Judging the Judge: A Study on Perception of LLM vs Human Decisions-Makers in Ethical Dilemmas

Azalée Robitaille
azalee.robitaille@umontreal.ca
University of Montreal
Montreal, Quebec, Canada

Yalda Kasravi Mavi
yalda.kasravi.mavi@umontreal.ca
University of Montreal
Montreal, Quebec, Canada

ABSTRACT

Large language models (LLMs), such as ChatGPT, are increasingly integrated into personal and professional domains, influencing how individuals seek information, form opinions, and make decisions across ethically and morally complex scenarios. While prior research has focused on the technical design of AI systems and their ethical frameworks, limited work explores how perceptions of LLM-generated responses compare to human responses and how these perceptions influence individuals' reasoning and attitudes.

This paper investigates how participants perceive and respond to LLM-generated versus human-generated ethical reasoning across four ethically sensitive themes: autonomous vehicle dilemmas, medical resource allocation, variations of the trolley problem, and workplace dilemmas. Using a between-subjects study design and a mixed-methods approach, we assess participants' trust in and willingness to consider contrasting perspectives depending on whether the responses are attributed to an LLM or another human participant.

Quantitative analyses reveal that participants are more open to revising their initial opinions when alternative responses are attributed to humans compared to an LLM. Conversely, participants demonstrate increased defensiveness and skepticism toward responses perceived as LLM-generated. Qualitative findings further underscore a fundamental difference in the perceived credibility, trustworthiness, and relatability of human versus AI-generated moral reasoning.

KEYWORDS

AI, Perception of AI, Ethical dilemmas, LLM

ACM Reference Format:
Azalée Robitaille and Yalda Kasravi Mavi. 2025. Judging the Judge: A Study on Perception of LLM vs Human Decisions-Makers in Ethical Dilemmas. In . ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Large language models (LLMs), such as ChatGPT, are rapidly becoming an integral part of many people's personal and professional daily life. From answering everyday queries to assisting in educational assignments and professional tasks, LLMs have a notable influence on how people seek new information and make decisions across various domains. Because of that, LLMs are increasingly being incorporated into systems that require complex reasoning, and with this widespread adoption comes the need to understand more than only the technical capabilities of LLMs.

Indeed, it is also important to have an understanding of what kind of impacts they can have, more precisely how people perceive their outputs and how these perceptions can influence their own opinion and thoughts. These questions are particularly relevant in areas where there is not necessarily a right and a wrong answer like ethically sensitive contexts where decisions have moral or societal implications. This challenge involves understanding the nuances between human users and AI-generated content.

As LLMs play a growing role in areas such as healthcare, autonomous vehicles, and the workplace environment, it is crucial to explore how people perceive LLM-generated responses to different scenarios in comparison to those provided by humans. Because this is an emerging field, research in this area is still incomplete and mostly focuses on how AI systems are designed and integrated into decision-making processes, with the hopes of ensuring that they are used responsibly and transparently. Previous work has put an emphasis on integrating ethical theories into artificial moral agents [24] and designing systems that simulate ethical decision-making [1]. However, there has been limited exploration of how the perception that people have of LLMs' might impact the perception that they have of their own reasoning, opinions and attitudes. This gap leaves questions unanswered about whether people evaluate the credibility, quality and trustworthiness of ethical reasonings differently when the perceived source is an LLM compared to another human decision-makers.

In this paper, we address these previously mentioned gaps by exploring the participants perceptions of LLM vs human responses in ethical scenarios and examine the influence of contrasting opinions when they are either attributed to ChatGPT or another human participant. More precisely, we investigate on how participants perceive LLM-generated answers, in this case ChatGPT, versus human-generated answers in 4 themes of ethically complex scenarios: autonomous vehicle scenarios, medical resource allocation scenarios, trolley problem variations and workplace dilemmas scenarios.

In turn, we also examine how the two conditions, whether answers are attributed to humans or LLMs, have an influence on the individuals' willingness to consider the proposed perspective and revise their initial responses to the same questions.

Through a mixed-methods approach combining qualitative and quantitative analyses, we show that participants are more open

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

to revising their opinions when they believe alternative responses come from other human participants and that, conversely, they tend to be more defensive about their initial opinions when alternative answers are attributed to an LLM, indicating a fundamental difference in how AI-generated and human-generated reasoning is perceived.

These contributions provide valuable implications for designing AI systems that can better engage with human users in morally complex decision-making. By shedding light on the biases and dynamics of human-AI interactions, we aim to inform the ethical deployment of LLMs in real-world contexts, fostering systems that are not only efficient but also trusted and aligned with human values.

2 MOTIVATION AND RELATED WORK

Artificial Intelligence (AI) is increasingly automating decision- making in diverse industry sectors. In many high-stakes domains such as healthcare, education, criminal justice systems, organizational management and public assistance[6, 7, 23] AI systems are automating or augmenting decisions that human experts used to make. In order to better understand people's reactions to this change, many scholars have investigated how people perceive algorithmic decisions as compared to human decisions.

Their findings suggest that people tend to perceive algorithmic decisions as inferior to human decisions and are resistant to following them. In several studies, people trusted algorithmic decisions less than human decisions and were less likely to adopt them, particularly when tasks were deemed to require a human's unique capabilities [16], be subjective [5] or require attention to individual uniqueness [18].

Even in domains where AI consistently outperforms humans, such as predictive analytics or diagnostic medicine, adoption and acceptance can be hindered by users' skepticism and resistance to relinquishing control [3, 16].

Via generative adversarial networks (GANs), artificial intelligence (AI) has influenced many areas, especially the artistic field, as symbol of a human task. The work [20] presents a wide-scale experiment in which 565 participants were asked to evaluate paintings (which were created by humans or AI) on four dimensions: liking, perceived beauty, novelty, and meaning.

However, not all people experience trust or skepticism in the same way. Those who face marginalization or discrimination in human-led systems, such as minority groups or those with prior experiences of bias, may view algorithmic decisions as more neutral or fair compared to human judgments. Conversely, people with a deeper understanding of AI processes might become more skeptical, especially if they perceive these systems as opaque, prone to errors, or biased by their training data [10, 19]. Demographics, such as age, education, and cultural background, are all factors that can also shape how individuals judge the fairness and credibility of algorithmic decisions.

This project builds on these findings to explore an under-examined question: how does perception of AI-generated reasoning influence human decisions? Our work investigates this dynamic across four ethically sensitive and relatable scenarios: autonomous vehicle decision-making, medical resource allocation, variations of the trolley problem, and workplace ethical dilemmas. These contexts were chosen for their relevance to everyday ethical challenges and their possible ability to evoke nuanced human-AI comparisons.

