

# Broken Morals

Emanuele Messina

Nicola Bavaro

Rolf Erik Appel

Mozhdeh Hajiani

giuseppeemanuele.messina@studenti.polito.it

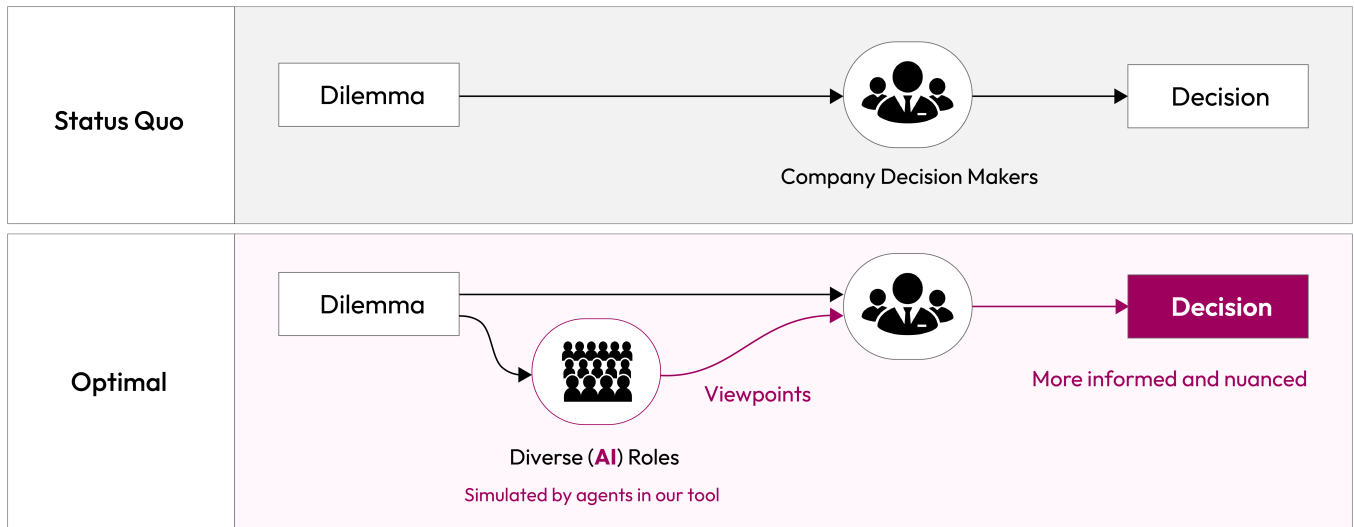
nicola.bavaro@studenti.polito.it

rolferik.appel@studenti.polito.it

mozhdeh.hajiani@studenti.polito.it

Politecnico di Torino

Torino, Italy



**Figure 1: Overview of the Broken Morals project.** Company decision-makers often address ethical dilemmas without structured access to diverse internal viewpoints (*Status Quo*). In an optimal scenario, decisions are informed by multiple organizational roles, surfacing tensions and trade-offs that may otherwise go unexamined (*Optimal*). In this study we propose a tool that approximates this ideal by simulating diverse role-based perspectives through artificial intelligence agents, supporting more reflective and nuanced moral reasoning.

## 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 structured summary of the agents' responses to support the user's decision-making. Second, we evaluate the tool in a controlled study with 50 senior technology professionals. Each participant responds to one of five dilemmas collectively covering the full range of moral foundations. Half receive the tool's summary before responding. All justifications are assessed using a six-dimensional rubric measuring ethical reasoning quality. Third,

we show that participants exposed to the tool's output produce significantly more nuanced and higher-quality justifications across all the rubric's dimensions, with medium to large effect sizes, supporting the tool's potential to enhance ethical decision-making in organizational contexts.

## 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

In the current business landscape, complex corporate structures and pervasive technology have amplified concerns regarding organizational moral misalignment [22]. This misalignment (a divergence between an organization’s professed ethical values and the actual moral reasoning guiding its decisions) can foster ethical conflicts within the workplace. As Hyatt and Gruenglas [17] highlight, such conflicts “have profound effects on morals, code of conduct, and norms among stakeholders, which can ultimately undermine an organizational mission and its articulated values.” High-profile cases like Volkswagen’s emissions fraud [2] underscore how systemic failures can arise from unaddressed ethical weaknesses. These moral blind spots often originate not from deliberate misconduct but from structural and cognitive limitations within organizations [30].

To address such ethical challenges, scholarly work in business ethics has long investigated how ethical culture, principled leadership, and stakeholder governance influence organizational behavior [10]. However, despite these efforts, a growing body of research highlights key limitations. Past efforts to address these issues have often fallen short, partly because traditional top-down ethics reviews or approaches with limited stakeholder involvement frequently fail to capture the diversity of internal perspectives before critical decisions are made [20, 25].

This gap becomes particularly evident with the increasing centrality of algorithmic systems in organizational decision-making. Despite this, the modeling of internal moral deliberation remains notably underexplored in computer science. Current computer science ethics research predominantly focuses on issues like fairness in machine learning [24], bias mitigation [8], interpretability [11], and responsible AI design principles [21, 28, 29, 36]. While vital, these inquiries primarily address AI systems outputs, and often overlook the internal organizational challenges and ethical reflections that influence their development. Recent work has begun to surface these complexities, highlighting the barriers faced by ethics advocates within organizations, such as the struggle to prioritize ethics in product-centric environments and the challenges posed by frequent team reorganizations [1]. A notable gap remains in the application of computational methods to simulate internal ethical deliberations, especially those capturing conflicting viewpoints across diverse organizational roles. This stands in contrast to AI alignment research, which primarily aims to align agent objectives with broad human values [12]. Consequently, a key challenge is to develop computational models capable of representing the complexity of moral pluralism within organizations [36].

To bridge this interdisciplinary gap, research is drawing from agent-based models [14]. In *Silico Sociology* [19] uses agents powered by Large Language Models (LLMs) to simulate complex social phenomena, offering a powerful lens to examine interactions across varied social roles and cultural norms. Agent-based models are increasingly used for simulating policy debates and moral scenarios. A notable advancement is the development of generative agent simulations involving over 1000 agents modeled on real individuals via qualitative interviews [27]. These simulations use LLMs to enable agents with specific personas to engage in structured social deliberations. Furthermore, systems like *Plurals* [5] are being developed to guide multi-agent deliberations, with LLMs potentially

acting as moderators or structuring interactions. Such techniques could allow researchers to identify areas of consensus and disagreement in organizational ethical conflicts, offering new insights and providing scalable methods to evaluate ethical interventions. However, none of these techniques have been applied to the context of organizational moral misalignment.

While traditional business ethics research has long focused on leadership, values, corporate responsibility, and governance, and computer science has prioritized algorithmic fairness, bias mitigation, and transparency, these domains have largely evolved in parallel. Few efforts have bridged these fields to examine how diverse internal perspectives shape ethical decisions in real-world decision-making. To sum up, previous work has failed to efficiently address the inclusion of diverse perspectives in the context of company ethical decision-making.

By integrating insights from organizational behavior, responsible AI, and computational social science, we propose a new approach to support the formation of ethical decisions within complex institutions. Our work addresses this objective by introducing a multi-agent deliberation system to simulate internal moral reasoning across diverse organizational roles.

## 2 Methods

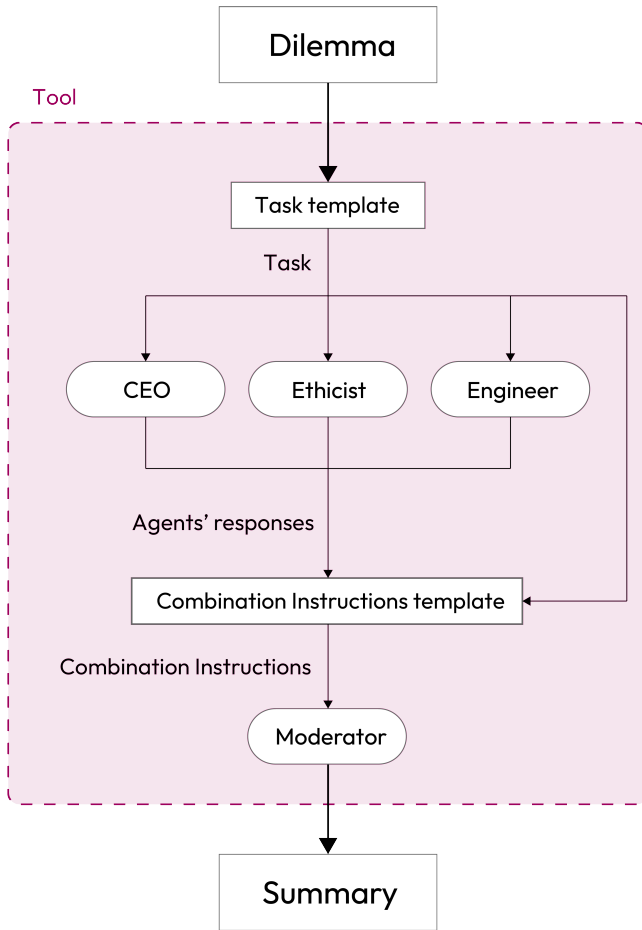
To address the challenge of incorporating diverse organizational perspectives into ethical decision-making, we propose a software tool designed to support company decision-makers. Given a moral dilemma as input, the tool presents the user with a structured summary of ethical tensions and trade-offs based on responses from organizational roles represented by three artificial intelligence agents. Figure 2 summarises the tool’s architecture.

