Broken Morals

Emanuele Messina
Nicola Bavaro
Rolf Erik Appel
Mozhdeh Hajiani
giuseppeemanuele.messina@studenti.polito.it
nicola.bavaro@studenti.polito.it
rolferik.appel@studenti.polito.it
mozhdeh.hajiani@studenti.polito.it
Politecnico di Torino
Torino, Italy

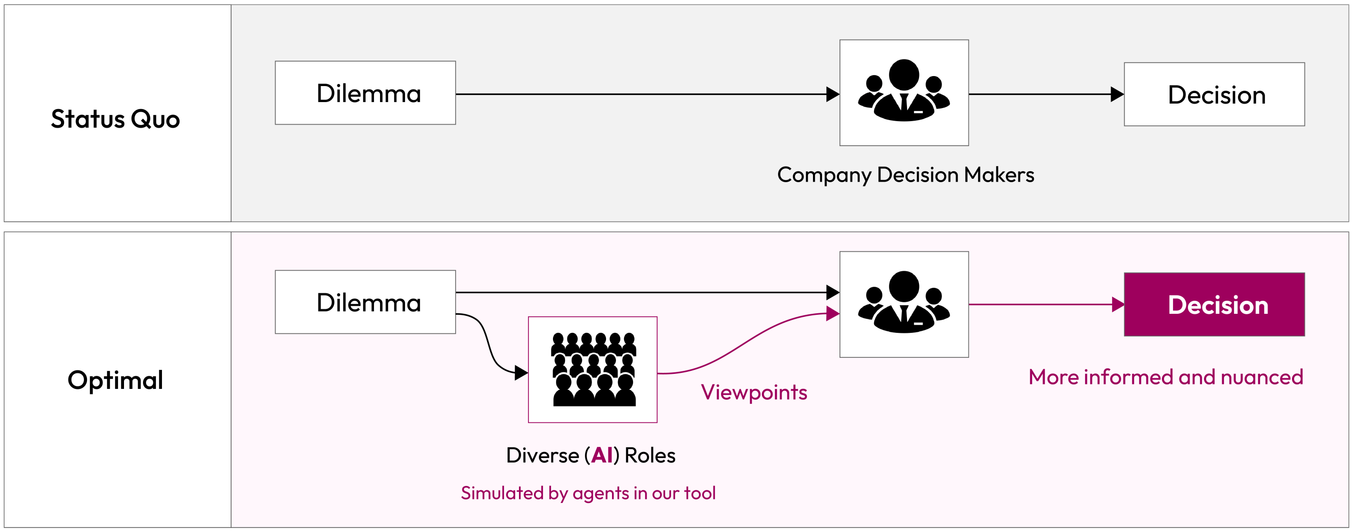

Figure 1: Overview of the Broken Morals project. (*Status Quo*) Usually, company decisions are taken without including the perspectives of different roles. (*Optimal*) We show that the exposure to those perspectives and viewpoints leads decision-makers to take decisions that are more relevant, useful and aware of contextual conditions.

Abstract

Organizational moral misalignment is very important because ethical blind spots can lead to reputational damage, internal conflict, and poor decision-making. In tackling organizational moral misalignment, related work often failed to incorporate diverse perspectives within the company before decisions are made, instead relying on static top-down ethics reviews or individual assessments. To partly tackle this limitation, we make three contributions. First, we develop a software tool that uses AI agents to simulate the perspectives of a CEO, an ethicist, and an engineer. Given a moral dilemma, the tool generates a summary of the agents' responses and presents it to the user to support their decision-making. Second, we evaluate the effectiveness of our tool through a controlled study with 50 senior technology professionals, from director level to CEO. Each participant is shown one of five business ethics dilemmas that collectively cover all the foundations declared in moral foundation theory, and tasked with providing their justification. Half of the participants are also shown the tool's summary for the given dilemma. All justifications are evaluated using a six-dimensional scoring rubric designed to assess the quality and sophistication of ethical reasoning. Third, we show that exposure to our tool leads to more nuanced ethical justifications [...].

