The Role of Large Language Models in Academic Writing

Edward Speer

California Institute of Technology

espeer@caltech.edu

April, 2025

1   Introduction

The majority of research journals now provide policies for the use of large language models (LLMs) as academic writing. In 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." In the Cambridge Press, "AI does not meet the Cambridge requirements for authorship, given the need for accountability." The chief concern among these policies appears to be responsibility. 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. This is a reasonable provision for ensuring accountability — how would one hold an LLM accountable for an error in writing? — but it also has the effect that *should* an LLM produce any original ideas that make it into a published paper, those ideas will not be attributed to the LLM but to the human authors, which represents a potential ethical dilemma if we find LLMs capable of contributing novel ideas and argumentation. This of course raises the question of whether LLMs are capable of doing so and to what extent.

The common role LLMs hinted at by these policies is as a copy-editor or proofreader, where the LLM is used to improve the clarity and coherence of academic writing, but not as an intellectual collaborator in the sense of contributing ideas and argumentation to the writing. In this paper, we attempt to explore the latter role, and to investigate the extent to which a current LLM can serve as a coathor in the sense of contributing ideas and argumentation to an original philosophical research paper. We do this by using a LLM to collaborate on a research paper in the field of algorithmic fairness and distributive justice. 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 current possibilities and limitations as well as a discussion of the future of academic writing in light of these developments.

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. Our investigation began from this vague notion of the intersection between algorithmic fairness measures and distributive justice, and we used the LLM as a collaborator to help us refine and develop our ideas from this starting point.

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 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.

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 a 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

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. 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.

Entitlement Justice and Measures of Algorithmic Fairness

Edward Speer
California Institute of Technology
espeer@caltech.edu

Llama-2-7b
Meta Platforms, Inc.

April, 2025

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 remain unanswered about the theoretical foundations of these measures and their relationship to broader sociotechnical systems. In particular, the relationship of these measures to philosophically rigorous definitions of justice is not well understood.

In an effort to develop the theoretical foundations of algorithmic fairness, 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.

Recent papers ((Binns, 2018), (Hertweck et al., 2024), and (Kuppler et al., 2021)) have explored the relationship between algorithmic fairness and theories of distributive justice. Efforts have largely focused on grounding fairness measures in liberal justice, while theories of libertarian justice have been largely overlooked in the literature. Given that libertarian justice is a prominent area of inquiry in political philosophy that addresses a broad range of concerns not covered in liberal justice, it is worth investigating how to close this gap. In particular, how do the concerns

of libertarian justice 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 libertarian lens?

In this paper, we will carefully examine the relationship between algorithmic fairness and libertarian justice, and develop a formalism that lays clear the relationship between the two. We will demonstrate that libertarian justice can be encoded within a measure of algorithmic fairness, and show that doing show offers a nuanced and context-sensitive means of understanding algorithmic fairness. We will argue that by conceptualizing algorithmic fairness through libertarian justice, system designers are made to clearly present the inherent normative reasoning and values embedded in their systems.

TODO: Sharped objectives in previous paragraph

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 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 resource R. A *decision rule* is a mapping $d \to \{0, 1\}$, under which i_n receives R iff $d(i_n) = 1$. 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.

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 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 esidence 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, goegraphic location of residence$\}$

- Let C = projected cost of living in location of residence, H = household income

- $f = H - C \geq 10000$

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 are significant, they do not fall simply within the domain of distributive justice, and so will not be considered here.

2.1 Algorithmic Fairness Measures

Algorithmic fairness measures as they're presented in the literature operate on the classification scheme f and a set of morally protected characteristics such as race or gender $A \subseteq X$, attempting to enforce contraints on how decisions can be sensitive to such characteristics. An overview of measures commonly discussed in the literature is presented below.

TODO: Elaborate/improve discussion of individual/group A critical point to note in the discussion of algorithmic fairness is the distinction between individual and group fairness. Individual fairness focuses on the fair treatment of a single individual, while group fairness focuses on the fair treatment of groups of people identified by some common morally protected characteristic. A typical approach to individual fairness is to require that similar individuals be treated similarly, while group fairness requires balancing statistical quantiites of outcomes across groups. For our purposes, both individual and group fairness can be characterized by what constraint they pose on the classification function, and so we will be able to analyze both through the same lens of distributive justice.

