Overclocked: AI gender recognition and trans people

Genevieve Clifford
g.clifford.2004599@swansea.ac.uk
Swansea University
Swansea, Wales, UK

ABSTRACT

Automatic gender recognition (AGR) systems have been shown to be trans-exclusionary, and sentiment around them has been judged to be overwhelmingly negative by transgender people. The purpose of this paper is to judge the state of the art, determining if biases perpetuated by authors in the AGR-space still exist within the systems they produce. Considering an idealised "model paper" written to be reflexive, trans-inclusive, and with ethical qualms addressed, it was found through a reflexive content analysis methodology that there has been some shift in trans-inclusivity in the past five years, with 21% of papers being trans-inclusive to some extent (cf. 2% previously), but no technical papers were found taking any reflexive approach towards gender modality, nor discussing ethical issues around AGR systems from a trans perspective.

Author Keywords

automatic gender recognition; transgender; misgendering; design justice; ethics; positionality; reflexivity

CCS Concepts

•**Social and professional topics** → **Gender;** Governmental surveillance; •**Security and privacy** → Social aspects of security and privacy; •**Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence; Machine learning;**

INTRODUCTION

In their 2018 paper 'The Misgendering Machines' [8], Keyes investigates algorithmic bias in automatic gender recognition (AGR) systems. They find through evaluation of papers published on the topic in high H-index papers between 1997 and 2017 ($n = 58$) that authors in the AI-space operationalised (without self-acknowledgement) gender in three different (and incorrect) ways, that: gender is a binary construct (94.8% of papers); gender is immutable (72.4% of papers); and gender is purely physiological (60.3% of papers). They further state that such authorial bias that seeps into technology (that are created by these authors) manifests in perpetuating violence against transgender and non-binary people through enforcing gendered spaces, misgendering, and reinforcing erasure.

In the five years that have elapsed since Keyes published 'The Misgendering Machines', I[1] wish to ask has there been any change in the field? Do authors think reflexively about gender or consider the ways in which their algorithms might affect transgender people?

LITERATURE REVIEW

Trans perceptions of AGR technology

> Sci-Fi Author: In my book I invented the Torment Nexus as a cautionary tale. Tech Company: At long last, we have created the Torment Nexus from the classic sci-fi novel Don't Create The Torment Nexus [1]

The above Tweet by Alex Blechman highlights the position that many transgender people feel when confronted by AGR technology. It is this perception of technologists, engineers, and others implementing software that is harmful in this paradigm including trans people that Hamidi et al. [7] investigate through qualitative methods. They found that trans people they interviewed ($n = 10$) had overwhelmingly negative perceptions of AGR technology, citing fears of misgendering and mischaracterisation by machines, as well as questioning the role such machines have in society, along with the potential to oppress and marginalise.

Such fears are well-founded, Costanza-Chock [2, pp. 1-4] (a nonbinary trans* femme researcher and designer) detail their experience navigating airport security in the United States; an experience of biometric bordering and humiliation at the hands of the state. They describe the TSA's millimetre wave scanning technology (exemplar outputs of which are shown in Figure 1, originally from Costello [3]) and having anatomy that isn't recognised by it; their genitals are anomalous on the 'female' setting, and their breasts are anomalous on the 'male' setting. Both readouts from the system prompt an escalation to a TSA officer through a pat-down; Costanza-Chock concludes *"I can't win"*.

Outside the realm of biometric bordering, AGR technology has been used in social media. Pilipets and Paasonen [10] reflect

Permission to make digital or hard copies of all or part of this work for classroom use is granted as per Swansea University's policies on this. Intellectual property is likewise defined by the university and applies in the same way that it does for all coursework assessments.
CSDM03 '23, Feb 02–May 05, 2023, Swansea, Wales, UK
ACM xxx-x-xxxx-xxxx-x/xx/xx. . . $00.00
DOI: http://dx.doi.org/xx.xxxx/xxxxxxx.xxxxxxx

