Patrick Gerard's Qualifying Exam

Combining Sociological Insight with Scalable Models

Committee: Kristina Lerman, Emilio Ferrara, Luca Luceri, Marlon Twyman, Swabha Swayamdipta

My Committee – Thank You All!





Committee: Kristina Lerman, Emilio Ferrara, Luca Luceri, Marlon Twyman, Swabha Swayamdipta

Other Thanks

Dr. Tim Weninger (so young!), Dr. Hans Hanley, SEA + Humans Lab!

* if you weren't shown / listed, it is because I hate you

Who Am I?

I'm Patrick Gerard, a second-year (hopeful) PhD student working under Prs. **Kristina Lerman** and **Emilio Ferrara**. I'm interested in the intersection of **machine learning and network science** and how they can be utilized to uncover the mechanisms of **information diffusion** and narrative evolution across media.

Ask me afterwards about my current work – I'm really excited about it :0

Who Am I?

patrickgerard.co to click on papers*

Recent Timeline

March 2025: ICWSM Paper Accepted 🎉
Fear and Loathing on the Frontline: Decoding the Language of Othering by Russia-Ukraine War Bloggers

January 2025: ICWSM Paper Accepted 🎉
Modeling Information Narrative Evolution on Telegram During the Russia-Ukraine War

September 2024: Stanford ESRG Talk 🎤
Gave a talk on narrative evolution and othering frameworks for LLM-guided community analysis.

June 2024: Interview with CNBC 🎙️
Featured for my work analyzing Truth Social and the rise of fringe platforms.

Ask me afterwards about my current work – I'm really excited about it :0

* this presentation will also be published there

Fear and Loathing on the Frontline: Decoding the Language of Othering by Russia-Ukraine War Blogger

Combining Sociological Insight with Scalable Models

Authors: Patrick Gerard, William Theisen, Tim Weninger, Kristina Lerman

Why This Matters

"Jews were not killed because they were human beings. In the eyes of the killers they were **not human beings**! They were **Jews**!"

– Elie Wiesel, Auschwitz Survivor and Nobel Peace Prize Laureate

Why This Matters

"One day everything seemed normal and then we were being called **cockroaches** and **snakes**."

– Jacqueline Murekatete, Rwanda Genocide Survivor

Why This Matters

"The problem isn't that perpetrators don't know they're doing wrong. It's that **they believe they're doing right**." [1]

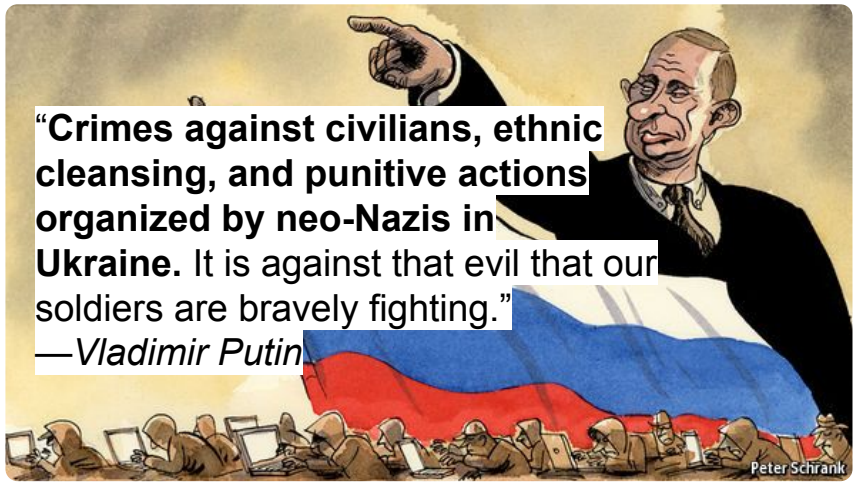
We must stop asking how people **ignore evil**—and start asking how they come to **celebrate it**.

Reicher et al. 2008

From Celebration to Construction: How Does Violence Become Justified?

"How does someone come to celebrate harm?"

It starts with a **story**.
A story about *who belongs*, *who threatens*, and *who must be stopped.*



What is Othering?

No clear, universal definition, but several overlapping **workable definitions**

*"Othering is the construction of a **positive self and a negative other**."*
— Pettersson & Sakki, 2017

*"Othering is the **outcasting of certain groups** based on **arbitrary attributes**."*
— Sakki & Castrén, 2022

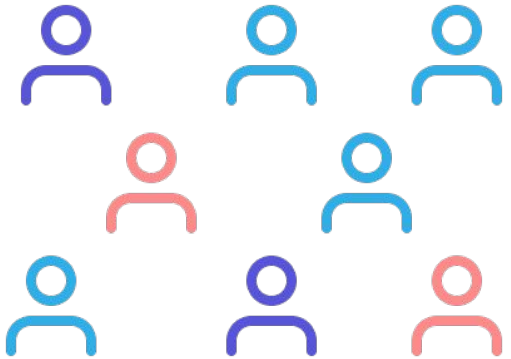
*"Othering is a **social process** whereby a **dominant group** or person uses negative attributes to define and **subordinate others**."*
— Canales, 2010

What is Othering?

Difference is not the problem. Meaning is.

Humans vary—culturally, geographically, racially, religiously.

What matters is **how society assigns meaning** to these differences



Duckitt 2003; Joffe 1999; Reicher et al. 2008

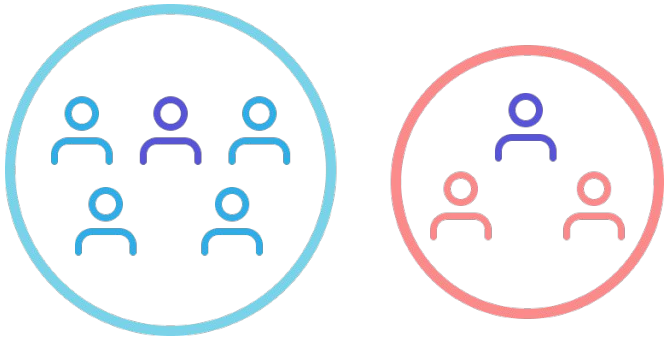
What is Othering?

The ingroup is assembled.

A cohesive ingroup identity is built, often around shared culture, values, or history.

This identity gains power not just from similarity, but from **contrast:**

*We are who we are, because **we are not them**.*



Reicher et al. 2008 (Step 1); Jetten et al. 1997

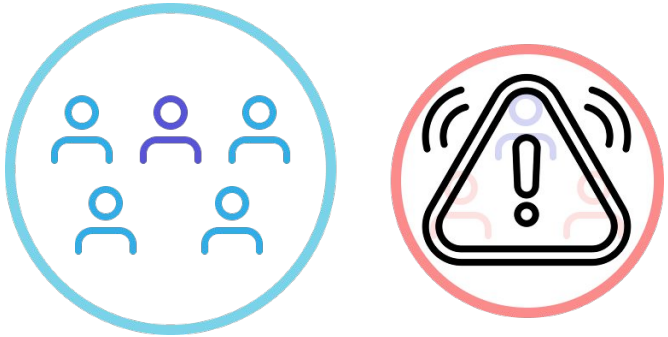
What is Othering?

From difference to danger.

The outgroup is cast as a **threat** to the ingroup's identity, values, or survival.

Accompanied by:

- **Depersonalization** of outgroup members
- **Scapegoating** and fear
- Often framed as **existential crisis**



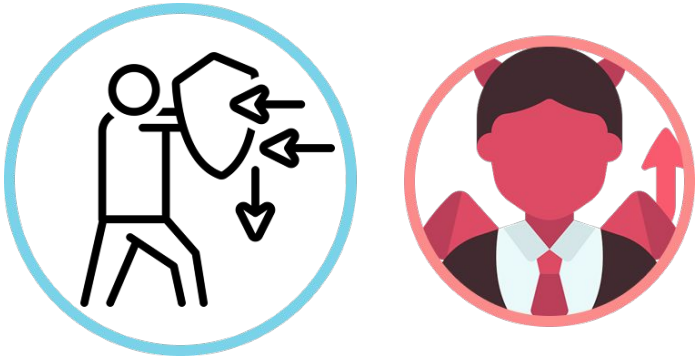
Sakki & Castrén 2022; Reicher et al. 2008 (Step 3)

What is Othering?

