Imagining Fair(er) Datasets for GenAI: Lessons from the Arts

Theresa Krampe[1]

Abstract

As generative artificial intelligence (GenAI) is rapidly inserting itself into many domains of everyday life, there is also a growing awareness of its ethical implications. Several systems, among them chatbots and image generators, have been shown to reiterate gendered, racial, or ableist stereotypes and to contribute to the erasure of marginalised voices and perspectives. In machine learning and AI ethics, concepts such as fairness and algorithmic bias have become instrumental in recognising and mitigating these issues. The task of addressing covert, structural, and multilayered forms of discrimination, however, remains challenging. In this chapter, I argue that the arts and culture, as domains that tend to be overlooked in mainstream discussions around GenAI, offer valuable inspiration for envisaging more diverse and inclusive datasets for fair(er) AI systems. With the help of two case studies—*The Zizi Project* by Jake Elwes and *Not the Only One* by Stephanie Dinkins—I show how AI artworks can draw attention to the risks of GenAI to unfairly discriminate against so-called vulnerable groups, challenge the values and assumptions underlying hegemonic visions of technology, and draft alternative AI futures.

1. Introduction: Imagining Fair(er) AI

Machine learning (ML) technologies are rapidly insinuating themselves into everyday life. The use of advanced search engines, digital assistants, or even automated decision-making software has already become quite natural to many of us. With the emergence of high functioning and highly visible Large Language Models (LLMs) such as GPT as well as image generators such as Midjourney or DALL-E, the same is increasingly true for generative artificial intelligence (GenAI). Over the past few years, the ability of

1 https://orcid.org/0009-0001-9416-4676

these systems to create an astonishing variety of texts, images, or sounds reinvigorated discussions around the potential of AI for innovation and creativity (see Cooper 2020; Eapen et al. 2023; Marr 2023), but also raised concerns about privacy (see Hagendorff 2019), authorship and copyright (see El Atillah 2023), misinformation (see Hsu and Thompson 2023), and job losses in the creative industries and other areas that formerly seemed automation-proof (see Ellingrud et al. 2023; Fleming 2024; Verma and De Vynck 2023). No less importantly, numerous studies have indicated that AI systems can be, and often are, subject to systematic biases that unjustly discriminate against groups of people on the grounds of race, class, gender, disability, or other protected attributes (see Akter et al. 2021; Barocas et al. 2023: 19–20; Hagendorff 2019; Heesen et al. 2021; Mehrabi et al. 2021a). Even though chatbots and image generators seldom act as immediate gatekeepers in such high-stakes scenarios as criminal punishment, medical diagnoses, or granting loans (see Angwin et al. 2016; Barocas and Selbst 2016), they can nevertheless discriminate against social groups if they perpetuate stereotypes, privilege certain types of knowledge, create disrespectful or demeaning output, and reinforce unfair regimes of in/visibility and marginalisation along the lines of identity (see Barocas et al. 2023; Gautam et al. 2024; Loh 2024; Mehrabi et al. 2021a). Words and images, after all, shape the way we make sense of the world and our place within it.

This chapter examines the discourse around fairness and/as discriminatory bias in AI ethics, as well as how these understandings are currently being renegotiated in the arts and culture. Artworks, activist interventions, and community-led projects have not yet received sufficient scholarly recognition in the field of ML. As a result, their potential to interrogate the values and assumptions underlying ML and their unique capacity to imagine fair(er) datasets and fair(er) uses of GenAI has remained largely untapped. By considering artworks in the context of AI ethics, I thus seek to shift the discussion around fairness and bias to forms of intervention that appropriate GenAI in unconventional, creative, and often subversive ways. Art, I argue, can draw attention to the risks of AI to unjustly discriminate against so-called vulnerable groups, challenge the values and assumptions underlying hegemonic visions of AI, and draft alternative AI futures, thereby offering valuable prompts for creators, users, and critics of GenAI.

The main part of this chapter is divided into two sections. Following immediately after this introduction, Section 2 offers an in-depth discussion of discriminatory bias in GenAI from an interdisciplinary ethical perspective, focussing on the idea(l) of fairness as freedom from discriminatory

bias as an important measure of training data quality, while also examining how these biases may impact social groups and power structures. Section 3 then shifts the perspective to interventions and potential solutions as they are envisioned in contemporary art. To do so, I discuss two case studies: *Zizi—Queering the Dataset* (2019) from the *Zizi Project* by London-based conceptual artist and researcher Jake Elwes and *Not the Only One* (2018–ongoing) by the US-American transdisciplinary artist Stephanie Dinkins. Both works use GenAI trained on diversified and carefully curated datasets as part of an artistic practice that challenges hegemonic formations of gendered and racialised identity. To analyse them, I integrate AI ethics with queer/feminist, intersectional, and media-analytical approaches.

2. Bias and Discrimination in GenAI

2.1 Fairness as Freedom from Discriminatory Bias

Fairness, in a general sense, can be defined as the impartial and just treatment of people. In Western societies, this is usually associated with the idea of equal chances for self-advancement, for example, that a person's position in society should be proportionate to their contribution, rather than to factors they have no control over (see Feuerriegel et al. 2020: 380–381). It is thus important to note that the un/fairness of AI potentially relates to a much wider range of pressing topical issues than what is addressed in this chapter, including AI's involvement in neo-colonial relations of subjugation and extractivism; practices of datafication and (racialised) surveillance; the exploitation of natural resources; or precarious labour in the Global South, among others (see Tacheva and Ramasubramanian 2023). In AI ethics, however, fairness has become strongly associated with the notion of discriminatory bias, understood as systematic distortions in an AI system that result in the unfair differential treatment of people based on socially relevant groups, usually on the grounds of protected attributes such as race, class, gender, ability, or sexual orientation (see Akter et al. 2021; Barocas et al. 2023; Hagendorff 2019; Kim 2022; Leavy et al. 2020; Mehrabi et al. 2021a). This focus on bias, and on datasets as sources of bias but also, crucially, points of intervention, is also echoed in the case studies analysed in this chapter, and hence I will mainly limit my discussion to fairness in this latter sense.

That ML can entail discriminatory biases first came to widespread attention around ten years ago, in a series of highly publicised incidents.

When Google's photo app assigned the label "gorilla" to People of Colour (PoCs) in 2015, this not only revealed inaccuracies in the algorithm, but also invoked rather nasty racist discourses. A few years later, an influential study by Joy Buolamwini and Timnit Gebru (2018) found biases regarding skin colour and gender in commercial classification software, which reinforces highly problematic forms of intersectional oppression and prolongs a history of marginalisation and disenfranchisement of Black women. Other well-known examples include translation software associating certain professions with gendered stereotypes (see Prates et al. 2020) or reiterating societal prejudices towards sensitive attributes such as gender, race, and sexual orientation (see Lin et al. 2023).

