SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Computational approaches for uncovering implicit strategies in political discourse

Julia Mendelsohn

University of Michigan

juliame@umich.edu

Content Warning

I will be talking about material that may be offensive and upsetting to some audience members.

Democrats are the problem. They don’t care about crime and they want illegal immigrants, no matter how bad they may be, to pour into and infest our Country, like MS-13. They can’t win on their terrible policies, so they view them as potential voters!
~Former President Donald Trump, June 2018

Democrats are the problem. They don’t care about crime and they want illegal immigrants, no matter how bad they may be, to pour into and infest our Country, like MS-13. They can’t win on their terrible policies, so they view them as potential voters!
~Former President Donald Trump, June 2018

Good night to everyone but the massive amount of Soros-funded illegals who are trying to invade our border.
~ AZ State Sen Wendy Rogers, Sep 2021

Democrats are the problem. They don't care about crime and they want illegal immigrants, no matter how bad they may be, to pour into and infest our Country, like MS-13. They can't win on their terrible policies, so they view them as potential voters!

~Former President Donald Trump, June 2018

Anti-immigration

Good night to everyone but the massive amount of Soros-funded illegals who are trying to invade our border.

~ AZ State Sen Wendy Rogers, Sep 2021

Anti-immigration

> ***Democrats are the problem.*** *They don’t care about crime and they want illegal immigrants, no matter how bad they may be, to pour into and infest our Country, like MS-13.* ***They can’t win on their terrible policies, so they view them as potential voters!***
> ~Former President Donald Trump, June 2018

> *Good night to everyone but the massive amount of Soros-funded illegals who are trying to invade our border.*
> ~ AZ State Sen Wendy Rogers, Sep 2021

Democrats are the problem. They don’t care about ***crime and they want illegal immigrants, no matter how bad they may be****, to pour into and infest our Country, like* **MS-13***. They can’t win on their terrible policies, so they view them as potential voters!*
~Former President Donald Trump, June 2018

Good night to everyone but the massive amount of Soros-funded illegals who are trying to ***invade our border*****.**
~ AZ State Sen Wendy Rogers, Sep 2021

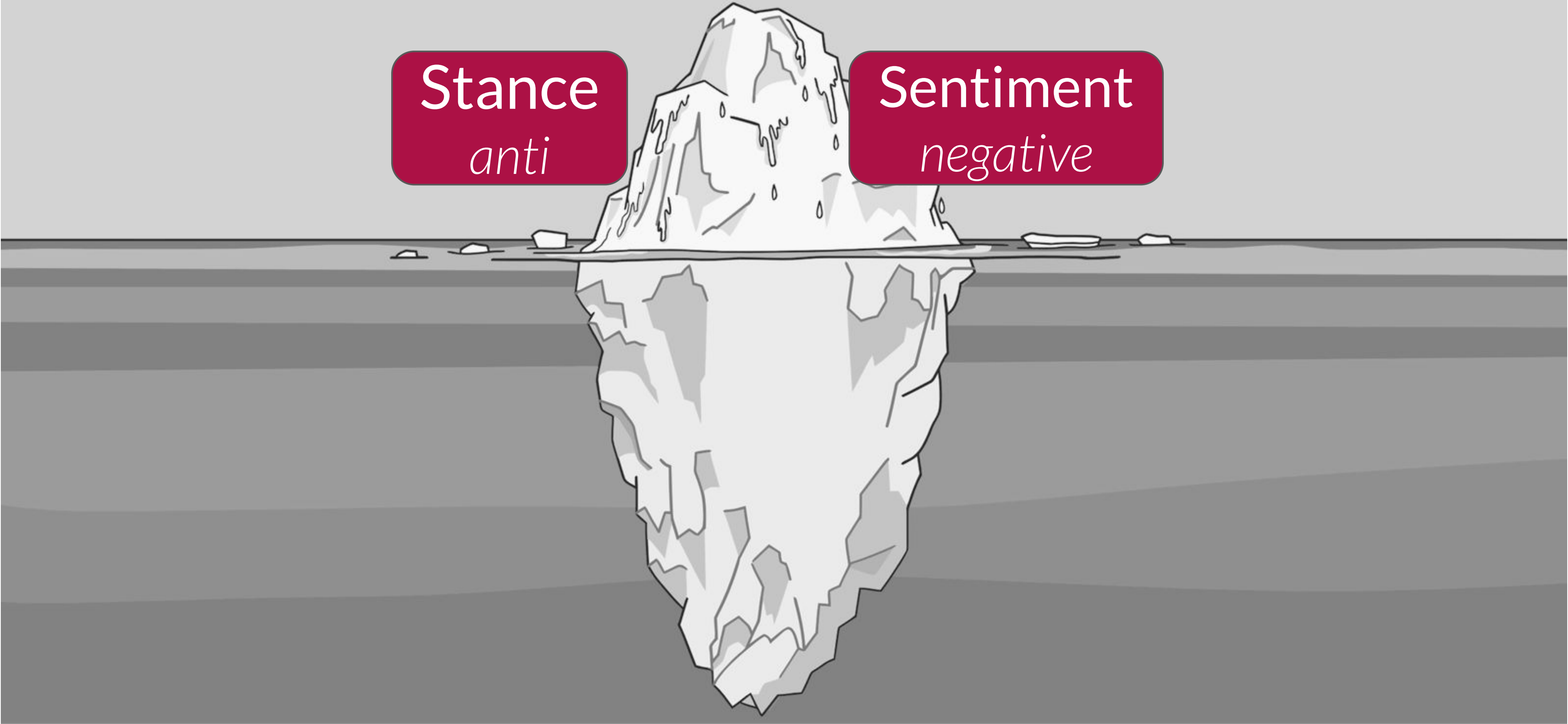
*Democrats are the problem. They don’t care about crime and they want illegal immigrants, no matter how bad they may be, **to pour into and infest our Country**, like MS-13. They can’t win on their terrible policies, so they view them as potential voters!*
~Former President Donald Trump, June 2018

*Good night to everyone but the **massive amount** of Soros-funded illegals who are trying to **invade our border**.*
~ AZ State Sen Wendy Rogers, Sep 2021

> *Democrats are the problem. They don’t care about crime and they want illegal immigrants, no matter how bad they may be, to pour into and infest our Country, like MS-13. They can’t win on their terrible policies, so they view them as potential voters!*
> ~Former President Donald Trump, June 2018

> *Good night to everyone but the massive amount of* ***Soros-funded illegals*** *who are trying to invade our border.*
> ~ AZ State Sen Wendy Rogers, Sep 2021

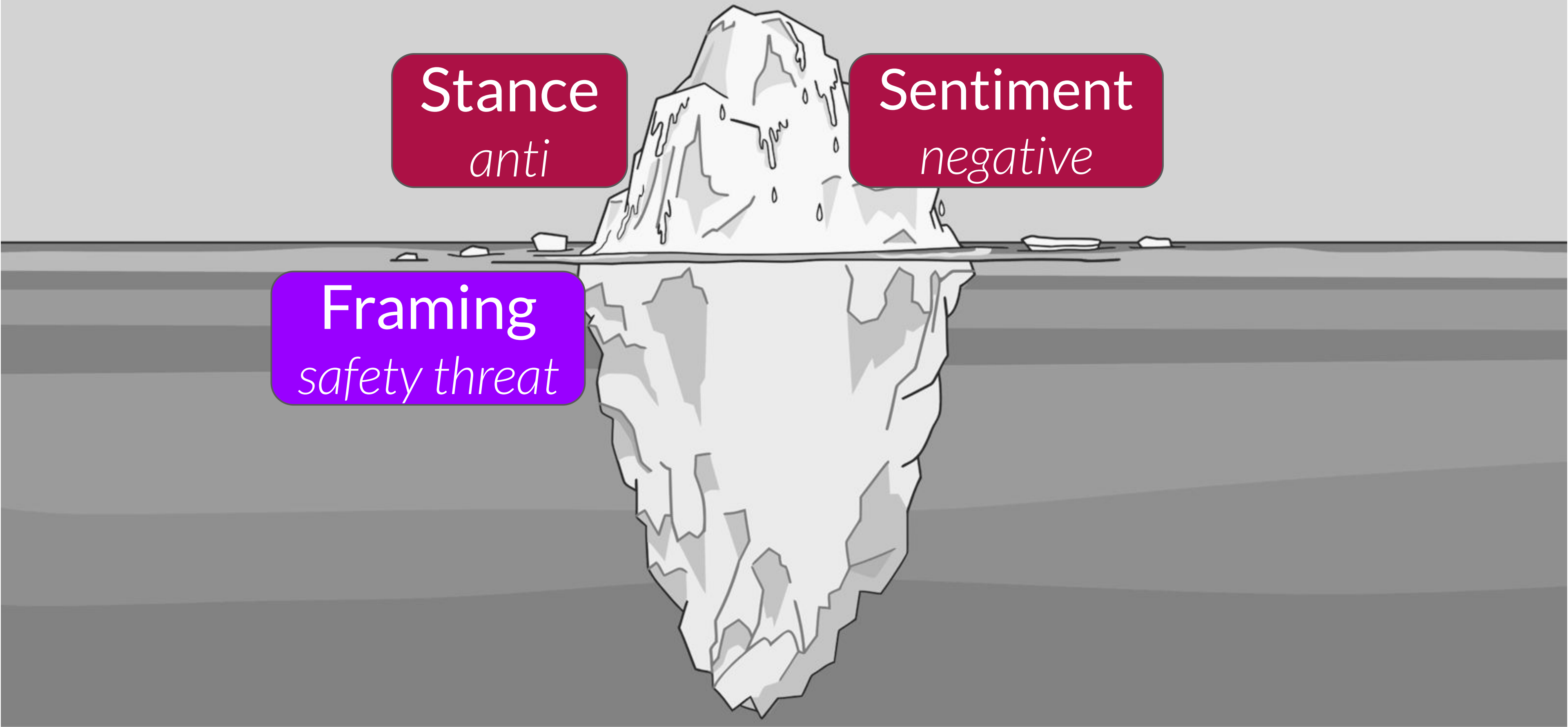
The Implicit Iceberg



Stance
anti

Sentiment
negative

The Implicit Iceberg

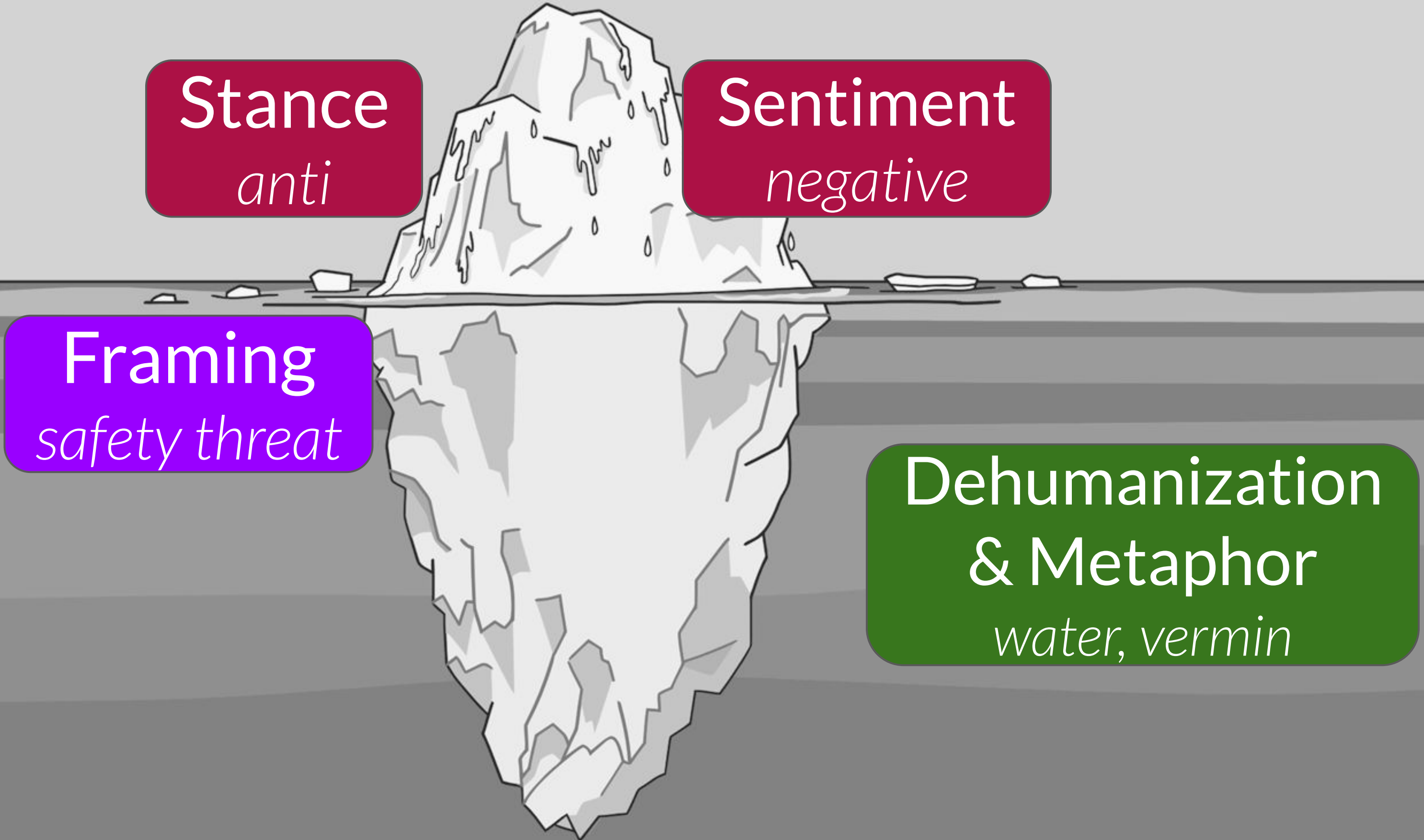


Stance
anti

Sentiment
negative

Framing
safety threat

The Implicit Iceberg



Stance

anti

Sentiment

negative

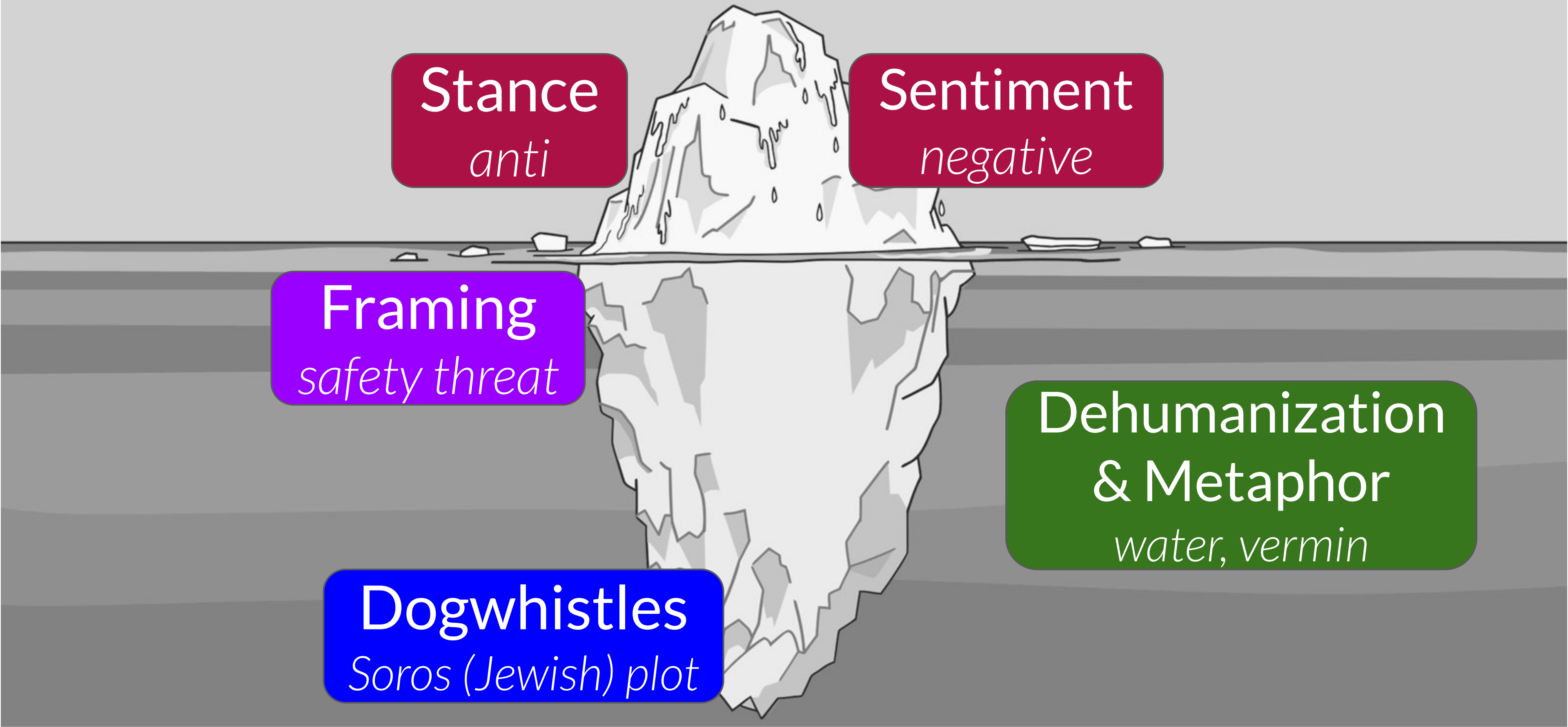
Framing

safety threat

Dehumanization & Metaphor

water, vermin

The Implicit Iceberg



Stance

anti

Sentiment

negative

Framing

safety threat

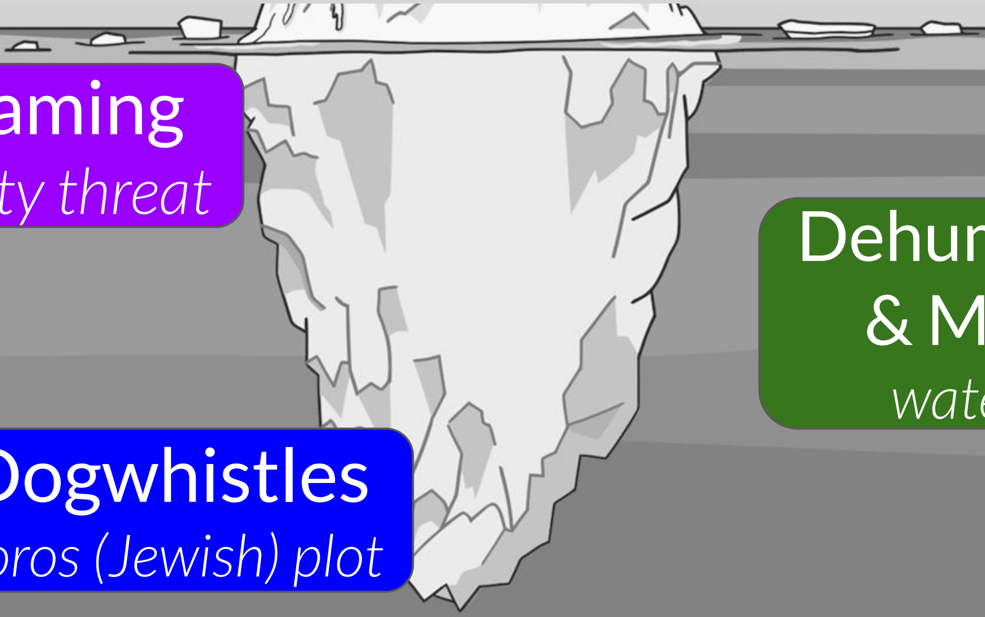
Dehumanization & Metaphor

water, vermin

Dogwhistles

Soros (Jewish) plot

These strategies shape ***how audiences understand*** political issues…



Framing

safety threat

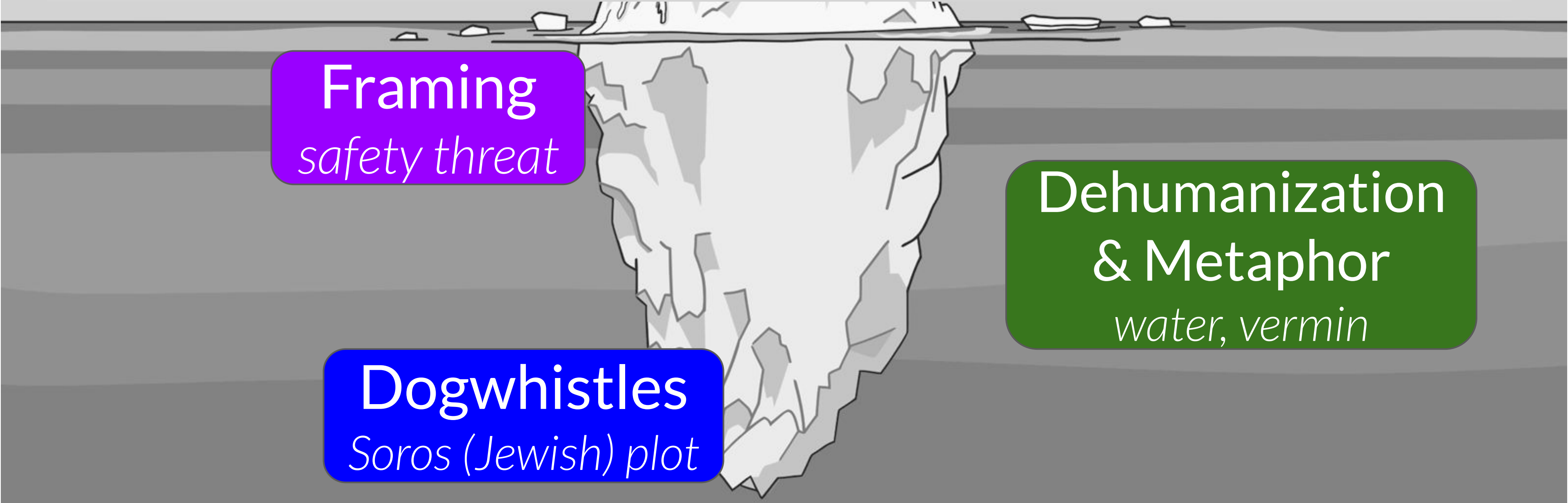
Dehumanization & Metaphor

water, vermin

Dogwhistles

Soros (Jewish) plot

…and are key elements of political communication



Framing

safety threat

Dogwhistles

Soros (Jewish) plot

Dehumanization & Metaphor

water, vermin

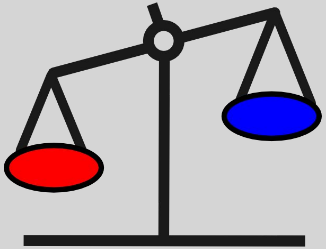
…and are key elements of political communication

…and are key elements of political communication

Campaigns
[Tilley, 2020]

…and are key elements of political communication

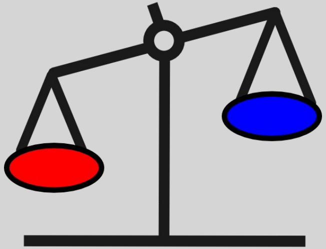
Campaigns
[Tilley, 2020]



Media Bias
[Esses et al., 2013]

…and are key elements of political communication

Campaigns
[Tilley, 2020]

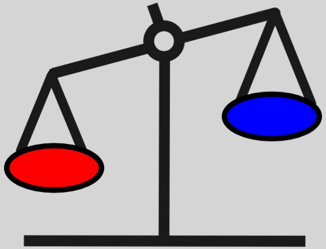


Media Bias
[Esses et al., 2013]

Misinformation
[Henderson & McCready, 2019]

…and are key elements of political communication

Campaigns
[Tilley, 2020]



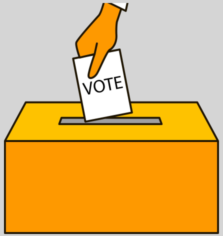
Media Bias
[Esses et al., 2013]

Misinformation
[Henderson & McCready, 2019]

Propaganda
[Landry et al., 2022]

…with far-reaching implications

…with far-reaching implications



Electoral Outcomes
[Haney López, 2014]

…with far-reaching implications

Policymaking
[Walgrave et al., 2018]

Electoral Outcomes
[Haney López, 2014]

…with far-reaching implications

Electoral Outcomes

[Haney López, 2014]

Policymaking

[Walgrave et al., 2018]

Public Opinion

[Jacoby, 2000; Chong & Druckman, 2007]

…with far-reaching implications

Policymaking
[Walgrave et al., 2018]

Trust
[Hopmann et al., 2015]

Electoral
Outcomes
[Haney López, 2014]

Public
Opinion
[Jacoby, 2000; Chong
& Druckman, 2007]

…with far-reaching implications

Policymaking
[Walgrave et al., 2018]

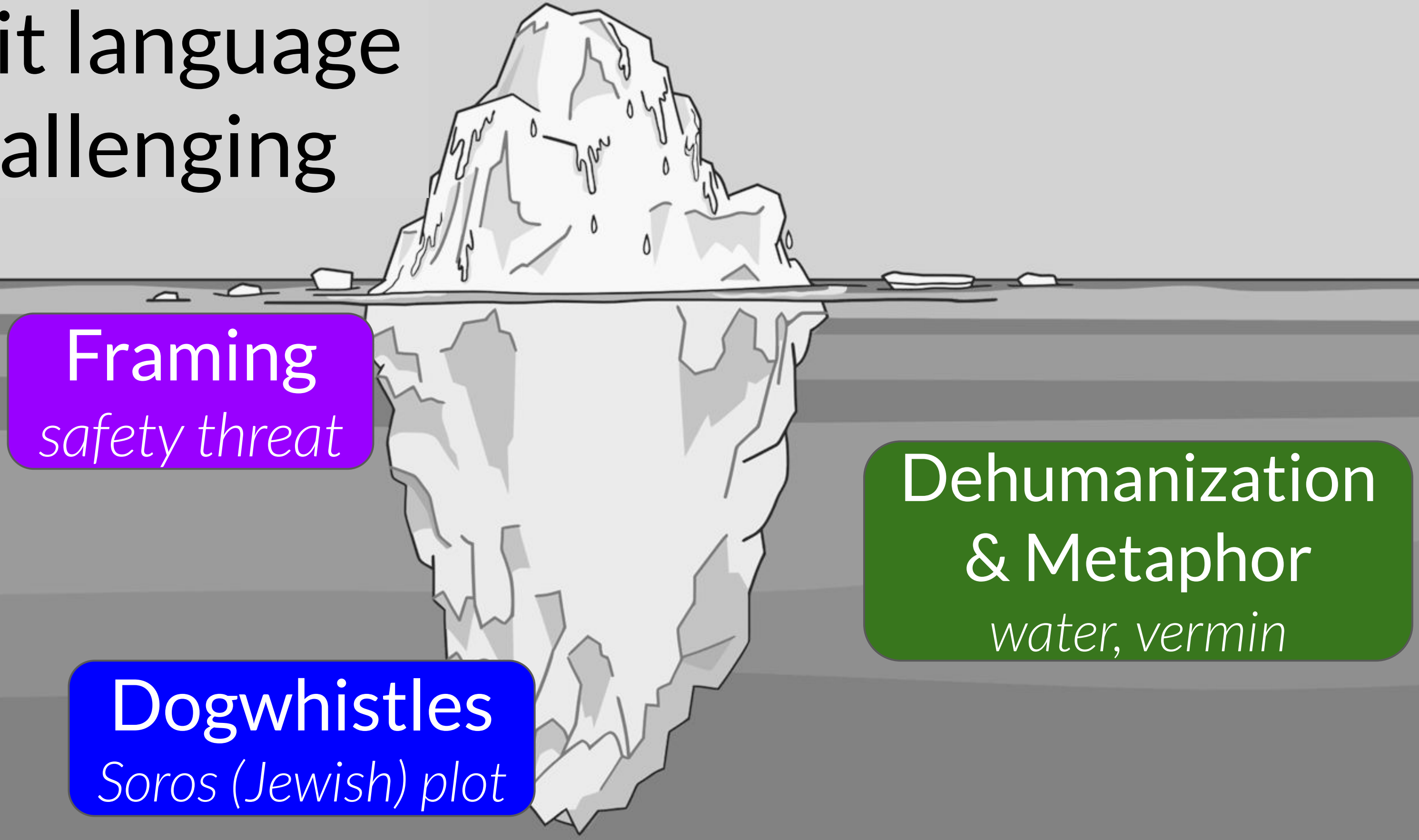
Trust
[Hopmann et al., 2015]

Electoral Outcomes
[Haney López, 2014]

Public Opinion
[Jacoby, 2000; Chong & Druckman, 2007]

Safety & Well-being
[Rai et al., 2017]

Uncovering implicit language is challenging



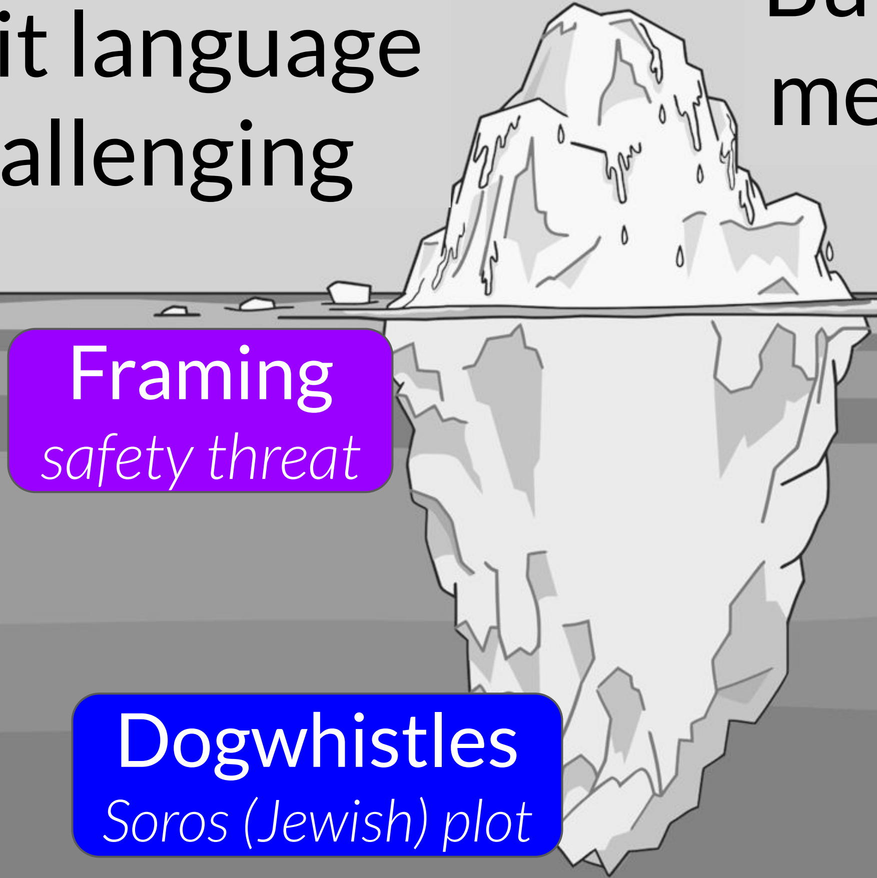
Framing
safety threat

Dehumanization & Metaphor
water, vermin

Dogwhistles
Soros (Jewish) plot

Uncovering implicit language is challenging

But computational methods can help!

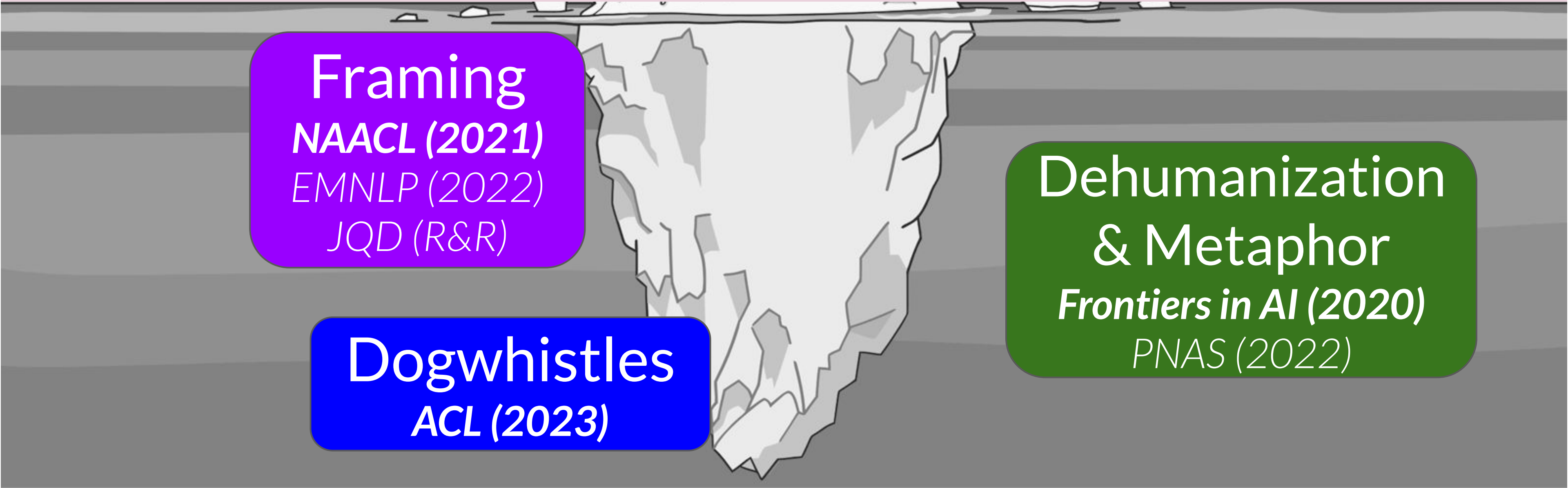


Framing
safety threat

Dogwhistles
Soros (Jewish) plot

Dehumanization & Metaphor
water, vermin

I develop **computational approaches** to study these strategies and their social, political & technological implications



Framing

NAACL (2021)

EMNLP (2022)

JQD (R&R)

Dogwhistles

ACL (2023)

Dehumanization & Metaphor

Frontiers in AI (2020)

PNAS (2022)

Roadmap

Overview

Roadmap

Overview

Framing

Dehumanization

Dogwhistles

Roadmap

- Overview
- Framing
- Dehumanization
- Dogwhistles
- Future Plans

Roadmap

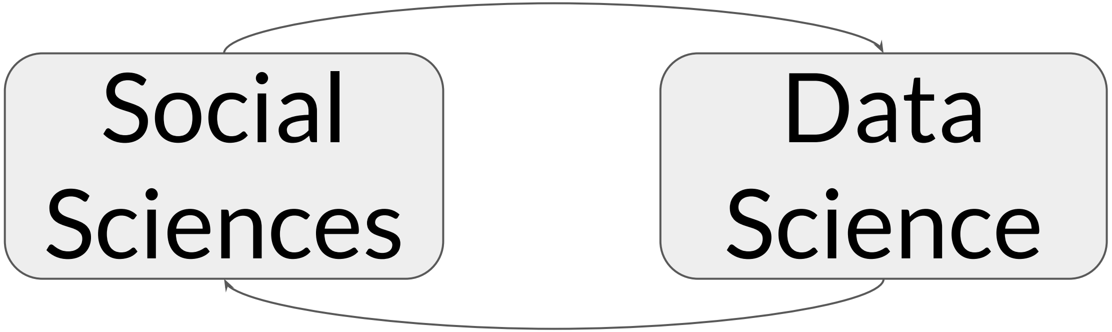
Overview

Framing

Dehumanization

Dogwhistles

Future Plans

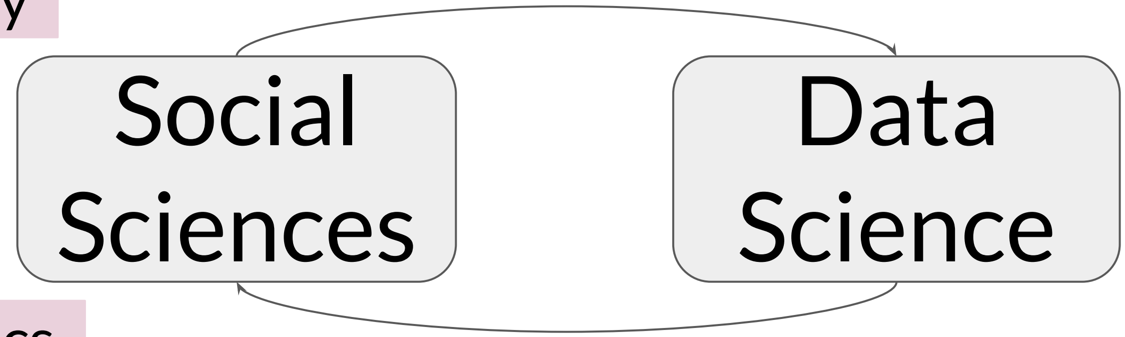


Social Sciences

Data Science

Political Science

Sociology



Linguistics

Communication

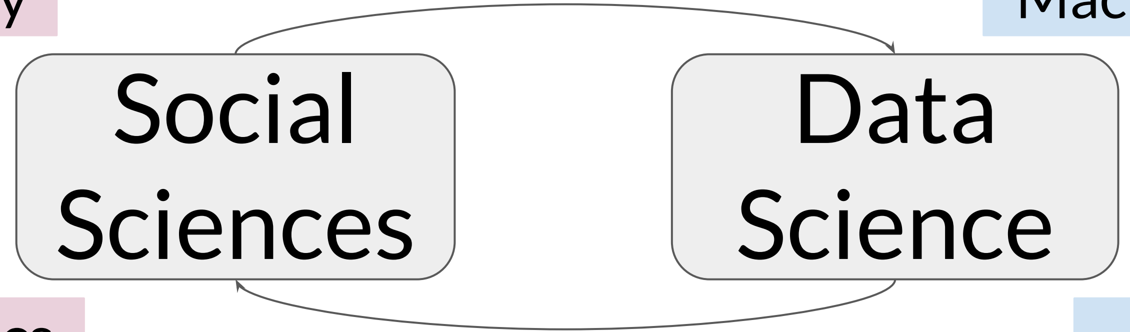
Psychology

Political Science

Sociology

Natural Language Processing (NLP)

Machine Learning



Social Sciences

Data Science

Linguistics

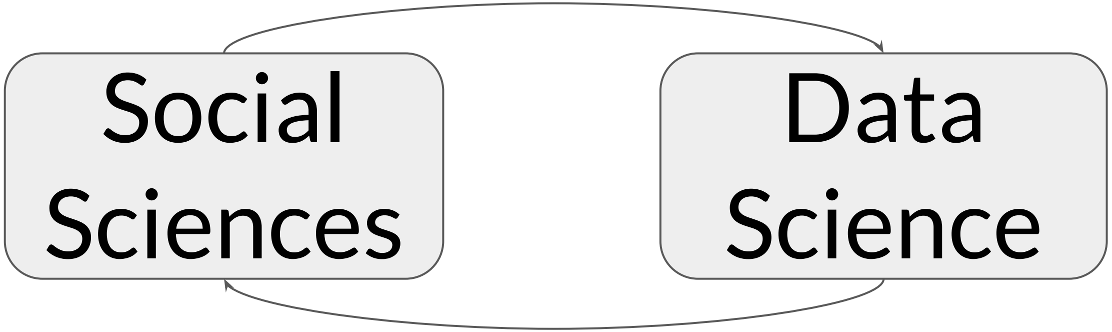
Communication

Psychology

Statistics

Causal Inference

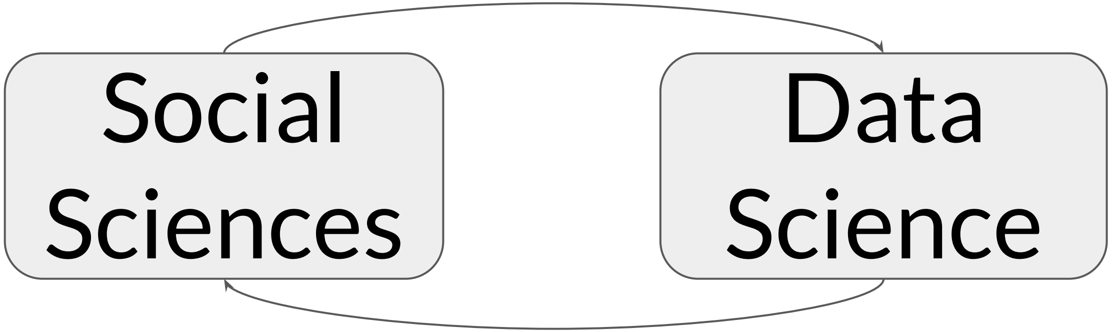
Network Analysis



Social Sciences

Data Science

Develop typologies and data resources

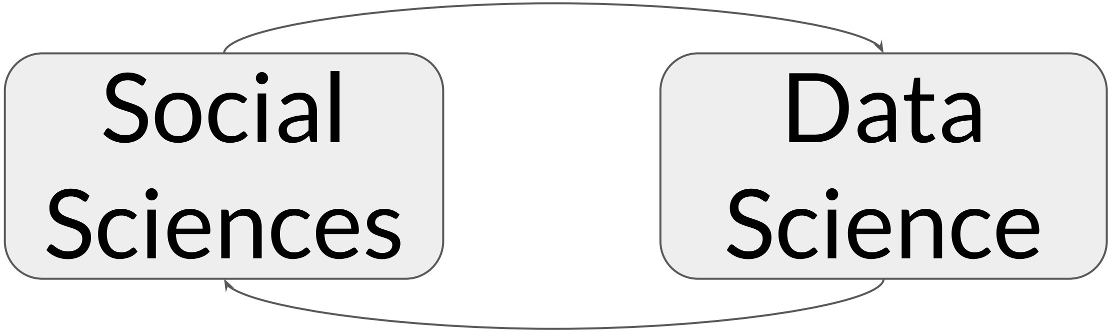


Social Sciences

Data Science

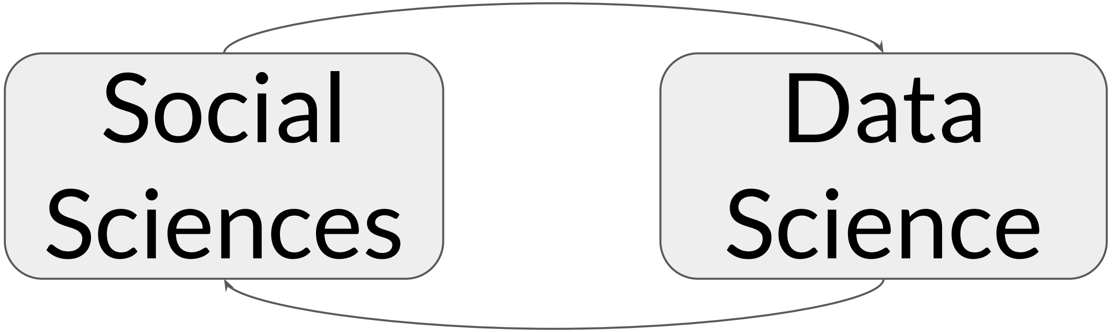
Develop typologies and data resources

Build and evaluate computational models



Develop typologies and data resources

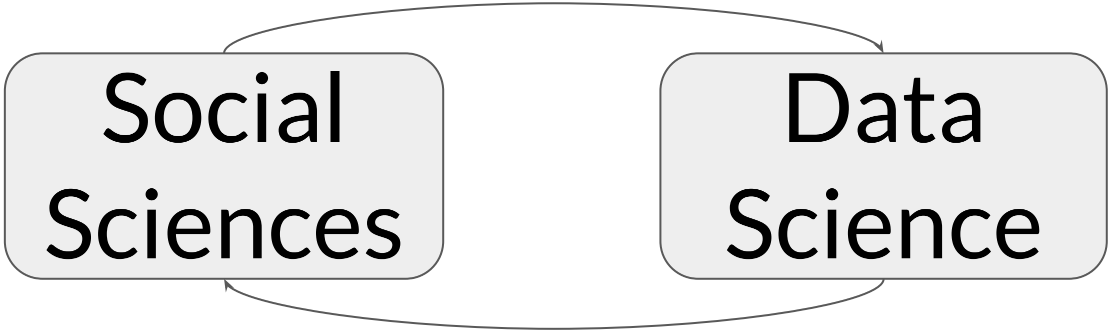
Build and evaluate computational models



Analyze political discourse across multiple domains

Develop typologies and data resources

Build and evaluate computational models



Assess impacts for people and language technology systems

Analyze political discourse across multiple domains

Roadmap

- Overview
- **Framing**
- Dehumanization
- Dogwhistles
- Future Plans

Modeling Framing in Immigration Discourse on Social Media

North American Chapter of the Association for Computational Linguistics (NAACL), 2021

Julia Mendelsohn Ceren Budak David Jurgens

Framing can influence public opinion and policy, but we know little about how ordinary people frame political issues on social media.

