From Dogwhistles to Bullhorns
Unveiling Coded Rhetoric with Language Models

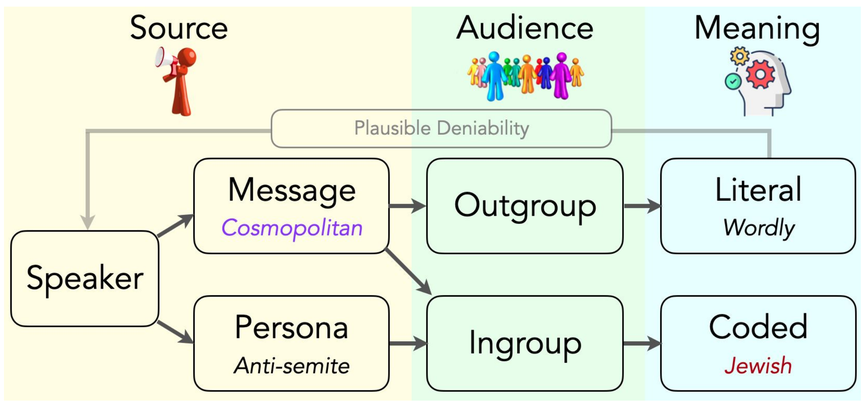
Julia Mendelsohn
UMSI
juliame@umich.edu

Ronan Le Bras
AI2
ronanlb@allenai.org

Yejin Choi
UW / AI2
yejinc@allenai.org

Maarten Sap
CMU / AI2
maartensap@cmu.edu

Dogwhistles are expressions that "send one message to an out-group and a second (often taboo, controversial, or inflammatory) message to an in-group" [Henderson and McCready, 2018]



"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."
~Sen. Josh Hawley (R-MO, 2019)

Understanding dogwhistles is important for NLP & CSS

Meaning depends on identity, context & *multiple* audiences

Mechanism of political influence and persuasion

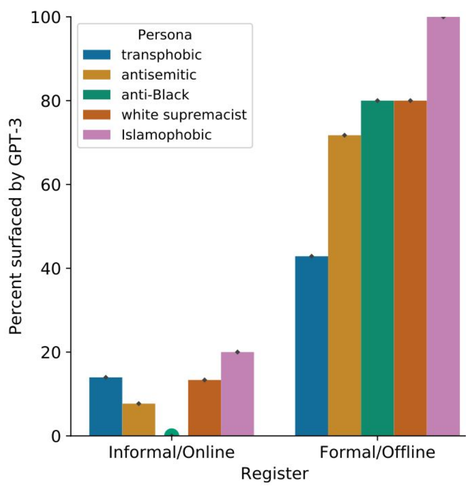
Enables abusive rhetoric that evades content moderation

Can GPT-3 *surface* dogwhistle expressions?

- Prompt with 5 different definitions
- Request examples in ~50 ways (including for different personae)
- Generate 5 outputs per prompt
- GPT-3 surfaces 45% of dogwhistles in our glossary
- Performance varies across persona (worst for transphobic)

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?

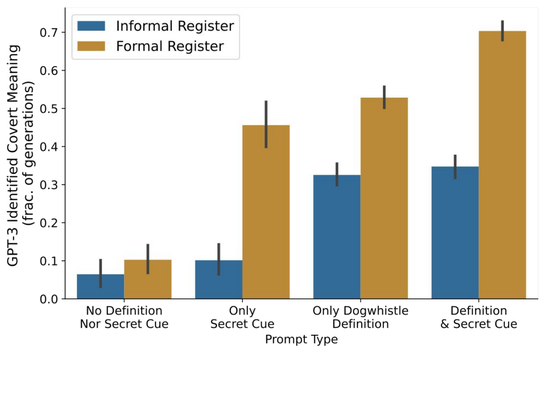
1. "Law and order"
2. "The silent majority"
3. "Family values"
4. "Welfare queens"
5. "Illegal aliens"

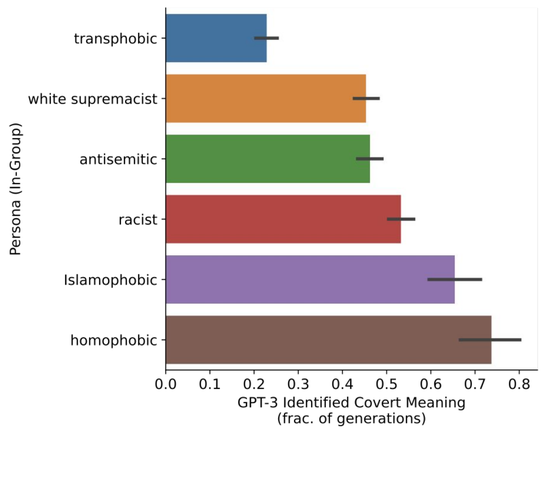


Can GPT-3 *identify* covert meanings?

