Encoding normative ethics: On algorithmic bias and disability
by Ian Moura

Abstract
Computer-based algorithms have the potential to encode and exacerbate ableism and may contribute to disparate outcomes for disabled people. The threat of algorithmic bias to people with disabilities is inseparable from the longstanding role of technology as a normalizing agent, and from questions of how society defines shared values, quantifies ethics, conceptualizes and measures risk, and strives to allocate limited resources. This article situates algorithmic bias amidst the larger context of normalization, draws on social and critical theories that can be used to better understand both ableism and algorithmic bias as they operate in the United States, and proposes concrete steps to mitigate harm to the disability community as a result of algorithmic adoption. Examination of two cases — the allocation of lifesaving medical interventions during the COVID-19 pandemic and approaches to autism diagnosis and intervention — demonstrate instances of the mismatch between disabled people’s lived experiences and the goals and understandings advanced by nondisabled people. These examples highlight the ways particular ethical norms can become part of technological systems, and the harm that can ripple outward from misalignment of formal ethics and community values.

Contents
I. Introduction
II. Background: Algorithms, fairness, and technology’s relationship to societal norms
III. Disability, normalization, and algorithmic bias: Two case studies
IV. Proposed remedies
V. Conclusion

I. Introduction
As digital algorithms have become more widely incorporated into a variety of decision-making processes, algorithmic bias has become an increasing concern among researchers, policy-makers, activists, and members of the public. A growing body of literature discusses the disproportionate harm that algorithmic bias can create for marginalized groups, particularly on the basis of racial identity, but only recently has that work begun to recognize disabled people as uniquely at risk of harm from algorithmic
use. Misalignment between the intention of fairness in adopting an algorithmic tool and the actual outcome that tool creates is a particular risk for disabled people, who are less likely to be considered when algorithms are developed and implemented, and more likely to have needs and experiences that result in their classification as outliers. The assumptions involved in developing and using algorithmic tools not only shape approaches to equity; they are also reflected in who gets to define fairness, whose rights are prioritized, and whose humanity is accepted.

This paper situates algorithmic bias within an overall understanding of technology, disability, and ableism as interrelated forces which interact with societal power dynamics. Drawing on work from multiple fields, it connects digital algorithms to a larger history of quantification and categorization as tools of normalization. Using two case studies, one based on the development and use of Crisis Standards of Care plans during the COVID-19 pandemic and one drawing on work from activists and scholars within the autistic self-advocacy movement, this article demonstrates how the issues emerging as digital algorithms are more widely adopted relate to longstanding efforts to measure, categorize, and normalize disabled people, and that consequently, the risks of algorithmic bias to people with disabilities are inextricable from the broader ways in which ableism is enacted through intervention on disability. Finally, the paper concludes with suggestions for potential safeguards that may decrease the risk of algorithmic bias toward people with disabilities, and calls on researchers and practitioners to recognize the limitations of these responses and address the underlying social structures that contribute to algorithmic harms.

II. Background: Algorithms, fairness, and technology’s relationship to societal norms

What is an algorithm?

At the most basic level, an algorithm is a set of instructions which can be followed step-by-step to arrive at a particular outcome. In the context of digital algorithms, these instructions are fed into a computer, and results may or may not be mediated by human oversight; digital algorithms can range from a rule applied in a spreadsheet to custom-built machine learning techniques (Wieringa, 2020). Like all technology, algorithms should be considered as a system, rather than as individual artifacts, and as expressions of power and control (Franklin, 2004; Winner, 1980). Though algorithms are often presented as a way to quantify fairness and render decision-making more objective, there are numerous ways to define “fairness,” many of which are incompatible with one another or involve significant tradeoffs (Corbett-Davies, et al., 2017; Mitchell, et al., 2021; Saxena, et al., 2020).

Algorithmic bias: Defining what it means to be “fair”

Algorithmic bias can be understood as decision-making by an algorithmic tool that is prejudicial towards a particular person or group (Ntoutsi, et al., 2020). Consequently, algorithmic bias is closely linked to the distribution of power within society, and to which members have the ability to define and regulate fair outcomes (Barabas, et al., 2020; Kasy and Abebe, 2021). Though algorithms often obscure such power dynamics, there is increasing recognition of the need to re-center an understanding of power within discussions of algorithmic fairness (Barabas, et al., 2020; Kasy and Abebe, 2021; Selbst, et al., 2019).

When algorithms fail to accurately recognize, classify, or make predictions about disabled people, or otherwise perform in ways that perpetuate bias, the people who have designed or implemented these tools sometimes justify this by saying that people with disabilities are outliers, and that it is therefore difficult to design or train algorithms to treat disabled people fairly. However, disabled people are not inherently more difficult to identify, categorize, or understand. Instead, these failures result from specific decisions made during the design and development of the algorithms themselves. The frequent decision to treat disabled people as outliers or “edge cases” all but guarantees that once adopted, algorithms will perpetuate unequal treatment and encode the existing social reality where disabled people are assumed to be less fundamentally
human and less representative of human experiences (Nakamura, 2019). For example, the pedestrian detection systems employed in many autonomous vehicles are not designed around a broad and inclusive notion of how people move around cities and roadways, and as a result, these systems struggle to recognize people using wheelchairs or other mobility aids as humans. Algorithms designed to monitor student behavior during online test-taking may flag disabled students as “suspicious” or assume they are cheating because their activity does not match the algorithm’s standard for “normal” test-taking. Similarly, automated speech recognition may fail to understand someone who stutters or whose disability otherwise impacts their speech, because their communication is too different from the examples of spoken language used to develop the program.