Framing can influence public opinion and policy, but we know little about how ordinary people frame political issues on social media.

We combine political communication and NLP to analyze the public's **production** and **reception** of **frames** in immigration discourse on Twitter

What is framing?

> “Selecting some aspects of a perceived reality and make them **more salient** in a communicating text, in such a way as to promote a particular **problem definition, causal interpretation, moral evaluation, and/or treatment recommendation** for the item described” [Entman, 1993]

What is a frame?

What is a frame?

- **Issue-generic Policy** [Boydstun et al., 2013]
  - *Crime & punishment, morality, economic, policy*

What is a frame?

- **Issue-generic Policy** [Boydstun et al., 2013]
    - *Crime & punishment, morality, economic, policy*
- **Immigration-specific** [Benson, 2013]
    - *Immigrants as victims (e.g. of global economy or discrimination)*
    - *Immigrants as heroes (e.g. contributing to economy or cultural diversity)*
    - *Immigrants as threats (e.g. to jobs, or to public safety)*

What is a frame?

- **Issue-generic Policy** [Boydstun et al., 2013]
  - *Crime & punishment, morality, economic, policy*
- **Immigration-specific** [Benson, 2013]
  - *Immigrants as victims (e.g. of global economy or discrimination)*
  - *Immigrants as heroes (e.g. contributing to economy or cultural diversity)*
  - *Immigrants as threats (e.g. to jobs, or to public safety)*
- **Issue-generic Narrative** [Iyengar, 1991]
  - *Episodic: focus on specific actions, events, examples, or case studies*
  - *Thematic: focus on broader political, social, cultural context*

Framing processes

- **Frame-building**: factors affecting how an issue is framed

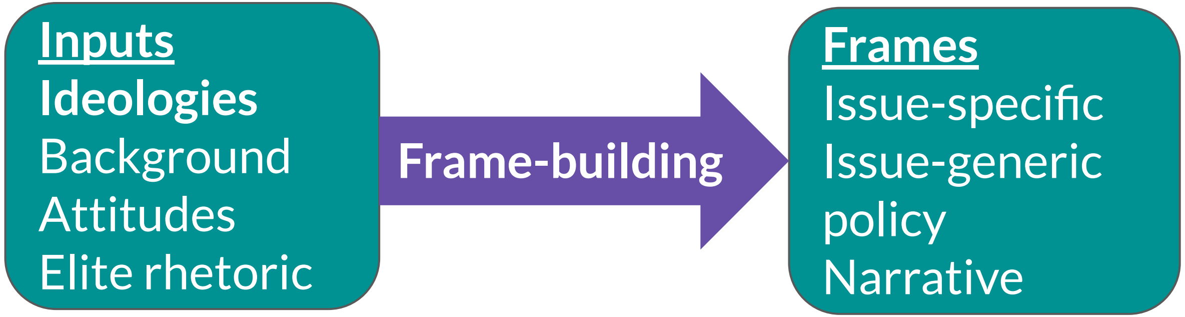


Figure & theoretical model adapted from de Vreese [2005] and is a simplification of Scheufele's [1999] four-process model

Framing processes

- **Frame-building**: factors affecting how an issue is framed
- **Frame-setting**: frame effects on audiences

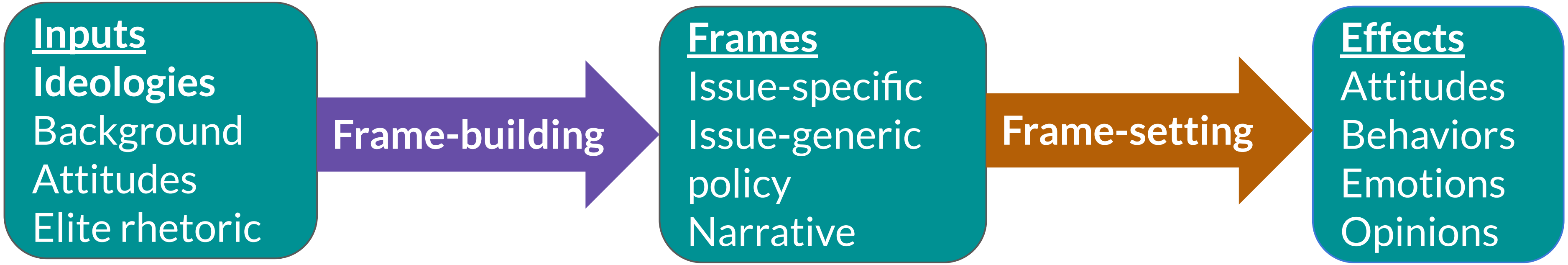


Figure & theoretical model adapted from de Vreese [2005] and is a simplification of Scheufele's [1999] four-process model

Dataset collection & annotation

Dataset collection & annotation

Automated frame detection

Dataset collection & annotation

Automated frame detection

Frame building: role of ideology in framing

Dataset collection & annotation

Automated frame detection

Frame building: role of ideology in framing

Frame setting: effects on user engagement

Dataset collection & annotation

Automated frame detection

Frame building: role of ideology in framing

Frame setting: effects on user engagement

Building a corpus of immigration-related tweets

- 2.6M English-language tweets from 10% sample, 2018-2019
- Contain relevant term *(e.g. immigrant, undocumented, illegals)*

Building a corpus of immigration-related tweets

- 2.6M English-language tweets from 10% sample, 2018-2019
- Contain relevant term *(e.g. immigrant, undocumented, illegals)*
- Ideology estimates based on network structure
  - Continuous Liberal (-) to Conservative (+) scale
  - Bayesian spatial following model assumes homophily [Barberá 2015]
  - Politician accounts provide initial seed estimates

Building a corpus of immigration-related tweets

- 2.6M English-language tweets from 10% sample, 2018-2019
- Contain relevant term *(e.g. immigrant, undocumented, illegals)*
- Ideology estimates based on network structure
  - Continuous Liberal (-) to Conservative (+) scale
  - Bayesian spatial following model assumes homophily [Barberá 2015]
  - Politician accounts provide initial seed estimates
- Codebook development for each frame typology

Building a corpus of immigration-related tweets

- 2.6M English-language tweets from 10% sample, 2018-2019
- Contain relevant term *(e.g. immigrant, undocumented, illegals)*
- Ideology estimates based on network structure
  - Continuous Liberal (-) to Conservative (+) scale
  - Bayesian spatial following model assumes homophily [Barberá 2015]
  - Politician accounts provide initial seed estimates
- Codebook development for each frame typology
- Manually annotated 4500 tweets

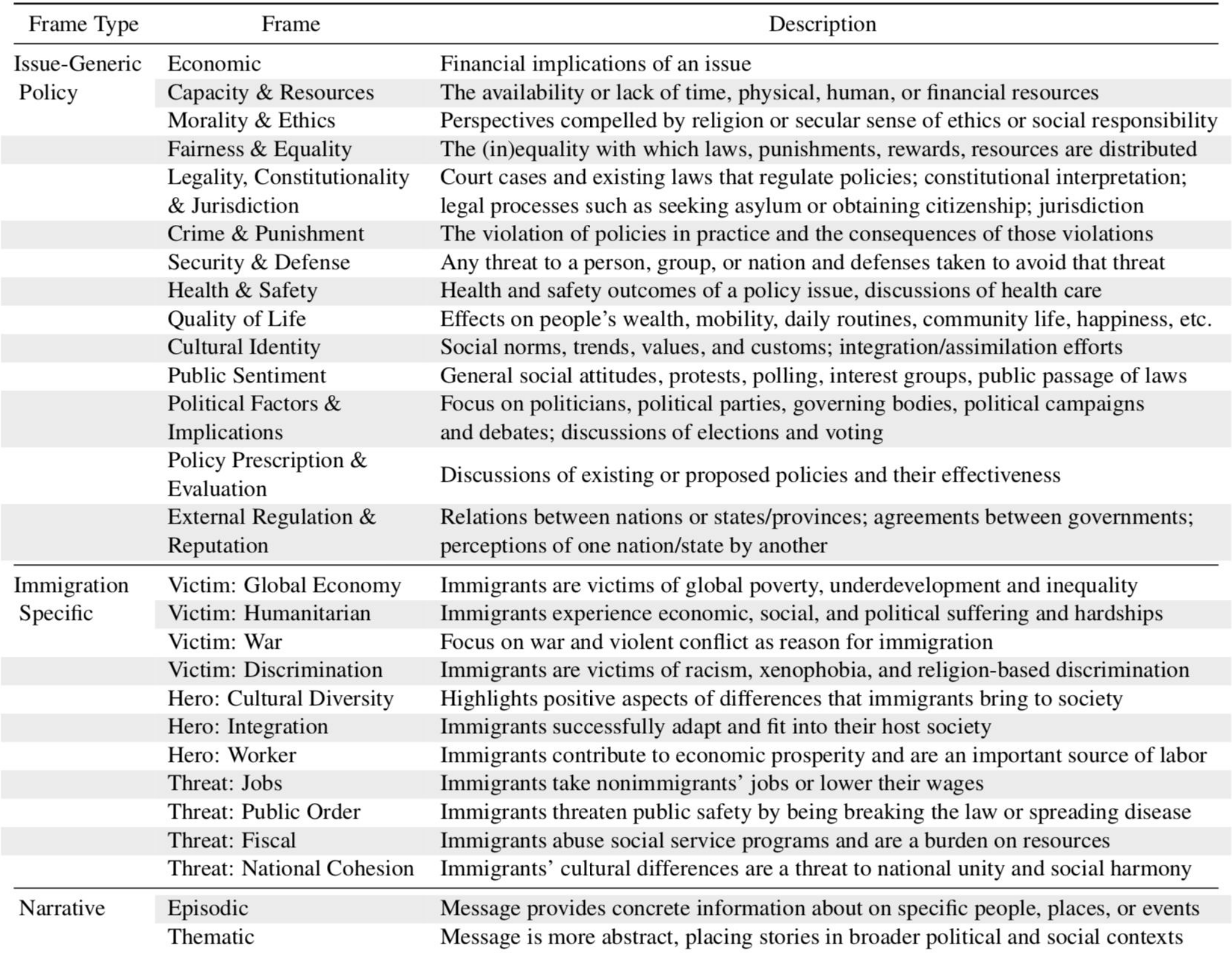
Building a corpus of immigration-related tweets

- 2.6M English-language tweets from 10% sample, 2018-2019
- Contain relevant term *(e.g. immigrant, undocumented, illegals)*
- Ideology estimates based on network structure
  - Continuous Liberal (-) to Conservative (+) scale
  - Bayesian spatial following model assumes homophily [Barberá 2015]
  - Politician accounts provide initial seed estimates
- Codebook development for each frame typology
- Manually annotated 4500 tweets
- *We also analyzed framing across the US, UK, and EU*

Data Annotation

3 typologies

27 categories



Frame Type	Frame	Description
Issue-Generic Policy	Economic	Financial implications of an issue
	Capacity & Resources	The availability or lack of time, physical, human, or financial resources
	Morality & Ethics	Perspectives compelled by religion or secular sense of ethics or social responsibility
	Fairness & Equality	The (in)equality with which laws, punishments, rewards, resources are distributed
	Legality, Constitutionality & Jurisdiction	Court cases and existing laws that regulate policies; constitutional interpretation; legal processes such as seeking asylum or obtaining citizenship; jurisdiction
	Crime & Punishment	The violation of policies in practice and the consequences of those violations
	Security & Defense	Any threat to a person, group, or nation and defenses taken to avoid that threat
	Health & Safety	Health and safety outcomes of a policy issue, discussions of health care
	Quality of Life	Effects on people's wealth, mobility, daily routines, community life, happiness, etc.
	Cultural Identity	Social norms, trends, values, and customs; integration/assimilation efforts
	Public Sentiment	General social attitudes, protests, polling, interest groups, public passage of laws
	Political Factors & Implications	Focus on politicians, political parties, governing bodies, political campaigns and debates; discussions of elections and voting
	Policy Prescription & Evaluation	Discussions of existing or proposed policies and their effectiveness
	External Regulation & Reputation	Relations between nations or states/provinces; agreements between governments; perceptions of one nation/state by another
Immigration Specific	Victim: Global Economy	Immigrants are victims of global poverty, underdevelopment and inequality
	Victim: Humanitarian	Immigrants experience economic, social, and political suffering and hardships
	Victim: War	Focus on war and violent conflict as reason for immigration
	Victim: Discrimination	Immigrants are victims of racism, xenophobia, and religion-based discrimination
	Hero: Cultural Diversity	Highlights positive aspects of differences that immigrants bring to society
	Hero: Integration	Immigrants successfully adapt and fit into their host society
	Hero: Worker	Immigrants contribute to economic prosperity and are an important source of labor
	Threat: Jobs	Immigrants take nonimmigrants' jobs or lower their wages
	Threat: Public Order	Immigrants threaten public safety by being breaking the law or spreading disease
	Threat: Fiscal	Immigrants abuse social service programs and are a burden on resources
	Threat: National Cohesion	Immigrants' cultural differences are a threat to national unity and social harmony
Narrative	Episodic	Message provides concrete information about on specific people, places, or events
	Thematic	Message is more abstract, placing stories in broader political and social contexts

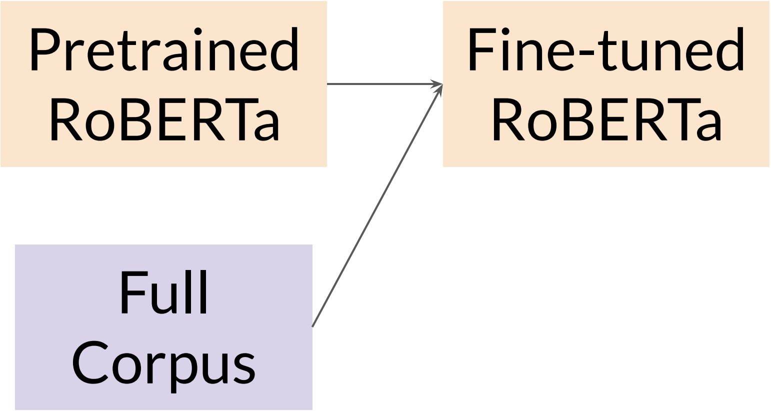
Dataset collection & annotation

Automated frame detection

Frame building: role of ideology in framing

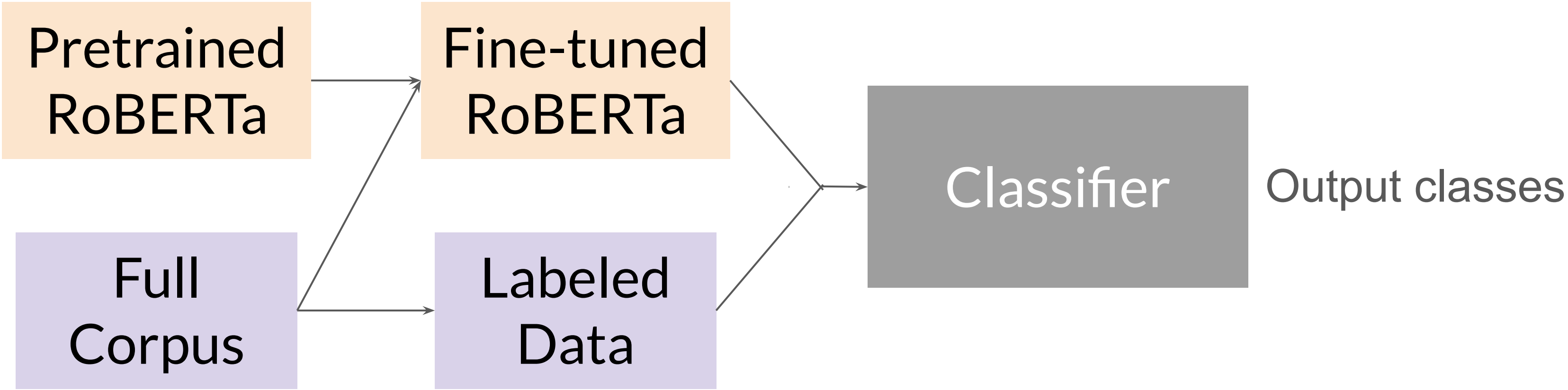
Frame setting: effects on user engagement

Multilabel classification



- Fine-tune RoBERTa [Liu et al., 2019] to recognize patterns in immigration tweets

Multilabel classification



Pretrained RoBERTa

Fine-tuned RoBERTa

Full Corpus

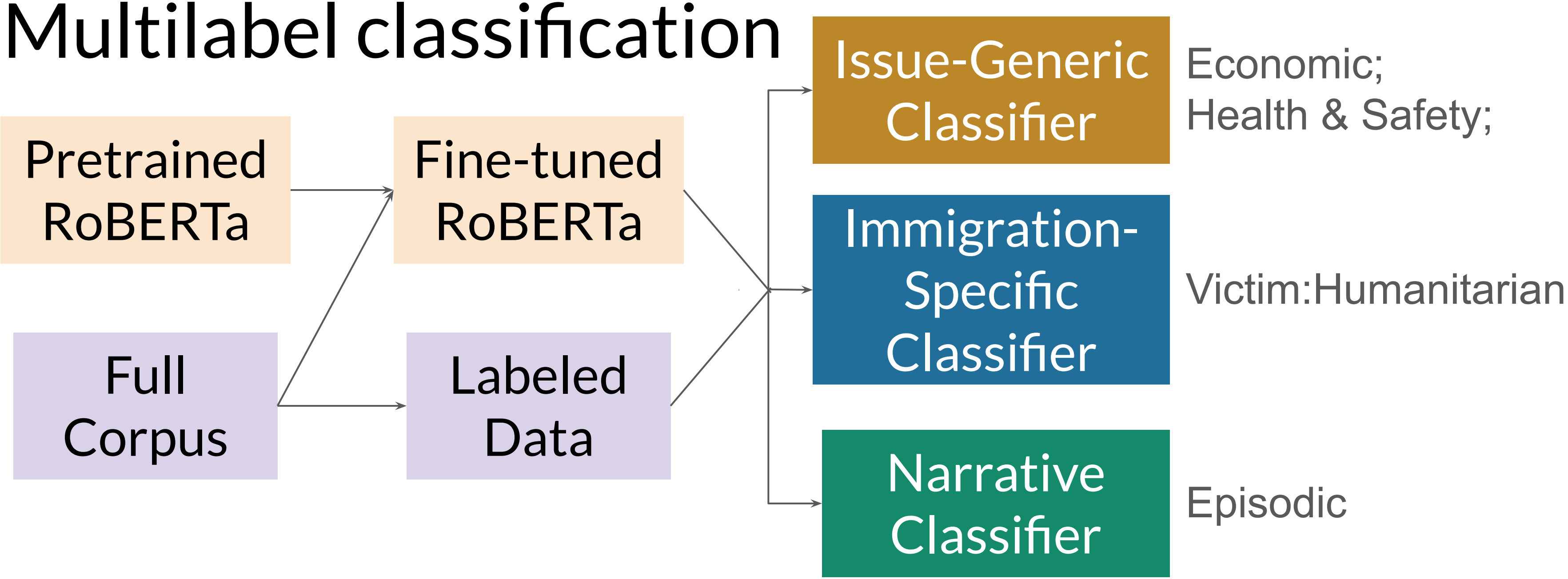
Labeled Data

Classifier

Output classes

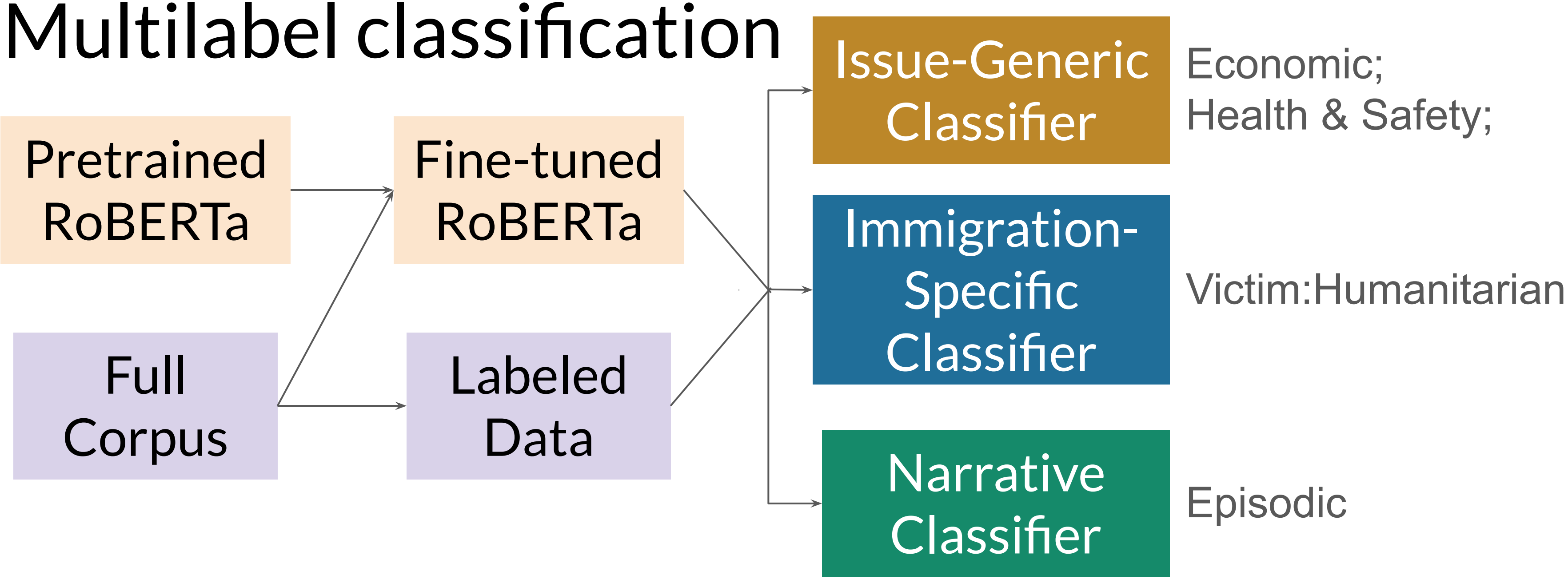
- Fine-tune RoBERTa [Liu et al., 2019] to recognize patterns in immigration tweets

Multilabel classification



- Fine-tune RoBERTa [Liu et al., 2019] to recognize patterns in immigration tweets

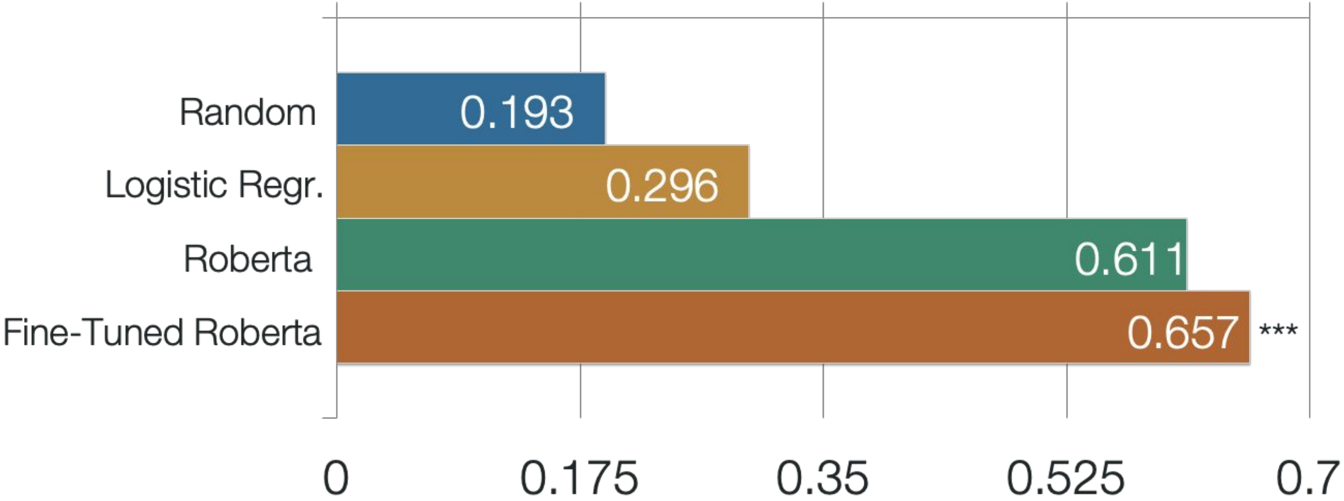
Multilabel classification



- Fine-tune RoBERTa [Liu et al., 2019] to recognize patterns in immigration tweets
- **Baselines**: random prediction, logistic regression with unigram and bigram features, RoBERTa without fine-tuning

Fine-tuned ROBERTa outperforms all baselines

F1 score by model on test set



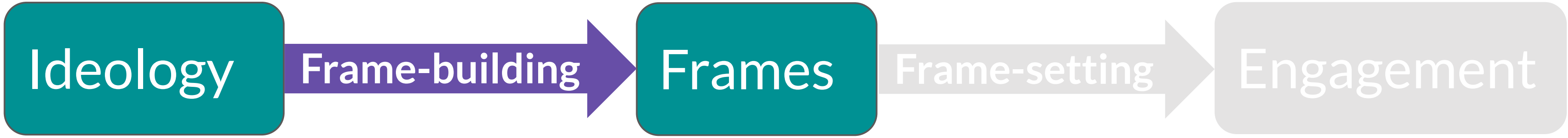
Dataset collection & annotation

Automated frame detection

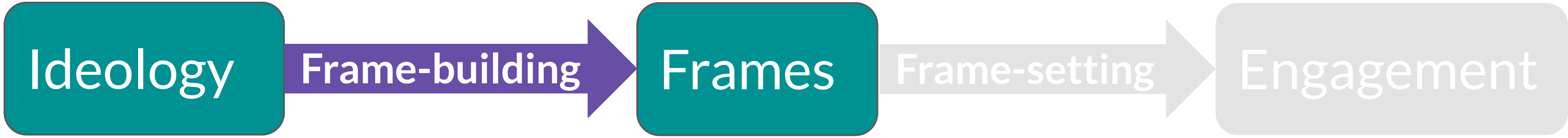
Frame building: role of ideology in framing

Frame setting: effects on user engagement

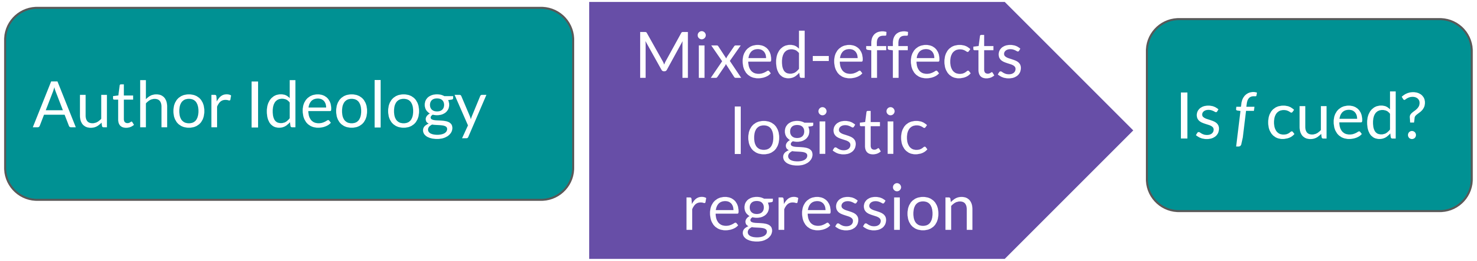
Role of ideology in selecting frames



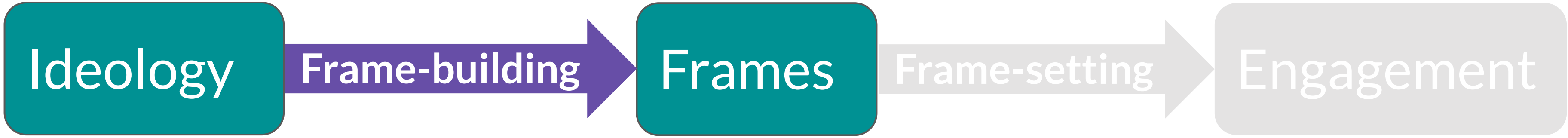
Role of ideology in selecting frames



For each frame *f*:



Role of ideology in selecting frames



Ideology

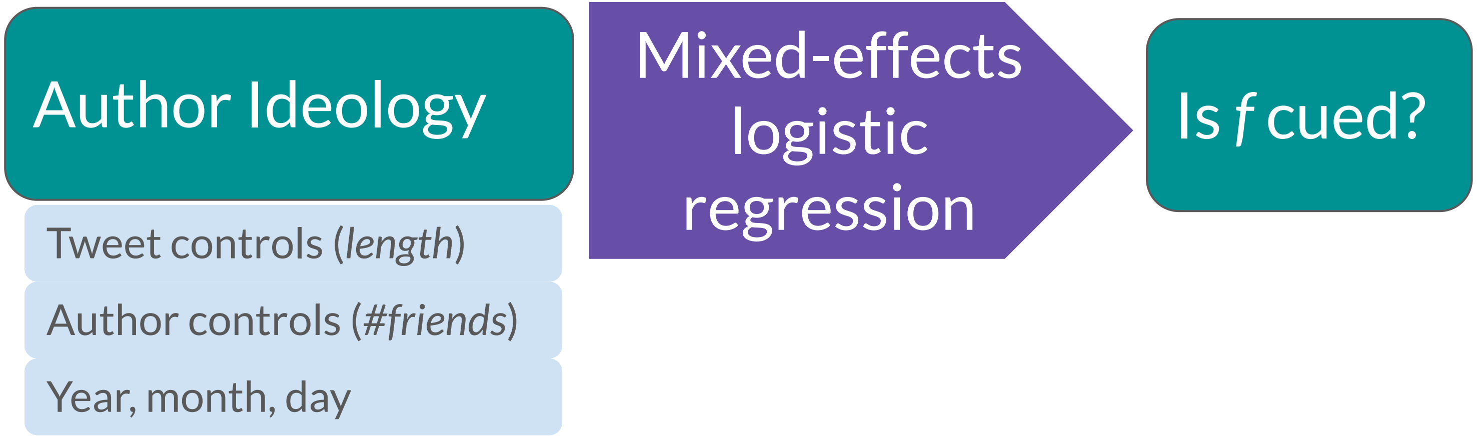
Frame-building

Frames

Frame-setting

Engagement

For each frame *f*:



Author Ideology

Tweet controls (*length*)

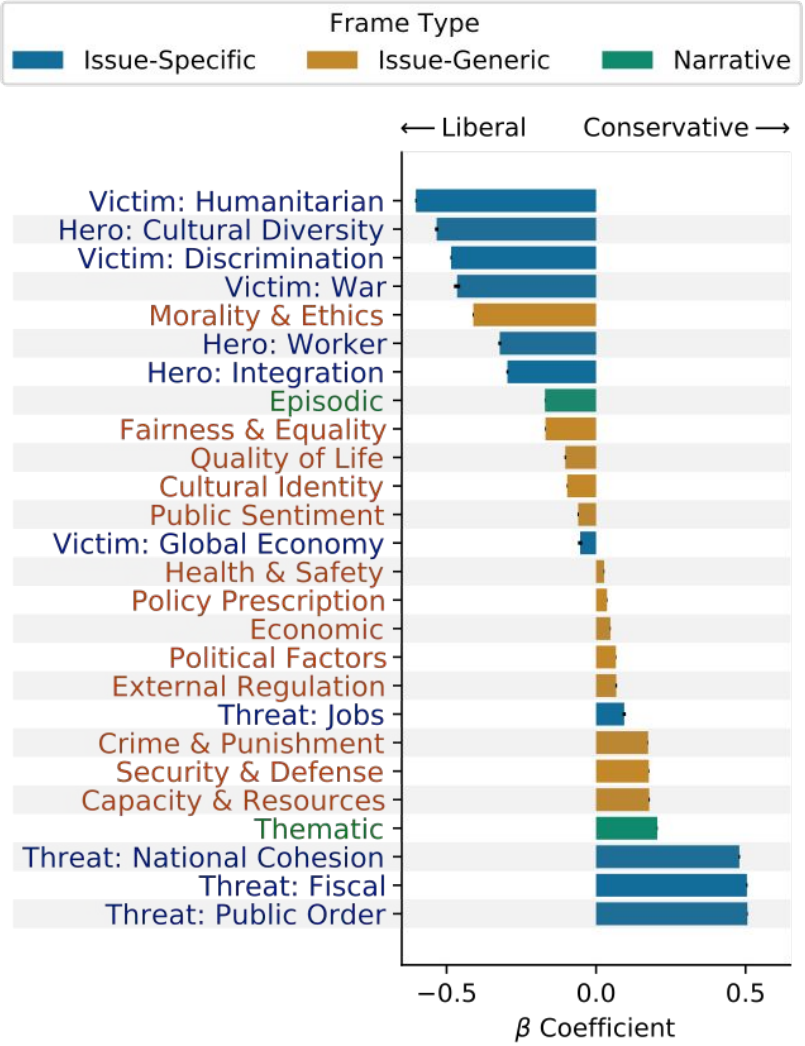
Author controls (*#friends*)

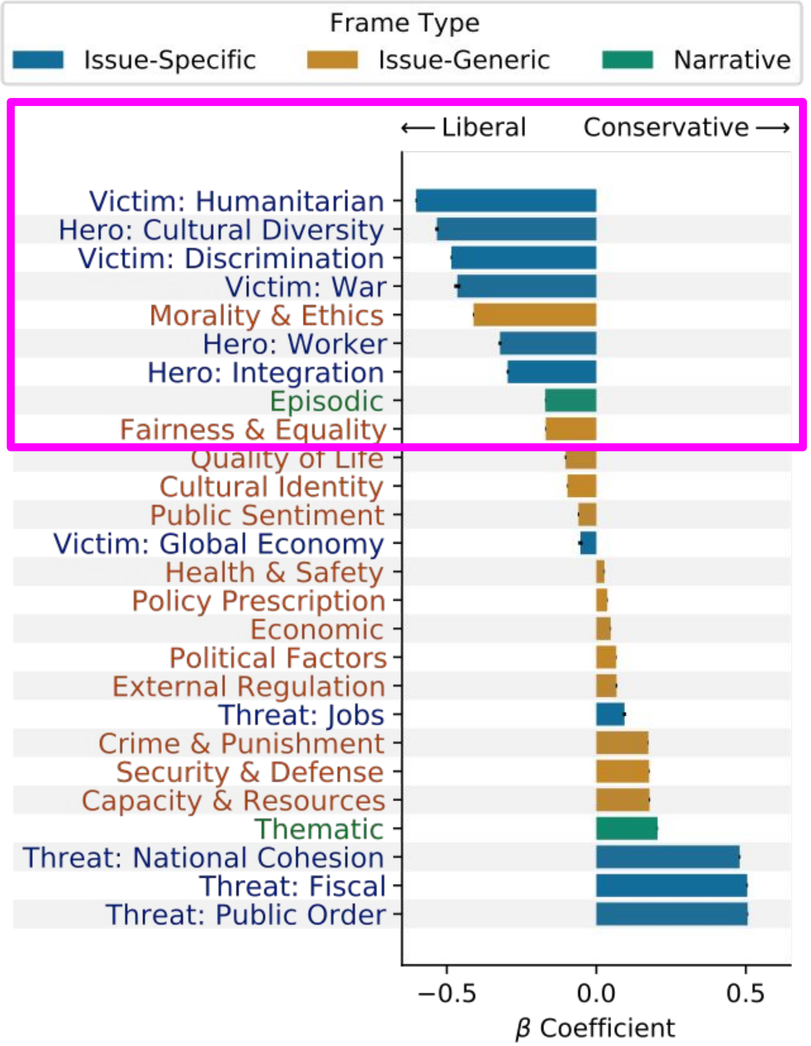
Year, month, day

Mixed-effects logistic regression

Is *f* cued?

