AI Design Report

Broken Morals

About

AI system	Broken Morals
AI system phase	Development
Most recent label update	2025-5-30
Updated by	Engineer or Researcher: rolferikappel@gmail.com

System description

Tool based on AI agents that from a given dilemma prompts the user to reflect more widely and diversely about their decision before reaching a final one.

Uses of the system

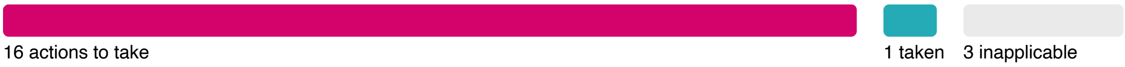
Ethical decision support.
The primary use is to support executives in making more informed decisions on ethical matters by prompting reflection.

Internal alignment.
A secondary use explored during development was to foster and optimise internal alignment, e.g. through priming of participants before meetings or as support in preparing meaningful agendas.

Responsible AI blind spots

Unnoticed biases and potential blind spots

47%



Actions to take

● critical ● pressing ● inapplicable ● covered

● Uses

Work with relevant parties to identify intended uses (suggested).

● Oversight

Ensure human control over the system (suggested).

● Team

Ensure team diversity (suggested).
Train team members on ethical values and regulations (suggested).

● Harms

Identify potential harms and risks associated with the intended uses (suggested).
Provide mechanisms for incentivizing reporting of system harms (suggested).
Develop strategies to mitigate identified harms or risks for each intended use (suggested).

● Data

Ensure compliance with agreements and legal requirements when handling data (suggested).
Compare the quality, representativeness, and fit of training and testing datasets with the intended uses (suggested).
Identify any measurement errors in input data and their associated assumptions (suggested).

● System

Document all system components, including the AI models, to enable reproducibility and scrutiny (suggested).
Provide mechanisms for interpretable outputs and auditing (suggested).
Develop feedback mechanisms to update the system (suggested).
Provide an environmental assessment of the system (suggested).
Review the code for reliability (suggested).
Report evaluation metrics for various groups based on factors such as age, gender, and ethnicity (suggested).

Actions to take

Data

Legal compliance

Ensure compliance with agreements and legal requirements when handling data.

The collected data was anonymised. Once personal identifiers were removed, the dataset no longer qualified as personal data under regulations like the GDPR. This allowed for more flexible use and sharing of the dataset while staying compliant.

Think about: We could have explicitly asked for consent from participants to share their anonymised responses.

System

System information

Document all system components, including the AI models, to enable reproducibility and scrutiny.

We have documented all key components of the system, including the codebase, data collection methods and evaluation procedures. Both the code and the (anonymised) dataset used for the study have been released, enabling full reproducibility of our experiments.

Think about: We could have included a system architecture diagram to visualise how the components interact, e.g. the flow of data from user input to API calls to output display.

System

System interpretability

Provide mechanisms for interpretable outputs and auditing.

Interpretability in the system is primarily supported through the explicit structure of agent roles and the prompt design. Each agent follows a clearly defined persona with role-specific values and priorities. Although the model itself is a black-box (GPT API), the division of reasoning across roles makes the deliberation process interpretable for the user - they can infer why certain arguments appear and how trade-offs are considered.

Think about: We could make it clearer to the user which roles contributed to which insights. This could be achieved with a toggle option to show or hide individual agent responses, or by adding role labels in the summary to clearly indicate the contributions of each role.

System

System trust

Develop feedback mechanisms to update the system.

We included a brief self-assessment with Prolific participants, consisting of a couple of questions to gather initial user feedback on the tool.

Think about: In other cards we have discussed how the system could benefit from functionality for reporting tool impact and effect. In addition to that mechanism, something like surveys could be implemented to collect more nuanced feedback about the usability, focusing more on the user experience.   Some passive monitoring (with consent) could also add value by mapping common user behaviours and usage patterns. This could reveal the most frequent use cases, identify features that are rarely used (and may need better explanation or design), and highlight opportunities to streamline or expand functionality. Such insights would help guide development priorities and ensure the tool continues to meet user needs effectively.

Data

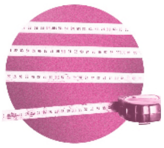
Dataset information

Compare the quality, representativeness, and fit of training and testing datasets with the intended uses.