[1]Throughout this paper, I use first-person active language (as opposed to third-person passive), this is in keeping with the reflexive approach I have taken in this paper — I discuss this more in later sections

on Tumblr's algorithmic approach to 'tackling'[2] adult (often tagged `#nsfw` — not safe for work) content on the website following Apple's decision to remove it from the App Store for this reason. In 2018, Tumblr implemented an algorithmic system to censor NSFW content, both text and images. The latter category yielded the somewhat bleakly corporate-/HR-esque phrase *"female presenting nipples"*; Pilipets and Paasonen write *"Nipples — female, male, trans, and gender-fluid — became central mediators of the porn ban critique, channeling (sic) resistance against the new policy."* [10, p. 1466]. Haimson et al. [6] present the argument that Tumblr's flawed algorithm for determining between what is pornographic and medical/educational[3] means that it *"limits erotic, medical, and/or educational trans knowledge, stripping power from trans online communities"* [6, p. 353].

Positionality in the AGR-space and trans studies

Davies [4] introduces the concept of reflexivity (and thereby positionality) in the context of ethnography in "Reflexive Ethnography", though it is still relevant here; it is the *"ways in which the products of research are affected by the personnel and process of doing research"* [4, p. 4] and is present in all phases of the research process, in both the social and natural sciences. Positionality in this case refers to the 'personnel' in research, and how researchers are situated in relationship to the phenomena or social construct they are researching.

Research centring positionality in AGR systems has precedent; Keyes and Austin embrace this in 'Feeling Fixes' [9] where they perform an audit on the HRT Transgender Dataset, noting many ethical concerns — using an opt-out consent model and using deeply personal transition timeline videos to train systems to detect *"terrorists [that] might undergo hormone replacement therapy to sneak across the US border"* [9, p. 3]. The authors trans positionality is central to their work, given that they could have been *"simultaneously subject and object"*, they embrace the *"mess"* of auditing, incorporating their own experiences of auditing into their audit proper [9, p. 4].

The importance of trans leadership in research is discussed by Rosenberg and Tilley [11] with their stepwise model of trans insider-outsider[4] (IO) inclusion in research projects. In increasing order of inclusion, this model considers projects that have: no trans IO input; consultancy with trans IO; trans IOs as RAs and peers; and trans IO led research. Through Rosenberg's own master's thesis work, she found that being explicit about her transfeminine positionality led to participants having greater levels of trust, knowing their perceptions were less likely to be misconstrued by her.

[2] I include this word in quotes in part to challenge the anti-sex worker narrative imposed by this framing

[3] Some trans Tumblr users will upload images of their bodies after surgical procedures — i.e., after top surgery (breast implant procedures and mastectomies) or perhaps bottom surgery (vaginoplasties and phalloplasties)

[4] The insider-outsider status rejects the universality of the 'trans experience', positing that while trans people share many experiences, the minutiae of life for each individual will be different

STUDY DESIGN

Reflexive content analysis

Content analysis is to be conducted on a number of papers spanning a series of years; these fall into two categories: papers published before 2017 and those published after. During the content analysis, I will also take reflexive notes, which will be used to inform my analysis.

Inclusion criteria for papers published before 2017

1. Paper analysed in "The Misgendering Machines" [8]

Inclusion criteria for papers published after 2017

1. Journal included in "The Misgendering Machines" [8]
2. Paper published after 2018 (assume 1st January)
3. "gender recognition" OR "gender classification" OR "gender clasifier" OR "classify gender" OR "infer gender" OR "determine gender" OR "gender determination" OR "recognize gender" contained in paper full-text
4. Contains empirical research (technical and/or social) about automatic gender recognition using facial recognition technology

On the selection of papers

The two inclusion criteria I have presented are not exactly equal. Keyes *"examined the entire archive of papers from each venue"* [8, p. 5] and manually selected papers that used AGR technology (and categorised if the paper's sole focus was gender recognition or not). I believe this yields a carefully curated list of papers to perform content analysis on, but it comes with one caveat: it is a laborious approach. In testing for selecting search terms, I accidentally neglected to include the quote marks around search terms on IEEE Xplore, this yielded some 400 papers across each of the publications hosted there. With a liberal underestimate of 2 minutes to categorise each paper, it would take > 13 hours of repetitive unbroken labour in total; the total time taken for Keyes' approach would have been much longer. The search terms I have selected are those that I most frequently encountered while reading Keyes' set of papers; this yielded 86 total papers, as seen in Table 1.