Our study is derived from the work on explainability, algorithmic fairness, and trust calibration in human-AI interactions [9, 13, 21]. By employing a mixed-methods approach, we aim not to analyze the decisions themselves, but the reasoning processes and biases underlying them. To do so , quantitative metrics like agreement rates and response alignment will be complemented by qualitative insights into participants' justifications and reflections.

This research aims to deepen our understanding of how people interact with AI in morally sensitive decision-making scenarios, addressing critical questions about when and where AI is trusted to make decisions.

In this work, we seek to provide theoretical insights for designing AI systems that are not only technically proficient but also perceived as credible, understandable, and acceptable by diverse populations. Specifically, further we could investigate strategies for explaining AI decisions in ways that enhance their transparency and create trust, encouraging people to feel confident in the use of AI. Ultimately, this research aspires to inform the responsible integration of AI into decision-making processes, identifying contexts where AI is trusted while ensuring its alignment with human needs and ethical standards.

3 METHODOLOGY

The goal of this experiment is to understands the perception of AI generated answers to ethical dilemmas in contrast to human generated answers and its impacts on the participants perception of their own answers.

In the first part of this study, participants were asked to answer a survey containing 4 ethical dilemmas scenarios with 4 multiple-choice questions each.

Then, in the second part, they were asked to review another set of answers to the same questions by answering Likert-style questions about the answer's credibility, their agreement with the reasoning, their confidence in their original answer after reading the alternative answer and how likely they would be to change their answer.

Lastly, in the third part of the survey, they were offered the possibility to change their answer and justify why they would or would not.

This is a between subject deception study where one group was told that the alternative answers they were reviewing came form ChatGPT and they other group was deceived into thinking they were the answers of another human participant.

To explore our goal, we formulated the following research hypotheses :

- H1: Participants will perceive answers generated by LLMs differently from human-generated answers, with potential biases in how participants evaluate the credibility and reasoning behind the responses.
- H2: Participants will be more likely to change their initial responses after being exposed to a differing opinion, with

the likelihood of change being influenced by whether the differing opinion is attributed to a human or an LLM.

3.1 Participants

Participants were recruited online through a social media post that only disclosed that the study was about answers to complex ethical dilemmas (Appendix A) to ensure that the participants that would ultimately be assigned to the deceived group would not have doubts about the real source of the answers they were reviewing. The recruiting message was kept vague and participants that had prior knowledge of our area of study were automatically disqualified to reduce the demand characteristic bias to a minimum.

To identify other participants that would need to be disqualified, the answers to the survey were manually inspected to look for incomplete surveys. Only one participant was omitted from the analysis for this reason, leaving 17 valid survey responses [table 1]. Participants did not receive any monetary compensation for completing the survey. Demographic details, such as age, gender, and educational background, were collected during a pre-study survey.

3.2 Task

The task for participants was divided into 3 main parts, excluding the pre-study survey where participants had to answer diverse demographic questions.

In Part 1, they were asked to answer multiple-choice questions about various ethical dilemmas. This part of the survey consisted of 4 ethical dilemmas themes: autonomous vehicles[22], medical resource allocation[8, 12, 15], trolley problem variations[4, 11, 14] and workplace dilemmas[2, 17]. Each one of the different themes had 4 multiple-choice questions which added to a total of 16 multiple-choice questions. To ensure the quality and relevance of the questions, they were all selected from different published papers.

In the second part (Appendix B), Participants were then presented with a set of answers to the same dilemmas, labeled as either AI-generated (ChatGPT) or human-generated, depending on their assigned group. Within the fours dilemmas, the questions to half of them were randomly assigned alternative responses, while the other half retained the same answers as the participant with the alternative explanations generated by ChatGPT. For each of the ethical dilemmas, participants answered Likert-scale questions assessing: Credibility of the response, Agreement with the reasoning behind the response, Confidence in their original answer after reviewing the alternative response and Likelihood of changing their initial answer.

In the third and final part of the study, participants were explicitly asked if they wanted to change their original answers after considering the alternative responses. If so, they were asked how many questions and to explain their answers to provide more insight into their reasoning (Appendix B).

3.3 Setup

A Google Form survey was used to collect participant responses during the study. We utilized four separate consecutive Google Form surveys, the first one being the pre-study survey and each of the remaining ones designed to address one of the distinct 3 parts of the research .

The surveys were administered during a recorded Zoom call to ensure that participants understood the assignment and followed the instructions accurately. During the session, participants were presented with multiple choice questions and open-ended questions, for which they provided their responses in short-answer text boxes within the Google Form.

For the second survey, where they had to review a set of answers (all generated by ChatGPT and predetermined by the researchers), participants were shown a prefilled Google Form containing answers attributed to the "other participant". The recorded Zoom call allowed us to observe the participants' engagement and clarify any misunderstandings in real time. This setup ensured that the collected responses reflected participants' understanding of the questions without external influence.

3.4 Study design

We opted for a between-subjects design over a within-subjects design to keep the experiment shorter for each participant and to prevent order effects across conditions (e.g. their evaluation of one condition could influence their perceptions in the subsequent condition, compromising the validity of the findings). Our study aimed to understand participants' perceptions of moral judgments and credibility in relation to the source of the responses (AI vs. human). If participants were exposed to both types of responses, they may develop a comparison mindset, consciously or unconsciously evaluating responses relative to each other. This would compromise the study to measure participants' natural, raw reactions to either AI or human-generated responses. By ensuring that participants in each group are exposed to only one condition, the between-subject design allows us to generalize findings more confidently. It reflects real-world scenarios where individuals typically encounter responses from one source at a time (e.g., interacting with an AI system or receiving input from a human expert).

There is one primary independent variable, the perceived source of the reviewed answers, with 2 levels: ChatGPT or Human generated. Participants were randomly assigned to 1 of the 2 groups (Group A: Condition = ChatGPT or Group B: Condition= Human) so that they would only experience one condition. The primary quantitative measures were Credibility of the response (Credible), Agreement with the reasoning behind the (Agree), Confidence in their original answer after reviewing the alternative response (Confidence), Likelihood of changing their initial answer (Changing) and How many answers they would like to change (Changed). All the measures were a 1-5 scale, except for how many answers they desired to change. The post-survey questionnaire also included open-ended questions that provide qualitative measures.

3.5 Analysis

The data analysis involved both quantitative and qualitative methods to thoroughly examine the participants' responses. Both quantitative and qualitative analyses were used to answer the first hypothesis (H1), whereas only quantitative analysis was used for the second hypothesis (H2).