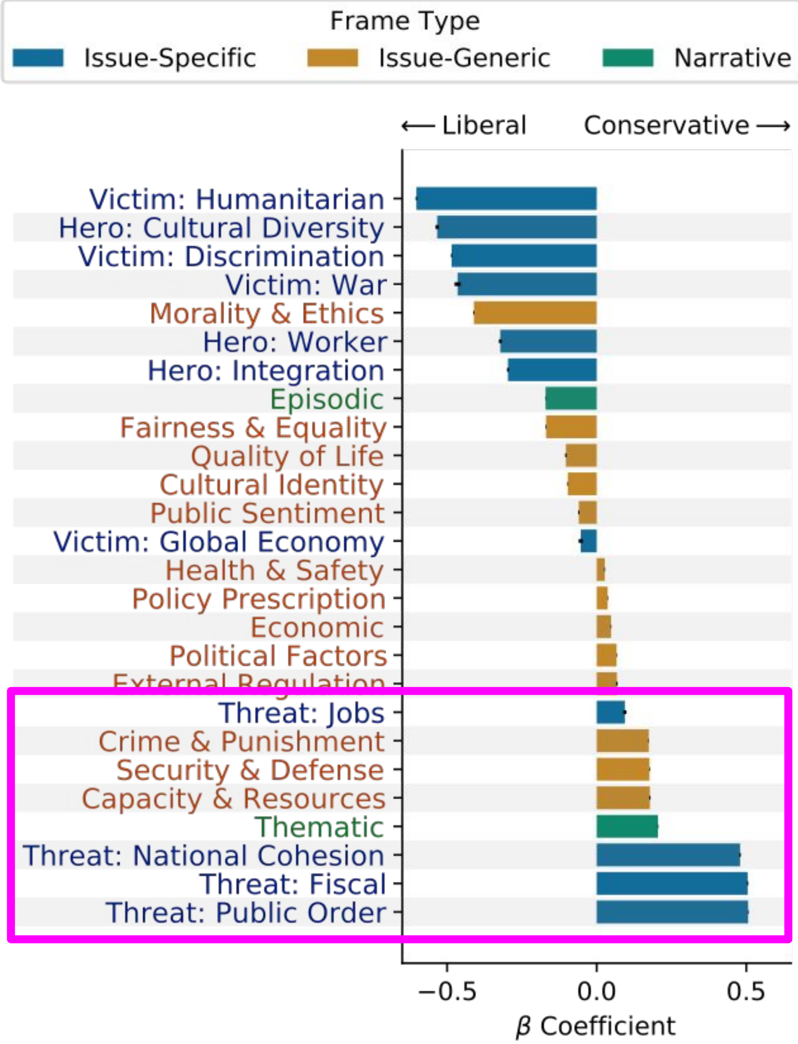
Frame Type
Issue-Specific
Issue-Generic
Narrative
⟵ Liberal
Conservative ⟶
Victim: Humanitarian
Hero: Cultural Diversity
Victim: Discrimination
Victim: War
Morality & Ethics
Hero: Worker
Hero: Integration
Episodic
Fairness & Equality
Quality of Life
Cultural Identity
Public Sentiment
Victim: Global Economy
Health & Safety
Policy Prescription
Economic
Political Factors
External Regulation
Threat: Jobs
Crime & Punishment
Security & Defense
Capacity & Resources
Thematic
Threat: National Cohesion
Threat: Fiscal
Threat: Public Order
−0.5
0.0
0.5
β Coefficient


Frame Type
Issue-Specific
Issue-Generic
Narrative
⟵ Liberal
Conservative ⟶
Victim: Humanitarian
Hero: Cultural Diversity
Victim: Discrimination
Victim: War
Morality & Ethics
Hero: Worker
Hero: Integration
Episodic
Fairness & Equality
Quality of Life
Cultural Identity
Public Sentiment
Victim: Global Economy
Health & Safety
Policy Prescription
Economic
Political Factors
External Regulation
Threat: Jobs
Crime & Punishment
Security & Defense
Capacity & Resources
Thematic
Threat: National Cohesion
Threat: Fiscal
Threat: Public Order
−0.5
0.0
0.5
β Coefficient



Liberals frame immigrants as **heroes** and **victims**

- Liberals cue ***fairness*** and ***morality***, framing immigrants as ***victims of discrimination*** and ***inhumane*** policies.



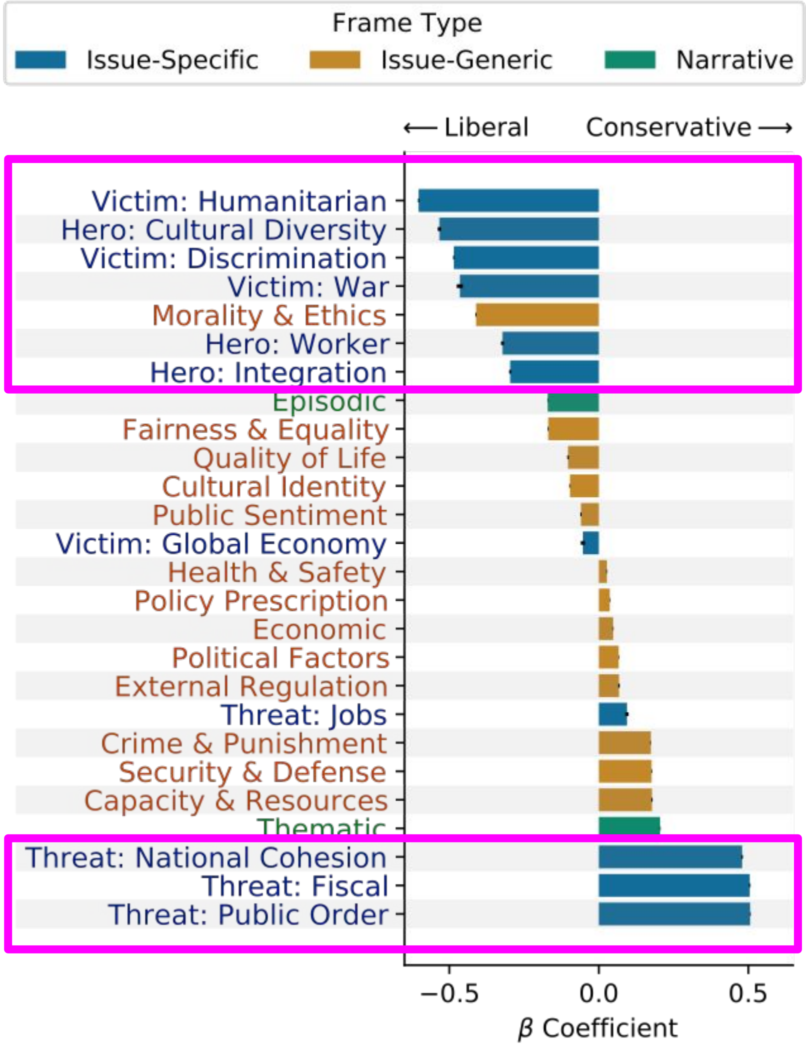
Liberals frame immigrants as **heroes** and **victims**

- Liberals cue ***fairness*** and ***morality***, framing immigrants as ***victims of discrimination*** and ***inhumane*** policies.

Conservatives frame immigrants as **threats**

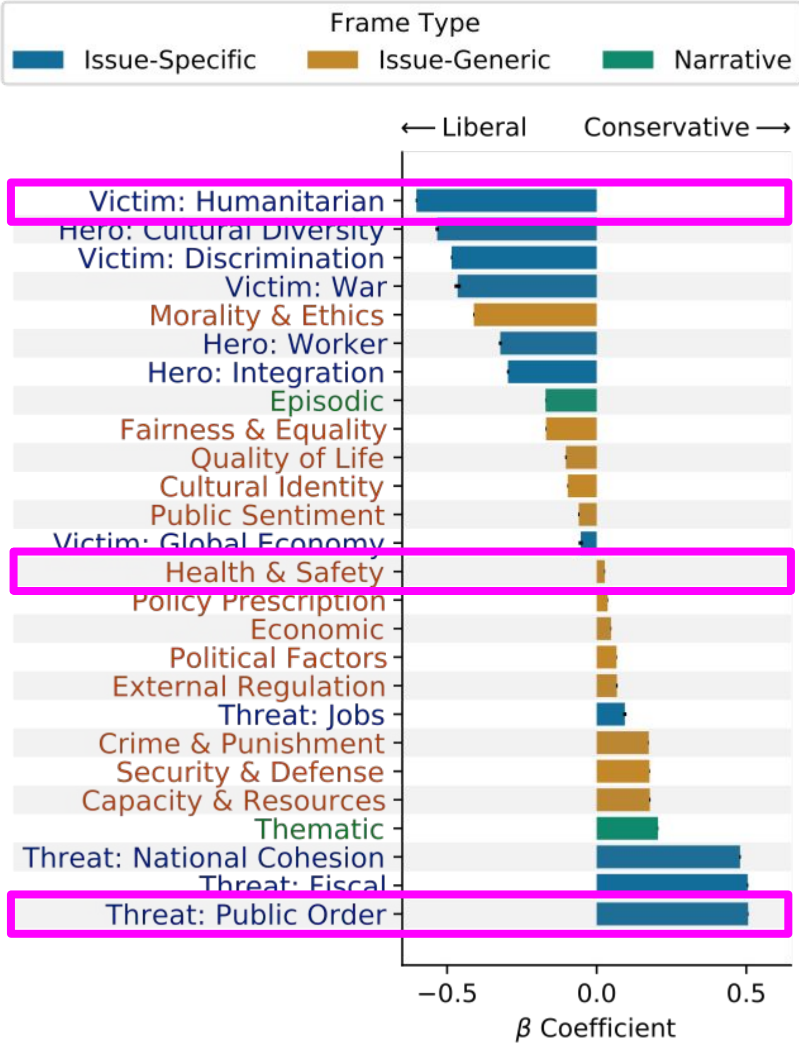
- Conservatives cue ***threat to public safety***, ***burden on taxpayers & government programs***

Each frame typology offers value



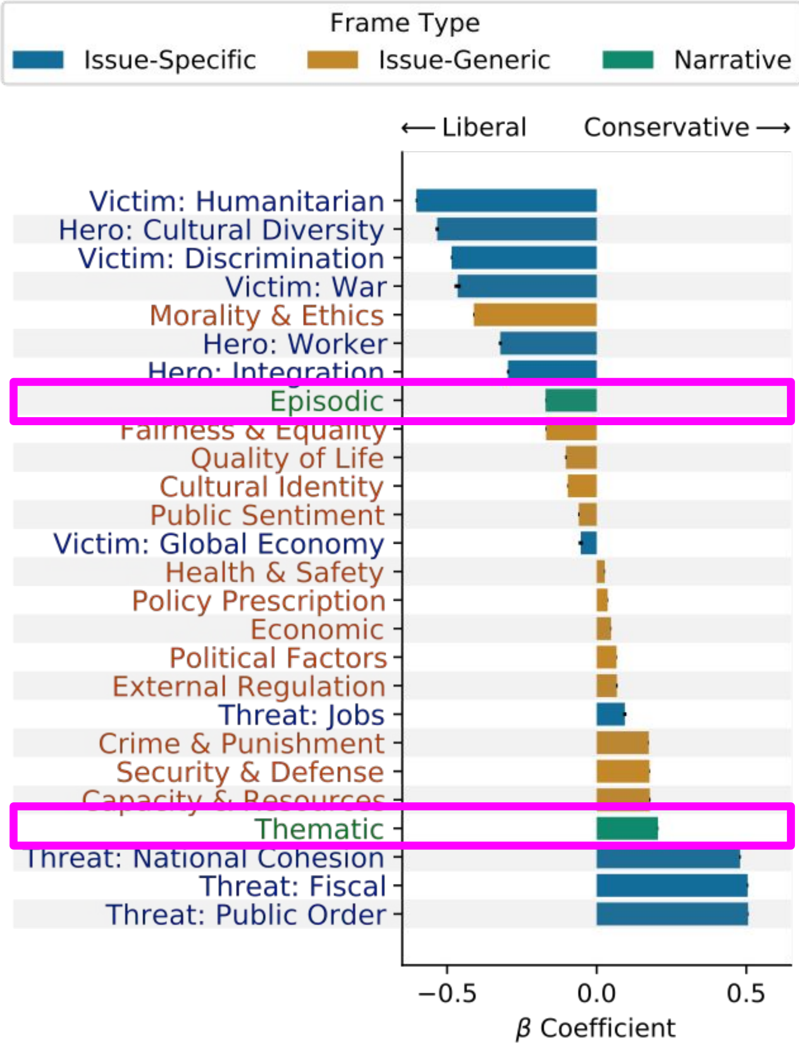
Immigration-specific frames reveal ideological differences obscured by issue-generic policy frames

Each frame typology offers value



Immigration-specific frames reveal ideological differences obscured by issue-generic policy frames

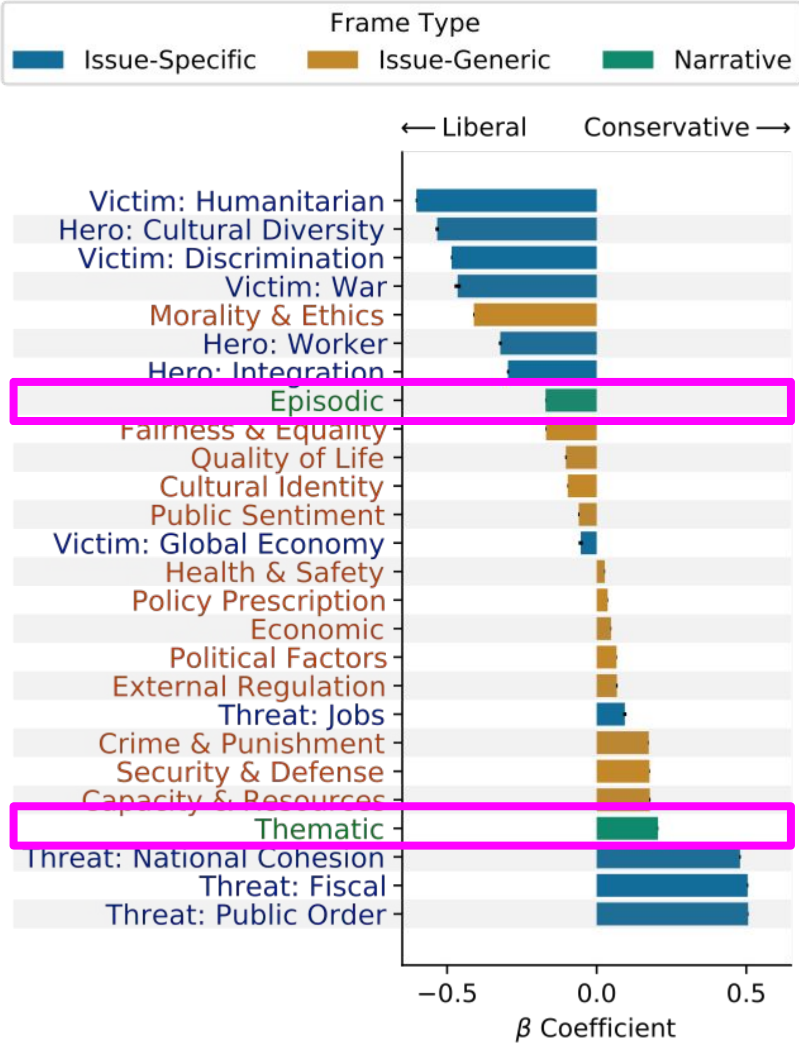
(e.g. *health & safety*)



Each frame typology offers value

We uncover ideological variation in narrative framing

- Liberals →episodic frames
- Conservatives →thematic frames



Each frame typology offers value

We uncover ideological variation in narrative framing

- Liberals →episodic frames
- Conservatives →thematic frames
- Similar to immigration news [Somaini, 2019]
- Role of emotion? [Iyengar 1991, Pliskin et al., 2014]

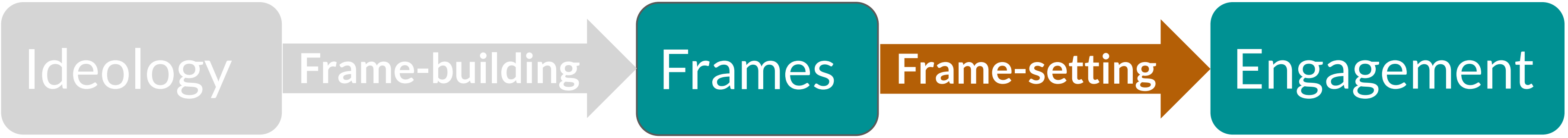
Dataset collection & annotation

Automated frame detection

Frame building: role of ideology in framing

Frame setting: effects on user engagement

How does framing impact a message's audience?



Ideology

Frame-building

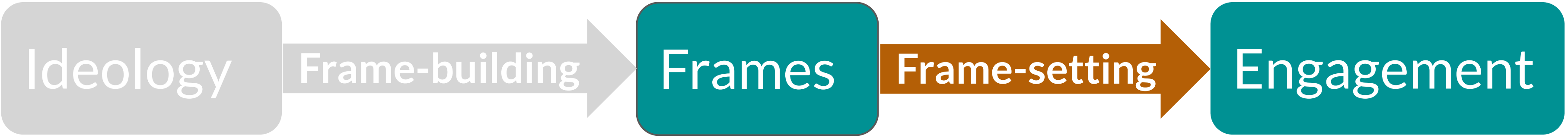
Frames

Frame-setting

Engagement

Framing impacts readers' opinions about immigration [Lecheler et al., 2015]

How does framing impact a message’s audience?

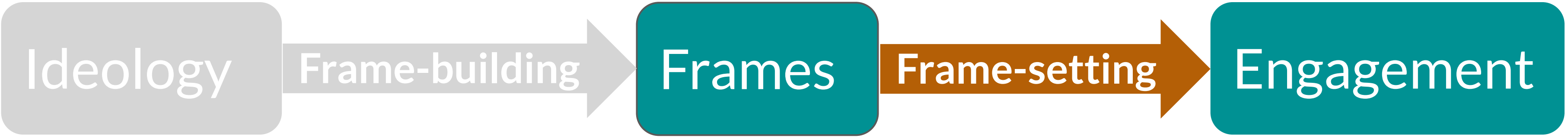


Ideology → Frame-building → Frames → Frame-setting → Engagement

Framing impacts readers’ opinions about immigration [Lecheler et al., 2015]

Twitter provides insight into frame-setting via interactive signals

How does framing impact a message’s audience?

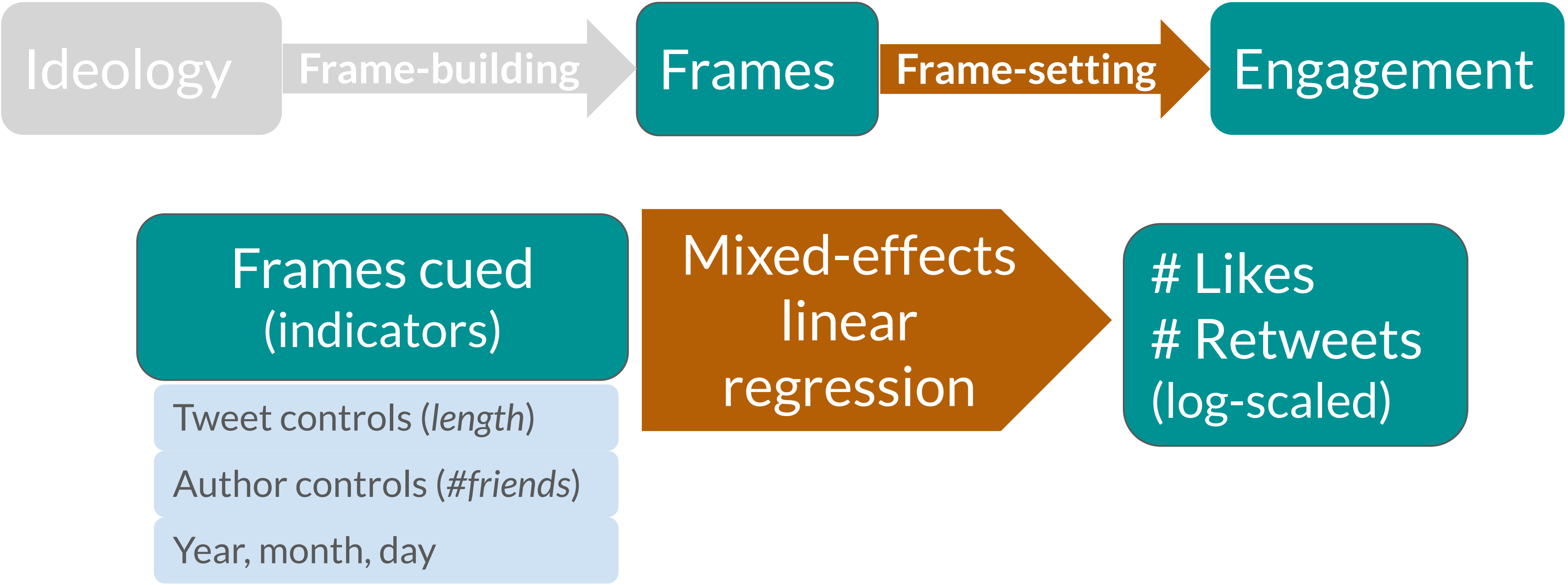


Framing impacts readers’ opinions about immigration [Lecheler et al., 2015]

Twitter provides insight into frame-setting via interactive signals

- ♥ **Favoriting**: endorsement, reader aligns with author’s message
- 🔁 **Retweeting**: amplification, diverse motivations, e.g. desire to inform or entertain others [boyd et al., 2010]

How does framing impact a message’s audience?



Ideology → Frame-building → Frames → Frame-setting → Engagement

Frames cued (indicators)

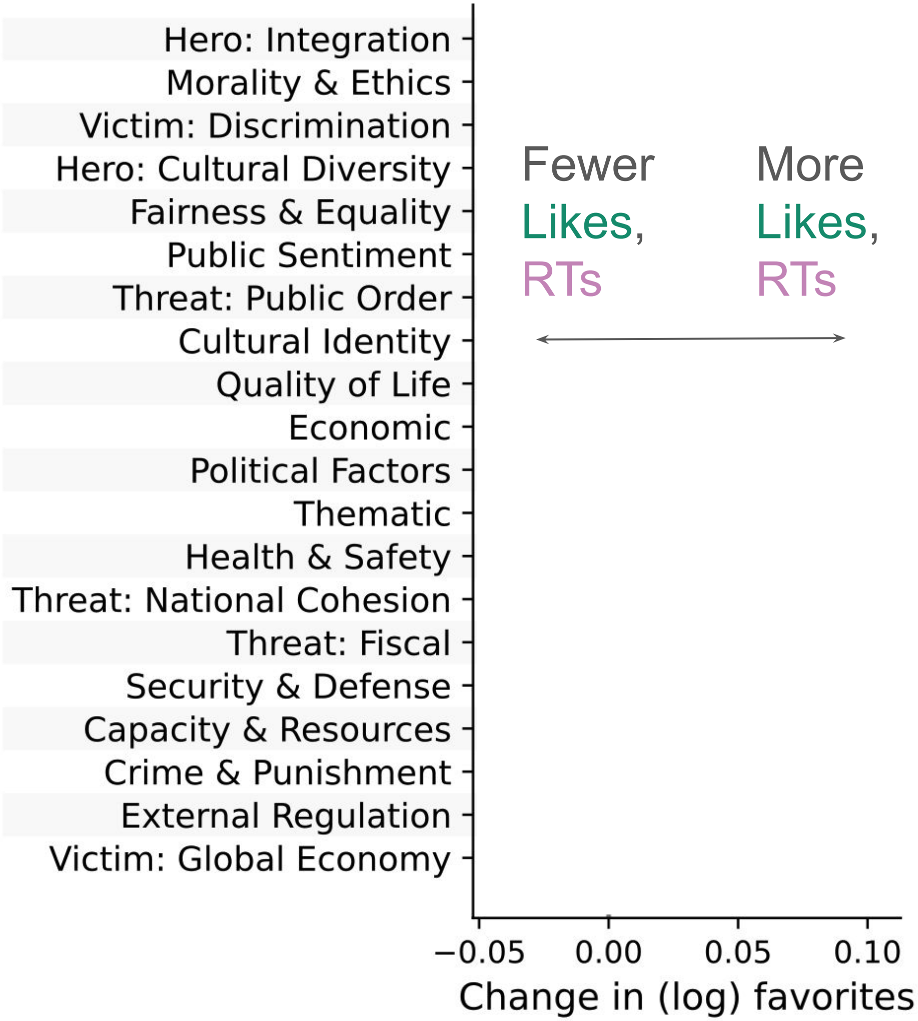
Tweet controls (*length*)

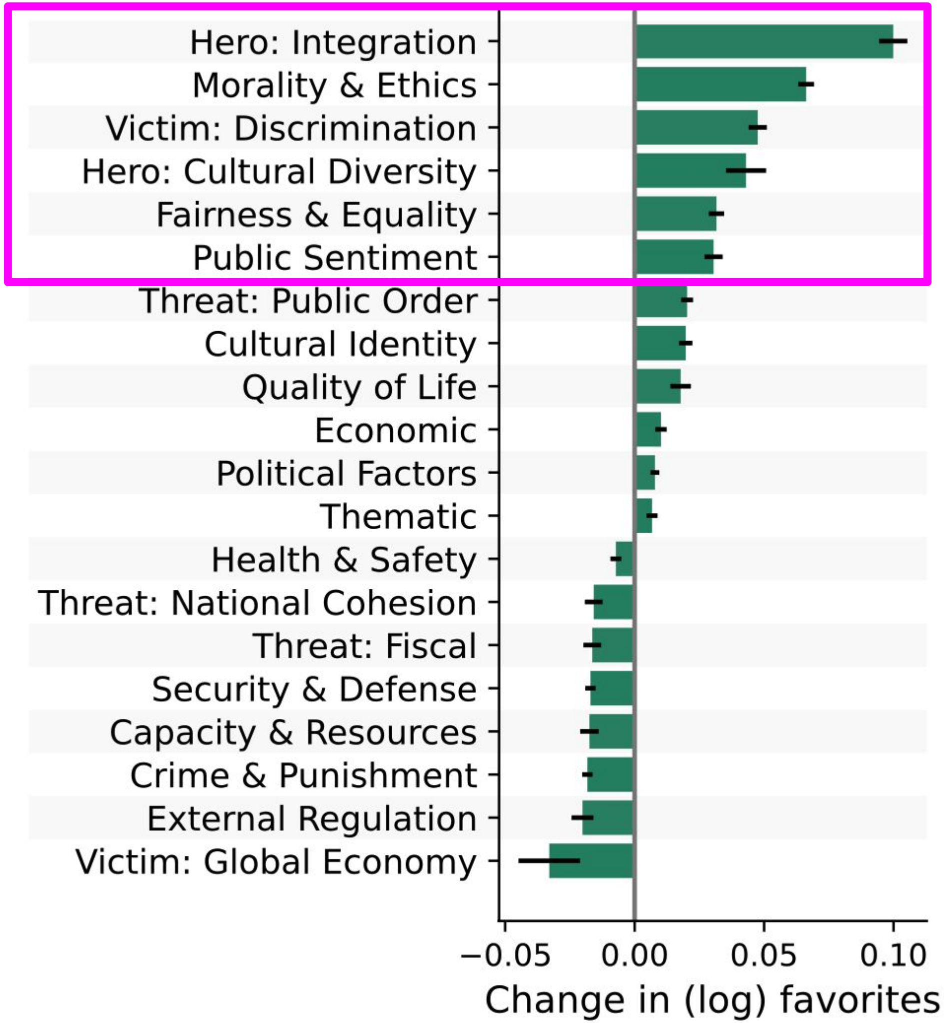
Author controls (*#friends*)

Year, month, day

Mixed-effects linear regression →

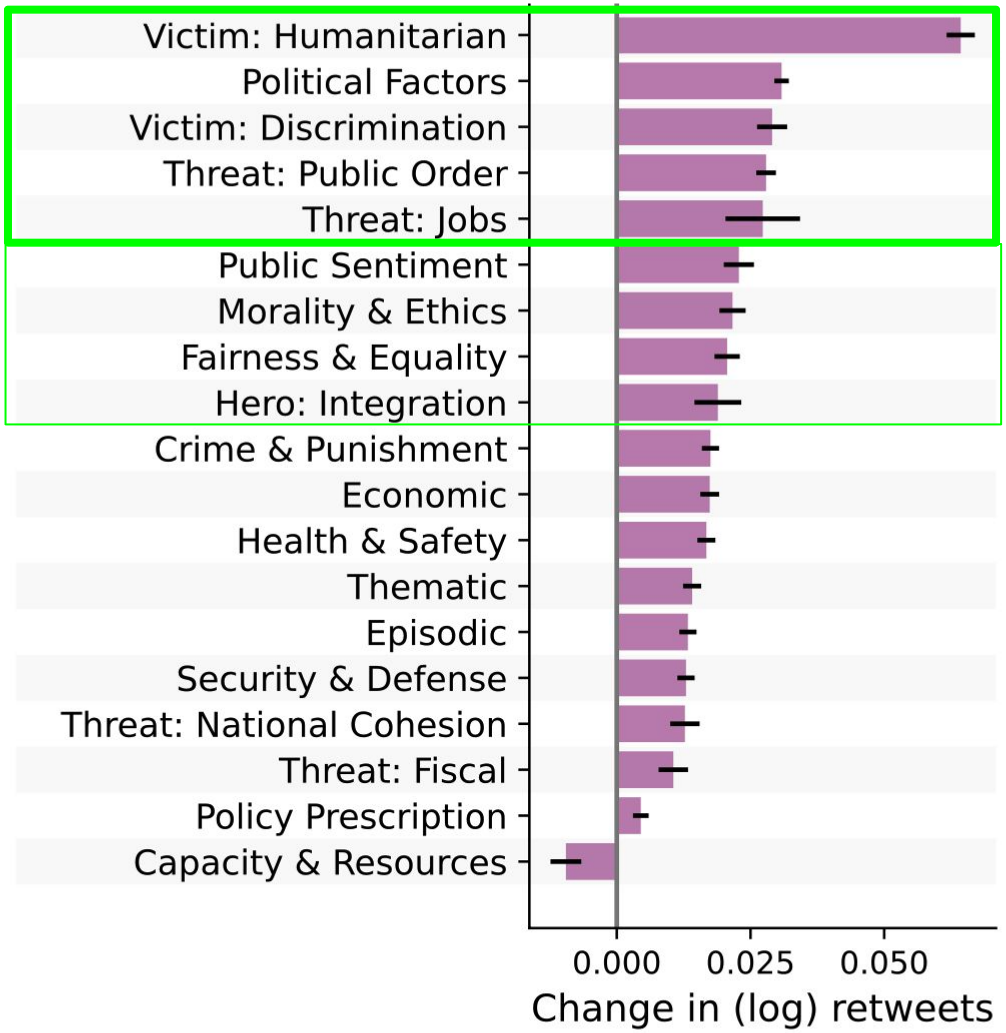
# Likes
# Retweets
(log-scaled)


Hero: Integration
Morality & Ethics
Victim: Discrimination
Hero: Cultural Diversity
Fairness & Equality
Public Sentiment
Threat: Public Order
Cultural Identity
Quality of Life
Economic
Political Factors
Thematic
Health & Safety
Threat: National Cohesion
Threat: Fiscal
Security & Defense
Capacity & Resources
Crime & Punishment
External Regulation
Victim: Global Economy
Fewer
Likes,
RTs
More
Likes,
RTs
−0.05
0.00
0.05
0.10
Change in (log) favorites



❤

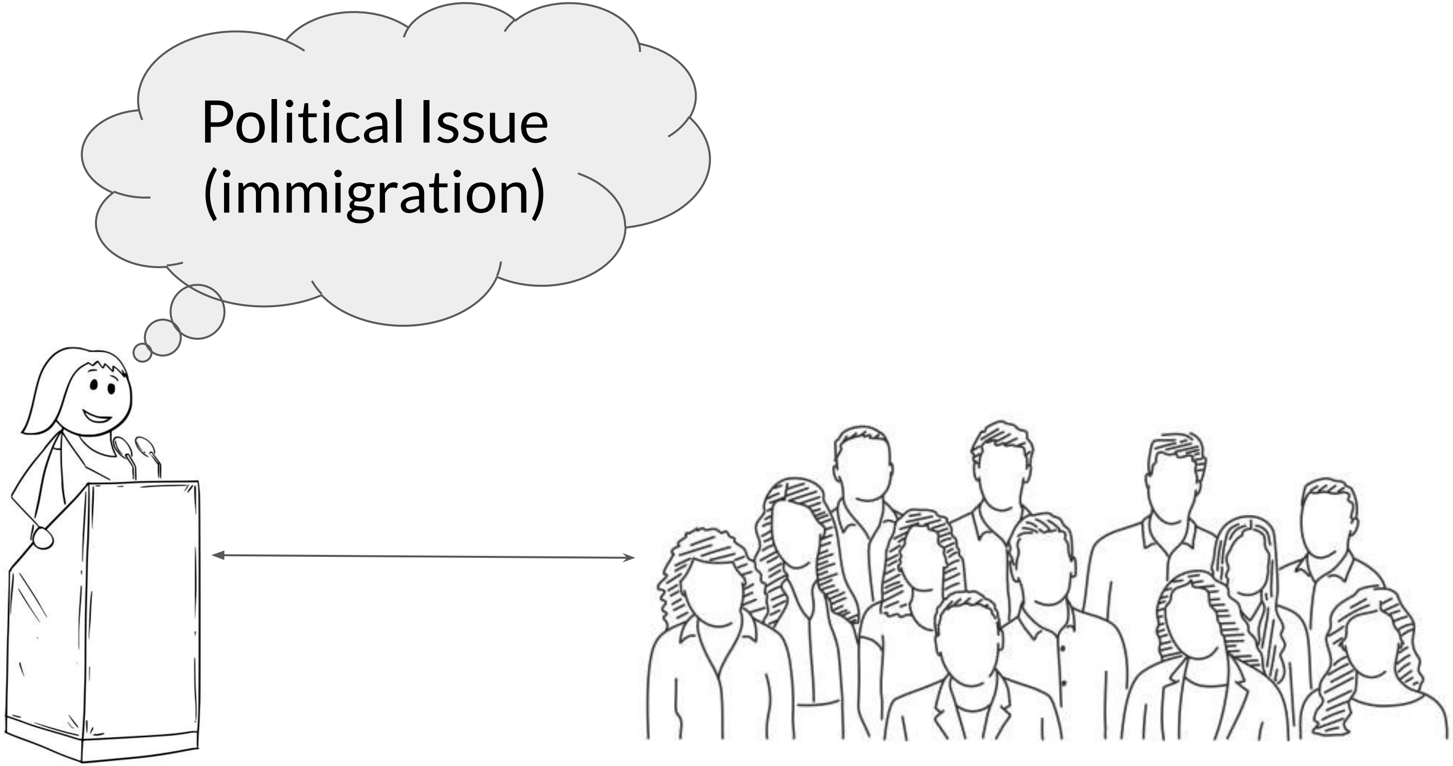
Cultural (*hero: integration*) and **human interest** (*morality, fairness, victim: discrimination*)



***Issue-specific* security & safety** (*threat: public order, victim: humanitarian*), **political**, **human interest**

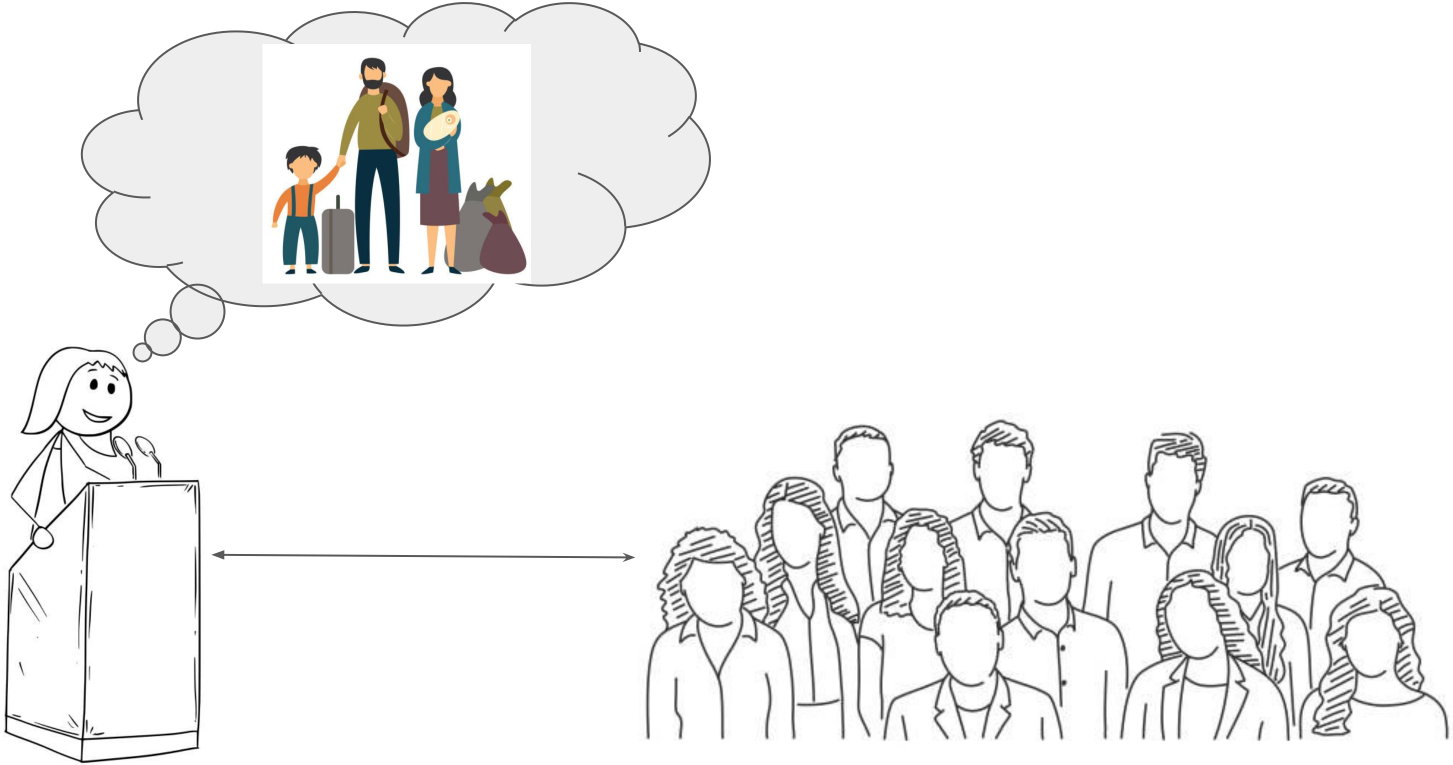
- Dataset collection & annotation
- Automated frame detection
- Frame building: role of ideology in framing
- Frame setting: effects on user engagement

