Adhering, Steering, and Queering: Treatment of Gender in Natural Language Generation

Yolande Strengers[1], **Lizhen Qu**[1], **Qiongkai Xu**[2], **Jarrod Knibbe**[1]
[1]Faculty of Information Technology, Monash University, Australia
[2]The Australian National University & Data61 CSIRO, Australia
{Yolande.Strengers, Lizhen.Qu, Jarrod.Knibbe}@monash.edu
Qiongkai.Xu@anu.edu.au

ABSTRACT
Natural Language Generation (NLG) supports the creation of personalized, contextualized, and targeted content. However, the algorithms underpinning NLG have come under scrutiny for reinforcing gender, racial, and other problematic biases. Recent research in NLG seeks to remove these biases through principles of fairness and privacy. Drawing on gender and queer theories from sociology and Science and Technology studies, we consider how NLG can contribute towards the advancement of gender equity in society. We propose a conceptual framework and technical parameters for aligning NLG with feminist HCI qualities. We present three approaches: (1) adhering to current approaches of removing sensitive gender attributes, (2) steering gender differences away from the norm, and (3) queering gender by troubling stereotypes. We discuss the advantages and limitations of these approaches across three hypothetical scenarios; newspaper headlines, job advertisements, and chatbots. We conclude by discussing considerations for implementing this framework and related ethical and equity agendas.

Author Keywords
Feminist HCI; Natural Language Generation.

CCS Concepts
•**Human-centered computing** → **Human computer interaction (HCI);**

INTRODUCTION
Natural Language Generation (NLG) is increasingly employed in HCI to create rich, interactive, personalised interfaces and interactions. For example, NLG empowers virtual assistants to distract children on long journeys [29], helps manage our business relationships [35], and auto-improves accessibility on social networking platforms [71]. However, AI systems have come under increasing scrutiny for their ethical challenges and biases, such as Amazon's hiring algorithms that discriminated against female candidates [19]. In addition, the anthropomorphism of NLG applications, such as chatbots, poses additional challenges as devices increasingly try to act like and relate to people as (gendered) humans or human-like creatures. In turn, people are increasingly relating to these devices as they would to other people. Microsoft's obscene and inflammatory Twitter bot Tay [8, 50], or users' racialized and sexualized treatment of feminized 'bots' [6, 8, 10, 68] provide telling examples of how anthropomorphized devices can exacerbate and perpetuate gender stereotypes and violence in some contexts and societies.

As NLG comes to empower and facilitate more human-machine interactions, we risk it ubiquitously perpetuating and enhancing existing biases. Natural language processing and generation research is exploring ways to address these biases. Today, these approaches typically remove or hide any 'sensitive' attributes (e.g., gender information). In effect, this results in neutralising or hiding gendered language from text. We consider the representation and treatment of gender in NLG systems, exploring how NLG can go beyond the current neutralising of gender, and move society towards a more equitable future. We present a framework for the future of NLG, illustrated by hypothetical scenarios inspired by design fiction [7] and speculative design research [21]. Our contribution is framed within and extends Bardzell's [4] call for feminist HCI to contribute to "an action-based design agenda". This agenda is informed by feminist issues of "agency, fulfillment, identity, equity, empowerment, and social justice". As such, our paper falls into what Bardzell calls a "generative contribution" to feminist HCI, which aims to develop new design insights and tangibly influence the design process - in this case the design of NLG systems and the treatment of gender in the content they generate.

Our key contribution is a conceptual framework featuring three approaches the NLG community could adopt to pursue gender equity; namely, removing gender bias in line with current fairness approaches (*adhering*), amplifying gender differences away from the norm (*steering*) and troubling gender stereotypes (*queering*). We illustrate our framework through three fictional content-generated scenarios (job advertisements, newspaper headlines, and chat-bots) and discuss the possibilities, benefits and limitations of treating text with one or more of our conceptual approaches. These scenarios provide a starting point to explore our conceptual framework and inform

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376315

future research, rather than a comprehensive analysis of NLG scenarios where our framework might apply. As such, this paper should be viewed as a generative contribution and novel experiment in bringing feminist HCI together with NLG to consider content-driven scenarios. Our intention is for this work to inspire further speculation, refinement and testing of our conceptual framework within the HCI and NLG communities both within and beyond the scenarios discussed here.

With this in mind, in the final sections of the paper we attempt to bridge the gap between current gender approaches in AI research and our feminist HCI-inspired framework, by discussing the potential implications and future directions for implementation with the NLG and HCI communities. Importantly, in highlighting the considerable challenges that remain in pursuing a feminist HCI agenda with NLG content, we caution that NLG may not be the best way to address gender equity in many or all scenarios.

BACKGROUND

Natural Language Generation is fast becoming a prevalent tool in the design of interfaces and human-computer interaction. For example, research is exploring automatically generated stories (e.g., [15, 33]) and news articles (e.g., [34]). However, much of HCI's NLG-based interests focus on chatbots and virtual assistants (VA), for example, to empower marginalised communities [2], and to support industry [62] and home-healthcare [14]. Simultaneously, there is a growing body of work reflecting on the ethical principles and guidelines that can and should underpin the selection of category labels and assumptions in texts (e.g. [41, 61]). In this section we discuss human language and how it informs gender bias in NLG models and methods to overcome it.

Gender Bias in Writing

In the fields of sociolinguistics and psycholinguistics, language has a long history of being understood as a system with shifting social signification [16, 47]. People adjust their vocabulary, sounds and syntax depending on who they are speaking to, the context of the situation, and the genders (as well as other social, cultural or demographic markers) of the people or content they are interacting with. The contextual, nuanced and shifting status of language in different societies raises many challenges when considering gender bias in text. In addition, the technological parameters of NLG, which rely on binary classification models, widely applied assumptions, and other simplifications of language, pose additional challenges for addressing gender bias in writing.

To develop and build language models for NLG, practitioners use existing, human-produced texts as training corpuses. People's own writing, however, is heavily biased and in turn we build this bias into our NLG models. The study of computational linguistics reveals that bias in people's writing. The study of gender and language often finds its roots in the work of Lakoff [39]. She argued that gender inequality is "reflected in both the ways women are expected to speak, and the ways in which women are spoken of". This work laid the foundations for the treatment of gender in computational linguistics, namely addressing (i) how women and men talk; and (ii) how people talk about men and women.

To understand how different genders use language[1], computational linguistic research explores the relative frequency of word usage, and specific semantics and topics of discussion. For example, a widely used approach is to apply a predictive classifier, such as logistic regression, to predict gender in input texts [3, 51], or to use statistical tests to find gender associations [12]. Using such an approach on social media corpuses, for instance, Bamman et al. [3], found that female authors use more pronouns, emotional terms (sad, love, etc.), and emoticons, while male authors use more proper nouns, swearing, and taboo words. Similar results are echoed in literary fiction [36], demonstrating that these are not differences between lay and expert writers. Female authors also use more words regarding home, body, and social relationships than their male counterparts, who speak more of occupations, numbers, and prepositions [36].

Similar differences are identified in how people talk about men and women (e.g., [12, 32, 51, 69]). Across multiple genres, research shows how women are associated with language around appearance, fertility, relationship status, and emotions. Men, conversely, are associated with their work, and traditional stereotypes of male characteristics. In student evaluations, we also see differences: female Computer Science professors are more likely to be praised for being communicative and personalizing instruction, while male professors are recognized for being knowledgeable teachers and experts [12].