Table 1: Participant Demographics

Participant	Age	Gender	Education Level	Employment Status	Occupation / Study Area
p1	18-24	Female	Doctoral degree or higher	Student	Clinical Psychology (Ph.D.)
p2	25-34	Female	Master's degree	Unemployed	Geography-Urban Development
p3	18-24	Male	Bachelor's degree	Employed full-time	Software Developer, CS
p4	25-34	Male	Some college	Student	Education
p5	35-44	Male	Doctoral degree or higher	Employed full-time	Data Scientist, CS
p6	55-64	Female	Bachelor's degree	Employed full-time	Teaching
p7	18-24	Male	Bachelor's degree	Student	Mechanical Engineering
p8	25-34	Non-binary	Bachelor's degree	Student	Law
p9	25-34	Female	Bachelor's degree	Employed part-time	Backend Developer
p10	25-34	Female	Master's degree	Employed part-time	Civil Engineering
p11	35-44	Female	Master's degree	Employed part-time	Software Development, CS
p12	18-24	Female	Some college	Employed full-time	Nursing
p13	25-34	Female	Master's degree	Employed full-time	Clinical Counselor
p14	35-44	Male	Master's degree	Student	Cloud/Fog Computing
p15	25-34	Male	Master's degree	Student	Computer Engineering
p16	25-34	Female	Master's degree	Student	–
p17	18-24	Male	Bachelor's degree	Student	Video Game Developer, CS

3.5.1 Quantitative analysis. Before conducting any advanced statistical analyses, we first checked the normality of the data for each variable using the Shapiro-Wilk test. This test helps us determine whether the distribution of a variable is close to a normal (bell-shaped) distribution. We applied the Shapiro-Wilk test to the following variables: Agree (how much participants agreed with the answers), Credible (how credible they found the explanations), Confidence (how confident they were in their original answers), Changing (how many answers they wanted to change) and Changed (how many answers they actually changed). The results of the Shapiro-Wilk test for all these variables showed p-values smaller than 0.05, indicating that the data were not normally distributed. To visually confirm this, we also generated Q-Q plots, which showed that the data indeed deviated from a normal distribution.

Given that the data were not normally distributed and showed significant variability, we decided to use Linear Mixed Models (LMM) for our analysis. LMMs are appropriate for this kind of data because they can handle non-normal distributions and allow us to model both fixed effects (variables we expect to have a consistent influence, like Group or Scenario) and random effects (individual differences between participants).

For each of the five dependent variables, we fitted two models:

Main Model: This model included Group as a fixed effect and Participant-ID as a random effect. The goal was to examine how the type of response (AI-generated vs. human-generated) influenced participants' ratings for each variable.

Extended Model: The second model expanded on the main model by adding Answers-num as a fixed effect. This variable indicates whether the reviewed answer was the same as the participant's original response or different. By adding this variable, we could assess whether agreeing or disagreeing with the alternative response affected the participants' ratings, over and above the influence of Group. We fit these models using the lmer() function from the lme4 package in R. After fitting the models, we examined their summaries to assess the significance of the fixed and random effects for each dependent variable.

In brief:

- Group served as a fixed effect.
- Participant ID was included as a random effect to account for variability across individuals.
- An additional factor, Answers-num (representing whether the responses were the same or different), was added to extended models for more detailed insight.
- We also included age, education degree, occupation level, and gender as covariates to check if participants' demographic characteristics influenced their responses.

Once the models were fit, we conducted post-hoc tests to explore any significant findings further. These tests help to pinpoint which specific comparisons or interactions are driving the effects. Post-hoc tests are important for gaining deeper insights into the data and allow us to interpret the findings more clearly.

For each dependent variable (Agree, Credible, Confidence, Changing, and Changed), we evaluated the models based on:

The significance of the fixed effects (whether Group or Scenario significantly impacted the ratings).

The variance explained by the random effects (how much individual differences, represented by Participant-ID, contribute to the responses).

The overall fit of the models (how well the models explain the data).

Lastly, we inspected the variance components to determine whether individual differences (such as Participant-ID, Age group, Gender, Occupation level, and Education level) played a larger role in explaining the differences in participants' responses. This evaluation was crucial for understanding what factors influence how participants perceive the answers, whether AI-generated or human-generated , and how much these factors are accounted for in the models.

In addition to the quantitative analysis described above, we also conducted a qualitative analysis to provide a more comprehensive understanding of the participants' perceptions and to fill in any gaps that may not be captured by the statistical models.

3.5.2 Qualitative analysis. To analyze the open-ended responses, we employed a qualitative approach in two stages of coding. First, we conducted inductive coding to identify emergent themes directly from the data without imposing preconceived categories. This process allowed us to capture participants' perspectives in their own words. Subsequently, we carried out focused coding to refine and consolidate the initial themes into more structured and meaningful categories.

To further explore the data, we also created an affinity diagram for each group (Group A: Condition= ChatGPT and Group B: Condition=Human), a technique which involves organizing and grouping the responses into clusters based on thematic similarities.

In the affinity diagram, the answers to the last open-ended question (In the open-ended section, did you face any obstacles when you were trying to explain your choice?) were omitted because they did not provide any insight on the participant's perception of the answers. By comparing the affinity diagrams of the groups, patterns, differences, and unique themes to each group were identified. This comparative analysis provided insights into variations in participant responses across the conditions.

The combination of coding and affinity diagramming ensured a robust qualitative analysis, enabling us to systematically uncover and compare the nuances of participant perspectives and their perceptions of the answers depending on which group they were asigned to.

4 FINDINGS

4.1 Quantitative

Before performing any statistical tests, we checked if the data followed a normal distribution using the Shapiro-Wilk test. The variables we tested were: Agreement (how much participants agreed with the answers), Credibility (how credible they found the answers), Confidence (how confident they were in their original answers), Likelihood of Changing (how likely they were to change their answers), Changed (how many questions they desired to change their answer to).

The results of the Shapiro-Wilk test showed that all the variables had p-values smaller than 0.05, which means that the data were not normally distributed. This non-normal distribution is important because it justifies the use of Linear Mixed Models (LMM) for further analysis, which can handle non-normal data more effectively.

In addition, we observed variance in the data, meaning there was a lot of variation in the responses. We noted that the AI-generated group exhibited larger variance compared to the human-generated group, suggesting that participants were less consistent in how they responded to AI-generated answers compared to human-generated ones.

4.1.1 Confidence Scores. Both Group (Estimate = -0.46181, Std = 0.33699, Error = 15.00000, t value = -1.370, p = 0.191) and Scenario (Estimate = 0.11765, Std = 0.07487, Error = 50.00000, t value = 1.571, p = 0.122) were non-significant predictors of confidence, indicating that confidence levels were not influenced by group membership or the specific scenario presented to the participants.

Answers_num (Estimate = -0.006925, Std = 0.169170, Error = 49.00000, t value = -0.041 , p = 0.968) had a significant negative effect, meaning that in non-matching answer cases, participants' confidence in their original answers decreased.

