# Language (Technology) is Power: A Critical Survey of “Bias” in NLP

Su Lin Blodgett   Solon Barocas   Hal Daumé III   Hanna Wallach

# A wave of research on “bias”

- Recent vital work demonstrates that NLP systems exhibit “bias”

# A wave of research on “bias”

- Recent vital work demonstrates that NLP systems exhibit “bias”
- This work struggles to define “bias”

# A wave of research on “bias”

- Recent vital work demonstrates that NLP systems exhibit “bias”
- This work struggles to define “bias”
- As a field, we must be careful and precise about what we mean by “bias”

# We take careful stock of work on “bias” in NLP

- We survey 146 papers on “bias” in NLP, focusing on text
- For each paper, we categorize:
    - its stated motivations
    - its proposed techniques

# We take careful stock of work on “bias” in NLP

- We survey 146 papers on “bias” in NLP, focusing on text
- For each paper, we categorize:
    - its stated motivations
    - its proposed techniques
- Finding: papers often lack clear and consistent conceptualizations of “bias”

# We take careful stock of work on “bias” in NLP

- We survey 146 papers on “bias” in NLP, focusing on text
- For each paper, we categorize:
    - its stated motivations
    - its proposed techniques
- Finding: papers often lack clear and consistent conceptualizations of “bias”
- Afterward: a potential path forward

# The landscape of “bias” in NLP

| NLP task | Papers |
|---:|---:|
| Embeddings (type-level or contextualized) | 54 |
| Coreference resolution | 20 |
| Language modeling or dialogue generation | 17 |
| Hate-speech detection | 17 |
| Sentiment analysis | 15 |
| Machine translation | 8 |
| Tagging or parsing | 5 |
| Surveys, frameworks, and meta-analyses | 20 |
| Other | 22 |

# The landscape of “bias” in NLP

| NLP task | Papers |
|---:|---:|
| Embeddings (type-level or contextualized) | 54 |
| Coreference resolution | 20 |
| Language modeling or dialogue generation | 17 |
| Hate-speech detection | 17 |
| Sentiment analysis | 15 |
| Machine translation | 8 |
| Tagging or parsing | 5 |
| Surveys, frameworks, and meta-analyses | 20 |
| Other | 22 |

| | Papers |
|---:|---:|
| Category | Motivation |
| Allocational harms | 30 |
| Stereotyping | 50 |
| Other representational harms | 52 |
| Questionable correlations | 47 |
| Vague/unstated | 23 |
| Surveys, frameworks, and meta-analyses | 20 |

# The landscape of “bias” in NLP

| NLP task | Papers |
|---|---|
| Embeddings (type-level or contextualized) | 54 |
| Coreference resolution | 20 |
| Language modeling or dialogue generation | 17 |
| Hate-speech detection | 17 |
| Sentiment analysis | 15 |
| Machine translation | 8 |
| Tagging or parsing | 5 |
| Surveys, frameworks, and meta-analyses | 20 |
| Other | 22 |

| | Papers |
|---|---|
| **Category** | **Motivation** |
| Allocational harms | 30 |
| Stereotyping | 50 |
| Other representational harms | 52 |
| Questionable correlations | 47 |
| Vague/unstated | 23 |
| Surveys, frameworks, and meta-analyses | 20 |

# The landscape of "bias" in NLP

| NLP task | Papers |
|---|---|
| Embeddings (type-level or contextualized) | 54 |
| Coreference resolution | 20 |
| Language modeling or dialogue generation | 17 |
| Hate-speech detection | 17 |
| Sentiment analysis | 15 |
| Machine translation | 8 |
| Tagging or parsing | 5 |
| Surveys, frameworks, and meta-analyses | 20 |
| Other | 22 |

| | Papers |
|---|---|
| Category | Motivation |
| Allocational harms | 30 |
| Stereotyping | 50 |
| Other representational harms | 52 |
| Questionable correlations | 47 |
| Vague/unstated | 23 |
| Surveys, frameworks, and meta-analyses | 20 |

# The landscape of “bias” in NLP

| NLP task | Papers |
|---|---|
| Embeddings (type-level or contextualized) | 54 |
| Coreference resolution | 20 |
| Language modeling or dialogue generation | 17 |
| Hate-speech detection | 17 |
| Sentiment analysis | 15 |
| Machine translation | 8 |
| Tagging or parsing | 5 |
| Surveys, frameworks, and meta-analyses | 20 |
| Other | 22 |

| | Papers |
|---|---|
| Category | Motivation |
| Allocational harms | 30 |
| Stereotyping | 50 |
| Other representational harms | 52 |
| Questionable correlations | 47 |
| Vague/unstated | 23 |
| Surveys, frameworks, and meta-analyses | 20 |

# The landscape of “bias” in NLP

| NLP task | Papers |
|---:|---:|
| Embeddings (type-level or contextualized) | 54 |
| Coreference resolution | 20 |
| Language modeling or dialogue generation | 17 |
| Hate-speech detection | 17 |
| Sentiment analysis | 15 |
| Machine translation | 8 |
| Tagging or parsing | 5 |
| Surveys, frameworks, and meta-analyses | 20 |
| Other | 22 |

| | Papers |
|---:|---:|
| Category | Motivation |
| Allocational harms | 30 |
| Stereotyping | 50 |
| Other representational harms | 52 |
| Questionable correlations | 47 |
| Vague/unstated | 23 |
| Surveys, frameworks, and meta-analyses | 20 |