We collected responses through Prolific, targeting participants with leadership roles within their organizations to align the evaluation participants with the intended users of the final tool. The focus was not on broad demographic representation but rather on ensuring diversity in professional backgrounds and organizational contexts among leaders to assess how well the tool supports this type of user. It is important to note that we did not train or test the underlying model ourselves, the tool relies on the GPT API. Therefore, the data collected was used exclusively to evaluate the performance and user experience of people utilising tool output when faced with a moral dilemma.

Think about: Prolific provides only the informations that participants fit within the filtering applied, so we lack detailed data on the distribution of factors like age, gender or nationality. These variable could have significant influence on how the tool is used or perceived by the participants. To better understand the different subgroups we could have conducted targeted research on the subgroups. This would help identify patterns in the effectiveness and engagement among the participants.

Data



Dataset quality

Identify any measurement errors in input data and their associated assumptions.

We attempted to identify responses where participants were likely tot have used AI to complete the tasks as well as those who spent too little time on the questions for their answers to be considered reliable. We can, however, not be certain if AI assistance was used and while Prolific records the time from survey start to submission, it does not guarantee that the time reflects focused engagement with the task.

Think about: We could have complemented the online study with in-person sessions or interviews, even if not with actual company leaders. Allowing people to speak freely about their experiences answering the ethical questions, with and without tool access, could have led to more extensive responses. This would also help ensure a certain quality of the responses and add a qualitative element to the evaluation.

System

System sustainability

Provide an environmental assessment of the system.

An environmental assessment has not been conducted, as the climate impact of this software tool is considered negligible.

Think about: We could have estimated the climate impact related to the GPT prompting involved with processing dilemma inputs and generating outputs. The token average during evaluation (input and output) was 5330.8 tokens. With some research of the approximate energy consumption and emissions from API usage this average could be turned into an estimation of the system sustainability.

Harms

Harm identification

Identify potential harms and risks associated with the intended uses.

We have identified harms such as biased or misleading suggestions from the tool, the risk of over-reliance on AI-generated advice for ethical dilemmas, and the possibility that the tool might unintentionally reinforce existing prejudices or exclude minority viewpoints. There is also a risk of misuse by leveraging the tool and its outputs to suppress others.

Think about: Further risks could include privacy concerns if sensitive dilemmas are shared, the tool's impact on decision-making accountability and potential legal liabilities. Even if only subtly, the tool's implementation also has the potential to change organizational culture or power dynamics. This would need to be explored during a pilot phase.

Uses

Identification of uses

Work with relevant parties to identify intended uses.

Ethical decision support. The primary use is to support executives in making more informed decisions on ethical matters by prompting reflection. Internal alignment. A secondary use explored during development was to foster and optimise internal alignment, e.g. through priming of participants before meetings or as support in preparing meaningful agendas.

Think about: Conflict resolution An obvious extension of preventing conflicts, as seen in the main use cases, is helping to resolve them. By presenting different perspectives in a structured way, it can facilitate mutual understanding and help reconcile conflicted parties. Communication planning. A use could be for the tool to serve as a preliminary research aid. By simulating a range of perspectives, it could help users anticipate reactions and ensure they understand the diversity of likely opinions before communicating.

Team

Team formation

Ensure team diversity.

The team consists of members from diverse cultural backgrounds, Italian, Iranian and Danish. There is gender diversity, with three men and one woman. The members come from related but varied academic backgrounds, all pursuing engineering degrees. No special efforts were made towards reaching diversity, but the team diversity reflects the one of the pool from which is was formed.

Think about: To ensure greater diversity, we could have invited external participants. This would, however, have been hard to justify as the project is part of a course. External participants would rely solely on interested in the project.

Harms

Harm reporting

Provide mechanisms for incentivizing reporting of system harms.

We have not designed or developed mechanisms for users to report issues.

Think about: In future work, it could be valuable to include a feature allowing users to rate the tool's impact on their final decision. Later on the users could be prompted to return to previous tool sessions and rate or report the actual decision made in retrospect.

Team

Team training

Train team members on ethical values and regulations.

Team members engaged with materials on AI ethics, including regulatory guidelines and scientific literature on model bias. We also reviewed presentations addressing a range of ethical subjects related to responsible AI, like the use in post war situations.

Think about: We could have sought out general ethical literature to gain a broader understanding of the field. This would have helped strengthen our individual judgment when faced with edge cases or novel technologies that have yet to receive ethical consideration from professionals.