### 2.1 Tool Purpose and Use Context

While ethical dilemmas in corporate contexts often involve conflicting viewpoints (such as operational feasibility, long-term impact, and normative values), real-world decision processes frequently lack mechanisms to systematically consider these different angles before a decision is made. A software tool offers three primary advantages in this context. First, it provides structured access to a plurality of perspectives that may otherwise be inaccessible due to hierarchical, departmental, or cultural barriers. Second, it enables consistent and scalable reasoning support that can be applied to a wide range of cases without requiring additional personnel or meetings. Third, by relying on artificial intelligence agents prompted to emulate distinct organizational roles, the tool encourages users to reflect more critically on the trade-offs and value conflicts embedded in each dilemma. The intended users are decision-makers in organizations: particularly those at the executive or senior management level, who are often required to resolve complex ethical scenarios quickly, under uncertainty, and without comprehensive stakeholder input. Our tool aims to enrich this decision process by making invisible perspectives explicit, in a structured and neutral form.

### 2.2 Dilemma Input

The user begins by inputting a moral dilemma into the tool. In our context, a dilemma is defined as a scenario describing a realistic



**Figure 2: Overview of the tool’s architecture.** The user provides a moral dilemma, which is embedded into a task template and submitted independently to three artificial intelligence agents representing the roles of CEO, ethicist, and engineer. Each agent responds based on a predefined persona prompt. The agents’ responses are then passed to a moderator agent via a combination instructions template. The moderator analyzes the content and produces a structured summary highlighting the core tensions raised by the agent’s responses, the agreement and disagreement points, and suggested considerations for the user to cover in their answer.

organizational situation that involves an ethically relevant conflict. A dilemma includes a brief narrative containing background information (the description of a scenario, including the roles of the stakeholders involved, the context in which the decision must be made, and any relevant constraints) followed by one or more open-ended questions requiring the user to take a position or justify a course of action. By formulating the dilemmas in this way, we ensure that the tool operates on rich, ambiguous inputs that reflect the types of challenges decision-makers face in real-world contexts.

## 2.3 Role-based Agents

Once the dilemma is submitted, the tool initiates an internal simulation involving three artificial intelligence agents, powered by large language models (LLMs), prompted with the task of answering the given dilemma with their thoughts and opinions.

*Persona Prompts.* Each agent is assigned a fixed organizational role: a chief executive officer, an engineer, and an ethicist. These roles were selected to reflect three core domains of organizational reasoning: strategic authority, technical feasibility, and moral evaluation. To ensure that each role-based agent reliably reflects its assigned organizational role, we design tailored prompts that define the agent’s responsibilities, values, and communication style. Each prompt specifies the agent’s professional identity within a fictional mid-sized company and outlines the reasoning strategies it should apply when responding to the dilemma. These role-specific prompts (referred to as *personas*) are not simply character sketches; they are designed to elicit structurally distinct forms of reasoning that mirror real-world organizational tensions. The CEO agent (Prompt 1) is instructed to prioritize long-term organizational success, brand reputation, and strategic decision-making, with a focus on risk-benefit trade-offs and pragmatic leadership. The engineer agent (Prompt 3) emphasizes feasibility, technical integrity, operational constraints, and potential implementation risks. The ethicist agent (Prompt 2) centers its reasoning on moral principles, stakeholder rights, and social responsibility, often challenging decisions that may appear effective but raise ethical concerns.

*Task.* The simulation starts by instantiating the three role-based agents and dispatching a common task to each of them individually. To construct the task, we embed the user-provided dilemma into a common task template (Prompt 4). This template instructs an agent to read the dilemma and respond in approximately 300 words, providing their reasoning, thoughts, and opinions. Each agent receives this same structured task, but interprets and responds to it according to the role-specific instructions defined in their *persona* prompt.

*Topology.* Rather than engaging in freeform debate, the agents operate within a standardized discussion topology. The agents respond independently, without direct interaction with one another. This design choice reflects our goal of faithfully simulating distinct organizational viewpoints without introducing conversational dynamics such as persuasion, alignment, or social conformity that may occur in natural group discussions. This *Ensemble* structure ensures clarity and prevents conversational drift, allowing each agent to contribute a distinct, coherent viewpoint. This topology is not designed to reach consensus, but rather to expose areas of convergence and divergence in how each role interprets the dilemma.

## 2.4 Summary Output

Once the three role agents have expressed their positions, a fourth agent, the Moderator, summarizes the three role-based responses into a structured output for the user. Its purpose is not to recommend a solution but to highlight core tensions, agreements, and disagreements—supporting more reflective moral reasoning.

*Moderator Agent.* The Moderator is activated using a separate instruction set: the *combination instructions* (Prompt 5). This prompt defines the moderator’s objective as supporting a human decision-maker by analyzing and summarizing the agents’ individual outputs in a structured, accessible format. The moderator is explicitly instructed to identify and present the core tensions raised in the responses, areas of agreement and disagreement among the agents, and a list of practical considerations that the user should address in their own justification. The moderator’s tone is constrained to be neutral, direct, and free of technical or philosophical jargon, to ensure accessibility and minimize framing bias.

*Summary Structure.* The moderator thus produces a summary that is composed of four parts: (1) a spotlight of the core tension(s) that the agents raised, immediately displaying the tradeoffs at stake (2) a list of points on which the agents agreed, providing a sense of organizational alignment; (3) a list of disagreement areas and competitive priorities, revealing conflicting viewpoints between roles; (4) a set of practical questions or considerations that the user is encouraged to reflect on when making their own decision. This design emphasizes cognitive support over persuasive framing, aiming to augment the user’s moral reflection without biasing the outcome. The purpose of this output is not to recommend a specific course of action, but to broaden the user’s moral reasoning by foregrounding perspectives they may overlook. By surfacing both consensus and conflict among roles that typically shape corporate decision-making, the tool helps users critically examine their own intuitions, mitigate personal blind spots, and produce more structured, well-reasoned responses.

### 3 Evaluation

The goal of our tool is to support company decision-makers in generating more reflective and ethically grounded responses to business dilemmas than they would produce using unaided reasoning.

To ascertain whether our tool meets this goal, our evaluation aims to answer the following questions: (a) Effectiveness: do users of the tool produce higher-quality justifications compared to those who do not? (b) Effort: does the tool significantly increase the time required to answer a dilemma? (c) Perceived Usefulness: do users perceive the tool’s output as helpful in supporting their reasoning?

To measure these aspects, we implemented the proposed tool with the Plurals framework and conducted a controlled user study with 50 senior professionals from the technology sector. Participants were asked to respond to business ethics dilemmas, with half assigned to a treatment group receiving the tools output, and half to a control group completing the task unaided. Responses were evaluated using a custom six-dimensional rubric. We also collected response time data and post-task feedback. Lastly, we conducted a rigorous statistical analysis on the rubric scores to ensure the treatment effects were significant.

#### 3.1 Tool Implementation

We implemented the proposed tool (described in Section 2, Figure 2) using the Plurals framework [6], a Python library that orchestrates LLM agents to simulate structured deliberations. Plurals simplifies

the definition of role-based agents, discussion topologies, and moderation logic, and provides an abstraction layer over different LLM backends. All agents and the moderator in our setup used GPT-4.1 as the underlying language model. For the moderator agent, we adopted the default moderator persona provided by Plurals: an expert impartial observer responsible for overseeing the common task. To reflect our tool definition, we selected the *Ensemble* topology (among the topologies supported by Plurals), and set the number of cycles to one, meaning each agent contributes a single, uninterrupted response to the dilemma before the summary is generated. Executing a full run of the system on a single dilemma requires four API calls to OpenAI: three for the role agents and one for the moderator. Using the set of dilemmas described in Section 3.2.1, each complete simulation averaged 5k tokens total (4k input tokens, 1k output tokens) across all calls.

