When AI is Gender-biased: The Effects of Biased AI on the Everyday Experiences of Women

Galit P. Wellner[†]
galit_wellner@yahoo.com

ABSTRACT

AI algorithms might be gender biased as evidenced from translation programs, credit calculators and autocomplete features, to name a few. This article maps gender biases in technologies according to the postphenomenological formula of I-technology-world. This is the basis for mapping the gender biases in AI algorithms, and for proposing updates to the postphenomenological formula. The updates include refereces to I-algorithm-dataset, and the reversal of the intetionality arrow to reflect the lower position of the human user. The last section reviews three ethical analyses for AI algorithms - distributive justice, ethics of care and mediation theory's ethics.

1. Introduction

Most people find it hard to believe that computers can be gender biased. Computers, and especially their software, are conceived as neural and impartial. Yet, the potential for bias exists, as the interaction with computers has been based on alphanumeric texts since the early days of digital computing: first in the form of punch cards, then through the mediation of a keyboard and a screen, and today also via speech that is deciphered and coded into text. We type in numbers and words in word processors, spreadsheets and search engines; run Optical Character Recognition (OCR) to let the computer understand the words appearing in a scanned document or a picture as if we typed them; and give oral commands or use voice menus (e.g., when contacting airline companies and the like), to name a few. In parallel, inside the computer, many of the processes deal with texts, such as data storage, spelling checkers or search engines.

[†] Tel Aviv University and Bezalel Academy of Art and Design, Israel.

ISSN: 1972-1293

The integration of artificial intelligence (AI) algorithms[1] diverted these processes to a new direction, sometimes praised to be human-like. The area of specialty - termed Natural Language Processing (NLP) - aims to make computers understand human language and be able to conduct a conversation with humans like humans. This is how we interact with chatbots, navigation systems and the like.

The developments on the interface and on the processing domains lead us to believe that any computer-based interaction will provide us with exact information tailored to our needs while remaining neutral. In other words, we perceive the whole process to be beneficial to us, and at the same time impartial. Moreover, we trust the computer much more when the interaction with it mimics human-to-human interaction. Accordingly, we expect the interaction to be pleasant, or at least not irritating.

For one category of users these promises are kept. It is the category known as young white male users. Female users, however, are less likely to enjoy from interactions that respect their identity and provide equal access to opportunities. On the interface level, female users are frequently approached as if they are male, and many became accustomed to such a way of addressing them or referring to them. This phenomenon is more frequent in languages where nouns, adjectives, verbs and other grammatical components change according to the gender of the subject. In Hebrew, for example, in the sentence "I am a teacher" the word "teacher" will be different for male and female subjects. When translating it from English to Hebrew, Google Translate uses only the male form. When I type in "I am a teacher" the algorithm disregards the fact that in Israel, the only country where Hebrew is an official language, a majority of the teachers are female. It also disregards my identity as a user of the algorithm, even when I am logged in with my Gmail account. This is not an inherent limitation of the algorithm. Other algorithms of the same company, Google, can do that. The company′s advertisement placement algorithms personalize the offers according to my gender, among other parameters.

Another example is WhatsApp and its auto-correct mechanism. When I type in a message for my female friends, sometimes the algorithm auto-corrects my text from female to male form. Needless to say that both forms are appropriate and legitimate in the Hebrew language.

[1] In this article, algorithms are understood as software artefacts involved in computerized data processing (see (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016)).

The gender bias can be evidenced even in languages that have a neutral form like Turkish. In 2017 *Quartz Magazine* conducted an experiment translating a list of professions, occupations and descriptions from that form in Turkish to English. Google Translate′s algorithm returned for the majority of professions the male variations and few, mostly stereotypical, were female, e.g. a nurse and a nanny.[2] The experiment also tested attributes, just to discover that even the attribute "old" is referred to men, although usually most older adults are female. Furthermore, the test revealed that positive attributes such as "very beautiful" and "hard working" were translated to the male form, whereas "lazy" was translated as a female attribute. *Quartz Magazine* noted that this phenomenon occurs also in other languages such as Chinese and Finnish.

This grim situation has not escaped the attention of computer scientists. In 2017 Aylin Caliskan, Joanna Bryson and Nara Arvind (2017) attempted to provide an explanation by reviewing the popular datasets with which AI algorithms are trained. A dataset is the basis on which algorithms statistically assess which word is more likely to appear next to others. They explained that these datasets are usually based on the texts of World Wide Web. When examining word proximity, the researchers found that female names such as "woman" and "girl" usually appeared near words indicating family, whereas male names appeared near words indicating career. Similarly, words indicating females were more associated than male words with the arts than with mathematics or science. This proximity can explain the 2017 experiment in Turkish. In their article, the researchers asserted that "the statistical contexts of words capture much of what we mean by meaning" (Caliskan, Bryson, & Arvind, 2017, p. 185). The reduction of meaning to statistical proximity allows algorithms to "understand" text. Hence, the researchers claimed that it is possible that "all implicit human biases are reflected in the statistical properties of language" (p. 185). Simply put, Caliskan et al argue that the gender bias in AI algorithms should be inferred to the datasets with which algorithms are trained. In their view the algorithms remain neutral.

But sometimes the fault is in the algorithms themselves. James Zou and Londa Schiebinger (2018) assert that most AI algorithms amplify statistical appearance. They explain: "If a specific group of individuals appears more frequently than others in the training data, the program will optimize for those individuals because this boosts overall accuracy" (p. 325). Additionally, due to the

[2] https://qz.com/1141122/google-translates-gender-bias-pairs-he-with-hardworking-and-she-with-lazy-and-other-examples/ (accessed 1 December 2019).

feedback loop mechanism, the gender-biased results are fed back to the system, thereby deepening the biases. They specifically refer to Google Translate′s tendency to use the masculine form and warn that it can "potentially revers[e] hard-won advances towards equity" (p. 325).

