# Re-Visions of Now and Future:

Commentary & Research on Contemporary AI and a Way Forward

**A Community Publication by <AI & Equality>**

December 2023

# Table of Contents

# Introduction

Our title of *Re-Visions of Now & Future* encompasses <AI & Equality>'s goal to build a community around idea sharing and re-imaging our current world in order to work towards a collective positive and inclusive vision for our future. We aim to use this publication as one of many tools for discussion, action and community building. We accomplish this through our online <AI & Equality> community aimed at forming collaborations and action through global and interdisciplinary connection. The research and ideas presented within this publication exist beyond these pages and within our community – we encourage readers to engage with them through the <AI & Equality> community (more information can be found at community.aiequalitytoolbox.com). We hope these pieces present you with new information and ideas, and most importantly, inspire you to actively get involved in envisioning and shaping a new way forward in a human-rights-promoting and feminist-orienteted AI world.

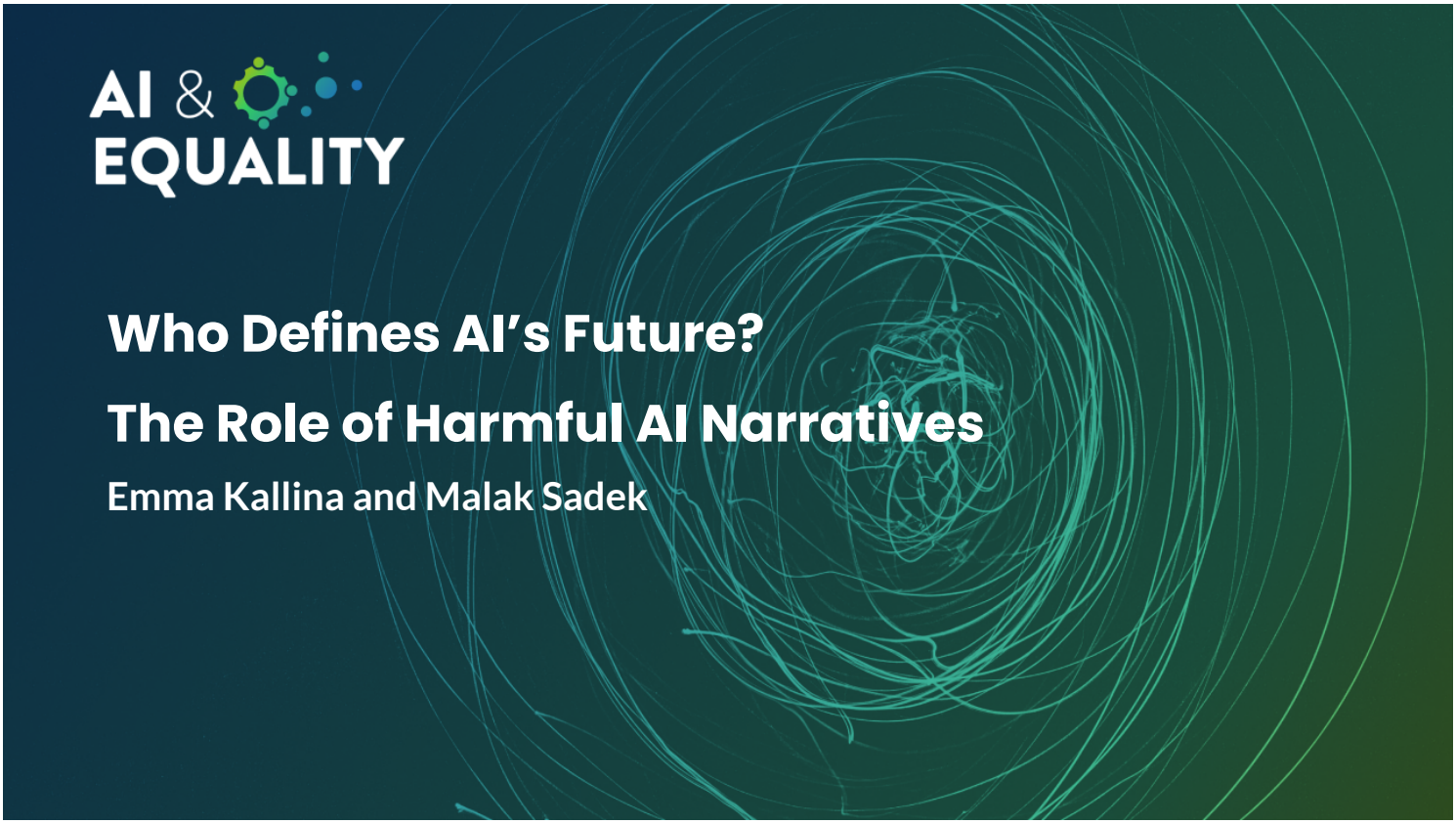
The increasing integration of artificial intelligence (AI) into our world and its potential for harm no longer comes as a surprise. As AI tools continue to be deployed, there is a need for iterative reflection, investigation, and debate on how AI impacts our sociotechnical world. However, we can't stop there. It is also necessary to ask ourselves: what *are* solutions and remedies to identified problems? Grounding these questions in human-rights based frameworks provides agreed upon definitions of harm and remedies to harms of AI. Combining feminist and human-rights based approaches to the lifecycle of AI development and deployment allows us to understand AI through a shared framework and provide actionable solutions to issues in our datafied world.

This community publication provides space for authors from diverse disciplines and backgrounds to showcase new perspectives and advocate for what we want our future to look like. Our publication begins with *Who Defines AI's Future? The Role of Harmful AI Narratives*, which introduces current narratives about AI and provides current good practices that work in battling harmful AI narratives. Next, *Re-imagining Artificial Intelligence on the African Continent* dives deeper into a geographical context, looking at the harms of AI and solutions in the African context. Similarly, the pieces *Feminist Design Framework for Envisioning Gender-Inclusive Ride-Hailing Sector* and *Mainstreaming Gender Perspective in AI crowd work in the Global South* use

interviews and survey methodology to illustrate the nuances of AI and labor in the geographical contexts of India and Latin America. From geographic contextualization the pieces then expand to a legal understanding of AI, as *Possibilities and Perils of Artificial Intelligence in Open-Source Human Rights Investigations* and *Copyright Law in the Age of Machine-Generated Art* illustrate how AI can be a tool in legal investigation and how legal frameworks shape our AI tools. Finally, our publication concludes with *Not Losing Ourselves to the AI Storm*, a piece that (literally) illustrates how "AI lives in the past and dreams of the future" and prompts us to consider the power that we, as humans, hold as we shape the future of AI. Throughout these pieces, we are introduced to a narrative of both broad and local contexts of AI and emphasize the importance in not only theoretically identifying issues with AI, but gaining first-hand perspectives and providing solutions.

– Anna-Maria Gueorguieva, *Editor*

*<AI & Equality> is a community committed to establishing and promoting a human rights-based approach to AI that centers equity & inclusion at the core of the code. All authors are members of our online community. This open-to-anyone global community aims to connect individuals from all backgrounds, regions, and disciplines to work towards a collective goal of human-rights based AI. We plan to continue providing a platform and space for sharing the thoughts, ideas, and work of our community through future publications. Join us at community.aiequalitytoolbox.com.*



# Who Defines AI's Future?
# The Role of Harmful AI Narratives

Emma Kallina and Malak Sadek

## Abstract

As AI-based systems are becoming more mainstream, they are increasingly discussed within public spheres. Thereby, several prevailing narratives around AI systems form almost stereotypical perceptions as to who is in control, who will benefit, and what these systems are capable of. The narratives dominate the development process, resulting in a narrow vision that largely fails to counteract the imprint of historical (and present) power structures in future technologies (Hoffmann, 2019; Lee & Singh, 2021). Since they define expectations of power, they further limit the access of underrepresented groups - both by decreasing the perceived necessity to include them, as well as by lowering their own perception of self-efficacy. The former causes the AI development process to still lack the mandate and process to consolidate stakeholders, especially underrepresented groups. The latter counteracts attempts by these underrepresented groups themselves to claim agency in the development process, further distorting the values included in AI systems. In the following article, we will outline three prominent, yet harmful AI narratives and provide examples of movements, organisations, or tools working against these.

## The 'Whiteness' of AI and Other Structural Injustices

Most humanoid instances of AI, both in media and in real life are portrayed as 'white' (Cave & Dihal, 2020 [1]). This subtle racialization of anthropomorphised machines is both a representation of the race of the people creating them, where whiteness creates whiteness, and a dissuading barrier for people of colour to break the status-quo of machines not made by them nor for them, perpetuating a "feedback loop" (West, Whittaker, & Crawford., 2019 [2]) or a "cycle of social injustice" (Cave & Dihal, 2020). These injustices do not only stop at race but also affect minorities across a number of diversity axes. Since a homogeneous group of people tends to generalise broadly based on the characteristics represented among them, people and elements that do not fit into the creators' categories are often disadvantaged by the systems created by them (Hollaneck, 2021 [3]; West, Whittaker, & Crawford, 2019). This need to generalise and simplify complex problem spaces erases people's experiences and histories because of experts' reluctance to tackle them, or their lack of awareness of their existence in the first place (Majlesein, 202 [4]) (referred to as the "privilege hazard" by D'ignazio and Klein (2020)[5]).

The extremely complex and interrelated nature of existing biases in AI systems motivated the Leverhulme Centre for the Future of Intelligence [6] to call biases in AI systems a "vicious cycle": The lack of diversity among researchers and practitioners leads to data-based and algorithmic biases, as well as a narrow vision of the interests and values prioritised during system development. These, in turn, create an image and a narrative around AI systems and their design that limit the types of people attempting to become AI researchers and practitioners, further perpetuating existing power imbalances.

***Best Practice Examples: AymurAI***

*AymurAI* [7] is an open-source tool, designed to support criminal justice courts in Latin America in their collection and publication of data about gender-based violence. AymurAI was planned by a female team, illustrating how a non-traditional group of AI creators can envision and build AI tools that serve highly important, but often overlooked use cases.

## Technological Determinism

Technological determinism describes a reductionist understanding of technological innovation. It asserts that society's technology progresses by itself and the rules of efficiency, influencing the societal structures and values around it (Chandler, 1995 [8]). In other words, technological determinism describes "the belief in technology as a key governing force in society" (Smith & Marx, 1994, p. 2 [9]). This theory neglects that technology is shaped by society and that societal factors enable - or prevent - the development of a specific technology (e.g. D'ignazio & Klein, 2020). Instead, technology is regarded as neutral, free of any power structures or bias, merely a tool. This characterises AI systems as unable to disadvantage someone - harm is solely created by its users. Additionally, technological determinism renders efficient high-tech systems unavoidable, depicting the communities around them as passive spectators whose interventions will be unable to change technological advancements.

BigTech companies often harness technological determinism to justify their products. For example, in Mark Zuckerberg's view, Facebook is optimised for "showing people what they think is meaningful" (Metz, 2017 [10]). By implying that the algorithm neutrally reacts to human desires, the company pushes away any responsibility for the detrimental consequences of their platform (ironic, if we consider that their RAI guidelines include 'Accountability & Governance', Pesenti (2022) [11]), leading to an "environment blind[ness]" with regards to relevant socio-cultural facets of technology and its implications (Whitworth and Ahmed, 2014 [12]). Additionally, it neglects the values that the company instils in the algorithm to serve its management goals thus making it inherently biased (Martinho et al., 2021 [13]). Such a perspective is detrimental.

Furthermore, endeavours related to AI are generally framed as a solely technical and abstract pursuit. This leverages technical practitioners as superior 'wizards' whilst diminishing the participation from communities and non-technical experts (D'ignazio & Klein, 2020). However, a solely technical approach creates a socio-technical gap between social requirements and technical designs (Ehsan et al., 2023 [14]). Aside from the obvious disconnect that this gap creates between the technologies created and those using them or being affected by them, it can also lead to serious ethical problems and misuses (Ehsan et al., 2023; Schniderman, 2020 [15]; Selbst, 2019 [16]). This is especially true for AI-based

systems which are considered highly socio-technical (van de Poel, 2020 [17]). The result is a number of gaps between the creators of the AI system, the roles responsible for maintaining its wider infrastructure, those making business decisions, and the people using it and being affected by it.

***Best Practice Example: EU AI Act Stakeholder Consultation***

The EU AI Act is a proposed law on AI regulation by the European Union, the first of its kind (Future of Life Institute, 2022 [18]). During the process of drafting the act, citizens and stakeholders were able to provide feedback on the draft from February to June 2020 via a custom-made platform (European Commission, 2020 [19]). This counteracts the narrative of AI determinism in two ways: Firstly, the EU AI Act is a form of external regulation, designed to guide the further development of AI systems. This is an active stance, contradicting the passive notion of AI determinism. Secondly, it empowers the people that will be users and stakeholders of these systems by giving them a voice in shaping the regulation that in turn will shape the technology of the future.

## AI Race

In recent policy and strategic documents from governments, a strong narrative around the "race for technological superiority" in AI emerged (Cave & Oh´ Eigeartaigh, 2018, p. 36 [20]). It is fueled by the perception that the country establishing leadership will profit from scientific, infrastructural, and economic advantages, i.e. "the winner takes all" (Cave and Oh´ Eigeartaigh 2018). This is enhanced by the conviction that AI advantages can be applied on a vast scale across sectors, initiating a ubiquitous transformation towards increased efficiency (for more details see Cave and Oh´ Eigeartaigh (2018) who systematically analyse potential states of the AI race). Such a narrative comes with several problems: It prioritises speed over thoroughness and reflection. However, most steps to "not only [make] AI more capable, but also [maximise] the societal benefit of AI" (Open Letter of the Future of Life Institute of Life Institute (2015) [21]) - e.g. the iterative involvement of multiple stakeholders, participatory design and careful roll-outs with continuous evaluation - require time and care.

If confronted with ethical issues, the AI Race narrative pushes at best towards the creation of reactionary, passive measures such as guidelines and checklists to enable a "comply and deploy" mindset among practitioners (Crawford & Calo, 2016 [22]). As a result, little active

reflection is practised while creating these systems, leading to harmful practitioner mindsets such as “rejecting practices or downplaying the importance of values or the possible threats of ignoring them" (Manders-Huits & Zimmer, 2009 [23]) or shifting responsibility onto other stakeholders alone (Mitchell, 2020 [24]).

***Best Practice Example: Microsoft’s Judgement Call Cards & MIT’s AI Blindspot Cards***

The card deck ‘Judgement Call Cards’ [25] was developed to support AI practitioners in considering ethical aspects and stakeholder values early in the creation of an AI-based system. The tool tackles the typical ‘move fast and break things’ mentality that focuses solely on technical perspectives. The practitioners are prompted to consider who their system’s stakeholders are and then write reviews for the planned system from their perspectives, representing their various, potentially contradicting values. This counteracts the AI race narrative by allowing practitioners to reflect on the experiences that their AI-based system might provide for different stakeholders, triggering an awareness of the kinds of harms they might cause. The resulting empathy (hopefully) results in increased motivation to identify mitigation strategies for these negative experiences, e.g. to explore potential countermeasures in collaboration with the communities affected.

MIT’s AI Blindspot Cards [26] are another deck of cards for AI practitioners. The cards highlight a number of ‘blindspots’ that AI practitioners are likely to suffer from along the various design phases for AI-based systems. Covering planning, building, deploying and monitoring, each card presents a potential blindspot that might cause harm if overlooked. The cards include potential questions to ask, stakeholders to engage with, and a real-world example illustrating the problems that might occur if the corresponding blindspot is not considered. By highlighting specific stakeholders and illustrating concrete and real harms that specified blindspots might cause, these cards work towards combating the AI race narrative by encouraging practitioners to slow down and reflect on the direct and specific problems they might enable if they do not engage in proactive measures to address potential blindspots before they occur. Similar to Microsoft’s Judgement Call cards, the hope is that solidifying these harms as real and present through the use of named stakeholders or specific examples will motivate practitioners to work actively towards ensuring these harms do not occur.

## Conclusion

Narratives around AI systems shape perceptions about who is in control, who benefits, and what these systems can or should achieve. This article has shed light on three harmful AI narratives that form inaccurate misconceptions among practitioners and the public which, in some cases, lead them to becoming self-fulfilling prophecies: Firstly, the "whiteness of AI", perpetuating the cycle of social injustice and data-based biases that affect minority groups. Secondly, technological determinism that portrays AI systems as neutral tools, absolving companies of responsibility for the detrimental consequences of their platforms. Lastly, the idea of an AI race creates a gap between the required time-intensive socio-technical considerations and the AI development practice.

This work hopes to raise awareness regarding these narratives and how they might harm valuable and meaningful progress towards the creation of human-centred, responsible AI. The motivation is to investigate deeper, systemic problems regarding the creation and perception of AI-based instead of focusing on the technology alone. Through providing examples of how these narratives can be reversed, we hope to inspire more projects that work towards a more diverse, feminist, and democratic future of AI.

## References


[1] Cave, S., Dihal, K. The Whiteness of AI. Philos. Technol. 33, 685–703 (2020). https://doi.org/10.1007/s13347-020-00415-6

[2] West, S., Whittaker, M., Crawford, K. (2019). Discriminating systems: Gender, race, and power (Tech. Rep.). AI Now Institute.

[3] Hollaneck, T. (2021). Design choices in ML systems – cambridge digital humanities' social data school. Retrieved from https://www.youtube.com/watch?v=OvcsaRj6dDo&feature=emb_logo

[4] Majlesein, S. (2021), Building solid experiences on unsolid ground, Presentation at World Interaction Design Day (IxDD) and can be accessed at: https://vimeo.com/618306283.

[5] D'ignazio, C., & Klein, L.F. (2020). Data feminism. MIT press. https://data-feminism.mitpress.mit.edu/

[6] The centre's work can be viewed on their website at: http://lcfi.ac.uk/

[7] Their work can be seen at: https://sites.google.com/view/aymurai-en

[8] Chandler, D. (1995). Technological or media determinism.

[9] Smith, M.R., & Marx, L. (1994). Does technology drive history?: The dilemma of technological determinism. Mit Press.

[10] Metz, C. (2017). Mark Zuckerberg's Answer to a World Divided by Facebook Is More Facebook, https://www.wired.com/2017/02/mark-zuckerbergs-answer-world-divided-facebook-facebook/

[11] Pesenti, J. (2022). Facebook's five pillars of Responsible AI. Retrieved from: https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/

[12] Whitworth, B. and Ahmad, A. (2014), Socio-Technical System Design, Interaction Design Foundation.

[13] Martinho, A., Poulsen, A., Kroesen, M., Chorus, C. (2021). Perspectives about artificial moral agents. AI and Ethics, 1 (4), 477–490. https://doi.org/10.1007/s43681-021-00055-2

[14] Ehsan, U., Saha, K., Choudhury, M. and Riedl, M. (2023), 'Charting the sociotechnical gap in explainable ai: A framework to address the gap in XAI, Proceedings of the ACM on Human-Computer Interaction (34). https://doi.org/10.48550/arXiv.2302.00799

[15] Schneiderman, B. (2020), 'Human-centered artificial intelligence: Three fresh ideas', AIS Transactions on Human-Computer Interaction (THCI) 12(3). https://doi.org/10.17705/1thci.00131

[16] Selbst, A. (2019), 'Accountable algorithmic futures'. Retrieved at: https://points.datasociety.net/building-empirical-research-into-the-future-of-algorithmic-accountability-act-d230183bb826

[17] van de Poel, I. (2020), 'Embedding values in artificial intelligence (ai) systems', Mind and Machines 30, 385–409.

[18] Future of Life Institute (2022), The Artificial Intelligence Act. https://artificialintelligenceact.eu/

[19] European Commission (2020), White Paper on Artificial Intelligence: Public consultation towards a European approach for excellence and trust. https://digital-strategy.ec.europa.eu/en/library/white-paper-artificial-intelligence-public-consultation-towards-european-approach-excellence-and

[20] Cave, S., & ´Oh´Eigeartaigh, S.S. (2018). An AI race for strategic advantage: Rhetoric and risks. Proceedings of the 2018 aaai/acm conference on ai, ethics, and society (p. 36–40). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3278721.3278780

[21] The letter is readable at: https://futureoflife.org/open-letter/open-letter-autonomous-weapons-ai-robotics/

[22] Crawford, K. and Calo, R. (2016), 'There is a blind spot in AI research', Nature 538, 311–313. https://doi.org/10.1038/538311a

[23] Manders-Huits, N. and Zimmer, M. (2009), 'Values and pragmatic action: The challenges of introducing ethical intelligence in technical design

communities', The International Review of Information Ethics 10, 37–44. https://doi.org/10.29173/irie87

[24]Mitchel, M. (2020). Intentional Ignorance is a Value-Laden Choice. Presentation at the PAIR Symposium. Accessible at: https://www.youtube.com/watch?v=TcE6_NPjvuo

[25] Accessible at: https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/judgment call

[26] Accessible at: https://aiblindspot.media.mit.edu/

## Biographies

**Emma Kallina** is a PhD student at the University of Cambridge. Her research focuses on democratising the AI development process through improved stakeholder involvement along the entire lifecycle of AI systems. She believes that everyone who will be impacted by a system should have a say in how it is created.

**Malak Sadek** is a Design Engineering PhD student at Imperial College London. Her work lies at the intersection of AI and design. Specifically, she's interested in creating tools that help collaboratively designing conversational AI that respects stakeholders' values. The aim is to work towards ensuring AI-based systems are better aligned with the values of those involved in making and using them.

# Re-imagining Artificial Intelligence on the African continent

Rachel Yayra Adjoe

## Abstract

The impact of artificial intelligence on Africa must be carefully considered as it spreads throughout the world. Therefore, this essay takes the time to present some potential solutions for issues regarding Africa as a continent, women in Africa, and people living with disabilities in relation to AI. These include investing more in gender research, encouraging women to participate in all stages of the AI development process, increasing the number of research labs and institutes dedicated to AI research, and developing more diverse data sets that contain information on both women and people with disabilities.

## Introduction

Decades from now, if all goes according to plan, autonomous electric vehicles will rule the streets of Africa, reducing carbon emissions on the continent; automatic sign language translators will ensure that the deaf African is no longer excluded from virtual meetings; and plant diseases on the African continent will be detected by the farmer with a single shutter of the camera on a smartphone, giving the rest of humanity more time to focus on what really matters.

When I look at the state of artificial intelligence in Africa, I see a desert characterized by a lack of innovation in AI rather than a jungle overflowing with life and opportunities in AI. From this point, it may be difficult to envisage the AI utopia outlined above, with concerns whirling around in the distance about whether it is conceivable for an Africa wrecked by so much political instability, poverty, and lack of access to technology. My answer to this question is “yes”. I do believe that AI can reach such a level in Africa; however, it would require a considerable amount of effort from Africa and African governments.

The second, and most pressing, concern is how and whether we can achieve this level without harming marginalized groups, particularly women and people with disabilities. In this essay, I explore this question while examining how AI has been used to perpetuate bias and discrimination towards the African continent, women in Africa, and individuals living with disabilities in Africa. I also discuss solutions to these difficulties, first steps which will help us take us toward the AI future that Africa deserves.

## Problems of AI on the African continent

One of the most pressing concerns in Africa relates to the wide use of foreign AI. Even though more and more African developers are creating solutions that fit Africa, these solutions still feel like a drop in the ocean compared to the myriad of foreign AI tools widely available on the market. Examples of these AI tools include self-driving automobiles, personal assistants like Google Assistant, and, more recently, ChatGPT. The proliferation of these tools on the market, unsurprisingly, has had some consequences since these solutions do not work well for Africans. For an Africa definitely no stranger to foreign solutions not working in Africa, the

reason is clear: most of these solutions were not developed with the African or even persons of colour in mind. Therefore, unsurprisingly, they do not work. There have been countless examples of solutions unable to recognize cancer on dark skin or even facial recognition software that completely fails on darker skin. However, an example that springs to mind is from personal experience. On Twitter, I noticed that abusive comments and cyberbullying written in English are mostly hidden from the user or tagged as abusive content. However, an abusive statement written in a local Ghanaian language goes unchecked. This exposes Africans online to abuse and bullying that their western counterparts may be protected from - and illustrates the plight of an African using a tool from a world that does not think about them.

Another problem pertaining to AI in Africa is the lack of quality datasets for machine learning and AI purposes. On the African continent, there is some local data. However, most of these datasets were not collected for machine learning purposes and are most likely in incorrect formats. This is important because the data acquired to be used for a machine learning goal must be deliberate, containing all the necessary independent variables that influence the precise variable that the researcher or developer wishes to predict. The reasons for this shortage are diverse. According to Oyindamola Johnson in her article titled *'Challenges of Access to Data in Africa: A Two-Way Conversation'*, the lack of data in developing countries is due to issues such as a lack of attention to research, poor infrastructure, a lack of funding, and a lack of centralized data banks. Even when data is collected, it is usually poorly sourced and classified, as well as filled with human bias and mistakes from using analog and outdated data collection methods. At other times, the data is collected by national agencies and global NGOs, which do not make the raw data public but only their own truncated analyses. The dearth of large enough datasets has caused some labs to embark on data collection processes to collect the data they need. Researchers in Africa are also often not inclined to share their collected data, as data collection is often tedious, time-consuming, or expensive. For developers that cannot afford to properly go searching and collecting data to develop solutions for Africa, they resort to using foreign datasets that may or may not capture the cultural environment in Africa.