I proceed knowing that the total set of papers to analyse are a bifurcated set with slightly different methods of data collection, but operating under the assumption that they are approximately equivalent (they cover the same topics, and those excluded follow Keyes' methodology exactly) and are able to be compared.

As seen further in Table 1, the 'top 7' journals analysed by Keyes remain, although their comparative rank has changed and in general, their H-indices have increased[5]. Publications from the 'Journal of the Optical Society of America A' are excluded as I lack institutional access.

Quantitative scoring

In proposing a quantitative score for each paper, I am in essence laying out an idealised paper based on criteria I believe are important for appropriate trans inclusivity in AGR; these

[5] H-index information sourced from Scimago, settings: Computer Science / Computer Vision and Pattern Recognition / 2021

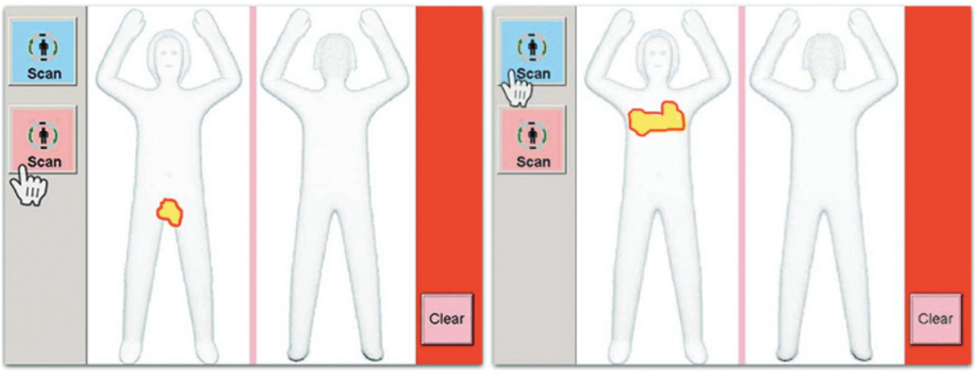

Figure 1. Millimetre wave scan outputs, showing "anomalies" have been flagged when selecting different gender options

Table 1. Journals with publications to be analysed with H-indices from 2018 and 2021, as well as number of publications to be analysed

Publication	H 2018	H 2021	N^o $<$ 2017	N^o $>$ 2017 (analysed/total)
IEEE Trans Pattern Analysis and Machine Intelligence	288	377	8	12 (16)
Proc IEEE Conf Computer Vision and Pattern Recognition	192	408	7	15 (21)
Pattern Recognition	160	218	19	7 (22)
Int J Computer Vision	160	201	4	4 (8)
Proc IEEE Int Conf on Computer Vision	138	280	5	9 (12)
J Optical Society of America A	132	162	— (1)	—/—
Pattern Recognition Letters	122	163	14	0 (7)
$\sum$	—	—	57 (58)	47 (86)

Table 2. Paper evaluation criteria

	Score for category		
	1.0	0.5	0.0
IO-status	IO-Step > 0	IO-Step $= 0$	No reference to IO-status
Reflexivity	Evidence of author reflexivity on gender modality	n/a	No evidence of author reflexivity on gender modality
Ethics	Trans-inclusive ethics section	References to other papers, limited introspection	No reference beyond institutional ethics

Table 3. Keyword search terms

IO-status	Ethics	Reflexivity
trans	ethic	reflex
binary		posit
queer		we*
sexual		I*
fluid		

topics are those identified through the literature review: inclusion of trans stakeholders in research and design, being open about author positionality regarding their gender modality, and a clear reflection on ethical issues regarding trans people and AGR technology.

