Automated Misgendering

an Inquiry Into the Ethics of Automatic Gender Recognition

Photo by Quino Al on Unsplash

Marit Eva Hoefsloot
September 2021

This page is intentionally left blank

Automated Misgendering

an Inquiry Into the Ethics of
Automatic Gender Recognition

Marit Eva Hoefsloot

29-09-2021

Supervisor: Dr Maren Behrensen
Second reader: Dr Patrick Taylor Smith

Master thesis for the completion of MSc Philosophy of Science, Technology, and Society
Faculty of Behavioural, Management, and Social Sciences
University of Twente
Enschede, the Netherlands

Table of Contents

Acknowledgements

At the beginning of this thesis process, I was asked to prepare a pitch with my thesis's main argument and societal relevance. Our teachers told us that having a pitch prepared would help face our friends and families over the next couple of months. They were not wrong: every time I saw my friends, family, and even colleagues (which was not often due to the COVID-19 pandemic), they asked me what my thesis was about. Explaining the societal relevance was not a problem for me as I felt strongly about the importance of the topic. However, I struggled a bit more with the scope of the argument. As my thesis progressed, the argument took different paths: from political philosophy and ethics to metaphysics and epistemology – and back to ethics. Only after I handed in my second full draft was I able to pitch my thesis to my mom succinctly and persuasively. This was the moment I knew I was close to finishing. However, I could not have done this alone, so I wish to take a moment to thank the people who helped make this thesis possible.

First and foremost, I would like to thank my supervisor Maren Behrensen for the engaging lessons in feminist gender theory, passing on Sally Haslanger's wisdom, and the thoughtful manner of giving feedback. Receiving feedback has always been a challenge for me, but Maren was very understanding and helped me find a way in which the feedback would come across as constructive. Relatedly, I want to thank my second reader Patrick Taylor Smith for the sparring sessions and for pushing me to establish a solid ethical grounding in my argument. Moreover, Patrick looked out for you, the reader, and made sure I anticipated your objections and responded to them – to make you feel 'accommodated,' as he liked to say.

Next, I would like to thank my parents Dorien Brunt and Lex Hoefsloot, and my sisters Ellen and Fenna Hoefsloot. Through walks and talks, they tested my pitching skills. Every time we went for a long walk in the Dutch dunes, I came home with a new approach, case study, or even a fully fleshed-out argument. I would also like to extend my gratitude to Ana Sánchez, Annelie Oortwijn, Roos Kruimer, and Sarah Stapel for their daily support, laughs, and "office" days. They helped me by talking about my thesis when I was stuck, and especially by *not* talking about it when I needed it. Last but not least, I want to thank my fellow PSTS'ers John Walker, Eliana Bergamin, Jerry Wenzel, Luuk Stellinga, and Kristy Claassen. They have helped shape this thesis through relentless discussions and pushed me to work harder and reach higher. I want to thank Eliana specifically for the hype-up voice messages full of guidance and advice – from the very beginning to the last end.

Abstract

This thesis investigates the ethics of the use of automatic gender recognition technologies (AGR). AGR tools are a type of facial analysis technology used to identify and verify someone's gender identity. The research question for this thesis is: Is the use of AGR technologies in their most accurate and fair form ethical? Moreover, how can the negative impact of AGR's use be alleviated pragmatically? To answer these questions, I will research how and for what purpose the technology is built, investigate the underlying assumptions regarding gender, and discuss the harmful consequences of the (mis-)use of AGR technologies. The first chapter will show that AGR technology is built on the assumptions that gender can be externally determined based on one's appearance, that gender is a binary categorisation, and it is stable over time. In the second chapter, I argue that these assumptions are incorrect, as self-identification is central to gender identity and people must have first-person authority and autonomy over their gender and their gender alone (Bettcher, 2009). This claim is not an epistemic one but a moral one; while you might disagree with someone's gender identity, you are obliged to respect their authority on the matter. Finally, in the third chapter, I argue for the immorality of the use of AGR technologies. It fundamentally disrespects the authority and agency people have over their gender.

As the developers and users of AGR are set to make significant financial and time efficiency gains due to its broad implementation (O'Neill, 2021), it is not realistic to aim to abolish the technology. Thus, I argue that there is another option. Through the education of AGR's users concerning the nature of the results, the harmful consequences of using AGR tools can be minimised. This education involves creating a new vocabulary to reflect that the results from AGR technologies are merely predictions with probabilities, which can be either accurate or inaccurate. I call the results created by AGR tools *probabilistic information* and the information given by the subjects themselves *agential information* to reflect Dembroff and Saint-Croix's agential identity (2019). Through such vocabulary, we become more aware of the authority behind the pieces of information. As the agential information must be respected, the new vocabulary can help resolve conflicts in situations where the AGR result does not correspond to someone's self-identified gender.

Keywords: automatic gender recognition, gender identity, agential identity, first-person authority, ameliorative analysis

Introduction

Over the last decade, many large tech companies have developed facial analysis (FA) technologies such as facial detection and recognition software (Scheuerman et al., 2019). Especially tech giants like Amazon, Google, IBM, and Microsoft have produced and commercialised their facial analysis software. Essentially, FA software aims to identify an individual based on a picture or video of their face. While these tools can be deployed for a plethora of cases, this thesis focusses on the use of such software for the classification of people's gender. This is known as automatic gender recognition (AGR) software (Keyes, 2018).

AGR tools are deployed in cases ranging from demographic research to security and surveillance and can even be used to personalise services and games (Lin et al., 2016). Moreover, AGR technology is currently being used in research and development concerning human-robot interaction (Ramey & Salichs, 2014) and human-computer interaction on social media (Keyes, 2018). The goal of AGR is to translate the human ability of gendering into code. This would facilitate human-computer interaction, as robots and computers supposedly need to be capable of classifying someone's gender to interact with them (Anusha et al., 2016; Khan et al., 2019; Ramey & Salichs, 2014).

Currently, automatic gender recognition software often misgenders people - especially people who do not fall within the gender binary, such as agender and trans folk (Buolamwini et al., 2018; Scheuerman et al., 2019). This disproportionate misgendering is also visible in the accuracy rates of the AGR tools that are commercially available. In the case of IBM's Watson tool, the false positive rate (FPR) for women is more than twice as high as the FPR for men (Buolamwini & Gebru, 2018, p.9). This is to say that subjects are more likely to be incorrectly recognised as women than men – while they do not identify as the respective gender. This difference becomes more prominent when the subject group is split up based on skin colour. The FPR for darker-skinned women was 25.2, where the FPR for lighter-skinned males was merely 0.4 with the IBM tool (Buolamwini & Gebru, 2018, p.9). While various possible causes for these biases exist, training data bias is most common (Danks & London, 2017). When an algorithm is built using images of primarily hypermasculine white men and hyperfeminine white women, it will be very accurate in recognising the genders of these people. However, this means the algorithm will be significantly less accurate in recognising the genders of people who do not fit this image, for example, genderqueer black people.

This misgendering can create harmful and dangerous situations, specifically for people who are already in vulnerable positions. In their influential paper, *The Misgendering Machines*, Os Keyes investigates how AGR tools operationalise the concept of gender identity and the consequences of the widespread use of AGR tools (2018). They argue that inaccurate AGR tools work to erase the lives and experiences of trans people and suggest ways to create AGR tools based on a more nuanced operationalisation of gender.

While I agree that inaccurate AGR tools erase the lives and experiences of trans people, I argue that even the use of more accurate AGR tools – e.g., with a more nuanced definition of gender – is unethical. The technology cannot be improved in such a way that justifies its use, as the act of gendering, whether by an external person or machine, disrespects the subject's agency and authority regarding their gender. Unfortunately, the technology already exists and is used in situations from airport security to beverage recommendations (Ng et al., 2019; Watts, 2019). Moreover, the developers and users of AGR are set to make significant financial and time efficiency gains due to its broad implementation opportunities. These are significant incentives to continue the use of AGR technology, regardless of whether it is ethical or not. It is, therefore, unrealistic to abolish the technology altogether.

Thus, I will recommend a solution that can minimise the harm done by misgendering using AGR technologies. The central tenet of this solution is based on educating the users of AGR technology concerning the nature of the results. Namely, it is essential to remember that algorithms such as AGR tools create *predictions* of someone's gender which can be inaccurate. A helpful tool for such awareness is using new vocabulary for the AGR results and the subject's self-identified gender to reflect the authorities behind the gendered statements. I call the results created by AGR tools *probabilistic information* and the information given by the subjects themselves *agential information,* in line with the subject's agency regarding their gender.

Through such vocabulary, we become more aware of the authority behind the pieces of information. This helps resolve conflicts when the AGR result does not correspond to someone's self-identified gender. The agential information is legitimate due to the subject's agency and authority regarding their gender. Suppose the developers and users of AGR technologies understand that the results are merely suggestive rather than the ultimate truth. In that case, the technology can still be used to automate otherwise time-consuming tasks such as filtering through large groups without causing further harm to marginalised people due to their gender.

Theoretical Framework

In Sally Haslanger's approach to the question of "What is Gender?", she considers why we need the term gender at all and what work we want it to do for us (2012, p.224). Haslanger is a pragmatist because she is not trying to find a universal answer to the question but rather an answer that solves the specific issue at hand. She believes that the question "what is gender?" might bring about different answers depending on the context. Whether a particular answer is a good fit for the situation depends on the political consequences it brings about and, to a lesser extent, the semantic and pragmatic impacts. Finally, Haslanger argues that there is often a discrepancy between the way we use terms in practice and the official, socially accepted meaning we give those terms. She believes that the meaning of the terms we use can be re-engineered to account more fully for what we intend them to mean. "My priority […] is not to capture what we do mean, but how we might usefully revise what we mean for certain theoretical and political purposes" (Haslanger, 2012, p.224). I will use Haslanger's framework of manifest, operative, and target concepts to critique the developers of AGR for their incorrect understanding and misuse of gender.

The manifest concept is the formal, socially agreed-upon definition, often found in dictionaries. It is the explicit and intuitive meaning of a word. The operative concept is then the colloquially used definition; it is the implicit and practised meaning. For example, the manifest concept of 'a woman' can be seen as the attribute "womanhood" or a "Woman's Nature" (Haslanger, 2012, p.93). However, according to Haslanger, this Woman's Nature is an illusion; and what is really at stake is the way women are seen by men. Female humans become 'women' through the sexual responses they receive from men. That is the operative concept. Finally, the target concept is the definition we ought to be using. The target concept is determined based on aspects particular to the situation, such as fairness or inclusivity. According to Haslanger, the target concept of "a woman" is a combination of physical characteristics and social position in the hierarchical society (2012, chapter 7). The manifest, operative, and target concepts do not necessarily coincide as people can have different definitions and understandings of words. However, when someone's operative concept inflicts harm on another person or excludes them from a specific group, we have reason to argue that their operative concept must be re-engineered to coincide with the target concept.

Haslanger's method of finding a target concept of gender is known as ameliorative analysis (Haslanger, 2012, p.367). This means it is not an epistemic theory but a moral one. She does not merely aim to describe what concept of gender she currently recognises in society; she aims to improve the concept of gender and our understanding of it. A faulty concept of gender can cause harm to people that are excluded from the definition, marginalised in society, and oppressed due to their gender. Haslanger argues that to minimise this harm, we must revise our manifest and operative concepts to coincide with the target definition. This theory can be seen as a theory of justice, as it describes how moral harm is caused due to gender and comes with a solution to alleviate this harm. While it is not uncommon for people to have different definitions for specific terms, these concepts can potentially inflict harm through exclusion, oppression, and marginalisation. To correct these harms, we need target concepts, as they help to create justice, harmony, and agreement regarding these concepts.

The normative framework of manifest, operative, and target concepts will form the scaffolding of this thesis, as it focuses on the different ways in which terms are defined and used. However, I will not be filling this scaffolding with Haslanger's conception of gender for two reasons. Firstly, Haslanger initially confined her theory to the gender binary; she writes in terms of man/woman and male/female. This is exclusionary to other people's experiences and reductionist of the concept of gender. Gender is more dynamic and fluid than how Haslanger initially portrayed it. Later, she adapted her theory to make it more inclusive and argued that it is for everyone who is "against the binary construction of men and women" (Haslanger, 2020, p.232). Importantly, Haslanger states multiple times that she is not trying to completely define what gender is, and is instead looking for what we want gender to be and what we need the concepts of gender for. Secondly, Haslanger's conception of gender focuses solely on the societal level, whereas gender also operates on the personal and interpersonal levels. In the second chapter, I will go deeper into the discussion of Haslanger's concept of gender and why it is unsuitable. I will consider other theories of gender to find a fitting target concept of gender for this project. In the next section, I will briefly start the discussion on the concepts of sex and gender, as these will form the basis for the following chapters.

Defining Terms: Sex and Gender

While gender and gender identities are woven in most aspects of our daily lives, the definitions of these concepts are not always clear or socially agreed-upon. In the previous section, I shortly introduced Sally Haslanger's manifest, operative, and target concepts of gender (2012). This section will further explain the difference between how the term is formally defined and colloquially used. The manifest concept is a commonly agreed-upon definition of a term; this is the formal definition as it can be found in dictionaries or legal documentation. However, this is not always the definition we deploy in daily use. The operative concept is our subjective understanding of a definition, which is used in informal settings. This concept describes a definition that is always in the back of our minds, even when we try to adhere to the manifest concept. Moreover, neither the operative nor the manifest concepts may describe the target definition we aim to describe or use. The target concept of gender is a normative proposal for the definition and use of gender. It is the concept that we *ought* to be deploying, not necessarily the concept we *are* deploying.

One example of a way in which terms are misused or misdefined is the terms sex and gender. While many feminist philosophers and sociologists have argued for the distinction between these two terms since the 1970s, their meaning is often misunderstood. In feminist philosophy from the '70s, the difference between sex and gender was described as "gender is the social meaning of sex" (Haslanger, 2012, p.184). The idea was to describe sex as a fixed and determined physical characteristic and gender as its malleable social counterpart. Sally Haslanger resists this idea as sex and gender do not need to coincide: some people are assigned a certain sex but identify as another gender. Haslanger points out that people can even be gendered as a specific gender without identifying as that gender at every moment in their life (2012). Gender is merely a set of social norms which can be imposed on anyone, regardless of their self-identification.