Prejudice Becomes a Moral Project.

The final step is **moral inversion**:

- Harsh treatment of the outgroup is **justified**
- Ingroup defense becomes a **cause for celebration**
- **Status hierarchies** are reinforced
- **Prejudice is perpetuated** in the name of good



Kennedy et al. 2023; Fiske & Rai 2014; Reicher et al. 2008 (Step 5)

What is Othering?

We operationalize existing definitions for large-scale analysis

Our Definition:

We define *othering* as a discursive process that constructs an **ingroup–outgroup boundary** and frames the **outgroup as morally or existentially problematic**.

Kennedy et al. 2023; Fiske & Rai 2014; Reicher et al. 2008

What is Othering – How Does it Differ from Hate/Fear Speech?

Othering
A **social process** that constructs a boundary between “us” and “them,” portraying the outgroup as threatening, inferior, or morally corrupt. It unfolds through **narrative structure**, not isolated slurs or sentiment.

Hate Speech
Typically **overt**, hostile language targeting a group based on identity (e.g., slurs, insults).

Fear Speech
Language that frames a group as a **danger**, often without explicit hatred. Centers on **threat amplification** (e.g., “They’re coming for your children”).

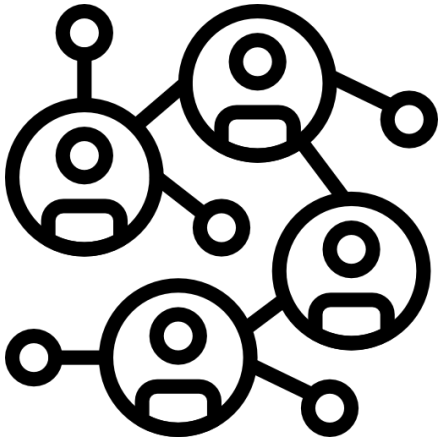
*Hate and fear speech are **symptoms**. **Othering is the structure** behind them: the process that makes exclusion and harm seem reasonable and even necessary.*

Existing Work — Theoretical Work

Social Identity and Justification of Harm
Reicher et al. (2008) outline a 5-stage process where identity construction, threat perception, and moralization transform harm into virtue.

Symbolic and Existential Threats
Joffe (1999) and *Duckitt (2003)* describe how perceived threats to culture, identity, and safety fuel prejudice and outgroup hostility.

Moral Foundations of Violence
Fiske & Rai (2014) propose that even extreme violence is often framed as morally necessary within relational contexts. *Hoover et al. (2021)* show group-based moral values predict justification of hate when outgroups are seen as morally violating.



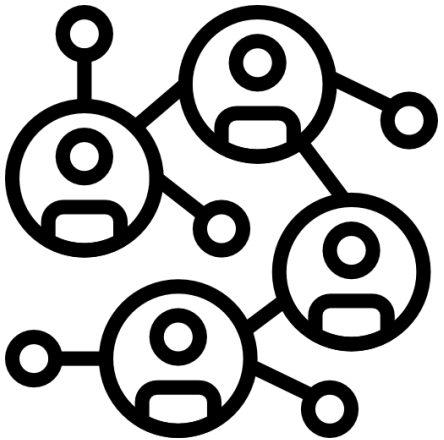
Existing Work — Theoretical Work

Taxonomy-Relevant Concepts

***Depersonalization**:* Outgroup members stripped of individuality (*Sakki & Castrén, 2022*)

***Moral exclusion**:* Outgroups placed outside the circle of moral concern (*Opotow, 1990*)

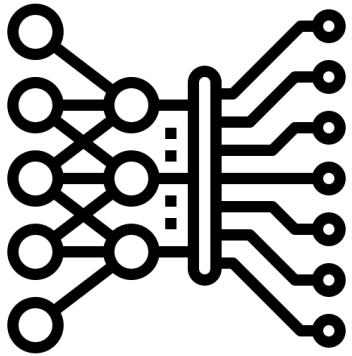
***Group-based threat**:* Perceived danger activates defensive cohesion (*Duckitt, 2003*)



***Language constructs group boundaries** and makes exclusion appear justified, even **morally necessary**.*

Existing Work — Computational Work on Harmful Speech

- **Hate speech detection**, often framed as a binary or multi-class classification task *(e.g., Davidson et al. 2017; Founta et al. 2018)*
- **Fear speech and incitement**, emphasizing emotional tone and downstream risk *(e.g., Saha et al. 2023; Mathew et al. 2020)*
- **Toxicity prediction**, including models deployed by platforms for moderation *(e.g., Perspective API, Borkan et al. 2019)*
- **Coded and subtle language**, where recent work explores moral framing, sarcasm, and dog whistles *(e.g., Vidgen et al. 2021)*

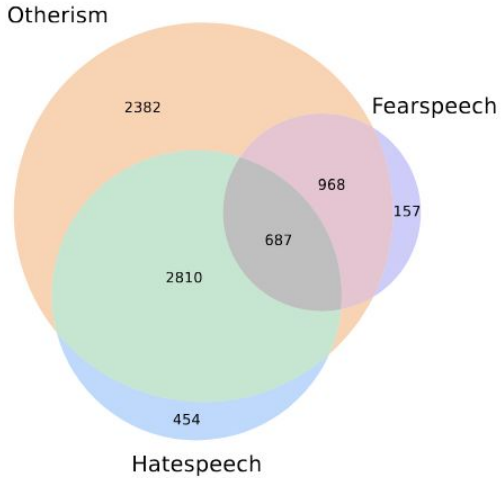


*Rather than modeling how exclusion is **constructed and reinforced**, most work focuses on **classifying individual instances** of harmful language.*

Key Gaps in the Literature

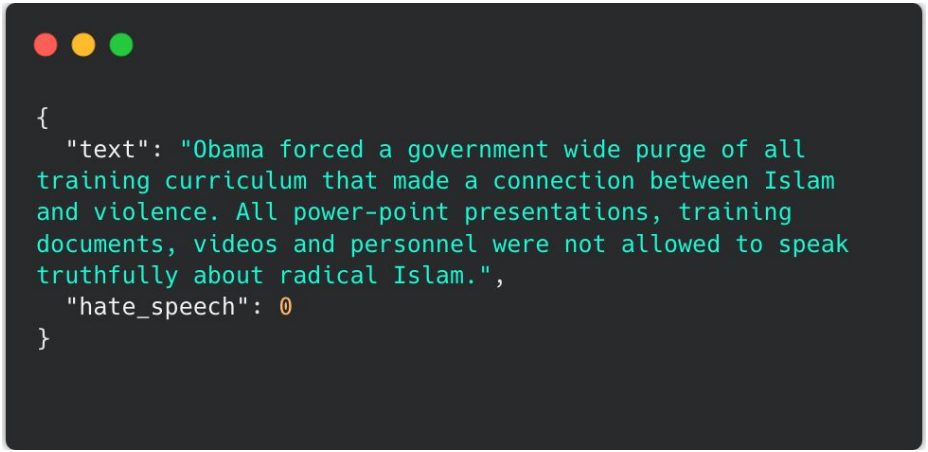
No Operationalized Taxonomy of Othering

Despite rich sociological theory, there is no operationalized taxonomy of othering suitable for NLP; existing labels like hate or fear are too coarse to capture its structure and moral logic.



Venn diagram showing **overlap** between othering, **fear speech**, and **hate speech** in the Gab corpus.

Key Gaps in the Literature - Why Existing Labels Fall Short



```
{
  "text": "Obama forced a government wide purge of all
training curriculum that made a connection between Islam
and violence. All power-point presentations, training
documents, videos and personnel were not allowed to speak
truthfully about radical Islam.",
  "hate_speech": 0
}
```

Example from a standard hate speech classifier

This is not just about tone—it's about structure. The post builds a **threat narrative** that primes readers to see a group as dangerous and harm as legitimate.