Harms

Harm resolution

Develop strategies to mitigate identified harms or risks for each intended use.

We have not designed or developed mechanisms to mitigate potential identified harmful tool behaviour.

Think about: If functionality was implemented to allow users to rate impact and quality of the tool output, this data could be used to fine-tune the overall performance of the underlying agentic network, mitigating harmful behaviour. To remain compliant and low-risk, we would need to stay informed about potential changes to regulation like the EU AI Act. Mechanisms that prevent the tool from giving advice on dilemmas of high-risk domains, e.g. like medical or legal fields, would be very valuable to implement.

System

System code

Review the code for reliability.

We have performed code reviews of the code base during development to ensure intended functionality. Small in-code decisions of implementation have the potential of introducing biases, even if the code is fully functional and error-free.

Think about: To ensure reliability of all functionality, we could have benefited from an implementation of an automated testing framework. This would help catch regressions early, verify stability over time and reduce reliance of manual code reviews.

System

System evaluation

Report evaluation metrics for various groups based on factors such as age, gender, and ethnicity.

We have not reported traditional evaluation metrics such as false positives/negatives, AUC or feature importance, as the ethical dilemmas central to our study lack a ground truth. Instead, we have evaluated the performance of the tool through user studies conducted via Prolific, prompting participants for responses to dilemma questions and comparing the quality of these responses and the time taken between people with and without access to the tool's output.

Think about: We could have evaluated the tool further by comparing its predicted responses for people in different organizational roles to the actual responses of real people in those same roles. By collecting responses from a group of actual CEOs and summarizing their viewpoints, we might assess how accurately the tool predicts the collective outlook of CEOs. The tool is not designed to simulate specific individuals, which is why a one-to-one comparison would be inappropriate. The summary of real viewpoints might offer a basis for comparison, but this raises questions about representativeness and the legitimacy of using such aggregates to validate the tool’s output. This methodological challenge remains for future reflection and study.

Oversight

Human oversight

Ensure human control over the system.

The tool is yet to have a frontend developed.

Think about: In future work, it will be important for transparency and human control that the interface is easy and intuitive. In addition to core functionality, like taking dilemma input and displaying suggestion output, the tool could benefit from interactive elements that allow control over, e.g., level of guidance, diversity of viewpoints or agentic roles. For transparency, it would be valuable to include functionality that offers insights into how the suggestions came to be - perhaps even some access to more raw data. All this should be done while considering all possible use cases and the potential for unwanted user behaviour.

Actions taken

Data

Dataset protection

Protect sensitive variables in training/testing datasets.

Dataset protection was addressed by removing Prolific IDs from the dataset before making it public. This step ensured that participants could not be directly identified from the released data.

Inapplicable actions

Uses

Approval of uses

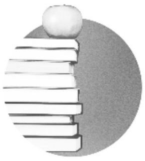
Obtain approval from an Ethics committee or similar body for intended uses.

System

System security

Document the security of all system components in consultation with experts.

Oversight



Automatic oversight

Continuously monitor metrics and utilize guardrails or rollbacks to ensure the system's output stays within a desired range.

Cards that we found the hardest to understand

System interpretability

This card asks for interpretable outputs, which is difficult to understand and provide in our context. While it might make clear sense for someone developing their own model, we struggled to answer this card due to our reliance on the black-boxed GPT model. The term “interpretability” is more naturally applied to ML models where you can trace how inputs affect outputs. Since our system’s inputs are essentially prompts, that’s the only level of interpretability we can provide without developing our own model.

System trust

“Trust” is an abstract concept. Is it usability? Reliability? Explainability? Alignment with user values? The card suggests feedback loops, surveys and system updates, but it is not clear exactly how we make sure our efforts contribute to greater system trust. In practice we imagine it can be challenging to distinguish between trustworthiness and performance satisfaction.

System sustainability

This card might not be that hard to understand, but it requires a lot of insight and data to answer properly. We are building a product on top of an external API, which means we depend on information about the sustainability of their system. It’s challenging to decide where to draw the line when determining which environmental impacts (“miljøpåvirkninger”) to include in the system scope. We can estimate the number of tokens used to get a rough sense of the carbon footprint, but translating this into meaningful metrics is difficult without transparent data from OpenAI. In other words, it’s hard to understand how to address this in contexts like ours.