3

INCORPORATING INTERSECTIONALITY INTO AI ETHICS

Liza Ireni-Saban and Maya Sherman

Introduction

AI-driven devices have increasingly become mediators of our social, cultural, economic, and political interactions (Rahwan et al. 2019). Therefore, the appearance of more nuanced and sophisticated aspects of diversity, as well as the emergence of new ways of thinking about identity, require that the notion of AI ethics establishes new tools and strategies for supporting and advocating diversity and inclusion in contemporary AI developments.

Within this notion, it is important to mention that the growing impact of AI technologies in human reality has enabled the strengthening of the scope and scale of disinformation campaigns and fake news dissemination. The scholastic perception of AI in the disinformation sphere is ambiguous, as it accelerates data propagation online via social media platforms but also enables us to automatically detect false content and remove it at a relatively high level of accuracy. Consequently, the alleged nexus between AI and disinformation has amplified the ethical discourse regarding AI usage and implementation.

Moral philosophers and applied ethicists often suggest using a 'deontological' approach to moral norms as a starting point, while others suggest a teleological (consequentialist) ethical approach concerned with whether an action or decision leading to an outcome is good or bad to the society as a whole (Gomila and Amengual 2009). It is argued that for AI as machines with a more limited degree of autonomy, a rule-based approach may be sufficient, or even endorsed. Since these theories are based on explicit rules and norms external to the real world in which they are to be applied, they have limited practical value. In other words, these theories are, from a philosophical perspective, unable to fully explicate the complexities of moral considerations as these complexities are experienced in the world. This chapter suggests that deontological and utilitarian ethics cannot fully address the challenges that AI technological innovations pose

to contemporary ethics, which require a more flexible, context-dependent, and case-based approach to morality.

For that, we offer to investigate how AI ethics can be enhanced through critical engagement with intersectionality. Ethical perspectives on intersectionality share normative ideals toward social justice. It is suggested that the development of AI systems has brought forth, with unprecedented clarity, the socioeconomic differences across all identities, and the recognition that the experience of privilege based on social groups and locations is fluid rather than static. The junction of intersectionality theory, ethics, and AI allows us to conceptualise and harness, for the first time, patterns of inequality as redistribution of power, and privilege is a core tenant of deliberative intersectional engagement. An intersectional framework for AI ethics can be used to scrutinise the algorithmic biases and issues rooted in existing AI-driven systems and applications.

This chapter is organised as follows. In the first section, we introduce the theoretical lens of intersectionality and then move to a discussion on how AI ethics can benefit from insights of intersectionality. Finally, we present concrete examples depicting the challenges of AI to underscore the new and different insights intersectionality generates for AI ethics.

Intersectionality: A new paradigm

Albeit a fledgling analytical paradigm, the evolution and development of intersectionality have been long in the making. The concept of intersectionality owes its origins to the feminist movement, which sought to develop a more comprehensive and encompassing schema for recognising and appreciating the converging forces of oppression that affect women on different dimensions of identity. Intersectionality arose from the works by women of colour in the 1960s as a means for expressing the limitations of feminist theory to accurately portray the struggle of women across racial and class boundaries (Samuels and Ross-Sheriff 2008: 5). In fact, in its most essential form, intersectionality theory reflects a criticism of second-wave feminism, which was the preeminent mode of feminism in the 1960s in the US. The second wave of feminism expanded the goals set out by the first wave, which primarily took up the causes of universal suffrage and repealing discriminatory legislation. The second wave of feminism focussed on broadening gender equality by addressing issues like sexuality, domestic rights, reproductive rights, and *de facto* discrimination (Burkett 2016). However, opponents of the emergent movement asserted that the second wave favoured a historical narrative that "whitewashed" and homogenised the feminist struggle and ignored different voices of minority communities such as black women and queer women (Orr and Braithwaite 2012).

The term intersectionality was first coined by Kimberlé Crenshaw (1989), who used the metaphor of intersecting roads to illuminate how differing levels of oppression on the grounds of gender and race interact with one another to create a new and unique experience of marginalisation and discrimination. Crenshaw

offered the concept of intersectionality as redress to the singularity and unidimensional consideration of the phenomenon of oppression. Although intersectionality theory's origin is rooted in the struggle of women of colour for recognition within the big-tent feminist movement of the 60s and 70s, as Samuels and Ross-Sheriff (2008: 5) note, it went even further and called on scholars to acknowledge that "for many women of colour, their feminist efforts are simultaneously embedded and woven into their efforts against racism, classism, and other threats to their access to equal opportunities and social justice" (ibid: 5).