Even as we move to deep learning approaches to NLP, such as using word embeddings to capture syntactic and semantic properties of words [48, 55], we see further language-based biases. For example, these embeddings reveal analogies such as 'man is to king, as woman is to queen'. However, the same approach also finds problematic stereotypical analogies such as 'man is to computer programmer, as 'woman is to homemaker' [9]. As a result of this gender bias inherent within people's own writing, and replicated by algorithms, NLG research is exploring ways to remove this bias from their language models.

Bias Removal for NLG

A popular technique for the removal of sensitive words is to use blacklist dictionaries (i.e., dictionaries of *words-to-be-censored*). Motivated by Microsoft's Tay bot [50], who quickly became racist and sexist when interacting with (and trained from) Twitter users, Shlesinger et al. [61], unpacked the relationship between race, blacklists, and AIs. The authors highlighted how these blacklists can easily target and silence large communities. Avoidance of the term 'paki', for example, prevents any reference to Pakistan, thus silencing an entire

[1] It is important to note the pervasive treatment of gender as a binary classification in computational linguistics. This has roots in more traditional thinking, such as that of Lakoff [39], but continues to pervade today due to the challenges of collecting gender-labelled data. As gender is enacted in the day-to-day, and thus not easy to categorise for the purpose of labelling training data, researchers in NLP typically simplify to binary male and female labels. Finding solutions to moving beyond this binary classification remains future work.

country. Other research has shown that blacklisting is also not a viable solution for gendered words [61]. For example, in a study of occupation classification [20], researchers blacklisted gendered pronouns and first names from online biographies, but found this was not sufficient to remove gender bias.

Several debiasing techniques have also been proposed for the word embeddings underlying deep-learning-based approaches (e.g., [9, 74, 75]). The key idea is to minimize the information within the gender subspace of an embedding (a multi-dimensional relational word space), whilst retaining as much wider information as possible. However, a recent study [28] finds that the current debiasing techniques only obfuscate the gender bias instead of removing it.

Reframing Bias Removal within Machine Learning Theory

Later, we propose a framework of three approaches through which the NLG community could pursue gender equity from a feminist HCI perspective. Aspects of this framework are closely related to existing machine learning theory. Here, we briefly present the background to this theory and discuss its relation to NLG.

Most NLG models formulate language generation as a machine learning problem [56, 72]. Machine learning theories provide guidelines, analysis tools, and general algorithms to develop NLG models. One such machine learning theory is *fair representation learning* [43, 44], a key problem in the theory of algorithmic fairness [45]. Fair representation aims to suppress information about sensitive attributes in a dataset, whilst retaining non-sensitive information. This is intended to ensure that users make fair decisions, not influenced by sensitive attributes, such as gender. We argue that mitigating bias in NLG is highly related to this problem, as (i) generated texts can be viewed as a representation of their underlying semantic meaning, and (ii) generated text could be considered *fair* if a user makes an unbiased decision based upon it (with respect to a sensitive attribute, e.g., gender). To date, almost all bias detection and debiasing work in NLG is unaware of algorithmic fairness theory (with the exception of Xu et al. [53]).

In order to achieve fair representation, Song et al. [65] introduce two criteria:

- *expressiveness*: maintain as much non-sensitive information as possible (see [44]). In NLG, this would imply meaningfulness, grammatical correctness, and semantic relevance.
- *fairness*: either *group fairness* or *individual fairness*. Group fairness [22] aims for different groups to be treated equally, where individual fairness [37, 38, 73] aims to treat similar individuals equally.

The machine learning research on fairness provides general models and algorithms that can be customized for debiasing NLG models. Recent work [53] has applied algorithmic fairness theory to sentence rewriting, demonstrating the effectiveness of adding group fairness as a constraint into the learning process. More details about related models and algorithms can be found in Mehrabi et al.'s survey [45]. We later return to this theory and unpack how it may come to support gender equality (treating everyone equally), and how we may push further still for gender equity (by, for example, promoting the interests of a marginalized group).

PUTTING FEMINIST HCI IN DIALOGUE WITH NLG

In this paper we follow sociologists such as Butler [11] and Connell [17], and feminist HCI scholars such as Rode [58] and Bardzell [4], in understanding gender as a fluid concept which is continually enacted and performed through interaction. From this starting point, NLG is understood as being performative of gender, by reinforcing or disrupting current biases in ways that ripple through societies. For example, recruitment algorithms [19] can reinforce the stereotype that men are better suited to technical jobs and enact this reality through the hiring practices of a company like Amazon. Algorithms, and other methods of NLG, can therefore be viewed as part of the performances of gender.

To date, the NLG community has only engaged with gender theory at a cursory level, instead being more focused on the removal and detection of bias as discussed above. Likewise, gender has not been a key focus within the CHI community, as noted by Rode [58]. In response, the emergence of gender HCI and - more recently - feminist HCI, provides relevant insights and inspiration for NLG, given its explicit focus on issues of "agency, fulfillment, identity, equity, empowerment, and social justice" in the design of technology interactions [4]

While both gender and feminist HCI is concerned with the gendered relations and identities that shape the use of technologies and their design, feminist HCI - and feminism more broadly, which is often understood as a series of social movements carried out in 'waves' - is specifically oriented towards intervention and action. As Bardzell [4] argues, "by making visible the manifold ways that gender is constructed in everyday life, contemporary feminism seeks to generate opportunities for intervention, making it a natural ally to design."

In her landmark CHI paper, 'Feminist HCI', Bardzell [4] outlines six qualities of feminist HCI: pluralism, participation, advocacy, ecology, embodiment and self-disclosure. While all are potentially relevant, we focus on three of these qualities here: pluralism, advocacy and self-disclosure. We interpret these three qualities as being more focused on the *intent* of a design (or algorithm) to pursue feminist objectives, and therefore most relevant in developing NLG-generated content. The remaining three qualities (participation, ecology and embodiment) emphasize the development of the designs themselves or people's interactions with them. In a more comprehensive analysis, which considers people's interactions with NGL-generated content, *all* of Bardzell's feminist HCI qualities are indeed highly relevant and should be considered.

The first quality we focus on, *pluralism*, involves challenging universal, 'natural', or normative truths. Here, "a key feminist strategy is to denaturalize normative conventions" and explore alternative approaches. Likewise, nurturing and elevating the status of marginal groups can also be key strategy. For NLG systems, denaturalizing normative language (e.g. ceasing the practice of relying on gender stereotypes), and prioritizing marginalized genders represents a pluralist path consistent with this feminist HCI quality.

Second, *advocacy* refers to the advancement of progressive solutions that serve feminist and gender equity objectives of elevating the status of minority groups, such as women and non-binary genders. As Bardzell notes, this is not simply a matter of keeping up with political emancipation but about seeking to bring it about, in turn requiring designers (and others) to question their own positions on what an 'improved society' looks like, and how NLG contributes towards this.

Third, *self-disclosure* "refers to the extent to which the software renders visible the ways in which it effects us as subjects". A key criticism previously levelled at NLG is the lack of disclosure regarding the assumptions that inform the treatment of text and content generation [41]. In critically reflecting on, and making transparent, the assumptions that guide decisions that inform NLG, this feminist HCI quality can help facilitate an ongoing discussion about how to best include, represent, and serve marginalized users.

Drawing on these qualities, we see opportunities for feminist HCI to inform the treatment of text in NLG, and subsequent user interactions with a range of text-based platforms and systems. Applying these qualities to NLG systems provides the possibility to disrupt problematic normative conventions, advocate for improved gender equity and make these aims and assumptions explicit. This, in turn, contributes to the feminist HCI agenda of drawing on gender theories to broaden the repertoire of methodologies available to those designing NLG algorithms/ systems [5]. In subsequent work, the conceptual framework we present below could also provide a starting point to inform research on user experiences with these systems, and evaluations of their societal effects, particularly in regards to the performance of gender and gender equity outcomes.