TODO CHECK PAPER WRITING CHECKLIST

CCS Concepts

• **Applied computing** → **Sociology**; • **Computing methodologies** → *Multi-agent systems.*

Keywords

business ethics, organizational decision-making, large language models, moral misalignment, AI deliberation, Plurals framework, ethical reasoning, prototype evaluation, decision support systems

1 Related Work

2 Methods

TODO METHODS

2.1 Proposal

To incorporate perspectives of different organizational roles when making company decisions, we propose a novel software tool. The tool is intended to be used by company decision makers when taking decisions on internal dilemmas, with the goal of making more informed decisions. The user inputs a moral dilemma in the tool, and the tool gives the user a summary as output. Internally, the tool simulates a deliberation among AI agents representing different organizational roles, on the given dilemma. A moderator agent poses the dilemma to the other agents, oversees the discussion and finally generates the summary drawing from the deliberation transcript. The moderator generates the final summary by extracting the key points in which the role-based agents agreed and disagreed during the discussion, and a list of key points and/or questions that the end user should cover when answering the dilemma. The summary thus contains three paragraphs: areas of agreement, areas of disagreement and practical points for the user to cover when answering the dilemma. The purpose of this summary is to make the user think critically about the dilemma without forcing or inducing a particular point of view, but rather aiding their thought process and enriching their final response with aspects they might not consider due to lack of efficient communication with other work figures, as well as other psychological reasons such as underestimating or misunderstanding the dilemma itself and its implications. The proposed behavior for this tool is general and can be implemented in various ways: varying the number of agents and their roles, the topology in which they discuss, and the type of moderation they are subjected to. The following section describes specifically the implementation we developed and tested in the context of this study.

2.2 Implementation

For this pilot study we decided to implement a simple prototype of the proposed tool for the purpose of testing. We implemented the AI agents and their deliberation using the Plurals framework: a Python library that orchestrates AI deliberation. Plurals consists of Agents (LLMs, optionally with personas) which deliberate within customizable Structures, with Moderators overseeing deliberation. Plurals is a generator of simu- lated social ensembles. We defined three AI agent personas to represent three specific roles inside a generic technology company: a Chief Executive Officer (CEO), an Ethicist and a generic Engineer. The moderator persona is

-the tool uses plurals to run the simulations -the moderator poses the dilemma -the topology is ensemble, everyone is at the same level, they respond independently to get each viewpoint without undermining anyone based on their hierarchy level for this study we chose 3 roles ceo ethicist engineer -the moderator generates a summary with agreement disagreement points between the agents justifications and a list of practical points/questions for the user to cover when answering the original dilemma this way the user is not biased towards an artificial opinion but instead is pushed towards considering the different perspectives thus to take more informed decisions.

see appendix for the specific prompts

3 Evaluation

TODO EVALUATION

The goal of our tool is to help company decision-makers make better decisions with respect to the status quo. To ascertain the effectiveness of our tool at meeting this goal, we conducted a user study with people in tech to verify that the answers in the treatment group scored better than control with respect to a rubric we created, for a set of dilemmas we selected. Additionally, we wanted to see if, overall, there was a time difference in the responding between the two groups, and if participants in the treatment group appreciated the tool summary as an aid when answering.

To measure our metrics, we 1) selected the dilemmas 2) recruited people 3) divided into control and treatment, with minutes and pay and task 4) created a rubric 5) scored the answers

3.1 Dilemma Selection

scu dataset, mfc classifier

3.2 Participants Recruitment

prolific, roles

3.3 Participant Groups and Tasks

control, treatment, form format, pay, time, task, questionnaires

3.4 Evaluation Rubric

To compare the quality of answers across the Control and Treatment groups, we needed a way to evaluate responses in a consistent and meaningful way. Since moral dilemmas invite complex, subjective reasoning, we could not rely on simple or automatic measures. Instead, we designed an evaluation rubric that breaks down each response into clear dimensions, allowing us to assess specific aspects of reasoning and communication.