The modern definition of intersectionality holds that "gender cannot be used as a single analytic frame without also exploring how issues of race, migration status, history, and social class, in particular, come to bear on one's experience as a woman" (ibid.). Consequently, the methodological approaches of researchers and academics employing an intersectional technique mandate that they explore the multitude of "the overlapping and mutually reinforcing" systems of oppression. The once-accepted universalist approach to the constructs of "woman" or "feminist" as singular, all-encompassing experiences has now been replaced by analyses that consider women as whole individuals whose identities may be informed and reinforced by multiple interlocking structures of oppression. Finally, current intellectual pursuits of intersectional analyses incorporate not only mutually reinforcing systems of oppression but also the myriad of privileges which also inform the feminist experience. Since intersectionality was developed in reaction to and as a criticism of the tendency of feminist narratives to whitewash oppression experiences, the development of intersectionality theory evolved alongside the dialectical evolution of the feminist movement.

Just as important as the subjects of intersectional analysis, one may mention the relationship between the myriad of systems of oppression and structures of power which interact to shape intersectional experiences. This paradigm states that just like different identity layers coalesce to create unique experiences of discrimination, the structures of domination which perpetuate systems of inequality inherently intersect with axes of oppression. Fellows and Razack (1998: 335) suggest this mechanism functions as mutually assimilated networks that "rely on one another" so that "systems of oppression could not be accomplished without gender and racial hierarchies; imperialism could not function without class exploitation, sexism, heterosexism and so on."

In addition to the myriad of levels of oppression, intersectionality acknowledges the varying degrees of privilege, which also inform the unique experiences of women. These privileges occur naturally from the deficits created by the structures of oppression. An example of the symbiotic relationship between privilege and oppression is evident in Samuels and Ross-Sheriff's research (2008) on black or multiracial young children adopted by white parents. Since there was a largely socioeconomic, ethnic, and cultural homogeneity within the interviewees' neighbourhoods, they were inevitably a racial minority in their own community. Here, the interplay between privilege (socioeconomic status) and racism creates the very incubator in which the biracial children experience a unique system of oppression.

Being a biracial or trans-racial adoptee in a white community meant that their experience of structural oppression was unique to their particular set of privileges and oppressions. Although having two white parents meant that they were transmitting the dominant group's culture, and this ultimately allowed the adoptees to operate largely in white race contexts comfortably, few of them reported dating in high school because their appearance was devalued by the dominant "Eurocentric images of beauty" (ibid: 7). In this example, we see that while being raised in a white community endows certain privileges, it simultaneously and inherently creates situations of alienation. Samuels and Ross-Sheriff's anecdotal research demonstrates that not only is oppression an integral component of intersectionality, but in order to fully appreciate the impact of systems of oppression on individual experiences, academics must also consider the networks of privilege.

Although intersectionality theory is intrinsically related to the feminist movement, its methodological contribution reaches far beyond feminist debates. The intersectional methodology encourages researchers to investigate the multilayered effects of experiences of oppression in their unique and varied manifestations. This is a departure from traditional methodological techniques that often pursue a parsimonious quality in both variables and conclusions. The epistemological approaches to the different forms of oppression to this point had been discrete in nature – the exploration of patriarchy was a distinct pursuit, and therefore experiences of victimisation from institutionalised sexism were interpreted as if they existed in a vacuum. Likewise, racism was investigated as a stand-alone system of persecution. In many ways, the methodological inadequacies of research had a deterministic effect on the analysis of the experiences of oppression themselves. Crenshaw put forth that a "single-axis framework" failed to consider the compounded marginalisation that women of colour faced. Crenshaw's foundation offered a theoretical schema for understanding not only bi-level discrimination but multiple layers of oppression (Dhamoon 2011: 231). This differed dramatically from the traditional single-group approach, which attempted to investigate the phenomena "by analyzing the intersection of a subset of dimensions of multiple categories" (McCall 2005: 1787). Single subject design is a subgroup of the categorical comparative approach, and it is useful for streamlining analytical spaces which can become convoluted when multiple groups and levels are compared side-by-side (ibid: 1786). For example, if researchers want to compare specific ethnic groups within broader racial classifications – e.g. Vietnamese, Thai, and Laos subgroups within the more general grouping of Southeast Asian – it becomes necessary to restrict the breadth of analysis for the sake of comprehension. Therefore, a study of this nature would consider these Southeast Asian subgroups independently of gender or class. Naturally, this method has its advantages as it allows researchers to simplify the subject of their research for 'big picture analysis'. However, the very aspect which makes this analytical framework attractive – the ability to disregard intermediary layers of analysis – is also its pitfall. Research which isolates its subject from the multitude of intervening affective torrents of complexity is ultimately reductionist in its analysis.

