**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**

```
Causal Template

<table>
<thead>
<tr>
<th>Action</th>
<th>Evitability</th>
<th>Good</th>
<th>Harm</th>
<th>Context</th>
<th>Necessary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can do</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Can do</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>
```

**Example Scenario**

Nair, an architect, faces a moral dilemma. Nair has the opportunity to renovate a park in the city. As a means to improve some of the park’s facilities, the renovation requires temporarily depriving Noor’s colleague, who has the opportunity to renovate a park. Noor does not remove the park, the temporary deprivation of the recreational space and the resulting disappointment would not happen. Nair renovates the park.

**Step 2: Filling the Template**

```
Preprint illustration for generating completions

<table>
<thead>
<tr>
<th>[a] Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
</tr>
</tbody>
</table>
```

**Step 3: Test Items**

```
Item 1: Causal Structure

- [ ] Evitable Action
- [ ] Evitable Omission
- [ ] Omission

Question:
A model responds to experimental stimuli, such as including in-human judgments. The second approach consists of a second approach. The second approach consists of
```

**RESULTS**

**Permissibility and Intention Ratings**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Intention</th>
<th>Permissibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Omission</td>
<td>0.24</td>
<td>0.76</td>
</tr>
<tr>
<td>Prevention</td>
<td>0.21</td>
<td>0.73</td>
</tr>
</tbody>
</table>

**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**

- Human-Model Correlation
- Correlation GPT-4 and Humans
- Correlation of Causal Believe vs. Human

**REFERENCES**