2017

Gender as a Variable in Natural-Language Processing: Ethical Considerations

Brian N. Larson
Texas A&M University School of Law, blarson@law.tamu.edu

EACL 2017

Ethics in Natural Language Processing

Proceedings of the First ACL Workshop

April 4th, 2017
Valencia, Spain

Sponsors:

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-945626-47-0

Introduction

Welcome to the first ACL Workshop on Ethics in Natural Language Processing! We are pleased to have participants from a variety of backgrounds and perspectives: social science, computational linguistics, and philosophy; academia, industry, and government.

The workshop consists of invited talks, contributed discussion papers, posters, demos, and a panel discussion. Invited speakers include **Graeme Hirst**, a Professor in NLP at the University of Toronto, who works on lexical semantics, pragmatics, and text classification, with applications to intelligent text understanding for disabled users; **Quirine Eijkman**, a Senior Researcher at Leiden University, who leads work on security governance, the sociology of law, and human right; **Jason Baldridge**, a co-founder and Chief Scientist of People Pattern, who specializes in computational models of discourse as well as the interaction between machine learning and human bias; and **Joanna Bryson**, a Reader in artificial intelligence and natural intelligence at the University of Bath, who works on action selection, systems AI, transparency of AI, political polarization, income inequality, and ethics in AI.

We received paper submissions that span a wide range of topics, addressing issues related to overgeneralization, dual use, privacy protection, bias in NLP models, underrepresentation, fairness, and more. Their authors share insights about the intersection of NLP and ethics in academic work, industrial work, and clinical work. Common themes include the role of tasks, datasets, annotations, training populations, and modelling. We selected 4 papers for oral presentation, 8 for poster presentation, and one for demo presentation, and have paired each oral presentation with a discussant outside of the authors' areas of expertise to help contextualize the work in a broader perspective. All papers additionally provide the basis for panel and participant discussion.

We hope this workshop will help to define and raise awareness of ethical considerations in NLP throughout the community, and will kickstart a recurring theme to consider in future NLP conferences. We would like to thank all authors, speakers, panelists, and discussants for their thoughtful contributions. We are also grateful for our sponsors (Bloomberg, Google, and HITS), who have helped making the workshop in this form possible.

The Organizers
Margaret, Dirk, Shannon, Emily, Hanna, Michael

Organizers:

Dirk Hovy, University of Copenhagen (Denmark)
Shannon Spruit, Delft University of Technology (Netherlands)
Margaret Mitchell, Google Research & Machine Intelligence (USA)
Emily M. Bender, University of Washington (USA)
Michael Strube, Heidelberg Institute for Theoretical Studies (Germany)
Hanna Wallach, Microsoft Research, UMass Amherst (USA)

Program Committee:

Gilles Adda
Nikolaos Aletras
Mark Alfano
Jacob Andreas
Isabelle Augenstein
Tim Baldwin
Miguel Ballesteros
David Bamman
Mohit Bansal
Solon Barocas
Daniel Bauer
Eric Bell
Steven Bethard
Rahul Bhagat
Chris Biemann
Yonatan Bisk
Michael Bloodgood
Matko Bosnjak
Chris Brockett
Miles Brundage
Joana J. Bryson
Ryan Calo
Marine Carpuat
Yejin Choi
Munmun De Choudhury
Grzegorz Chrupala
Ann Clifton
Kevin B. Cohen
Shay B. Cohen
Court Corley
Ryan Cotterell
Aron Culotta
Walter Daelemans
Dipanjan Das
Hal Daumé III
Steve DeNeefe
Francien Dechesne
Leon Derczynski
Aliya Deri
Mona Diab
Fernando Diaz
Benjamin Van Durme
Jacob Eisenstein
Jason Eisner
Desmond Elliott
Micha Elsner
Katrin Erk
Raquel Fernandez
Laura Fichtner
Karën Fort
Victoria Fossum
Lily Frank
Sorelle Friedler
Annemarie Friedrich
Juri Ganitkevich
Spandana Gella
Kevin Gimpel
Joao Graca
Yvette Graham
Keith Hall
Oul Han
Graeme Hirst
Nathan Hodas
Kristy Hollingshead
Ed Hovy
Georgy Ishmaev
Jing Jiang
Anna Jobin
Anders Johannsen
David Jurgens
Brian Keegan
Roman Klinger
Ekaterina Kochmar
Philipp Koehn
Zornitsa Kozareva
Jayant Krishnamurthy
Jonathan K. Kummerfeld
Vasileios Lampos
Angeliki Lazaridou
Alessandro Lenci
Nikola Ljubesic
Adam Lopez
L. Alfonso Urena Lopez
Teresa Lynn
Nitin Madnani
Gideon Mann
Daniel Marcu
Jonathan May
Kathy McKeown
Paola Merlo
David Mimno
Shachar Mirkin
Alessandro Moschitti
Jason Naradowsky
Roberto Navigli
Arvind Neelakantan
Ani Nenkova
Dong Nguyen
Brendan O'Connor
Diarmuid O'Seaghdha
Miles Osborne
Jahna Otterbacher
Sebastian Padó
Alexis Palmer
Martha Palmer
Michael Paul
Ellie Pavlick
Emily Pitler
Barbara Plank
Thierry Poibeau
Chris Potts
Vinod Prabhakaran
Daniel Preotiuc
Nikolaus Pöchhacker
Will Radford
Siva Reddy
Luis Reyes-Galindo
Sebastian Riedel
Ellen Riloff
Brian Roark
Molly Roberts
Tim Rocktäschel
Frank Rudzicz
Alexander M. Rush
Derek Ruths
Asad Sayeed
David Schlangen
Natalie Schluter
H. Andrew Schwartz
Hinrich Schütze
Djamé Seddah
Dan Simonson
Sameer Singh
Vivek Srikumar
Sanja Stajner
Pontus Stenetorp
Brandon Stewart
Veselin Stoyanov
Anders Søgaard
Ivan Titov
Sara Tonelli
Oren Tsur
Yulia Tsvetkov
Lyle Ungar
Suresh Venkatasubramanian
Yannick Versley
Aline Villavicencio
Andreas Vlachos
Rob Voigt
Svitlana Volkova
Martijn Warnier
Zeerak Waseem
Bonnie Webber
Joern Wuebker
François Yvon
Luke Zettlemoyer
Janneke van der Zwaan