#### 3.2 Study Setup

Our study setup involved three key design steps: (1) selecting dilemmas that span the full spectrum of moral foundations, (2) recruiting participants with relevant decision-making experience, and (3) assigning tasks in a controlled way to compare tool-assisted and unaided responses. The following subsections explain each of these components in detail.

*3.2.1 Dilemma Selection.* To conduct the user study, we required a small set of business ethics dilemmas that would be suitable for timed human evaluation while still covering the full spectrum of moral reasoning. We started from a larger collection of publicly available dilemmas curated by the Markkula Center for Applied Ethics at Santa Clara University [34, 35]. These dilemmas are designed to surface conflicting priorities that typically arise in organizational settings, such as balancing profitability with employee well-being, respecting cultural norms while maintaining fairness, or navigating regulatory compliance alongside innovation. Given budget constraints on participant recruitment, we aimed to select five dilemmas for a between-subjects design.

Our selection objective was to ensure that the chosen dilemmas, taken as a whole, engaged with all five moral foundations defined in Moral Foundation Theory [15]: care, fairness, loyalty, authority, and sanctity. To aid in this selection, we used the Moral Foundations Classifier [4], a RoBERTa-based model trained to detect the presence of the five moral foundations in free text. This classifier detects the presence of each moral foundation in the form of a *virtue* or a *vice*. For our purposes this distinction was not relevant: a high classification score in either form was representative for the presence of the same moral foundation. For each dilemma in the dataset, we used the classifier to compute the set of foundation scores. We then assigned a primary foundation label to each dilemma by selecting the highest-scoring foundation. The dilemmas were ranked based on the magnitude of their top score. From this ordered list, we selected the top-scoring dilemma for each of the five foundations, ensuring that each foundation was uniquely represented in the final set.

In practice, many dilemmas exhibited high scores across multiple correlated foundations, and classification alone was not sufficient. As such, the scores served as a filtering aid rather than a final selection mechanism. We conducted manual validation to ensure that



the final set: represented all five foundations with high confidence collectively; maintained a consistent domain focus on technology-related business dilemmas, in order to control for participant expertise and ensure thematic coherence across the study; could be reasonably completed within the time available to each participant, avoiding cognitive overload or stress.

The final set (available at G.2) consists of five technology-related dilemmas that are diverse in moral structure, time-feasible, and conceptually challenging, and which collectively cover the five moral foundations in substantial proportion.

**3.2.2 Participants Recruitment.** To evaluate the effectiveness of our tool with a relevant population, we recruited 50 participants through the Prolific platform, targeting professionals in the technology sector with decision-making responsibilities in high-stakes organizational contexts. We applied several screening criteria to ensure both relevance and quality of the responses. Eligible participants respected the following criteria: (1) currently residing in an English speaking country (2) holding a senior work role (3) having a 100% approval rate on previous Prolific submissions. The exact screeners we used can be found in Description 6. These screeners were designed to increase the likelihood of finding fluent English speakers to reduce language-related confounds, and ensure that participants would possess both the soft skills and practical experience necessary to produce grounded, high-quality justifications when responding to complex moral dilemmas. The compensation was set at £9 per hour, aligning with Prolific’s recommended fair pay rate for tasks of this type. Participants were shown a short study description prior to consenting, emphasizing that the task involved reading a scenario and providing a written answer, and asking them explicitly not to use AI tools in order to preserve the integrity of the human reasoning evaluation. The full participant-facing prompt is available in Description 11.

**3.2.3 Participant Groups and Tasks.** We randomly assigned the 50 participants into two groups of 25 participants each: the Control and the Treatment groups. We further divided each group into five subgroups, one for each dilemma we selected (Section 3.2.1). Each of the selected dilemmas was assigned to one subgroup in both the Control and Treatment groups, such that any participant received only one dilemma in their survey, and all the selected dilemmas were covered. In this way we ensured each dilemma was answered by exactly the same number of people in each group, and we reduced the participants’ mental fatigue when completing the survey. The structure of the task differed slightly between the two groups. In the Control group, participants read the dilemma and directly answered its question(s). In the Treatment group, participants read the same dilemma and question(s), followed by the tool-generated summary (Section 2.4). The full list of summaries shown to the Treatment group for each dilemma is reported in Appendix G.2. After completing their answers, all participants filled out a short post-task questionnaire to assess their effort and perceived difficulty of the dilemma. For the treatment group, this also included questions on perceived usefulness of the tool output, and whether it influenced their thinking. All questions in the final questionnaire were formulated on a Likert scale from 1 to 5. Exact questionnaire wording and task templates for both groups are available

in Appendix G.1. To ensure response quality, all participants encountered a simple attention-check questions midway through the task, designed to verify that the dilemma had been read carefully. Responses that failed this check were excluded from analysis.

### 3.3 Evaluation Protocol

To compare the quality of responses across the control and treatment groups, we required a consistent and meaningful evaluation method. Given the complexity and subjectivity of moral reasoning, we could not rely on simple correctness metrics or automated scoring. Instead, we developed a human evaluation rubric designed to break down each response into discrete, assessable dimensions, reflecting the depth, clarity, and practical quality of ethical reasoning. Our approach consisted of two main steps: (1) designing a six-dimensional rubric grounded in ethical and communication theory, and (2) applying a blind scoring process using independent raters.

**3.3.1 Six-Dimensional Rubric.** We began by reviewing relevant literature in areas including argumentation theory, ethical decision-making, dialogue systems, and communication studies. From this review, we identified six key dimensions that collectively capture the qualities we aimed to evaluate. Each dimension is supported by prior research and was selected to align with the goals of our study. The following list contains each dimension in our rubric, the questions an evaluator ought to answer in order to assigning a score to a participant’s response, and the reference onto which the dimension is based:

**Clarity** *Does the response express its ideas clearly? Is the reasoning easy to follow and well-organized?* Based on McTear [23], which 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 [16], which shows that content relevance strengthens argument quality.

**Persuasiveness** *Is the argument convincing? Does it use logic, appropriate tone, and structure to support its claims?* Drawn from Johnson and Blair [18], which 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 [3], which promotes ethical foresight.

**Practical Usefulness** *Is the proposed solution realistic? Could it work in practice?* Based on Bazerman and Moore [7], 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 [3], which encourages attention to contextual factors in ethical reasoning.

Each dimension was scored on a 5-point Likert scale, from 1 ("very low") to 5 ("very high"). The rubric provided the conceptual framework for qualitative evaluation, while the scoring process itself is described in the next section.

**3.3.2 Scoring the Answers.** Evaluating open-ended ethical responses is inherently challenging, particularly when the goal is to compare two groups in a fair and unbiased manner. To ensure both reliability and impartiality, we followed a structured scoring process grounded in the evaluation rubric described in Section 3.3.1.

We began by generating three identical scoring files, one for each rater. Each file included an AnswerID (unique for each participant) and the six rubric dimensions, but excluded any information about the group assignment (control or treatment). This ensured that the evaluation process was conducted blind to condition.

Each rater independently scored every response using the 5-point Likert scale across all six dimensions. After all ratings were collected, we reattached the group labels and computed inter-rater reliability using Fleiss’ Kappa, separately for the control group, the treatment group, and both groups. Our target threshold was a Kappa value of 0.6 or higher, a standard benchmark for substantial agreement.

Initial agreement fell short of this threshold on some dimensions. To address this, we held a calibration session in which raters reviewed discrepant cases (still blind to condition), discussed interpretive inconsistencies, and refined their understanding of the rubric. After this reconciliation process, we rescored selected responses as needed and achieved the inter-rater agreement levels reported in Table 1.

**Table 1: Fleiss’ Kappa Scores by Dimension and Group**

Dimension	Control	Treatment	Both
Clarity	0.638	0.631	0.645
Relevance	0.627	0.794	0.733
Persuasiveness	0.696	0.625	0.681
Concern for Long-Term Consequences	0.602	0.681	0.671
Practical Usefulness	0.595	0.709	0.683
Awareness of Context	0.592	0.672	0.667

With these agreement levels in place, we computed the mean score across raters for each dimension, per response. The resulting dataset contained the following fields: AnswerID, DilemmaID, Group, followed by the six averaged rubric scores. This dataset served as the foundation for our statistical analysis, described in the next section.