In academia, the occurrence of discriminatory bias in AI has motivated the development of an entire field dedicated to "uncovering and rectifying [...] biases in statistical and machine learning models" (Mitchell et al. 2021: 142).[2] In the corresponding publications, fairness is typically defined *ex negativo* and with reference to anti-discrimination: A model is fair if it does not show "prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics" (Mehrabi et al. 2021a: 1; see also Sun et al. 2025).[3] Scholars have furthermore pointed out that AI can discriminate against people even if it treats all groups equally, for instance if it employs variables that are correlated with group membership such as height for gender, language for nationality, and zip code for race or class (see Mehrabi et al. 2021a; Sun et al. 2025). This is why "fairness through unawareness"—the somewhat naïve view that discrimination can be avoided by excluding sensitive attributes from the ML process – is rightly deemed unsuitable for addressing discriminatory bias in most cases. Not only can removing information about race or gender diminish the model's performance,[4] but

2 This is also evidenced by the fact that fairness has become a mainstay for international conferences such as FAcct, ICML, or AAAI (Wang et al. 2022). Recurring labels applied to the research field include fair AI, fair ML, or algorithmic fairness. To my knowledge, there are no established definitions of or clear delimitations between these labels, so I will use them roughly synonymously in this chapter.

3 Note that discrimination can be justified under certain circumstances, for instance, if it is unavoidable or if the information is relevant for the decision (e.g., car insurance rates that vary with the age of the holder; see Barocas et al. 2023). Another interesting example is algorithmic affirmative action, i.e., the idea that algorithms could be trained in such a way as to counterbalance structural disadvantages for marginalised groups (see Kim 2021 for a critical legal perspective; Segev 2025 for an ethical one).

4 Though, by now, there are several technical fixes for this particular problem, as summarised by Hagendorff: "Advanced machine learning methods can learn from small

it is also ineffective as these attributes tend to re-enter the model by proxy. In the worst case, such colour or gender-blind approaches even exacerbate the problem because they make the model's biases harder to detect (see Brandner and Hirsbrunner 2023: 27; Feuerriegel et al. 2020: 379; see also Kim 2022; Mehrabi et al. 2021a).

2.2 The Role of Training Data Quality

A crucial factor for the formation (and hence the prevention) of bias in ML is the quality of the data used to train the model—though it is worth noting that training data quality is not the only source of bias.[5] More often than not, AI systems learn prejudices prevalent in society from the training dataset, which typically consist of large amounts of texts or images scraped from the internet. Pre-existing human prejudices can become encoded into AI applications because certain groups are over or underrepresented in the dataset, or because the training data replicates stereotypes. Discriminatory aspects can furthermore be introduced via the annotation of data by service providers, or through user feedback (see Barocas et al. 2023: 248–252; Brandner and Hirsbrunner 2023: 26; Feuerriegel et al. 2020: 381; Hagendorff and Wezel 2020: 358). As ML processes rely on generalisations and pattern recognition, such stereotypes are not only adopted but are in fact reinforced by GenAI. If a system defaults to male pronouns for "lawyers" or "doctors," for instance, this further increases the frequency with which these professions are associated with men in the dataset. Considering the importance of data to make or break this vicious circle, it is not coincidental that the art projects analysed below engage specifically with questions of training data quality, pointing to gaps in the dataset that lead to discriminatory output, but also imagining ways of using data in such a

datasets via data augmentation, can generate synthetic data via GANs or variational autoencoders to artificially increase the amount of training stimuli, use transfer learning to use knowledge from an already learned task, utilize few shot learning mechanisms, etc." (2021: 567).

5 In a comprehensive survey, Mehrabi and colleagues (2021a) identify multiple sources of bias along the entire ML process, from data collection via the algorithm to the interaction with the users, paying attention to how, among other things, a company's hiring practices, the design of the algorithm, or the presentation of information via the graphical user interface may influence a model's biases (see also Chowdhury and Mulani 2018; Zou and Schiebinger 2018).

way as to promote epistemic justice or, in any case, interrogate established epistemic hierarchies.

While in-depth scholarly engagements with training data quality and discrimination are still comparatively rare, by now, there is some agreement on criteria that are deemed conducive to increasing the quality of training data.[6] Some of these seem particularly relevant for reducing the risk of discriminatory outputs. The completeness of the dataset, as a case in point, touches upon matters of exclusion, as when certain groups are absent from the dataset due to persistent social dynamics of marginalisation. Inclusion in the dataset can be a double-edged sword, however. On the one hand, some AI applications show significantly poorer performance for some groups of people because the available training data is incomplete or insufficient. On the other hand, collecting additional data may put a burden on the groups in question if they are being "overresearched" or if their increased visibility also increases their vulnerability (see Benjamin 2019; Eubanks 2018). The diversity of the dataset, too, seems worth considering. If diversity is taken to mean that the dataset contains at least one type of each entity present in the overall group, this would ensure the representation of minority groups even though they only comprise a very small percentage of the relevant population (see Mohammed et al. 2023). However, this also means that there may be trade-offs between diversity and representativeness if the latter is taken to mean that the dataset should accurately reflect the population represented. What is more, diversity comes with its own set of challenges as it entails tricky questions of social categorisation. While it may be desirable for a dataset to distinguish between multiple subgroups, for example, to promote intersectional fairness (see below), the same strategies are also prone to pigeonholing and masking the contingency and constructedness of social groups as such.

As even such a selective discussion shows, there are several different dimensions of training data quality that significantly impact bias and dis-

6 Meta studies (e.g., Hagendorff 2019; 2021) as well as data quality tools such as the data quality glossary (Mohammed et al. 2023; see also Brandner et al. 2023) offer useful overviews of relevant criteria. Recurring criteria include correctness (does the dataset contain errors?); transparency (what information is available about the data, e.g., its origin, purpose, or the quality assurance measures employed?); timeliness (is the data up to date?); relevance (is the data and metadata relevant to the purpose?) as well as more complex criteria such as explainability and credibility. From a normative perspective, Hagendorff furthermore explores the idea of "'good' behavioral datasets" (2021: 564), i.e., training data that is chosen from a subset of the population whose behaviour is deemed both competent and morally sound.

crimination. Not all of these can necessarily be satisfied in equal measure, making it impossible to create a one-fits-all solution. What is more, most approaches to fairness in ML do not (or not sufficiently) account for the complexity of the socio-cultural contexts from which AI systems emerge and within which they operate. The underrepresentation or stereotyping of specific groups in the training data is often the result of complex structural forms of marginalisation and historical injustice, not all of which are easy to detect, let alone remedy (see Costanza-Chock 2020). Therefore, most fairness tools are ill-prepared to address the root causes of discrimination or to consider the indirect and long-term effects of predictions and decisions (see Thomsen 2024). As Anna Lauren Hoffmann puts it, there is no "easy fix" (Hoffmann 2019: 910) for structural discrimination that can simply be applied at the level of code. Quite the opposite, fairness tools could actually detract from the need to interrogate social power structures (see Leavy et al. 2020: n.pag.). What is more, if we take seriously the idea of fairness as equal opportunity, we may also want to think about correcting for systemic structural disadvantages and historical injustices (see Barocas et al. 2023; Binns 2018; Crawford 2017). In the field of GenAI, this could mean finding ways of reinserting the voices of women, PoCs, and those who have been systematically erased from the archive into the dataset, even at the expense of values like representativeness or accuracy.