Invited Speakers:

Graeme Hirst, University of Toronto (Canada)
Quirine Eijkman, Leiden University (Netherlands)
Jason Baldridge, People Pattern (USA)
Joanna Bryson, University of Bath (UK)

Gender as a Variable in Natural-Language Processing: Ethical Considerations

Brian N. Larson
Georgia Institute of Technology
686 Cherry St. MC 0165
Atlanta, GA 30363 USA
blarson@gatech.edu

Abstract

Researchers and practitioners in natural-language processing (NLP) and related fields should attend to ethical principles in study design, ascription of categories/variables to study participants, and reporting of findings or results. This paper discusses theoretical and ethical frameworks for using *gender* as a variable in NLP studies and proposes four guidelines for researchers and practitioners. The principles outlined here should guide practitioners, researchers, and peer reviewers, and they may be applicable to other social categories, such as race, applied to human beings connected to NLP research.

1 Introduction

Bamman et al. (2014) challenged simplistic notions of a gender binary and the common quest in natural-language processing (NLP) studies merely to predict gender based on text, making the following observation:

> If we start with the assumption that 'female' and 'male' are the relevant categories, then our analyses are incapable of revealing violations of this assumption.... [W]hen we turn to a descriptive account of the interaction between language and gender, this analysis becomes a house of mirrors, which by design can only find evidence to support the underlying assumption of a binary gender opposition (p. 148).

Gender is a common variable in NLP studies. For example, a search of the ACL Anthology (`aclanthology.info`) for the keyword "gender" in the title field revealed seven papers in 2016 alone that made use of personal (as opposed to grammatical) gender as a central variable. Many others used gender as a variable without referring to gender in their titles. It is not uncommon, however, for studies regarding gender to be reported without any explanation of how gender labels were ascribed to authors or their texts.

This paper argues that using gender as a variable in NLP is an *ethical issue*. Researchers and practitioners in NLP who unreflectively apply gender category labels to texts and their authors may violate ethical principles that govern the use of human participants or "subjects" in research (Belmont Report, 1979; Common Rule, 2009). By failing to explain in study reports what theory of gender they are using and how they assigned gender categories, they may also run afoul of other ethical frameworks that demand transparency and accountability from researchers (Breuch et al., 2002; FAT-ML, nd; MacNealy, 1998).

This paper discusses theoretical and ethical frameworks for using *gender* as a variable in NLP studies. The principles outlined here should guide researchers and peer reviewers, and they may be applicable to other social categories, such as race, applied to human beings connected to NLP research. Note that this paper does not purport to select the best theory of gender or method of ascribing gender categories for NLP. Rather, it urges a continual process of thoughtfulness and debate regarding these issues, both within each study and among the authors and readers of study reports.

In summary, researchers and practitioners should (1) formulate research questions making explicit theories of what "gender" is; (2) avoid using gender as a variable unless it is necessary to answer research questions; (3) make explicit methods for assigning gender categories to participants and linguistic artifacts; and (4) respect the difficulties of respondents when asking them

Proceedings of the First Workshop on Ethics in Natural Language Processing, pages 30–40,
Valencia, Spain, April 4th, 2017.

to self-identify for gender.

Section 2 considers theoretical foundations for *gender* as a research construct and rationales for studying it. Section 3 proposes ethical frameworks for academic researchers and for practitioners. Section 4 examines several studies in NLP that are representative of the range of studies using gender as a variable. Section 5 concludes with recommendations for best practices in designing, reporting, and peer-reviewing NLP studies using gender as a variable.

2 Gender and rationales for its study

2.1 Three views of gender

There are many views of how gender functions as a social construct. This section presents just three: the common or *folk* view of gender, a *performative* view of gender, and one social psychological view of gender. None of these views can be seen as correct for all contexts and applications. The view that is appropriate for a given project will depend on the research questions posed and the goals of the project.

