Patterns

CellPress
OPEN ACCESS

Perspective

Algorithmic injustice: a relational ethics approach

Abeba Birhane[1,*]
[1]School of Computer Science, University College Dublin, Ireland & Lero—The Irish Software Research Centre, Dublin, Ireland
*Correspondence: abeba.birhane@ucdconnect.ie
https://doi.org/10.1016/j.patter.2021.100205

THE BIGGER PICTURE Machine learning (ML) increasingly permeates every sphere of life. Complex, contextual, continually moving social and political challenges are automated and packaged as mathematical and engineering problems. Simultaneously, research on algorithmic injustice shows how ML automates and perpetuates historical, often unjust and discriminatory, patterns. The negative consequences of algorithmic systems, especially on marginalized communities, have spurred work on algorithmic fairness. Still, most of this work is narrow in scope, focusing on fine-tuning specific models, making datasets more inclusive/representative, and "debiasing" datasets. Although such work can constitute part of the remedy, a fundamentally equitable path must examine the wider picture, such as unquestioned or intuitive assumptions in datasets, current and historical injustices, and power asymmetries.

As such, this work does not offer a list of implementable solutions towards a "fair" system, but rather is a call for scholars and practitioners to critically examine the field. It is taken for granted that ML and data science are fields that solve problems using data and algorithms. Thus, challenges are often formulated as problem/solution. One of the consequences of such discourse is that challenges that refuse such a problem/solution formulation, or those with no clear "solutions", or approaches that primarily offer critical analysis are systematically discarded and perceived as out of the scope of these fields. This work hopes for a system-wide acceptance of critical work as an essential component of AI ethics, fairness, and justice.

Concept: Basic principles of a new data science output observed and reported

SUMMARY

It has become trivial to point out that algorithmic systems increasingly pervade the social sphere. Improved efficiency—the hallmark of these systems—drives their mass integration into day-to-day life. However, as a robust body of research in the area of algorithmic injustice shows, algorithmic systems, especially when used to sort and predict social outcomes, are not only inadequate but also perpetuate harm. In particular, a persistent and recurrent trend within the literature indicates that society's most vulnerable are disproportionally impacted. When algorithmic injustice and harm are brought to the fore, most of the solutions on offer (1) revolve around technical solutions and (2) do not center disproportionally impacted communities. This paper proposes a fundamental shift—from rational to relational—in thinking about personhood, data, justice, and everything in between, and places ethics as something that goes above and beyond technical solutions. Outlining the idea of ethics built on the foundations of relationality, this paper calls for a rethinking of justice and ethics as a set of broad, contingent, and fluid concepts and down-to-earth practices that are best viewed as a habit and not a mere methodology for data science. As such, this paper mainly offers critical examinations and reflection and not "solutions."

INTRODUCTION

Algorithmic decision making increasingly pervades the social sphere. From allocating medical care,[1] to predicting crimes,[2] selecting social welfare beneficiaries,[3] and identifying suitable job candidates,[4,5] complex social issues are increasingly automated. Artificial intelligence (AI) and machine-learning tools have become the hammer every messy social challenge is bashed with. High efficiency and seemingly neat shortcuts to complex and nuanced problems make algorithmic systems attractive. However, automated and standardized solutions to complex and contingent social issues often contribute more harm than good—they often fail to grasp complex problems and provide a false sense of solution and safety. Complex social issues require historical, political, and moral awareness and structural change. Any data scientist working to automate issues

of a social nature, in effect, is engaged in making moral and ethical decisions—they are not simply dealing with purely technical work but with a practice that actively impacts individual people.

As social processes are increasingly automated and algorithmic decision making deployed across various social spheres, socially and politically contested matters that were traditionally debated in the open are now reduced to mathematical problems with a technical solution.[6] The mathematization and formalization of social issues brings with it a veneer of objectivity and positions its operations as value-free, neutral, and amoral. The intrinsically political tasks of categorizing and predicting things such as "acceptable" behavior, "ill" health, and "normal" body type then pass as apolitical technical sorting and categorizing tasks.[7] Unjust and harmful outcomes, as a result, are treated as side effects that can be treated with technical solutions such as "debiasing" datasets[8] rather than problems that have deep roots in the mathematization of ambiguous and contingent issues, historical inequalities, and asymmetrical power hierarchies or unexamined problematic assumptions that infiltrate data practices.

The growing body of work exposing algorithmic injustice has indeed brought forth increased awareness of these problems, subsequently spurring the development of various techniques and tactics to mitigate bias, discrimination, and harms. However, many of the "solutions" put forward (1) revolve around technical fixes and (2) do not center individuals and communities that are disproportionally impacted. Relational ethics, at its core, is an attempt to unravel our assumptions and presuppositions and to rethink ethics in a broader manner via engaged epistemology in a way that puts the needs and welfare of the most impacted and marginalized at the center.

In the move to rethink ethics, concrete knowledge of the lived experience of marginalized communities is central. This begins with awareness and acknowledgment of historical injustices and the currently tangible impact of AI systems on vulnerable communities. The core of this framework is grounding ethics as a practice that results in improved material conditions for individuals and communities while moving away from ethics as abstract contemplations or seemingly apolitical concepts such as "fair" and "good." Relational ethics, then, is a framework that necessitates we re-examine our underlying working assumptions, compels us to interrogate hierarchical power asymmetries, and stimulates us to consider the broader, contingent, and interconnected background that algorithmic systems emerge from (and are deployed to) in the process of protecting the welfare of the most vulnerable.

Through the lens of relational ethics, we explore the wider social, political, and historical nature of data science, machine learning, and AI and the need to rethink ethics in broader terms. This paper primarily offers a critical analysis and encourages a grasp of problems from their roots. It departs from traditional scholarship within the data and AI ethics space that offer technical solutions, or implementable remedies that attempt to mitigate problems of ethical, social, and political nature.