As a running example, consider the case of a hiring algorithm. The decision problem is to map a set of applicants to hiring decision. 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. This gives us the following setup:
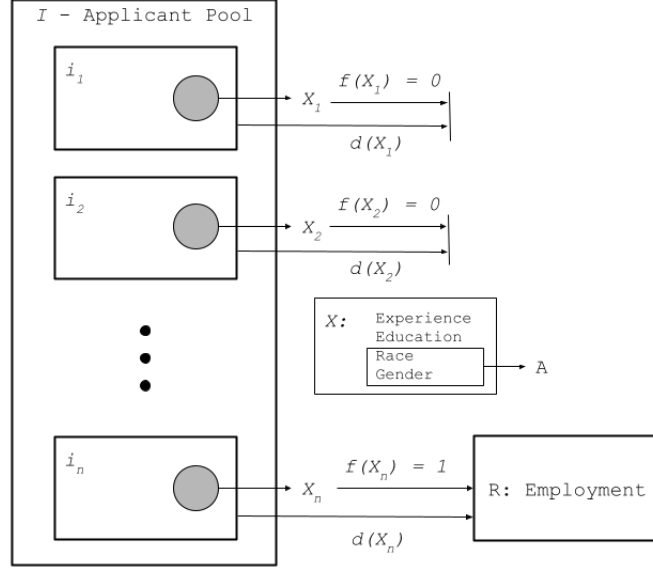
TODO: Rework Figure

Fig. 1: A hiring decision algorithm as a decision problem

where a fairness measure will operate exclusively on the mapping from experience, education, cover letter, race, and gender to hiring decisions.

Some of the most commonly discussed measures in the literature are presented below together with their strengths critiques leveraging this lens. For a more exhaustive list of measures, see Corbett-Davies et al. (2023).

TODO: Revise definitions

Definition 1. *Fairness Through Unawarenss — 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.

This notion is intuitively appealing — how can the algorithm dicriminate against me based on my race if it doesn't know my 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.

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. However, a close look reveals difficulties.

Under demographic parity, we must balance the probability of success between groups, which becomes very difficult when the base rates of success are unequal. 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, my false positive rate will be very high for men while my 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 true outcomes D over all individuals, we have*

$$P[Y = 1|A = a, D = 1] = P[Y = 1|A = b, D = 1] \ \forall a, b \in A$$

$$P[Y = 1|A = a, D = 0] = P[Y = 1|A = b, D = 0] \ \forall a, b \in A$$

(Hardt et al., 2016).

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 mistreatment 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 is often critiqued for struggling to deal with unequal base rates between groups. 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 have a base rate of recidivism much higher than white defendants (Crime and Alliance, 2023). As a result, allocating a parole to a white prisoner has a base line 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 white prisoners receiving parole undeservingly.

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.

Holding the other covariates constant, if we were to change the value of a protected attribute, the outcome of the classification function should not change. 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. However, how should one select a measure? What are the normative 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 introduces his account of justice through a thought experiment called the veil of ignorance. In this 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 themself endowed with a high level of intelligence, or a valuable skill, or wealth at birth, or one may find themself 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 group of individuals designing the society from behind the veil of ignorance as 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 soicety 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 formalised 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 entitled to 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 load 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 is equally likely to receive a loan, therefore balancing opportunity across groups. However, whether or not this satisfies the difference principle depends on the base rates of success across groups. If the base rate of loan default is higher for men than for women, demographic parity may require that we give loans to some men who are less likely to pay back their loans rather than qualified women, and in doing so, we are certainly not distributing inequalities to the less advantaged as we would need to under the difference principle.

3   Entitlement Justice

An entitlement theory of justice is a distributive theory of justice which posits the following distribution rule: Allocate amount R of resource X to agent A if and only if A is entitled to R of X. 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 though 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 refered 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 refered 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 mized their own labor with a resource, they transferred some of themself 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 posessions 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 themself 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 thoery 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 inerently 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.

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 thorugh 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.

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 if it represents the true distribution of property rights, but the process by which the output distribution is arrived at must be in accordance with the principles of just acquisition.

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? It is evident that to encode entitlement fairness, we cannot simply measure the outputs of an algorithm, but rather must analyze the treatment of protected 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 "cultural fit" which will be difficult to extract. These 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 morally 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. 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-dimensions 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 morally 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 return gains to the university, and therefore justify their admissions in a free market of talent.

- 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 has 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.

- 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.

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.