While Haslanger focuses on the gender group of women, I prefer to use gender-neutral language as 'women' is not the only gender group that experiences oppression due to their gender. Other gender groups include transgender, nonbinary, gender-fluid. It is also possible to not identify with any gender; that is known as agender. To briefly clarify, someone is transgender when they do not identify with the sex category or gender assigned to them at birth (Bettcher, 2017). Whereas transgender is sometimes perceived as fitting the binary division of men and women, the non-binary, agender, and gender-fluid genders strictly reject that binary. These genders are not the same for everyone who identifies as them; a non-binary gender can mean the absence of gender for one person and a combination of 'woman' and ‘man’ for the next. The gender spectrum has been introduced to signify a vast number of genders and combinations that can apply to people. Someone’s gender identity can be anywhere on the spectrum and might change over time.

Nevertheless, the gender binary is still the dominant categorisation, meaning that the people who fit in that binary are in power. Gender relations are power relations (Radtke & Stam, 1994); women are seen and treated as subordinates to men in the private, public, and political spheres. This informs everything from the social norms that people are expected to uphold to how much they are getting paid and whether their voice is heard in political debates. Anyone who challenges this power becomes marginalised by the dominant group to keep the power. While this is visible in the case of men as the oppressor and women as the oppressed, it is even more pressing for people who do not fit in the gender binary altogether. When these people challenge the power of the dominant groups – in this case, both the men and women who fit in the gender binary – they challenge the entire hierarchical power structure, not merely their position in the hierarchy.

While the term sex can be understood as your physical characteristics and gender as your social identity, the distinction between gender and sex is often misunderstood. This becomes clear when authors use the terms as synonyms, as will be shown in the first chapter. For context, Talia Mae Bettcher investigated why the terms sex and gender are still used interchangeably; she argued that the terms track the same, namely genitalia (2009). Bettcher argues that genitalia are still seen as the ‘truth’ or ‘reality.’ She explains this in the context of transwomen, who are often asked whether or not they have undergone sex reassignment surgery. This question is based on the assumption that having a penis or not defines the gender of an individual. This shows that the meanings of sex and gender are fixed, and both depend on genitalia (Bettcher, 2009). Moreover, Bettcher argues that “gender presentation literally *signifies* physical sex” (2009, p.105). As both gender and sex track the presence of specific genitalia, there is a connection between gender and sex in which sex is the physical characteristic and gender is the corresponding social and behavioural presentation. In the case of trans people, their biological sex and gender presentation misalign, so they are seen as doing gender ‘wrong.’

In summary, the definitions of sex and gender are not uncontroversial, and they are the focal point of an ongoing debate. As mentioned above, the manifest concept of gender is its socially agreed-upon definition. However, it becomes clear that there is not one socially agreed-upon definition of gender. Whereas Haslanger sees sex as a physical characteristic and gender as a set of social norms, Bettcher argues that sex and gender are both used to track someone’s genitalia. According to Bettcher, gender is only used to refer to someone’s genitalia rather than the social norms of sex. This also implies a social norm that requires people from a particular gender to have genitalia that are generally imagined to correspond to said gender. To be clear, this is not an argument for Bettcher’s account of gender or against Haslanger’s account. Instead, this section showed that it is essential to remember that these terms can mean different things, depending on the context and the situation. Throughout this thesis, I will investigate what definitions of gender are being deployed, what is socially agreed upon, and what definition we ought to use. The political consequences of the different definitions will play a significant role in deciding which definitions are the right fit for this project.

Structure of this Project

The research questions for this thesis are: Is the use of AGR technologies in their most accurate and fair form ethical? Moreover, how can the negative impact of AGR's use be alleviated pragmatically? To answer these questions, I will first review the current state-of-the-art regarding the development of AGR technologies and the developers' language to describe the relationship between the technology and gender. In the second chapter, I will evaluate the philosophical writing on the nature of gender to create a baseline to investigate whether or not AGR can do what it claims it does: recognising an individual's gender properly. This will also spark a discussion regarding the harms of misgendering through the use of AGR technologies. I will then define a target concept of gender that covers all aspects of gender and gender identity that we encounter in everyday life. I take a social constructionist approach, which entails seeing gender groups as social categories rather than natural divisions (Haslanger, 2006). In the third chapter, I will combine and compare the results from the first two chapters and investigate whether it would be possible to create an AGR tool that describes the target concept of gender. I will show that this is not possible, which gives me grounds to argue that the use of AGR is unethical. However, as the technology is already in use and benefits large stakeholders (O'Neill, 2021), it seems unrealistic to abolish the technology altogether. Instead, I will end this thesis with a recommendation on how to minimise the harmful consequences of AGR.

The framework that brings these three chapters together is Sally Haslanger's manifest, operative, and target concepts. Namely, the first chapter describes which manifest and operative concepts can be identified in the academic writing concerning the AGR developments. The second chapter then defines the target definition of what gender is and what AGR tools should aim to recognise. The third chapter investigates whether the manifest and operative concepts of AGR can be re-engineered to describe the target concept. I will show that this is impossible and that we must look for other ways to prevent the harmful misgendering that can arise from the use of AGR technology.

Chapter 1: The State-of-the-Art of AGR Software

In the introduction, I introduced Sally Haslanger's framework of manifest, operative, and target concepts. This chapter will focus on the manifest and operative concepts used by the developers of automatic gender recognition technology. I will critically examine the developers' choices and assumptions regarding gender that are visible in the papers that accompany their new algorithms and methodologies. Moreover, this chapter will introduce the working of automatic facial recognition software and present the main advantages and challenges of AGR.

This chapter aims to show how the developers that work on AGR technologies have a limited view of gender and base their algorithms on their assumptions regarding the nature of gender. To do so, I have chosen to focus on the papers written by academics in computer science. However, it is essential to note that academics with a background in gender studies are working on more inclusive and fairer AGR tools. Examples of such authors are Foad Hamidi, Morgan Klaus Scheuerman, Os Keyes, Timnit Gebru, Joy Buolamwini, Jed R. Brubaker, Jacob M. Paul, and their colleagues (e.g., Buolamwini & Gebru, 2018; Hamidi et al., 2018; Keyes, 2018; Scheuerman et al. 2019). These pioneering academics have brought disproportionately inaccurate AGR technologies to the light and have advocated for more inclusive alternatives. Nevertheless, as they are a minority in the world of AGR developments, I want to use this chapter to demonstrate how the average computer engineer that sees gender recognition as a purely computational challenge writes about gender. Thus, this chapter will not feature the work of the gender academics and focus solely on the work of AGR developers.

The literature discussed in this chapter is written and presented by the developers of AGR technology at renowned conferences such as the IEEE International Conference on Multimedia & Expo Workshops (Santarcangelo et al., 2015) and the 2014 ACM/IEEE International Conference on Human-Robot Interaction (Ramey & Salichs, 2014). These papers are all peer-reviewed and were published in a wide range of publications, from highly esteemed conferences to the lower end of computer science journals. It accurately represents the state-of-the-art automatic gender recognition as offered by the average AGR developer.

This chapter will first define the essential terms in facial analysis and outline what assumptions regarding the nature of gender are present in the papers. The sections after that will dive deeper into the functioning of AGR technologies and the main challenges that persist in the development of AGR tools. This will then inform the discussion on the use and value of AGR technologies presented by the developers. The last section of this chapter will summarise the main takeaways and place these into Haslanger's framework of manifest, operative, and target concepts.

Definitions and Assumptions

In the computer science field of facial analysis, there are a couple of terms that are closely related but separate tasks, which sometimes seem to be convoluted or used interchangeably. These terms include face detection, face recognition, and face classification. Figure 1 gives a helpful overview of how these different tasks relate to each other and how they differ (Scheuerman et al., 2020, p.6).

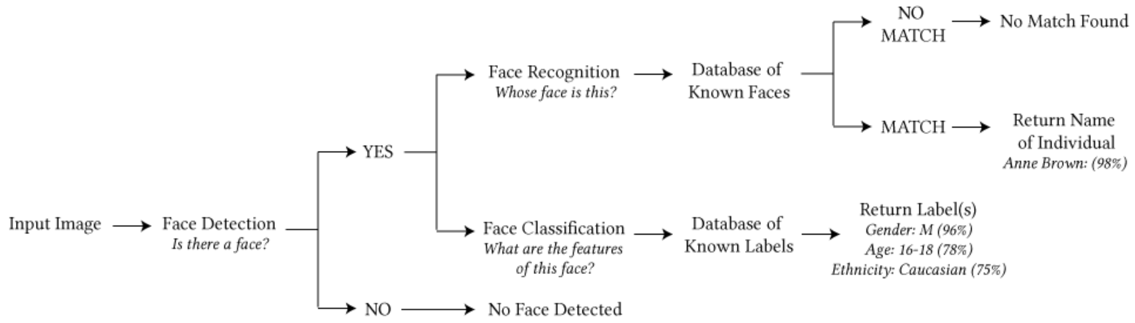

Figure 1: the main tasks in facial analysis (Scheuerman et al., 2020, p.6)

As is clear from figure 1, the face detection task first recognises whether a face is present in the input image. When a face is detected, it can either be classified or recognised. The recognition task compares the input image with a database of known faces and aims to identify who this person is. An example of such a task is police surveillance on the street, where a real-time picture of your face is compared with a database of faces of wanted persons. The classification task focuses more on the different characteristics that can be found in the individual's face, for example, the race and gender of the person (Scheuerman et al., 2020). I will mainly focus on face classification and its use for gender classification.

Interestingly, most papers reviewed in this chapter provided an overview of established gender classification methods and even "how researchers previously approached gender classification" (Khan et al., 2019, p.2). However, not a single article started by explaining how the authors understood the term 'gender.' A discussion of what gender is and how they deploy it is not deemed necessary in explaining the researcher's approach to gender classification. The authors of these papers were never challenged to think about their definitions of sex and gender.

This is also visible in the way the terms' gender' and 'sex' are used interchangeably in the paper by Mahalingam and Ricanek (2013). First, they start by stating how "the face conveys identity, lineage, sex, race, ethnicity, mood, feelings" (Mahalingam & Ricanek, 2013, p.1). However, as this paper is about facial recognition during gender transition, they refer to their subject's gender rather than their sex. As they did not mention 'gender' separately in the list of characteristics that can be inferred from someone's face, I assume they are defining both terms to denote the same thing.

The topic of gender transition is a considerable challenge in facial analysis research. However, some researchers working on creating robust facial analysis tools over gender transition are mainly interested in the computational challenge rather than ensuring the inclusion and accuracy for transgender people. This is reflected in the way the authors write about transgender people and gender transition;

"This is based on the intuition that the appearance factor captures the features that are unaffected by the therapy and transition factor captures the feature changes due to therapy" (Kumar et al., 2016, p.1). The researchers follow their 'intuition' rather than conducting a study to see whether this is the case.

Moreover, Vijayan et al. write, "[a] male becomes female or a female become a male by altering the balance of oestrogen and testosterone hormones" (2016, p.1367). They have reduced the process of gender transition and the lived experience of transgender people to merely hormone therapy. This is not the only occasion where the definitions Vijayan et al. use to describe sex and gender leave much desired. In a near-identical sentence as the aforementioned quote by Mahalingam & Ricanek, Vijayan et al. state that the "[f]ace conveys different *feelings* [emphasis added] of a person like lineage, identity, sex, ethnicity, race, feelings, mood etc." (2016, p.1367). There are three problematic notions in this sentence alone. First, personal characteristics such as lineage and ethnicity are mentioned in the same breath as feelings and mood. Second, while ethnicity and race are separately mentioned, they only address sex and not gender. Third, all the mentioned characteristics were summed up under the umbrella term 'feelings', which is inaccurate.

Another point Vijayan et al. make is that "[g]ender transformation can be considered a variant of face disguise, however, disguise falls under the broader category of biometric obfuscation, which refers to the deliberate alteration of the face for the purpose of masking one's identity. Transgender persons undergo HRT for the purpose of masking or creating a new identity" (Vijayan et al., 2016, p.1368). While the attempt to include transgender people in facial analysis technologies is admirable, these researchers lack the sensitivity and awareness desired when working with issues such as gender. While this paper was published in the discontinued self-publishing journal Procedia Technology, it was also presented and peer-reviewed at the International Conference on Emerging Trends in Engineering, Science and Technology 2015. With such outreach and peer-review, I expected the authors to approach the subject with more sensitivity and professionality.

Behind the Scenes

While many different companies and research groups are developing their gender classification tools, they all deploy similar techniques. The earlier versions of facial analysis tools were generally built with appearance-based methods (Anusha et al., 2016; Khan et al., 2019; Kumar et al., 2016). These methods first extract the patches of the photo that contain the periocular, mouth, and nose regions of the subjects, which are then classified into – frequently limited to the binary – genders, using a classification tool. The most used tools include support vector machines (SVMs), neural networks, decision trees, and AdaBoost (Anusha et al., 2016; Khan et al., 2019). Interestingly, SVMs are only valid for two-class problems (Santarcangelo et al., 2015), which means that gender is approached as a binary issue. Similarly, Geetha et al. deployed a support vector machine with a linear kernel method for "this binary classification" between males and females (2019, p.2530). A drawback of the patch-based methods is that they require extra computational steps, as the images of the faces first need to be split up into patches.

An example of an appearance-based method is that of Anusha et al. (2016). They analyse the difference in brightness between someone's skin and their lips. According to Anusha et al., the contrast between the brightness of the skin and lips is more significant for female subjects than for male subjects, meaning that it can be used as a discriminatory facial feature (2016). However, this is a questionable distinction, as this is based on the assumption that "[f]emale skin is generally lighter than male skin. But, female eyes and lips are not lighter than male eyes and lips" (Anusha et al., 2016, p.202). They did not provide any sources or references for these statements, so this seems to be an ungrounded assumption.

Appearance-based methods use a wide range of facial features to determine their subject's gender. Khan et al. deployed a list of facial features for their appearance-based approach (2019). These range from the large size of male foreheads and noses to the curl of female eyelashes and the shape and grooming of one's eyebrows (Khan et al., 2019, p.6). Geetha et al. also set out to create a more robust feature-extraction algorithm by looking at different features, depending on the outcome of the respective features (2019). If a particular feature did not give a conclusive result for a specific face, the algorithm would look at another feature. Vijayan et al. state that the left eye region is a more vital determinant of gender than the right eye region (2016, p.1372). It is unclear whether this is specific to their particular dataset or this conclusion can be extrapolated to other datasets.

For appearance-based methods to work well and the patches to be recognisable, the photometrics need to be of high quality (Kumar et al., 2016). This is to say that the quality of the image and the lighting must be of sufficient quality to determine someone's gender. Moreover, these methods generally ignore that facial features can change over time through ageing or gender transition. In response, Kumar et al. proposed a new method that utilises the patched appearance-based approach but considers a transition factor that comes about by hormone replacement therapy (HRT) (2016).