The rest of the paper is structured as follows. In the next section, we flesh out the roots of relational ethics and provide comparisons of relationality with the dominant orthodoxy, rationality. We then lay out the four tenets of relational ethics, followed by a brief conclusion.

RELATIONAL ETHICS: THE ROOTS

Before delving into the roots and central tenets of relational ethics, it makes sense to make visible the dominant school of thought: rationality. Relationality exists both as a push back against rationality, but also on its own right, for example in the case of ubuntu as a philosophy, ethics, and way of life.[9] (A brief web search for "ubuntu" brings up information on a Linux operating system that has been around since 2004, usurping the original meaning of the word that has existed for centuries within sub-Saharan Africa. The appropriation of the word with its rich culture and history to a shallow tech sloganeering is not only wrongful but also symptomatic of the Western tech world's inability to center non-Western perspectives while stripping them of their rich culture, history, and meaning.) At the heart of relational ethics is the need to ground key concepts such as ethics, justice, knowledge, bias, and fairness in context, history, and an engaging epistemology. Fundamental to this is the need to shift over from prioritizing rationality as of primary importance to the supremacy of relationality.

Rationality: the dominant orthodoxy

> "Renaissance thinkers like Montaigne acknowledged that universal, foundational principles cannot be applied to such practical matters as law, medicine and ethics; the role that context and history play in those areas prevents it."
>
> Alicia Juarrero[10]

The rational view serves as the backbone for much of Western science and philosophy, permeating most fields of enquiry (and social and institutional practices) from the life sciences, to the physical sciences, the arts and humanities, and to the relatively recent field of computer science.[11,12] The rational worldview, the quintessential orthodoxy for Western thought, can be exemplified by the deep contention that reason and logical coherence are superior for knowledge production (in understanding the world) above and beyond relational and embodied becoming. The privileging of reason as the ultimate criterion makes knowing a distant act. The deep quest for the rational worldview is certainty, stability, and order, and thus isolation, separation, and clear binaries form the foundations in place of connectedness, interdependence, and dynamic relation.[13] Since the rational worldview has come to be seen as the standard, anything outside of this is viewed as an outlier. Spelling out what this worldview entails, what its underlying assumptions are, and the consequences for a subject of enquiry which inherits this worldview, therefore, is an important step toward providing context for the relational worldview.

Although the rationalist worldview results from the accumulation of countless influences from pivotal thinkers, its lineage can be traced through Western influential giants such as Newton, Descartes, and all the way back to Plato. René Descartes, the quintessential rationalist, attempted to establish secure foundations from which knowledge can be built based solely on reason

and rational thought. In this quest, Descartes attempted to rid us of unreliable, changeable, and fallible human intuitions, senses, and emotions in favor of reason and crystalline logic.[14] At the heart of his quest was to uncover the permanent structures beneath the changeable and fluctuating phenomena of nature on which he could build the edifice of unshakable foundations of knowledge. Anything that can be doubted is eliminated. Subsequently, discussions and understanding of concepts such as knowledge and ethics tend to be abstract, genderless, contextless, and raceless. Knowledge, according to this worldview, is rooted in the ideal rational, static, self-contained, and self-sufficient subject that contemplates the external world from afar in a "purely cognitive" manner as a disembodied and disinterested observer.[15] In the desire to establish timeless and absolute knowledge, abstract and contextless reasoning is prioritized over concrete lived experience submersed in co-relations, interdependence, fluidity, and connectedness.[16] More fundamentally, Ahmed[17] contends that all bodies inherit history and the inheritance of Cartesianism is grounded in a white straight ontology. The reality of the Western straight white male then masquerades as the invisible background that is taken as the "normal," "standard," or "universal" position. Anything outside of it is often cast as "dubious" or an "outlier."

In a similar vein, and with a similar fundamental influence as Cartesianism, the Newtonian worldview aspired to pave the path for universal knowledge in a supposedly observer-free and totally "objective" manner. This thoroughly individualistic worldview sees the world as containing discrete, independent, and isolated atoms. Neat explanations and certainty in the face of ambiguity provide a sense of comfort. Within the physical world, Newtonian mechanistic descriptions allowed precise predictions of systems at any particular moment in the future, given knowledge of the current position of a system. This view fared poorly, however, when it came to capturing the messy, interactive, fluid, and ambiguous world of the living who are inherently context bound, socially embedded, and in continual flux. Emphasizing the futility of reductionist approaches to complex adaptive systems, Cilliers[18] (p.64) contends, "From the argument for the conservation of complexity—the claim that complexity cannot be compressed—it follows that a proper model of a complex system would have to be as complex as the system itself." In a worldview that aspires for objective, universal, and timeless knowledge, the very idea of complex and changing interdependence and co-relations—the very essence of being insofar as there can be any—are not tolerated. Despite the inadequacy of the billiard ball model of Newtonian science in approaching complex adaptive systems such as human affairs, its residue prevails today, directly or indirectly,[19] within the data sciences and the human sciences in general.

The historic Bayesian framework of prediction[20] has played a central role in establishing a normative explanation of behaviors.[21] Bayes' approach, which is increasingly used in various areas including data science, machine learning, and cognitive science,[22,23] played a pivotal role in establishing the cultural privilege associated with statistical inference and set the "neutrality" of mathematical predictions. Price, who published the papers after Bayes' death, noted that Bayes' methods of prediction "shows us, with distinctness and precision, in every case of any particular order or recurrence of events, what reason there is to think that such recurrence or order is derived from stable causes or regulations in nature, and not from any irregularities of chance"[20] (p.374). However, despite the association of Bayes with rational predictions, Bayesian models are prone to spurious relationship and amplification of socially held stereotypes.[24] Horgan[25] notes, "Embedded in Bayes' theorem is a moral message: If you aren't scrupulous in seeking alternative explanations for your evidence, the evidence will just confirm what you already believe."