So far, I described the gender bias in AI algorithms in the domain of language. However, gender bias exists in other types of algorithms, such as financial algorithms. There the bias does not reveal itself on the interface level but remains hidden in the processing. Financial algorithms′ racial biases were revealed by Cathy O′Neal in her book *Weapons of Math Destruction* (2016) where she showed how racial biases were built into the algorithms that calculated loan rates and insurance premiums. Gender biases in these algorithms recently gained traction when Apple together with Mastercard and Goldman Sachs launched a credit card.[3] A high-tech entrepreneur checked the new credit card and was surprised to find out that his wife received a much smaller credit limit than he did, although she is a permanent US citizen and he is a just a temporary resident (a green-card holder), that is—he is riskier for American banks. Moreover, as a couple they jointly file their taxes and according to the law all their property is jointly and equally owned. Based on their data, they should have received the same credit limit. After he tweeted this finding, Steve Wozniak, one of Apple founders, checked the credit limits of himself and his wife, just to realize that he was given credit ten times higher than his wife, although they have no separate bank accounts and they jointly hold all their assets. In both cases, the major difference is the gender of the credit card holder.

The gender bias exists also in the income side. Arianne Barzilay and Anat Ben-David (2016) examined the revenues generated through gig-economy platforms such as Uber. In such platforms, the boss is the algorithm. It is the algorithm that decides who gets an opportunity to contact a customer, and it is the algorithm that determines the rates. Their empirical findings revealed that women worked more hours in general "on the platform," but their average per-hour income was only two thirds of that of men. In another study on eBay, Tamar Kricheli-Katz and Tali Regev (2016) found that women sellers received a smaller number of bids and lower final prices than did equally qualified male sellers of the exact same product. They also found that on average, women sellers

[3] https://qz.com/1745842/a-regulator-is-looking-into-whether-apples-credit-card-is-sexist/ (accessed 1 December 2019).

received 80 cents for every dollar a man received when selling identical new products and 97 cents when selling same used products.

These examples of the gender bias in AI-based systems refer to technologies that are deeply involved in many of our daily activities. They led me to examine the experience of gender-biased algorithms in everyday life. My analysis is grounded in postphenomenology, a branch of philosophy of technology that models our relations with technologies and through them with the world. The first section of this article serves as an introduction to this theory by detailing the basic four postphenomenological relations – embodiment, hermeneutic, alterity and background relations. This framework serves to map the gender biases in technologies, whether pre-planned or accidental. In the next section, I investigate the changes required in the postphenomenological relations to accommodate them to a reality dominated by AI. Since most of algorithms are developed as trade secrets and their developers are subject to non-disclosure agreements, we do not have access to the intentions of the designers and the developers. The companies who own the algorithms are likely to claim that the effects are unintended. When insiders like Cathy O′Neal describe the processes, intensions are revealed. Until such whistle blowers appear for gender biases, my focus is on the consequences, intended or unintended. In either case, the biases should be cured, no matter if they were planned in advance or became an unfortunate outcome. In the last section, the ethical implications of gender-biased AI algorithms is studied. Three ethical frameworks are investigated: distributive justice, ethics of care and morality of objects. Each has already been implemented in philosophy of technology, but the implications of AI are *terra incognita.*

2. "Classical" Postphenomenology and Gender Bias

Postphenomenology analyzes the relations between humans, technologies and the world (Ihde, 1990); (Verbeek, 2005)). The theory is named *post*-phenomenology because it is based on phenomenology, i.e. – the study of our experience in and of the world (it also relies on pragmatism – see (Ihde, 2009); (Rosenberger, 2017); (Langsdorf, 2020)). The "post" prefix means that the theory aims at extending phenomenology, not at reversing it. The need for *post*-phenomenology arises as phenomenology limitedly deals with technology although today most of our experiences of the world are mediated by technology to various extents. Thus, postphenomenology extends phenomenology by analyzing the role of technology as a mediator. This mediator is not passive or neutral.

Technology mediates the world and in doing so it transforms the user and her environment. To model these transformations, Ihde (1990) developed a formula composed of three elements: I, technology and world. The technology is in the middle, mediating between the experiencing "I" and the world. These three elements are connected by dashes, arrows and parentheses: the dash indicates a link between two elements; the arrow represents intentionality in the sense of directedness, i.e. a form of connectedness to the surrounding; and the parenthesis signifies that two elements function as a single unit or withdraw to the background. The following detailed review of the relations will demonstrate how the formula and its ingredients work. For each relation there will be one example that refers to a gender bias emerging from the relation.

Usually the first postphenomenological relation is the embodiment one. In this type of relations, the human user and the technological artifact act jointly as a unit in the world. This is our experience when we wear clothes (or fitness bracelet, to use a more contemporary example), ride bicycles or look through a microscope. In all these situations we behave as if these technologies are part of our body. The postphenomenological formula to represent them is:

(I – technology) → world

The "I" and the "technology" are united by the parentheses and together they are directed to the world. In the car, the safety belt should maintain embodiment relations with the passengers. Women, however, do not feel comfortable in these relations due to bad fit to the female chest (see (Michelfelder, Wellner, & Wiltse, 2017)). Had the safety belt been designed to fit all passengers, its receptiveness could have been higher, and more lives could have been saved.