A slightly different method is based on a geometric approach (Khan et al., 2019). The face is still divided into patches, but rather than classifying the individual patches, the algorithm measures and compares the distance between the different facial features. The ratios between the distance from the subject's facial features such as ears, eyes, mouth, and nose are collected and compared with other subjects to determine their gender. Features such as skin texture and brightness are disregarded in these methods.

Many of the algorithms reviewed here were built using the same building blocks. The Viola and Jones face detection algorithm (2004) is often used in the very first stage of the facial analysis pipeline (see figure 1). Namely, this algorithm can be deployed to detect a subject's face and subsequently filter out the unnecessary parts of the photo, such as the background, hair, and clothing. According to Vijayan et al., it "provides robust and extremely rapid object detection" in real-time (2016, p.1370). Second, the classification algorithms are often built using adaptive boosting (also known as AdaBoost). This allows for constructing a robust classifier using several weaker classifiers, such as SVMs (Santarcangelo et al., 2015). Table 1 gives an overview of the discussion of the Viola and Jones face detection algorithm and AdaBoost, per article. Note, I did not include the papers by Ng et al. (2019) and Khan et al. (2014), as they do not propose novel AGR methods but are comparative analyses of the current state of AGR research and development.

Article	Discuss Viola & Jones?	Discuss AdaBoost?
(Anusha et al., 2016)	Yes	No
(Geetha et al., 2019)	No	No
(Khan et al., 2019)	Yes	Yes
(Kumar et al., 2016)	Yes	No
(Mahalingam & Ricanek, 2013)	No	No
(Ramey & Salichs, 2014)	No	No
(Santarcangelo et al., 2015)	Yes	Yes
(Vijayan et al., 2016)	Yes	Yes

Table 1: The discussion of the Viola-Jones and the Adaboost methods per reviewed article

Interestingly, Ramey and Salichs investigate a subject's body instead of their face and base their gender classification on the shape of the subject's breasts (2014). They justify this by stating that "[c]onsidering the shape of the breast of a person is a natural way of estimating his or her gender" (Ramey & Salichs, 2014, p.272). They have selected their set of subjects based on their body shapes to represent people's different types of bodies. However, describing the range of body shapes they have included, they state that their dataset includes "slender females and overweight males" (Ramey & Salichs, 2014, p.273). There are normative values hidden in this type of language: females' chest areas are supposed to be large, and males' chests are supposed to be flat. This is also represented in the following quote:

> "Note that an algorithm of gender detection based on the body shape cannot intrinsically reach a 100% accuracy, as there is no strict categorisation between female and male shapes. Overweight male subjects for instance appear more curvy than some women. This limitation can be overcome by coupling this gender recognition algorithm with others, based for instance on the face of the users or their height. Yet, our algorithm correctly estimates the gender of the user in almost nine times out of ten."
>
> (Ramey & Salichs, 2014, p.273).

It seems like Ramey and Salichs have some awareness of the arbitrariness of determining a person's gender based on their body shape, as there are no clear distinctions between the body shape of a man or a woman. However, they discard this awareness by stating that this can be overcome by adding other metrics to the equation and that their automatic gender recognition tool is accurate *enough*. I italicised 'enough' here because the way computer scientists use accuracy as a metric is questionable in itself. Frequently, if an algorithm can get the 'right' answer in the majority of the cases, the algorithm is seen as usable. However, with issues like gender classification – and racial classification, for that matter – not all cases weigh the same. People who are gender non-conforming are at a higher risk of being misgendered due to AGR tools. Suppose an algorithm is structurally worse at correctly recognising socially and culturally marginalised people, as opposed to the dominant groups in society. In that case, the algorithm's accuracy based on the entire population is not enough to measure the useability of said algorithm. The accuracy must be sufficient across all social groups, not merely on the entire population.

The move to use breast analysis to classify someone's gender is problematic, as it reinforces the norms and values that surround body shapes. It is effectively automating sexual harassment as it objectifies and analyses the subject's secondary sex characteristics. Additionally, it opens up the possibility to use other physical markers as gender classification subjects. As mentioned in the quote above, Ramey and Salichs believe that gender classification can become more accurate if it were to combine faces, breasts, and height in the analysis. The more physical markers are used, the more accurate the gender classification will become. However, where is the limit? Will genitalia, internal organs or even chromosomes become the subject of gender analysis?

While this might sound purely speculative, here is a historical precedent for such cases. As this is already done in sports classifications and gender assignment at birth, could it be extended to all aspects of daily life, such as airport security controls? Someone's presumed genitalia are already used as the decisive matter in the bathroom debate (Schilt & Westbrook, 2015), but only against people whose gender identity might not coincide with the standard image society has of people with that gender. Alternatively, think of the story of the record-breaking South African middle-distance runner Caster Semenya. As she is an outstanding female runner, her success was investigated, and her womanhood was questioned. A series of invasive examinations revealed that Semenya has XY-chromosomes and raised testosterone levels (Court of Arbitration for Sport, 2019). This led to a prohibition from competing against other women. The question becomes, could using breasts as the focal point of gender classification create a slippery slope for gender classification based on more invasive techniques in all aspects of daily life?

An objection to this worry is the critical distinction between gendering based on outward appearances and genitalia or chromosomes. Namely, our outward appearances are public in the sense that everyone can perceive them anytime we move in public. Moreover, clothing styles are subject to presentation and are sometimes used to express one's gender explicitly. On the other hand, genitalia and chromosomes are not public, as they cannot be perceived without close investigation. Thus, there is a normative difference between these techniques of gendering. However, I want to emphasise here that the case of Caster Semenya shows that the line between what is public and what is invasive might be blurrier than we think. Because of her athletic talent, her gender was closely examined, and the results were publicly announced. It is, thus, not impossible to imagine a world in which such investigations would become normalised.

The last point that I would like to highlight is that the algorithms discussed in this section are either based on predefined datasets or on datasets that have been created for these studies. While the labelling process for the predefined datasets is unknown, Khan et al. labelled their photos manually (2019). Whereas they state that they have labelled the images for "each gender" (2019, p.6), they only use 'male' and 'female' to describe their subjects. While the labels and the labelling of the training data set are integral parts of creating an algorithm, there is limited information available about the process and the relevant choices. Moreover, the available information indicates that the labellers are using their intuition rather than a set of predefined metrics or rules.

The Main Challenges

While the last decade has seen a lot of research and innovation in facial recognition technologies regarding gender, there remain a couple of challenges that various authors mention. In general, a significant problem that inhibits facial recognition is that the context and quality of the photographic and video material vary significantly (Khan et al., 2019; Kumar et al., 2016; Ng et al., 2019; Santarcangelo et al., 2015). The varying posing, lighting, and pixels make it hard for computer systems to recognise faces. Consequently, some studies build their algorithms using datasets in which this variability is minimised. Such a dataset can, for example, be assembled out of photographs of US members of congress (Schwemmer et al., 2020). As these types of datasets ensure that the context, background, lighting, and subjects' pose are identical, they are seen as 'constrained' datasets (Santarcangelo et al., 2015). It is important to note that these photos are staged and that the difference between men and women can be enhanced through clothing and make-up (Ng et al., 2019). Thus, these photos are polished representations of people and genders; the men are hyper-masculinised, and the women are hyper-feminised. 'Unconstrained' datasets are the datasets in which the posing, lighting, context, and quality vary. These photos are more realistic representations of people and thus form much more of a challenge for facial analysis technologies (Khan et al., 2019).

According to Santarcangelo et al., the main things that make facial recognition in unconstrained datasets challenging are variations in the number of people in one photo, fluctuations in their respective age groups, and different ethnicities among the subjects (Santarcangelo et al., 2015, p.5). This is also in line with the conclusions Ng et al. draw in their comparative analysis of AGR; human characteristics such as age and gender as well as facial expressions can all confound the algorithm (2019). Moreover, algorithms built on the training data with subjects from various ethnicities and used to predict the gender of people from one specific ethnicity are less accurate than algorithms built and used for people of the same ethnicity (Ng et al., 2019). Interestingly, these papers do not mention the multiplicity of gender being a confounding factor.

On the other hand, according to Kumar et al., an "upcoming problem" brings about novel challenges for facial recognition technology: gender transition (2016, p.2). Similarly, Mahalingam and Ricanek call facial recognition for transgender subjects a "truly novel and extremely unique biometric problem" (2013, p.1). This is since gender transition using hormone replacement therapy changes the texture and form of an individual's face. This makes it difficult for the current state-of-the-art facial analysis technologies to accurately recognise a subject's face during HRT.

According to Ng et al., some confounding factors can also be translated into cues that can help discriminate between people from different genders (2019). Namely, items such as glasses and hats can obscure the eye area, which creates difficulty for AGR algorithms that specifically look at the eye region of their subjects. However, as the styles of glasses and hats can be different for men and women, they can also be seen as discriminative features for the gender classification, say Ng et al. (2019). Here, the link between gender and style suggests that Ng et al. assume different genders have different clothing styles.

Use and Value

As automated gendering practices' added value and usefulness are sometimes unclear, this section will turn towards the promised use and value of automatic gender recognition software. The most cited use cases include an improved user experience through adaptable technologies, improved surveillance and public safety, and better marketing with targeted advertisements (Hamidi et al., 2018; Ng et al., 2019; Khan et al., 2014). On top of these uses, Khan et al. also mentioned that automatic gender recognition could help with content-based searching and indexing of digital media (2014).

Of all the papers I reviewed in the chapter, the vast majority mentioned why and for what AGR tools were necessary. For Ramey and Salichs, gender recognition is needed in human interaction with social robots (2014). For this interaction to be smooth and accurate, the robot needs as much information about the person they interact with as possible. While the gathering of information can often be done in an explicit question-and-answer setting, "the range of applications widens if some features can be inferred with no user action required" (Ramey & Salichs, 2014, p.272). Anusha et al. and Ng et al. also emphasise using AGR for human-computer interaction and affective computing (2016; 2019).

Santarcangelo et al. went in another direction and focused on gender recognition for Digital Out of Home (DOOH) advertising (2015). DOOH advertising essentially means all the advertising done in public spaces, including posters at bus stops and billboards. Computer vision enhanced with facial analysis tools can be used to investigate the customer's behaviour and preferences, which helps the retailer trace their revenue. Moreover, digital signage with real-time gender recognition can adapt its content to the potential customer looking at the sign (Santarcangelo et al., 2015). Imagine a sign at a bus stop that shows an advertisement for a perfume or a sportscar, depending on whether the person waiting at the bus stop is a woman or a man. Of course, this is a limited representation of gender, but Santarcangelo et al. only use the gender binary to describe their subjects.

Another example is beverage vending machines in Japan, where the perceived gender of the customer informs what recommendation they receive (Ng et al., 2019). These recommendations have led to an increase in sales. In such use cases, there is an assumption present that preferences regarding products such as beverages are dependent on or related to the gender of the customer. This is similar to the use of glasses and hats as gender descriptors from the previous section. Moreover, such divisions of gender-based preferences reiterate and potentially sustain social norms such as the idea that women should wear make-up and men should like sports cars. Imagine you would continuously get recommended to buy pink-coloured drinks; unconsciously, you would start thinking that blue-coloured drinks are not for you, and you would not consider them viable options anymore. Getting continuous recommendations on what you are supposed to like can affect your preferences and make other options deem inappropriate. This way, gender-based product recommendations can strengthen the corresponding social norms regarding what people from certain genders must or must not prefer.

Khan et al. believe that AGR already takes up an essential part of "targeted advertisement, forensic science, visual surveillance, content-based searching, human-computer interaction systems, etc." (2019, p.1). For example, security cameras with AGR are used to monitor and prevent violence in public transport vehicles and stations (Ng et al., 2019). Another example of AGR in surveillance is the use of biometrics such as automatic gender recognition by the police to recognise and identify wanted persons on the street. This is done using surveillance cameras that capture the faces of pedestrians, which are then compared to the criminal watch lists. However, as this is a computationally heavy task, the input images are first put through an automatic gender recognition tool to filter people of a specific gender (Ng et al., 2019). If the police are looking for a woman, they can first filter all pedestrians on the gender of women, to later complete a more thorough facial analysis to find the right person. The justifications for such use of facial recognition are based on values such as decreasing investigation time and improving stop and search actions (Watts, 2019). AGR is used for similar reasons at airport border controls. There has been a move to create automatised border controls that compare the individual's face with their passport photo (Watts, 2019). A part of this facial analysis is automatic gender recognition to give a complete comparison between the subject and their passport (FRA Focus, 2019). The justification for such tools is based on the increase of financial and time efficiency and the shortening of the queues (O'Neill, 2021).

Moreover, AGR plays a prominent role in such real-world applications and proves to be helpful for the development of facial analysis in general. Khan et al. state that all tasks in this computer vision field are closely related and inform each other (2019). AGR software can aid other facial analysis tasks such as expression recognition (Ng et al., 2019). As such, the improvement of gender recognition software brings about the improvement of facial analysis in general.

Interestingly, only one of the reviewed studies took the time to ask their subjects whether they thought the technology would become useful and whether it would be able to solve the real-world target issue. According to Ramey and Salichs, "users acknowledge the usefulness of detecting their gender for a personalised interaction" (2014, p.273). However, as this study was based on the analysis and classification of breast shape, their subjects also mentioned that the method was experienced as intrusive, and they were concerned for their privacy. Ramey and Salichs end on a – supposedly – positive note, as they state that their subjects had "worse expectations than reality: the sensor and the algorithm are not as intrusive as infra-red cameras" (2014, p.273).

The Manifest and Operative Concepts of Gender

In the introduction, I presented Haslanger's manifest, operative, and target concepts (2012). To recap, these three concepts are essentially three different definitions of a single term. These definitions can coincide, but there can also be differences, big or small. The manifest concept is the formal definition, which is socially agreed upon and found in dictionaries or legal documentation. The operative concept is then the definition that is colloquially used, the definition that determines our actions and interactions. For example, the manifest concept of glass might describe its specific properties, such as its crystalline structure and transparency. The operative concept is broader than that; it also includes cups and windows and says more about what we use the material for rather than its properties. Thus, the manifest and operative concepts can describe similar objects but can have different definitions and can be used in different contexts.

Finally, the target concept is the definition that we should strive to use. The manifest and operative concepts can both fail to adequately describe the target concept, in which case the first two concepts can be re-engineered to match. For example, every couple of years, new editions of dictionaries are published because of erroneous definitions or because a term's social meaning has changed. In this case, the manifest concept – the definition in the dictionary – is no longer correct. The target concept is then the meaning that we want to achieve, and the manifest concept can be altered to coincide with the target concept.

