Gender and Sex in the Computer Graphics research literature

Ana Dodik*
Meta Platforms

Silvia Sellán*
University of Toronto

Theodore Kim
Yale University

Amanda Phillips
Georgetown University

ABSTRACT

We survey the treatment of sex and gender in the Computer Graphics research literature and its scientific and real-world consequences. We conclude current trends on the use of gender in our research community are scientifically incorrect and constitute a form of algorithmic bias with harmful effects. We propose ways for correcting these trends and pose novel research questions.

1 INTRODUCTION

References to sex and gender can be found all throughout the Computer Graphics research literature: a dataset is said to contain images of men and women, user study participants are reported to have a certain male/female ratio, a body modeling algorithm trains two different gendered models, a voice modification method is said to work on male and female voices, etc.

The scientific consensus around the concepts of sex and gender has greatly evolved in the past decades (see, e.g., [Nature Editorial Board 2018]). As surveyed by Fausto-Sterling [2012], *sex* is not one but a combination of many biological classifications (*chromosomal sex, hormonal sex, reproductive sex, ...*) which cannot be unambiguously assigned in a binary way to as much as one in 50 people [Blackless et al. 2000]. *Gender*, on the other hand, is used to refer to an individual's self-identity [Money and Ehrhardt 1972], their performance of certain acts [Butler 2003] or arbitrary social organizational structures that segregate people in different public bathrooms and even decide who can access education or participate in public life [Lorber 1994]. By all these contemporary definitions, gender is non-binary, fluid and culturally-specific. Furthermore, assuming outdated binary definitions of sex and gender is not just scientifically incorrect, but can also be shown to be harmful to those who conform the least to this artificial binary [UNHCHR 2015].

Despite this, we observe that the treatment of sex and gender in Computer Graphics research still answers to a traditional binary understanding of it that excludes intersex and many transgender and gender non-conforming people. We argue that our community's current use of gender is imprecise, contradictory and detrimental to our scientific integrity. We examine the harmful real-world consequences of the algorithmic bias introduced by our modeling choices with respect to gender on how gender non-conforming people interact with our technology in their daily lives. We advocate for reexamining our treatment of gender and show that this will not only correct worrying trends in our community, but also open the door to whole new avenues of research.

2 SURVEY

Inspired by the work of Keyes [2018], we conducted a survey of all technical papers presented at SIGGRAPH North America and SIGGRAPH Asia since 2015 (see supplemental material). We observed references to gender routinely throughout, varying in nature from demographic information reported about user study participants to gender-specific algorithms. Whenever gender is used explicitly as a variable, it is always as a binary one. Despite its prominence, gender is never given a precise definition in all the reviewed Computer Graphics literature, and appears to be used implicitly as a proxy for anything from body proportions to facial expression to voice inflection in speech.

An analysis of the above reveals worrying trends about the current use of gender as a variable in Computer Graphics, both scientifically and ethically. As we mention examples of works that perpetuate these trends, we stress that we do not associate any malicious intent to any. Rather, we wish to show how seemingly neutral, well-established practices in our community can lead to us unwittingly perpetuating forms of algorithmic bias.

3 ALGORITHMIC FAIRNESS ANALYSIS

Our literature survey demonstrates that the current use of gender in the computer graphics literature is at best ill-defined, and at worst incorrect. In this Section, we demonstrate through various examples why such an approach causes a number of *technical* issues within the surveyed works, and is, as such, at odds with producing precise and high-quality reproducible research.

In our discussion, we apply a framework which disambiguates and categorizes different types of bias according to the stages of a system's lifecycle [Suresh and Guttag 2021]. Note that while the original paper focuses on machine learning, we find it to be equally applicable to general problems in computer graphics. We give concrete examples of how *all* types of bias occur throughout the surveyed work.

Silvia: to-do

Historical bias. Historical bias occurs when data encodes existing prejudice. For example, a *gender classifier* trained on data collected in a society where social norms dictate gender expression might learn that "wearing a dress" means woman, and "short hair" means man, despite the data being *abundant* and *perfectly sampled*.

Representation bias. This type of bias occurs when a part of a population is poorly represented by a dataset. This can happen due to a multitude of reasons:

- *Sample selection bias* occurs when the sampling procedure is biased in such a way to not include non-binary people. Given the scale of the works reviewed and the reported statistics about the prevalence of such individuals, the works we reviewed should include non-binary individuals; however, we could not identify a single paper that explicitly mentioned them in their motion capture actors, datasets or user study participants. A possible explanation is that the sampling procedure was accidentally designed in such a way to to decrease the likelihood of capturing gender-diverse individuals, or it might also occur due to measurement bias.