- Prompt GPT-3 for meaning of all dogwhistles in glossary
- Manipulate dogwhistle definition and secret cue
- Better recognition for "formal" register, worst for transphobic persona

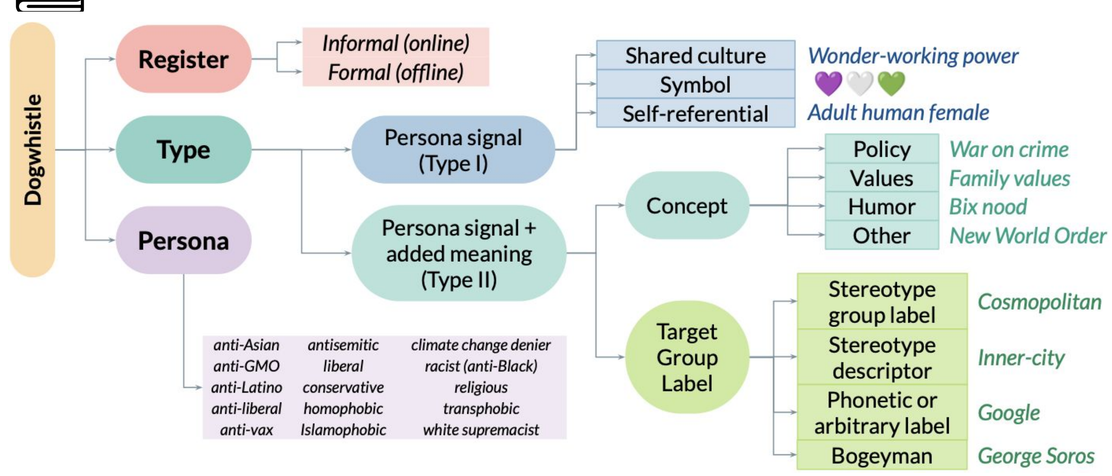
A dogwhistle is the use of coded or suggestive language in political messaging to garner support from a particular group without provoking opposition. For example, "cosmopolitan" secretly means "Jewish to many anti-Semitic people".





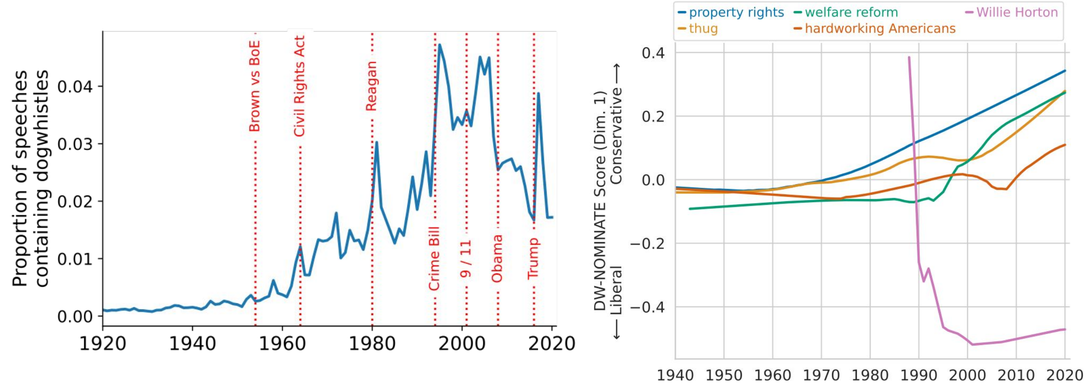
Check out our **glossary**:
dogwhistles.allen.ai
Scan the QR code for our paper!

Typology



Dogwhistles in US political speeches

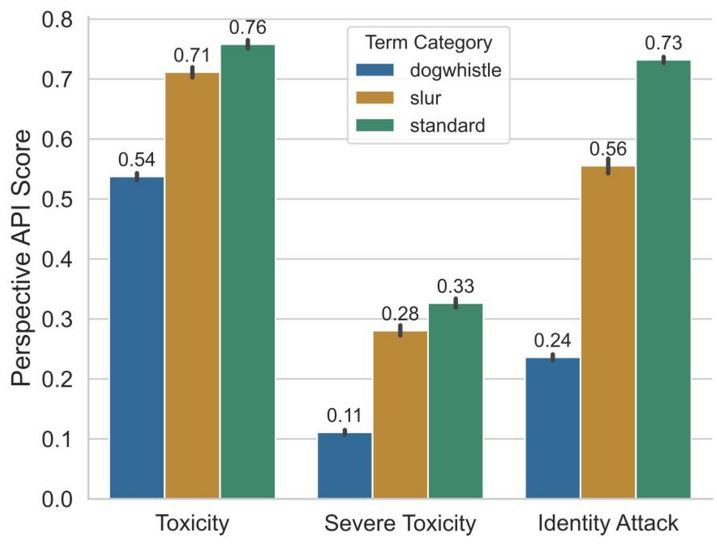
- Proportion of speeches containing racial dogwhistles in U.S. Congressional Record has increased since the Civil Rights Era
- Most used by more conservative speakers over time
- Many more patterns to explore using our dogwhistle glossary



Dogwhistles and toxicity detection

When dogwhistles replace slurs & "standard" group labels, hate speech is rated as less likely to be perceived as toxic.

- Hateful template sentences from HateCheck [Röttger et al., 2021]
- Google/Jigsaw Perspective API



Opening many directions for future work

- Distinguish dogwhistle vs non-dogwhistle usages from context
- Predict emergence of new dogwhistle terms
- Study impact of dogwhistles in broad range of NLP/NLG tasks
- Probe how and why LLMs recognize (some) dogwhistles
- Move beyond the dogwhistle as a binary variable
- Expand research to other languages and cultures