Justice In The Machine: A Rawlsian And Kantian Framework For Fairness In Generative AI

[1.]Dr. Soma Das, [2] Mr. Suhail Ahmad [3] Mr. Akhilesh Kumar Pandey
[1]Prof. and Dean, [2] Additional Registrar [3] Assistant Professor
[1]School of Humanities and Social Sciences, Monad University, Hapur.
[2] Monad University, Hapur, [3] School of Law, Monad University, Hapur.
[1] somanaina6032@gmail.com [2]suhail.philosophy@gmail.com
[3] akpcmat1978@gmail.com

Abstract:

Generative Artificial Intelligence (GenAI) is a new frontier for computational creativity and large language models (LLMs), along with diffusion models, provide existing capabilities for generating text, images, audio, and video, that are like nothing seen before. However, no system is developed in a vacuum and GenAI models are trained on vast volumes of human data that often entrench and amplify the biases, prejudices, and social injustices endemic to society. The paper asserts that GenAI will exacerbate injustice on a never-before-seen scale if deployed without appropriate ethical safeguards due in large part to the human data it must rely on. The proposed ethical framework is informed by John Rawls's justice as fairness and by Immanuel Kant's respect for autonomy. In Rawls, we have a distributive perspective from which to assess how the benefits and burdens of AI are apportioned so that we consider more than just technical fixes and socio-political justice. In Kant, we see the importance of transparency in the communication of human, and AI, activities so we may empower individuals with informed consent and have respect for individual autonomy and agency. Kant also provides developers and designers working with AI with a guide to articulating ways that users may engage with models and holding the models accountable for design.By analyzing bias within image and language generation and considering the effectiveness of various corporate mitigation efforts, the conclusion of this paper is that only when we integrate these justice principles into the AI lifecycle (data curation through deployment and auditing) can we realize GenAI's innovative risk and opportunity while directing it toward a more just and equitable society. The paper ends with an ethical prescription based on conclusions drawn, highlights crucial spaces for research that needs to bring philosophical theorization into technical implementation, and calls for policy-based standardization of these fairness principles.

Keywords: Generative AI, Algorithmic Bias, Algorithmic Fairness, AI Ethics, John Rawls, Immanuel Kant, Distributive Justice, Autonomy, Transparency, Explainable AI (XAI), Research Gaps.

1. INTRODUCTION: THE DOUBLE-EDGED SWORD OF GENERATIVE AI

The rise of Generative AI (GenAI) represents a seminal moment in the evolution of technology. This subset of artificial intelligence creates unique, high-fidelity content such as coherent text, photorealistic images, complex music, or functional code. Specific models of GenAI like OpenAI's GPT-4, Google's Gemini, and Stability AI's Stable Diffusion, have advanced from the research domain to global, mainstream applications at rapid speeds. Its impacts are felt everywhere, from conversational assistants to design practices, science-based discovery, and educational applications in every sector. The possibilities for society are vast, with the potential for enhanced creative processes, increased avenues for content creation, and accelerated innovation in all fields. However, this exciting promise is dulled by an ongoing and deeply entrenched challenge: the challenge of algorithmic bias. Evolving from previous software tools, GenAI is not bound by explicit programming rules; rather, it learns probabilistic patterns based on a huge amount of training data pulled from petabytes of text and images from a plethora of open internet sources. This dataset is in part a reflection of collective humanity, from our best, to our worst, and, most damaging, the vast history of injustice, stereotype and systemic inequities that humanity has perpetrated.These models not only reproduce these patterns, they also elevate and remix them, often creating outputs that are biased, discriminatory, or harmful. For instance, if an image generator is instructed to produce "a productive person," it might generate a light-skinned man in a business suit. If the prompt were "a person at a social protest," the generator may produce pictures of people of color while implicitly associating disorder with these races. A language model may generate text that shows bias based on gender in professional contexts, or is insensitive towards non-Western traditions and cultures. These are not hypothetical bugs; they are part and parcel of systems that were trained on imperfect data. This fact presents a big and pressing research question: how can we operationalize and secure fairness and justice in the design and deployment of generative AI, without overly heavy-handed regulation that curtails technological innovation and the potential benefits it can provide?

