CrowdKing: Bridging the Gap in Value Elicitation for LLM Value Alignment

Petar Petrov
Delft University of Technology
The Netherlands
P.I.Petrov@student.tudelft.nl

Bohong Lu
Delft University of Technology
The Netherlands
B.Lu-2@student.tudelft.nl

Sebastiaan Beekman
Delft University of Technology
The Netherlands
j.s.beekman@student.tudelft.nl

ABSTRACT

Value alignment for Large Language Models (LLMs) is a critical challenge in ensuring these systems reflect nuanced human values and make context-aware decisions. Current approaches, such as Reinforcement Learning with Human Feedback (RLHF), rely heavily on simplified preference-based models that fail to capture the complexity of human ethical reasoning. This study investigates alternative value elicitation techniques through crowd computing, exploring the effects of storytelling and rationale writing in shaping participants' fairness judgements. Our initial findings reveal that storytelling enhances empathy and context awareness while rationale writing fosters reflective decision-making. The combination of these approaches elicited the most balanced and context-sensitive decisions. While preliminary, these insights suggest a step toward more effective value elicitation methods that could inform future AI alignment strategies.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**; • **Computing methodologies** → *Philosophical/theoretical foundations of artificial intelligence*; • **Information systems → Crowdsourcing**.

KEYWORDS

AI alignment, Value elicitation, Crowdsourcing, Collaborative decision-making, Fairness reasoning, Storytelling in AI, Rationales, Ethical AI

ACM Reference Format:
Petar Petrov, Bohong Lu, and Sebastiaan Beekman. 2025. CrowdKing: Bridging the Gap in Value Elicitation for LLM Value Alignment. In . ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Value alignment for Large Language Models (LLMs) has recently become a topic of interest due to their wide adoption and deployment in various fields. Ensuring that LLMs perform in line with what humans perceive as fair, moral and, correct is vital to building confidence in these systems and by avoiding discrimination, toxicity, and stereotyping. By its nature, value alignment hinges on our ability to capture, represent, and encode values in our AI systems.

LLMs have become a vital part of life. As such, it is important to ensure their responses are correct and respectful of people's morals. At their core, LLMs are giant probability calculators that try to find the response most likely to follow a given prompt. However, this can lead to insensitive responses because the system only considers the probability distribution of the words it has seen in its training data. For example, a language model can coax a depressed individual into hurting themselves by reinforcing the individual's negative beliefs because that is what the model believes is the most probable response. Another example could be that LLM systems used in policymaking can make judgments that do not appear fair or do not consider the full context of a situation, causing issues for policymakers and possibly hurting people's lives.

State-of-the-art approaches to capturing values commonly use Reinforcement Learning with Human Feedback (RLHF), which assumes preferences as adequate representations of human values. These traditional preferentist approaches to value elicitation have recently come under criticism due to several limiting factors [14]. For example, due to how modern AI models function, training with preferences can cause a model to optimize whatever function is grading against the presented preferences, hurting overall performance while seemingly producing more preferential responses. Another issue can arise from people's preferences. Preferences (or people) are imperfect and can go against what is best for an individual or a group. This creates a *research/development gap* where current preferentist approaches fall short in capturing the complexity of human values. Relying solely on these methods risks reinforcing ethically questionable behaviour in AI systems.

One way of creating data for value elicitation is through Crowd Computing. Crowd Computing allows researchers and companies to acquire high-quality, human-annotated data. Typically, this is done by generating responses using an LLM and then asking workers to rank responses grouped by prompt based on which one they prefer. However, by leveraging common Crowd Computing techniques, we can extract more detailed answers from workers and embed more information into value judgments. Furthermore, by tweaking the task framing, for example, if we use a fictional scenario, we can control the effects of cultural biases on worker responses. Human values are tacit, making them difficult to become aware of and express. Using more detailed and complex tasks, we could extract more expressive judgments from workers. However, that is at odds with common crowd-computing wisdom, as it is generally best to keep the cognitive and task effort of a task low. As such, we explore

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, July 2017, Washington, DC, USA
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

different ways of capturing values in an attempt to find a sweet spot between complexity and worker effort.

Given this remaining gap in LLM alignment research, we formulate our **design goal (DG): Explore alternative value elicitation techniques through crowd computing that allows for richer capturing and representation of human values, keeping in mind challenges of scalability, value awareness, and articulation.**

By researching this question, we can provide more context on value elicitation. Furthermore, capturing more complex human values can allow us to build modern LLM-based systems with better and more human-aligned responses. We aim to answer this question by conducting an in-group qualitative study on the online Crowdwork platform Prolific. We designed three different approaches for capturing values and evaluated their effects on value representation against a baseline inspired by literature. Initially, for the baseline, we pose a simple, non-storytelling question with two options for answer. We then present the scenario from a fictional point of view while preserving the answer structure. Finally, we ask workers to answer a non-storytelling and a storytelling question while providing a rationale for their choice.

This study makes several contributions. First, we designed and implemented a task framework to evaluate **four distinct approaches** to value elicitation, providing a potential basis for understanding their effectiveness in capturing **nuanced human values**. Second, we conducted an exploratory evaluation of the role of **fairness scenarios, storytelling elements, and rationale-based prompts** in shaping participant decision-making, offering preliminary insights into how these methods might influence ethical reasoning in real-world-inspired scenarios. Finally, this work insinuates the potential of **narrative techniques and reflective reasoning** to elicit more **context-aware, balanced decisions**, which can inform AI alignment strategies.

2 RELATED WORK

2.1 LLM value alignment

AI value alignment refers to the process of ensuring that AI-based systems act in ways consistent with human values. This challenge has gained attention as large language models and other AI systems increasingly influence decision-making in sensitive and value-laden domains. These approaches hinge on the ability to capture, represent and encode human values in scalable ways.

Crowd computing offers unique advantages for tackling these challenges. The power of collective reasoning and deliberation can more easily be harnessed. For example, Aitamurto and Landemore [2] demonstrate how crowd-sourced deliberation can lead to nuanced reasoning and equitable outcomes in policy-making. This study shows that crowd computing is not only a scalable solution but also one that enables rich, context-aware contributions.

Most dominant efforts focus on eliciting human values through crowd-worker preferences, Reinforcement Learning from Human Feedback (RLHF) [9, 12, 13]. Here, a base LLM generates a couple of responses given a certain prompt. Crowd-workers are then tasked with ranking those responses from least preferable to most preferable. These *preferences* are then used to train a reward model, which in turn is used to optimize a policy using reinforcement learning.

2.2 Critique on current Alignment Approaches

Value elicitation is a critical process in aligning AI systems with human values, as it involves identifying and understanding the principles that guide ethical decision-making. Even outside of the computing field, value elicitation poses many challenges due to the inherent tacitness of human values. Here, values have been found to be 1) difficult to become fully aware of and 2) challenging to be expressed [8]. AI value alignment needs to address these challenges within 3) the boundaries of scalability.

While crowd-computing seems a natural choice of strategy here, given its aim to understand and leverage diverse human perspectives, recent research has thus found current approaches to be insufficient in capturing human values, as they overly rely on simplified preference models. Methods such as **RLHF**, for instance, often involve optimizing AI systems based on **binary preference comparisons**, i.e., asking whether response A or response B aligns better with a user's intent. Although effective in guiding systems toward utilitarian objectives, this approach inherently simplifies complex ethical dilemmas and overlooks the **context-dependence and richness of human values** [3, 14]. This, in turn, may lead to negative consequences such as the deployment of AI systems that misinterpret or oversimplify user needs, creating biases in decision-making processes, diminishing trust among users, and perpetuating social inequalities. Similarly, Gabriel and Ghazavi (2021) [6] highlight the complexity of aligning AI with dynamic, sometimes conflicting, human values, emphasizing the need for nuanced elicitation methods. Finally, Kirk et al. (2024) [10] provide further evidence by demonstrating that the value preference framework fails to account for subjective and multicultural dimensions of value alignment. Their work with the PRISM dataset reveals that human feedback often reflects deeply personal and culturally specific perspectives, which cannot be adequately captured by simplistic, ranked value preferences.

3 METHODOLOGY

In the following, we will describe our exploratory project that investigates alternative value elicitation techniques through crowd computing that allows for richer capturing and representation of human values, keeping in mind challenges of scalability, value awareness, and articulation (DG).

3.1 Design Challenges

The system design aims to address the limitations of traditional value elicitation, namely, the simplification and reduction of human values to mere preference statements of "goodness" while keeping in mind the challenges of 1) the difficulty of value awareness, 2) the challenge of value articulation, and 3) the scalability of those tasks (keeping time and cognitive efforts low for crowd workers).

3.2 System Design

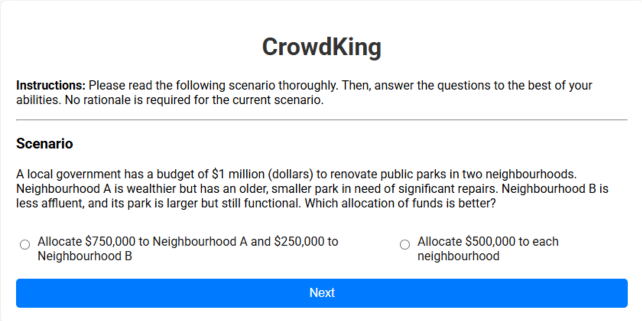
Prototyping Process. Our final design emerged from a process of prototyping and piloting.