The criteria are given in Table 2; the papers were skim-read checking for any of the criteria matching the table, and a set of keywords were used to check against the full-text of the paper. Keywords are given in Table 3; all entries were checked using fuzzy-find algorithms (i.e., part of word can appear anywhere within text) apart from those marked with an asterisk.

Software used

To reduce the administrative burden of content analysis, I have created a Python TUI application: *Content Zapper*[6]. As seen in Figure 2, the application allows users to dismiss irrelevant papers and assign a numerical score for each of the three categories previously discussed. The application accepts a number of outputs from online database searches (in CSV format), which are deserialised and stored in a database[7]. Basic statistical analysis will be undertaken using the SQLite command-line interface (to export CSV files) and Microsoft Excel.

RESULTS AND ANALYSIS

In total 104 papers were analysed according to the criteria as proposed previously ($n_{\text{pre-2017}} = 57$; $n_{\text{post-2017}} = 47$), a set-by-set summary is given. Of the 86 papers yielded from the database searches for the post-2017 set, 39 (45%) were excluded due to not being empirical research about facial recognition software that includes AGR technology. No publications

[6]Software made available on request

[7]Libraries used: Textual (for TUI), Orator (for object-relational mapping), and pandas (for serialisation)

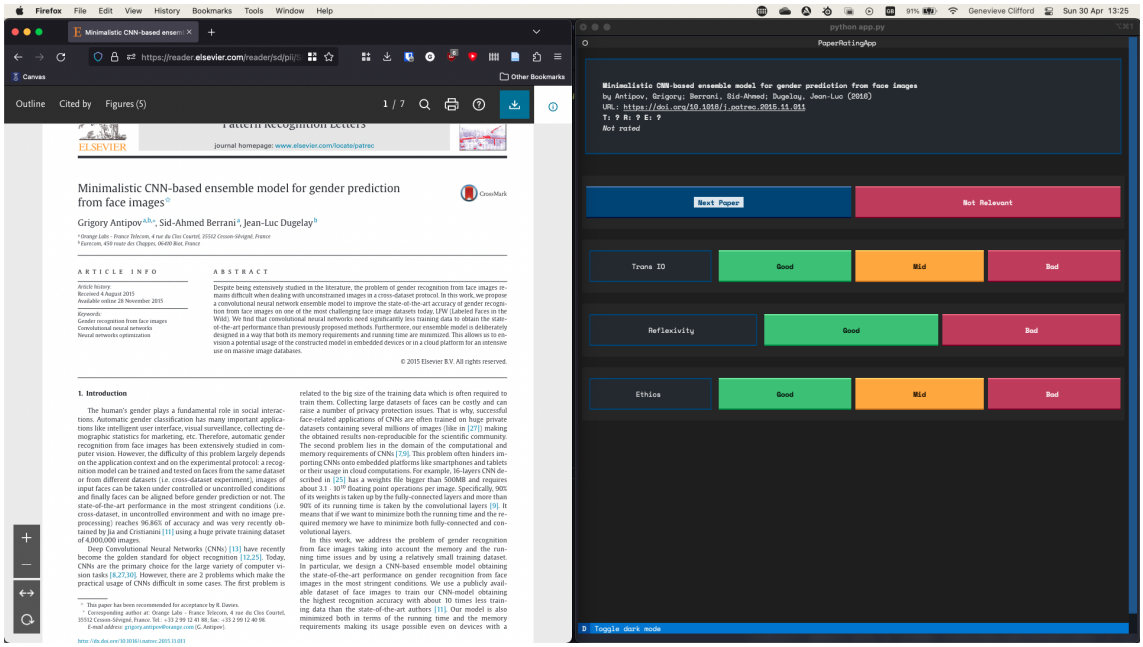

Figure 2. A screenshot of "Content Zapper", an application created to assist with content analysis

analysed had any reflexive statements other than author biographies (none of which contained any references to authorial gender modality).

Pre-2017 publications

> . . . just even a simple *'oh, you know, trans people exist'* would suffice at this point.
>
> You can't really build a gender detector, but these authors seem to think you can.