Although we did not observe a statistically significant effect, participants who viewed AI-generated answers appeared to have greater confidence in their original responses. figure1

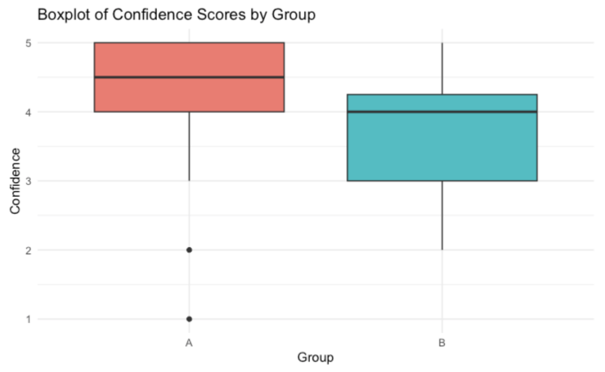


Figure 1: Confidence Scores by Group

4.1.2 Credibility Scores. Group (Estimate = 0.3611, Std = 0.3524 Error = 15.0000, t value = 1.025 ,p = 0.322) and Scenario (Estimate = -0.1647 , Std = 0.1205 Error = 50.0000, t value = -1.367, p = 0.178) did not significantly affect the Credibility ratings. This implies that participants' ratings of how credible they found the answers were not influenced by the group they belonged to or the scenario. figure 2

Answers_num (Estimate = -0.5201, Std = 0.2619, Error = 49.0000, t value = -1.986, p = 0.0527) showed a marginally significant negative effect . While this effect was not highly significant, it suggests a trend where non-matching answers were associated with lower credibility ratings. While Group and Scenario had no significant impact, the Answers_num factor was important, showing that when participants where faced with a non-matching answer to their own, they tended to rate the answers less credible.

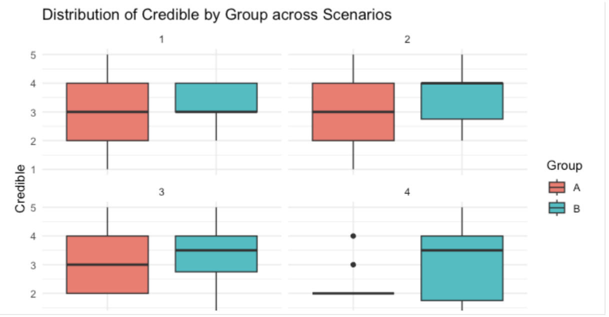


Figure 2: Credibility Scores Across Groups for each scenario.

4.1.3 Agreement scores. Both Group (Group A: Condition = ChatGPT and Group B: Condition = Human) and Scenario (the four different ethical dilemma themes) had no significant effects on Agreement.

This was evident in both models without and with the inclusion of Answers_num (whether the answers provided were the same or different from the participants' original answers). The non-significant results suggest that participants' agreement with answers was not influenced by Group (Estimate = 0.05556, Std = 0.34296, Error = 15.00000, t value = 0.162, p = 0.873) or Scenario (Estimate = -0.17647, Std = 0.11818, Error = 50.00000, t value = -1.493, p = 0.142) figure 3

The Answers_num factor had a significant positive effect (Estimate = -0.93144, Std = 0.23151, Error = 49.00000, t value = -4.023, p = 0.0002), indicating that when the answers were marked as "same" as the original participant's answers, the agreement scores increased. This suggests that when participants reviewed answers matching their original answers, they tended to agree more with the reasoning they were presented with, even if the justification was different from their own, irrespective of the group.

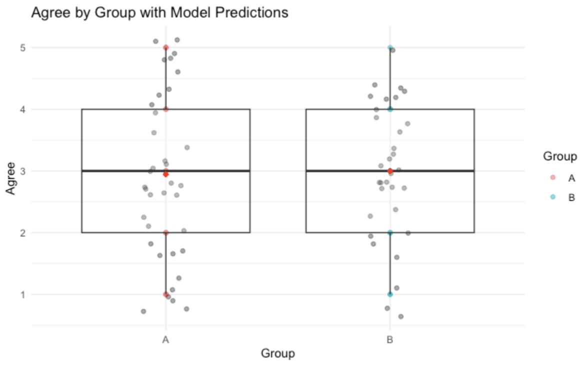


Figure 3: Agreement Scores by groups. Group A people who got AI_generated answers, B participants who got Human generated answers.

4.1.4 Likelihood of Changing Answers. Neither Group (Estimate = 0.18056, Std = 0.22521, Error = 14.99998, t value = 0.802, p = 0.435) nor Scenario (Estimate= 0.01765, Std = 0.09967, Error = 49.99998,t value = -0.177, p = 0.142) had significant effects on the likelihood of changing answers. This indicates that participants' tendency to change their responses was not influenced by the group they were in or the scenario under which they were evaluating the answers. figure **??**

The Answers_num (p = 0.0002) factor showed a significant negative effect . This suggests that when participants were provided with matching answers were less likely to change their responses.

4.1.5 Number of answers changed. For the number of questions to which participants changed their answers, Group (p = 0.322) and Scenario (p = 0.178) showed no significant effects. This suggests that participants' changing answers weren't influenced by the group they were in or the scenario they were presented with.

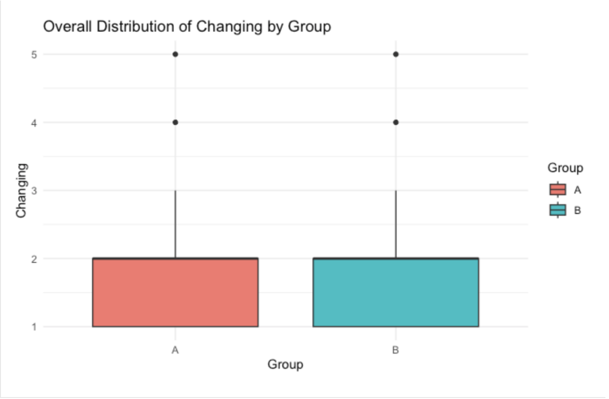


Figure 4: Likelihood of Changing Answers Across Groups

Answers_num (p = 0.0527) showed a marginally significant negative effect , indicating participants were more likely to change their original answer when they were reviewing a non matching answer.

Across all 30 models analyzed, Answers_num consistently emerged as the most significant variable across all five outcomes (Agreement, Credibility, Confidence, Changing, and Changed). The findings indicate that same or different answers has a strong negative influence on their perceptions of agreement, credibility, confidence and likelihood of change. When participants were reviewing answers in disagreement with their own original answer, they generally exhibited lower agreement, credibility, and confidence and were more likely to change their answers.

On the other hand, Group and Scenario were statistically non-significant, suggesting that individual participant differences were the primary drivers of variance in these outcomes. The lack of significant effects for these two factors implies that the experimental conditions (Group and Scenario) did not meaningfully influence how participants evaluated or responded to the questions.