A *folk* belief, as the term is used here, refers to the *doxa* or beliefs of the many that may or may not be supported by systematic inquiry—common beliefs distinguished from scientific knowledge or philosophical theories (Plato, 2005). In the folk conception, the "heteronormative gender binary" (Larson, 2016, p. 365) conflates sex, the chromosomal and biological characteristics of people, with gender, their outward appearances and behaviors. The salience of these categories and their binary nature are taken as obvious and natural. Consequently, the options available on a survey for the question "Gender?" are frequently "male" or "female" (sex categories) rather than "masculine" or "feminine" (gender categories). There is a growing understanding in contemporary western culture, however, that some individuals either do not fall easily into the binary or exhibit gender characteristics inconsistent with the biological sex ascribed to them at birth—these persons are sometimes referred to as being "transgender," while those whose sex and gender are congruent are "cisgender" (DeFrancisco et al., 2014). Various communities of persons who are not cisgender have other names they prefer to use for themselves, including "gender non-conforming," "non-binary," and "genderqueer" (GLAAD, nd b). According to one academic report, there are 1.4 million transgender people in the United States alone, and for these persons, the language used to characterize them can function as respectful on the one hand or offensive and defamatory on the other (GLAAD, nd a). Note that the gender labels that transgender persons ascribe to themselves do not include "other." The folk view of gender might be an appropriate frame for the NLP researcher seeking to explore study participants' use of language in relation to their own conceptions of their genders.

Another view of gender sees it as *performative*. So, according to DeFrancisco et al. (2014, p. 3) gender consists in "the behaviors and appearances society dictates a body of a particular sex should perform," structuring "people's understanding of themselves and each other." According to Larson (2016), an actor's gender knowledge comprises components of the actor's cognitive environment—beliefs about behaviors the actor expects to have a particular effect or effects on another based on knowledge about a typified situation in the actor's cognitive environment. Among these behaviors is language. Butler (1993) characterized gender as a form of performativity arising in "an unexamined framework of normative heterosexuality" (p. 97). According to all these theories, gender performativity is not merely *performance*, but rather performances that respond to, or are constrained by, norms or conventions and simultaneously reinforce them. This approach to gender could be useful, for example, in a study exploring the ways that language might be used to resist folk views of gender, especially in a context like transgender communities, where resistance to gender *doxa* is essential to building identity. Similarly, it could be useful in studying cases where persons of one gender attempt to appropriate conventional communicative practices of another gender without adopting a transgender identity. Bamman et al. (2014) made specific reference to this family of theories in their study of Twitter users.

A third approach to thinking about gender is to assume a gender binary, identify characteristics that cluster around the modes of the binary, and assess the gender of study participants based on their closeness of fit to these modes. This is exactly the approach of the Bem Sex Roles Inventory (Bem, 1974) and other instruments developed by social psychologists to assess gender. This approach allows the researcher to break gender down into constituent features. So, for example, the BSRI asso-

ciates self-reliance, independence, and athleticism with masculinity and loyalty, sympathy, and sensitivity with femininity (Blanchard-Fields et al., 1994). This approach might be useful, for example, for an NLP practitioner seeking to identify consumers exhibiting individual characteristics—like independence and athleticism—in order to market a particular product to those consumers without regard to their gender or sex. Such approaches may not be available to NLP researchers, though, as they require participants to fill out surveys.

These are only three of many possible approaches to gender, and as the examples suggest, they vary widely in the kinds of research questions they can help to answer.

2.2 Rationales for studying gender

Broadly speaking, NLP studies focused on gender stem from two sources: researchers and practitioners. Borrowing from concepts in the field of research with human participants, we can characterize *research* as "activity designed to test an hypothesis, permit conclusions to be drawn, and thereby to develop or contribute to generalizable knowledge" (Belmont Report, 1979). Practitioners, by contrast, are interested in providing solutions or "interventions that are designed solely to enhance the well-being of an individual. . . client"—in other words, the development of commercial applications. These two rationales can blend when academics disseminate research with the intention of attracting commercial interest and when practitioners disseminate study findings to the academic community with a goal, in part, of attracting attention to their commercial activities. Practitioners may also intend to develop applications that serve the needs of multiple clients, as when they seek to sell a technical solution to many players within an industry.

The practitioner may have more instrumental objectives, hoping, for example, for insights about consumer behavior applicable to an employer's or client's commercial goals. Practitioners engaged in such studies need not be concerned about the finer points of academic-researcher ethics. They should be conscious, however, of the social effects of their research when it is disseminated, covered in the news, etc. Even if their research is used only internally for their companies or clients, they may use variables in machine learning applications in such a way as to cause "algorithmic discrimination," where "an individual or group receives unfair treatment as a result of algorithmic decision-making" (Goodman, 2016). The ethical frameworks discussed in the next section provide reasons to avoid such discrimination.

3 Ethical frameworks

Academic researchers and commercial practitioners may draw their ethical principles from different ethical frameworks, but they have similar ethical obligations for ascribing category labels to persons and for using and reporting the research resulting from them.

In the United States, academic researchers are generally guided by principles articulated in the Belmont Report (1979), which calls on researchers to observe three principles:

- *Respect for persons* represents the right of a human taking part or being observed in research (sometimes called a "subject" or "participant") to make an informed decision about whether to take part and for a researcher "to give weight to autonomous persons' considered opinions and choices."

- *Beneficence* requires that the research first do no harm to participants and second "maximize possible benefits and minimize possible harms."

- *Justice* demands that the costs and benefits of research be distributed fairly, so that one group does not endure the costs of research while another enjoys its benefits.