We began by reviewing previous research in areas such as argumentation, dialogue systems, ethical reasoning, and decision-making. Based on this review, we identified six dimensions that best reflect the qualities we wanted to measure. Each dimension is supported by existing literature and was selected to match the goals of our study.

- *Clarity and Structure.* Does the response express its ideas clearly? Is the reasoning easy to follow and well-organized?
  Based on McTear (2005), who highlights the importance of structure and readability in effective communication.
- *Relevance.* Does the response stay focused on the moral dilemma? Are the points made directly connected to the question?
  Informed by Habernal and Gurevych (2016), who show that content relevance strengthens argument quality.
- *Persuasiveness.* Is the argument convincing? Does it use logic, appropriate tone, and structure to support its claims?
  Draws from Johnson and Blair (2006), who emphasize the role of logical and well-structured reasoning in persuasion.
- *Concern for Long-Term Consequences.* Does the response consider future or societal effects of the proposed action?
  Inspired by the Impact Assessment Card (CSCW '25), which promotes ethical foresight.

- *Practical Usefulness.* Is the proposed solution realistic? Could it work in practice?

  Based on Bazerman and Moore (2012), who stress the importance of actionable and realistic decision-making.
- *Awareness of Context.* Does the response take into account the social, cultural, or situational context of the dilemma?

  Also from the Impact Assessment Card (CSCW '25), which encourages attention to contextual factors in ethical reasoning.

Each dimension was defined using a 5-point Likert scale, ranging from 1 ("very low") to 5 ("very high"). The rubric itself provided the conceptual framework for evaluating responses, while the actual scoring procedure is described in the following section.

3.5 Scoring the Answers

Evaluating subjective responses is inherently challenging, especially when the goal is to compare two groups fairly. To ensure our ratings were both unbiased and reliable, we followed a structured scoring process based on the evaluation rubric described above.

We began by generating three identical scoring files—one for each rater. Each file contained an `AnswerID` and the six rubric dimensions, but no information about whether the response came from the Control or Treatment group. This ensured that all evaluations were blind to condition.

Each rater independently assigned scores using the 5-point Likert scale across all dimensions. Once all ratings were completed, we reattached the group labels to the responses and calculated inter-rater reliability using Fleiss' Kappa for each dimension. We computed agreement scores separately for Control, Treatment, and the full dataset. Our target was to achieve a Kappa score above 0.6 for all dimensions—commonly accepted as the threshold for substantial agreement.

Initial scores fell short of this threshold in some cases. To address this, we held a calibration session where raters discussed discrepancies, reviewed selected responses, and refined their understanding of the rubric. After this reconciliation, we repeated the scoring where needed and achieved the following agreement levels:

Table 1: Fleiss' Kappa Scores by Dimension and Group

Dimension	**Control**	**Treatment**	**Total**
Clarity	0.669	0.617	0.645
Relevance	0.746	0.701	0.733
Persuasiveness	0.756	0.600	0.681
Concern for Long-Term Consequences	0.735	0.599	0.671
Practical Usefulness	0.754	0.606	0.683
Awareness of Context	0.715	0.612	0.667

With these agreement levels in place, we computed the average score across raters for each dimension. This resulted in a final dataset with the following structure: `AnswerID`, `DilemmaID`, `Group`, followed by the six rubric scores. This dataset served as the foundation for the subsequent statistical analysis.

