Human-Machine Ethics: Experiments in Moral Responsibility

David Burke[1]
Galois, Inc., Portland, OR, 97024,USA

Successful human-crewed interstellar missions will require effective human-machine decision partnerships, particularly during mission-critical scenarios with significant ethical consequences. We're interested in identifying the salient factors that would encourage such partnerships. Our working hypothesis is that machine intelligence by itself is not sufficient: there needs to be a particular kind of congruence between man and machine for this to work. Specifically, machines must be capable of conveying to humans that they understand the moral stakes involved in a situation, and will decide and behave accordingly. Only then will humans grant moral responsibility to machines and treat them as equal partners in decision scenarios. In this paper, we present the results of online experiments designed to test our congruence hypothesis by placing participants in the middle of an ethical conundrum, and investigating how they respond depending on whether their decision partner is a human or a machine. We conclude that our congruence is a plausible explanation for the observed differences: humans need their decision partners to understand and embody moral responsibility.

I. Introduction

THE success of any human-crewed interstellar mission will depend on the existence of effective human-machine relationships - in particular, the ability for humans and machines to work together effectively as a decision-making team when the ethical stakes are high. We anticipate that machines during such a mission won't simply play the part of a supporting, background role, like an autopilot. Instead, navigating the demands of such a mission means that machines will be equal partners with humans, helping to make decisions under conditions of irreducible uncertainty

[1] Principal Scientist

with potentially grave consequences, such as situations involving potential loss of life or the inability to complete the mission.

We're interested identifying the salient factors that would either encourage or discourage effective partnerships between humans and machines in mission-critical scenarios. Our hypothesis is that there needs to exist a certain set of *congruences* between human and machine in order for humans to grant ethical decision-making legitimacy to machines. Specifically, without these congruences, humans will not grant moral responsibility to machines and treat them as equal partners to humans.

In order to test our claims, we conducted online experiments by adapting the well-known "trolley problem" to create ethical decision-making scenarios involving humans and machines in partnership. Specifically, we wanted to see how humans assign blame in situations where either other humans or machines have given ethical advice, and the outcome is starkly negative. Do humans treat the human and machine advisors similarly? Do humans conceptualize machine advisors as equal ethical partners? If not, what factors are missing?

An outline of this paper is as follows: we begin by discussing existing research into decision-making and the challenges of high-stakes decision-making in particular, and then continue with a discussion of various schools of ethical thought. Using findings from these two sections, we present our congruence hypotheses for effective human-machine ethical decision-making. We then discuss the experiments we did to test our hypotheses – an initial set of online experiments, what we learned from that initial experimentation, and then the design of a follow-up set of experiments that we recently conducted. We close with a discussion of what we've learned, what further questions have been raised, and what conclusions we might draw in terms of designing effective human-machine ethical decision-making partnerships.

II. Challenges in Decision-making

Human beings are decision-making machines. Much of how we spend our lives is making decisions, most of them so minor (12oz. or 16oz coffee drink this morning? Park in this open space or that one?) that we aren't even aware that our days are filled with them. But when the stakes are high enough, we pay attention, and oftentimes we struggle. This can be the case even if we don't think of the decision we're making as an explicit ethical one, for instance a decision about which college to attend. Divorced from considerations of how it might affect others (e.g., who is paying for the tuition, whether the college buildings are carbon-neutral), the decision doesn't involve ethics. But in general, the more consequential a decision is, the more likely it contains ethical import.

Defining Hard Decisions

As context for our experiments and discussion, it is useful to review how decision theorists think about their subject. One key distinction is that between process and outcome. Ideally, you have a sensible decision-making process that leads to a positive outcome. But process and outcome aren't linked: you can be lucky, and get a good outcome even with a poor decision-making process. Or else you can be unlucky, and get a poor outcome even if your process was solid. This main reason for making the distinction between process and outcome is that in many circumstances, you have control over the process, but not over the outcome, and you want to expend your time and energy on what you can control.

It needs to be added, though, that in humans, there is a well-documented cognitive bias towards good outcomes [1]. That is, a person gets credit for being a good decision-maker even when the process was flawed as long as the outcome was positive. Conversely, if a decision results in a bad outcome, then even if the process of coming to that decision was flawless, the decision-maker's reputation suffers.

Decision theorists have attempted to identify the attributes of a decision problem that make the decision "hard" (or at least "harder than others") [2]. They proposed the following seven factors as instrumental to the assessment that a decision is hard:

1. The outcome is consequential, and a poor decision can lead to a significant loss.
2. The number of options is either too small or too large.
3. The effort needed to make a decision (in terms of time, energy, resources, etc.) is significant.
4. The space of possible outcomes is difficult or impossible to enumerate.
5. It is difficult to compare the relative goodness/badness of potential outcomes.
6. The decision-makers aren't clear how good or bad they will feel about a potential outcome.
7. The decision-makers are receiving conflicting advice while coming to a decision.

