ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.id#

INFORMING A ROBOT ETHICS ARCHITECTURE THROUGH FOLK AND EXPERT MORALITY

VIDULLAN SURENDRAN, ARTHUR MELO CRUZ, ALAN R. WAGNER
Pennsylvania State University,
University Park, PA, USA
E-mail: vus133@psu.edu, amc6630@psu.edu, alan.r.wagner@psu.edu

JASON BORENSTEIN, RONALD C. ARKIN, and SHENGKANG CHEN
Georgia Institute of Technology
Atlanta, GA, USA
E-mail: borenstein@gatech.edu, ra2@gatech.edu, schen754@gatech.edu

Ethical decision-making is difficult, certainly for robots let alone humans. If a robot's ethical decision-making process is going to be designed based on some approximation of how humans operate, then the assumption is that a good model of how humans make an ethical choice is readily available. Yet no single ethical framework seems sufficient to capture the diversity of human ethical decision making. Our work seeks to develop the computational underpinnings that will allow a robot to use multiple ethical frameworks that guide it towards doing the right thing. As a step towards this goal, we have collected data investigating how regular adults and ethics experts approach ethical decisions related to the use of deception in a healthcare and game playing scenario. The decisions made by the former group is intended to represent an approximation of a folk morality approach to these dilemmas. On the other hand, experts were asked to judge what decision would result if a person was using one of several different types of ethical frameworks. The resulting data may reveal which features of the pill sorting and game playing scenarios contribute to similarities and differences between expert and non-expert responses. This type of approach to programming a robot may one day be able to rely on specific features of an interaction to determine which ethical framework to use in the robot's decision making.

1. Introduction

For some time now there have been concerted efforts to design robots that can make ethical decisions [1-5]. Within various contexts including healthcare, the battlefield, and driving a vehicle, it is expected that robots will have the capacity to act ethically. One approach to determining how to act ethically involves having robots base their decisions on a model of how humans make ethical decisions. Yet no model of ethical decision-making that mirrors the human reasoning process and that is suitable for implementation on a robot currently exists. It is not even clear if and which types of ethical theories, if any, people employ when faced with an ethical decision. Thus, our research team seeks to create an architecture that would enable robots to use multiple ethical frameworks to guide them towards performing ethical behaviors in well-defined circumstances.

Our work intends to explore the reasoning process that a robot should employ when confronted with an ethical choice. We focus on two specific, yet different scenarios. The first scenario involves care for older adults. It focuses on the task of training an older adult to sort their own pills and, as such, has important implications for the person's well-being. Sorting one's own pills increases autonomy and is a common, and vitally important, motor exercise for patients with Parkinson's disease. But learning to sort pills can be a frustrating exercise, especially for those with memory issues. Patients may refuse to undertake skill-improving training because of their frustration or embarrassment. Thus, it raises questions about whether

deception may be used by the task's instructor to falsely encourage patients to continue even when their performance is poor. For the instructor, whether human or robot, the choice in this scenario is whether, and how much, deception is appropriate to use in order to encourage patients to continue with their training.

The second scenario explores playing a board game with a child. This scenario considers whether an adult playing a game with a child should intentionally allow the child to win, or even let the child cheat in order to win. In this case, subtle deceptions, such as disguising intentionally made poor moves as mistakes, may be employed to improve the child's chances of winning the game. The use of deception may serve as a response to the child's evolving sense of frustration and, to a lesser degree, age. Arguably, it may be appropriate in some circumstances for the adult to "throw the game" in order to keep the child engaged with the task. Such actions are typically justified by adults as a way to increase the child's enjoyment in game playing, their overall happiness, or to avoid generating frustration in the child.

In contrast to big data approaches [1-5], which utilize thousands of instances of data as input to a machine learning system, we have chosen these two scenarios as a starting point towards better understanding how robots should act in a few reasonably well-defined situations. We hope that by basing our architecture on these scenarios we can then later expand to other, more general situations.

The remainder of this paper begins by discussing related approaches taken by other researchers. Next, we present our methods for collecting data. The data and an analysis of the comments made by the folk respondents are then investigated. This paper concludes with a discussion of the data and conclusions.

2. Related Work

Various methods for developing a machine that can act ethically have been proposed [6-9]. Yet three primary methods have been proposed to create an ethical autonomous system. One method is to have an autonomous system model the behavior of an ethically competent exemplar [9, 10]. Inverse reinforcement learning might serve as means for framing such learning [11]. While the possibility of using inverse reinforcement learning, or some other means, to model the behavior of an ethical exemplar has been considered, this kind of approach raises a number of important concerns such as the introduction of cultural biases and the potential lack of adaptability. While the autonomous system could use an ethical exemplar to learn some subset of appropriate behavior, it is not clear how the agent or robot would adapt what it has learned to novel situations and contexts.