Table 2: Frequency of Special Characters

Non-English or Math	Frequency	Comments
Ø	1 in 1,000	For Swedish names
π	1 in 5	Common in math
$	4 in 5	Used in business
Ψ_1^2	1 in 40,000	Unexplained usage

3.6 Results

discuss rater agreement differences if present discuss statistical tests and what they show, both rubric and questionnaires

4 Discussion

TODO DISCUSSION

Here you discuss how your results are: (1) in-line with previous work; and (2) differ (expand) on previous work. Also, you can list the limitations of your work (link to future work)

(a) Which results match previous findings in the literature? (b) Which results differ from previous findings, and why

5 Future Work

Study with more participants Participants from different roles, countries, industries Use more dilemmas from dataset Test more Plurals configurations Different topologies Different/more roles Compare with single vanilla LLM Study to codesign the tool Understand which exact requirements a tools like ours should have to support real world decisions

6 Tables

The "`acmart`" document class includes the "`booktabs`" package — https://ctan.org/pkg/booktabs — for preparing high-quality tables.

Table captions are placed *above* the table.

Because tables cannot be split across pages, the best placement for them is typically the top of the page nearest their initial cite. To ensure this proper "floating" placement of tables, use the environment **table** to enclose the table's contents and the table caption. The contents of the table itself must go in the **tabular** environment, to be aligned properly in rows and columns, with the desired horizontal and vertical rules. Again, detailed instructions on **tabular** material are found in the *LaTeX User's Guide*.

Immediately following this sentence is the point at which Table 2 is included in the input file; compare the placement of the table here with the table in the printed output of this document.

To set a wider table, which takes up the whole width of the page's live area, use the environment **table*** to enclose the table's contents and the table caption. As with a single-column table, this wide table will "float" to a location deemed more desirable. Immediately following this sentence is the point at which Table 3 is included in the input file; again, it is instructive to compare the placement of the table here with the table in the printed output of this document.

Always use midrule to separate table header rows from data rows, and use it only for this purpose. This enables assistive technologies to recognise table headers and support their users in navigating tables more easily.

Table 3: Some Typical Commands

Command	A Number	Comments
`\author`	100	Author
`\table`	300	For tables
`\table*`	400	For wider tables

7 Math Equations

7.1 Inline (In-text) Equations

A formula that appears in the running text is called an inline or in-text formula. It is produced by the **math** environment, which can be invoked with the usual `\begin . . . \end` construction or with the short form `$ . . . $`. You can use any of the symbols and structures, from α to ω, available in LaTeX [1]; this section will simply show a few examples of in-text equations in context. Notice how this equation: $\lim_{n\rightarrow \infty}x=0$, set here in in-line math style, looks slightly different when set in display style. (See next section).

7.2 Display Equations

A numbered display equation—one set off by vertical space from the text and centered horizontally—is produced by the **equation** environment. An unnumbered display equation is produced by the **displaymath** environment.

Again, in either environment, you can use any of the symbols and structures available in LaTeX; this section will just give a couple of examples of display equations in context. First, consider the equation, shown as an inline equation above:

$$\lim_{n\rightarrow \infty}x=0 \tag{1}$$

Notice how it is formatted somewhat differently in the **displaymath** environment. Now, we'll enter an unnumbered equation:

$$\sum_{i=0}^{\infty} x + 1$$

and follow it with another numbered equation:

$$\sum_{i=0}^{\infty}x_i=\int_{0}^{\pi+2} f \tag{2}$$

just to demonstrate LaTeX's able handling of numbering.

8 Figures

The "`figure`" environment should be used for figures. One or more images can be placed within a figure. If your figure contains third-party material, you must clearly identify it as such, as shown in the example below.

Your figures should contain a caption which describes the figure to the reader.

Figure captions are placed *below* the figure.

Every figure should also have a figure description unless it is purely decorative. These descriptions convey what's in the image to someone who cannot see it. They are also used by search engine crawlers for indexing images, and when images cannot be loaded.

Figure 2: 1907 Franklin Model D roadster. Photograph by Harris & Ewing, Inc. [Public domain], via Wikimedia Commons. (https://goo.gl/VLCRBB).