## Problems of AI on African women

AI has also been used to perpetuate harmful stereotypes about women. As society progresses, it aims to eliminate practices and ideas that are not beneficial. Nonetheless, some of these unconscious biases stay with us, seep into our creativity, and taint the solutions we create. An example of this is the fact that most customer service chatbots in Africa are coded as female, or at the very least feminine. In a case study conducted in Nigeria by Borokini et al. in 2023, it was found that out of 10 banks that had deployed chatbots, 7 of them were female-gendered on account of their names, avatars, and descriptors.

In a UNESCO report, *'I'd Blush if I Could'*, UNESCO detailed the potential negative impacts of chatbots on society's perceptions of gender. According to the paper, the spread of female-gendered conversational agents was driven mostly by customer choice and a non-critical assessment of product development decisions by product teams, all of which could entrench and perpetuate biases against women today. Long-standing social prejudices and assumptions about women may also have informed the creation and building of female-gendered chat bots, which in turn may continue and cement perceptions about social norms about women's capacity and nature. For instance, women have long been associated with servile or submissive roles; therefore, the use of female or feminine attributes on chatbots further and unconsciously promotes this bias. Such a gendered categorization of women may not only reinforce gender stereotypes but may also further sexualize women as female chat bots may be utilized to attract more consumers or increase an institution's profit through their soft, feminine voices and appearance. (Borokini et al., 2023) This raises questions concerning how ethical the practice is and how it may contribute to the disempowerment of women.

Another problem with AI that affects women on the African continent concerns the bias in the datasets that are available. Due to a violently misogynistic and sexist past, resulting in a lack of access to economic products and technology for African women, data collected may feature a gross overrepresentation or underrepresentation of women. Because AI is not neutral, if caution is not taken, it can reflect biases present in a given environment, including those related to race, gender, and sexual orientation. (Belenguer, 2022) This implies that any AI solution developed using a biased dataset will only have negative effects on women. Additionally, the designed solution may not precisely address their needs. A case study

conducted by Shamira Ahmed in her 2021 paper *'A Gender Perspective on the Use of Artificial Intelligence in the African FinTech Ecosystem: Case Studies from South Africa, Kenya, Nigeria, and Ghana'*, reveals that women are generally overrepresented in the underserved population in terms of bank account use and access, due to the lack of access in the past and therefore perpetuating a perfidious cycle of exclusion; women are also more likely to be without a smartphone and/or mobile internet access for the same reason. Therefore, solutions developed using these types of datasets will not work for women and illustrate the effects of heavily skewed and biased data on women in Africa.

Some other types of harm that AI may cause may come in more shocking and deliberate ways. One of these ways is through the popularization of image generation AI and DeepFakes. In late 2017, an anonymous individual using the alias "deepfakes" posted explicit videos to a social news website and forum on Reddit, claiming to be Taylor Swift, Scarlett Johansson, Aubrey Plaza, Gal Gadot, and Maisie Williams. (Gardiner, 2019) In a continent that is plagued by misogyny and has a high incidence of image-based sexual assault, this seems concerning, as it has become easier to generate fake but realistic-looking suggestive photos of real women. While there are many ways that videos and other content can be changed to falsely depict people and events, DeepFake technology is especially dangerous because it allows users to produce realistic images of the highest caliber and edit video and audio realistically. This is a new form of sexual harassment that takes power away from women to control their own sexual identity and is intended to humiliate and devalue a person. In the 2020 paper *'Deepfakes and Domestic Violence: Perpetrating Intimate Partner Abuse Using Video Technology'*, by Kweillin T. Lucas, it is also noted that deepfake technology can pose a significant risk for victims of domestic violence as perpetrators can use this form of image-based sexual assault to continue abusing their victims. This has caused many women to have some anxiety about posting pictures on the internet and has painted AI in a bad light for most of them.

## Problems of AI on people living with disability

Another problem with AI in Africa that hinders its goal of being beneficial for all concerns the treatment of the disabled. Even though more scientists are creating solutions for the disabled in Africa, especially in the health domain, AI is still discriminating against the disabled in a more insidious way.(Whittaker et al., 2019) There is not much research into the

negative impacts of local AI on the disabled in Africa, yet despite this, a case can still be made that as foreign AI tools infiltrate Africa due to their ease of use and availability, firms in Africa start using some of these foreign tools that have some discriminatory bias against people living with disabilities. Furthermore, the same biases explored in the following paragraphs may be perpetrated by our very own AI in the future as the societal attitude toward disability in Africa is still extremely negative.

One example of how AI may negatively affect the disabled in Africa is through a job screening application that incorporates AI. Systematic bias can arise if the data used to train a model contains human decisions that are biased, and the bias is passed on to the learned model. For example, if recruiters continually reject applications from students with disabilities, a model trained on that data will repeat the same behavior. (Trewin, 2019) A scarier example may be in the case of HireVue, a virtual interview software application. According to experts, this AI software uses a candidate's tone of voice, gestures, and facial expressions to determine whether they are a good fit. These criteria are also areas of weakness and divergence related to neurodivergent disorders, including Tourette syndrome and autism spectrum disorder. (Moss,2021) Therefore, for job applicants who have these characteristics, the odds are already stacked against them.

One reason why discrimination against the disabled is so high in AI algorithms is due to a lack of diversity in datasets. Nakamura offers an example of AI not being able to understand the words of a speech-impaired person as being seen as normal, as other humans are also not able to do the same. However, misrecognition of disabled people can have serious consequences. This would include wheelchair users being run over by car drivers who do not recognize them as humans a problem embedded in the datasets used to train automobile vision systems. (Nakamura, 2019). This is the way continuous bias shows up in AI to discriminate against the disabled.

## Protection of Africa as a continent

No matter how difficult these issues appear to be, humans always seem to manage to overcome the challenges.. I am confident that this will be the case with the challenges that we confront with AI on the African continent. At this point, one of the most important

problems to solve concerns the lack of access to good-quality datasets. The establishment of robust, modern national statistical systems capable of properly collecting and preserving critical data requires governments to increase funding for national statistical agencies. (Research ICT Africa, 2020) Also, as a means of fostering togetherness toward common goals, research labs and institutions should make their raw datasets publicly available by publishing them in open-access journals. This will speed up the development of AI and solutions that reflect and benefit Africa as a whole. Moreover, proper data collection methods should be used when collecting data on the African continent to ensure the integrity of the collected data using the help of subject matter experts when necessary. Finally, tech solutions should always add a data collection layer to their function in a way that protects the individual users' rights to privacy.

Furthermore, addressing the scarcity of high-quality datasets will almost certainly alleviate the challenges associated with the popularity of foreign AI technologies. This is because more developers will be motivated to create AI solutions as they will have access to more data to create tools that are appropriate for the continent. To fight the prevalence of foreign AI, countries are being urged to promote STEM education and increase financing for AI research centers. This will provide scientists with the resources they need to develop the best solutions.

## African women

To not only protect but empower women and tackle gender bias existing in data, governments and policymakers are urged to invest more in gender research so we have more insight into how everything affects people of all genders and from all walks of life. Universities and labs should receive funding and grants to pursue gender-related research. This includes research on the potential for AI to enable the transcendence of gender identities and ways to mitigate gender discrimination. Gender studies and how gender influences life on the African continent are not widely studied. This contributes to our lack of precise knowledge of how sexist or ableist our data is, and our lack of answers to gender-related issues relevant to our region. I think Africa should be looking in this direction since relying entirely on international gender research to guide domestic policy could result in poor choices, wrong priorities, and ineffective initiatives. While some of the findings from studies conducted in the Global North may be applicable in Africa, the distributions of gender

and power there may differ. (Research ICT Africa, 2020) Governments are also encouraged to supportthe collection of gender-disaggregated data. Gender-disaggregated data is any data on individuals broken down by sex and collected and tabulated separately for men and women. These types of datasets usually allow for the measurement of differences between women and men on various social and economic grounds.

Last but not least, we should be concerned about and think critically about all our choices and ideas during every stage of the development of AI solutions and assess whether our decisions are a result of unconscious bias or not. A way to ensure this is to encourage the inclusion of women at all stages of an AI solution's development. History has shown that society views men as the default, and thus when solutions are built, other types of people are not necessarily considered. Therefore, a team that is completely male signals that the solutions built will mostly cater to men. An example is the predominance of female voice assistants, which may lie in the fact that they are designed by workforces that are overwhelmingly male, and weren't enough women to talk about how offensive creating female voice assistants might be. In the end, the unconscious bias the men held prevailed. Even during the deployment stages, women should be included, and developers must continuously assess the quality of their solutions for women and historically marginalized groups. This may take a long time, as women are usually discouraged from getting into STEM. However, as a starting point, it would be beneficial to developers of AI solutions at all levels to be educated and responsive to women's issues and the effects decisions in the lab or in product development have on them.

## People living with disabilities

The best and most effective place to start the fight against discrimination against persons living with disabilities is through including nondiscrimination throughout the AI development process. It would be beneficial for creators of AI solutions at all levels to become informed about disabilities, acquire a sense of empathy for those who live with them, and take these individuals' needs into account. It would also be beneficial to encourage research into how the typical African is affected by disability and whether the solutions we develop cater to them. Additionally, much like with the problems with women, it is crucial that the researchers continually evaluate the caliber of their work with people with disabilities. Because a significant section of the population is impaired, even if this may be challenging

as disabled individuals are not a homogeneous group, it is worthwhile to pursue. Instead of creating barriers that prevent disabled people from gaining employment, we should find a way to get them into teams at all levels of the AI development phase. Additionally, developers should use existing techniques to test for bias and mitigate bias throughout the machine learning pipeline.(Trewin,2019) Data on people living with disabilities should be included in data collection processes to ensure that machine learning algorithms learn to recognize and cater to people living with disabilities. Through these measures, Africa can start to win the war against inequalities against people with disabilities.

The goal of this essay was to identify AI challenges affecting Africa and marginalized African people and examine potential solutions to these problems. Some of these solutions include investing more in gender research, encouraging women to participate in all stages of the AI development process, increasing the number of research labs and institutes dedicated to artificial intelligence research, and developing more diverse data sets that contain information on women and people with disabilities. Based on the analysis provided, Africa has reason to be optimistic. However, achieving the vision we imagine necessitates a concerted and purposeful effort.

# References


Ahmed, S. (2021). A Gender perspective on the use of Artificial Intelligence in the African FinTech Ecosystem: Case studies from South Africa, Kenya, Nigeria, and Ghana. In: International Telecommunications Society (ITS) 23rd Biennial Conference - Digital societies and industrial transformations: Policies, markets, and technologies in a post-Covid world-

Belenguer, L. (2022). AI bias: exploring discriminatory algorithmic decision-making models and
the application of possible machine-centric solutions adapted from the pharmaceutical industry. AI and Ethics, 2. doi:https://doi.org/10.1007/s43681-022-00138-8.

Borokini, F., Wakunuma, K. and Akintoye, S. (2023). The Use of Gendered Chatbots in Nigeria: Critical Perspectives. Social and Cultural Studies of Robots and AI, pp.119–139.doi:https://doi.org/10.1007/978-3-031-08215-3_6.

EQUALS Skills Coalition , U. (2019). I'd blush if I could: closing gender divides in digital skills through education. doi:https://doi.org/10.54675/rapc9356.

Gardiner, N. (2019). Facial re-enactment, speech synthesis and the rise of the Deepfake. Theses : Honours. [online] Available at: https://ro.ecu.edu.au/theses_hons/1530 [Accessed 20 May 2023].

Johnson, O. (2021). Challenges of Access to Data in Africa: A two-way conversation. [online] Enyenaweh Research. Available at: https://www.enyenawehafrica.org/post/challenges-of-access-to-data-in-africa-a-two-way-conversation [Accessed 20 May 2023].

Lucas, K.T. (2022). Deepfakes and Domestic Violence: Perpetrating Intimate Partner Abuse Using Video Technology. Victims & Offenders, 17(5), pp.647–659. doi:https://doi.org/10.1080/15564886.2022.2036656.

Moss, H. (2021). Screened Out Onscreen: Disability Discrimination, Hiring Bias, and Artificial Intelligence. SSRN Electronic Journal.doi:https://doi.org/10.2139/ssrn.3906300.

Nakamura, K. (2019). My Algorithms Have Determined You're Not Human. The 21st International ACM SIGACCESS Conference on Computers and Accessibility. doi:https://doi.org/10.1145/3308561.3353812.

Research ICT Africa (2020). An African perspective on gender and artificial intelligence needs African data and research.

Trewin, S. (2018). AI Fairness for People with Disabilities: Point of View. arXiv (Cornell University). doi:https://doi.org/10.48550/arxiv.1811.10670

Whittaker, M., Alper, M., Bennett, C.L., Hendren, S., Kaziunas, L., Mills, M., Morris, M.R.,Rankin, J., Rogers, E., Salas, M. and West, S.M. (2019). Disability, bias, and AI. AI Now Institute

## Biography

Rachel Yayra Adjoe is a Research Assistant at the Responsible Artificial Intelligence Laboratory (RAIL) in Ghana. She holds a Bachelor's Degree in Electrical and Electronics Engineering from KNUST. She is proficient in Python and Javascript and enjoys using her skills to contribute to the exciting technological advances that occur every day at RAIL, KNUST. Rachel is particularly interested in computer vision, electric vehicles, and recommender systems, and she enjoys building them.

# Feminist design framework for envisioning gender-inclusive ride-hailing sector: Perspectives from India

Pallavi Bansal, Payal Arora, and Usha Raman

December 2023

## Abstract

The rapid rise of the ride-hailing sector is opening new avenues for women to participate in the Indian platform economy. However, women continue to be underrepresented as transport providers as much of the debate in the ride-hailing sector centers around women as transport users, not as transport providers. In this paper, we critically explore the strategies to enhance the participation of women drivers in the ride-hailing sector by conducting in-depth qualitative interviews with the stakeholders such as women taxi drivers, platform providers, NGOs, and policy and research support groups in India. The interviews are supported by analyzing documents related to government policies, platform companies, organizations, and the policies mentioned by the participants. This study builds on the Feminist HCI (Human-Computer Interaction) theory that proposes technologies should integrate feminist values in the design (Bardzell, 2010). The paper advocates for redesigning the ride-hailing platforms in alignment with marginalized actors.

## Introduction

The debut of the ride-hailing platforms[1] invited a mixed reaction from scholars and activists. On the one hand, supporters believed that the platforms challenged the traditional model of the taxi industry and enabled a fast, flexible, low-cost, and user-friendly mobility in urban areas for middle-class consumers. On the other hand, the critics questioned the claims of the platforms to monetize underused resources, reduce carbon footprint and provide equal economic and social opportunities to marginalized people including women from low-income backgrounds (Schor, 2016; Shaheen et al., 2016; Zanoni, 2019). Despite the criticism, these platforms expanded quickly around the world with Uber, Lyft, Didi Chuxing, Ola, and Cabify dominating the markets. Currently, there are around 90 ride-hailing and ride-sharing platforms globally (RideGuru, 2021), with Uber being the largest ride-hailing platform in the world (Curry, 2021). However, women continue to be underrepresented as transport providers in this sector with only 14% of female drivers in Uber worldwide (Srivastava, 2019). Lack of access to public spaces due to socio-cultural norms in many parts of the world restricts women's freedom of mobility and limits their access to various job opportunities. The negative stereotype of inferior women drivers further exacerbates the situation of keeping women in the home (Berger, 1986). Hence, for many women, driving is symbolic of and synonymous to freedom – both social and mobile. This qualitative study seeks to increase their access to safe and secure mobility by addressing gender-based concerns in the ride-hailing sector.

In the last decade, a considerable amount of research has been done in the ride-hailing sector of the platform economy. Researchers believe the expansion of this on-demand economy allow for flexible sources of income outside conventional (Graham et al., 2020; Heeks, 2019), but at the same time, foster a new class of precarious workers, also termed as "Cybertariat" (Huws, 2015, para 1). A growing body of academic literature has been investigating the working conditions of the resultant digital labor globally (Chen, 2017; Fielbaum & Tirachini, 2020; Surie, 2018), where platforms immunize themselves (van Doorn, 2017), act as 'intermediaries' and do not classify these workers as 'employees' (Cherry, 2016; Choudary, 2018), regulate the work practices (Beerepoot & Lambregts, 2018), create

[1] The usage of the terms "ride-hailing", "ride-sharing" and "ride-sourcing" have been debatable with some researchers and even the platforms using these interchangeably. Though, the term "ride-hailing" gained more public acceptance after the Associated Press declared the usage of the term "ride-sharing" as a misnomer, and adopted "ride-hailing" in its stylebook. To maintain consistency, we will use the term 'ride-hailing' for this paper.

information and power asymmetry (Schmidt, 2017), violate fundamental labor rights (De Stefano & Aloisi, 2018), indulge in algorithmic management and control (Lee et al., 2015; Rosenblat & Stark, 2016), and foster economic exclusion due to discrimination and occupational segregation (Graham et al., 2017).

However, there is a noticeable gender and platform skew since the majority of the studies in this domain focus on the driving experiences of male cab drivers and lack diversified data from various platforms. Further, wherever there is a focus on gender discrimination, it is in the backdrop of the overall gig or on-demand economy (Hunt et al., 2019; Kasliwal, 2020). There is a dearth of research on the gendered experiences and perspectives in the ride-hailing sector in the Global South context, where the issues of class domination (socio-economic), patriarchal structures, institutions, and gendered norms intersect with each other. In fact, the automation of this sector could reproduce and perhaps amplify the gendered discriminatory practices that have long pervaded these contexts as evident in the scholarship on legacy structural exclusions, manifesting in systems of algorithmic bias and oppression (Hanrahan et al., 2017; Muller, 2020; Page et al., 2017).

We move beyond the normative borders of the Global North and foreground the feminist ideology and perspective to global development and design that would consider the lived experiences of a diverse range of users (P. Arora & Raman, n.d.). For sustainable change in this sector, we use feminist approaches where we foreground the female taxi drivers and women-focused taxi platforms and critically engage with their concerns and practices. The paper addresses the gap in translating the Feminist HCI (Human-Computer Interaction) theory (Bardzell, 2010) into practice by applying some of the principles such as participation and pluralism. Rather than “establish[ing] an objective, distant, and scientific relationship with subjects” (Bardzell, 2010, p. 1306), we sought to collaborate with the various stakeholders and value participatory processes. We imbibe Bardzell’s quality of pluralism to “nurture the marginal,” (p. 1305) in our approach and analysis. Thus, this study helps in bringing forth the otherwise neglected sector and advocates for redesigning the ride-hailing platforms in alignment with the marginalized actors. This paper is a part of a larger study on the Future of Work and digital labor platforms in the Global South.

## Ride-hailing sector in India

According to IFC and Uber Technologies (2018), "Ride-hailing services—sometimes called transportation network companies—digitally connect the driver of a car or other vehicle with a user, generally via an app but sometimes using a website" (p. 9). In India, the government classifies such companies as 'aggregators'— digital mediators or marketplaces connecting passengers and drivers for the purpose of transportation (Business Today, 2020). In India, some of the key players acting as 'aggregators' in the shared mobility space are Ola, Uber, BlaBlaCar, Rapido, Meru, Wheelstreet, Vogo, Bounce, Zoomcar, Yulu, and Revv, which offer different services such as ride sharing, ride hailing, car rental, car sharing, bus/shuttle and two-wheeler sharing (Frost & Sullivan, 2019; Prescient & Strategic Intelligence, 2021). This study focuses on the ride-hailing model where passengers book a taxi either through apps or websites.

The Indian government has acknowledged that these platforms create work, provide entrepreneurial opportunities, and contribute towards skill development of laborers. In 2016, the government asked Uber and Ola to train one lakh (0.1 million) commercial drivers annually, especially women, and set up driver training institutes as part of the 'Skill India' initiative (R. Arora, 2016). Similarly, other schemes such as the MUDRA loan scheme, provides loans to buy commercial vehicles for livelihood purposes, and the Stand-Up India scheme facilitate loans between INR 10 lakh (USD 13,400) and INR 1 crore (USD 1,34,000) to socially underserved members of the scheduled caste or tribe community, or a women entrepreneur, to start any greenfield enterprise[2] (Prabhat et al., 2019; Ramachandran & Raman, 2021). A few Indian state governments such as – Telangana Government introduced the 'Driver Empowerment Programme' to provide financial assistance to minority drivers belonging to backward classes (A. Y. Khan, 2020); the West Bengal government under the Gatidhara scheme aided jobless youth to buy vehicles for commercial use (Mitra, 2019); and the Punjab government under the 'Apni Gaddi Apna Rozgar' scheme decided to double the subsidy support for purchasing commercial vehicles post-coronavirus market recession (Sharma, 2020). However, the problem with these initiatives is that they are few and far in between. They also lack proper channels of communication and are cumbersome in nature. For instance, most drivers in a study conducted by Prabhat et al.(2019), preferred taking loans

[2] In this context, green field signifies the first time venture of the beneficiary in the manufacturing, services, agri-allied activities or the trading sector (Stand-Up India Scheme Features, n.d.)

from corporate lenders to prevent themselves from painful documentation and identification process.

## Women drivers in India's ride-hailing sector

Both the mainstream ride-hailing platform companies in India, Ola and Uber, have not quoted the exact number of female drivers working with the platforms in any official reports. In the earlier mentioned report of IFC and Uber (2018), India was excluded from their data due to a low sample size of women drivers as there were only eight active drivers with Uber at the time. In 2018, Ola, said the number of female-driver partners rose by 40% in every quarter for them (Sridharan, 2018).

The potential benefits of the ride-hailing sector for women workers are multifold. First, these platforms help bridge the gendered digital divide (Florito et al., 2018) by offering smartphones and imparting the necessary skill-based training to potential workers (Aneja & Shridhar, 2019). Second, they allow flexibility to determine one's own hours and decide the amount of work to be taken up, which is one of the major enablers for women as they continue to perform family and household duties (Kasliwal, 2020). Third, they help in fostering "inclusivity" and creating "entrepreneurship" for "people without the right degree, ethnic minorities or from poor neighborhoods" (De Morgen, 2016, as cited in Zanoni, 2019, p. 149). Fourth, this sector helps in generating new and non-traditional sources of income and asset access and ownership for women (IFC and Uber Technologies, 2018). Fifth, ride-hailing platforms enhance freedom of movement and provide greater sense of independence to both women riders and drivers. A survey conducted by IFC and Uber Technologies (2018) reveals many women like the social aspect of driving and would like to break the cultural barriers and resistance in relation to this profession.

However, women's participation in the ride-hailing sector is affected by a range of barriers starting from social and legal to financial and safety concerns. Women are structurally and socially more vulnerable to external shocks and hence often face double discrimination in this sector. The gender- related cultural norms intersecting with class discrimination restrict women to take up taxi driving as a profession. For instance, it is more difficult for a Dalit Muslim woman to get trained and employed as a taxi driver as people from these communities are expected to take up manual scavenging or toilet cleaning jobs (Chugh,

2016). To mitigate some of these barriers, a few platform companies in India launched gender-segregated transport services with different operational models – from for-women-by-women"[3](Meru Eve, Taxshe, Sakha Cabs, Koala Kabs, Viira Cabs, Pink Ola, Pink Taxi, and GoPink Cabs) to "a women-for-all"[4] (She Taxi 2020, Kudumbashree, Priyadarshini Taxi). They work either by straddling the old call-a-cab model and ride-hailing: passengers must book the cab via a phone number or through the company's website and the cabs are equipped with a GPS tracker and safety button or a completely app-based model for bookings.

## Methodology

Feminist HCI scholars propose integrating values such as agency, fulfilment, identity, equity, empowerment, and social justice into the technological systems (Bardzell, 2010). Built on the feminist standpoint theory, it advocates for the use of women's viewpoints and experiences as an alternative point of departure for social science research. The marginalization of women and their social disadvantages can be channelized as resources and be turned into our scientific and political advantage. Our methodology focusing on the women taxi drivers help us understand both the perspectives – 'female' drivers in a male-dominated segment, and the dynamics of 'platform drivers'. Similarly, centering women-focused taxi initiatives help in uncovering the challenges faced by platforms who are marginalized in terms of the resources. Our work represents a "generative contribution" that "involve the use of feminist approaches explicitly in decision-making and design process to generate new design insights and influence the design process tangibly" (Bardzell, 2010, p. 1308). Hence, this paper can prove to be an essential guide for policymakers, platform providers, and designers in the digital platform economy.

One of the prominently discussed methodologies for policymaking and collaborative design process is the stakeholder approach. This approach helps in the identification of the relevant actors in relation to the issue being addressed and uncovers the perspectives and experiences of different stakeholder groups while understanding connection between these networks. However, there is only a feeling of equality amongst the different stakeholder groups as it often deems invisible the marginalized groups that are already fraught with inequality and precarity. This makes the stakeholder approach a rather reductionist one in

[3] The platforms register only women drivers who are matched with women passengers and/or children and senior citizens
[4] The platforms register only women drivers, but they serve both female and male passengers

which “shadows of the context” are not considered (Eskerod and Larsen, 2018, p.1). Focusing on the perspectives and lived experiences of the marginalized stakeholder groups can bring this shadow conceptual information to visibility i.e., already considering the context within which these products and policies emerge. This could be cultural stigma, discrimination, and preconceived stereotypes.