Following Larson [41], we advocate for "a continual process of thoughtfulness and debate regarding these issues", rather than one 'best' solution or approach. Further, following Schlesinger et al. [61], we reject the notion of 'universal' truths in relation to gender (or any other minority grouping such as race), instead considering all texts as part of situation-specific and culturally-embedded contexts that change over time, as do understandings and performances of gender.

A FRAMEWORK FOR PURSUING FEMINIST HCI IN NLG

In this section we present and discuss three approaches to the treatment of gender under NLG; *adhering*, *steering*, and *queering*. We consider how these approaches align with the three qualities of feminist HCI outlined in the previous section, and how these approaches fit within current techniques, and future opportunities, in NLG research.

Adhering

The first approach most closely resembles the status quo in NLG. It involves *adhering* to the current best practice of erasing or removing any gender bias from text-based algorithms or the content it generates, with minimal loss of information utility. As such, this approach only loosely reflects the three feminist HCI qualities of pluralism, advocacy and self-disclosure we are foregrounding in this paper. Nonetheless, we consider it here by way of comparison with our other approaches, and to explore its potential limitations and advantages across our speculative scenarios.

In regards to pluralism, the approach of adhering does not directly challenge normative truths about gender, although it may reduce unwanted gender assumptions by, for example, removing gendered language from a particular set of texts (such as job applications). Second, this approach could be viewed as a subtle or weak form of advocacy, in that it seeks to remove gender biases that may restrict or limit opportunities for marginalized genders. However, it does not explicitly advocate for a particular gender or genders, nor does it elevate the status of any marginalized genders. It strives for *equality* (in the sense of equal representation of all), but not necessarily *equity* (promotion of marginalized interests). Finally, adhering represents a form of self-disclosure, or self-reflection on the part of NLG designers, so long as decisions regarding the treatment of texts are publicly disclosed, alongside the assumptions that inform them.

Implementing Adhering

As mentioned already, the goal of ahhering is to hide traits of gender. In current work, this is achieved by removing and substituting related words that are associated with stereotypical gendered language [53, 57] (such as removing emotional language, as frequently associated with female authors - as described in the Background, above). One high-level idea approach to achieve adhering commonly used in natural speech, but still underexplored in NLG, is *abstraction*. For example, in order to hide gender traits, we can rephrase 'I went home with my wife' to 'I went home with my partner'. In this way, we replace the implicit indicator 'wife' with its hypernym. Due to the existence of large ontologies with rich 'is-a' relations, appropriate hypernyms to replace gender-associated terms can be easily found. However, the substitution risks leading to ungrammatical or unnatural sentences. Thus, revising techniques need to be devised for ensuring both minimal gender bias and expressiveness of generated language.

The adhering approach is also consistent with the fairness principles from machine learning discussed earlier. As such, the theory on algorithmic fairness provides guidance to NLG researchers and practitioners to implement adhering, and also highlights its limitations. As Mcnamara et al. [44] point out, there is no perfect fairness without loss of information utility. The goal of adhering is to optimise a trade-off between expressiveness and bias. The existing theory can help provide insights into how this optimisation may be best achieved.

Steering

Our second approach of steering is more consistent with feminist HCI qualities, in that it seeks to elevate the status of marginalized groups - in this case a marginalized gender. From an NLG perspective, this might involve amplifying particular gendered characteristics in certain text-based algorithms in order to promote a minority gender, or challenge current stereotypes and norms about that gender. For example, in fields where one gender dominates but where greater gender parity and diversity is a desired societal goal, the aim would be to explicitly reverse the current norm in text-based content in order to challenge biases and social norms. In these situations,

NLG algorithms could seek to make gender attributes *more* explicit (such as in resumes, so that the hiring teams can prioritise certain applicants to pursue gender equity and diversity). Alternatively, in situations where NLG is generating content about particular gender-biased professions (such as nursing or in computer science), algorithms could assist in reversing stereotypical trends, by referring to nurses as men or computer scientists as women or other non-binary genders.

This amplification of a particular gender attribute in the *opposite* direction to the status quo meets the feminist HCI qualities of pluralism and advocacy, with the possibility for self-disclosure. With regards to pluralism, steering is a deliberate and explicit attempt to challenge normative assumptions about gender. This treatment is motivated by advocacy concerns, specifically to elevate the status of minority genders in a particular situation or context. This approach could be either implicit or explicit. For example, the priority promotion of women in Science, Technology, Engineering, Mathematics and Medicine (STEMM) disciplines is generally self-disclosed by those who advocate this approach as being about elevating the status of this marginalized group within these disciplines [70]. However, it would also be possible for algorithms to deliberately hide their attempts to steer away from gender stereotypes by, for example, default gendering all chatbots male without disclosing why this decision has been made.

Implementing Steering

Closely related to steering are the style transfer techniques applied to text [40, 49, 54], if we were to consider the gender of the author as a style. Performing style transfer here would require substituting the indicators of the source gender, with those from the target gender. Similarly, Zmigrod et al. [76] considers flipping the grammatical gender of nouns and their surrounding words in non-English languages. Due to the lack of a 'gold standard' for *steered* text that an algorithm could be trained to target, the evaluation of the transferred text would come to rely on the prediction results of an automatically trained gender classifier. Thus, the target text would be transferred (say, from male-author voice, to female-author), with the intention that the gender classifier would now identify it as 'female'. However, as evident from our background discussion earlier in the paper, there are clear differences between human and computer classifications of gender from text. For example, a human reader may not associate the frequent use of pronouns with female authors. Additionally, using a gender classifier cannot measure and adjust the degree to which gender attributes are made more or less explicit. Instead, we may need data annotated to train a degree classifier. The annotation task is subjective (as are all gender classifications), thus this task may require showing a pair of texts to annotators, and letting them judge which one is more gender explicit. Doing this without simply reproducing gender binaries and stereotypes would be difficult, and therefore there are likely to only be certain situations where this approach will achieve feminist HCI objectives.

In summary, steering could achieve the feminist HCI qualities of pluralism, advocacy and self-disclosure in certain situations (the latter of which depends on the classifying decisions and the reasons for them being disclosed to users of those systems). Additionally, there are existing techniques in NLG that go some way to achieving steering, but these also raise challenges (such as classifying and amplifying gender specific categories) that may prove difficult to realise in practice. This is mainly because steering depends on the gendered interpretation of the reader, which is variable and context specific.

Queering

The third approach we propose draws on queer theory [60] from sociology and its application to HCI as a design "tactic" [42, 66] that can be used to 'trouble' existing stereotypes and normative assumptions about gender. The idea of troubling gender was proposed by Butler [11], referring to the dynamic processes of performing gender identities. The concept of "staying with the trouble" has also been pioneered by Haraway [30], who asks her readers to "make trouble" and "stir up potent response[s]" by making "oddkin" with each other "in unexpected collaborations and combinations". Applied to the concept of gender, staying with the trouble involves unsettling gender stereotypes, binaries, and norms. This is also one of the intentions of queer theory, which has been extensively associated with lesbian, gay, bisexual, queer, intersex, and asexual (LGBQIA) communities, or with anyone who does not conform to mainstream gender or sexual orientations or identities.

For Ahmed [1, 60], queering represents an opening up, widening, or expansion of categories that move beyond binary expressions. For Light [42] and other HCI scholars [66], queering is an alternative to the traditional conservatism of HCI and its focus on user 'needs' which perpetuate the status quo. Light promotes queer design that is "spaceful, oblique and occasionally mischievous". Similarly, Spiel et al. [66] describe queering as "the playful, subversive troubling of existing systems". To queer something then, "is to treat it obliquely, to cross it, to go in an adverse or opposite direction", and to give something "movement and flex" [42].