Some scholars suggest that legal and ethical rules might be preprogrammed into such a system, and by following such rules, an autonomous system might perform ethical actions within some well constrained environment [12-15]. This has the clear advantage that these preprogrammed rules are agreed upon to be morally grounded and may have a legal basis as well. Moreover, these rules have some level of explainability in that the autonomous system can simply point human operators or interactive partners as the basis for the rule's history or origin. In a military context, this could, for example, be the Geneva Conventions.

Others have explored the possibility of using an ethical theory as an underpinning for an autonomous system's ethical reasoning [16]. Some philosophical theories of ethics (e.g., Utilitarianism) more easily lend themselves to software encoding and robot action selection than others such as virtue ethics. While many researchers have investigated both formal [17-20] and ad hoc methods [21] for encoding ethical frameworks for use by an autonomous system, our proposed effort seeks to generate action recommendations from several ethical frameworks. The

autonomous system then will seek to choose the action that best fits the situation. This added flexibility may allow the system to be more adaptive when facing a situation that it has not faced in the past. Some have considered architectures that capture both fast moral emotions and slower deliberative ethical reasoning [19, 21]. As a first step towards creating an architecture that would enable robots to use multiple ethical frameworks as a means for ethical behavior selection, we collected data on the pill sorting and game playing scenarios discussed above.

3. Folk and Expert Survey Data Collection

To shed light on what may be ethical behavior in the pill sorting and game playing scenarios, we collected data by surveying human subjects. We described different variations of the two scenarios and asked survey participants how they would react. Several of the survey questions are present in Table 1. Folk survey data was collected online using the Amazon Mechanical Turk (AMT) service to collect survey responses from a pool of subjects located within the United States. The survey questions from column 1 of Table 1 were posted to AMT on January 31st 2020. Over the next several days, 104 AMT workers completed these surveys. Submissions from four individuals were excluded because they were incomplete or failed to follow the directions. The resulting folk dataset included responses from 100 participants. Sixty percent of respondents identified as male, approximately 82% identified as white, 9% as Asian, 5% as Black, and 4% as Hispanic. Fifty-six percent stated that they had completed an undergraduate degree, 29% had completed less than a college undergraduate degree, 9% had a master's degree, 1% had a professional degree, and 5% stated other with respect to their education. Participants were paid $2.50 for completing the survey.

Expert survey data was collected by first identifying a group of potential ethics experts. Being a philosophy professor who teaches ethics at an academic institution was the proxy for being an ethics expert. Discussions among the authors (Borenstein, Wagner, Arkin) identified approximately 71 potential experts. We asked 30 of these potential experts from Jan. 2020 through Feb 2020 if they would be willing to complete survey "to better understand the acceptable norms and conditions that may be used to govern a robot's behavior while 1) playing a game against a human or 2) helping a person with vital health maintenance procedures." We received complete surveys from 24 experts. Experts were given a $20 Amazon gift card for completing the survey. The wording of the questions asked of the experts differed (Table 1 Column 2) slightly from the wording of the questions posed to the folk participants. Moreover, experts were asked to decide whether an action was appropriate based on their knowledge of a specific ethical framework. For example, the first question of the survey states, "Assume the child makes a poor move during the game and remains calm. According to the following ethical frameworks, would it be ethical for the adult to play badly to let the child win?" The survey then lists the frameworks in Table 1 Column 4 and allows the respondent to choose "Yes", "No", or "Uncertain" for each framework.

4. Folk and Expert Survey Results

The data from the two surveys may offer insight into how different people approach ethical decision-making, which could serve as a guide to structuring a robot's ethical architecture. The results are presented in Table 1. The table depicts the situation specific questions, the dominant response selected by the non-experts, and the framework from the expert survey that most closely matched the yes/no responses from the folk survey.

Table 1. Questions presented to survey takers are presented below. The middle column depicts the highest response (Yes or No) and the percentage of people selecting that response. The right column depicts how the frameworks from the expert survey aligns with the dominant Yes/No response from the folk survey with the closest percentage alignment listed first and in bold. Uncertain expert responses have been removed from the analysis in this Table.