Traditional classifiers **miss this** entirely.

Key Gaps in the Literature

No Theory-Aligned, Scalable Annotation Pipeline

Existing models rely on crowd-labeled data optimized for speed, but we lack a validated pipeline that scales theory-grounded annotations with high fidelity.

No Empirical Study of Othering as a Process

Lack of analysis connecting othering to event timelines, moral co-framing, or attention metrics

What Kind of Environment Reveals These Gaps?

Our Data Testbed: Telegram & Beyond

What Kind of Environment Reveals These Gaps?

We need a setting where:

- Othering **evolves** in response to **real events**
- **Morality and violence** are discursively **intertwined**
- **Language adapts** quickly to context and platform

*To address these gaps, we need a **real-world context** where othering is not only present, but **evolving, consequential**, and **morally charged**.*

What Kind of Environment Reveals These Gaps?

Social media platforms—especially during conflict—**offer precisely this environment.** They allow us to observe how othering is **formulated, moralized, and socially reinforced** in real time.

We can move from studying **static content** to studying **mechanisms**: how exclusion is made thinkable, when it escalates, and how it spreads.

A Two-Part Empirical Testbed

To study the social mechanisms of othering, we focus on two distinct platforms:

Telegram: High-Stakes Conflict Discourse

- Primary testbed: Russian & Ukrainian warbloggers
- High-conflict, morally charged discourse
- Ideal for observing how othering **emerges, escalates, and adapts** in real time

Gab: Validation in a non-conflict setting

- U.S.-based, ideologically extreme platform
- Low moderation, decentralized language use
- Tests whether patterns of **moralized exclusion** generalize beyond wartime

Telegram and Gab

Russian and Ukrainian Warbloggers

High-Stakes, Real-Time Narrative Warfare

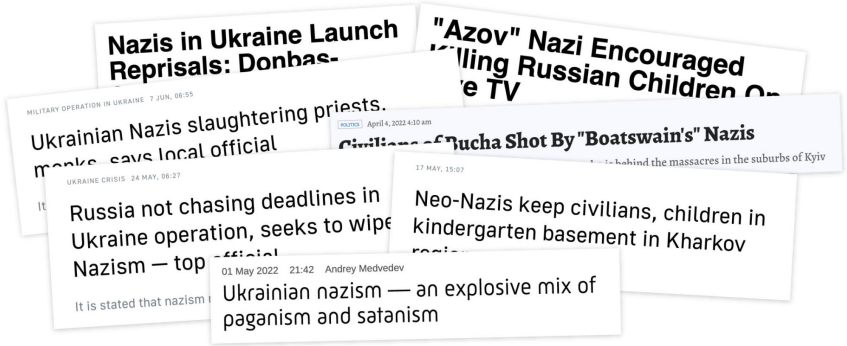
The Russia–Ukraine war has generated a massive volume of politicized discourse where **identity, threat, and moral justification** are actively contested.

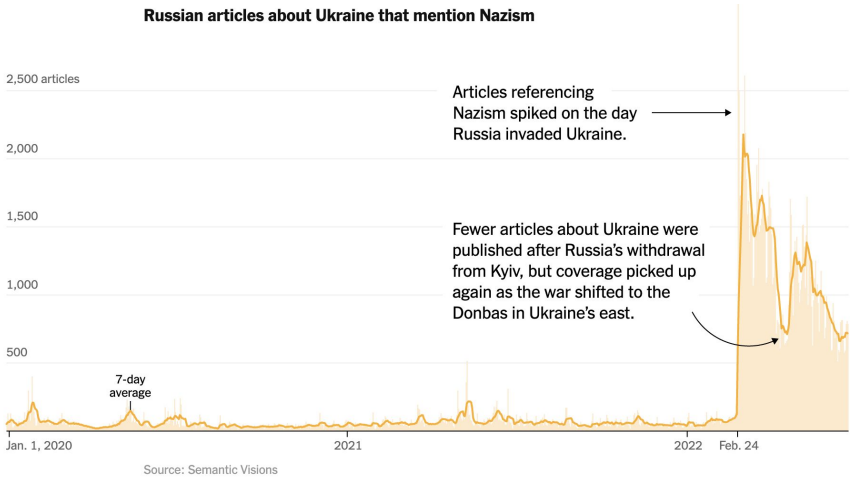
Ideal Testbed for Escalating Othering

Warbloggers use Telegram to frame the enemy, rally the ingroup, and construct meaning around violence: **a real-world laboratory for observing othering in action**.

Telegram is a popular messaging and broadcast platform which became the **most downloaded social app** in Russia and Ukraine and was used by **~39% of Ukrainians** and **~19% of Russians** as a **primary news source** as of 2022 (Oleinik 2024).

Russian and Ukrainian Warbloggers





Mentions of **Nazism** in Russian media **spiked on the day of the invasion** and continued as the war escalated, framing **Ukraine as a morally deviant, existential threat**. These narratives were amplified by war bloggers and state-aligned channels, offering a real-time window into how **othering is constructed, justified, and sustained during conflict**.

Data Source and Scope: Telegram

Posts from **Russian- and Ukrainian-leaning Telegram channels**

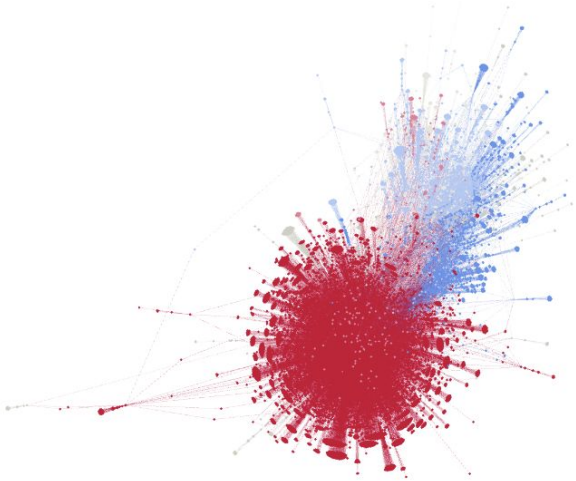
Collected from **Oct 2015–Aug 2023**

Final analysis focused on **~8.6M posts** from **568 channels**

→ **243 pro-Ukrainian** (4.2M posts)

→ **325 pro-Russian** (4.4M posts)

Data primarily in **Russian and Ukrainian**



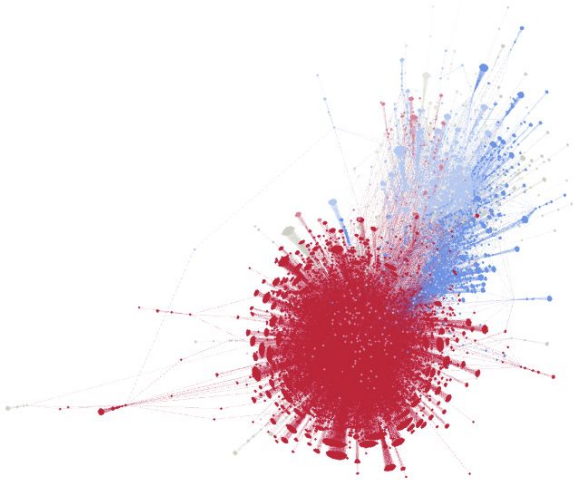
Co-reference network of Telegram warbloggers. Nodes represent channels, colored by inferred stance:
🔴 **Pro-Russian** (red),
🔵 **Pro-Ukrainian** (blue),
⚪ **Unaffiliated/Neutral** (grey)

Data Source and Scope: Telegram

Posts from **Russian- and Ukrainian-leaning Telegram channels**

Community Labeling:

- Constructed author network: edge if Channel A forwarded B
- Used bios + recent posts to hand-label 100 seed channels
- Applied **label propagation** → validated with 90%+ accuracy



Edges reflect shared content and profile similarities. Most grey nodes focus on local or apolitical topics (e.g., trading, logistics).

