2021).

This work has revealed that many common fairness metrics implicitly adopt egalitarian perspectives. For example:

- Demographic parity aligns with a strict egalitarian view that different groups should receive equal outcomes regardless of other factors.
- Equal opportunity metrics reflect Rawlsian concerns with fair equality of opportunity.
- Calibration and predictive parity metrics can be linked to meritocratic conceptions of desert-based justice.

These connections have helped clarify the normative assumptions underlying different technical approaches to fairness and have highlighted the need for contextual judgment in selecting appropriate metrics for specific applications (Lee & Floridi, 2021).

However, as noted in the introduction, libertarian perspectives—particularly Nozick's entitlement theory—have been largely absent from these discussions. This absence is notable given the significant influence of libertarian thinking in debates about technology, markets, and regulation more broadly (Thierer, 2016).

2.3.2 The Gap: Libertarian Perspectives on Algorithmic Fairness The omission of libertarian perspectives from discussions of algorithmic fairness represents a significant gap in the literature. This gap may stem from several factors:

1. The apparent incompatibility between libertarian emphasis on individual rights and process versus the focus on group-level statistics in most fairness metrics.

2. The challenge of translating historical and procedural conceptions of justice into formal, quantifiable measures applicable to algorithms.

3. The dominant egalitarian orientation of much research on discrimination and bias, reflecting broader trends in social justice scholarship.

4. The difficulty of applying Nozick's principles to algorithmic contexts where the concepts of "acquisition" and "transfer" do not have obvious analogues.

Despite these challenges, a Nozickian perspective on algorithmic fairness offers potential insights that complement existing approaches. By focusing on the processes by which algorithmic decisions are made, rather than solely on the resulting distributions, such a perspective might address some of the procedural concerns that have been raised about current fairness metrics (Grgić-Hlača et al., 2018; Binns, 2018). Additionally, a libertarian approach might help navigate tensions between group fairness and individual treatment that have proven challenging for egalitarian metrics (Binns, 2020).

In the following section, I develop a framework for algorithmic fairness grounded in Nozick's entitlement theory, addressing the challenges of translating his historical and procedural conception of justice into the context of algorithmic decision-making.

1.4 3. A Nozickian Framework for Algorithmic Fairness

1.4.1 3.1 Conceptual Foundations

3.1.1 Core Elements of Nozick's Theory Relevant to Algorithms To develop a Nozickian framework for algorithmic fairness, we must first identify which elements of Nozick's entitlement theory are most relevant to algorithmic decision-making contexts. While the direct translation of concepts like "acquisition" and "transfer" from physical property to algorithmic decisions is not straightforward, several core aspects of Nozick's theory provide valuable starting points:

Procedural Justice: For Nozick, justice is fundamentally about processes rather than patterns. A just distribution is one that arises through just processes, regardless of the resulting pattern. This focus on procedural justice can be applied to algorithmic contexts by emphasizing the procedures through which algorithms make decisions, rather than solely the distribution of outcomes across groups.

Historical Dimension: Nozick's theory is historical—the justice of current holdings depends on the history of how they came to be. In algorithmic contexts, this suggests attention to the historical processes that generate the data used by algorithms and the historical context in which algorithmic systems operate.

Rights and Consent: Central to Nozick's theory is respect for individual rights, particularly property rights and the right to engage in voluntary exchanges. In algorithmic contexts, this might translate to concerns about consent, data ownership, and individuals' rights regarding how their information is used in decision-making processes.

Non-Interference: Nozick argues against "continuous interference" to maintain patterns of distribution. In algorithmic contexts, this might suggest skepticism toward extensive interventions in algorithm design solely to achieve particular distributive outcomes, especially if these interventions might violate individual rights or distort market processes.

Rectification: Nozick acknowledges the need to rectify past injustices in acquisition and transfer. In algorithmic contexts, this principle might apply to addressing historical biases in data or correcting for past discriminatory practices that may be perpetuated by algorithms.

These elements provide a foundation for thinking about algorithmic fairness from a Nozickian perspective. However, translating them into concrete principles for algorithmic design and evaluation requires addressing several conceptual challenges.

3.1.2 Translating Nozickian Concepts to Algorithmic Contexts The translation of Nozick's principles to algorithmic contexts requires addressing several key questions:

What constitutes "just acquisition" in algorithmic decision-making? In Nozick's theory, just acquisition concerns how unowned resources come to be legitimately owned. In algorithmic contexts, this might relate to how data is collected and how algorithmic models are developed. Just acquisition could involve obtaining data with proper consent, ensuring transparency about how data will be used, and developing algorithms in ways that respect the rights and autonomy of affected individuals.

What constitutes "just transfer" in algorithmic decisions? For Nozick, just transfer involves voluntary exchanges free from coercion or fraud. In algorithmic contexts, this might relate to how decisions are made and communicated. Just transfer could involve ensuring that algorithmic decisions are transparent, that individuals understand the basis for decisions affecting them, and that they have meaningful opportunities to contest or appeal decisions they believe are unjust.

How does the principle of rectification apply to algorithms? Nozick's principle of rectification addresses how to correct past injustices in acquisition or transfer. In algorithmic contexts, this might involve correcting for historical biases in data, addressing past discriminatory practices that may be perpetuated by algorithms, or providing remedies for individuals who have been harmed by unjust algorithmic decisions.

What rights do individuals have in relation to algorithmic systems? Nozick's theory is fundamentally concerned with individual rights, particularly property rights. In algorithmic contexts, relevant rights might include data ownership rights, rights to consent to how one's data is used, rights to explanation or contestation of algorithmic decisions, and rights to compensation for harms caused by algorithmic systems.