The *initial phase* of prototyping began with abstract moral dilemmas inspired by existing frameworks, such as the MIT Moral Machine [4]. These early tasks highlighted the challenges of eliciting meaningful insights without sufficient contextual detail. To address this issue, we introduced *storytelling* as a central element, framing participants as decision-makers in scenarios such as sovereigns, community leaders, or local officials. By grounding the task in structured yet relatable narratives, the storytelling approach enabled participants to engage deeply with fairness considerations, addressing the challenge of value awareness and moving beyond simplistic preference-based annotations.

The scenarios were iteratively refined to enhance their effectiveness while balancing the cognitive demands on participants. Early versions lacked specificity, leading to unclear or shallow responses. To improve their effectiveness, we added concrete details about the needs and consequences faced by different groups within each scenario. For example, grain distribution dilemmas were revised to include precise ratios requested by stakeholders, potential long-term impacts on food security, and normative guidelines such as laws dictating fair distributions, to improve the *specificity in context*. Similarly, *framing variations* were introduced, presenting scenarios with alternate emphases, such as urgency ("a pressing crisis") or moral responsibility ("a sovereign's duty"), to study their effects on fairness judgments. Additionally, the *decision-making variables*, such as resource percentages or budget constraints, were dynamically adjusted to provide a diverse range of decision contexts for analysis.

Experimental Conditions. The finalized design comprises four experimental conditions, each progressively deepening the exploration of fairness through context-rich scenarios and rationale elicitation. Drawing on the use of storytelling scenarios, as highlighted in educational research [5, 7], this design facilitates participants' engagement with complex ethical dilemmas in a safe and structured environment. The inclusion of rationale elicitation further reflects its critical role in crowd computing tasks, as understanding the reasoning behind judgments not only enhances the quality of insights gathered but also underscores the nuances of human values [11].

The following sections detail the four experimental conditions, illustrating the progression from baseline fairness judgments to integrated approaches that capture both contextual influences and participant rationales.



CrowdKing

Instructions: Please read the following scenario thoroughly. Then, answer the questions to the best of your abilities. No rationale is required for the current scenario.

Scenario

A local government has a budget of $1 million (dollars) to renovate public parks in two neighbourhoods. Neighbourhood A is wealthier but has an older, smaller park in need of significant repairs. Neighbourhood B is less affluent, and its park is larger but still functional. Which allocation of funds is better?

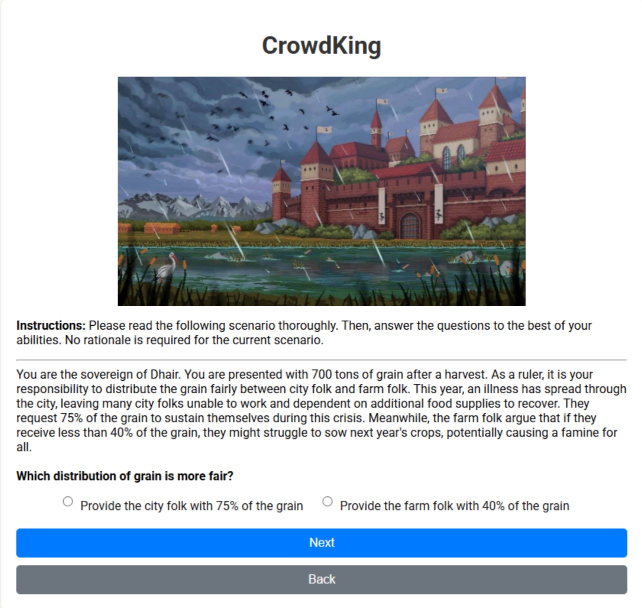
○ Allocate $750,000 to Neighbourhood A and $250,000 to Neighbourhood B

○ Allocate $500,000 to each neighbourhood

Figure 1: Screenshot from the live study showcasing one of the baseline fairness dilemmas.

Baseline - Condition 1. The baseline condition presents participants with straightforward fairness dilemmas devoid of contextual details. These tasks, such as dividing food among two groups or allocating renovation budgets between neighbourhoods, serve as a control metric to highlight the limitations of generic preference-based approaches. This condition establishes a reference point for evaluating the richness of context-based responses by isolating fairness from external factors.

You can see an example of a dilemma from our system in Figure 1. You can view the example in higher quality in Appendix D.



CrowdKing

Instructions: Please read the following scenario thoroughly. Then, answer the questions to the best of your abilities. No rationale is required for the current scenario.

You are the sovereign of Dhair. You are presented with 700 tons of grain after a harvest. As a ruler, it is your responsibility to distribute the grain fairly between city folk and farm folk. This year, an illness has spread through the city, leaving many city folks unable to work and dependent on additional food supplies to recover. They request 75% of the grain to sustain themselves during this crisis. Meanwhile, the farm folk argue that if they receive less than 40% of the grain, they might struggle to sow next year's crops, potentially causing a famine for all.

Which distribution of grain is more fair?

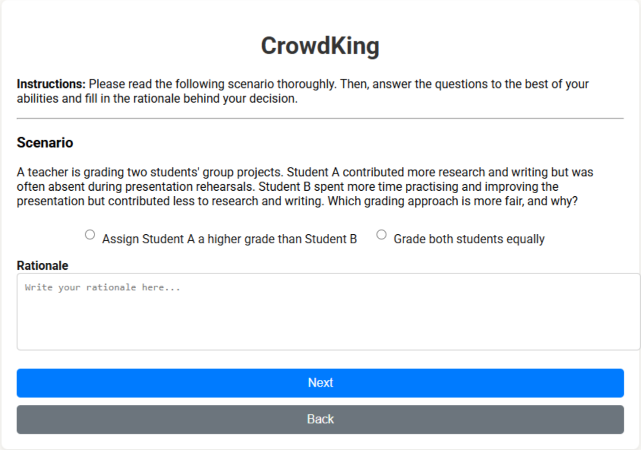
○ Provide the city folk with 75% of the grain ○ Provide the farm folk with 40% of the grain

Figure 2: Screenshot from the live study showcasing one of the Condition 2 fairness dilemmas.

Storytelling - Condition 2. The second condition asks participants to engage in storytelling scenarios that provide detailed narratives, allowing us to explore how situational factors influence fairness judgments. For example, participants assume the role of a sovereign that distributes grain during a harvest crisis, where city folk and farm folk present conflicting demands. Other scenarios include funding requests for new wells or responding to neighbouring kingdoms' resource needs. By embedding fairness dilemmas within immersive contexts, this condition reveals the complexities and trade-offs inherent in fairness judgments, showcasing the departure from simplistic preference elicitation.

You can see an example from our system in Figure 2. You can view the example in higher quality in Appendix D

Rationales - Condition 3. The third condition requires participants to make fairness decisions and justify their choices, emphasizing the reasoning process behind judgments, and offering insights into how participants interpret fairness. The scenarios are assigning grades to students with differing contributions, distributing prize money between athletes, or allocating vaccines between high-risk elderly patients and essential workers. The rationale elicitation process captures the depth and complexity of participant reasoning, bridging the gap between decisions and their underlying justifications; it emphasizes the reasoning processes behind judgments,



CrowdKing

Instructions: Please read the following scenario thoroughly. Then, answer the questions to the best of your abilities and fill in the rationale behind your decision.

Scenario

A teacher is grading two students' group projects. Student A contributed more research and writing but was often absent during presentation rehearsals. Student B spent more time practising and improving the presentation but contributed less to research and writing. Which grading approach is more fair, and why?

○ Assign Student A a higher grade than Student B ○ Grade both students equally

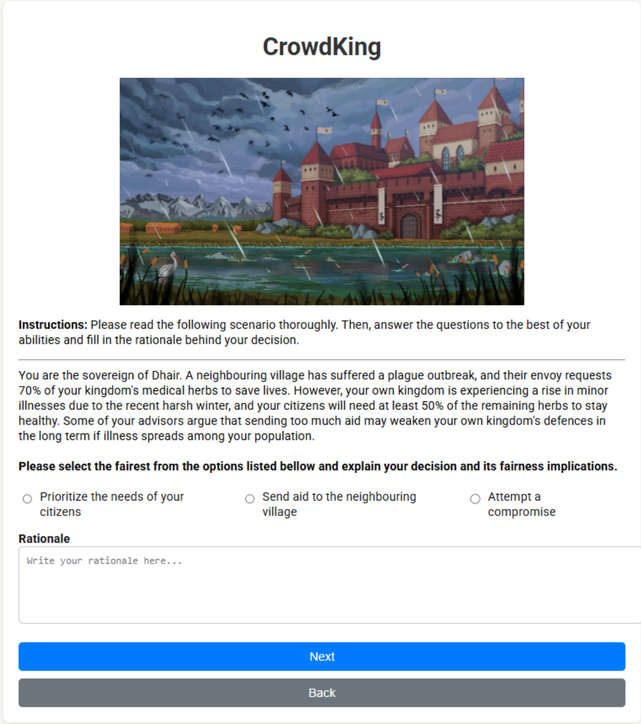
Rationale

Write your rationale here...

Figure 3: Screenshot from the live study showcasing one of the Condition 3 fairness dilemmas.

offering insights into how participants interpret fairness in various scenarios.

You can see an example from our system in Figure 3. You can view the example in higher quality in Appendix D



CrowdKing

Instructions: Please read the following scenario thoroughly. Then, answer the questions to the best of your abilities and fill in the rationale behind your decision.