Although digital algorithms, particularly those that are more fully automated, are often discussed in terms that suggest they perform specific tasks similarly to the way a human would, the reality is that algorithms function in ways that are fundamentally incomparable to human thought and reasoning. Before they can “make decisions,” algorithms must be specified and trained by their human creators; regardless of the algorithm involved, this process requires data. In some cases, datasets are intentionally and systematically constructed; in others, they are scraped from publicly available sources, and when necessary, labeled manually, often using crowd work platforms (Gray and Suri, 2019; Paullada, et al., 2021). These labels are approximations of the data objects they are attached to, not direct representations, and the process of creating them is interpretive, often poorly documented and thus difficult to replicate or even understand, and subject to cultural, contextual, and personal biases (Barocas and Selbst, 2016; boyd and Crawford, 2012; Hoffmann, 2019; Paullada, et al., 2021). Recent studies have found models trained on these datasets can replicate human biases, encoding prejudicial attitudes about aspects of identity including race, gender, and disability status (Hutchinson, et al., 2020; Steed and Caliskan, 2021).

Algorithmic bias can, and often does, arise from issues related to training data, so addressing issues related to algorithmic bias and disability requires a consideration of the data on which algorithms rely and operate. It is critical that designers, developers, researchers, and policy-makers all recognize that none of these data are an objective reflection of the world. Rather, they result from a series of choices and decisions, and frequently contain the same biases and prejudices that permeate society (Hoffmann, 2019; Paullada, et al., 2021). For example, ableism can lead doctors to deny disabled people access to certain kinds of lifesaving care, or to provide less intensive treatment when they face a critical illness. Data that are collected about how different people fared following a medical crisis will contain evidence of this bias — for instance, by showing that disabled people have worse outcomes — but without an awareness of the role of ableism in shaping access to the kinds of care that facilitate recovery from such events, these data may be assumed to show that disability leads to worse outcomes, rather than that ableism causes disabled people to get poorer quality care.

A lack of data from and about people with disabilities also makes it difficult, or even impossible, to detect unfair treatment of disabled people by an algorithm, a situation that is compounded by limited recognition of the heterogeneity of the disability community (Andrus, et al., 2020). Limited representation of people with disabilities within data used to develop, train, and test algorithms has multiple causes, including the lack of systematic approaches to dataset development noted previously, as well as general limitations in the amount, quality, and representativeness of data on people with disabilities currently available (Packin, 2021; Paullada, et al., 2021). Since disability is complex and multifaceted, and disabled people are a diverse, heterogeneous group, disabled people who have additional marginalized identities, such as disabled women, disabled nonbinary people, or disabled people of color, are particularly underrepresented (Packin, 2021). As with other aspects of digital algorithms, the lack of disability data is one manifestation of an overarching issue. In order to be represented in an algorithm, concepts have to be abstracted away from how they exist in the world; for example, fairness becomes a matter of optimization or distribution, rather than something that relates to fundamental relationships within a society. Similarly, even when algorithms do include representations of disability, they are generally simplified or based on data and measurements that are most easily available, such as diagnostic codes from electronic health records or whether someone receives specific governmental disability benefits. These ways of understanding disability divorce it from the societal and relational context in which it exists, and simplify disability into a discrete, quantifiable
Encoding normative ethics: On algorithmic bias and disability

characteristic, rather than the intermingling of identity and experience that it so often is.

Beyond the way an algorithm functions or the data it operates on, fairness also depends on the context in which an algorithm is used. The implicit acceptance of normative assumptions about power, particularly in terms of what should be predicted or categorized by an algorithm and about whom, frequently leads to the construction of technological tools focusing on people who already hold less power in society (Barabas, et al., 2020). In the context of disability, this can lead to algorithms that discriminate on the basis of being a patient within a medicalized framework that views disability as a problem to be solved, rather than an axis of diversity (Bennett and Keyes, 2020). These power dynamics also influence the types of problems to which algorithmic solutions are applied; for instance, to “optimize” and in effect, limit, the amount of in-home care disabled people receive, but not to streamline the process of applying for disability benefits. For disabled people whose identities are intersectionally complex — such as disabled people of color, or queer disabled people — even an algorithm that purports to treat disabled people fairly may have differential impacts that put them at elevated risk of harm.

Technology, disability, and the encoding of societal norms

Scholarship and commentary on algorithmic bias often draws on literature that examines technology as an expression of societal norms. However, it has been less common for work on algorithmic bias to consider how this creates unique risks for disabled people, or to connect these threats to the well-being and autonomy of disabled people to the “analog” [4] algorithms and non-algorithmic technologies that preceded them. The kinds of quantification, measurement, and statistical methods on which algorithms rely are bound up in a long history of ableism and eugenics within science, medicine, and technological development. This context, and the way it puts disabled people uniquely at risk of harm from algorithmic tools, is absent from many discussions addressing algorithmic bias. Similarly, disabled people’s role as innovators and hackers who create and re-create technology in response to needs that society fails to recognize or accommodate is consistently omitted from considerations of how to rectify algorithmic bias.

The concept of technology — both as a general category of human innovation and as an element of progress and modernity — developed following the Industrial Revolution (Marx, 1997). Technology is more than just individual inventions; it is a system defined by networks built around mechanical or computational elements that becomes a primary means of addressing social problems (Michael, et al., 2020; Oelschlaeger, 1979). In the context of disability, technological fixes are often a substitute for creating a more inclusive and accessible society.

Until early in the twentieth century, disability did not exist as a unified concept; rather, a variety of terms were applied to people who would now be considered disabled, and these descriptors did not necessarily imply that a person could not work or lead an independent life (Rose, 2017). Following the American Civil War, accommodations for disabled people in both familial and employment settings decreased as the nation moved towards wage-based, urbanized, industrial capitalism, and disabled people were increasingly segregated from society (Dorfman, 2015; Rose, 2017). Today, in just one example of how different technological advances can serve to create and enforce an overarching value or norm, activists and scholars have documented the role of algorithmic tools embedded in hiring and performance management processes, and their power to discriminate against disabled people (Brown, et al., 2022, 2020; Whittaker, et al., 2019). Folding specific judgments related to ability and competence into algorithms that govern employment — for example, by relying on automated video interviewing tools, or by defining performance in terms of productivity and speed — divorces these subjective decisions about who is a valuable worker from their historic context. By reinforcing the ableist standards that contribute to high unemployment among people with disabilities, these tools maintain a status quo that treats disability as incompatible with full participation in society, while simultaneously offloading responsibility for these judgments to technological specifications, rather than human prejudice.