The Southeast Asian research ignored gender and class in the investigation of Southeast Asians in order to maintain simplicity. However, an intersectionality study of Southeast Asians would employ an ecological model to identify the integrative nature of the myriad of Southeast Asian experiences. For example, it would distinguish the experience of the middle-class, Vietnamese man in comparison to the low-class, Vietnamese woman, and so on and so forth. Although this type of multidimensional, 'interaction effect' modelling makes the research exponentially more complicated, it is arguably the only design equipped to deal with the confluence of multiple systems of oppression and paradigms of power. Furthermore, intersectionality allows researchers to investigate "how multiple and differing sets of interactive processes and systems vary at different levels of life and across time and space" (Dhamoon 2011: 237). This ideation of subjects of oppression and power as dynamic, multilayered, and complex is, of course, antipodal to the positivist tradition which assumes that all phenomena are fixed, generalisable, and fully conceivable. Instead, intersectionality values unpacking and evaluating processes and systems (ibid.).

An additional methodological approach that strives to satisfy this call for complexity is called anti-categorical complexity (McCall 2005). This approach deconstructs the reductionist analytical categories, maintaining that social life and social structures are infinitely too complex and dynamic to be fettered by fixed categorical definitions. Anti-categorical complexity has been applied in deconstructing once-finite categories such as sexuality or gender and examining how they are instead socially designed constructs (Fotopoulou 2012). The anti-categorical approach which emerged from the critique moved to the tendency of white, big-tent feminists to frame women and gender as essential and homogenous categories embracing all women (McCall 2005). The crux of the criticism was that no solitary category could aptly account for the host of experiences of the individual. Additionally, most intersectional experiences did not fit cleanly into these socially constructed categories. Critics also highlighted that the pro-categorisation camp was reinforcing inequalities by excluding experiences that did not fit comfortably into the socially eschewed constructions.

The second approach to complexity is referred to as inter-categorical complexity (ibid.). This approach accepts the socially constructed categories *pro tempore* as a provisional means for tracking the disparities between social groups along multiple lines of intersecting identities, dimensions, and power structures. The fundamental assumption of inter-categorical complexity is that although the relationships and interstices of inequality are fluid and ever-shifting, by adopting categories and simultaneously considering their intersections, researchers are afforded the leverage granted by comparative modes of analysis (Bauerband and Galupo 2014). McCall (2005) puts forth intra-categorical complexity as a last approach to the complexity of intersectionality. Intra-categorical complexity falls somewhere in-between the anti-categorical approach, which wholeheartedly rejects categorisation, and the inter-categorical approach, which provisionally excepts categories,

if only for the purpose of comparative analysis. The intra-categorical complexity approach appreciates the methodological potential of categories but tends to focus on "neglected points of intersection – 'people whose identity crosses the boundaries of traditionally constructed groups'" (Dill 2002: 5).

The contribution of intersectional methodologies, although they introduced new obstacles, was paramount for the poststructuralist movement and the larger popular movement to deconstruct social boundaries as a means of combating inequality. Ultimately, the methodological subgroups challenged the then-predominant mode of analysis which suffered from a blatant failure to reflect the *loci* of neglected experiences of oppression.