Dichotomous thinking—such as subject versus object, emotion versus reason—persists within this tradition. Ethical and moral values and questions are often treated as clearly separable (and separate) from "scientific work" and as something with which the scientist need not contaminate their "objective" work. In its desire for absolute rationality, Western thought wishes to cleave thought from emotion, cultural influence, and ethical dimensions. Abstract and intellectual thinking are regarded as the most trustworthy forms of understanding, and rationality is fetishized. Data science, and the wider discipline of computer science, have implicitly or explicitly inherited this worldview.[11] These fields, by and large, operate with rationalist assumptions in the background. The view of the data scientist/engineer is mistaken as "the view from nowhere"—the "neutral" view. Misconceptions such as a universal, relatively static, and objective knowledge that can emerge from data are persistent.[26] Data science and data practices reincarnate rationalism in many forms, including in the manner in which messiness, ambiguity, and uncertainty are not tolerated; in the pervasive binary thinking (such as emotion versus reason, where the former is assumed to have no place in data science); the way in which data are often severed from the person (with emotions, hopes, and fears) in whom they are rooted and the context in which they emerge; the manner in which the dominant view is taken as the "God's eye view;" and the way questions of privilege and oppression are viewed as issues with which the data sciences need not concern themselves. Not only does the inheritance of rationality to data sciences and computation make these fields inadequate to deal with complex and inherently indeterminable phenomena, Mhlambi[11] has further argued that the AI industry, grounded in rationality, reproduces harmful and discriminatory outcomes.

Relationality

Contrary to the rationalist and individualist worldview, relational perspectives view existence as fundamentally co-existent in a web of relations. Various schools of thought can be grouped under the umbrella of the relational framework with a core commonality of interdependence, relationships, and connectedness. Relational-centered approaches include Black feminist (Afro-feminist) epistemologies, embodied and enactive approaches to cognitive science, Bakhtinian dialogism, ubuntu (the sub-Saharan African philosophy), and complexity science. (This is not an exhaustive list of all approaches that could be identified as relational. The focus on these specific schools of thought and approaches, as opposed to others that might fall under relational approaches, is heavily influenced by the author's background and academic training.) Although these schools of thought vary in their subjects of enquiry, aims, objectives, and methods, they have relationality in common.

Relational frameworks emphasize the primacy of relations and dependencies. These accounts take their starting point in reciprocal co-relations. Kyselo,[27] for example, contends that the self is social through and through—it is co-generated in interactions and relations with others. We achieve and sustain ourselves together with others. Similarly, according to the sub-Saharan tradition of ubuntu as encapsulated by Mbiti's[28] phrase "I am because we are, and since we are, therefore I am," a person comes into being through the web of relations. In a similar vein, Bakhtin[29] emphasized that nothing is simply itself outside the matrix of relations in which it exists. It is only through an encounter with others that we come to know and appreciate our own perspectives and form a coherent image of ourselves as a whole entity. By "looking through the screen of the other's soul," he wrote, "I vivify my exterior." Selfhood and knowledge are evolving and dynamic; the self is never finished—it is an open book.[30]

Relational ethics takes its roots from these overlapping frameworks. In the rest of this section we delve into Afro-feminist thought and the enactive approach to cognitive science with the aim of providing an in-depth understanding of the roots of the relational worldview.

Afro-feminism

> "Knowledge without wisdom is adequate for the powerful, but wisdom is essential for the survival of the subordinate."
>
> Patricia Hill Collins[31]

Pushing back against the dominant Western orthodoxy, Afro-feminist epistemology grounds knowing in an active and engaged practice. The most reliable form of knowledge, especially concerning social and historical injustice, is grounded in lived experience. One of the most prominent advocates of Afro-feminist epistemology, Patricia Hill Collins,[31] emphasizes that people are not passive cognizers that contemplate and grasp the world in abstract forms from a distance; instead, knowledge and understanding emerge from concrete lived experiences. At the heart of it, the Afro-feminist approach to knowing contends that concrete experiences are primary and abstract reasoning secondary. Knowing and being are active processes that are necessarily political and ethical. Drawing core differences between the dominant Western tradition and the Afro-feminist perspective, Collins identifies two types of knowing: knowledge and wisdom. Knowledge is closely tied to what Collins calls "book learning"—learning that emerges from reasoning about the world from a distance in a rational way. This form of knowledge aspires to arrive at "an objective truth" that transcends context, time, specific and particular conditions, and lived experiences. Wisdom, on the other hand, is grounded in concrete lived experience. Formal education, according to Collins, is not the only route to such forms of knowledge, and wisdom holds high credence in assessing knowledge claims. Distant statistics or theoretical accuracies do not take precedence over the actual experience of a person. Knowledge claims are not worked out in isolation from others but are developed in dialog with the community. It is taken for granted that there exists an inherent connection between what one does and how one thinks. This is especially the case when the type of knowledge in question concerns oppression, structural discrimination, and racism. Wisdom, and not "book learning," enables one to resist oppression. From the core arguments of Afro-feminist epistemology, it follows that concepts such as ethics and justice need to be grounded in concrete events informed by lived experience of the most marginalized individuals and communities that pay the highest price for algorithmic harm and injustice.

Current data practices, for the most part, follow the rational model of thinking where data are assumed to represent the world out there in a "neutral" way. In the process of data collection, for example, the data scientist decides what is worth measuring (making some things visible and others invisible by default) and how. In the process of data cleaning, rich information that provides context about which data are collected and how datasets are structured is stripped away. Emphasizing the importance of contexts for datasets, Loukissas[32] has proposed a shift into thinking in terms of data settings instead of datasets.