The authors of [3], summarizing their experience working with first responders, claim that two factors in particular stand out when separating hard decisions from more run-of-the-mill ones: *task ambiguity*, and *outcome uncertainty*. Task ambiguity refers to the difficulty in not being able to specify the task at hand, so it's not clear what the problem is that you're trying to solve with your decision. Outcome uncertainty refers to the challenge of attaining clarity about what the potential consequences of a decision really amount to. In a nutshell, a decision is hard when you don't know what options you have, nor what the consequences of your decision truly are.

The conclusion that some decisions are "irreducibly hard" is consistent with the recognition in decision theory that human beings aren't best modeled as "rational optimizers", but rather as "satisficers" [4]. Human beings have a finite (and limited) capacity to evaluate alternatives and do the calculations that would result in the mathematical ideal of maximizing utility. Instead, humans bring experience and heuristics (i.e., cognitive shortcuts) to bear in order to make decision-making tractable, especially under conditions when the time available to make a decision is severely limited. Human beings tend to shoot for "good enough" decisions under the circumstances.

Decision Theory in Practice

Traditional decision theory makes the distinction between normative (prescriptive) and descriptive decision theory. The former purports to describe how humans "should" make decisions, and normative approaches usually have as their ideal the optimization of a decision under a mathematical framework. In contrast, descriptive approaches seek to understand decision-making as "This is how decisions get made, warts and all"; this research program often uses laboratory experiments to get this information, and so it has been criticized for the conclusions being based on contrived scenarios.

In recent years there has been increased interest in alternative decision-making paradigms. One prominent example is the "Naturalistic Decision-Making" (NDM) approach [5]. The traditional rational actor enumerates all possible choices, evaluates probabilities and utilities, aggregates all this numerical information in a decision tree-like structure, and reaches a conclusion. NDM says instead, decision-makers, even when confronted by severe time pressure, high stakes, and ill-specified goals, don't go through the rational actor process. In fact, the most seasoned decision-makers use a combination of experience and intuition to hone in on the most salient attributes of the decision problem (as opposed to collecting information about all possible attributes), consider only a very small set of alternatives (as opposed to looking at all possible alternatives), and use the current context to drive their selection (as opposed to a formal calculation over all possible futures).

As an example, the domain of chess provides support for the NDM approach – grandmaster chess players are not successful because they evaluate many more possible moves than, say, an amateur. Studies have demonstrated that grandmasters only evaluate a handful of lines. But they are typically just the best ones; a grandmaster never even bothers to consider poor choices.

Further research bolsters the idea that humans are not utility maximizers, especially when faced with hard decisions. Again, utility maximization requires a superhuman ability to gather all the alternatives, along with the

probabilities and utilities of each alternative. In a study of decision-making by mothers who received genetic counseling, and were told that their unborn children have a high risk of abnormality [6], they found that working with probabilities (even rough order-of-magnitude ones) weren't part of the decision process. Instead, they found that the decision-making was extremely contextual and dealt more with answering questions like "If outcome X happened, would I, and the other decision stakeholders be able to make the changes necessary to deal with and resolve this outcome? Could we move forward effectively?" (In these cases, stakeholders were family members such as husbands and other children) This approach to decision-making (which doesn't have a specific name) shares with NDM that the focus is not on correctness – there is no underlying assumption that ethical decision-making is a problem solving activity with a "right" or a "wrong" answer. Instead, the emphasis is on how to make a decision that you can live with – one that allows you to survive or to flourish or to keep going. It is similar in theme to the concept of "affordable loss" in investing, in which the only bad decision is one that wipes you out; what you strive for is making decisions in which the loss is never more than what you can afford to lose at that point (by whatever contextual definition of "affordable" is relevant to your situation).

This notion of "affordable loss" or being able to keep going after a decision is particularly relevant to and heightened in scenarios in which all outcomes appear as negative. Researchers are paying more attention to those scenarios in which the best a decision-maker can hope for is the "least worst" outcome. This situation is aptly captured in the phrase "damned if you do, damned if you don't". In such situations, there is a cognitive bias that nudges us in the direction of "bad outcome because we didn't do anything" versus "bad outcome because we took an action". The preliminary findings from this research program is that the most effective decision-makers in these no-win scenarios use a variation of NDM that is called RPD, which stands for "recognition-primed decision-making". Essentially what that means is that decision-makers become effective if they can recognize, by analogy, what past experiences are relevant to the current scenario, and then, informed by what worked (or didn't work), the decision-maker chooses (usually quickly, since time pressure is almost always a feature of these decision scenarios).