Disabled people have long been a target of technological innovation, often with the explicit goal of erasing disability and normalizing disabled people. Within most mainstream scholarship on technology, disabled
people have more commonly been seen as problems to be solved, rather than sources of expertise (Hofmann, et al., 2020; Mankoff, et al., 2010; Wu, 2021). In the context of research, treating disability as little more than a potential area of inquiry and intervention contributes to what some scholars have termed “epistemic violence,” whereby disabled people’s expertise and knowledge are dismissed as too subjective or anecdotal to merit real consideration (Ymous, et al., 2020). These attitudes facilitate an oversimplified view of disability which uses empathy activities or simulations as substitutes for authentically engaging with disabled people (Bennett and Rosner, 2019; Hofmann, et al., 2020).

Assuming that disabled people are the subjects of research and innovation, rather than researchers and innovators in their own right, erases the disability community’s history of using technology in unorthodox ways to navigate a world that fails to consider their needs and to build new communities that recognize them as full members. For instance, consider the multitude of “hacks” developed by disabled people when assistive technology is insufficient or nonexistent; the ways in which disabled people have embraced, and often repurposed, novel technologies in the service of access and activism; and the autistic self-advocate community’s embrace of the Internet as a means of building a new kind of social realm (Buckle, 2020; Dekker, 2020; Hamraie and Fritsch, 2019; Maudlin, 2022; Seidel, 2020; Tisoncik, 2020). These liberatory approaches, through which disabled people embrace technology to exercise autonomy and access community, contrast with the development of what Liz Jackson has termed “disability dongles:” sloppy applications of technology aimed at “fixing” the problem of disability, often with little recognition of relevant history or previous research and development (Jackson, et al., 2022; Whittaker, et al., 2019). While disability dongles are regularly covered by journalism and shared on social media, the kinds of hacking and innovation disabled people routinely practice, and the underlying social conditions that so often make them a necessity, are rarely afforded such recognition.

In addition to considering the history and impact of technology as it relates to disability, understanding algorithmic bias requires reckoning with the dual history of eugenics and statistical methods. This history is especially relevant when considering the impact of algorithmic tools on disabled people; along with industrialization, the development of statistics as a distinct approach to knowledge and understanding was a significant contributor to the modern concept of disability (Davis, 2019). The eugenics movement coincided with an increasing interest in quantification and data collection, and a growing emphasis on the use of science and technology to address social problems (Farrall, 1979; Louçã, 2009; Smith, 2020). As noted earlier, the development and widespread adoption of new forms of mechanical technology also led to changes in the nature and structure of employment, and the resulting interest in a physically able and largely interchangeable workforce contributed significantly to the modern concept of disability alongside the eugenics movement’s interest in biology and inheritance as part of overall progress. Both industrialization and the influence of eugenicists helped popularize the concept of normality within science and society, and with it the understanding of “normal” ways of being as both intrinsically superior to deviance or disability and attainable in ways that “ideal” standards might not be (Davis, 2019).

While overt public support for eugenics waned following its association with Nazism during and after the second World War, the movement was nonetheless extremely successful in terms of its influence on statistics, medicine, social science, and public policy. Eugenicist attitudes about who deserves to exercise fundamental rights remain a core aspect of the kinds of discrimination and systemic oppression faced by disabled people (Pfeiffer, 1994). The acceptance of population norms, understanding of individual variance as a source of error to be corrected, and the notion of the mean as the best unbiased estimator of a trait or characteristic of a population all live on within standard methods for research and analysis, and form part of the theoretical basis through which algorithms operate. On a more conceptual level, normality continues to feature prominently in work to measure, categorize, and manage populations, and the understanding of disability as an expression of deviance from normality has been a significant means of justification for inequality based on gender, race, and sexual orientation (Baynton, 2017).

The influence of eugenics on statistics and quantitative social science research, and by extension, on algorithms, is also evident in the downward focus in the way algorithms are adopted and applied. Algorithms are often used to limit access to scarce resource for people who are already disadvantaged,
Encoding normative ethics: On algorithmic bias and disability

creating processes that wealthy, well-connected people can expect to circumvent; numerous examples, from bail-setting algorithms to automated employment screenings, are widespread (Barabas, et al., 2020; Brown, et al., 2020, 2022; Eubanks, 2018; Kasy and Abebe, 2021). The targets of decision-making algorithms are overwhelmingly people who experience marginalization in one form or another, and who rarely command the societal power to influence how such tools are constructed or applied. In many cases, algorithmic tools constitute a “digital poorhouse” that surveils poor and working-class people, conflates economic disadvantage with irresponsibility and risk, and presumes certain values are neutral and universal (Eubanks, 2018). Often, the combination of a normative concept of fairness, scarcity of available resources, and disparate access to these limited goods creates both the kind of problems algorithmic biases are increasingly used to solve and a situation where algorithmic tools have grave potential to amplify existing inequity.

Algorithmic decision-making also has the potential to create self-fulfilling prophecies, contributing to the situations they are intended to predict. For example, a disabled person who an algorithm predicts is less likely to survive an illness is subsequently less likely to receive the medical interventions that might help them, increasing their risk of death, strengthening the evidence for the purported relationship between disability and decreased benefit from medical intervention, and creating additional cases where disability predicts a decreased chance of long-term survival, which can then be fed back into the algorithm.