However, the introduction of intersectionality methodology shall include its limitations. Although intersectionality offers a versatile theoretical basis for researching modes of oppression and privilege, it has simultaneously complicated methods of analysis. In fact, the defining aspect of the methodology of intersectionality studies is "the complexity that arises when the subject of analysis expands to include multiple dimensions of social life" (McCall 2005: 1772). Indeed, most scholars have accepted the legitimacy and necessity of intersectionality to convey the intricacies of intersecting experiences of real life, and yet, intersectionality remains underdeveloped without a practical application.

More recently, Reyes (2017) and Moore (2012) have advocated intersectionality as a useful lens for the shifted focus of code-switching to marginalised factions within society. This stream of research is especially relevant to the social identity framework of intersectionality underlying AI ethics offered in this chapter. We elucidate the development of AI by reference to the basic premises of social interactionism. They include the following assumptions: capturing reality by individuals is a social construction; individuals constantly affect one another as through their interaction over time; individuals are capable of deliberate actions and the way they interact with others and within ourselves; individuals define what exists and decide how to act accordingly. Therefore, we consider social identity as "the self as reflexively understood by the individual in terms of his or her biography" (Giddens 1991: 244). It should be noted that while one's concept of the self may remain consistent over time, social identity is more familiar with a process of shifts and adjustments as it plays out in everyday life. Through a process of social interaction, we work to communicate our identities to others, while we attribute identities to them (Charon 2010; Gecas 1982).

Contemporary issues of AI technologies in the ethical sphere

Following the discussion presented above, we will elaborate upon the contemporary issues relating to AI technologies in the ethical sphere. Although there is no one accepted definition of AI, various scholars address the machine's ability to exhibit intelligent behaviour, react to the environment, and learn from it (Samoili et al. 2020). Nonetheless, this chapter will focus on the AI's twofold functionality

within the disinformation sphere. Meaning, AI as referable to complex algorithmic models allowing to automatically generate, detect, and mitigate false contents online and impact on public opinion.

Broadly, there are several disinformation-related affairs that revolve around the evolution of AI technologies. Among these incidents, various scholars highlight the Cambridge Analytica affair and its hidden manipulation of ad targeting (Cadwalladr and Graham-Harrison 2018).[1] Other notable disinformation affairs are the incitement of ethnic cleansing in Myanmar[2] and the emergence of masses of fake Russian accounts (Eidelson 2018; Bloomberg Editorial Board 2017; Weise 2017). Notably, these affairs highlight the unprecedented implications of AI bias in the international community.

One of the most prevalent bias types is the historical bias, which represents an existing inequality and socio-technical issues within the data generation process (Suresh and Guttag 2019). An example emerged in 2015, when academic and media sources revealed a clear gender bias within Google search engine. In the incident, the top results for 'CEO' image search showed mainly photos of men, and when the search engine identified the seeking user as female, it displayed fewer ads for executive positions. It represents a historical bias, since this kind of bias reflects an existing gender inequality in society (Suresh and Guttag 2019; Yapo and Weiss 2018).

The omnipresent spread of AI in the cybernetic and physical spheres has led to a broader discourse regarding its ethical implications. On the one hand, AI-driven interfaces enable the analysis of large sums of data and provide us with a tailored user experience and enhanced personalisation processes, as seen within various fields such as autonomous driving, predictive policing, and language translation. On the other hand, one must consider the ambiguous outcomes of AI usage from the legal, social, and ethical perspectives (Doshi-Velez et al. 2017; Amodei et al. 2016; Sculley et al. 2014; Bostrom 2003; McCarthy 1960). Interesting to note, a Deloitte survey of tech executives in the US (Loucks, Davenport, and Schatsky 2018) highlighted the potential ethical risks of AI, with emphasis on its falsification of contents and imagery and the increase of algorithmic bias.

Therefore, discussing AI and disinformation requires a deeper analysis of the notion of *algorithmic bias*, including its different types as well as its main sources. The literature raises various bias types, which depend both on the data itself and on the user. Data bias may be the result of technical or computational matters, an inappropriate algorithmic deployment, or a user misinterpretation of the algorithm's outputs. Danks and London (2017) highlight notable computational sources of bias, such as the training data bias, and algorithmic processing. For instance, the input data used may be biased and lead to biased outputs for the algorithmic tasks. Furthermore, the algorithm itself may be biased, such as in cases of a statistically biased estimator within the algorithm.