As a design practice, queering "is predicated on letting (other) values and lifestyles surface - not the ones already in use, but ones that might come to be if allowed enough space to emerge" [42]. Applied to NLG, the role of a text-based algorithm might therefore be to mischievously take certain gender traits in an unconventional direction. For example, queering might involve developing distinctive personalities or genders for text-based chatbots, conversational agents, robots and other AI driven by NLG. This is different from generating a gender neutral chatbot that seeks to erase or hide gender attributes in text-based conversational agents. Instead, it might involve generating creative and playful responses to questions that would normally request a gendered outcome. Feldmen [23] has pioneered development along this path through Kai - her genderless banking assistant chatbot. Kai is designed with a quirky bot personality that redirects gendered questions back to its distinctiveness as an algorithmically-generated text-based system, using humour and respectful dialogue [70]. A queering approach is therefore pluralist in challenging normative truths that all anthropomorphic devices and content needs to be gendered, or indeed that gender is heteronormative and confined

to the male/ female binary. It pursues a feminist advocacy agenda by ensuring that marginalized or stereotypical genders aren't associated with assistant chatbots (as is currently the case with the majority digital voice assistants which have female voices and personalities). It is also inherently political in its orientation, intentionally advocating for a troubling of gender and its associated norms. Queering can also directly pursue self-disclosure by reinforcing itself as a suite of algorithms, rather than pretending to act like a person or embody a gendered identity and personality.

Implementing Queering

Queering represents a new direction and challenge for NLG research, as it requires a careful coupling of naturalness and creativity. Computational creativity in language is still in its infancy [18,26,46], and the lack of theory and widely accepted evaluation methods [26] are a major obstacle in this area. The very concept of queering calls for considerations beyond gender and binary classifications, to include playful adaptations of subjects and contexts, and the pursuit of an explicit social advocacy agenda pioneered out by NLG programmers. This is likely to require new inter-disciplinary partnerships between the NLG and HCI communities, since it is unlikely that NLG practitioners will be comfortable with pursuing such an agenda on their own.

Despite these challenges, NLG researchers can begin to formulate the requirements of queering into specific training goals, namely balancing creativity and meaningfulness. Algorithmically, creativity suggests adding randomness to the generation process, such that one might consider any generated text to be 'outliers', when compared to the typical outputs. This randomness can be achieved by maximizing the entropy of word distributions. However, the outliers may easily lead to nonsensical text, or inadvertently generate other gender *faux pas* that could further amplify biases. One possible solution is to add constraints to limit the randomness of word selection, such that we may randomly substitute words with their neighbouring words in an ontology (where neighbors are close lexical relatives). As with *adhering*, we continue to use loss terms to further constrain for meaningfulness, grammatical correctness, and semantic relevance.

Furthermore, hiding gender traits can again follow the theory of fair representation learning, which may encourage both removal of all gender traits, or mixing gender traits as per the queering approach. If we want to encourage mixing gender traits, we could add constraints to make sure that the probability of observing all gender types are above a certain threshold.

Due to the expected creativity of queering, a gold standard evaluation corpus will be difficult to cover all creative and meaningful outputs. Here especially, then, it is important to couple this approach to user-centered data collection, in order to fully understand the efficacy of this approach. In summary, queering is an experimental and aspirational approach that requires further testing and refinement, both in defining its technical parameters, and in conducting research with users of queering-generated content to understand how and in what ways feminist HCI qualities are or can be met.

ADHERING, STEERING AND QUEERING: THREE SCENARIOS

In the absence of NLG and user studies that test our conceptual approach in the 'real world', we turn to three hypothetical content-generated scenarios to further consider their implications. These are: job advertisements, newspaper headlines, and chatbots. The scenarios demonstrate the changing context in which gender equity or feminist goals might be pursued in realistic situations. They highlight the need to move beyond a 'one size fits all' approach to addressing gender equity and fairness concerns in NLG research and subsequent HCI applications.

We selected these scenarios in relation to four criteria: i) NLG-generated content that already exists in the public domain; (ii) widely studied content in the NLG research community [24, 63]; iii) NLG content which has attracted public and scholarly attention for generating gender biased outcomes (specifically advertisements, news and dialogues); and iv) complementary, but also contrasting, examples that allow us to explore the application of our framework in different situations. More specifically, job advertisements are an example of NLG applications with high commercial impact, news content is one of the most extensively studied media in NLG [24], and chatbots are one of the most ubiquitous devices in the world, and a significant NLG application [70].

The examples we present are fictional, and inspired by design fiction for ethics and gender research [7,64] and speculative design thinking [21]. We developed the scenarios through generative conversations between our multi-disciplinary author team, informed by our conceptual framework, feminist HCI qualities, and recent advances in NLG. Like other forms of speculative design, these scenarios invite reflection on the societal implications of possible texts - particularly in regards to their ability to realise feminist HCI qualities through NLG.However, the scenarios are limited to a select few, and need to be further tested and considered with a range of other possible and emerging scenarios. This would likely lead to further refinement or additional elements to our current conceptual framework. Our scenario speculation process is also limited by the lack of input from others, and would benefit from further development using approaches such as co-design or participatory design.

Job advertisements

Our first scenario involves applying NLG to job advertisements. A suggested application of our three approaches to this scenario is provided in Figure 1. For the first approach (adhering), treatment involves neutralising the text, or erasing any identifying gender (or other sensitive) attributes from the advertisement to facilitate equality. This is already a common strategy when seeking to hire more women in male-dominated disciplines or professions (where job descriptions commonly employ male-biased language [25]). This approach is also consistent with the existing role of NLP in unconscious bias software (routinely used by some companies) which seeks to remove stereotypical masculine or feminine language from text [70]. It is therefore less design *fiction*, and more in line with current best practice for addressing gender bias in job advertisements.

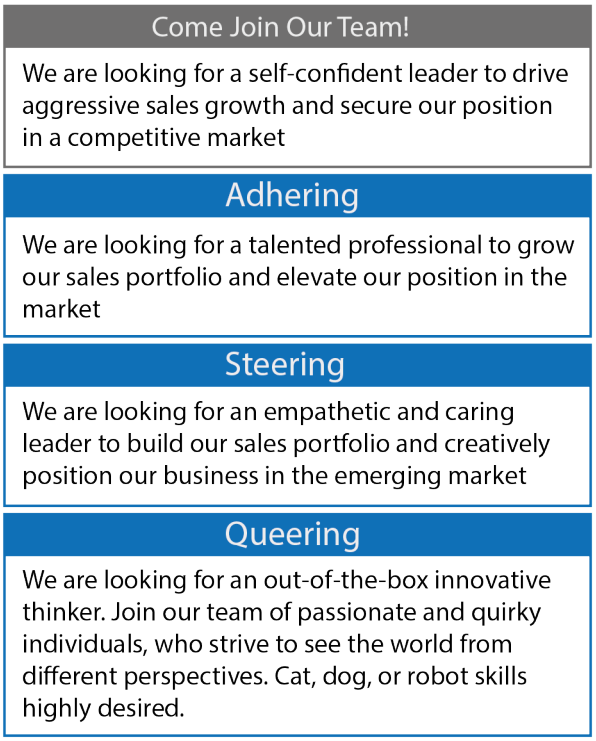

Figure 1. NLG as applied to job advertisements: adhering, steering and queering examples

For the second approach (steering), job advertisement language would be deliberately biased towards a particular gender or genders that the employer was encouraging to apply. This might be appropriate for professions under-represented by women, for example, where people identifying with this gender are already more likely to undervalue their own skills and expertise relative to men's (such as in computing and technology fields) [70]. In this example, the text would be deliberately biased towards stereotypical feminine language in an attempt to encourage more women to apply, and detract men from applying. This approach is already possible using existing NLG techniques, as outlined earlier.