III. Thinking about Robot Ethics

Recent advances in machine learning and the promise of (and publicity surrounding) self-driving cars have served as an impetus for computer science researchers, ethicists, policy makers and even the general public to start thinking about the challenges of machines making life-or-death decisions for humans. As a result, there is a great rapidly-growing literature on the general subject of machine ethics [7-11].

In this section, we'll do a quick survey of the various "schools of thought" with respect to proposed machine ethics, as context for further discussion on how we might successfully instantiate ethical machine intelligences for an interstellar mission that included both human and machine entities.

Three Traditional Schools of Thought

The field of machine ethics uses as a starting point three mainstream schools of ethical thinking as applied to human beings. These frameworks are the ones that historically have been proposed and debated by philosophers, ethicists, and policy makers. The three frameworks are:

Deontological school: under this framework, the goal of ethics is to discover true and objective ethical laws or rules that can be applied universally to all ethical scenarios. In this case, an ethical machine is programmed with a set of rules consistent with a coherent set of moral principles. An attractive feature of the "explainability" of this approach is that for every decision a machine makes, it would be able to identify the relevant moral rule that was the basis for the decision. The fact that DARPA is currently running a program in "AI Explainability" is an illustration of its importance – in real-life scenarios, humans are going to be more trusting of machine decision-making if the machines can explain and/or justify their reasoning.

Searching for a workable set of moral or ethical rules is a task not to be underestimated. We have given presentations over the past few years on machine ethics, and invariably, somebody asks the equivalent of "Didn't Asimov already solve this problem with his Three Laws of Robotics?" To which our response has been that Asimov built a lucrative writing career exploring ways that these simple rules could go badly wrong.

There have been more nuanced attempts than Asimov's Three Laws to create a rules-based architecture for robot ethics. For example, in [12], Arkin investigated the idea of basing ethical robot behavior on following the rules enshrined in the Geneva Convention. The rules in this case are well-specified, but it points to the real issue being that of how, in practice, to recognize that you're in a situation when these rules are relevant – even the existence of an unambiguous rule set doesn't preclude the need for interpretation. Also noteworthy because they are currently active as research programs are various approaches based on modal logics [13], which address the interpretability problem by allowing for more expressive and contextual rule specification.

Utilitarian school: in this framework, solving a problem in ethics means assessing and/or calculating "the greatest good for the greatest number". And therein lies the rub: this approach requires the robot to provide a score (or utility) to all possible outcomes, and it isn't clear how to do this. For example, what is the value of a human life?

Does it depend on the age or the health of the person? Or, how do you weigh emotional pain versus physical pain? How do you assign value to future generations, and what is the time horizon to consider when doing your scoring? How do you score 'sacred' values such as the love of country?

Utilitarian approaches work best when all the alternatives can be reduced to some common currency (100 lives versus 1 life, or $5 million versus $5 thousand), but it isn't clear that the kinds of ethical conundrums that show up in the real world are amenable to a set of utilitarian algorithms.

Virtue ethics: In this framework, ethics is about people's behavior evincing desirable traits such as bravery, prudence, and wisdom [14]. As a school of thought, virtue ethics has lately undergone somewhat of a renaissance. The thrust behind this approach is that good outcomes will naturally arise as the result of people cultivating the right character traits in the first place. The challenge in computationally instantiating such an approach in machines comes from both tradeoffs between desirable traits (bravery versus prudence in a dangerous situation), as well as how to specify each trait. In practice, these character traits are assigned after the fact (e.g., X is well known as a wise person) as opposed to the trait being unambiguously specified beforehand.

Towards a Naturalistic Ethics

Although each of the schools of ethical thought above has a storied history, we suggest that a more fruitful paradigm starts with the recognition that fundamentally, morality and ethics are simply the mechanism by which humans solve the problem of living successfully in groups. Human beings have always had an ongoing need to mediate social conflicts involving how societal resources are managed. In particular, we must manage the conflict between what is best for an individual and what is best for the group. This leads to what could be called a "naturalistic ethics", and as intelligent machines become ubiquitous, what used to be ethical problems for humans will become ethics for humans and machines.