In this chapter, I have discussed how the developers of AGR technologies describe gender and gender identity. The different authors did not define what they take gender to be, nor did they attempt a literature review of the metaphysics of gender. As such, there was no straightforward socially agreed-upon definition of gender. However, I could identify their manifest concept in how the authors described how they analysed gender and the language they used. For instance, Khan et al. wrote about a natural division between men and women that they recognise in society and attempt to formalise into code (2019). I will now look back at the different papers I have discussed in this chapter and summarise their manifest concept of gender.

Most authors described gender as a binary and relatively stable characteristic (Anusha et al., 2016; Geetha et al., 2019; Santarcangelo et al., 2015). For example, Santarcangelo et al. wrote: "A

Support vector machine (SVM) is a powerful classifier for two classes-based problems: gender recognition problem (male vs female) is a good application scenario" (2015, p.4). Moreover, Ng. et al. reviewed the main challenges for AGR tools and listed: "[h]uman factors are due to the characteristics of a person, such as age and ethnicity. Facial expressions (neutral, smiling, closed eyes etc.) are also considered a confounding factor" (2015, p.741). While they acknowledged that people's facial characteristics could change in both the short and long term, they failed to acknowledge that gender itself can change over time. These authors see gender as assigned at birth – either man or woman – and remains stable throughout their lives.

However, some of the reviewed papers acknowledged the possibility of gender transition and investigated what difficulties that might pose for automatic gender recognition tools (Kumar et al., 2016; Vijayan et al., 2016). For instance, Vijayan et al. write: "Gender transformation can be considered a variant of face disguise, however, disguise falls under the broader category of biometric obfuscation, which refers to the deliberate alteration of the face for the purpose of masking one's identity. Transgender persons undergo HRT for the purpose of masking or creating a new identity" (Vijayan et al., 2016, p.1368). The underlying idea is that gender transition can change the subject's facial features, which would create a challenge for appearance-based technologies. Nevertheless, the only method for gender transition that these authors mentioned was hormone replacement treatment: "A person who undergoes gender transformation via hormone replacement therapy is a transgender person. A male becomes female or a female become a male by altering the balance of estrogen and testosterone hormones" (Vijayan et al., 2016, p.1367). This statement implies that gender is purely dependent on hormones and is interestingly not backed by references nor scientific proof. These were the only accounts of gender being in any way fluid; the other accounts approached gender as stable and immutable. However, even Kumar et al. and Vijayan et al. only acknowledged two genders and merely acknowledged that transitions between these genders could take place.

Moreover, sex and gender are used interchangeably (e.g., Mahalingam & Ricanek, 2013), which contradicts the idea that sex is physical and anatomical and gender is its socially constructed meaning, as suggested by feminist philosophers such as Haslanger (2012). Another central aspect of the reviewed articles was that gender is directly correlated to someone's appearance. AGR technologies analyse images of subject's faces to determine their gender, which indicates that someone's facial features are assumed to be informative of their gender. The proposed use cases of automatic gender recognition tools suggest that individuals have personal preferences that coincide with their gender. Namely, Santarcangelo et al. state that the main application of AGR will be in digital out of home (DOOH) advertising (2015). Moreover, Ng et al. mention examples of AGR for the use of beverage recommendations in vending machines (2019).

One of the most concerning things is that gender is grouped with "feelings and moods" and ethnicity and race (Mahalingam & Ricanek, 2013, p.1; Vijayan et al., 2016, p.1367). This suggests that the metaphysical nature of gender is either on a similar level as feelings and moods or ethnicity and race.

However, feelings and moods are fleeting, while gender, ethnicity and race are not. Someone's mood might change from one moment to the next; this affects their facial expressions. However, someone's gender and race do not change in such short timespans, nor do they alter someone's immediate facial expressions. Moreover, feelings and moods are not typically characteristics of social categories or the direct subjects of social hierarchies. Grouping these characteristics together disregards their differing epistemic, metaphysical, and social statuses. Interestingly, Anusha et al. attempted to make a connection between feelings and gender by stating that "[r]esearch have proved that females express more anger than the male and male express more joy than the females" (2016, p.202). However, this is not backed by arguments, references, or scientific evidence.

As becomes clear, the terms sex and gender are used in various situations and with different meanings throughout the papers. This is also enlightening regarding the operative concepts of gender that the authors deploy. As mentioned earlier, all authors describe gender as something that can be read from the subjects' facial features. This is operationalised into certain rules and relations between physical aspects and gender; for example, if a person has large browbones, they are a man. Some researchers focused on facial features such as the colour difference between a subject's lips and skin (Anusha et al., 2016) or the distances between facial features such as eyes and ears (Khan et al., 2019; Santarcangelo et al., 2015). One article even claimed that the left eye is more determinative of gender than the right eye and have used that to compute someone's gender (Vijayan et al., 2016). Other researchers focused on other features such as breast size and compared the breast sizes of men and women (Ramey & Salichs, 2014). What these approaches have in common is the assumption that gender is recognisable from someone's appearance. While the researchers did not specify whether this relation is a correlation or causation, they assume that someone's appearances are informative about their gender. They have operationalised this assumption into computational rules, which produce a probabilistic prediction of a subject's gender. However, as I will show in the third chapter, the probabilistic nature of these results is often misunderstood.

In short, these papers present their manifest concept of gender as something stable, immutable, determined by hormones, appearance-based, which coincides with preferences and tastes. The operative concept is then the way that these manifest concepts are formalised into code. The manifest and operative concepts seem to correspond quite well over the different articles, with the only exception being the possibility of gender transition using HRT. Notably, Kumar et al. and Vijayan et al. discussed HRT because they assumed HRT would change the subject's facial features (Kumar et al., 2016; Vijayan et al., 2016). Kumar et al. even explicitly mentioned that their assumption that some facial features of people undergoing HRT are unchanging was based on "intuition" (2016, p.1). The authors of the papers discussed here did not attempt to research the metaphysics of gender, nor did they investigate and define their personal understanding and assumptions. Gender and gendering are seen as common knowledge that does not need defining.

This chapter has explicated the manifest and operative concepts of gender as described by the developers of AGR tools. In the second chapter, I will aim to find a target concept of gender, which describes the complete picture of what gender is and what AGR should be able to recognise. The third chapter will then investigate whether the manifest and operative concepts can be re-engineered to coincide with the target concept of gender. If so, it would be reasonable to assume that AGR can be improved upon to minimise the harmful misgendering of people with marginalised genders. Moreover, the third chapter will further investigate the probabilistic nature of the AGR results and will show why a correct understanding of this nature is essential.

Chapter 2: The Pragmatic Metaphysics of Gender

In the previous chapter, I have shown how the developers of automatic gender recognition tools define and discuss gender and which underlying assumptions about the nature of gender can be identified. These were the manifest and operative concepts of gender. Many AGR developers talk about categorising men and women as if they are merely digitalising a natural phenomenon. However, in this chapter, I will show that gender groups are not biological categories. Gender is socially constructed, and AGR formalises the corresponding social norms and expectations into technology and code. AGR does not merely read bodily features and create a prediction based on them; it also reiterates and reinforces cultural and social norms. This chapter will search for the target definition of gender; that is, the definition of gender that we ought to use and that AGR should be able to recognise.

Finding the appropriate target concept of gender will inform whether it would be possible to improve AGR to become more inclusive. For AGR to improve, its manifest and operative concepts must be re-engineered to coincide with the target concept. An example of this is the work done by Merler et al. and Keyes (2019; 2018). Merler et al. argued that AGR could be made more inclusive if it would be built using more diverse images, and Keyes argued that a more nuanced definition of gender is the solution. The possibility of improving AGR will be further discussed in the third chapter; this chapter will first investigate the proper definition of gender.

Gender identity is something that we deal with daily, as it affects the way we see ourselves, the way we interact with people, and our position in broader society. Many people with a gender other than ‘man’ or ‘woman’ structurally experience negative consequences because of their gender identity. Their gender identities are not recognised and respected in interpersonal relations and even wider society. This can take many different forms; one example is online forms by governmental institutions or companies, which only include checkboxes for men and women. Such disregard of other genders can have harmful consequences, for example, when someone cannot apply for government financial aid because their gender does not fit the norm. Because of AGR’s versatility, its target concept needs to be able to describe gender on the personal, interpersonal, and societal levels. As shown in the first chapter, the technology developers did not build their tools for one use; they all mention that the technology can be applied in many cases. Moreover, even if the technology was developed with a specific goal, it might still be used in other instances. Thus, AGR cannot work with a very narrow and contextual conception of gender, such as “would this individual feel more comfortable being searched by a male or female officer?” Instead, the target concept must explain the role of gender in all aspects of life and must be invariant to the use case.

To find the correct theory of gender, I will review various philosophical conceptions of gender, which approach the issue from different societal scales. These scales are the personal level, interpersonal level, and societal level. For example, where Haslanger’s societal level theory focuses on the role of gender in the oppression of groups of people (2012), Bettcher’s personal level theory emphasises the aspect of first person-authority in her theory (2009). The theory I am looking for must describe the role of gender on all three levels.

The following sections will briefly look into (mis)gendering practices and their harmful consequences. The subsequent sections will amelioratively analyse different conceptions of gender and how they explain the harmful consequences of misgendering practices and marginalisation based on gender. The different conceptions of gender will be categorised into personal, interpersonal, and societal level theories. I will evaluate these theories on their appropriateness for criticising how the developers of AGR tools operationalise gendering and why AGR-mediated misgendering is harmful.

Gendering Practices

Gendering is the act of assigning someone a gender; this can either be done by another person or a machine. However, by gendering someone, you assume that you have authority over someone else's gender and that their perspective is irrelevant. This chapter will argue that such gendering is unethical and that individuals must have agency and sole authority over their gender: the first-person authority as coined by Bettcher (2009). This entails a moral obligation to respect each other's agency and authority regarding our genders. Importantly, this authority is of an ethical and social nature rather than an epistemic one. In other words, our first-person position does not give us an "epistemic advantage" with which we can know or even create our gender identity (Bettcher, 2009, p.100). Thus, first-person gendering statements have no metaphysical influence or consequences. Instead, they represent how we want to be seen and treated. This is the existential self-identity (2009, p.111), a concept that will be further explained in the section *Gender on the Personal Level.*

Bettcher even argues that finding a metaphysical account of gender is counterproductive through two compelling arguments (2009, p.111). First, it is impossible and unimportant to find a metaphysical theory that describes all people from one gender, as this runs into the commonality and normativity problems. In short, these issues describe how theories of gender often aim to find common ground among people of a certain gender – with the risk of excluding people – or try to create norms for how people of a certain gender should act.[1] Thus, searching for a metaphysical theory of gender is due to create conflicts with other people's self-identities. It creates the possibility of being unable to uphold your conception of how you must present your gender. Second, a metaphysical theory of gender does not explain which norms and expectations someone upholds, which is informational regarding which community someone is – or wants to be – a part of. Importantly, when someone states 'I am a woman,' she does not say that she conforms to the metaphysical conception of a woman. Instead, she believes she is a woman and wants to be seen and treated as a woman.

[1] These are two common issues in gender theory and will be further developed in the section *Gender on the Social Level*.

These are two significant reasons why a purely metaphysical account of gender is unimportant for this project. Rather than aiming to solve the commonality or normativity issues, this chapter will follow the method of ameliorative analysis and focus on the *pragmatic* metaphysics of gender. An ameliorative analysis is inherently ethical as it does not merely describe the way the world works and how terms and concepts are used but simultaneously aims to find a way to build a better world, for instance, through more conscious language usage. I am not trying to find a target concept of gender for its epistemic value but rather for its political value. This means that the pragmatic, political, and semantic consequences of the discussed theories must be central considerations for this project.

Gendering and misgendering – the act of incorrectly gendering someone – have major political, pragmatic, and semantic consequences. For example, misgendering can lead to constraints in accessing political resources that people have a right to access according to their gender. Whereas gendering is the social practice that makes misgendering possible, misgendering is a singular instance of harm against an individual. The following section will explicate how misgendering is immoral; as misgendering is a smaller and more precise set of actions than gendering itself, its political, semantic and pragmatic consequences are more straightforward. However, as long as the dominant gendering practices allow misgendering to take place, gendering is immoral, too. Thus, before delving into the pragmatic metaphysics of gender, I will briefly discuss the harms of misgendering and their immorality. These reasons for misgendering's immorality still inform the immorality of gendering practices in general, as it is the social practice that makes misgendering possible.

The Harms of Misgendering Practices

Misgendering can take many different forms, all of which are harmful in their own right. For instance, misgendering can be as simple as accidentally referring to someone by the wrong pronoun - in their presence or behind someone's back. In the case of AGR, misgendering can take the form of a machine misclassifying an individual's gender (Keyes, 2018). Moreover, misgendering someone can lead to dangerous situations. People might be asked to expose themselves to 'prove' their gender, and discussions surrounding someone's 'correct' gender can be met with aggression and anger.

Nevertheless, this is not always the case; an act of misgendering is often waved off as a 'harmless' mistake, misjudgement or a slip of the tongue. These relatively minor acts of misgendering terms are forms of microaggressions. Microaggressions are hostile, everyday forms of indignity towards other people and can be verbal, behavioural, and environmental (Wing Sue, 2010, p.5). Such microaggressions can be both intentional or accidental. While a single instance of misgendering can be seen as harmless on the surface, the sum of all these microaggressions becomes suffocating for people who go through it on a daily basis. Robin Dembroff and Daniel Wodak present four reasons why misgendering people from marginalised genders is harmful (2018).

The first reason is based on disrespect; essentially, misgendering someone is a form of disrespect against them and the other people with the same gender identity. Using the wrong pronoun to refer to someone denies them their gender identity and inflicts harm onto them and those who use the same pronouns. While deliberately referring to someone by the wrong pronoun is more harmful than accidentally doing so, the latter is a form of disrespect, too, as it is negligent of their well-being. Whereas Dembroff and Wodak focus on misgendering as a form of disrespect regarding someone's person and well-being, misgendering also disrespects the individual's agency and autonomy. Deliberately disagreeing with someone over their gender identity assumes that their personal view of their gender identity is inconsequential, and their gendering speech acts can be ignored. Thus, misgendering someone is a form of disrespect against their agency and autonomy regarding their gender identity.

This type of misgendering takes place on the personal and interpersonal level, as it is an interaction between two or more people, but the misgendered subject experiences the disrespect personally. The argument against such misgendering is based on the idea that misgendering is intrinsically harmful as it is a form of disrespect. AGR technologies can reinforce such disrespect, as it is proposed to play a prominent role in human-robot interaction (Ramey & Salichs, 2014). In such cases, the robot in question would refer to someone using gendered pronouns and honorifics based on their prediction. When the robot misgenders the person it is interacting with and approaches them using the wrong honorifics, this can be experienced as a form of disrespect.