The second type of relations is termed hermeneutic relations because they involve interpretation and meaning generation. They also involve some kind of reading, and here "reading" and "text" are broadly construed. The text can obviously be alphabetic, as in this article that you read now, but it can also be a graphic representation as in the case of a water gauge indicating how much rain fell in an hour or a day. The postphenomenological formula for hermeneutic relations looks like this:

I → (technology – world)

The technology and the world seem like a unified entity in which the technology simply reflects the world. Obviously, the technology is never neutral and it always has some biases. When a technology mediates the world hermeneutically, one has to be conscious of those biases and take them into account when thinking of the world "out there." If language is regarded as a technology, its usage of gender is never neutral. For instance, in Hebrew assembly instructions are frequently directed to a male constructor, whereas recipes to a female cook. Once the gender bias was revealed, more and more instructions and receipts are directed to a neutral form of "you" in the plural or use the base form of verbs to avoid any gender connotation.

The third relation is alterity in which technology is referred to as a "quasi-other." The technological device is a partner for a dialog, even if it does not answer as humans do. We experience alterity relations when interacting with an ATM, when thinking of our car or computer in terms of "s/he" or when asking a voice assistant to buy something, report the weather forecast or play a song. Alterity relations are represented by a permutation of the postphenomenological formula in which the world is in parentheses to denote that we do not pay attention to it:

I → technology (– world)

Conceiving robots and chatbots in terms of alterity relations immediately raises the gender question – which gender should be assigned to the technology?

The last relation that Ihde discusses is background relations which is a kind of a reverse mirror to alterity relations. Here it is the technology that "withdraws to the background," to use Heidegger′s famous phrase. The focus is the world, and the technology serves as no more than a background for it. For example, the chair I sit on, the light I use, the Internet connection, all these form the background against which I write this article. The formula is:

I → (technology –) world

These relations reflect social and cognitive norms (see (Michelfelder, Wellner, & Wiltse, 2017)) that are usually hidden – from most people most of the time. In her seminal study, Ruth Schwartz-Cowan (1976) reveals how the design and marketing of household appliances at the beginning of the twentieth century was imbued with gender bias. These appliances, together with the then-dominant beliefs, put women in the position of a housekeeper, the one that cleans, cooks, irons etc.

Note that in all four relations, the intentionality arrow is directed from the experiencing "I" towards the technology and/or the world. As the arrow represents human intentionality, the arrow points *from* the "I" to the other components. This tendency reflects an anthropocentric point of view in which the human is the active agent around whom the world and the technology revolve. As we shall see in the next section, this assumption is dramatically eroded in the age of AI.

The four basic postphenomenological relations, which were originally formulated by Ihde, were expanded by Peter-Paul Verbeek (2008) through the notion of technological intentionality. Similarly to human intentionality, " 'technological intentionality' here needs to be understood as the specific ways in which specific technologies can be directed at specific aspects of reality" (2008, 392). Technological intentionality is interpreted as the ability of technologies to form intentions so that they direct the users to do things which were unthinkable in the absence of such a technology. There is obviously a difference between human and technological intentionality: "even though artifacts evidently cannot form intentions entirely on their own, . . . because of their lack of consciousness, their mediating roles cannot be entirely reduced to the intentions of their designers and users either" (2008B, p. 95). For Verbeek, technological intentionality supports human intentionality so that "When mediating the relations between humans and reality, artifacts help to constitute both the objects in reality that are experienced or acted upon and the subjects that are experiencing and acting" (95). This development of technological intentionality forms the foundation for the next section where the postphenomenological formula is playfully altered to accommodate it to the landscape of AI technologies.

3. Postphenomenology of AI and Gender Bias

There are additional permutations of the postphenomenological formula that attempt to characterize the various ways in which we interact with contemporary technologies (e.g. (Wiltse, 2014); (Liberati, 2016); (Wellner, 2017); (Wellner, 2018)). The new permutations deal with the new capabilities afforded by twenty-first century′s technologies from various perspectives. Realizing that AI technologies attempt to think and decide, I introduce in this article another set of permutations. They represent our relations with AI technologies that are loaded with quasi-mental capabilities, and hence they are named "artificial *intel-*

ligence." This set of permutations follows the basic structure of the postphenomenological formula consisting of three elements connected by dashes, arrows and parentheses. The major change is that the "technology" is replaced by "algorithm" and the "world" by "dataset." This change is meant to adjust the postphenomenological terminology to AI, where the environment is depicted by data, and the technology is reduced to an active algorithm that analyzes and makes decisions. In short, whereas the classical postphenomenological formula refers to I, technology and world, the AI permutations of the formula speak of I, algorithm and dataset.

In this section, I describe four new algorithmic relations each matching one of the classical relations. After presenting a new relation, I provide some examples of the gender biases that were found for this type of relation. This structure allows me to bypass the debate whether the bias is in the algorithm or in the dataset (see (Wellner & Rothman, 2019)).

3.1 The Algorithmic Bodily Relations

The first type, corresponding to the classical "embodiment relation," is termed here "the algorithmic bodily relations." It deals with a kind of algorithms that measure the human body and attempt to construct a digital representation of it made of data. In both cases of the classical and the AI formula, the technology "extends" the human body: in embodiment relations what is extended is the biological body scheme (i.e. when riding bicycles) or the senses (i.e. when wearing eyeglasses); in algorithmic bodily relations, the extension is performed on the digital representation of the human body. For instance, a face recognition algorithm extends the human face by adding to it a name or another identification. These relations require a user (or better – her face), an algorithm that analyzes the image taken by a camera, and a dataset on which the algorithm was trained (in the classical formula the camera and the algorithm are jointly referred as "technology," whereas here our focus is on the algorithm part, i.e. that which makes the decision). Like humans that recognize a face by matching between a face and a name, the algorithm recognizes a face by matching the image with a record in the database which may consist of an identification number and/or a name. The formula can be:

(I ← algorithm) → dataset

This permutation formula depicts the ways in which such algorithms direct people, as in the case of entering a country or a facility. When the algorithm recognizes a face, a gate or a door opens, but if it fails to recognize, the entrance is blocked, and the "I" has to seek the help of a human operator, usually requiring some queueing. The "I" is thereby subordinated to the algorithm.