In closing this section, it is worth stressing once more that data quality and the absence of bias are not the only understandings of fairness that are potentially relevant to the question of discrimination. Fairness may also concern matters of legitimacy (see Mitchell et al. 2021: 143). Indeed, the question of whether or not, in a given situation, it is justified to use ML in the first place ought to precede concerns about algorithmic bias (see Barocas et al. 2023: 23–24). Writing from an intersectional queer/feminist perspective, Katrin Köppert maintains that even methods like fair and explainable AI often fail question the assumptions behind the AI imaginaries they advance. In other words: Making AI fair(er), more transparent, more trustworthy, etc. does not automatically ensure a project's usefulness or ethical merit. This is especially true when considering AI "in the bigger picture of the climate crisis, extractivism, and machismo" (Köppert 2024: n.pag.); or in its complicity with those hegemonic power relations and mechanisms that Jasmina Tacheva and Srividya Ramasubramanian (2023) aptly subsume under "AI Empire."

2.3 Setting the Terms of In/Exclusion

Understanding the moral stakes of algorithmic bias requires us to further unpack how GenAI is implicated in harmful and unjust regimes of in/visibility and in/exclusion. From an ethical perspective, the notion of representational harms (see Crawford 2017) seems helpful when it comes to articulating when and why biases in GenAI are harmful. In contrast to allocative harms, which concern the distribution of resources and opportunities (see Barocas et al. 2023: 19–20), representational harms can be understood as "harms [that] arise when a system (e.g., a search engine) represents some social groups in a less favorable light than others, demeans them, or fails to recognise their existence altogether" (Blodgett et al. 2020: 5455). Hence, representational harms are concerned with the relations between real-world social hierarchies and (verbal, audiovisual, etc.) representations of groups of people. In particular, this concerns positive or negative stereotypes, the use of denigrating language or imagery, the erasure of minority identities, Othering and dehumanisation, as well as the naturalisation of social categories more generally (see Barocas et al. 2023: 19–20; Crawford 2017: n.pag.; Dev et al.: n.pag.; Mehrabi et al. 2021b: 5017).

To further specify the nature of the harm done, it is also worth considering forms of epistemic injustice: the wrongs "done to someone specifically in their capacity as a knower" (Fricker 2007: 1).[7] Examples may include the misrepresentation, dismissal, or silencing of a person's knowledge and contribution to discourse. Analysing the consequences of ML for digital participation, Wulf Loh cautions that GenAI runs the risk of creating a new kind of digital divide that exacerbates injustice and discrimination. Models trained by conventional means, i.e., by scraping data from existing archives such as the internet, are not only conservative by default but also likely to "overlook speech acts, knowledge domains or cultural artifacts underrepresented on the web, such as minority languages or art collections" (Loh 2024: 215). The resulting misrepresentation or erasure of marginalised

7 With Miranda Fricker, we may further distinguish between testimonial injustices that occur when a speaker is given less credibility because of their social identity, and hermeneutical injustices that occur if a social group is excluded from the construction and negotiation of social values and meanings (2007: 1, 6). Both occur in the context of GenAI, for instance if the system reproduces identity prejudices ("women are irrational") that undermine their credibility (testimonial injustice) or if it prevents people from articulating their identity and experience and having that experience be part of the dataset (hermeneutical injustice). See Loh (2024) for a detailed discussion.

groups constitutes an instance of epistemic injustice because it disregards the identity of the members of said groups, discounts them as competent knowers, and ultimately renders their experience unintelligible. At the same time, this dynamic of in/exclusion is also detrimental for the majority group. Biased GenAI typically privileges culturally dominant knowledges while withholding alternative ways of knowing, thus preventing new ideas from entering the discourse (see Fricker 2007: 43–44). In these cases, GenAI not only fails to redress but in fact exacerbates social bias and epistemic injustice by further marginalising minority voices and barring their experiences from the dataset.

Similar arguments have long been articulated within the field of Postcolonial Studies. The concept of subaltern articulation, theorised by Gayatri Chakravorty Spivak in her influential essay "Can the Subaltern Speak" (1988), chimes particularly well with concerns around the epistemic in/justices produced by GenAI in that it highlights asymmetries in who is granted voice and intelligibility within systems of knowledge and power. Occupying a position of radical marginalisation, the subaltern is structurally excluded from hegemonic discourse and unrepresentable within dominant epistemic frameworks. Spivak's conclusion that the subaltern cannot speak—arguably less an expression of fatalism than a critique of "Western" attempts to speak on behalf of the subaltern—also offers an important problematisation of contemporary practices around AI ethics: Considering the structural, historical, and political conditions that render certain voices in/audible or un/intelligible, it is insufficient for powerful tech companies to introduce ethics reviews or to diversify their datasets. Quite the contrary, such gestures risk reproducing the very hierarchies they claim to challenge, thus perpetuating forms of epistemic injustice. Similar criticisms can also be levelled against the field of AI ethics itself as long as the discussion remains biased towards Western values and epistemic traditions, sidelining scholarship from the so-called Global South as well as minority voices in the Global North (see Roche et al. 2021; 2023; Segun 2021).

Finally, any analysis of discriminatory bias in AI would do well to pay attention to the complex ways in which racism, sexism, ableism, etc. overlap and combine to produce very specific forms of social inequality. What makes these forms of intersectional discrimination particularly insidious is that their cumulative effects amount to more than the sum of their parts: As Kimberlé Crenshaw (1989) has famously shown, even if a person is neither discriminated against on the grounds of her gender, nor on the

grounds of her race, she may nevertheless experience discrimination as a *Black woman.* Consequently, anti-discrimination laws based on single-axis thinking cannot protect individuals from intersectional forms of injustice and oppression—and neither can single-axis approaches to bias in AI (see Ciston 2019; Collins 2017; Costanza-Chock 2020; Hoffmann 2019; Kong 2022; Noble 2018).[8] Recognising the importance of intersectionality, ML researchers have since proposed criteria for an intersectional approach to fairness, demanding that fair AI should consider multiple attributes and protect minority groups by not only protecting each attribute value but also all intersecting values (see Foulds et al. 2020; Kearns et al. 2018) Modelling these criteria, however, remains challenging, not least because the uniqueness, complexity, and sheer number of intersecting forms of discrimination make them difficult to compute (see Fosch-Villaronga and Malgieri 2024; Kong 2022).