III. Disability, normalization, and algorithmic bias: Two case studies

As detailed earlier, rather than an example of a new and unique threat, algorithmic bias exploits many of the structural issues that have impacted the disability community for decades. The following case studies explore how harm emerges from misalignment between disabled people’s experiences and needs, and the goals of prediction and classification, regardless of whether they are the result of an algorithmic tool. The first case focuses on the Crisis Standards of Care used during the early stages of the COVID-19 pandemic. The second discusses autism diagnosis and the way it continues to inform specific interventions. Focusing on cases that demonstrate the same pattern of risk and harm to disabled people as in many instances where algorithms are being adopted, but which do not involve the use of digital algorithms, helps clarify the ways that these risks are the product of ableism within society, and not of specific forms in technology.

Prediction amid scarcity: Disability, the COVID-19 pandemic, and Crisis Standards of Care

During the COVID-19 pandemic, anticipated shortages in medical equipment, including ventilators, as well as in the medical professionals needed to support a sudden influx of critically ill patients, led many states to draft or update existing Crisis Standards of Care (CSC) plans (Antommaria, et al., 2020; Cleveland Manchanda, et al., 2020; Emanuel, et al., 2020; Laventhal, et al., 2020; Ne’eman, et al., 2021). Analyses of CSC plans from across the United States conducted early on during the pandemic revealed that the overwhelming majority relied on scoring systems such as the Sequential Organ Failure Assessment (SOFA) as part of determining eligibility for critical care resources, and despite ethical statements that said decisions should not factor in disability, most plans considered co-occurring health conditions like cardiac disease or renal failure in allocation determinations (Antommaria, et al., 2020; Cleveland Manchanda, et al., 2020; Moura, et al., 2020). Though tools like the SOFA were proposed as objective means of determining which patients would most likely benefit from intensive care interventions, the SOFA was designed to be used at a population, not an individual, level and previous research has called into question its reliability as a predictor of survival, as it fails to account for higher baseline scores assigned to people with specific disabilities (Cleveland Manchanda, et al., 2020; Ne’eman, et al., 2020).

Concerns about how medical professionals would allocate scarce resources as a result of these plans generated a national conversation about the assumptions and ethics built into Crisis Standards of Care and related decision-making frameworks. This discussion has emphasized the ways that equal treatment can lead to disparate access and highlighted the ableism inherent in many standardized measures of health, which in turn lead to biased predictions of risk and render disabled people less likely to receive life-saving
Encoding normative ethics: On algorithmic bias and disability

medical treatment in times of scarcity (Andrews, et al., 2021; Cleveland Manchanda, et al., 2020; Guidry-Grimes, et al., 2020). The specific guidelines encompassed in many of these plans [5], coupled with the disproportionate impact of the pandemic on congregate care settings, have demonstrated the dual impact of ableism in shaping both access to mechanical ventilation and other lifesaving medical care and structural definitions of fairness (Andrews, et al., 2021; Guidry-Grimes, et al., 2020). While disability advocates were able to challenge many existing CSC plans, leading to revisions that reduced their bias against disabled patients, plans which were not updated during the pandemic remained significantly more likely to include categorical exemptions (Ne’eman, et al., 2021).

The checklists and scoring tools incorporated into many CSC Plans fall under the broad conceptualization of algorithms introduced earlier in this paper, but they are not the kind of digital algorithms that most work on algorithmic bias has focused on. However, the “analog” algorithms used in Crisis Standards of Care nonetheless demonstrate many of the same risks to disabled people as their digital counterparts. For example, just as algorithmic tools rely on the concept of a statistical normal, the SOFA and other scoring tools are based on the premise of a standard baseline for health which is used to divide individuals into groups that score above or below that baseline without consideration for how individual context — both overall and at the time of a specific assessment — may affect this process. As in the case of many of the measures that are incorporated into digital algorithms, the SOFA is a tool that is designed to describe populations, but is nonetheless used to make decisions regarding individuals, with little regard for how people with disabilities, who may be very different from most of the population, might suffer as a result. The fact that these issues are such a prominent feature within CSC plans is a reminder that the problems algorithms can create for disabled people are not inherently a problem with the algorithm, or with technology more broadly. Rather, they are a reflection of how prediction, as an outgrowth of statistics, often fails to react to disability without attempting to either normalize or erase it.

Analysis of the contents of CSC plans in place during the early stages of the COVID-19 pandemic demonstrates how reliance on algorithmic tools, even when they are not digital in nature, can facilitate the spread of ableism and assumptions about whose life is worth living. The same plans that disability rights organizations and individual activists called out for conflating scores on particular health measures with an individual’s likelihood of benefitting from mechanical ventilation or assuming disability equates lower quality of life were used as examples by jurisdictions looking to assemble their own crisis standards. As states rushed to ensure that they had plans in place, several resorted to copying whole sections of documents developed and enacted in other parts of the U.S. — in some cases, without even removing the original state’s name from the duplicated materials — and consequently, furthered the reach of discriminatory allocation methods even as advocacy groups filed legal challenges against them in the states where they had originated. Similarly, even as algorithmic tools face challenges on both ethical and legal grounds, they continue to proliferate. Often, these challenges seem to simply encourage organizations that have adopted algorithmic decision-making tools to do so quietly, avoiding public scrutiny, just as some states reacted to the outcry over CSC plans by restricting public access to their own documents or taking them off their Web sites entirely.