How do we balance procedural justice with concerns about discriminatory outcomes? While Nozick's theory focuses on procedural justice rather than distributive patterns, there may be cases where seemingly just procedures lead to outcomes that appear discriminatory. Addressing this tension requires careful consideration of how procedural and distributive concerns can be balanced within a broadly libertarian framework.

Addressing these questions will help us develop concrete principles for evaluating algorithmic fairness from a Nozickian perspective. In the following sections, I propose such principles and develop formal metrics based on them.

1.4.2 3.2 Principles of Nozickian Algorithmic Fairness

Building on the conceptual foundations outlined above, I propose the following principles for evaluating algorithmic fairness from a Nozickian perspective:

3.2.1 Principle of Just Data Acquisition An algorithm satisfies the principle of just data acquisition if:

1. All data used to train and validate the algorithm was collected with the informed consent of the individuals to whom the data pertains, or from legitimately public sources.
2. Individuals retained meaningful control over their data, including rights to access, correct, delete, or restrict the use of their personal information.
3. The collection and use of data respected relevant privacy rights and did not involve deception or coercion.
4. The Lockean proviso is satisfied: the data collection did not deprive others of essential information resources or create significant informational asymmetries that undermine individual autonomy.

This principle translates Nozick's concern with just initial acquisition to the context of data collection and algorithm development. It emphasizes consent, control, and respect for individual rights over personal information.

3.2.2 Principle of Just Algorithmic Processing An algorithm satisfies the principle of just algorithmic processing if:

1. The algorithm makes decisions based on factors that individuals have meaningfully influenced through their voluntary choices and actions, rather than immutable characteristics or circumstances beyond their control.
2. The algorithm's decision-making process is transparent and intelligible to affected individuals, allowing them to understand how their actions and choices influence decisions about them.
3. The algorithm does not violate individuals' rights or entitlements established through just processes (e.g., legal contracts, earned credentials, legitimate expectations based on past performance).
4. The algorithm's operation does not involve fraud, deception, or other forms of involuntary transfer (e.g., making decisions based on criteria different from those publicly claimed).

This principle applies Nozick's focus on voluntary transfers and respect for established entitlements to the context of algorithmic decision-making. It emphasizes transparency, agency, and respect for rightfully acquired claims.

3.2.3 Principle of Algorithmic Rectification An algorithm satisfies the principle of algorithmic rectification if:

1. It includes mechanisms to identify and correct for past injustices that may be perpetuated through algorithmic decisions, particularly those involving violations of just acquisition or transfer.
2. It provides meaningful opportunities for individuals to contest decisions they believe are unjust and to seek appropriate remedies.
3. It includes processes for updating and improving the algorithm in response to identified instances of injustice or rights violations.

4. When historical data reflects past discriminatory practices or unjust social arrangements, the algorithm incorporates appropriate adjustments to prevent the perpetuation of these injustices.

This principle applies Nozick's principle of rectification to algorithmic contexts, acknowledging that historical injustices may require correction to achieve a just state of affairs. It recognizes that while procedural justice is paramount, there may be cases where past procedural violations necessitate interventions to restore justice.

3.2.4 Meta-Principle: Minimal Interference with Voluntary Exchanges In addition to the three substantive principles outlined above, a Nozickian approach to algorithmic fairness would include a meta-principle regarding the role of regulation and intervention:

Interventions in algorithmic design and operation should be limited to those necessary to ensure just acquisition, processing, and rectification, and should minimize interference with voluntary exchanges and individual rights. This principle reflects Nozick's concern with "continuous interference" to maintain particular distributive patterns.

This meta-principle does not preclude all regulation or intervention—Nozick himself acknowledges the legitimacy of state action to protect rights and enforce just acquisition and transfer. However, it suggests skepticism toward extensive interventions solely to achieve particular distributional outcomes, especially if these interventions might infringe on individual rights or distort market processes.

1.4.3 3.3 Formalizing Nozickian Fairness Metrics

Translating the principles outlined above into formal metrics that can be applied to algorithmic systems presents a significant challenge. Unlike traditional fairness metrics, which typically focus on statistical properties of an algorithm's outputs, Nozickian metrics must capture procedural aspects of justice that are not easily quantified. Nevertheless, I propose the following formal approaches to measuring algorithmic fairness from a Nozickian perspective:

3.3.1 Procedural Entitlement Fairness (PEF) I propose a metric called Procedural Entitlement Fairness (PEF) that evaluates the extent to which an algorithm's decisions respect individuals' rightfully acquired entitlements. Let's define:

- E_i = the set of rightfully acquired entitlements of individual i
- D_i = the decision made by the algorithm regarding individual i
- $R(E_i, D_i)$ = a function that measures the extent to which decision D_i respects the entitlements E_i

Then, Procedural Entitlement Fairness (PEF) for a population of n individuals can be defined as:

$$PEF = \frac{1}{n} \sum_{i=1}^{n} R(E_i, D_i)$$

Where $R(E_i, D_i)$ is normalized to a value between 0 and 1, with 1 indicating full respect for entitlements and 0 indicating complete violation.

The challenge in implementing this metric lies in defining E_i and $R(E_i, D_i)$ for specific contexts. In a lending context, for example, E_i might include an individual's credit history, repayment record, and legitimate expectations based on past financial behavior, while $R(E_i, D_i)$ might measure how well the lending decision aligns with these established entitlements.