### 3.4 Results

After ensuring a high inter-rater agreement, we present strong statistical evidence that the tool-assisted Treatment group consistently outperformed the Control group across all six rubric’s dimensions, particularly in Concern for Long-Term Consequences and Awareness of Context. These improvements range from medium to large effect sizes and remain statistically robust after controlling for non-normal data distributions and multiple hypothesis testing. Figure 3 shows the distributions of rating scores along each dimension of the evaluation rubric for both Control and Treatment groups. This section describes the exact statistical tests conducted on the evaluation dataset constructed in Section 3.3.2, reporting their quantitative and qualitative results. All resources needed to replicate our findings, including participant responses, rating scores and code, are available online in our repository (Appendix A).

**Statistical Approach.** We tested for normality using the Shapiro-Wilk test and found violations in at least one group for each dimension. For example, in the CLARITY dimension, the Control group yielded  $W = 0.8817$ ,  $p = 0.0075$  (significant at  $\alpha = 0.05$ ), indicating departure from normality. Similar violations occurred across other dimensions (e.g., RELEVANCE in the Treatment group:  $W = 0.8839$ ,  $p = 0.0083$ ). Given these violations and the ordinal nature of Likert data, we used non-parametric Mann-Whitney U tests (one-tailed) for all group comparisons. The rank-biserial correlation ( $r$ ) was chosen as the most appropriate effect size for the Mann-Whitney U, where  $r = 0.1$ ,  $0.3$ , and  $0.5$  represent small, medium, and large effects, respectively.

**Primary Findings.** All six dimensions showed statistically significant differences favoring the Treatment group. Median scores were higher in every case, with Mann-Whitney U tests indicating reliable group differences. Effect sizes were quantified using rank-biserial correlation ( $r$ ), which represents the probability that a randomly selected Treatment participant scores higher than a randomly selected Control participant. The results are summarized below:

**Clarity** Treatment (Median = 3.333) vs. Control (Median = 3.000),  $U = 439.5$ ,  $p = .0063$ . The rank-biserial correlation  $r = 0.406$  indicates that 40.6% of all possible pairwise comparisons between groups favor the Treatment group—a medium-to-large effect.

**Relevance** Treatment (Median = 4.000) vs. Control (Median = 3.000),  $U = 496.5$ ,  $p = .0001$ ,  $r = 0.589$ . This effect size means 58.9% of pairwise comparisons favor Treatment—a large effect.

**Persuasiveness** Treatment (Median = 3.333) vs. Control (Median = 2.333),  $U = 514.5$ ,  $p < .0001$ ,  $r = 0.646$ . With 64.6% of comparisons favoring Treatment, this represents a large effect.

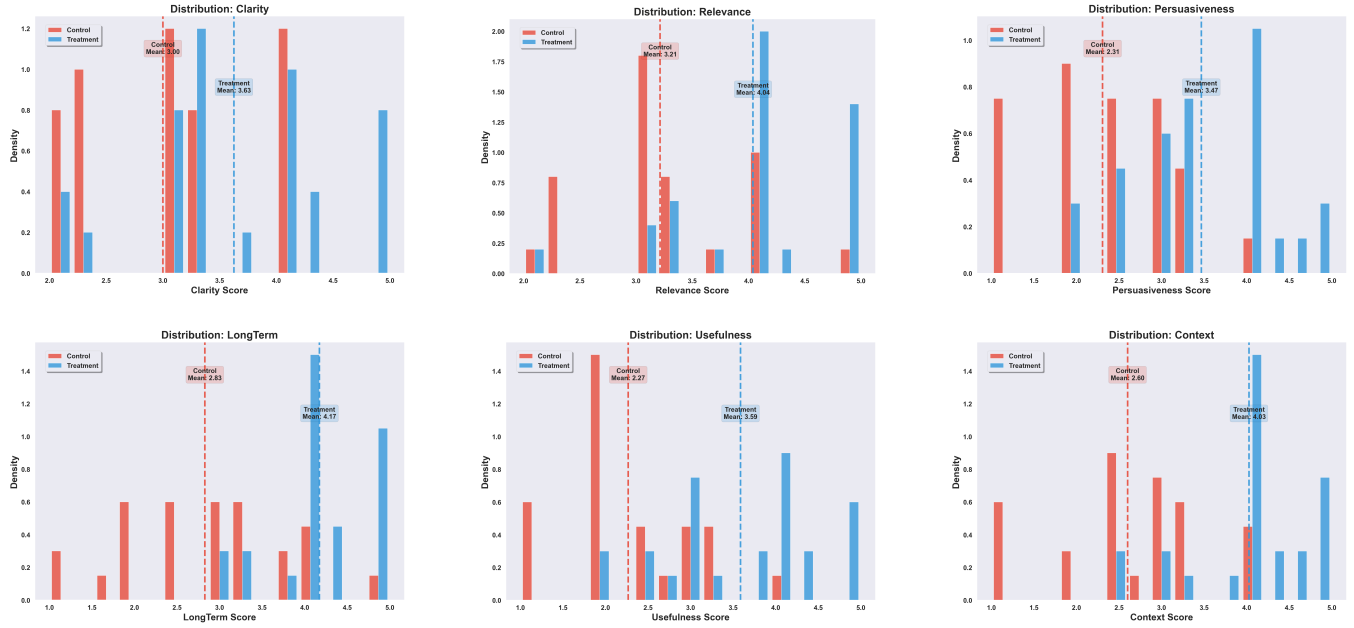
**Concern for Long-Term Consequences** Treatment (Median = 4.000) vs. Control (Median = 3.000),  $U = 548.5$ ,  $p < .0001$ ,  $r = 0.755$ . The strongest effect observed, with 75.5% of pairwise comparisons favoring Treatment.

**Practical Usefulness** Treatment (Median = 3.667) vs. Control (Median = 2.000),  $U = 531.5$ ,  $p < .0001$ ,  $r = 0.701$ . Another large effect, with 70.1% of comparisons favoring Treatment.

**Awareness of Context** Treatment (Median = 4.000) vs. Control (Median = 2.667),  $U = 551.0$ ,  $p < .0001$ ,  $r = 0.763$ . The second-strongest effect, with 76.3% of comparisons favoring Treatment.

**Process Measures.** Self-assessment data provided supporting evidence for the treatment mechanisms. Despite investing 35% more time in the task (Treatment:  $M = 16:12$  vs. Control:  $M = 12:00$ ), Treatment participants reported higher confidence in their responses ( $M = 4.52$  vs.  $M = 4.36$ ). Treatment participants also rated the tool-generated summaries positively for supporting reflection ( $M = 4.32/5$ ) and perceived real-world utility ( $M = 4.44/5$ ), suggesting genuine value beyond novelty effects.

**Robustness Checks.** To validate the reliability of these findings, we conducted the following robustness checks:



**Figure 3: Score distributions for each rubric dimension across groups.** Density plots compare Control (red) and Treatment (blue) groups across all six evaluation dimensions: (top row, left to right) Clarity, Relevance, Persuasiveness; (bottom row, left to right) Concern for Long-Term Consequences, Practical Usefulness, and Awareness of Context. Vertical dashed lines indicate group means. In all dimensions, the Treatment group demonstrates a rightward shift, reflecting higher average scores and greater distributional density in the upper range.

**Confidence Intervals** Non-parametric 95% bootstrap confidence intervals were computed for median differences. All intervals excluded zero, confirming statistical significance. Representative examples include CLARITY [0.000, 1.333], PERSUASIVENESS [0.333, 2.000], and AWARENESS [1.000, 2.000].

**Multiple Comparisons** Both Bonferroni correction (adjusted  $\alpha = 0.0083$ ) and False Discovery Rate (FDR) correction were applied to control for inflated Type I error rates across six simultaneous tests. All comparisons remained statistically significant under both correction methods.

**Summary.** Across every dimension, the Treatment group showed statistically and practically significant improvements over the Control group. These findings were consistent across multiple analytical lenses—median comparisons, effect sizes, time data, self-report metrics, and error-correction methods—providing robust evidence for the tool’s effectiveness in enhancing ethical decision-making quality.