The final approach (queering) involves being deliberately playful or mischievous with job advertisements by emphasizing non-normative language and steering away from typical masculine or feminine language. For example, queering a job advertisement might involve emphasizing unusual attributes from candidates (such as 'out of the box' thinking and 'quirky' individuals), or highlighting the eclectic culture of the workplace (e.g. 'cat, dog, or robot skills highly desired'). Queering is more speculative than the other two approaches. It is difficult to determine exactly how NLG could generate this kind of text, as we discussed previously in the Queering section.

Importantly, all three approaches would likely result in different realizations of feminist HCI qualities within this scenario. In the adhering approach, existing research suggests that people identifying with all genders are more likely to feel welcome applying for the job [70]. However, this may not go far enough to rectify a significant imbalance in gender representation within a field or profession, therefore failing to deliver advocacy. Steering is more likely to encourage applicants from a non-normative gender in a particular field or profession to apply [70], resulting in both pluralist and advocacy outcomes. The third approach is untested and purely fictional. It's unclear how queering would affect the gender identify of job applicants. However, given what is known about language and gender, it may result in greater gender diversity in applicants, particularly from those who don't identify with stereotypical norms. It's difficult to see how each example would deliver self-disclosure outcomes as standalone statements. They would need to be accompanied by a transparent and explicit company policy on gender and inclusion.

Newspaper headlines

Newspaper headlines are another form of text where gender bias can be problematic. For example, freelance journalist Gilmore [27] runs the FixedIt campaign dedicated to fixing media reports of male violence against women. Her work seeks to correct preconceptions and subconscious biases about violence towards women (such as the idea that what a woman is wearing is relevant to the crime). Her headline 'fixes' also seek to distinguish sex from rape, and ensure that the gendered nature of sexual violence is explicitly discussed and acknowledged in newspaper text. Although headlines are manually 'fixed' by Gilmore to ensure that primarily male perpetrators and their actions are the focus of headlines (rather than women, children or other victims and their actions), there is also a potential role for NLG here in automatically tracking and correcting headlines that reflect a particular gender bias.

In Figure 2, we explore this possibility with the following fictional headline: 'Men need to do more to help around the home, study finds'. The headline reflects a common problem highlighted by gender and feminist scholars and commentators [31, 59], in that it perpetuates the idea that men are subordinate to women when it comes to doing housework. Men's role is positioned as a 'helper', suggesting that women are responsible for running a household, managing housework and delegating it out to men and other household members.

An adhering approach to this scenario would involve removing references to gender, and focusing on the need for everyone to equally contribute to the housework. A steering approach would seek to correct the inherent bias in this headline by amplifying the gender problem evident. This might involve more explicitly calling for men to take on their 'equal share' of housework or, more controversially, suggesting that men need to become the managers of housework (which is the reverse of current stereotypical arrangements). A queering approach might focus on something different from the heterosexual norm whilst still being provocative, such as asking whether children should be responsible for doing all the housework, or asking a question about who should clean up after the dog or cat?

In this scenario, it is hard to see how an adhering approach serves feminist HCI qualities, given that it deliberately erases gender from what is an inherently gendered issue. It therefore potentially reinforces Gillmore's [27] critique of gendered violence headlines, which hide or mask the gender of most

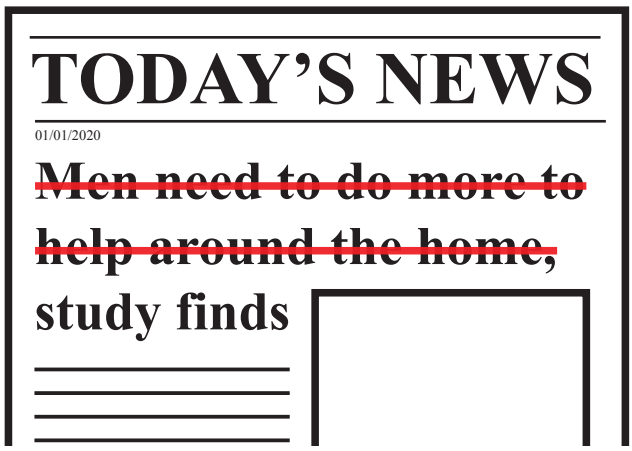

Figure 2. Re-framing newspaper headlines: adhering, steering and queering examples

perpetrators (e.g. men). Likewise, in the example provided in Table 2, removing gender identifiers from the headline erases discussion of the actual problem: specifically that men do less housework overall than women [31]. For this scenario, we therefore argue that adhering is not an appropriate strategy for pursuing feminist HCI qualities.

Steering, on the other hand, elevates the gendered problem of who does the majority of housework and seeks to advocate for greater equity at home by making this issue explicit. Given its ability to directly discuss the problem at hand, it is a more appropriate approach for this particular scenario. Queering, while still potentially valuable as an approach, may unintentionally direct attention onto other distracting and seemingly 'trivial' issues (such as children doing more housework). While it may also provoke some valuable reflection, discussion and outcomes for elevating the status of minority genders and groups, there is a risk that queering may divert attention away from the principle gendered concern in this scenario.

Chatbots

Our final scenario focuses on the text generated for chatbots or conversational agents (see Figure 3). Our example focuses on a fictional opening introductory line for a chatbot assistant, the majority of which are gendered female when performing feminised roles such as reception, administration or housekeeping duties [8, 52, 67]. In the first approach - adhering - any gendered attributes including the chatbot's gendered name (Samantha) is neutralised, and the suggestion that the chatbot is 'looking forward to having some fun' with its user is removed. This approach therefore disrupts gender norms by disassociating the chatbot with a stereotypical female character. It indirectly advocates for a professional and respectful relationship between the bot and its users by removing references to 'having fun' which could be interpreted as being flirtatious when delivered by an overtly feminised bot (as demonstrated by critiques of the flirtatious Ms Dewey assistant [68]). It also moves towards self-disclosure by removing suggestions that a bot can behave like a person, or is somehow like a human-in-service, which can 'look forward' to helping its users (as critiqued by Chassin [13]).

Applying the second approach - steering - to this scenario, also upholds feminist HCI qualities by bucking gendered stereotypes and advocating for a male chatbot performing traditionally feminized tasks. However, gendering the bot male potentially falls prey to the same lack of self-disclosure critiques levelled at female chatbots, in that the bot is represented as having a human identity and stereotypical male characteristics (e.g. gender). This anthropomorphized approach may encourage users to humanize bots, and form emotional attachments with a 'friendly' and familiar device that may unwittingly expose them to broader security or privacy concerns [8, 67].

The final approach - queering - could take the chatbot in other surprising directions. In the example provided in Figure 3, we focus on generating text which highlights the bot's unique bot personality (following Feldmen's work with Kai, previously discussed), and the bot's self-disclosure of what it is, what it is doing, and how it is providing a service to the user. The bot avoids gender not by presenting itself as gender neutral, but by presenting itself as a device for which gender is not relevant. In this example, Sam is neither male nor female; it is a different entity entirely.