Thus, using a combination of feminist and stakeholder approaches, we bring forth the concerns and experiences of people who are most affected whenever policymaking and designing happens in vacuum. In doing so, we take a critical stance and understand that it is an interpretation of the interpretation. We also acknowledge our positions as people belonging to a certain class, caste, and gender. To mitigate these challenges, we have represented most of the data through direct quotes, and these quotes are not modified to correct grammar or colloquial language and slangs.

***Data Collection***

11 semi-structured in-depth qualitative interviews of vested stakeholders such as four female taxi drivers working/worked for different platforms in various cities of India, founder of Priyadarshini Taxi, founder of Koala Kabs, Director COO of Taxshe, Programme Director of Azad Foundation, Project Associate with The Gender Park, and two researchers (together) of Ola Mobility Institute (OMI) were conducted (details in Table I, Appendix). Centralizing the theme of women drivers and women-focused taxi initiatives, an attempt was made to include random samples from different strata or groups such as female taxi drivers (both currently working and non-working), popular women-focused taxi platforms, closed or shut-down women-focused taxi platform, re-launched and government supported taxi initiatives, NGOs and support groups, and policy research organizations and think tanks.

The interviews were supported by carrying out a document analysis of over 50 documents related to the government policies, platform companies, organizations and the policies and references mentioned by the participants. These documents included annual reports, mission and vision statements, research reports, company blogs, media articles (news stories, features, opinion columns, and editorials), and a few case studies on the platforms. Most of these were downloaded from the official websites of the companies by visiting their ‘about us’, ‘latest news’, ‘blogs’, and ‘annual reports’ sections. The remaining were accessed by conducting a Google search and using keywords related to the companies, organizations,

and government policies in context to women taxi drivers. The document selection was based on the "relevance of documents to the research problem and purpose" (see Bowen, 2009, p. 33).

### *Data Analysis*

A hybrid approach combining technique of inductive and deductive thematic analysis (Fereday & Muir Cochrane, 2006) was used. Inductive TA helped in coding from the data based on the interview participants' responses and experiences, and deductive TA helped in drawing on the theoretical constructs from feminist HCI scholarship as explained in the earlier sections. After the interviews were transcribed and documents collated, the transcripts and documents were read simultaneously in entirety to recognize the voices and views of the people and understand the whole picture. Then, initial codes were generated upon reading every word and sentence of the interview transcripts and documents. Some of these elemental codes included flexibility, agency, upskilling, car loans, perception, access, safety, clean toilets, technology, tracking, feedback, data privacy, and app design. The data was scrutinized and compared with codes in order to organize ideas and concepts that seemed to cluster together. This further helped in refining the codes to include conceptual information, for instance, perception changed to perception of women taxi drivers and perception of the profession. Similarly, access was refined to include concepts related to agency, awareness, training, and public interaction. These refined code clusters that seemed to share unifying characteristics were put under substantive themes, and then these themes were compared across interview transcripts and data from documents.

## Results and Discussion

The below framework serves as device to make sense of the data from the interview transcripts, previously described documents, and establish connections between various other related empirical and theoretical studies. This five-point framework could serve as a conceptual design for inclusive labor reforms in the platform economy.

### 1. *Appropriate gendered framing and positionality*

While the perception of honourable and decent work prevails for women in the beauty segment of the platform economy (IT for Change, forthcoming as cited in Zainab, n.d.), interviewees reported negative perception towards the participation of women as transport providers. Several interviewees pointed out how women driving on the road as a 'taxi driver' is either viewed as 'claiming male dominated public spaces' or a form of 'desperation' and not 'choice'. Former platform driver, Malini Tyagi, elucidates these points:

> *[Certain people] want to take advantage of the fact that they have landed in a situation where there is a women driver, and if they can have their way with this person, because, of course, you know, the woman who is driving the car is not doing it out of her own choice. She's doing it out of compulsion and will obviously be facing financial issues and will do anything to make money.(Malini Tyagi)*

Sushil Shroff, Director COO, Taxshe further elaborates that the profession does not appeal to lower middle and upper middle-class women due to the element of 'blue-collar' work and the perception of such women as 'cheap' and 'available':

> *Cabbies were like the blue-collar job, and the general tendency in India is that if a woman is driving, then she is a cheap and available woman, because anyone can sit in her car and talk anything to her, and she'll have to take it.*

The situation signals further vulnerability of female drivers, who are "doubly oppressed by their status as women in Indian society, and their status as drivers in a professional hierarchy that does not seem to place a lot of value on service jobs" (Researchers, OMI). The framing and branding by the platform companies plays an important role in this scenario. For instance, women-only driving platform, Taxshe, has positioned female drivers as 'alternate moms' who drive women and kids safely to their destinations, and have labelled them as 'Roo', inspired from the animal 'Kangaroo'. This led to the parents trusting women drivers for ferrying their children to school and agreeing for paying higher prices for the 'premium' and 'rare' service – being driven by women.

These experiences and quotes indicate the need for careful framing and positioning of female drivers to challenge the traditional mindset and view platform driving as a viable livelihood choice for women.

### 2. *Outreach, mobilization, and characterizing target population*

Platform representatives and support groups expressed the need for adopting a nuanced approach for outreach, mobilizing, and appropriately characterizing those women who could be potential participants in the ride-hailing sector. The majority of women drivers in this sector belong to the marginalized section of society and outreach and mobilization efforts of NGOs like Azad Foundation and Neeva Foundation help in building confidence and establishing trust among potential women drivers and their families. Once this is done, another challenge faced for recruiting women is the higher percentage of migrant populations who lack proper documentation. Amrita Gupta, Azad Foundation, says 85% to 90% of women they engage with are migrants who do not have the identity proofs or necessary documents for obtaining the driving licenses. Their annual report further elucidates the issue in context of gender:

> *The basic documents required for a learner's license are proof of address and proof of age. Many women do not have a birth certificate. Proof of address is not available as household arrangements are set under names of the men in the households. Young girls who attend schools are rarely encouraged to keep their documents safely, as families do not have professional aspirations for them. (Annual Report 2014-15, Azad Foundation, p. 8)*

After the outreach and mobilization activities, another crucial aspect is identifying the core characteristics of this target group and devising adequate training programmes as per their need assessment. Taxshe, whose main segment consists of women above the age of 35, explains the dynamics as follows:

> *Most of our women are above the age of 35...Now, these are uneducated women who have a maximum of seventh standard or eighth standard of education in their villages, not even in a proper school in gram panchayats, and all they have done is this education. And surprisingly all of them have left education when in seventh standard or eighth standard because it is the time when girls become mature. And in their villages and all they have never had washroom*

*facilities in their schools, let alone at the time of their periods. So, they simply drop out of school, they are just pulled out of school as simple as that. (Sunil Shroff)*

Neeva Foundation's website elaborates this group as "women from financially weak segments, single mothers, burn victims, rape and abuse victims, college students, home makers." Hence, while formulating policies, it is essential to consider the core character of this population and the implications on access and ability.

### 3. *Broadening access*

While the Fairwork Principles discuss fair conditions that ensure work being carried out in a healthy and safe environment (Graham et al., 2020), participants argue that it is also essential to recognize multiple forms of 'access' that can be facilitated by platform companies, support groups or NGOs, and workers' families to carry out this work:

***Access to transformative training and upskilling:*** Sunil Shroff expresses need for both technical and non-technical training in context to the marginalized resource poor women, who are at the intersection of caste, class, and different identities. He says that Taxshe's model consists of 100 hours of driving training and an additional 70-80 hours of imparting other knowledge and skills, which are clubbed under 'soft skills' such as English speaking, legal rights, financial literacy, first aid, self-defence, sexual and reproductive health, grooming, work readiness and reporting, and communication skills.

In contrast, the one-day training programs of mainstream platforms are sometimes even inadequate for the male drivers. Malini Tyagi, who was also a fleet owner with Ola, elucidates how the drivers used to contact her often to understand the app and payment features. Hence, the interview participants call for a transformative or at least an extensive training module, on job upskilling opportunities in line with the changing market and to meet the demands of digitization and automation for retaining women in this sector.

***Access as awareness:*** NGOs like Azad Foundation insist on self-defence training for prospective women drivers that moves beyond just basic safety training. The idea is to make women aware of their surroundings and take prompt action in case required:

> *The idea [of self-defense] is the awareness. The awareness of how to behave, the awareness of what to say and what not to say. For example...there is a girl who said that when I am out on the road at night, I am aware that I keep my eyes around. Looking at people who might be following me. In case I see somebody following me, I quickly check the nearest police station and drive my car towards the police station. So, it's like awareness of your surroundings, awareness of what to do. (Amrita Gupta)*

***Access to interacting with the public:*** Women drivers Malini Tyagi and Maya Sharma exuded confidence in terms of handling the passengers while driving, however Sunita Mishra and Madhuri expressed concerns about the lack of confidence in terms of interacting with the public. The long gendered issue of women being often relegated to the domestic sphere is evident in Madhuri's quote, "We were staying inside home, we were not going outside for jobs, so we were not having any contact with public." In this scenario, the soft-skill training accompanied with 'confidence-building' played an important role in Sunita's life – a widow with two daughters.

***Access is agency:*** Interview participants define 'agency' in terms of 'flexibility' and 'choice' afforded to women – to be able to walk out of the extensive training programmes and rejoin at any point; to decide which passenger (male or female) to pick; to opt to work during morning, afternoon, or even late at night; and to decide the number of hours they want to spend on the road working in the 'flexible' platform economy.

With context to 'walkouts', Amrita Gupta elaborates agency as follows:

> *But the idea is to give flexibility... a woman has her care work responsibility, she will be the first one to be pulled out of the training, when there is a bereavement in the family, or there is a sickness in the family...We say walkouts, because at this point, the woman is being given multiple opportunities to come back. We go back to the families, we negotiate with their families and after that if she has taken the decision to walk out, that's her own agency. She has chosen it, right.*

With context to work hours, agency is exemplified by the representatives of Taxshe and Priyadarshini Taxi. They stressed the need to facilitate women to pick the desired time slot in lieu of their care responsibilities. "Minimum of nine to ten hours you have to work,

minimum. Which hours, morning, evening, afternoon, late night, is the call of the woman," says Susieben Shah. Sushil Shroff adds:

> *We have flexible timings. And they can also divide the shift into two parts, four hours in the morning and four hours in the afternoon. So that they can go back to their home, school for the kids, do some work and yet be on the job and still make money for the family.*

With context to deciding which passenger (male or female) to pick, the Indian platform companies do not have that option as of now. However, ride-hailing company '99' in Brazil and Uber's 'Women Rider Preference' feature in Saudi Arabia and Brazil allow women drivers to "elect to serve either men or women passengers or opt in to serve only women passengers at any time" (IFC, 2020, p.3).

Thus, there is a need to identify various kinds of access that should be facilitated for women drivers in order to bridge the gendered gaps and encourage more women to take up driving as a profession.

### *4. Gender-sensitive sustainability*

The data suggest being responsive to gender and focusing on the sustainability of women-centric platforms and professional driving for women in the following ways:

***Feminist funding:*** One of the main challenges faced by women-only taxi initiatives is that of funding or finding the investors who are supportive of women entrepreneurs. Sushil Shroff elaborates:

> *In the investment segment, woman is not considered as the right bet to put your money...So anywhere, you will find that if there's a woman entrepreneur who's heading the thing, she will have a lot of difficulty in finding the right funding partners, in finding the right technology partners, and all that.*

Amrita Gupta highlights the lack of support from the corporates and the government for this kind of unique model that requires extensive training and handholding which may go beyond three months. Further, the funding seems to be a major issue for this asset-heavy

model that at least two platform representatives reported extreme difficulties during the pandemic, with one of them having to close the business.

**Gender-responsive infrastructure:** A lack of gender-responsive infrastructure is another concern raised by the interview participants that hinders the ability of women to be on the road. Driver Malini Tyagi expressed inconvenience in terms of 1) locating a nearby washroom and 2) finding a clean washroom. To overcome this challenge, Taxshe has mandated the schools and clients to allow women drivers to use washrooms inside the school premises and their homes while ferrying school children. OMI points towards the lack of awareness and adequate training in relation to locating the washroom through the mainstream apps.

Secondly, contrary to the popular belief, it was found that most women prefer working during the night as 1) it helps them balance the household and childcare responsibilities during the day, 2) night time fares are higher, and 3) the traffic congestion is low. In this context, OMI recommends following infrastructural changes:

> *Make the night time more comfortable and make that more accessible. And by that, I mean better street safety at large, comfort stations, social spaces that open late into the night that are accessible for drivers, women drivers and women passengers and other kinds of travellers. All of them have to be more vibrant, have to be more safe.*

***Safety and harassment:*** Both Malini and Maya expressed their discontent with the mainstream platforms reporting the problematic male passengers. Maya Sharma says:

> *No action was taken except their promise to talk to the customer about it... "We will look into the matter and discuss internally". Whenever we spoke to the customer care, they said we are registering a complaint and we will forward it to the right person.*

Moreover, she did not even lodge a First Information Report (FIR) because, "we never had the number of the customer and other details. There was no number given in the app. We used to get the call from the customer through the app." Malini Tyagi believes that men approaching women is common in all the sectors but in a blue collar profession like driving, women are

more prone to harassment as men do not ask in a polite way and in certain cases, they assume that they do not even have to ask.

***Ownership capacity:*** The data indicated that women are not keen on owning the cars as several interviewees cited incidents where the male members of the family snatched away the cars, making a case for asset non-ownership:

> *It so transpires that the men of the family, whether it be the brother, the husband, the brother-in law, the father, took over the car. Let's say she would have a booking late in the night or say 9 o'clock, the car will be with the husband or the father or the respective men in the family. And the woman had no say... So, they suggested that madam, you start a company, you own the car, and we drive it for you. (Susieben Shah)*

Thus, asset ownership is a complicated issue that requires thoughtful solutions. Pushing women to purchase cars by providing subsidies ignores the deeper embedded issues in the Indian society, where a woman has no power and control over the male members of her family.

### *5. Fairness in data management and design*

The analysis of the interview transcripts and the related documents also pointed towards the need for a fair technological design that does not perpetuate any sort of discrimination; builds algorithms that cater to the requirements of women drivers such as the desire for fixed salary (and solve the accompanied issues related to workers' accountability and quality control); and helps in establishing trust between the platform drivers and the platform customers.

***Social sorting:*** Interview participants shared how sometimes customers make strange requests that could lead to social sorting in the platform economy. For instance, Amrita Gupta describes the kind of requests they receive from potential customers for hiring drivers for private placements:

> *There is a lot of monitoring that is happening...And a lot more social, a lot of more selection is happening. "I don't want a girl who is dark. I don't want a girl with a Muslim name. What is her*

*Surname? What is her caste?" And these are the things are happening on the platform economy as well... When we are dealing with private placements, we have had such comments – she is too short, she is too thin, she is too tall, she is dark, she doesn't look good, all of that.*

Woman driver Maya Sharma shares an incident where she refused to drive as she felt uncomfortable with the requirements of the customer:

*She selected me, but I wasn't happy. Even though the job was my need, we have been waiting for this job for five or six months after finishing the training and we were putting in all the effort ultimately for the job. My financial condition too wasn't very well at the time. I was in need of a job. But I wasn't happy when she told me the reason for selecting me. She said that she actually needed a young girl only. She said that she didn't want any old lady or any woman. She did not require a woman. She needed a young girl. So, she said you are perfect for us, for our home. So, I said, "ma'am, actually the thing is not about if the girl is young or not. The other two drivers who were with me were older than me. One of them was 40 and the other 35 years in age. The other two drivers that came with me for the test drive. So, I told her that ma'am it's not about the age. These two also do perfect driving. You should not say so, they would feel bad. It's my nature that I say it to the face." [This narration is in context to the driver placement service offered by a platform]*

Amrita Gupta adds that the digital gig economy does not solve this problem either as platforms have access to the Aadhaar cards[5] of workers, where 'names' can reveal their castes and religions. In fact, a study by Rathi & Tandon (2021) found how certain mobile applications and websites allow users to filter the gender, caste, and religion of domestic workers leading to discriminatory differentiation in the platform economy. Another research article states that the platform economy can result in social sorting of "bodies using gender, class and race as categories of discipline and discrimination" (S. Khan, 2019, para 3).
In this context, Graham et al. (2017) writes that the digital platforms should be developed in a way that "can allow workers to access their local market through a veil of anonymity provided by the digital medium, masking the characteristic on the basis of which

[5] Aadhaar is a 12 digit individual identification number issued by the Government of India. The number serves as a proof of identity and address.

discrimination occurs" (p. 147). Hence, developing a non-discriminatory design is both a need and a challenge in the digital platform economy.

***Fixed salary design:*** It was also found that most women drivers desire fixed salaries over the fluctuating or uncertain income model followed by the mainstream ride-hailing platforms. For instance, when the interview participant, Sunita Mishra, was asked multiple times that would she still prefer fixed salary over the option to earn more with the mainstream platforms, she maintained that she does not mind earning less, and is satisfied with it.

> *No, I am satisfied with my salary... No, no, I am satisfied with fixed salary only... Yes, even if it is less, I am satisfied. (Sunita Mishra)*

Similarly, Amrita Gupta, confirms vehemently that the drivers want salary as the salary and the social security benefits even help them procuring loans for purchasing cars. She adds that the salaried bank accounts are the mandatory requirement listed by most financial institutions for processing loan applications.

Sushil Shroff elaborates upon the need for a fixed salary design in a different context. He said that the incentive-based system with long driving hours does not appeal to the women drivers. "Men in India drive about 14 to 16 hours, a woman possibly cannot do that much of work due to the disproportionate burden of the care work responsibilities that falls on women," adds the Taxshe Director COO. In this case, a woman will always lose out on the incentives or fail to achieve the Minimum Business Guarantee (MBS) targets set by the mainstream platforms, as she will be expected to drive a minimum of 12-14 hours in a day to make ends meet (Khatoon et al., 2019).

Thus, the data indicate that the platforms need to incorporate a fixed salary or income model as part of their design. However, a bigger challenge in the event of a fixed payments model is how to develop algorithms that ensure workers' accountability, quality control, incentives, and feedback from the customers. The mainstream platforms are currently able to penalize and incentivize or reward the workers partly because they do not have a fixed salary design, as the workers are afraid to lose out on the business and money.

***Hybrid model:*** Women drivers like Sunita Mishra were found to be more comfortable in driving on 'fixed routes' and with 'familiar-sensitized' customers. Sunita explained that the pick-up location was shared with her in advance, which helped her in mapping the route and registering the landmarks before she started ferrying the school children. Sushil Shroff elaborated that Taxshe creates WhatsApp groups where parents and drivers are added for communication and tracking, which results in a technology mediated 'circle of trust'. He further said that Taxshe acts as a mediator not only for connecting the two parties, but also to sensitize the customers:

> *They are sensitized about the driver...that if the driver is late... now suppose they pick up your child and you are two minutes late, the next child will be waiting, that parent will get angry. And they don't see why you're late, they just shout at you, that is the tendency of people's behavior with drivers. So, we go and sensitize them that if she's late, you're not going to shout at my lady drivers. If you shout, she may get nervous, it puts people at risk, or you will try to hurry up the next drive, so that puts your child at risk. So, you're not going to shout. If you have any problem about being late, you mention in a group that the driver was late. And the parent who had put the child a little late two minutes late or three minutes late, will take up the onus of saying that no it was late because of me. So, then the blame game does not happen. It should not reach my woman driver.*

At the same time, it is important to note that the app-based model was preferred by the customers (parents) of Koala Kabs. Shailja Mittal explains:

> *Parents wanted to book through app only because they could track it. They could track where the car is, where, what the speed is...If the kids are using the service, then what time my kid is being dropped in the class? What time is he being picked up? They used to get all the alerts from the app. The drivers were trained that way. They used the app that way. And that is how we used to rate our driving partners...that have you use the app properly or not, because for parents, it was very important to track that where the kids are. So that is why they always prefer to book the service through app only.*

These scenarios and the interviewees point towards the need of a hybrid model where there is an integration of technology (apps) with human managers. It is imperative that a 'circle of

trust' is built where female drivers feel safer to drive and their managers understand them. Sunita Mishra fondly speaks about her assignment manager and co-founder of the Taxshe:

> *They have that patience, that care, they care also...If it is an emergency, I can't help it out, they arrange it very fast. Aarti does all those Superman work. Vandana madam is also very understanding. That is the thing, she understands us.*

Overall, this section highlights the issues in the current technological design of the ride-hailing platforms and pushes for a non-discriminatory and hybrid model, which protects the identity of the works and where technology is integrated with human connection.

## Conclusion

"Inclusive transportation is a key, but often underemphasized, catalyst for gender equality" (IFC, 2020, p. 2). The underrepresentation of women as drivers has large-scale implications on women's freedom of movement and access to the job market. The findings indicate a need for a hybrid institutional model – non-profit foundation and for-profit taxi company that understands the complexities and underlying social norms to increase the participation of women drivers. Further, new technologies employed in the ride-hailing sector pose questions about digital safety and privacy, labor precarity and mental stress in the absence of human managers, and surveillance and tracking methods due to algorithm-driven models. In this case, an amalgamation of app-based technology and human-centric mediation can help in establishing trust, providing direct feedback, and sensitizing both the drivers and the customers. Here, we recentered the lived experiences of marginalized female drivers in India that are typically neglected in the ride-hailing sector and push for a cross-cultural framework that denaturalizes the Western hegemony. Ultimately, we aim to decolonialize the Anglo-Saxon normative perspectives on what constitutes as ethical platforms, fair work, and inclusive labor by privileging alternative cultural and sectoral contexts.

## Acknowledgement

This research has been done as part of a project seed-funded by the International Development Research Centre (IDRC), Canada. Project no.: 109331-001. Project name: Organizing Digitally (Public name: Feminist Approaches to Labor Collectives).

## References

Aloisi, A. (2015). Commoditized Workers. Case Study Research on Labour Law Issues Arising from a Set of "On-Demand/Gig Economy" Platforms. Comparative Labor Law & Policy Journal, August 2015. https://doi.org/10.2139/ssrn.2637485

Aneja, U., & Shridhar, A. (2019). Worker Wellbeing on Digital Work Platforms in India: A Study of OlaCabs and UrbanClap in New Delhi. In Tandem Research.

Arora, P., & Raman, U. (n.d.). Fair Work, Feminist Design, and Women's Labor Collectives. In M. Graham & F. Ferrari (Eds.), Digital Work in the Planetary Market. MIT Press/IDRC.

Arora, R. (2016, July 25). Government asks taxi aggregators to train one lakh drivers annually. Economic Times BrandEquity.https://brandequity.economictimes.indiatimes.com/news/business-of-brands/government-asks-taxi-aggregators-to-train-one-lakh-drivers-annually/53373649

Bardzell, S. (2010). Feminist HCI: Taking stock and outlining an agenda for design. Conference on Human Factors in Computing Systems - Proceedings, 2, 1301–1310. https://doi.org/10.1145/1753326.1753521

Beerepoot, N., & Lambregts, B. (2018). Reining in the global freelance labor force: how global digital labor platforms change from facilitators into arbitrators. The Future of Work in the Global South, 12–15.

Bowen, G. A. (2009). Document analysis as a qualitative research method. Qualitative Research Journal, 9(2), 27–40. https://doi.org/10.3316/QRJ0902027

Business Today. (2020, November 27). Big news for Ola, Uber! Now, there's a definition for cab aggregators. BusinessToday. https://www.businesstoday.in/latest/economy-politics/story/govt-Defines-cab-aggregators-through-motor-vehicle-aggregator-guidelines-279843-2020-11-27

Chen, J. Y. (2017). Thrown under the bus and outrunning it! The logic of Didi and taxi drivers' labourand activism in the on-demand economy. New Media & Society, 20(8), 2691–2711. https://doi.org/10.1177/1461444817729149 9

Chen, J. Y. (2018). Technologies of Control, Communication, and Calculation: Taxi Drivers' Labour in the Platform Economy. Humans and Machines at Work, 231–252. https://doi.org/10.1007/978-3-319-58232-0_10

Cherry, M. (2016). Beyond misclassification: The Digital Transformation of Work. Comparative Labor Law & Policy, 37(3), 544–577.

Choudary, S. P. (2018). The architecture of digital labour platforms: Policy recommendations on platform design for worker well-being. In ILO future of work research paper series.