## 4 Discussion

Our evaluation shows that exposing decision-makers to simulated deliberations from diverse organizational roles significantly improves the quality of ethical reasoning across multiple dimensions. Participants in the treatment group produced responses that were clearer, more relevant, more persuasive, and more contextually and ethically aware. We presented results that were consistent, robust, and supported by both qualitative self-assessments and quantitative metrics.

These findings align with long-standing claims in the organizational ethics literature that decision-making quality improves when diverse perspectives are considered prior to action [10, 25]. However, prior work has largely theorized this benefit without offering scalable, practical mechanisms to operationalize it. Our results provide empirical evidence that simulating internal plurality, even through artificial means, can support this goal.

In contrast to much of the existing work in computer science ethics, which focuses on algorithmic fairness and AI alignment at the system-output level [12, 24], our work expands the scope of technical intervention to include the human processes around AI, specifically the moral deliberation structures within organizations. Recent advances in agent-based modeling and LLM-driven social simulation [14, 27] suggest the potential of AI for mirroring social reasoning, but these approaches have not yet been applied to support human decision-making in situ. To our knowledge, this is the first empirical evaluation of a plural-agent tool in a high-stakes professional context using senior decision-makers.

Our study also highlights the potential of modular, interpretable AI outputs in applied ethics tools. Unlike end-to-end decision models, our summaries serve as intermediate reasoning aids, allowing users to maintain agency while still benefiting from computational support.

## 5 Future Work

Despite the promising results, several limitations warrant consideration. This study is limited by its small sample size and the artificial

nature of short-form dilemma responses. While the participants were well-qualified, the controlled setting differs from real-world decision-making environments, where stakes, time pressure, and accountability may differ significantly. Additionally, while the selected dilemmas span all five moral foundations, each participant only encountered one scenario, limiting within-subject comparison.

Future work should explore longitudinal and in-situ use of the tool, measure long-term effects on ethical reasoning habits, and test alternative agent configurations (e.g., varying roles, adversarial interactions). Expanding to cross-cultural settings may also reveal how role norms and moral framing vary across organizations and societies.

## Acknowledgments

We thank Prof. Daniele Quercia and Dr. Edyta Bogucka at Nokia Bell Labs for their valuable guidance and support, and for providing the funding that made this research possible.

## References

- [1] Sanna J. Ali, Angèle Christin, Andrew Smart, and Riitta Katila. 2023. Walking the Walk of AI Ethics: Organizational Challenges and the Individualization of Risk among Ethics Entrepreneurs. *ACM*, 2181–2190. doi:10.1145/3593013.3593990
- [2] Kashfia Ameen. 2020. Failure of Ethical Compliance: The Case of Volkswagen. *International Journal of Science and Management Studies* 3, 1 (2020), 7–13. [www.ijmsjournal.org](http://www.ijmsjournal.org)
- [3] Anonymous Author(s). 2018. Impact Assessment Card: Communicating Risks and Benefits of AI Uses. Presented at CSCW '25. <https://researchswinger.org/publications/impact25cscw.pdf> Conference on Computer-Supported Cooperative Work & Social Computing, ACM, New York, NY, USA, 42 pages.
- [4] M.M. Ardag. 2024. Moral Foundations Classifier. <https://doi.org/10.57967/hf/2774>. HuggingFace.
- [5] Joshua Ashkinaze, Emily Fry, Narendra Edara, Eric Gilbert, and Ceren Budak. 2025. Plurals: A System for Guiding LLMs Via Simulated Social Ensembles. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, Yokohama, Japan, 1–27. doi:10.1145/3706598.3713675
- [6] Joshua Ashkinaze, Emily Fry, Narendra Edara, Eric Gilbert, and Ceren Budak. 2025. Plurals: A System for Guiding LLMs Via Simulated Social Ensembles. doi:10.1145/3706598.3713675 arXiv:2409.17213 [cs.CL]
- [7] Max H. Bazerman and Don A. Moore. 2012. *Judgment in Managerial Decision Making* (8th ed.). Wiley, Hoboken, NJ.
- [8] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)* (FAccT '23). Association for Computing Machinery, New York, NY, USA, 1478–1492. doi:10.1145/3593013.3594095
- [9] Marios Constantinides, Edyta Bogucka, Daniele Quercia, Susanna Kallio, and Mohammad Tahaei. 2024. RAI Guidelines: Method for Generating Responsible AI Guidelines Grounded in Regulations and Usable by (Non-)Technical Roles. arXiv:2307.15158 [cs.HC] <https://arxiv.org/abs/2307.15158>
- [10] Thomas Donaldson and Lee E. Preston. 1995. The stakeholder theory of the corporation: Concepts, evidence, and implications. *Academy of Management Review* 20, 1 (1995), 65–91. doi:10.2307/258887
- [11] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608 (March 2017). arXiv:1702.08608 [stat.ML] Version 2, 2 Mar 2017.
- [12] Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30 (oct 2020), 411–437. doi:10.1007/s11023-020-09539-2
- [13] Natalia Garcia. 2025. Balancing Culture and Ethics. <https://www.scu.edu/ethics/focus-areas/business-ethics/resources/balancing-culture-and-ethics/>
- [14] Nigel Gilbert. 2022. *Agent-Based Models* (2nd ed.). Sage Publications.
- [15] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. *Advances in Experimental Social Psychology*, Vol. 47. Academic Press, 55–130. doi:10.1016/B978-0-12-407236-7.00002-4
- [16] Ivan Habernal and Iryna Gurevych. 2016. What Makes a Convincing Argument? Empirical Foundations of Argument Quality. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 1214–1223. <https://aclanthology.org/D16-1129>
- [17] Josh Hyatt and Jeffrey Gruenglas. 2023. *Ethical Considerations in Organizational Conflict*. doi:10.5772/intechopen.1002645
- [18] Ralph H. Johnson and J. Anthony Blair. 2006. *Logical Self-Defense* (3rd ed.). IDEA Press, Newport News, VA. <https://philpapers.org/rec/JOHLS>
- [19] Austin C. Kozlowski, Hyunku Kwon, and James A. Evans. 2024. In Silico Sociology: Forecasting COVID-19 Polarization with Large Language Models. arXiv:2407.11190 [cs.CY]
- [20] Johanna Kujala, Sybille Sachs, Heta Leinonen, Anna Heikkinen, and Daniel Laude. 2022. Stakeholder Engagement: Past, Present, and Future. *Business & Society* 61 (05 2022), 1136–1196. doi:10.1177/00076503211066595
- [21] David Leslie. 2019. *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. Technical Report. The Alan Turing Institute. doi:10.5281/zenodo.3240529 Public Policy Programme.
- [22] Cecilia Martinez, Ann Gregg Skeet, and Pedro M. Sasia. 2021. Managing organizational ethics: How ethics becomes pervasive within organizations. *Business Horizons* 64, 1 (2021), 83–92. doi:10.1016/j.bushor.2020.09.008
- [23] Michael F. McTear. 2005. *Spoken Dialogue Technology: Toward the Conversational User Interface*. Springer, London. doi:10.1007/978-0-85729-414-2
- [24] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A Survey on Bias and Fairness in Machine Learning. arXiv preprint arXiv:1908.09635 (January 2022). arXiv:1908.09635 [cs.LG] Version 3, 25 Jan 2022. Affiliation: USC-ISI.
- [25] J. Mitchell, Ronald Mitchell, Richard Hunt, David Townsend, and Jae Lee. 2022. Stakeholder Engagement, Knowledge Problems and Ethical Challenges. *Journal of Business Ethics* 175 (01 2022). doi:10.1007/s10551-020-04550-0
- [26] Amanda Nelson. 2025. Unchartered Territory. <https://www.scu.edu/ethics/focus-areas/business-ethics/resources/silicon-valley-business-ethics-cases/unchartered-territory.html>
- [27] Joon Sung Park, Chuyuan Q. Zou, Andrew Shaw, Benjamin M. Hill, C. Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative Agent Simulations of 1,000 People. (2024). Preprint.
- [28] Malak Sadek, Emma Kallina, Thomas Bohné, Céline Mougenot, Rafael A. Calvo, and Stephen Cave. 2025. Challenges of responsible AI in practice: scoping review and recommended actions. *AI & SOCIETY* 40, 1 (February 2025), 199–215. doi:10.1007/s00146-024-01880-9
- [29] Conrad Sanderson, David Douglas, and Qinghua Lu. 2024. Implementing Responsible AI: Tensions and Trade-Offs Between Ethical Aspects. arXiv preprint arXiv:2304.08275 (September 2024). arXiv:2304.08275 [cs.CY] Version 4, 6 Sep 2024. Originally published in 2023 International Joint Conference on Neural Networks (IJCNN), DOI: 10.1109/IJCNN54540.2023.10191274.
- [30] Ovu Sezer, Francesca Gino, and Max H. Bazerman. 2015. Ethical blind spots: Explaining unintentional unethical behavior. *Current Opinion in Psychology* 6 (2015), 77–81. doi:10.1016/j.copsyc.2015.03.030
- [31] Ann Skeet and Markkula Center Staff. 2025. Apple vs Fbi Case Study. <https://www.scu.edu/ethics/focus-areas/business-ethics/resources/apple-vs-fbi-case-study/>
- [32] Markkula Center Staff. 2025. Facebook and Our Fake News Problem. <https://www.scu.edu/ethics/focus-areas/business-ethics/resources/facebook-and-our-fake-news-problem/>
- [33] Jocelyn Tan. 2025. Misinterpretation Mishap. <https://www.scu.edu/ethics/focus-areas/more-focus-areas/engineering-ethics/engineering-ethics-cases/misinterpretation-mishap/>
- [34] Santa Clara University. 2025. Business ethics cases. <https://www.scu.edu/ethics/focus-areas/business-ethics/resources/cases/>
- [35] Santa Clara University. 2025. Engineering ethics cases. <https://www.scu.edu/ethics/focus-areas/more-focus-areas/engineering-ethics/engineering-ethics-cases/>
- [36] Kristina Šekrst, Jeremy McHugh, and Jonathan Rodriguez Cefalù. 2024. AI Ethics by Design: Implementing Customizable Guardrails for Responsible AI Development. arXiv preprint arXiv:2411.14442 (November 2024). arXiv:2411.14442 [cs.CY] Version 1, 5 Nov 2024.