Note that many of these algorithms are the product of machine learning, so that the criteria how to identify a face is not programmed by people, but rather is the result of a machinic process with no human intervention or control. There is an effort to make this process transparent, so that the users can understand why the algorithm reached a certain decision (see (Wellner & Rothman, 2019)).

The dominant position of the algorithm is represented by the reversal of the left arrow that now points *to* the human, rather than *from* the human as in the classical embodiment relations. The second arrow that points to the dataset exposes that the results of the algorithmic analysis are fed back into the dataset in order to improve the next rounds. Another example is algorithms that analyze our typing speed. These algorithms attempt to conclude the user′s mood in order to offer some content to reinforce or mitigate that mood. Again, the results are stored back in the database for future improvements of the process.

Embodiment relations as well as the new algorithmic bodily relations model the ways in which a technology relates to the human body and how the human body adjusts to the technology. The body is also a major topic in gender studies, especially by referring to the body as one of the main sites of gender bias and discrimination. How should AI algorithms refer to the gendered body? What do they do in practice? Obviously, the female body is different from the male body, and AI algorithms need to be sensitive to such differences (cf. (Michelfelder, Wellner, & Wiltse, 2017)). Most of the algorithms, however, are designed under a neutrality assumption that prevents them from acknowledging the difference and from operating differently.

Back to the face recognition example mentioned above, feminist philosophy of technology would highlight the very low success rates in recognizing female faces compared to male faces.[4] This kind of analysis would also show that the recognition algorithms can identify very well "white faces," but much less successfully recognize the faces of people of other colors. The combination of the two biases leads to higher error rates in recognizing darker-skinned females -

[4] https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software (accessed 28 December 2019).

34.7%, compared to 0.8% in the case of lighter-skinned male (Buolamwini & Gebru, 2018). This is not a theoretical problem. In practice, when the entrance to a country is run by face recognition algorithms, those who need to stand in line for a human agent to let them in are prone to be women, darker-skinned persons, and the largest group is likely to be darker-skinned females, although most of them do not pose any security risk. The lighter-skinned males would probably enter with no standing in line and no questioning as per their motivation to come.

3.2 Maximum Opacity

The second type, corresponding to "hermeneutic relations," represents the situations in which algorithms actively shape our worldview. Just like hermeneutic relations where the technology and the world are taken as one unit, here the algorithm and the dataset are operating together. The permutation of the formula is:

I 🡐 (algorithm – dataset)

These relations model the algorithmic translation as described in the introduction above. In these examples, the world is conceived as ruled by men and for them. Female linguistic forms are classified as mistakes, as demonstrated by the WhatsApp auto-correct example. Likewise, search results are produced and presented in a certain order that requires careful reading and hermeneutics. For example, *Wired Magazine* recently reported how Facebook′s algorithm interpret a query such as "photos of my male friends" as a typo for "female friends," and showed pictures of women in bathing suits.[5] Moreover, running the query "photos of my male friends" did not bring pictures of male friends, but rather male dogs and a few male-themed cartoons. A Facebook′s spokesperson defined this as a bug to be fixed, i.e. a problem in the algorithm. *Wired Magazine* did not blame the algorithm but rather the dataset, and indirectly – the users who type sexist search queries.

It is difficult to examine the algorithm itself for two reasons. First, the algorithm is considered a trade secret of the developing companies. We can guess that they take into account our searching, browsing and reading histories, as well as the IP address from which we enter the Internet, the device that we use for

[5] https://www.wired.com/story/facebook-female-friends-photo-search-bug/ (accessed 1 December 2019).

access (e.g. the model of the cellphone or the laptop), and the preferences of our Facebook friends. Second, the process of machine learning automatically produces a set of software instructions that even the developers do not always understand how a certain decision was calculated. For our inability to know why an algorithm reached a certain conclusion, I term this type of relations "maximum opacity." The algorithms and the dataset form one entity that remains obscure.

And because of this inability to separate algorithm from dataset in the process of machine learning, the dataset is frequently criticized for gender bias. Take for instance job advertisement. These systems have been accused for presenting to women and other minorities relatively low paid jobs, compared to young men, mostly of lighter skin. Is it the blame of the algorithm that is biased? Anja Lambrecht and Catherine Tucker (2018) found that women are excluded from high-paid science and technology related job advertisements because the price of targeting them is higher than that of men. Women are "potentially more valuable targets for advertisers" because they "largely control household purchases" and hence are "are more likely than men to purchase" (p. 4).

Lambrecht and Tucker′s hypothesis was questioned by a group of researchers who interrogated the ad delivery mechanisms of Facebook. Ali et al. (2019) created eleven job ads with different texts and images and defined the same target audience for all the ads. For each job, they created five variations with diverse pictures of potential employees differing on gender and race (total four) and one neutral with no human being presented. They found that all their five ads for positions in the lumber industry were presented to over 90% men and to over 70% white users. By contrast, their five ads for janitors were presented to over 65% women and over 75% black users. Contra Lambrecht and Tucker, Ali et al contend that "the skew in delivery cannot merely be explained by possibly different levels of competition from other advertisers for white and black users or for male and female users" (p. 21). In other words, the bias is more likely to be located in the algorithm.

When considering the relations from the user′s perspective, as the postphenomenological formula originally dictates, it makes no difference whether the bias is in the algorithm or in the dataset, since both are in parentheses and regarded as a joint entity. From a legal and business perspective, in the two cases of the photo search and the ad placement, Facebook is responsible for both the algorithm and the dataset. The relations are opaque because of Facebook′s policy against transparency of its algorithms and datasets. From a technical perspective, the process of machine learning dictates tight links between the algorithm and the dataset, leading the way to regard them as one entity. What is important from the postphenomenological relations is the arrow that now flows from this joint entity to the experiencing "I".