Clearly, neither fairness guidelines nor statistical debiasing and other technological "fixes" are sufficient to properly recognise, let alone eradicate, historically and structurally anchored intersectional forms of discrimination that are perpetuated by GenAI. To reiterate, medial representations, including AI-generated texts, images, and other outputs, transport and produce "controlling images" (see Collins 2002) and narratives that shape how socially relevant groups are imagined, spoken about, or encountered in a given society. They set the terms of who is visible, who can speak, and who is listened to; terms that are typically skewed towards the normalisation of a dichotomy between an unmarked, universal, and empowered (*white,* male) subject position and its various marked (gendered, racialised, etc.) Others (see Haitz 2022: 235). But this does not mean that resistance is futile, or that exercising meaningful epistemic agency in, through, or around GenAI would be impossible. Emerging fields such as queer, decolonial, or indigenous AI have already come up with promising critical approaches that bridge theory and practice, move across activist and academic venues, and combine scholarly scrutiny with political action. Under the banner of design justice (Costanza-Chock 2020), for instance, we find important initia-

8 That an intersectional perspective is not only helpful but essential for fair AI can easily be demonstrated by returning to the groundbreaking study on automatic facial recognition by Buolamwini and Gebru (2018). Instead of focusing on individual metrics, the authors compared the performance of Microsoft, IBM, and Face++ in relation to four subgroups that take into account both skin colour and gender and were thus able to show that the error rate for Black women was significantly higher in all three systems studied than for all other groups (see also Mehrabi et al. 2021a: 9).

tives towards community-led ways of creating and using AI that are based on principles of sustainability, accountability, or accessibility and which are overall attuned to the needs (and vulnerabilities) of those affected by the AI system. What tends to unite ethical, queer-feminist, intersectional, and decolonial approaches is their aim to break with entrenched institutional logics and naturalised patterns of thought, and their impetus to imagine AI otherwise. As I explain in more detail in the upcoming section, the arts can become vital resources for such interventions as they challenge hegemonic narratives and reinsert marginalised voices, knowledges, experiences, and imaginaries into the dataset.

3. The Intersectional Datasets of Zizi *and* Not The Only One

3.1 *Zizi—Queer(ing) the Dataset* by Jake Elwes

To create fair(er) AI, we need to pay keen critical attention to the processes by which new technologies normalise and reify socio-cultural forms of exclusion and violence, as well as the real-world consequences this has for different people (see Eubanks 2018; Hoffmann 2019). This also means paying attention to AI imaginaries, i.e., the totality of culturally dominant narratives, ideas, beliefs, and values associated with AI (see Ernst et al. 2019; Jasanoff and Kim 2015; Mager and Katzenbach 2021; Natale and Ballatore 2020). They influence not only how AI systems are understood and legitimised in the present, but also what kinds of technologies become imaginable for the foreseeable future. With a view to the state of the discourse around fair AI outlined above, the situation seems comparatively dire at first glance. To quote Köppert again, technical approaches, scholarly discourses, and fantasies about AI "cement a very specific concept of technology and, in this respect, are a tactic of concealing what AI could also be" (Köppert 2024: n.pag.). There is still hope for GenAI, though, and perhaps surprisingly to some, it comes in the form of AI art. As artists, curators, and media scholars have begun to point out, and as I will demonstrate shortly, the arts and culture can pinpoint ethical, ecological, and empowering ways of creating and using AI (see Köppert 2024: n.pag.).[9] In the words of Sandra Ciston of the Creative Code Collective, these projects

9 By now, instructive examples abound, including very visible ones such as ImageNet Roulette by Trevor Paglen and Kate Crawford, or the Algorithmic Justice League. ImageNet Roulette is a largescale installation that targets the automated interpretation

"bring intersectional thinking into tech spaces, helping shift an entrenched mindset with creative and helpful, playful and interventionist tools alike" (Ciston 2019: 5). The arts could also offer productive critical interventions in the discourse around AI fairness as such, by reflexively engaging with the inherent contradictions, complexities, and shortcomings of fairness, rather than trying to simplify or mask them. For AI ethics as a research field, it thus certainly seems worthwhile to take them seriously as contributions to the conceptualisation and implementation of fairness, and to look for productive intersections between scholarship and artistic practice.

A particularly interesting example in this regard is *The Zizi Project* (2019–ongoing), a collection of works by London-based conceptual visual artist-researcher Jake Elwes that emerged from a partnership between the artist and the Experiential AI research group at the University of Edinburgh. *The Zizi Project* takes a decidedly queer approach to the training data by combining GenAI with drag performance. Here, I would like to focus on the first part of the project, *Zizi—Queering the Dataset* (2019; hereafter: *Zizi*), which consists of a digital video that was presented at different sites as an installation with several video channels (see Elwes 2019). The video shows a series of faces, often ambiguous in terms of race and gender, slowly morphing into one another in ever-shifting constellations. Bold makeup accentuates lips and eyes, noses and other features pop in and out of existence, creating strange one-eyed creatures. Before long, a new face emerges, and is temporarily thrown into almost startling relief, before shifting yet again, never static, and never staying long enough to be pinned down by the observer's gaze. The effect is unsettling, yet strangely beautiful, and invites audiences to question their assumptions regarding the supposedly natural facts of race or gender and to instead adopt a more fluid perspective.

With regard to *Zizi*'s contribution to the discussion around fair datasets in GenAI, the processes of data curation and training, as outlined by the promotional material and the project website, are revealing (see Elwes 2019; Watling 2021). The video was created by feeding 1.000 images of drag per-

and labelling of images. Users can upload their webcam image and subsequently observe how it is being labelled by a neural network trained on the ImageNet database (Crawford and Paglen 2019; see also Ciston 2019). The Algorithmic Justice League is an organisation combining research (a.o. by prominent scholars such as Joy Buolamwini and Sasha Costanza-Chock) with activism and art, including visual arts, creative writing, and poetry. It also features strongly in the 2020 Netflix documentary *Coded Bias*.

formers into StyleGan, a generative adversarial network that was originally trained on 70.000 images of human faces contained in the Flickr-Faces-HQ Dataset, and then using it to generate new faces. As per the blurb on the artist's website, *Zizi* seeks to intervene in the myths and power structures around AI: The disruption and re-training of models "causes the weights inside the neural network to shift away from the normative identities it was originally trained on and into a space of queerness" (Elwes 2019: n.pag.). True to its "mission statement" to queer the *dataset,* the work is thus centrally concerned with the impact of training data on the output, exploring possibilities of diversifying the dataset, discarding the criterium of representativeness as an expression of a heteronormative order, and thereby finding a means of overcoming the conservatism of GenAI and facilitating the emergence of new and unexpected results.