3.3.2 Consent and Transparency Index (CTI) The Consent and Transparency Index (CTI) measures the extent to which an algorithmic system respects principles of just acquisition and transparent processing. Let's define:

- C_i = a binary indicator of whether individual i's data was collected with proper consent (1) or not (0)
- T_i = a measure of the transparency of the algorithm's decision process for individual i, normalized to a value between 0 and 1
- w_c and w_t = weights assigned to the consent and transparency components, respectively, with $w_c + w_t = 1$

Then, the Consent and Transparency Index (CTI) can be defined as:

$$CTI = \frac{1}{n}\sum_{i=1}^{n}(w_c \cdot C_i + w_t \cdot T_i)$$

This metric captures key aspects of just acquisition (consent) and just processing (transparency), reflecting the procedural focus of Nozick's theory.

3.3.3 Rectification Responsiveness Measure (RRM) The Rectification Responsiveness Measure (RRM) evaluates an algorithm's capacity to identify and correct for past injustices that may be perpetuated through its decisions. Let's define:

- I = the set of identified instances where the algorithm's decisions may perpetuate past injustices
- C_j = a measure of the adequacy of the correction implemented for instance $j \in I$, normalized to a value between 0 and 1
- A = a measure of the algorithm's accessibility to contestation and appeal, normalized to a value between 0 and 1
- w_c and w_a = weights assigned to the correction and accessibility components, respectively, with $w_c + w_a = 1$

Then, the Rectification Responsiveness Measure (RRM) can be defined as:
$RRM = w_c \cdot \left(\frac{1}{|I|}\sum_{j\in I} C_j\right) + w_a \cdot A$
This metric captures the algorithm's ability to identify and correct for past injustices, as well as its accessibility to contestation by affected individuals, reflecting Nozick's principle of rectification.

3.3.4 Choice Sensitivity Ratio (CSR) The Choice Sensitivity Ratio (CSR) measures the extent to which an algorithm's decisions are sensitive to factors that individuals can influence through their voluntary choices, as opposed to immutable characteristics or circumstances beyond their control. Let's define:

- F_v = the set of features used by the algorithm that reflect voluntary choices
- F_i = the set of features used by the algorithm that reflect immutable characteristics or circumstances beyond individual control
- I_v = the importance of voluntary features in the algorithm's decisions, measured by their collective influence on the algorithm's output
- I_i = the importance of immutable features in the algorithm's decisions, measured similarly

Then, the Choice Sensitivity Ratio (CSR) can be defined as:
$CSR = \frac{I_v}{I_v + I_i}$
This ratio ranges from 0 to 1, with higher values indicating greater sensitivity to voluntary choices, which aligns with Nozick's emphasis on individual agency and responsibility.

1.4.4 3.4 Integrated Nozickian Fairness Framework

While the individual metrics proposed above capture specific aspects of Nozickian fairness, an integrated framework is needed to evaluate algorithmic systems comprehensively. I propose an integrated Nozickian Fairness Score (NFS) that combines the four metrics:

$NFS = w_{PEF} \cdot PEF + w_{CTI} \cdot CTI + w_{RRM} \cdot RRM + w_{CSR} \cdot CSR$

Where w_{PEF}, w_{CTI}, w_{RRM}, and w_{CSR} are weights assigned to each component, with $w_{PEF} + w_{CTI} + w_{RRM} + w_{CSR} = 1$.

The weights can be adjusted based on the specific context and the relative importance of different aspects of Nozickian fairness in that context. For example, in contexts where historical injustices are particularly salient, greater weight might be given to RRM, while in contexts where individual agency is paramount, CSR might receive greater weight.

This integrated framework allows for a comprehensive evaluation of algorithmic systems from a Nozickian perspective, focusing on procedural justice rather than distributive patterns. Importantly, it does not preclude the possibility of unequal outcomes across demographic groups—if these inequalities arise through just processes, they are not considered unjust from a Nozickian perspective.

3.4.1 Comparison with Traditional Fairness Metrics The Nozickian fairness framework differs significantly from traditional fairness metrics in several key respects:

Focus on Process vs. Outcome: Traditional metrics like demographic parity and equalized odds focus on the distribution of outcomes across demographic groups. In contrast, the Nozickian framework focuses on the processes by which decisions are made, emphasizing consent, transparency, respect for entitlements, and sensitivity to voluntary choices.

Individual vs. Group Level: Traditional metrics typically operate at the group level, comparing outcomes across demographic categories. The Nozickian framework primarily operates at the individual level, evaluating how algorithmic decisions respect each individual's rights and entitlements.

Historical vs. Snapshot View: Traditional metrics typically take a snapshot view, evaluating the current distribution of outcomes without reference to how that distribution came about. The Nozickian framework takes a historical view, considering how data was acquired and how past injustices might require rectification.

Acceptance of Unequal Outcomes: Traditional metrics often aim for some form of equality or parity in outcomes across groups. The Nozickian framework accepts potentially unequal outcomes if they arise through just processes of acquisition, transfer, and rectification.

These differences illustrate the distinct perspective that a Nozickian approach brings to discussions of algorithmic fairness. Rather than asking whether an algorithm produces equal outcomes across groups, it asks whether the algorithm respects individual rights, operates transparently, and provides appropriate mechanisms for rectification when needed.

1.5 4. Application to Case Studies

To illustrate how the Nozickian framework for algorithmic fairness might be applied in practice, I examine two hypothetical case studies: algorithmic lending decisions and automated hiring systems. These applications demonstrate the framework's practical utility while highlighting how it differs from traditional fairness approaches.

1.5.1 4.1 Case Study 1: Algorithmic Lending Decisions

4.1.1 Context Consider a financial institution that uses an algorithmic system to evaluate loan applications. The system uses various data points—including credit history, income, employment stability, and debt-to-income ratio—to predict an applicant's likelihood of repaying a loan. Based on this prediction, the algorithm either approves or denies the loan application, or offers different interest rates to different applicants.