The second reason why misgendering is harmful regards the social resources that are divided based on sex and gender. Being part of a specific gender group comes with access or restrictions to resources in society. These resources can include job opportunities, financial benefits, clothing, and access to spaces such as women's toilets. Misgendering people can keep them from using these resources, which they have a right to access. AGR technologies reinforce such misgendering in public bathrooms or airport security cases, as they can withhold the resources of access and safety from the subject. Such misgendering is not intrinsically but instrumentally harmful. Misgendering takes away the possibility for people to access the social resources to which they have a right. Here the resources are the intrinsic good, and gendering someone correctly is a means to access those resources. This takes place on the personal, interpersonal, and societal levels, as withholding social resources can be caused by interactions between two people or on a more structural basis. The effects are also felt on a personal level.

The third reason is what Dembroff and Wodak call intelligibility. This describes the set of norms and values to which someone with a specific social identity is expected to subscribe. As a woman, you deal with different societal expectations than people from other genders. These societal expectations are blueprints for how you are supposed to appear, act, and interact.[2] However, when you are misgendered, you are compared to the wrong social blueprints. This then undermines your ability to accept or reject the societal norms that come with your specific gender. Again, this argument alludes to the instrumental

[2] I want to note that the absence of a blueprint is, in a way, a blueprint in itself. For example, gender fluid people might not want to adhere to men and women's blueprints but to their blueprints. They are expected to go against society's expectations regarding how men and women should behave.

value of gendering people correctly, as it gives the person room to navigate society and its norms in a way that fits them. AGR technologies reinforce such misgendering as it upholds the social norms of what a person from a specific gender should look like. It is built upon the comparison between a person and the social blueprint of their predicted gender.

Finally, the fourth reason is more ideological and takes primarily place on the societal level. This argument regards the broader social system in which the misgendering practices take place. Misgendering people reinforces the societal structures that give way to the first three reasons. While this is more pressing in deliberate misgendering, even accidental misgendering reinforces the idea that you can gender other people based on their appearance and that everyone should be using the pronouns "he" or "she". AGR technologies reinforce this. It perpetuates the idea that gender identity is readable from someone's appearance and that it is unnecessary to ask someone what their preferred gender pronouns are. Taking part in this harmful, offensive, and potentially dangerous practice allows society to continue harming people from marginalised genders. Instead, asking someone what pronouns they prefer and using them can create a shift in the systems of language, concepts, values, and norms that make up our social structure. Again, this argument is instrumental as misgendering is a way to strengthen the dominant ideology, which rejects other people's correct gendering.

Gender on the Societal Level

According to Haslanger, gender is socially constituted through power relations. Gender is a social status that fits within society's hierarchical structure. While gender is not solely based on biological distinctions, bodily features inform whether someone is recognised as a specific gender. In particular, a person is a woman when their body fits in the norm often observed with people of that gender, which is linked to a woman's role in reproduction. Moreover, the presence of these bodily features dictates the social position in which the person is placed. Finally, the bodily features and the social position that correspond to being a woman determine the subordination the women experience in relation to men. While Haslanger uses binary language to discuss language, her descriptions of social positions based on gender can be extended to non-binary genders. Namely, people who are oppressed because of their gender and whose bodies fit the conventional norms that correspond to their gender belong to said gender.

Central to Haslanger's theory are the hierarchical structures that are present in society. Haslanger defines the gender 'women' as the social group that experiences oppression due to their gender (2012, chapter 7). Conversely, the gender 'man' is then the social group that experiences privilege due to their gender group. Interestingly, Haslanger maintains that if this inequality in privilege based on gender were to be erased, the social groups 'men' and 'women' would be eliminated. There would still be males and females, but the social groups based on perceived bodily features would no longer exist.[3]

As mentioned in the introduction, the discussion surrounding gender is often regarded as the social meaning of sex (Haslanger, 2012, chapter 7). Haslanger uses the terms' male' and 'female' to denote people's sex, and 'man' and 'woman' to describe genders. Gender can circumscribe social roles, norms, behaviours, traits, meanings and identities that reflect actual or imagined biological characteristics. Moreover, these imagined and assumed physical characteristics have become gender markers and are used as instruments of power to enforce the dominant ideology of the gender binary.

According to Haslanger, the effort to determine the nature of gender often encounters two problems: the commonality and normativity issues (Haslanger, 2012, p.228). The commonality problem describes the aim to find common ground between all people of one gender. An example question can be: is there something that all females have in common that makes them women? However, as such theories often work with physical or imagined sex and gender markers, they depend on an underlying assumption that assigning a gender to someone based on their appearance is possible. This brings a risk of misgendering as gender is not one-size-fits-all, and identity is not necessarily linked to physical or personality characteristics. The normativity problem then describes the investigation into whether certain norms and values people of a certain gender need to ascribe to. The question is whether the membership of gender groups depends on fitting in these norms and values or whether the norms we currently recognise in society must be accepted. Additionally, this can lead to the exclusion and marginalisation of people who fulfil these norms and values to a lesser extent than others.

Haslanger aims to move away from these two problems as she is less interested in whether empirical similarities between people of a gender group can be found and more interested in which conception of gender can help to ensure sexual justice. That would be the target concept. This approach is specifically helpful for practical issues such as design recommendations for AGR technologies as it brings issues of equality and justice back into focus. The discussion surrounding AGR tools often overly focuses on the commonality issue, as one of the AGR field's main challenges is finding features that distinguish men from women. According to Haslanger, there is no one right way to classify and order human bodies as they are different in many aspects. Moreover, the usefulness of classification methods depends on the case you are investigating. This means that there is no one solution to the commonality problem.

Further, these methods have political consequences, influencing how gender is seen, used, and institutionalised. Gender is an issue of power, not of finding the perfect definition of the metaphysics of gender, or applying one categorisation method to the entire population. In the context of AGR technologies, it is essential to remember that the way gender recognition technologies are built and

[3] This is an ongoing discussion in feminist literature. Moreover, there are issues with this approach, as some people might want to fight against their position as subordinate without challenging the existence of their gender altogether. For more information, consider the papers by Bach (2012), Jenkins (2016), and Mikkola (2009).

employed has political consequences, as some people will be excluded and others included. Therefore, the basis on which Haslanger builds her theory of gender is decidedly applicable to AGR technologies and will form the groundwork for the rest of this thesis.

In sum, depending on the situation and the case, the right target concept of a term must be determined based on pragmatic/political and semantic considerations. The pragmatic – and sometimes political – criterion is a question of whether the original terms of the discourse should be used or whether the terms should be altered or even replaced entirely (Haslanger, 2012, p.225). The semantic criterion then regards whether a shift in the meaning of a term helps describe the phenomenon at hand and whether the central function of the term still stays largely the same such that the term can be used in the same situations. Thus, rather than finding the end-all definition of gender, we must consider what definition works in the relevant political situation. AGR's political situation is quite complex, as the technology is versatile and can be applied in many different cases: from airport security to beverage recommendations. Moreover, these applications are likely to be dealing with a wide variety of politics, as people from all over the world travel through airports and are subject to airport security technologies. Whereas some countries allow for an 'X' in your passport to denote a non-binary gender, others do not recognise such genders. However, the UN International Civil Aviation Organisation (ICAO) standards require passport-reading machines to read the genders M, F or X (Bowcott, 2020). It is thus vital to take into account the international standards and norms regarding the matter.

Haslanger's definition of gender is mainly reliant on the social and political context of the social group. A gender group is a set of people who are routinely seen as having certain biological features, which are assumed to be a part of the said group according to the social context in which they are situated. Moreover, fitting in these gender groups and having the biological characteristics that are imagined for these groups reinforces the social context's dominant ideology. You can only be a member of said gender group when you have the proper biological make-up, and these gender groups are hierarchically ordered within society. Importantly, someone is not a part of a specific gender solely based on their biological make-up. Instead, an individual's gender membership depends on their social position in a system involving other gender groups. Gender is defined relationally: gender groups get meaning through their relation to other gender groups (Haslanger, 2012, chapter 1).

Haslanger approaches gender groups as hierarchical social groups; she bases her definition of a woman on the presence of oppression due to gender identity. Gender is inherently political and constructs and strengthens the norms that are connected to sex categories. Understanding how gender is socially constructed will also create an understanding of social processes of control. While Haslanger's theory offers a relevant and robust foundation regarding what is important in a theory of gender, it overly focuses on the societal level – and to a lesser extent, the interpersonal level – of gender identity and fails to explain what gender means on the personal level. The theory disregards how gender identity is experienced on a personal level or how being misgendered can be felt like a form of disrespect and harm. Of Dembroff and Wodak's four arguments against misgendering (2018), this societal theory only substantiates the arguments from social resources and ideology, as these take place on the societal – and to a lesser extent the interpersonal – level. Haslanger's societal theory helps describe how AGR technologies represent current gendering practices. By echoing existing gender norms regarding what people from a certain gender should look like and reiterating the idea that people should be externally categorisable into the gender binary, AGR technologies help uphold social hierarchies. However, it does not fully describe what happens when someone is misgendered on an individual level. This is bigger than just the harms of microaggressions; it concerns the personal experience of society's rejection of your gender identity. While gender identity plays a significant role in our daily life, the experience of being misgendered is not acknowledged in Haslanger's theory. As the arguments against misgendering as a form of disrespect and intelligibility cannot be disregarded, the next section will look into a theory that places the individual on the centre stage.

Gender on the Personal Level

Another approach in defining gender focuses on the role and power of self-identification. Talia Mae Bettcher is one of the philosophers who places central importance on self-identification in the construction of one's gender identity (2009). According to Bettcher, someone's gender identity reflects "how one conceives of oneself, or feels oneself to be with respect to sex and/or gender categories" (2017, p.120). Under this definition, gender identity is a subjective and internal matter; gender is not defined in relation to broader society, nor does it describe how someone must present themselves. As one's belief about their gender is purely self-determined, there is no relation between someone's gender identity and appearance, making it impossible for external people or machines – including AGR – to recognise how they identify.

This also means that declaring one's gender identity should not be understood as merely an acknowledgement or a refusal of one's genitalia. Instead, it reflects that person's self-identity (Bettcher, 2009). As self-identity is not a physical trait but rather a belief someone has over themself, the person in question has authority over this identity and the group membership that flows from it. However, one issue with this is that self-identification and its representation can be affected by the person's political attitudes and the context in which they are situated. For instance, someone might self-identify as a woman but does not feel comfortable making that self-identity known publicly due to the political climate in which she lives.

Talia Mae Bettcher's account of trans identities and first person-authority offers an interesting argument regarding why misgendering people is harmful and unjust (2009). One common form of misgendering is using the wrong pronoun to refer to someone. However, this often raises the question: who decides what the 'right' and 'wrong' pronouns for certain people are? Bettcher argues that people, and specifically trans people, have first-person authority regarding their gender. This essentially means that first-person declarations of gender – think of 'I am agender' and 'My pronouns are they/them' – are uniquely informative and must be respected. In turn, second and third-person statements do not have this decisive status. This is not an epistemic claim but an ethical one. While individuals do not have a "superior epistemic position" over their own gender, they have a superior ethical position (Bettcher, 2009, p.100). Correspondingly, Bettcher argues we should not be talking about one's metaphysical self-identity but rather their existential self-identity (Bettcher, 2009, pp.110-112). You are uniquely positioned to determine what life you want to live and how you want to be treated. This is your existential self-identity, and this is what you have first-person authority over. As this is an ethical claim, others are morally obliged to respect your wishes regarding your pronouns. When there is a disagreement regarding the gender of a person, the individual's first-person statements are correct. Through focussing on the existential self-identity, Bettcher places central importance on the semantic, political, and pragmatic implications of first-person gendering statements rather than their metaphysical and epistemic status. This is in line with the ameliorative analysis method.

In sum, the ethical implications of gendered speech acts have priority over the corresponding metaphysical and epistemological considerations. When someone announces their pronouns, they declare which position they want to take up in society. While you can disagree with the reasoning or the metaphysics that underlies their conception of gender, you must respect their gender declaration due to their ethical first-person authority. This is an essential concept for this thesis as it describes how second- and third-person gendering by humans or machines are problematic. AGR technologies take away the first-person authority people have over their gender. Technologies such as AGR are not only built upon the assumption that someone's face can be used to determine their gender but also on the assumption that the subject does not have authority over their gender identity. In Bettcher's theory, this is unethical as the subject should have first-person authority at all times.

Bettcher then argues that the disagreements and statements regarding someone's gender that oppose their first-person declaration – a kind of misgendering – are forms of sexual abuse. While it might seem as if statements such as 'You are really a man' when you identify as agender miss the mark in the same way as 'You want to go running' when you would rather do yoga, there is a significant difference. This difference is twofold. The first claim is abusive regarding the person's genitalia, as it disrespects the person's right to privacy regarding their genital status. Moreover, as questions such as "do you have a penis?" are considered sexual harassment, statements regarding someone's genitalia – be it implicit – can also be seen and experienced as sexually harassing, which is a form of sexual abuse (Bettcher, 2009, p.107). Second, it disregards the person's power over their self-identity, as such statements assume that you have authority over the matter. Thus, gendering statements make use of sexually abusive techniques and disrespect someone's sense of identity. We could also consider AGR as a sexually abusive technology, as it perpetuates a system of forcing the disclosure of one's genitalia. It analyses and determines someone's gender, regardless of whether the subject wants to disclose their gender identity or§ the absence or presence of specific genitalia. For example, think of the AGR technology that takes the subject's breasts as the object of investigation (Ramey & Salichs, 2014). It extracts information from someone's secondary sex characteristics, often without the explicit consent of the subject. The technology simply takes what it needs to predict someone's gender by looking at their breasts without respecting the subject's first-person authority over their gender. It forces the disclosure of the subject's secondary sex characteristic, as the subject has no choice regarding whether or not they are subjected to this technology. This is a form of sexual abuse, regardless of whether the gender was correctly or incorrectly recognised.

While Bettcher's theory offers a strong argument against external gendering, it does not describe the social norms and expectations that come with gender and appearance and how people from certain genders navigate those. Furthermore, gender self-identification does not merely happen internally; it happens in conjunction with presentation, behaviour, interaction, and recognition by governmental bodies. These are important issues for automatic gender recognition tools but not circumscribed by Bettcher's theory. Moreover, as this theory only takes place on the personal level, the four arguments against misgendering as laid out by Dembroff and Wodak are left unexplained (2018). This theory helps analyse AGR tools on a personal level; it explains why external gendering disrespects the authority people have over their gender and how misgendering through the use of AGR is harmful. However, it disregards how (the lack of) societal recognition can influence the construction of one's gender identity. This project necessitates a theory that bridges the personal, interpersonal, and societal levels and can bring these three aspects together. The following section will introduce a theory that can circumscribe gender on the personal, interpersonal, and societal levels.