You are the sovereign of Dhair. A neighbouring village has suffered a plague outbreak, and their envoy requests 70% of your kingdom's medical herbs to save lives. However, your own kingdom is experiencing a rise in minor illnesses due to the recent harsh winter, and your citizens will need at least 50% of the remaining herbs to stay healthy. Some of your advisors argue that sending too much aid may weaken your own kingdom's defences in the long term if illness spreads among your population.

Please select the fairest from the options listed bellow and explain your decision and its fairness implications.

○ Prioritize the needs of your citizens ○ Send aid to the neighbouring village ○ Attempt a compromise

Rationale

Write your rationale here...

Figure 4: Screenshot from the live study showcasing one of the Condition 4 fairness dilemmas.

Combined - Condition 4. The final condition integrates context-rich scenarios with rationale elicitation, allowing for a comprehensive exploration of fairness. In this condition, participants engage with complex dilemmas, which are distributing medical herbs during a plague, allocating water between settlements reliant on a shared oasis, or prioritizing disaster recovery efforts for towns with differing economic and social significance. The combined approach enables the examination of fairness judgments under intricate, multifaceted conditions, capturing the interplay of context and rationale.

You can see an example from our system in Figure 4. You can view the example in higher quality in Appendix D

3.3 Implementation Details

All of the code used for this experiment can be found at [1].

Database Design. The database stores the participants' responses across multiple stages of the experiment, with tables tracking each participant's progress and responses. In particular, the database includes a `Users` table that holds participant progress, identifiers, and session details. We also take advantage of individual tables for each experiment stage, which track the questions given to workers and their responses. Finally, we use an `Evaluations` table that stores feedback from participants.

API Design. The API, built using `Flask`, serves as the primary interface between the participants and the back-end system. Key API routes include a `Consent Route` to collect user consent before starting the experiment, a `Tutorial Route` to ensure participants understand the task instructions and multiple routes for serving different stages of the experiment. Each stage involves presenting the participant with questions, recording their answers, and updating their progress in the database. Additionally, we included `Evaluation Routes` to collect feedback after each stage, with all responses stored for analysis. Error handling is a critical part of our system, especially after we experienced minor technical difficulties which necessitated a more comprehensive error managing setup.

Quality Control. Quality control is essential in crowd-sourced experiments to uphold the integrity of the collected data. A key method deployed is the use of **attention checks**. During the experiment, participants are required to fill in a form (in addition to the default questions) with a predefined response (e.g., "Neutral") to verify their attentiveness. Failure to provide the correct response identifies potential issues, allowing for the exclusion of those who appear distracted, disengaged, or provide random answers. Additionally, the rationales participants must provide for conditions 2 and 3 act as a 'passive attention check', ensuring the rationale aligns both independently and with the given answer. Participants who fail to meet the quality standards are redirected to a Prolific completion page displaying the appropriate failure code. The overall **approval rate** for this study was 8/9, reflecting the effectiveness of the quality control measures implemented.

3.4 Prolific Platform Setup

The study was conducted using Prolific, a crowd-sourcing platform that allows efficient participant recruitment and quality control. Participants were provided with a unique *completion code* upon finishing the study and passing the attention check, which they entered on Prolific to confirm their participation. Those who failed the attention check received a different code and were not compensated.

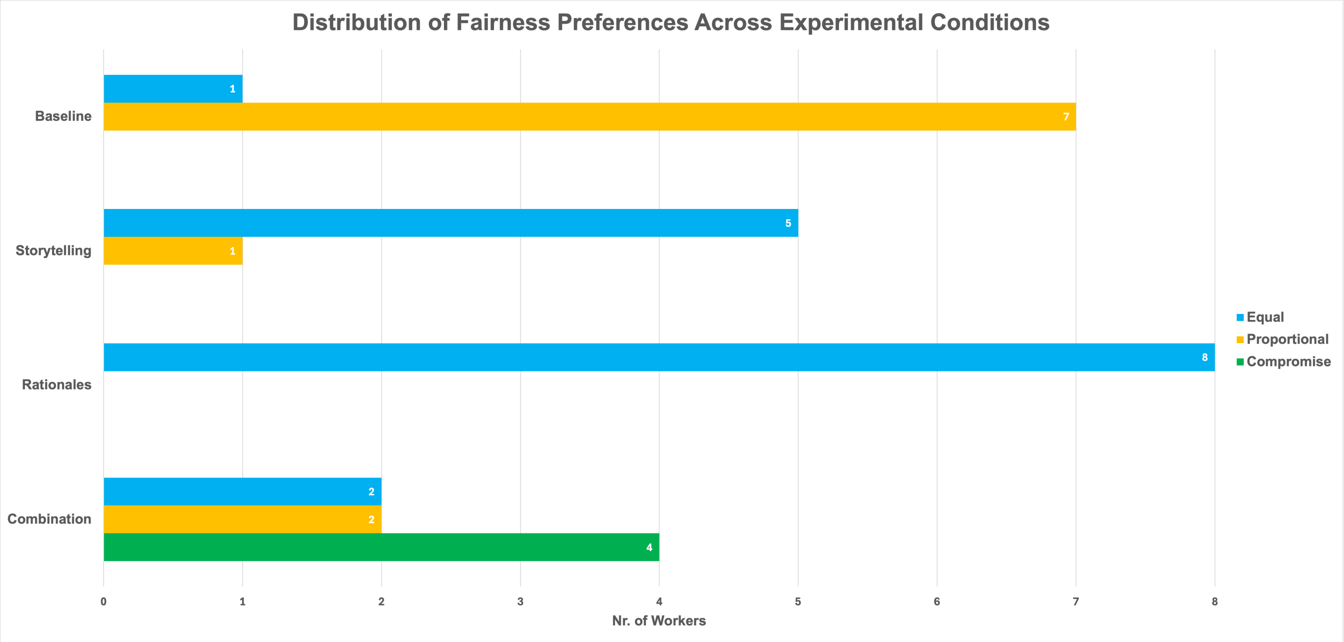


Figure 5: The figure illustrates the number of participants (total of 8) choosing each fairness option (Equal, Proportional, and Compromise) across four experimental conditions: the baseline, storytelling, rationale, and combination experiments. It should be noted that due to an error in the code base, two 'storytelling' responses were deemed invalid.

We used manual submission approval to verify that there were no issues with responses before paying workers. Recruitment targeted a diverse and *representative sample*, with participants required to be fluent in English and excluding those who had previously participated in similar studies. This ensured a high-quality participant pool and reliable data.

Study Procedure. Prolific sent out links to participants, allowing them to start the study. We set up the links to hold the worker's participant ID and the relevant study and session IDs. Initially, we estimated it would take five minutes for workers to complete the study, and we conducted a trial run with ten workers. During the trial run, we experienced technical difficulties due to a lack of proper error handling, where our study could not save responses to the database without error. We pinpointed the issue as a problem with the server setup, which we resolved by changing hosting providers. Ultimately, only one of the ten trial participants finished the study, while the rest could not due to technical difficulties. The trial round showed us that the mean time to complete the study is ten to fifteen minutes rather than five. We adjusted the setup on Prolific and ran a second run with additional error handling, better hosting, and only five participants. The second round had a median time of ten minutes, with all participants successfully completing the study. We collected extra data points by running a third batch with only three workers and a mean time of fifteen minutes.

Evaluation. To evaluate our system, we set up individual questionnaires for each stage of the study that workers completed after finishing the respective step. The questionnaires captured information about the worker's experience and their opinions on the conditions and question(s) presented. We modelled our evaluation after a Qualitative exploratory comparison. So, after we collected the data, we first identified change patterns (i.e., how did rationale alter a worker's response to a given problem). We then investigated what caused any detected changes by examining the rationales and survey questions.

4 RESULTS AND DISCUSSION

This qualitative exploratory study investigated richer value elicitation methods through crowd computing. Drawing inspiration from the Research through Design (RtD) approach, our findings reveal several preliminary insights into participant behaviour and design challenges, with implications for future AI alignment research.

4.1 Conditions

4.1.1 ***Baseline.*** In the baseline experiment, where participants evaluated scenarios purely based on the 'betterness' without added contextualization, the majority opted for a **Proportional** distribution (Figure 5). Proportional refers to the principle of "you get what you work for" or "Larger groups get more resources". This trend suggests that participants leaned toward a utilitarian approach, prioritizing effort-based rewards over equal distribution.

The absence of narrative context or a requirement to justify decisions likely led participants to default to this straightforward fairness heuristic. Post-scenario evaluations corroborated this, with most participants reporting that they found it "easy" to decide, often justifying their choices with statements like, *"more individuals, more resources."* This simplicity in reasoning suggests that proportional fairness emerges as a dominant, easily applied framework in the absence of context.

4.1.2 ***Storytelling - Condition 1.*** When participants were provided scenarios that incorporated storytelling elements, the preference for **Proportional** fairness dropped by more than 50%, while **Equal** choices rose noticeably, from one participant to five. This shift may indicate that the inclusion of narrative context enables participants to empathize with affected groups, prompting them to weigh fairness more holistically, and potentially override purely effort-based reasoning.

Storytelling appeared to facilitate a clearer understanding of the scenario's stakes and a stronger connection with the individuals involved, helping participants recall details and construct nuanced perspectives. Evaluation data supported this, with participants describing the narratives as "more engaging," "better for providing perspective," and "easier to reason and express an opinion" compared to the baseline.