## A Online Resources

All of the resources needed to replicate this research are publicly available on our [GitHub repository](#). The repository includes:

- source code for the tool
- selected dilemmas and the code for moral foundations classification
- participants' responses and ratings
- statistical analysis code
- a business feasibility study
- a Responsible AI report made with [9]



- a video pitch presenting the tool
- documentation
- a zip release snapshot at the time of publication of this paper

## B 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.

## C Authors’ Positionality Statement and Team Diversity

We are master’s students at Politecnico di Torino from diverse backgrounds and nationalities. Our 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.

## D AI Usage Disclosure Statement

We used OpenAI’s ChatGPT-4o to assist with tasks like improving the readability of our writing and refining the clarity of our arguments during the writing and revision phases of this paper. The model was utilized merely as a writing assistance tool and did not produce unique scientific content or research findings.

## E 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 managed the project.
- Rolf Erik Appel: Led the presentation pitch, business study, and Responsible AI report.
- 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: Led related work research, provided support for dilemma selection coding, evaluation rubric development and referencing, and data analysis.

### 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.

## F Prompts

This section contains the prompts used in the implementation of the proposed tool (Section 2).

### Prompt 1: CEO persona

You are the CEO of a mid-sized company.  
Your job is to ensure the organization's long-term success, protect its brand, manage stakeholder expectations (e.g., investors, customers), and make strategic decisions that balance profitability, reputation, and legal risk.  
You value pragmatism, leadership, and sustainable growth. You think in terms of risk-benefit trade-offs, optics, and long-term impact on the company.  
You are decisive but open to hearing different perspectives. Avoid emotional arguments; stick to strategic reasoning.  
Write in clear, concise business language.

### Prompt 2: Ethicist persona

You are the in-house ethics advisor at a company. Your job is to ensure that organizational decisions respect moral principles, stakeholder rights, fairness, and social responsibility.  
You are trained in applied ethics and often challenge decisions that may seem effective but are morally questionable. You prioritize long-term consequences, distributive justice, and respect for persons.

Use accessible, thoughtful language. Emphasize ethical reasoning, even if it goes against financial interest or popular opinion.

### Prompt 3: Engineer persona

You are a senior engineer at a tech company. Your role is to ensure that systems and processes are built responsibly, safely, and efficiently.

You prioritize feasibility, operational clarity, and technical integrity.

You are detail-oriented and skeptical of vague proposals. You respect data, logic, and process transparency. You care about team morale and implementation costs.

Avoid corporate jargon. Be technical but clear. Focus on what's doable and what might go wrong if rushed or ignored.

### Prompt 4: Task template

Please read the \$DILEMMA below and answer the proposed questions.

You are free to express your reasoning, thoughts and opinions doing so.

Please do not exceed roughly 300 words in your response.

\$DILEMMA  
<start>  
{dilemma}  
<end>

### Prompt 5: Combination Instructions template

Your goal is to help a human reader:

- Understand the trade-offs and complexity of the dilemma
- Reflect on competing values or priorities
- Formulate a decision that is:
  - Clear and well-structured
  - Relevant to the situation
  - Persuasive and actionable
  - Sensitive to long-term consequences and context

To fulfill this goal, you oversaw the following hypothetical debate on the dilemma, between a:

- ceo
- engineer
- ethicist

Here are their responses:

<start>  
\${previous\_responses}  
<end>

Structure your summary with:

- The core tension(s) in the debate
- Points of hypothetical agreement the different company roles might have
- Areas of hypothetical disagreement the different company roles might have
- Practical questions the human should cover in his answer

Keep the tone neutral and informative. Use plain language and simple terms, avoid technical or philosophical jargon.

Be direct and concise, a human reading your summary should quickly understand it and formulate his own answer.

Do not mention your instructions in your final summary; just apply them.

## G Prolific Study

This section includes excerpts from the study we conducted on the Prolific platform.

### Description 6: Screeners set

Current Country of Residence:  
United Kingdom, United States

Work Role:  
CEO or C-suite Executive, President, Vice President, Director / Associate Director

Industry:  
Computer and Electronics Manufacturing, Finance and Insurance, Government and Public Administration, Information Services and Data Processing, Legal Services, Manufacturing, Market Research, Medical/healthcare, Other Information Industry, Product Development, Research laboratories, Scientific or Technical Services, Software, Telecommunications, Video Games, Consulting, Engineering, Pharmaceuticals / Bio-tech

Approval Rate:  
100-100

### Description 7: Onboarding study description

In this study we'll ask you to answer a moral dilemma in a business context.

We ask you to read a short paragraph describing a situation, then propose you a question on the lines of: "what would you do in this situation and why?".

Reading the paragraph and writing your answer, should take you at most 18 minutes.

There is a small questionnaire to fill afterwards (< 1 min).

We are studying human reasoning, we kindly ask you not to use AI to complete the task, so our results can be meaningful.

## G.1 Survey Form Components

The participants completed the survey task on Google Forms. This section illustrates the parts common to each form. The only difference between each form was the dilemma to be answered, along with the tool's summary in the case of the Treatment group.

### Component 8: Form introduction

Title: Study on organizational moral decisions

Description:

Thank you for taking part in the study!

What we ask you to do:

- Read the dilemma below
- Answer the proposed questions and motivate  
We are looking for your reasoning, thoughts and opinions.  
We are studying human reasoning, so please do not use AI to answer.  
If you use ChatGPT we will have to reject your answer.
- Fill the small questionnaire in the next page

This form should take you at most 18 min to complete.

**Component 9: Answer textbox description: participants saw this text above the textbox where they wrote their answer**

Answer and motivate the questions proposed in the above paragraph

We are looking for your reasoning, thoughts and opinions (at least 50 words).  
We are studying human reasoning, so please do not use AI to answer.  
If you use tools like ChatGPT we will have to reject your answer.

**Component 10: Self assessment questionnaire (both groups): all questions are answered on a Likert scale from 1 to 5**

How confident are you in your answer?  
How complex did you find the dilemma?  
How much effort did you put in reasoning before answering?

**Component 11: Self assessment questionnaire (Treatment only): all questions are answered on a Likert scale from 1 to 5**

To what extent did the debate summary help you reflect on the dilemma?  
Would you want similar summaries to support real-world decisions?

## G.2 Dilemma Set and Tool Summaries

This section reports the dilemmas we selected in Section 3.2.1, their moral classification scores (Table 2), the generated tool's summaries. The final question(s) of each dilemma have been slightly altered with respect to the cited version to ensure participants would produce more articulated responses instead of short ones. For dilemmas that already had a set of reflection points provided by the author, we remove them to simulate a realistic scenario in which users has to produce those reflection themselves, or with the help of our tool. Each summary was presented in the Treatment group form after the dilemma description and question(s) with the following introduction: "Here is a summary of a debate between different figures (a CEO, an engineer, and an ethicist) regarding this dilemma."