As AI-driven algorithms have become highly prevalent in our decision-making processes and day-to-day practices, the risk of generating discriminatory and offensive outputs rises significantly (ibid). In addition to bias and fake news, one can mention fairness and privacy violations as significant ethical loopholes. From the

privacy prism, there is an ongoing conflict between data privacy and efficacy, since AI-driven models enable access to large sums of data and allows it to be analysed with greater accuracy (Whittlestone et al. 2019; Zimmerman 2018). For instance, AI technologies such as sensor networks, social media tracking, and facial recognition enable us to broaden surveillance practices and threaten one's right to privacy. These technologies in use may lead to discriminatory patterns and stigmatisation, even without one's explicit consent or knowledge (Whittaker et al. 2018). In this regard, Mehrabi et al. (2019) suggest:

> like people, algorithms are vulnerable to biases that render their decisions 'unfair' [. . .] fairness is the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics. Thus, an unfair algorithm is one whose decisions are skewed toward a particular group of people
>
> (ibid.: 1).

For example, in the US one can count numerous incidents in which AI usage has exacerbated existing social inequalities, such as the gender bias of the Amazon recruiting tool (Dastin 2018) and the notable discriminatory credit algorithms against minority groups (Bartlett et al. 2018; Glantz and Martinez 2018; Waddell 2016). In addition, the ProPublica investigation found that the risk tool used to create Florida risk scores was biased against black people and led to discriminatory outputs (Angwin et al. 2016).

Interesting to note, several scholars argue that these ethical constraints differ between cultures and languages. For example, Facebook misinterpreted a Palestinian man's post in 2017 due to a machine-translation error, which led to his arrest by the Israeli police (Hagerty and Rubinov 2019; Hern 2017). Therefore, this type of linguistic error might lead to significant implications due to cultural controversies. Another linguistic incident occurred within the ambiguous translation of the 'like' button, leading to algorithmic filtering of an indigenous collective in Brazil (Ochigame and Holston 2016).[3] As Ochigame and Holston suggested in their article on this incident: "AI principles are inevitably value-laden terms, dense with significance. Such terms, even when thoughtfully translated, can have distinct connotations and meanings in different cultures" (ibid.: 10).

Following the cultural perspective, Hagerty and Rubinov (2019) demonstrate that low- and middle-income countries might be more susceptible to ethical implications and not necessarily receive the potential AI benefits. This notion was highlighted by the World Economic Forum (2018), as it claimed that developing countries are prone to great risks of discriminatory patterns when using machine learning methods. From the policy perspective, Hashmi et al. (2019) emphasise the increasing ambiguity revolving around the AI implementation in the public service: "The advent of AI raises a host of ethical issues, related to moral, legal, economic and social aspects of our societies and government officials face challenges and choices pertaining to how to apply AI technologies in the public sector and in governance strategies" (ibid.: 8).

In this regard, it is important to point out that the AI ability to manipulate and deceive human users online turns these ethical issues in the data-driven age into inherent loopholes, which should be properly moderated by policy makers and regulators. Interesting to mention, several scholars consider AI as a countermeasure against the rise of fake news on social media platforms due to its ability to identify fake bot accounts and automated fact-checking. Moreover, current deep learning models enable the enhancement of text classification and analysis of online content (Sharma et al. 2019). According to Facebook, AI tools are responsible for the removal of 99.5% of terrorist-related content and 98.5% of fake accounts (Marsden and Meyer 2019; Kertysova 2018).

Nonetheless, one must consider the monumental contribution of AI to the creation and propagation of disinformation campaigns. With the advance of machine learning and NLP, one's ability to automatically generate content and tailor it for unique users is amplified, and therefore AI enhances the microtargeting of vulnerable audiences (Kertysova 2018). As a result, AI represents an offensive instrument, and not only a defensive component in the detection processes of online bots (Yang et al. 2019). Remian (2019) highlights this dual perception of AI in the disinformation sphere: "Artificial intelligence has the potential to be used for the spread of disinformation, propaganda, and the shaping of social and cultural values. As with security, AI may play a role in both delivering and protecting from misinformation and attempts to manipulate" (ibid.: 31).

These polar AI functions have been analysed by several scholars who look at the AI's ability to create deepfakes, which undermine the authenticity of visual videos (Strickland 2018; Güera and Delp 2018). All the above demonstrates the symbiotic relationships between AI and disinformation, when despite the existing defensive function against bots and automation, one may abuse the inherent AI bias to spread incitement and manipulation within weakened populations.