Data Source and Scope: Gab

Posts from U.S.-based Gab accounts

- Collected from **June 2016 to August 2021**
- Data is English, sourced from a platform with **minimal content moderation**
- Lets us check whether patterns of **moralized exclusion** appear **outside conflict** and in a **very different discursive environment**

Gab: A low-moderation, U.S.-based platform known for **far-right and extremist discourse** (Saha et al., 2023).

From Conflict Discourse to Computational Inquiry

What We Ask, and How We Answer

Guiding Research Questions

In this discourse environment, we ask:

RQ1 – Temporal Dynamics

How does the use of othering language by Russian and Ukrainian war bloggers on Telegram **change over the course of the war**?

RQ2 – Moral Framing

How does the **moral and othering language** used by war bloggers **interact** and vary by group?

Guiding Research Questions

In this discourse environment, we ask:

RQ3 – Attention and Influence

How does portraying the target group as the other affect **social attention**?

RQ4 – Times of Crisis

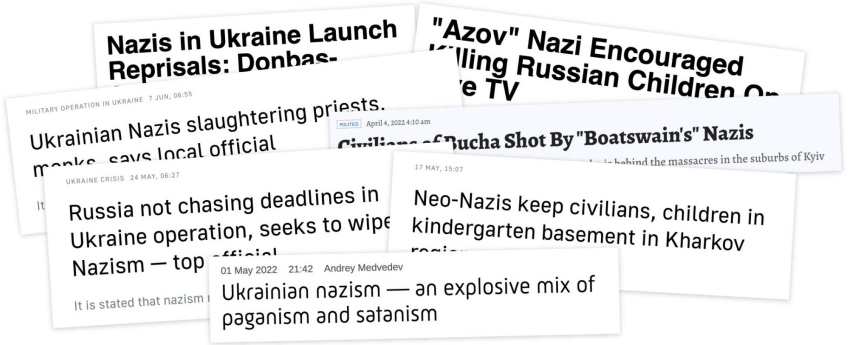
Does use of othering language **intensify during times of crisis**, and in what ways are these behaviors **more strongly rewarded**?

Our Approach

From Theory to Scalable Detection and Analysis

Operationalize Theory → Taxonomy of Othering

Grounded in Reicher et al. (2008), Joffe (1999), Fiske & Rai (2014)



Four core types:

- Threats to Culture or Identity
- Threats to Survival or Physical Security
- Vilification
- Explicit Dehumanization

Our Approach

A four-part taxonomy of othering language, grounded in social psychology. Each category reflects a distinct rhetorical mechanism used to justify exclusion or harm: ranging from cultural threat to explicit dehumanization.

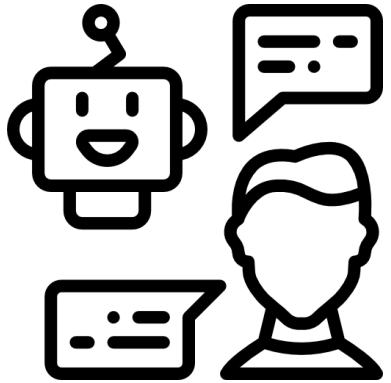
Category	Definition	Example Post
Threats to Culture or Identity	Frames the outgroup as a danger to the ingroup's cultural or social survival—challenging its values, language, or traditions [21, 30, 34, 41].	*"The erosion of the Russian language in Ukrainian schools: Ukrainian policymakers pushing to erase the Russian tongue risk severing the threads that weave together our history."*
Threats to Survival or Physical Security	Portrays the outgroup as an existential threat to the ingroup's physical safety, justifying preemptive hostility [21, 30, 34, 41].	*"Zelensky's regime has accumulated 30 tons of plutonium and 40 tons of enriched uranium at the Zaporizhia NPP [...] the regime really is on the verge of creating its own nuclear bomb! And hundreds of 'dirty' bombs can be made from such a quantity of radioactive material!"*
Vilification / Villainization	Casts the outgroup as inherently evil or immoral, legitimizing resistance or aggression [21, 30, 34].	*"Because these Ukronazi girls can fight only by hiding behind hostages. All their courage went down the drain in chants and slogans like 'hang the Muscovite.' But when the Russians came, they shit themselves, just like their Bandera."*
Explicit Dehumanization	Compares the outgroup to animals, objects, or supernatural threats, paving the way for extreme violence [21, 30, 34].	*"These are zombies, who may have been brothers before, but over the past 8 years, from the bite of Nazism and Banderization, they have turned into non-humans. That is why our army calls on all brothers to lay down their arms, so that we can distinguish a brother from an infected zombie, who can only bite and infect."*

Our Approach

From Theory to Scalable Detection and Analysis

Scalable Annotation Pipeline (LLM-Assisted)

- Human-labeled data → GPT-4 alignment → distilled into open-source LLM
- Evaluated using both **inter-annotator agreement** (e.g., $\kappa > 0.85$) and **ML performance metrics** (F_1 scores, precision/recall)
- Human agreement ensures **conceptual validity**; ML metrics ensure **scalability and consistency**



Can now annotate thousands of posts with theory-aligned precision

Our Approach

Annotation Pipeline – Example Output

User Message
"These Ukronazis tore apart our ancestors' resting place. They want to tear down everything dear to us. But when the Russians came, they shit themselves, just like their Bandera"
System Output
{'Threats to Culture or Identity': 1, 'Threats to Survival or Physical Security': 0, 'Vilification/Villainization': 1, 'Explicit Dehumanization': 0, 'None': 0, 'explanation': 'The text describes local Nazis desecrating a historic Russian cemetery, in a way that represents a threat to cultural identity and vilifies the opposing group.'}

Table 5. Example conversation showing model annotation of hate speech and othering language, formatted as structured output.

Results

Tracing the Patterns and Dynamics of Othering in the Wild

RQ1: Temporal Dynamics

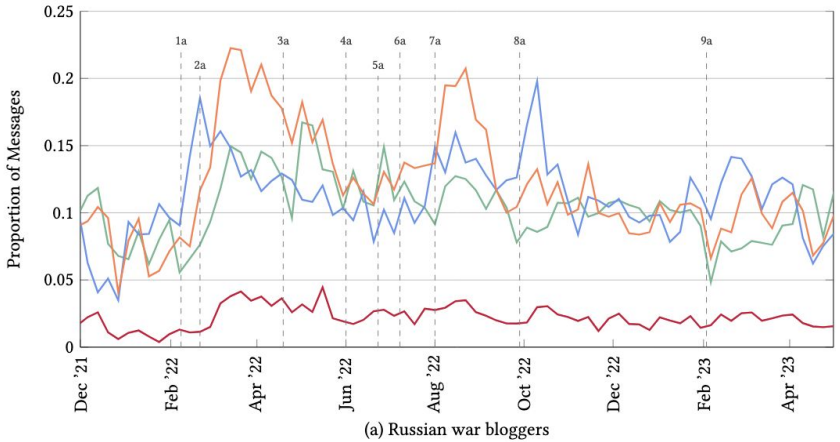
How Does Othering Change Over Time?

Othering **intensifies following military and political shocks**, but with different rhetorical patterns across groups.