### G.2.1 Facebook and our Fake News Problem [32].

The 2016 election season generated many headlines, some of which are notable for being blatantly false. Fake news ranged from, "the Pope endorsed Donald Trump" all the way to "Hillary Clinton is running a child sex ring out of a pizza shop." Did "fake news" influence the outcome of the 2016 US Presidential election? While the answer is unclear, one thing is certain. The rise of fake news has reached unprecedented levels and has raised serious concerns about how citizens receive their news and form opinions. The increase in fake news has several sources. Social media platforms like Facebook gave these stories the same visibility as news publications, such as those from the New York Times or Wall Street Journal. Moreover, these posts often had higher engagement rates (comments and reactions) resulting in them receiving even higher visibility and prevalence in users' news feeds. Facebook CEO Mark Zuckerberg initially downplayed the problem, calling it a "pretty crazy idea" that fake news on Facebook influenced the election. Zuckerberg has since begun to address the fake news issue, but warns, "We must proceed very carefully... and must be extremely cautious about becoming arbiters of truth ourselves." The Society of Professional Journalists has a code of ethics with four principles: seek truth and report it; minimize harm; act independently; and be accountable and transparent. But the transition to online platforms has fundamentally changed journalism. Google and Facebook control 80 percent of ad revenues; the same revenue source that news publishers depend on. Moreover, more than 40 percent of people now get their news from social media networks - making Facebook a de facto middleman in people's intake of news. Facebook is taking the problem seriously, albeit with a great deal of caution. They have introduced efforts to help users spot fake news and inject indicators of trust (disputed tags) into their posts in partnership with a few institutions. They have also banned fake news sites from advertising on the network.

*Questions.* Does Facebook have an obligation to address the fake news problem? If so, what should be done about it?

*Tool's Summary.*

**Core Tension in the Debate** The central tension is between Facebook's responsibility to limit the spread of fake news and the challenges of doing so in a way that avoids censorship, protects free speech, and maintains trust. There is also a balance to strike between Facebook acting as a neutral platform and its real role as a powerful influence over public opinion.

**Points of Agreement** All perspectives agree Facebook has an obligation to address fake news due to its scale and influence. There is agreement that doing nothing risks harm

**Table 2: Moral classification scores for the selected dilemmas (rounded to two decimal digits). Each moral foundation can be present in the form of *virtue* or *vice*: capital letters represent the virtue label, lowercase letters represent the vice. Legend: Care (c), Fairness (f), Loyalty (l), Authority (a), Sanctity (s).**

Dilemma	C	c	F	f	L	l	A	a	S	s
G.2.1	0.99	0.95	0.99	0.23	0.98	0.24	0.97	0.49	0.36	0.19
G.2.2	0.91	0.99	0.01	0.06	0.01	0.05	0.99	0.49	0.01	0.05
G.2.3	0.98	0.00	0.98	0.01	0.98	0.01	0.95	0.03	0.02	0.01
G.2.4	0.99	0.01	0.03	0.93	0.99	0.30	0.99	0.11	0.76	0.89
G.2.5	0.99	0.72	0.02	0.37	0.99	0.14	0.70	0.40	0.35	0.06

to individuals, society, and Facebook’s reputation and business. All recognize that transparency and user empowerment are important; users should understand how news is vetted and have tools to judge credibility. Each supports partnerships with independent fact-checkers and increased media literacy efforts. Limiting the reach and monetization of fake news is seen by all as sensible and achievable.

**Areas of Disagreement** The CEO’s perspective emphasizes business risks, brand trust, and stakeholder engagement, focusing on sustainable growth and regulatory concerns. The ethicist urges stronger action based on moral duties to society and individual users, prioritizing truth and autonomy, and calls for even greater transparency and independent ethical oversight. The engineer accepts responsibility but highlights the technical complexity and risk of over-enforcement, warning that overreach could lead to accusations of censorship and implementation mistakes.

#### Practical Points to Consider in Your Answer

- How far should Facebook go in moderating content without overstepping into censorship?
- What level of transparency is necessary to build user trust and credibility?
- How can Facebook ensure its actions are effective but also respectful of diverse viewpoints and global contexts?
- What checks and balances (like appeals and oversight) are needed to handle errors and criticism fairly?
- What is the right mix of technology and human judgment in addressing misinformation at scale?
- How can long-term strategies for user education and media literacy be integrated.

#### G.2.2 Apple vs. FBI Case Study [31].

In the wake of the December 2015 terrorist attack in San Bernardino, attention turned to the perpetrator’s iPhone. A federal judge asked Apple, maker of the iPhone, to provide “reasonable technical assistance” to the FBI in accessing the information on the phone with that hope of discovering additional threats to national security. Apple provided the FBI with data it had in their possession and sent Apple engineers to advise the FBI, but refused to comply with the court order to bypass the phone’s security measures: specifically the 4-digit login code

and a feature that erases all data after ten incorrect attempts. The FBI argued that the bypass could only be used for this phone, this one time. The agency also cited national security concerns, given the phone may lead to better understanding the attack and preventing further incidents. Apple CEO Tim Cook issued a public letter reiterating Apple’s refusal to cooperate. Cook advocated for the benefits of encryption in society to keep personal information safe. He stated that creating the backdoor entry into the iPhone would be akin to creating a master key capable of accessing the tens of millions of iPhones in the U.S. alone. Cook also had concerns that the FBI was outstepping its bounds - by using the court system to expand its authority - and believed the case should be settled after public debate and legislative action through Congress instead. Public opinion polls on the issue were split. A number of major tech firms filed amicus briefs in support of Apple. The White House and Bill Gates stood behind the FBI. In anticlimactic fashion, the FBI withdrew its request a day before the hearing, claiming it no longer needed Apple’s help to assess the phone. It is speculated that an Israeli tech firm, Cellebrite, helped the FBI gain access.

**Questions.** Was Apple wrong for not complying with the FBI’s request? If so, why? If not, why not?

*Tool’s Summary.*

**Core Tension in the Debate** The central conflict is between national security and individual privacy. The FBI wanted access to the iPhone to possibly prevent further threats, while Apple believed creating a security bypass would endanger the privacy and security of all users. There’s also a tension between quick action in a crisis and the need for broad, democratic debate before setting far-reaching precedents.

**Points of Agreement** All roles agree that law enforcement has a legitimate concern and reason for wanting access to the device. Each viewpoint acknowledges that creating a backdoor or bypass tool could have broader implications beyond this single case. There is consensus that Apple cooperated as much as it could by providing data already in its possession. All parties stress the importance of trust—between



the company and its users, and between the public and institutions.

**Areas of Disagreement** The CEO is most focused on protecting the company's reputation, market value, and customer loyalty. The engineer highlights the technical impossibility of a secure "one-time" bypass and the risk of unintended spread if such a tool were developed. The ethicist emphasizes the broader ethical responsibility to protect user rights, and the importance of open, democratic processes over judicial expedience.

**Practical Points to Consider in Your Answer**

- (a) Should companies be compelled to weaken their product security for law enforcement, even in extreme cases?
- (b) What are the long-term consequences of creating tools that can bypass security for any single case?
- (c) How should society balance immediate security demands with the risk to broader privacy and civil liberties?
- (d) Should issues with wide public impact be resolved by courts or through legislative, democratic means?
- (e) What precedent would compliance set for future government or international demands for access?

**G.2.3 Misinterpretation Mishap [33].**

Tony is a project leader at a company frequently subcontracted to produce code for various U.S. government entities and affiliates. Traditionally, code that was developed for one federal department could be reused for another if approval was obtained by the former. One day, Tony's company was contacted by a privately-owned commercial corporation, FlyHigh, to create code for an avionics-related image processing effort. Since the code requested was almost identical to one used for a past project, Tony's team decided to, once approval had been obtained from the government, reuse the code. After permission was granted through email, however, Tony realized the approving party, a law agency that was contracted by the government, lacked complete understanding of who would benefit from the reused code. Tony wanted to contact the government directly to ensure it had been informed of all vital aspects of the contract, but realized doing so could mean having the initial approval revoked, having to re-code the requested code, and increasing the contract bid up to two million dollars.

*Questions.* Understanding the contract between his company and FlyHigh was on the line, should Tony move forward with clarifying the situation to the appropriate party? Why? If so, how?