All publications in this set were technical papers[8], all had scores of 0.0 in both Reflexivity and Ethics metrics; and all except one had scores of 0.0 for IO-status. The exception to this was a paper by Demirkus et al. who state *"the difficulty in estimating gender is largely due to the fact that gender is a continuous trait"* [5, p. 1200]. I do however note (paraphrased) in my reflexive log: *"award 0.5, it's marginal, but this is the closest any paper has come to mentioning that transgender people exist yet"*.

To emphasise, no paper in the pre-2017 set had: any kind of reflexivity regarding author gender modality; explicit mention of any ethical approval; nor used the words *'transgender'*, *'non-binary'*, nor any other mention of any gender modality that isn't cisgender or variations of terms already mentioned. In addition, no paper engaged with transgender people in any meaningful way beyond the zeroth step of Rosenberg and Tilley's *'trans-IO staircase'* model of researcher reflexivity.

Post-2017 publications

> A social-oriented paper! It talks about trans and non-binary people! This is great! References Hamidi, seeks views from trans people. Fantastic, genuinely! (In reference to Su and Crandall [12])
>
> Authors seem to be aware of operationalisation of gender (i.e., mapping of attributes related to gender, clothing, makeup, etc.) but don't critically reflect on this.
>
> Acknowledgement of non-binary gender, but still work relies on gender binary. (In reference to Wang et al. [13] amongst others)

All papers in this set except one (as reflected in the first quote) were technical papers, and all technical papers scored 0.0 on both Ethics and Reflexivity. In contrast with the pre-2017 set a greater proportion of papers had a score for IO-status > 0.0 (21% cf. 2%); such papers in general referred to gender not being a binary concept, thereby showing at least some awareness of the existence of trans people.

As discussed, the one major exception to this was Su and Crandall's *"The Affective Growth of Computer Vision"* [12] which scored 1.0 in both Ethics and IO-status. This is not a technical paper, they collect empirical qualitative research from practitioners in computer vision, seeking their perceptions on the field through writing short stories. This was the only paper analysed to explicitly refer to transgender people, seeking trans-IO input at a level higher than zeroth, as well as explicitly making an informed reference to ethical issues for these systems regarding trans people.

DISCUSSION

From percentages alone, it would be interesting to conclude that authorial bias regarding operationalisation of gender through a binary system has reduced in quantity, the results Keyes and I have collected are not comparable. I looked for explicit references to trans and non-binary people in text, whereas I believe Keyes looked for lack of reference to a binary gender system. It is possible to conclude however, that a greater proportion of technical papers published post-2017

[8] i.e., they implement AGR systems and discuss their results

are more inclusive of non-binary people. This comes with a caveat however, all such references are tokenistic. An example of such is my last reflexive log for post-2017 and also a quote from Wang et al. [14, p. 8917]*"we consider the two color/grayscale domains as purely binary and disjoint whereas the concept of gender is more fluid"*. This is to say, although authors personally acknowledge that non-binary people exist, the research they perform still relies upon perpetuating a binary gender system, given the existence of labelled datasets with binary categories. No papers in either set attempted to classify non-binary genders, although the creation of such a system would be fraught with problems, as it assumes that gender is purely physiological and should be measured through external attributes instead of asking a person about their gender; or as Keyes would describe *"a trans-inclusive system for non-consensually defining someone's gender is a contradiction in terms"* [8, p. 13].

Reflecting on my own positionality as a transfeminine PGR student in computer science, is somewhat depressing to see that not a single paper mentioned concerns raised by trans people besides the one non-technical paper. Although I agree with Keyes' position on the contradiction of trans-inclusion in AGR and the nature of systems that infer gender disregarding self-knowledge, this seems to not have permeated through to what I have read. However, I can see how compromising this positionality may be for authors in the AGR space, as a belief in self-knowledge having primacy in determining gender [8, p. 13, paraph.] would contradict one's continuing existence in the space. It is however difficult to draw conclusions about this, my methodology didn't allow for seeing if individual authors' positions have changed, rather it is more the collective position of a subset of the total authors that are still researching AGR systems. Analysis of this might be conducted using a bibliometric approach, looking at if individual authors have left the field, or stopped publishing as much, followed up with qualitative methods seeking how they now perceive the field. It is this analysis that I feel would be useful to address in future work.