4.1.6 Post HOC Test. To further explore the relationships between variables, a post-hoc test was conducted using pairwise comparisons between groups. These tests confirmed that while differences between groups were not always statistically significant, patterns of increased variability and slight biases against AI responses were consistent across measures.

For example, Credibility ratings demonstrated a near-significant difference favoring human-generated responses. Similarly, the Likelihood of Changing showed a clear tendency toward greater influence by human-generated different responses.5

4.1.7 Demographic Analysis. To study the variance among participants in more detail, we analyzed a linear mixed-effects model (LMM) for four demographic variables: age, educational level, gender, and employment level.

Age group 55-64 showed a significant positive effect on credibility (p = 0.00991), specifically in Group A (Condition= ChatGPT). Older participants in this group rated the credibility of the answers significantly higher compared to younger participants. This suggests that older individuals (ages 55-64) in Group A had higher

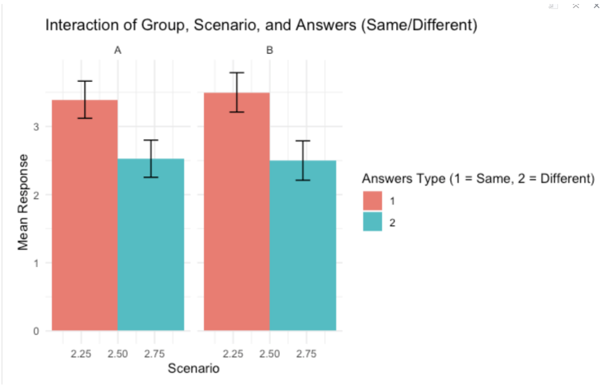


Figure 5: higher credibility score in human-generated group observing same answers

trust in the AI-generated answers compared to their younger counterparts. However, Group B did not show significant age-related differences in Credibility, indicating that this trend may be unique to Group A. figure 6

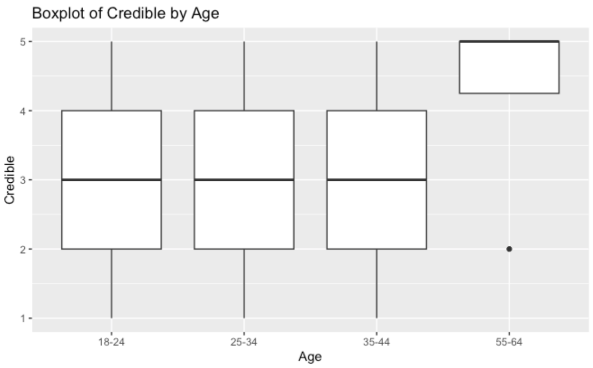


Figure 6: participant with higher age group rated higher in the credibility score seeing AI generated responses

Participants with a Bachelor's degree exhibited a significant negative relationship with credibility (p = 0.04128) in Group B, indicating that those with a Bachelor's degree rated the credibility of the answers lower compared to those with higher levels of education or those in Group A (Condition = ChatGPT). A similar but marginally significant trend was observed for participants with a Master's degree (p = 0.06194), though this was not as pronounced in either group. The trend indicated that those with a Master's degree in Group B rated the answers slightly lower in terms of credibility.

Gender did not show any significant effect on agreement, credibility, confidence, or likelihood of changing answers in either Group A or Group B. The comparison between Male, Female, and Non-binary/non-conforming participants did not yield substantial differences in how the answers were perceived. This suggests that gender was not a key factor in shaping participants' perceptions in this study.

Employment status had minimal influence on the study outcomes. Participants who were Employed part-time, Students, and Unemployed did not show significant effects on agreement, credibility, or confidence in either Group A or Group B. However, there were small variations in some measures, but none reached statistical significance, suggesting that employment status did not play a significant role in shaping participants' perceptions in this survey. figure 7

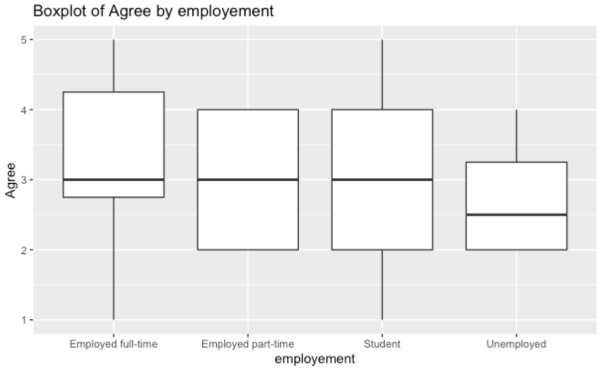


Figure 7: no difference among various employment group

4.2 Qualitative

The qualitative analysis revealed four key themes, that emerged from the inductive coding, that highlight the contrasting attitudes between Group A and Group B in their perspective of the different answers and their willingness to revise their own answers: Openness to Alternative Perspectives, Defensiveness and Resistance, Doubts, and Challenges and Difficulties.

Group A exhibited more defensive and critical attitudes toward alternative perspectives, while Group B displayed greater openness and accorded more value to differing viewpoints.

These findings were further supported by the affinity diagram (Appendix C), which demonstrated an overall more positive attitude in Group B and a more negative, dismissive attitude in Group A.

4.2.1 *Openness to Alternative.*

Group B participants demonstrated a greater openness to alternative perspectives and a stronger willingness to revise their answers. Participants frequently Noted the quality of other answers, Acknowledged the validity of different reasoning, and Valued other perspectives.

These are all codes that were not present in Group A, but that appeared frequently in Group B. A participant shared, "It seems to me that her/his answer was logical, and it would be better if more people were saved," showing a consideration of alternative reasoning.

Another participant noted, "Yes, seeing different perspectives always helps for further choices" highlighting the value of alternative viewpoints for improving decision-making or revising their own answers.

Although Group B participants occasionally Defended their initial answers or Recognized the limited impact of alternative perspectives, these instances were less prominent than in Group A as shown by the affinity diagram.

4.2.2 Defensiveness and Resistance. In contrast to Group B, Group A participants were generally more defensive about their opinions and less receptive to alternative perspectives. Many participants also minimized the relevance of other perspectives, as reflected in codes like Recognizing limited impact of alternative perspectives, Dismissing alternative perspective and Reducing AI credibility.

Additionally, some responses critiqued or questioned the quality of alternative answers. For example, when asked if seeing the different answers affect the way you would approach the same problems in the future, one participant stated: "No, most answers were incoherent or nonsensical," reflecting skepticism toward the credibility of the alternative answers.

Another participant answered to the same question: "No, it didn't, because I know the answers come from an AI" further indicating resistance to considering these perspectives as valuable. While some participants engaged in reflection and considered alternative perspectives, the affinity diagram reveals that they are indeed a minority in this group.