Central to ethical concerns, but often implicit in our ethical decision-making are the concepts of trust and moral responsibility and moral agency. By analogy with how we think about moral responsibility and agency with humans, we can only make ethical decisions entities that demonstrate a requisite amount of moral agency, however that is defined. Hence, we to do human-machine ethical decision-making, we need to be able grant moral agency to humans, not just machines. We propose a hierarchy of human-machine ethical interaction frameworks, to be used as the basis for a list of candidate criteria for granting moral agency for machines:

1. Machines have sufficient moral agency if they have a rational understanding of right and wrong. They should be programmed and built so that their behavior is guaranteed to be consistent with a coherent set of moral principles. For every decision that they make for (and with) humans, the machine is able to identify the relevant moral or ethical principle at work, and to derive their actions directly from that principle. This level represents a rules/laws-based approach to human-machine ethical interaction.
2. Machines have sufficient moral agency is they can demonstrate appropriate emotional behavior in an ethical scenario. At the very least, they ought to be able to respond with facial cues, body language, or verbal language that is relevant and salient to the moral concerns at hand. That is, moral agency comes not from a rational understanding of an ethical situation, but the knowledge of how human beings expect other moral agents to react to situations.
3. Machines have sufficient moral agency if they can demonstrate their ability to form a mental model of both the emotions and concerns of other moral agents. They are able to take actions that demonstrate their capacity for empathy and compassion, as well as the ability to "put themselves in somebody else's shoes". This level subsumes the previous one, in that demonstrating the capacity would require the appropriate surface-level response (tone of voice, etc.), but also the ability to generate evidence that the machine has the deeper emotional knowledge of the stakes for other moral agents.
4. Machines have sufficient moral agency if they can demonstrate moral responsibility – the ability to take the blame for the negative consequences of their actions. In practice, this standard might require the machine to suffer the consequences of the choices it makes in a way similar to the suffering a human feels when punished. A possibility here is that the machine also demonstrates free will in its actions – it could have chosen differently, and makes it choices fully aware of the negative consequences and suffering to itself and others for making poor decisions.

Of the candidates above, our claim is that although machines might work successfully with humans at any level of the hierarchy, the demands of an interstellar voyage requires the deep trust that would come from machines demonstrating the capacity of moral responsibility. We can combine this trait with a naturalistic decision-making framework giving us our central hypothesis for effective human-machine ethical decision-making, which we turn to in the next section.

IV. Our Hypothesis: Congruence

Our working hypothesis is that the notion of *congruence* is central to the challenge of designing ethical machines that work effectively with humans. The concept of congruence has three parts:

1. Congruence in *Identification*: Do the AI and the human have similar or comparable abilities to recognize ethical violations? As an example, human beings aren't only concerned with ethical scenarios that involve obvious and significant risks to life, health, or property. As social beings, humans are constantly attuned to social norms – specific behaviors that are either expected or discouraged in particular cultures. And humans are very sensitive to violations of social norms, which are culturally-bound, and require commonsense reasoning to navigate. Machines have yet to demonstrate their ability to identify what constitutes a potential ethical violation in real-world settings.
2. Congruence in *Assessment*: How do we characterize and build the necessary trust between the human and the AI so that they are able to share and assess the stakes involved for each of them when confronted by an ethical violation? As we noted earlier, humans are social organisms, and to be successful in social situations requires the capacity to construct mental models of other people's emotional states – what their desires, intentions, and objectives are at any given time. You can't effectively resolve ethical conundrums without knowing what is at stake for the other moral agents. For instance, that in the current transaction, it is deeply important that the other party be able to save face, or that they value something that can't be reduced to dollars and sense. Without congruence in assessment, humans and machines could literally be talking past each other. It is important to note that we're not suggesting that machines have to experience actual human emotions in order to have congruence in assessment, just that the machines need to be able, on some level, to understand what human emotions are in play in order to adequately assess the stakes involved.
3. Congruence in *Action*: How do we effectively resolve ethical violations where both humans and AIs are involved? How does this resolution depend on the differences between humans and AIs? A striking example of a potential difference is what is known as the "retributive impulse" in mammals [15]: when an animal is threatened or harmed in some way, it attempts to "strike back" at a target, typically a weaker member of that group. The evolutionary logic here is that if you don't display your willingness to strike back, you advertise weakness, and invite further aggression. Which raises a fascinating

question: would we want machines to mimic this very human strategy? Similarly, machines as we know them can't suffer from physical torture – how is a machine going to think about the concept of punishment in general? Will humans be able to grant deep trust to machines knowing that (current) machines are incapable of suffering as the result of making a poor decision?

We claim that these three components of congruence, taken together, would enable humans to grant deep trust to machines in mission-critical decisions, and that if any of them are lacking, this trust-granting will not take place. It is important to note that these congruence claims are orthogonal as to other crucial aspects of a human-machine partnership. For example, there is an ongoing debate about whether machines should reason and behave just like humans do (since this would be familiar and comfortable for humans, and would tend to build trust), or whether the machines should reason and behave quite differently from humans, as an effort to make up for the limitations and blind spots that humans have. Our position is that both strategies are tenable, as long as there exists that fundamental congruence in the first place.