Further on the topic of future work, a literature search that has greater scope (similar to Keyes' orignal methodology) would seek to find work that didn't use the exact set of search terms that I had used. It would also be beneficial to see Keyes' original operationalisation of gender work carried out with this new data set, to allow for better comparisons to be made.

ACKNOWLEDGEMENTS

This paper would not have been possible without Os Keyes; *"The Misgendering Machines"* has clearly influenced this paper, without its publication this paper would not exist. They also graciously provided the set of 58 papers they analysed in that paper.

REFERENCES

[1] Alex Blechman. 2021. Sci-Fi Author: In My Book I Invented the Torment Nexus as a Cautionary Tale Tech Company: At Long Last, We Have Created the Torment Nexus from Classic Sci-Fi Novel Don't Create The Torment Nexus. (Nov. 2021). https://twitter.com/alexblechman/status/1457842724128833538

[2] Sasha Costanza-Chock. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. The MIT Press. DOI: http://dx.doi.org/10.7551/mitpress/12255.001.0001

[3] Cary Gabriel Costello. 2016. The TSA: A Binary Body System in Practice. (Jan. 2016). https://www.transadvocate.com/the-tsa-a-binary-body-system-in-practice_n_15540.htm

[4] Charlotte Aull Davies. 2008. *Reflexive Ethnography: A Guide to Researching Selves and Others* (2nd ed.). Routledge, London ; New York.

[5] Meltem Demirkus, Doina Precup, James J. Clark, and Tal Arbel. 2016. Hierarchical Spatio-Temporal Probabilistic Graphical Model with Multiple Feature Fusion for Binary Facial Attribute Classification in Real-World Face Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 6 (June 2016), 1185–1203. DOI: http://dx.doi.org/10.1109/TPAMI.2015.2481396

[6] Oliver L. Haimson, Avery Dame-Griff, Elias Capello, and Zahari Richter. 2021. Tumblr Was a Trans Technology: The Meaning, Importance, History, and Future of Trans Technologies. *Feminist Media Studies* 21, 3 (April 2021), 345–361. DOI: http://dx.doi.org/10.1080/14680777.2019.1678505

[7] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition Systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–13. DOI: http://dx.doi.org/10.1145/3173574.3173582

[8] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–22. DOI:http://dx.doi.org/10.1145/3274357

[9] Os Keyes and Jeanie Austin. 2022. Feeling Fixes: Mess and Emotion in Algorithmic Audits. *Big Data & Society* 9, 2 (July 2022), 205395172211137. DOI: http://dx.doi.org/10.1177/20539517221113772

[10] Elena Pilipets and Susanna Paasonen. 2022. Nipples, Memes, and Algorithmic Failure: NSFW Critique of Tumblr Censorship. *New Media & Society* 24, 6 (June 2022), 1459–1480. DOI: http://dx.doi.org/10.1177/1461444820979280

[11] Shoshana Rosenberg and P. J. Matt Tilley. 2021. 'A Point of Reference': The Insider/Outsider Research Staircase and Transgender People's Experiences of Participating in Trans-Led Research. *Qualitative Research* 21, 6 (Dec. 2021), 923–938. DOI: http://dx.doi.org/10.1177/1468794120965371

[12] Norman Makoto Su and David J. Crandall. 2021. The Affective Growth of Computer Vision. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9287–9296. DOI: http://dx.doi.org/10.1109/CVPR46437.2021.00917

[13] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Seoul, Korea (South), 5309–5318. DOI: http://dx.doi.org/10.1109/ICCV.2019.00541

[14] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 8916–8925. DOI: http://dx.doi.org/10.1109/CVPR42600.2020.00894