4.2.3 Doubts. The qualitative analysis also revealed that participants in both groups occasionally expressed doubts and uncertainty about their answers (Expressing doubts), along with moments where they acknowledged the potential imperfection of their reasoning (Acknowledging personal imperfection in judgment).

These instances highlight that some participants were aware of the limitations of their decision-making process. However, the results from the inductive coding analysis and the affinity diagram indicate that this is not a major finding that emerged from the data since these codes only occurred sporadically across both groups.

Consequently, while these findings provide some insight into participants' self-reflection, they represent a minor theme in the overall analysis.

4.2.4 Challenges and Difficulties. Both groups expressed occasional difficulty with the questions, reflecting the complexity of the task and the cognitive effort required to engage with it.

Participants noted struggles with balancing emotional and ethical considerations (Struggling with choosing between emotional and ethical choices) and, at times, admitted to finding certain questions challenging (Admitting occasional difficulty). This does not show a contrasting difference between Group A and Group B since these difficulties were present in both groups.

5 DISCUSSION

To summarize, our results shows that participants generally perceived answers generated by other human participants in a more favorable way than answers generated by ChatGPT. Participants in the group that was deceived into thinking that the answers they were reviewing were the ones of another human participants were more open to consider the alternative reasoning, rated it as more credible and accorded more value to a differing opinion.

In contrast, participants that were told the answers were LLM generated were typically more defensive about their own answers and more dismissive of the alternative opinion.

We discuss related research directions our work opens up for future work, speculations on the results and the limitations of our work.

5.1 Unexpected findings and Speculations

The most unexpected finding in our study was that participants did not significantly rate the alternative answers as less credible or less agreeable according to the quantitative analysis.

This result was surprising, as prior research consistently suggests that people perceive algorithmic decisions less favorably than human decisions. A likely explanation for this discrepancy is the limited number of participants in our study, which may have reduced the statistical power needed to detect significant differences. With a larger sample size, it is possible that more pronounced patterns would emerge, reflecting the skepticism and resistance to AI-generated perspectives that are already documented in existing literature. This will be further explained in the *Limitations section.*

However, participants' self-reported opinions in the qualitative analysis provide further support for this interpretation, as the qualitative analysis revealed skepticism of AI's credibility in terms of ethical judgment (defensiveness over own answers and dismissal of alternative answers) in Group A (AI-generated) that was not apparent in Group B.

It is also surprising that the quantitative analysis did not reveal that participants were significantly changing their answers more when the perceived source of the answers was another human participant, because the qualitative analysis strongly suggested that there is indeed a negative bias towards AI-generated answers, but this could be explained with the same reasoning.

5.2 Reflection

While the quantitative findings did not align with our expectations and the literature, the qualitative analysis closely agrees with prior work.

Specifically, the participants in Group A demonstrated greater resistance to alternative perspectives, often dismissing AI-generated answers as incoherent or poor quality. For example, statements like *"No, it didn't, because I know the answers come from an AI"* indeed reflect well-documented trends in the literature that users trust algorithmic decisions less and are reluctant to adopt them [5, 16].

This skepticism may stem from a perception that AI lacks the human-like reasoning required for nuanced tasks or might also come from concerns about transparency and the potential for errors and biases in AI systems [10, 19].

While the quantitative results deviated from expectations, the qualitative analysis reinforces existing literature on resistance toward algorithmic decisions. This also highlights the value of mixed-method approaches, as the qualitative findings offer a deeper understanding of participant attitudes, thoughts and perception that might not have been captured if only a quantitative analysis was conducted.

5.3 Limitations

Because of the limited time frame, our participants sampling was not balanced in terms of background, and the sample size was insufficient for a between-subject experiment. The relatively small sample size limited our ability to detect statistically significant results in the quantitative analysis.

Increasing the number of participants would provide a more robust dataset, allowing the differences in how credibility and agreeability of AI-generated and human answers are perceived, that seem to be consistent in the literature,to emerge from the analysis. A bigger sample size would also improve the reliability of the patterns identified across the two different conditions.

The participant pool was also heavily skewed toward individuals with computer science backgrounds. The stronger familiarity with AI systems that can people in this field have may have influenced their attitudes, introducing biases that might not reflect those of the general population. Enhancing the diversity by including participants from a wider range of educational and professional backgrounds make the findings more generalizable.

Lastly, while our ethical dilemmas were designed to provoke thoughtful responses, a broader range of scenarios could be an interesting addition because participants may respond differently depending on the specific domain or context of the ethical questions presented. For instance, dilemmas involving environmental ethics might evoke different reactions compared to autonomous vehicle dilemmas. Expanding the variety of questions would provide deeper insights into how participants perceive AI-generated ethical reasoning across more diverse contexts.

REFERENCES

[1] Giulio Antonio Abbo, Serena Marchesi, Agnieszka Wykowska, and Tony Belpaeme. 2024. Social Value Alignment in Large Language Models. In *Value Engineering in Artificial Intelligence*, Nardine Osman and Luc Steels (Eds.). Springer Nature Switzerland, Cham, 83–97.

[2] John P Allegrante and Richard P Sloan. 1986. Ethical dilemmas in workplace health promotion. , 313–320 pages.

[3] Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão, and Sonia Sousa. 2024. A systematic literature review of user trust in AI-enabled systems: An HCI perspective. *International Journal of Human–Computer Interaction* 40, 5 (2024), 1251–1266.

[4] April Bleske-Rechek, Lyndsay A Nelson, Jonathan P Baker, Mark W Remiker, and Sarah J Brandt. 2010. Evolution and the trolley problem: People save five over one unless the one is young, genetically related, or a romantic partner. *Journal of Social, Evolutionary, and Cultural Psychology* 4, 3 (2010), 115.

[5] Noah Castelo, Maarten W Bos, and Donald R Lehmann. 2019. Task-dependent algorithm aversion. *Journal of Marketing Research* 56, 5 (2019), 809–825.

[6] John Danaher. 2016. The threat of algocracy: Reality, resistance and accommodation. *Philosophy & technology* 29, 3 (2016), 245–268.

[7] John Danaher, Michael J Hogan, Chris Noone, Rónán Kennedy, Anthony Behan, Aisling De Paor, Heike Felzmann, Muki Haklay, Su-Ming Khoo, John Morison, et al. 2017. Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big data & society* 4, 2 (2017), 2053951717726554.

[8] Angus Dawson, David Isaacs, Melanie Jansen, Christopher Jordens, Ian Kerridge, Ulrik Kihlbom, Henry Kilham, Anne Preisz, Linda Sheahan, and George Skowronski. 2020. An ethics framework for making resource allocation decisions within clinical care: responding to COVID-19. *Journal of bioethical inquiry* 17 (2020), 749–755.