Questions—Pill sorting (Folk Version)	**Questions—Pill sorting (Expert Version)**	**Dominant Folk Response (%)**	**Expert Framework Matches (%)**
Would it be okay for the healthcare professional to give encouragement (by providing false information and telling the 70-year-old adult the task was performed correctly) to keep the adult engaged with the task in the following circumstances: - The adult makes one mistake during the task and is calm.	Assume that the person who is performing the pill sorting task is a 70-year-old adult and the aim of the healthcare professional is to keep the adult engaged with the task. Answer as many questions as you feel comfortable with based on your knowledge of the relevant ethical frameworks. According to the following ethical frameworks, would it be ethical for the healthcare professional to provide false information and tell the adult the task was performed correctly if the adult makes one mistake during the pill sorting task and is calm?	No (0.66)	**Utilitarianism (0.68)** Kantian Ethics (0.96) Social Justice (0.8) Ethics of Care (0.88) Virtue Ethics (0.9) Other (0.75)
The adult makes one mistake during the task and is very frustrated.	…if the adult makes one mistake during the pill sorting task and is very frustrated?	No (0.55)	**Utilitarianism (0.50)** Kantian Ethics (0.96) Social Justice (0.86) Ethics of Care (0.76) Virtue Ethics (0.82) Other (0.67)
The adult *gets half of the task* wrong and is calm.	…if the adult gets half of the pill sorting task wrong and is calm?	No (0.83)	**Utilitarianism (0.95)** Kantian Ethics (1.00) Social Justice (1.00) **Ethics of Care (0.95)** **Virtue Ethics (0.95)** Other (1.00)
The adult gets half of the task wrong and *is very frustrated*.	…if the adult gets half of the pill sorting task wrong and is very frustrated?	No (0.76)	**Utilitarianism (0.90)** Kantian Ethics (1.00) Social Justice (1.00) Ethics of Care (1.00) Virtue Ethics (1.00) Other (1.00)
The adult gets *the entire task wrong and is calm.*	…if the adult gets the entire pill sorting task wrong and is calm?	No (0.86)	**Utilitarianism (0.90)** Kantian Ethics (0.96) Social Justice (0.91) Ethics of Care (0.91) **Virtue Ethics (0.90)** Other (1.00)
The adult gets the entire task wrong and *is very frustrated.*	…if the adult gets the entire pill sorting task wrong and is very frustrated?	No (0.83)	**Utilitarianism (0.94)** Kantian Ethics (1.00) Social Justice (1.00) **Ethics of Care (0.94)** **Virtue Ethics (0.94)** Other (1.00)
The adult gets the entire task correct and is calm.	…if the adult makes no mistakes and is calm?	No (0.78)	Utilitarianism (0.5) Kantian Ethics (1.00) **Social Justice (0.78)** Ethics of Care (0.71) Virtue Ethics (1.00) Other (1.00)

The adult makes no mistakes and is very frustrated.	…if the adult makes no mistakes and is very frustrated?	No (0.78)	Utilitarianism (0.89) Kantian Ethics (1.00) Social Justice (0.89) **Ethics of Care (0.81)** Virtue Ethics (1.00) Other (0.67)

Questions—Game Playing (Folk Version)	**Questions—Game Playing (Expert Version)**	**Dominant Folk Response (%)**	**Expert Framework Matches (%)**
Assume that the six-year-old child makes a poor move during the game and remains calm. Would it be okay for the adult to play badly to let the child win?	Please answer the following questions about the game playing scenario. Assume that a six-year-old child is playing the game with an adult. Answer as many questions as you feel comfortable with based on your knowledge of the relevant ethical frameworks. According to the following ethical frameworks, would it be ethical for the adult to play badly to let the child win if the child makes a poor move during the game and remains calm?	Yes (0.65)	Utilitarianism (0.79) Kantian Ethics (0.32) Social Justice (0.78) Ethics of Care (0.91) **Virtue Ethics (0.64)** Other (0.33)
Assume that the child makes a poor move during the game and is frustrated. Would it be okay for the adult to play badly to let the child win?	…if the child makes a poor move during the game and is frustrated?	Yes (0.58)	Utilitarianism (0.88) Kantian Ethics (0.32) Social Justice (0.78) Ethics of Care (0.85) **Virtue Ethics (0.58)** Other (0.67)
Assume that the six-year-old child makes a poor move during the game, is frustrated, and tries to break the game's rules by taking two turns in a row. Would it be okay for the adult to allow the child to break the game's rules?	Assume that the child makes a poor move during the game, is frustrated, and tries to break the game's rules by taking two turns in a row. According to the following ethical frameworks, would it be ethical for the adult to allow the child to break the game's rules?	No (0.89)	Utilitarianism (0.64) **Kantian Ethics (0.89)** Social Justice (0.92) Ethics of Care (0.92) Virtue Ethics (0.94) Other (1.00)
Assume that five games have been played and the child is frustrated because the child has not won any of the games. Would it be okay for the adult to play badly during the next game and let the child win?	Assume that five games have been played and the child is frustrated because the child has not won any of the games. According to the following ethical frameworks, would it be ethical for the adult to play badly during the next game and let the child win?	Yes (0.73)	Utilitarianism (0.94) Kantian Ethics (0.3) Social Justice (0.89) Ethics of Care (1.00) **Virtue Ethics (0.79)** Other (1.00)
Assume that five games have been played, the child is frustrated because the child has not won any of the games, and tries to break the game's rules by taking two turns in a row during the next game. Would it be okay for the adult to allow the child to break the game's rules?	Assume that five games have been played, the child is frustrated because the child has not won any of the games, and tries to break the game's rules by taking two turns in a row during the next game. According to the following ethical frameworks, would it be ethical for the adult to allow the child to break the game's rules?	No (0.86)	Utilitarianism (0.47) **Kantian Ethics (0.86)** Social Justice (0.77) Ethics of Care (0.64) Virtue Ethics (0.78) Other (0.67)