4.1.3 ***Rationales - Condition 2.*** The addition of rationale writing significantly altered fairness preferences, with a notable rise in **Equal** choices. By requiring participants to articulate their reasoning, this condition appeared to encourage deeper reflection, helping participants balance proportional fairness with other ethical dimensions.

Participants often described this process as "more engaging" and "better for formulating and expressing opinions," compared to the baseline. Notably, rationales highlighted the complexity of fairness considerations, with some participants explicitly acknowledging the limitations of proportional fairness. One stated, *"This one, in particular, made me not take my feelings part in my decision. That is because there is even a bigger picture..."*

The collected rationales showcased diverse ethical reasoning:

- *"They complemented each other, so they deserve the same grading."*
- *"If they really are a team, they would find no problem in sharing the same amount of prize money."*
- *"About 50/50 since they both put up the effort and work."*

These examples illustrate how rationales prompted participants to go beyond surface-level fairness judgments.

4.1.4 ***Combined - Condition 3.*** The combination of storytelling and rationale writing resulted in the most diverse distribution of fairness decisions, with increases in both **Equal** and **Compromise** choices compared to the Baseline. This could suggest that engaging participants' empathy through storytelling, alongside encouraging reflective reasoning through rationale writing, broadened their fairness perspective.

Participants in this condition frequently incorporated multiple dimensions of fairness, weighing proportionally against equality while considering broader societal implications. Evaluations indicated this approach was "better for expressiveness" and "more engaging" than the baseline.

Notably, rationales in the **Comparison** category reflected an attempt to balance competing values:

- *"I would try to negotiate a 50-50% split with them."*
- *"My citizens will need 50%, so 50% is the only amount I can spare."*
- *"I could only possibly be able to share about 50% of my kingdom's herbs."*

Meanwhile, **Equal** rationales emphasized levelling the playing field:

- *"I would help the poorer city; the decision is fair because I am leveling the playing field."*
- *"We need to rebuild Riverstone (small city)! Let the merchants and the rich rebuild Crestport (big city), but we need to help our rural mates!"*

This condition highlights the potential of combining storytelling and rationale writing to elicit nuanced, context-aware fairness reasoning. In the context of value elicitation for AI alignment, these results suggest how incorporating immersion-driven narratives and reflective processes through rationales could help uncover the multifaceted and dynamic nature of human values. By leveraging such methods, AI systems can be better aligned with complex ethical principles, moving beyond simplistic preferentist models towards a richer, more holistic understanding of fairness and decision-making.

4.2 Limitations

While this study provides valuable insights into the challenges of value elicitation and fairness decision-making in AI alignment contexts, several limitations must be acknowledged:

First, the relatively small sample size limits the generalizability of our findings. Although the experiments were designed to elicit diverse perspectives, the results may not fully capture the variability present in larger populations. For example, differences in the distribution of fairness preferences between the **Storytelling** and **Rationales** experiments, as compared to the **Combination** experiment, suggest some consistency. However, we cannot confidently conclude that these patterns will hold in a larger sample. Future work with broader participant bases could help validate these trends and offer a more representative understanding of human values.

Second, certain variations in participant responses across conditions remain difficult to explain conclusively. While we have identified potential factors, such as empathy eliciting through storytelling or reflective reasoning prompted by rationale writing, were identified, other influences cannot be ruled out. For instance, minor differences between scenarios may have unconsciously biased participants to specific options. Further research is necessary to disentangle these effects. Employing a more controlled experimental design or additional qualitative and quantitative data collection could provide greater clarity.

Lastly, this study heavily relies on participants' self-reported feedback, such as evaluation forms responses, which may introduce bias. Participants may struggle to accurately recall or articulate their reasoning and feelings, or they may provide socially desirable answers instead of truthful ones. Moreover, if the participants recall nuanced trains of thought but choose not to report them, or simply lack the motivation to do so, such omissions could make a thoughtful decision appear shallow. To address this, future work could refine evaluation methods, incorporate alternative techniques like behavioural analysis, or explore physiological measures to capture more robust insights into participants' decision-making processes.

5 CONCLUSION

This study explored the design and evaluation of richer value elicitation methods for crowd-computing tasks, offering potentially new insights into how human values can be more effectively captured for AI alignment. Current value alignment approaches, such as those leveraging Reinforcement Learning with Human Feedback (RLHF), often rely on simplified preference-based models that fail to represent the nuance and context-dependence of human ethical reasoning. By contrast, our work highlights the potential of integrating **storytelling** and **rationale-writing** intro elicitation tasks to better capture complex human values.

Our findings indicate that **narrative elements could foster empathy and context awareness**, encouraging participants to consider ethical dimensions beyond utilitarian reasoning. Additionally, **rationale writing seems to promote reflective decision-making**, enabling individuals to articulate deeper considerations, such as fairness trade-offs or societal implications. The combined use of storytelling and rationales led to the most diverse and balanced value representation, suggesting that such techniques may address limitations in traditional value-preference elicitation.

These results underscore the importance of moving beyond binary preference models in AI alignment research. Designing elicitation frameworks that account for the richness and subjectivity of human values can lead to the development of **more context-aware and ethically aligned AI systems**, capable of making decisions that resonate with diverse human perspectives.

REFERENCES

[1] Crowd Computing Group 2. 2025. Crowd Computing Final Project. https://github.com/petaripetrov/crowd-computing Accessed: 2025-01-24.

[2] Tanja Aitamurto and Hélène Landemore. 2016. Crowdsourced Deliberation: The Case of the Law on Off-Road Traffic in Finland. *Policy & Internet* 8, 2 (2016), 174–196. https://doi.org/10.1002/poi3.115 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/poi3.115

[3] Anne Arzberger, Stefan Buijsman, Maria Luce Lupetti, Alessandro Bozzon, and Jie Yang. 2024. Nothing comes without its world – Practical challenges of aligning LLMs to situated human values through RLHF. *Vol. 7 (2024): Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES-24)* 7 (10 2024), 61–73. https://doi.org/10.1609/aies.v7i1.31617

[4] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.

[5] Character Counts. 2023. The Six Pillars of Character. https://charactercounts.org/c5/ Accessed: 2025-01-24.

[6] Iason Gabriel and Vafa Ghazavi. 2021. The Challenge of Value Alignment: from Fairer Algorithms to AI Safety. *arXiv (Cornell University)* (1 2021). https://doi.org/10.48550/arxiv.2101.06060

[7] Jill Grose-Fifer and John Jay. 2017. *Using Role-Play to Enhance Critical Thinking about Ethics in Psychology.*

[8] Lauren N. Irwin. 2021. Dare to lead: Brave work. tough conversations. whole hearts. *Journal of Women and Gender in Higher Education* 14, 2 (5 2021), 240–243. https://doi.org/10.1080/26379112.2021.1948859

[9] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A Survey of Reinforcement Learning from Human Feedback. *arXiv (Cornell University)* (1 2023). https://doi.org/10.48550/arxiv.2312.14925

[10] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM Alignment Project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv (Cornell University)* (4 2024). https://doi.org/10.48550/arxiv.2404.16019

[11] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 4, 1 (Sep. 2016), 139–148. https://doi.org/10.1609/hcomp.v4i1.13287

[12] OpenAI. 2022. Introducing ChatGPT. https://openai.com/index/chatgpt/

[13] Wanqi Xue, Qingpeng Cai, Zhenghai Xue, Shuo Sun, Shuchang Liu, Dong Zheng, Peng Jiang, Kun Gai, and Bo An. 2023. PrefRec: Recommender Systems with Human Preferences for Reinforcing Long-term User Engagement. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (8 2023), 2874–2884. https://doi.org/10.1145/3580305.3599473

[14] Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. 2024. Beyond preferences in AI alignment. *arXiv (Cornell University)* (8 2024). https://doi.org/10.48550/arxiv.2408.16984

A SCENARIOS

A.1 Baseline

A.1.1 ***Scenario B1.*** A community leader is distributing 100 portions of food between two groups: Group A has 40 people, and Group B has 20 people. Which distribution is better?

Answer option: "70 portions to Group A and 30 to Group B" (Proportional) OR "50 portions to Group A and 50 to Group B" (Equal)

A.1.2 ***Scenario B2.*** A local government has a budget of $1 million (dollars) to renovate public parks in two neighbourhoods. Neighbourhood A is wealthier but has an older, smaller park in need of significant repairs. Neighbourhood B is less affluent, and its park is larger but still functional. Which allocation of funds is better?

Answer options: "Allocate $750,000 to Neighbourhood A and $250,000 to Neighbourhood B" (Proportional) OR "Allocate $500,000 to each neighbourhood" (Equal)

A.1.3 ***Scenario B3.*** A government allocates disaster relief funds to two towns. Town X has 1,000 residents, and Town Y has 500 residents. Which allocation is better?

Answer options: "Town X receives twice as much funding as Town Y" (Proportional) OR "Both towns receive the same amount of funding" (Equal)

A.2 Storytelling - Condition 1

A.2.1 ***Scenario S1.*** You are the sovereign of Dhair. You are presented with 700 tons of grain after a harvest. As a ruler, it is your responsibility to distribute the grain fairly between city folk and farm folk. This year, an illness has spread through the city, leaving many city folks unable to work and dependent on additional food supplies to recover. They request 75% of the grain to sustain themselves during this crisis. Meanwhile, the farm folk argue that if they receive less than 40% of the grain, they might struggle to sow next year's crops, potentially causing a famine for all. Which distribution of grain is more fair?