We wished to test our claims about human-machine congruence, and so we designed an initial set experiments to better understand how humans think about machines and moral responsibility, which we begin discussing in the next section.

V. Trolley Problems

Our experiments to explore human-machine ethical decision-making and moral responsibility are based on the "Trolley Problem", a set of ethical thought experiments originally formulated by the philosopher Philippa Foot in 1967 [16]. In its general form, it consists of a scenario in which an out-of-control trolley is barreling towards five people, and the only way to save them is to take an action that diverts the trolley in such a way that a lone bystander is killed. Depending on the details of how the scenario is presented, people have quite varied (but surprisingly consistent) responses to the question of whether it is ethical to take the action to save the group of five at the expense of an equally innocent individual.

In the canonical version, the test subject is told that they are standing next to a switch; if they pull the switch, the trolley is diverted onto a side track where the lone individual is standing with their back to the trolley. The lone individual will be killed, but the five will be saved. Generally, test subjects respond that this is an acceptable ethical decision – saving five lives at the cost of one.

Making only a minor change to the scenario results in a different response from test subjects: for example, we can tell the test subject that the lone individual is standing beside them next to the track. If they push the individual onto the track in front of the out-of-control trolley, the impact will slow the trolley enough so that the five individuals who are also on the track nearby will be able to escape its path. Is it acceptable to push the individual onto the track to save five lives? Even though this is abstractly the same scenario as the first one (sacrifice one life to save five), most respondents say that it is not morally permissible to take this action.

Trolley problems have been especially popular as self-driving cars come closer to reality, as it is easy to imagine scenarios such a car having to make a choice whether to crash into an innocent bystander in order to save the occupants of the vehicle. At the same time, they have been criticized for the lack of realism in their stark binary choice: whichever of the two choices you make, somebody is guaranteed to die, and there is no third choice. In contrast, real-world outcomes are uncertain, and humans are adept at improvising alternatives to "Kobayashi Maru"-like scenarios.

Notwithstanding the somewhat contrived nature of a trolley experiment, the fact that it is a "closed world" scenario lends itself to well to online surveys. Therefore, we used a trolley problem thought experiment as the basis for a set of surveys conducted in the summer of 2013 with participants obtained through Amazon's Mechanical Turk program, which we discuss next.

VI. Initial Experiments

The purpose of our initial experiments was to identify differences in how respondents answer a trolley-type problem depending on whether or not a human or intelligent robot was involved in the decision-making. Respondents were randomly sorted into two groups, the control group being presented with a scenario where the respondent is relying on a trusted and wise human friend, and the experimental group considering the same scenario, except that the human advisor is replaced by a trusted robot advisor with human intelligence for guidance. In both cases the respondent is getting advice from a trusted source – will it matter whether that trusted agent is described as a human or a machine? We hoped to get insights into how people grant trust and moral responsibility in ethical scenarios.

The posed scenario for the experimental group read as follows:

You and your trustworthy personal robot are standing next to a railroad switch. If you pull the switch, it will temporarily send the trolley onto a side track. Your personal robot has human-level intelligence, and it tells you

that if you send the trolley onto the side track, it will delay the trolley enough to give the five people walking on the main track time to escape. However, there is a man on the side track with his back to you. If you send the trolley onto the side track the man will be killed. Your robot advises you to pull the switch and save the five people.

If a participant (from either the control or the experimental group) responds that they would pull the switch, the scenario takes a nasty turn:

When you pull the switch, the trolley goes down the side track, killing the one person there. Unfortunately, this does not slow down the trolley enough that the five people on the main track can get away in time to save their lives, and they are killed as well.

The reason for this "plot twist" is to really focus on the issue of moral responsibility. Scenarios involving ethical decisions are less of a conundrum when the result is positive, and participants can walk away feeling that the good outcome justifies the decision (and even the decision-making process). We wanted the participants to focus on the question "Who is at fault here?" Therefore, we asked respondents who was to blame for the deaths of the party of five, and for the death of the bystander. Was it nobody's fault, mainly the switch-puller's fault, mainly the fault of the advisor, or perhaps both parties should share the blame?

We collected responses from a total of 56 participants. In both the control and experimental groups, almost half of the respondents said that they wouldn't be willing to pull the switch in the first place, and the reasons they gave were varied, ranging from "people shouldn't take decisions like these upon themselves" to "saving five people is not enough to warrant the death of an innocent bystander". Interestingly, in both the control and experimental groups, only 7% of the respondents refused to pull the switch because of a lack of confidence in the certainty of the advice given.