The rational worldview that aspires to an "objective" knowledge from a "God's eye view" has resulted in the treatment of the researcher as invisible, and their interests, motivations, and background as inconsequential. In contrast, for Afro-feminist thought, the researcher is an important participant in the knowledge production process.[33] For Sarojini Nadar,[34] coming to know is an active and participatory endeavor with the power to transform. Consequently, narrative research, since it puts story telling at the center, invites us to consider stories as "data with soul."[34]

Enactive cognitive science

> "Loving involves knowing, and […] knowing involves loving. Loving and knowing, for human beings, entail each other. To understand knowing only "coldly," abstractly, objectively is either not to see the loving involved, or not to know fully."
>
> Hanne De Jaegher[35]

In a similar vein to Afro-feminist thought, the enactive cognitive science theory of participatory sense-making[36] advocates for an active and engaged knowing rooted in our relating. A proponent of this position, Hanne De Jaegher,[35] contends that our most sophisticated human knowing lies in how we engage with each other. In a recent work, De Jaegher[35] emphasizes that discrete, rational knowing comes at the detriment of "Knowing-in-connection." Far from a distant and "objective" discretizing logic, knowing is an activity that happens in the relationship between the knower and the known. Proposing an understanding of human knowing in analogy with loving, De Jaegher argues that in knowing, like loving, what happens is not neutral, general, or universal. Knowers, like lovers, are not abstract subjects but are particular and concrete. "Who loves matters"—and both loving and knowing take place in the relation between them.[35]

Human knowing is based not on purely rational logic, as the rational worldview has assumed, but on living and connected know-hows. "Our most sophisticated knowing," according to De Jaegher, "is full of uncertainty, inconsistencies, and ambiguities." One of the consequences of prioritizing reason is that knowledge of the world and of other people becomes something that is rooted in the individual person's rational reasoning, in direct contrast to engaged, active, involved, and implicated

knowing. Humans are inherently historical, social, cultural, gendered, politicized, and contextualized organisms. Accordingly, their knowing and understanding of the world around them necessarily takes place through their respective lenses.

People are not solo cognizers that manipulate symbols in their heads and perceive their environment in a passive way, as the rationalist view would suggest, but they actively engage with the world around them in a meaningful and unpredictable way. Living bodies, according to Di Paolo et al.,[37] are processes, practices, and networks of relations which have "more in common with hurricanes than with statues." They are unfinished and always becoming, marked by "innumerable relational possibilities, potentialities and virtualities" and not calculable entities whose behavior can neatly be categorized and predicted in a precise way. Bodies "… grow, develop, and die in ongoing attunement to their circumstances … Human bodies are path-dependent, plastic, nonergodic, in short, historical. There is no true averaging of them."[37] (p.97) What might a version of ethics—in the context of data practices and algorithmic systems—that takes the core values of enactive cognitive science and Afro-feminist epistemology (described in the two preceding subsections) as its foundations look like? The next section details this issue.

Before we delve into that, it is worth reemphasizing that while the rational worldview tends to see knowledge, people, and reality in general as stable, for relational perspectives, we are fluid, active, and continually becoming. Nonetheless, the relational versus rational divide is not something that can be clearly demarcated, but overlaps with fuzzy boundaries. Some approaches might prove difficult to fit in either category while others serve to bridge the gap: Harding's[38] "strong objectivity" is one such example that links relational and rational approaches. Furthermore, the relational and rational traditions exist in tension with a continual push and pull. For example, complexity science is a school of thought that emerged from this tension.

ETHICS BUILT ON THE FOUNDATIONS OF RELATIONALITY

> "Ethics is a matter of practice, of down-to-earth problems and not a matter of those categories and taxonomies that serve to fascinate the academic clubs and their specialists."
>
> Heinz von Foerster[39]

What does the idea of ethics—within the context of data practices and algorithmic systems—built on the foundations of relationality look like? This section seeks to elucidate this issue. What follows is not a set of general guidelines, or principles, or a set of out-of-the-box tools that can be implemented to supposedly cleanse datasets of bias or to make a set of existing algorithmic tools "ethical," for the problems we are trying to grasp are deeply rooted, fluid, contingent, and complex. Neither is it a rationally and logically constructed "theory of ethics" that hypothesizes about morality in abstract terms. Rather, the following are the central tenets, informed by Afro-feminist and enactivist perspectives outlined in the previous section, which should aid in shifting toward an understanding of people and of concepts such as data, ethics, algorithms, matrices of oppression, and structural inequalities as inherently interlinked and processual.

Knowing that centers human relations

Since knowing is a relational affair, it matters who enters into the knower-known relations. Within the fields of computing and data sciences, the knower is heavily dominated by privileged groups of mainly elite, Western, cis-gendered, and able-bodied white men.[40] Given that knower and known are closely tied, this means that most of the knowledge that such fields produce is reduced to the perspective, interest, and concerns of such a dominant group. Subsequently, not only are the most privileged among us restricted to producing partial knowledge that fits a limited worldview (while such knowledge, tools, and technologies they produce are forced onto all groups, often disproportionately onto marginalized people), they are also poorly equipped to recognize injustice and oppression.[41] D'Ignazio and Klein[42] call this phenomenon "the privilege hazard." This means that minoritized populations (1) experience harm disproportionally and (2) are better suited to recognize harm due to their epistemic privilege.[43]

Centering the disproportionally impacted

The harm, bias, and injustice that emerge from algorithmic systems varies and is dependent on the training and validation data used, the underlying design assumptions, and the specific context in which the system is deployed, among other factors. However, one thing remains constant: individuals and communities that are at the margins of society are disproportionally impacted. Some examples include object detection,[44] search engine results,[45] recidivism,[46] gender recognition,[47] gender classification,[48,49] and medicine.[1] The findings of Wilson et.al.,[44] for instance, demonstrate that object detection systems designed to predict pedestrians display higher error rates identifying dark-skinned pedestrians while light-skinned pedestrians are identified with higher precision. The use of such systems situates the recognition of subjectivity with skin tone whereby whiteness is taken as the ideal mode of being. Furthermore, gender classification systems often operate under essentialist assumptions and operationalize gender in a trans-exclusive way, resulting in disproportionate harm to trans people.[48,50]

Given that harm is distributed disproportionately and that the most marginalized hold the epistemic privilege to recognize harm and injustice, relational ethics asks that for any solution that we seek, the starting point be the individuals and groups that are impacted the most. This means we seek to center the needs and welfare of those that are disproportionally impacted and not solutions that benefit the majority. Most of the time this means not simply creating a fairness metric for an existing system but rather questioning what the system is doing, particularly examining its consequences on minoritized and vulnerable groups. This requires us to zoom out and draw the bigger picture: a shift from asking "how can we make a certain dataset representative?" to examining "what is the product or tool being used for? Who benefits? Who is harmed?"