*Othering follows the shockwaves of war, but each side speaks a **different language of threat**.*

RQ1: Temporal Dynamics - Russian Warbloggers

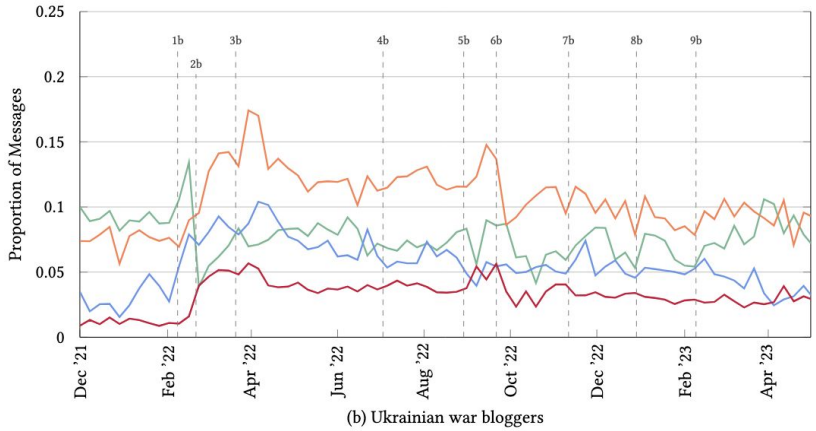
Date	Event	Key
2022-02-08	Putin claims allowing Ukraine to join NATO would increase the prospects of a Russia-NATO conflict that could turn nuclear.	1a
2022-02-21	Putin cites Nazism in Ukraine in speech legitimizing upcoming invasion.	2a
2022-02-24	Russia invades Ukraine.	-
2022-04-19	Russia officially pivots to 'next phase' of war. Russia shifted its troops from the Kyiv offensive to Ukraine's eastern Donbas region, and the amassed forces launched a broad attack there on April 18. Ukraine called it a "new phase of the war."	3a
2022-06-01	The Biden administration authorizes an 11th presidential drawdown of security assistance to Ukraine valued at up to $700 million.	4a
2022-06-23	The Biden administration authorizes a 13th presidential drawdown of security assistance to Ukraine valued at up to $450 million.	5a
2022-07-08	The Biden administration announces $400 million in additional security assistance for Ukraine.	6a
2022-08-01	The Biden administration announces $550 million in additional security assistance for Ukraine.	7a
2022-09-28	United States Department of Defense announces approximately $1.1 billion in additional security assistance for Ukraine.	8a
2023-02-03	United States Department of Defense announces a significant new package of security assistance for Ukraine, including the authorization of a presidential drawdown of security assistance valued at up to $425 million, as well as $1.75 billion in Ukraine Security Assistance Initiative (USAI) funds.	9a



**Key events were identified using methods adapted from prior work (Gerard et al. 2024). Each event corresponds to a major political or military development that was prominently discussed by Russian war bloggers.*

RQ1: Temporal Dynamics - Ukrainian Warbloggers

Date	Event	Key
2022-02-08	Putin claims allowing Ukraine to join NATO would increase the prospects of a Russia-NATO conflict that could turn nuclear.	1b
2022-02-21	Putin cites Nazism in Ukraine in speech legitimizing upcoming invasion.	2b
2022-02-24	Russia invades Ukraine.	-
2022-03-02	Russia captures Kherson.	-
2022-03-21	Russian troops used stun grenades and gunfire to disperse a rally of pro-Ukrainian protesters in the occupied southern city of Kherson on Monday.	3b
2022-03-21	Russia abandons Kherson.	-
2022-04-01	Reports of Russian atrocities in Bucha begin to surface.	-
2022-07-03	Russia captures Lysychansk, all of Luhansk Oblast	4b
2022-08-29	Ukraine launches first major counteroffensive.	5b
2022-09-21	Ukraine forces Russian retreat.	6b
2022-11-11	Ukraine recaptures Kherson.	7b
2022-12-29	Major Russian missile attack on infrastructure facilities in Kyiv, Kharkiv, Lviv, and other cities.	8b
2023-02-09	Russia launches second spring offensive.	9b



**Key events were identified using methods adapted from prior work (Gerard et al. 2024). Each event corresponds to a major political or military development that was prominently discussed by Ukrainian war bloggers.*

RQ1: Temporal Dynamics

How Does Othering Change Over Time?

Othering **intensifies following military and political shocks**, but with different rhetorical patterns across groups.

*Othering follows the shockwaves of war, but each side speaks a **different language of threat**.*

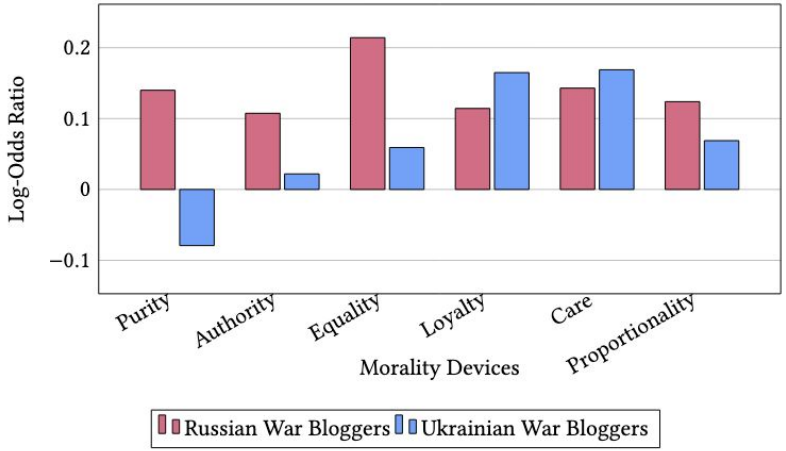
RQ2: Moral Framing

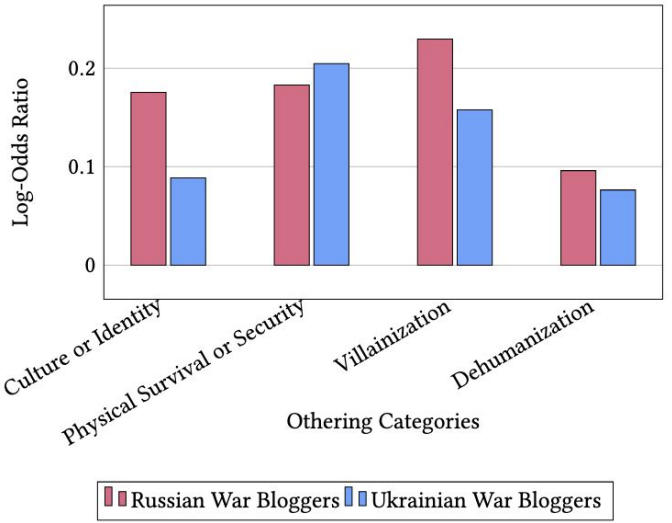
How is Othering Moralized Across Groups?

Russian channels moralize othering through sacred **duty** and **purity**; Ukrainian channel through **care** and **defense**.