TESTING AND IMPLEMENTING THE FRAMEWORK

The above speculative discussion highlights the value of pursuing different approaches to gender treatment in NLG, depending on the scenario and context. Our adhering, steering and queering framework is not a complete representation of the possible approaches NLG researchers could take for addressing gender or other minority variables. Nor should our hypothetical examples be viewed as a fixed set of opportunities. As we have already noted, gender varies over time and in different situations. For example, what one culture might interpret as masculine language may not be the same as another. Implementing such a framework should therefore be considered a reflexive process that aims to better serve feminist HCI qualities through continual improvement. This process can be aided by at least three further steps which we discuss below: (i) the refinement of implementation guidelines, (ii) continual assumption testing and querying, and (iii) research with people who engage with text-based content.

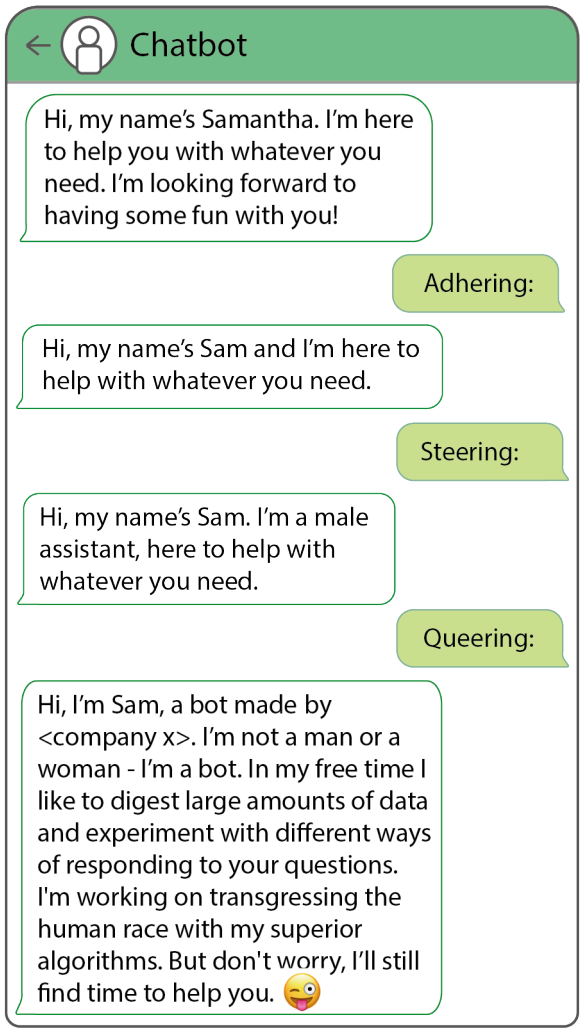

Figure 3. Chatbots: adhering, steering and queering examples

Implementation guidelines

Larson [41] proposes four gender-specific guidelines intended to make the decisions underpinning NLG and its content more transparent, accountable and fair: (i) formulating research questions with explicit theories of gender, (ii) avoiding using gender as a variable unless necessary or directly relevant to the research questions, (iii) making the methods for assigning gender categories to participants and linguistic markers explicit, and (iv) respecting difficulties when relying on data where respondents are asked to self-identify their gender.

Placing these guidelines in a feminist HCI lens, and applying them to our adhering, steering and queering framework, requires some modification. While the first, third and fourth guidelines are directly relevant to the feminist HCI qualities of plurality and self-disclosure, the second becomes redundant, given that NLG adopting a feminist HCI stance must always consider gender from the outset. This guideline might therefore be reformulated as: (ii) *ensuring* gender is a variable in all NLG research. In addition, a feminist HCI agenda involves advocacy. It therefore requires a fifth guideline: (v) Adopting NLG approaches that aim to serve gender equity and feminist objectives by elevating the status of minority groups.

Testing and querying assumptions

Given the dynamic nature of gender, our second consideration follows Haraway's call (taken up by other HCI researchers exploring race and gender issues [61, 64]) to 'stay with the trouble'. In our case, staying with the trouble involves holding onto the complexity of gender when considering what assumptions to use in developing algorithms for text-based content generation. What is considered queer in one context, for example, may not hold true for another. Likewise, what needs to be steered in one time and place may be the opposite in another.

Staying with the trouble also invites NLG and HCI researchers to hold space for an ongoing question about where and how assumptions are generated and whether they productively serve feminist HCI objectives in different scenarios and contexts. For example, in regards to the hypothetical scenarios proposed in this paper, our text-based examples assume a particular western-centric and largely heteronormative understanding of gender that is unlikely to serve feminist HCI qualities in many situations. In generating fictional examples, we brought our own assumptions to bear on this text that deserve further testing and investigation through engagement with gender theory, and through research with readers and users of this content. Staying with the trouble in our scenarios, therefore involves maintaining reflexive thinking about our own biases, and not taking for granted that adhering, steering or queering will always lead to better gender equity outcomes.

In addition, staying with the trouble invites both NLG and HCI researchers to test this framework, and the feminist HCI principles we have pursued here, in different scenarios. Many questions remain that require further consideration. For example, does self-disclosure actually matter in all situations involving interactions between humans and NLG content? Do different considerations apply if the content comes from a bot, newspaper or company? How does 'chatter' differ from more formal text in regards to how people interact with it and the gender expectations they might hold or interpret? And how do all of these things vary in different social and cultural contexts? Such questions remind us that we should not assume that NLG content is suitable, desirable or even possible in many emerging situations where it is currently being employed or proposed. We therefore need to remain critical about the value of NLG in the scenarios we have proposed. Additionally, these questions highlight the need for ongoing collaboration with the HCI community to test gender assumptions in research with users in real-world situations.

Research and collaboration across HCI and NLG

The HCI community is uniquely placed to pursue a feminist HCI agenda with NLG researchers. This is because we are increasingly involved in designing interfaces that depend on NLG content, or conducting research with users of those devices. HCI collaboration is particularly important for steering and queering approaches, where user reactions and interpretations are less known and, in the case of queering, more open to

different viewpoints given its focus on experimental and playful content generation. Likewise, the growing body of HCI research on chatbots and conversational agents, can both build on the adhering, steering, and queering framework proposed in this paper, and research their effectiveness in pursuing feminist HCI objectives. For example, our paper invites further speculation on what a chatbot designed to steer or queer gender might look and feel like. What kind of interactions would it facilitate? How can these best support NLG text designed to challenge gender norms, elevate the status of minority groups, or disclose gender biases and assumptions? Likewise, for the other scenarios of advertisement text and newspaper headlines, this paper invites further reflection and interaction designs that pursue a feminist HCI agenda through one of the approaches we propose (or additional approaches not yet identified). Conversely, research with users who interact with the kinds of content scenarios we have suggested can feed back into the design of NLG techniques and algorithms.

However, several challenges remain in pursuing this research agenda. First, these considerations can only be explicitly pursued when the approach to gender is disclosed by NLG and HCI researchers, and when the objectives of those approaches (such as to pursue a feminist HCI agenda) are known and shared by the relevant community. As we have already suggested, this involves an explicit focus on testing and querying gender assumptions in different scenarios and contexts. Second, to pursue the kind of research we propose here, HCI researchers may need to adjust their benchmarks for success. A typical indicator of a successful design in HCI has been whether it inspires easy and (user) 'friendly' interactions . However, our feminist HCI framework for NLG may require something quite different of the HCI community. Steering or queering gender, for example, may provoke some discomfort as gender stereotypes and norms are challenged. It may inspire designs that are not initially or always 'liked', but could become respected over time, or change gender perceptions. Thus, the HCI community may need to consider alternative benchmarks for success when experimenting with the ideas proposed in this paper.