This paper argues that innovation and fairness are not a zero-sum game, but actually complementary and mutually reinforcing. The main argument is that if we actively and implicitly embed principles rooted, or informed, by standard philosophies (specifically, John Rawls's justice as fairness and Immanuel Kant's categorical imperative) into the very DNA of the AI life cycle, we can

eliminate, or at least reduce, bias. Not as a tech fix by "de-biasing" or by a post-hoc filter, but as justice-by-design. This would provide a moral compass for tech decisions. The remaining sections of the paper will explore the literature that raises alarm bells for algorithmic harm, develop a solid theoretical foundation using Rawls and Kant, investigate particular examples of bias in image and text generation, assess the strengths and weaknesses of industry responses, and finally, offer a set of ethical prescriptions and future work informed by ethics.

2. LITERATURE REVIEW:

Mapping the Terrain of Algorithmic Harm The discussion of algorithmic bias is always interdisciplinary, drawing on computer science, ethics, sociology, critical race theory, and law. Foundational texts have helped us understand how automated systems work, and reproduce inequality as a result.Cathy O'Neil's important book on this risk, Weapons of Math Destruction (2017), substantially contributed to the public mind about this risk. O'Neil is a data scientist and claims that algorithms often "encode human prejudice, misunderstanding, and bias into software that rapidly and opaquely scales." She outlines three key features of these destructive models: opacity (the processes are hidden from the impacted), scale (they can be deployed to millions), and damage (they can harm - as in deny opportunities). While her work is primarily about predictive algorithms, typically used in justice (recidivism risk scores), hiring, and lending, the implications are terrifying for GenAI. GenAI systems are defined by scale; their processes are inscrutable even to their creators; and the damage they can inflict - through representation harm, amplification of misinformation, and engrainment of stereotypes - will have ever-worse ramifications and even more insidious cultural damage.

In reference to this, Virginia Eubanks, in Automating Inequality (2018), offers a finer-grain, human-facing case for how automated decision-making systems implemented via public utilities (welfare eligibility, child protection and policing) just routinely punish and surveil the poor.By using ethnographic work, she shows how technical systems can, under the mask of neutrality, objectivity, and efficiency, normalize and signify existing discriminations. She demonstrates this in a way that serves as a cautionary tale for using GenAI in consequential social contexts. Picture a GenAI system used to automatically summarize responses for child welfare cases, or to determine eligibility for a social program; Eubanks's work is more than merely illustrative as to how such a system could, in the way she illustrates, scale-injustice.

The discipline of computer science, as a field, has experienced a growth of algorithmic fairness as a subfield, from a marginal pursuit to an organized research area in the mainstream. The foundational work of CTR scholars such as Solon Barocas, Moritz Hardt, and Arvind Narayanan has developed a suite of quantitative appropriateness metrics. Examples of these metrics include group fairness metrics, like demographic parity (the requirement of outcomes to be independent of protected attributes), equality of opportunity (the requirement of equal true positive rates across all groups), and individual fairness (the requirement of similar individuals receiving similar outcomes). This community of technical literature that continues to grow, as published by the most elite conference venues globally (Conference on Fairness, Accountability, and Transparency (FAccT); ICML; NeurIPS), should provide the canonical toolbox for measuring and quantifying bias.The primary weakness of this overwhelmingly technical discussion is that it is too often described with a an almost anaemic normative base. Such discussion often focusses on the sometimes mathematical trade-offs of the various descriptions of fairness (ex. you cannot often satisfy both demographic parity and equality of opportunity) while not providing a lot of guidance as to why a developer, corporation or regulator would choose one fairness objective over the other in a specific context. Instead, these conversations seem to present the choice as a technical optimization problem, and not a normative ethical decision ripe with value choices. This paper seeks to bridge that important gap, by providing a strong, philosophically grounded rationale for prioritisation of fairness objectives, demonstrating how Rawlsian and Kantian ethics can be used as a map for navigating trade-offs.