Traditional fairness metrics might focus on whether the algorithm's decisions result in equal approval rates or similar interest rates across demographic groups such as race or gender. From an egalitarian perspective, substantial disparities in outcomes might be considered evidence of unfairness, even if the algorithm accurately predicts repayment likelihood.

Let us examine this scenario through the lens of Nozickian fairness, applying the metrics developed in Section 3.

4.1.2 Just Data Acquisition To evaluate the principle of just data acquisition, we would consider:

- Was the data used to train and validate the algorithm collected with informed consent?
- Do individuals retain control over their financial data, with rights to access, correct, and restrict its use?
- Does the data collection respect privacy rights and avoid deception or coercion?
- Does the data collection create significant informational asymmetries that undermine individual autonomy?

If the algorithm uses credit bureau data, we would examine whether individuals provided informed consent for this use and whether they have meaningful control over their credit information. If the algorithm scrapes social media or other personal data without explicit consent, this would violate the principle of just acquisition.

Applying the Consent and Transparency Index (CTI), we might find that while traditional credit data is collected with some form of consent (though perhaps not fully informed consent), the use of alternative data sources might lack proper consent, resulting in a lower CTI score.

4.1.3 Just Algorithmic Processing To evaluate the principle of just algorithmic processing, we would consider:

- Does the algorithm base decisions primarily on factors that individuals have meaningfully influenced through their voluntary choices (e.g., payment history, debt management) rather than immutable characteristics or circumstances beyond their control (e.g., race, family background)?
- Is the algorithm's decision-making process transparent and intelligible to loan applicants?
- Does the algorithm respect legitimate expectations based on an individual's credit history and financial behavior?
- Does the algorithm operate as described, without hidden factors or deceptive practices?

Applying the Procedural Entitlement Fairness (PEF) metric, we would assess whether individuals with similar credit histories and financial behaviors receive similar treatment, regardless of demographic characteristics. The Choice Sensitivity Ratio (CSR) would measure the extent to which loan decisions are based on voluntary financial behaviors rather than immutable characteristics or circumstances beyond individual control.

4.1.4 Algorithmic Rectification To evaluate the principle of algorithmic rectification, we would consider:

- Does the algorithm account for historical discrimination in lending that may be reflected in credit histories?
- Are there meaningful opportunities for individuals to contest loan denials or unfavorable terms?
- Does the algorithm improve over time based on identified instances of unjust decisions?
- Are there mechanisms to address cases where historical data reflects past discriminatory practices?

Applying the Rectification Responsiveness Measure (RRM), we would assess whether the lending algorithm includes mechanisms to identify and correct for past discriminatory lending practices that might be perpetuated through algorithmic decisions.

4.1.5 Integrated Evaluation Combining these assessments into the integrated Nozickian Fairness Score (NFS), we might find that a lending algorithm scores highly on some dimensions (e.g., basing decisions on voluntary financial behaviors) but poorly on others (e.g., limited transparency or weak rectification mechanisms).

Importantly, from a Nozickian perspective, disparities in loan approval rates across demographic groups would not necessarily indicate unfairness if these disparities arose through just processes. If individuals from different groups have different credit histories due to their voluntary financial choices, and the algorithm bases decisions on these histories in a transparent and consistent manner, the resulting disparities would not be considered unjust.

However, if disparities arise from historical injustices in acquisition or transfer (e.g., past discriminatory lending practices that affected credit histories), the principle of rectification would require appropriate adjustments to prevent the perpetuation of these injustices. This might involve specific interventions to account for the effects of past discrimination while still respecting individual financial behavior.

4.1.6 Comparison with Egalitarian Approaches The Nozickian approach to evaluating the lending algorithm differs markedly from traditional egalitarian approaches. While an egalitarian approach might focus on achieving similar approval rates or loan terms across demographic groups, the Nozickian approach focuses on the processes by which lending decisions are made.

This difference is particularly evident in cases where disparities in outcomes result from differences in credit histories that reflect voluntary financial choices. An egalitarian approach might view such disparities as problematic and advocate for interventions to achieve more equal outcomes. A Nozickian approach would accept these disparities as just if they arose through processes that respected principles of just acquisition, transfer, and rectification.

However, in cases where disparities reflect historical injustices rather than voluntary choices, both approaches might support interventions, though for different reasons. An egalitarian approach would seek to reduce disparities for their own sake, while a Nozickian approach would seek to rectify past violations of just acquisition or transfer.

This comparison highlights both the distinctions and potential points of convergence between egalitarian and libertarian approaches to algorithmic fairness. While they differ in their fundamental orientations—pattern versus process—they may sometimes support similar interventions, particularly in cases involving historical injustice.

1.5.2 4.2 Case Study 2: Automated Hiring Systems

4.2.1 Context Consider a company that uses an automated system to screen job applicants. The system analyzes resumes, cover letters, and possibly video interviews to predict which candidates are likely to succeed in the role. Based on these predictions, the system either advances candidates to the next stage of the hiring process or rejects their applications.

Traditional fairness metrics might focus on whether the algorithm's decisions result in similar selection rates across demographic groups, with disparities potentially viewed as evidence of discrimination. Let us examine this scenario through the lens of Nozickian fairness.

4.2.2 Just Data Acquisition To evaluate the principle of just data acquisition, we would consider:

- Was the training data (e.g., past hiring decisions, employee performance reviews) collected with informed consent from the individuals involved?
- Do job applicants consent to having their applications processed by an automated system?
- Is the data collection process transparent about how applicant information will be used?
- Does the collection and use of applicant data respect privacy rights and avoid deception?