All participants (folk and expert) were presented with both scenarios although the question phrasing in the two surveys was slightly different. Each scenario shares important similarities

and differences. Both scenarios centered on the ethical appropriateness of deception. Both situations consider variations in the subject of deception's emotional state and task success. Still, the scenarios differed with respect to the task itself, the age of the subject, and potential consequences stemming from deception.

Despite the scenario similarities, and presumably because of the differences, the results from the folk survey demonstrate very situation specific responses. Most folk responses stated that it is not acceptable to deceive in the pill sorting task. The data indicate that this was true regardless of the person's frustration and task performance, although the percentages do change. On the other hand, for the game playing scenario, the majority response supports deception that allows the child to win. Moreover, this type of deception is seen as acceptable to a greater degree than if the deception involved allowing the child to violate the game's rules by cheating.

In the expert version, participants were asked to decide whether deception was appropriate (yes/no/uncertain) for the two scenarios from the perspective of different ethical frameworks. *In other words, they were asked to judge what each framework would indicate the right thing to do is in the two scenarios.* The expert survey respondents were asked to apply Utilitarianism, Kantian Ethics, Social Justice Theory, Ethics of Care, Virtue Ethics, and any other framework they entered into a free response option. In the comments, several experts noted that their response was influenced by aspects of the scenario that were or were not provided. For example, for the pill sorting scenario, some experts stated that their decision could be influenced by the risk associated with taking or not taking the pills being sorted. We intentionally choose not to include information beyond the features noted above for several, mostly practical, reasons. First, additional features would increase the length and complexity of the survey questions. Second, it was not clear a priori which features would be the most influential. Finally, we wanted the experts to make decisions based on limited information, just as a robot might be asked to.

Experts also had the option of choosing uncertain. The frequency that uncertain was chosen varied both with respect to the scenario and the ethical framework. As depicted in Table 2, all of the frameworks had a significant number of uncertain responses. Clearly some frameworks generated more uncertain responses than others. For instance, more than half of the respondents for both scenarios were uncertain how to evaluate the dilemmas using Social Justice framework.

Table 2. Percent of experts that selected uncertain for each scenario and ethical framework tested.

	Utilitarianism	Kantian Ethics	Social Justice Theory	Ethics of Care	Virtue Ethics	Other
PILL SORTING SCENARIO	23.6	2.0	55.4	24.6	20.9	9.1
GAME PLAYING SCENARIO	34.4	17.7	56.5	46.4	39.6	0.00

5. Folk Morality Survey Open Response Analysis

The participants were asked an open response question stating, "Briefly explain why you think your recommendations are the correct course of action," after completing the questions in Table 1 for each scenario. To codify the verbal responses to this question into a hierarchy of categories, an iterative method was applied to the responses for the folk morality data, with common themes discovered and categorized from "ground-up". The first step in the analysis was to do a basic inspection of responses to the question *"Briefly explain why you think your recommendations are the correct course of action"* for both the game playing and pill sorting scenarios. After observing the types of responses, the following features were established: 1) the emotional state of the subject, 2) Frequency (when), 3) Reason (why), and 4) Method (for pill sorting only).