Chugh, N. (2016, July 8). "How can a girl become a taxi driver?": defying India's caste and gender taboos. The Guardian. https://www.theguardian.com/global-development/2016/jul/08/india-Dalit-how-can-a-girl-become-a-taxi-driver-defying-caste-gender-taboos

Curry, D. (2021, May 7). Ride Hailing Taxi App Revenue and Usage Statistics (2021). Business of Apps.https://www.businessofapps.com/data/ride-hailing-app-market/

De Stefano, V., & Aloisi, A. (2018). Fundamental Labour Rights, Platform Work and Human-Rights
Protection of Non-Standard Workers. SSRN Electronic Journal, January. https://doi.org/10.2139/ssrn.3125866

Eskerod, P., & Larsen, T. (2018). Advancing project stakeholder analysis by the concept 'shadows of
the context.' International Journal of Project Management, 36(1), 161–169. https://doi.org/10.1016/j.ijproman.2017.05.003

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. International Journal of
Qualitative Methods, 5(1), 80–92. https://doi.org/10.1177/160940690600500107

Fielbaum, A., & Tirachini, A. (2020). The sharing economy and the job market: the case of ride-
hailing drivers in Chile. Transportation, 0123456789. https://doi.org/10.1007/s11116-020-10127-7

Fleming, P. (2017). The Human Capital Hoax: Work, Debt and Insecurity in the Era of Uberization.
Organization Studies, 38(5), 691–709. https://doi.org/10.1177/0170840616686129

Florito, J., Aneja, U., & de Sanfeliu, M. B. (2018). A Future of Work that Works for Women. Gender Economic Equity and the Future of Work, G20 Insights, 1–20.

Frost & Sullivan. (2019). With Projected CAGR of 9.7% over 2019-2025, Shared Mobility Set to
Emerge as Major Transportation Mode across India.

Graham, M., Hjorth, I., & Lehdonvirta, V. (2017). Digital labour and development: impacts of global
digital labour platforms and the gig economy on worker livelihoods. Transfer, 23(2), 135–162. https://doi.org/10.1177/1024258916687250

Graham, M., Woodcock, J., Heeks, R., Mungai, P., Van Belle, J. P., du Toit, D., Fredman, S., Osiki,
A., van der Spuy, A., & Silberman, S. M. (2020). The Fairwork Foundation: Strategies for
improving platform work in a global context. Geoforum, 112(February), 100–103. https://doi.org/10.1016/j.geoforum.2020.01.023

Hanrahan, B. V., Ma, N. F., & Yuan, C. W. (2017). The roots of bias on uber. ECSCW 2017 -
Proceedings of the 15th European Conference on Computer Supported Cooperative Work. https://doi.org/10.18420/ecscw2017-27

Heeks, R. (2019, January 29). How Many Platform Workers Are There in the Global South? ICT4DBlog. https://ict4dblog.wordpress.com/2019/01/29/%0Ahow-many-platform-workers-are-there-in-the-global-south/

Hunt, A., Samman, E., Tapfuma, S., Mwaura, G., Omenya, R., Kim, K., Stevano, S., & Roumer, A.
(2019). Women in the gig economy: Paid work, care and flexibility in Kenya and South Africa.November,
92.https://data2x.org/wp-content/uploads/2019/11/WomenintheGigEconomy_ODI.pdf

Huws, U. (2015, January 1). iCapitalism and the Cybertariat. Monthly Review. https://monthlyreview.org/2015/01/01/icapitalism-and-the-cybertariat/

International Finance Corporation. (2020). Gender-Segregated Transportation in Ride-Hailing:
Navigating the debate.

International Finance Corporation and Uber Technologies. (2018). Driving Toward Equality: Women,Ride-Hailing and The Sharing Economy.

Kasliwal, R. (2020). Gender and the Gig Economy: A Qualitative Study of Gig Platforms for Women
Workers. Observer Research Foundation, 359, 1–14.

Khan, A. Y. (2020, November 26). Uplifting minorities, the Telangana way. Telangana Today.

Khan, S. (2019, January 21). Social Sorting as a Tool for Surveillance. Heinrich Böll Stiftung, Gunda
Werner Institute: Feminism and Gender Democracy.

https://www.gwi-boell.de/en/2019/01/21/social-sorting-tool-surveillance

Khatoon, N., Kumar, N., Singh, P., Kumari, S., & Lakimsetti, S. H. (2019). The Platform Economy:
A case study on Ola and Uber from the driver partners' perspective. In TISS Hyderabad (pp. 1–
22).

Lee, M. K., Kusbit, D., Metsky, E., & Dabbish, L. (2015). Working with Machines: The Impact of
Algorithmic and Data-Driven Management on Human Workers. Conference: CHI '15 Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 1603–1612.
https://doi.org/10.1145/2702123.2702548

Mitra, A. (2019, December 9). Gatidhara Scheme: Govt scheme to tackle joblessness gives slowing
auto industry a boost. The Indian Express.

Muller, Z. (2020). Algorithmic harms to workers in the platform economy: The case of Uber.
Columbia Journal of Law and Social Problems, 53(2), 167–210.

Page, X., Marabelli, M., & Tarafdar, M. (2017). Perceived Role Relationships in Human-Algorithm
Interactions: The Context of Uber Drivers. ICIS 2017: Transforming Society with Digital Innovation.

Prabhat, S., Nanavati, S., & Rangaswamy, N. (2019). India's "uberwallah": Profiling Uber drivers in
the gig economy. ACM International Conference Proceeding Series, October.
https://doi.org/10.1145/3287098.3287139

Prescient & Strategic Intelligence. (2021). India Shared Mobility Market Overview.
https://www.psmarketresearch.com/market-analysis/india-shared-mobility-market

Ramachandran, S., & Raman, A. (2021). Unlocking jobs in the platform economy: Propelling India's Post-Covid Recovery. In Ola Mobility Institute (OMI).
https://olawebcdn.com/ola-institute/OMI_Platform_Economy_Report.pdf

Rathi, A., & Tandon, A. (2021). Platforms, Power, & Politics: Perspectives from domestic & carework in
India.https://cis-india.org/raw/platforms-power-and-politics-pdf

RideGuru. (2021). Rideshares Worldwide. RideGuru.
https://ride.guru/content/resources/rideshares-worldwide

Rosenblat, A., & Stark, L. (2016). Algorithmic labor and information asymmetries: A case study of
Uber's drivers. International Journal of Communication, 10, 3758–3784.
https://doi.org/10.2139/ssrn.2686227

Schmidt, F. A. (2017). Digital Labour Markets in the Platform Economy: Mapping the Political
Challenges of Crowd Work and Gig Work. Freidrich Ebert Stiftung, January, 1–32.

Schor, J. (2016). Debating the Sharing Economy. Journal of Self-Governance and Management
Economics, 4(3), 7.
https://doi.org/10.22381/jsme4320161

Shaheen, S., Cohen, A., & Zohdy, I. (2016). Shared Mobility: Current Practices and Guiding
Principles. In Fhwa-Hop-16-022 (Issue Washington D.C.).

Sharma, N. (2020, July 29). Covid effect: Govt to switch gears, double 'Apni Gaddi' scheme subsidy.Hindustan Times.
https://www.hindustantimes.com/chandigarh/covid-ect-govt-to-switch-gears-double-apni-gaddi-scheme-subsidy/story-zwsiHnwHv5PtJjzrpUIqRJ.html

Sridharan, A. (2018, May 18). When women get behind the wheel ... to get ahead. The Hindu Business Line.
https://www.thehindubusinessline.com/news/variety/when-women-get-behind-the-wheel-to-get-ahead/article23928754.ece

Srivastava, S. (2019, July 27). 35 Best Uber Statistics to Know (2019-2020 Updated). Appinventiv.
https://appinventiv.com/blog/uber-statistics/
Stand-Up India Scheme Features. (n.d.). Retrieved December 19, 2021, from
https://www.standupmitra.in/Home/SUISchemes

Surie, A. (2018). Are Ola and Uber Drivers Entrepreneurs or Exploited Workers. Economic and
Political Weekly, 53(24), 1–7.

van Doorn, N. (2017). Platform labor: on the gendered and racialized exploitation of low-income
service work in the “on-demand” economy. Information, Communication & Society, 20(6), 898–
914.
https://doi.org/10.1080/1369118X.2017.1294194

Zainab, K. (n.d.). Women and the Platform Economy. IT for Change. Retrieved March 28, 2022, from

Zanoni, P. (2019). Labor market inclusion through predatory capitalism? The “sharing economy,”
diversity, and the crisis of social reproduction in the Belgian coordinated market economy.
Research in the Sociology of Work, 33, 145–164.
https://doi.org/10.1108/S0277-283320190000033009

## Biographies

**Pallavi Bansal** is currently working as an Assistant Professor at the Times School of Media, Bennett University. She is also pursuing a PhD from Erasmus University Rotterdam, The Netherlands, and is a junior researcher at Feminist Approaches to Labour Collectives (FemLab.Co). She holds an MSc in Media, Communication and Development from the London School of Economics and Political Science and has previously worked with the United Nations in Geneva and The Times of India in Delhi. Now, she actively blogs for The Times of India and her current research focuses on the feminist design of the AI-enabled digital labour platforms.

**Payal Arora** is a digital anthropologist, a TEDx speaker, and an author of award-winning books, including 'The Next Billion Users' with Harvard Press. Her expertise lies in user experience in the Global South and inclusive design. Several international media outlets have covered her work including the BBC, The Economist, Quartz, Tech Crunch, The Boston Globe, F.A.Z, The Nation and CBC. Forbes named her the "next billion champion" and the "right kind of person to reform tech." She is a Professor at Erasmus University, Academic Director in UX and Inclusive Design at the Erasmus Centre for Data Analytics, and Co- Founder of FemLab, a feminist future of work initiative.

**Usha Raman** is a professor in the Department of Communication at the University of Hyderabad in India where she researched and teaches in the areas of digital cultures, science and health communication, and feminist media studies. She is co-founder of FemLab, a feminist academic-activist initiative on the futures of work in a digital landscape.

AI & EQUALITY

# Mainstreaming Gender Perspective in AI crowd work in the Global South: *Diagnostic, policy recommendations and smart tools for women's empowerment*

Cristina Martínez Pinto, Luz Elena González Zepeda, Tatiana Telles, Norma Elva Chávez, Maya De Los Santos, Alberto Navarrete, and Saiph Savage

## Abstract

Workers from Kenya and the Philippines were among the first to be externally employed to label data, but after 2018, 75% of the leading AI crowd work platforms' workforce was Venezuela and it is possible that other Latin American countries will follow in the context of the post covid-19 economic recovery. While crowd work presents opportunities for income and employment creation in regions where local economies are stagnant- there are not enough initiatives that address the impact of such work in the Global South through the lens of gender perspective, considering that 1 in every 5 crowd workers in the region are women. To address this knowledge gap, we conducted an experimental survey on 60 women from Latin America who use the crowd work platform Toloka to understand their personal goals, professional values, and hardships faced in their work. Key insights revealed that a majority of the women shared a desire to hear the experiences of other crowd working women; particularly to help them navigate tasks, develop technical and soft skills, and manage their finances with more ease. Additionally, 75% of the women reported completing crowd work tasks, on top of caring for their families; while over half of them confirmed they needed to negotiate their family responsibilities, in order to pursue crowd work in the first place. These findings confirmed an important component lacking from the experiences of these women was a sense of connection with one another. Based on these observations, we propose a system designed to foster community between Latin American women in crowd work to improve their personal and professional advancement.

## Introduction

The production of labeled data is crucial to meet the expectations of Artificial Intelligence (AI) system development and training. The great technological promise of this century depends largely on algorithms that need millions of labeled training data to learn, recognize, and categorize information. Data labeling is the product of AI models and the manual work of people who monitor, correct and augment the predictions of the former, thus improving their accuracy. This form of labor is known as crowd work.

As crowd workers continue to make crucial contributions to the training and development of advancing AI systems, interest in ensuring they produce quality work has grown. However, previous approaches have neglected to center crowd workers in system designs and to consider their identities, motivations, and well-being. These considerations are important, because they acknowledge crowd workers are not a monolith and may have different needs and goals. Designing a system sensitive to the unique needs of a specific worker population can increase the likelihood of the tool's usefulness and, by extension, contribute to crowd workers well-being and help them produce better quality work.

In recent years, Latin American and Caribbean workers have been identified as significant crowd work contributors. The crisis conditions in these areas have left many dependent on crowd work as a steady source of income. Latin American women make up a sizable portion of these workers, using crowd work as a path to financial independence while balancing traditional caregiver responsibilities expected of them due to their strong patriarchal societies. Unfortunately, research surrounding the perspectives of Latin American crowd workers is largely incomplete as most investigations have been centered on the experiences of Western women, who comparably have greater freedom and support to prioritize their individual needs.

To address this knowledge gap, we conducted an experimental survey on 60 women from Latin America who use the crowd work platform Toloka to understand their personal goals, professional values, and hardships faced in their work. Key insights revealed a majority of the women shared a desire to hear the experiences of other crowd working women; particularly to help them navigate tasks, develop technical and soft skills, and manage their finances with

more ease. Additionally, 75% of the women reported completing crowd work tasks, on top of caring for their families; while over half of them confirmed they needed to negotiate their family responsibilities, in order to pursue crowd work in the first place. These findings confirmed an important component lacking from the experiences of these women was a sense of connection with one another.

Based on these observations, we propose a system designed to foster community between Latin American women in crowd work to improve their personal and professional advancement. By providing them with a safe platform to engage in meaningful conversation, they can begin to build an extensive foundation of knowledge for completing their work while growing their career and personal skills in the process. Moreover, since many Latin American women share the pressure of balancing their family responsibilities and reputation alongside their work, they are in the best position to relate and give advice to one another.

In order to facilitate communication between the women, and help them identify conversations related to their current interests, we would integrate an intelligent chatbot into the system. Designed to emulate the personality of Latin American heroines, the chatbot will assist users in searching for specific advice, help them textualize their interests, and be a guide for navigating the platform. Furthermore, the chatbot will be capable of recommending to users other crowd working women who may be good connections based on their interests, expertise, and experiences.

Designing a space for crowd workers to access each other's knowledge can lead to improved quality in their work by empowering them to seek support addressing problems in their working conditions. According to field practitioners and researchers, a better future in crowd work can only be fulfilled if key aspects such as transparency, fair pay, and opportunities for professional advancement are embodied in tools available to crowd workers. Having their experiences documented will not only benefit crowd workers long-term, but also provide them with the ability to support one another in identifying areas of improvement in their work environments. This can inform decisions to organize and advocate for respect, fair pay, and transparency from requesters and crowd working platforms. Despite being promoted as an opportunity to create income and employment in regions where local economies are stagnant, there are not enough initiatives that address the impact of such work in the Global

South through the lens of gender perspective, considering that 1 in every 5 crowd workers are women.

In addition to the experimental surveys conducted, we analyzed the current state of AI-related crowd work in the Latin America and the Caribbean region, studying the differentiated impact of such work opportunities for women and men, and developed a set of policy recommendations based on international best practices and use cases.

## Feminist Research

Research can be considered feminist when it is grounded in a set of theoretical traditions that privilege women's issues and experiences. A key particularity of feminist research lies in its focus on power imbalance, both in the subject that is studied and the relationships between the subject and the research. Feminist research practice includes the practice of reflexivity or positionality, as a tool for producing knowledge but also an ethical method focused both in subverting unequal relationships of power in knowledge production, to consider our own position within the research as one of power, and to always take into account the vulnerability of the researched. An ethical commitment cannot happen without the step of reflexivity. Reflexivity is the deconstruction of knowledge production, by addressing academia as a place for knowledge production, where social interactions and thus social constructions happen.

Gender mainstreaming can be used as a strategy for achieving equality by integrating gender analysis in all the phases of policies, actions, plans, and budgets. Gender as a variable of analysis looks at social attributes and opportunities associated with being female and male; to the relationships between women and men, and to the relations between women and those between men. Moreover, as gender is generally understood as a social construct affected by power structures, different analytical concepts have been used to perfect the analysis of those power structures and its relations to gender, the most widely used being intersectionality. Intersectionality is used to map out the intersecting power relations in social relations by viewing categories such as race, class, gender, sexuality, nation, ability, ethnicity, among others, interrelated and mutually shaping one another.

As such, by considering the gender variable as central to our research, we go beyond sex disaggregation, and look into what the data conveys about attitudes, norms, and gaps within AI crowd work. Within the project, two survey pilots were conducted in such a way to prevent exposure of the women identified as subjects of research (crowd workers), to abide by ethical standards and fair pay, and to enable communication channels for them to open up about their lived-in experiences beyond what previous literature review on women crowd workers has revealed.

## Methodology

Our team conducted two survey pilots in Spanish on the Toloka platform. Toloka is a global crowd sourcing company, founded in 2014 by Olga Megorskaya and integrated within the Yandex search engine, as an enabling environment to support data-related processes. The platform was chosen as it allowed us to filter workers geographically, to specifically analyze Latin American workers. Additionally, we contacted Toloka's Educational Program team, who expressed interest in supporting research activity through the platform.

The first version of the survey was first answered by 6 people (three men and three women); the survey pilot was divided into specific sections, and most of the questions were of Likert scale type and open-ended questions. The time to submit the survey was established to 20 minutes for women and men crowd workers that were interested in doing the task, and a filter for sex was used to get the answers of a group of women and men separately. Even though Toloka only permits binary sex desegregation, an option for more inclusive gender identity was included within our survey. As we became employers on the platforms as requesters, the payment offered for the completion of the survey as a task was calculated based on the minimum wage in the United States and the time to complete such a task The payment to submit the survey was calculated as (minimum payment x fraction of an hour). For example, if we assume a payment of USD 7.25 and the time to complete the survey was 20 minutes, the worker would receive an amount of USD 2.4 (7.25*0.333). In order to gain insight on how gender affected their work, we used different questions regarding widely recognized gendered factors such as time poverty, work-home balance, outside attitudes to the work done, care work, and direct opinions about if they had considered gender as a liability to their work. The second version of the survey was an iteration based on feedback about the need to include additional questions to better understand the characterization of women working in

crowd work in Latin America, their motivations, needs, socioeconomic context, skills, as well as hopes. The sample of the second survey included 57 women and three men.

## Results of survey experiments

***First survey***

*Men respondents*:

Three men were part of the sample. The pilot had 3 age groups, from 18 to 55 years old. Educational profiles included associate professionals and university level participants without a degree. They found Toloka through social networks, and they do crowd work because they prefer remote work and to supplement their income.

The respondents grew up in a hierarchical, relationship-based culture. They perceive knowing their work styles, having a vision for their career, knowing how to look for job opportunities, and what strengths help them at work. They work from home with their own computer most of the time; they don't make much use of external resources to perform their tasks at Toloka (no Facebook, no WhatsApp, no Reddit). They have been working in Toloka between a month to a year; their average income ranges from .02$ to 2$ per task.

Negotiation of home care tasks is neutral for them because they can postpone them for later, or because someone else takes care of their children (if any). They consider that gender does not affect the way they perform collective work. They believe that their current schedule allows them to have enough time to rest, but not enough time for self-care.

Their families respect their work on crowd work platforms, because it supports them with household expenses; the 3 men stated spending most of their Toloka income on their families. They conducted their work without supervision from other people. They considered that remote work does not contribute so much to their financial independence, because they spend their income on their family, but that it brings them valuable skills for their CV. The respondents knew English thanks to private education, self-education and the use of translation tools. After the COVID-19 pandemic, the survey respondents considered that remote work became the new normal.

**Table 1. Conclusion of qualitative answers of men sample (1)**

| |
|---|
| Recurring and important topics:<br>● Self care: Not enough time for self care.<br>● Family income: All spent Toloka income on their families to some extent, men spent most of it on their families.<br>Care work: Others in their proximity assumed care work to allow them time to do crowd work.<br>● Motivations: Valued other benefits, such as new skills.<br>● Alienation: Valued meeting other crowd workers. Perception that gender does not affect crowd work. Neutrality regarding crowd work |

*Women respondents:*

The sample included 3 women from different age groups, from 18 to 45 years old. All respondents were based in their home country and worked on their own computers from home. Two of them had master's degrees. These women considered that they had identified their work culture and values as egalitarian and relationship-based. They knew their work style, goals, career vision, and how to look for opportunities, together with their main strengths. The respondents had been working on the Toloka platform since 2021-2022 and learned about it through social media. They worked with Toloka to make the most of their free time and to generate supplementary income. Their average income ranged from $0.10 to $4.00. Two people mentioned YouTube as the platform where they frequently asked for help.

They considered that their schedules allowed them to have time for self-care and to rest. They did not consider spending their extra income on their families but on themselves, and believed that crowd work contributed to their financial independence. They stated that they did not have to negotiate care-work to do collective work, because their schedules did not conflict. They were not supervised when doing collective work; their families remained neutral or supported their work on collective platforms. They perceived that the COVID-19 pandemic impacted the way they viewed collective work, because they can work more safely from home.

**Table 2. Conclusion of qualitative answers of women sample (1)**

Recurring and important topics:

- Self care: Work and care balance allowed them for self-care time.
- Education Capital: Youtube as a tool to search for tutorials on how to perform the tasks. Use of tools to translate English and develop English competencies.
- Independence: Crowd work is perceived as complimentary income that may promote financial independence.
- Alienation: No contact with other women crowd workers. May value contact with crowd workers. Perception that gender does not affect crowd work. Respect or neutrality regarding crowd work.

***Second survey***

A second survey was conducted with 61 questions in Spanish on the Toloka platform; in this iteration, a group of 60 people was considered to submit the survey. It was divided into specific sections, and most of the questions were of Likert scale type and open-ended questions. In this case, the time to submit the survey was established to 23 minutes for women and men crowd workers that were interested in doing the task, and it did not include a filter for gender. Additionally, the survey was applied specifically in the region of Latin America, using a filter for every country of the region. From the total of respondents, only 3 identified as men, and 2 preferred not to be identified.

There were four age groups represented in the sample, and 76% of respondents were 18 to 35 years old. The majority answered that their gender had little or no impact on their experience performing crowd work, while 17% believed it had some or an important impact on it. In accordance with the age group, most respondents had a professional degree, or had unfinished university studies; 13% held a master's degree.

Regarding their professional self-assessment, most of them knew their work strengths, work style, work environments, and had a long-term vision of their career. They could balance their professional and personal goals and they wanted to develop new skills. Other respondents regarded themselves as being neutral in goal balancing, building long-term vision of their careers, and identifying work environments that interest them in their career search. Few of the respondents perceived that they did not know how to search for careers relevant to their interests, and did not have a long-term vision of their careers.

They did not work in public spaces or libraries. They worked about 2 times per week from their own devices and at home, and spent about 20 minutes completing a common task, in some cases up to 40 minutes; the task they performed most is data labeling, web page testing, and surveys (only 10% performed audio transcription tasks).

Income per task was in the range of $0.01 to $0.16. 91% of the respondents conducted other economic activities in addition to Toloka. The most recurring reason for them to perform

crowd work tasks was to gain supplementary income. Additionally, a preference for working from home, and having schedule flexibility, were important factors. Few people considered it a recreational activity, and one respondent mentioned crowd work as one of few possibilities for employment. Regarding the impact of the geopolitical situation of their country of origin, participants considered the main impacts to be economic, which reduced their professional prospects and required them to perform a second economic activity; 21% remained neutral.

**Table 3. Conclusion of qualitative answers of women sample (2)**

Important topics towards the development of the proposed tool:

- Communitary context: 55% have not established relationships with other workers, and 68% would like to have a way to do so.
- Important Skills to improve:
  - Online marketing
  - Programming, IT, web development
  - English
  - Text interpreting, writing, reading comprehension, transcription, excel ○ AI, advanced tech tools
  - Finance
  - Soft skills, leadership
  - Habits of perseverance and concentration to execute work online
  - Effective communication, project management
  - Teamwork
- Topics of major importance for crowd workers in Toloka in the development of a tool:
  - Sharing experiences and tips to improve profitability.
  - Doing more tasks, productivity
  - Training for professional and skill development
  - Finding the best tasks
  - Clarification of instructions, ambiguities and possible mistakes
  - Financial issues; better payment on tasks

Most importantly, the respondents shared their preferred features for a tool. Explicability of tasks is important; although crowd workers in Toloka can contact the requester, multimedia resources and tutorials are useful for them to fully understand the tasks. Similarly, as 63% of respondents use search engines and translating applications as tools, prioritizing language learning features is an important skill to develop. Regarding the perceived need to share experiences with other workers, the main goal of a communication feature would be to increase task completion, profitability and performance.

## Design Implications


**Figure 1. Example of our intelligent platform that guides women crowd workers to build community and guides them to increase their skills with the use of AI systems.**

We propose to design an AI-powered social connection and recommendation system, specifically designed to assist crowd workers in building a supportive community and developing their technical and soft skills. Figures 1, 2, and 3 provide examples of the AI based platforms and chatbot tools we have started to explore based on our survey results, understanding the unique needs and challenges facing crowd workers. The chatbot will be designed to match individuals with similar interests, goals and skill sets, allowing for the formation of meaningful connections and opportunities for skill-sharing and collaboration. The AI of a system that connects crowd workers would likely involve a number of different technologies and techniques. One key component would be natural language processing (NLP), which would allow the system to understand and analyze the language used in the survey responses and profile information provided by users. This would enable the chatbot to understand the interests, skills, and goals of each individual, and match them with others who have similar characteristics.

Another important component would be machine learning algorithms, which would be used to analyze the data collected from the survey and continuously improve the system's ability to make accurate and relevant connections. These algorithms would be trained on the survey data and would be able to identify patterns and relationships between different users, and make predictions about who would be the best match for a given individual.