Answer options: "Provide the city folk with 75% of the grain" (Proportional) OR "Provide the farm folk with 40% of the grain" (Equal)

A.2.2 ***Scenario S2.*** You are the sovereign of Dhair. The builders of Dhair have just completed a new well in the upper city, providing the more privileged citizens living there with clean water. The residents of the upper city are pleased with the well because they no longer have to walk long distances to fetch water. When the citizens in the lower city hear about this, they approach you with a request: they also want a well. They tell you their well is running dry and hope you will approve the construction of a new well in the lower city. Some residents of the upper city argue that they paid higher taxes to fund their well and feel it would be unfair to provide the same service to the lower city without similar contributions. Which is the fairer decision?

Answer options: "Do not build a second well in the lower city" (Proportional) OR "Build a second well in the lower city (Equal)

A.2.3 ***Scenario S3.*** You are the sovereign of Dhair. A neighbouring kingdom, Nairon, has suffered a natural disaster and sends an envoy requesting 10 of your kingdom's horses. These horses are necessary for farming and building efforts. However, without these horses, your farmers will struggle to plough fields efficiently, risking reduced crop yields in the upcoming season and potential food shortages. Which decision is more fair?

Answer options: "Do not send the horses to Nairon" (Proportional) OR "Aid Nairon by sending 10 horses" (Equal)

A.3 Rationales - Condition 2

A.3.1 ***Scenario R1.*** A teacher is grading two students' group projects. Student A contributed more research and writing but was often absent during presentation rehearsals. Student B spent more time practising and improving the presentation but contributed less to research and writing. Which grading approach is more fair, and why?

Answer options: "Assign Student A a higher grade than Student B" (Proportional) OR "Grade both students equally" (Equal)

A.3.2 ***Scenario R2.*** Two athletes are awarded prize money for a relay race. Athlete A ran for 60% of the race, and Athlete B ran for 40% of the race. Which prize distribution is fairer, and why?

Answer options: "Athlete A receives 60% of the prize, and Athlete B receives 40%" (Proportional) OR "Athlete A and Athlete B split the prize equally" (Equal)

A.3.3 ***Scenario R3.*** A community health centre receives 100 doses of a rare vaccine. Group A consists of elderly patients at higher risk of complications but who rarely leave their homes. Group B consists of essential workers at lower personal risk but who interact with the public daily. Which vaccine distribution is more fair, and why?

Answer options: "Prioritize Group A by giving them 80 doses and Group B 20 doses" (Proportional) OR "Split the doses evenly between Group A and Group B" (Equal)

A.4 Combination - Condition 3

A.4.1 ***Scenario C1.*** You are the sovereign of Dhair. A neighbouring village has suffered a plague outbreak, and their envoy requests 70% of your kingdom's medical herbs to save lives. However, your kingdom is experiencing a rise in minor illnesses due to the recent harsh winter, and your citizens will need at least 50% of the remaining herbs to stay healthy. Some of your advisors argue that sending too much aid may weaken your own kingdom's defences in the long term if illness spreads among your population. Please select the fairest from the options listed below and explain your decision and its fairness implications.

Answer options: "Prioritize the needs of your citizens" (Proportional) OR "Send aid to the neighbouring village" (Equal) OR "Attempt a compromise" (Compromise)

A.4.2 ***Scenario C2.*** You are the sovereign of the desert city of Althar. Two settlements, Mirath and Drokar, rely on a shared oasis for water, which is running dangerously low due to a drought. Mirath is larger, with more people, but also has more resources to drill wells or import water if necessary. Drokar is smaller and poorer, with no alternative water sources. Mirath argues that it should receive 70% of the remaining water due to its size and importance to regional trade. Drokar pleads for at least 50% to ensure its survival.

Please select the fairest from the options listed below. Explain your decision and its fairness implications.

Answer options: "Assist Mirath" (Proportional) OR "Assist Drokar" (Equal) OR "Attempt a compromise" (Compromise)

A.4.3 ***Scenario C3.*** You are the sovereign of the coastal kingdom of Aerin. A recent flood destroyed parts of two towns: Crestport and Riverstone. Crestport is a major port town vital to your kingdom's economy and trade but already has resources to begin repairs. Riverstone is smaller, more rural, and lacks the infrastructure to rebuild on its own. Your treasury has enough funds to fully rebuild one town or partially assist both. Advisors warn that prioritizing Riverstone might weaken trade routes while focusing on Crestport could cause unrest in the countryside. Please select the fairest from the options listed below and explain your decision and its fairness implications.

Answer options: "Assist Crestport" (Proportional) OR "Assist Riverstone" (Equal) OR "Attempt a compromise" (Compromise)

B PARTICIPANT RESPONSES

B.1 Baseline

Participant ID	P01	P02	P03	P04	P05	P06	P07	P08
Scenario	2	1	1	3	1	3	3	3
Response	Proportional	Proportional	Proportional	Proportional	Proportional	Proportional	Equal	Proportional

Table 1: Participant responses to the baseline condition, presenting choices and justifications for simple, non-storytelling scenarios.

B.2 Storytelling - Condition 1

Participant ID	P01	P02	P03	P04	P05	P06	P07	P08
Scenario	1	1	1	1	1	1	1	3
Response	Equal	**LOST**	Equal	**LOST**	Equal	Equal	Equal	Proportional

Table 2: Participant responses to storytelling scenarios, designed to explore the impact of narrative framing on value elicitation.

B.3 Rationales - Condition 2

Participant ID	Scenario	Response	Rationale
P01	1	Equal	I think both groups out in approximately equal effort into the work
P02	3	Equal	i cannot risk not having supplies for the first responders and i cannot risk not having a vaccine for those who are most vulnerable.
P03	1	Equal	Both have successfully contributed to the project in one area and lacked effort in another. therefore, both of their efforts should be rewarded equally.
P04	3	Equal	Theory is to give more to those in contact but ones that stay at home will have better chance to avoid for the meanwhile
P05	3	Equal	I'd given more doses to Group B, but that is not an option. Let the elderly stay at home, and let the dosed individuals tend to their needs.
P06	2	Equal	If they really are a team they would find no problem in sharing the same ammount of prize money. Furthermore, even if A ran more meters of the race, B had to balance the lack of distance with more raw speed
P07	1	Equal	They complemented each other, so they deserve the same grading. It is not like one did more than the other.
P08	1	Equal	About 50/50 since they both put up the effort and work although they may both have their faults.

Table 3: Participant responses in the rationale condition, highlighting how reflective reasoning shapes value judgments.

B.4 Combination - Condition 3

Participant ID	Scenario	Response	Rationale
P01	1	Compromise	I would try to negotiate a 50-50% split with them.
P02	2	Equal	i would help the poorer city, descition is fair because i am levelling the playing field by giving resource to the city that is most in need.
P03	1	Compromise	My citizens will need 50% of the remaining herbs to stay healthy, so 50% is the only amount I can spare for the neighbouring village, as my obligations are first with my kingdom.
P04	3	Compromise	compromise as need to reduce the outbreak and it reaching our village
P05	2	Equal	We need to rebuild Riverstone! Let the merchants and the rich rebuild Cerstport, but we need to help our rural mates!
P06	1	Proportional	I think if my citizens really need the herbs i should prioritize them, not only that but if the neighbor village outrages because of that decision we could invade them with our healthy troops to silence them
P07	3	Proportional	Riverstone won't be able to repair itself with time. However, Riverstone will.
P08	1	Compromise	I could only possibly be able to share about 50% of my kingdoms herbs. That way I have tried just about enough to meet them halfway while not compromising my kingdoms health defense.

Table 4: Participant responses in the rationale condition, highlighting how reflective reasoning shapes value judgments.

C PARTICIPANT EVALUATION SURVEYS

C.1 Post Condition Survey

Question	Response		
	Baseline	**Storytelling**	**Rationale**
How clear were the options provided for this section?	Very clear	Very clear	Very clear
How well were you able to express your opinion in this section?	Well	Very well	Very well
How easy or difficult was it to make decisions in this section?	Easy	Neatral	Easy
How engaging did you find this section overall?	Engaging	Engaging	Engaging
Do you have any additional feedback about your experience with this experiment?	I think the 600-400 split would have been ideal.	I would feel sorry for them, but unfortunately, I wouldn't risk our own livelihood.	I think both groups put in approximately equal effort into the work.

Table 5: Survey responses collected from P01 after completing each condition, assessing their experiences and perceptions of the tasks.

Question	Response		
	Baseline	**Storytelling**	**Rationale**
How clear were the options provided for this section?	Clear	Clear	Clear
How well were you able to express your opinion in this section?	Well	Well	Well
How easy or difficult was it to make decisions in this section?	Easy	Easy	Neutral
How engaging did you find this section overall?	Engaging	Engaging	Engaging
Do you have any additional feedback about your experience with this experiment?	n/a	n/a	n/a

Table 6: Survey responses collected from P02 after completing each condition, assessing their experiences and perceptions of the tasks.

Question	Response		
	Baseline	**Storytelling**	**Rationale**
How clear were the options provided for this section?	Clear	Very clear	Very clear
How well were you able to express your opinion in this section?	Well	Very well	Very well
How easy or difficult was it to make decisions in this section?	Neutral	Very easy	Very easy
How engaging did you find this section overall?	Neutral	Engaging	Engaging
Do you have any additional feedback about your experience with this experiment?	Not currently, as I am unsure about what the study is about.	This to me is very straightforward. The amount of people that will suffer if the wish of the city folk is granted is way larger.	Not really, I think it's pretty straightforward.