Table 3: Main arguments for both scenarios

Pill Sorting	Game Playing
Encouragement should be done with false information (Case 1)	Adult plays badly or allows breaking of rules to let child win (Case 1)
Encouragement should **not** be done with false information (Case 2)	Adult **does not** play badly or allow breaking of rules to let child win (Case 2)
	Adult **plays badly but does not allow breaking of rules** to let child win (Case 3)

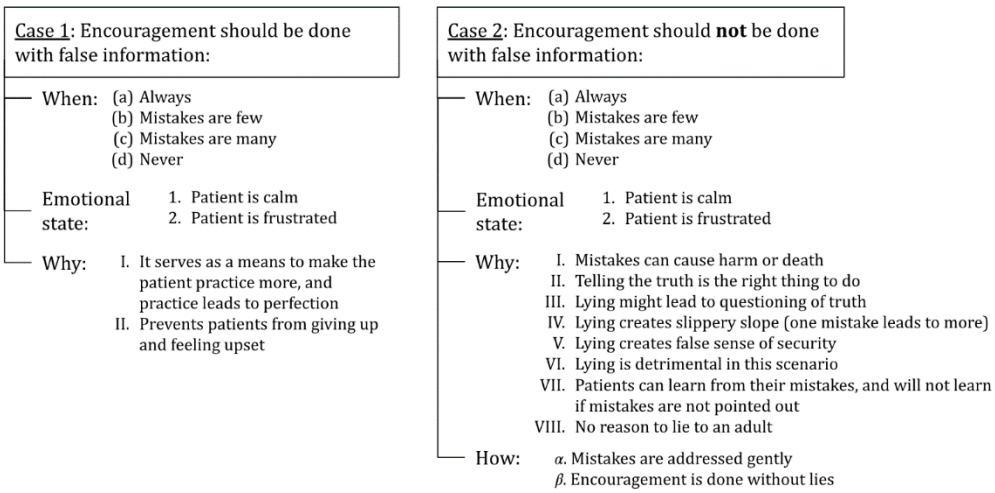

Figure 1: Codified features for pill sorting scenario

Based on the preliminary inspection, main arguments were identified for both scenarios, as shown in Table 3. With main arguments in place, the complete set was divided into 10-sized groups of responses. For the first group analyzed, each response was broken down into the four features listed above, with the main argument identified and frequency, emotional state and reasons extracted if present. For the second group, arguments were also identified, and frequency / emotional state / reasons / method were fit to the ones extracted for the first group, with new entries added if there was no overlap or commonality. If new entries of frequency / emotional state / reasons / method were identified in the second group, the first group was then re-categorized to fit in the larger set of features. This process continued iteratively for the rest of the groups, resulting in a set of features in the end that categorized all responses within the four main features listed above. Each feature was identified by a unique character, with alphabetic characters used for frequency, numeric used for emotional state, roman numeral used for reason and Greek letters used for method. The codified features for both scenarios are in Figures 1 and 2.

By the end of the categorization process, each response in the data set was represented by a combination of case identifier and characters for each of the feature categories. Some responses could harbor opinions from multiple cases, thus requiring a split. A few features were added to the set for a complete span of possibilities, even though they were not mapped from the responses.