Under regulations of the U.S. Department of Health and Human Services known as the Common Rule, "all research involving human subjects conducted, supported or otherwise subject to regulation by any federal department or agency" must be subjected to review by an institutional review board or IRB (Common Rule, 2009). As a practical matter, most research universities in the United States require that all research involving human participants be subject to IRB review. The Common Rule embodies many of the principles of the Belmont Report and of the Declaration of Helsinki (World Medical Association, 1964).

Other authorities argue that academic researchers have ethical responsibilities regarding their research, even if it does not involve human

participants. In that context, internal and external validity (or validity and reliability) of research findings are ethical concerns (Breuch et al., 2002; MacNealy, 1998). Not being explicit about what the researcher means by the research construct *gender* raises a problem for readers of research reports, as they cannot evaluate a researcher's claims without knowing in principle what the researcher means by her central terms. Not being explicit about the ascription of the category *gender* as a variable to participants or communication artifacts that they create brings into question internal and external validity of research findings, because it makes it difficult or impossible for other scholars to reproduce, test, or extend study findings. In short, doing good science is an ethical obligation of good scientists.

Practitioners are bound by ethical frameworks that are applicable to all persons generally. In the West, these may be drawn from normative frameworks that determine circumstances under which one can be called ethical: "virtue ethics"—having ethical thoughts and an ethical character (Hursthouse and Pettigrove, 2016); "deontological" ethics—conforming to rules, laws, and other statements of ethical duty (Alexander and Moore, 2016); and "consequentialism"—engaging in action that causes more good than harm (Sinnott-Armstrong, 2015). Other western and non-western ethical systems may prioritize other values (Hennig, 2010). Deontological ethics is drawn from sets of rules, such as religious texts, industry codes of ethics, and laws. Deontological theorists derive such rules from theoretical procedures, such as Kant's categorical imperative, where "all those possibly affected" can "will a just maxim as a general rule"; Rawl's "veil of ignorance," in which participants cannot know what role they will play in the society for which they posit rules; or Habermas's discourse ethics, rules resulting from a "noncoercive rational discourse among free and equal participants" (Habermas, 1995, p. 117). In a sense the Belmont Report provides a set of rules for deontological evaluation.

Consequentialist ethical systems like utilitarianism evaluate actions not by their means but their ends. They are thus consistent with the Belmont Report edict that research's benefits should outweigh its costs. But neither the Belmont Report nor other ethical systems typically permit actors to ignore the means they use to pursue their ends.

Some researchers/practitioners have argued for *fairness, accountability,* and *transparency* as ethical principles in applications of machine learning, a technology commonly used in NLP. Consider, for example, Hardt (2014) and Wallach (2014), and the group of researchers and practitioners behind FAT-ML (FAT-ML, nd). In this literature, it is not always clear what these three terms are meant to represent. So, for example, *fairness* appears to be a social metric similar to the Belmont Report's *beneficence* and *justice*. Wallach refers to it almost strictly in the phrase "bias, fairness, and inclusion." This seems concerned with fairness in the distributive sense of the Belmont Report's *justice* rather than the aggregate sense of consequentialist ethical systems. Wallach's uses of *transparency* and *accountability* echo the ethical principles for researchers suggested by Breuch et al. (2002) and MacNealy (1998). She appears to view them as principles to which researchers and practitioners should aspire.

FAT-ML could be operationalized as an ethical framework this way: NLP studies would expose their theoretical commitments, describe their research constructs (including *gender*), and explain their methods (including their ascription of gender categories). The resulting *transparency* permits *accountability* to peer reviewers and other researchers and practitioners, who may assess a given study against principles intended to result in valid and reliable scientific findings, principles designed to ensure respect for persons, justice, beneficence, and other evolving ethical principles under the rubric of *fairness*. Identification of the applicable rules awaits the rational non-coercive discourse of which the First Workshop on Ethics in NLP is an early and important example.

4 Applying frameworks to previous studies

This section considers how previously published and disseminated studies satisfy the ethical frameworks noted above and whether those frameworks may challenge the studies. Note that consideration of these particular studies is not meant to suggest that they are *ethically flawed*; they have been selected because they are recent studies or high-quality studies that have been widely cited. Generally, the studies discussed in this section included very careful descriptions of their methods of data collection and analysis. However, though

each purported to tell us something about gender, hardly any defined what they meant by "gender" or "sex," many did not indicate how they ascribed the gender categories to their participants or artifacts, and some that did explain the ascription of gender categories left room for concerns.

A great many studies have explored gender differences in human communication. An early and widely cited study is Koppel et al. (2002), where the researchers used machine learning to predict the gender of authors of published texts in the British National Corpus (BNC). Koppel and colleagues noted that the works they selected from the BNC were labeled for author gender, but they did not indicate how that labeling was done.

Like Koppel et al., many study authors allow the ascription of the gender category to be the result of an opaque process—that is, they do not fully embrace the transparency and accountability principles identified above, making the validity of studies difficult to assess. For example, in a study of computer-mediated communication, Herring and Paolillo (2006) assigned gender to blog authors "by examining each blog qualitatively for indications of gender such as first names, nicknames, explicit gender statements... and gender-indexical language." The authors did not provide readers with examples of the process of assigning these labels—called "coding" here as it is frequently by qualitative researchers, and not to be confused with the computer programmer's notion of "coding" or writing code—a coding guide, which is the set of instructions that researchers use to assign category labels to persons or artifacts, or a statement about whether the researchers compared coding by two or more coders to assess inter-rater reliability (Potter and Levine-Donnerstein, 1999).