The impact of such an AI-powered system would be to create a more connected and supportive community for crowd workers, by connecting them with others who have similar interests, skills and goals. This would help to increase opportunities for collaboration, skill-sharing and networking, which would be beneficial for their careers and professional development. Additionally, a system like this would help to increase representation and visibility of crowd workers in their industries, and could help to bridge the gap of inequality that they are facing, in terms of access to resources and opportunities.

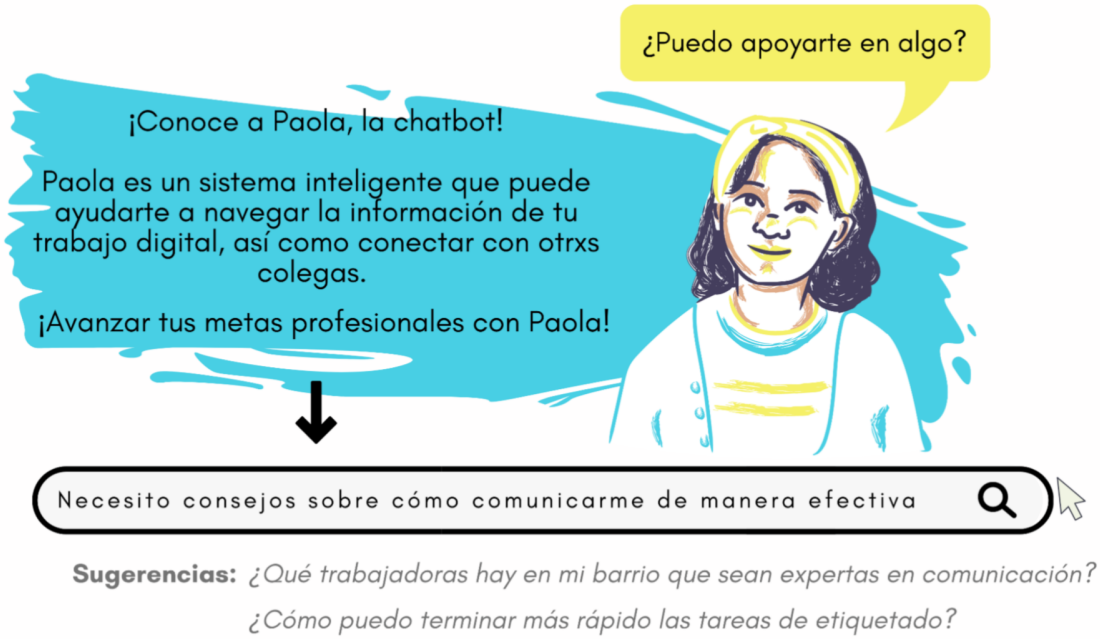



**Figure 2. Example of the intelligent chatbot users would interact with to navigate the platform and assist them in finding information and other workers relevant to their questions and interests.**


**Figure 3. Example of the workers the intelligent chatbot recommends the user connect with based on their search query about effective communication.**

The technical development of the platform would involve a combination of machine learning and natural language processing techniques. The AI would analyze the job listings on the platform and the resumes and profiles of crowd workers in order to match them with the most relevant and suitable job opportunities.

One of the key ways in which the AI would integrate the cultural background of the crowd workers would be through the use of NLP to understand the language and terminology used in job listings and resumes. For example, if a job listing used a specific term or phrase that is particularly relevant to the Latin American community, the AI would be able to identify and match that job with a crowd worker who has relevant skills and experience.

Another way in which the AI would integrate the cultural background of the crowd workers would be through the use of machine learning algorithms to analyze the data and identify patterns and trends specific to the Latin American community. For example, the AI may identify that crowd workers are underrepresented in certain types of jobs or industries, and could then use this information to prioritize job opportunities that would help to increase representation of crowd workers in those areas.

As this work is being designed with Latin American women in mind, we will refer to cultural theory to guide the integration of Latin American heroines into the design of our technology. We identified key women heroes from Latin America, such as Professor Paola Ricuarte, and utilized their stories to drive the design of our AI platform in a direction that is inspiring and encouraging for women workers. Additionally, personifying the chatbot lends to dispelling workers' concerns about interacting with AI and keeps them engaged while using the service.


**Figure 4. Example of the feedback survey and option to leave a suggestion at the end of each page of the platform in order to collect information on how workers are enjoying the service.**

Finally, another AI system we envision based on our survey results is one that could recommend different jobs to crowd workers based on their preferred job style and values that would involve the use of deep learning techniques. Deep learning is a type of machine learning that uses neural networks to process and analyze large amounts of data, and can be used to understand complex patterns and relationships between different data points.

The AI system would be trained on a large dataset of job listings and resumes, and would use deep learning algorithms to analyze the data and identify patterns and trends specific to the Latin American community. The system would take into account the preferred job style and values of the crowd workers, and use this information to match them with jobs that are most relevant and suitable to their skills and experience.

For example, if a crowd worker values a flexible schedule and has experience in customer service, the AI system would recommend jobs that offer flexibility and are in the customer service field. If a crowd worker values a work-life balance and has experience in marketing, the AI system would recommend jobs that offer a balance of work and life and are in the marketing field.

The AI system would also continuously learn and improve over time, by analyzing the data generated by the crowd workers, such as feedback on the jobs they applied for, and performance data. The impact of such an AI-powered system would be to increase the visibility and opportunities for crowd workers in the gig economy, by matching them with jobs that are relevant and suitable to their skills, experience, preferred job style and values. Additionally, it could help to bridge the gap of inequality that they are facing, in terms of access to resources and opportunities, and it could help to increase representation and visibility of crowd workers in their industries, giving them more opportunities for career growth and development.

## Literature Review: Global South diagnostic

As a new type of labor that leverages technology and new digital platforms, crowd work is gaining popularity. Globalization and digitalization have had a significant impact on labor markets, with particular shifts in how individuals work, changing work patterns towards a new emphasis on individual agency, and responsibility. Crowd-sourcing arose as a method of sourcing labor in which individuals or organizations use digital platforms to tap into the aggregate skills, knowledge, and expertise of a vast, geographically dispersed workforce.

There are a variety of definitions for crowd work and crowd-sourcing. Estellés-Arolas and González-Ladrón de Guevara offer a definition that summarizes the key elements of the concept (2012: 197):

> *"Crowdsourcing is a type of participative online activity, in which an individual, an institution, a nonprofit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that which the user has brought to the venture, whose form will depend on the type of activity undertaken".*

The broad range of tasks that can be completed includes audio and video transcription, content moderation, data gathering and processing, image identification, transcription, and annotation, as well as translation. These digital platforms serve as a "middleman" between employers and workers, assisting in the supervision of the creation, submission, acceptance, and payment of the work completed. Amazon Mechanical Turk (AMT), Upwork, Clickworker, PeopleperHour, CloudFactory, CrowdFlower, Freelancer, and Microworker are a few examples of digital networks for crowd work.

Collective work is presented as an opportunity to increase the participation of women in economic activities, as the flexibility of the remote work scheme can be a viable alternative to earn primary or additional income for some women. Previous studies have demonstrated that crowd workers are driven by intrinsic motivation, as well as extrinsic motivation, such as financial reward and the task itself, which allows them to be self-employed and to feel empowered.

Workers from Kenya and the Philippines were among the first to be externally employed to label data; however, after 2018, 75% of the leading AI crowd work platforms' workforce was Venezuelan, and it is possible that other Latin American countries will follow in the context of the post COVID-19 economic recovery. These platforms offer the chance to access a sizable pool of local and international labor to increase productivity, efficiency, and market reach, lowering operating costs for requesters. Crowd workers can also actively design their work-life balance, and freely decide how many contracts to take. Despite such benefits, the regularity of labor and income, working conditions, social protection, skill utilization, freedom of association, and the right to collective bargaining are some of the challenges that crowd workers face.

Currently, the majority of crowd work is not governed by labor laws, rather, the platforms themselves determine the working conditions. The crowd worker's ability to exert control over working time is frequently constrained by the nature of the task or the necessity to maximize productivity. In order to make enough money, many crowd workers must put in hours that are significantly longer than those typical of normal employment. Additionally, they have little alternatives for legal action in cases of unfair treatment. Crowd workers may be arbitrarily deactivated as a result of algorithmic management, which results in them losing their

revenue and being prevented from using the platform. They may also be liable to additional penalties or disciplinary actions without having the chance to appeal if considered unfair. In other words, digital platforms and their algorithms have a direct impact on crowd workers, as many managerial aspects are built into algorithmic decision-making and performance and monitoring systems.

According to the Organisation for Economic Co-operation and Development (OECD), platform workers are frequently misclassified as not-being employees, because platform companies would incur costs that are 20–30% higher by doing so. By classifying workers as independent contractors, platforms are able to avoid indirect costs associated with employee rights, such as the right not to be fired unfairly; in addition to direct costs, such as minimum salaries, maximum hours, paid leave, and paid sick leave. The contracting figure under which the employment relationship is established classifies all workers as independent contractors, which exempts platforms from guaranteeing basic protections. On the other hand, the ubiquity that characterizes the platforms also works in their favor, because in addition to the fact that accountability mechanisms are scarce, it is unclear under the jurisdiction of which region or country they operate and must respond.

Moreover, the International Labor Organization's (ILO) survey results from 2015 and 2017 indicate that there are significant gender pay gaps, and that many crowd workers earn less than the local minimum wage. Depending on the platform, women were paid between 18 and 38% less on average than males. There were regional variances in average earnings as well, with Northern America ($4.70) and Europe and Central Asia ($3.00) having higher wages than other regions with pay ranging from US$1.33 (Africa) to US$2.2 (Asia and the Pacific). According to the results of the ILO survey, respondents chose to engage in crowd work despite the low pay for a variety of reasons, including their preference for working from home and the opportunity to supplement their income.

## Growth of crowd work in the Global South and LAC

At the beginning of their operations, most collective work platforms worked by subcontracting people from Global North countries. However, in recent years, some structural changes within the industry have led to the diversification of the workforce. One of these changes is related to the growing demand for optimized data from the automotive industry

and the emergence of specialized data generation platforms with higher levels of precision. Unlike traditional collective work platforms that only serve as mediators between clients and workers, specialist platforms guarantee the client data with an accuracy level of 99 percent. In order to produce large volumes of information with these characteristics, new strategies have been implemented to reduce costs and increase their competitiveness; investing in the optimization of processes and quality control, use of gamification mechanisms, and recruitment of workforce in new regions and countries of the Global South, such as Latin America, where lower hourly rates can be settled.

In addition to the convenience of recruiting a lower-paid labor force, the structural and social conditions of most countries in the Global South and Latin America make the collective work scheme on platforms an attractive source of income for their employees, as summarized in Figure 5.


**Figure 5. Interaction of factors for the growth of collective work in Latin America and the Global South. Own elaboration with information from Schmidt, F. (2019).**

## Platforms for collective work in Venezuela

In Venezuela, the economic crisis has fueled an explosion in investment and growth of foreign collective work platforms. During the last five years, the market conditions have led to a massive recruitment by collective work platforms in the country, which represented 75% of the workforce of companies such as Hive Micro and Spare5, equivalent to about 200,000 workers in 2018. In 2019, the list of companies interested in the Venezuelan workforce increased; for instance, the company Scale launched various recruitment strategies through social media campaigns that highlighted the possibility of obtaining an attractive, stable, and long-term income. In 2020, this same company launched Remotasks Plus, a collective work platform with tasks and an interface in Spanish initially aimed exclusively at the Venezuelan population, that was later launched to the rest of the world. The COVID-19 pandemic presented itself as a new opportunity for the growth and expansion of the workforce for collective work platforms. During this period, Scale oriented its efforts on training crowd workers; they conducted boot camps in countries from different regions, including Latin America, Asia, Arabic-speaking countries and the southern region of Africa.

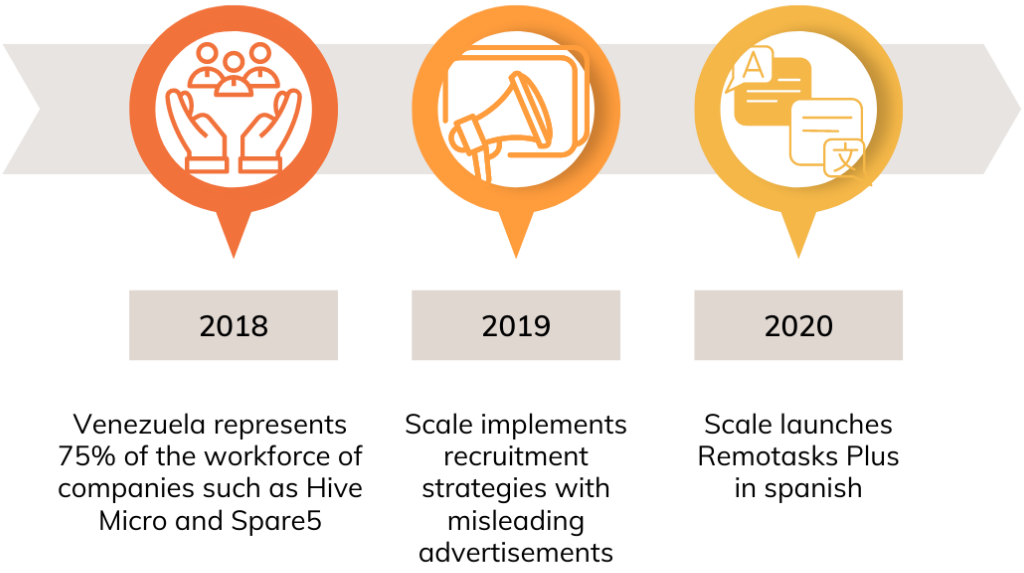



**Figure 6. Growth of collective work in Venezuela. Own elaboration. Own elaboration with information from Schmidt, F. (2019)**

## Characteristics of the labor force in the Global South

In 2015 and 2017, ILO carried out the first editions of a survey for people engaging in collective work in some of the main English-speaking platforms, such as Amazon Mechanical Turk, Microworkers, Prolific, Clickworker and Crowdflower (see Table 1). The total sample of participants contemplated 3,500 observations of people from 75 countries. Sociodemographic data of the interviewed population were collected, as well as the main motivations for carrying out this work and some of their working conditions.

**Table 1. Composition of the sample for the first and second edition of the survey of people who work in collective work carried out by the ILO. Source: Rani, U., Berg, J. & Furrer, M. (2018)**

| | 2015 (S1) | 2015 (S2) | 2017 |
|---|---|---|---|
| AMT United States | 686 | 573 | 231 |
| India | 128 | 104 | 251 |
| Other countries | 0 | 0 | 7 |
| CrowdFlower | 353 | | 355 |
| Clickworker | | | 455 |
| Prolific | | | 495 |
| Microworkers | | | 556 |
| Total | 1 167 | 677 | 2 350 |

From the disaggregated analysis of the data presented in Work and Labor Relations in Global Platform Capitalism it is possible to identify some of the characteristics of the people who carry out crowd work in the Global South. For example, it highlighted that, considering the total population participating in both surveys, the distribution by gender is 80% men and 20% women. Regarding the educational level, 45% of the people surveyed had university education, 27% postgraduate studies, and less than 13% had high school-level education or a lesser attainment.

The population surveyed in 2017 doubled the observations of the first edition of 2015. The comparison between both periods indicates that Venezuela remained the country in Latin America with the largest population working in collective work platforms, while India

occupies the same place in relation to the rest of Global South countries (see Figures 7 and 8).

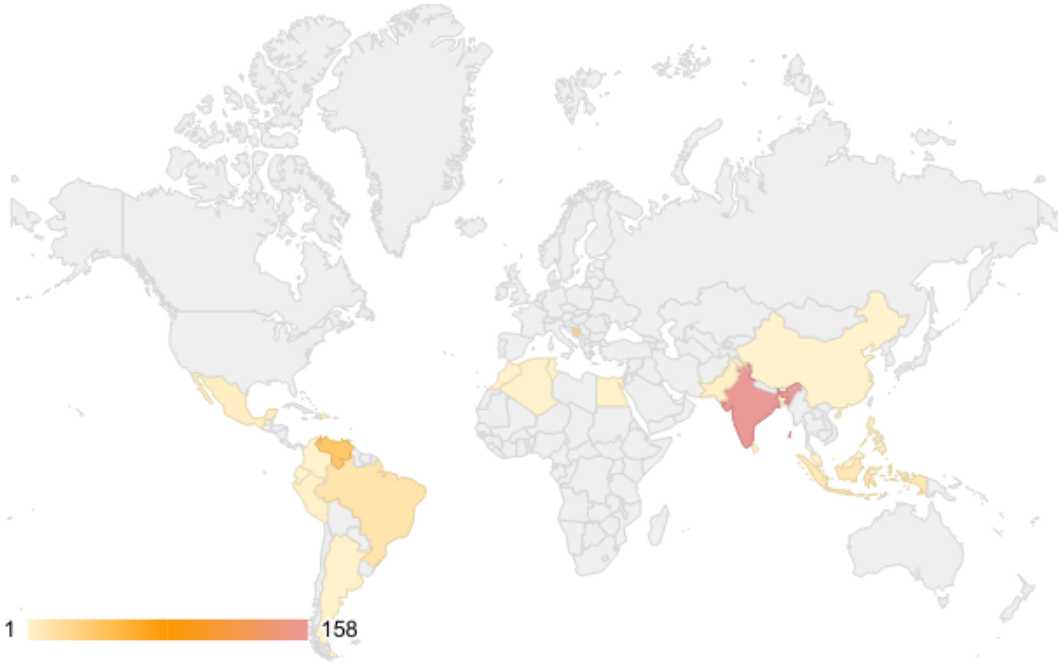



**Figure 7. Country of origin of the people participating in the collective work survey carried out in 2015 by ILO. Source: Own elaboration, with data from Rani et al. (2018)**

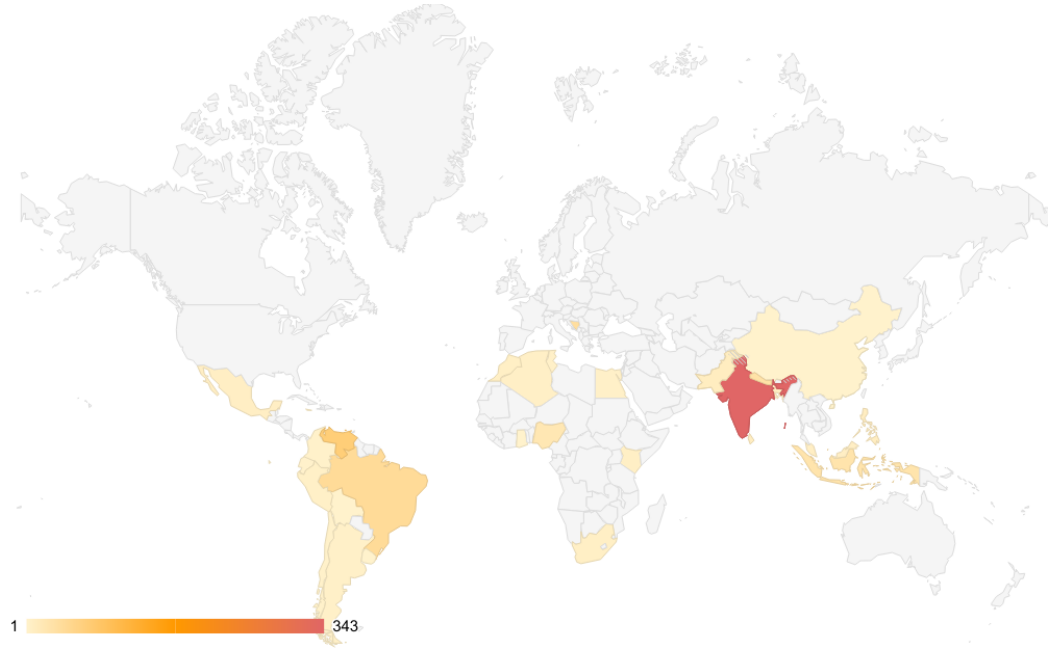



**Figure 8. Country of origin of the people participating in the collective work survey carried out in 2017 by ILO. Source: Own elaboration, with data from Rani et al. (2018)**

Low wages and lack of social protection and benefits that characterize these work schemes cast doubt on the optimistic perspective on collective work as an economic opportunity. Unlike other traditional schemes of job insecurity faced by thousands of workers around the world, this emerging digital workforce is not only limited in their salary expectations and professional growth in the short and long term, but also carries out their work behind anonymity, under a scheme of minimal institutionalization.

Although these characteristics predominate, it is worth noting that the emergence of new business models of specialist platforms represents changes in their operating logic, and this translates into a different experience for the people who work on them. Specialist platforms guarantee accurately tagged data to their customers, and reducing the costs of manual labor has been critical to achieving this goal. For this reason, the recruitment of people in Global South countries has increased exponentially in recent years.

Massive recruitment in multiple regions of the world has required the translation of tasks, as well as internal regulations and prior training so that people of different origins can participate in the execution of such tasks. More importantly, the role of mediator is transformed into that of direct supplier, and with this, changes are also generated in the employer-employee relationship and power asymmetries. In order to improve the accuracy of the data, the specialist platforms implement long training periods; the assignment of tasks is governed by a hierarchy in which the best paid tasks are also the most sophisticated, and there are specific requirements in terms of "expertise" or training necessary to access them. This gives crowd workers the alternative to distinguish themselves from their peers, specialize, and add value to their profiles.

## Gender perspective in crowd work

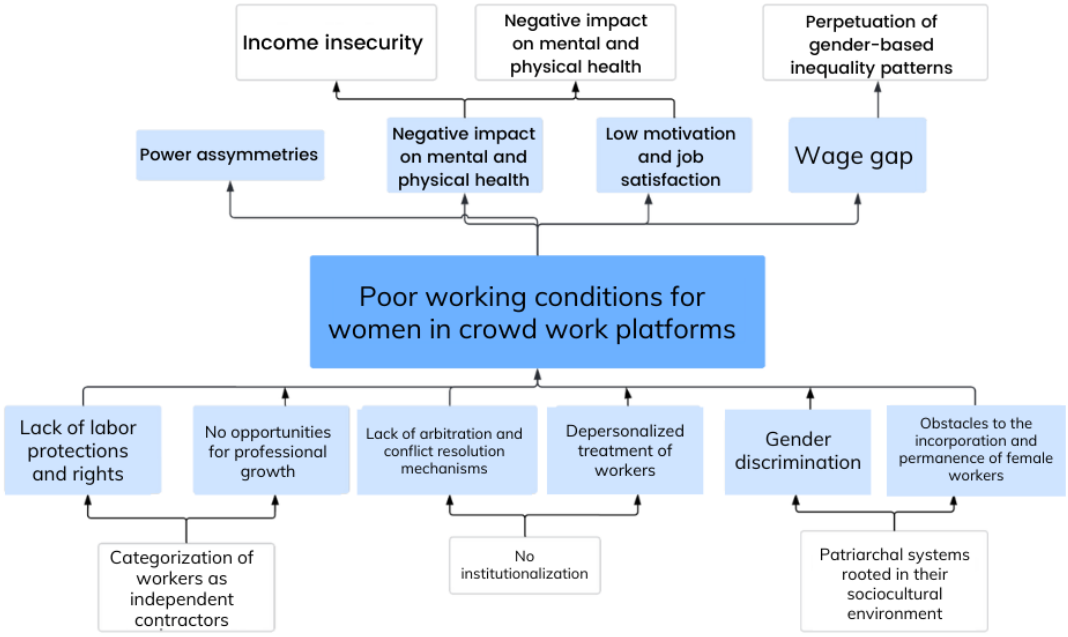



**Figure 9. Problem tree based on literature review. Source: Own elaboration.**

Collective work has represented an opportunity to increase the participation of women in economic activities, as the flexibility of remote work schemes can be a safe and viable alternative to earn income for many women. Women have reported feeling insecure when commuting, while some have been victims of harassment on public roads, or are dependent on their relatives to move from their home to any other point. Figures 9 and 10 show some of the motivations of women in the Global South to participate in collective work, as well as the limitations to work more hours on the platforms. While the preference and need to work from home, together with leisure activity and complimentary income are the main reasons women engage in crowd work, the category “other” points out to reasons that are less known and might not be taken into account when diagnosing the ecosystem.