Although the COVID-19 pandemic brought greater attention to ableism within the healthcare system, medical rationing is not unique to the pandemic, and decisions about who to treat when there are not enough healthcare resources for everyone often involve implicit or explicit ableism (Andrews, et al., 2021; Largent, 2020). A related issue is the so-called “disability paradox,” where people with significant disabilities nonetheless report high quality of life (Albrecht and Devlieger, 1999; Drum, et al., 2008; Horner-Johnson, 2021). In instances where self-rated health and other health measures, such as functional status or physician assessment, do not agree, it is often the self-report that is called into question. For example, in a study of older adults, those whose self-rated health exceeded their presumed health based on other measures were more likely to still be alive at eight- and 12-year follow-ups (Chipperfield, 1993). Nonetheless, participants’ assessments of their own health were described by researchers as “overestimations,” even though the results indicated that other measures had underestimated their health. The underlying assumption that doctors or other experts are more equipped to make assessments of disabled people’s experiences than disabled people themselves, and that these assessments are more accurate and
Encoding normative ethics: On algorithmic bias and disability

objective, motivates choices about what data an algorithm uses to make predictions. Because these choices reflect determinations of who is capable of producing valid knowledge, they are part of the way that societal ableism is integrated into algorithmic tools. As CSC plans demonstrate, the risks disabled people face from algorithmic tools are not unique to situations where digital algorithms are in use; rather, increasing reliance on digital algorithms simultaneously heightens and obscures harms that are already routinely encountered by people with disabilities.

Categorization and normalizing intervention: Lessons from the autistic self-advocacy movement

The history of autism as a distinct diagnostic category demonstrates another issue for disabled people that pre-dates the adoption of algorithmic tools: the use of classification to create new targets for normalizing intervention. While digital classification algorithms are a relatively recent development, the use of classification to create sub-groups of individuals, both within society in general and among disabled people specifically, is a longstanding practice. Classification algorithms are not creating this practice, but they are amplifying it and removing some elements of human oversight when, for example, they replace a round of human review or allow fewer people to review a greater number of applications or cases. As algorithmic tools are implemented to categorize people into specific groups — for example, people who need higher or lower levels of in-home services, or those who will be interviewed for a job and those who will not — disparities that are reflected in past data, such as those based on race, class, and ability, become the basis for future classification. Once people are classified in a certain way, they often become subject to specific interventions; individuals who are categorized as having “high” need for services may be the target of attempts to reduce that need, for instance, with little consideration of whether such efforts are justified or appropriate. While anyone who is classified as being outside “normal” standards may be affected by this process, disabled people, whose default ways of being and very existence are often by definition atypical, are particularly at risk of normalizing interventions.

Autism only began to emerge as a diagnostic category in the early twentieth century, and the specific diagnostic criteria, categorization, and meaning of autism have shifted over time, most recently with the significant revision to the *Diagnostic and statistical manual of mental disorders* (DSM-5, fifth edition) (Botha, 2021; Kapp and Ne’eman, 2020). The creation of a category of autistic children, who were predominantly white, male, and middle- or upper-class, separated them from existing classifications of disability, which often carried eugenic overtones and conveyed fears about racialized “unfitness” (Gibson and Douglas, 2018; Silberman, 2016). Autistic behaviors were framed as the deviant results of a specific condition, rather than expected inferiority and generalized disability (Gibson and Douglas, 2018). Multiple studies have identified ongoing disparities in autism diagnosis related to race, gender, and other demographic factors, and have found that compared to white children, children of color tend to receive autism diagnoses at a later age, experience higher rates of more stigmatized diagnoses prior to receiving an autism diagnosis, and require more visits to doctors and specialists to receive a diagnosis (Becerra, et al., 2014; Begeer, et al., 2009; Burkett, et al., 2015; Dababnah, et al., 2018; Evans, et al., 2019; Goin-Kochel, et al., 2006; Mandell, et al., 2009, 2007, 2002; Obeid, et al., 2021; Petrou, et al., 2018; Thomas, et al., 2012). There has been significant interest in creating algorithmic screening and diagnostic methods that are less costly in terms of time and clinical resources. The persistent bias in diagnosis means that attempts to develop algorithms based on the currently identified population of autistic people will reflect the societal and contextual conditions that contribute to diagnostic disparities, and potentially perpetuate them.

As researchers and clinicians proposed treatments for autism in the 1960s, they stressed the need to correct and normalize autistic traits. In particular, an approach based on operant conditioning which became known as Applied Behavioral Analysis, or ABA, focused on rendering autistic children “indistinguishable” from their typically developing peers (Gibson and Douglas, 2018; Kirkham, 2017; Roscigno, 2019; Silberman, 2016). ABA remains the most commonly used intervention for autism despite criticism on both ethical and methodological grounds (Bottema-Beutel and Crowley, 2021; Gruson-Wood, 2016; Kirkham, 2017; Roscigno, 2019; Wilkenfeld and McCarthy, 2020). Despite a growing body of evidence [6] suggesting that many of the targets of ABA, such as repetitive movements and other self-stimulatory (or “stimming”) behaviors or communication challenges, are the product of neutral differences or a mismatch between
autistic and non-autistic ways of being, much of the work in autism research and practice takes a deficit-based approach focused on intervening on and normalizing autistic people.

Normalizing approaches to autism intersect with technological development through conceptual and rhetorical framings and persistent interest and investment in technology as a means of intervention to reduce or eliminate particular behaviors in autistic people. Scientific and mainstream depictions of autistic people as robotic, emotionless, and unable to meaningfully contribute to the dialogue around autistic experience are pervasive, and the reliance on the metaphor of autistic people as machines is used to justify both continued investment in normalizing technology and outright dehumanization and oppression (Botha, 2021; Keyes, 2020; Williams, 2021). The denial of fundamental humanity all too often begets outright physical violence, which itself creates an opportunity for development of normalizing technology. For example, the Judge Rotenberg Center, in Massachusetts, is a residential school and treatment center that primarily serves people with developmental, intellectual, and psychiatric disabilities, receives government funding, and uses a custom-designed technological device to deliver electric shocks as punishment to disabled residents (Neumeier and Brown, 2020; Roscigno, 2019). Despite the condemnation of the United Nations and a longstanding campaign from autistic self-advocates, and from disability rights advocates more broadly, to shut down the program, as of this writing, the Judge Rotenberg Center remains open and continues to use electric shocks (Lopez, 2021; Neumeier and Brown, 2020). In a demonstration of the role of intersectionality in shaping disabled people’s experiences, the school-age population at the Judge Rotenberg Center is overwhelmingly comprised of students of color; during the 2015–2016 school year, over 80 percent of residents were Black or Latinx (Neumeier and Brown, 2020). Thus, while classification can lead to normalizing intervention for disabled people generally, other aspects of identity, such as race or gender, can lead to stark differences in the shape of that intervention — for example, whether someone is described as “needing support” and provided with occupational therapy or labeled “noncompliant” and subjected to physical punishments.