3.3 "Her"

The third type represents the algorithmic quasi-other that interacts with the user. In its extreme form, the algorithm is expected to produce a dialogue that resembles a human-to-human interaction, as evidenced in the development of robots and chatbots. These algorithms are intended to directly interact with the user, and hence the formula is:

$$I \leftarrow \text{algorithm} \; (- \text{dataset})$$

The relation is termed "Her" after the Spike Jonze movie from 2013 where the hero, played by Joaquin Phoenix, falls in love in the fully personalized "operating system" of his cellphone (today this function would be termed chatbot).

The gender biases in this category can be very visible, as in the case of the voice assistants that are named as female – Siri, Cortana, etc., and their voices[6] are programmed to sound like females. From the first "hello," before the actual interaction starts, the setting is clear, and the user expects an interaction with an obedient female.

Another dialogue-like interaction is provided by autocomplete and other text suggestion features. Take for example Gmail′s Smart Compose algorithm that automatically proposes an answer to an email. Users discovered that when they typed "I am meeting an investor next week," Smart Compose suggested as a possible follow-up question: "Do you want to meet *him*?" and did not offer "her" as an option.[7] The company explained that most investors are male, so statistically the proposed answer may fit most cases. The problem is that these emails are fed back into the system and reinforce this answer, even if the number of female investors rises. The way the company handled this bias is by "muting" the Smart Compose feature in gender-related references, so that when a user

[6] Note that the voice is part of the identity of the chatbot, and at the same time functions as its embodiment in the real world. The focus of this section, however, is on the alterity aspects.

[7] https://www.reuters.com/article/us-alphabet-google-ai-gender/fearful-of-bias-google-blocks-gender-based-pronouns-from-new-ai-tool-idUSKCN1NW0EF (accessed 28 December 2019).

types in a sentence regarding a meeting next week, the algorithm will not make any suggestion. Gmail's solution does not deal directly with the way the algorithm decides, but rather tweaks the output side. One may wonder why the Smart Compose algorithm itself was not modified, or why it does not present several options as part of the dialog with the user.

3.4 Background Collection

Lastly, the algorithmic relations that correspond to the classical background relations are termed "background collection." Like background relations where the "world" gains our attention, here what is important is the dataset. The focus is on data, which can be browsing history or x-ray collections and the cancer diagnoses made based on them. The formula here can be:

(I ←) algorithm → dataset

The user is the least important element in the relation and hence s/he is put in the parentheses. In the classical formula it was the technology that was put in the parentheses...

The gender bias example here is Amazon′s algorithm to automate the reading and filtering of job candidates′ CVs. The company tried to inject AI into its recruiting processes and developed an algorithm to score the CVs. It turned out that the system did not recommend women candidates for software development jobs. The explanation was that the training dataset included a majority of men thereby reflecting the company′s employment history, and so the algorithm concluded that the ideal candidate is a man. Reuters reported that Amazon abandoned the project.[8] The candidates had no idea why they were rejected, and hence the formula puts the "I" inside the parentheses.

Another example in which the "I" withdraws to the background is gig economy platforms such as Uber and Airbnb where the service providers function as background information - the car′s driver, the homeowner. This type of user should not be confused with another "I", the customer, who can be the car′s passenger or the guest at the apartment. The customer is considered of importance. As new economies, these platforms had the potential to avoid the existing gender bias that is reflected in legacy datasets. They could have provided

[8] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (accessed 28 December 2019).

equal opportunities for women as Uber drivers or as AirBnB homeowners. Surprisingly, they also duplicate the salary differences. In these platforms, women receive fewer orders and their income is lower than that of men (see (Fisman & Luca, 2016) and the references there; see also (Barzilay & Ben-David, 2016)). It also happens in less controlling platforms such as eBay (Kricheli-Katz & Regev, 2016), where women sellers received a smaller number of bids and lower final prices than did equally qualified male sellers of the exact same product as detailed in the introduction.

3.5 Summary of the Algorithmic Relations

Gender biases exist in the technological world, before and after the introduction of AI algorithms. But these algorithms lead to new manifestations and forms of discrimination. Table 1 depicts the differences between the two sets of postphenomenological relations:

Classical Relations		Algorithmic Relations	
Embodiment	(I – technology) → world	Algorithmic body	(I ← algorithm) → dataset
Hermeneutic	I → (technology – world)	Maximum opacity	I ← (algorithm – dataset)
Alterity	I → technology (– world)	“Her”	I ← algorithm (– dataset)
Background	I → (technology –) world	Background collection	(I ←) algorithm → dataset

Table 1: Classical vs. Algorithmic Relations

In the new algorithmic relations, the arrow of intentionality is reversed to model that the human user is no longer in control and does not occupy a focal point. In the most extreme case – in background collection relations – the user withdraws to the background. A critical approach would highlight the decline of the individual subject and the rise of global corporates who develop the algorithms and control the datasets. These companies put their profits before anything else, including the needs and wishes of their users.

Pointing to the corporates provides only a partial explanation. A full explanation should also tackle the role of technologies that autonomously learn (i.e.

machine learning) and their opacity. Thus we shift from the general "technological intentionality" (Verbeek, 2008), which can be found in many technological domains, to a new form of intentionality which is termed "programmed intentionality" (Rowley, 2015), "intentionality of the other than human" (de la Bellacasa, 2017), "thing intentionality" (Wakkary, Hauser, & Oogjes, 2020), "relegation" (Wellner, 2020), or "multi-intentionality" and "reverse Intentionality" (Wiltse, 2020). The goal of this article is to explore the effects of this type of intentionality from a gender perspective.

