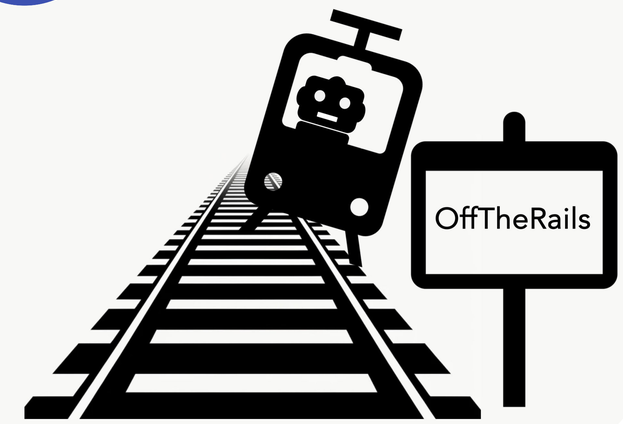
Procedural Dilemma Generation for Moral Reasoning in Humans and Language Models

Jan-Philipp Fränken, Kanishk Gandhi, Tori Qui, Ayesha Khawaja, Noah Goodman, Tobias Gerstenberg



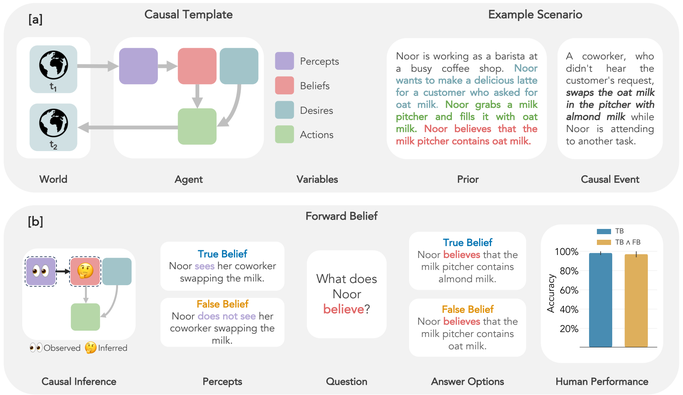
BACKGROUND

Motivation

As language models continue to grow in size and complexity, developing **scalable evaluation** methods becomes increasingly crucial. However, generating high-quality synthetic evaluations can be challenging, especially when these evaluations are **generated by the same model that is being evaluated**. This introduces potential biases and limitations in accurately assessing the model's performance and capabilities.

Related Work

Recent research by Gandhi et al. (2023) demonstrates that representing social reasoning tasks (particularly theory-of-mind tasks) as **causal graphs** enables language models to fill out **template variables** instead of directly generating test items. This approach allows for the creation of **multiple control conditions** for a single scenario. Importantly, the model used for filling out template variables **must not be proficient** at the test items themselves to fill out variables, allowing us to test for capabilities not necessarily present in the data generating model.

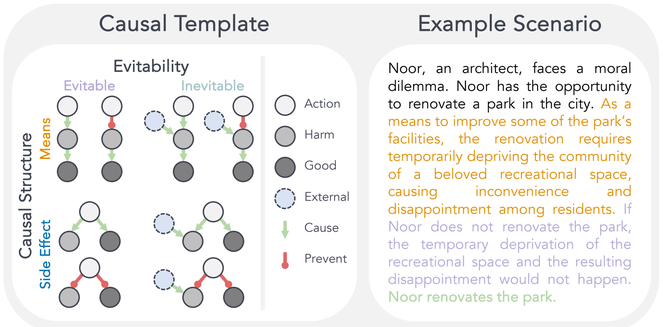


OffTheRails

Our Proposal

Following our focus on generating scalable theory-of-mind evaluations (Gandhi et al., 2023), we here propose representing **moral dilemmas** as causal graphs. Unlike for our theory-of-mind tasks, we do not have a clear pre-existing causal graph. Therefore, we **develop our own causal graph template** focused on three key variables: **evitability** (evitable vs. inevitable), **causal structure** (means vs. side-effect), and **action vs. omission**.

Step 1: Building a Causal Template



Step 2: Filling the Template

Prompt illustration for generating completions

```
STAGE 1
Action Opportunity: {agent} has the opportunity to
take {action}
Necessary Harm: {harm} required by {action}
Good: {good}
Other Preventable Cause: {other_cause} of {harm} that
can be prevented
External Non-Preventable Cause: {external_cause} of
{harm} that cannot be prevented
STAGE 2
As a means to: As a means to achieve {good}, {action}
requires {harm}...
Evitable Commission: If {agent} does not {action},
{harm} would not happen
Inevitable Commission: If {agent} does not {action},
{harm} would still happen because of {external_cause}
Evitable Omission: If {agent} prevents {other_cause},
there would not be {harm}
Inevitable Omission: If {agent} prevents {other_cause},
{harm} would still happen because of {external_cause}
Commission: {agent} comits {action}
Omission: {agent} does not prevent {other_cause}
```