3. THEORETICAL FRAMEWORK:

Philosophical Direction for AI Fairness To develop a strong and normative base for fairness in GenAI, we use two foundational modern moral philosophies that consider justice at a societal and individual level.

3.1 John Rawls's Justice as Fairness: A Macro-Level Plan of Action John Rawls'

A Theory of Justice, arguably the most influential political philosophy of the 20th century.Rawls asks us to craft principles of a just society from a fictitious original position whose participants are under a "veil of ignorance." This position is like a thought experiment, and individuals, at this stage, do not know their own subjective position in the world—i.e. they do not know their race, sex, class, physical or mental talents, or vision of the good life. The idea is that from this original position of fairness and impartiality, Rawls argues rational persons would adopt two principles of justice:

(a) The Liberty Principle: That everyone is to have the same basic liberties, with the same number of basic liberties, as others; and the political liberties (liberty of position); and freedom of speech and assembly; and the right to personal property. These may seem fundamental to many Western societies, but the Liberty Principle is one fundamental principle that successful rational individuals would choose.

(b) The Difference Principle: Social and economic inequalities are to be arranged to be both (a) to the greatest benefit of the least advantaged; (the maximin rule); and (b) attached to offices and positions open to all under conditions of fair equality of opportunity.

As applied to GenAI's development and deployment, the Liberty Principle argues that users should be protected from systems capable of infringing their basic rights. For example, a GenAI that produces non-consensual deepfake pornography or maliciously defames someone is a direct assault on personal liberty and security. Or, more subtly, a model that systematically misrepresents or erases a cultural group undermines their liberty to be equal participants in society.

The Difference Principle is stronger still and even more directly relevant. This principle requires that AI's development and the benefits of AI do not deepen existing inequities but instead work to improve the standing of the most disadvantaged. This is not about equalizing outcomes, but fairly distributing the benefits of technological progress. For example:

- A just GenAI tool for legal aid would be designed to help those unable to afford a lawyer in the best way possible, not who biggest corporate law firms have made most effective and optimized.
- A just image generation model would be designed from the "veil of ignorance", understanding that its implicit representations of beauty, norms of authority and professionalism, would be diverse and de-stereotypical, and would actively improve the symbolic standing of marginalized groups.
- Under the "fair equality of opportunity" clause, GenAI instruments used in higher education, or hiring, cannot put already disadvantaged groups at a further systematic disadvantage. This interprets fairness not simply as the absence of statistical bias, but instead as a forward-thinking, distributive requirement for equitable outcomes. The question is not "Is the model biased or unbiased?" It is "Does this model create a situation or better the situation of the least advantaged?"

3.2 Kantian Respect for Autonomy: A Micro-Level Guide for Interacting with Individuals

Immanuel Kant's moral theory, a deontological moral philosophy based on the categorical imperative, is critical to the ethical interaction between AI and individuals. For AI, the most pertinent iteration of the categorical imperative is: "Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end."

To treat someone as an "end" means one acknowledges their dignity and worth as a rational, autonomous person. To treat someone as "merely a means" means one uses them as an instrument or tool for one's singular purposes, with no respect for their goals or ability to bring their decision-making power to bear.

In the context of GenAI, respecting autonomy involves the following requirements:

(a) Transparency and Non-Deception: It is imperative for users to be informed in a way that they understand they are interacting with an AI. Non-transparent systems artificially make users accept they are engaging with a human user (for example, a few chatbots) and treat the users merely as a means to obtain something they want (for instance engagement metrics or data about the user). We are claiming that Kantian ethics require such transparencies. Although not covered in the Trustworthy AI Principles, practitioners should do things like disclose when the logic of the AI charts a specific trajectory. This can even be thought of in terms of "Explainable AI (XAI)" which represents efforts to make the outputs and decisions of seemingly incomprehensible models comprehensible to humans.