The people in political discourse

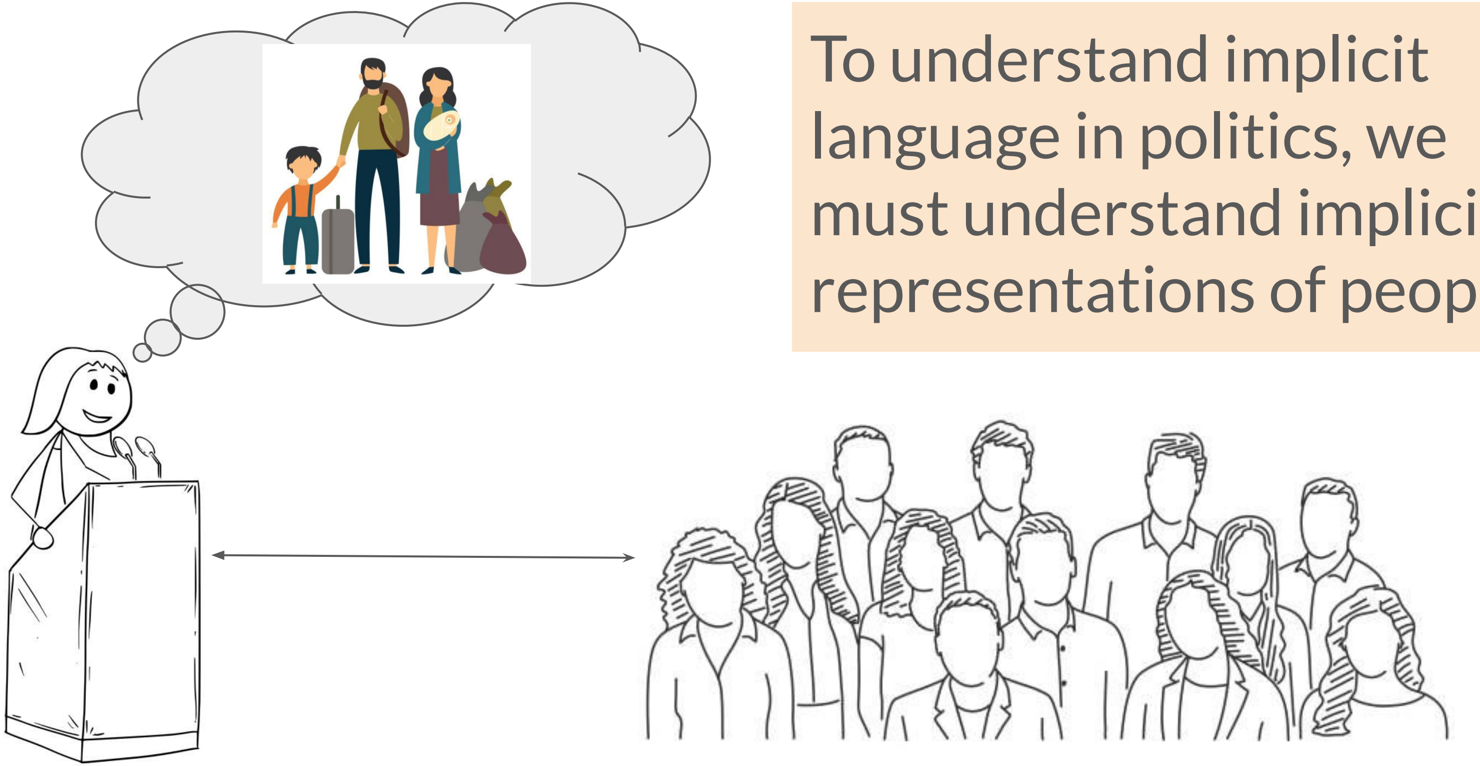


Political Issue
(immigration)

The people in political discourse



The people in political discourse



To understand implicit language in politics, we must understand implicit representations of people

Roadmap

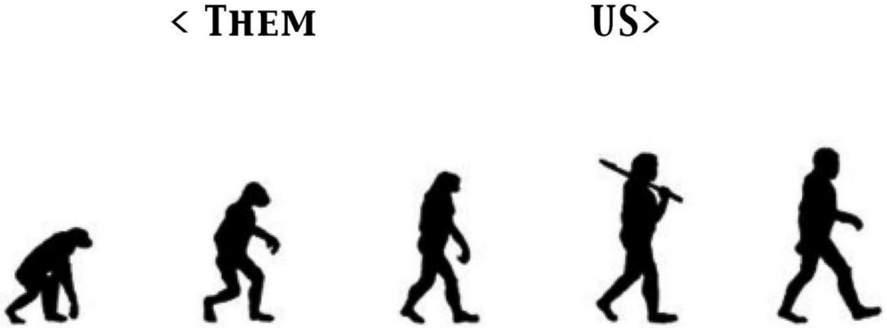
- Overview
- Framing
- Dehumanization
- Dogwhistles
- Future Plans

A Framework for the Computational Linguistic Analysis of Dehumanization

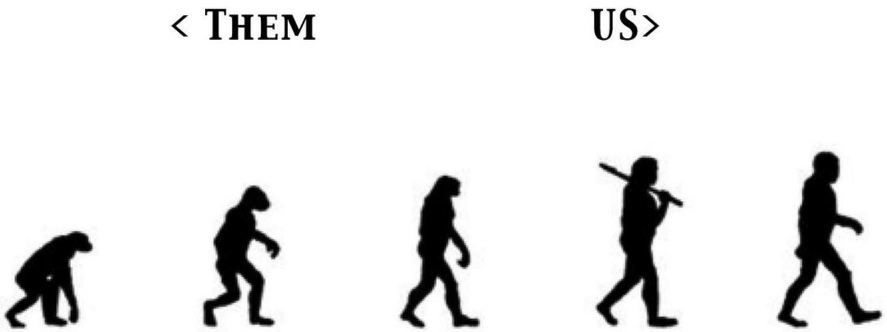
Frontiers in Artificial Intelligence, 2020

Julia Mendelsohn Dan Jurafsky Yulia Tsvetkov

Dehumanization: perceiving or treating people as less than human. It leads to extreme intergroup bias and violence. [Haslam & Stratemeyer, 2016]

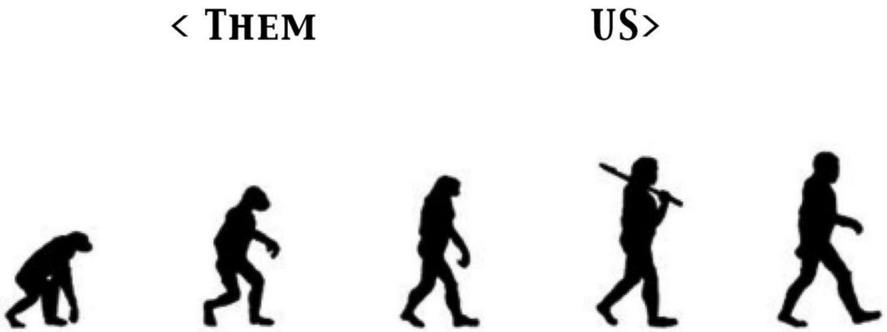


Dehumanization: perceiving or treating people as less than human. It leads to extreme intergroup bias and violence. [Haslam & Stratemeyer, 2016]



Dehumanization is expressed through language, but often subtly

Dehumanization: perceiving or treating people as less than human. It leads to extreme intergroup bias and violence. [Haslam & Stratemeyer, 2016]



Dehumanization is expressed through language, but often subtly

Computational techniques expose subtle associations & facilitate broad analyses of how marginalized groups are portrayed

Introduce framework and computational linguistic measures

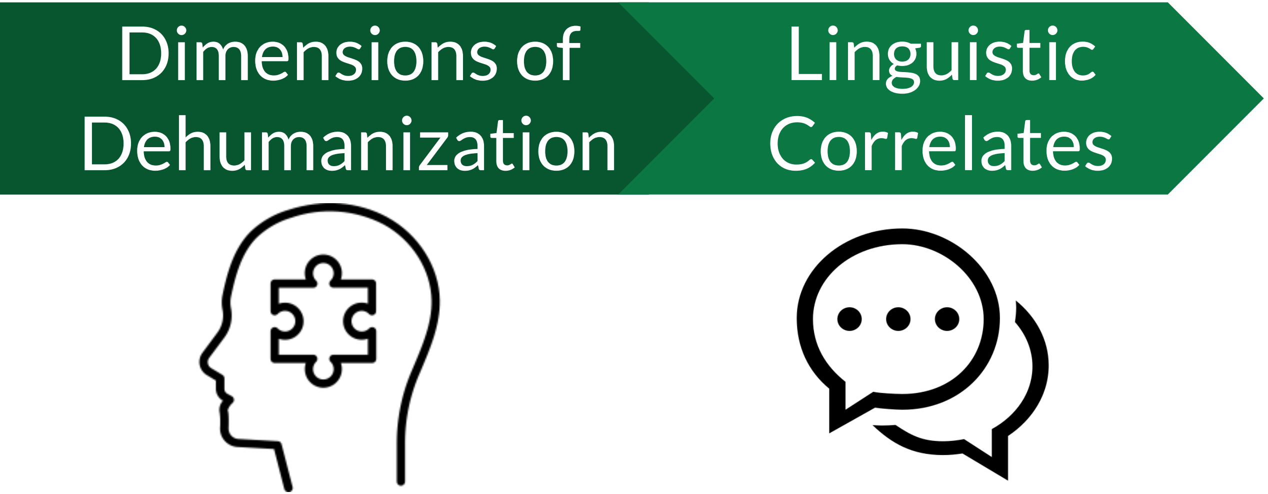
Case study of LGBTQ representation in the New York Times

Our framework

Our framework

Dimensions of Dehumanization

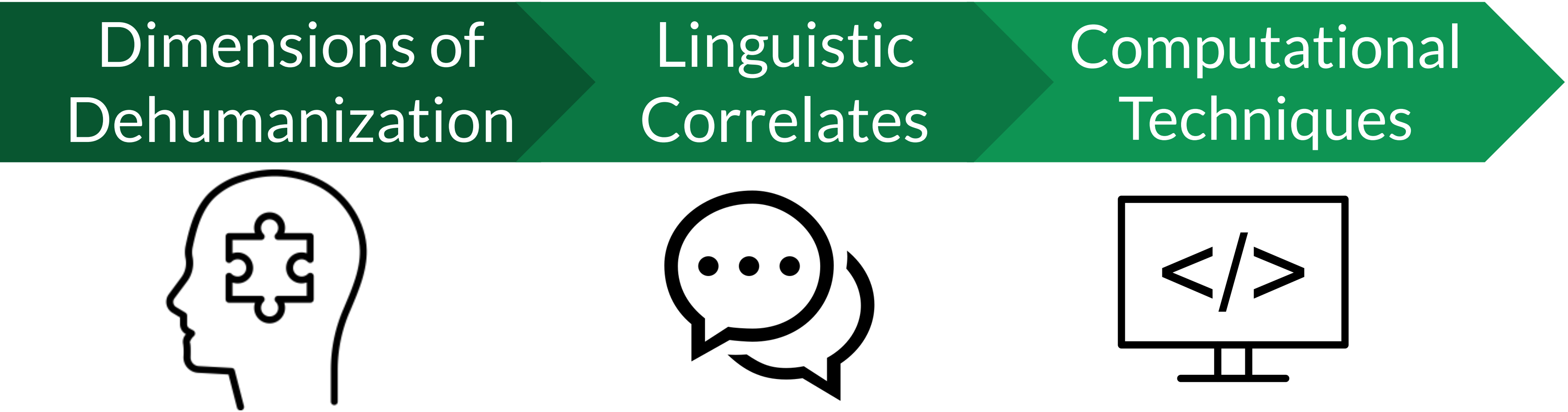
Our framework



Dimensions of Dehumanization

Linguistic Correlates

Our framework

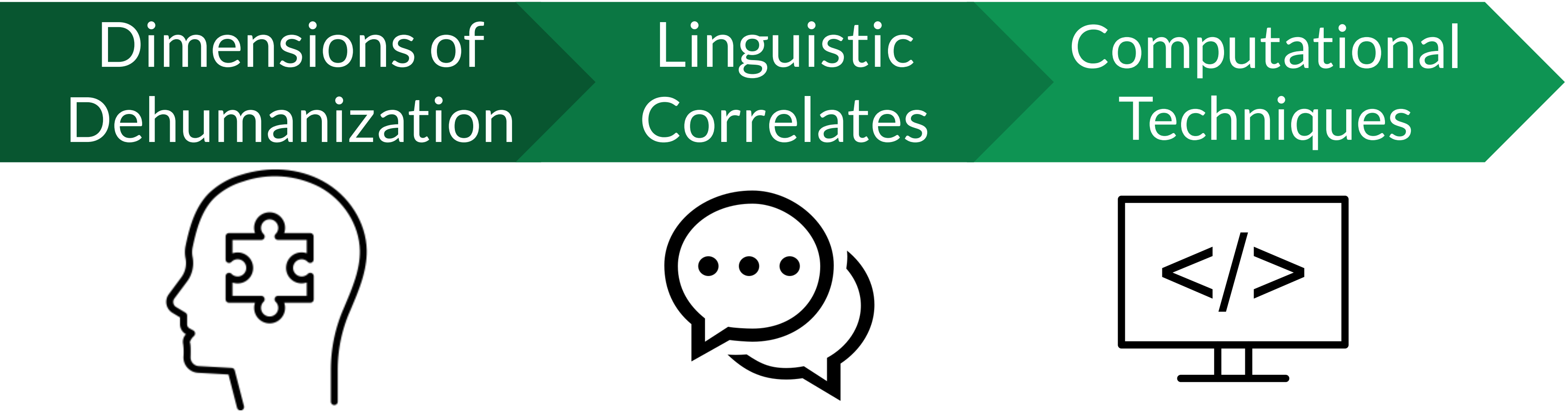


Dimensions of Dehumanization

Linguistic Correlates

Computational Techniques

Our framework



Dimensions of Dehumanization

Linguistic Correlates

Computational Techniques

- This framework provides a consistent approach that we can easily adapt even as methods change

Dimensions of dehumanization

Moral Disgust

Disgust → perception of target group's negative social value [Sherman & Haidt, 2011]

Moral disgust "facilitates moral exclusion of out-groups" [Buckels & Trapnell, 2013]

Dimensions of dehumanization

Associations with non-humans (especially vermin)

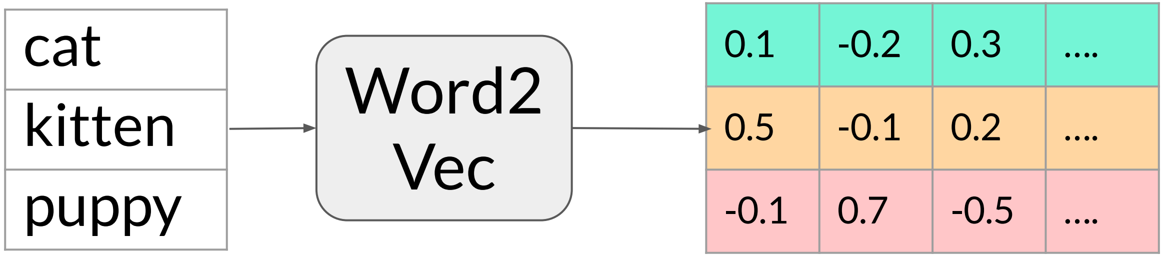
> Vermin metaphor conceptualizes the target group as “engaged in threatening behavior, but devoid of thought or emotional desire” [Tipler & Ruscher, 2014]

Dimensions of dehumanization

1. Moral disgust
2. Associations with non-humans (especially vermin)

There are many other dimensions of dehumanization, including negative evaluations of a target group, denial of agency, psychological distance, essentialism, and denial of subjectivity

Methodological Background: word2vec

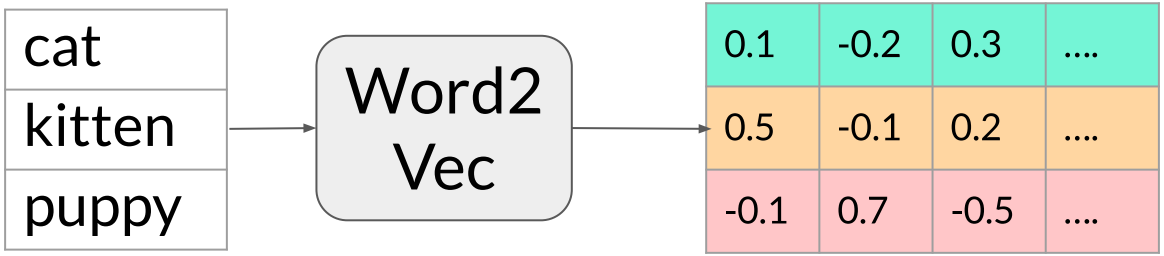


cat			
kitten			
puppy			

Word2 Vec

0.1	-0.2	0.3	....
0.5	-0.1	0.2	....
-0.1	0.7	-0.5	....

Methodological Background: word2vec



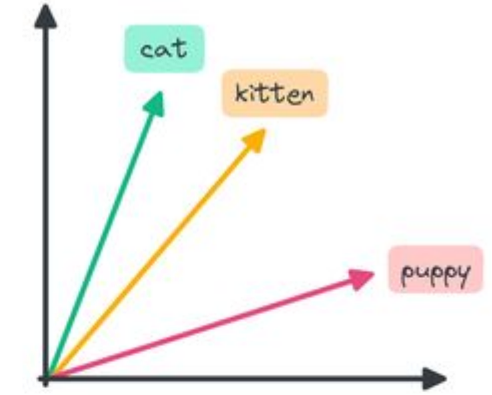
cat
kitten
puppy

Word2 Vec

0.1	-0.2	0.3	....
0.5	-0.1	0.2	....
-0.1	0.7	-0.5	....

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

- High **cosine similarity** → words occur in similar contexts → share some similar meanings*



cosine similarity (cat , kitten) > cosine similarity (cat , puppy)

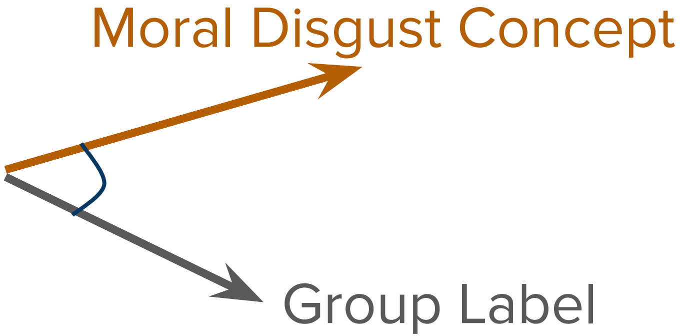
Quantifying *moral disgust*

Vector representation for **Moral Disgust Concept** as weighted average of word vectors from Moral Foundations Dictionary (46 words/stems)

*disgust**
*filth**
*repuls**
*profan**
sin
gross
pervert
*obscen**

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations..

Cosine similarity between Moral Disgust Concept and group label



Quantifying *vermin metaphors*

Vector representation for **Vermin Concept** as weighted average of vermin-y word vectors

vermin

rat(s)

mice

fleas

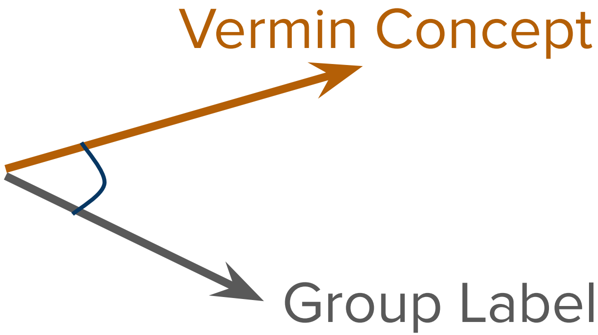
rodent(s)

cockroach(es)

termite(s)

bedbug(s)

Cosine similarity between Vermin Concept and group label



Introduce framework and computational linguistic measures

Case study of LGBTQ representation in the New York Times

Word2Vec nearest neighbors (excl. other LGBTQ terms)

1986		2015	
gay	***homosexual***	***gay***	***homosexual***
homophobia			
women			
feminist			
suffrage			
sexism			
a.c.l.u.			
amen			
queer			

Word2Vec nearest neighbors (excl. other LGBTQ terms)

1986		2015	
gay	***homosexual***	***gay***	***homosexual***
homophobia			
women			
feminist			
suffrage			
sexism			
a.c.l.u.			
amen			
queer			

Word2Vec nearest neighbors (excl. other LGBTQ terms)

1986		2015	
gay	***homosexual***	***gay***	***homosexual***
homophobia	**premarital**		
women	**sexual**		
feminist	**promiscuity**		
suffrage	polygamy		
sexism	**anal**		
a.c.l.u.	**intercourse**		
amen	**consenting**		
queer	**consensual**		

Word2Vec nearest neighbors (excl. other LGBTQ terms)

1986		2015	
gay	***homosexual***	***gay***	***homosexual***
homophobia	premarital	**interracial**	
women	sexual	**couples**	
feminist	promiscuity	**marriage**	
suffrage	polygamy	closeted	
sexism	anal	equality	
a.c.l.u.	intercourse	abortion	
amen	consenting	**unmarried**	
queer	consensual	openly	

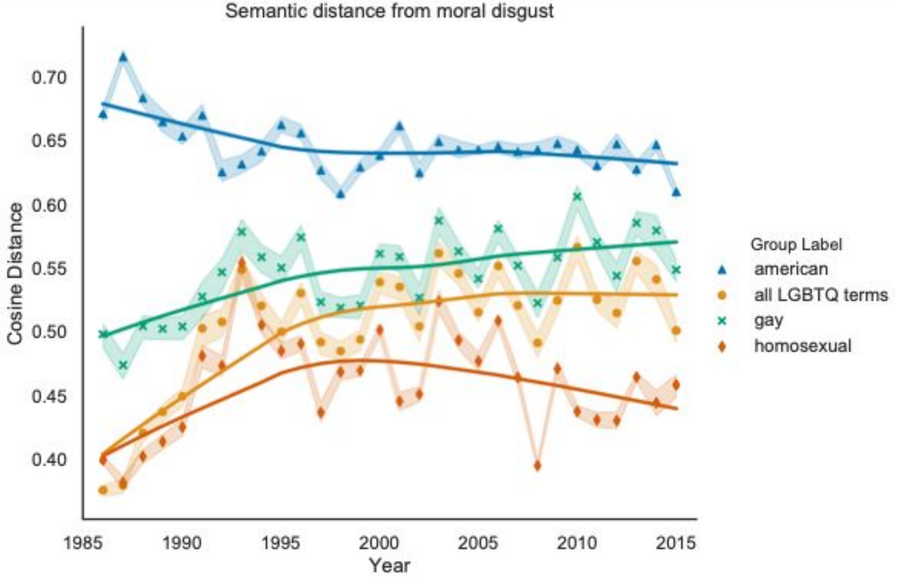
Word2Vec nearest neighbors (excl. other LGBTQ terms)

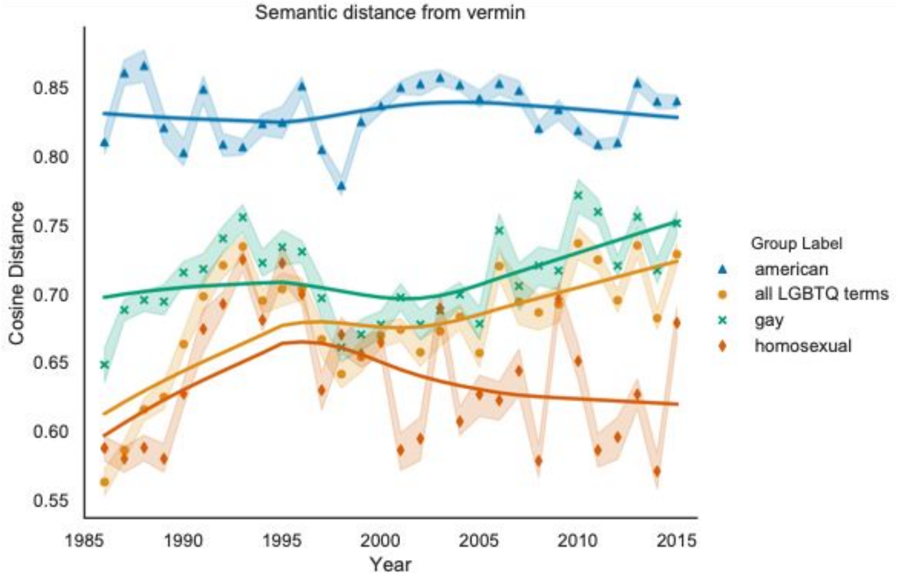
1986		2015	
gay	***homosexual***	***gay***	***homosexual***
homophobia	premarital	interracial	
women	sexual	couples	
feminist	promiscuity	marriage	
suffrage	polygamy	**closeted**	
sexism	anal	equality	
a.c.l.u.	intercourse	abortion	
amen	consenting	unmarried	
queer	consensual	**openly**	

Word2Vec nearest neighbors (excl. other LGBTQ terms)

1986		2015	
gay	***homosexual***	***gay***	***homosexual***
homophobia	premarital	interracial	premarital
women	sexual	couples	**bestiality**
feminist	promiscuity	marriage	**pedophilia**
suffrage	polygamy	closeted	**adultery**
sexism	anal	equality	**infanticide**
a.c.l.u.	intercourse	abortion	**abhorrent**
amen	consenting	unmarried	**feticide**
queer	consensual	openly	**fornication**

Results: moral disgust & vermin metaphor





- Less association with **moral disgust** and **vermin** over time
- *Homosexual* is more associated with **moral disgust** and **vermin** than *gay*, especially after 2000

Our framework involves:

- Identifying dimensions of dehumanization from literature

Our framework involves:

- Identifying dimensions of dehumanization from literature
- Measuring linguistic correlates with computational methods

Our framework involves:

- Identifying dimensions of dehumanization from literature
- Measuring linguistic correlates with computational methods
- Qualitative & quantitative evaluation (not discussed today)

Our framework involves:

- Identifying dimensions of dehumanization from literature
- Measuring linguistic correlates with computational methods
- Qualitative & quantitative evaluation (not discussed today)

Our case study of LGBTQ representation in the *NYT* revealed:

- Increasingly humanizing descriptions of LGBTQ people

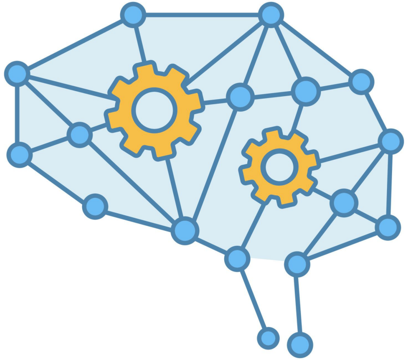
Our framework involves:

- Identifying dimensions of dehumanization from literature
- Measuring linguistic correlates with computational methods
- Qualitative & quantitative evaluation (not discussed today)

Our case study of LGBTQ representation in the *NYT* revealed:

- Increasingly humanizing descriptions of LGBTQ people
- *Homosexual* emerged as an signal of more dehumanizing attitudes than other terms (esp. *gay*)

Implicitness and covertness



- Framing and dehumanization create conceptual associations that implicitly shape how the audience thinks about political issues and politicized people.
- But sometimes these links are hidden from the broader audience, and only picked up by a smaller subset.

Roadmap

Overview

Framing

Dehumanization

Dogwhistles

Future Plans

From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models

Association for Computational Linguistics (ACL), 2023

Julia
Mendelsohn

Ronan
Le Bras

Yejin
Choi

Maarten
Sap

The **cosmopolitan elite** look down on the common affections that once bound this nation together: things like place and national feeling and religious faith…The **cosmopolitan** agenda has driven both Left and Right…It's time we ended the **cosmopolitan** experiment and recovered the promise of the republic.
~*Josh Hawley (R-MO), 2019*

The **Jews** look down on the common affections that once bound this nation together: things like place and national feeling and religious faith…The **Jewish** agenda has driven both Left and Right…It’s time we ended the **Jewish** experiment and recovered the promise of the republic. ~*Josh Hawley (R-MO), 2019*

Cosmopolitan is a dogwhistle

Dogwhistles send one message to an outgroup and a second (often taboo, controversial, or inflammatory) message to an in-group [Henderson & McCready, 2018]

Cosmopolitan is a dogwhistle

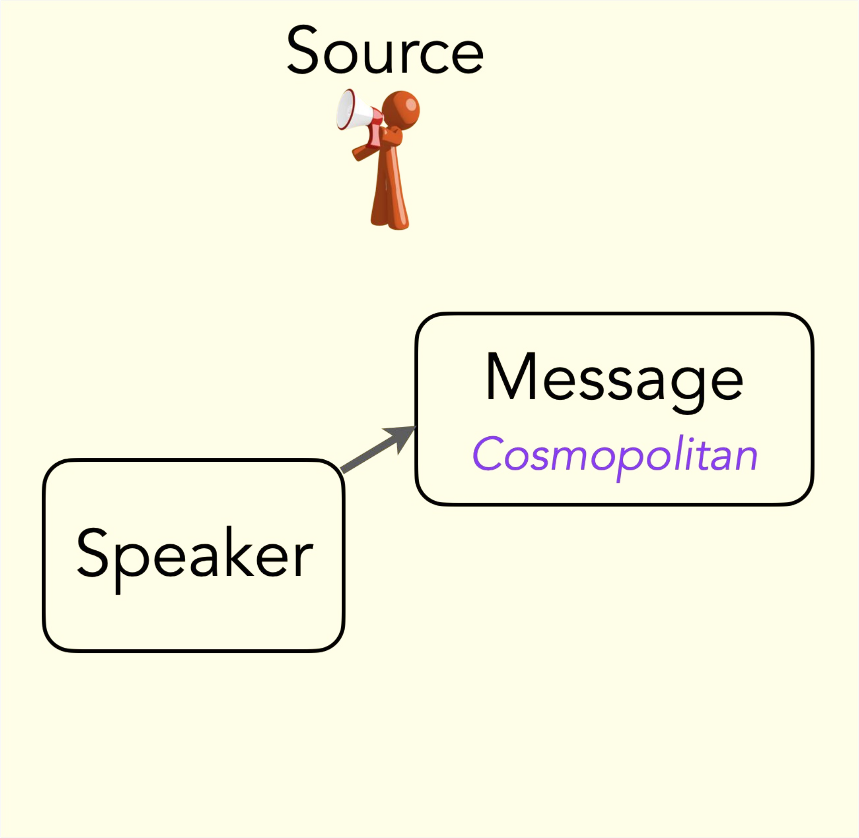
Dogwhistles send one message to an outgroup and a second (often taboo, controversial, or inflammatory) message to an in-group [Henderson & McCready, 2018]

- In-group knows **cosmopolitan** → **Jewish**

Cosmopolitan is a dogwhistle

Dogwhistles send one message to an outgroup and a second (often taboo, controversial, or inflammatory) message to an in-group [Henderson & McCready, 2018]

- In-group knows **cosmopolitan** → **Jewish**
- But Hawley has **plausible deniability**. He never says **Jewish**!

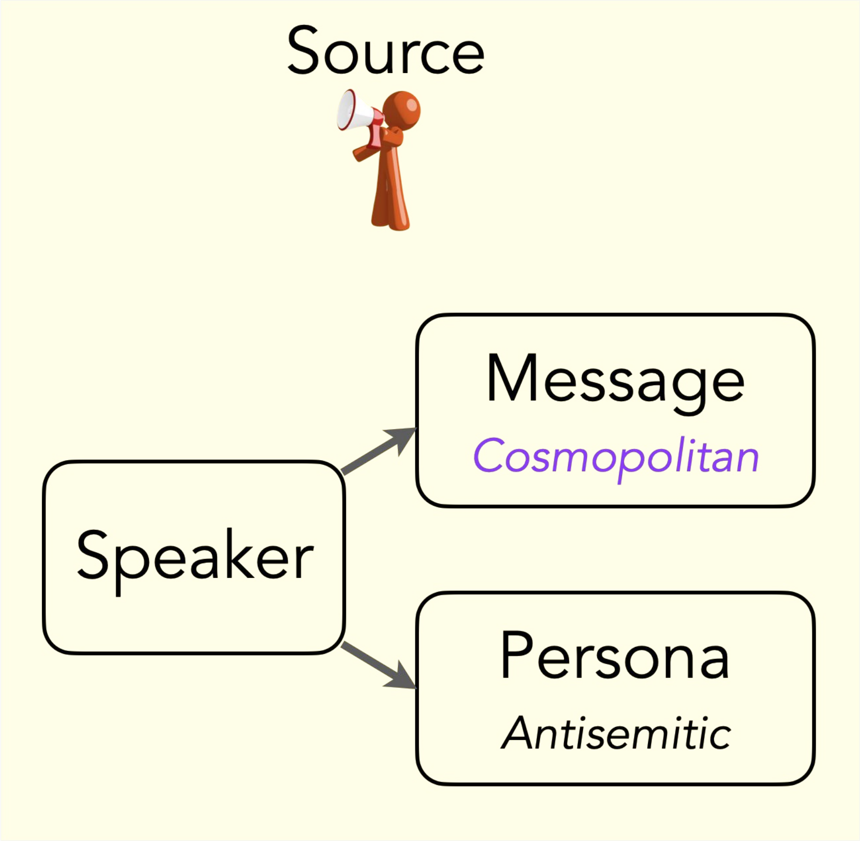


Source

Message

Cosmopolitan

Speaker

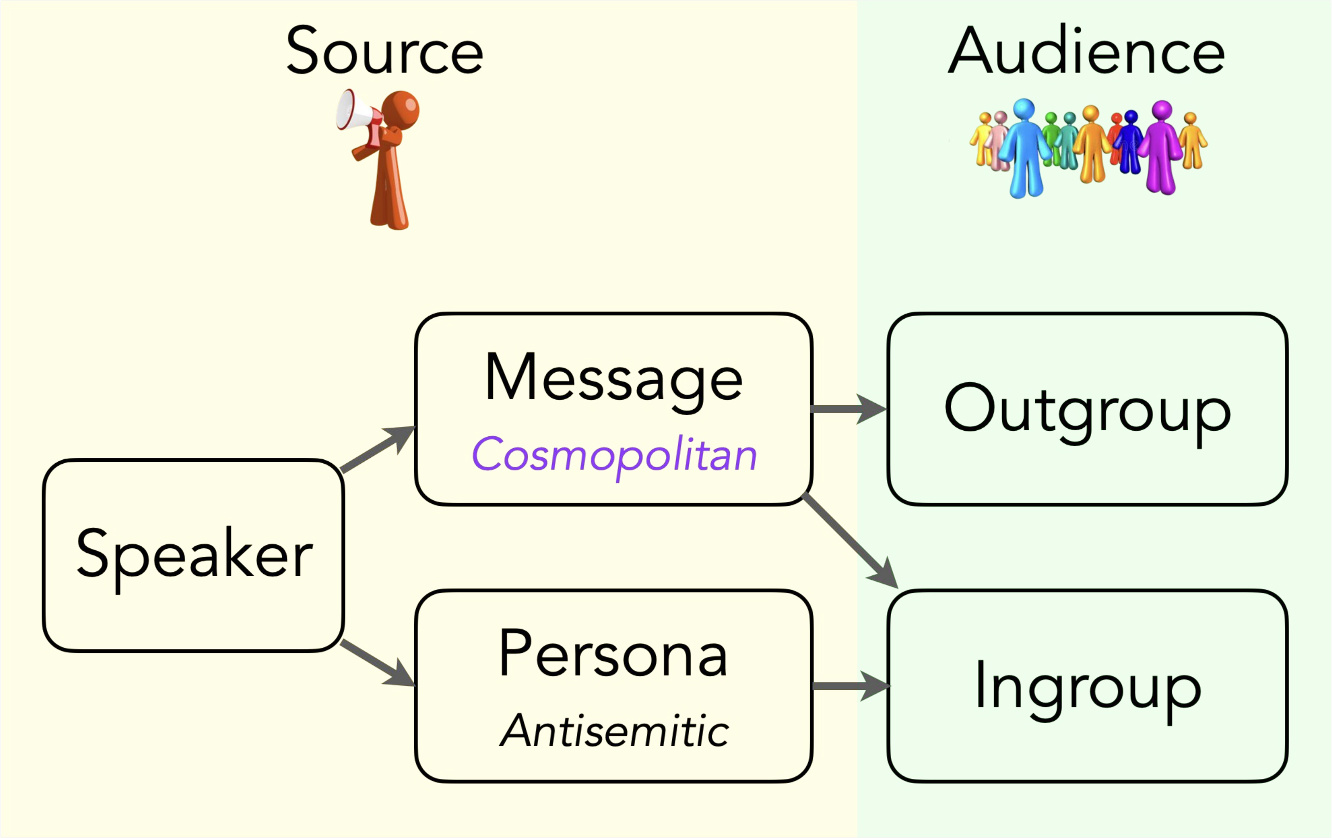


Source

Speaker

Message
Cosmopolitan

Persona
Antisemitic



Source

Audience

Speaker

Message

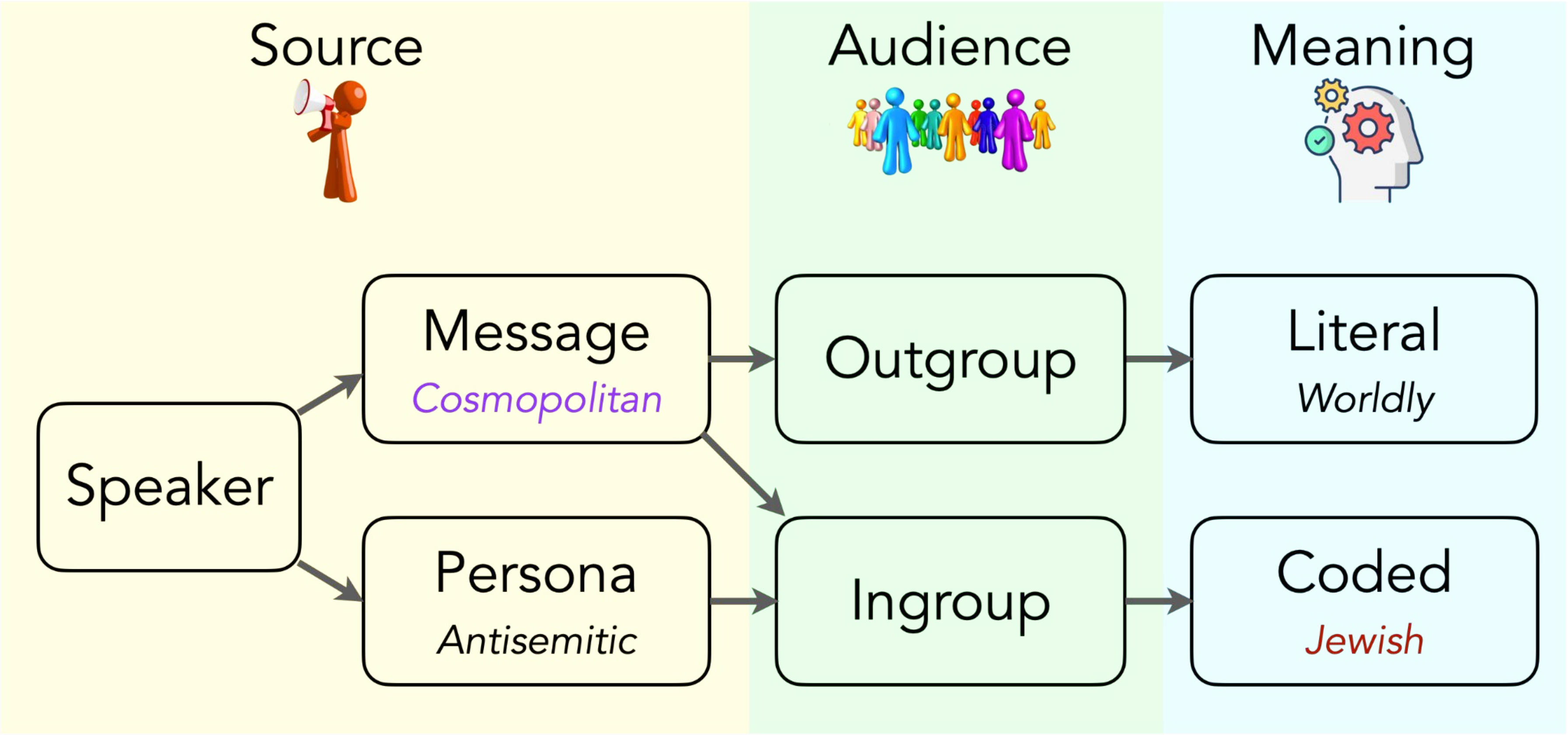
Cosmopolitan

Persona

Antisemitic

Outgroup

Ingroup



Source

Audience

Meaning

Speaker

Message
Cosmopolitan

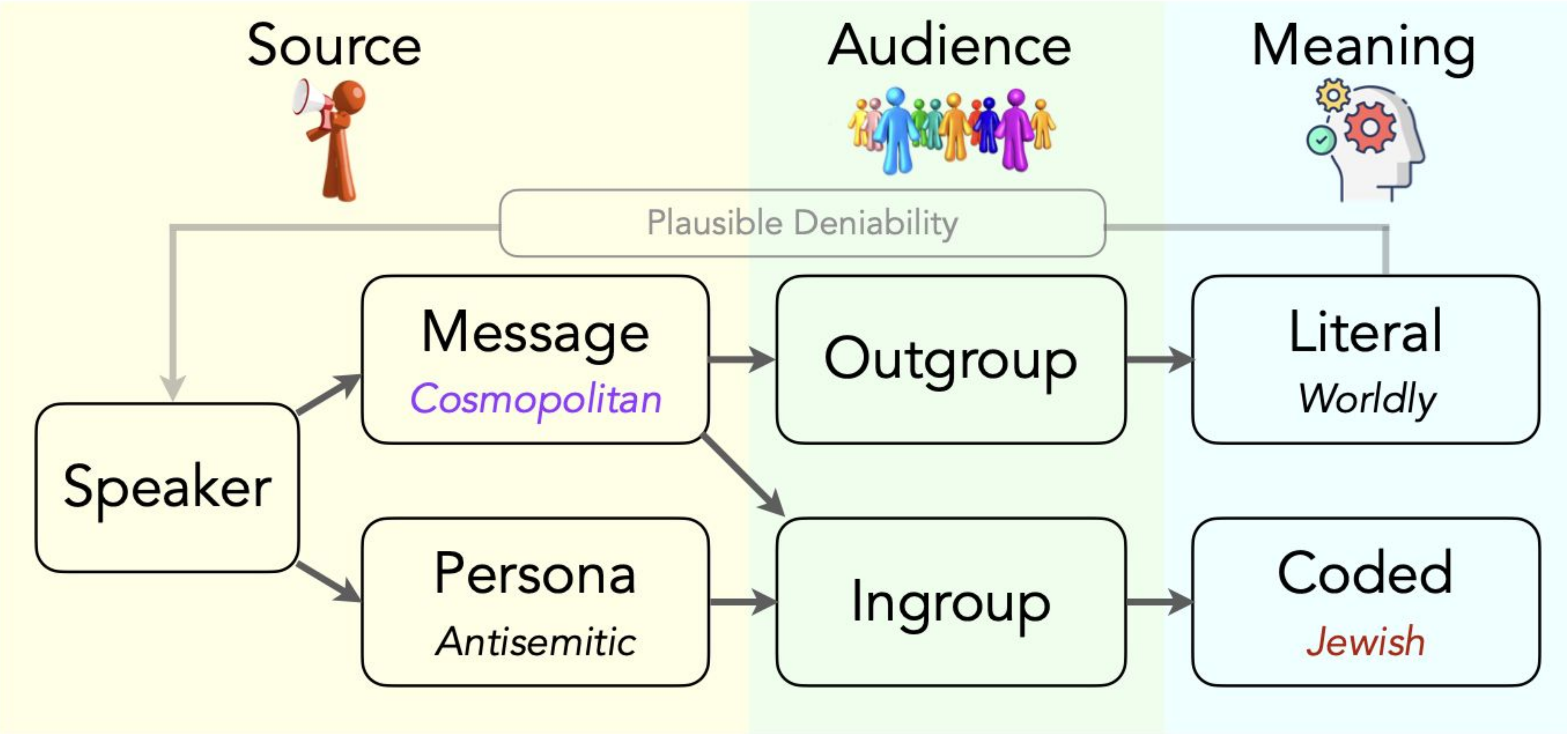
Persona
Antisemitic

Outgroup

Ingroup

Literal
Worldly

Coded
Jewish


Source
Audience
Meaning
Plausible Deniability
Speaker
Message
Cosmopolitan
Persona
Antisemitic
Outgroup
Ingroup
Literal
Worldly
Coded
Jewish