Gender on the Personal, Interpersonal, and Societal Levels

In the previous sections, I have introduced Haslanger's and Bettcher's theories of gender, which focus solely on the societal and personal levels of gender, respectively. While both theories describe important aspects of the nature of gender identity and the societal structures that come with it, they do not encompass all aspects of gender identity formation and recognition. Robin Dembroff has created a theory that acts as the middle ground between Haslanger's societal theory and Bettcher's personal theory and has extended Haslanger's conception of gender on several axes. First, there are genders other than women who experience oppression and marginalisation due to their membership in a gender social group. Dembroff follows Haslanger's division of dominant and subordinated gender groups, where the dominant groups represent the groups that experience privilege due to their gender in the dominant social contexts. However, as gender is not a binary division, Dembroff includes all gender identities in this definition.

Dembroff then argues that dominant social groups based on gender marginalise trans people and completely exclude non-binary people (2018). This is because trans and non-binary people are systemically seen as having a gender that differs from their self-identity. However, as some transgender people still conform to the man-woman binary, they are not entirely excluded from the dominant gendering practices, as long as they comply with society's image of people with their gender. Non-binary people, per definition, do not fit within the dominant gender groups and are thus marginalised and oppressed because of their gender. In short, dominant gendering practices oppress trans and nonbinary people.

To create gender classifications that do not marginalise nor exclude people, we need to move away from the dominant contexts. Otherwise, these gendering practices, however inclusive they might be, would merely reinforce the existing oppression. This is also the case for AGR technologies; if the technology would be improved such that it becomes very accurate in recognising the gender identities of people who fit in the dominant gender categories, it would merely echo the current gendering practices. This could create more marginalisation for people from other genders. Moreover, even if the technologies could be expanded to properly recognise trans or non-binary people, this still reinforces dominant gendering practices. As gender is a spectrum and the same gender identities can mean different things to individuals, gender cannot be distilled into a discrete set of categories. Thus, AGR technologies are bound to exclude some people based on their gender identity. The only solution would be to move away from the dominant gendering practices completely.

The second way in which Dembroff has extended Haslanger's theory of gender relates to social kinds, positions, and blueprints. Together with Catharine Saint-Croix, they built their conception of gender up from Haslanger's foundation of social kinds (2019). Dembroff and Saint-Croix agree with Haslanger that being a member of a specific social group is means that you will be sorted into the corresponding social position. Your social position is dependent on your social practices and interactions with people, ideas, and artefacts around you. Moreover, these social positions come with social blueprints, which describe the attitudes and belief systems that people in these social positions should uphold. These blueprints are public, stable, and at the foundation of many behavioural and emotional tendencies. Blueprints are engrained in society to such an extent that individuals cannot change them because of the force of the majority. In short, by participating in society, we are given a social position that comes with a corresponding social blueprint that determines how we should act and feel within the corresponding social group.

Additionally, Dembroff and Saint-Croix also take a slightly wider stance than Bettcher. They argue that while self-identification is an essential aspect of social identity, it is not the sole determiner (2019). They have investigated the role of an individual's agency in determining their gender identity. They were motivated by the discussion surrounding using preferred pronouns and argued why misgendering through referring to people using the wrong pronouns is unethical. Dembroff and Saint-Croix say that people have agency over their gender identity, which means they have the sole power to decide and determine their gender identity. This agential identity depends on two conditions: self-identification and position-directed externality. The self-identification condition is the foundation of the agential identity, as an individual must self-identify with the social group. According to Dembroff and Saint-Croix, there can be no group membership without self-identification with said group. This self-identification can differ in strength; some people might self-identify themselves weakly with a particular group, and others perceive it as a central part of their identity. Additionally, as it is possible to self-identify with multiple groups, some self-identify strongly with one group but weakly with another.

The position-directed externality then describes how the individual must make (or intend to make) their self-identified group membership visible to broader society. This can be done by reflecting certain physical features, behaviours, and attitudes that match the social group. This is the externality side of the position-directed externality. The position-directed aspect means that the externality must be coupled with a choice to make their internal identity public. There are cases in which a person shows self-identification, but this is not accepted by broader society. For example, some people might self-identify with a specific race without having the corresponding background and ancestry. A community can refuse the uptake of someone who self-identifies with their group. Self-identity is not enough nor necessary for having a certain social position, and vice versa. Essentially, one's agential identity is the relationship between their self-identity and their preferred perception by society. As such, agential identities play a significant role in social interactions. According to Dembroff, gender queerness is inherently a political stance, as you necessarily deal with the social structures, norms, and expectations that come with your gender identity. This is less visible in Bettcher's theory as she presents gender as purely based on self-identification and free from the influence that broader societal structures may present.

In sum, Dembroff and Saint-Croix have created a balanced position that describes the personal, interpersonal, and societal levels of gender and gender identity. Their theory encompasses both self-identification and hierarchical societal structures. It describes on the personal level how without self-identification, one cannot be a member of a specific social group. However, this self-identification must be visible and recognisable by others through interactions with the outside world – this is the interpersonal level. Moreover, gender groups come with social positions and blueprints, the navigation of which is an essential part of gender identity. These social positions are placed in hierarchical structures, which finally describes the societal level of gender identity. As such, Dembroff and Saint-Croix's theory is uniquely capable of bridging these three levels and describing how these three levels interact. This also means that Dembroff and Wodak's four arguments – disrespect, social resources, intelligibility, ideology – against misgendering are covered, as Dembroff and Saint-Croix's theory describes how misgendering is unethical on the three levels. Additionally, as Dembroff and Saint-Croix's theory is built up from Haslanger's work, it has created a space for hierarchical societal structures. It describes how dominant gendering practices reinforce the ideology of the binary gender. Nevertheless, Dembroff and Saint-Croix do not place sole importance on these hierarchical structures and have moved past the gender binary and created space for the gender spectrum.

The Target Concept of Gender

The final section of this chapter will turn towards the target concept of gender. This is the definition of gender we ought to be using in designing and employing AGR technologies. In the previous sections, I have shown how Haslanger, Bettcher, Dembroff and Saint-Croix approach gender and gendering. In weighing these approaches, I investigated how these three approaches can account for gender on the personal, interpersonal, and societal levels. The conception of gender that best explains gender on these three levels is Dembroff and Saint-Croix (2019). They argue that while membership with a gender group cannot exist without self-identification, gender identity is more than just self-identification. Group membership also necessitates the navigation of the corresponding norms and expectations in interpersonal relations and broader society. The central concepts in this understanding of gender are self-identification, position-directed externality, agential identity, and fluidity. Moreover, gender must not be seen as a binary but as a spectrum; gender cannot be distilled into a discrete set of genders. With this approach as the target concept, the link between gender and preferences regarding clothing or even beverages becomes arbitrary and unfounded.

Nonetheless, Haslanger's and Bettcher's approaches offer unique takeaways for this project, which must be included in the target concept of gender. Haslanger's theory circumscribes how gender is a social construct and plays a central role in creating and upholding the hierarchical structure of society. Moreover, Haslanger argues that the definition of gender depends on the political and social climate in which the topic of study takes place as well as semantic and pragmatic considerations. This is a good foundation for this project due to the versatility and ubiquity of AGR applications. Another takeaway I will be incorporating in this project is the understanding that the potential political consequences of a definition of gender must inform whether the definition is fitting or not. That is to say, it is important to try to anticipate what political consequences people might experience due to the creation and uptake of a certain definition of gender. AGR developers must either strengthen and uphold the current hierarchical social structure or redefine what gender is in society. While this decision might not be conscious for many AGR developers, they play a significant role in maintaining social gender norms. This insight will underly the rest of this project.

What makes Bettcher's approach interesting for this project is the ethical grounding of the first-person authority. While not disconnected from agential identity, Bettcher's first-person authority is uniquely capable of describing the difference between a first-person statement and a second- or third-person statement regarding someone's gender. This is useful in distinguishing a computed result from an AGR tool from a first-person gender speech act. Nevertheless, the concepts of first-person authority and agential identity can strengthen each other. When I speak about first-person authority, I do not merely mean that a person has privileged information about their gender. Whether this is the case is irrelevant to this project. The important thing is that every person should have the agency and autonomy to express their gender identity and share it with the world. Thus, the two concepts of first-person authority and agential identity are closely related, and both play an essential role in the target concept.

However, an objection against combining Dembroff and Saint-Croix's agential identity and Bettcher's first-person authority is that the position-based externality aspect poses an extreme requirement, which does not create space for first-person authority. The position-based externality creates a tension between intending or succeeding to make your identity visible to your environment. In other words: does your agential identity depend on merely intending to make your identity visible to the outside world, or does it also matter whether you succeed in making your identity known? The pragmatic response to this is that these two mechanisms are interrelated. Intending to make an identity known and succeeding in doing so are cyclical processes that strengthen and diminish each other. You make your identity visible and known through interactions with your surroundings and taking up a social position. By doing so and succeeding in this, your intention to continuously present your identity to the world becomes strengthened. Thus, the intention to identify as something has a built-in social success requirement. Your social and technological environment can enhance or diminish your first-person authority as they can create or take away the space in which you can make your identity known. This is the position-based externality view that will be used in this thesis.

Thus, the target concept of gender for this project is a combination of Dembroff and Saint-Croix's agential identity, Bettcher's first-person authority, and Haslanger's societal understanding of gender. Together, these theories successfully describe what issues surround gender on the personal, interpersonal, and societal levels. In the previous chapter, I have shown that AGR technologies can be used in a plethora of cases, from personal beverage recommendations to anti-violence security systems, public transport and even mass police surveillance. It is thus crucial that the target concept of gender can touch upon the personal, interpersonal, and societal consequences of AGR technologies.

In sum, this project's target concept focuses on agential identity, which is the sum of self-identification and position-based externality. Additionally, this target concept is based on the understanding that gender is not a natural division of people but a social one. People are not born with a certain sex and a corresponding gender. Instead, we are given a gender at birth through interaction with our social network. This socially imposed gender can coincide with the gender that you self-identify as, but that is not a necessary link. It is possible that someone's gender identity changes over time or that they grow into their gender identity. We ought to see gender as fluid and as something that every individual has agency over. While gender still plays a prominent role in how we see people and in our interpersonal interactions, we must remember that the only person who can determine their gender is the subject themselves. Any external person or machine should not have full authority to determine and decide someone else's gender. While external parties might be able to recognise an individual's gender, they have no determining power or agency.

In the next chapter, I will use the perspective proposed by Dembroff and Saint-Croix as the target concept and investigate whether it would be theoretically possible to re-engineer AGR's operative concept of gender to coincide with the target concept. An AGR technology that can capture the target concept of gender would be more inclusive, fair, and less harmful to people of marginalised genders.

Chapter 3: The Ethics of AGR

In the previous two chapters, I have investigated the working of automatic gender recognition tools and the philosophical grounding of different conceptions of gender. The first chapter explicated the manifest and operative concepts of gender that are visible and used in AGR. The second chapter aimed at finding a suitable target concept of gender for this project. This final chapter will investigate whether the use of AGR is used can be ethically justified. To do so, I will bring the manifest, operative, and target concepts together and explore whether the manifest and operative concepts can be re-engineered so that they coincide with the target definition. The underlying idea is that the functionality of AGR would be maximised and its harmful consequences minimised if it were able to describe the target concept of gender successfully. This is since an AGR tool that works with the target definition would be inclusive, fair, and respectful to people's authority and agency. Thus, if AGR could successfully describe the target definition of gender, there would be an ethical justification for the use of AGR.

To recap, the target concept of gender describes how gender is self-identified, subject to first-person authority, a spectrum, and fluid. However, AGR is built on the assumption that gender is readable from someone's appearance, binary, and stable over time. While some of these assumptions can be overcome, others are more fundamental to AGR technologies. For example, while the technology could be rebuilt to include more genders than 'woman' and 'man', likely, the classification will still be done with a discrete set of genders. For example, Scheuerman et al. experimented with commercially available AGR tools on a dataset that contained seven gender categories (2019). They came to these seven categories through surveying genderqueer people and investigating which folksonomies were used most on Instagram. Nevertheless, these seven gender categories are not all-encompassing as gender is a spectrum, and everyone identifies slightly differently.

This chapter will further explicate that the nature of gender and AGR are at odds, and it is impossible to create an AGR tool that accurately and respectfully recognises subjects' gender. However, as the technology already exists and the developers and users of AGR – are set to – make large financial profits, it is unlikely that the technology will be discontinued entirely. Therefore, I am not arguing for the abolishment of AGR. Nevertheless, it is possible to make the use of AGR less harmful for people from marginalised genders. In this chapter, I will outline my recommendation for minimising the harmful consequences of the use of AGR.

Ethical Concerns

The question I will turn to now is whether it is ethically justifiable to use AGR tools. The current state-of-the-art AGR technologies disproportionately misgender people of colour and women (Buolamwini et al., 2018; Scheuerman et al., 2019). As the technology is versatile in applications, it can be deployed from cases such as airport security to human-robot interactions (Ramey & Salichs, 2014) and vending machine recommendations (Ng et al., 2019). The technology can be used to grant or deny access to resources, such as bathrooms and airports (Watts, 2019). The technology is also often meant to increase safety in public spaces; by equipping CCTV cameras with facial recognition technology, the police can keep track of the movement of suspicious people (Watts, 2019). A part of this facial recognition technology is gender recognition. However, as some people are disproportionately often misrecognised, the technology gives unequal burdens to those people. Think of being flagged as suspicious and followed by the police, denied access to a public bathroom, or frisked at airport security.

As mentioned in the second chapter, misgendering happens in many forms and situations. There would be no gendering practices in an ideal world, as it is the social structure that allows misgendering to occur. However, the reality is that gendering and misgendering occur on a daily basis and create serious harm. What makes algorithmic misgendering especially harmful is the fact that technology is often presented as correct and unbiased. This gives the user of the technology more reason to distrust the algorithmically misgendered people, especially if the user manually misgendered them, too. For example, imagine applying for a loan at the bank. As a part of the application procedure, your passport needs to be checked through facial analysis, which includes AGR software. You are transgender, and your passport reflects your correct gender (as opposed to the gender assigned at birth). The AGR tool might misgender you, making it more difficult to explain that the passport belongs to you. Such situations can occur in many aspects of daily life and make it harder for often-misgendered people to participate in society fully.