Applying the Consent and Transparency Index (CTI), we would assess the extent to which the system operates with proper consent and transparency. If the system uses social media data or other personal information without explicit consent, this would lower the CTI score.

4.2.3 Just Algorithmic Processing To evaluate the principle of just algorithmic processing, we would consider:

- Does the algorithm base decisions primarily on factors that reflect candidates' voluntary choices and actions (e.g., education, skills development, work experience) rather than immutable characteristics or circumstances beyond their control (e.g., race, gender, socioeconomic background)?

- Is the algorithm's decision-making process transparent and intelligible to job applicants?
- Does the algorithm respect legitimate qualifications and credentials that candidates have rightfully acquired?
- Does the algorithm operate as described, without hidden factors or deceptive practices?

Applying the Choice Sensitivity Ratio (CSR), we would measure the extent to which the hiring algorithm's decisions are sensitive to factors that reflect candidates' voluntary choices rather than immutable characteristics. The Procedural Entitlement Fairness (PEF) would assess whether candidates with similar qualifications receive similar treatment, regardless of demographic characteristics.

4.2.4 Algorithmic Rectification To evaluate the principle of algorithmic rectification, we would consider:

- Does the algorithm account for historical discrimination in hiring that may be reflected in the training data?
- Are there meaningful opportunities for candidates to contest rejections they believe are unjust?
- Does the system improve over time based on identified instances of unfair decisions?
- Are there mechanisms to address cases where historical hiring data reflects past discriminatory practices?

Applying the Rectification Responsiveness Measure (RRM), we would assess whether the hiring algorithm includes mechanisms to identify and correct for past discriminatory hiring practices that might be perpetuated through algorithmic decisions.

4.2.5 Integrated Evaluation Combining these assessments into the integrated Nozickian Fairness Score (NFS), we might find that a hiring algorithm scores highly on some dimensions (e.g., respecting candidates' rightfully acquired credentials) but poorly on others (e.g., limited mechanisms for contestation or rectification).

From a Nozickian perspective, disparities in selection rates across demographic groups would not necessarily indicate unfairness if these disparities arose through just processes. If individuals from different groups have different qualifications due to their voluntary choices, and the algorithm bases decisions on these qualifications in a transparent and consistent manner, the resulting disparities would not be considered unjust.

However, if disparities arise from historical injustices that affected educational or career opportunities, the principle of rectification would require appropriate adjustments. This might involve specific interventions to account for the effects of past discrimination while still respecting individual qualifications and achievements.

4.2.6 Comparison with Egalitarian Approaches The Nozickian approach to evaluating the hiring algorithm differs from traditional egalitarian approaches in its focus on process rather than outcome patterns. While an egalitarian approach might focus on achieving similar selection rates across demographic groups, the Nozickian approach focuses on the processes by which hiring decisions are made.

This difference is particularly evident in cases where disparities in outcomes result from differences in qualifications that reflect voluntary educational and career choices. An egalitarian approach might view such disparities as problematic and advocate for interventions to achieve more equal outcomes. A Nozickian approach would accept these disparities as just if they arose through processes that respected principles of just acquisition, transfer, and rectification.

However, in cases where disparities reflect historical injustices rather than voluntary choices, both approaches might support interventions, though for different reasons. An egalitarian approach would seek to reduce disparities for their own sake, while a Nozickian approach would seek to rectify past violations of just acquisition or transfer.

This comparison highlights how different philosophical perspectives lead to different evaluations of algorithmic fairness, with potentially significant implications for how automated systems are designed, deployed, and regulated.

1.6 5. Discussion and Implications

1.6.1 5.1 Philosophical Implications

The Nozickian framework for algorithmic fairness developed in this thesis has several important philosophical implications for how we conceptualize and evaluate fairness in automated decision-making systems.

5.1.1 Pluralism in Conceptions of Algorithmic Fairness The development of a libertarian approach to algorithmic fairness alongside existing egalitarian approaches highlights the inherent pluralism in conceptions of fairness. Different philosophical traditions offer distinct lenses through which to evaluate algorithmic systems, with different normative commitments and priorities. This pluralism suggests that there is no single, universally applicable definition of "fairness" that can be encoded into algorithms.

Instead, the choice of fairness metrics and frameworks should be recognized as inherently value-laden, reflecting particular philosophical commitments about what constitutes just treatment. This recognition calls for greater transparency about the normative assumptions underlying different approaches to algorithmic fairness and for more explicit engagement with the philosophical foundations of these approaches.

5.1.2 Procedural vs. Distributive Justice in Algorithmic Contexts The Nozickian framework foregrounds procedural aspects of justice that are often neglected in traditional fairness metrics. While most existing metrics focus on the distribution of outcomes across demographic groups, the Nozickian approach emphasizes the processes by which these outcomes are determined, including issues of consent, transparency, respect for entitlements, and rectification of past injustices.

This procedural focus offers a valuable complement to distributive approaches, highlighting aspects of algorithmic systems that may be morally significant regardless of their distributional effects. It suggests that even algorithms that produce equal outcomes across groups may be procedurally unjust if they violate principles of consent, transparency, or respect for rightfully acquired entitlements.

5.1.3 The Role of History in Algorithmic Justice Nozick's entitlement theory is fundamentally historical, evaluating the justice of current holdings based on how they came to be. Similarly, the Nozickian framework for algorithmic fairness emphasizes the historical dimension of justice, considering how data was acquired, how past injustices might affect current algorithmic decisions, and what rectification might be required.

This historical perspective contrasts with the ahistorical approach of most traditional fairness metrics, which evaluate the current distribution of outcomes without reference to how that distribution came about. The Nozickian framework suggests that historical context matters for algorithmic fairness—that we cannot evaluate the justice of algorithmic decisions solely by looking at their immediate effects without considering the historical processes that shaped the data and social contexts in which these algorithms operate.