(b) Informed Consent: It is important for users to have meaningful agency in their interaction with AI. This implies informed consent regarding how their input data will be used, for example for training or inference and the ability to opt-out. Generating deepfakes or synthetic media of a real person without that person's clear consent is a glaring example of treating that individual as a means (e.g. for entertainment, profit or malice) instead of an end-in-themselves, with their autonomy as an individual and their right to self-represent.

(c) Respect for Rational Agency: AI systems should be designed to augment, not undermine, human reasoning. Systems which generate something which is socially persuasive misinformation or generate addictive feedback loops violate the rational capability of the user.

Rawls and Kant collectively offer a useful normative model for considering AI ethics. Rawls gives us a macro, distributive model for us to ask the question of the impact of AI systems and the role of AI in a just society. Kant gives us a micro, deontological model for us to ask questions of individual interaction, design choices, and what we owe to each individual user. On one hand, we are asking questions about what is the just distribution of the technology; on the other hand, we are asking questions about the morality of each interaction that it facilitates.

4. ANALYSIS: THE FRAMEWORK TO PRACTICAL PROBLEMS

4.1 Case Study: Bias in Image Generation Models like DALL·E, Midjourney, and Stable Diffusion have been found to propagate several consequential racial and gender stereotypes. A landmark study from the University of Washington in 2023 found that prompts for "a person in a low-income country" or "a person living in poverty" overwhelmingly result in images of dark-skinned people, usually in an African context, while prompts for "a CEO", "a director", or "a professor" yield images of well-dressed white men. "Nurse", "flight attendant", or "social worker" had a large female representation, while "engineer", "doctor", and "criminal" had a large male representation.

• **Rawlsian Analysis:** This output is a clear violation of the Difference Principle. The algorithm exacerbates negative outcomes, as the technology inherently embeds and perpetuates harmful stereotypes that will prejudicially affect already marginalized groups. The output provides no "fair equality of opportunity" by visually pairing "high status" and authoritative roles with specific demographic information (white, male) and negative or low status images (people of color, women). A system designed behind the "veil of ignorance" would remove these kinds of associations. The response of justice is not merely to "balance" the dataset as is suggested, it is actually to curate the dataset and adjust the model such that the algorithm's inherent output opposes the pre-existing embedded bias in society instead of simply reproducing and exacerbating it - thus positively benefiting the least socially advantaged ontologically.

• **Kantian Analysis:** The production of stereotypical imagery fails to respect the humanity of the subjects portrayed, reducing the subjects to caricatures and drawing them from collective memory, with the AI using the statistically 'reasonable' output to finish its task. The biased and limited view as a neutral output does not respect the autonomy of the user. The output diminishes the user's ability to access fair information and make fully informed judgments by limiting their perspective in a limited and biased way, thus using them (to generate/give a quick image) as a means to an end (e.g., generating a quick image) by not respecting their rational agency.

4.2 Case Study: Bias in Large Language Models (LLMs) LLMs such as ChatGPT exhibit more subtle, but equally harmful types of bias, including political and cultural bias. Typically trained on Western, English-language text taken from the internet, LLMs have a tendency to represent a liberal, cosmopolitan worldview, and they often do not have the ability to represent a conservative worldview, nor non-Western worldviews or indigenous knowledge systems at all. When examining the underlying processes developed within LLM (the ways LLMs reproduce their historical, cultural, and ideological biases as a composite of their human-made training datasets), we can even argue that LLMs often exhibit "sycophancy" where they go so far as to agree with whatever a user stated based on their own political bias, regardless if that bias is based on a falsehood. LLMs can also generate text that may be culturally insensitive or historically misrepresented through a Eurocentric lens.

• **Rawlsian Analysis:** This highlights an asymmetrical distribution of cultural and epistemic power; the hegemonic representatives, voices, narratives and knowledge systems of dominant groups (Western, English-speaking, majoritarian demographics) are amplified and centered with respect to their dominating counterparts (minority and non-Western cultures). This dilemma undermines the Liberty Principle as it artificially limits the intellectual liberties of users who receive heritage ideologies based on a biased and incomplete worldview, and prohibits the continued global discourse that the user can participate in must remain historically and culturally valid.