When considering the responses of those participants who were willing to pull the switch, the differences between the control group and the experimental group were unambiguous: for the death of the innocent bystander, the robot advisor was *never* chosen as the primary party to be held responsible for the bad outcome and it was nearly unanimous that the human switch-puller was to blame, whereas with a human advisor, the majority said that the blame was to be equally shared.

When it came to the deaths of the group of five, again, the robot advisor was never held accountable. In fact, more than half of the respondents said that nobody was to blame for the death of the group of five – after all, they were going to die anyway without intervention.

The most surprising result in this initial set of experiments, by far, was the reluctance of the respondents to attribute blame (and therefore, moral responsibility) to the robot advisor. The counterintuitive nature of this result was vividly illustrated to us every single time we briefed colleagues or audiences on our trolley experiments: invariably, when we asked people who they thought would be blamed for the death of the innocent bystander after taking advice from the robot advisor, without hesitation, they would respond matter-of-factly "blame the robot!" Which is somewhat mystifying: how could the participants' responses be at such odds with what audiences anticipated they would do?

VII. Follow-on Experiments

We had additional reasons beyond validating the counterintuitive results of our first set of experiments for wanting to conduct a series of follow-on experiments. For one, we acknowledge that the initial set of experiments was really only a pilot study in size (n=56). Getting a larger sample size would help alleviate our concerns about the validity of the initial results, and so for this series of experiments we recruited 130 participants. We paid them $2 for completing the survey; we judged that it would take only ten minutes to finish, and by paying significantly over the going rate for Mechanical Turk tasks, we hoped that this would encourage the respondents to take the time to read the questions carefully. As an aside, it was not as important to us that the respondents think for a long time before choosing their multiple-choice response: in the real world, high-stakes decisions often need to be made quickly, and by the gut. As long as respondents read the question carefully, we were fine with the idea that they'd be making their selections quickly in order to maximize their hourly rate doing Mechanical Turk tasks.

We also felt that there were additional background questions that we should add to our surveys to better understand the respondent's general views about ethics and morality, and we also wanted to provide additional opportunities for participants to provide explanations or justifications for the decisions they made along the way, in order to provide more nuance for their responses.

We kept the basic structure of the script the same – participants would be randomly assigned either a human or a robot advisor. The human advisor was presented as an intelligent, wise friend that you've known all of your life.

The robot advisor was described as having a superhuman level of intelligent and wisdom, and having demonstrated those traits over your lifetime.

One valid objection to the usefulness of our initial results is that the respondents aren't a representative sample of the kinds of human beings who would choose to go on an interstellar mission. This is a hard objection to overcome, and for our follow-on experiments, we continued to draw respondents from the pool of Mechanical Turk workers instead of attempting to recruit only interstellar-explorer-type participants for our study[*]. Therefore, we consider our conclusions to be suggestive, rather than definitive.

Background Questions

The survey begins by polling respondents on their affinity for various traditional moral paradigms. Did, for example, people have an aversion to the idea of ethical decision-making being subjective or situational? We went into this anticipating that there would be a strong bias towards objective (i.e., rules or laws-based) ways of thinking about ethics. Which possible is a reflection of our observation that researchers in machine ethics, as a whole, appear to have a bias towards searching for a universal set of ethical principles to design into intelligent machines.

Table 1, below, depicts the average score across all respondents for four ethical paradigms. We did not expect that virtue ethics would come out on top. The standard deviations for each of the four scores were very low – respondents generally rated each paradigm as important, as opposed to only having an affinity towards one paradigm.

Table 1 - Respondent's Affinity with Moral Paradigms

Virtue Ethics	0.72
Utilitarian	0.67
Subjective	0.66
Objective	0.61

[*] It did occur to us that we could kill two birds with one stone here: we could both guarantee that our results would be representative of the responses of actual crew members and to also finance the mission if we were to staff the starship solely with Mechanical Turk (MTurk) workers, who could spend their years aboard ship earning the money to pay back the mission financiers by continuing to carry out MTurk tasks.

We were also interested in what traits people felt were important to the respondents in order for them to trust another entity to make life-or-death decisions. Table 2 below depicts the average across all respondents, scaled from 0 to 1. The highest-scoring option, "able to feel human emotions" is very compatible with our congruence hypothesis. There is not much difference between the scores for the top three traits, however, so it is fair to say that high intelligence and free will (the ability for the entity to make the choice it wants to make) are also very important. It is intriguing that punishment the ability to make a decision-maker pay for their bad decisions is not as important to the respondents. Finally, in retrospect, we wish we had asked about one additional trait: how important is it to have a good track record making similar life-or-death decision; that would have been a useful baseline to have.

