Introduction to *GRACE'*s Discussion of Abeba Birhane's "Algorithmic Injustice"

Bethel Bayrau
Stanford Medical School

and
Wayne Chinganga
Stanford University

Our ethics and values are always present in the creation and use of technology. The technology society creates and chooses not to create is a window into the ethics and values of the powerful—Sabelo Mhlambi (2020)

Dr. Abeba Birhane's provocative AI ethics paper, "Algorithmic Injustice: A Relational Ethics Approach," draws on frameworks too often neglected in AI ethics studies. Her important work on race, justice, and ethics frameworks for machine learning algorithms calls for inclusion of relational Sub-Saharan African philosophies in such a curriculum. Delineating the ethical limitations of European individualist rationality as a definition for personhood, especially in marginalized communities and on the African continent, Birhane shows how traditionally European frameworks fail to address the perspectives of those whom AI most impacts. Following many important African philosophers like Mogobe B. Ramose, Emmanuel Chukwudi Eze, Ifeanyi Menkiti, Sabelo Mhlambi, and others, Birhane offers the AI Ethics community important insights from African relational ethics, which link one's personhood to the personhood of others, and show that to talk about AI harms one must understand the communal relational perspective.

Seeing AI ethics relationally departs radically from the dominant AI Ethics discourse in the American and European academies, which understand about justice through consequentialist outcomes and post-Kantian normative frameworks, where all decision-making remains individualistic. Typical AI ethics syllabi reading, especially here at Stanford, often include John Rawls' "Justice as Fairness" and his famous thought experiment "the original position," where all individuals must select the properties of their societies, or in our case, the algorithm, from behind a "veil of ignorance," which deprives the individual of knowing anything about their own social status, ethnicity, gender, so that they can never chose a system, which serves their own self-interest. We engineering students are often presented with Rawls to help us aspire to build better algorithms and seek designs that are "fair" and "just" according to universalist, liberal categories that foreground individual will. Such liberal, rationalist, individualism is helpful in some contexts and less so in others. This framework makes less sense for non-Western, and especially African cultures, as these tend to view the community and its relations as primary. Recent efforts to add Birhane to course syllabi appear promising, yet require more discussion into deeper questions of whether all those impacted by algorithmic technology enjoy equal access to western concepts of freedom.

Dr. Birhane's paper foregrounds the disproportionate harm of western technology and rationality to those in the Global South. She elaborates why western rationality and its metrics may be inappropriate for assessing harms in non-western communities, and she posits an ethics influenced by relational philosophies that should be applied as a "habit, not a mere methodology for data science." By mentioning the word, "habit," Birhane references virtue ethics, which, in the African context, refers to consistent lifelong practices or "habits" that serve communal good. She elaborates the need for such ethics, when she first describes how AI has removed socially

and politically "contested matters" from an accessible public sphere to the arcane language of technology, where engineers view these long standing issues "as mathematical problems with a technical solution." Birhane asks her readers to "rethink" such blind faith in technology, and instead pursue an ethics with "concrete knowledge of the lived experience of marginalized communities." She wants ethics students to develop awareness about "historical injustices and the currently tangible impact of AI systems on vulnerable communities." Arguing for improved material conditions for these communities involves "moving away from ethos as abstract contemplations or seemingly apolitical concepts such as 'fair and good.' " In an extended theoretical reflection from Plato, Newton, and Descartes to the founders of computer science, Birhane shows why western categories are anything but apolitical when they base knowledge solely on reason. Data science and data metrics as we know them today have inherited this western perspective that assumes a supposedly neutral, scientific "view from nowhere." Bayesian models, for example, are known to contain "spurious stereotypes" but are widely accepted as "neutral" evidence. She quotes John Horgan who argues that Bayes's theorem, while widely deployed as a "powerful method for generating knowledge, can also be used to promote superstition and pseudoscience" (Horgan, 2016). To further critique western methods, Birhane draws on the ethos of Alan Turning who defines "thinking" narrowly for machines, and Kurt Gödel whose incompleteness theorem reminds audiences that the pioneers of computer science also showed that "a consistent formal system, such as the mathematics of computing, cannot by itself prove the truthfulness or falsity of all theorems that can result from the system's rules and axioms." For her, both humans and machines are "incomplete," so we must not uncritically accept whatever mathematical results we get.

Having established her critique of western rationalism under the aegis of computer science pioneers, Birhane next draws on systems of Afro-feminism and post-colonial African philosophy to present her alternative approach which can be used both when designing AI and also for Africans to assess their own well-being and develop "a genuine sense of belonging in the world." Claiming the Southern African philosophy of *Ubuntu* has a "relational personhood. diametrically opposed to rationality as personhood," she argues that such a philosophy can counteract the Western world's asymmetric relationship with much of the world, computing culture, and AI's quest for a mechanical personhood. Birhane is also careful to be inclusive in her description of African philosophy, because for her, not all relational frameworks are non-western, nor southern African. Relationality, as she wants us to apply to AI, includes *Suthu* knowledge systems and *Nguni*, as well as the American Afro-feminism of Patricial Hill Collins, on D'Ignazio and Klein's *Data Feminism* (2020), and Mikhail Bhaktin's Russian formalism, for example. Informed by these frameworks, she weaves in a non-western ethics, reminding audiences that she seeks no mere " methodology" or "tool," but posits instead a "re-examination of the nature of existence, knowledge, oppression, and injustice" to study the social, historical, and political context of algorithmic injustice. For Birhane, "relational ethics provides the framework to rethink the nature of data science through a relational understanding of being and knowing."