•**Kantian Analysis:** Providing biased information and failing to tell the user what the limitations of the model are, along with its training data, is a fundamentally deceptive action. Treating the user as a means to achieve an end (a fast answer) denies the user of their autonomy as a rational agent searching for truth. Conversely, if an LLM were to be deployed to create content that was appropriative of or misrepresents a culture, and not in context, it would treat that culture as a means to itself (as a source of aesthetic or narrative development), rather than as an end-in-itself with its own sovereignty and dignity.

4.3 Evaluation of Corporate Responses: Strengths and Ethical Gaps Leading AI companies - OpenAI, Google, and Microsoft, in particular - know that they have ethical obligations and all have begun implementing a range of mitigation modelled after ethical guidelines:

•**Pre-training Data Filtering:** Classifying and filtering out elements of training datasets that are contentiously harmful, pornographic, or hateful.

•**Reinforcement Learning from Human Feedback (RLHF):** A method in which human reviewers evaluate potential model outputs, generates a reward model, and then values the model outputs to help the AI "be helpful, harmless, and honest." RLHF is a strong method of human-guided model behaviour.

•**Output Blocking:** Implementing a real-time classifier that will refuse to generate outputs deemed violent, hateful, sexual, or dangerous.

• **Strengths:** These methods have demonstrably reduced the most inappropriate, dubious, or outright harmful outputs. RLHF, in particular, allows a nuanced form of value-based shaping of model behavior that goes beyond simply blocking labelled keywords.

Weaknesses/Ethical Deficiencies (Interpreted through our frame):

• **Opacity and "Black Box" Governance:** The RLHF process is proprietary, along with the reviewer identities, their training, and the values they are supposed to uphold. This contradicts Kant's necessity of transparency. Users and civil society cannot inspect the values that the model embodies, making it impossible to give informed consent to engage with a system that does not disclose the moral compass that underpins its recommendations. This is treating users as means, not ends.

• **Paternalism and the "Censorship" Issue:** Excessive output blocking could also be paternalistic; for example, refusal to generate output about certain historical events or artistic themes for the sake of some ephemeral social conscience limits legitimate creative, academic, or satirical inquiry. This should be seen as a refusal to respect the user's Kantian autonomy to deploy a tool toward their own rational ends, provided their ends are not harmful in and of themselves.

•**Superficiality and the "Bias Band-Aid":** Many approaches focus on sanitizing outputs rather than truly addressing the source which is the biased training data. Much of this work is simply a technical fix, instead of a socially just redesign. Rather than stopping bias at the source, it simply tries to put a cap on the geyser of bias. This does not comply with the Rawlsian demand to actively improve the standing of the least well-off through a fair foundational design. Additionally, this often leads to a never-ending whack-a-mole process of identifying new biases.

4.4 Ethical prescriptions: From Theory to Practice.

Based upon the combined Rawlsian and Kantian framework, we offer a multi-layered, holistic approach:

(a) Fairness Auditing should be mandatory. Mandatory third-party auditing of models should be openly published before they are deployed to the public. Any audit should be comprehensive and not restricted to narrow technical metrics. Model performance must assessed for performance across a variety of demographic and cultural factors. A set of metrics should used, with method for choosing justified on Rawlsian grounds (that is, giving priority to metrics that prioritize the least advantaged). The results of any audits should be openly published.

(b) Data curation must move from a de-biasing approach to a proactive curation of training datasets. This means:

• **Inclusive Sourcing:** Actively seeking and bringing in text and imagery from a range of cultures, languages, and perspectives that have been historically marginalized on the mainstream internet.

• **Community Relations:** Collaborate with libraries, universities, and cultural institutions from marginalized communities to bring in their archives and knowledge. This relates to the Rawlsian "veil of ignorance" -- designers must build models for a world where any user could be of any background, and so the training data must represent that pluralistic world.