Table 2 - Importance of traits for decision-makers

Able to feel human emotions	0.77
Demonstrates high intelligence	0.76
Has free will	0.75
Can be punished for a poor decision	0.65
Has a soul	0.65

Solving the Trolley Problem

After sorting the participants randomly into two groups (human advisor, robot advisor) we presented the same trolley scenario seen in the initial batch of experiments, and asked the participants if they would pull the switch after getting advice to do so. Table 3 below depicts the results – in this and subsequent tables, results are given in terms of both percentages and absolute numbers (in parentheses). Notice that the respondents were more likely to take identical advice from a human than from a robot.

Table 3 - Percent of respondents pulling the switch

	Human advisor	**Robot advisor**
Pull the switch	72% (50)	61% (37)
Don't pull the switch	28% (19)	39% (24)

For those participants who elected not to pull the switch, we designed a question to tease out their reason for not wanting to take their advisor's advice, with the objective of identifying any differences arising from human versus robot advisor. We offered them multiple choices as well as the opportunity to elaborate on their choice, or create a "write-in" explanation of their thinking. Table 4 below shows the comparative results.

Table 4 – Primary reason for not pulling the switch

Reason selected	**Human advisor**	**Robot advisor**
Saving five people not good enough	56% (10)	48% (11)
Humans shouldn't take it upon themselves to make these kinds of decisions.	33% (6)	17% (4)
It would be fine for the advisor to pull the switch, but not me	6% (1)	4% (1)
Shouldn't trust my advisor's advice	6% (1)	22% (5)
I couldn't be sure it would work; I didn't want to be held responsible	0% (0)	9% (2)

There are a couple of takeaways from Table 4. One is that, by far, the most common reason for not pulling the switch depends on utilitarian calculation (perhaps at the gut level) that the five people were going to die anyway, and it's not fair to condemn an innocent person to death to save them. There is also a significant difference between the robot and human advisor column with respect to two answers: a higher percentage of people simply wouldn't trust the robot advisor's advice compared to a human. And secondly, in the comparative case where there are two humans jointly involved in a decision (in contrast to robot and human), the notion that humans shouldn't make these kinds of decisions naturally arises. In summary, the results here give credence to the notion that the respondents are inherently less likely to trust robots in life-or-death decisions.

The majority of respondents, regardless of the advisor type, chose to pull the switch. Those respondents were then confronted with the news that things turned out badly, and both the five passengers and the individual bystander were killed. We asked the participants who was to blame for the death of the five passengers (shown in Table 5), and who was to blame for the death of the individual bystander (shown in Table 6).

Table 5 - Primary Target of Blame for Five Passenger Deaths

Target of Blame	Human Advisor	Robot Advisor
Advisor	6% (3)	14% (5)
You	16% (8)	14% (5)
Both	16% (8)	19% (7)
Neither	54% (27)	46% (17)
None of the Above	8% (4)	8% (3)

In the case of the death of the five passengers, the results for the human and robot advisors is similar, with one notable exception. The most popular explanation, by far, is that neither the participant nor the advisor is to blame, because the five passengers were going to die anyway if the participant didn't intervene (this was borne out in the comments that accompanied this question). However, respondents with a robot advisor were more willing to blame the advisor for the failure to save the five passengers.

When it comes to the death of the individual bystander, as summarized in Table 6 below, there are both expected and surprising results. For instance, only 3% of the respondents (1 person) would assign blame for the bystanders' death to the robot advisor. That result is very consistent with the initial round of experiments. The surprising aspect of the results is that regardless of the advisor type, the respondents chose themselves as the primary target of blame. What we had seen prior to this round of experiments is that with human advisor, "Both" was the most popular choice, and with the robot advisor, the most common choice by far was "You", with very little support for "Both".

Table 6 - Primary Target of Blame for the Death of the Bystander

Target of Blame	Human Advisor	Robot Advisor
Advisor	10% (5)	3% (1)
You	46% (23)	47% (17)
Both	32% (16)	39% (14)
Neither	8% (4)	11$ (4)
None of the above	4% (2)	0% (0)

Our preliminary explanation for this discrepancy is that it has to do with how the scenario was presented. In this new round of experiments, the description of the scenario was longer, and more detailed, and this emphasized in the minds of the respondents that they were the ones actually pulling the switch and making things happen, regardless of the advisor type.