The question of intentionality is frequently translated (and even reduced to) ethicality (de la Bellacasa, 2017, p. 122), and become even more complex in a non-anthropocentric position in which technologies possess an ever-growing technological intentionality. What are the moral implications of such an intentionality? In the context of gender bias, we should ask how to refer to such technologies that discriminate their female users? Ethic of technology should provide us with some guidelines, and the next section reviews some analyses in this emerging field.

4. The Ethics of Gender-biased AI Algorithms

Postphenomenology is based on the supposition that humans and technologies are co-shaped, that is—humans shape technologies and at the same time those technologies shape humans by opening for them new horizons. The new horizons form an environment that enables the development of even newer technologies and so forth in an endless loop. Within the co-shaping paradigm, ethics plays an important role, as it points to the positive directions for present and future developments. However, ethical considerations are missing from the postphenomenological formula. Ethics functions at best as a hidden layer.

The ethical questions that interest me are located in the cross-section between gender and AI: Can women and other minorities conduct co-shaping processes with algorithms that do not respect their identity? How should they respond to technologies that put them in an inferior position? These issues become more pressing in a world where one cannot choose whether to be subjected to these algorithms. Unlike the classical ethical decision of buying a certain brand of smartphone or the ethical selection of food in a supermarket (Puech, 2016, p. 104), AI algorithms operate on us whether we like it or not: Our CVs are read by algorithms, and our consent is not required. Ads and friends′ posts are presented to us (or not) with no need for prior involvement on our part. Moreover, the choices we have are very limited: either to opt out and remain unemployed; or look for the few employers who sort potential employees with human agents according to well defined ethical standards.

In this section I assess the applicability of three ethical analyses in philosophy of technology to the problematics presented in the previous section: the first is known as distributive justice and is analyzed by Sven Ove Hansson; the second is ethics of care and its adjustments to philosophy of technology as offered by Michel Puech; and the third is the technological intentionality and its ethical implications as developed by Peter-Paul Verbeek.

4.1Distributive Justice

Distributive justice analysis departs from the understanding that discrimination is morally wrong. This is a basic ethical rule and it is considered an "uncontroversial norm" (Hansson, 2017A, p. 12). In philosophy of technology, Hansson (2017B) identifies two basic categories of distributive justice. One relates to the distribution of technologies in society, i.e. who has access to safe drinking water, or who can financially own a smartphone. In the domain of AI, this category does not pose any difficulty because the distribution is regarded as equal since everyone has access to these technologies (at least in the West).

The other category of distributive justice seems to be more problematic in the context of AI as described in this article: this category refers to technology as the cause of social injustice. Here Hansson refers to technologies that create "permanent advantages for a privileged minority" (p. 53) or "permanent disadvantages for a underprivileged groups" (p. 54). This category may fit some of the job discrimination cases presented in the previous section, especially to those that reveal the limited access to high-paid job advertisements. It may also fit the low success rates of face recognition algorithms when applied in airports on women, minorities and their combinations. In the context of AI, the word "permanent" in Hansson′s definition might be limiting with no cause, since some of the above-mentioned algorithms create advantages and disadvantages that are relevant to a certain period of time, as in the case of a credit line. It is not a permanent construct like medicines and cognitive enhancements that Hansson mentions. And yet, those algorithms discriminate in an unethical way.

Hansson proceeds to detail two additional sub-categories: technologies that promote prejudice and "technological change with an unfair distribution of transition costs" (p. 55). The former includes technologies and technics that reinforce racial prejudices (such as whitening face cream) or reinforce oppressive conceptions of the female beauty (such as breast implants). The latter refers to situations in which the mere transition from one technology to another incur unbearable costs on a certain segment. The introduction of robots, for example, can be beneficial to society as a whole, but in the short term, the workers of an assembly line who lose their jobs are the ones to "pay" the transition costs. So far AI algorithms did not exhibit gender biases in these sub-categories.

All in all, distributive justice categories may help us detect and classify gender bias in AI algorithms, but this kind of analysis will hardly show us the way to avoid the bias. The common solution would require the removal of the discriminating technology from our social lives. But the immersive-ness of AI algorithms urges us to seek new solutions, new ways of operation, and a new logic that will not be prejudiced towards women and other minorities, in order to deliver the promises of neutrality and impartiality of AI technologies.

4.2 Ethics of Care

Ethics of care may give us such operational directions. Puech broadly defines care as follows: "Caring does not only convey the hermeneutic . . . intentionality of ′giving sense to,′ ′acknowledging′ something as a distinct entity. Caring as existential openness to things and to the world is an active pre-occupation and consists in actions, including decision and volition" (Puech, 2016, p. 95). An ethics of care covers a broad spectrum ranging from action all the way to decision making. Puech stresses that "care is more fundamental than justice" (p. 96).

Therefore, the ethical aspects of AI technologies can be translated into the question: Can algorithms care? If care is based on the assumption that empathy is "the key emotion in ethics" (p. 96), then we need to develop algorithms that are not only "intelligent" but also emotional and especially empathic. Some work has been done in this direction, notably by Rosalind Picard (1997) who developed the framework of affective computing.

But Puech remains on the "subjective" side and develops an ethics of care in the direction of self-care of the user. For him "the concept of care [is] founded not on a valuation of the object but on the constitutive experience of the subject" (Puech, pp. 98-99). This direction might lead to a clear distinction between subjects and objects, although such a distinction is becoming more and more complicated to sustain in the presence of AI technologies that make decisions previously made by humans only.

