Comments on Panel 5

Isaac D. Mehlhaff

The University of Chicago

June 25, 2025

González-Bustamante

Comprehensive assessment of toxicity detection performance and factors affecting it

But most platforms use in-house models—how to make findings informative for them?

- Patterns in training, architecture, etc. to inform model development?

González-Bustamante

Comprehensive assessment of toxicity detection performance and factors affecting it

But most platforms use in-house models—how to make findings informative for them?

- Patterns in training, architecture, etc. to inform model development?

Challenge is not just porting models across languages, but across **cultures** (Lu et al. 2025)

- More info on TextDetox corpus—coding done in a way that is sensitive or agnostic to cultural differences?
- Is the prompt truly language-agnostic? Always in English? Adjectives could vary in tone/meaning/connotation across languages, translation could affect interpretation
- Variation across languages could partially explain low R^2 in meta-analysis?

González-Bustamante

Comprehensive assessment of toxicity detection performance and factors affecting it

But most platforms use in-house models—how to make findings informative for them?

- Patterns in training, architecture, etc. to inform model development?

Challenge is not just porting models across languages, but across **cultures** (Lu et al. 2025)

- More info on TextDetox corpus—coding done in a way that is sensitive or agnostic to cultural differences?
- Is the prompt truly language-agnostic? Always in English? Adjectives could vary in tone/meaning/connotation across languages, translation could affect interpretation
- Variation across languages could partially explain low R^2 in meta-analysis?

Reproducibility!

- More rigorous, systematic analysis here could be valuable

Henry et al.

Demonstrates important use of AI in research: controlling experimental environments

Theory/hypotheses

- Conceptual difference between **semantic space** (H2) and **distortion** (H3)?
- Tough to perceive semantic distortion in real world b/c relative to a baseline
- Justification for correct $>$ over: over-moderation infringes on psych safety by signaling viewpoints are unwelcome, similar for procedural fairness(?)

Henry et al.

Demonstrates important use of AI in research: controlling experimental environments

Theory/hypotheses

- Conceptual difference between **semantic space** (H2) and **distortion** (H3)?
- Tough to perceive semantic distortion in real world b/c relative to a baseline
- Justification for correct > over: over-moderation infringes on psych safety by signaling viewpoints are unwelcome, similar for procedural fairness(?)

Platform/treatment

- How to determine "correct" moderation when mostly subjective?
- Mild swearing treatment seems weak and disconnected from theory/practice, which centers on viewpoints
- Treatment depends on users seeing messages before they are removed—how to ensure compliance?

Lim et al.

Thorough framework for using AI to conduct sensitive experiments

Do we know enough about how AI systems process info?

- Findings related to debunking could be due to Bayesian reasoning or motivated cognition
- Models can mimic human **responses**, but can they mimic human **tought processes**?

Lim et al.

Thorough framework for using AI to conduct sensitive experiments

Do we know enough about how AI systems process info?

- Findings related to debunking could be due to Bayesian reasoning or motivated cognition
- Models can mimic human **responses**, but can they mimic human **tought processes**?

Need for multi-agent framework?

- Why not do the same thing by querying individual models?
- Computational cost benefit?

Lim et al.

Thorough framework for using AI to conduct sensitive experiments

Do we know enough about how AI systems process info?

- Findings related to debunking could be due to Bayesian reasoning or motivated cognition
- Models can mimic human **responses**, but can they mimic human **tought processes**?

Need for multi-agent framework?

- Why not do the same thing by querying individual models?
- Computational cost benefit?

Perils of proprietary models in artificial experiments

- Variation is artificially reduced, not appropriate for statistical inference (Bisbee et al. 2024)
- Does this framework help with any of that?

Lim et al.

Thorough framework for using AI to conduct sensitive experiments

Do we know enough about how AI systems process info?

- Findings related to debunking could be due to Bayesian reasoning or motivated cognition
- Models can mimic human **responses**, but can they mimic human **tought processes**?

Need for multi-agent framework?

- Why not do the same thing by querying individual models?
- Computational cost benefit?

Perils of proprietary models in artificial experiments

- Variation is artificially reduced, not appropriate for statistical inference (Bisbee et al. 2024)
- Does this framework help with any of that?

Validation of model responses

Berk et al.

Clever application of hierarchical transformers to debate-level measures

Grounding in deliberation (or debate?) literature (theoretical and empirical)

- If existing work doesn't emphasize debate-level qualities, why care about this method?
- Could operationalize DQI more directly to show importance of debate-level features

Berk et al.

Clever application of hierarchical transformers to debate-level measures

Grounding in deliberation (or debate?) literature (theoretical and empirical)

- If existing work doesn't emphasize debate-level qualities, why care about this method?
- Could operationalize DQI more directly to show importance of debate-level features

Concept $\rightarrow$ measure

- Not justification unless every participant uses it—high bar to clear, easier for small groups
- Reciprocity: staying on-topic and not asking rhetorical questions

Berk et al.

Clever application of hierarchical transformers to debate-level measures

Grounding in deliberation (or debate?) literature (theoretical and empirical)

- If existing work doesn't emphasize debate-level qualities, why care about this method?
- Could operationalize DQI more directly to show importance of debate-level features

Concept $\rightarrow$ measure

- Not justification unless every participant uses it—high bar to clear, easier for small groups
- Reciprocity: staying on-topic and not asking rhetorical questions

Binary debate-level annotation artificially constrains info comments can provide by comparison

- More appropriate would be coding comment-level and aggregating
- Human coders implicitly determining some cutpoint when reading full debate

Berk et al.

Clever application of hierarchical transformers to debate-level measures

Grounding in deliberation (or debate?) literature (theoretical and empirical)

- If existing work doesn't emphasize debate-level qualities, why care about this method?
- Could operationalize DQI more directly to show importance of debate-level features

Concept $\rightarrow$ measure

- Not justification unless every participant uses it—high bar to clear, easier for small groups
- Reciprocity: staying on-topic and not asking rhetorical questions

Binary debate-level annotation artificially constrains info comments can provide by comparison

- More appropriate would be coding comment-level and aggregating
- Human coders implicitly determining some cutpoint when reading full debate

Validation of [CLS] vectors

Hobolt et al.

Refocuses on role of bottom-up processes in generating political emotions

Conceptual clarity

- "Negative affect" could mean different things—focus on emotion instead?
- Discussion, debate, or deliberation?
- Negative emotive language does not necessarily imply affective polarization
  - Pre-registered tests getting at polarization more closely—why not test those?

Hobolt et al.

Refocuses on role of bottom-up processes in generating political emotions

Conceptual clarity

- "Negative affect" could mean different things—focus on emotion instead?
- Discussion, debate, or deliberation?
- Negative emotive language does not necessarily imply affective polarization
  - Pre-registered tests getting at polarization more closely—why not test those?

Why not use previously validated models for sentiment analysis?

Hobolt et al.

Refocuses on role of bottom-up processes in generating political emotions

Conceptual clarity

- "Negative affect" could mean different things—focus on emotion instead?
- Discussion, debate, or deliberation?
- Negative emotive language does not necessarily imply affective polarization
  - Pre-registered tests getting at polarization more closely—why not test those?

Why not use previously validated models for sentiment analysis?

Uncertainty propagation for causal tests with learned proxies

- Error enters in sampling of statements and in train/val/test split (Knox et al. 2022)