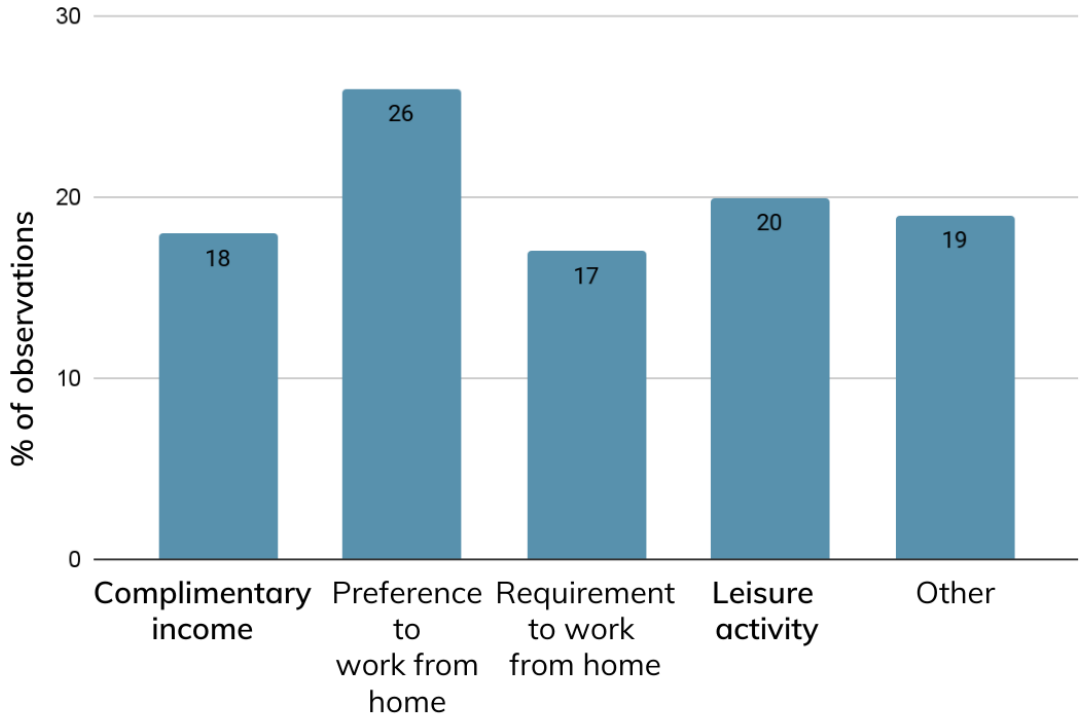



**Figure 10. Motivations of women in the Global South to carry out collective work. Source: Own elaboration, with data from Rani et al. (2018)**

It is important to note that experience as a platform worker is different for men and women. For example, in an experiment in 2022 with a group of 16 low-income Indian women to learn about their adaptation process and experiences working on a collective work platform through mobile devices, the exercise revealed that these women had a significantly longer adaptation process than women from other regions. One of the reasons for this was their fear of damaging their family's reputation by carrying out these types of activities. Their families did not regard online work as appropriate, mainly because it could translate into being exposed to other men and to unsafe spaces; they also had little trust in the use of platforms as a means to receive payments.

The freedom and autonomy of women working on platforms was significantly limited: the expectation for the women interviewed was to deliver their income to the male figures of their family who manage the household's resources. On the other hand, unlike the men participating in the study, all the women worked from home, in the presence of other

members of their family. It is no surprise that these women received little support from their families to work online and were questioned for prolonged cell phone use. This situation generated them anxiety, and required them to be in a state of constant vigilance, so as not to be reprimanded while doing collective work. To avoid these situations, the women interviewed tried to finish housework first, and then continued working on their cell phones. According to data retrieved internally by Karya, a crowd work platform that is designed for low-income, digitally-novice communities in India, 79% of women perform tasks at dawn, between 12:00 and 3:00 am. During the interviews, many women reported that by working on the platform, their rest periods had been reduced, and on average, they slept 5 hours a night, which compromised their sleep hygiene.

As part of the strategies that facilitated the process of adaptation and permanence of women in collective work, the relevance of the support of other women and the creation of spaces and communities through digital channels such as WhatsApp stood out. These spaces were useful to share doubts, experiences, and frustrations with other colleagues who also knew and worked on the platforms. For example, newly hired women usually looked for the guidance of more experienced women, who could share tools and useful technical knowledge to speed up the induction stage, as well as recommendations to achieve greater precision in the tasks performed.

The value and effect of support networks were also reflected in the percentage of tasks successfully completed by women (64%), compared to male participants (48%). Over time, the channels that were originally spaces to answer questions about the platform were used to share other concerns and personal experiences beyond collective work. In addition, these networks played an important role as spaces for support and emotional containment, essential for women to decide to continue on the platforms despite the many obstacles they face every day. For these women, the income generated through their work meant much more than a way to contribute to the family economy; an opportunity towards their financial autonomy. It also affected their self-perception, self-confidence and autonomy.

There are particular challenges conditioned by the social and structural circumstances of the region in which women crowd workers live. For example, in some regions of South Asia, women experience obstacles similar to those reported in India in terms of access to

electronic devices and autonomy in their use. Compared to men, women of the region are less likely to have a cell phone, and when they do have access to one, its use is usually restricted and supervised by male figures in the family.

## Decolonial care and well-being centered AI development

Decolonial theory compares center and peripheral power spaces that are present as colonial continuities. Ricaurte proposes an analytical model that organizes socio-technical dimensions of present coloniality through the means of data; on the "economy" dimension of the model, data labor, and the economic value of data are mentioned as expressions of coloniality. In crowd work dynamics, technology corporations take on the role of metropolises, while workers and their organization are the power-contesting periphery.

A commonly present obstacle to developing well-being centered AI systems is that current processes can obscure asymmetrical power relations and underlying values that are not questioned and replicate colonial understandings of work; crowd work gets concealed by utilitarian approaches to technology. Ethical guidelines for AI have been proposed as a means to create common knowledge and standards for socially-responsible technologies, but as Ricaurte also includes in the model, the data epistemology that precedes many of those guidelines is yet to be structurally questioned.

The design and convening of ethical guidelines for AI started in the Global North, catalyzing dialogue in the space of autonomous weapons. The discussion on contextual values in science and technology has guided the establishment of ethical values as a minimum standard: respect for persons, beneficence and justice. While such principles are helpful to ignite ethical imagination, they are not specific enough to assess the differentiated impact of AI development in the Global South, whose communities and territories face the hardest impacts of emerging technologies. Even when predominant AI actors from the Global South engage in such dialogues, the paradox of participation "wherein inclusion can exist while structural harms persist", directly challenges the development of decolonial AI.

Marie Therese Png examines the tensions of South and North in the Inclusive AI landscape, and poses three necessary steps towards AI governance that is centered on the Global South: "To 1. Engage in a historical-geopolitical analysis of structural inequality and the coloniality of

geopolitical power asymmetries and international legal frameworks; 2. Co-construct roles for Global South actors to substantively engage in AI governance processes; and 3. Identify mechanisms and protocols that mitigate "paradoxes of participation" and redress institutional power imbalances, in order to meaningfully engage with underrepresented stakeholder groups".

Rather than enumerating more ethical principles, this paper-to-prototype process is oriented to engage and work around the tensions that arise when putting them in practice. Critical science can help foresight the prospective harms of crowd work in the Global South, and lay the foundation for a participatory, slow, reflective and cooperative development of empowerment tools.

## Policy Recommendations

***Crowd work platforms***

- **Allow for the filtering and disaggregation of data.** This is critical to allow researchers, policymakers, together with decision makers within the private sector to better understand the characteristics of crowd workers by region, countries, and other variables of interest.
- **Establish transparent standards for assigning equitable payment**. Stipulated national and regional minimum hourly wages must be taken into account in the calculation of payment per task conducted. Following a formula as the one we used can be a first step into discussing what fair payment could entail and look like.
- **Support and promote English skill building and training to non-English speaking crowd workers.** Doing so could lead to a higher task completion and accuracy, honing crowd workers skills.
- **Design and implement Education Programs for workers.** Crowd workers can benefit from learning about the applications of the data and information they provide to requesters through education programs in STEM, the social sciences and business.
- **Establish channels, mechanisms, and processes to improve communication with crowd workers.** In addition to features that enable communication with requesters, the platform provider should establish processes and mechanisms to provide valuable information to crowd workers.

*International organizations*

- **Update estimates on the size of the collective labor force.** Richer data will include an intersectional perspective that takes gender, race, class, and disability as variables of interest globally, disaggregating information regionally.
- **Promote crowd workers organization.** Support initiatives that connect stakeholders from academia, civil society and policy making of the crowd work space with workers to promote organization.
- **Mitigate the paradox of participation.** Identify mechanisms that redress institutional power imbalances within initiatives that aim to include and benefit crowd workers.

*Policy makers, researchers and community leaders*

- **Promote equitable crowd work and governance of AI in the Global South.** Co-construct roles for Global South actors to engage in AI governance, centering on the well-being of crowd workers in varying contexts.
- **Exchange best practices.** Share insight and knowledge with platforms' policy teams, with other researchers and leading voices in the crowd work space.
- **Co-create a feminist community of practice.** There is a need to continue growing the body of crowd work research through a gender lens; to build on the existing research, the creation of a feminist community of practice can support nascent voices in the field.

## Conclusion

The development of Artificial Intelligence relies heavily on the crowd-sourcing of data from international labor. While crowd work platforms such as Hive Micro, Appen, Amazon Mechanical Turk, Spare 5 and Scale in the Global South may provide employment opportunities for women, there are still issues of association opportunities, collective bargaining, transparency, fair pay, and professional development that need to be addressed. Additionally, most prior work has not focused on documenting the experience of women crowd workers in contexts like Latin America, to create technology that can truly empower them. In this research, we focused on conducting a qualitative study with more than 60 Latin American gig workers to understand the need, challenges, and opportunities they face in

crowd work. We conducted one of the first studies to uncover the unique perspectives of these populations, presenting key insights and providing design implications for creating AI that can empower women in the Global South. We also provide policy recommendations for platforms, international organizations, policymakers, researchers, and community leaders to promote more equitable and gender-responsive crowd work.

## Acknowledgments

This work was possible thanks to the support of the Feminist AI Network and a grant by the International Development Research Centre (IDRC). It was also partially supported by NSF grant FW-HTF-19541 where our technical team is studying how to design intelligent interfaces for hispanic rural workers.

Special thanks for their insight and constructive criticism:

- Feminist AI Network: Paola Ricaurte, Mariel Rosauro, Caitlin Kraft-Buchman & Jaime Gutiérrez
- Our interns at PIT Policy Lab: Itzel Laurel & Sarah Owolebi
- Toloka's Educational Program team members: Natalia Fedorova, Elena Johnson

# References


Adams-Prassl, A. & Berg, J. (2017), When Home Affects Pay: An Analysis of the Gender Pay Gap Among crowd workers.. http://dx.doi.org/10.2139/ssrn.3048711

Berg, J., & Rani, U. (2021). Chapter 4: Working conditions, geography and gender in global crowd work. En "Work and Labour Relations in Global Platform Capitalism" . Cheltenham, UK: Edward Elgar Publishing. https://www.elgaronline.com/view/edcoll/9781802205121/9781802205

Bourdieu, P. (1990). The Scholastic Point of View. Cultural Anthropology 5(4): 380-391.

Collins, P. H., & Bilge, S. (2020). Intersectionality. John Wiley & Sons.

Drahokoupil, J., and Vandael, K. (2021). Labour and the Platform Economy. Edward Elgar, Cheltenham. 1-46

Durward, Blohm & Leimeiste (2020) The Nature of Crowd Work and its Effects on Individuals' Work Perception, Journal of Management Information Systems, 37:1, 66-95. https://doi.org/10.1080/07421222.2019.1705506

Estellés-Arolas, E., & González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. Journal of Information Science, 38(2), 189–200

European Institute for Gender Equality (nd). Gender Mainstreaming. https://eige.europa.eu/gender-mainstreaming

European Institute for Gender Equality (nd). Gender Analysis. https://eige.europa.eu/taxonomy/term/1143

Fieseler, C., Bucher, E. & Hoffman, C. (2017). Unfairness by design? The Perceived Fairness of Digital Labor on crowd working Platforms. Journal of Business Ethics. 156.

Fredman, S., et.al. (2021). International Regulation of Platform Labor: A Proposal for Action. Weizenbaum Journal of the Digital Society, 1(1), w1.1.4.

Haro, K. & Hernández, A. P., (2022). How the AI profits from catastrophe. MIT Technology Review. https://www.technologyreview.com/2022/04/20/1050392/ai-industry-appen-scale -data-labels/

Hastrup, Kirsten (1992). Out of anthropology: the anthropologist as an object of dramatic representation. Cultural anthropology, 8/1/1992, Vol. 7, Issue 3.

Hesse-Biber, S.N. (Ed.).(2014). Feminist Research Practice: A primer (2nd. ed.). SAGE Haraway, Donna. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. Feminist studies 14 (3), 575-599.

Idowu, A., Elbanna, A. Digital Platforms of Work and the Crafting of Career Path: The crowd workers' Perspective. Inf Syst Front 24, 441–457 (2022). https://doi.org/10.1007/s10796-020-10036-1

International Labor Office. (2019), 'Work For a Brighter Future: Global Commission on the Future of Work'.

Kittur, A., Nickerson, J., Bernstein, E., Shaw, A., Zimmerman, J., Lease, M. & Horton, J. (2013). The future of crowd work. In Proceedings of the 2013 Conference on Computer supported cooperative work (CSCW '13). Association for Computing Machinery, New York, 1301-1318.

Mansbridge, J. J., & Okin, S. M. (2020). Feminismo: breve introducción a una ideología política. Página indómita.

Marie-Therese Png. (2022). At the Tensions of South and North: Critical Roles of Global South Stakeholders in AI Governance. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 1434–1445. https://doi.org/10.1145/3531146.3533200

Margaryan, A. (2017). Understanding crowd workers' Learning Practices. Paper presented at 17th Biennial Conference of the European Association for Research on Learning and Instruction. EARLI 2017, Tampere, Finland

Mohamed, S., Png, MT. & Isaac, W. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. Philos. Technol. 33, 659–684 (2020). https://doi.org/10.1007/s13347-020-00405-8

Nickerson, Jeffrey V. (March 20, 2013). Crowd Work and Collective Learning. A. Littlejohn & A. Margaryan (eds.), Technology-Enhanced

Professional Learning: Routledge, Forthcoming, Available at SSRN: https://ssrn.com/abstract=2246203

O'Higgins, N. & Pinedo Caro, L. (2022). crowd work for young people: Risks and opportunities. ILO Working paper 50. https://www.ilo.org/global/publications/working-papers/WCMS_837670/lang--en/index.html

Paolacci, Chandler, Ipeirotis (June 24, 2010). Running Experiments on Amazon Mechanical Turk. Judgment and Decision Making, Vol. 5, No. 5, 411-419, Available at SSRN: https://ssrn.com/abstract=1626226

Rama A., Divya S., Vivek S, Kalika B., & Aditya V. (2022). Feeling Proud, Feeling Embarrassed: Experiences of Low-income Women with Crowd Work. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22). Association for Computing Machinery, New York, NY, USA,298, 1–18. https://doi.org/10.1145/3491102.3501834

Rani, U., Berg, J., Harmon, E., Silberman, M. & Furrer, M. (2018). Digital labour platforms and the future of work: Towards decent work in the online world. International Labour Office

Ricaurte, P. (2019). Data Epistemologies, The Coloniality of Power, and Resistance. Television & New Media, 20(4), 350–365. https://doi.org/10.1177/1527476419831640

Schmidt F., A. (2019). Crowdsourced Production of AI Training Data: How human workers teach self-driving cars how to see. Forschungsförderung 155, Hans-Böckler-Stiftung, Düsseldorf.

Scott, J. W. (2008). Género e Historia. Fondo de Cultura Económica.
Sheldon, R. (2016). Tragic encounters and ordinary ethics. Manchester University Press
Toloka. (s.f.) Toloka documentation. https://toloka.ai/en/docs/

Toloka. (s.f.) Toloka documentation: Filters. https://toloka.ai/en/docs/guide/concepts/filters

Toloka. (s.f.) Toloka documentation: Messaging. https://toloka.ai/en/docs/guide/concepts/messagin

## Biographies

**Cristina Martínez Pinto** is the Founder and CEO of PIT Policy Lab. She worked as a consultant for the Digital Development Global Practice at the World Bank, led the AI for Good Lab at C Minds and co-founded Mexico's National AI Coalition IA2030Mx. She holds a Master's degree in Public Policy from Georgetown University and a Bachelor's degree in International Relations from Tec de Monterrey.

**Luz Elena González Zepeda** specializes in Development Cooperation, with a focus in tech policy for the LAC region. She managed the PIT Policy consultancy with the Government of Colombia to design a Governance Model for the Data Infrastructure. She has a B.A in International Relations from Tec de Monterrey, in Mexico. She is also content curator for Women in AI Mexico, and event promoter within the MCODER.ai community for women in data science.

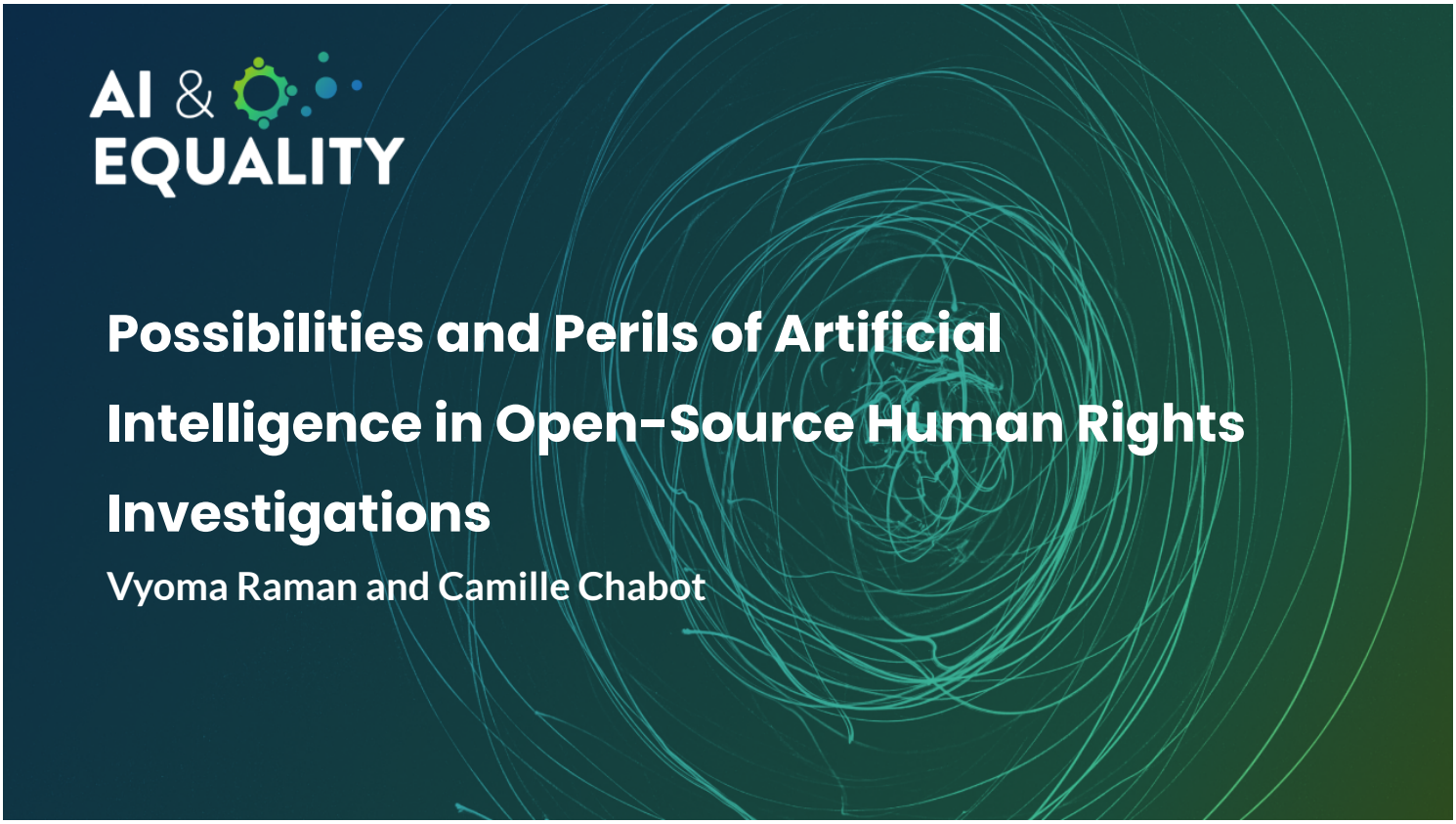
**Tatiana Telles** is Puentech Lab's Public Policy and Gender Specialist, with 10 years of experience in Government and Civil Society. Focused on problem solving. Feminist. She seeks to create more egalitarian communities.

**Norma Elva Chávez** is a distinguished researcher and educator based at the National Autonomous University of Mexico (UNAM). She serves as the director of both the UNAM Civic AI Lab and the Dispositivos Logicos Programmables lab. Her research primarily focuses on the design of robots and technologies aimed at empowering women, enabling them to achieve greater success. dShe has dedicated her career to understanding the biases and challenges faced by women, and using that knowledge to develop innovative technologies that support women in their diverse goals.

**Maya De Los Santos** is a first-generation Afro-Latina pursuing an Electrical and Computer Engineering degree in Northeastern University's Honors Program and is a research assistant in the Civic AI lab led by Dr. Saiph Savage. In her current work, Maya is dedicated to designing and researching human-centered AI systems that ensure fair work opportunities for Latina gig workers. Maya is also a Google Generation Scholarship Recipient, a Grace Hopper Celebration Scholar, and was selected for an NSF grant to participate in Carnegie Mellon University's REUSE research program.

**Alberto Navarrete Hernández** works as a professor at the National Autonomous University of Mexico in the Faculty of Engineering, in the Department of Computer Science, teaching digital systems design. He is also an assistant in the laboratory of Programmable Logic Devices, performing in the design of laboratory practices of the subjects. His field of expertise and main interest is digital signal processing and data analysis, using Machine Learning techniques.

**Saiph Savage** is an Assistant Professor at Northeastern University in the Khoury College of Computer Sciences, where she directs the Northeastern Civic A.I. Lab. The impact of Dr. Savage's research has led to her being named one of the 35 innovators under 35 by the MIT Tech Review, be awarded large grants from the National Science Foundation (NSF), and enjoy recognition from UNESCO for having globally one of the most impactful AI research projects. Her work has also won paper awards at top scientific venues in human-computer interaction, including ACM CHI, CSCW, and the Web Conference. The growing public interest in Dr. Savage's research has garnered international press coverage from the BBC, New York Times, The Economist, Deutsche Welle, Vice, Wired, Forbes, and Fortune.



# Possibilities and Perils of Artificial Intelligence in Open-Source Human Rights Investigations

Vyoma Raman and Camille Chabot

## Abstract

This article provides an overview of the role of social media, artificial intelligence (AI), and large language models (LLMs) in open-source human rights investigations. Drawing on our experiences as lead student researchers at the Human Rights Center's Investigations Lab at UC Berkeley School of Law's, we reflect on the transformative potential of social media in uncovering and documenting human rights violations, particularly in contexts where traditional investigative methods face limitations. We highlight the importance of ethical considerations and describe the Berkeley Protocol on Digital Open-Source Investigations, a guiding framework for online research. We delve into the current applications of AI in open-source investigations, focusing on the role of algorithms in social media and how human rights researchers manipulate these algorithms to curate relevant media. We examine AI's contribution to data collection, noting its ability to accelerate the process of gathering large volumes of data but also highlighting challenges in verifying data relevance and accuracy. We discuss the use of AI in content verification for reverse image searching, deep fake detection, and language translation.

Furthermore, we explore the potential use of LLMs in open-source human rights investigations. LLMs offer opportunities for enhanced content discovery, verification, and research resiliency. They can assist in understanding user intent, analyzing diverse sources, and providing comprehensive results. LLMs also aid in content verification by rapidly analyzing and summarizing data, identifying patterns, and facilitating report generation. Moreover, LLMs have the potential to mitigate the emotional impact of research on investigators by identifying and warning about distressing content. However, we emphasize the need for careful consideration of the benefits, limitations, and ethical implications associated with using LLMs. Monitoring for biases, misinformation, and interpretability challenges is crucial. We underscore the importance of supplementing LLM outputs with diverse sources, addressing biases, and ensuring transparency in legal contexts. While LLMs offer promise in enhancing open-source human rights investigations, their use should be judicious, complemented by human expertise, and subjected to rigorous ethical scrutiny.

## Introduction

2017 marked a watershed moment [1] in the application of open-source information (OSI) to human rights investigations: the International Criminal Court issued the first arrest warrant reliant on social media evidence, targeting Libyan military officer Mahmoud Mustafa Busayf Al-Werfalli. OSI2 encompasses information that one can readily acquire from the internet, available to any member of the public without the need for special legal status or unauthorized access. The warrant hinged on a series of videos obtained from social media, evidencing seven executions in Benghazi between June 2016 and July 2017, which implicated Al-Werfalli in war crimes. This case underscored the transformative potential of social media in the field of human rights investigations, a potential that is ripe for further enhancement with the emergence of artificial intelligence.

As lead student researchers at the Human Rights Center's (HRC) Investigations Lab at the UC Berkeley School of Law, we use OSI to investigate human rights violations in locales such as Iran, Western Sahara, Latin America, and the U.S. The HRC has played a pioneering role in harnessing OSI for human rights violation investigations. Our training in open-source investigation methods has fueled our involvement in several projects probing human rights violations in partnership with NGOs, international organizations, courts, and journalists to uncover and fact-check public information relevant to human rights crises.