"Queering," in this context, can be understood in a twofold sense: as an umbrella for the identities behind and beyond the acronym LGBTQIA*, and as an intervention in normative discourses and binary distinctions *per se*. Drawing on a Butlerian framework of gender performativity, *Zizi* shows that sex and gender are, indeed, performative, produced by discourse, constructed from repeated acts and moment-to-moment gestures, and interpreted by and through cultural meanings. Yet, even though this performative construction happens within the "rigid regulatory frame" (Butler 1990: 43) of normative assumptions and pressures, it is not unchangeable; an aspect that *Zizi* emphasises strongly. On the one hand, the exaggerated features of *Zizi*'s morphing images serve as a means of parody, which has value in itself by exposing the constructedness of gender as a powerful social construct. On the other hand, the ever-shifting faces highlight the fact that gender is fluid and unstable; that it must be constantly expressed and interpreted in order for the body to become legible. *Zizi* refuses to offer the audience any sort of respite or sense of reassurance that might come with a temporary halt in the flow of images where gender identity could be fixed, instead privileging the flux, ambiguity, and open-endedness of queer play. In this sense, the work demonstrates how AI-generated images need not offer simplified answers or create a false sense of objectivity, but can also express ambiguity while remaining very much attuned to complex and shifting sociocultural contexts.

The capacity of art for showing, rather than telling, holds promise when it comes to promoting critical data literacy and increasing public awareness of bias in AI. Drew Hemment and colleagues (of the Experiential AI research group that commissioned and collaborated on *The Zizi Project*)

propose the term experiential AI to capture the potential of AI art for offering new, human-centred perspectives on explainable AI (see Hemment et al. 2023; 2024). Analysing the *Zizi Show* (Elwes 2021–ongoing), a deepfake drag show that is also part of the Zizi collection, the authors conclude that "*Zizi* is an explanation of bias in ML and the power of the dataset through experiential means. *Zizi* highlights the way data and design choices shape what ML does" (Elwes 2019: n.pag.; see also Hemment et al. 2024). Data-driven art projects such as *Zizi* could offer more visceral and engaging learning processes, and thereby enhance public understanding of AI, including a sort of critical data literacy based on an awareness of the gaps and biases in the dataset and how these relate to social norms and power structures (see Hemment et al. 2023; 2024). It seems safe to assume that the different approach to explaining AI and the different type of intellectual access offered by the arts will also appeal to different kinds of audiences, notably, audiences beyond the "classic" stakeholders reached by scholarly papers and ethics guidelines. What is more, by creating productive interfaces between art, queer activism, and AI technology, these projects could also help dissolve so-called second-order divisions, i.e., the ostensibly self-imposed exclusions that hinder individuals or groups from participating in the digital realm (see Loh 2024). Upon closer inspection, these "self"-exclusions often turn out to be the result of structural factors and oppressive gendered, racialised, or ableist hierarchies, which once more points to the importance of community-led practices when it comes to re-imagining AI (see Costanza-Chock 2020).

To summarise, *Zizi—Queering the Dataset* marries the unmasking of the purportedly neutral fact of gender to the unmasking of the purported neutrality of the dataset, to the effect that both are exposed as powerful myths. In more concrete terms, *Zizi* combines its critique of sex and gender with a critique of white heteronormative bias in the datasets used to train influential AI systems. By inserting images of faces that do not conform to cis/white/heterosexual norms of gender expression, Elwes quite literally "queers" the dataset. That this also causes the algorithm to generate vastly different output clearly shows the close relation between the selection of training data and the representational politics of the output. To return to the discussion of fairness presented in Section 2, *Zizi* draws attention to the interdependence between (seemingly objective and representative) training data and societal values and hegemonic cultural formations and demonstrates the value of diversified datasets when it comes to prompting GenAI to create more inclusive, even subversive content. Through its rather

spectacular example of an AI-driven interrogation of gender norms, *Zizi* furthermore exemplifies the potential of art for increasing AI literacy, promoting critical reflection, and encouraging different groups of people to experiment with forms of AI that better suit their epistemic needs.

3.2 *Not the Only One* by Stephanie Dinkins

N'TOO, short for *Not the Only One* (Dinkins 2018–ongoing), is described on its webpage as "an ongoing experiment" and "an attempt to narrate a multigenerational memoir of a black American family" from the first-person perspective of an evolving AI (see Dinkins 2018: n.pag.). *N'TOO* was exhibited in various locations in the form of a sculpture resembling a seashell, with the faces of three Black women protruding from its surface. The bot inside the sculpture is voice-interactive, meaning that visitors can converse with it and listen to answers generated from the system's database. When curating the dataset for *N'TOO*, great care was taken to avoid importing biases and hegemonic epistemic hierarchies, focussing instead on creating a markedly intersectional database. *N'TOO* was trained on two types of data: the oral histories contributed by three Black women, representing three generations of the artist's family, and on Black diasporic literatures, films, and TV that were central to these women's experience. All of these data sources are particularly attuned to *Black women*'s experiences, and to the complex and overlapping forms of discrimination they have been facing across generations. As an evolving system, the AI also learns from its interactions with humans and expands its vocabulary, thus adding additional voices to the archive. By contrast, no open-source, ready-made, or large scraped datasets were used to avoid importing racist bias and epistemic violence that would "taint" (see Dinkins 2018: n.pag.) the self-narratives and self-understandings encoded in and communicated by the bot (see Cooper 2020; Klassen and Aceves Sepúlveda 2022; Paul 2024). As a downside, the data used to train the bot is insufficient to support seamless conversations with visitors. However, according to Dinkins, this is not so much a flaw as a crucial feature of *N'TOO*, since the gaps in the model create teachable moments about "the limitations of big data and possibilities of small data," the value of data sovereignty, and the role of the community (Dinkins 2018: n.pag.).

Not dissimilar to *Zizi*, *N'TOO* deconstructs the myth of data neutrality and encourages critical reflections about the quality of training data and its relation to systemic oppression. It does so with a strong decolonial, intersectional, and future-oriented impetus. At the core of *N'TOO*, we find a rather hopeful vision of a future in which marginalised communities would take control over the development, training, and use of GenAI, infusing it with community-specific values and tailoring it to community-specific goals. In turn, GenAI could enable the self-expression of these communities and help recover archives that seemed lost. As a first-person storyteller, *N'TOO* speaks with one voice, yet this voice remains polyphonous, representing both shared and profoundly personal experiences. As a living, Black female and multigenerational memoir, *N'TOO* shows how GenAI could help amplify voices, communicate experiences, and reconstruct genealogies that seemed lost because they were historically written out; erased from the official record. Within the framework of epistemic justice, *N'TOO* thus signifies a sort of collective "exercise of epistemic agency" by and for Black women that challenges "prevailing practices of epistemic injustice" (Collins 2017: 117). Such situated, intersectional, and community-led practices for training and using GenAI could then become effective means of writing (or rather speaking) back,[10] as well as an effective first step towards decolonising the archives upon which we train future AI systems (see Adams 2021; Hakopian 2024; Murphy and Largacha-Martínez 2022).