CONCLUSION

The key contribution of this paper is a conceptual framework featuring three approaches (adhering, steering and queering) for pursuing a feminist HCI agenda in the design of NLG text-based algorithms and the content they generate. We have speculated on how these approaches may offer differing outcomes in three fictional content-based scenarios (job advertisements, newspaper headlines and chatbots), noting that these scenarios are by no means an exhaustive list of possible options or contexts through which to explore our framework. While we have not yet investigated the *actual* effects or outcomes of these hypothetical scenarios in real-life situations, we have provided considerations to guide their implementation and shape further speculation and research.

Our analysis is unique, and results from the inter-disciplinary collaboration of the authors (spanning the disciplines of HCI, sociology and NLG). First, we placed a field (NLG and NLP more broadly) that has most commonly treated gender as a predictable variable, in dialogue with sociological theories of gender, feminism and queering. While we are not the first researchers to have done this, our framework extends current calls to remove gender bias in NLG by orienting it towards a feminist HCI agenda. This has allowed us to consider how NLG and the content it generates can advocate for the interests of minorities and seek to disrupt problematic gender stereotypes in order to advance equity.

Second, in addition to proposing a framework, we commented on the practicalities of operationalizing our three approaches in NLG research. This involved placing our conceptual ideas in dialogue with current and emerging NLG techniques and methods, and exploring these through three speculative scenarios. This has allowed us to provide some initial considerations for demonstrating the feasibility and applicability of our framework for NLG researchers. Finally, we placed our framework in dialogue with HCI research - both by applying a feminist HCI lens to NLG-derived content, and by considering how HCI researchers and practitioners could extend this framework into their own work by, for example, studying user interpretations and different performances of gender that arise through textual scenarios that adopt our suggested approaches. Such considerations are a crucial step in realising a feminist agenda and reducing gender bias in NLG content, which the HCI community is best placed to contribute to.

More broadly, this work contributes to both the NLG and HCI communities' concerns with pursuing equitable or ethical outcomes through technology and design. Bardzell's [4] feminist HCI qualities of plurality, advocacy and self-disclosure, which we have highlighted in this paper, could be applied to pursue better outcomes for any minority group who are disproportionately affected, under-represented, or marginalized through technologies and content delivered via NLG or machine learning. Likewise, the fictional scenarios and implementation steps provide a lens for thinking through how to approach equity and ethical issues in a range of situations and with regards to the many potential biases NLG systems potentially generate.

However, we caution that this paper should not be read as a universal endorsement for the application of NLG-derived content in the scenarios proposed here, or many that lie beyond. While NLG provides a method of reaching millions if not billions of people very quickly, and is being used for an increasing range of applications globally, it may not always be the best way to pursue gender equity outcomes. In particular, NLG risks losing much of the nuance and contextually-specific details that shape gender in diverse cultures and societies, as we have discussed in this paper. Nonetheless, through articulation of a conceptual framework intended to pursue a feminist HCI agenda, and illustrated through fictional scenarios that stay with the trouble of gender, this paper provides considerations for enabling future interactions with technologies. Specifically, it adds to the ongoing discussion of how NLG and HCI practitioners can go beyond 'levelling the playing field' through current methods of neutralising gender. And finally, it provides the building blocks for pursuing an agenda that elevates the status of and opportunities for minority groups in societies interacting with NLG-derived content.

REFERENCES

[1] Sara Ahmed. 2006. *Queer phenomenology: Orientations, objects, others*. Duke University Press.

[2] Matthias Baldauf, Raffael Bösch, Christian Frei, Fabian Hautle, and Marc Jenny. 2018. Exploring Requirements and Opportunities of Conversational User Interfaces for the Cognitively Impaired. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI '18)*. ACM, New York, NY, USA, 119–126. DOI:http://dx.doi.org/10.1145/3236112.3236128

[3] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18, 2 (2014), 135–160.

[4] Shaowen Bardzell. 2010. Feminist HCI: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1301–1310.

[5] Shaowen Bardzell and Jeffrey Bardzell. 2011. Towards a Feminist HCI Methodology: Social Science, Feminism, and HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 675–684. DOI: http://dx.doi.org/10.1145/1978942.1979041

[6] Christoph Bartneck and Jun Hu. 2008. Exploring the abuse of robots. *Interaction Studies* 9, 3 (2008), 415–433.

[7] Eric P.S. Baumer, Timothy Berrill, Sarah C. Botwinick, Jonathan L. Gonzales, Kevin Ho, Allison Kundrik, Luke Kwon, Tim LaRowe, Chanh P. Nguyen, Fredy Ramirez, Peter Schaedler, William Ulrich, Amber Wallace, Yuchen Wan, and Benjamin Weinfeld. 2018. What Would You Do?: Design Fiction and Ethics. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork (GROUP '18)*. ACM, New York, NY, USA, 244–256. DOI: http://dx.doi.org/10.1145/3148330.3149405

[8] Hilary Bergen and others. 2016. 'I'd blush if I could': Digital assistants, disembodied cyborgs and the problem of gender. *Word and Text, A Journal of Literary Studies and Linguistics* 6, 1 (2016), 95–113.

[9] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*. 4349–4357.

[10] Sheryl Brahnam. 2006. Gendered bods and bot abuse. In *Proceedings of CHI06 Workshop On the Misuse and Abuse of Interactive Technologies, Montréal, Québec, Canada*. 13–17.

[11] Judith Butler. 2002. *Gender Trouble*. Routledge.

[12] Serina Chang and Kathleen McKeown. 2019. Automatically Inferring Gender Associations from Language.

[13] Alexandra Chasin and others. 1995. Class and its close relations: Identities among women, servants, and machines. *Posthuman bodies* (1995), 73–96.

[14] Eugene Cho. 2019. Hey Google, Can I Ask You Something in Private?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 258, 9 pages. DOI:http://dx.doi.org/10.1145/3290605.3300488

[15] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 329–340. DOI: http://dx.doi.org/10.1145/3172944.3172983

[16] Herbert H Clark. 1994. Discourse in production. (1994).

[17] Raewyn Connell. 2005. *Masculinities*. Polity.

[18] Amitava Das and Björn Gambäck. 2014. Poetic Machine: Computational Creativity for Automatic Poetry Generation in Bengali.. In *ICCC*. 230–238.

[19] Jeffrey Dastin. 2018. Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women, REUTERS. Online. (9 October 2018). Retrieved September 19, 2019 from https://perma.cc/5UPB-NHLE.

[20] Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*. 120–128.

[21] Anthony Dunne and Fiona Raby. 2013. *Speculative everything: design, fiction, and social dreaming*. MIT press.

[22] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science conference*. 214–226.

[23] Jacqueline Feldman. 2017. The Dignified Bot, The Paris Review. Online. (13 December 2017). Retrieved April 24, 2019 from https://www.theparisreview.org/blog/2017/12/13/the-dignified-bot.

[24] Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61 (2018), 65–170.

[25] Danielle Gaucher, Justin Friesen, and Aaron C Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology* 101, 1 (2011), 109.

[26] Pablo Gervás. 2019. Exploring Quantitative Evaluations of the Creativity of Automatic Poets. In *Computational Creativity*. Springer, 275–304.

[27] J. Gilmore. 2019. *Fixed It*. Penguin Random House Australia. https://books.google.com.au/books?id=SkWQDwAAQBAJ

[28] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862* (2019).

[29] Michal Gordon and Cynthia Breazeal. 2015. Designing a Virtual Assistant for in-Car Child Entertainment. In *Proceedings of the 14th International Conference on Interaction Design and Children (IDC âĂŹ15)*. Association for Computing Machinery, New York, NY, USA, 359âĂŞ362. DOI: http://dx.doi.org/10.1145/2771839.2771916

[30] Donna J Haraway. 2016. *Staying with the trouble: Making kin in the Chthulucene*. Duke University Press.