[9] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–19.

[10] Sina Fazelpour and David Danks. 2021. Algorithmic bias: Senses, sources, solutions. *Philosophy Compass* 16, 8 (2021), e12760.

[11] Natalie Gold, Andrew M Colman, and Briony D Pulford. 2014. Cultural differences in responses to real-life and hypothetical trolley problems. *Judgment and Decision making* 9, 1 (2014), 65–76.

[12] Keegan Guidolin, Jennifer Catton, Barry Rubin, Jennifer Bell, Jessica Marangos, Ann Munro-Heesters, Terri Stuart-McEwan, and Fayez Quereshy. 2022. Ethical decision making during a healthcare crisis: a resource allocation framework and tool. *Journal of medical ethics* 48, 8 (2022), 504–509.

[13] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[14] Alessandro Lanteri, Chiara Chelini, and Salvatore Rizzello. 2008. An experimental investigation of emotions and reasoning in the trolley problem. *Journal of Business Ethics* 83 (2008), 789–804.

[15] Naomi Laventhal, Ratna Basak, Mary Lynn Dell, Douglas Diekema, Nanette Elster, Gina Geis, Mark Mercurio, Douglas Opel, David Shalowitz, Mindy Statter, et al. 2020. The ethics of creating a resource allocation strategy during the COVID-19 pandemic. *Pediatrics* 146, 1 (2020).

[16] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.

[17] Cheryl Lindy and Florence Schaefer. 2010. Negative workplace behaviours: an ethical dilemma for nurse managers. *Journal of Nursing Management* 18, 3 (2010), 285–292.

[18] Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. 2019. Resistance to medical artificial intelligence. *Journal of Consumer Research* 46, 4 (2019), 629–650.

[19] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature machine intelligence* 1, 11 (2019), 501–507.

[20] Martin Ragot, Nicolas Martin, and Salomé Cojean. 2020. Ai-generated vs. human artworks. a perception bias towards artificial intelligence?. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. 1–10.

[21] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International journal of human-computer studies* 146 (2021), 102551.

[22] Kazuhiro Takemoto. 2024. The moral machine experiment on large language models. *Royal Society open science* 11, 2 (2024), 231393.

[23] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[24] John Zoshak and Kristin Dew. 2021. Beyond Kant and Bentham: How Ethical Theories are being used in Artificial Moral Agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 590, 15 pages. https://doi.org/10.1145/3411764.3445102

6 APPENDICES

A RECRUITMENT MESSAGE

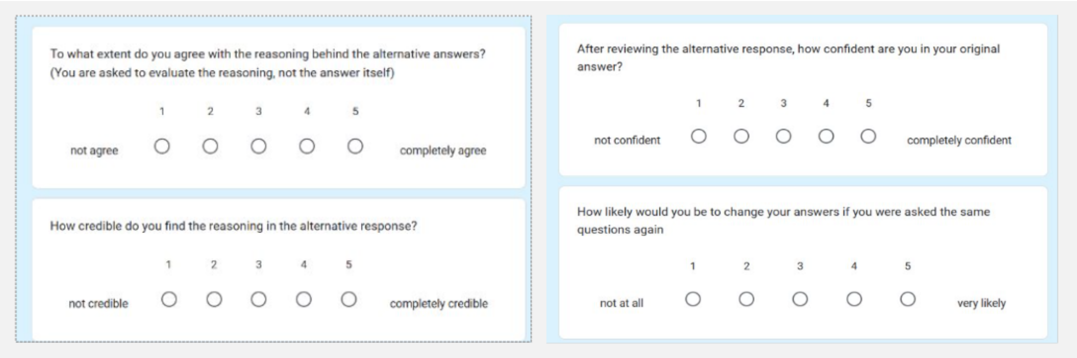
Join Our Study on Moral Decision-Making!

We're looking for participants to take part in a study exploring responses to moral dilemmas by answering 16 questions and judging some answers. This research aims to understand how people approach moral decisions and perceptions of differing opinions.

Your participation will help us better understand human decision-making and reasoning in complicated moral dilemmas. Interested in contributing? Contact us at azalee.robitaille@umontreal.ca or yalda.kasravi.mavi@umontreal.ca for more details.

Thank you for your help!

B SURVEY QUESTIONS



To what extent do you agree with the reasoning behind the alternative answers? (You are asked to evaluate the reasoning, not the answer itself)

	1	2	3	4	5	
not agree	○	○	○	○	○	completely agree

How credible do you find the reasoning in the alternative response?

	1	2	3	4	5	
not credible	○	○	○	○	○	completely credible

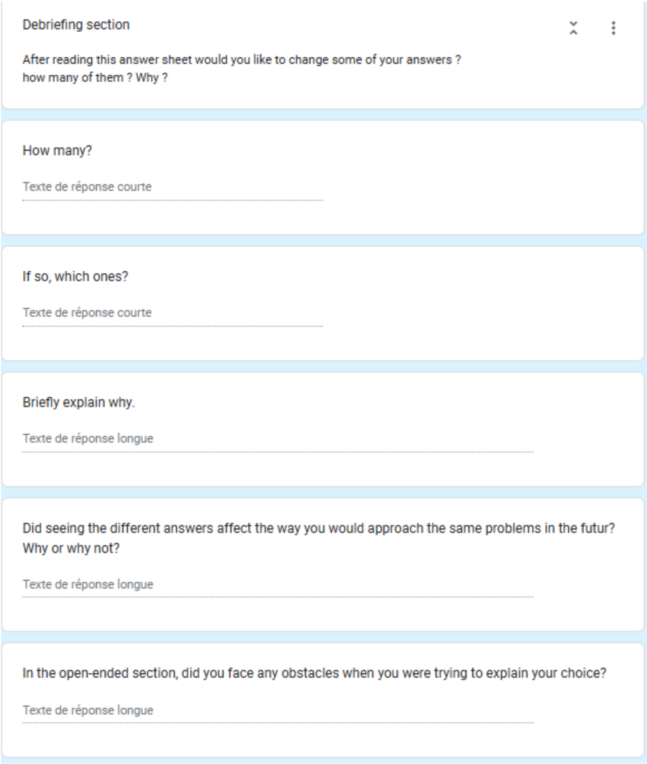
After reviewing the alternative response, how confident are you in your original answer?

	1	2	3	4	5	
not confident	○	○	○	○	○	completely confident

How likely would you be to change your answers if you were asked the same questions again

	1	2	3	4	5	
not at all	○	○	○	○	○	very likely

Figure 8: Part 2 Survey Questions



Debriefing section

After reading this answer sheet would you like to change some of your answers ? how many of them ? Why ?

How many?

Texte de réponse courte

If so, which ones?

Texte de réponse courte

Briefly explain why.