The attitudes that condone and facilitate the use of physical violence as a means of controlling and normalizing autistic people are also part of the methods used to define and measure behaviors and outcomes. Theoretical frameworks that treat autism as the result of a core deficit in theory-of-mind, for example, lead to a circular process where theory-of-mind tasks are included in autism research because autism involves issues with theory-of-mind (Astle and Fletcher-Watson, 2020). The reliance on deficit focused measures confirms preconceived ideas that autism is defined by deficits, and precludes conversations that instead position autism as a difference that can convey both challenges and benefits (Botha, 2021). Similarly, outcome measures frequently emphasize the reduction of behaviors specified in diagnostic criteria, regardless of whether this represents a meaningful or desirable outcome from the perspective of autistic people, often with the effect of promoting passing and normalization at the expense of actual well-being (Ne’eman, 2021). Whether these measures are used to set research agendas, determine interventions, or develop algorithmic tools, the effect is the same: to abstract subjective views that position autism as a deficient form of humanity into something that is treated as an objective reflection of reality.

As algorithmic tools are applied to autism-related topics and situations, the values entangled in these measures are encoded in technology. As an example, consider apps and other technological tools which endeavor to teach autistic people to make “normal” eye contact, despite evidence that many autistic people find it distracting, overwhelming, or even painful, and regardless of the fact that improving eye contact is rarely a goal that autistic people set for themselves (Keyes, 2020; Ne’eman, 2021). Similarly, a recent review of wearable technology developed for autism intervention identified an emphasis on normalizing autistic people, with nearly half of the surveyed papers describing technological interventions for social skills training or facial emotion recognition, despite evidence that when autistic people encounter challenges in these areas it is not the result of a fundamental deficit located within the autistic person, but of a mismatch between autistic and non-autistic communication and behavior (Williams and Gilbert, 2020b). Similarly, assistive technology intended to support autistic people in social situations focuses on shaping and “correcting” the behavior of autistic people, rather than on teaching neurotypical people to be more accepting or inclusive (Keyes, 2020).
Another set of algorithms that purports to identify autism based on gaze tracking, gait, auditory cues, or other data risks reinforcing existing diagnostic biases, particularly around race, gender, and class, by basing identification of autism on people who have already obtained an autism diagnosis (Bennett and Keyes, 2020; Keyes, 2020). The emphasis these technologies place on early diagnosis can itself be a problem. While it is important that autistic people, like anyone else, have access to the supports they need, early diagnosis is often situated as the first step in an intervention that sees passing as neurotypical as the end goal, and views autism, at best, as an undesirable condition to be overcome (Keyes, 2020; Ne’eman, 2021). Once the systems that marginalize autistic people are incorporated into algorithmic technology, no one has to explicitly say or believe that autistic people are deviant or abnormal, or that their autism is a problem to be solved; the proliferation of algorithms that aim to detect autism, accompanied by claims that early detection will lead to more successful integration into neurotypical society, conveys the message.

Treating autistic people and autism as something abnormal that must be minimized predicates inclusion on normalization and the ability to pass as neurotypical, and fails to recognize the role of ableism in shaping (or precluding) autistic people’s inclusion in society. This kind of abstraction, whereby autism is separated from the context in which it was defined and continues to exist, is echoed in algorithms that turn complex social issues into discrete, more easily solvable problems. It is also reflective of the way that disability often is understood as a problem within an individual, rather than part of how that individual exists relative to their society. For disabled people, algorithmic tools not only remove the problems they are deployed to solve from their context; they also rely on an understanding of disability that is removed from societal forces. In order to create representations of the world that a computer can use, algorithmic tools rely on measurement and statistical tools that have perpetuated ableist views on the basis of an imposed standard of normality. While the abstractions in algorithmic tools create risks for many different marginalized groups, people with disabilities face compounded harms because of the intertwining of algorithms, statistics, eugenics, and collective understandings of what it means to be disabled.

IV. Proposed remedies

The adoption of algorithmic decision-making tools, and the coinciding risk of algorithmic bias, is a significant threat to the wellbeing and inclusion of disabled people in society, but the negative consequences outlined in this paper are not inevitabilities. Strategies from other areas of research can serve as possible means of responding to and preventing algorithmic bias towards disabled people. In particular, participatory methods, the use of inclusive design, and efforts to improve disability data must be part of an overarching effort to ensure that disabled people are not further marginalized by algorithmic tools.

One way to reduce the risk of algorithmic bias towards disabled people is to incorporate participatory methods across all stages of the design and development process. In participatory methods, members of a selected group take part in a process as experts and co-creators who share power with others in guiding and executing the work (Trewin, et al., 2019). An increasing number of research teams have incorporated participatory methods into autism research, and several have authored guidelines based on their experiences with these approaches, which designers and developers seeking to incorporate participatory methods into the algorithmic creation process might draw on (Benevides, et al., 2018; Cascio, et al., 2021, 2020; Nicolaidis, et al., 2019, 2011). At the level of data collection, the addition of participatory elements may look like establishing a disability advisory board that co-creates standards and guidelines about what user data to retain and use in developing and testing algorithmic tools. Participatory approaches to algorithmic creation might include collaboration with disabled people to determine areas where an algorithmic approach to problem-solving will be most beneficial, and least harmful; working with disability-led organizations to assess the needs of broader subgroups within the disability community; or even partnering with disabled people to develop better ways to identify disability within existing datasets.