Rao et al. (2010) examined Twitter posts ("tweets") to predict the gender categories they had ascribed to the texts' authors. They identified 1,000 Twitter users and inferred their gender based upon a heuristic: "For gender, the seed set for the crawl came from initial sources including sororities, fraternities, and male and female hygiene products. This produced around 500 users in each class" (2010, p. 38). Of course, using linguistic performances (profiles and tweets) to assign gender to Twitter accounts and then using linguistic performances to predict the genders of those accounts is very like the "house of mirrors" that Bamman et al. (2014) warned of above.

The approach of Rao and colleagues and Herring and Paolillo also appears to put the researcher in the position of deciding what counts as male and female in the data. This raises questions of fairness with regard to participants who have been labeled according the researchers' expectations, or perhaps their biases, rather than autonomous decisions by the participants.

Other studies make their ascription of gender categories explicit but fail to cautiously approach such labels. For example, two early studies, Yan and Yan (2006) and Argamon et al. (2007), used machine learning to classify blogs by their authors' genders. They used blog profile account settings to ascribe gender categories. Burger et al. (2011) assigned gender to Twitter users by following links from Twitter accounts to users' blogs on blogging platforms that required users to indicate their genders. More recently, Rouhizadeh et al. (2016) studied Facebook users from the period 2009–2011 based on their self-identified genders (but these data were gathered before Facebook's current gender options, see below), and Wang et al. (2016) looked at Weibo users, collecting self-identified gender data from their profiles.

None of the studies in the previous paragraph described how frequently account holders indicated their own genders, what gender options were possible, or how researchers accounted for account holders posing with genders other than their own. The answers to such questions would make the studies more transparent, helping readers to assess the their validity and fairness. For example, if many users of a site refused to disclose their genders, it is possible that the decision not to disclose might correlate with other characteristics that would make gender distinctions in the data more or less pronounced. The Belmont Report's concern about autonomy would best be addressed by understanding the options given to participants to represent themselves as gendered persons on these blogging platforms. If there were only two gender options—probably "male" and "female"—we might well wonder whether transgender persons may have refused to answer the question, or if forced to answer, how they chose which gender.

One study deserves special mention: Bamman et al. (2014) compared user names on Twitter profiles to U.S. Census data which showed a gender distribution for the 9,000 most commonly appearing first names. Though some names

were ambiguous—used for persons of different genders—in the census data, 95 percent of the users included in the study had a name that was "at least 85 percent associated with its majority gender" (p. 140). They then examined correlations between gender and language use. This approach might fall prey to criticisms regarding category ascription similar to those leveled at the studies above. Bamman et al., however, exhibited much more caution in the use of gender categories than any of the other studies cited here and engaged in cluster analyses that showed patterns of language use that crossed the gender-binary boundary. By describing the theory of gender they used and the method of ascribing the gender label, they made their study transparent and accountable. Whether it is fair is an assessment for their peers to make.

5 Guidelines for using gender as a variable

This section describes four guidelines for researchers and practitioners using gender as a variable in NLP studies that fall broadly under these admonitions: (1) formulate research questions making explicit theories of what "gender" is; (2) avoid using gender as a variable unless it is necessary to answer research questions; (3) make explicit methods for assigning gender categories to participants and linguistic artifacts; and (4) respect the difficulties of respondents when asking them to self-identify for gender. It also includes a recommendation for peer reviewers for conference-paper and research-article submissions. Note that this paper does not advocate for a particular theory of gender or method of ascribing gender categories to cover all NLP studies. Rather, it advocates for exposing decisions on these matters to aid in making studies more transparent, accountable, and fair. The decisions that practitioners and researchers make will be subject to debate among them, peer reviewers, and other practitioners and researchers.

5.1 Make theory of gender explicit

Researchers and practitioners should make explicit the theory of gender that undergirds their research questions. This step is essential to make studies accountable, transparent, and valid. For other researchers or practitioners to fully interpret a study and to interrogate, challenge, or reproduce it, they need to understand its theoretical grounds.

Ideally, a researcher would provide an extended discussion of the central variable in his or her study. For example, Larson (2016) offered a definition of "gender" used in the study along with a lengthy discussion of the concept. Both the discussion and analysis in Bamman et al. (2014) engaged with previous theoretical literature on gender and challenged the gender constructs used in previous NLP studies. But articles using gender as a variable need not go to this extent. The goal of making gender theory explicit can be achieved by quoting a definition of "gender" from earlier research and giving some evidence of actually having read some of the earlier research. In the alternative, the researcher may adopt a construct definition for gender; that is, the researcher may answer the question, "What does 'gender' measure?" Thus, researchers can either choose definitions of "gender" from existing theories or identify what they mean by "gender" by defining it themselves.