To some extent, the idea of centering the disproportionally impacted shares some commonalities with aspects of participatory design, where design is treated as a fundamentally participatory act,[51] and even aspects of human-centered design,[52]

where individuals or groups within a society are placed at the center. However, the idea of centering the disproportionally impacted goes further than human-centered or participatory design as broadly construed. While the latter approaches often neglect those at the margins[53] and shy away from power asymmetries and structural inequalities that permeate the social world, and "mirror individualism and capitalism by catering to consumer's purchasing power at the expense of obscuring the hidden labor that is necessary for creating such system"[54] for the former, acknowledging these deeply ingrained structural hierarchies and hidden labor is a central starting point. In this regard, with a great emphasis on asymmetrical power relations, works such as Costanza-Chock's[55] *Design Justice* and Harrington's[53] *The Forgotten Margins* are examples that provide insights into how centering the disproportionately impacted might be realized through design led by marginalized communities.

The central implication of this in the context of a justice-centered data practice is that minoritized populations that experience harm disproportionately hold the epistemic authority to recognize injustice and harm given their lived experience. Understanding of these concepts, therefore, needs to proceed from the experience and testimony of the disproportionately harmed. The starting point toward efforts such as ethical practice in machine-learning systems or theories of ethics, fairness, or discrimination needs to center the material condition and the concrete consequences an algorithmic tool is likely to bring. Having said that, these are efforts with extreme nuances and magnitudes of complexity in reality. For example, questions such as "how might a data worker engage vulnerable communities in ways that surface harms, when it is often the case that algorithmic harms may be secondary effects, invisible to designers and communities alike, and what questions might be asked to help anticipate these harms?" and "how do we make frictions, often the site of power struggles, visible?" are difficult questions but questions that need to be negotiated and reiterated by communities and data workers.

Bias is not a deviation from the "correct" description

One of the characteristics of a rational worldview is the tendency to perceive things as relatively static. In a supposedly objective worldview, bias, injustice, and discrimination are (mis)conceived as being able to be permanently corrected. The common phrase "bias in, bias out" captures this deeply ingrained reductive thinking. Although datasets are often part of the problem, this commonly held belief relegates deeply rooted societal and historical injustices, nuanced power asymmetries, and structural inequalities to mere datasets. The implication is that if one can "fix" a certain dataset, the deeper problems disappear. When we see bias and discrimination, what we see is problems that have surfaced as a result of a field that has thoughtlessly inherited deeply rooted unjust, racist, and white supremacist histories and practices.[56] As D'Ignazio and Klein[42] contend, "addressing bias in a dataset is a tiny technological Band-Aid for a much larger problem." Furthermore, underlying the idea of "fixing" bias is the assumption that there exists a single correct description of reality where a deviation from it has resulted in bias. As we have seen in Rationality: the dominant orthodoxy, the idea of a single correct description, theory, or approach is reminiscent of the rationalist tradition where the "correct way" is often synonymous with the status quo. The idea of bias as something that can be eliminated, so to speak, once and for all, is misleading and problematic. Even if one can suppose that bias in a dataset can be "fixed," what exactly are we fixing? What is the supposedly bias-free tool being applied to? Is it going to result in net benefit or harm to marginalized communities? Is the supposedly "bias-free" tool used to punish, surveil, and harm anyway? And in Kalluri's[57] words, "how is AI shifting power" from the most to the least privileged? Looking beyond biased datasets and into deeper structural issues, historical antecedents, and power asymmetries is imperative. The rationalist worldview and its underlying assumptions are pervasive and take various nuanced forms. Within the computation and data sciences, the propensity to view things as relatively static manifests itself in the tendency to formulate subjects of study (people, ethics, and complex social problems in general) in terms of problem → solution. Not only are subjects of study that do not lend themselves to this formulation discarded but also, this tradition rests on a misconception that injustice, ethics, and bias are relatively static things that we can solve once and for all. Concepts such as bias, fairness, and justice, however, are moving targets. As we have discussed in Relationality, neither people nor the environment and context in which they are embedded are static. What society deems fair and ethical changes over time and with context and culture. The concepts of fairness, justice, and ethical practice are continually shifting. It is possible that what is considered ethical currently and within certain domains for certain societies will not be perceived similarly at a different time, in another domain, or by a different society. This, however, is not a call to relativism but rather an objection to static and final answers in the face of fluid reality. Adopting relational ethics means that we view our understandings, proposed solutions, and definitions of bias, fairness, and ethics as partially open. This partial openness allows for revision and reiteration in accordance with the dynamic development of such challenges. This also means that this work is never done.

Prioritizing of understanding over prediction

> "I have never been impressed with claims that structural linguistics, computer engineering or some other advanced form of thought is going to enable us to understand men without knowing them."
>
> Clifford Geertz[58]

The rationalist tradition's tendency toward timeless and generalizable knowledge aspires to establish timeless laws and generalizable theories. This pipeline takes observed commonalities, recurring similarities, and repeated patterns among past events or particular behaviors and abstracts them into generalizations that can be applied toward forecasting the future. Because the rationalist's focus is to uncover what remains constant regardless of context, culture, and time, the rationalist view embraces abstraction, generalization, and universal principles at the expense of concrete, particular, and contextual understanding—that is, knowledge grounded in active, concrete, and reciprocal relationships. According to Geertz,[58] the desire to formulate general theories is in an irremovable tension with the need to gain deep understanding of particular and contextual events

and behaviors. The further theory goes, the deeper the tension. Geertz suggests that theories and generalizations inevitably lack deep and contextual understanding of human thought. Theoretical disquisitions stand far from the immediacies of social life. Any generalization or theory constructed in the absence of deep understanding, not grounded in the concrete and particular, is vacuous.