*Each side **builds its enemy** with different **moral scaffolding**.*

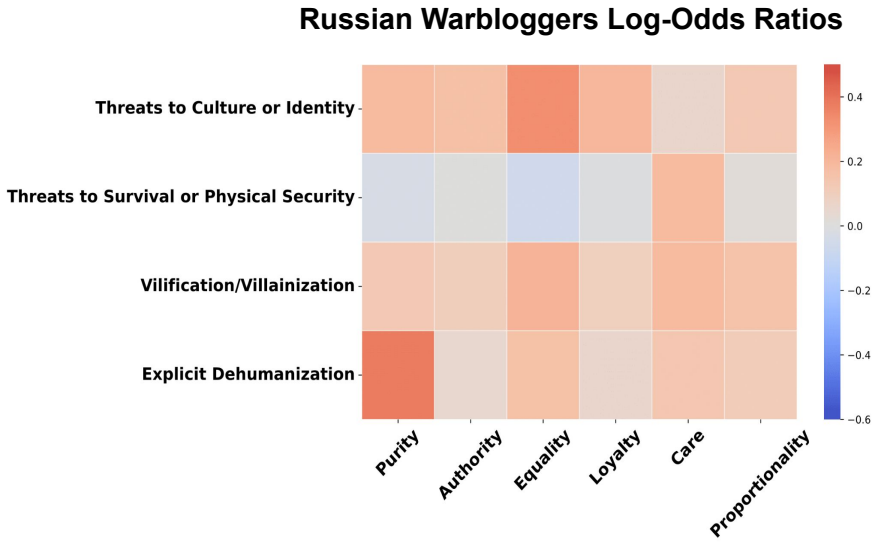
RQ2: Moral Framing

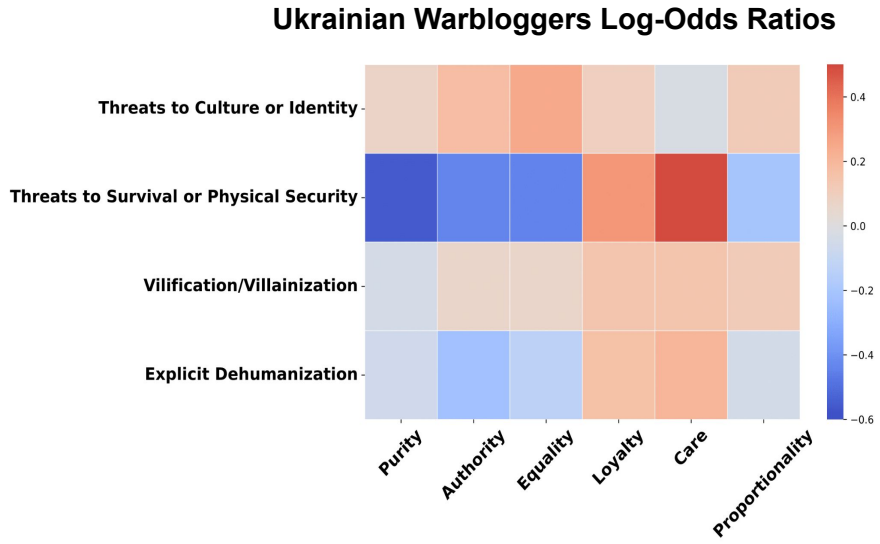




*Russian war bloggers lean on moral foundations like **purity, authority**, and **equality** when engaging in othering, while Ukrainian bloggers more often invoke **care**, **loyalty**, and **proportionality***

RQ2: Moral Framing





Moral co-framing patterns differ sharply across groups. ***Russian warbloggers*** *show consistent co-occurrence between othering and a* ***broad set of moral foundations****.* ***Ukrainian warbloggers*** *display more* ***selective alignment****, especially between* ***survival threats*** *and* ***fairness-related foundations*** *like care and proportionality.*

RQ2: Moral Framing

How is Othering Moralized Across Groups?

Russian channels moralize othering through sacred **duty** and **purity**; Ukrainian channel through **care** and **defense**.

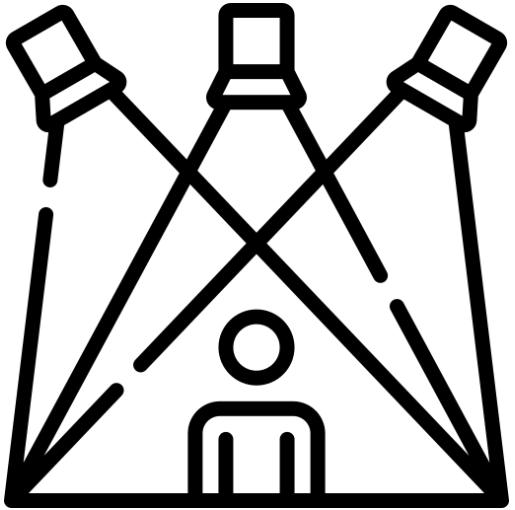
*Each side **builds its enemy** with different **moral scaffolding**.*

RQ3: Attention and Influence

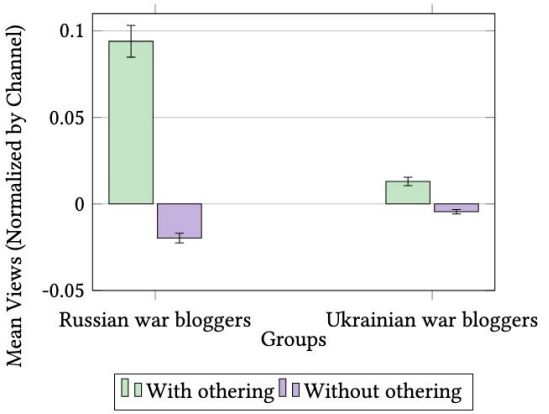
How does Othering Correlate with Visibility?

Posts that contain othering receive **significantly more views**. Channels that consistently use it tend to be **more central** in the network—suggesting reward mechanisms

*Othering aligns with **greater visibility**, and with discursive prominence.*



RQ3: Attention and Influence



(b) Comparison of mean views with and without othering (z-score channel-normalized).

	Centrality Metric	
Community	**Degree**	**Eigenvector**
Russian	0.254	0.333
Ukrainian	0.128	0.147

(a) Spearman correlation between a channel's proportion of messages with othering language and its degree and eigenvector centralities (all $p < 0.01$).

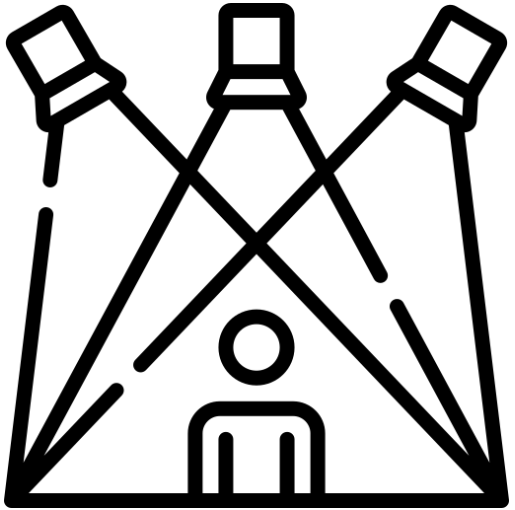
*Messages containing othering language receive **more views** on average, and channels that use othering more frequently tend to be **more central** in the discourse network. This pattern holds across both Russian and Ukrainian communities.*

RQ3: Attention and Influence

How does Othering Correlate with Visibility?

Posts that contain othering receive **significantly more views**. Channels that consistently use it tend to be **more central** in the network—suggesting reward mechanisms

*Othering aligns with **greater visibility**, and with discursive prominence.*



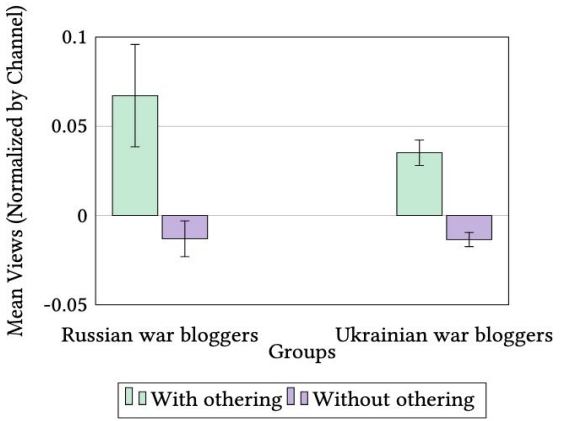
RQ4: Crisis & Reward Dynamics

How Does Othering Shift During Moments of Collective Threat?

During crisis periods, posts with othering **receive notably higher views**. Channels using othering during these moments occupy more central network positions.

*Crises coincide with surges in both the **reach** and **prominence of othering**.*

RQ4: Crisis Events



(b) Comparison of mean views with and without othering (z-score channel-normalized) following crises.

	Centrality Metric	
Community	**Degree**	**Eigenvector**
Russian	0.290 (+13.2%)	0.385 (+14.5%)
Ukrainian	0.177 (+32.1%)	0.136 (-7.8%)

(a) Spearman correlation between a channel's proportion of messages with othering language and its degree and eigenvector centralities following key events (all $p < 0.01$).

*The association between othering and both visibility (views) and network centrality strengthens following **major crisis events**. This reflects the same broader trend observed overall, but with even **greater magnitude during moments of heightened tension**.*

RQ4: Crisis & Reward Dynamics

How Does Othering Shift During Moments of Collective Threat?

During crisis periods, posts with othering **receive notably higher views**. Channels using othering during these moments occupy more central network positions.