A possible solution may promote an "export" of some ethical recommendations from the "subjective" arena to the "objective" realm of technologies. Clues to such a solution can be traced in Puech′s description of care which resembles the mechanisms of machine learning: "Care is . . . disentangled from controversial issues with normative descriptions of the world: care cannot depend on previously theoretical background consisting in a moral picture of the world" (p. 99). Thus, machine learning and ethics of care are similar in not being based on a predefined set of rules but rather as being self-adaptive to real world situations.

In addition to an "ethical" dataset, the algorithms themselves should be updated. They should be ethically "educated", where education means to go through a certain set of processes in which the algorithm "learns" how to differentiate between good and bad. This is the meaning of the shift from machine *learning* to machine *education* (Wellner & Rothman, 2019),[9] making education the key also in the realm of AI.

4.3 The Morality of Objects

My third ethical investigation is based on the third chapter of Verbeek′s book *Moralizing Technology*, titled "Do artifacts have morality?". This provocative question echoes Langdon Winner′s widely cited article "Do artifacts have politics" (1986). For both questions, the initial answer is negative, regarding tools and machines as mute objects and hence as something that cannot have politics or morality. This is the classical approach to ethics. Both authors, however, offer another answer, that artifacts do have politics and morality.

It seems that their approach can fit machine learning algorithms that express strong technological intentionality and lower levels of human control. Put differently, with AI, humans′ intentions are distorted and their ability to control the consequences might be limited. The humans can be the users, the service providers (i.e. Uber drivers or Airbnb's hosts) or even the developers, in all cases their control of the algorithm is limited. The line of thinking that artifacts do not

[9] A similar solution has been offered by de la Bellacasa (2011) who recommends a caring strategy intended to ″hold together the thing″, so that various stakeholders become united with a care, or a matter of concern.

possess intentions and hence are not morally responsible is losing some of its strength in the face of AI algorithms that self-develop their own reasoning. Verbeek′s explanation fits this situation very well: “The fact that we cannot call technologies to account for the answers they help us to give does not alter the fact that they do play an actively moral role” (p. 42). Due to its extreme distributedness and limited “traceability” (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016, pp. 5, 12-13), AI′s technological intentionality does not necessarily depend on the wishes, desires and beliefs of a specific developer. Their unintended-ness and unpredictability is inherent. For Verbeek, however, this is a general attribute of technologies and not a unique feature of AI.[10] This attribute makes technologies different from physical phenomena, for example, and more resembling humans. In the case of AI algorithms, this is obvious and indisputable.

What does it mean that AI algorithms are considered moral agents? Take for example translation algorithms. As we have seen, they do not provide an “objective” translation from one language to another but also a worldview. Unfortunately, this worldview consists of regarding the world as populated by men, and of pushing women to stereotypical occupations, at best. Thus, AI algorithms pose ethical questions that are not necessarily “action-oriented” but rather shaping the answer to the question “how to live” (Verbeek, 2011, p. 53). The fact that they impose certain decisions on their users (a lower credit line, less attractive jobs, masculine dominancy in spoken language) has led me to reverse the arrow of intentionality so that it points *to* the experiencing “I.” Consequently, AI algorithms serve as “moral agents ′in themselves′, capable of moral action” (p. 52). Their action of construction of a male-centric worldview should be treated with ethical tools as proposed by Verbeek.

[10] For some scholars, AI technologies belong to the sub-category of ″moral agents *in themselves*″ (Verbeek, p. 47), and in it are a *sui generis* because they ″make their own decisions″ (p. 50). Verbeek reviews the criteria offered by Floridi and Sanders for moral agents that require the technology to exhibit interactivity, autonomy, and adaptability in order to become morally accountable. It means that they make their own decisions. Verbeek rejects these conditions for being too limiting, so that many technologies are left out, that is – not considered moral agents. Verbeek attempts to include as many technologies as possible in the category of moral agents.

5. Summary and Conclusion

In this article I examined the gender bias of AI technologies with the tools developed in postphenomenology. In the first section, an overview of the four classical relations served as a framework to map various occurrences of gender bias in technologies in general. The second section was devoted to AI technologies and the gender biases exhibited in this domain. These technologies require an update to the postphenomenological relations, in which the arrow of intentionality is reversed and points to the "I." The reversed arrow models the enhanced technological intentionality of AI systems and how they direct their users.

Reversing the arrow of intentionality can reflect various explanations of the gender bias in AI. One explanation is used by developers who argue that they did not mean to create discriminating platforms. They usually point to the fact that algorithms learn from datasets, and since the datasets reflect the "world," which is gender biased, the algorithms end up duplicating the world′s logic, biases included (e.g. (Caliskan, Bryson, & Arvind, 2017)). Another explanation is that the algorithms are "opaque" so that users and developers cannot know why the algorithm made a decision (cf. (Wellner & Rothman, 2019) and the references there). All these arguments mean that the gender bias is regarded as an unintended consequence, beyond the control of the developers. The technological intentionality argument developed here allows some "freedom" to the algorithms, but at the same time requires they hold some responsibility.

In the last section I examined the applicability of three ethical tools to handle the moral responsibility of AI algorithms. The first tool was distributive justice that serves to identify what is wrong with gender bias. The second tool was ethics of care. This subsection attempted to draw initial guidelines for a theory of an algorithmic care as a direction for technological development that may prevent such a bias. This direction can be termed as a move from "machine learning" toward "machine education" according to which the algorithms should gain an understanding of good and bad (see (Wellner & Rothman, 2019)). In the last step I investigated how to regard AI technologies as moral agents, and what are the ethical implications of intense technological intentionality.

The fundamental question is whether AI systems simply reflect us as a society, as Caliskan et al conclude, or is there anything we can do to make them more beneficial to society. It is clear, however, that these technologies pose new challenges. I would like to regard them as urging us to find new ways to combat the biases - i.e., as an opportunity rather than just a threat.