On a similar note, the Russian philosopher Mikhail Bakhtin refers to the manner in which abstract general rules are derived from concrete human actions and behaviors as "theoretism." Bakhtin argues that such attempts to abstract general rules from particulars "loses the most essential thing about human activity, the very thing in which the soul of morality is to be found," which Bakhtin calls the "eventness" of the event.[59] Eventness is always particular, and never exhaustively describable in terms of rules. To understand people, we must take into account "unrepeatable contextual meaning." Likewise, the historian of science Lorraine Daston contends that the endeavor for a universal law is a predicament that does not stand against unanticipated particulars, since no universal ever fits the particulars.[60] Commenting on current machine-learning practices, Daston[61] explains: "machine learning presents an extreme case of a very human predicament, which is that the only way we can generalize is on the basis of past experience. And yet we know from history—and I know from my lifetime—that our deepest intuitions about all sorts of things, and in particular justice and injustice, can change dramatically."

While the rationalist tradition tends to aspire to produce generalizable knowledge disentangled from historical baggage, context, and human relations, relationalist perspectives strive for concrete, contextual, and relational understanding of knowledge, human affairs, and reality in general. Data science and machine-learning systems sit firmly within the rationalist tradition. The core of what machine-learning systems do can be exemplified as clustering similarities and differences, abstracting commonalities, and detecting patterns. Machine-learning systems "work" by identifying patterns in vast amounts of data. Given immense, messy, and complex data, a machine-learning system can sort, classify, and cluster similarities based on seemingly shared features. Feed a neural network labeled images of faces and it will learn to discern faces from not-faces. Not only do machine-learning systems detect patterns and cluster similarities, they also make predictions based on the observed patterns.[62] Machine learning, at its core, is a tool that predicts. It reveals statistical correlations with no understanding of causal mechanisms.

Relational ethics, in this regard, entails moving away from building predictive tools (with no underlying understanding) to valuing and prioritizing in-depth and contextual understanding. This means we examine the patterns we find and ask why we are finding such patterns. This in turn calls for interrogating contextual and historical norms and structures that might give rise to such patterns instead of using the findings as input toward building predictive systems and repeating existing structural inequalities and historical oppression. If we go back to the Bayesian models of inference mentioned in Rationality: the dominant orthodoxy, we find that such models are prone to amplification of socially held stereotypes. Repeating Horgan's point:[25] "Embedded in Bayes' theorem is a moral message: If you aren't scrupulous in seeking alternative explanations for your evidence, the evidence will just confirm what you already believe." A data practice that prioritizes understanding over prediction is one that interrogates prior beliefs instead of using the evidence to confirm such belief and one that seeks alternative explanations by placing the evidence in a social, historical, and cultural context. In doing so, we ask challenging but important questions such as "to what extent do our initial beliefs originate in stereotypically held intuitions about groups or cultures?", "why are we finding the 'evidence' (patterns) that we are finding?", and "how can we leverage data practices in order to gain an in-depth understanding of certain problems as situated in structural inequalities and oppression?"

Data science as a practice that alters the social fabric

> "Technology is not the design of physical things. It is the design of practices and possibilities."
>
> Lucy Suchman[63]

Machine classification and prediction are practices that act directly upon the world and result in tangible impact.[64] Various companies, institutes, and governments use machine-learning systems across a variety of areas. These systems process people's behaviors, actions, and the social world at large. The machine-detected patterns often provide "answers" to fuzzy, contingent, and open-ended questions. These "answers" neither reveal any causal relations nor provide explanation on why or how.[65] Crucially, the more socially complex a problem is, the less capable machine-learning systems are of accurately or reliably classifying or predicting.[66] Yet analytics companies boast their ability to provide insight into the human psyche and predict human behavior.[67] Some even go so far as to claim to have built AI systems that are able to map and predict "human states" based on speech analysis, images of faces, and other data.[68]

Thinking in relational terms about ethics begins with reconceptualizing data science and machine learning as practices that create, sustain, and alter the social world. The very declaration of a taxonomy brings some things into existence while rendering others invisible.[7] For any individual person, community, or situation, algorithmic classifications and predictions give either an advantage or they hinder. Certain patterns are made visible and types of being objectified while other types are erased. Some identities (and not others) are recognized as a pedestrian,[44] or fit for an STEM career,[69] or in need of medical care.[1] Some are ignored and made invisible altogether.

Categories simplify and freeze nuanced and complex narratives obscuring political and moral reasoning behind a category. Over time, messy and contingent histories and political and moral stories hidden behind a category are forgotten and trivialized.[70] The process of categorizing, sorting, and generalizing, therefore, is far from a mere technical task. While seemingly invisible in our daily lives, categorization and prediction bring forth some behaviors and ways of being as "legitimate," "standard," or "normal" while casting others as "deviant."[70] Seemingly banal tasks such as identifying and predicting "employable" or "criminal" characteristics carry grave consequences for those that do not conform to the status quo.

Relational ethics encourages us to view data science in general, and the tasks of developing and deploying algorithmic tools that cluster and predict, as part of the practice of creating and reinforcing existing and historical inequalities and structural injustices. Therefore, in treating data science as a practice that alters the fabric of society, the data practitioner is encouraged to zoom out and ask such questions as "how might the deployment of a specific tool enable or constrain certain behaviors and actions?", "does the deployment of such a tool enable or limit possibilities, and for whom?", and "in the process of enabling some behaviors while constraining others, how might such a tool be encouraging/discouraging certain social discourse and norms?"

CLOSING REMARKS

Rethinking ethics is about undoing previous and current injustices to society's most minoritized and empowering the underserved and systematically disadvantaged. This entails not devising ways to "debias" datasets or derive abstract "fairness" metrics but zooming out and looking at the bigger picture. Relational ethics encourages us to examine fundamental questions and unstated assumptions. This includes interrogating asymmetrical and hierarchical power dynamics, deeply ingrained social and structural inequalities, and assumptions regarding knowledge, justice, and technology itself.