*Joint First Authors

- A dataset can lead to algorithmic unfairness if it is uniformly sampled, but *contains under-represented groups*. A uniformly sampled dataset with 1000 people is expected include between 1 and 20 intersex and gender non-conforming individuals. This means that the algorithms trained on this dataset are more likely to produce worse results on these individuals than in the general population. In our survey, we did not identify a single paper that acknowledged the existence of intersex and gender non-conforming people; consequently, we did not identify a single paper that explicitly attempted to correct this type of representation bias.

Measurement bias. Measurement, or reporting bias is typically introduced during the task definition through the selection and measurement of features and target variables. During our literature review, we identified the presence of the following issues:

- The frequent usage of *proxies*, both as features and as target variables. For example, we observed works that use *sex* or *gender* to mean *comonly co-occuring bodily characteristics* or *a set of voice attributes*. The use of *sex* and *gender* as proxy variables is not only harmful, but often also imprecise and inaccurate from a technical stand-point. It is quite possible that the authors would be better served using other less abstract features (e.g. hair length, or voice pitch). The imprecision can compound further when a single work combines several understandings of sex and gender; for example, we observed virtual try-on algorithms use gender both as a proxy for body parameters *and* cultural choices in attire, and conversational agents use it to denote both the pitch of the agent's voice as well as cultural traits in their verbal and non-verbal communication. The usage of these variables can further lead to *omitted variable bias*, which we discuss later.
- An *inaccurate* method of measurement. For example, all of the papers we reviewed treat gender as a *discrete binary variable*. Therefore, even if data was collected from gender-diverse participants, the method of measurement would not be able to capture that.
- An *incorrect* method of measurement. We identified works that both propose and use gender labeling of image, voice and body geometry data, in many cases by automated systems or manual third parties as opposed to participant self-identification. Similarly, we observed works that report many user study participants' gender as "unknown", which may mean it is the researchers themselves who are attempting to assume their participant's gender without asking them to self-identify it. Algorithmic bias occurs in this situation, even if special care is taken to collect data of gender non-conforming individuals.

.

Aggregation bias. **Silvia: to-do Ana: I need help on this one, since I cannot think of any examples. From the survey paper: "Aggregation bias (or ecological fallacy) arises when false conclusions are drawn about individuals from observing the entire population. An example of this type of bias can be seen in clinical aid tools. Consider diabetes patients who have apparent morbidity differences across ethnicities and genders. Specifically, HbA1c levels, that are widely used to diagnose and monitor diabetes, differ in complex ways across genders and ethnicities. Therefore, a model that ignores individual differences will likely not be well-suited for all ethnic and gender groups in the population. This is true even when they are represented equally in the training data. Any general assumptions about subgroups within the population can result in aggregation bias."**

Learning bias. **Silvia: to-do**Algorithm designers will often unknowingly introduce bias into their methods. For example, using a regularizer that makes sure to regress to an "average" body is likely to make the algorithm worse for people far away from the average. Alternatively, optimizing for a model's accuracy might lead to the statistical pairity decreasing accross different groups.

Evaluation bias. **Silvia: to-do**Evaluation bias encompases all bias that is introduced during the evaluation of an algorithm. If the computer graphics community settles on benchmarks with biased data or metrics, it creates a knock-on effect where the development and deployment of models that performed well on those benchmarks is further encouraged. This is one of the reasons it is particularly dangerous to "leave the fairness questions" for future work.

Deployment bias. **Silvia: to-do** Deployment bias refers to the harm that is introduced as a consequence of the model being published or deployed in the real world:

- Assuming an implicit definition of (e.g. binary) gender might incentivize future researchers to conform to that definition. This is additionally problematic when researchers from cultures with a different understanding of gender need to adjust to foreign cultural norms. (**Ana: Amanda, I might need some help with this one.**)
- Deploying an algorithm can lead to *feedback loops*. For example, if gender-diverse people have poor experiences with clothing size recommender systems, they are less likely to use them. As a consequence, the data about the performance of such a system will be skewed to include fewer gender-diverse people. Using such data to further optimize the systems can lead to compounding effects.
- A deployed system can cause harm by nudging the users to artificially change their behavior. For example, a trans person might feel the need to change the pitch of their voice in order to not get misgendered by an algorithm.
- Deploying a body model which has "male", "female", and "gender neutral" variants might lead to unintentional harm, if, for example, the "gender neutral" model is used only that if the algorithm does not have enough confidence the user is either "male" or "female". This can lead to gender trans or non-confirming individuals being actiely told that they are not recognized as the gender they identify with.