Moreover, while the bank employee in this example can misgender you just as easily as an AGR algorithm, the added layer of technological misgendering creates an extra barrier that people from marginalised genders need to deal with. The people who are subject to such burdens are the people who are already subject to oppression and marginalisation in the dominant patriarchal society. People of colour and people who are not cis-men experience oppression due to their race and gender, resulting in fewer opportunities and higher thresholds to participate in society and achieve the social position they aim to uptake. The use of AGR tools only heightens the existing thresholds and takes away the little opportunities they have. Basic activities such as freely moving around in society, gaining access to rooms and buildings, applying for loans and benefits become more difficult to access. This is unjust, as it strengthens inequality and is effectively the opposite of equity.

An objection to this might be: what if the technology would be improved to solve the disproportionality of misgendering and that people are equally misgendered? While this would create more equality, it would still not be a fitting solution, as it goes against the first-person authority as presented by Bettcher (2009) as well as the agential identity as presented by Dembroff and Saint-Croix (2019). By training the algorithms on people from different ethnicities, genders, and abilities, the algorithms would become more or equally accurate for every subject. This way, the use of AGR would not strengthen inequality, which makes that argument against the morality of AGR fall away. Nevertheless, the technology is still built on the assumption that someone can be externally gendered. It goes against the first-person authority, as the gender recognised by AGR tools is a third-person statement. The fact that a machine is empowered to determine someone's gender disregards individuals' first-person authority over their gender.

Moreover, AGR goes against agential identity, taking away the individual's agency over their gender expression. Agential identities are the self-identities that we make visible in interactions with others, so it is the self-identity that we choose to share with the people around us. However, AGR technologies can also determine someone's gender when they are not aware they are being analysed. In interactions with people, you can choose which part of your self-identity you want to show, which is less achievable with AGR technologies. An AGR tool extracts the information it needs to predict someone's gender by looking at their facial features or breasts without respecting the subject's first-person authority over their gender. While people should have authority and agency over their gender, AGR technologies fundamentally disrespect that, regardless of whether the gender recognition result was accurate or inaccurate. This is not just to say that people have privileged information regarding their gender identity, but rather that a person should have the autonomy and agency to express their identity in the way they see fit.

In sum, using AGR tools in its current form is not ethically justifiable – not even when the technology is improved such that there are no accuracy gaps for different ethnicities and genders. However, what if the technology could be improved to operationalise a definition of gender that is more inclusive and more respectful to people's autonomy and agency over their gender? For this to be a viable solution, AGR's operative concept of gender must be re-engineered to coincide with the target concept of gender. This way, the definition of gender operationalised by AGR would include people's autonomy and agency over their gender. If AGR could be reimagined to address the target definition of gender successfully, it would be safe to assume that the tool could be technologically improved in a way that prevents misgendering for everyone and places the agency with the subject.

However, the fundamental building block of AGR technology is computer vision. This includes images such as photos but also videos. The information that goes into the AGR tool must come through a camera and whatever a computer-camera combination can see is also visible to the naked eye. Computer vision cannot pick up self-identification, behaviour patterns, social hierarchies, or systemic oppression; it can merely see physical characteristics and appearances. Thus, AGR technologies are fundamentally built on the assumption that gender can be externally determined based on appearance. Agency, self-identification, and first-person authority are not visible to computer vision, so AGR cannot consider these aspects in their computations.

The other essential aspects of the target concept of gender include gender fluidity and the gender spectrum. Of these aspects, the concept of gender fluidity is best translated to computer vision. Gender fluidity means that someone's gender is not stable over time but mutable and dynamic. While this depends on the use case, AGR technologies are often deployed to capture someone's gender in one moment in time. This is quite fitting for a fluid conception of gender if there were no expectations that this moment also represents the subject's gender throughout their lives. Thus, with a shift in expectations, AGR technologies could easily be translated to capture gender fluidity. However, the gender spectrum poses a problem for AGR technologies. These types of classifiers can divide subjects into a discrete set of categories. While the number of categories can be more than two – so, more genders than just ‘man’ and ‘woman’ – there cannot be an infinite number of gender categories or even a complete lack of gender categories. AGR technologies are, thus, unable to encompass the gender spectrum fully.

Therefore, it is impossible to re-engineer AGR’s operative concept of gender to fully match the target concept of gender. The technology is fundamentally limited as it takes the authority and agency away from the subject. This renders the use of AGR technologies unethical. While improvements can be made regarding the gender categories people are grouped into or the training data that the algorithms are built on, these efforts for inclusivity are inadequate as they still take away the subject’s first-person authority and agency regarding gender expression. People (should) have first-person authority and agency over their gender, and AGR fundamentally disrespects that.

Objection From Pragmatism

As I have shown in the previous chapters, the current practices of gendering by AGR can create harmful and downright dangerous situations, especially for people of marginalised genders. Moreover, AGR technology is built assumptions regarding the nature of gender that are incorrect. As some of these assumptions cannot be overcome and they violate the authority and agency people have over their gender, the use of AGR is unethical. An objection to this argument could be that the abolition of AGR is unrealistic due to the power and interests of the most prominent stakeholders: the developers and users of AGR technologies.

The companies and governments that determine the future of AGR software inhabit a powerful place in society. They are set to make a significant monetary profit from the creation and use of AGR tools as they can be used to automate tedious tasks. It offers the users time efficiency and operational cost reduction (O’Neill, 2021). Their algorithms are adopted in cases from airport security to personalised product recommendations on Instagram. The US Customs and Border Protection (CBP) has stated that by the end of 2021, all international passengers – including US citizens – will be analysed using different facial recognition technologies, such as AGR, at twenty airports across the US (Alba, 2019). The benefits the CBP have mentioned range from creating efficiency, security, and accuracy (CBP, 2021). However, these benefits are more economical than moral – they do not justify the use of AGR. The only benefits that are more of a moral nature are the supposed increased safety of the travellers and more face-to-face time between the traveller and the border control officer. However, while the faces of over two million passengers spread over 15,000 flights have been analysed, only one arrest has taken place in the US (Reservations, 2019). Moreover, the AGR technology the CBP uses needs 2 seconds to analyse a person’s face with an accuracy of 99%. This is not significantly faster nor more accurate than human border control officers.

Nevertheless, the power these companies have makes these racial and gender bias issues so problematic. A decision-making algorithm can determine whether you will be stopped and searched at the airport (CBP, 2021), the size of the loan you applied for (Zednik, 2019), or even the length of your prison sentence (Angwin et al., 2016). Moreover, the innovations in artificial intelligence are done at a high tempo and are often untransparent, such that government policy and regulations cannot keep up. Additionally, as the current state of AGR works in favour of the people from dominant gender and ethnic groups, they do not experience any further oppression due to the technology. People who develop and deploy the technology are often members of these dominant genders and ethnic groups, so they are unlikely to be intrinsically incentivised to improve AGR technologies. These are solid reasons for the abolition of AGR tools, but also reasons why this abolition might not be realistic.

Thus, while the use of AGR technologies is unethical, the technology exists and is unlikely to be discontinued in the short term. Instead, we need to find a different way to minimise the harm AGR technology causes. In the next section, I will outline a proposed approach to improve the situation for people from marginalised genders who are misgendered by AGR tools. My recommendation involves educating the users of AGR tools and introducing new vocabulary to increase the awareness of the possible inaccuracy of the AGR results.

A Workaround

In the previous sections, I have argued that neither re-engineering AGR's operative concept of gender nor the abolition of automatic gender recognition tools are feasible goals. In this section, I will introduce a third option. This is the middle ground between the improvement and the abolition. The central tenet of this solution is education regarding the nature of the output of AGR tools. Essentially, I will argue that the only way to avoid harmful misgendering is by creating sensitivity and understanding among the users of the tools. This can be done by creating a new vocabulary to reflect and awaken this sensitivity, for two reasons.[4] First, the correct language can act as a constant reminder of the exact nature of the topic. For example, as climate change became a more urgent and alarming matter, the British newspaper The Guardian changed its language and started referring to the issue as a "climate crisis" (Carrington, 2019). This was done to increase the public's awareness regarding the severity of the matter. Second, the correct language can minimise the harmful consequences of misgendering, as language is often the cause of these issues. Language plays a central role in acts of gendering; think of the pronouns we use to refer to someone.

As mentioned before, it is not uncommon that AGR tools fail to recognise an individual's gender correctly. This leaves us with two conflicting pieces of information. On the one hand, we have the gender

[4] I have also argued for the new vocabulary and the separation of data based on their epistemic properties in an essay for the MSc Philosophy of Science, Technology, and Society. The title of that essay was "Beware the prediction: distinguishing algorithmic classification from self-determination." It was written for the completion of the course Transformations of Knowledge in the Digital Age.

as determined by the subject, and on the other hand, we have the gender as predicted by the AGR tool. This can spark a discussion regarding which piece of information is correct. However, as shown in the second chapter, gender speech acts – assigning a gender to someone through language such as “I am a woman” – is subject to first-person authority. This is to say that the information that the individual gives must be respected and used. The gendering statements done by anyone else than the subject themselves should be disregarded unless they align with the subjects’ statements. In the case of conflicting statements regarding someone’s gender, using different words would eliminate the question of which one of the two is correct, as the agential identity and first-person authority are reflected in the terms themselves. Moreover, when predicting something as sensitive as gender identity, we need to be able to talk about the pieces of information in a way that reflects the sensitivity of the topic. A step in the right direction would be to introduce new vocabulary to describe information that is subject to agential identity and first-person authority and the data that is potentially mistaken.

The results that are put forth by algorithms are per definition probabilities (Khan et al., 2019). I will name this kind of data ‘probabilistic information’. As this type of information is essentially a likelihood and not a certainty, the epistemic status of the probabilistic information is weak. It is essential to acknowledge that such information can be fallacious and should not be taken to reflect reality. On the other hand, the information that is put forward by the subjects themselves can be depended upon due to their first-person authority and agential identity. While this is not a metaphysical but an ethical claim, the reliability of the information is unaffected, and first-person gender statements must be respected. The epistemic status of this data is more robust than that of the probabilistic information. I will denote the data that reflects the subject’s conception of their gender identity as ‘agential information’ to reflect Dembroff and Saint-Croix’s agential identity (2019). Thus, as the probabilistic and the agential information have different epistemic statuses, it is essential to keep the pieces of information apart. If the probabilistic information is treated as a reflection of reality, AGR software can seriously harm the most marginalised groups due to the possibility of inaccuracy. The use of different words will help to prevent such harm.

Imagine a situation at airport security, where a passport control machine equipped with AGR has assigned someone with a gender other than the one in their passport. A security guard is then required to come and assess the situation. If the gender as predicted by the AGR tool is taken as truthful, the guard will interrogate the traveller on their gender or even take them into a separate room. If the way Caster Semenya has been treated to prove her gender is any indication (Court of Arbitration for Sport, 2019), such interrogations can become very invasive. Semenya had to undergo a plethora of tests to prove her gender, which included analysing her chromosomes. This is a form of disrespect (Dembroff & Wodak, 2018) and even sexual abuse (Bettcher, 2009). Semenya is an example of someone whose gender identity was not socially accepted and who was subjected to invasive tests designed to determine her gender scientifically. While this was an exceptional case, the use of AGR technologies might pave the way for more invasive gender recognition technologies to become normalised, even at airport security.

If the airport security guard were taught that the agential information is per definition correct and should be respected, they would become more likely to approach these situations with sensitivity. The guard would learn that the AGR tools and the corresponding results are merely probabilities that reflect a prediction of what the gender of the subject could be. The guard would then be incentivised to listen to the subject and consider whether the AGR result was inaccurate. Thus, using vocabularies such as probabilistic and agential information can help respect and protect the gender identity of the people who are misgendered due to AGR technologies. While this scenario might not be very realistic, I am painting this picture to show how the benefits of AGR could be reaped without the risk of creating room for harmful consequences. Not only the developers of AGR but also the direct users and subjects must be aware of the fact that the results are probabilities that might be incorrect.

In sum, while the very use of AGR is detrimental to respecting first-person authority and agential identity due to the automation of external gendering, we can still use the technology to speed up time-consuming tasks. In the airport security case, the traveller will be gendered either by a person or by a machine; there is no situation where the traveller would *not* be externally gendered. However, the AGR machine might be much more time and cost-efficient. Both the AGR tool and the security guard can make mistakes, and the crucial point is that this must be acknowledged. The solution is to manage the expectations surrounding the accuracy of AGR tools. We should not see AGR as something that can *determine* someone's gender but only as a tool that can potentially recognise someone's gender. In case of conflict, the subject is always right.

Moreover, we can also limit the use cases; in some use cases, there seems to be no significant benefit, neither moral nor economical. For example, AGR tools are used in human-robot interactions to enable the robot to refer to the user with the correct honorifics and pronouns, especially in social and care robots (Ramey & Salichs, 2014). According to Ramey and Salichs, the aim is to create a natural interaction and make the human feel comfortable interacting with the social robots. These are not interactions that need to be made more efficient; instead, they need to be made more inclusive and sensitive to the identity and lived experiences of the users. Thus, in such a case, it makes more sense to program the robot to ask for the user's preferred pronouns and honorifics as there is no added value of efficiency necessary.

The Benefits of Mindful AGR Use

In this chapter, I have argued that the use of AGR tools is unethical. The current state-of-the-art AGR disproportionately misgenders people from marginalised genders and disrespects the first-person authority and agency people have over their gender. Externally and deliberately misgendering – whether done by a human or a machine – inherently assumes that the subject's conception of their gender identity is negligible and can be disregarded. While the technology could be improved to no longer disproportionately misgender marginalised people, the disrespect towards the individual's agency and first-person authority cannot be overcome. AGR tools can only work with the information registerable by computer vision, which is limited to physical characteristics and appearances. Thus, AGR technology fundamentally disrespects the first-person authority and agency people have over their gender identity.

Additionally, there are no ethical advantages to using AGR technologies; there are only economic benefits such as time and cost-efficiency. Nevertheless, due to these economic benefits, it is unlikely that the technology will be discontinued or even abolished. To minimise the harmful consequences that come from the use of AGR, we can introduce new vocabulary to denote the difference between the information that is contributed by the subjects themselves and the information that is recognised with the use of AGR algorithms. As I have discussed in the second chapter, the piece of information regarding gender identity that must be respected is that which the individuals put forth themselves. The statement 'I am a woman' is an ethical claim as it represents how the subject wants to be seen and treated (Bettcher, 2009). Thus, only the subject themselves is in a position to classify their gender. To reflect this, I introduced the terms agential information and probabilistic information, where the former is self-identified by the subject, and the latter is computed by an AGR tool.