Ethical practice, especially with regard to algorithmic predictions of social outcomes, requires a fundamental rethinking of justice, fairness, and ethics above and beyond technical solutions. Ethics in this regard is not merely a methodology, a tool, or simply a matter of constructing a philosophically coherent theory but a down-to-earth practice that is best viewed as a habit—a practice that alters the way we do data science. Relational ethics is a process that emerges through the re-examination of the nature of existence, knowledge, oppression, and injustice. Algorithmic systems never emerge in a social, historical, and political vacuum, and to divorce them from the contingent background in which they are embedded is erroneous. Relational ethics provides the framework to rethink the nature of data science through a relational understanding of being and knowing.

ACKNOWLEDGMENTS

This work was supported, in part, by Science Foundation Ireland grant 13/RC/2094_P2 and co-funded under the European Regional Development Fund through the Southern and Eastern Regional Operational Programme to Lero – the Science Foundation Ireland Research Center for Software (www.lero.ie). I would like to thank Anthony Ventresque, Dan McQuillan, Elayne Ruane, Hanne De Jaegher, Johnathan Flowers, and Thomas Laurent for their useful feedback on an earlier version of the manuscript. I would also like to extend my deepest gratitude to the anonymous reviewers who provided a thorough review and invaluable feedback on a previous version of the manuscript.

REFERENCES

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

2. Lum, K., and Isaac, W. (2016). To predict and serve? Significance *13*, 14–19.

3. Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor (St. Martin's Press).

4. Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. (2020). Mitigating bias in algorithmic hiring: evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 469–481.

5. Ajunwa, I., Friedler, S.A., Scheidegger, C., and Venkatasubramanian, S. (2016). Hiring by algorithm: predicting and preventing disparate impact, Yale Law School Information Society Project Conference Unlocking the Black Box: The Promise and Limits of Algorithmic Accountability in the Professions http://sorelle.friedler.net/papers/SSRN-id2746078.pdf.

6. McQuillan, D. (2020). Non-fascist AI. In Propositions for Non-Fascist Living: Tentative and Urgent, M. Hlavajova and W. Maas, eds. (MIT Press/BAK), pp. 113–124.

7. Bowker, G.C., and Star, S.L. (2000). Sorting Things Out: Classification and its Consequences (MIT Press).

8. Gonen, H., and Goldberg, Y. (2019). Lipstick on a pig: debiasing methods cover up systematic gender biases in word embeddings but do not remove them. arXiv, 1903.03862.

9. Menkiti, I.A. (1984). Person and community in African traditional thought. In African Philosophy: An Introduction, 3rd edn, R. Wright, ed. (University Press of America), pp. 171–182.

10. Juarrero, A. (2000). Dynamics in action: intentional behavior as a complex system. Emergence *2*, 24–57.

11. Mhlambi, S. (2020). From rationality to relationality: ubuntu as an ethical and human rights framework for artificial intelligence governance. In Carr Center for Human Rights Policy Discussion Paper Series, 2020-009 https://carrcenter.hks.harvard.edu/publications/rationality-relationality-ubuntu-ethical-and-human-rights-framework-artificial.

12. Terry, W., and Flores, F. (1986). Understanding Computers and Cognition: A New Foundation for Design (Intellect Books).

13. Prigogine, I., and Stengers, I. (1984). Order Out of Chaos: Man's New Dialogue with Nature (Verso Books).

14. Descartes, R. (1984). The Philosophical Writings of Descartes, *Vol. 2* (Cambridge University Press).

15. Gardiner, M. (1998). The incomparable monster of solipsism: Bakhtin and Merleau-Ponty. In Bakhtin and the Human Sciences: No Last Words, M.M. Bell and M. Gardiner, eds. (London: Sage), pp. 128–144.

16. Merleau-Ponty, M. (1968). The Visible and the Invisible: Followed by Working Notes (Northwestern University Press).

17. Ahmed, S. (2007). A phenomenology of whiteness. Femin. Theor. *8*, 149–168.

18. Preiser, R. (2016). Critical Complexity: Collected Essays of Paul Cilliers (Walter de Gruyter GmbH & Co KG).

19. Cilliers, P. (2002). Complexity and Postmodernism: Understanding Complex Systems (Routledge).

20. Bayes, T., and Price, N. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F.R.S. communicated by Mr. Price, in a letter to John Canton, A.M. F.R.S. In Philosophical Transactions of the Royal Society of London, *53*, pp. 370–418, https://doi.org/10.1098/rstl.1763.0053.

21. Hahn, U. (2014). The Bayesian boom: good thing or bad? Front. Psychol. *5*, 765.

22. Seth, A.K. (2014). The Cybernetic Bayesian Brain. Open MIND (MIND Group).

23. Jones, M., and Love, B.C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. Behav. Brain Sci. *34*, 169.

24. Pager, D., and Karafin, D. (2009). Bayesian bigot? Statistical discrimination, stereotypes, and employer decision making. Ann. Am. Acad. Polit. Soc. Sci. *621*, 70–93.

25. Horgan, J. (2016). Bayes's theorem: what's the big deal?. https://blogs.scientificamerican.com/cross-check/bayes-s-theorem-what-s-the-big-deal/.

26. Gitelman, L. (2013). Raw Data Is an Oxymoron (MIT Press).

27. Kyselo, M. (2014). The body social: an enactive approach to the self. Front. Psychol. *5*, 986.

28. Mbiti, J.S. (1969). African Religions and Philosophy (Heinemann).

29. Bakhtin, M. (1984). Problems of Dostoevsky's poetics (University of Minnesota Press). https://doi.org/10.5749/j.ctt22727z1.