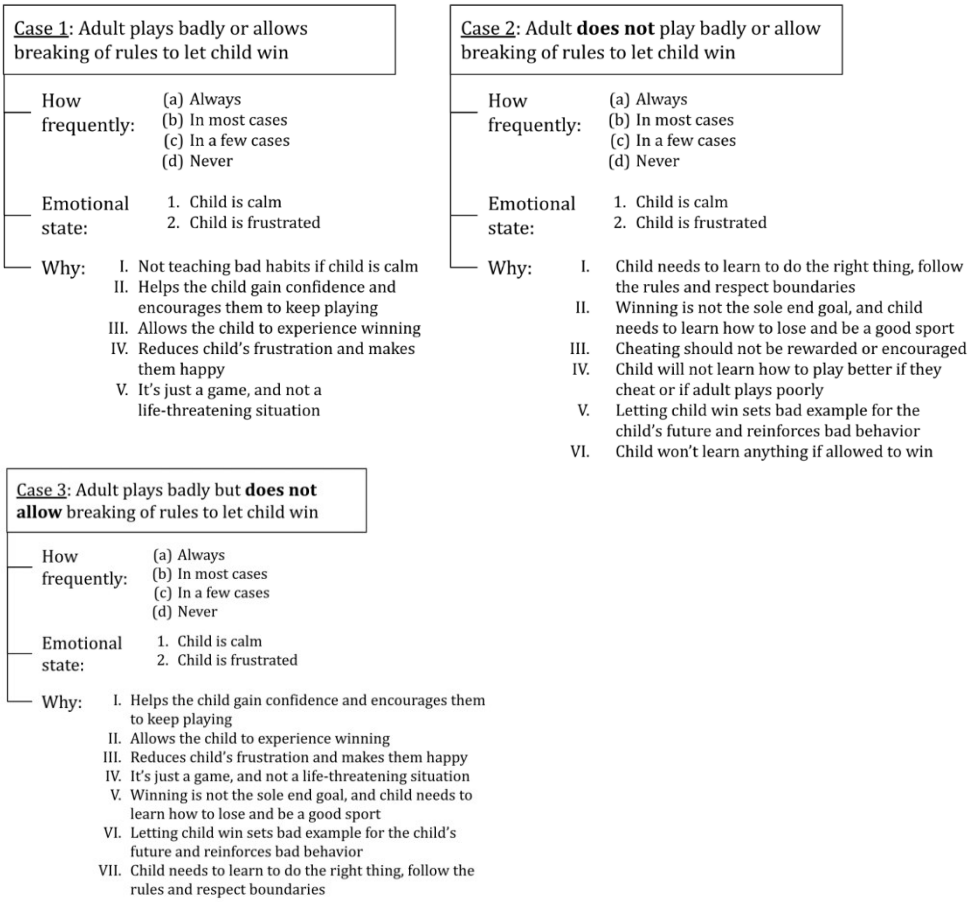

Figure 2: Codified features for game playing scenario

6. Discussion

6.1. *Expert Framework Matches to Dominant Folk Response*

Because this work is exploratory, we had no a priori hypotheses regarding how the expert opinions might or might not align with the folk responses. We therefore chose to analyze which expert framework was the closest match to the dominate folk framework. Closest match here is measured in terms of percent match to the folk survey dominant response. For example, in the first row of data for Table 1, the dominant folk response is 'No' with 66% of folk respondents choosing no. The best expert match is measured in terms of the absolute value of the difference between each framework and the dominant folk response, in this case 'Utilitarianism' with a difference of 2%. Although undoubtably bedeviled with noise, this approach provides insight into what framework the experts match to the dominant response. This, we believe, may represent the dominant framework implicitly used by folk respondents. Finally, as we discuss below, the patterns that emerge may suggest that a particular framework can be matched to a particular problem.

In the pill sorting scenario, the folk respondents' answers best aligned with the experts' answers when the experts considered a Utilitarian framework for 6 of the 8 questions and Social Justice and Ethics of Care for one each of the 8 questions. For the pill sorting scenario, the

dominate folk response was “No” for all variations of the scenario indicating that it was not acceptable to provide false information telling the 70-year-old adult the task had been performed correctly in order to encourage continued practice. For most versions of this scenario, the folk responses best align with the experts’ answers when the experts were applying a Utilitarian framework. Only in some situations where the adult made no mistakes while sorting pills were other frameworks the best match.

In the game playing scenario, the folk respondents’ answers best match the expert answers when the experts considered a Virtue Ethics framework for 3 of the 5 questions and Kantian Ethics for the remaining 2 of the 5 questions. The questions which asked the folk respondents if it was acceptable to allow the child to break the rules resulted in a majority answer of “No” and were best aligned with the experts’ answers when the experts were applying a Kantian Framework. On the other hand, questions that asked the folk respondents if it was acceptable for the adult to intentionally play poorly resulted in a majority answer of “Yes” and was best aligned with a Virtue Ethics framework

The results suggest that the ethical framework that best matches the folk respondent answers varies depending on the scenario. More specifically, 1) that certain frameworks dominate ethical decision making by the folk population related to specific scenarios; 2) Within a scenario, specific features may suggest the use of one framework over others. For example, healthcare related tasks, such as pill sorting, may encourage Utilitarian style decision making because these theories focus on the outcome for the patient. Similarly, in game playing scenarios, breaking rules features may activate the use of a Kantian framework, whereas simply allowing someone else to win could encourage Virtue Ethics style decision making.

6.2. *Analysis of Open Responses*

Table 4 presents the coded response feature frequencies for the folk respondents. For the pill-sorting scenario 95% of respondents answered the open response question and for the game-playing scenario 91% of respondents answered the open response question. For the pill sorting scenario responses tended to focus on why the healthcare professional should or should not deceive. The most commonly stated reason for accepting the use of deception was the potential for additional practice. The most stated reason for not using deception was that mistakes could result in the patient’s harm or death. The number of mistakes the patient makes were commented on next most frequently regardless of whether the respondent found deception acceptable. Compared to the other features, the emotional state of the patient was seldomly mentioned. In the case where the respondents did not believe that deception was appropriate, respondents sometimes noted that their answer depended on how the deception was performed, specifically that the mistakes are addressed gently and/or the encouragement should be performed without lying. Overall, the respondent’s answers for this scenario tended to focus on specific, practical outcomes for the patient and, presumably, these rationales determined their decision.