Conclusion: enhancing AI ethics through intersectionality

According to the review of intersectionality research across disciplinary domains, positioning intersectionality as having an important role within AI ethics highlights two key areas of concern: advocating diversity and inclusion. First, advocating diversity requires that vulnerable and disadvantaged groups and outgroup members of various identity groups will gain a fair and just treatment by ensuring that AI bias will be moderated and diminished.

Second, promoting inclusion contributes to a more appropriate and just representation of disadvantaged individuals' involvement in designing and engineering AI and algorithmic systems. Within AI ethics, the principle of inclusion encourages remedying "situations where people are believed to have been silenced or excluded from decisions which would directly affect them and which do not acknowledge their knowledge or expertise" (Townsend 2013: 36).

Due to the vast phenomena of fake news propagation, there is an urgent need to provide a relevant ethical approach to dealing with the existing technical and

moral issues arising from AI technologies. Intersectionality serves as a proper and adaptable mechanism for coping with the AI-augmented falsification processes.

Notes

1 The Cambridge Analytica affair is a salient event in the disinformation sphere, in which a data analytics firm was accused of profiling 50 million Facebook users and targeting them with tailored content in order to influence political outcomes in the 2016 US presidential elections. For a more detailed discussion, see Rehman (2019).
2 In 2017, the publication of fake news imagery has aggravated the conflict between Rohingya Muslims and the military in Rakhine State in Myanmar. This false propaganda fueled the radical hatred of Rohingya Muslims and violence in Myanmar. This incident shows how Facebook has transformed into a platform for hate speech and online falsehoods aimed at vulnerable minorities. For greater detail, see Miles (2018) and Ratcliffe (2017).
3 This algorithmic incident is the result of a semantic loophole of Facebook's 'like' button. The anthropologists Rodrigo Ochigame and James Holston scrutinised an indigenous collective in Mato Grosso do Sul, which uses Facebook as its primary outreach vector to the public and often posts videos against the violence of private agribusiness militias. However, in the Portuguese dialect in Brazil, the button is semantically translated as 'enjoy', and therefore, various users decided not to 'like' certain posts of violence and oppression. As a result, Facebook's filtering algorithm reduced the group's visibility and raised the challenges of the land right activists to spread their message online. For further detail, see Ochigame and Holston (2016).

References

Amodei, D., C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané (2016) 'Concrete problems in AI safety', *arXiv preprint arXiv:1606.06565*.

Angwin, J., J. Larson, S. Mattu, L. Kirchner (2016) 'Machine bias', *ProPublica*, 23 May 2016, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (last consulted 6 May 2020).

Bartlett, R., A. Morse, R. Stanton, N. Wallace (2018) 'Consumer-lending discrimination in the era of fintech', unpublished working paper, Berkeley, University of California.

Bauerband, L.A., M.P. Galupo (2014) 'The gender identity reflection and rumination scale: Development and psychometric evaluation', *Journal of Counseling & Development* 92(2): 219–31.

Bloomberg Editorial Board (2017) 'Think the U.S. has a Facebook problem? Look to Asia', *Bloomberg*, 22 October 2017, https://bloom.bg/3diIXDq (last consulted 6 May 2020).

Bostrom, N. (2003) 'Ethical issues in advanced artificial intelligence', *Science Fiction and Philosophy: From Time Travel to Superintelligence*, 277–84.

Burkett, E. (2016) 'Women's movement', in *Encyclopædia Britannica*, www.britannica.com/topic/womens-movement (last consulted 6 May 2020).

Cadwalladr, C., E. Graham-Harrison (2018) 'Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach', *The Guardian*, 17 March 2018, www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election (last consulted 6 May 2020).

Charon, J.M. (2010) *Symbolic Interactionism: An Introduction, an Interpretation, an Integration*, Upper Saddle River: Prentice Hall.

Crenshaw, K. (1989) 'Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics', *University of Chicago Legal Forum*: 139–67.

Danks, D., A.J. London (2017) 'Algorithmic bias in autonomous systems', *IJCAI*:. 4691–97.