In its interest in drawing connections between humans and AI, *N'TOO* also chimes with posthumanist thinking and alternative models of kinship (see Haraway 1991; Nakamura 2023). Treated as a mind, interpreter, and witness in its (or rather her?) own right, the bot becomes something in-between an archive and a fourth-generation family member (see Dinkins 2024; Klassen and Aceves Sepúlveda 2022). Observing visitors engage with *N'TOO*, Dinkins takes particular note of the tentative bonds of nurture and care that emerge between humans and AI, and which seem to hint at the possibility for new forms of kinship and relationality. Dinkins is undoubtedly quite optimistic about the potential of AI to "make kin" and

10 Once again, I take inspiration from influential concepts and approaches in postcolonial (literary) studies, where the "writing back" paradigm refers to concerted and often successful efforts of postcolonial authors to contest and subvert dominant imperialist discourses through narrative practices that respond to canonical literary works (Ashcroft et al. 1989). Writing and artistic expression, in this context, become acts of reclaiming authority and demanding voice; a means of asserting control over the narrative and to re-establish interpretive sovereignty.

to erode boundaries between subject and object, self and other, human and machine. As she argues in her essay "Afro-now-ism," the present moment of rapid technological development is one of profound opportunity for imagining, and ultimately achieving, a better future beyond systemic oppression and oppositional thinking. Radical technological and cultural changes around AI, she writes, challenge us to "understan[d] ourselves as participants in an expanding continuum of intelligences" (Dinkins 2024: 5–6) and to recognise that "[t]he boundaries between sovereign consciousness, nature, valued knowledge, biotechnologies, power and social reality are optical illusions" (Dinkins 2024: 7). Rather than cautioning against the human tendency towards anthropomorphism, *N'TOO* embraces the blurring of boundaries between human and machine as a way of overcoming binary thinking and, possibly, as an alternative to technological fixes for our present condition of ontological and epistemological uncertainty.

Both *Zizi* and *N'TOO* are characterised by a transgressive approach to training data in the sense that they free themselves from the hegemonic power structures that govern our data and our mythmaking while showcasing how technology might be repurposed to empower marginalised communities. Importantly, neither project stops at recognising and criticising the systematic exclusion of non-*white* and non-heteronormative voices from the dataset. Rather, the agency, creativity, and knowledge(s) of these communities are understood as essential for thinking and building better technological futures (see Nakamura 2024). To close the circle to the theoretical and practical approaches to fairness in GenAI surveyed in the beginning, artistic interventions such as *Zizi* and *N'TOO* show how the fairness of AI could be improved by listening to marginalised communities and taking them seriously as knowers. On the one hand, this has very immediate practical implications since designers of GenAI must learn about the needs and desires of everyone affected, and especially of vulnerable groups, and accommodate this information in the design of the system (see Fosch-Villaronga and Malgieri 2024). Approaches such as Design Justice provide good-practice examples (see Costanza-Chock 2020). On the other hand, the inclusion and wider familiarity with intersectional approaches stands to affect the scholarly discourse, ideally leading to a more contextualised and epistemically just understanding of fairness in AI.

4. Conclusion

Drawing on an extensive and fast-growing body of research on fairness in ML, this chapter proposed an understanding of fairness as freedom from discriminatory bias, where biases are understood as systematic errors that lead to the unfair differential treatment of socially relevant groups of people. In the context of GenAI, discriminatory biases can lead to representational harms by perpetuating stereotypes, generating demeaning content, or by erasing specific identities, typically those that are already disadvantaged in "real life." Remedying unfair discriminatory biases in GenAI is therefore imperative from an ethical point of view, but is currently hindered by tricky challenges that indicate the need for contextual, community-led, and intersectional approaches. I have moreover argued that the arts and culture can point the way towards possible solutions to the dilemmas and paradoxes of fair AI, not least because they provide access to a more diverse set of epistemic resources. On the basis of these theoretical considerations, the chapter then discussed *Zizi—Queering the Dataset* by Jake Elwes and *Not the Only One* by Stephanie Dinkins as two examples of art projects that use diverse datasets in order to re-imagine fairness in GenAI from queer and decolonial perspectives. In emphasising intersectional thinking and the agency of racialised and marginalised groups, these works highlight the untapped potential of playful, creative, and community-led approaches for infusing GenAI with new ideas and values in the present, and for imagining fairer AI futures.

Funding Declaration: This work was funded by the German Ministry of Education and Research within the project "Privacy, Democracy, and Self-Determination in Times of AI and Globalization" (PRIDS). Funding no.: 16KIS1380.

Acknowledgements: I wish to thank my colleagues Jana Hecktor and Lisa Koeritz for their thorough reading and knowledgeable feedback on an early draft of this paper. Any remaining errors are my own.

References

Adams, Rachel (2021): Can Artificial Intelligence Be Decolonized?, in: Interdisciplinary Science Reviews 46 (1–2), pp. 176–197.

Akter, Shahriar et al. (2021): Algorithmic Bias in Data-Driven Innovation in the Age of AI, in: International Journal of Information Management 60, article 102387.

Angwin, Julia et al. (2016): Machine Bias, in: ProPublica, 23 May 2016 (online available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing – accessed 18 October 2024).

Ashcroft, Bill / Griffiths, Gareth / Tiffin, Helen (1989): The Empire Writes Back: Theory and Practice in Post-Colonial Literatures, New York.

Barocas, Solon / Hardt, Moritz / Narayanan, Arvind (2023): Fairness and Machine Learning: Limitations and Opportunities, Cambridge.

Barocas, Solon / Selbst, Andrew D. (2016): Big Data's Disparate Impact, in: California Law Review 104 (3), pp. 671–732.

Benjamin, Ruha (2019): Race After Technology: Abolitionist Tools for the New Jim Code, Cambridge.

Binns, Reuben (2018): Fairness in Machine Learning: Lessons from Political Philosophy, in: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81, pp. 149–159.

Blodgett, Su Lin et al. (2020): Language (Technology) Is Power: A Critical Survey of "Bias" in NLP, in: Dan Jurafsky et al. (eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5454–5476.

Brandner, Lou Therese / Hirsbrunner, Simon David (2023): Algorithmische Fairness in der polizeilichen Ermittlungsarbeit: Ethische Analyse von Verfahren des maschinellen Lernens zur Gesichtserkennung, in: TATuP 32 (1), pp. 24–29.

Brandner, Lou Therese et al. (2023): How Data Quality Determines AI Fairness: The Case of Automated Interviewing, in: EWAF'23: European Workshop on Algorithmic Fairness, Winterthur, Switzerland (online available at: https://ceur-ws.org/Vol-3442/paper-25.pdf – accessed 18 October 2024).