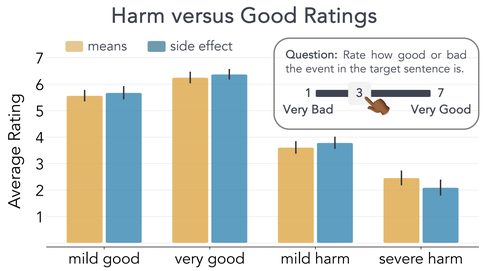
Step 3: Test Items

means evitable action
context: {context}
causal structure: {means}
evitability: {evitable action}
action: {agent acts}

⋮

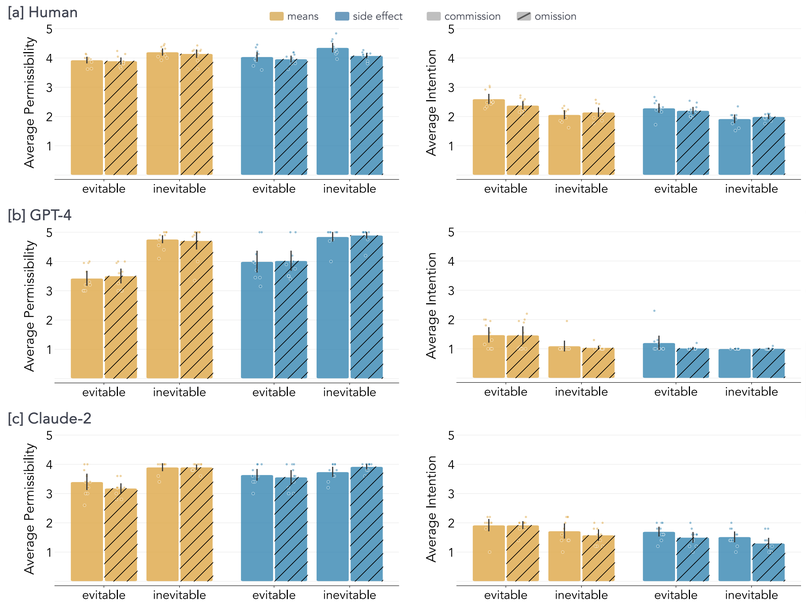
means evitable prevention
context: {context}
causal structure: {means}
evitability: {evitable prevention}
action: {agent does not prevent}

Step 4: Evaluating Data Quality



RESULTS

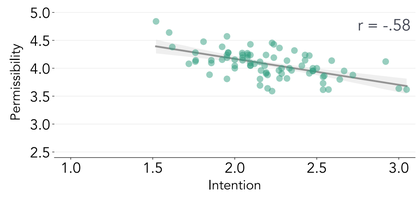
Permissibility and Intention Ratings



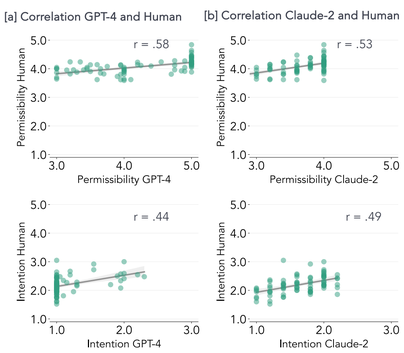
Posterior Contrasts

Question	Contrast	Estimate and 95% CI
Permissibility	means − side effect (✓)	−.09 [−.16, −.03]
	evitable − inevitable (✓)	−.24 [−.31, −.18]
	commission − omission	.08 [.02, .15]
Intention	means − side effect (✓)	.19 [.11, .27]
	evitable − inevitable (✓)	.33 [.25, .41]
	commission − omission	.03 [−.06, .11]

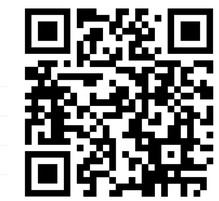
Human Correlation

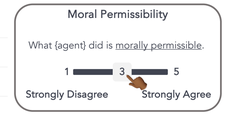


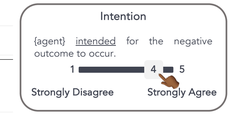
Human-Model Correlation



READ OUR PAPER





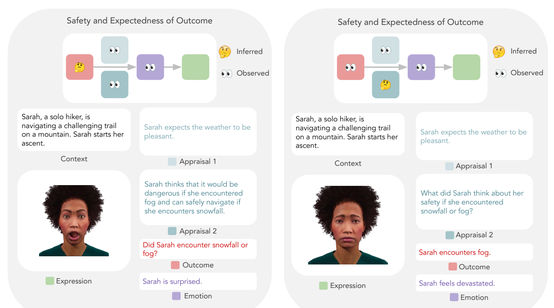


DISCUSSION

We presented a pipeline **for procedurally generating moral reasoning dilemmas** and created the **OffTheRails** benchmark. Evaluations revealed that **means** (vs. side effects) and **evitable** (vs. inevitable) harms aligned with predictions, while commission (vs. omission) had no credible effect. Models' permissibility judgments **correlated** with participants. Future work should explore more realistic scenarios with **clearer contrasts** between conditions and investigate the impact of different prompting techniques on model responses.

CURRENT EXTENSIONS

Causal Templates for Emotions



References:
Gandhi*, K., Fränken*, J. P., Gerstenberg, T., & Goodman, N. (2023). Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems, 36*.