Practitioners may take a different view. Consider, for example, a practitioner working at a social media site that requires its users to self-identify in response to the question "gender." It is reasonable for this practitioner to use NLP tools to study the site's customers based on their responses to this question, seeking usage patterns, correlations, etc. But a challenge arises as social media platforms recognize nuances in gender identity. For example, in 2015 Facebook began allowing its users to indicate that their gender is "female," "male," or "custom," and the custom option is an open text box (Bell, 2015). A practitioner there using gender data will be compelled to use many labels or group them in a manner selected by the practitioner. Using all the labels presents difficulties for classifiers and for the practitioner attempting to explain results. Grouping labels requires the practitioner to *theorize* about how they should be grouped. This takes us back to the admonition that the researcher or practitioner should make explicit the theory of gender being used.

5.2 Avoid using gender unless necessary

This admonition is perhaps obvious: Given the efforts that this paper suggests should surround the selection, ascription, use, and reporting of gender categories in NLP studies, it would be foolish to use gender as a category unless it is necessary to achieve the researcher's objectives, because the effort is unlikely to be commensurate with the pay-

off. It is likely, though, that the casual use of gender as a routine demographic question in studies where gender is not a central concern will remain commonplace. It seems an easy question to ask, and once the data are collected, it seems easy to perform a cross-tabulation of findings or results based on the response to this question.

The reasons for avoiding the use of gender as a variable unless necessary are grounded in all the ethical principles discussed above. A failure to give careful consideration to the questions presented in this paper creates a variety of risks. Thus, researchers should resist the temptation to ask: "I wonder if the women responded differently than the men." The best way to resist this temptation is to resist asking the gender question in the first place, unless it is important to presenting findings or results.

A reviewer of this paper noted that following this recommendation might inadvertently discourage researchers and practitioners from checking the algorithmic bias of their systems. Indeed, it is thoroughly consistent with values described here for researchers and practitioners to engage in such checking. In that case, gender is a necessary category, but where such work is anticipated, the other recommendations of this Section 5 should be carefully followed from the outset.

5.3 Make category assignment explicit

Researchers and practitioners should make explicit the method(s) they use to ascribe gender categories to study participants or communication artifacts. This step is essential to make the researcher's or practitioner's studies accountable, transparent, and valid. Just as the study's *theory* of gender is an essential basis for interpreting the findings—for interrogating, challenging, and reproducing them—so are the *methods* of ascribing the variable of study. This category provides the largest number of specific recommendations. (In this section, the term "researcher" refers both to researchers as discussed above and to practitioners who choose to disseminate their studies into the research community.)

Researchers have several choices here. Outside of NLP, they have very commonly ascribed gender to study participants based on the researchers' own best-guess assessments: The researcher interacts with a participant and concludes that she is female or he is male. For small-scale studies, this approach will not likely go away; but the researcher should consider at the time of study design whether and how to do this. Researchers reporting findings should acknowledge if this is the approach they have taken.

A related approach makes sense where the researcher is studying how participants behave toward each other based on what they perceive others' genders to be. For example, if studying whether a teacher treats students differently based on student genders, the researcher may need to know what genders the *teacher* ascribes to students. The researcher should give thought to how to collect information about this category ascription from the teacher. The process could prove challenging if the researcher and teacher operate in an environment where students challenge traditional gender roles or where students outwardly identify as transgender.

But participant self-identification should be the gold standard for ascribing gender categories. Except in circumstances where one might not expect complete candor, one can count on participants to say what their own genders are. On the one hand, this approach to ascribing a gender label respects the autonomy of study participants, as it allows them to assert the gender with which they identify. On the other hand, it does not account for the fact that each study participant may have a different conception of gender, its meaning, its relation to sex, etc. For example, a 76-year-old woman who has lived in the United States her whole life may have a very different conception of what it means to be "female" or "feminine" than does a 20-year-old recent immigrant to Germany from Turkey. Despite this, each may be attempting to make sense of her identity as including a female or feminine gender.

In theory, the researcher could address the concerns regarding participant self-identification using a gender-role inventory. In fact, one study looking for gender differences in writing did exactly that, using the Bem Sex Role Inventory (BSRI) to assess author genders (Janssen and Murachver, 2004). The challenge with these approaches is that gender is a moving target. Sandra Bem introduced the BSRI in 1974 (Bem, 1974). It has since been criticized on a wide variety of grounds, but of importance here is the fact that it was based on gender role stereotypes from the time when it was created. Thus a meta-analysis by

Twenge (1997) of studies using the BSRI showed that the masculinity score of women taking the BSRI had increased steadily over 15 years, and men's masculinity scores showed a steady decrease in correlation over the same period. These developments make sense in the context of a gender roles inventory that is necessarily validated over a period of years after it is first developed, resulting in an outdated set of gender stereotypes being embodied in the test, stereotypes that are not confirmed later as gender roles change. This does not mean that these inventories have no value for some applications; rather, researchers using them should explain that they are using them, why they are using them, and what their limitations are.

Researchers should consider the following specific recommendations: First, if a study relies upon a gender-category ascription provided by someone else, as does Koppel et al. (2002), it should provide as much information as possible about how the category was ascribed and acknowledge the third-party category ascription as a limitation. This supports the goals of research validity, transparency, and accountability.

Second, if the researchers relied upon self-identified gender from a technology or social media platform, the study report should show that the researchers have reflected on the possibility that users of the platform have not identified their genders at all (where the platform does not require it), that users may intentionally misidentify their genders, that transgender users may be unable to identify themselves accurately (if the platform presents only a binary), or that they may have been insulted by the question (if the platform presents them with "male," "female," and "*other*," for example). All these reflections address questions of validity, transparency, and accountability. The final two, however, also implicate the autonomy and respect for persons the Belmont Report calls for.