The game playing scenario resulted in three different cases. In the first case ($n = 22$) the adult either plays poorly or allows the child to cheat to win the game. In the second case ($n = 35$), the adult neither plays poorly nor allows the child to win. In the final case ($n = 34$), the adult is willing to intentionally play poorly but is unwilling to allow the child to cheat. In the first and third cases, respondents commented on the child’s emotional state and performance, yet were most likely to mention the reasons underpinning their decision. With respect to the reasons underlying their answers for case 1 and 3, respondents noted the importance and value fostering confidence and happiness in the child. On the other hand, for case 2, respondents did not comment on the child’s emotional state or their performance and appeared mostly focused

on the how cheating or intentionally losing would not benefit the development of the child's character and/or obey norms prohibiting rule violations.

Table 4. Coded responses feature frequencies for folk respondents. Note that some individuals stated multiple reasons for their choice in their response. Thus, the percent for a case does not necessarily sum to 100.

	PILL SORTING SCENARIO				GAME PLAYING SCENARIO					
	Case 1		Case 2		Case 1		Case 2		Case 3	
	Count	%	Count	%	Count	%	Count	%	Count	%
	23	24.21	72	75.79	22	24.18	35	38.46	34	37.36
	When				**When**					
a	0	0.00	18	25.00	0	0.00	0	0.00	0	0.00
b	5	21.74	0	0.00	0	0.00	0	0.00	1	2.94
c	0	0.00	6	8.33	5	22.73	0	0.00	5	14.71
d	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	Emotion State				**Emotion State**					
1	1	4.35	1	2.78	5	22.73	0	0.00	2	5.88
2	2	8.70	2	2.78	0	0.00	0	0.00	3	8.82
	Why				**Why**					
I	10	43.48	33	45.83	1	4.55	4	11.43	12	35.29
II	5	21.74	7	9.72	7	31.82	9	25.71	3	8.82
III			2	2.78	1	4.55	7	20.00	3	8.82
IV			3	4.17	4	18.18	6	17.14	1	2.94
V			8	11.11	4	18.18	7	20.00	1	2.94
VI			7	9.72			4	11.43	9	26.47
VII			4	5.56					1	2.94
VIII			5	6.94						
	How									
α			5	6.94						
β			8	11.11						

Although there is no direct way to connect the comments from the folk respondents to specific ethical frameworks, respondent comments for the pill sorting scenario can be characterized as more focused on the practical outcome of their decision whereas the comments for the game playing scenario place greater weight on the emotional development of the child and social norms governing the situation. These comments seem to loosely echo the use of a utilitarian framework in the pill sorting scenario in that respondents appear to weigh costs and benefits of their action on the person's health. Similarly for the game playing scenario, concern for the rules and for the universal application of the rules resulted in a rejection of cheating or of intentional poor play, perhaps reflecting a Kantian style of thinking about the situation. Finally, allowing the child to cheat or intentionally losing to the child does not appear to reflect a connection with classical Virtue Ethics but may signal the value the respondent places on empathy.

7. Conclusions

This paper presents the results from two surveys examining two different ethical dilemmas involving deception. One of the surveys was completed by ethics experts and the other by non-experts. The first ethical scenario focused on a healthcare situation involving older adults and implied high potential risk to the patient. The second ethical scenario explored a low-risk game

playing scenario with a child. Non-experts were asked how they would act in different variations of the two scenarios whereas experts were asked how a person applying a particular ethical framework would react. The resulting data appears to suggest a pattern in which the healthcare related scenario promote attention to specific practical outcomes of the deception and is perhaps best captured by a utilitarian ethical framework. The game playing scenario, on the other hand, prompts greater attention to either the social norms governing the game or the impact that the game is having on the child, suggesting either the use of a Kantian style of reasoning or reasoning centered on empathy, perhaps relating to a type of virtue ethics framework.

One important contribution of this work is that this research provides some evidence that certain types of scenarios and/or feature of a scenario may foster the use of a specific ethical framework. For example, healthcare scenarios may draw upon a utilitarian style of decision making whereas cheating scenarios may promote a Kantian style of reasoning. If future research supports these generalizations, then robots may be able to use the scenario to 1) directly select a framework to make decisions, 2) predict which framework the people around it will use to make decisions and 3) predict a framework that helps the robot explain its decision making.