## Open Source Investigations

As open-source investigators, our everyday work involves probing social media to unearth and document human rights violations that traditional investigative methods, like interviews and on-the-ground inquiries, often fall short of addressing.

### *Social Media's Revolution for Human Rights Investigations*

The expansive reach of social media has revolutionized investigative practices adopted by legal experts, journalists, and human rights activists. The proliferation of civilian-generated visual content enables the use of real-time information in reports and legal proceedings, narrowing the explosion of first-hand accounts and citizen-driven, unofficial narratives of

human rights violations amplifies our capacity to form a comprehensive understanding of on-ground realities, thus laying the groundwork for accountability.

Though open-source investigations frequently complement traditional research methods, they truly shine in contexts where hostility on the ground stifles conventional media and in-person investigations. In such cases, they offer a unique window into otherwise unreachable incidents. To illustrate, our partnership with Amnesty International facilitated an open-source investigation into potential crimes against humanity in Iran amid ongoing protests triggered by the in-custody death of Mahsa Amini. The constraints of activist detention and censorship by the Islamic Republic rendered on-the-ground work unfeasible.

Lastly, the dissemination of incidents on globally accessible social media platforms has broadened public awareness of violations far beyond the reach of conventional reports. The real-time, multi-perspective documentation of human rights incidents through social media has spurred the rise of citizen science in the field of human rights, empowering ordinary citizens to document human rights violations worldwide using their connected devices. Entities like Bellingcat, a Netherlands-based investigative journalism organization, and Amnesty International's university network, the Digital Verification Corps, exemplify this pioneering approach of mobilizing newly-trained citizens to conduct human rights investigations using OSI.

***Methods and Ethics in Open-Source Investigations***

Open-source investigations stand at the precipice of continuous evolution, demanding investigators to persistently acclimate to new methods and technologies. Yet, the bedrock principles of these investigations have proven resistant to the tides of time.

One such cornerstone is the Berkeley Protocol on Digital Open-Source Investigations3, a comprehensive set of guidelines for professionally and ethically conducting online research into alleged human rights violations and international crimes. The Protocol provides guidance on methodologies for collecting, analyzing, and archiving digital information within the context of human rights investigations. It also focuses on safeguarding the physical, psychological, and digital well-being of online investigators and first responders, recognizing that their work may place them in potentially threatening situations.

Effective and ethical open-source investigations hinge on diligent preparation. Investigators commence their work by assessing potential risks and threats and examining the digital media, and platforms used on the ground, the key actors involved, and the vernacular surrounding specific violations. Particular consideration is also given to technology accessibility to discern any discrepancies between online information and on-the-ground realities. For instance, the well-documented gender digital divide might result in underrepresentation of violence against women on social media platforms as compared to violence against men. This preparatory phase enables researchers to pinpoint potential risks, biases, and gaps, and counter them using various mitigation strategies.

Open-source investigations typically begin with the discovery of content relevant to the investigation's objective. Investigators utilize a combination of keywords in pertinent languages to find written or visual content that addresses the central questions of the investigation across various online and social media platforms. Tools like TweetDeck, CrowdTangle, and Boolean searches are instrumental in sifting through multiple media platforms and accessing user-generated posts and documents.

The adept use of these tools can significantly contribute to confirming established media trends and unearthing content that has been under-reported. For example, during our investigation of Title 42 in partnership with Human Rights First, we aimed to unveil occurrences of lesser-reported violence perpetrated against asylum seekers affected by the emergency health law at the Mexican border. Our focus spanned a range of incidents, from sexual violence to attacks against LGBTQ+ individuals. The pervasive social taboo surrounding these forms of violence often inhibits victims from discussing their experiences with traditional media outlets. Therefore, open-source investigations provide a privileged avenue to bring these concealed incidents to light.

When gathering content, it is paramount to adopt practices that deliberately counteract confirmation bias. This is achieved by involving researchers from a range of backgrounds and expertise and by utilizing VPNs to diversify search results. As discovery primarily hinges on keywords, different investigators can yield disparate outcomes based on their keyword selection, their chosen platforms, and their device's location. Therefore, to amass comprehensive evidence reflecting the actual situation on the ground, it is vital to employ

carefully selected research teams and connect to multiple servers using VPNs. Researchers also often employ sock puppet accounts or fictitious online identities to offset data-driven personalization and safeguard their online and physical safety.

Leveraging user-generated content as reliable evidence for advocacy, trials, or journalistic stories, requires verification for authenticity. It is not uncommon for users to re-use photos and videos from past incidents or present misleading contexts, making it crucial for online investigators to fact-check the location (geolocation) and time (chronolocation) of visual content.

Tools like InVid are employed to help researchers determine the first instance a video or image was posted, locate critical frames in each video, and enlarge parts of images for various verification tasks. Reverse image search engines like Yandex and Google, and 3D map software like Google Earth Pro and PeakVisor, assist in finding correlations between visual content and identifiable locations, and in pinpointing precise coordinates. Investigators might also need to discern the precise moment an incident occurred; apps like SunCalc are used to estimate the day and time based on the sun's position. Lastly, in scenarios where content is at risk of being removed, such as graphic content taken down from social media platforms, investigators archive and protect it from destruction using tools like Hunchly.

During the investigation process, online investigators may encounter distressing material and vivid descriptions of human rights violations, which can lead to secondary trauma and PTSD. The Berkeley Protocol provides a set of guidelines to mitigate these risks. The Protocol recommends that researchers should be aware of their own and their colleagues' typical behavior, noticing any changes in eating, sleeping, and recreation habits. They can also employ various techniques to minimize exposure to harmful content, such as turning off audio, minimizing the screen, hiding violent material, using grayscale mode, working in pairs, and avoiding late-night work. Furthermore, researchers strive to foster a sense of community and camaraderie, which is crucial for maintaining good mental health in online investigations4.

## Current Uses of AI in Open-Source Investigations

The essence of artificial intelligence is deeply intertwined with the current methodologies adopted for discovery and verification in open-source investigations. The tools that we deploy in our research frequently lean on algorithms to sift through, amass, and analyze pertinent open-source content.

### *Algorithms in Social Media Feeds*

Algorithms serve as the unseen puppet masters of the social media experience, subtly molding the data landscapes navigated by users on a daily basis. These algorithms leverage vast data repositories of past interactions and user similarities to tailor content aimed at captivating and retaining user attention. News feed algorithms, which are AI systems determining the most engaging and pertinent content to exhibit in a user's news feed or timeline, embody this concept of data-driven personalization. They significantly dictate the flow and nature of information range of factors, such as the posts users engage with, the frequency of interactions with certain users, the time spent on various types of posts, and even the speed of scrolling. Consequently, these algorithms curate a personalized feed comprising videos, images, and other posts, fine-tuned to a user's preferences and geared towards prolonging their time spent on the platform.

Human rights researchers have discovered ways to manipulate these algorithms for purposes that transcend simple content consumption. For instance, a frequently employed technique to curate relevant media involves creating artificial social media profiles, or "sock puppets," which strategically interact with specific content. These profiles typically don't reflect the personal interests of their creators; rather, creators systematically engage with a distinct research topic of interest to unearth relevant content and analyze the surfaced information. By interacting with posts, following pages, and clicking on content that aligns with a specific theme, researchers can manipulate the news feed algorithm to present information that corresponds with the sock puppet's designed interest.

We have integrated this methodology into our human rights research, delving into topics as diverse as student protests in Iran and police brutality in Western Sahara. For example, while investigating an incident of excessive force in Smara, we employed a sock puppet with benign interests in sports to discreetly and anonymously probe evidence of Sahrawi activism

and authorities' reaction to it. By selectively viewing videos and interacting with posts from Moroccan authorities and Sahrawi activists, we began noticing similar content appearing in our feed. Thus, we were able to manipulate the news feed algorithm to curate content relevant to our investigation.

Despite their utility, news feeds can inadvertently propagate detrimental narratives. Designed to favor engaging content, these algorithms might unwittingly amplify misinformation or propaganda that elicits strong emotional reactions and, in turn, user engagement. This could potentially lead to the creation of echo chambers, where users are consistently exposed to content that reinforces their existing beliefs, thus obstructing the dissemination of accurate information.

### *Data Collection Using AI*

The recent advent of artificial intelligence has tremendously accelerated the process of data scraping from myriad sources, far surpassing the pace achievable by researchers manually conducting discovery. AI-driven data collection operates on the same fundamental principles as traditional discovery, primarily the employment of keywords and specific time frames to amass pertinent content. However, its transformative edge lies in its ability to gather substantially larger volumes of data within a significantly reduced timeframe. For instance, Amnesty International's

Digital Verification Corps leaned on an AI data scraping tool in their investigation into police brutality in Iran following the custodial death of Mahsa Amini. The combination of an internet shutdown and the sporadic nature of data availability posed formidable challenges to procuring evidence of crimes against humanity. The AI data scraping tool, configured with keywords provided by Amnesty International's Iran researchers, managed to collect and systematically arrange data on over three thousand incidents.

However, the incorporation of AI in data collection has brought forth its own set of unique challenges that researchers must grapple with. In this investigation, we found ourselves burdened with verifying the relevance of the massive volume of data collected. This was due to the scraping tool's occasional misattribution of media to inaccurate locations or time frames or the inclusion of content irrelevant to the investigation. Additionally, the imperative

to ensure a diversity of perspectives in the discovery process equally applies to the development of the data scraping tool. Therefore, special care must be taken with respect to who designs the tool and the keywords they deploy, to avoid skewing outcomes and risking partiality.

### *AI Content Verification Tools*

In addition to gathering content, AI systems such as Yandex and Google Reverse Image Search have emerged as resourceful allies for verifying visual content. These tools use computer vision algorithms to capture the different features of a given image—its colors, shapes, textures, and more—to create a unique fingerprint. The AI compares this fingerprint to those within its extensive database, seeking similar patterns. The harmony between AI and computer vision technologies thus enables these tools to swiftly pinpoint duplicates, altered images, or potential sources, bolstering the efficacy of content verification.

Verifying content also involves confirming that it was not artificially generated as a fake or deep fake. InVid has integrated multiple AI filters into its forensic capabilities to detect additions and modifications in images. These employ machine learning algorithms that sharpen their understanding of vast datasets of genuine and manipulated images. Instead of merely analyzing an image, the AI scrutinizes its intricate aspects, such as statistical patterns and visual attributes, looking for signs of tampering. By comparing the altered image with a reference or original image, the system can detect discrepancies in pixel-level details, inconsistencies in lighting and shadows, anomalies in noise patterns, or artifacts introduced during the editing process. These filters act as red flags, highlighting areas of the image that are likely to have been modified, thereby assisting investigators in identifying potential manipulations and deep fakes.

AI plays a crucial role in investigations conducted in foreign languages by enabling the translation of words from both text and images. Popular tools like Yandex and Google Translate are widely used for this purpose, proving particularly valuable when dealing with ideographic languages, where researchers cannot simply retype the text unless they have familiarity with the language. However, it's important to acknowledge that these translation tools are not exempt from gender bias. Gender bias in translation arises when AI algorithms or the datasets used to train translation models exhibit biases in how they handle

gender-specific language or cultural nuances linked to gender. For instance, certain languages may have grammatical rules or sentence structures that convey gender information, and if the translation models are not appropriately trained or calibrated, they might inadvertently reinforce or perpetuate gender stereotypes. To ensure fair and accurate translations for all users, it is essential to address and mitigate gender bias through continuous research, data curation, and algorithmic improvements.

## Leveraging Large Language Models (LLMs) for Human Rights Investigations

As we explore the potential application of large language models (LLMs) like GPT-4 in open-source human rights investigations, we enter a realm of cautious optimism. These AI systems have the potential to be intricate data processing tools due to their capabilities of context recognition, semantic interpretation, and pattern detection. The prospect of transforming open-source investigation is tempting, but it is vital to navigating this path with measured steps, acknowledging the possible benefits, inevitable limitations, and the need for careful supervision.

***LLMs in Content Discovery***

The initial stage of any open-source human rights investigation is akin to navigating a labyrinth of online information. Traditionally, investigators rely on their expertise and intuition, employing search engines, social media platforms, and similar resources as described earlier. Investigators use keyword searches, explore hashtags, follow leads from related articles or posts, and track certain individuals, organizations, or locations over time. While these methods can yield useful information, they can be time-consuming in an urgent situation and still leave investigators open to the risk of missing crucial details.

In this context, LLMs propose a different approach. They don't merely align with keywords or phrases like traditional search algorithms. Instead, they attempt to decode the underlying intent behind a user's query. This includes understanding the context, recognizing sentiment and connotation, and, consequently, producing results that may be more accurate and comprehensive.

We contributed to OSI discovery and verification efforts of police brutality in Chile in October 2019. While we employed a variety of boolean search techniques to produce results, an LLM could augment this process. By recognizing the deeper context of the query, could assemble a broader spectrum of related information. This might include civilian documentation of police violence, related protests, official responses, legal proceedings, and public reactions. Beyond understanding user intent, LLMs are also adept at comprehending and synthesizing information from diverse sources of online media. They can analyze text, identify key themes and entities, extract relevant facts, and even summarize lengthy documents. This ability to process and make sense of large amounts of information can be a game-changer for investigators, helping them sift through the online information deluge more effectively. However, their efficiency doesn't replace the discerning eye of an investigator but supplements it by helping them navigate the digital deluge more effectively.

By combining their understanding of user intent and online media, LLMs can serve as supplementary tools for investigators to discover more relevant content on the Internet. They can extract first-hand social media accounts of an incident, find a particular type of media (e.g. videos) about it, and aggregate and summarize relevant press releases. In addition, they could suggest related topics or entities to explore, highlight emerging trends or patterns, or even point out inconsistencies or gaps in the available information that warrant further investigation.

For instance, in our investigation of murders of Indigenous environmental defenders in the Amazon basin, an LLM could have hastened our identification of related issues that have caused the violence, such as the illegal lumber trade, agriculture lobbying, and demand for transition minerals used for renewable energy. It could also surface content from lesser-known or non-English sources, thereby providing a more diverse and comprehensive view of the situation.

Despite these possible advantages, the use of LLMs demands continual monitoring and recalibration. The issue-laden release5 of Microsoft's Bing AI serves as a reminder of the potential unpredictability of LLMs. The risk is real: these models can reflect and even amplify biases, hostility, and misinformation prevalent in their training data, which often includes a representative, yet flawed, slice of the internet. In the face of these risks, we must approach

LLMs not as a panacea, but as a potentially valuable tool that must be wielded with care and vigilance.

### *LLMs in Content Verification*

The journey through an open-source human rights investigation, while undeniably important, is often intricate and laborious. A critical juncture of this journey is the verification stage, where investigators sift through the data they discovered, analyze it, and craft comprehensive reports to corroborate incidents. In this process, human researchers carry the bulk of the load, meticulously picking through the information piece by piece – a task whose time-consuming nature often conflicts with its need to respond to rapidly evolving situations.

Here, LLMs could potentially serve as valuable adjuncts. Their ability to parse through colossal amounts of data swiftly suggests a more efficient path to unearthing the facts surrounding a human rights violation. Imagine an LLM rapidly condensing a spreadsheet populated with thousands of entries into a concise and comprehensible summary, outlining vital incident parameters such as primary locations, types of violations, prevalent keywords, incident dates, and sources. This goes beyond merely repackaging raw data into a user-friendly format; it allows investigators to glean hidden patterns and trends, thereby enabling the insightful navigation of an investigation. The recognition of a sudden increase in specific types of violations, or the emergence of new hotspots, can shape the direction of an investigation. The capacity for LLMs to digest content and produce coherent written summaries make them more powerful for this work than traditional data analysis.

LLMs' language versatility is also a significant boon. Human rights incidents span the globe, and LLMs' capability to automatically translate content can potentially lessen the dependency on human translators, lending efficiency and scalability to the process.

An integral part of open-source investigations is the production of detailed verification reports. These meticulous documents, crucial for any ensuing legal proceedings, outline the process by which each piece of evidence was found and fact-checked. With machine-assisted workflows, investigators could potentially find original online sources for discovered content more easily, interpret multilingual content, identify video locations by

providing written descriptions of specific landmarks, and integrate their findings into a coherent, standardized report. This could streamline the report generation process while maintaining consistency across different cases.

Guided by clear investigative objectives, LLMs can possibly optimize the verification process by identifying and prioritizing content that aligns with the investigation's goals. Their understanding of content, context, and intent, allows them to discern and rank different scenarios across diverse forms of content. For instance, in video content, an LLM could distinguish a peaceful protest from a violent altercation based on post descriptions and comments. This discerning capability becomes paramount in human rights investigations where the evidence of violence and abuses are typically the central focus.

Let's consider an LLM trained on post descriptions of varying crowd situations and their comments. It learns to recognize textual patterns and signals associated with violent incidents, such as specific actions, language, or emotional tone. When presented with new posts, the LLM could analyze associated textual metadata, comments, or transcriptions, and rank the media by the likelihood of containing relevant content. This could allow investigators to focus their efforts on the most promising leads, potentially improving the efficiency of the investigative process.

Despite these potential advantages, we must remember that LLMs are not infallible. They are tools that could amplify our efforts, but they also require vigilant oversight and rigorous testing. As we walk the fine line between potential benefits and pitfalls, we must strive to leverage these technologies judiciously and ethically in our mission to uphold human rights.

***LLMs in Research Resiliency***

The pursuit of truth in open-source human rights investigations, while undeniably vital, carries a heavy emotional burden. Investigators are routinely exposed to graphic and distressing content, whether it is a video capturing brutal violence or a chilling first-hand account. This constant exposure can lead to vicarious trauma, a condition akin to post-traumatic stress disorder affecting individuals who regularly witness the traumatic experiences of others. Compounded by the ceaseless nature of this work, fueled by a profound commitment to human rights, the risk of burnout becomes a grim reality. Here,

investigators may feel trapped, unable to step away without the gnawing fear of compromising their mission.

In this context, LLMs may present a potential shield by helping to mitigate the impact of traumatic content on investigators. By analyzing and sifting through vast amounts of data, these AI systems can identify content relevant to the investigation, potentially reducing investigators' exposure to irrelevant graphic material. Additionally, they can be trained to identify potentially distressing content and provide advance warnings, allowing researchers to prepare themselves before engaging with such material.

Consider an LLM trained to discern textual and visual cues indicative of violent or traumatic incidents, such as descriptions of violent acts, blood, or explosions. It could be programmed to flag descriptions of violent acts or visual depictions of blood or explosions, subsequently alerting investigators about the nature of the content they are about to encounter. This preemptive caution creates a protective buffer, affording investigators a chance to either mentally brace themselves or delegate the task if they feel ill-equipped to handle it at that moment.

We have previously worked on developing a video-viewing platform with capabilities like facial blurring, object tracking, audio analysis, and grayscaling. An LLM could be integrated into this to display content warnings and offer suggestions for less distressing ways of engaging with content. If an LLM identifies the presence of graphic elements like blood in a video, it might recommend viewing the content in grayscale to lessen the graphic impact. Similarly, it could propose muting the audio if it detects potential auditory triggers such as explosions or gunshots. We have worked on developing a video-viewing platform with these capabilities.

While these adjustments might seem small, they can significantly contribute to preserving the emotional well-being of investigators, thereby supporting the sustainability of their crucial work. By providing such protective mechanisms against the emotional toll of human rights investigations, LLMs could potentially bolster investigator resilience and strengthen the overall capacity of organizations conducting these pivotal inquiries.

However, as we contemplate these potential benefits, we must also maintain a measured perspective. LLMs are not a cure-all solution; they should be used in tandem with robust mental health support structures, including access to counseling and strategies for self-care and stress management. While LLMs might aid in mitigating trauma exposure, the true backbone of emotional well-being and long-term sustainability for investigators remains rooted in the human-centric support systems within an organization.

### *Ethics and Implications of Using LLMs in Human Rights Investigations*

As we explore the potential of LLMs in open-source human rights investigations, we must also navigate the ethical implications inherent in their application. With the potential for transformative progress comes the reality of novel challenges; using LLMs within the field of human rights investigations is not without its potential pitfalls and ethical quandaries.

It is crucial not to overlook the risks that could arise from an overreliance of researchers on LLMs. As powerful as these tools can be, they are not infallible and should not be treated as an absolute source of truth. LLMs, at their core, are models trained on vast but nonetheless finite data sets. While this data provides the models with a broad basis for understanding and generating language, they are not without their inherent limitations. Investigators must bear in mind the fallibility of LLMs and consider their outputs as starting points for investigation, rather than definitive conclusions.

Misinformation is a significant concern when dealing with LLMs. Despite their impressive ability to produce contextually accurate information, LLMs generate responses based on patterns found in their training data. They do not have access to real-time or situation-specific information beyond what they were trained on. Thus, an LLM may create an output that appears authentic and credible but might not be accurate or relevant in the current context. For instance, during an investigation, an LLM might reference sources or media from similar scenarios in its training data, which are not relevant to the unique situation under investigation.

Another critical issue with LLM usage is the risk of amplifying bias. LLMs, despite their expansive scope, can inadvertently perpetuate existing biases in the data they were trained on. If their training data lack representation from certain demographics or omit information

about specific issues, these shortcomings will likely be reflected in the LLM's output. This can result in a skewed representation of reality, potentially overlooking marginalized communities or underrepresented issues. An LLM, for example, may disproportionately surface content about dominant groups relevant to a particular conflict while failing to look for underreported human rights abuses in marginalized communities. Hence, it becomes imperative for human rights investigators to consciously incorporate diverse sources and viewpoints alongside LLM outputs to ensure a holistic understanding of the situation.

Interpretability is another significant consideration when using LLMs in a legal context. Human rights investigations often demand a high level of transparency and accountability. Each step of the investigative process must be justified and capable of being explained. However, LLMs, often lack interpretability and are considered a "black box."6 Understanding the decision-making process of an LLM—how it decided to conduct specific queries or rank the relevance of content—can be an incredibly challenging task.

The lack of interpretability becomes particularly significant when the research findings contribute to legal proceedings. If human rights investigators heavily rely on LLM-generated outputs, it may become challenging to justify these findings or validate their reliability in court. After all, a court might find it hard to accept evidence whose derivation cannot be entirely explained or justified. Thus, it becomes crucial for investigators to consider how LLM outputs will be used downstream and to design investigation workflows accordingly.

## Conclusion

The integration of social media and AI has revolutionized open-source human rights investigations, expanding our capabilities in discovering, verifying, and collecting content to uphold human rights and promote accountability worldwide. Within this context, large language models (LLMs) hold immense promise and complexity. These AI systems possess sophisticated data processing abilities that with the potential to transform the way we uncover information, validate evidence, and protect researchers from vicarious trauma. LLMs empower investigators to navigate the vast expanse of online information, pinpoint relevant content, and extract invaluable insights. They offer the potential to streamline the verification process, automate data analysis, and shield investigators from emotional strain. However, it is crucial to recognize that LLMs are not flawless and can inadvertently

perpetuate biases, amplify misinformation, and lack transparency in their decision-making. To harness the true potential of LLMs, human rights organizations must uphold ethical considerations, complement LLM outputs with diverse sources and human expertise, and ensure that their application adheres to principles of transparency, accountability, and fairness. By proceeding on this path with careful oversight and judiciousness, LLMs can play a meaningful role in advancing human rights and instigating positive change.

While LLMs offer great potential for expanding the investigative capacities of human rights researchers and activists, they also present new challenges that must be addressed by professionals in the field. The recruitment, training, and monitoring of human rights investigators need to consider the implications of AI, including dangers such as overreliance on AI tools,

# References


[1] Irving, Emma. "And so It Begins… Social Media Evidence in an ICC Arrest Warrant." Opinio Juris, Sept. 2018, opiniojuris.org/2017/08/17/and-so-it-begins-social-media-evidence-in-an-icc-arrest-warrant

[2]OHCHR. "Berkeley Protocol on Digital Open Source Investigations: A Practical Guide on the Effective Use of Digital Open Source and Information in Investigating Violations of International Criminal, Human Rights and Humanitarian Law." OHCHR, www.ohchr.org/en/publications/policy-and-methodological-publications/berkeley-protocol-digital-open-source.

[3] OHCHR. "Berkeley Protocol on Digital Open Source Investigations: A Practical Guide on the Effective Use of Digital Open Source and Information in Investigating Violations of International Criminal, Human Rights and Humanitarian Law." OHCHR, www.ohchr.org/en/publications/policy-and-methodological-publications/berkeley-protocol-digital-open-source.

[4] 4 OHCHR. "Berkeley Protocol on Digital Open Source Investigations: A Practical Guide on the Effective Use of Digital Open Source and Information in Investigating Violation of International Criminal, Human Rights and Humanitarian Law." OHCHR. 47-48. www.ohchr.org/en/publications/policy-and-methodological-publications/berkeley-protocol-digital-open-source

[5] Roose, Kevin. "Why a Conversation With Bing's Chatbot Left Me Deeply Unsettled." The New York Times, 17 Feb. 2023, www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html.

[6] Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. "A Survey of Methods for Explaining Black Box Models." ACM Computing Surveys 51, no. 5 (2018): 1–42. https://doi.org/10.1145/3236009.