# The landscape of "bias" in NLP

| NLP task | Papers |
|---|---|
| Embeddings (type-level or contextualized) | 54 |
| Coreference resolution | 20 |
| Language modeling or dialogue generation | 17 |
| Hate-speech detection | 17 |
| Sentiment analysis | 15 |
| Machine translation | 8 |
| Tagging or parsing | 5 |
| Surveys, frameworks, and meta-analyses | 20 |
| Other | 22 |

| | Papers | |
|---|---|---|
| Category | Motivation | Technique |
| Allocational harms | 30 | 4 |
| Stereotyping | 50 | 58 |
| Other representational harms | 52 | 43 |
| Questionable correlations | 47 | 42 |
| Vague/unstated | 23 | 0 |
| Surveys, frameworks, and meta-analyses | 20 | 20 |

# Papers' motivations may be vague or imprecise

Biased embeddings can perpetuate systematic biases in society, discriminate against different groups of users, and promote social injustice.

Biased outputs or discriminatory behaviors might offend users or result in negative user experiences.

Biased algorithms risk taking problematic actions, affecting important downstream applications such as hiring.

# Papers' motivations may be vague or imprecise

Biased embeddings can perpetuate systematic biases in society, discriminate against different groups of users, and promote social injustice.

Biased outputs or discriminatory behaviors might offend users or result in negative user experiences.

Biased algorithms risk taking problematic actions, affecting important downstream applications such as hiring.

# Papers' motivations may be vague or imprecise

Biased embeddings can perpetuate systematic biases in society, discriminate against different groups of users, and promote social injustice.

Biased outputs or discriminatory behaviors might offend users or result in negative user experiences.

Biased algorithms risk taking problematic actions, affecting important downstream applications such as hiring.

# Papers sometimes give no normative reasoning

Models should not rely on demographic attributes expressed in the text to make predictions.

Models that rely on demographic attributes in the text yield higher error rates.

# Papers for the same task may conceptualize “bias” differently

- Gender/racial “bias” sometimes looks at text written ***about*** different groups, and text written ***by*** different groups

# Papers for the same task may conceptualize “bias” differently

- Gender/racial “bias” sometimes looks at text written ***about*** different groups, and text written ***by*** different groups
- Word embeddings papers have been motivated by hiring/résumé filtering, stereotyping, under-representation/under-recognition of women, and more
  - but generally all actually measure stereotyping

# Papers' motivations and techniques may not be well-matched

- Many papers are motivated by *allocational* harms—hiring, credit, etc.
  - ...but rarely ever measure them

# Papers' motivations and techniques may not be well-matched

- Many papers are motivated by *allocational* harms—hiring, credit, etc.
  - …but rarely ever measure them
- Therefore, we still know little about what allocational harms NLP systems give rise to

# A potential path forward

# Recommendation 1:
# Analyze language and social hierarchies together

- Social hierarchies: those resulting from unjust distributions of resources and power

# Recommendation 1:
# Analyze language and social hierarchies together

- Social hierarchies: those resulting from unjust distributions of resources and power
- A vast literature outside NLP shows us that language plays a role in maintaining social hierarchies

# Recommendation 1:
# Analyze language and social hierarchies together

- Social hierarchies: those resulting from unjust distributions of resources and power
- A vast literature outside NLP shows us that language plays a role in maintaining social hierarchies
  - language names social groups and transmits stereotypes [Maass 1999]
  - language choices shape narratives and discourses [Rosa 2019]
  - language ideologies enable linguistic discrimination and justify existing social hierarchies [Lippi-Green 2012, Rosa and Flores 2017, Craft 2020]

# Recommendation 1:
# Analyze language and social hierarchies together

- Social hierarchies: those resulting from unjust distributions of resources and power
- A vast literature outside NLP shows us that language plays a role in maintaining social hierarchies
- Ask: How are social hierarchies, language ideologies, and NLP systems coproduced? [Benjamin 2020]

# Recommendation 2:
# Articulate conceptualizations of “bias”

- Provide explicit statements of why system behaviors that are described as “bias” are harmful, in what ways, and to whom

# Recommendation 2:
# Articulate conceptualizations of “bias”

- Provide explicit statements of why system behaviors that are described as “bias” are harmful, in what ways, and to whom
- Be explicit about normative reasoning

# Recommendation 2:
# Articulate conceptualizations of “bias”

- Provide explicit statements of why system behaviors that are described as “bias” are harmful, in what ways, and to whom
- Be explicit about normative reasoning
- Explicit conceptualizations and normative reasoning:
  - ensure that motivations and quantitative techniques are well-matched
  - enable open community discussions of inherently normative questions
  - enable reflection on what researchers identify as “bias”

# Recommendation 3:
# Examine language use in practice

- Language is necessarily situated, and different social groups have different experiences

# Recommendation 3:
# Examine language use in practice

- Language is necessarily situated, and different social groups have different experiences
- Center work on the lived experiences of members of communities affected by NLP systems

# Recommendation 3:
# Examine language use in practice

- Language is necessarily situated, and different social groups have different experiences
- Center work on the lived experiences of members of communities affected by NLP systems
- Interrogate the power relations between technologists and affected communities

# Conclusion

- Papers often lack clear, consistent conceptualizations of “bias”
- Motivations and techniques may not always be well-matched
- Recommendations:
  - Reorient around relationships between social hierarchies, language ideologies, and technology
  - Articulate conceptualizations of “bias”, including normative reasoning
  - Examine language use in practice by centering communities