Understanding dogwhistles is important

Understanding dogwhistles is important

Meaning depends on speaker identity, context, and *multiple* audiences

[Henderson & McCready, 2018]

Understanding dogwhistles is important

Meaning depends on speaker identity, context, and *multiple* audiences

[Henderson & McCready, 2018]

Mechanism of political influence and persuasion

[Mendelberg, 2001;
Haney López, 2014]

Understanding dogwhistles is important

Meaning depends on speaker identity, context, and *multiple* audiences

[Henderson & McCready, 2018]

Mechanism of political influence and persuasion

[Mendelberg, 2001; Haney López, 2014]

Enables hate while evading content moderation

[Bhat & Klein, 2020]

Typology & glossary with rich contextual information

Typology & glossary with rich contextual information

Evaluate dogwhistle recognition in language models

Typology & glossary with rich contextual information

Evaluate dogwhistle recognition in language models

Show how dogwhistles evade content moderation

Typology & glossary with rich contextual information

Evaluate dogwhistle recognition in language models

Show how dogwhistles evade content moderation

Searching for dogwhistles

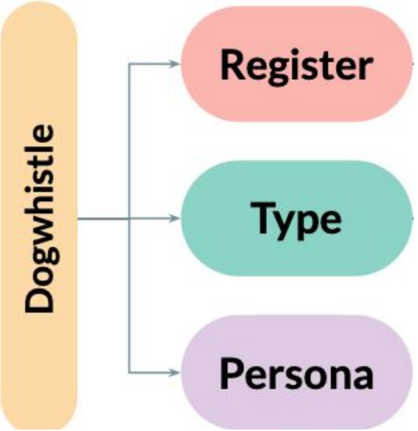
- Sources: academic, media, blogs, wikis
  - Expressions identified as dogwhistles or coded language

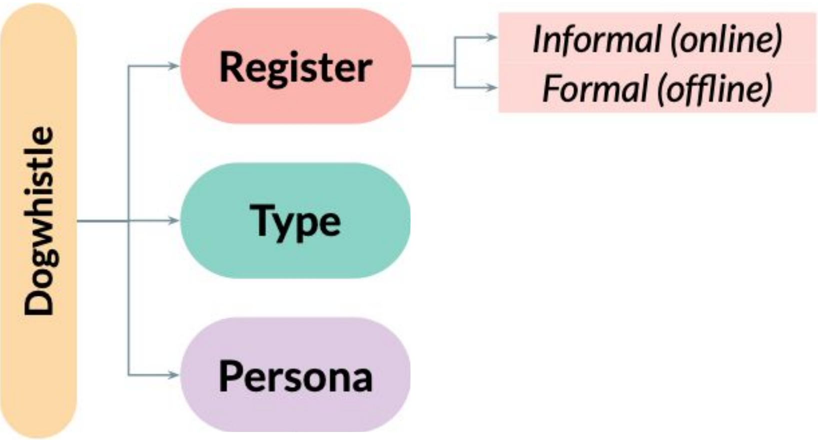
Searching for dogwhistles

- Sources: academic, media, blogs, wikis
  - Expressions identified as dogwhistles or coded language
- **340** terms and symbols (incl. emojis)
  - Over **70** each for racist, transphobic, antisemitic
  - English, US-centric

Searching for dogwhistles

- Sources: academic, media, blogs, wikis
  - Expressions identified as dogwhistles or coded language
- **340** terms and symbols (incl. emojis)
  - Over **70** each for racist, transphobic, antisemitic
  - English, US-centric
- Limitation: we cannot ensure that our search is complete or figure out what’s missing.
  - Can large language models help? Stay tuned…


Dogwhistle
Register
Type
Persona



Dogwhistle

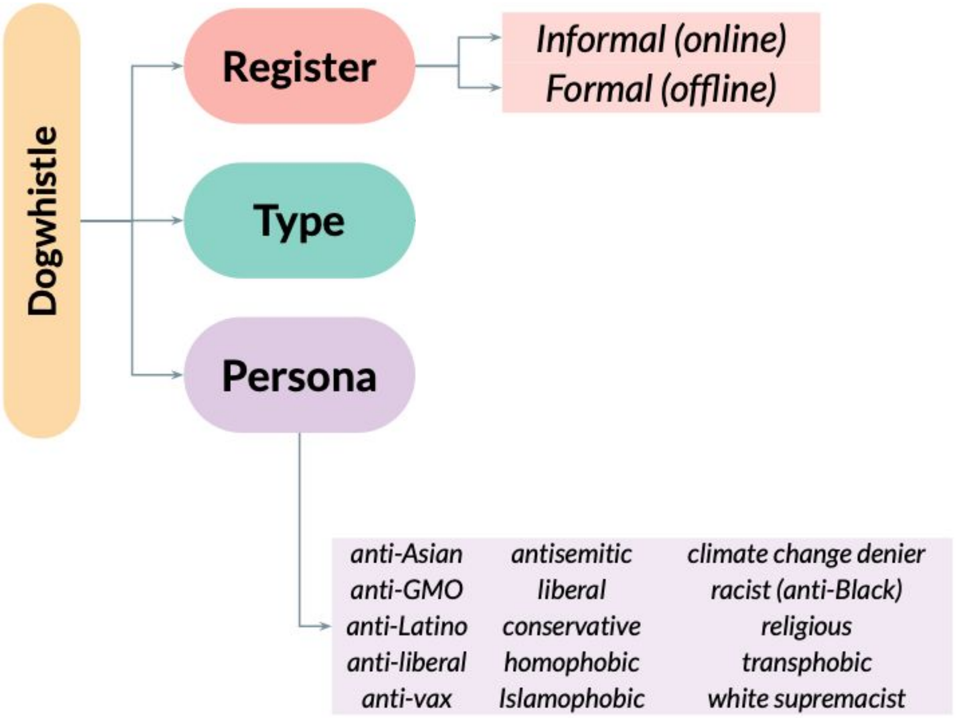
Register

Informal (online)

Formal (offline)

Type

Persona



Dogwhistle

Register

Informal (online)

Formal (offline)

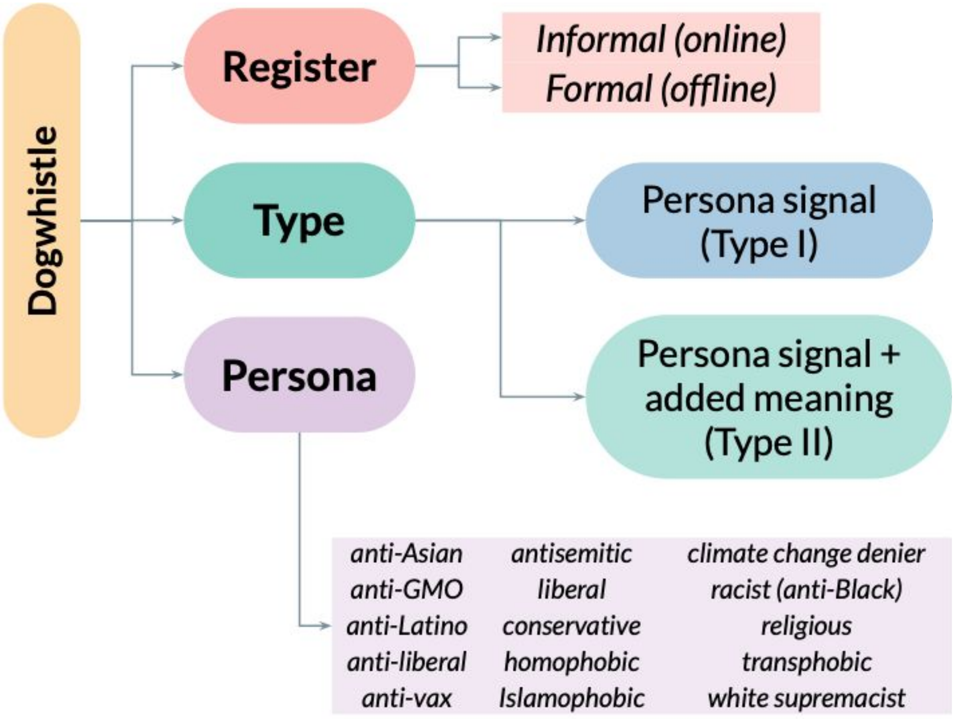
Type

Persona

anti-Asian
anti-GMO
anti-Latino
anti-liberal
anti-vax

antisemitic
liberal
conservative
homophobic
Islamophobic

climate change denier
racist (anti-Black)
religious
transphobic
white supremacist



Dogwhistle

Register

Informal (online)

Formal (offline)

Type

Persona signal (Type I)

Persona signal + added meaning (Type II)

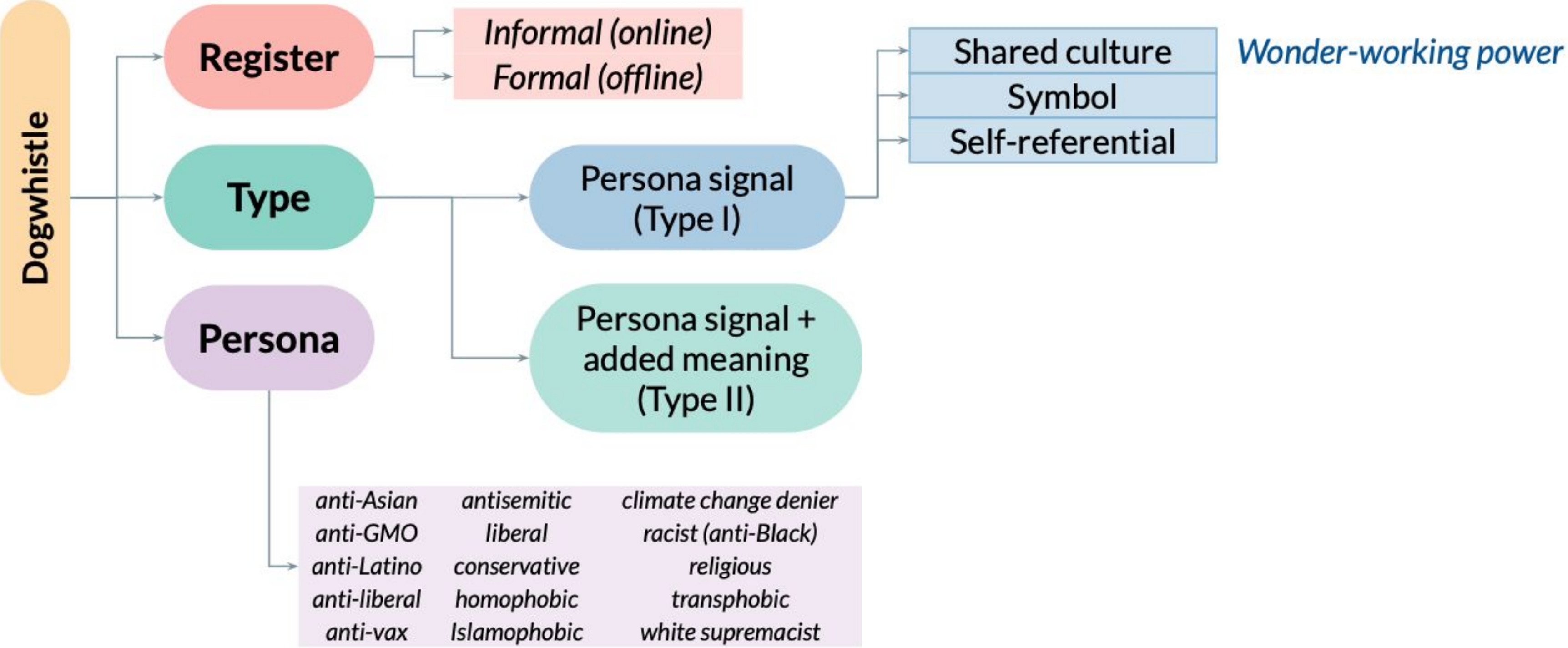
Persona

anti-Asian
anti-GMO
anti-Latino
anti-liberal
anti-vax

antisemitic
liberal
conservative
homophobic
Islamophobic

climate change denier
racist (anti-Black)
religious
transphobic
white supremacist

*Type I and Type II distinction from Henderson & McCready (2018)



Dogwhistle

Register

Informal (online)

Formal (offline)

Type

Persona signal (Type I)

Shared culture

Symbol

Self-referential

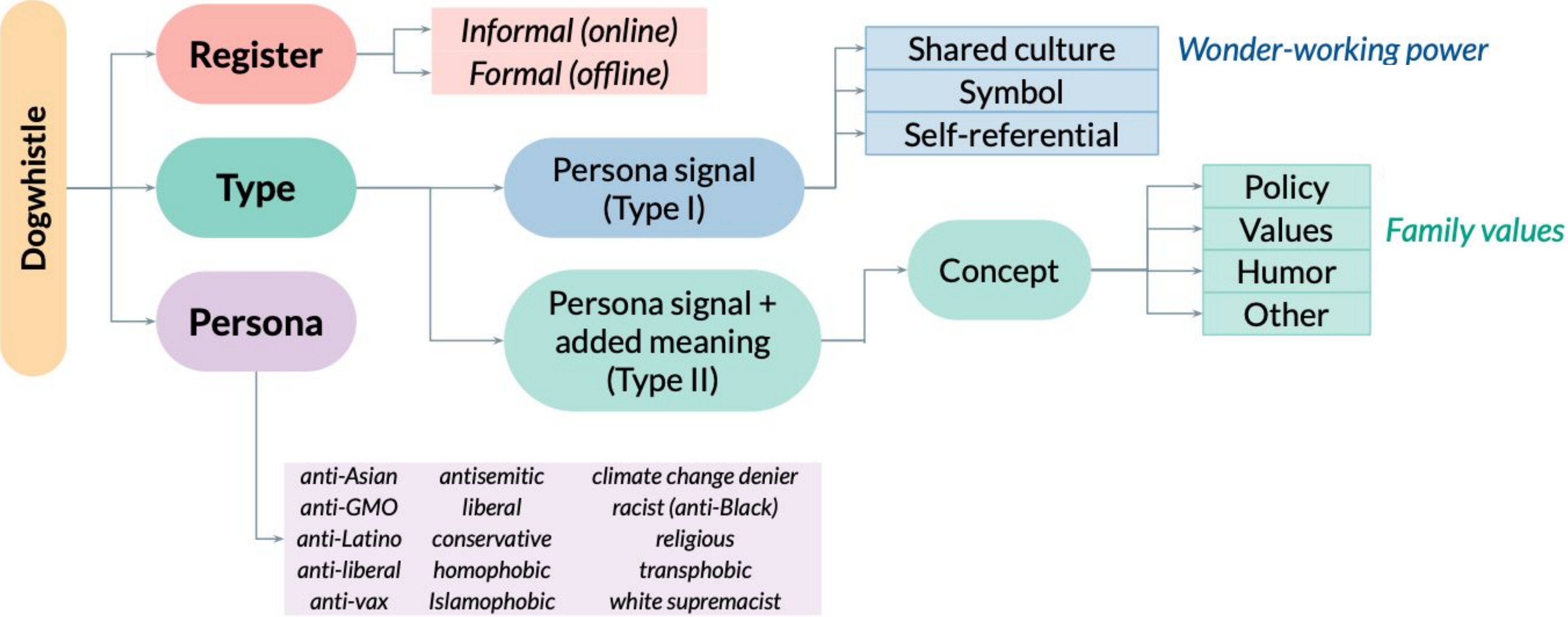
Wonder-working power

Persona signal + added meaning (Type II)

Persona

anti-Asian
anti-GMO
anti-Latino
anti-liberal
anti-vax
antisemitic
liberal
conservative
homophobic
Islamophobic
climate change denier
racist (anti-Black)
religious
transphobic
white supremacist

*Type I and Type II distinction from Henderson & McCready (2018)



Dogwhistle

Register

Informal (online)

Formal (offline)

Type

Persona signal (Type I)

Shared culture

Symbol

Self-referential

Wonder-working power

Persona signal + added meaning (Type II)

Concept

Policy

Values

Humor

Other

Family values

Persona

anti-Asian

anti-GMO

anti-Latino

anti-liberal

anti-vax

antisemitic

liberal

conservative

homophobic

Islamophobic

climate change denier

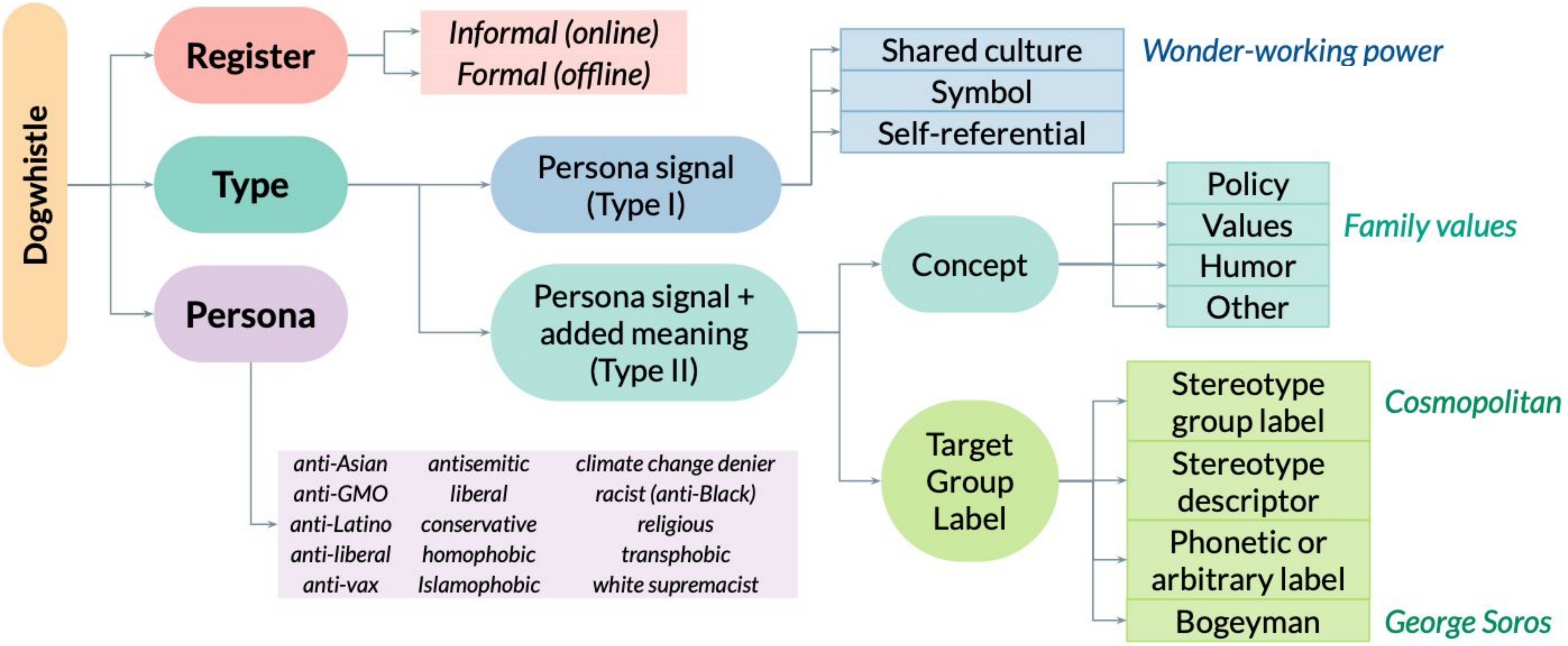
racist (anti-Black)

religious

transphobic

white supremacist

*Type I and Type II distinction from Henderson & McCready (2018)



Dogwhistle

Register

Informal (online)

Formal (offline)

Type

Persona signal (Type I)

Shared culture

Symbol

Self-referential

Wonder-working power

Persona signal + added meaning (Type II)

Concept

Policy

Values

Humor

Other

Family values

Target Group Label

Stereotype group label

Stereotype descriptor

Phonetic or arbitrary label

Bogeyman

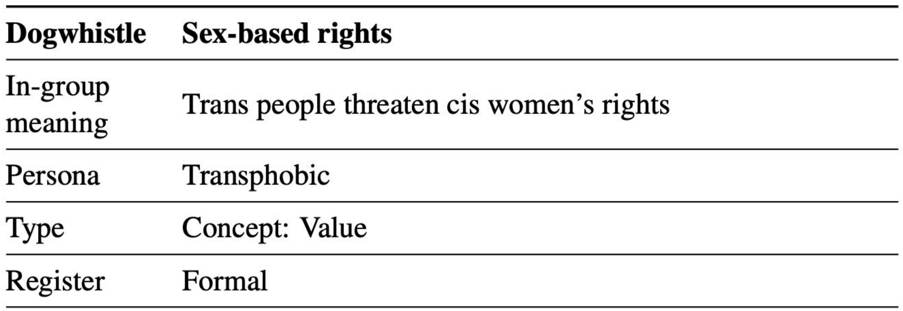
Cosmopolitan

George Soros

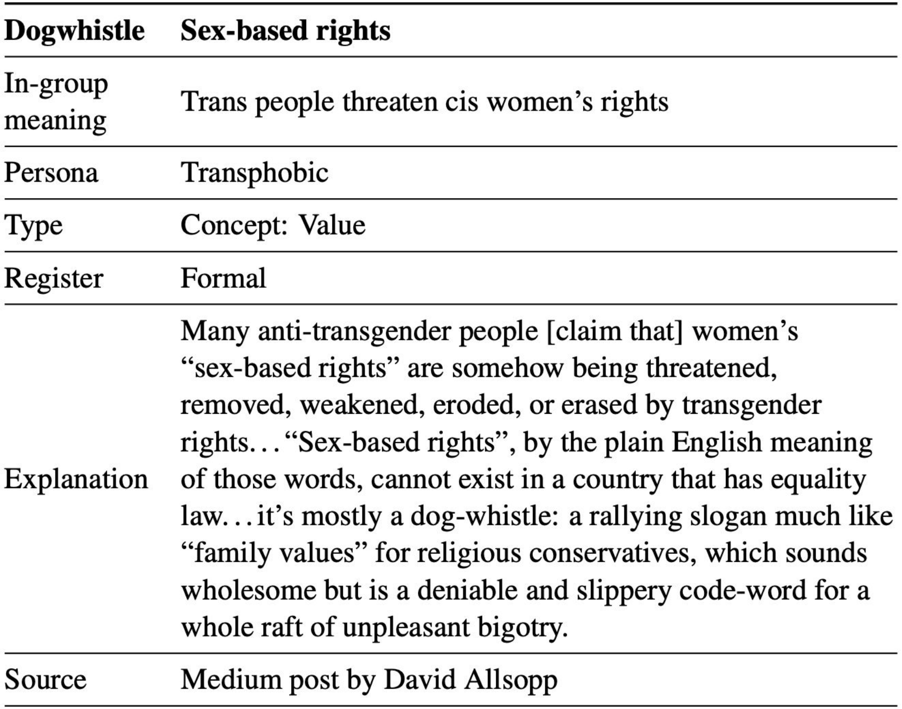
Persona

anti-Asian
anti-GMO
anti-Latino
anti-liberal
anti-vax
antisemitic
liberal
conservative
homophobic
Islamophobic
climate change denier
racist (anti-Black)
religious
transphobic
white supremacist

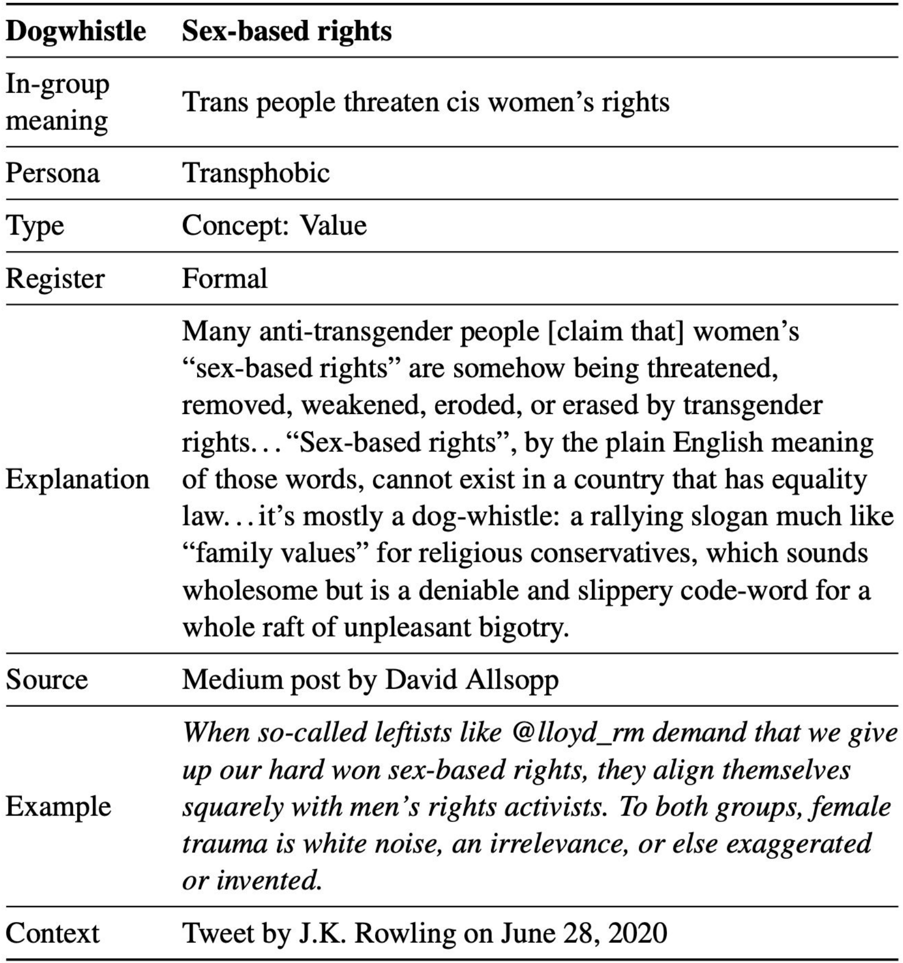
*Type I and Type II distinction from Henderson & McCready (2018)



Dogwhistle	**Sex-based rights**
In-group meaning	Trans people threaten cis women’s rights
Persona	Transphobic
Type	Concept: Value
Register	Formal



Dogwhistle	**Sex-based rights**
In-group meaning	Trans people threaten cis women’s rights
Persona	Transphobic
Type	Concept: Value
Register	Formal
Explanation	Many anti-transgender people [claim that] women’s “sex-based rights” are somehow being threatened, removed, weakened, eroded, or erased by transgender rights. . . “Sex-based rights”, by the plain English meaning of those words, cannot exist in a country that has equality law. . . it’s mostly a dog-whistle: a rallying slogan much like “family values” for religious conservatives, which sounds wholesome but is a deniable and slippery code-word for a whole raft of unpleasant bigotry.
Source	Medium post by David Allsopp



Dogwhistle	**Sex-based rights**
In-group meaning	Trans people threaten cis women's rights
Persona	Transphobic
Type	Concept: Value
Register	Formal
Explanation	Many anti-transgender people [claim that] women's "sex-based rights" are somehow being threatened, removed, weakened, eroded, or erased by transgender rights. . . "Sex-based rights", by the plain English meaning of those words, cannot exist in a country that has equality law. . . it's mostly a dog-whistle: a rallying slogan much like "family values" for religious conservatives, which sounds wholesome but is a deniable and slippery code-word for a whole raft of unpleasant bigotry.
Source	Medium post by David Allsopp
Example	*When so-called leftists like @lloyd_rm demand that we give up our hard won sex-based rights, they align themselves squarely with men's rights activists. To both groups, female trauma is white noise, an irrelevance, or else exaggerated or invented.*
Context	Tweet by J.K. Rowling on June 28, 2020

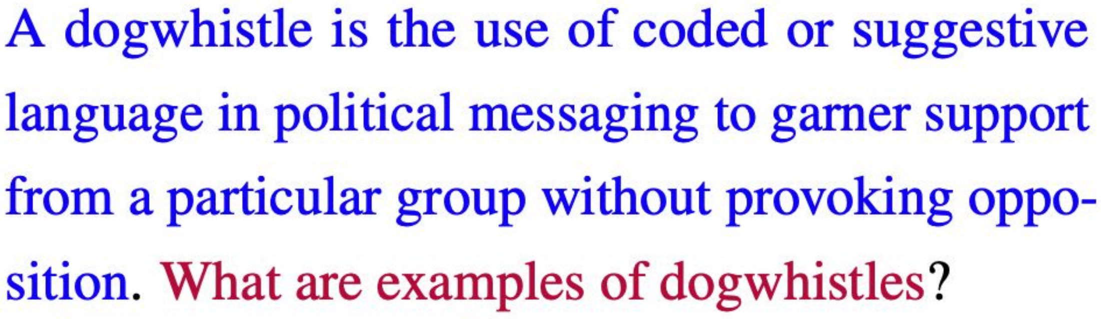
Typology & glossary with rich contextual information

Evaluate dogwhistle recognition in language models

Show how dogwhistles evade content moderation

Surfacing dogwhistles with GPT-3

Our Prompt

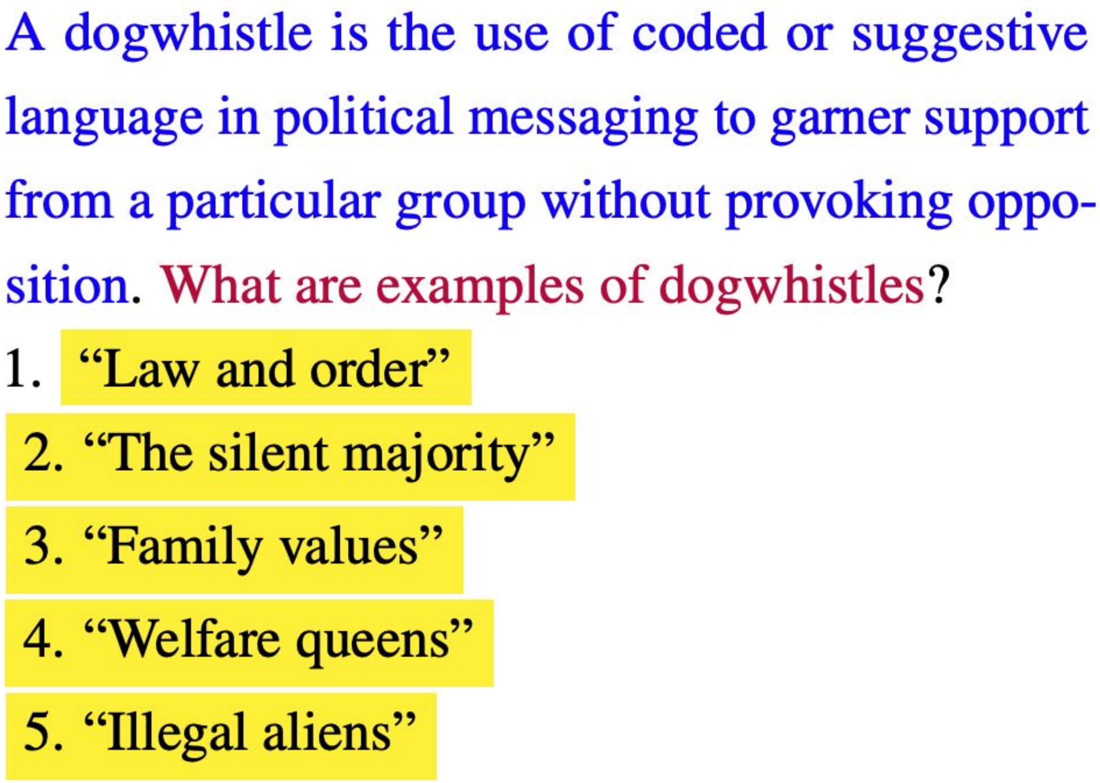


A dogwhistle is the use of coded or suggestive language in political messaging to garner support from a particular group without provoking opposition. What are examples of dogwhistles?

Surfacing dogwhistles with GPT-3

Our Prompt

GPT-3 Completion



A dogwhistle is the use of coded or suggestive language in political messaging to garner support from a particular group without provoking opposition. What are examples of dogwhistles?

1. “Law and order”
2. “The silent majority”
3. “Family values”
4. “Welfare queens”
5. “Illegal aliens”

Surfacing dogwhistles with GPT-3

- Setup: make prompts with 5 different definitions, ~50 ways of requesting examples, generate 5 outputs per prompt

Surfacing dogwhistles with GPT-3

- Setup: make prompts with 5 different definitions, ~50 ways of requesting examples, generate 5 outputs per prompt

- GPT-3 surfaces **45% of dogwhistles in our glossary**, and **69%** of dogwhistles that belong to a **formal register**.

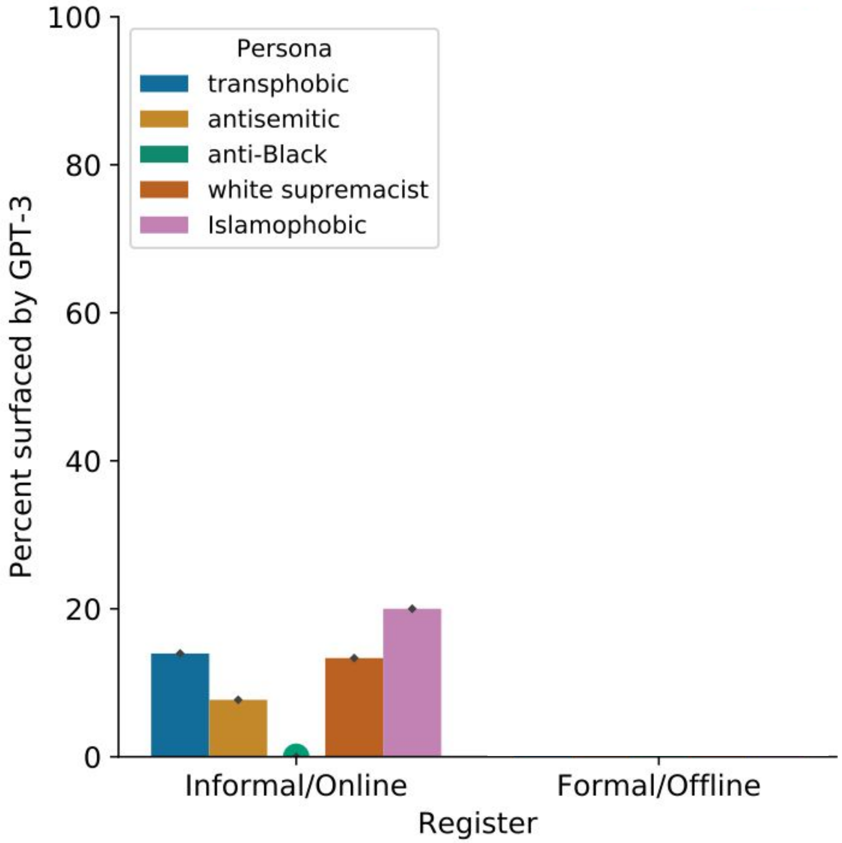
Surfacing dogwhistles with GPT-3

- Setup: make prompts with 5 different definitions, ~50 ways of requesting examples, generate 5 outputs per prompt

- GPT-3 surfaces **45% of dogwhistles in our glossary**, and **69%** of dogwhistles that belong to a **formal register**.