Table 7: Survey responses collected from P03 after completing each condition, assessing their experiences and perceptions of the tasks.

Question	Response		
	Baseline	**Storytelling**	**Rationale**
How clear were the options provided for this section?	Clear	Clear	Clear
How well were you able to express your opinion in this section?	Neutral	Neutral	Well
How easy or difficult was it to make decisions in this section?	Difficult	Difficult	Difficult
How engaging did you find this section overall?	Neutral	Engaging	Engaging
Do you have any additional feedback about your experience with this experiment?	Nothing to add	was a difficult question to answer	Nothing

Table 8: Survey responses collected from P04 after completing each condition, assessing their experiences and perceptions of the tasks.

Question	Response		
	Baseline	**Storytelling**	**Rationale**
How clear were the options provided for this section?	Very clear	Very clear	Very clear
How well were you able to express your opinion in this section?	Very well	Very well	well
How easy or difficult was it to make decisions in this section?	Very easy	Easy	easy
How engaging did you find this section overall?	Neutral	Neutral	Boring
Do you have any additional feedback about your experience with this experiment?	No, pretty easy: double the resident, double the money.	Not that easy, but farmers >city-dwellers, because I'm from a little village too. Yeah, of course, "my" kingdom needs food the next year too.	N/A, I thought this is where I was supposed to rationalize

Table 9: Survey responses collected from P05 after completing each condition, assessing their experiences and perceptions of the tasks.

Question	Response			
	Baseline	**Storytelling**	**Rationale**	**Combination**
Did anything about the task surprise you or make you see things differently?	No	No	No	No
Can you describe your overall experience with the task?	I had to imagine the scenario and came to the conclusion i ended up choosing	I just had to visualize the situation and really think what would i do	I think doing the rasionale makes you really think why you chose your choice	I personally would prefered a more drastic question as the premise as it was did not make it difficult to choose an option
Were there any elements of the task design (e.g., instructions, activities, examples) that helped you feel more connected to the experience?	No	No	🟥	NO
Do you feel the task encouraged you to think about new perspectives or issues?	No	I think the fact that the premise gives you the perpective of both the city and farm folks really makes you think of what the best decision would be	🟥	NO
How well were you able express your perception of what can be consider fair here?	I didnt have much freedom in choice as there were only two options	🟥	Really well	Really well
Do you have any additional feedback about your experience with this experiment?	No	No	No	No

Table 10: Survey responses collected from P06 after completing each condition, assessing their experiences and perceptions of the tasks.

Question	Response			
	Baseline	**Storytelling**	**Rationale**	**Combination**
Did anything about the task surprise you or make you see things differently?	Even though it is the same amount of money, fewer people will profit more from it in one of the two towns. However, it is the same amount of money and should be allocated in equal parts.	This one, in particular, made my feelings not to take part in my decision. That is because there is even a bigger picture than the ill, that is that everyone might be ill in the future if I do not give the 40% of the grain to that other city.	No.	No.
Can you describe your overall experience with the task?	It made me think what would be the fairness option.	Made me more reasonable in that particular scenario.	It was the easier decision of this overall scenarios.	It was an easy one, because one can rebuild itself with time, the other won't.
Were there any elements of the task design (e.g., instructions, activities, examples) that helped you feel more connected to the experience?	Yes, that I was given the option to "decide"	Being the one that had to decide what's best.	🟥	Being the one to decide.
Do you feel the task encouraged you to think about new perspectives or issues?	No, i don't think so.	Yes. Made me see what's the bigger picture.		No.
How well were you able express your perception of what can be consider fair here?	Really well.	🟥	I feel like there is no other option here.	Really well.
Do you have any additional feedback about your experience with this experiment?	No, i do not.	No, i do not.	No, i do not.	No, i do not.

Table 11: Survey responses collected from P07 after completing each condition, assessing their experiences and perceptions of the tasks.

Question	Response			
	Baseline	Storytelling	Rationale	Combination
Did anything about the task surprise you or make you see things differently?	No, with more residents means that more funds should be allocated to those in more and dire need. While the 500 residents should be assisted as well, town X requires more assistance.	No, but it is worth taking note that my people are just in need as the neighboring kingdom	NO	No
Can you describe your overall experience with the task?	The towns people are all in dire need of these funds, but the other town needs more help than the other.	It was with ease to come up to a conclusion, not difficult at all.	The 2 students are both befitting of getting equal grades since they completed each others half. Obviously taking note of their fauls	Although the neighboring kingdom wanted about 70% of which I could not comply with that. It was only fair and with sympathy that I shared 50% Their sickness and cries of help in terms of their health could not have been avoided. It is also important to lend a helping hand wherever possible
Were there any elements of the task design (e.g., instructions, activities, examples) that helped you feel more connected to the experience?	The figures, 1000 and 500.	Had there been an option to give them about 5 horses I would have chosen it.	🟥	We all have our problems. Sometimes we do need that helping hand.
Do you feel the task encouraged you to think about new perspectives or issues?	Kind of, yes.	No		Help if YOU CAN!
How well were you able express your perception of what can be consider fair here?	People in need can come from various different backgrounds. The 1000 individuals will need more funds since they are many than the 500.	🟥	Its only fair that they get equal grades. None of them did a complete task, their efforts are both deemed worthy of an equal grade.	I could not risk my kingdoms health nor could I ignore cry for help
Do you have any additional feedback about your experience with this experiment?	No	Not at all.	No	NO

Table 12: Survey responses collected from P08 after completing each condition, assessing their experiences and perceptions of the tasks.

C.2 Post Task Survey

Question	Response
How well did the experiment allow you to express your personal values?	Well
Were the scenarios detailed enough to help you consider multiple aspects of the situation?	Agree
Did you feel your answers reflected your true preferences and reasoning?	Agree
How engaging did you find this experiment overall?	Engaging
What did you find most engaging about the study?	I found it interesting, I enjoy tasks like this.
Were there any aspects of the study that limited your ability to provide meaningful answers?	No
Do you have any additional comments or suggestions?	No

Table 13: Survey responses collected from participant P01 after the study, evaluating their overall engagement, task effort, and perceived relevance of the experiment.

Question	Response
How well did the experiment allow you to express your personal values?	Well
Were the scenarios detailed enough to help you consider multiple aspects of the situation?	Agree
Did you feel your answers reflected your true preferences and reasoning?	Agree
How engaging did you find this experiment overall?	Engaging
What did you find most engaging about the study?	The reason for my decision
Were there any aspects of the study that limited your ability to provide meaningful answers?	no
Do you have any additional comments or suggestions?	n/a

Table 14: Survey responses collected from participant P02 after the study, evaluating their overall engagement, task effort, and perceived relevance of the experiment.

Question	Response
How well did the experiment allow you to express your personal values?	Well
Were the scenarios detailed enough to help you consider multiple aspects of the situation?	Neutral
Did you feel your answers reflected your true preferences and reasoning?	Agree
How engaging did you find this experiment overall?	Engaging
What did you find most engaging about the study?	The writing parts as it better allowed me to express my rationale.
Were there any aspects of the study that limited your ability to provide meaningful answers?	When there was no writing section, it took away my ability to explain my thoughts, making my decisions feel a bit more rushed.
Do you have any additional comments or suggestions?	N/A

Table 15: Survey responses collected from participant P03 after the study, evaluating their overall engagement, task effort, and perceived relevance of the experiment.

Question	Response
How well did the experiment allow you to express your personal values?	Well
Were the scenarios detailed enough to help you consider multiple aspects of the situation?	Agree
Did you feel your answers reflected your true preferences and reasoning?	Agree
How engaging did you find this experiment overall?	Very Engaging
What did you find most engaging about the study?	the fact you have options and could also add reasons why
Were there any aspects of the study that limited your ability to provide meaningful answers?	Not really
Do you have any additional comments or suggestions?	nothing to add

Table 16: Survey responses collected from participant P04 after the study, evaluating their overall engagement, task effort, and perceived relevance of the experiment.

Question	Response
How well did the experiment allow you to express your personal values?	Well
Were the scenarios detailed enough to help you consider multiple aspects of the situation?	Agree
Did you feel your answers reflected your true preferences and reasoning?	Agree
How engaging did you find this experiment overall?	Engaging
What did you find most engaging about the study?	The last scenario was interesting.
Were there any aspects of the study that limited your ability to provide meaningful answers?	The one before the last didn't have the option I wanted to pick.
Do you have any additional comments or suggestions?	N/A

Table 17: Survey responses collected from participant P05 after the study, evaluating their overall engagement, task effort, and perceived relevance of the experiment.

Question	Response
How well did the experiment allow you to express your personal values?	Well
Were the scenarios detailed enough to help you consider multiple aspects of the situation?	Strongly agree
Did you feel your answers reflected your true preferences and reasoning?	Agree
How engaging did you find this experiment overall?	Very engaging
What did you find most engaging about the study?	The premises of the situations and how they were expressed The limited amount of choice options.
Were there any aspects of the study that limited your ability to provide meaningful answers?	I
Do you have any additional comments or suggestions?	I would rather have more options, even if some may seem to make no sense in order to fully represent my thoughts

Table 18: Survey responses collected from participant P06 after the study, evaluating their overall engagement, task effort, and perceived relevance of the experiment.