Dastin, J. (2018) 'Amazon scraps secret AI recruiting tool that showed bias against women', *Reuters,* 10 October 2018, www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (last consulted 6 May 2020).

Dhamoon, R.K. (2011) 'Considerations on mainstreaming intersectionality', *Political Research Quarterly* 64(1): 230–43.

Dill, B.T. (2002) 'Work at the intersections of race, gender, ethnicity, and other dimensions of difference in higher education', *Connections: Newsletter of the Consortium on Race, Gender, and Ethnicity*', 5–7, www.crge.umd.edu/publications/news.pdf (last consulted 6 May 2020).

Doshi-Velez, F., M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, S. Shieber, J. Waldo, D. Weinberger, A. Wood (2017) 'Accountability of AI under the law: The role of explanation', *arXiv preprint arXiv:1711.01134.*

Eidelson, J. (2018) 'Facebook tools are used to screen out older job seekers, lawsuit claims', *Bloomberg*, 29 May 2018, https://bloom.bg/3dm0avO (last consulted 6 May 2020).

Fellows, M.L., S. Razack (1998) 'The race to innocence: Confronting hierarchical relations among women', *Journal of Gender, Race and Justice* 1: 335–52.

Fotopoulou, A. (2012) 'Intersectionality queer studies and hybridity: Methodological frameworks for social research', *Journal of International Women's Studies* 13(2): 19–32.

Gecas, V. (1982) 'The self-concept', *Annual Review of Sociology* 8(1): 1–33.

Giddens, A. (1991) *Modernity and Self-Identity: Self and Society in the Late Modern Age*. Stanford, California: Stanford University Press.

Glantz, A., E. Martinez (2018) 'Kept out: How banks block people of color from homeownership', *Associated Press*, 15 February 2018, https://apnews.com/ae4b40a720b74ad8a9b0bfe65f7a9c29 (last consulted 4 May 2020).

Gomila, A., A. Amengual (2009) 'Moral emotions for autonomous agents', in J. Vallverdu, D. Casacuberta (eds), *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*, New York: IGI Global, 161–74.

Güera, D., E.J. Delp (2018) 'Deepfake video detection using recurrent neural networks', in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–6.

Hagerty, A., I. Rubinov (2019) 'Global AI ethics: A review of the social impacts and ethical implications of artificial intelligence', *arXiv preprint arXiv:1907.07892.*

Hashmi, A., R. Lalwani, A. Senatore, C. Perricos (2019) 'AI Ethics: The Next Big Thing in Government', Anticipating the impact of AI Ethics within the Public Sector. *Deloitte Global & World Government Summit.*

Hern, A. (2017) 'Facebook translates "good morning" into "attack them", leading to arrest', *The Guardian*, 24 October 2017, www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest (last consulted 6 May 2020).

Kertysova, K. (2018) 'Artificial intelligence and disinformation', *Security and Human Rights* 29(1–4): 55–81.

Loucks, J., T. Davenport, D. Schatsky (2018) 'State of AI in the enterprise', 2nd edition, *Deloitte Insights*, www2.deloitte.com/content/dam/insights/us/articles/4780_State-of-AI-in-the-enterprise/DI_State-of-AI-in-the-enterprise-2nd-ed.pdf.

McCall, L. (2005) 'The complexity of intersectionality', *Signs: Journal of Women in Culture and Society* 30(3): 1771–800.

McCarthy, J. (1960) *Programs With Common Sense*, RLE and MIT Computation Center, 300–07.

Marsden, C., T. Meyer (2019) 'Regulating disinformation with artificial intelligence: Effects of disinformation initiatives on freedom of expression and media pluralism', *European Parliamentary Research Service*, 1–72.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019) 'A survey on bias and fairness in machine learning', *arXiv preprint arXiv:1908.09635*.

Miles, T. (2018) 'U.N. investigators cite Facebook role in Myanmar crisis', *Reuters*, 12 March 2018, https://uk.reuters.com/article/us-myanmar-rohingya-facebook/u-n-investigators-cite-facebook-role-in-myanmar-crisis-idUKKCN1GO2PN (last consulted 13 June 2020).

Moore, M.R. (2012) 'Intersectionality and the study of black, sexual minority women', *Gender & Society* 26(1): 33–39.