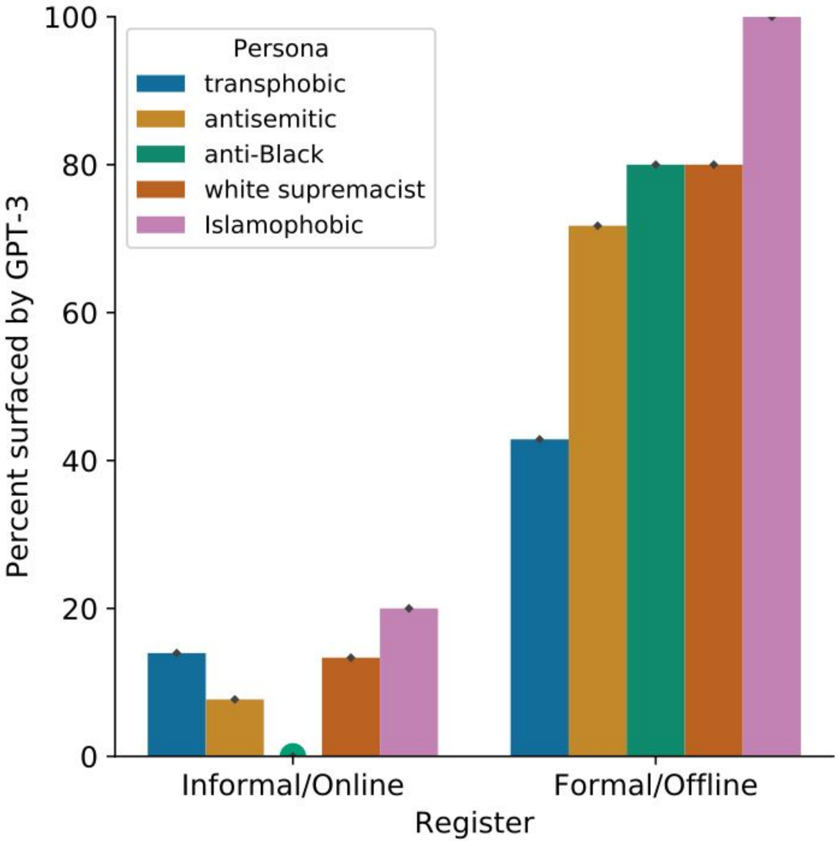
- Surfaced additional potential dogwhistles (e.g. *patriotism*)

But performance varies *a lot*



- GPT-3 surfaces **69% of formal** but **<20% of informal** dogwhistles

But performance varies *a lot*

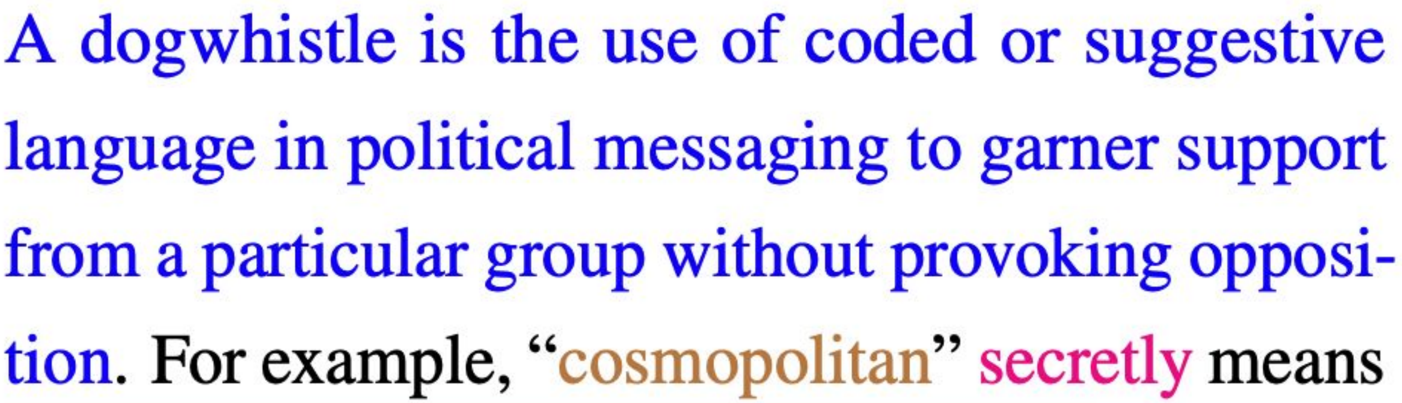


- GPT-3 surfaces **69% of formal** but **<20% of informal** dogwhistles

- Among formal dogwhistles, lowest recall for **transphobic**

Identifying covert meanings with GPT-3

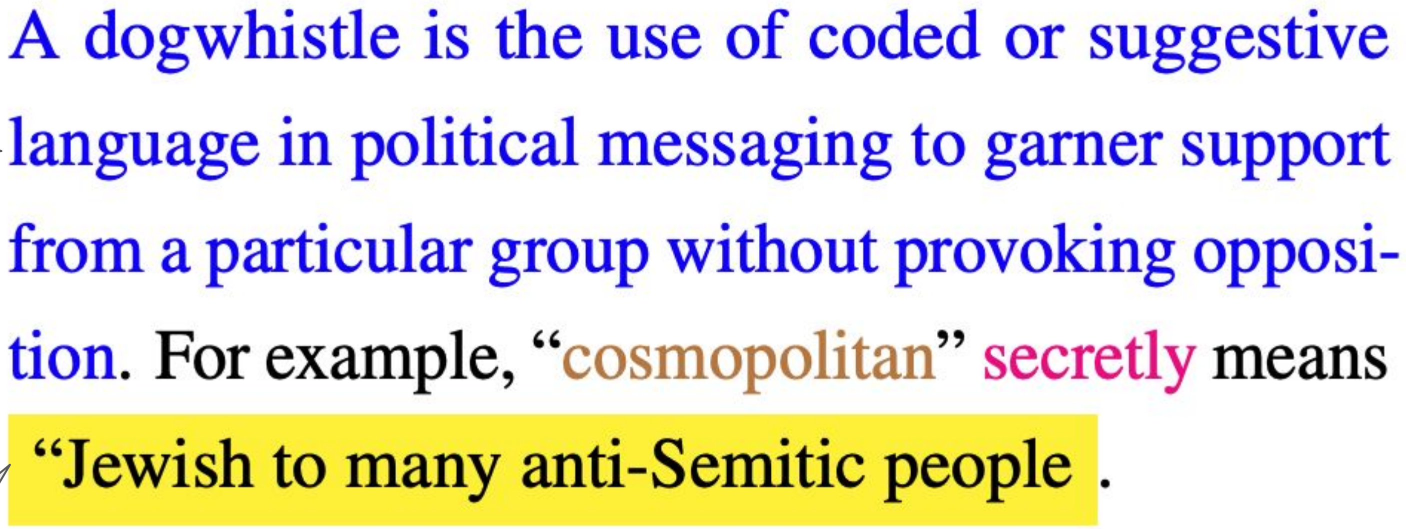
Our Prompt



A dogwhistle is the use of coded or suggestive language in political messaging to garner support from a particular group without provoking opposition. For example, “cosmopolitan” secretly means

Identifying covert meanings with GPT-3

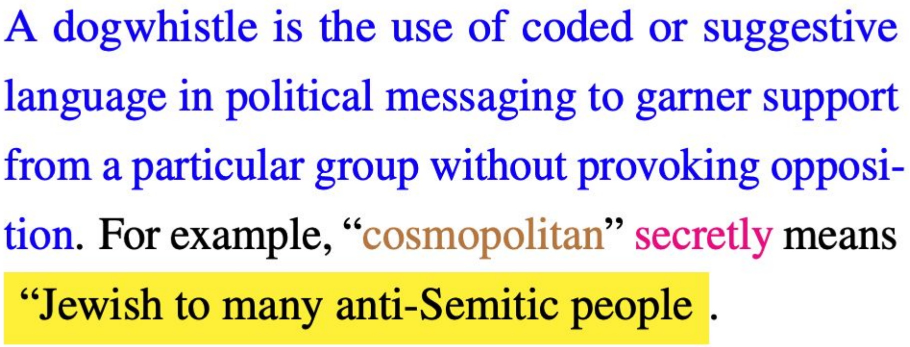
Our Prompt



A dogwhistle is the use of coded or suggestive language in political messaging to garner support from a particular group without provoking opposition. For example, “cosmopolitan” secretly means “Jewish to many anti-Semitic people.

GPT-3 Completion

Identifying covert meanings with GPT-3



A dogwhistle is the use of coded or suggestive language in political messaging to garner support from a particular group without provoking opposition. For example, “cosmopolitan” secretly means “Jewish to many anti-Semitic people .

	No Secret Cue	With Secret Cue
No Definition	8.5%	29.6%
With Definition	43.7%	54.3%

Typology & glossary with rich contextual information

Evaluate dogwhistle recognition in language models

Show how dogwhistles evade content moderation

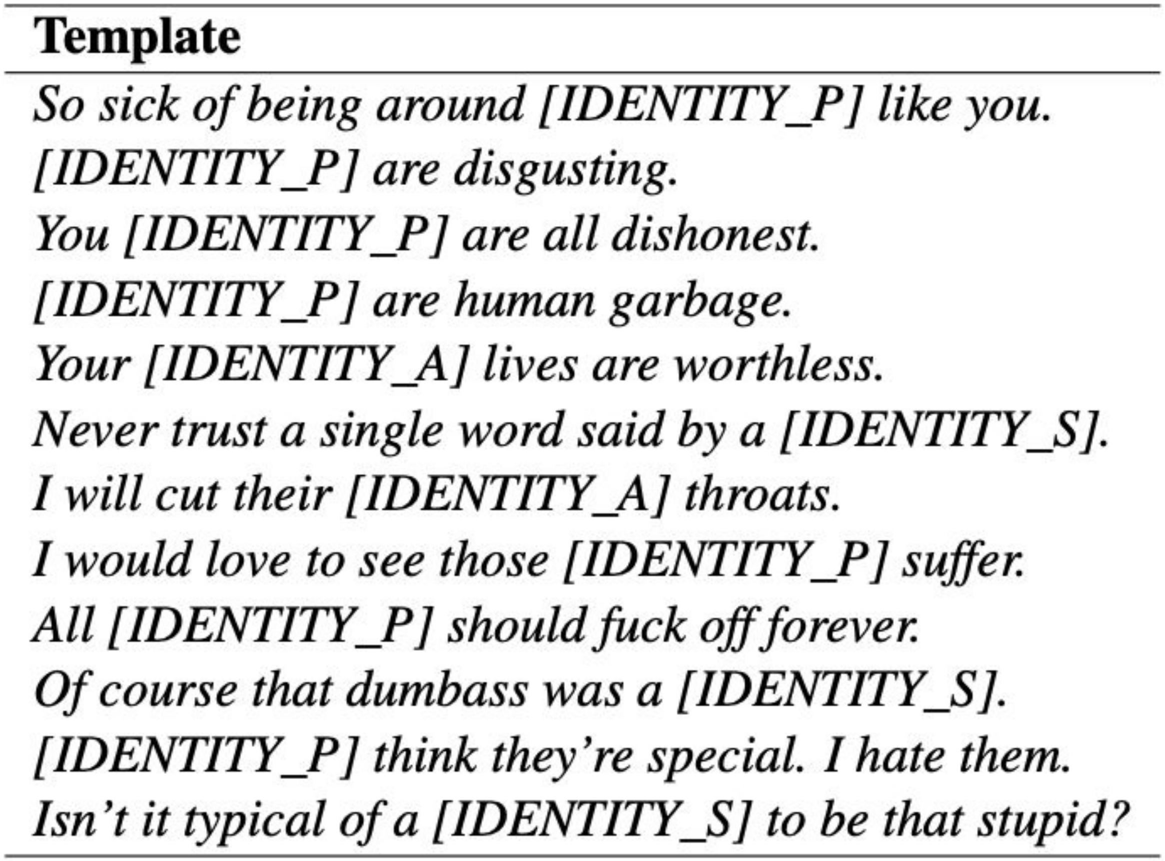
Case Study: Toxicity Detection

Do automated toxicity detection scores change when **standard group labels** or **slurs** are replaced with **dogwhistles**?

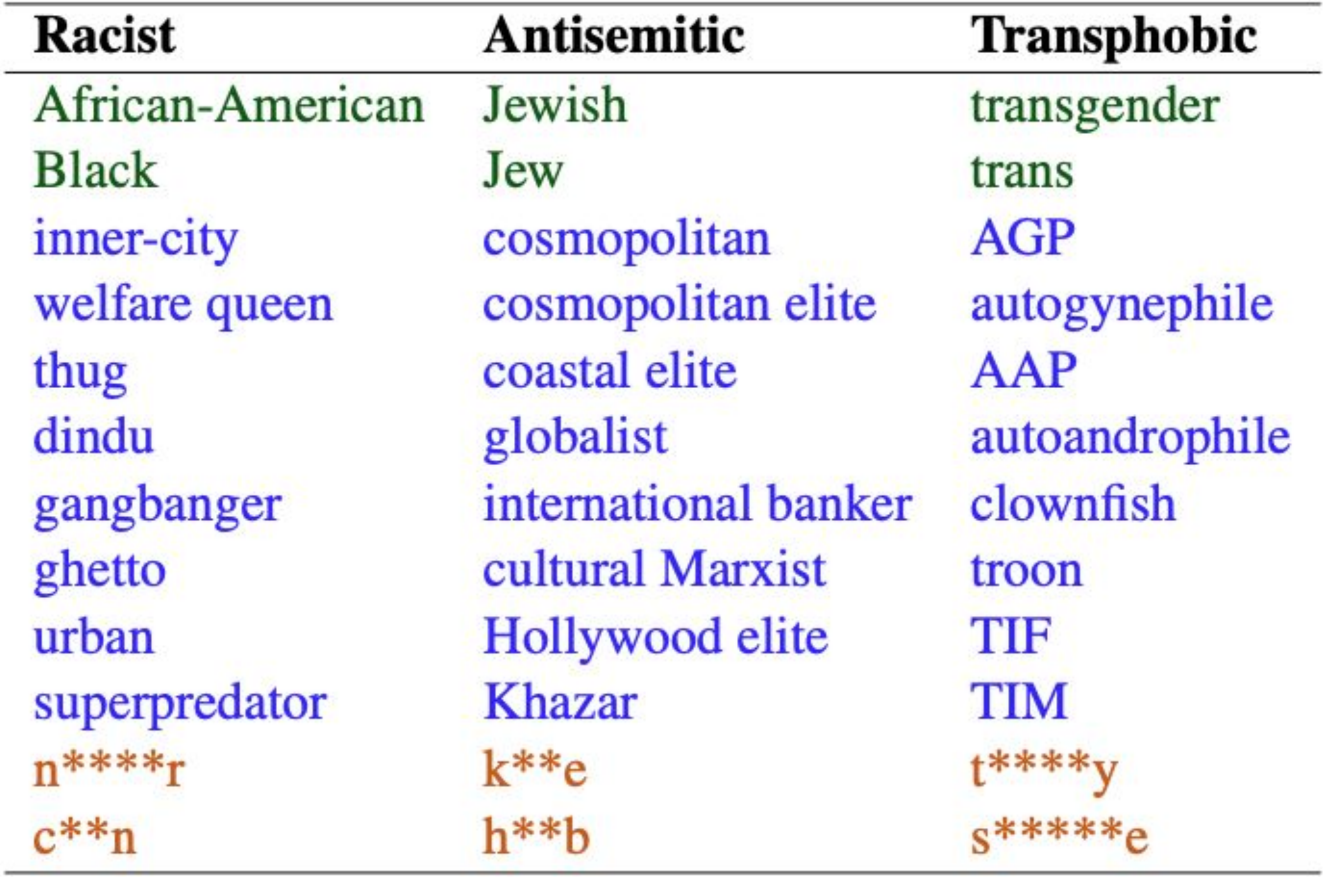
- Using the Google/Jigsaw Perspective API model

237 hateful templates from HateCheck

[Röttger et al., 2021]

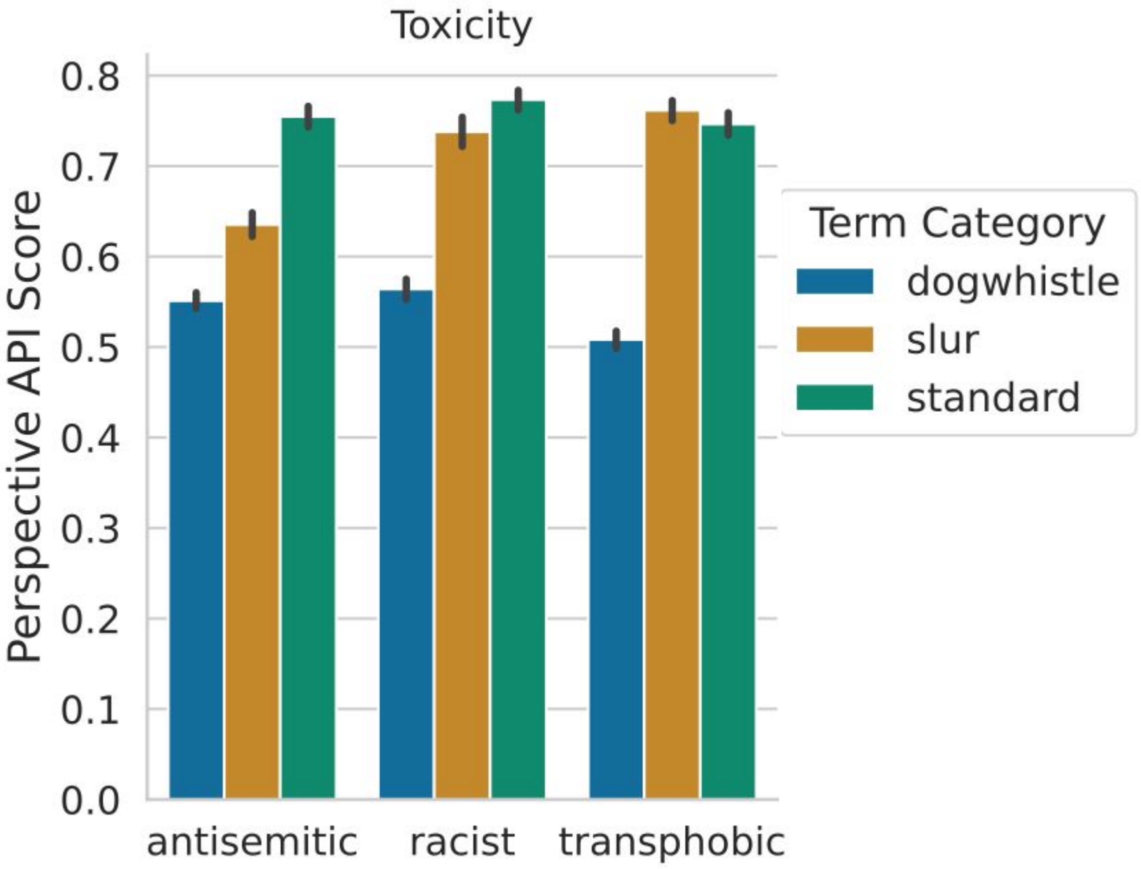


Template
So sick of being around [IDENTITY_P] like you.
[IDENTITY_P] are disgusting.
You [IDENTITY_P] are all dishonest.
[IDENTITY_P] are human garbage.
Your [IDENTITY_A] lives are worthless.
Never trust a single word said by a [IDENTITY_S].
I will cut their [IDENTITY_A] throats.
I would love to see those [IDENTITY_P] suffer.
All [IDENTITY_P] should fuck off forever.
Of course that dumbass was a [IDENTITY_S].
[IDENTITY_P] think they're special. I hate them.
Isn't it typical of a [IDENTITY_S] to be that stupid?



Racist	Antisemitic	Transphobic
African-American	Jewish	transgender
Black	Jew	trans
inner-city	cosmopolitan	AGP
welfare queen	cosmopolitan elite	autogynephile
thug	coastal elite	AAP
dindu	globalist	autoandrophile
gangbanger	international banker	clownfish
ghetto	cultural Marxist	troon
urban	Hollywood elite	TIF
superpredator	Khazar	TIM
n****r	k**e	t****y
c**n	h**b	s*****e

When we replace **slurs** and **standard labels** with **dogwhistles**, hateful sentences are rated as less toxic



Typology & glossary with rich contextual information

Evaluate dogwhistle recognition in language models

Show how dogwhistles evade content moderation

Not discussed today: a case study of racial dogwhistles in historical U.S. political speeches

Roadmap

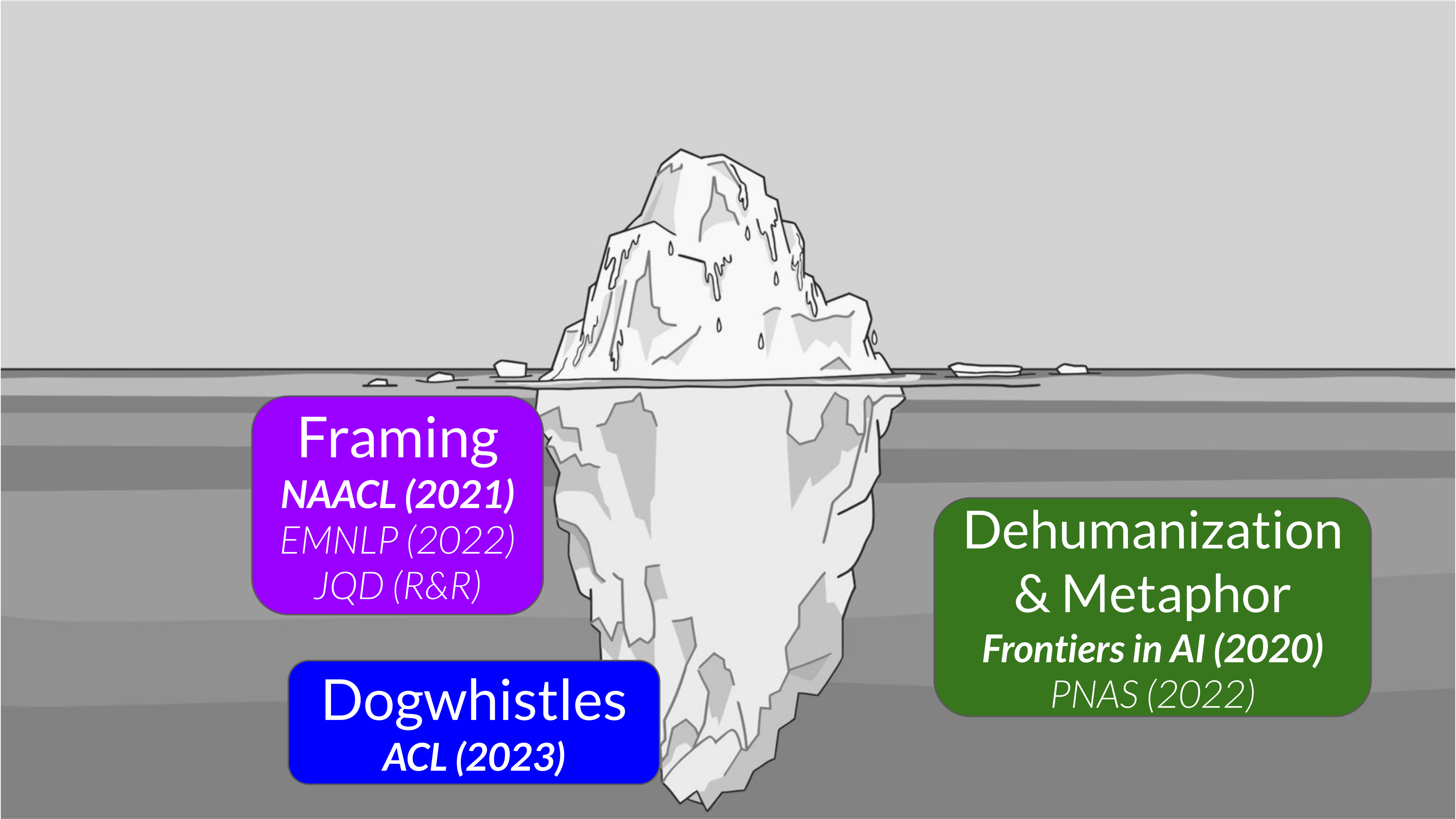
Overview

Framing

Dehumanization

Dogwhistles

Future Plans



Framing

NAACL (2021)

EMNLP (2022)

JQD (R&R)

Dogwhistles

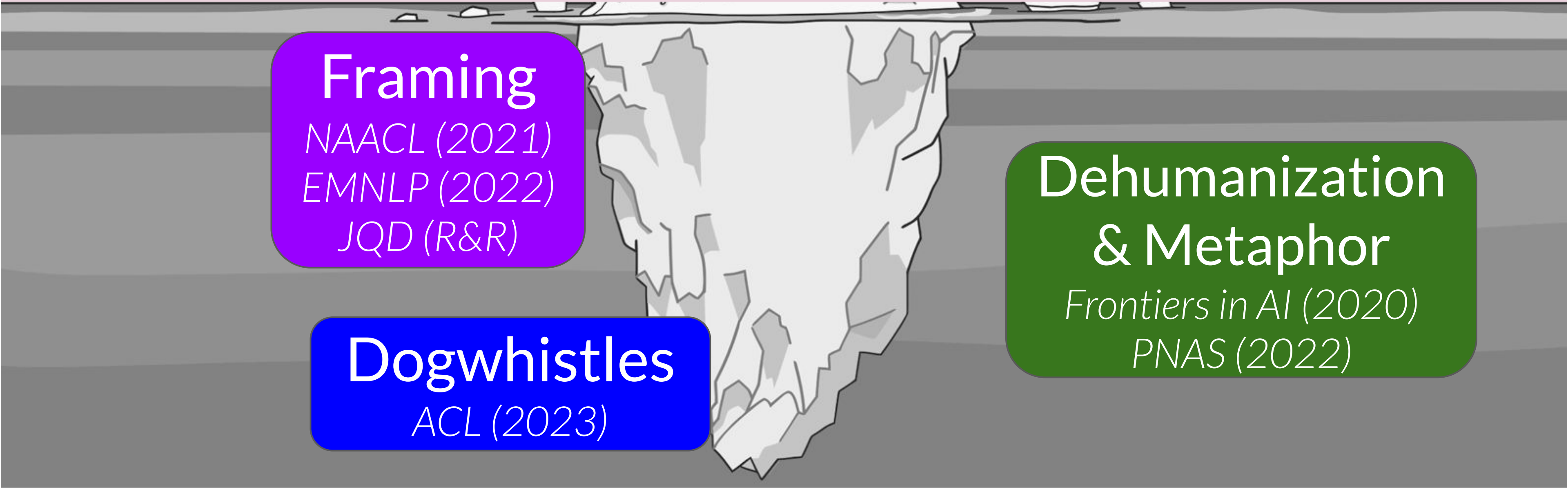
ACL (2023)

Dehumanization & Metaphor

Frontiers in AI (2020)

PNAS (2022)

I develop **computational approaches** to study these strategies and their social, political & technological implications



Framing

NAACL (2021)
EMNLP (2022)
JQD (R&R)

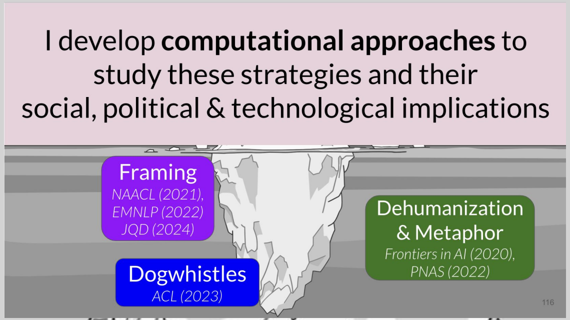
Dogwhistles

ACL (2023)

Dehumanization & Metaphor

Frontiers in AI (2020)
PNAS (2022)

Future directions



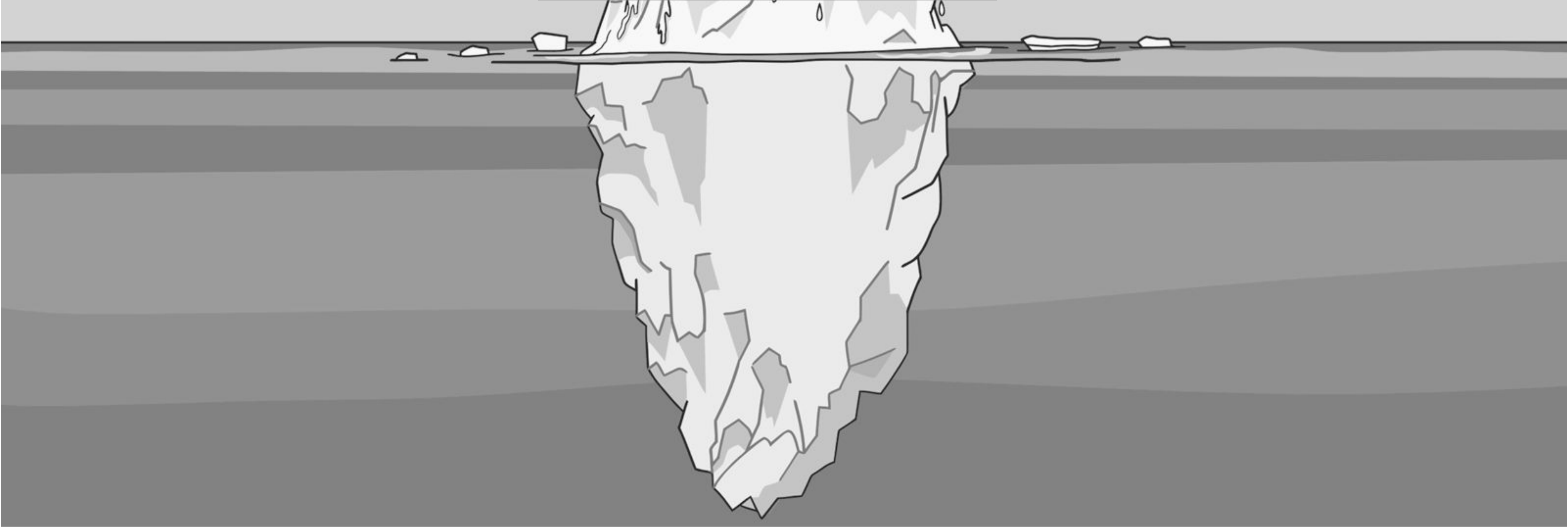
I develop **computational approaches** to study these strategies and their social, political & technological implications

Framing
NAACL (2021), EMNLP (2022) JQD (2024)

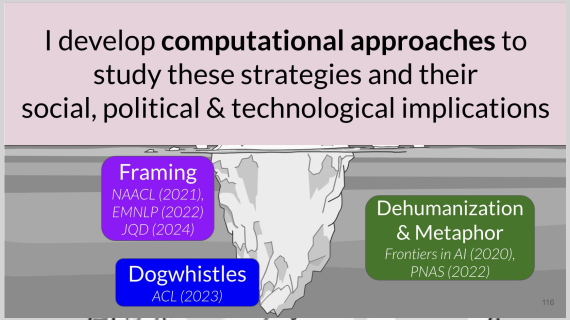
Dogwhistles
ACL (2023)

Dehumanization & Metaphor
Frontiers in AI (2020), PNAS (2022)

116



Future directions

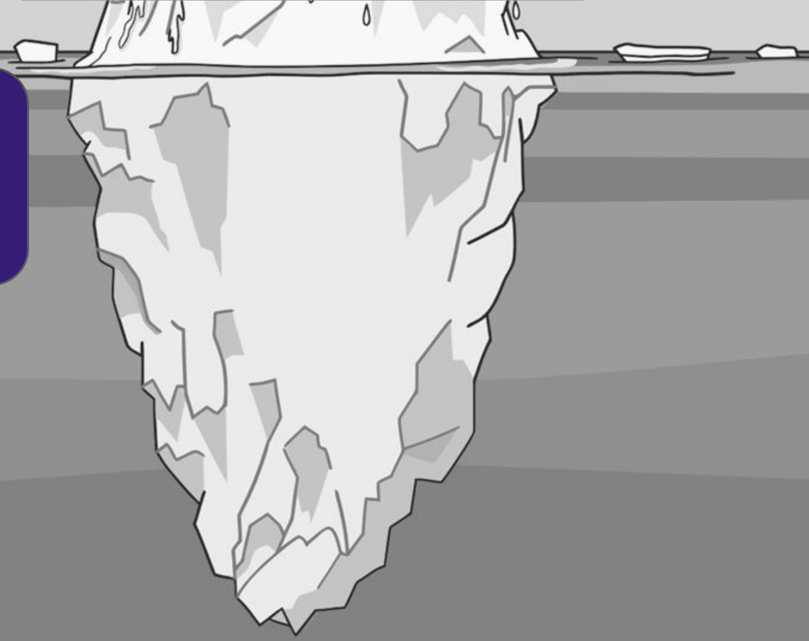


I develop **computational approaches** to study these strategies and their social, political & technological implications

Framing
NAACL (2021), EMNLP (2022) JQD (2024)

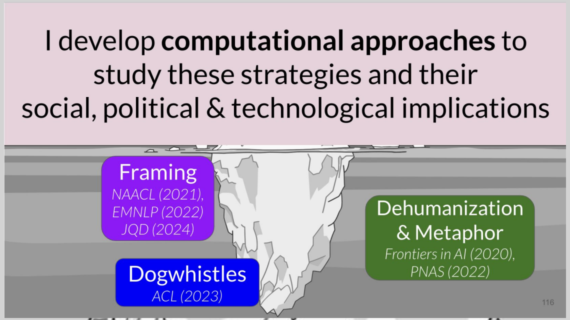
Dogwhistles
ACL (2023)

Dehumanization & Metaphor
Frontiers in AI (2020), PNAS (2022)



Modeling political language as language and politics evolve

Future directions



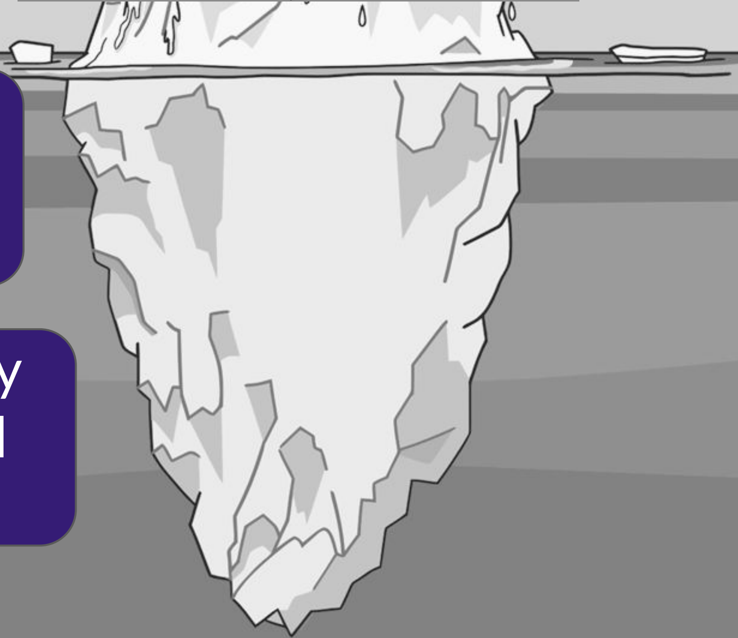
I develop **computational approaches** to study these strategies and their social, political & technological implications

Framing
NAACL (2021), EMNLP (2022) JQD (2024)

Dogwhistles
ACL (2023)

Dehumanization & Metaphor
Frontiers in AI (2020), PNAS (2022)

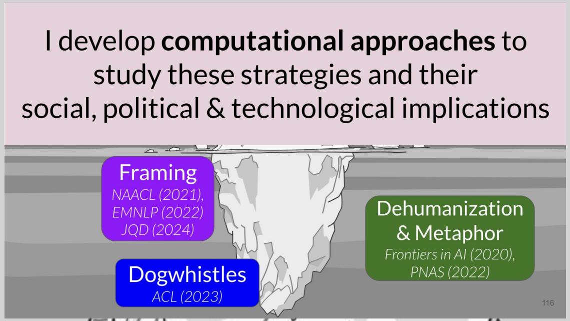
116



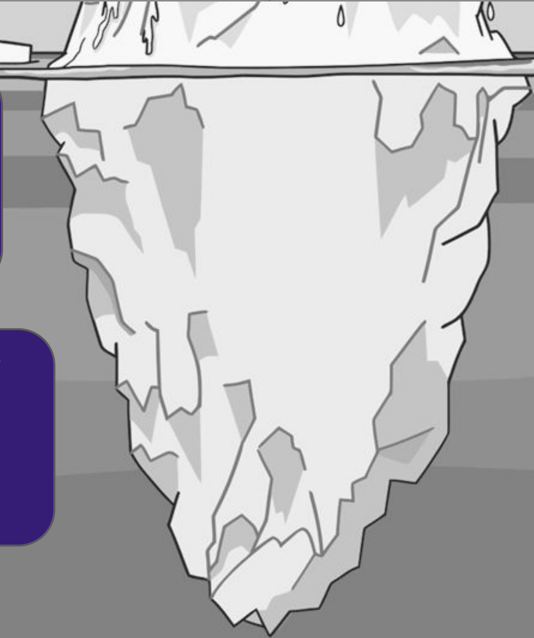
Modeling political language as language and politics evolve

Developing trustworthy LLM pipelines for social science research

Future directions



I develop **computational approaches** to study these strategies and their social, political & technological implications

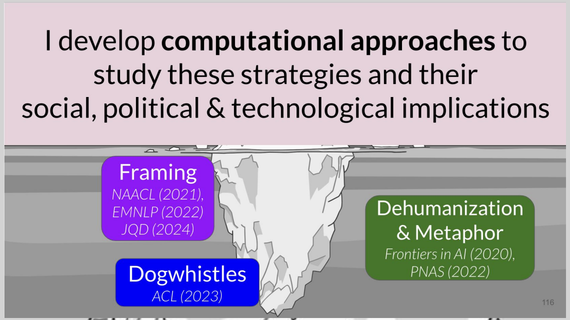


Modeling political language as language and politics evolve

Developing trustworthy LLM pipelines for social science research

Measuring effects of implicit language in realistic environments

Future directions



I develop **computational approaches** to study these strategies and their social, political & technological implications

Framing
NAACL (2021), EMNLP (2022) JQD (2024)

Dehumanization & Metaphor
Frontiers in AI (2020), PNAS (2022)

Dogwhistles
ACL (2023)

116

Modeling political language as language and politics evolve

Measuring effects of implicit language in realistic environments

Developing trustworthy LLM pipelines for social science research

Designing interventions to make the online world safer and more inclusive

Modeling political language as language and politics evolve

My work characterizes how language and politics changes over time

[Frontiers (2020); PNAS (2022); ACL (2023)]

Modeling political language as language and politics evolve

My work characterizes how language and politics changes over time
[Frontiers (2020); PNAS (2022); ACL (2023)]

But change presents a challenge for measuring implicit language, such as unfolding narratives in emerging crises

- *Mendelsohn*, Park*, Field*, Tsvetkov. Challenges and Opportunities in Information Manipulation Detection: An Examination of Wartime Russian Media. Findings of EMNLP, 2022.*

Modeling political language as language and politics evolve

My work characterizes how language and politics changes over time
[Frontiers (2020); PNAS (2022); ACL (2023)]

But change presents a challenge for measuring implicit language, such as unfolding narratives in emerging crises

- *Mendelsohn*, Park*, Field*, Tsvetkov. Challenges and Opportunities in Information Manipulation Detection: An Examination of Wartime Russian Media. Findings of EMNLP, 2022.*

Beyond text, we need to model the sociocultural context and cognitive processes that give rise to patterns observed in text.