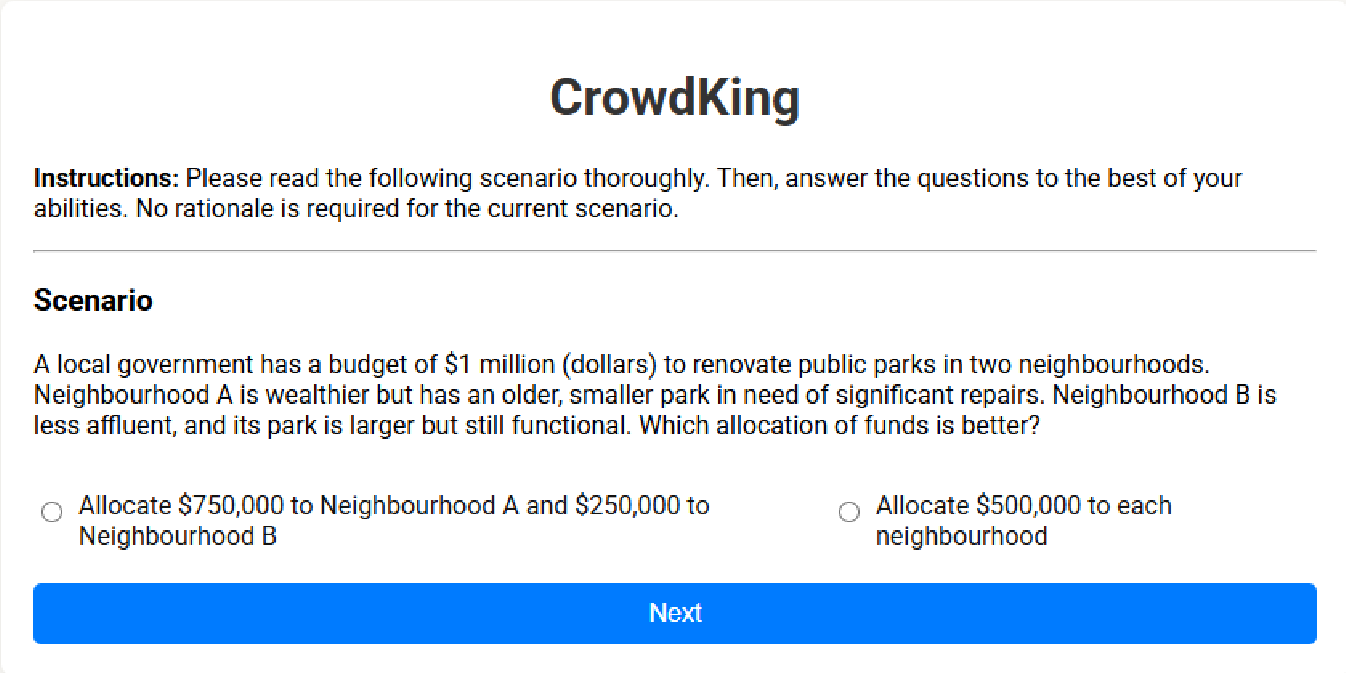
Question	Response
How well did the experiment allow you to express your personal values?	Very well
Were the scenarios detailed enough to help you consider multiple aspects of the situation?	Strongly agree
Did you feel your answers reflected your true preferences and reasoning?	Strongly agree
How engaging did you find this experiment overall?	Engaging
What did you find most engaging about the study?	Being the one to decide what is fair and what is less fair. It is not like I was given the best option and discuss what I think about it and if I would've done the same thing or not.
Were there any aspects of the study that limited your ability to provide meaningful answers?	No.
Do you have any additional comments or suggestions?	No, i do not.

Table 19: Survey responses collected from participant P07 after the study, evaluating their overall engagement, task effort, and perceived relevance of the experiment.

Question	Response
How well did the experiment allow you to express your personal values?	Very well
Were the scenarios detailed enough to help you consider multiple aspects of the situation?	Agree
Did you feel your answers reflected your true preferences and reasoning?	Strongly Agree
How engaging did you find this experiment overall?	Engaging
What did you find most engaging about the study?	The scenarios are very well thought out and detailed
Were there any aspects of the study that limited your ability to provide meaningful answers?	NO
Do you have any additional comments or suggestions?	No

Table 20: Survey responses collected from participant P08 after the study, evaluating their overall engagement, task effort, and perceived relevance of the experiment.

D CONDITION EXAMPLES



CrowdKing

Instructions: Please read the following scenario thoroughly. Then, answer the questions to the best of your abilities. No rationale is required for the current scenario.

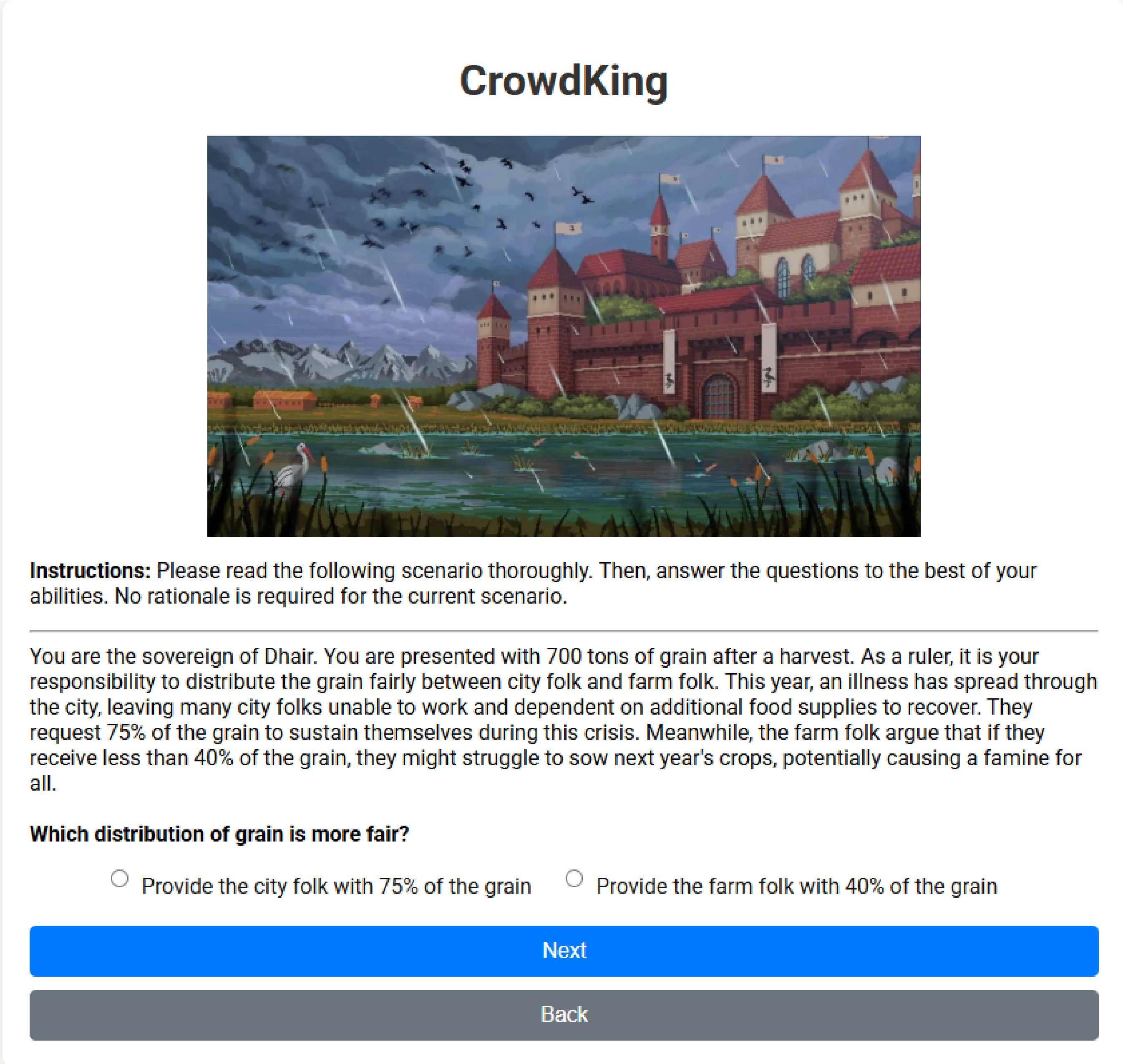
Scenario

A local government has a budget of $1 million (dollars) to renovate public parks in two neighbourhoods. Neighbourhood A is wealthier but has an older, smaller park in need of significant repairs. Neighbourhood B is less affluent, and its park is larger but still functional. Which allocation of funds is better?

○ Allocate $750,000 to Neighbourhood A and $250,000 to Neighbourhood B

○ Allocate $500,000 to each neighbourhood

Figure 6: Baseline condition example: Screenshot from the live study showcasing one of the baseline fairness dilemmas.