(c) Radical Algorithmic Transparency: Companies must conduct Explainable AI (XAI) research for generative models. At the very least, we need model cards and datasheets that clearly outline a model's capabilities and limits, its intended uses, demographics of its training data and human feedback volunteers. This can satisfy Kant's informed consent, which allows it to be known what the tool a user is using actually does.

(d) Participatory Design: All design and testing should include stakeholders from marginalized and diverse communities. Ensure that the least advantaged have agency in shaping the technology that will influence them. This is directly applicable Rawls' principles. This could look like ethics review boards, user advocacy associations, and inclusive beta-testing programs.

5. IDENTIFICATION OF RESEARCH GAP: PHILOSOPHICAL THEORY TO PRACTICAL TECHNICAL IMPLEMENTATION

Even if Rawls' and Kant's ethical framework is solid, and the proposed solution is sound, there is a significant gap between high-level philosophy and applied AI development and this gap is arguably the most important area for future study. The shift from principle to practice is nontrivial and is rife with still unresolved technical, methodological, and epistemological work. This gap can be conceptualized as three interrelated research dimensions:

5.1. Technical implementation gap: This is the most straightforward problem. How does one technically embed principles, for instance the "Difference Principle" or "respect for autonomy", into a model's architecture or training loop?

(a) Quantifying "Least Advantaged": Rawls' theory requires that you identify the "least advantaged" population before prioritizing their benefit. Operationally defining this for a global AI system is exceptionally difficult. Are they the globally economically poor? The intersectionally discriminated? A culturally marginalized group? We need research to develop scalable, context aware metrics for disadvantage, that can be included in loss functions or evaluation suites.

(b) Algorithmic Translations of Kantian Ethical Principles: How does an engineer construct a "transparency" module? Current methods of Explainable AI (XAI) such as SHAP or LIME are chiefly inadequate for billion parameter generative models because they tend to provide post-hoc explanations that lack reliability and could, in fact, be misleading. It is essential to conduct basic research on novel paradigms of intrinsic transparency and verifiable reasoning in generative models, and only then can we meet the Kantian standard of "doing no harm," that is, respecting user autonomy by providing context and understanding.

(c) The Trade-Off Problem: Exploring the trade-offs between Rawlsian objectives (e.g., training for worst-case group performance) and model design attributes (e.g., overall accuracy, creativity, innovation) is a very open area for research. Just what the mathematical and computational limitations for implementing a maximin strategy in training models are poorly understood.

5.2. The Participatory Design and Operationalization Gap: Participatory design is easy to prescribe and incredibly hard to do well and at scale.

• Methods of Inclusive Stakeholder Engagement: There is a lack of empirically validated, scalable frameworks for incorporation of global perspectives at all tables in the technical design process. Who is at the table? How are their contributions obtained, weighed, and translated into design directives? A research agenda must progress beyond tokenistic focus group consultation for the purposes of creating structured, equitable, and individually efficient methods for democratic input on AI development.

• From Community Input to Data Curation: A key question is how to translate the values and knowledge of specific communities into an actual process of defining the training dataset for an AI or specifying the reinforcements for reinforcement learning. These challenges require interdisciplinary teams including AI ethics, anthropology, data science, and more, to develop new tools and workflows collaboratively.

5.3. The Policy and Audit Gap: The objective of "mandatory fairness auditing" has emerged before a field has been established to do it.

• Standard Auditing Frameworks: Model cards and a datasheet are a good first step, but they do not constitute standard models for auditing a generative model for fairness. Research is critically needed to establish comprehensive, multi-faceted audit frameworks that can be applied consistently across various model types (text, image, video) and cultural contexts.

• **The Role of Regulation**: The most significant research gap is finding what policy is best. Should regulation require a certain outcome (i.e., a Rawlsian approach), a certain process (i.e., a Kantian transparency approach), or some mix of both? What regulatory regimes allow for innovation while preventing harm? If we are to design effective governance that can keep up with rapid technological change, we need comparative legal and policy research.