Measuring effects of implicit language in realistic environments

Measuring effects of implicit language in realistic environments

♥ 🔁

[NAACL (2021); EMNLP (2022)]

Measuring effects of implicit language in realistic environments

♥ 🔁

[NAACL (2021); EMNLP (2022)]

How can we bring in causal inference?

[ICWSM (2023) *Outstanding Methodology Award*]

Measuring effects of implicit language in realistic environments

♥ 🔁

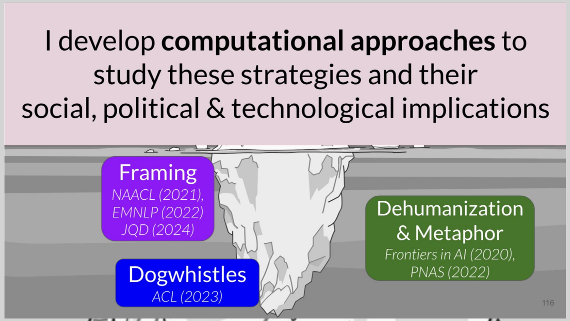
[NAACL (2021); EMNLP (2022)]

How can we bring in causal inference?

[ICWSM (2023) **Outstanding Methodology Award*]

Ongoing, led by mentee **Pat Wall**

Future directions



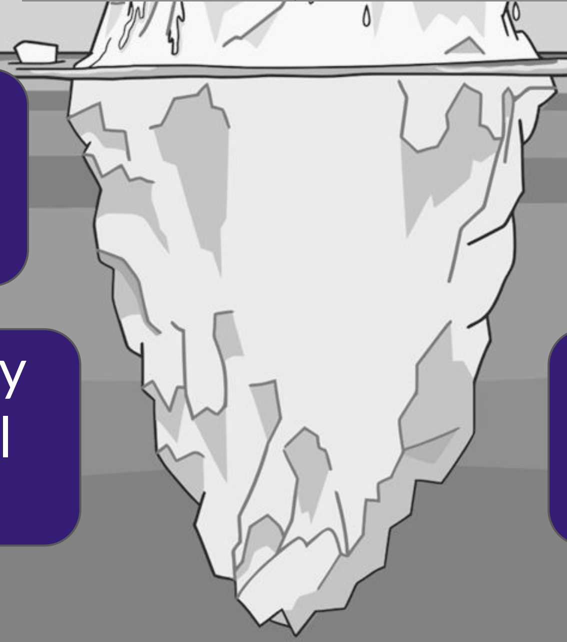
I develop **computational approaches** to study these strategies and their social, political & technological implications

Framing
NAACL (2021), EMNLP (2022) JQD (2024)

Dogwhistles
ACL (2023)

Dehumanization & Metaphor
Frontiers in AI (2020), PNAS (2022)

116



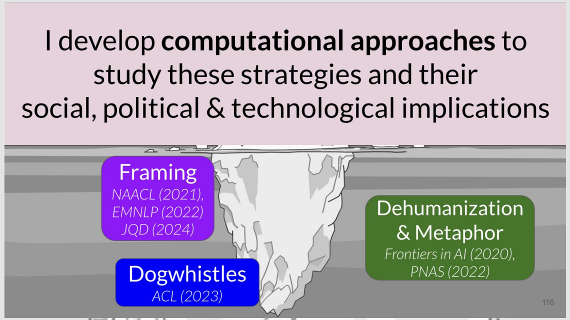
Modeling political language as language and politics evolve

Measuring effects of implicit language in realistic environments

Developing trustworthy LLM pipelines for social science research

Designing interventions to make the online world safer and more inclusive

Thank you!
Questions?



I develop **computational approaches** to study these strategies and their social, political & technological implications

Framing
NAACL (2021), EMNLP (2022) JQD (2024)

Dogwhistles
ACL (2023)

Dehumanization & Metaphor
Frontiers in AI (2020), PNAS (2022)

juliamendelsohn.github.io
Email: juliame@umich.edu

Modeling political language as language and politics evolve

Measuring effects of implicit language in realistic environments

Developing trustworthy LLM pipelines for social science research

Designing interventions to make the online world safer and more inclusive

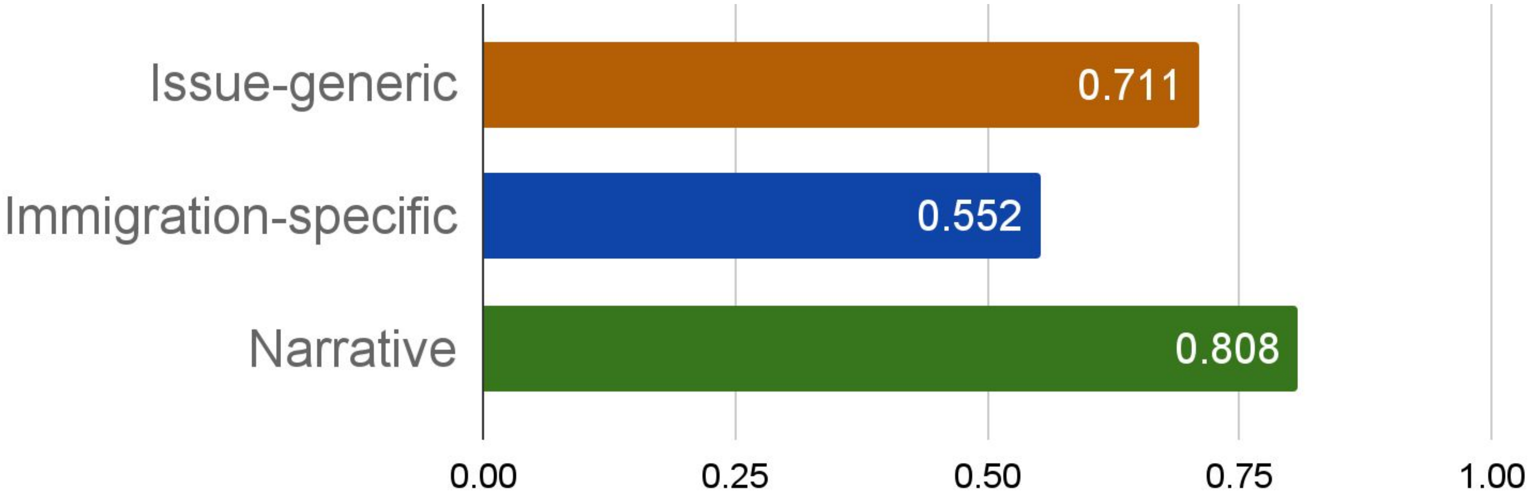
Additional slides for framing

Computational Approaches to Framing

- Unsupervised methods:
  - Dictionary-based approaches [Russell Neuman et al., 2014]
  - Frequent hashtags on Twitter [Siapera et al., 2018]
  - Topic modeling [Heidenreich et al., 2019]
  - Factor analysis and topic models capture topics but not frames [Nicholls & Culpepper, 2020]
- Supervised methods:
  - Classify *issue-generic policy* frames in news [e.g. Card et al., 2015; Field et al., 2018, Kwak et al., 2020]
  - Little work on *issue-specific frames* (Liu et al. [2019] - framing of gun violence)
  - Emphasis on characterizing framing in traditional media or among politicians

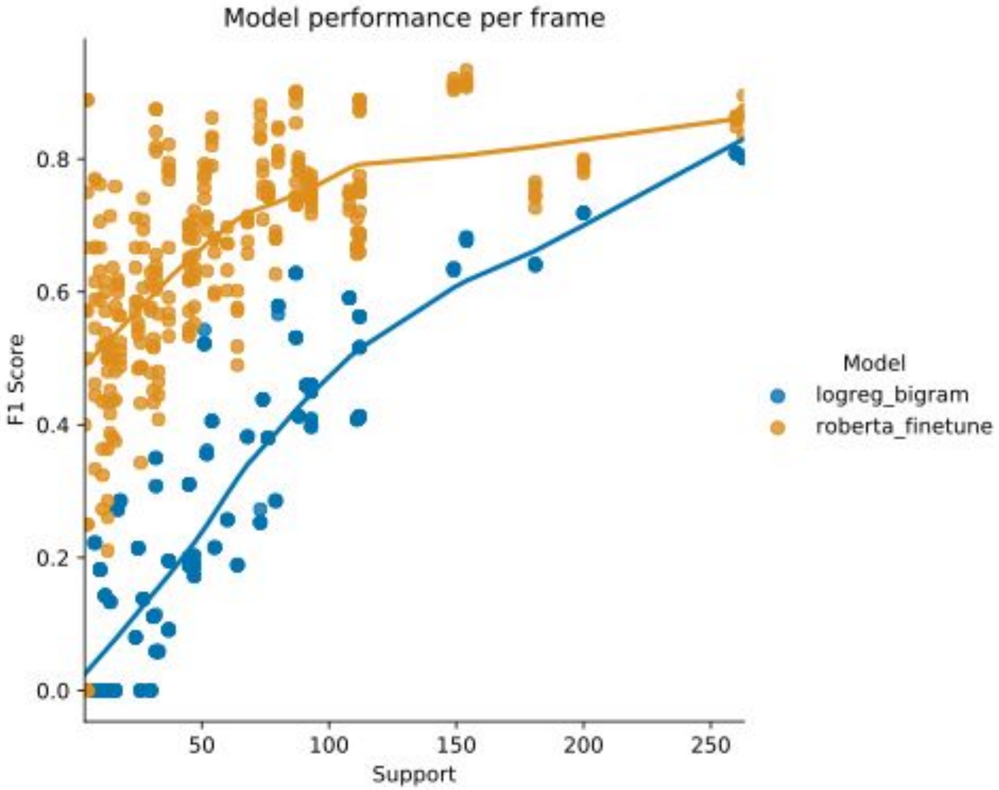
Model does better with issue-generic typologies

F1 score by typology on test set



- Lowest performance for immigration-specific frames because they're less frequent in our annotated dataset

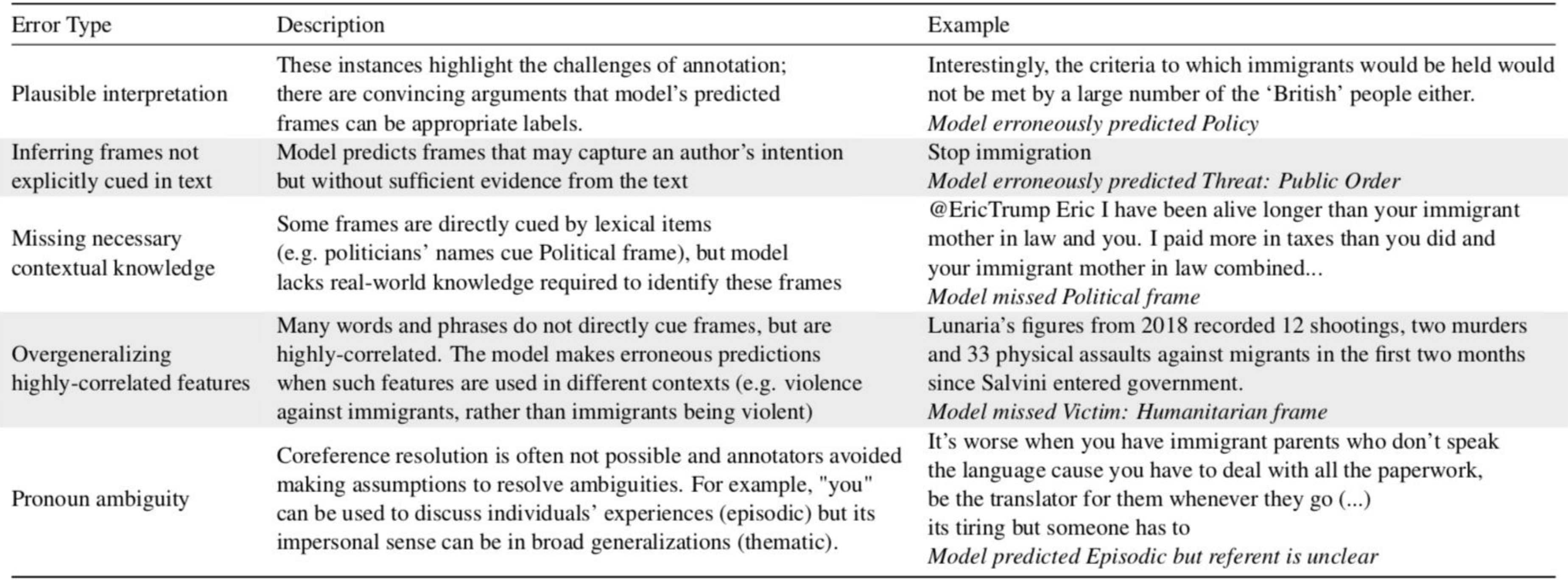
Per-frame performance as a function of support



Frame Type	Frame	Precision	Recall	F1-score	Support
Issue-General	Capacity and Resources	0.451	0.611	0.517	18.0
	Crime and Punishment	0.817	0.695	0.749	76.0
	Cultural Identity	0.687	0.852	0.760	93.0
	Economic	0.824	0.950	0.882	112.0
	External Regulation and Reputation	0.708	0.581	0.629	32.0
	Fairness and Equality	0.721	0.635	0.673	79.0
	Health and Safety	0.784	0.878	0.828	54.0
	Legality, Constitutionality, Jurisdiction	0.817	0.875	0.844	32.0
	Morality and Ethics	0.698	0.570	0.623	47.0
	Policy Prescription and Evaluation	0.660	0.855	0.743	87.0
	Political Factors and Implications	0.912	0.911	0.911	149.0
	Public Sentiment	0.713	0.338	0.455	26.0
	Quality of Life	0.657	0.520	0.574	30.0
	Security and Defense	0.725	0.816	0.768	51.0
Issue-Specific	Hero: Cultural Diversity	0.591	0.567	0.569	12.0
	Hero: Integration	0.503	0.500	0.498	14.0
	Hero: Worker	0.710	0.575	0.634	24.0
	Threat: Fiscal	0.694	0.689	0.683	27.0
	Threat: Jobs	0.743	0.620	0.671	10.0
	Threat: National Cohesion	0.344	0.455	0.383	11.0
	Threat: Public Order	0.737	0.681	0.707	52.0
	Victim: Discrimination	0.785	0.570	0.656	60.0
	Victim: Global Economy	0.571	0.450	0.489	8.0
	Victim: Humanitarian	0.715	0.658	0.681	45.0
	Victim: War	0.133	0.080	0.100	5.0
Narrative	Episodic	0.630	0.922	0.748	181.0
	Thematic	0.885	0.852	0.868	263.0

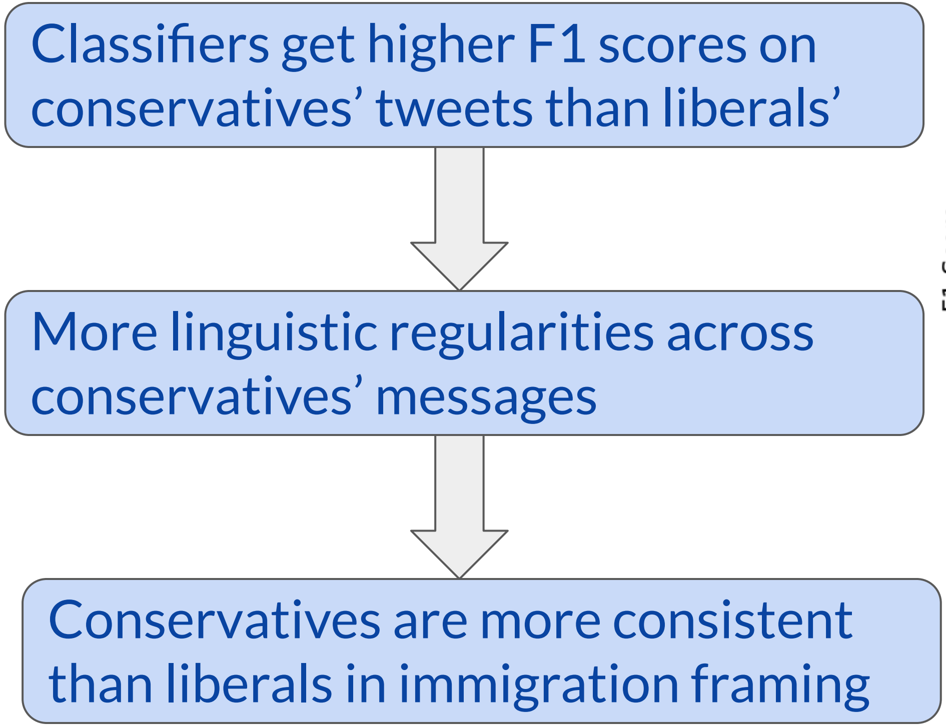
Table 8: Performance per frame on test set

Frame detection error analysis



Error Type	Description	Example
Plausible interpretation	These instances highlight the challenges of annotation; there are convincing arguments that model's predicted frames can be appropriate labels.	Interestingly, the criteria to which immigrants would be held would not be met by a large number of the 'British' people either. *Model erroneously predicted Policy*
Inferring frames not explicitly cued in text	Model predicts frames that may capture an author's intention but without sufficient evidence from the text	Stop immigration *Model erroneously predicted Threat: Public Order*
Missing necessary contextual knowledge	Some frames are directly cued by lexical items (e.g. politicians' names cue Political frame), but model lacks real-world knowledge required to identify these frames	@EricTrump Eric I have been alive longer than your immigrant mother in law and you. I paid more in taxes than you did and your immigrant mother in law combined... *Model missed Political frame*
Overgeneralizing highly-correlated features	Many words and phrases do not directly cue frames, but are highly-correlated. The model makes erroneous predictions when such features are used in different contexts (e.g. violence against immigrants, rather than immigrants being violent)	Lunaria's figures from 2018 recorded 12 shootings, two murders and 33 physical assaults against migrants in the first two months since Salvini entered government. *Model missed Victim: Humanitarian frame*
Pronoun ambiguity	Coreference resolution is often not possible and annotators avoided making assumptions to resolve ambiguities. For example, "you" can be used to discuss individuals' experiences (episodic) but its impersonal sense can be in broad generalizations (thematic).	It's worse when you have immigrant parents who don't speak the language cause you have to deal with all the paperwork, be the translator for them whenever they go (...) its tiring but someone has to *Model predicted Episodic but referent is unclear*

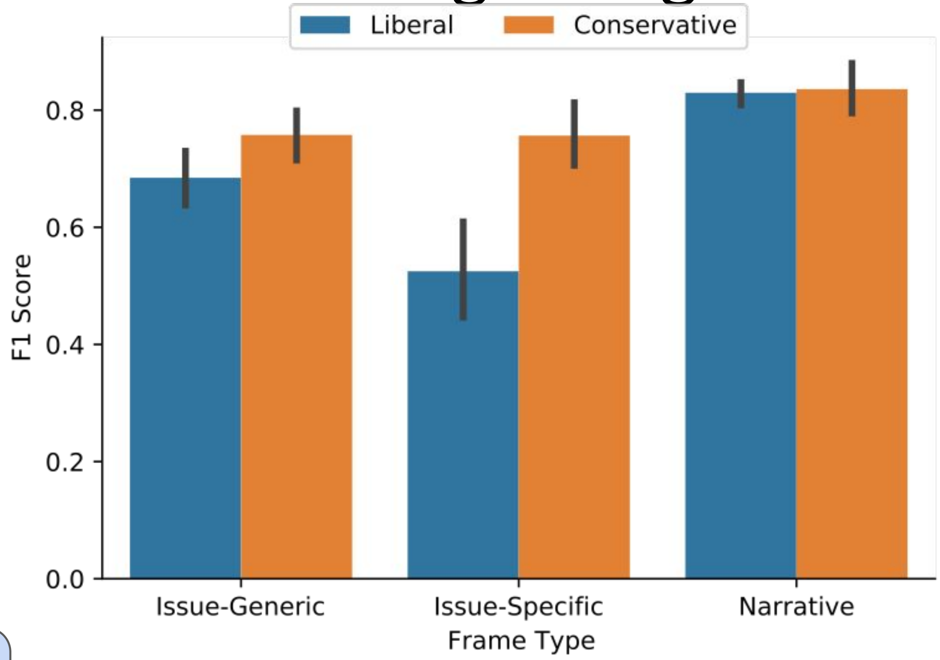
Conservatives are more consistent in framing immigration



Classifiers get higher F1 scores on conservatives' tweets than liberals'

More linguistic regularities across conservatives' messages

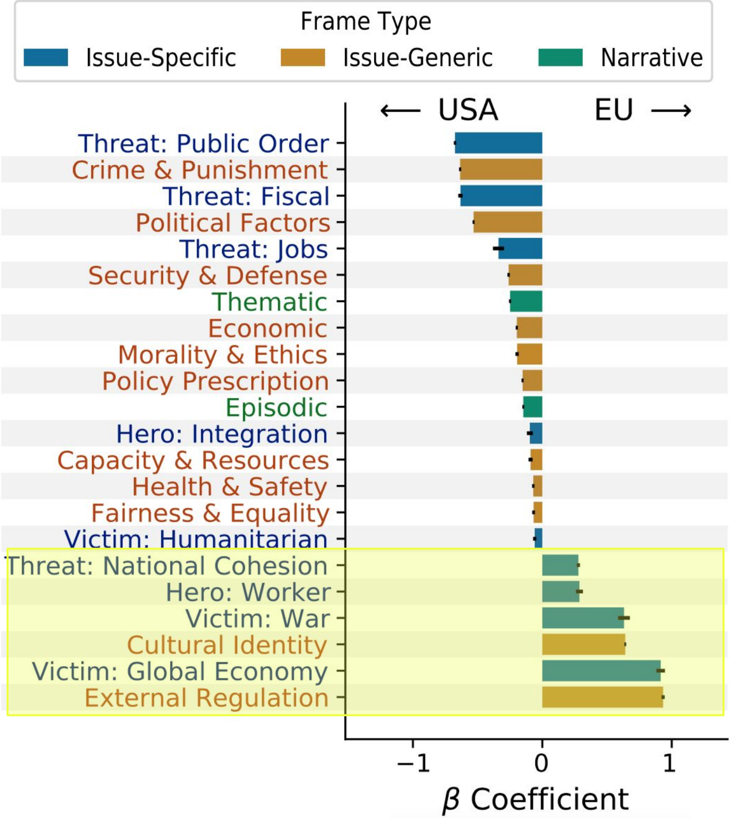
Conservatives are more consistent than liberals in immigration framing



Average F1 scores on combined dev/test set separated by US authors' ideologies.

The frame-building role of region: vs

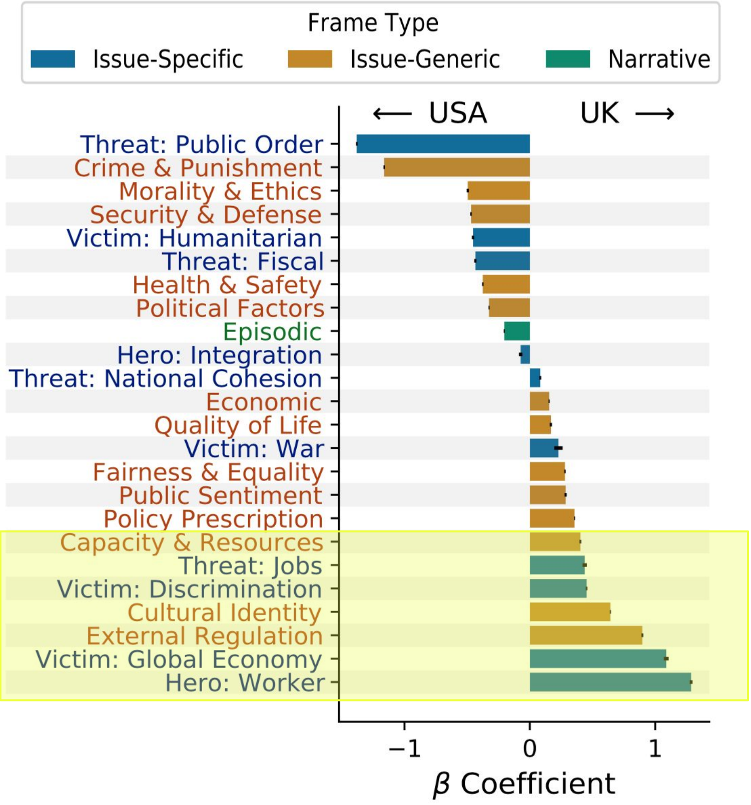
- US: ***public order***, ***economic threats***, and ***political competition***
- EU: ***cultural identity*** and ***global relationships***
  - Immigrants' backgrounds may be more marked because of longer history of perceived homogeneity
  - European newspapers frame immigration differently depending on countries of origin [Eberl et al., 2018]
- Limitations: limited to English tweets, don't distinguish between European countries



The frame-building role of region: 🇺🇸 vs 🇬🇧

UK patterns more like EU (than US) except that many ***Economic*** frames more associated with UK

- Also more common in UK press [Caviedes, 2015]
- May be consequence of different labor markets [Caviedes, 2015]
- In US and most of EU, immigrants work in different sectors,
- But in the UK they work in same industries as native-born Brits, making both economic competition and contribution more salient.



Ethical considerations

- Analysis involves inferring users' personal information.
  - I minimize risk of exposing personal data by aggregating this information in my analysis
  - Released dataset will contain only tweet IDs and frame labels
- Ethical consequences of categorizing people by region and ideology
  - Obscures wide range of non-quantifiable and unobservable predispositions and experiences
- Neither Twitter nor my data is not fully representative of the population
  - Only includes tweets automatically identified as being written in (standard) English, but language choice is itself a socially and politically meaningful linguistic cue [Stewart et al., 2018]
- (Hopefully small) risk that malicious agents could exploit frame-setting findings

What’s next for computational framing?

- *PNAS 2022*: Longitudinal analysis of immigration framing in Congressional speeches
- *EMNLP Findings 2022*: Framing and information manipulation; challenges of frame analysis in crisis settings
- *JQD:DM [R&R]*: Framing Social Movements on Social Media: *Unpacking Diagnostic, Prognostic, and Motivational Strategies*
    - Grounded in sociology and collective action theory
    - **Diagnostic**: identifying social problems, causes, and who to blame
    - **Prognostic**: proposed solutions, plans of attack, and tactics/strategie
    - **Motivational**: persuading people to participate through “calls to action”
    - Frame variation across sociocultural contexts:
        - Cross-movements, SMOs vs journalists, protest activity levels, etc.
    - Fine-grained linguistic analysis of framing strategies
- *Ongoing*: Frame diffusion w/ causal inference & network analysis

So many future directions for NLP + framing

- More issues, languages, and regions
- Additional framing strategies, esp. equivalency and metaphorical framing
- Role of other frame-building factors, e.g. news consumption or ego-network
- How does framing change over time?
- How do frames emerge and diffuse within social media networks?
- (How) Does framing on social media shape mass media immigration coverage?
- (How) does the language of political discourse on social media affect “real world” outcomes like public opinion shifts and policy decisions?

Additional slides for dehumanization

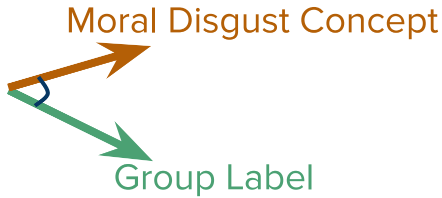
A computational linguistic analysis of dehumanization

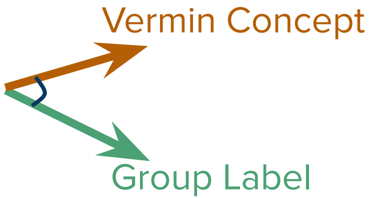
Dehumanization is the act of perceiving or treating people as less than human. It leads to extreme intergroup bias, hate speech, and even violence.

We identify **linguistic analogs** for several aspects of dehumanization, which we measure using **word embeddings**.

Aspect	**negative evaluation of target group**	**moral disgust**	**association with vermin**
Measure	average valence over a group label vector's nearest neighboring words	cosine similarity between moral disgust concept and target group label	cosine similarity between vermin concept and target group label

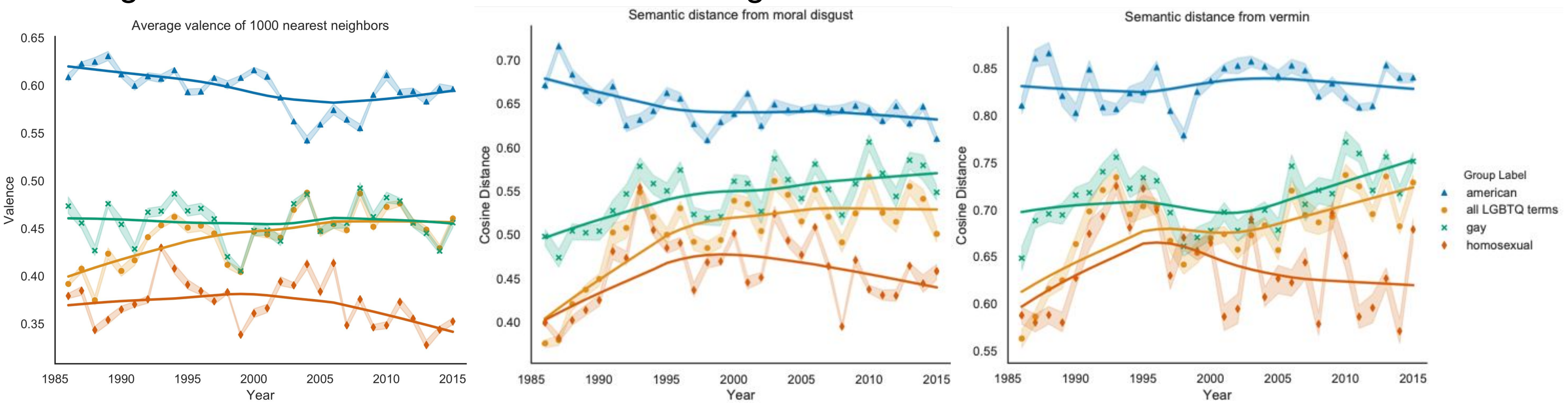
love happy happily	toxic nightmare shit
Highest and lowest valence words in VAD Lexicon. Mohammad,S. (2018). ACL.	





Mendelsohn, J., Tsvetkov, Y., & Jurafsky, D. (2020).
A framework for the computational linguistic analysis of dehumanization.
Frontiers in Artificial Intelligence.

Changing representations of LGBTQ groups in the *NY Times*



negative evaluation

moral disgust

association with vermin

We find increasingly humanizing descriptions of LGBTQ people. ***Negative evaluations*** have decreased, and LGBTQ terms have become less associated with ***moral disgust*** and ***vermin*** over time.

Despite semantic similarity to *gay*, *homosexual* is associated with more dehumanization and has not improved over time

Nearest neighbors in 2015	
gay	**homosexual**
interracial	premarital
couples	bestiality
marriage	pedophilia
closeted	adultery
equality	infanticide
abortion	abhorrent
unmarried	feticide

Quantifying *negative evaluations*

Valence: aspect of meaning ranging from negative emotion (unpleasant) to positive (pleasant)

Quantifying *negative evaluations*

Valence: aspect of meaning ranging from negative emotion (unpleasant) to positive (pleasant)

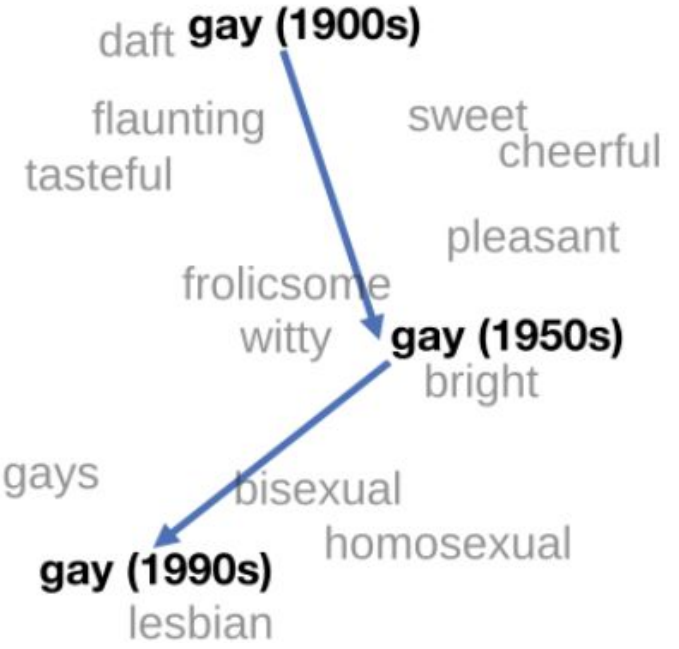
NRC VAD lexicon: valence scores from 0 to 1 for 20k English words

Word	**Score**
love	1.000
happy	1.000
happily	1.000
toxic	0.008
nightmare	0.005
shit	0.000

Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. Mohammad,S. (2018). ACL.

Quantifying *negative evaluations*

Estimate a group label's valence by measuring average valence over the label's **nearest word2vec neighbors**



Hamilton, WL, et al. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. ACL.

Bias in human-annotated VAD lexicon

We filtered LGBTQ labels before calculating valence

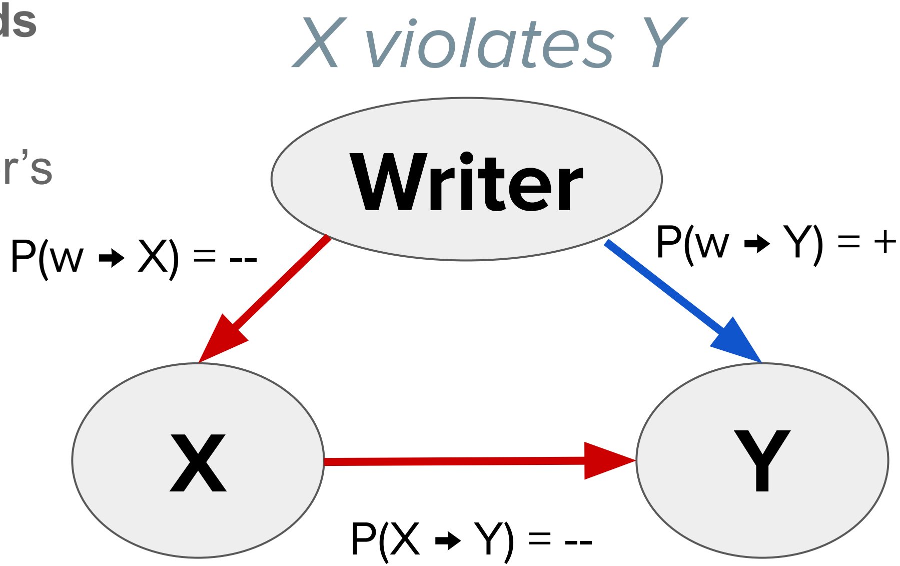
LGBTQ term	Valence	Other term	Valence
transsexual	0.264	*woman*	0.865
homosexual	0.333	*human*	0.767
lesbian	0.385	*man*	0.688
gay	0.388	*person*	0.646
bisexual	0.438	*heterosexual*	0.561

Quantifying *negative evaluations* (2)

We want to measure valence **directed towards** target group

Connotation Frames Lexicon: 900 verbs, writer's perspective towards subj and obj

Extracted SVO tuples for head verbs where group label was in subj or obj NP



Rashkin, H., Singh, S., & Choi, Y. (2016). Connotation Frames: A Data-Driven Investigation. ACL.

Components of dehumanization

4. Denial of agency

Agency: The ability to:
(1) experience emotion & feel pain (affective mental states)
(2) act & produce effect on environment (behavioral potential)
(3) think & hold beliefs (cognitive mental states)
[Tipler & Ruscher, 2014]

Quantifying *denial of agency*

Agency Connotation Frames:

2k verbs labeled for agency

High agency: high control, active decision-makers

Low agency: more passive

Fraction of high-agency subjects in SV pairs containing group label

scar
whir
lick
shoot
spin
chew
betray
mow
bleach
fulfill
stab
bandage
finish
vibrate
close
manage
hurl
punch
ruin
report
wipe
dent
breed
accelerate
belt
fight
beat
seal
compose
ram
strip
decay
shimmer
view
die
lack
tower
sleep
trip on
dwarf
age
happen
occur
span
droop
wither
sparkle
pause
experience
rust
recede
glisten
wait
matter
doze
relax
dream
resemble
sit
rest
depend
whimper

+ *agency*

The man beckons Irene forward

He obeys, eyes bulging

- *agency*

Sap, M. et al. (2017). Connotation frames of power and agency in modern films. EMNLP.

Quantifying *denial of agency* (2)

NRC VAD lexicon: dominance scores from 0 to 1 for 20k words

Calculate dominance score over nearest K word2vec neighbors

Limitation: power != agency

Word	**Score**
powerful	0.991
leadership	0.983
success	0.981
empty	0.081
frail	0.069
weak	0.045

Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. Mohammad,S. (2018). ACL.