A figure description must be unformatted plain text less than 2000 characters long (including spaces). **Figure descriptions should not repeat the figure caption – their purpose is to capture important information that is not already provided in the caption or the main text of the paper.** For figures that convey important and complex new information, a short text description may not be adequate. More complex alternative descriptions can be placed in an appendix and referenced in a short figure description. For example, provide a data table capturing the information in a bar chart, or a structured list representing a graph. For additional information regarding how best to write figure descriptions and why doing this is so important, please see https://www.acm.org/publications/taps/describing-figures/.

Acknowledgments

The authors thank Prof. Daniele Quercia and Dr. Edyta Bogucka for their valuable guidance and support of this research. We also acknowledge Nokia Bell Labs for providing the funding that made this research possible.

References

[1] Leslie Lamport. 1986. *LaTeX: A Document Preparation System.* Addison-Wesley, Reading, MA.

A Ethical considerations

This study was conducted in accordance with general ethical guidelines for research with human participants, ensuring respect for participant autonomy, privacy, and well-being. Participants were recruited voluntarily through Prolific, a reputable online platform for academic studies, using its standard informed consent procedures and screening tools to select senior technology professionals ranging from director level to chief executive officers. No coercion or undue influence was involved in recruitment. The study exposed participants to automatically generated summaries created by artificial intelligence agents simulating organizational roles. These agents do not represent real individuals; their responses were generated based on predefined prompts and models. Participants were informed that these simulated perspectives were intended to assist their ethical decision-making and should not be considered definitive answers. All ethical dilemmas presented were non-sensitive business ethics scenarios drawn from publicly available sources. No personally identifiable or health-related information was collected beyond Prolific's internal IDs, which were removed before publishing the anonymized study data to further protect participant anonymity. Participants were free to withdraw from the study at any time without penalty. The materials and procedures posed no known risks or distress.

B Authors' Positionality Statement

We are master's students at Politecnico di Torino from diverse backgrounds and nationalities. Out teams is comprised of two Italians, one Iranian, and one Danish. Our specializations are Data Science, Computer Engineering, Artificial Intelligence and Design and Innovation. Our team's strength lies in its global and interdisciplinary makeup. This diversity in ethnic and interdisciplinary expertise let us leverage varied approaches and viewpoints to ensure a thorough and well-rounded study. We are aware that our backgrounds influence how we conducted the research and interpreted the data. We acknowledge that our perspectives might shape the way we framed ethical dilemmas and analyzed responses. We have critically reflected on these potential biases throughout the project to promote transparency and ethical rigor. We encourage readers to consider how our positionality may affect the findings.

C Division of Labour

While certain areas of the project had clear leads to leverage each team member's expertise, many tasks were collaboratively shared across the team.

Major Contributions:

- Emanuele Messina: Developed the Plurals software code and prompts, conducted the Prolific study, created graphics, and coordinated overall project management.
- Rolf Erik Appel: Led the presentation pitch and business study, and compiled the RAI cards.
- Nicola Bavaro: Developed code for dilemma selection and data analysis, led data analysis and interpretation, and developed the evaluation rubric with supporting references.
- Mozhdeh Hajiani: Performed dilemma selection coding, developed the evaluation rubric with supporting references, led related work research, and provided analysis support.

Shared Efforts:

- All members participated in reviewing and refining the evaluation rubric.
- All members collaborated in refining dilemma selection to ensure clarity, feasibility, diversity, and validation of automatic classification.
- The team jointly contributed to writing and revising the manuscript, ensuring a cohesive and polished final paper.

The diverse ideas and viewpoints of all team members were essential and carefully considered throughout the project, making its successful completion possible.

D More stuff

TODO

D.1 That we didn't put

In the 8 pages, like the tables with the exact numbers of answers and evaluations.

E Online Resources

TODO github repo, video pitch, business feasibility study, rai guidelines

Received 2 June 2025