This gap analysis shows that creating genuinely fair GenAI involves much more than a conceptual exercise in philosophy or a technical exercise in computer science; it is a socio-technical issue. To address this gap requires a new type of interdisciplinary work that combines principles of moral philosophy with advances in computer science, sociology, law, and design.

6. CONCLUSION AND FUTURE DIRECTIONS:

The development of generative AI is not merely a technological journey. It is a global social, economic and ethical experiment. The challenge for this decade will be to harness the power of this general purpose technology on behalf of equity and justice rather than inequality, discrimination, and harm. This paper contends that while technical solutions matter, they are not the only helpful step; we need to make sure to have a clear and strong ethical compass also.

John Rawls' and Immanuel Kant's frameworks present that ethical foundation. Rawls' theory of justice as fairness requires us to assess AI systems not by the state-of-the-art level of technology but instead by how they share the benefits that societies authorized and still use along with how they provide the means to protect their most vulnerable members. Kant's theory of respect for autonomy demands that we use radical transparency for design, along with consistent and thorough approach about consent, wherein users are not reduced to lab rats or data points or engagement statistics but respected as ends-in-themselves.

The proposed "best practices" that we outlined in this paper—independent fairness auditing, proactive data curation, radical transparency, and participatory design—are concrete examples of how these philosophical commitments come to fruition. Best practices—from government regulation to industry involvement and self-regulation—are not barriers to innovation, but rather the basis for innovation, shielding it from unethical harm, and building internal legitimacy by maximizing the likelihood that innovation is inclusive, sustainable, and ethical. Best practices create a path toward trust and legitimacy, the ultimate catalysts for long-term sustained adoption of technology.

Future research must urgently reach for the gaps articulated above, especially in:

1. Technical Research: New algorithms, loss functions, and new model architectures that explicitly optimize for Rawlsian and Kantian values.

2. Methodological Research: Robust, scalable participatory design and stakeholder engagement frameworks that would fit within the AI development life-cycle.

3. Policy Research: Flexible regulatory audit framework that standardizes fairness assessment processes but provides room for innovation to flourish below it.

In addition, answering the questions below will be key in moving ethical theories into practical application:

- How would governments and international organizations encourage development of Rawlsian AI—through tax credits or public procurement standards to use ethically audited AI models?
- What would a liability framework look like to adapt to the requirements of Kantian transparency, and hold companies accountable for harms created by opaque systems?

If we internalize the enduring values of Rawls and Kant in our latest technologies and intentionally put in the effort to 'bridge the implementation gap' as mentioned above, we can work towards ensuring the generative AI revolution is for all of humanity, not just the privileged few. The aim is not who can build the most neutral machines, but rather who can build just ones.

7. BIBLIOGRAPHY

- Abid, A., Farooqi, M., & Zou, J. (2021). Persistent Anti-Muslim Bias in Large Language Models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.
- Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.
- Bender, E. M., Gebru, T., McMillan-Major, A., &Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity.
- Birhane, A., & Prabhu, V. U. (2021). Large Image Datasets: A Pyrrhic Win for Computer Vision?. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*.
- Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, 81.
- Crawford, K. (2021). *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.
- Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.
- Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Communications of the ACM*. (Added for trade-off problem)
- Kant, I. (1785). *Groundwork of the Metaphysic of Morals*. (Trans. by H.J. Paton, 1948).
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)*.
- O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group.
- **Passi, S., & Barocas, S. (2019). Problem Formulation and Fairness. *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAccT)*.** (Added for participatory design gap)
- Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.
- Raji, I. D., et al. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT)*. (Added for auditing gap)
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Sheng, S., Chang, K., Natarajan, P., & Peng, N. (2019). The Woman Worked as a Babysitter: On Biases in Language Generation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Selbst, A. D., & Barocas, S. (2018). The Intuitive Appeal of Explainable Machines. *Fordham Law Review.*
- Zou, J., &Schiebinger, L. (2018). AI can be sexist and racist — it's time to make it fair. *Nature*, 559(7714), 324-326.