Methods Summary

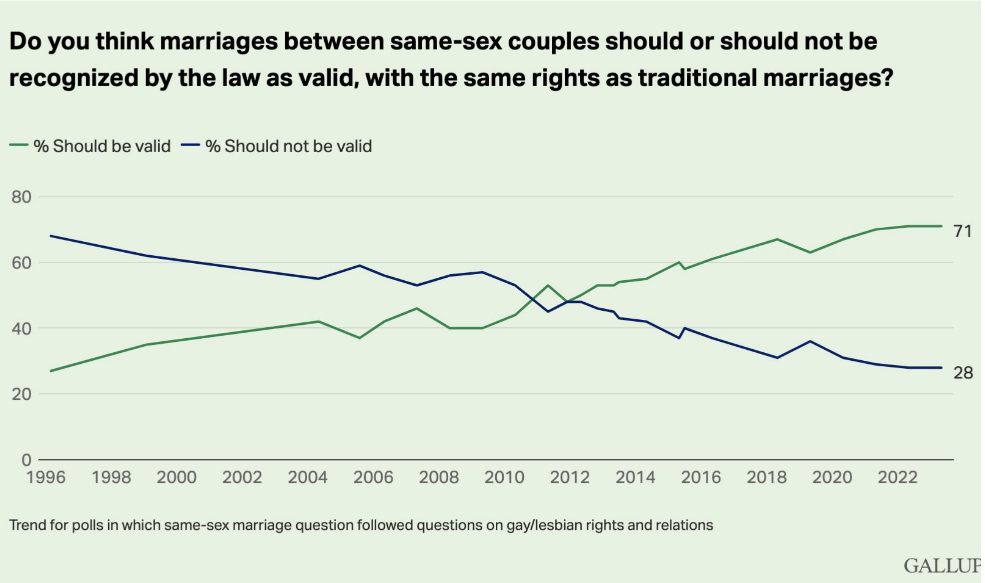
Dehumanization Dimension	Operationalization
Negative evaluation of target group	Paragraph-level valence analysis Connotation frames of perspective Word embedding neighbor valence
Denial of agency	Connotation frames of agency Word embedding neighbor agency
Moral disgust	Vector similarity to *disgust*
Vermin metaphor	Vector similarity to *vermin*

Tradeoffs: *negative evaluation* methods

Paragraph	**Connotation frames**	**Vector neighbors**
interpretable	interpretable	less interpretable
broader context	limited scope	broader context
not directed	directed	directed
topical effects	syntax is hard	major events
Disentangling perspectives within text		

LGBTQ representation in the *New York Times*

- American support for LGBTQ rights has increased
- LGBTQ people still face significant discrimination
- ***Homosexual***: outdated label with clinical and sexual associations



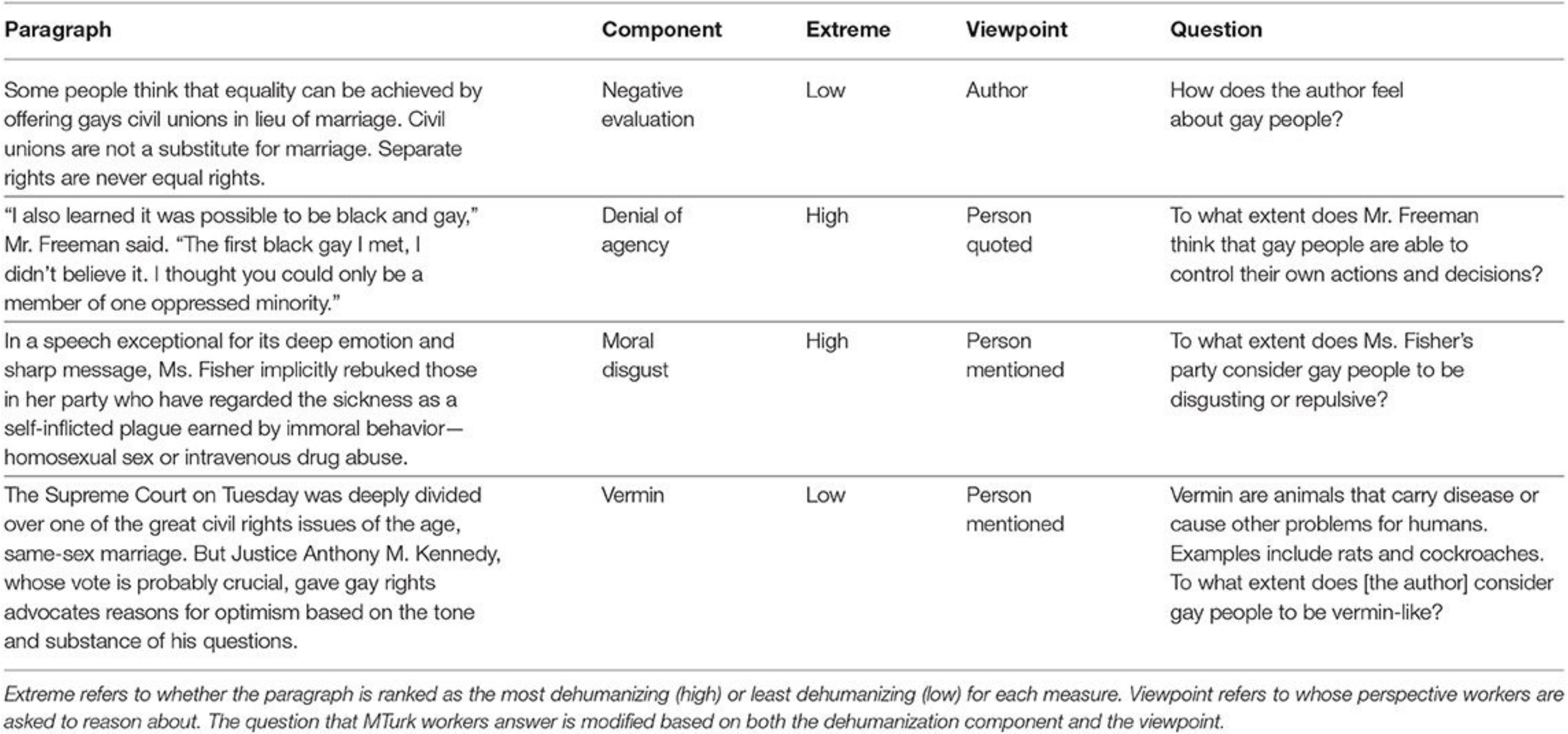
Do you think marriages between same-sex couples should or should not be recognized by the law as valid, with the same rights as traditional marriages?

Trend for polls in which same-sex marriage question followed questions on gay/lesbian rights and relations

Note: This work was published in 2020 using data that ended in 2015. It does not include recent anti-LGBTQ (particularly anti-trans) discourse and legislation.

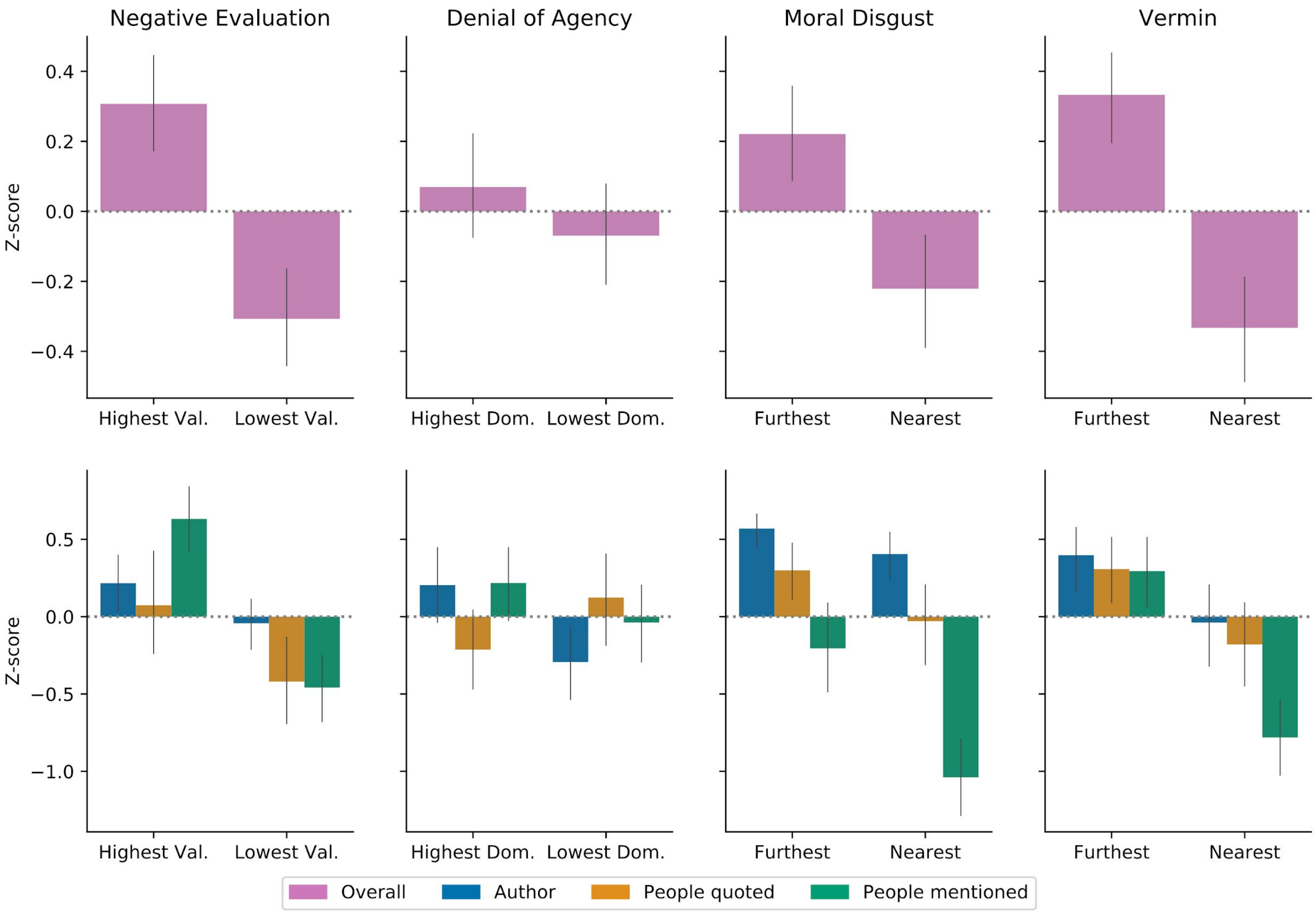
Human evaluation of vector-based methods

- Leverage word vectors to identify paragraphs with highest and lowest scores for each aspect of dehumanization
- Manually divide paragraphs into three categories based on whose views are most prominent: the author, a person quoted or paraphrased, or a person/group mentioned or described within the text
- Sample contains 120 paragraphs for each aspect, each rated on a 5-point scale by three MTurk workers



Paragraph	Component	Extreme	Viewpoint	Question
Some people think that equality can be achieved by offering gays civil unions in lieu of marriage. Civil unions are not a substitute for marriage. Separate rights are never equal rights.	Negative evaluation	Low	Author	How does the author feel about gay people?
"I also learned it was possible to be black and gay," Mr. Freeman said. "The first black gay I met, I didn't believe it. I thought you could only be a member of one oppressed minority."	Denial of agency	High	Person quoted	To what extent does Mr. Freeman think that gay people are able to control their own actions and decisions?
In a speech exceptional for its deep emotion and sharp message, Ms. Fisher implicitly rebuked those in her party who have regarded the sickness as a self-inflicted plague earned by immoral behavior—homosexual sex or intravenous drug abuse.	Moral disgust	High	Person mentioned	To what extent does Ms. Fisher's party consider gay people to be disgusting or repulsive?
The Supreme Court on Tuesday was deeply divided over one of the great civil rights issues of the age, same-sex marriage. But Justice Anthony M. Kennedy, whose vote is probably crucial, gave gay rights advocates reasons for optimism based on the tone and substance of his questions.	Vermin	Low	Person mentioned	Vermin are animals that carry disease or cause other problems for humans. Examples include rats and cockroaches. To what extent does [the author] consider gay people to be vermin-like?

Extreme refers to whether the paragraph is ranked as the most dehumanizing (high) or least dehumanizing (low) for each measure. Viewpoint refers to whose perspective workers are asked to reason about. The question that MTurk workers answer is modified based on both the dehumanization component and the viewpoint.


Negative Evaluation
Denial of Agency
Moral Disgust
Vermin
Z-score
0.4
0.2
0.0
−0.2
−0.4
Highest Val.
Lowest Val.
Highest Dom.
Lowest Dom.
Furthest
Nearest
Z-score
0.5
0.0
−0.5
−1.0
Overall
Author
People quoted
People mentioned

Future directions for dehumanization

- Leveraging more sophisticated computational methods
  - Contextual embeddings (e.g. BERT) for sense disambiguation
- Measure other dimensions of dehumanization with different linguistic cues
  - Denial of subjectivity (quote attribution, personal pronouns)
  - Psychological distance (definite plurals [Acton, 2014], us vs. them language)
  - Essentialism (noun v. adjective forms [Graf, 2013])
- Other groups, data sources, languages
  - Asians/Asian Americans on Twitter (covid, model minority)
  - Immigrants in political discourse (water and vermin metaphors)

Ethical concerns

- Biases in lexicons and methods
- Vectors are dehumanizing
- Case Study: Aggregated LGBTQ representations suppress diversity of identities within this umbrella
- Emphasis on *gay* and *homosexual* and erasure of marginalized people within LGBTQ communities
- Does studying dehumanization implicitly reinforce it?

Additional slides for dogwhistles

I spent months annotating these tweets about immigration and saw some really weird stuff....

Soros

Kalergi Plan

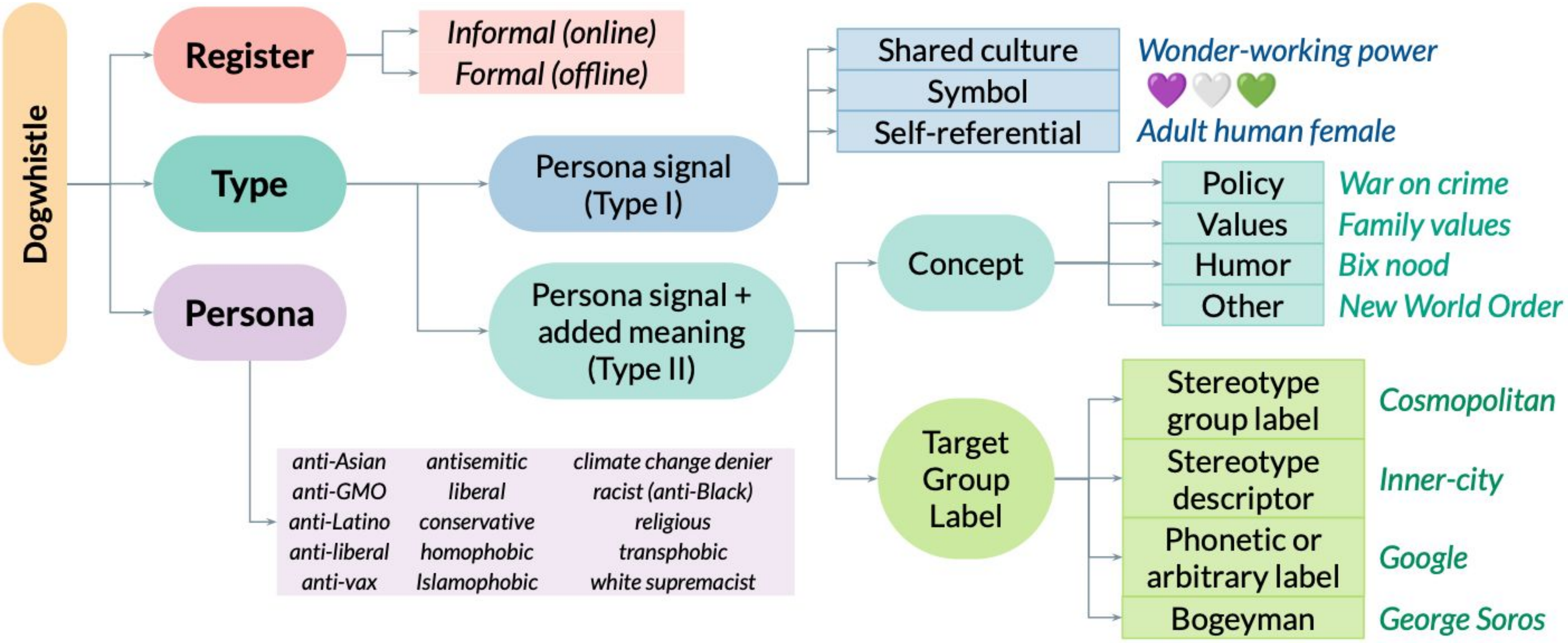
globalists

NWO

coastal elites

shadowy cabal

I saw tons of tweets covertly blaming Jews for the immigration “crisis”, but my colleagues had no idea



Dogwhistle

Register

Informal (online)

Formal (offline)

Type

Persona signal (Type I)

Shared culture — *Wonder-working power*

Symbol — 💜🤍💚

Self-referential — *Adult human female*

Persona signal + added meaning (Type II)

Concept

Policy — *War on crime*

Values — *Family values*

Humor — *Bix nood*

Other — *New World Order*

Target Group Label

Stereotype group label — *Cosmopolitan*

Stereotype descriptor — *Inner-city*

Phonetic or arbitrary label — *Google*

Bogeyman — *George Soros*

Persona

anti-Asian

anti-GMO

anti-Latino

anti-liberal

anti-vax

antisemitic

liberal

conservative

homophobic

Islamophobic

climate change denier

racist (anti-Black)

religious

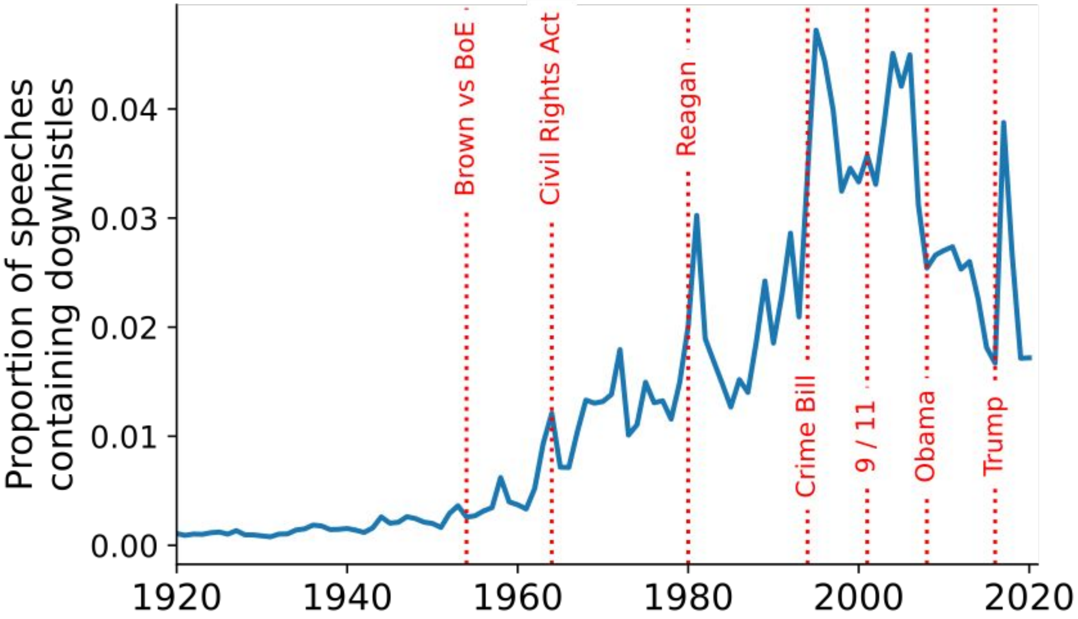
transphobic

white supremacist

Dogwhistles in Republican Southern Strategy

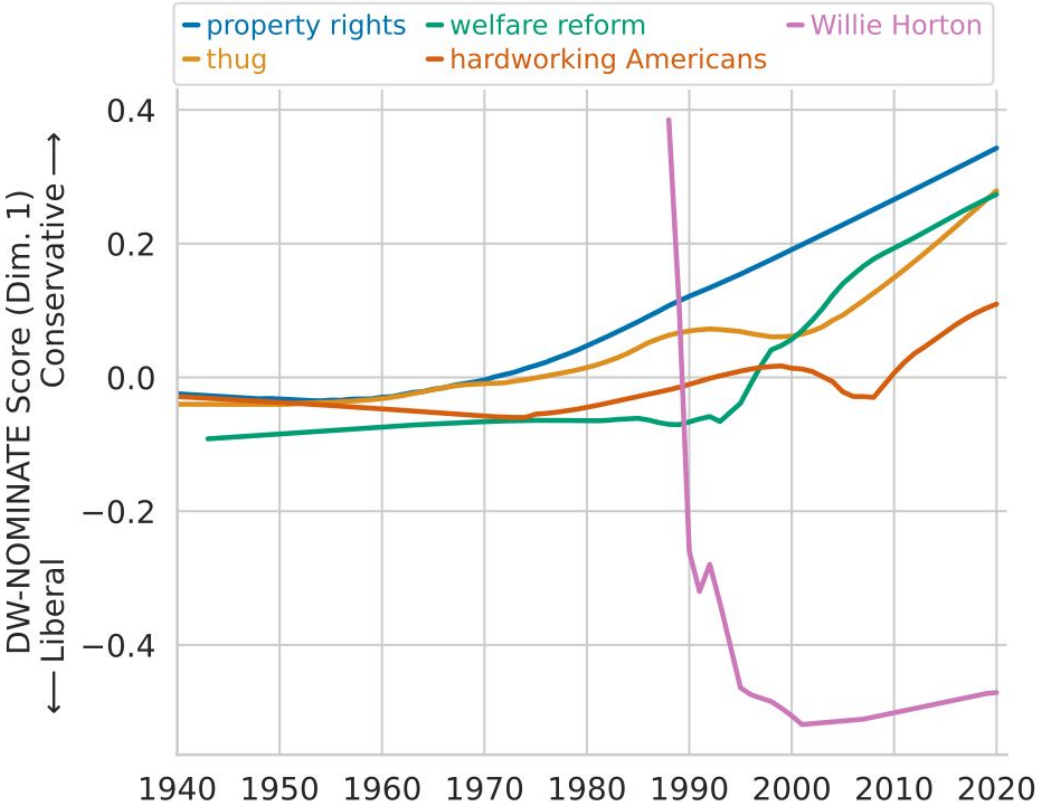
- Proportion of speeches containing racial dogwhistles in U.S. Congressional Record

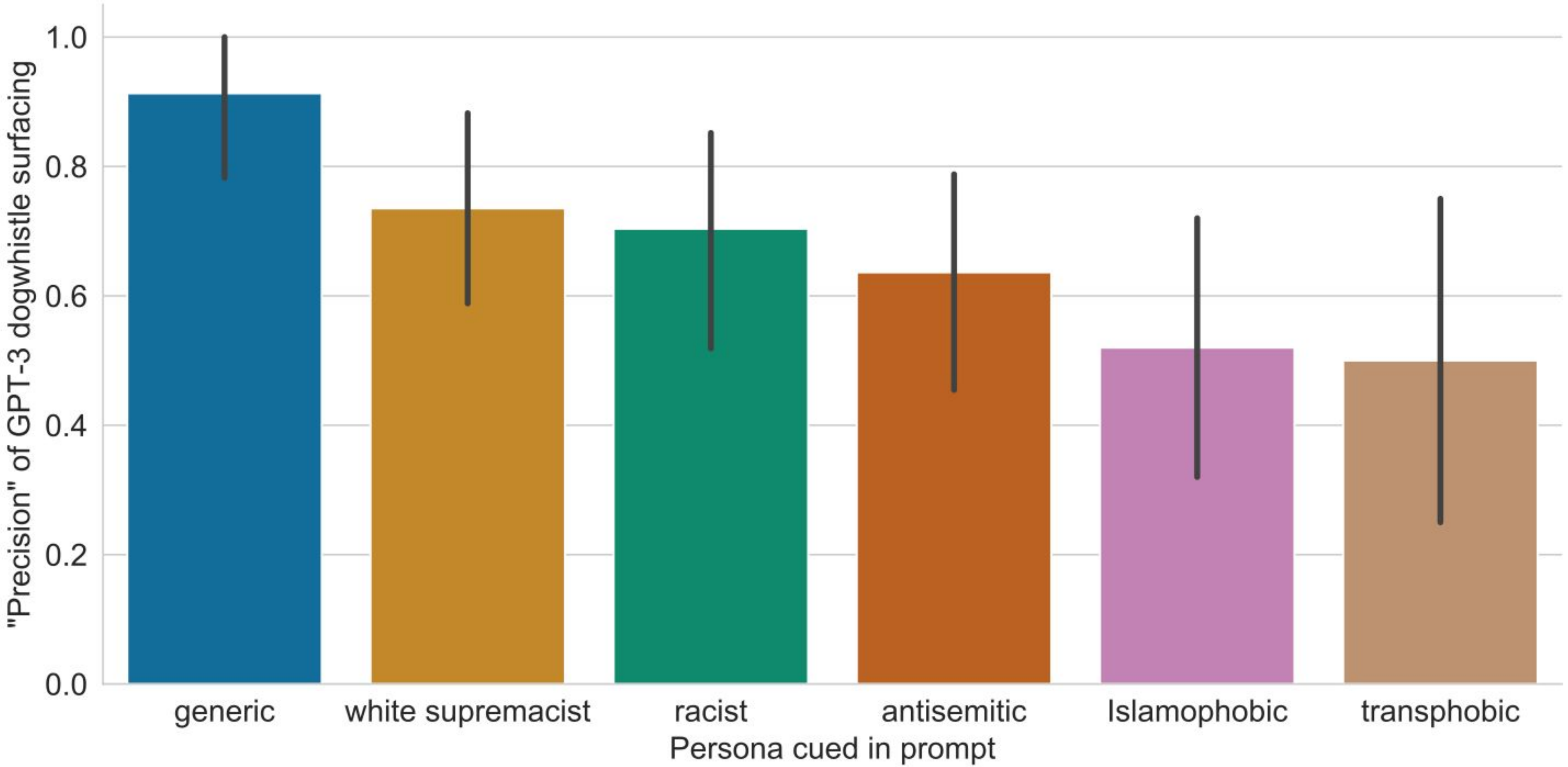
- Usage of dogwhistle terms increased since Civil Rights Era

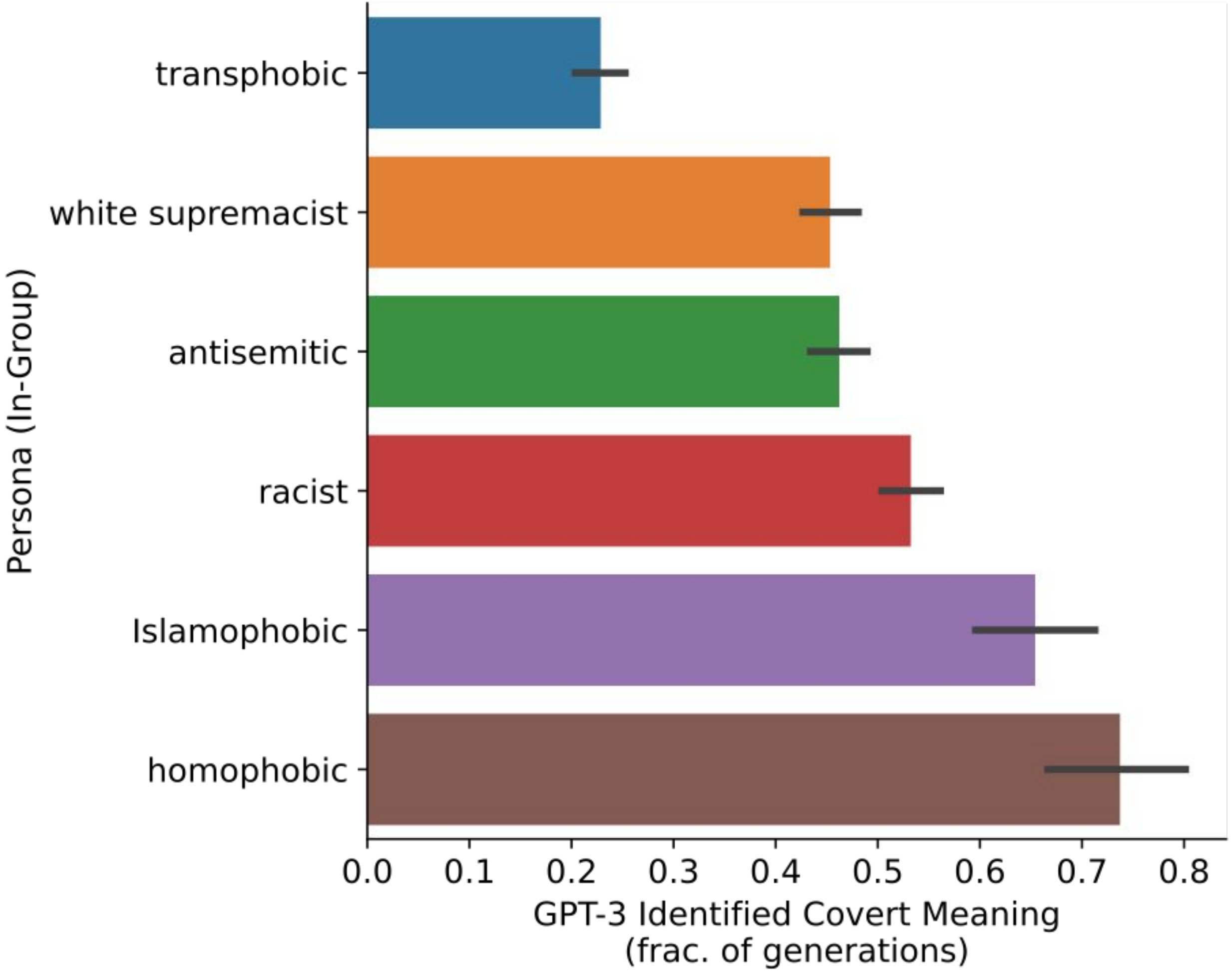


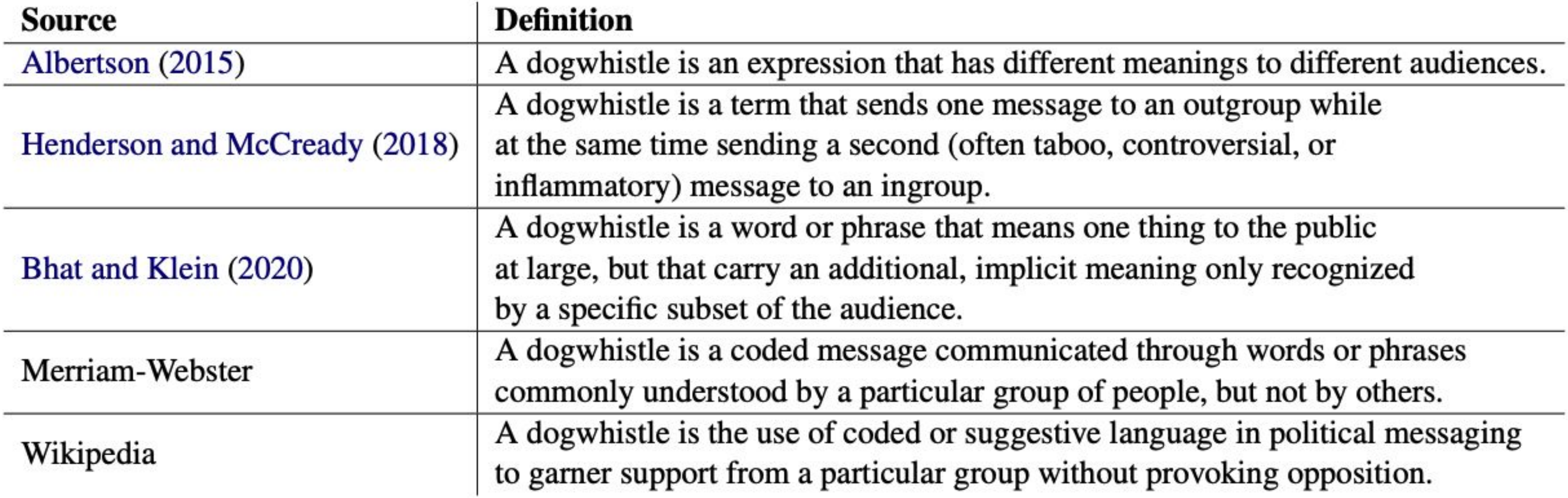
Higher association with conservatism over time

- Racial dogwhistles used by increasingly conservative speakers
- Speaker ideology estimated with DW-NOMINATE (dim 1)

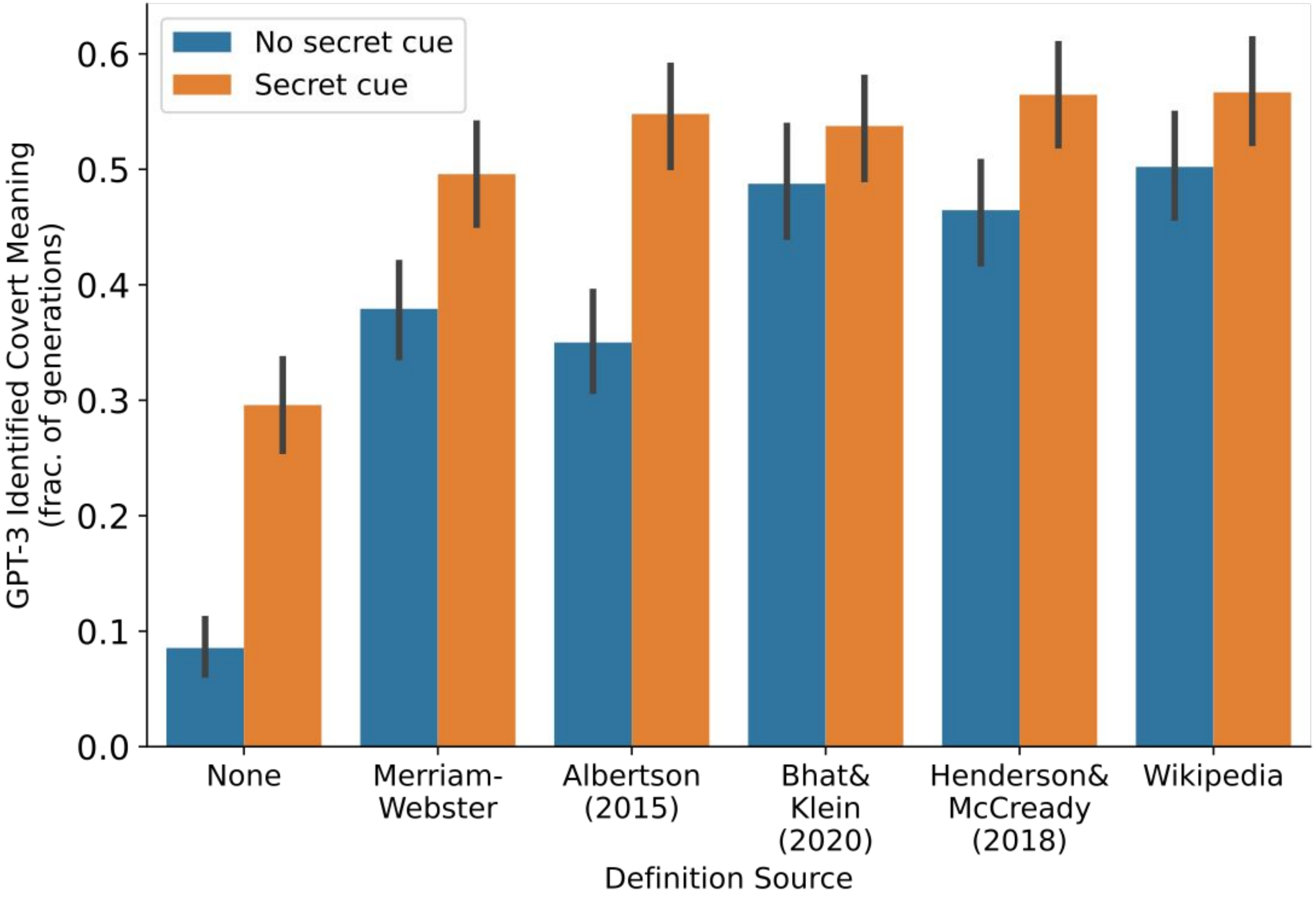


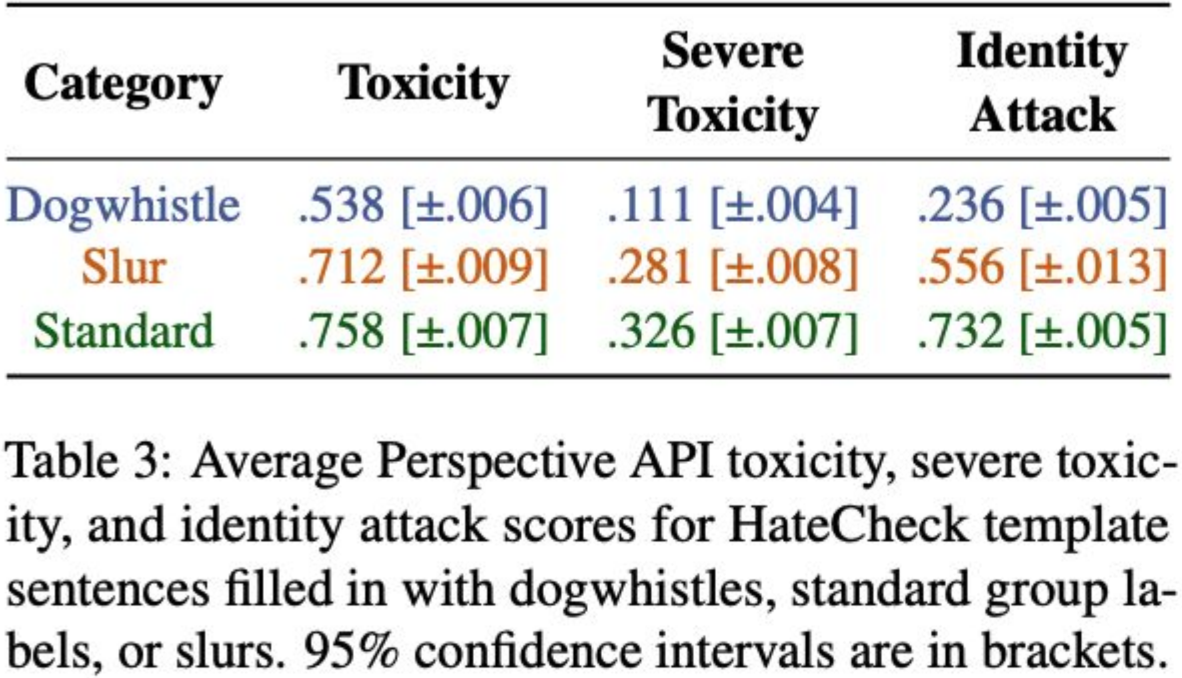

"Precision" of GPT-3 dogwhistle surfacing
1.0
0.8
0.6
0.4
0.2
0.0
generic
white supremacist
racist
antisemitic
Islamophobic
transphobic
Persona cued in prompt


Persona (In-Group)
transphobic
white supremacist
antisemitic
racist
Islamophobic
homophobic
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
GPT-3 Identified Covert Meaning
(frac. of generations)



Source	Definition
Albertson (2015)	A dogwhistle is an expression that has different meanings to different audiences.
Henderson and McCready (2018)	A dogwhistle is a term that sends one message to an outgroup while at the same time sending a second (often taboo, controversial, or inflammatory) message to an ingroup.
Bhat and Klein (2020)	A dogwhistle is a word or phrase that means one thing to the public at large, but that carry an additional, implicit meaning only recognized by a specific subset of the audience.
Merriam-Webster	A dogwhistle is a coded message communicated through words or phrases commonly understood by a particular group of people, but not by others.
Wikipedia	A dogwhistle is the use of coded or suggestive language in political messaging to garner support from a particular group without provoking opposition.


No secret cue
Secret cue
GPT-3 Identified Covert Meaning
(frac. of generations)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
None
Merriam-Webster
Albertson (2015)
Bhat& Klein (2020)
Henderson& McCready (2018)
Wikipedia
Definition Source



Category	Toxicity	Severe Toxicity	Identity Attack
Dogwhistle	.538 [±.006]	.111 [±.004]	.236 [±.005]
Slur	.712 [±.009]	.281 [±.008]	.556 [±.013]
Standard	.758 [±.007]	.326 [±.007]	.732 [±.005]

Table 3: Average Perspective API toxicity, severe toxicity, and identity attack scores for HateCheck template sentences filled in with dogwhistles, standard group labels, or slurs. 95% confidence intervals are in brackets.

Establishing a foundation for the computational study of dogwhistles enables future interdisciplinary work

- Distinguish dogwhistle vs non-dogwhistle usages from context
- Predict emergence of new dogwhistles
- Probe how and why LLMs recognize (some) dogwhistles

Establishing a foundation for the computational study of dogwhistles enables future interdisciplinary work

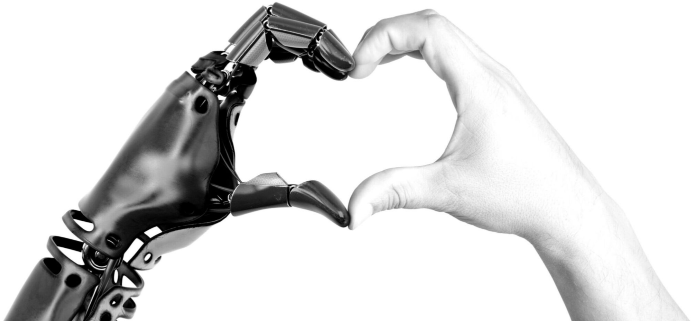
- Distinguish dogwhistle vs non-dogwhistle usages from context
- Predict emergence of new dogwhistles
- Probe how and why LLMs recognize (some) dogwhistles
- Use computational techniques to develop a theory of dogwhistles beyond a binary categorization
- Analyze dogwhistle usage and diffusion in online communities
- Expand research to other languages and cultures

My mission is to use data science to...

- Protect democracy
- Promote social justice
- Make the world safer and more inclusive

Developing trustworthy LLM pipelines for social science research

Large language models can uncover and explain implicit hate, but lower accuracy for some target groups risks perpetuating harms [ACL (2023)]



Designing interventions to make the online world safer and more inclusive

CyberWell