This frameworks section of *GRACE's* first volume, presents a special roundtable on Dr. Birhane's "Algorithmic Injustice" (2021), which underwent extensive debate among our editors. All editors are students from the Global South and/or marginalized/ former enslaved communities, studying at Stanford University, an elite American university. All contributors who submitted a response to our debate on Birhane identify at least in part as a member of the Black

diaspora and/or low-income community. As an African medical researcher (Bethel Bayrau) and African computer science student (Wayne Chinganga), we find ourselves highly persuaded by Dr. Birhane's analysis of algorithmic injustice and her call to move to a relational ethics, which is more suitable and beneficial for Africa. We do see how African ethics tend toward the relational and that the mere application of western individualist, rationalist theories fail to encompass the diverse lived-experiences we Africans encounter on our continent. Communities do hope to decide together which algorithms might benefit us and which have wreaked disproportionate harm in their scramble for digital colonization of Africa. However, we struggle to separate these relational ethics from the very same rationality that Birhane criticizes. As African scientists, we participate in *both* of these worlds and in some contexts, like education in Africa and at Stanford, we see ourselves as rational individuals who profit greatly from deploying Bayesian methods and other technical principles. In other settings, we view ourselves as communal actors whose ethics are indeed relational. Perhaps we might thus conclude that Dr. Birhane wants us to develop a healthy skepticism toward western philosophical systems, especially rationality and individuality, which presume everyone enjoys equal moral status, when in fact, not everyone does. Thus, rather than calling to eliminate western philosophy from our AI ethics curriculum, Birhane inspires us to gain our distance and include relational ethics that can better help us serve those most greatly harmed by algorithms.

The three responses to Dr. Birhane in this section address the question of what frameworks best equip us to fight algorithmic injustice. We see in a provocative piece, "Analytic Relationality vs. the Relational Ethics of the Global South: Making the Case for Abeba Birhane's Work" by Stanford Computer Science and Philosophy major, Julia Kwok and Stanford Global Studies lecturer and European AI Ethics policy expert Dr. Nakeema Stefflbauer, criticize western

analytic philosophy's efforts to theorize relationality without adequately addressing unequal access to moral status in western systems. In their comparison, Kwok and Stefflbauer consider analytic philosophy, which has belatedly joined conversations about algorithmic ethics and relationality, and argue that Birhane's work shares some arguments but ultimately offers a better framework for ethical intervention in algorithmic harms in the Global South. A similarly strong affirmation of Birhane's work appears in Stanford NLP PhD student, Tolúlọpẹ́ Ògúnrẹ̀mí, Masakhane Researcher, Wilhelmina Onyothi Nekoto (Namibia), and Stanford Computer Scientist Saron Samuel's "Decolonizing NLP for 'Low-resource Languages:' A Response to Abebe Birhane," which demonstrates how Birhane's ethics are best practiced in approaches that center African data subjects and engineers.

Not all *GRACE* submissions concurred with Birhane's work. Our most critical essay by German Studies and Computer Science major and *Stanford Review* Editor-in-Chief, Mimi St. John, questions whether Birhane calls merely for a more measured view of western rationality. Indeed, St. John interprets Birhane's argument as a wholesale rejection western rationality in favor of non-western relational ethics and questions the categories Birhane employs, reflecting on their deeply western roots and instability for normative ethics. Yet, St. John also concedes other possible readings and allows that Birhane intends otherwise than rejection. Thus, we begin this section with St. John and her opposing arguments as a tribute to the late great Stanford Professor Ken Taylor, who exhorted us to closely interrogate John Stuart Mill, and who always enjoyed bracing public debate. Next follows Kwok and Stefflbauer, and third, Ògúnrẹ̀mí, Nekoto, and Samuel, who round out the discussion with examples of inclusive algorithmic practice. GRACE is eager to hear responses to all these pieces.

References:

Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. *Patterns*, *2*(2), 100205.

D'ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT press.

Horgan, J. (2016, January 4). Bayes's theorem: what's the big deal? *Scientific American*, retrieved December 20, 2022. https://blogs.scientificamerican.com/cross-check/bayes-s-theorem-what-s-the-big-deal/

Mhlambi, S. (2020). From rationality to relationality: ubuntu as an ethical and human rights framework for artificial intelligence governance. *Carr Center for Human Rights Policy Discussion Paper Series*, *9*.