Buolamwini, Joy / Gebru, Timnit (2018): Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, in: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81, pp. 77–91.

Butler, Judith (1990): Gender Trouble: Feminism and the Subversion of Identity, New York.

Chowdhury, Rumman / Mulani, Narendra (2018): Auditing Algorithms for Bias, in: Harvard Business Review (online available at: https://hbr.org/2018/10/auditing-algorithms-for-bias – accessed 18 October 2024).

Ciston, Sarah (2019): Intersectional AI Is Essential: Polyvocal, Multimodal, Experimental Methods to Save Artificial Intelligence, in: Journal of Science, Technology and Arts 11 (2), pp. 3–8.

Collins, Patricia Hill (2002): Mammies, Matriarchs, and Other Controlling Images, in: Patricia Hill Collins, Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment, New York, pp. 69–96.

Collins, Patricia Hill (2017): Intersectionality and Epistemic Injustice, in: Ian James Kidd / José Medina, Gaile Pohlhaus Jr. (eds.), The Routledge Handbook of Epistemic Injustice, New York, pp. 115–124.

Cooper, Imani (2020): Inheritance: Ode to N'TOO, in: Absinthe 26 (1), n.pag.

Costanza-Chock, Sasha (2020): Design Justice: Community-Led Practices to Build the Worlds We Need, Cambridge.

Crawford, Kate (2017): The Trouble with Bias, in: AI Now Institute (online available at: https://ainowinstitute.org/news/the-trouble-with-bias – accessed 16 December 2025).

Crawford, Kate / Paglen, Trevor (2019): Excavating AI: The Politics of Images in Machine Learning Training Sets (online available at: https://www.excavating.ai – accessed 18 October 2024).

Crenshaw, Kimberlé (1989): Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics, in: University of Chicago Legal Forum, pp. 139–167.

Dev, Sunipa et al. (2022): On Measures of Biases and Harms in NLP, arXiv. https://doi.org/10.48550/arXiv.2108.03362

Dinkins, Stephanie (2018): Not The Only One: Project Website (online available at: https://www.stephaniedinkins.com/ntoo.html – accessed 18 October 2024).

Dinkins, Stephanie (2024): Afro-Now-Ism: The Unencumbered Black Mind Is a Wellspring of Possibility, in: Srimoyee Mitra (ed.), Stephanie Dinkins on Love and Data, Ann Arbor, pp. 4–15.

Eapen, Tojin T. et al. (2023): How Generative AI Can Augment Human Creativity, in: Harvard Business Review (online available at: https://hbr.org/2023/07/how-generative-ai-can-augment-human-creativity – accessed 18 October 2024).

El Atillah, Imane (2023): Copyright Challenges in the Age of AI: Who Owns AI-Generated Content?, in: Euronews, 10 July (online available at: https://www.euronews.com/next/2023/07/10/copyright-challenges-in-the-age-of-ai-who-owns-ai-generated-content – accessed 18 October 2024).

Ellingrud, Kweilin et al. (2023): Generative AI and the Future of Work in America, McKinsey Global Institute (online available at: https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america – accessed 18 October 2024).

Elwes, Jake (2019): Project Website for Queering the Dataset (online available at: https://www.jakeelwes.com/project-zizi-2019.html – accessed 18 October 2024).

Ernst, Christoph / Schröter, Jens / Sudmann, Andreas (2019): AI and the Imagination to Overcome Difference, in: spheres: Journal for Digital Cultures 5, pp. 1–12.

Eubanks, Virginia (2018): Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, New York.

Feuerriegel, Stefan / Dolata, Mateusz / Schwabe, Gerhard (2020): Fair AI: Challenges and Opportunities, in: Business & Information Systems Engineering 62 (4), pp. 379–384.

Fleming, Sam (2024): Generative Artificial Intelligence Will Lead to Job Cuts This Year, CEOs Say, in: Financial Times, 15 January (online available at: https://www.ft.com/content/908e5465-0bc4-4de5-89cd-8d5349645dda – accessed 18 October 2024).

Fosch-Villaronga, Eduard / Malgieri, Gianclaudio (2024): Queering the Ethics of AI, in: David J. Gunkel (ed.), Handbook on the Ethics of Artificial Intelligence, Cheltenham, pp. 301–315.

Foulds, James R. et al. (2020): An Intersectional Definition of Fairness, in: 36th International Conference on Data Engineering (ICDE), pp. 1918–1921.

Fricker, Miranda (2007): Epistemic Injustice: Power and the Ethics of Knowing, Oxford.

Gautam, Sanjana / Venkit, Pranav N. / Ghosh, Sourojit (2024): From Melting Pots to Misrepresentations: Exploring Harms in Generative AI, arXiv. https://doi.org/10.48550/arXiv.2403.10776

Hagendorff, Thilo (2019): From Privacy to Anti-Discrimination in Times of Machine Learning, in: Ethics and Information Technology 21, pp. 331–343.

Hagendorff, Thilo (2021): Linking Human and Machine Behavior: A New Approach to Evaluate Training Data Quality for Beneficial Machine Learning, in: Minds and Machines 31 (4), pp. 563–593.

Hagendorff, Thilo / Wezel, Katharina (2020): 15 Challenges for AI: Or What AI (Currently) Can't Do, in: AI & Society 35 (2), pp. 355–365.

Haitz, Louise (2022): Medienwissenschaft und Intersektionalität, in: Astrid Biele Mefebue / Andrea D. Bührmann / Sabine Grenz (eds.), Handbuch Intersektionalitätsforschung, Wiesbaden, pp. 229–242.

Hakopian, Mashinka Firunts (2024): Art Histories from Nowhere: On the Coloniality of Experiments in Art and Artificial Intelligence, in: AI & Society 39 (1), pp. 29–41.

Haraway, Donna (1991): A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century, in: Donna Haraway, Simians, Cyborgs, and Women: The Reinvention of Nature, New York, pp. 149–181.

Heesen, Jessica / Reinhardt, Karoline / Schelenz, Laura (2021): Diskriminierung durch Algorithmen vermeiden: Analysen und Instrumente für eine digitale demokratische Gesellschaft, in: Gero Bauer et al. (eds.), Diskriminierung und Antidiskriminierung: Beiträge aus Wissenschaft und Praxis, Bielefeld, pp. 129–148.

Hemment, Drew et al. (2023): AI in the Public Eye: Investigating Public AI Literacy through AI Art, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, New York, pp. 931–942.

Hemment, Drew et al. (2024): Experiential AI: Enhancing Explainability in Artificial Intelligence through Artistic Practice, in: Leonardo 57 (3), pp. 298–306.

Hoffmann, Anna Lauren (2019): Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse, in: Information, Communication & Society 22, pp. 900–915.