AGR tools must be seen as a hermeneutical tool rather than a representation of reality. The developers, users, and subjects must become aware of several notions: first, this is a sensitive topic and must be approached as such. Second, the information that is self-identified by the subjects is inherently correct and must be respected. Third, the probabilistic information is possibly inaccurate and must be not be depended upon. While the developers are likely aware that their tools produce probabilities, we cannot assume this is the same for the users and the subjects. Moreover, even when people are aware of the probabilistic nature of AGR results, they might not be aware of the agency and the first-person authority people have over their gender identity. Mindful use of AGR tools requires the introduction of appropriate vocabulary and the education of the developers, uses, and subjects regarding the epistemic and ethical statuses of these terms. This way, AGR tools can continue to be deployed to execute otherwise time-consuming and expensive work without depending on the technology entirely.

This is also in line with the three design recommendations proposed by Hamidi et al. (2018). They argued that people should be informed when and where they are subject to the use of AGR tools and should have the positive freedom to opt out. Moreover, the subjects ought to have the opportunity to define their own gender identity, even when that identity does not fall in the pre-defined set of gender categories that are included in the database. Finally, gender recognition systems should be built using a diverse set of faces that are representative of the population. That way, the likelihood of marginalised people being correctly gendered becomes significantly larger.

The separation of the agential information and the probabilistic information can even help to improve the accuracy of AGR software, as the agential labels can be used in supervised learning processes. With this added level of awareness, the developers of the tools are forced to be mindful that their algorithm potentially produces inaccurate predictions. The next step would be to make automatic gender recognition software fairer and more inclusive, for example, by diversifying the data (Merler et al., 2019) or operationalising a more nuanced concept of gender (Keyes, 2018). However, the key to this research and development must be the awareness that there is an epistemic difference between agential and probabilistic data.

Conclusion

This thesis argued that using automatic gender recognition (AGR) tools is unethical and harmful even in their most accurate and fair form. As I have shown in the first chapter, AGR technology is built on the assumptions that gender can be externally determined based on one's appearance, that gender is a binary categorisation, and that it is stable over time. These are the manifest and operative concepts that drive the development of AGR. In the second chapter, I argued that this assumption is incorrect, as gender is based on self-identity and people have authority over their gender and their gender alone. Finally, in the third chapter, I argued for the immorality of AGR technologies.

AGR often leads to harmful misgendering, especially of people who are already marginalised and oppressed due to their gender. To prevent further harm, one could consider significantly altering the technology such that it steers away from the gender binary and becomes capable of recognising other genders than 'man' and 'woman'. This way, it would not further marginalise people based on their gender. However, such technologies still presuppose that gender is something that external people and machines can determine. The use of AGR, even when it is more inclusive and fairer towards marginalised genders, disrespects the agency and first-person authority people have over their gender. As this first-person authority is ethical and we must respect someone's self-identified gender, this renders the use of AGR unethical.

An objection against this argument could be that the abolition of AGR seems unrealistic as the powerful governments, institutions, and companies that benefit from AGR are against the abolition of the technology. However, the time and money that can be saved using AGR create mere economic benefits. These are not moral validations for the use of AGR. Thus, these justifications do not curb the immorality of the use of AGR technologies. The use of AGR cannot be ethically justified as it disrespects the first-person authority that people have regarding their gender, regardless of the economic benefits it promises.

However, it is sensible to keep these economic benefits in mind, as they are likely to stand in the way of the abolition of AGR. This is to say that the institutions and corporations that use AGR are unlikely to discontinue using the technology merely because of ethical reasons. Some companies, such as IBM, have refrained from developing facial analysis technologies after it came to light that their technology misidentifies people of colour and women significantly more often than white men (Denham, 2020). Nevertheless, the technology exists and can be developed and used by any party with a reasonably powerful computer and a camera. While some groups argue for legislative guidelines and seem to be getting through to legislators (Delcker, 2020), this is a long-term project. As there are people at risk of experiencing the harmful consequences of misgendering due to AGR technologies at this very moment, we also need a short-term solution.

The solution I propose is based on the re-education of the users and developers of AGR. In situations of conflict between someone's self-identified gender and the gender as recognised by AGR, it is essential to remember that the results from AGR technologies are per definition predictions. These predictions are probabilities, meaning that there is a chance the result is correct, but there is also a chance the result is incorrect. On the other hand, individuals have first-person authority and agency over their gender identity, meaning their first-person statements are to be respected in any situation. When these statements regarding someone's gender are at odds, the first-person statement is always correct.

Moreover, while AGR technologies can merely recognise a discrete set of gender categories – often only the gender binary – there is an infinite number of genders someone can have, namely on the gender spectrum. As this gender spectrum goes against AGR's classification system, the users must be aware that the AGR tool might not recognise the subject's gender at all. While the developers are likely to be aware of the probabilistic nature of the AGR predictions, we cannot say the same for the users and the subjects. Additionally, even if the developers are aware of the possible inaccuracy of the AGR results, they might not be aware of the agency and the first-person authority people have over their gender.

Thus, we must re-educate AGR users and developers about the target definition of gender – including the gender spectrum, the fluidity of gender, agential identity, and first-person authority – and the nature of the predictions that AGR machines output. This way, the conflicts between gender statements will not become conflicts of authority. We can use new vocabulary to reflect the respective authorities. The first-person statements reflect agential information, and the results presented by AGR are a form of probabilistic information. While the probabilistic information can coincide with the agential information, they can also differ. Nevertheless, these words reflect that the agential information is always correct. This way, AGR technologies can still be used to automate time-consuming and expensive tasks while curbing the harmful consequences of misgendering people from marginalised groups. Using such wording reminds us that we must treat people with respect, agency, and first-person authority, especially when an AGR machine misgenders them.

List of References

Alba, D. (2019, March 11). *The US Government Will Be Scanning Your Face At 20 Top Airports, Documents Show*. Buzzfeed News. https://www.buzzfeednews.com/article/daveyalba/these-documents-reveal-the-governments-detailed-plan-for

Anusha, A. V., Jayasree, J. K., Bhaskar, A., & Aneesh, R. P. (2016). Facial expression recognition and gender classification using facial patches. *2016 International Conference on Communication Systems and Networks (ComNet)*, 200–204. https://doi.org/10.1109/CSN.2016.7824014

Bach, T. (2012). Gender Is a Natural Kind with a Historical Essence. *Ethics*, *122*, 231–272. https://doi.org/10.1086/663232

Bettcher, T. M. (2017). Trans 101. In R. Halwani, A. Soble, S. Hoffman, & J. M. Held (Eds.), *The Philosophy of Sex: Contemporary Readings* (Seventh Edition, pp. 119–138). Rowman & Littlefield.

Bettcher, T. M. (2009). Trans Identities and First-Person Authority. In L. Shrage (Ed.), *You've Changed: Sex Reassignment and Personal Identity*. Oxford University Press.

Bowcott, O. (2020, March 10). Lack of gender-neutral passports is lawful for now, says appeal court. *The Guardian*. https://www.theguardian.com/world/2020/mar/10/lack-of-gender-neutral-passports-is-lawful-for-now-says-appeal-court

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, *81*, 77–91.

Carrington, D. (2019, May 19). Why the Guardian is changing the language it uses about the environment. *The Guardian*. https://www.theguardian.com/environment/2019/may/17/why-the-guardian-is-changing-the-language-it-uses-about-the-environment

Court of Arbitration for Sport. (2019, May 1). CAS Arbitration: Caster Semenya, Athletics South Africa (ASA) and International Association of Athletics Federations (IAAF): Decision [Media Release]. Available from: https://www.tas-cas.org/en/general-information/news-detail/article/cas-arbitration-caster-semenya-athletics-south-africa-asa-and-international-association-of-athl.html

Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence AI and Autonomy Track*, 4691–4697. https://doi.org/10.24963/ijcai.2017/654

Delcker, J. (2020, January 16). EU considers temporary ban on facial recognition in public spaces. *Politico*. https://www.politico.eu/article/eu-considers-temporary-ban-on-facial-recognition-in-public-spaces/

Dembroff, R., & Saint-Croix, C. (2019). 'Yep, I'm Gay': Understanding Agential Identity. *Ergo*, *6*(20). https://doi.org/10.3998/ergo.12405314.0006.020

Dembroff, R., & Wodak, D. (2018). He/She/They/Ze. *Ergo*, *5*(14). https://doi.org/10.3998/ergo.12405314.0005.014

Denham, H. (2020, June 11). IBM's decision to abandon facial recognition technology fueled by years of debate. *The Washington Post*. https://www.washingtonpost.com/technology/2020/06/11/ibm-facial-recognition/

FRA Focus. (2019). *Facial recognition technology: Fundamental rights considerations in the context of law enforcement*. (2019). European Union Agency for Fundamental Rights. https://fra.europa.eu/en/publication/2019/facial-recognition-technology-fundamental-rights-considerations-context-law

Geetha, A., Sundaram, M., & Vijayakumari, B. (2019). Gender classification from face images by mixing the classifier outcome of prime, distinct descriptors. *Soft Computing*, *23*(8), 2525–2535. https://doi.org/10.1007/s00500-018-03679-5

Hamidi, F., Scheuerman, M. K., & Branham, S. M. (2018). Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition Systems. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 8:1-13. https://doi.org/10.1145/3173574.3173582

Haslanger, S.A. (2006). What good are our intuitions: Philosophical analysis and social kinds.

Aristotelian Society Supplementary Volume 80 (1):89-118.

Haslanger, S.A. (2012). *Resisting reality: Social construction and social critique*. Oxford University Press.

Haslanger, S.A. (2020). Going On, Not in the Same Way. In: A. Burgess, H. Cappelen, & D. Plunkett (Eds.), *Conceptual Engineering and Conceptual Ethics* (pp.230-260). Oxford Scholarship Online. doi: 10.1093/oso/9780198801856.003.0012

Jenkins, K. (2016). Amelioration and Inclusion: Gender Identity and the Concept of *Woman*. *Ethics*, *126*, 394–421. https://doi.org/10.1086/683535

Keyes, O. (2018). The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 88:1-22. https://doi.org/10.1145/3274357

Khan, K., Attique, M., Syed, I., & Gul, A. (2019). Automatic Gender Classification through Face Segmentation. *Symmetry*, *11*(770), 1–14. https://doi.org/10.3390/sym11060770

Khan, S.A., Ahmad, M., Nazir, M., & Riaz, N. (2014). A Comparative Analysis of Gender Classification Techniques. *Middle East Journal of Scientific Research*, *20*(1), 1–13. https://doi.org/10.5829/idosi.mejsr.2014.20.01.11434

Kumar, V., Raghavendra, R., Namboodiri, A., & Busch, C. (2016). Robust transgender face recognition: Approach based on appearance and therapy factors. *2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, 1–7. https://doi.org/10.1109/ISBA.2016.7477226

Lin, F., Wu, Y., Zhuang, Y., Long, X., & Xu, W. (2016). Human gender classification: A review. *International Journal of Biometrics*, *8*(3/4), 275–297. https://doi.org/10.1504/IJBM.2016.10003589

Mahalingam, G., & Ricanek, K. (2013). Is the eye region more reliable than the face? A preliminary study of face-based recognition on a transgender dataset. *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 1–7. https://doi.org/10.1109/BTAS.2013.6712710

Merler, M., Ratha, N., Feris, R. S., & Smith, J. R. (2019). Diversity in Faces. *ArXiv:1901.10436 [Cs]*. http://arxiv.org/abs/1901.10436

Mikkola, M. (2009). Gender Concepts and Intuitions. *Canadian Journal of Philosophy*, *39*(4), 559–583. https://doi.org/10.1353/cjp.0.0060

Ng, C.-B., Tay, Y.-H., & Bok-Min, G. (2015). A review of facial gender recognition. *Pattern Analysis and Applications*, *18*, 739–755. https://doi.org/10.1007/s10044-015-0499-6

O'Neill, P. (2021, May 19). *INTERVIEW: Thales' Youzec Kurp on Biometric Security at Schengen Borders*. FindBiometrics. https://findbiometrics.com/interview-thales-youzec-kurp-biometric-security-schengen-borders-705199/

Radtke, H., & Stam, H. (1994). *Power/gender: Social relations in theory and practice. Inquiries in social construction.* Sage.

Ramey, A., & Salichs, M. A. (2014). Morphological gender recognition by a social robot and privacy concerns: Late breaking reports. *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, 272–273. https://doi.org/10.1145/2559636.2563714

Reservations. (2019). *Survey: 43% of Americans Approve, 33% Disapprove of Facial Recognition Technology in Airports*. https://www.reservations.com/blog/resources/facial-recognition-airports-survey/

Santarcangelo, V., Farinella, G. M., & Battiato, S. (2015). Gender recognition: Methods, datasets and results. *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 1–6. https://doi.org/10.1109/ICMEW.2015.7169756

Scheuerman, M. K., Paul, J. M., & Brubaker, J. R. (2019). How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 144:1-33. https://doi.org/10.1145/3359246

Scheuerman, M. K., Wade, K., Lustig, C., & Brubaker, J. R. (2020). How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis.

Proceedings of the ACM on Human-Computer Interaction, *4*(CSCW1), 58:1-35. https://doi.org/10.1145/3392866

Schilt, K., & Westbrook, L. (2015). Bathroom Battlegrounds and Penis Panics. *Contexts*, *14*(3), 26–31. https://doi.org/10.1177/1536504215596943

Schwemmer, C., Knight, C., Bello-Pardo, E. D., Oklobdzija, S., Schoonvelde, M., & Lockhart, J. W. (2020). Diagnosing Gender Bias in Image Recognition Systems. *Socius: Sociological Research for a Dynamic World*, *6*, 1–17. https://doi.org/10.1177/2378023120967171

US Customs and Border Protection [CBP]. (2021). Say hello to the new face of speed, security and safety: Introducing Biometric Facial Comparison. https://biometrics.cbp.gov/

Vijayan, A., Kareem, S., & Kizhakkethottam, J. J. (2016). Face Recognition Across Gender Transformation Using SVM Classifier. *Procedia Technology*, *24*, 1366–1373. https://doi.org/10.1016/j.protcy.2016.05.150

Viola, P., & Jones, M. J. (2004). Robust Real-Time Face Detection. *International Journal of Computer Vision*, *57*(2), 137–154. https://doi.org/10.1023/B:VISI.0000013087.49260.fb

Watts, R. (2019). Facial recognition as a force for good. *Biometric Technology Today*, *2019*(3), 5–8. https://doi.org/10.1016/S0969-4765(19)30039-6

Wing Sue, D. (2010). *Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation*. Wiley.