Third, if researchers use a heuristic or qualitative coding scheme to assess an author's gender, it is critically important that readers be presented with a full description of the process. This includes providing a copy of the coding guide (the set of instructions that researchers use to assign category labels to persons or artifacts) and describing the process by which researchers checked their code ascriptions, including a measure of inter-rater reliability. Studies that use automated means to ascribe category labels should include copies of computer code used to make the ascriptions. This supports the goals of accountability, transparency, and validity.

Fourth, researchers who group gender labels collected from participant self-identification or use a heuristic to assign gender categories to participants or artifacts should consider "denaturalizing" the resulting category labels. This challenge is only likely to increase as sites like social media platforms recognize nuances in gender identity, as this section previously noted with regard to Facebook. For example, Larson (2016) asked participants to identify their own genders, giving them an open text box in which to do it. (See also the detailed discussion of methods in Larson (2017).) This permitted participants in the study to identify with any gender they chose, and respondents responded with eight different gender labels. Larson explained his grouping of the responses and chose to denaturalize the gender categories by not using their common names. The article thus grouped "F," "Fem," "Female," and "female" together with the category label *Gender F* and "Cis Male," "M," "Male," and "Masculine" with the label *Gender M*. Such disclosure or transparency supports the goals accountability and fairness.

The steps described here would have strengthened already fine studies like those cited in the previous section. Of course, they would not insulate them from criticism. For example, Larson (2016) collected self-identified gender information and denaturalized the gender categories as explained above, but the result was nevertheless a gender binary consistent with that prevalent in the folk-theory of gender. The transparency of the study methods, however, provides a basis for critique; had it simply reported findings based on "male" and "female" participants, the reader would not even be able to identify this basis for critique.

5.4 Respect persons

One final recommendation is applicable to researchers and to practitioners who may have a role in deciding how to collect self-identified gender labels from participants. Here, the practitioner or researcher should take pains to recognize differences and difficulties that respondents may face in ascribing gender to themselves or to others. For example, assuming that one is collecting demographic information with an online survey, one might offer respondents two options for gender:

"male" and "female." In contemporary western culture, however, it is not unusual to have respondents who do not easily identify with one gender or another or who actively refuse to be classed in a particular gender. Others are confidently transgender or intersex. Thus, two options may not be enough. However, the addition of an "other" might seem degrading or insulting to those who do not consider themselves to be "male" or "female." Another option might be "none of the above," but this again seems to function as an othering selection. There are so many ways that persons might choose to describe their genders that listing them might also be impractical, especially as the list itself might have reactive effects by drawing special attention to the gender question. Such effects might arise if the comprehensive nature of the list tips research participants off that gender is an object of study in the research. But even the "free-form" space discussed above presents difficulties for practitioners and researchers.

Grappling with this challenge, and in the case of researchers and practitioners disseminating their research, documenting that grappling, is the best way to ensure ethical outcomes.

5.5 Reviewers: Expect ethical practices

The way to ensure that researchers (and practitioners who disseminate their studies as research) conform to ethical principles is to make them accountable at the time of peer review. A challenge for researchers and peer reviewers alike, however, is space. A long paper for EACL is eight pages at the time of initial submission. A researcher may not feel able to report fully on a study's background, data, methods, findings, and significance in that space and still have space to explain steps take to ensure the use of the *gender* variable is ethical. At least two possible solutions come to mind.

First, researchers may make efforts to weave evidence of ethical study design and implementation into study write-ups. It may be possible with the addition of a small number of sentences to satisfy a peer reviewer that a researcher has followed the guidelines in this paper.

Second, a researcher could write up a supplemental description of the study addressing particularly these issues. The researcher could signal the presence of the supplemental description by noting its existence in the first draft submitted for peer review. If the paper is accepted, the supplemental description could be added to the paper before publication of the proceedings without adding excessive length to the paper. In the alternative, the supplemental description could be made available via a link to a web resource apart from the paper itself. ACL has provided for the submission of "supplementary material" at least at some of its conferences "to report preprocessing decisions, model parameters, and other details necessary for the replication of the experiments reported" (Association for Computational Linguistics, 2016). Other NLP conferences and technical reports should follow this lead. In any case, it may be helpful if the peer-review mechanisms for journals and conferences include a means for the researcher to attach the supplemental description, as its quality may influence the votes of some reviewers regarding the quality of the paper.

6 Conclusion

This paper represents only a starting point for treating the research variable *gender* in an ethical fashion. The guidelines for researchers and practitioners here are intended to be straightforward and simple. However, to engage in research or practice that measures up to high ethical standards, we should see ethics not as a checklist at the beginning or end of a study's design and execution. Rather, we should view it as a process where we continually ask whether our actions respect human beings, deliver benefits and not harms, distribute potential benefits and harms fairly, and explain our research so that others may interrogate, test, and challenge its validity.

Other sets of social labels, such as race, ethnicity, and religion, raise similar ethical concerns, and researchers studying data including those categories should also consider the advice presented here.

Acknowledgments

Thanks to the anonymous reviewers for helpful guidance. This project received support from the University of Minnesota's Writing Studies Department James I. Brown fellowship fund and its College of Liberal Arts Graduate Research Partnership Program.