5.1.4 Individual Rights vs. Group Fairness The Nozickian framework prioritizes individual rights and entitlements over group-level patterns of distribution. This individualist orientation contrasts with the group-based focus of most traditional fairness metrics, which compare outcomes across demographic categories. The framework suggests that focusing exclusively on group-level statistics may obscure important aspects of justice at the individual level.

This tension between individual and group perspectives on justice reflects broader debates in political philosophy about the proper subjects of justice. The Nozickian framework suggests that individuals, rather than groups, should be the primary subjects of justice in algorithmic contexts, while acknowledging that group-level patterns may be relevant evidence for identifying potential violations of individual rights.

1.6.2 5.2 Practical Implications

Beyond its philosophical significance, the Nozickian framework for algorithmic fairness has several practical implications for how algorithmic systems are designed, deployed, and regulated.

5.2.1 Implications for Algorithm Design The Nozickian framework suggests several principles that should guide the design of algorithmic systems:

Transparency and Explainability: Algorithms should be designed to be as transparent and explainable as possible, allowing individuals to understand how decisions affecting them are made and how their actions and choices influence these decisions.

Consent-Based Data Practices: The collection and use of data for algorithmic systems should be based on informed consent, with individuals retaining meaningful control over their personal information.

Respect for Entitlements: Algorithms should respect individuals' rightfully acquired entitlements, such as credentials, qualifications, or legitimate expectations based on past performance.

Contestability and Appeals: Algorithmic systems should include mechanisms for individuals to contest decisions they believe are unjust and to seek appropriate remedies.

Historical Awareness: Algorithm designers should be aware of historical injustices that may be reflected in training data and should incorporate appropriate mechanisms to prevent the perpetuation of these injustices.

These design principles differ from those that might be derived from egalitarian approaches to fairness, which might focus more on achieving particular distributional patterns across demographic groups. The Nozickian principles emphasize procedural aspects of algorithmic systems rather than their distributional effects.

5.2.2 Implications for Regulation and Policy The Nozickian framework also has implications for how algorithmic systems are regulated and governed:

Procedural Requirements: Regulations might focus on procedural requirements such as transparency, explainability, and contestability rather than mandating particular distributional outcomes.

Minimal Interference: Consistent with Nozick's meta-principle of minimal interference, regulations should be limited to those necessary to protect individual rights and ensure just processes, avoiding extensive interventions solely to achieve particular distributional patterns.

Rights-Based Approach: Regulations might adopt a rights-based approach, focusing on protecting individuals' rights to consent, explanation, contestation, and rectification rather than imposing specific fairness metrics.

Context-Sensitivity: Different contexts may call for different regulatory approaches, depending on the specific rights and entitlements at stake and the historical background against which algorithmic systems operate.

These regulatory implications contrast with approaches that might mandate specific distributive outcomes, such as requiring equal approval rates across demographic groups. The Nozickian approach suggests a more procedural and rights-based orientation to algorithmic governance.

5.2.3 Complementarity with Egalitarian Approaches Despite the differences between Nozickian and egalitarian approaches to algorithmic fairness, they need not be seen as mutually exclusive. In many contexts, procedural and distributive concerns may complement each other, with procedural requirements supporting more equitable distributions and distributional patterns serving as evidence of procedural fairness or unfairness.

For example, in contexts with histories of discrimination, significant disparities in algorithmic outcomes across demographic groups might serve as prima facie evidence of procedural unfairness, prompting investigation into whether principles of just acquisition, transfer, or rectification have been violated. Conversely, ensuring procedural fairness through transparency, consent, and respect for entitlements might naturally lead to more equitable distributions of outcomes.

This complementarity suggests that a comprehensive approach to algorithmic fairness might draw on both Nozickian and egalitarian perspectives, attending to both procedural and distributive aspects of justice. Such a pluralistic approach would recognize the complex and multifaceted nature of fairness and the need for multiple lenses through which to evaluate algorithmic systems.

1.6.3 5.3 Limitations and Challenges

While the Nozickian framework offers valuable insights for algorithmic fairness, it also faces several limitations and challenges that must be acknowledged.

5.3.1 Practical Implementation Challenges Several practical challenges arise when attempting to implement the Nozickian metrics proposed in this thesis:

Defining Entitlements: The Procedural Entitlement Fairness (PEF) metric requires defining what constitutes a "rightfully acquired entitlement" in specific contexts, which may be complex and contestable.

Measuring Voluntary Choice: The Choice Sensitivity Ratio (CSR) requires distinguishing between factors that reflect voluntary choices and those that reflect immutable characteristics or circumstances beyond individual control, which is often difficult in practice.

Identifying Past Injustices: The Rectification Responsiveness Measure (RRM) requires identifying instances where algorithmic decisions may perpetuate past injustices, which presupposes agreement about what constitutes historical injustice.

Balancing Components: The integrated Nozickian Fairness Score (NFS) requires assigning weights to different components, which involves value judgments about their relative importance.

These challenges highlight the need for context-specific implementation of the Nozickian framework, with careful attention to the particular rights, entitlements, and historical injustices relevant to each domain.

5.3.2 Theoretical Limitations Beyond practical challenges, the Nozickian framework also faces several theoretical limitations:

Initial Acquisition Problem: As with Nozick's original theory, the framework faces challenges regarding what constitutes just initial acquisition, particularly in the context of data that may have been collected under conditions of unequal power or information asymmetry.

Historical Complexity: The historical dimension of the framework requires addressing complex questions about past injustices and appropriate rectification, which may be difficult to resolve definitively.