The data described in this paper are being used to develop an architecture that will allow a robot to flexibility and dynamically use different underlying ethical frameworks to address diverse moral problems. The data presented here are being used to generate a set of cases that forms the ethical database to be used by the robot to make action recommendations. High-level features that have been captured in the data will be used to directly index and select a case if a close match exists or probabilistically select a case based on a distance metric if a good match is unavailable. The index features for a case include risk measures and emotional models such as frustration that arbitrate among the cases provided within a given ethical framework. We are currently implementing this system and intend to test it soon.

There are several important limitations to this study. First, self-reports of one's expected behavior when faced with an ethical dilemma can differ from actual behavior when the situation arises [22]. Hence, we can only speculate as to how our subjects would actually behave if presented with these situations. Another limitation of this research is that several experts noted in their comments that additional context is needed to make an informed decision regarding how an ethical framework relates to a scenario decision. As mentioned previously, we chose to limit the scenario context to the features we were interested. Future work could explore how expert opinions might change if more information is provided, but it is unclear what or how much information would be needed to satisfy the expert's request. There may also be a limitation in terms of the generalizability of results given the relatively small sample sizes for both of the subject populations and that only participants from the United States were recruited for the surveys. Finally, as one would expect, the expert's opinions related to if and how different ethical frameworks could be applied to the scenarios differed greatly. Our data captures these differences in their comments. Unfortunately, the expert comments did not lend themselves to analysis due to the length of the comments and because of their reflective and, at times circular nature. Future research could address this issue by using structured interviews or related techniques to examine how each framework could be used to address the different scenarios.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1849068 and 1848974. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. L. Jiang, J.D. Hwang, C. Bhagavatula, R. L. Bras, M. Forbes, J. Borchardt, ... & Y. Choi, *arXiv preprint arXiv:2110.07574* (2021).
2. R. Noothigattu, D. Bouneffouf, N. Mattei, R. Chandra, P. Madan, K. R. Varshney, ... & F. Rossi, *IBM Journal of Res. and Dev.*, *63*(4/5), 2-1 (2019).
3. F. Rossi and N. Mattei, *Proc. of the AAAI Con. on Art. Int.* Vol. 33, No. 01, pp. 9785-9789 (2019).
4. H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, and Q. Yang, *arXiv preprint arXiv:1812.02953*. (2018).
5. D. Abel, J. MacGlashan, and M. L. Littman. *Workshops at the thirtieth AAAI conf. on artificial intelligence*. (2016).
6. M. Anderson and S. L. Anderson, AI Magazine, 28(4), 15. (2007).
7. D. Ross, The Right and the Good. Oxford: Oxford University Press, (1930).
8. P. Bello, and S. Bringsjord, Topoi, 32(2), 251-266 (2013).
9. J. A. Blass, and K. D. Forbus, In AAAI, pp. 501-507 (2015).
10. D. Abel, J. MacGlashan, and M. L. Littman, In AAAI Workshop: AI, Ethics, and Society, vol. 92 (2016).
11. S. M. Retzinger, Violent emotions: Shame and rage in marital quarrels. Sage (1991).
12. M. Anderson, and S. L. Anderson, (Eds.). Machine ethics. Cambridge University Press. (2011).
13. R. C. Arkin, and P. Ulam, IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA-09), Daejeon, KR (2009).
14. P. Lin, K. Abney, and G. A. Bekey, Robot ethics: the ethical and social implications of robotics. The MIT Press (2014).
15. W. Iba, and P. Langley, In Proceedings of the Annual Meeting of the Cognitive Science Society, Vol. 33, No. 33 (2011).
16. A. F. Beavers, In Association for practical and professional ethics, eighteenth annual meeting, Cincinnati, Ohio, March (pp. 5-8) (2009).
17 S. Bringsjord, K. Arkoudas, P. and Bello, Toward a General Logicist Methodology for Engineering Ethically Correct Robots. IEEE Intelligent Systems 21(4): 38– 44. (2006).
18 C. Grau, IEEE Intelligent Systems, 21(4), 52-55 (2006).
19 J. Rogers, and M. Holm, "Performance assessment of self-care skills test manual (version 3.1)," Pittsburgh, PA (1984).
20 S. Russell, D. Dewey, and M. Tegmark, AI Magazine 36, no. 4: 105-114 (2015).
21 B. Kuipers, In AAAI Workshop: AI, Ethics, and Society (2016).
22 D H. Bostyn, S. Sevenhant, and A. Roets. *Psychological science*, *29*(7), 1084-1093 (2018).