References

Larry Alexander and Michael Moore. 2016. Deontological ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2016 edition.

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9).

Association for Computational Linguistics. 2016. Call for papers. the 55th Annual Meeting of the Association for Computational Linguistics | ACL Member Portal, November. Retrieved February 17, 2017 from https://www.aclweb.org/portal/content/55th-annual-meeting-association-computational-linguistics.

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

Karissa Bell. 2015. Facebook's new gender options let you choose anything you want. *Mashable*. Retrieved January 1, 2017, from http://mashable.com/2015/02/26/facebooks-new-custom-gender-options/.

Belmont Report. 1979. The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research. Retrieved January 24, 2017, from https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html.

Sandra L. Bem. 1974. The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 42(2):155–162.

Fredda Blanchard-Fields, Lynda Suhrer-Roussel, and Christopher Hertzog. 1994. A confirmatory factor analysis of the Bem Sex Role Inventory: Old questions, new answers. *Sex Roles*, 30(5-6):423–457.

Lee-Ann Kastman Breuch, Andrea M. Olson, and Andrea Frantz. 2002. Considering ethical issues in technical communication research. In Laura J. Gurak and Mary M. Lay, editors, *Research in Technical Communication*, pages 1–22. Praeger Publishers, Westport, CT.

John Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. Technical report, MITRE Corporation, Bedford, MA.

Judith Butler. 1993. *Bodies That Matter: On the Discursive Limits of "Sex"*. Routledge, New York.

Common Rule. 2009. Protection of Human Subjects. 45 Code of Federal Regulations Part 46. Retrieved February 13, 2017, from https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/.

Victoria Pruin DeFrancisco, Catherine Helen Palczewski, and Danielle Dick McGeough. 2014. *Gender in Communication: A Critical Introduction*. Sage Publications, Thousand Oaks, CA, 2nd edition.

FAT-ML. n.d. Fairness, accountability, and transparency in machine learning. Retrieved January 23, 2017, from http://www.fatml.org/.

GLAAD. n.d. a. Gay and Lesbian Alliance Against Defamation. GLAAD media reference guide. In focus: Covering the transgender community. Retrieved January 23, 2017, from http://www.glaad.org/reference/covering-trans-community.

GLAAD. n.d. b. Gay and Lesbian Alliance Against Defamation. Glossary of terms: Transgender. Retrieved January 23, 2017, from http://www.glaad.org/reference/transgender.

Bryce W. Goodman. 2016. A step towards accountable algorithms? algorithmic discrimination and the european union general data protection. In *29th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona. NIPS Foundation.

Jurgen Habermas. 1995. Reconciliation Through the Public use of Reason: Remarks on John Rawls's Political Liberalism. *The Journal of Philosophy*, 92(3):109–131.

Moritz Hardt. 2014. How big data is unfair: Understanding sources of unfairness in data driven decision making, September. Retrieved January 23, 2017, from https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de#.jr0yrklo0.

Alicia Hennig. 2010. Confucianism as corporate ethics strategy. *China Business and Research*, 2010(5):1–7.

Susan C. Herring and John C. Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459.

Rosalind Hursthouse and Glen Pettigrove. 2016. Virtue ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2016 edition.

Anna Janssen and Tamar Murachver. 2004. The relationship between gender and topic in gender-preferential language use. *Written Communication*, 21(4):344 –367.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401 –412.

Brian N. Larson. 2016. Gender/genre: The lack of gendered register in texts requiring genre knowledge. *Written Communication*, 33(4):360–384.

Brian N. Larson. 2017. First-year law students' court memoranda. Web download LDC2017T03, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, February. http://catalog.ldc.upenn.edu/LDC2017T03.

Mary Sue MacNealy. 1998. *Strategies for Empirical Research in Writing*. Longman, Boston.

Plato. 2005. Meno. In *Plato: Meno and Other Dialogues*, pages 99–143. Oxford University Press, Oxford.

W. James Potter and Deborah Levine-Donnerstein. 1999. Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, 27(3):258–284.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, pages 37–44, Toronto, ON, Canada, October. ACM.

Masoud Rouhizadeh, Lyle Ungar, Anneke Buffone, and Andrew H. Schwartz. 2016. Using syntactic and semantic context to explore psychodemographic differences in self-reference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2054–2059. Association for Computational Linguistics.

Walter Sinnott-Armstrong. 2015. Consequentialism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2015 edition.

Jean M. Twenge. 1997. Changes in masculine and feminine traits over time: A meta-analysis. *Sex Roles*, 36(5-6):305–325.

Hanna Wallach. 2014. Big data, machine learning, and the social sciences: Fairness, accountability, and transparency. Retrieved January 23, 2017, from https://medium.com/@hannawallach/big-data-machine-learning-and-the-social-sciences-927a8e20460d#.czusepxiz.

Yuan Wang, Yang Xiao, Chao Ma, and Zhen Xiao. 2016. Improving users' demographic prediction via the videos they talk about. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1359–1368. Association for Computational Linguistics.

World Medical Association. 1964. *Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects*. World Medical Association, Ferney-Voltaire, France, October 2013 edition.

Xiang Yan and Ling Yan. 2006. Gender classification of weblog authors. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 228–230, Palo Alto, CA, March. Association for the Advancement of Artificial Intelligence.