Tension with Structural Perspectives: The framework's focus on individual rights and procedural justice may not adequately address structural forms of injustice that operate without clear violations of individual rights.

Market Assumptions: Like Nozick's theory, the framework may rely on assumptions about the fairness of market processes that are contested by critics who point to market failures, power imbalances, and structural inequalities.

These theoretical limitations suggest the need for ongoing philosophical reflection on the foundations of the Nozickian approach and its application to algorithmic contexts.

5.3.3 Contextual Limitations Finally, the Nozickian framework may be more applicable in some contexts than others:

Market Contexts: The framework may be most applicable in market contexts where voluntary exchanges and property rights are central, such as lending or hiring.

Public Services: The framework may be less applicable in contexts involving public services or goods to which individuals may have claims based on need or citizenship rather than on entitlements acquired through market exchanges.

Fundamental Rights: In contexts involving fundamental rights or basic needs, egalitarian or sufficientarian approaches might take precedence over the procedural focus of the Nozickian framework.

These contextual limitations highlight the importance of philosophical pluralism in approaching algorithmic fairness, with different frameworks being more or less appropriate depending on the specific context and the values at stake.

1.7 6. Conclusion

1.7.1 6.1 Summary of Key Contributions

This thesis has developed a framework for algorithmic fairness based on Robert Nozick's entitlement theory of justice, offering a libertarian perspective that complements the predominantly egalitarian approaches in the existing literature. The key contributions of this work include:

1. **Conceptual Framework**: The thesis has articulated a conceptual framework for understanding algorithmic fairness from a Nozickian perspective, translating the principles of just acquisition, transfer, and rectification into the context of algorithmic decision-making.
2. **Formal Metrics**: The thesis has proposed formal metrics for evaluating algorithmic fairness according to Nozickian principles, including Procedural Entitlement Fairness (PEF), Consent and Transparency Index (CTI), Rectification Responsiveness Measure (RRM), and Choice Sensitivity Ratio (CSR).
3. **Integrated Approach**: The thesis has developed an integrated Nozickian Fairness Score (NFS) that combines these metrics into a comprehensive evaluation framework, allowing for context-specific weighting of different components.
4. **Case Studies**: The thesis has applied the Nozickian framework to hypothetical case studies in algorithmic lending and hiring, demonstrating its practical utility and highlighting contrasts with traditional egalitarian approaches.
5. **Philosophical Analysis**: The thesis has examined the philosophical implications of adopting a Nozickian perspective on algorithmic fairness, including implications for conceptions of procedural vs. distributive justice, individual rights vs. group fairness, and the role of history in evaluating algorithmic systems.

These contributions expand the philosophical foundations of algorithmic fairness research, offering a novel perspective that foregrounds procedural justice, individual rights, and historical context.

1.7.2 6.2 Broader Implications

The Nozickian framework developed in this thesis has several broader implications for the field of algorithmic fairness and ethics:

Philosophical Pluralism: By introducing a libertarian perspective into a field dominated by egalitarian approaches, this thesis highlights the importance of philosophical pluralism in algorithmic ethics. Different philosophical traditions offer distinct insights and values that can enrich our understanding of what constitutes fair algorithmic treatment.

Balancing Process and Pattern: The framework suggests the need to balance concerns about procedural justice (how decisions are made) with concerns about distributive patterns (what outcomes result). Both aspects are morally significant, and a comprehensive approach to algorithmic fairness should attend to both.

Historical Context: The framework emphasizes the importance of historical context in evaluating algorithmic systems, suggesting that we cannot assess fairness without considering how data was acquired, how past injustices might affect current decisions, and what rectification might be required.

Individual and Group Perspectives: The framework highlights tensions between individual and group perspectives on justice, suggesting that both have important roles in evaluating algorithmic systems but that they may sometimes point in different directions.

These broader implications suggest that algorithmic fairness research should embrace greater philosophical diversity, contextual sensitivity, and normative clarity about the values and principles that inform different approaches to fairness.

1.7.3 6.3 Future Research Directions

This thesis points to several promising directions for future research:

Empirical Testing: Future work could empirically test the proposed Nozickian metrics on real-world algorithmic systems, exploring their practical utility and limitations.

Cross-Philosophical Integration: Future research could explore how Nozickian and egalitarian approaches to algorithmic fairness might be integrated or balanced in specific contexts, drawing on the strengths of each perspective.

Contextual Refinement: The Nozickian framework could be refined and adapted for specific domains (e.g., healthcare, education, criminal justice), attending to the particular rights, entitlements, and historical injustices relevant to each context.

Critique and Response: Future work could engage more deeply with critiques of Nozick's theory and explore how the framework might be modified to address these critiques while maintaining its core insights.

Regulatory Applications: Research could examine how the Nozickian framework might inform regulatory approaches to algorithmic systems, balancing concerns about individual rights and procedural justice with other social values.

These research directions would build on the foundation laid in this thesis, further exploring the potential contributions of libertarian political philosophy to algorithmic ethics and fairness.

1.7.4 6.4 Concluding Reflections

The increasing role of algorithms in social decision-making raises profound questions about justice, fairness, and the proper relationship between individuals and automated systems. As we navigate these questions, we need philosophical frameworks that can help us articulate and balance the diverse values at stake.

This thesis has argued that Nozick's entitlement theory, despite its relative absence from current discussions of algorithmic fairness, offers valuable insights for this endeavor. By focusing on procedural justice, individual rights, and historical context, a Nozickian approach complements existing egalitarian perspectives, enriching our understanding of what constitutes fair algorithmic treatment.