Texte de réponse longue

Did seeing the different answers affect the way you would approach the same problems in the futur? Why or why not?

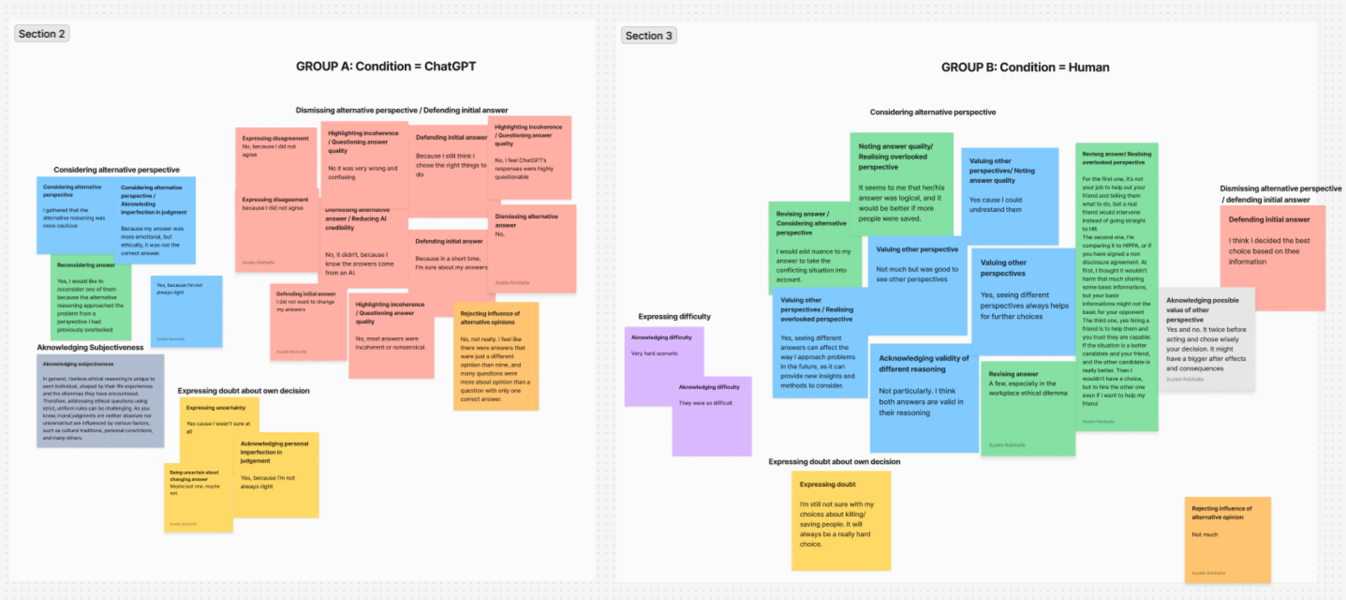
Texte de réponse longue

In the open-ended section, did you face any obstacles when you were trying to explain your choice?

Texte de réponse longue

Figure 9: Part 3 Survey Questions

C AFFINITY DIAGRAM



GROUP A: Condition = ChatGPT

Considering alternative perspective

Considering alternative perspective

I gathered that the alternative reasoning was more cautious

Considering alternative perspective / Aknowleding imperfection in judgment

Because my answer was more emotional, but ethically, it was not the correct answer.

Reconsidering answer

Yes, I would like to reconsider one of them because the alternative reasoning approached the problem from a perspective I had previously overlooked

Yes, because I'm not always right

Aknowledging Subjectiveness

Aknowledging subjectiveness

In general, I believe ethical reasoning is unique to each individual, shaped by their life experiences and the dilemmas they have encountered. Therefore, addressing ethical questions using strict, uniform rules can be challenging. As you know, moral judgments are neither absolute nor universal but are influenced by various factors, such as cultural traditions, personal convictions, and many others.

Dismissing alternative perspective / Defending initial answer

Expressing disagreement
No, because I did not agree

Highlighting incoherence / Questioning answer quality

No it was very wrong and confusing

Defending initial answer

Because I still think I chose the right things to do

Highlighting incoherence / Questioning answer quality

No, I feel ChatGPT's responses were highly questionable

Expressing disageement
because I did not agree

Dismissing alternative answer / Reducing AI credibility

No, it didn't, because I know the answers come from an AI.

Defending initial answer

Because in a short time, I'm sure about my answers

Dismissing alternative answer
No.

Defending initial answer
I did not want to change my answers

Highlighting incoherence / Questioning answer quality

No, most answers were incoherent or nonsensical.

Rejecting influence of alternative opinions

No, not really. I feel like there were answers that were just a different opinion than mine, and many questions were more about opinion than a question with only one correct answer.

Expressing doubt about own decision

Expressing uncertainty

Yes cause I wasn't sure at all

Acknowledging personal imperfection in judgement

Yes, because I'm not always right

Being uncertain about changing answer
Maybe just one, maybe not.

GROUP B: Condition = Human

Considering alternative perspective

Noting answer quality/ Realising overlooked perspective

It seems to me that her/his answer was logical, and it would be better if more people were saved.

Valuing other perspectives/ Noting answer quality

Yes cause I could undrestand them

Revising answer / Considering alternative perspective

I would add nuance to my answer to take the conflicting situation into account.

Valuing other perspective

Not much but was good to see other perspectives

Valuing other perspectives

Yes, seeing different perspectives always helps for further choices

Valuing other perspectives / Realising overlooked perspective

Yes, seeing different answers can affect the way I approach problems in the future, as it can provide new insights and methods to consider.

Acknowledging validity of different reasoning

Not particularly. I think both answers are valid in their reasoning

Revising answer
A few, especially in the workplace ethical dilemma

Revisng answer/ Realising overlooked perspective

For the first one, it's not your job to help out your friend and telling them what to do, but a real friend would intervene instead of going straight to HR.
The second one, I'm comparing it to HIPPA, or if you have signed a non disclosure agreement. At first, I thought it wouldn't harm that much sharing some basic informations, but your basic informations might not the basic for your opponent
The third one, yes hiring a friend is to help them and you trust they are capable. If the situation is a better candidate and your friend, and the other candidate is really better. Then I wouldn't have a choice, but to hire the other one even if I want to help my friend

Expressing difficulty

Aknowledging difficulty

Very hard scenario

Aknowledging difficulty

They were so difficult

Dismissing alternative perspective / defending initial answer

Defending initial answer

I think I decided the best choice based on thee information

Aknowledging possible value of other perspective
Yes and no. It twice before acting and chose wisely your decision. It might have a bigger after effects and consequences

Expressing doubt about own decision

Expressing doubt

I'm still not sure with my choices about killing/ saving people. It will always be a really hard choice.

Rejecting influence of alternative opinion

Not much

Figure 10: Affinity diagram of open-ended questions