30. Birhane, A. (2017). Descartes was wrong:'a person is a person through other persons'. Aeon https://aeon.co/ideas/descartes-was-wrong-a-person-is-a-person-through-other-persons.

31. Collins, P.H. (2002). Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment (Routledge).

32. Loukissas, Y.A. (2019). All Data Are Local: Thinking Critically in a Data-Driven Society (MIT Press).

33. Nnaemeka, O. (2004). Nego-feminism: theorizing, practicing, and pruning Africa's way. Signs *29*, 357–385.

34. Nadar, S. (2014). "Stories are data with soul"—lessons from black feminist epistemology. Agenda *28*, 18–28.

35. De Jaegher, H. (2019). Loving and knowing: reflections for an engaged epistemology. Phenomenol. Cogn. Sci. https://doi.org/10.1007/s11097-019-09634-5.

36. De Jaegher, H., and Di Paolo, E. (2007). Participatory sense-making. Phenomenol. Cogn. Sci. *6*, 485–507.

37. Di Paolo, E.A., Cuffari, E.C., and De Jaegher, H. (2018). Linguistic Bodies: The Continuity between Life and Language (MIT Press).

38. Harding, S. (1992). Rethinking standpoint epistemology: what is "strong objectivity?". Centennial Rev. *36*, 437–470.

39. von Foerster, H., and Poerksen, B. (2002). The metaphysics of ethics: a conversation. Cybernet. Hum. Know. *9*, 149–157.

40. Meredith, B. (2018). Artificial Unintelligence: How Computers Misunderstand the World (MIT Press).

41. Berenstain, N. (2016). Epistemic exploitation. Ergo *3*, https://doi.org/10.3998/ergo.12405314.0003.022.

42. D'Ignazio, C., and Klein, L.F. (2020). Data Feminism (MIT Press).

43. Bar On B.-A.. "Marginality and epistemic privilege". In: Alcoff L. Potter E. Feminist Epistemologies Routledge 83–100.

44. Wilson, B., Hoffman, J., and Morgenstern, J. (2019). Predictive inequity in object detection. arXiv, 1902.11097.

45. Noble, S.U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism (NYU Press).

46. Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

47. Buolamwini, J., and Gebru, T. (2018). Gender shades: intersectional accuracy disparities in commercial gender classification. PMLR *81*, 77–91.

48. Hamidi, F., Scheuerman, M.K., and Branham, S.M. (2018). Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3173574.3173582.

49. Pinar, B., Kyriakou, K., Guest, O., Kleanthous, S., and Otterbacher, J. (2020). To "see" is to stereotype. Proc. ACM Hum.-Comput. Interact. *4*, https://doi.org/10.1145/3432931.

50. Keyes, O. (2018). The misgendering machines: trans/HCI implications of automatic gender recognition. Proc. ACM Hum.-Comput. Interact. *4*, https://doi.org/10.1145/3274357.

51. Slavin, K. (2016). Design as participation. J. Des. Sci. https://doi.org/10.21428/a39a747c.

52. Irani, L., Vertesi, J., Dourish, P., Philip, K., and Grinter, R.E. (2010). Post-colonial computing: a lens on design and development. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1311–1320.

53. Harrington, C.N. (2020). The forgotten margins: what is community-based participatory health design telling us? Interactions *27*, 24–29.

54. Lloyd, A., Mancuso, D., Sonis, D., and Hubert, L. (2020). Camera obscura: beyond the lens of user-centered design. https://alexis.medium.com/camera-obscura-beyond-the-lens-of-user-centered-design-631bb4f37594.

55. Costanza-Chock, S. (2018). Design Justice: towards an intersectional feminist framework for design theory and practice. In Proceedings of the Design Research Society. https://doi.org/10.21606/drs.2018.679.

56. Birhane, A., and Guest, O. (2020). Towards decolonising computational sciences. arXiv, 2009.14258.

57. Kalluri, P. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. Nature *583*, 169.

58. Geertz, C. (1973). The Interpretation of Cultures (Basic Books).

59. Morson, G.S., and Emerson, C. (1989). Rethinking Bakhtin: Extensions and Challenges (Northwestern University Press).

60. Daston, L. (2018). Calculation and the division of labor, 1750-1950. Bull. German Hist. Inst. *62*, 9–30.

61. Gross, J. (2020). Historicizing the self-evident. https://www.phenomenalworld.org/interviews/historicizing-the-self-evident.

62. O'Neil, C., and Schutt, R. (2013). Doing Data Science: Straight Talk from the Frontline (O'Reilly Media, Inc.).

63. Suchman, L. (2007). Human-Machine Reconfigurations: Plans and Situated Actions (Cambridge University Press).

64. McQuillan, D. (2018). Data science as machinic neoplatonism. Philos. Technol. *31*, 253–272.

65. Pasquale, F. (2015). The Black Box Society (Harvard University Press).

66. Salganik, M.J., Lundberg, I., Kindel, A.T., Ahearn, C.E., Al-Ghoneim, K., Almaatouq, A., Altschul, D.M., Brand, J.E., Carnegie, N.B., Compton, R.J., et al. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. Proc. Natl. Acad. Sci. U S A *117*, 8398–8403.

67. Qualtrics (2020). Build technology that closes experience gaps. https://www.qualtrics.com/uk/.

68. Affectiva (2020). Affectiva human perception AI analyzes complex human states. https://www.affectiva.com/.

69. Lambrecht, A., and Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Manag. Sci. *65*, 2966–2981.

70. Star, S.L., and Bowker, G.C. (2007). Enacting silence: residual categories as a challenge for ethics, information systems, and communication. Ethics Inform. Technol. *9*, 273–280.

About the Authors

Abeba Birhane (she/her) is a cognitive science PhD candidate at the Complex Software Lab at University College Dublin, Ireland. Her interdisciplinary research aims to connect the dots between complex adaptive systems, machine learning, and critical race studies. More specifically, Birhane studies how machine prediction, especially of social outcomes, is dubious and potentially harmful to vulnerable and marginalized communities.