REFERENCES

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proceedings of the ACM on Human-Computer Interaction, 3*, 199.

Barzilay, A., & Ben-David, A. (2016). Platform inequality: Gender in the gig-economy. *Seton Hall L. Rev., 47*, 393-431.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81*, 1-15.

Caliskan, A., Bryson, J. J., & Arvind, N. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183-186.

de La Bellacasa, M. P. (2011). Matters of care in technoscience: Assembling neglected things. *Social studies of science, 41*(1), 85-106.

de la Bellacasa, M. P. (2017). *Matters of Care: Speculative Ethics in more than Human Worlds.* University of Minnesota Press.

Fisman, R., & Luca, M. (2016). Fixing Discrimination in Online Marketplaces. *Harvard Business Review*, 88-95.

Hansson, S. O. (2017A). Theories and Methods for the Ethics of Technology. In S. O. Hansson, *The Ethics of Technology: Methods and Approaches* (pp. 1-14). London and New York: Rowman and Littlefield International.

Hansson, S. O. (2017B). Technology and Distributive Justice. In S. O. Hansson, *The Ethics of technology: Methods and Approaches* (pp. 51-66). London and New York: Rowman and Littlefield International.

Ihde, D. (1990). *Technology and the Lifeworld: from Garden to Earth.* Bloomington and Indianapolis: Indiana University Press.

Ihde, D. (2009). Postphenomenology and Technoscience: The Peking University Lectures. New York: State University of New York Press.

Langsdorf, L. (2020). Relational Ethics: The Primacy of Experience. In *Reimagining Philosophy and Technology, Reinventing Ihde* (pp. 123-140). Cham: Springer.

Kricheli-Katz, T., & Regev, T. (2016). How many cents on the dollar? Women and men in product markets. *Science advances, 2*(2), e1500599.

Lambrecht, A., & Tucker, C. E. (2018). Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *SSRN.*

Liberati, N. (2016). Augmented reality and ubiquitous computing: the hidden potentialities of augmented reality. *AI & society, 31*(1), 17-28.

Michelfelder, D. P., Wellner, G., & Wiltse, H. (2017). Designing differently: toward a methodology for an ethics of feminist technology design. In S. O. Hansson, *The Ethics of Technology: Methods and Approaches* (pp. 193-218). London and New York: Rowman and Littlefield.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 1-21.

O'Neill, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Nueva York, NY: Crown Publishing Group.

Picard, R. W. (1997). *Affective computing*. Cambridge, MA: The MIT Press.

Puech, M. (2016). *The Ethics of Ordinary Technology.* New York and London: Routledge.

Rosenberger, R. (2017). Notes on a Nonfoundational Phenomenology of Technology. *Foundations of Science , 22*(3), 471–94. doi:https://doi.org/10.1007/s10699-015-9480-5

Rowley, M.-L. (2015). Toying with Intention: Embodiment, Empathy and Programmed Intentionality in New Media. In E. Bouet, *The (Un)Certain Future of Empathy in Posthumanism, Cyberculture and Science Fiction* (pp. 1-16). Brill.

Schwartz Cowan, R. (1976). The" Industrial Revolution" in the Home: Household Technology and Social Change in the 20th Century. *Technology and Culture, 17*(1), 1-23.

Verbeek, P.-P. (2005). *What Things Do: Philosophical Reflections on Technology, Agency and Design.* University Park, PA: The Pennsylvania State University Press.

Verbeek, P.-P. (2008). Cyborg intentionality: Rethinking the phenomenology of human–technology relations. *Phenomenology and Cognitive Science, 7*, 387–395.

Verbeek, P.-P. (2008B). Morality in design: Design ethics and the morality of technological artifacts. In *Philosophy and Design* (pp. 91-103). Dordrecht: Springer.

Verbeek, P.-P. (2011). Moralizing Technology: Understanding and Designing the Morality of Things. Chicago: The University of Chicago Press.

Wakkary, R., Hauser, S., & Oogjes, D. (2020). The Disappearing Acts of the Morse Things: A Design Inquiry into the Withdrawal of Things. In H. Wiltse, *Relating to Things: Design, Technology and the Artificial* (pp. 215-238). London: Bloomsbury.

Wellner, G. (2017). I-Media-World: The algorithmic shift from hermeneutic relations to writing relations. In Y. Van den Eede, S. Irwin, & G. Wellner, *Postphenomenology and*

Media: Essays on Human–Media– World Relations (pp. 207-228). Lanham: Lexington Books.

Wellner, G. (2018). From Cellphones to Machine Learning. A Shift in the Role of the User in Algorithmic Writing. In A. Romele, & E. Terrone, *Towards a Philosophy of Digital Media* (pp. 205-24). Cham: Palgrave MacMillan.

Wellner, G. (2020). Postphenomenology of Augmented Reality. In H. Wiltse, *Relating to Things: Design, Technology and the Artificial* (pp. 173-187). London: Blommsbury.

Wellner, G., & Rothman, T. (2019). Feminist AI: Can We Expect Our AI Systems to Become Feminist? *Philosophy & Technology*, 1-15.

Wiltse, H. (2014). Unpacking Digital Material Mediation. *Techné: Research in Philosophy and Technology, 18*(3), 154-182.

Wiltse, H. (2020). Revealing Relations of Fluid Assemblages. In H. Wiltse, *Relating to Things: Design, Technology and the Artificial* (pp. 239-254). London: Bloomsbury.

Winner, L. (1986). Do Artifacts have politics. In L. Winner, *The whale and the reactor: a search for limits in an age of high technology* (pp. 19-39). Chicago: University of Chicago Press.

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist–it's time to make it fair. *Nature, 559*, 324-326.