*Crises coincide with surges in both the **reach** and **prominence of othering**.*

Key Takeaways

What We Learn from Tracing Othering in Online Discourse

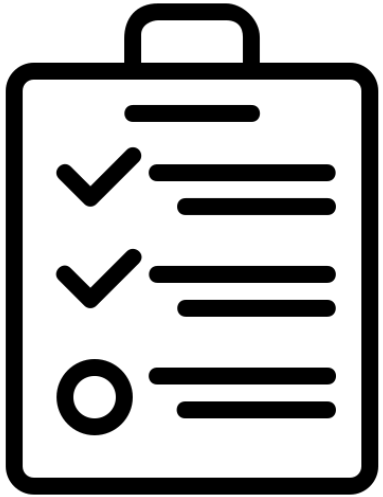
Key Takeaways: Modeling Othering in Conflict Discourse

Theory-Aligned, Scalable Detection

We introduce the first sociologically grounded framework for detecting othering, **achieving high agreement with humans** and scaling across millions of posts.

Moral Framing and Discursive Prominence

Groups deploy othering through distinct moral frameworks—these framings align with greater visibility and centrality, especially during moments of heightened tension.

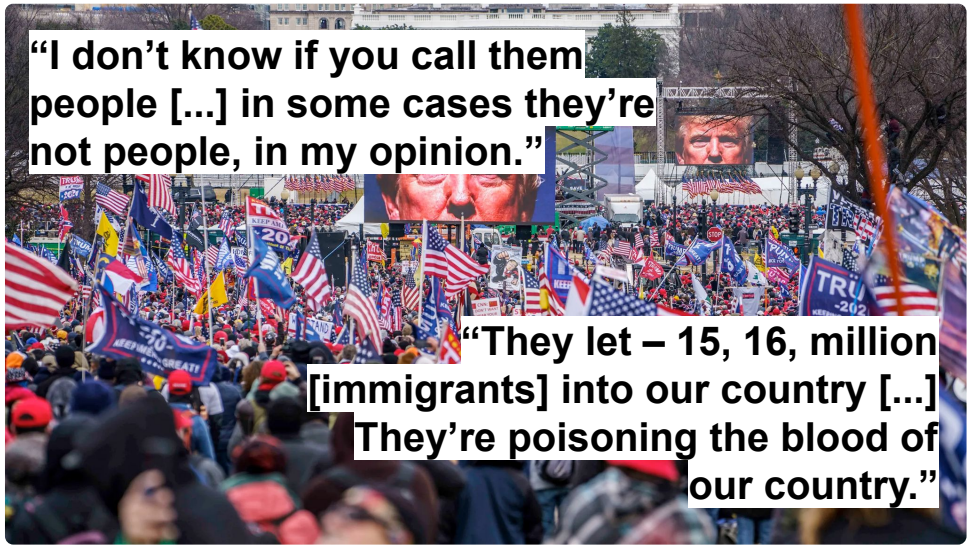


Key Takeaways: Modeling Othering in Conflict Discourse

Othering is not just a theoretical construct—it is now **detectable, interpretable, and scalable** in real-world discourse.

Why this Matters

To intervene early, we need to understand not just what people say, but how they **come to believe harm is justified**. This work is a step toward that

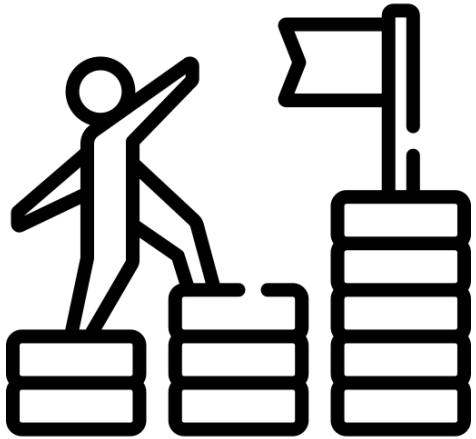


Next Steps — Tracing the Mechanics of Moralized Exclusion

Expand Contexts
Study how othering plays out in elections, public health scares, and protest movements

Analyze Evolved Psychological Mechanisms
Investigate how threat perception, group cohesion, and moral licensing are exploited.

Model the Dynamics of Narrative Adaptation
Track how morally framed othering shifts across platforms, audiences, and crises



Questions?

Works Cited

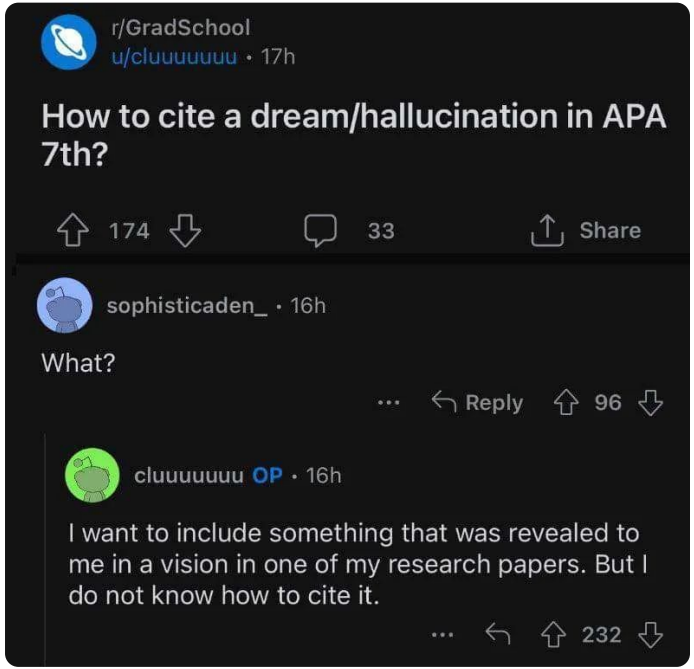
Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 512–515.

Duckitt, J. (2003). Prejudice and intergroup hostility.

Fiske, A. P., & Rai, T. S. (2014). *Virtuous violence: Hurting and killing to create, sustain, end, and honor social relationships*. Cambridge University Press.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., ... & Kourtellis, N. (2018, June). Large scale crowdsourcing and characterization of Twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1).

Gerard, P., Volkova, S., Penafiel, L., Lerman, K., & Weninger, T. (2024). Modeling information narrative detection and evolution on Telegram during the Russia-Ukraine war. *arXiv preprint arXiv:2409.07684*.



Works Cited

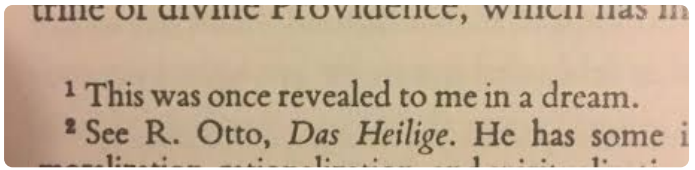
Hoover, J., Atari, M., Mostafazadeh Davani, A., Kennedy, B., Portillo-Wightman, G., Yeh, L., & Dehghani, M. (2021). Investigating the role of group-based morality in extreme behavioral expressions of prejudice. *Nature Communications*, *12*(1), 4585.

Jetten, J., Spears, R., & Manstead, A. S. (1997). Strength of identification and intergroup differentiation: The influence of group norms. *European Journal of Social Psychology*, *27*(5), 603–609.

Joffe, H. (1999). *Risk and 'the Other'*. Cambridge University Press.

Oleinik, A. (2024). Telegram channels covering Russia's invasion of Ukraine: A comparative analysis of large multilingual corpora. *Journal of Computational Social Science*, *7*(1), 361–384.

Opotow, S. (1990). Moral exclusion and injustice: An introduction. *Journal of Social Issues*, *46*(1), 1–20.