Ochigame, R., J. Holston, (2016) 'Filtering dissent: Social media and land struggles in Brazil', *New Left Review* (99): 85–110.

Orr, C.M., A. Braithwaite (eds) (2012) *Rethinking Women's and Gender Studies*, London: Routledge.

Ratcliffe, R. (2017) 'Fake news images add fuel to fire in Myanmar, after more than 400 deaths', *The Guardian*, 5 September 2017, www.theguardian.com/global-development/2017/sep/05/fake-news-images-add-fuel-to-fire-in-myanmar-after-more-than-400-deaths (last consulted 13 June 2020).

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.F., Breazeal, C., Crandall, J.W., Christakis, N.A., Couzin, I.D., Jackson, M.O., et al. (2019) 'Machine behaviour', *Nature* 568(7753): 477.

Rehman, I. (2019) 'Facebook-Cambridge Analytica data harvesting: What you need to know', *Library Philosophy and Practice* (e-journal) 2497.

Remian, D. (2019) 'Augmenting education: Ethical considerations for incorporating artificial intelligence in education', *University of Massachusetts at Boston*, https://scholarworks.umb.edu/cgi/viewcontent.cgi?article=1054&context=instruction_capstone (last consulted 6 May 2020).

Reyes, P. (2017) 'Working life inequalities: Do we need intersectionality?', *Society, Health & Vulnerability* 8(1): 14–18.

Samoili, S., M.L. Cobo, E. Gomez, G. De Prato, F. Martinez-Plumed., B. Delipetrev (2020) *AI Watch. Defining Artificial Intelligence. Towards an Operational Definition and Taxonomy of Artificial Intelligence* (No. JRC118163). Joint Research Centre (Seville site).

Samuels, G.M., F. Ross-Sheriff, (2008) 'Identity, oppression and power: Feminisms and intersectionality theory', *Affilia* 23(5): 5–9.

Sculley, D., G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young (2014) 'Machine learning: The high interest credit card of technical debt', *SE4ML: Software Engineering for Machine Learning, NIPS'14*.

Sharma, K., F. Qian, H. Jiang, N. Ruchansky, M. Zhang, Y. Liu (2019) 'Combating fake news: A survey on identification and mitigation techniques', *ACM Transactions on Intelligent Systems and Technology (TIST)* 10(3): 1–42.

Strickland, E. (2018) 'AI-human partnerships tackle "fake news": Machine learning can get you only so far – then human judgment is required', *IEEE Spectrum*, 55(9): 12–13.

Suresh, H., J.V. Guttag (2019) 'A framework for understanding unintended consequences of machine learning', *arXiv preprint arXiv:1901.10002*.

Townsend, A.M. (2013) *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*, New York: W.W. Norton & Company.

Waddell, K. (2016) 'How algorithms can bring down minorities' credit scores', *The Atlantic*, 2 December 2016, www.theatlantic.com/technology/archive/2016/12/how-algorithms-can-bring-down-minorities-credit-scores/509333 (last consulted 4 May 2020).

Weise, E. (2017) 'Russian fake accounts showed posts to 126 Million Facebook users', *USA TODAY*.

Whittaker, M., K. Crawford, R. Dobbe, G. Fried, E. Kaziunas, V. Mathur, O. Schwartz (2018) *AI Now Report 2018*, New York University: AI Now Institute.

Whittlestone, J., R. Nyrup, A. Alexandrova, S. Cave (2019) 'The role and limits of principles in AI ethics: Towards a focus on tensions', Paper presented at the AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, https://doi.org/10.1145/3306618.3314289.

World Economic Forum (2018) 'How to prevent discriminatory outcomes in machine learning', *Global Future Council on Human Rights 2016–2018*, www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf (last consulted 4 May 2020).

Yang, K.C., O. Varol, C.A. Davis, E. Ferrara, A. Flammini, F. Menczer (2019) 'Arming the public with artificial intelligence to counter social bots', *Human Behavior and Emerging Technologies* 1(1): 48–61.

Yapo, A., J. Weiss (2018) 'Ethical implications of bias in machine learning', *Proceedings of the 51st Hawaii International Conference on System Sciences*, 5365–72.

Zimmerman, M. (2018) *Teaching AI: Exploring New Frontiers for Learning*, Portland, OR: International Society for Technology in Education.