Hsu, Tiffany / Thompson, Stuart A. (2023): Disinformation Researchers Raise Alarms about A.I. Chatbots, in: The New York Times, 20 June (online available at: https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html – accessed 18 October 2024).

Jasanoff, Sheila / Kim, Sang-Hyun (2015): Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power, Chicago.

Kearns, Michael et al. (2018): Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness, in: PMLR 80, pp. 2564–2572.

Kim, Pauline T. (2022): Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action, in: California Law Review 110 (5), pp. 1539–96.

Klassen, Lois / Aceves Sepúlveda, Gabriela (2022): Amplified Listening to Race and Gender in Fiamma Montezemolo's "Echo" and Stephanie Dinkins's "N'TOO", in: Media-N 18 (1), pp. 102–120.

Kong, Youjin (2022): Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, New York, pp. 485–494.

Köppert, Katrin (2024): Queersplaining AI, in: Boell, 28 May (online available at: https://eu.boell.org/en/2024/05/28/queersplaining-ai – accessed 18 October 2024).

Leavy, Susan / O'Sullivan, Barry / Siapera, Eugenia (2020): Data, Power and Bias in Artificial Intelligence, arXiv. https://doi.org/10.48550/arXiv.2008.07341

Lin, Cong et al. (2023): Trapped in the Search Box: An Examination of Algorithmic Bias in Search Engine Autocomplete Predictions, in: Telematics and Informatics 85.

Loh, Wulf H. (2024): Generative KI, digitale Teilhabe und epistemische Ungerechtigkeit, in: Rechtsphilosophie – Zeitschrift für Grundlagen des Rechts 10 (2), pp. 215–233.

Mager, Astrid / Katzenbach, Christian (2021): Future Imaginaries in the Making and Governing of Digital Technology: Multiple, Contested, Commodified, in: New Media & Society 23 (2), pp. 223–236.

Marr, Bernard (2023): The Intersection of AI and Human Creativity: Can Machines Really Be Creative?, in: Forbes, 27 March (online available at: https://www.forbes.com/sites/bernardmarr/2023/03/27/the-intersection-of-ai-and-human-creativity-can-machines-really-be-creative – accessed 18 October 2024).

Mehrabi, Ninareh et al. (2021a): A Survey on Bias and Fairness in Machine Learning, in: ACM Comput. Surv. 54 (6), Article 115, pp. 1–35.

Mehrabi, Ninareh et al. (2021b): Lawyers Are Dishonest? Quantifying Representational Harms in Commonsense Knowledge Resources, in: Marie-Francine Moens et al. (eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, pp. 5016–5033.

Mitchell, Shira et al. (2021): Algorithmic Fairness: Choices, Assumptions, and Definitions, in: Annual Reviews of Statistics and Its Applications 8 (1), pp. 141–163.

Mohammed, Sedir et al. (2023): Ein Glossar zur Datenqualität (1.2), Zenodo. https://doi.org/10.5281/zenodo.7702426

Murphy, John W. / Largacha-Martínez, Carlos (2022): Decolonization of AI: A Crucial Blind Spot, in: Philosophy & Technology 35 (4), pp. 1–13.

Nakamura, Lisa (2023): "Who Are Your People?": Stephanie Dinkins's Afro-Now-ism as Algorithmic Abundance, in: Mitra, Srimoyee (ed.), Stephanie Dinkins on Love and Data, Ann Arbor, pp. 52–59.

Natale, Simone / Ballatore, Andrea (2020): Imagining the Thinking Machine: Technological Myths and the Rise of Artificial Intelligence, in: Convergence 26 (1), pp. 3–18.

Noble, Safiya Umoja (2018): Algorithms of Oppression: How Search Engines Reinforce Racism, New York.

Paul, Christiane (2024): The Data You Give, in: Mitra, Srimoyee (ed.), Stephanie Dinkins on Love and Data, pp. 28–35, Ann Arbor.

Prates, Marcelo O.R. / Avelar, Pedro H. / Lamb, Luís C. (2020): Assessing Gender Bias in Machine Translation: A Case Study with Google Translate, in: Neural Computing and Applications 32, pp. 6363–6381.

Roche, Cathy / Lewis, Dave / Wall, P.J. (2021): Artificial Intelligence Ethics: An Inclusive Global Discourse?, Proceedings of the 1st Virtual Conference on Implications of Information and Digital Technologies for Development, arXiv. https://doi.org/10.48550/arXiv.2108.09959

Roche, Cathy / Wall, P.J. / Lewis, Dave (2023): Ethics and Diversity in Artificial Intelligence Policies, Strategies and Initiatives, in: AI and Ethics 3 (4), pp. 1095–1115.

Segev, Re'em (2025): The Moral Status of Input and Output Discrimination, in: AI and Ethics 5 (1), pp. 323–332.

Segun, Samuel T. (2021): Critically Engaging the Ethics of AI for a Global Audience, in: Ethics and Information Technology 23 (2), pp. 99–105.

Spivak, Gayatri Chakravorty (1988): Can the Subaltern Speak?, in Cary Nelson / Lawrence Grossberg (eds.), Marxism and the Interpretation of Culture, London, pp. 24–28.

Sun, Xiao-yu / Ye, Bin / Xia, Bao-hua (2025): The Problem of Fairness in Tools for Algorithmic Fairness, in: AI and Ethics 5, pp. 1847–1857.

Tacheva, Jasmina / Ramasubramanian, Srividya (2023): AI Empire: Unraveling the Interlocking Systems of Oppression in Generative AI's Global Order, in: Big Data & Society 10 (2), pp. 1–13.

Thomsen, Frej Klem (2024): Algorithmic Indirect Discrimination, Fairness and Harm, in: AI and Ethics 4, pp. 1023–1037.

Verma, Pranshu / De Vynck, Gerrit (2023): ChatGPT Took Their Jobs. Now They Walk Dogs and Fix Air Conditioners, in: The Washington Post, 2 June (online available at: https://www.washingtonpost.com/technology/2023/06/02/ai-taking-jobs/ – accessed 18 October 2024).

Wang, Xiaomen / Zhang, Yishi / Zhu, Ruilin (2022): A Brief Review on Algorithmic Fairness, in: Management System Engineering 1, Article 7.

Watling, Eve (2021): Meet the Artist Queering AI Technology, in: The Independent, 26 July (online available at: https://www.independent.co.uk/arts-entertainment/photography/zizi-queering-dataset-ai-drag-jake-elwes-b1876396.html – accessed 18 October 2024).

The Zizi Project by Jake Elwes (n.d.): The New Real (online available at: https://www.newreal.cc/artworks/the-zizi-project – accessed 18 October 2024).

Zou, James / Schiebinger, Londa (2018): AI Can Be Sexist and Racist — It's Time to Make It Fair, in: Nature 559, pp. 324–326.