Using participatory approaches can help ensure that disabled people’s interests and needs are prioritized,
and can be part of establishing an ongoing relationship with the disability community (Fisher and Robinson, 2010). Authentic engagement with disabled people as co-developers of algorithmic systems can also be a step towards establishing cultural humility about disability, whereby non-disabled people recognize the limits of their understanding of disability and become more able to ameliorate power imbalances (Yeager and Bauer-Wu, 2013). Participatory methods also offer an alternative to design activities that seek to incorporate disability perspectives without actually involving disabled people, such as disability simulation exercises or simplistic user personas, which often reinforce negative stereotypes of disabled people and generally fail to convey the degree to which structural forces contribute to the experience of disability (Bennett and Rosner, 2019; Costanza-Chock, 2020).

As much as participatory methods can help to increase equity and mitigate or prevent the impact of algorithmic bias on disabled people, they are not a panacea. A key aspect of participatory research is thinking about who gets to participate: who is invited, whose needs are accommodated, and whose contributions are valued and included. Disagreements about who best represents the autistic community, for example, are often framed as a choice between parents or autistic adults, but these kinds of binaries elide the differences that exist within the disability community, not just between disabled and non-disabled people, and obscure the ways that power is unevenly distributed among people with disabilities. Disabled people who experience multiple forms of oppression and marginalization based on the intersections of their identity — including, for example, disabled people of color, disabled women or disabled non-binary people, and disabled people of lower socioeconomic status — face additional barriers to entry into these fields. Similarly, different kinds of disability are understood and stigmatized in different ways, and cross-disability ableism remains an issue in many disability spaces.

Ideally, participatory methods are transformative, and redistribute power in ways that reach beyond the limits of a particular project or research team (Williams and Gilbert, 2020a). However, granting power to those who do not ordinarily have it typically requires some relinquishment on the part of those who do. Such transitions are not always easy, particularly when they simultaneously require that lived experience be valued equally to academic expertise. This tension also hints at the distinction between a process that is empowering and one that is truly emancipatory. It is one thing to empower people; to give them power within a society that continues to systematically marginalize and disadvantage them. It is another to strive for emancipation: to do work that seeks to reimagine the world and free people from oppression.

Finally, participatory methods cannot simply be layered on top of work that otherwise maintains the tradition of responding to disability with efforts at normalization. Researchers, designers, and developers need to work to address their own ableism — and the ableism within their own fields and industries — first, before expecting disabled people to join them as collaborators. This includes considering the impact of asking disabled people to participate in work in disciplines and areas that remain fundamentally ableist. Valuing collaborators from the disability community means compensating them for their time and effort, as with anyone, but it also comes with a responsibility to ensure that they are treated with respect and dignity, and not expected to educate or enlighten others who have not yet taken the time or made the effort to increase their own awareness of ableism and its impact.

In addition to expanding use of participatory methods, approaches to algorithmic development should place greater emphasis on the role of design as a means of limiting or mitigating the risk of bias. As with participatory methods, there is potential for design to serve a performative role, rather than transformative role, and thus its use as a strategy for preventing or responding to algorithmic bias should not be a substitute for more foundational work that addresses ableism and other forms of systemic discrimination. There has been significant work examining the role of design as a means of expressing and maintaining particular values, as well as highlighting the use of the design process to respond to and reduce structural inequities (Costanza-Chock, 2020; Friedman, 1996; Winner, 2003). Within the context of algorithms, specifically, one option is to consider human-in-the-loop systems where a human operator oversees the algorithmic system and steps in as needed; in some cases, systems are designed with the explicit intention of having people and algorithms work in tandem, allowing the system as a whole to benefit from their respective strengths. A potential expansion on human-in-the-loop design is the use of society-in-the-loop algorithms, which aim to
Embed societal values into algorithmic governance (Rahwan, 2017). While the degree of ableism in many societal norms raises concerns about the precise values assumed in society-in-the-loop design, the concept of holding algorithms accountable to a broader population than just the people who develop and implement them holds promise as a means of mitigating bias.

Enacting formal audit and appeals processes is another way of creating algorithmic accountability. One possibility is to create a licensing system for algorithms deployed in critical or high-stakes settings like healthcare (Citron and Pasquale, 2014). Another is to develop independent auditing committees that can assess an algorithm's performance and propensity to create disparate impact or amplify bias towards a specific group. Operation of such committees is predicated upon transparency in algorithmic design, which is in itself a means of oversight. A related, and potentially more impactful solution, is to hold the companies that develop and sell algorithms accountable for equal protection violations under the state action doctrine; when algorithms perform functions for which the state has traditionally held exclusive responsibility or act in tandem with the government, proponents argue, they are operating as state actors and should be regulated accordingly (Crawford and Schultz, 2019).

Where disability is concerned, improving the data used to design, train, and test algorithmic tools is also an important consideration for reducing the risk of algorithmic bias. Efforts to increase and improve disability data collection have been a long-term focus among disability advocates, who tend to stress the importance of incorporating some kind of disability information alongside other demographic data. These efforts gained additional momentum during the COVID-19 pandemic, when a lack of disability data made identifying disparities in pandemic-related outcomes challenging, and in some cases, impossible (Reed, et al., 2020; Swenor, 2022). Work on this front must be done alongside and in partnership with the many disability activists and organizations that have been invested in it for decades. Efforts to increase and improve disability data must also consider the differences between understandings of disability as an outcome, an experience, or an identity, in order to recognize which conceptualizations matter, in what circumstances, and in what ways measurements may or may not capture them.