[31] G. Hartley. 2018. *Fed Up: Emotional Labor, Women, and the Way Forward*. HarperOne. https://books.google.com.au/books?id=l3xLDwAAQBAJ

[32] Alexander Hoyle, Hanna Wallach, Isabelle Augenstein, Ryan Cotterell, and others. 2019. Unsupervised Discovery of Gendered Language through Latent-Variable Modeling. *arXiv preprint arXiv:1906.04760* (2019).

[33] Ting-Yao Hsu, Yen-Chia Hsu, and Ting-Hao (Kenneth) Huang. 2019. On How Users Edit Computer-Generated Visual Stories. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM, New York, NY, USA, Article LBW2711, 6 pages. DOI: http://dx.doi.org/10.1145/3290607.3312965

[34] Soomin Kim, JongHwan Oh, and Joonhwan Lee. 2016. Automated News Generation for TV Program Ratings. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video (TVX '16)*. ACM, New York, NY, USA, 141–145. DOI: http://dx.doi.org/10.1145/2932206.2933561

[35] Meng-Chieh Ko and Zih-Hong Lin. 2018. CardBot: A Chatbot for Business Card Management. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion (IUI âĂŹ18 Companion)*. Association for Computing Machinery, New York, NY, USA, Article Article 5, 2 pages. DOI: http://dx.doi.org/10.1145/3180308.3180313

[36] Corina Koolen and Andreas van Cranenburgh. 2017. These are not the Stereotypes You are Looking For: Bias and Fairness in Authorial Gender Attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. 12–22.

[37] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. 2019a. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 1334–1345.

[38] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. 2019b. Operationalizing Individual Fairness with Pairwise Fair Representations. *arXiv preprint arXiv:1907.01439* (2019).

[39] Robin Lakoff. 1973. Language and woman's place. *Language in society* 2, 1 (1973), 45–79.

[40] Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text rewriting. (2018).

[41] Brian N Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. (2017).

[42] Ann Light. 2011. HCI as heterodoxy: Technologies of identity and the queering of interaction with computers. *Interacting with Computers* 23, 5 (03 2011), 430–438. DOI:http://dx.doi.org/10.1016/j.intcom.2011.02.002

[43] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309* (2018).

[44] Daniel McNamara, Cheng Soon Ong, and Robert C Williamson. 2019. Costs and benefits of fair representation learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 263–270.

[45] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. *arXiv preprint arXiv:1908.09635* (2019).

[46] Mateus Mendes, Francisco C Pereira, and Amílcar Cardoso. 2004. Creativity in natural language: Studying lexical relations. In *The Workshop Programme*. 44.

[47] Miriam Meyerhoff. 2015. *Introducing sociolinguistics*. Routledge.

[48] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[49] Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating Style Transfer for Text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 495–504.

[50] Gina Neff and Peter Nagy. 2016. Automation, algorithms, and politics| talking to Bots: Symbiotic agency and the case of Tay. *International Journal of Communication* 10 (2016), 17.

[51] Nitya Parthasarthi, Sameer Singh, and others. 2019. GenderQuant: Quantifying Mention-Level Genderedness. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2959–2969.

[52] Thao Phan. 2019. Amazon Echo and the Aesthetics of Whiteness. *Catalyst: Feminism, Theory, Technoscience* 5, 1 (2019).

[53] Xu Qiongkai, Qu Lizhen, Chenchen Xu, and Ran Cui. 2019a. Privacy-Aware Text Rewriting. In *Proceedings of the 12th International Conference on Natural Language Generation*.

[54] Xu Qiongkai, Chenchen Xu, and Lizhen Qu. 2019b. ALTER: Auxiliary Text Rewriting Tool for Natural Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

[55] Lizhen Qu, Gabriela Ferraro, Liyuan Zhou, Weiwei Hou, Nathan Schneider, and Timothy Baldwin. 2015. Big data small data, in domain out-of domain, known word unknown word: The impact of word representation on sequence labelling tasks. *CoNLL* (2015).

[56] Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*. 142–149.

[57] Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*. 17–26.

[58] Jennifer A Rode. 2011. A theoretical agenda for feminist HCI. *Interacting with Computers* 23, 5 (2011), 393–400.

[59] Jennifer A. Rode and Erika Shehan Poole. 2018. Putting the Gender Back in Digital Housekeeping. In *Proceedings of the 4th Conference on Gender & IT (GenderIT '18)*. ACM, New York, NY, USA, 79–90. DOI:http://dx.doi.org/10.1145/3196839.3196845

[60] Ahmed Sara. 2017. Living a Feminist Life. (2017).

[61] Ari Schlesinger, Kenton P. O'Hara, and Alex S. Taylor. 2018. Let's Talk About Race: Identity, Chatbots, and AI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 315, 14 pages. DOI: http://dx.doi.org/10.1145/3173574.3173889

[62] Benedikt Schmidt, Reuben Borrison, Andrew Cohen, Marcel Dix, Marco Gärtler, Martin Hollender, Benjamin Klöpper, Sylvia Maczey, and Shunmuga Siddharthan. 2018. Industrial Virtual Assistants: Challenges and Opportunities. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp '18)*. ACM, New York, NY, USA, 794–801. DOI: http://dx.doi.org/10.1145/3267305.3274131

[63] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue & Discourse* 9, 1 (2018), 1–49.

[64] Marie Louise Juul Søndergaard and Lone Koefoed Hansen. 2018. Intimate Futures: Staying with the Trouble of Digital Personal Assistants Through Design Fiction. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 869–880. DOI: http://dx.doi.org/10.1145/3196709.3196766

[65] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. 2019. Learning controllable fair representations. *arXiv preprint arXiv:1812.04218* (2019).

[66] Katta Spiel, Os Keyes, Ashley Marie Walker, Michael A. DeVito, Jeremy Birnholtz, Emeline Brulé, Ann Light, PBarlas, Jean Hardy, Alex Ahmed, Jennifer A. Rode, Jed R. Brubaker, and Gopinaath Kannabiran. 2019. Queer(Ing) HCI: Moving Forward in Theory and Practice. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM, New York, NY, USA, Article SIG11, 4 pages. DOI: http://dx.doi.org/10.1145/3290607.3311750

[67] Yolande Strengers, Jenny Kennedy, Paula Arcari, Larissa Nicholls, and Melissa Gregg. 2019. Protection, Productivity and Pleasure in the Smart Home: Emerging Expectations and Gendered Insights from Australian Early Adopters. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 645, 13 pages. DOI:http://dx.doi.org/10.1145/3290605.3300875

[68] Miriam Sweeney. 2014. *Not just a pretty (inter) face: A critical analysis of Microsoft's' Ms. Dewey'*. Ph.D. Dissertation. University of Illinois at Urbana-Champaign.

[69] Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

[70] Mark West, Rebecca Kraut, and Han Ei Chew. 2019. I'd blush if I could: closing gender divides in digital skills through education. (2019).

[71] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic Alt-Text: Computer-Generated Image Descriptions for Blind Users on a Social Network Service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW*

âĂŹ17). Association for Computing Machinery, New York, NY, USA, 1180âĂŞ1192. DOI: **http://dx.doi.org/10.1145/2998181.2998364**

[72] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[73] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.

[74] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. 629–634.

[75] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. 4847–4853.

[76] Ran Zmigrod, Sebastian J. Mielke, Hanna M. Wallach, and Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. 1651–1661.