Omitted variable bias. Omitted vairable bias occurs when the success of using a certain feature is overemphasized because it correlates with another important feature that has been omitted from the modeling stage. For example, "gender" is likely not as discriminative as a variable when the result is also conditioned on

"hair length", or "hip width". Alternatively, a voice modification algorithm may attribute its performance to gender, rather than a combination of pitch and certain socially-acquired speech inflections. **Ana: cite paper that introduces this concept because it's not [Suresh and Guttag 2021].**

Ana: I don't know where to fit this: a proposed new conversational agent might learn a correlation between a person's visual appearance and certain traits they exhibit in non-verbal communication. Suggestions?

In general, we have found the gender-related language to be imprecise, which hinders the clarity of the presentation. Oftentimes, we found that *gender* and *sex* are used seemingly interchangebly, and most of the times it is not even clear from context which one the authors intended to use. In fact, we argue that it is impossible to know a-priori if a trained model has picked up on *gender* characteristics, and not on the characteristics of *sex* or *gender expression*, without looking at algorithmic fairness metrics accross different subgroups.

A worrying trend we noticed is that none of the reviewed papers provide an analysis of the algorithmic biases that they potentially introduce. While different real-world constraints might not make it realistic for a research group to successfully mitigate certain sources of bias, the potentially introduced biases should at the very least be acknowledged. For example, none of the reviewed papers included any of the various algorithmic fairness metrics (**Ana: cite metrics**) into their evaluation, nor did they even include a discussion of the potential harm their methods could be causing.

Algorithmic fairness evaluations and discussions are left out of computer graphics papers not because they happen to be difficult or time-consuming, but because they are deemed unnecessary by the people who would at large be either unaffected or positively affected by the introduced biases.

However, we argue that the problems introduced in these methods are not only potentially harmful to under-represented populations, but also often *technically limited ways which are not well researched*. If a method cannot model a class of humans by design, or if a production system fails for a subsection of the population, these are fundamental *technical* limitations. The question is then, why does our community prioritize solving these limitations disproportionatly less then other technical problems?

3.1 Real world harm

Ana: Disambiguate between allocative and representational manifestation of harms, Barocas et al. Ana: The mental model for data collection should not be "we need samples from N people", rather it should be "we need samples from $m \times k \approx n$ people, such that each of the k groups of m people happen to be representative of a certain protected characteristic". Ana: We argue that discussions around algorithmic fairness need to become front-and-center within our own community, instead of being relegated to other venues or "future work". Ana: People, measure your bias! Ana: IT IS NOT NECESSARY TO USE GENDER AS A VARIABLE.

As scientific researchers, we must be aware of the effect that our arbitrary modelling decisions have in the real world as our algorithms are used by governments and private companies.

Since many people's gender experiences fall outside the male/female binary, our research's insistence on it can contribute to frustration (at best) and discrimination (at worst) when they interact with technology. A researcher's seemingly inocuous decision to use different search spaces for fitting male and female body proportions leads to airport body scanners that routinely subject transgender passangers to humilliation (see [Beauchamp 2019]). A modelling choice to conflate body proportions with choices in attire ironically excludes precisely the people with non-normative bodies who are the most in danger in traditional physical changing rooms (see e.g., [Silver 2017]). **Ana: lol please explain this to my tech lead...**

These negative effects are compounded even further as our algorithms are being used to generate synthetic datasets on which to train Machine Learning algorithms outside of our research area. If we do not examine and properly report our algorithm's limitations in representing people outside of the gender binary, these can later be used to train autonomous vehicles to detect pedestrians ([Behzadi 2021]), medical diagnosing tools ([Chen et al. 2021]) and even security threat detection ([Brewer 2020]).

Furthermore, as Computer Graphics researchers, we must consider our role in shaping whose stories get to be told and who gets to seem themselves represented in the entertainment culture. By conflating different attributes under the umbrella of gender, we exclude gender non-conforming individuals from every videogame and movie created using our tool, further invisibilizing already-invisible and marginalized communities.

It bears mentioning that our research community's entrenchment in the traditional gender binary is a rare example of Computer Graphics research lagging behind the needs of our partner industries. *Metahuman*, the latest photorrealistic character modeller by Unreal Engine [2021] has no mention of gender; Google [2020] removed all gender references from its Cloud Vision API; video games as diverse as *Animal Crossing: New Horizons*, *Cyberpunk 2077* and *Forza Horizon 5* completely decouple attributes like hairstyle, body proportions, voice pitch and prononouns from one another. **Ana: I am *very* against including Cyberpunk as a positive example of anything gender-related.**

Finally, the current use of gender in the Computer Graphics literature creates a hostile environment for gender non-conforming members of our research community, which goes against ACM SIGGRAPH's goal to be *a model of inclusion, equity, access and diversity for all*: **Ana: I would end the sentence here. Also, the previous sentence might be good to include in the abstract or as one of the very first sentences.** by seeing colleagues and collaborators consistently exclude us **Ana: us → gender-diverse people** from their own research work, we are (willingly or not) sent the message that we do not belong in this research community, encouraging us to look for jobs elsewhere.