Works Cited

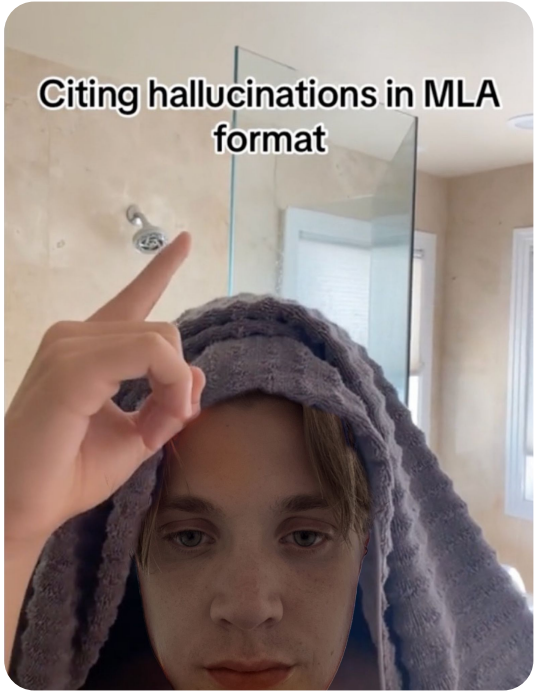
Pettersson, K., & Sakki, I. (2017). Pray for the fatherland! Discursive and digital strategies at play in nationalist political blogging. *Qualitative Research in Psychology*, *14*(3), 315–349.

Reicher, S., Haslam, S. A., & Rath, R. (2008). Making a virtue of evil: A five-step social identity model of the development of collective hate. *Social and Personality Psychology Compass*, *2*(3), 1313–1344.

Saha, P., Garimella, K., Kalyan, N. K., Pandey, S. K., Meher, P. M., Mathew, B., & Mukherjee, A. (2023). On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences*, *120*(11), e2212270120.

Sakki, I., & Castrén, L. (2022). Dehumanization through humour and conspiracies in online hate towards Chinese people during the COVID-19 pandemic. *British Journal of Social Psychology*, *61*(4), 1418–1438.

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019, August). Challenges and frontiers in abusive content detection. *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics.



Questions?

Appendix

Model Validation

Model Validation

Are We Capturing Othering Reliably – Russian Data

Category	Krippendorff's	Fleiss'
Threats to Culture or Identity	0.72	0.72
Threats to Survival or Physical Security	0.62	0.62
Vilification/Villainization	0.70	0.73
Explicit Dehumanization	0.68	0.65
None	0.70	0.73

(a) Krippendorff's & Fleiss' on Russian data

Category	Cohen's	Accuracy	F1
Threats to Culture or Identity	0.83	0.92	0.92
Threats to Survival/Security	0.75	0.82	0.80
Vilification/Villainization	0.80	0.90	0.90
Explicit Dehumanization	0.85	0.94	0.94
None	0.80	0.92	0.92

(b) Cohen's Kappa, Accuracy & F1 (GPT-4o vs. vote)

Table 2. (a) Inter-annotator agreement metrics on Russian war-blogger data; (b) model performance against majority vote

Model Validation

Are We Capturing Othering Reliably – Ukrainian Data

Category	Krippendorff's	Fleiss'
Threats to Culture or Identity	0.75	0.78
Threats to Survival or Physical Security	0.77	0.76
Vilification/Villainization	0.78	0.79
Explicit Dehumanization	0.80	0.80
None	0.77	0.78

(a) Inter-Annotator Agreement: Krippendorff's Alpha and Fleiss' Kappa for Ukrainian war bloggers data.

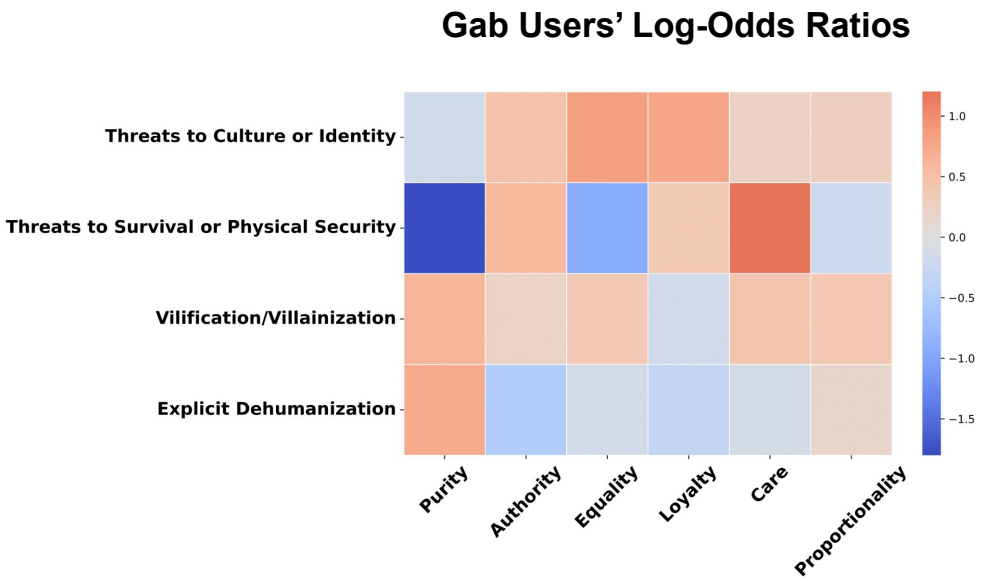
Category	Cohen's	Accuracy	F1
Threats to Culture or Identity	0.80	0.90	0.91
Threats to Survival or Physical Security	0.76	0.81	0.82
Vilification/Villainization	0.78	0.89	0.88
Explicit Dehumanization	0.81	0.96	0.96
None	0.83	0.93	0.92

(b) Inter-Annotator Agreement and Model Performance: Cohen's Kappa, Accuracy, and F1 between majority vote and HQ-LLM (GPT-4o) on Ukrainian war bloggers data.

Table 4. (a) Agreement metrics; (b) Model performance on Ukrainian war bloggers data.

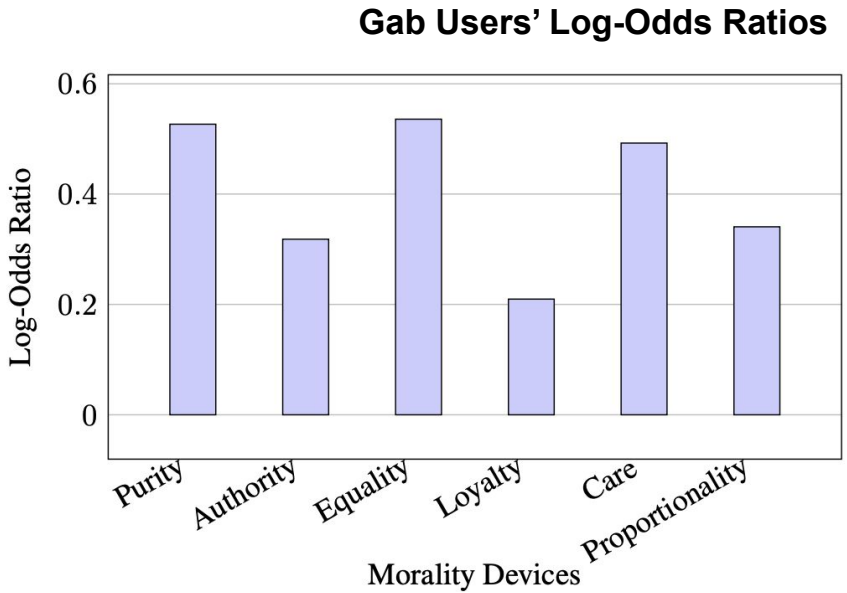
Gab Graphs

Gab Graphs: Moral Framing



Log-odds ratios for moral foundation use across othering categories in Gab user messages. Gab users tend to morally frame exclusionary language using ***purity, authority, and identity threat****; these patterns closely mirroring those observed in Russian war blogger discourse. This suggests similar rhetorical strategies may underlie othering in both conflict and extremist online communities.*

Gab Graphs: Moral Framing



Log-odds ratios of moral foundations in Gab messages containing othering. Gab users frequently frame their othering language through ***purity, equality, and care****—reinforcing patterns of moralization also seen in Russian war blogger discourse. The prominence of purity and authority may suggest a shared emphasis on sacredness and hierarchical order when justifying exclusion.*