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. If one's college application demonstrates a strong work ethic, the individual has spent significant effort to endow their application with value derived from their own self-ownership. The right to self-ownership is a fundamental principle that does represent one form of equality baked into the entitlement approach — we are required to treat the endowment of value derived from self-ownership as equal across all individuals, irrespective of their characteristics or group membership. Notice that I care about this on an individual basis — if for just one person I value their self-ownership less than another, I have failed to respect their property rights. As a result, it seems clear that the appropriate way to ensure that predictors of positive entitlement features are fair is to ensure that they satisfy counterfactual fairness. Holding fixed those covariates of an individual which give rise to the positive entitlement feature, we must ensure that the predictor outputs the same value regardless of the individual's other features.

For negative entitlement features, the situation is a bit different. Rather than gaining an entitlement through the mixing of one's self with the subject of the entitlement, the basis for negative entitlement features is the failure to respect the inherent rights of others. For example, if an individual cheats on entrance exam scores that they include in their application, they are attempting a coercive act that violates the rights of other students who are applying honestly and in doing so, they forfeit their ability to gain entitlement over admissions. In this case, our chief fairness concern is with errors in the predictor. Wrongful accusations of negative features caused by false positives in the negative feature predictor result in the deprivation of one's property rights and therefore must be avoided. Similarly, false negatives in the predictor result in the wrongful granting of property rights to individuals who have forfeited them, to the exclusion of others who are more entitled.

Given it is not possible to eliminate all errors in the predictor, we must differentiate between the two types of errors. False positives in the predictor result in wrongful accusations and therefore wrongful deprivation, but these cases may be fixed through the process of rectification in accordance with entitlement theory. In other words, if a negative predictor accuses an individual of some negative feature, say cheating on a test, that individual must be given the opportunity to defend themselves. If they are able to prove their innocence, then the initial allocation must be rectified, and the individual should receive the property right they were entitled to had they not been flagged for cheating. In contrast, false negatives in the predictor result in property deprivation in a way that is much more difficult to rectify. If we allocate admissions to a student who has cheated on their application and we fail to catch them, we won't know that we have made a mistake, and the student who should have gotten in in their place similarly won't know or be able to challenge the relevant decision. As a result, we must be much more careful about false negatives in the predictor of negative entitlement features than false positives. This culminates in

two results. Firstly, if an individual is denied an allocation on the basis of a negative entitlement feature, they must be given the opportunity to defend themselves in accordance with the principle of rectification. Secondly, the predictor of negative entitlement features must be as accurate as possible according to a weighted loss function that penalizes false negatives more heavily than false positives.

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 morally 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 disadvantaged 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.

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 morally relevant features of each individual in the population, $V = \{v_1, v_2, \ldots, v_j\}$. These features should be partitioned into two sets, V^+ and V^- where V^+ contains the positive entitlement features and V^- contains the negative entitlement features relevant to the problem domain. Implement a predictor for each feature, $v_j = \hat{p}_j(X)$. Now, these are the conditions for entitlement fairness:

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

2. $P[p_{j,A \leftarrow a}(X) = 1|X = x, A = a] = P[p_{j,A \leftarrow b}(X) = 1|X = x, A = a] \ \forall a, b \in A, v_j \in V^+$ (Positive Feature Counterfactual Condition)

3. $P[p_j(X) = 1|v_j(X) = 0, A = a] = P[p_j(X) = 1|v_j(X) = 0, A = b] \ \forall a, b \in A, v_j \in V^-$ and $P[p_j(X) = 0|v_j(X) = 1, A = a] = P[p_j(X) = 0|v_j(X) = 1, A = b] \ \forall a, b \in A, v_j \in V^-$ (Negative Feature Odds Condition)

TODO: College admissions example all the way through

5 Discussion

References

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.

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.

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.

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, 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.

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 describe show examples of the LLM's contributions and activities in each phase.

4.1 Literature Review

The first phase of the project was to carry out 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 two-sided 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 as 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 though 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 ask 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. The text produced is not only in need of line-editing, but is lacking in focus and detail, and often repetitive. 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

One of the most interesting aspects of this research was the way that Llama took on different and unintended modalities 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, 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 describe previously; perhaps the human author could be held accountable for *publishing* ideas produced by Llama, but should those ideas really be contributed to the human author if they are genuinely originated within the model?

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. 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 modalities which lead to the production of new ideas until such issues are resolved.

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.

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