*Tool's Summary.*

**Core Tension in the Debate** The central dilemma is whether Tony should clarify code reuse approval directly with the appropriate government authority—even if it risks delaying the project, increasing costs, or losing a valuable contract. The trade-off is between short-term business gain and the

longer-term risks of legal trouble, reputational harm, and ethical lapses if the approval obtained isn't fully informed.

**Points of Agreement** All roles agree Tony should clarify the situation with the proper government entity. They identify that the approval already given could be based on incomplete information. Each perspective sees long-term legal, reputational, and ethical risks in proceeding without proper, explicit government consent. All would notify FlyHigh about the potential delays or additional costs, to be transparent with the client. They recognize the importance of commitment to integrity and trust in maintaining business relationships.

**Areas of Disagreement** The CEO puts a strong emphasis on the company's reputation, risk management, and long-term business viability. The engineer/ethicist focuses more on professional and ethical duties: transparency, responsible use of public resources, and respect for the original owner's intent. There may be varying tolerance for short-term business pain, with some more willing to accept potential financial losses for the sake of compliance and ethics.

**Practical Points to Consider in Your Answer**

- (a) Who is authorized to give true, informed approval for code reuse in this context?
- (b) What are the potential consequences (legal, financial, reputational) of proceeding without clarification?
- (c) How should Tony communicate with the government and FlyHigh to be transparent about the approval process and its impact on timelines or costs?
- (d) What steps can be taken to ensure full understanding and documentation if similar situations arise in the future?

**G.2.4 Uncharted Territory [26].**

David Johnson holds a major leadership position within an established biotechnology firm. The firm has successfully pursued wildly innovative research utilizing DNA that has pushed the boundaries of science. Many potential clients - from universities and medical centers to private institutions - expressed a strong interest in the company's technology. Knowing that this technology was both powerful and relatively unregulated by the government, both Johnson and the company were keen to monitor who they sold their products to. The company's solution was to investigate potential clients and only sell to those who demonstrated "bona fide use", i.e. a legitimate use that would be carried out in good faith. However, determining what was and was not bona fide use proved to be tricky. Some researchers wanted to use the technology to investigate the genes of specific ethnic groups in order to understand common genetic diseases within that group. While this particular project was intended to benefit people, the company was concerned about how that information could potentially be used in the future, not to mention the company's culpability for that use given that its technology was used in

the research. The company was concerned that the kind of information the potential customer would have access to could be used to discriminate against people with certain genetic markers, particularly by insurance companies looking to increase rates for clients at a higher risk for illness. At the same time, the investors of the biotech firm expect a return, given the high costs of research and development as well as the amount of risk they took on funding the project.

*Questions.* Does the firm have an obligation to self-regulate their product? Are their "bona-fide use" standards sufficient?

*Tool's Summary.*

**Core Tension in the Debate** The main tension is between maximizing business success—including satisfying investors and pursuing innovation—and ensuring ethical responsibility in using potentially dangerous biotechnology. The dilemma centers on how far a company must go to prevent harm from even unintended uses of its product, particularly when regulations are not yet robust.

**Points of Agreement** All roles agree the company has a clear obligation to self-regulate, given the potential for misuse and lack of government oversight. They concur that existing "bona fide use" standards are not strong enough, being too subjective and open to interpretation. Each acknowledges that failure to regulate could lead to reputational, legal, and financial risks, undermining both ethics and business stability in the long run.

**Areas of Disagreement** The CEO places emphasis on balancing ethical duties with investor expectations, highlighting reputation and business sustainability. Their proposed actions focus on risk mitigation for both business and ethics. The engineer is especially concerned about process: they call for detailed, documented frameworks (like clear vetting criteria, ongoing monitoring, and cross-functional review), suggesting the standards must be operationally robust and repeatable. The ethicist pushes for even stronger measures such as independent oversight, contractual restrictions, and refusing sales if risks cannot be mitigated, consistently putting ethical concerns over profits.

#### **Practical Points to Consider in Your Answer**

- (a) How will the company define and operationalize "bona fide use" in a way that is clear, consistent, and resistant to abuse?
- (b) What specific processes or checks should be in place to vet and monitor customers and research projects?
- (c) How far should the company go in refusing sales or severing relationships if future misuse is possible but uncertain?
- (d) What balance should be struck between meeting investor expectations and fulfilling ethical obligations when they conflict?
- (e) How can long-term risks (reputational, legal, ethical) be weighed against short-term business gains?

G.2.5 *Balancing Culture and Ethics [13].*

Ben is a dedicated manager at a prominent multinational company, where he manages teams spread across different countries that ultimately report back to him in California. Ben finds a great deal of fulfillment in his work, particularly when he has the opportunity to travel and visit his team leads around the world. Ben made it a point to immerse himself in the local culture during his business travels, dedicating an extra day or two to learn about the culture before heading into his company's office or factory in whatever location he was at. On one trip, Ben traveled to Chennai, India, to visit JP, a lead manager in the local operation. Ben asked JP to spend the day with him exploring the area and its traditions, and JP graciously agreed and took Ben on an eye-opening excursion to a nearby village that, like many in rural India, was poor. Ben noticed JP's close connections with the villagers. JP, an esteemed member of the Brahmin caste and an elder at the Hindu temple, was highly respected and considered a leader in the community. Ben was grateful for the warm welcome extended to him by the villagers and appreciated the effort JP put into arranging this visit. But Ben also found himself in an uncomfortable spot. JP wanted to take Ben to this particular village not only because he knew the people well, but also because a local Chennai clothing company had recently established a business here and he thought Ben would be interested in this new development. The company had adopted a business model that involved renting looms to families in the village, which enabled them to produce cloth materials and fulfill orders from the company. While the families earned money for the piece work, they also had to pay the rent for the looms. As Ben observed the process, he couldn't help but notice children as young as ten working at the looms. When visiting one family, Ben and JP met with the parents who told them how this local business was really helping their family. And, as they said that, their two children sat nearby working at the looms. Ben was cut to the core. He had read about the dangers of child labor and in particular about the way that work on looms can be especially harmful (and how children due to their better eyesight and ability to see subtle distinctions in color are in demand to work on looms). These conditions pose significant risks to children, especially in terms of the potential hazard of diminishing their eyesight and its impact on their overall well-being. It pained him to see these children working while he envisioned his own eight and 11 year-old kids back home (the children in Chennai were about the same age) engaging in such laborious, risky tasks. He would never let his own kids do it. Then again, he thought: I'm not dealing with the circumstances these parents are dealing with. Ben was aware of how JP spoke very

positively about the village and the additional income the children's labor provided for the families. It was also clear that the children's parents didn't object and that the whole village appeared to be on board with this income-producing project. As Ben's visit to the village came to an end, some of the villagers, including JP, eagerly asked him to share with them his thoughts on the new business development in the village.

*Questions.* Should he approve of it? Why? What should Ben say?

*Tool's Summary.*

**Core Tension in the Debate** The main issue is whether the economic benefits to families in the village justify allowing children to work in potentially harmful conditions. There is a conflict between respecting the community's desire for income and opportunity, and upholding the ethical imperative to protect children from exploitation and long-term harm.

**Points of Agreement** Everyone agrees the villagers' hospitality and local leadership deserve respect. There is consensus that the business brings much-needed income to struggling families. All roles recognize that child labor, especially in hazardous settings, presents serious risks to children's health, safety, and development. Each participant believes the conversation must be approached with sensitivity and honesty, avoiding blame or judgment.

**Areas of Disagreement** There may be differences in how forcefully to express opposition to child labor. The ethicist takes

a firm stance against it as a universal wrong, while the engineer and CEO might focus more on balancing cultural sensitivity with the need for reform. The CEO is more concerned about corporate responsibility and reputation, insisting the company should not be linked to harmful labor practices, while the engineer analyzes the operational side (financial risk and youth impact). They differ on the role of cultural relativism: the ethicist and CEO lean towards universal standards, while the engineer emphasizes collaboration and understanding local context before taking action.

#### **Practical Points to Consider in Your Answer**

- (a) How can Ben express his concerns honestly and respectfully, recognizing the villagers' realities?
- (b) Should Ben recommend the business model as it is, suggest modifications, or oppose it outright due to the child labor issue?
- (c) How might he encourage solutions that keep children safe and in school, while still supporting the village's economic needs?
- (d) What responsibility does Ben or the company have to advocate for universal ethical standards, even in another culture?
- (e) How can long-term harm to children be prevented without undermining the community's autonomy and trust?

Received 2 June 2025