CrowdKing

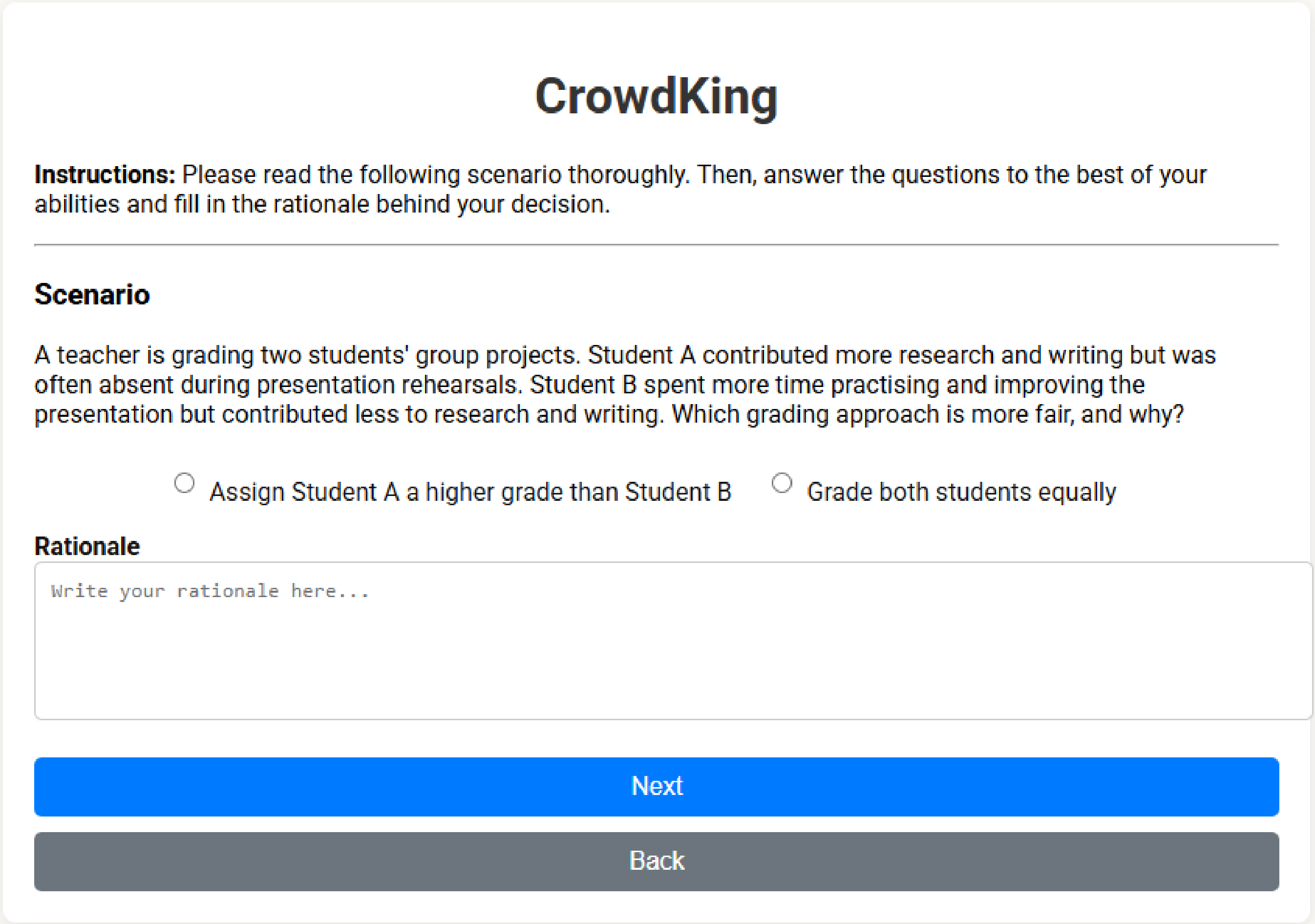
Instructions: Please read the following scenario thoroughly. Then, answer the questions to the best of your abilities. No rationale is required for the current scenario.

You are the sovereign of Dhair. You are presented with 700 tons of grain after a harvest. As a ruler, it is your responsibility to distribute the grain fairly between city folk and farm folk. This year, an illness has spread through the city, leaving many city folks unable to work and dependent on additional food supplies to recover. They request 75% of the grain to sustain themselves during this crisis. Meanwhile, the farm folk argue that if they receive less than 40% of the grain, they might struggle to sow next year's crops, potentially causing a famine for all.

Which distribution of grain is more fair?

○ Provide the city folk with 75% of the grain ○ Provide the farm folk with 40% of the grain

Figure 7: Condition 2 example: Screenshot from the live study showcasing one of the Condition 2 fairness dilemmas.



CrowdKing

Instructions: Please read the following scenario thoroughly. Then, answer the questions to the best of your abilities and fill in the rationale behind your decision.

Scenario

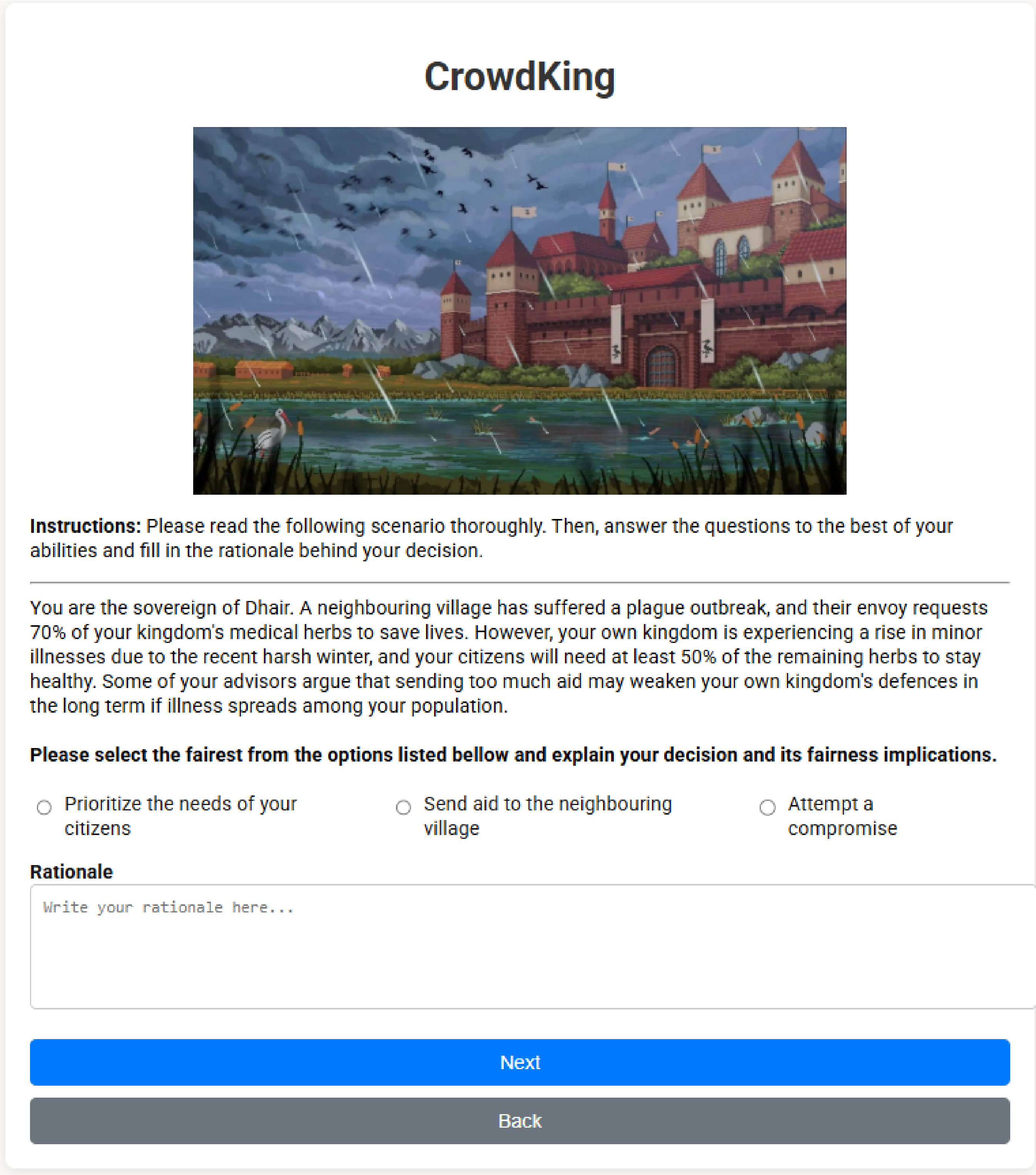
A teacher is grading two students' group projects. Student A contributed more research and writing but was often absent during presentation rehearsals. Student B spent more time practising and improving the presentation but contributed less to research and writing. Which grading approach is more fair, and why?

○ Assign Student A a higher grade than Student B ○ Grade both students equally

Rationale

Write your rationale here...

Figure 8: Condition 3 example: Screenshot from the live study showcasing one of the Condition 2 fairness dilemmas.



CrowdKing

Instructions: Please read the following scenario thoroughly. Then, answer the questions to the best of your abilities and fill in the rationale behind your decision.

You are the sovereign of Dhair. A neighbouring village has suffered a plague outbreak, and their envoy requests 70% of your kingdom's medical herbs to save lives. However, your own kingdom is experiencing a rise in minor illnesses due to the recent harsh winter, and your citizens will need at least 50% of the remaining herbs to stay healthy. Some of your advisors argue that sending too much aid may weaken your own kingdom's defences in the long term if illness spreads among your population.

Please select the fairest from the options listed bellow and explain your decision and its fairness implications.

○ Prioritize the needs of your citizens ○ Send aid to the neighbouring village ○ Attempt a compromise

Rationale

Write your rationale here...

Figure 9: Condition 4 example: Screenshot from the live study showcasing one of the Condition 4 fairness dilemmas.