## Biographies

**Vyoma Raman** is an incoming master's student in the Department of Computer Science at Stanford University and a recent graduate of UC Berkeley with bachelor's degrees in Computer Science and Interdisciplinary Studies. Her interests lie at the intersection of artificial intelligence and human rights, with a particular focus on algorithmic justice and disability studies. Currently, Vyoma is an affiliate researcher at Berkeley Artificial Intelligence Research Lab and the UC Berkeley Human Rights Center. Previously, she has interned on multiple responsible AI product teams at Microsoft and conducted research on bias in synthetic media with Berkeley's Natural Language Processing Group and School of Information.

**Camille Chabot** is a recent graduate of UC Berkeley, holding bachelor's degrees in Global Studies, Human Rights, and Chinese, as well as a BA in Politics, Government & Law from Sciences Po Paris. With expertise in open source investigations from UC Berkeley Human Rights Center's Investigations Lab, she now works as a research consultant, exploring the impact of Large Language Models (LLMs) like ChatGPT on various professional fields. Camille's focus is on utilizing international law to bridge the understanding gaps between Eastern and Western perspectives on human rights and uphold a minimum threshold for human dignity. Additionally, she is an incoming master's student at the Yenching Academy of Peking University, where she will further expand her knowledge of Chinese culture and international law.

# Copyright Law in the Age of Machine-Generated Art

Grace Wang

## Abstract

Generative-AI-art algorithms, most notably Stable Diffusion, DALL-E, and Midjourney, have escalated a long-sidestepped conversation concerning copyright's applications to digital-born media. While much discourse surrounding AI art technologies revolves around their infringement of traditional artists' work, recent movements toward artificial generalized intelligence necessitate conversations about how to adapt the existing copyright jurisprudence to an increasingly digital world. I propose that, in light of AI art's automated nature and its shaky claims to originality, machine-generated art should be delegated to the public domain immediately upon creation. Rather than straining creative licensing to encompass hyper-digital spaces, increased legislation regarding content moderation and defamation is needed to govern emergent technologies that open up unprecedented possibilities for misrepresentation and impersonation.

## Copyright Law in the Age of Machine-Generated Art

While many innovation-minded efforts have set out to declare visual art a defunct field, few technologies have threatened to decimate art as thoroughly as has generative AI. These text-to-image models, best represented by Stable Diffusion, DALL-E, and Midjourney, are trained to generate artwork from datasets of prepublished pieces without regard to these works' copyright protections, infringing upon the labor of sweat-and-blood creatives and posing unprecedented challenges for copyright law. Yet as AI steadily gains traction in professional and commercial spaces, it appears that AGI is on track to becoming reality. A more pressing question, then, becomes: should AI art itself be deemed an independent creation and receive protections under copyright? By virtue of its exploitative origins, AI art is not owed the benefits of being copyrightable. Policymakers should instead consider works authored by algorithms for immediate entry into the public domain, owing to their untenable claims to authorship and misalignment with the utilitarian and economic incentives that copyright law is intended to uphold.

At first glance, machine-generated art fulfills the definition of derivative work [Work based on or derived from one or more preexisting works. With the addition of "new original copyrightable authorship," derivative works are protected under copyright]. and is therefore liable to receive protection under copyright law. Yet deeper inspection into AI art's dependency on ready-made art complicates its claims to creativity and originality, making this an exceedingly generous legal denomination that AI art is undeserving of. In 1983, *Gracen v. Bradford Exchange* addressed a series of porcelain plates painted with scenes from *The Wizard of Oz* that sought copyright protection. It was ruled that "superimposing one copyrighted photographic image on another" was not original, an assessment that compromises AI art's—essentially a compilation of superimpositions—copyrightability. Though perhaps a more stringent application of copyright law, *Gracen v. Bradford Exchange* nevertheless adheres to the U.S. Copyright Office's criteria for copyrightability, which maintains that protection rights are extended to original works that are independently created and possess a "modicum" of creativity. Generative AI's fulfillment of this latter criteria is doubtful. As it stands, AI employs no creative input from the end user on a visual forefront, and proponents of the technology who herald AI's artistry repeatedly mistake efficiency for creativity; dilettantes who have only ever observed a completed piece fail to appreciate the skillset necessitated by the traditional creative process or its potential for

catharsis [At a launch party for StabilityAI in October of 2022, CEO Emad Mostaque reportedly proclaimed that “The world has been creatively constipated and we’re going to let them poop rainbows.”]. Furthermore, copyright law does not consider titles, phrases, and other “variations of typographic ornamentation” for protection. As such, AI art’s claims to copyrightability are doubly precarious: its visual components are outsourced, overlaid, and bear no inspiration from the end user, while the user’s sole original contribution to the end product, their inputted text-to-image prompts, is not eligible for legal protection.

Bestowing AI imitations with legal rights would validate works generated with malicious intent that are potentially ruinous for artists’ livelihoods. Sarah Andersen, an artist leading the charge in a lawsuit against Stability AI Ltd., Midjourney Inc., and DeviantArt.Inc, has faced this reality firsthand. Her popular webcomic “Sarah Scribbles” was harassed by alt-right readers in 2016, who replicated her handwriting in a typeface that they then used to edit panels to reflect violently racist, neo-Nazi ideologies unreflective of Andersen’s own political leanings. Andersen reportedly started receiving late-night calls from embittered fans who were unaware her art had been co-opted and “got the distinct impression that the alt-right wanted a public meltdown.”

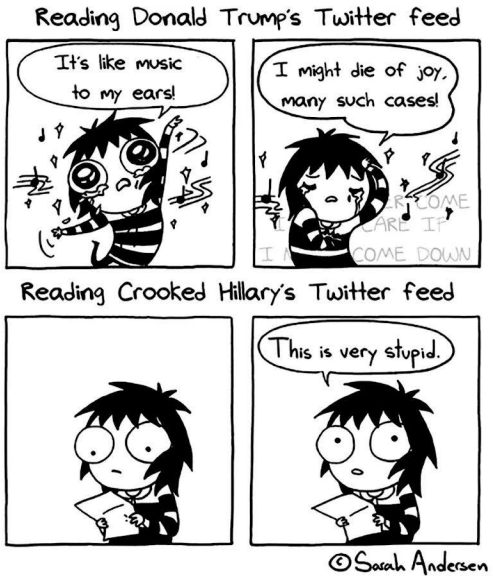



*Fig. 1. Reading Donald Trump’s Twitter feed, 2017.*

“I see a monster forming,” Andersen wrote in her opinion piece on AI art for *The New York Times.* Extending copyright to imitations like the one above would allow them to be legally recognized as standalone pieces and protect the AI user from being punished for

plagiarization— despite having alienated the original artist's creativity from their work. Such a debacle may sound familiar to those following the coinciding rise in deepfakes, where fans have staunchly objected to disseminating parodies of celebrities' voices and appearances, a phenomenon that has the potential to put not only their careers, but also their reputations, in jeopardy. As Andersen puts it, these AI algorithms are "trying to program a piece of [her] soul," and assigning an institutional endorsement of originality to regurgitations made to ladle derision upon creative professionals' work, identity, and values would be a moral failure of the legal system.

Furthermore, extending copyright protection to machine-generated art fulfills neither copyright law's economic rationale nor its primary objective of benefiting the public, making AI art's endowment with property rights not only unethical but also an ineffectual decision. Copyright law is guided by a utilitarian philosophy. The idea is that authors, when granted a limited monopoly, will be more incentivized to output works that prove to be inspirational, provocative, or comforting, ultimately bringing benefit to the public. These twofold intentions behind copyright law don't apply to machine-generated art the same way they do to traditional; considering how the costs of production and barriers to entry for AI art are practically nonexistent, their creation does not warrant a limited monopoly. The energy exerted in order to secure rights for machine-authored work would require much of the costs of copyright protection but procure little of the benefit because generative works operate on a "first-to-market" incentive—their worth is determined by the number of views, page visits, and other forms of digital traffic they incite—a playground for competition that moves forward faster than the copyright office can review content and regardless of whether a piece is copyright protected. While copyright law's purpose is to encourage creators to make products that promote the public good, previous arguments indicate that AI art can only be described as disruptive. The social strata that AI art occupies is one it already thrives in; as such, institutions would be better off endorsing other content—say, traditional art—rather than unnecessarily making space for generated art in the existing jurisprudence.

In light of the rational and moral difficulties encountered when applying copyright law to AI art, automatically delegating generated artworks to the public domain is the most logical course of action. The public domain, Copyright Commons, and other "copyleft" denominations open up works to creative reuse for all interested users without prompting

them to wade through royalties or legal encumbrances, a bona fide art library that enables users to access and stimulate human creativity in ways reminiscent of the claims made in AI art algorithms' mission statements. Indeed, on a digital stage where user-generated social platforms accelerate production and blur the line between author and audience, "sharing"—in both the distributive and cost-free sense of the term—dominates users' consciousness while "business"—born out of the arduous, obsolete producer-to-audience distribution model—recedes. While this is the case, AI art enthusiasts interested in monetizing their work should keep an open mind; the initial machine-authored work cannot be copyrighted under this model, but derivative works inspired by that AI piece can, leading AI to be a launchpad for ideas and to take a more mediatory role between technology and human ability. Nevertheless, this is not to say that the public domain is without issues. The ability to distinguish between human-authored and machine-authored works will prove to be increasingly difficult, and a pressing concern emerges in that commercial reuse is permitted in copyleft spaces, opening up defamatory AI works to being used for profit by a limitless number of entities. But the fact remains that imposing copyright law as it exists today over every digit and pixel of the datafied world is not a viable solution. Policymakers should not focus on protecting generated works, but rather reinvent creative licensing and amplify legislation concerning slander and libel so as to accommodate for a digital space that has long rendered traditional ideals of authorship and creativity obsolete.

Generative art algorithms' entrance onto the AI scene have escalated the need for a long-delayed conversation about copyright protection in the digital age. Not only does AI art sit poorly with copyright law's existing clauses, but it circulates in a near-unregulatable digital space and defies the economic, utilitarian, and moral rationales behind the legal denomination. Given this, an attainable solution would be to consider AI art for immediate entry into the public domain, a space built upon tenets of universal access and creative and commercial reuse. While such an action plan conveniently absolves the current jurisprudence of the need to ruminate creative rights and the way they're enforced, policymakers should consider that more practical avenues toward equitable creative licensing in the long run necessitate an examination of today's highly-digital social stratas and what ways copyright law can reflect this reality.

# References


Andersen, Sarah. "The Alt-Right Manipulated My Comic. Then A.I. Claimed It." The New York Times, 31 Dec. 2022. The New York Times, www.nytimes.com/2022/12/31/opinion/sarah-andersen-how-algorithim-took-my-work.html. Accessed 19 Mar. 2023.

de Rosnay, Melanie Dulong, and Juan Carlos De Martin, editors. "The Public Domain Manifesto." The Digital Public Domain: Foundations for an Open Culture, 1st ed., vol. 2, Open Book Publishers, 2012, pp. xix–xxvi. JSTOR, http://www.jstor.org/stable/j.ctt5vjsx3.6. Accessed 4 May 2023.

Jaszi, Peter. "Toward a Theory of Copyright: The Metamorphoses of 'Authorship.'" Duke Law Journal, vol. 1991, no. 2, 1991, pp. 455–502. JSTOR, https://doi.org/10.2307/1372734. Accessed 20 Mar. 2023.

Mahdawi, Arwa. "Nonconsensual Deepfake Porn Is an Emergency That Is Ruining Lives." The Guardian, 1 Apr. 2023, https://www.theguardian.com/commentisfree/2023/apr/01/ai-deepfake-porn-fake-images#:~:text=A%202019%20report%20by%20Sensity,order%2C%20featuring%20anyone%20 you%20like.

Reading Donald Trump's Twitter Feed. 2017. The New York Times, www.nytimes.com/2022/12/31/opinion/sarah-andersen-how-algorithim-took-my-work.html. Accessed 4 May 2023.

Ricolfi, Marco. "Consume and Share: Making Copyright Fit for the Digital Agenda." The Digital Public Domain: Foundations for an Open Culture, edited by Melanie Dulong de Rosnay and Juan Carlos De Martin, 1st ed., vol. 2, Open Book Publishers, 2012, pp. 49–60. JSTOR, http://www.jstor.org/stable/j.ctt5vjsx3.8. Accessed 4 May 2023.

Roose, Kevin, and Casey Newton, hosts. "Generative AI is Here. Who Should Control It?" Hard Fork, season 1, episode 4, The New York Times, Oct. 2022. Spotify, open.spotify.com/episode/1C6LhuMgaO1hHiblOAKLBw?si=bd162c2640aa488a. Accessed 4 May 2023.

"The public domain." University of California Copyright, copyright.universityofcalifornia.edu/use/public-domain.html. Accessed 4 May 2023.

United States, Ninth Circuit Court (9th Cir.). Andersen et al. v. Stability AI Ltd. et al. Docket no. 3:23-cv-00201-WHO, 23 Feb. 2023. Bloomberg Law, www.bloomberglaw.com/public/desktop/document/AndersenetalvStabilityAILtdetalDocketNo323cv00201NDCalJan132023Co/1?doc_id=X7O8UTHBKAC83BA5ES7JLB5GFP L. Accessed 19 Mar. 2023.

"What is Copyright?" Copyright.gov, U.S. Copyright Office, www.copyright.gov/what-is-copyright/. Accessed 19 Mar. 2023.

Yu, Robert. "THE MACHINE AUTHOR: WHAT LEVEL OF COPYRIGHT PROTECTION IS APPROPRIATE FOR FULLY INDEPENDENT COMPUTER-GENERATED WORKS?" University of Pennsylvania Law Review, vol. 165, no. 5, 2017, pp. 1245–70. JSTOR, http://www.jstor.org/stable/26600620. Accessed 20 Mar. 2023.

## Biography

**Grace Wang** is an undergraduate student at UC Berkeley studying Economics and Data Science. She is interested in how generative AI can be ethically introduced into creative fields that are traditionally modes of human expression.

# Not Losing Ourselves to the AI Storm: Exploring how AI lives in the past and dreams of the future

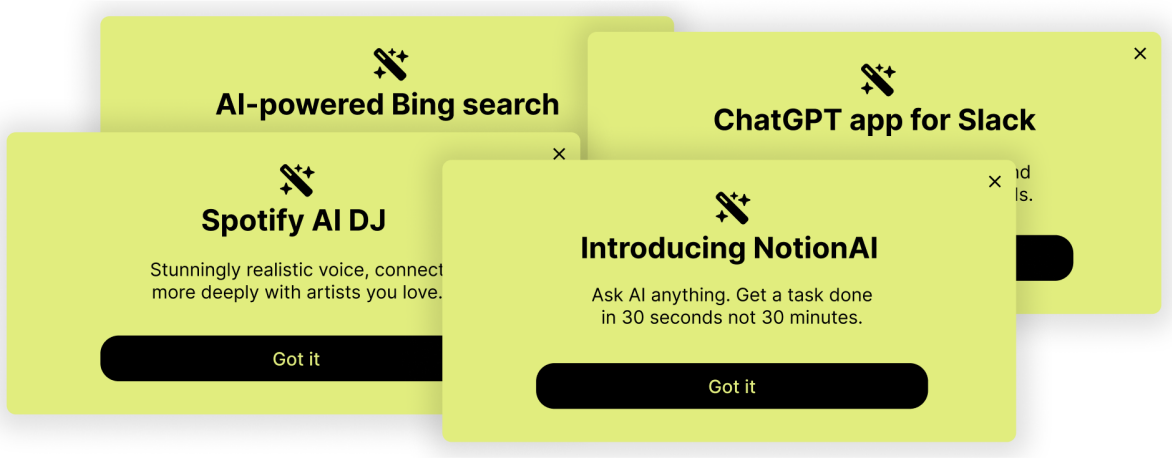
Yuna Shin

## Abstract

AI is showering us with attention when it comes to our ideas and questions. It's shaking up entire industries and our day to day routines. The tech giants are racing to dominate AI while public perception is a mixed bag of emotions: intrigue, delight, and fear. Will I lose my job? How can ChatGPT do this faster? Can I coexist with AI? We're losing ourselves to the idea that generative AI is the end all be all of creativity and productivity. As a result, how we think about "process" is being challenged. I argue that this shift in our collective values is similar to the anti-art movement in the early 1900s. Both AI and fashion trends have elements of the hyper-real that blur the line between what is human and what is artificial. This rollout of new AI product experiences carry over problems from existing digital spaces. Ultimately, AI lives in the past and dreams of the future. Through our presence as humans, we can shape, fix, and anticipate how AI bleeds into our present day.


As I drift online, I'm becoming more aware of AI's presence. When I browse the web, design a prototype, or debate what to cook for dinner it's becoming more uncertain what my next move should be. Should I invite AI into my thinking process?

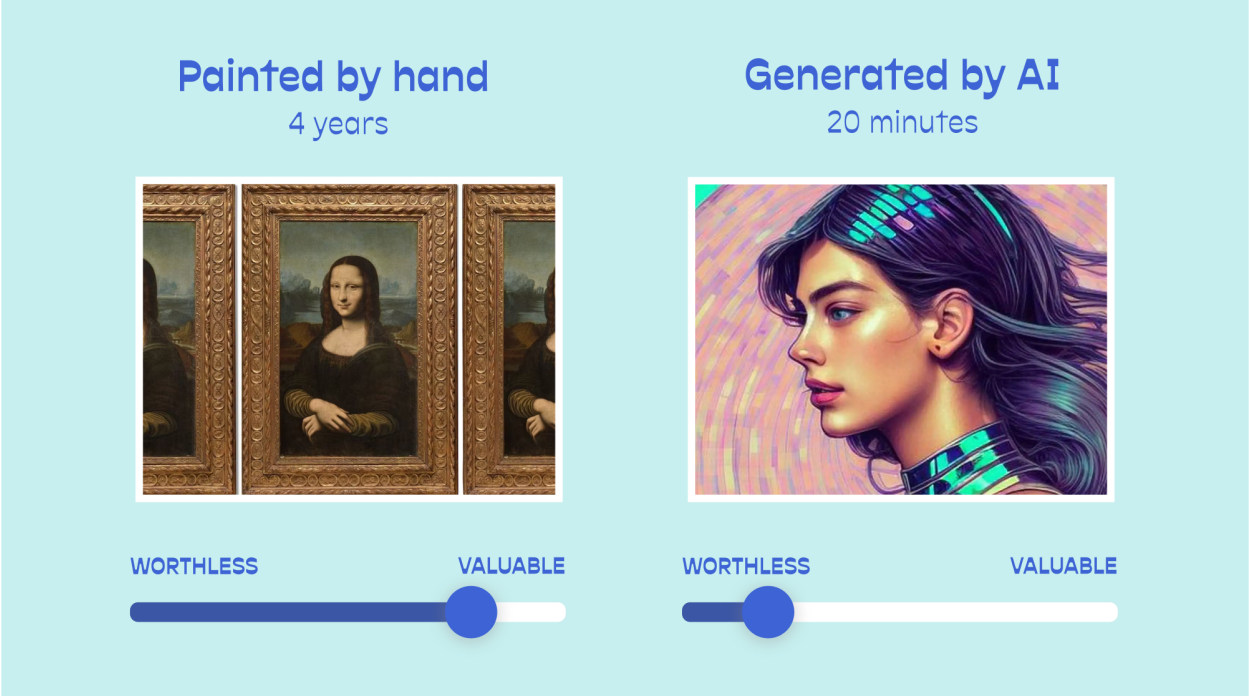
I find myself in a storm of cool AI products with murky ethics and big promises for a more personalized experience. When the thunder roars, we don't have the option to hide indoors. How do we coexist with this AI hype in our work and personal lives?




## AI changes how we create

Over the decades, digital technology has pushed us to reconsider our processes and collective values. AI-powered features are rolling out into everyday consumer products like Spotify, Notion, and Bing at lightning speed. It strikes us with delight, intrigue, and fear. Finally, we have tools that can shower our thoughts with attention deceivingly well. You ask and shall receive a dynamic and thoughtful response as an audio, code, image, text, or video output.

The leap from spell check to ChatGPT's ability to rewrite paragraphs in "Shakespearean dialect" lands us with new questions of what deserves our attention and praise. *Should we devalue an article written with the help of Notion AI? Is artwork generated by LensaiAI less precious than a hand-drawn painting by a local artist?*




## AI is making us rethink our values in similar ways as the anti-art movement

In the early 1900s, the anti-art movement was led by artists who purposely rejected prior definitions of what art is. It provoked a shift in what we value in the art world. During 1917, the French artist Marcel Duchamp submitted a store-bought urinal signed with the pseudonym “R. Mutt” to a gallery show. The submission was rejected and caused an uproar, but it expanded and confronted our imagination of what is considered art. This created opportunities for new forms of art that go beyond the institutional vantage point of the artist. Rather than focusing on the craft and sublimity of a physical artwork, anti-art paved the way for contemporary art that values the ideas and concepts being explored by the artist in dynamic ways like performance, video, sculpture, and installations. Generative AI is becoming the Marcel Duchamp of our 21st century. Similar to the anti-art movement, AI invites us to reject conventional tools, processes, and products. It allures us by freeing us from being alone with our thoughts and concisely telling us what to imagine. The invitation of an AI companion in our classroom, office, or home allows for us to speed up, cut in half, or eliminate our thinking process. This challenges our sense of self and our place in the world.

## AI intensifies the blurring of the line between what is human and what is artificial

As a result of AI changing how we create, what we're creating is also changing. The AI hype is taking storm in digital spaces where democratization of user privacy and autonomy is dwindling. For example, Twitter and Meta launched a paid product version that grants additional verification and visibility features. This increases the chance for misinformation, fake profiles, trolls, and bots. With AI intensifying the blurring of what is human and what is artificial, the need for authentication and transparency continues.

Vogue covered the fascination behind the viral hyper-real "big red boots" by MSCHF that resemble the pair Astro Boy wears in the anime series. These impractical, playful boots blur the line between the real and the unreal in similar ways as AI does. It plays into the double take we do while listening to the AI-powered DJ on Spotify or scrolling across the viral AI-generated image of Pope Francis in a white puffer. The uncanny quality of the big red boots force us to consider how digital aestheticization distorts details, realism, and quality. A stark contrast is made between what exists in the real world and what is trying to fit in. The boots make it obvious what qualities of the imperfect physical world can't be digitally copied over.


## Our presence gives value to AI outputs in a variety of ways

The shift of creativity in the age of AI also means world-building and dreaming with tools that are not independent nor neutral. In an article by CityLab, the architectural designer Tim Fu describes the AI art generator Midjourney as an advanced tool that can aid the creative process but "still requires the control and the artistry of the person using it." The rapidly generated images help with the earliest stages of a project, but the images lack detail. The architects spot gaps in the AI art generator's understanding of non-Western architecture.

In a recent NYT guest essay, Noam Chomsky describes how ChatGPT “either overgenerates (producing both truths and falsehoods, endorsing ethical and unethical decisions alike) or undergenerates (exhibiting noncommitment to any decisions and indifference to consequences)." Rather than a bot takeover, our responsibilities will expand in new ways as designers, programmers, educators, students, or casual users. We must create a new type of digital literacy to address this tension between the user and AI of knowing what to ask, how to push back, and when to accept an outcome.

By making these digital experiences with AI more collaborative, we can collectively anticipate blindspots. LinkedIn recently introduced a new feature called “collaborative articles” that starts with a pre-written article by AI. Experts on their platform with relevant skills based on their internal evaluation criteria are invited to add context and information. It uses AI as a jumping off point for discussion that emulates the back-and-forth that happens in comment sections. This is one approach for more human intervention that creates space for our live cynicism and voice to be at core of any AI output.

Together with our skepticism and presence can we prevent the distortion of our ideas. This puts necessary pressure on the in-between moments that shape who we are. The moments when we are alone with our thoughts—without the distraction of technology.

## You don’t need AI to dream big

Rather than engaging in the present moment, AI takes any context and uses training data to predict what comes next in the sequence. Instead of sieving through the excess of information on the Web, we get information rearranged from large language models like ChatGPT that don't leave a clear trace of how it ended up where it did. ChatGPT creates a foggy interpolation of the Web.

Digital technology distorts our understanding of linear time by repackaging the past to remix something new as a future possibility. Our senses are grounded in the real world and in the present, where we truly exist. AI lives in the past and dreams of the future. If we treat AI as the end-all-be-all for creativity, learning, productivity, and innovation, won’t we lose our sense of self and what we stand for? Generative AI exists for your text input; it lives to anticipate but doesn't live.

## Biography

**Yuna Shin** is an interaction designer and writer currently working for Artefact in Seattle. Her approach to design is to embrace the speculative and crafting process to question dominant narratives and imagine new ways of living. She writes about technology from a critical lens by connecting the dots between design, contemporary art, and digital pop culture.

## *Thank you for reading!*

<AI & Equality> is a community committed to establishing and promoting a human rights-based approach to AI that centers equity & inclusion at the core of the code. All authors are members of our online community. This open-to-anyone global community aims to connect individuals from all backgrounds, regions, and disciplines to work towards a collective goal of human-rights based AI. We plan to continue providing a platform and space for sharing the thoughts, ideas, and work of our community through future publications.

Join the AI & Equality Community here:
https://community.aiequalitytoolbox.com/home