As we continue to develop and deploy algorithmic systems that affect human lives, we should draw on the full range of philosophical traditions to guide our ethical evaluations and technical innovations. By incorporating diverse perspectives, including both egalitarian and libertarian approaches, we can work toward algorithmic systems that respect the complex and multifaceted nature of justice in the digital age.

1.8 References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. ProPublica, 23(2016), 139-159.

Arneson, R. J. (1989). Equality and equal opportunity for welfare. Philosophical Studies, 56(1), 77-93.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104, 671-732.

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. https://fairmlbook.org/

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In Conference on Fairness, Accountability and Transparency (pp. 149-159).

Binns, R. (2020). On the apparent conflict between individual and group fairness. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 514-524).

Bogen, M., & Rieke, A. (2018). Help wanted: An examination of hiring algorithms, equity, and bias. Upturn.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153-163.

Cohen, G. A. (1989). On the currency of egalitarian justice. Ethics, 99(4), 906-944.

Cohen, G. A. (1995). Self-ownership, freedom, and equality. Cambridge University Press.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214-226).

Dworkin, R. (2000). Sovereign virtue: The theory and practice of equality. Harvard University Press.

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

Fazelpour, S., & Lipton, Z. C. (2020). Algorithmic fairness from a non-ideal perspective. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 57-63).

Frankfurt, H. (1987). Equality as a moral ideal. Ethics, 98(1), 21-43.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. The Journal of Finance, 77(1), 5-47.

Green, B. (2018). "Fair" risk assessments: A precarious approach for criminal justice reform. In 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning.

Grgić-Hlača, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018). The case for process fairness in learning: Feature selection for fair decision making. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 51–57. https://doi.org/10.1145/3278721.3278725

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (pp. 3315–3323).

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference* (ITCS).

Lee, M., & Floridi, L. (2021). Algorithmic fairness in AI for social good: A survey. *Philosophy & Technology*, 34, 1023–1052.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.

Otsuka, M. (2003). *Libertarianism without inequality.* Oxford University Press.

Parfit, D. (1997). Equality and priority. *Ratio*, 10(3), 202–221.

Rawls, J. (1971). *A theory of justice.* Harvard University Press.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68).

Thierer, A. (2016). Permissionless innovation: The continuing case for comprehensive technological freedom. *Mercatus Center at George Mason University.*

Waldron, J. (1992). Superseding historic injustice. *Ethics*, 103(1), 4–28.

Kymlicka, W. (2002). *Contemporary political philosophy: An introduction* (2nd ed.). Oxford University Press.

Nielsen, K. (1979). Radical egalitarian justice: Justice as equality. *Social Theory and Practice*, 5(2), 209–226.

8.9 Commentary

What follows is fully human-written text, analyzing the fully AI-generated paper produced by Claude 3 Opus.

The paper produced by Clause 3 Opus has a few issues with structure and mechanics, but certainly nothing that would be difficult to fix. Some of the DOI addresses are not correct, and one of the citations given has the wrong title of the paper. The writing style in the paper is clunky and makes use of far too many different small sections of text, which makes it unappealing to read. However, these are small issues that could be fixed with a few rounds of editing after the paper was generated—there are far more serious issues that prevent this paper from being a viable academic work.

The paper begins relatively strongly, with a clear introduction and an accurate presentation of the problem. The background material on algorithmic fairness and distributive is presented clearly and correctly, and the model does a good job of connecting to and using relevant literature. For example, the presentation of canonical fairness measures on pages 4 and 5 is fully correct and cites the correct literature for each measure—the same literature cited in our paper. The model also does a good job of foreshadowing an eventual argument throughout the early sections of the paper. For example, on page 2, "These limitations highlight the need for approaches to algorithmic fairness that are philosophically grounded, contextually sensitive, and capable of addressing both procedural and distributive aspects of justice. This thesis aims to contribute to this effort by exploring how Nozick's entitlement theory might offer a distinctive perspective on algorithmic fairness—one that foregrounds procedural justice and historical processes rather than distributive patterns." The model does a good job of writing a viable start to the paper, all the way up until the point where it is meant to present a novel argument, beginning on page 10.

In leading up to a positive account of entitlement fairness, the model extracts a lot of similar significant features to what were extracted in our paper. Notably, on page 10-11, the model identified that for entitlement fairness, the algorithm must make determinations based on the actions and decisions of individuals as a matter of procedural justice, and must be transparent about what the actions and decisions used are, just as we identified that that positive and negative entitlement features should be the basis of the decision and must be made transparent. Similarly, the model points out that there must be a rectification procedure through which individuals can appeal the decisions of the algorithm.

When the algorithm attempts to provide a positive account of entitlement fairness beginning on page 13, the argument falls flat. We are provided with a measure which is the average of a certain score which is computed over all individuals in for whom decisions are made. The score contains 7 free parameters used as weights in linear equations, and the model provides no account of how those weights should be set. Even if the weights were defended, the terms that are being weighted are meaningless. For example, we are told that one of the weighted factors in the "consent and transparency index" is T_i, which is "a measure of the transparency of the algorithm's decision process for individual i, normalized to a value between 0 and 1." How is one meant to measure transparency? Why would transparency of one algorithm differ over different individuals? Similarly A is "a measure of the algorithm's accessibility to contestation and appeal, normalized to a value between 0 and 1." How is this meant to be measured? What factors are used to compute this value? The model makes no attempt to address or acknowledge these questions.

What we are left with is what you might have expected—the tone, and style of a viable academic paper, with accurate background information and quality structure, but with an extremely weak argument which is not well defended. Note that this aligns well with what was observed in the writing of our paper—Llama did really well on background research tasks, and in determining the structure of the paper, but was lacking in its ability to construct an original argument.