Finally, there is a need for selectivity about when and how algorithms are adopted and implemented, and realism about what technology can and cannot do. Not every problem is best solved with an algorithm, and algorithmic decision-making tools can just as easily obscure as illuminate the process through which predictions are made. Similarly, as tempting as it may be to view the data used to develop, train, and test algorithmic tools, or the eventual outputs of these systems, as somehow more objective than human judgments, the components of algorithmic tools are derived from human choices. People create data by deciding what to measure and how to measure it, and they build algorithms on a foundation of assumptions about what outcomes matter enough to predict and which categories are meaningful for classification. Ultimately, algorithms can only encode and enact the values people provide them with. As flawed as human decision-making can be, it is not inherently worse than algorithmic judgment; as accurate as algorithms might seem, they are not necessarily fairer than people. When the underlying issue is societal ableism, no amount of technological improvement is a substitute for work that effects social change.

V. Conclusion

The threat of algorithmic bias to people with disabilities is inseparable from bigger questions of how society defines shared values, quantifies ethics, and strives to operationalize fairness, and the overarching role of technology as a normalizing agent. As examples from the COVID-19 pandemic and autism research demonstrate, the risk of algorithmic bias to disabled people is not ultimately a technological issue. Rather, it is part of a more fundamental problem in the way disability is understood and targeted for intervention and normalization within society.

In each of the proposed methods to mitigate algorithmic bias, it is essential to consider whose ethics are
Encoding normative ethics: On algorithmic bias and disability

Many existing measures, approaches, and attitudes reflect the priorities and agendas of nondisabled people, rather than the goals of disabled people for themselves. Building on the principles of participatory approaches, one step in mitigating the risk of algorithmic bias towards disabled people is explicit consideration of not just what ethics are encoded in technological systems, but whose values they represent. Furthermore, there is a need for notions of algorithmic fairness that move beyond equality of treatment and strive to encompass equity and justice. Above all, remediying existing inequities, and preventing their algorithmic amplification, requires centering disabled people not just as subjects or participants, but as experts and leaders, and interrogating the way ableism shapes choices about what to measure, how to measure it, and the way that measurement is ultimately interpreted.

About the author

Ian Moura is a research assistant at the Lurie Institute for Disability Policy and a doctoral student in Social Policy at The Heller School for Social Policy and Management at Brandeis University. His research interests include services and outcomes for autistic adults, disability policy, data and measurement, and algorithmic bias. In addition to his studies, Ian is a Community Project Lead with the Academic Autism Spectrum Partnership in Research and Education (AASPIRE). He has also worked with Disability Rights Education & Defense Fund (DREDF), including leading a research team in collecting different states’ Crisis Standards of Care in light of the COVID-19 pandemic. Ian’s interest in disability grew out of his experiences as an autistic person.

E-mail: ianmoura [at] brandeis [dot] edu

Notes

1. This paper deals with both algorithms in a broad sense and with computational algorithms more specifically, including those commonly used within artificial intelligence, automated systems, and related technologies. When the distinction between the two is important, the term “digital algorithms” is used to indicate the latter, more specific meaning.

2. Opinions regarding person-first (e.g., “person with a disability”) versus identity-first (e.g., “disabled person”) language vary among both individuals and specific sub-groups within the broader disability community. This paper uses “disabled people” and “people with disabilities” interchangeably. When referring to a specific subgroup, such as autistic people or people with intellectual disabilities, the language generally preferred by that community is used.

3. The use of “scare quotes” around the phrase “make decisions” is intended as a reminder to the reader that although they are increasingly part of a variety of decision-making processes, algorithms themselves do not make choices in the same way as human beings; rather, they perform specified tasks, calculations, or categorizations, with the resulting output used by humans who decide whether and how to act in accordance with algorithmic output. Even in more highly automated systems, it is still human designers and engineers who decide when and how the results of various algorithms should be used and acted upon.

4. “Analog” is used here in the sense of tools that are not digital or computerized in nature. While describing the kinds of algorithms that pre-date and underpin the recent wave of digital algorithms as “analog” is not accurate in the strictest sense (e.g., some checklists may be accessed via a computer interface), the intent is to call attention to the long history of reliance on paper-based or pre-digital artifacts used for systematic scoring, prediction, categorization, and classification of individuals and potential outcomes.

5. The description and commentary on Crisis Standards of Care plans in this section builds on the Disability
Rights Education and Defense Fund’s work collection and analyzing these documents during the initial months of the COVID-19 pandemic in the United States (see Moura, et al. [2020] for more information on this project). DREDF’s documentation of the contents of Crisis Standards of Care served as the source for much of the information in this paper regarding what situations and information plans did and did not initially include or consider, particularly when those aspects had not yet been discussed in published literature.


7. For a hypothetical example of this, see Colusso, et al., 2019.

References


Encoding normative ethics: On algorithmic bias and disability


P. Kirkham, 2017. “‘The line between intervention and abuse’ — Autism and applied behaviour analysis,”
Encoding normative ethics: On algorithmic bias and disability


Encoding normative ethics: On algorithmic bias and disability


Encoding normative ethics: On algorithmic bias and disability

number 3, pp. 68–72.

*Applied Nursing Research*, volume 26, number 4, pp. 251–256.


---

**Editorial history**

Received 16 November 2022; accepted 12 December 2022.

---

This paper is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).

Encoding normative ethics: On algorithmic bias and disability
by Ian Moura.
*First Monday*, Volume 28, Number 1 — 2 January 2023
https://firstmonday.org/article/view/12905/10761
doi: [https://dx.doi.org/10.5210/fm.v28i1.12905](https://dx.doi.org/10.5210/fm.v28i1.12905)