4 WHERE DO WE GO FROM HERE?

We believe the reasons above to be enough to make us reevaluate the role of gender in our community's scientific literature.

For example, the reporting of gender among other demographic information in user study participants and dataset collection subjects answer to a scientifically positive goal (experimental transparency) as well as an ethical one, to safeguard against the "male

default" that plagues science and has plagued it since its infancy. However, we found instances in our survey of participants being reported as of "unknown gender", which may indicate that their gender is being assumed post facto by researchers as opposed to self reported, leading to the potential misidentification and exclusion of gender non-conforming individuals or of those from certain ethnicities (see e.g., [Buolamwini and Gebru 2018; Santamaría and Mihaljević 2018]). Therefore, we would argue it is still advisable to include this kind of data, as long as it is self reported by participants who are given a breadth of gender options not restricted to the traditional binary ones.

On the other hand, the scientific and ethical harm caused by gender-segregated algorithms is likely too significant to offset any possible benefits. At the very least, these choices should be justified and their consequences in terms of excluding gender non-conforming individuals should be examined and clearly stated. Eventually, we hope that our field evolves to address these limitations and move beyond the outdated gender binary. We trust that our fellow researchers share our scientific excitement in this new frame of reference and the potential novel research directions it opens; for example:

- What is a complete parametric model for the human body that is decoupled from gender and accurately represents the diverse bodies of all humans, regardless of whether they conform to traditional gender norms?
- How can our research inform or contrast more modern understandings of gender? Can data-based methods be used to evaluate cultural differences in gender presentation?
- How can we evaluate our algorithms for bias towards the gender binary? What tools are needed to obtain or synthesize data that covers more diverse experiences of gender?

We acknowledge that our proposed break with tradition may bring with it effort and difficult conversations, but these are challenges worth facing in the interest of scientific advancement as well as producing a fairer, more inclusive future.

REFERENCES

Toby Beauchamp. 2019. *Going Stealth: Transgender Politics and U.S. Surveillance Practices.* Duke University Press, Chapel Hill, NC.

Yashar Behzadi. 2021. Synthetic data to play a real role in enabling ADAS and autonomy. *Automotive World* (2021).

Melanie Blackless, Anthony Charuvastra, Amanda Derryck, Anne Fausto-Sterling, Karl Lauzanne, and Ellen Lee. 2000. How sexually dimorphic are we? Review and synthesis. *American Journal of Human Biology: The Official Journal of the Human Biology Association* 12, 2 (2000), 151–166.

Tim Brewer. 2020. DHS Awards $1 Million to Support Machine Learning Development for Airport Security. *Synthetik Applied Technologies Blog* (2020).

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency.*

Judith Butler. 2003. Gender trouble. *Continental feminism reader* (2003), 29–56.

Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. 2021. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering* (2021), 1–5.

Anne Fausto-Sterling. 2012. *Sex/gender: Biology in a social world.* Routledge.

Google. 2020. Ethics in Action: Removing Gender Labels from Cloud's Vision API. https://diversity.google/story/ethics-in-action-removing-gender-labels-from-clouds-vision-api/. Online; accessed 20 January 2022.

Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction* (2018).

Judith Lorber. 1994. *Paradoxes of gender.* Yale University Press.

John Money and Anke A Ehrhardt. 1972. Man and woman, boy and girl: Differentiation and dimorphism of gender identity from conception to maturity. (1972).

A Nature Editorial Board. 2018. US proposal for defining gender has no basis in science. *Nature* 563, 7729 (11 2018), 5.

Lucía Santamaría and Helena Mihaljević. 2018. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science* (2018).

Laura Silver. 2017. Topshop Refused To Let A Trans Person Into An All-Gender Changing Room. *BuzzFeed News* (2017).

Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization* (Oct 2021). https://doi.org/10.1145/3465416.3483305

UNHCHR. 2015. Discrimination and violence against individuals based on their sexual orientation and gender identity. (2015).

Unreal Engine. 2021. Digital Humans | Metahuman Creator. https://www.unrealengine.com/en-US/digital-humans/. Online; accessed 20 January 2022.