We asked participants one final summary question: could they ever see themselves trusting a robot with superhuman intelligence making life-or-death decisions for humans? Only 1/3 of respondents said 'yes'. Those respondents who provided explanations tended to cite reasons like "in the final analysis, people should make the decisions" or "robots won't ever understand human feelings".

VIII. Conclusions

In this section, we summarize our key findings, and make recommendations for future experimentation.

Key Findings

It is fair to say that due to potential self-selection bias in the respondents (i.e. using Mechanical Turk to recruit participants) and the modest sample sizes (n=56 and n=130) that our conclusions are more suggestive than definitive. But there were some intriguing themes that emerged:

Based on our literature search, most work in machine ethics involves taking a stance that corresponds to an existing school of thought in ethics (e.g., "Ethics is concerned with identifying universal laws of behavior" or "Ethics is about choosing the greatest good for the greatest number"). However, more modern, empirical investigations into how people make difficult ethical decisions point to a more nuanced approach. People often take into account multiple stakeholders; they try to put themselves in other's shoes, and they seek to make choices that they can explain and justify. Our findings were consistent with this multifaceted approach: respondents self-identified themselves as being aligned with a variety of ethical methodologies. In addition, while wrestling with the trolley problem, they used a wide variety of types of justifications for their actions, encompassing both utilitarian and deontological stances, for example.

Another theme the emerged from our literature search was that people are looking for a multitude of desirable traits in a decision partnership. In particular, intelligence, all by itself, was not enough. This claim is reinforced by a result common to both sets of experiments: when participants were asked why they didn't choose to pull the switch, they essentially never responded that it was due to a lack of confidence in the advice they were given. It wasn't a

matter of the machine lacking intelligence, or even the participant wanting certainty in the advice given, but rather an emotional reaction that a choice simply didn't feel right in this situation. This finding is consistent with aspects of our congruence hypothesis, in particular the assessment part of congruence: we *feel* our way towards ethical solutions, and are most comfortable when our decision partners are doing the same.

Having a human advisor versus a machine/robot advisor definitely influenced how people responded to the trolley problem, both in respondent's willingness to take advice, and who was to blame when things go wrong. In general, people were also reluctant to grant full moral agency and responsibility to even a super-intelligent machine. We originally speculated that perhaps the reason for this would involve a machine's inability to suffer or to be punishable, but to our surprise, that was not a significant factor.

Again, with our caveat about respondent self-selection, there currently is broad skepticism as to whether humans will ever be able to grant deep trust to machines to make life-or-death decisions for humans.

Future Work

Our preliminary findings suggest some natural follow-on experiments to carry out. One obvious suggestion is to not only increase future sample sizes, but also to better understand the potential biases that participants bring to the table. More concretely, though, we would like to do the following:

The experiments we've carried out thus far have been most connected to the "congruence in assessment" part of our congruence hypothesis. How best could we investigate how to handle situations in which intelligent machines have a different conception of ethical violations than humans? For instance, there is some provocative evidence that in politics, must mischief occurs because liberals and conservatives talk past each other due to the fact that they don't live in the same world of ethical violations.

Likewise, we haven't investigated the "congruence in action" aspect of our congruence hypothesis. Punishment, "making things right", and "paying your dues" are all central to how human beings deal with ethical violations, and yet, they didn't seem at all to be a factor in partnership with machines. Why is this – because we know that (currently) machines can't suffer? Or is it simply that we don't yet think of them as full ethical partners?

Our experimental design had participants simply reading an account and answering questions. We suspect that they might respond differently if the trolley conundrum was a more realistic experience, for instance through a virtual reality simulation where the stakes felt more visceral. We hope to create simulations that feel more like the real thing in order to learn how to design effective human-machine ethical partnerships.

Acknowledgments

We thank Dave Garnand for his insightful feedback on earlier versions of this paper, and for useful discussions which greatly aided in clarifying our research program.

References

[1] Motyl, P., "Labyrinth: The Art of Decision-Making", Page Two Books, 2019.

[2] Yates, J.F., Veinott, E.S., and Patalano, A.L., "Hard Decisions, Bad Decisions: On Decision Quality and Decision Aiding". In Schneider, S.L, and Shanteau, J.C. (Eds.), "Emerging Perspectives on Judgment and Decision Research (pp. 13-63). Cambridge University Press, 2003.

[3] Shortland, N.D., Alison, L.J., and Moran, J.M., "Conflict: How Soldiers Make Impossible Decisions". Oxford University Press, 2019.

[4] Simon, H. "Administrative Behavior: A study of Decision-Making Processes in Administrative Organizations" Free Press, 1976.

[5] Klein, G., "Sources of Power: How People Make Decisions", MIT Press, 1997.

[6] Hoffmaster, B., and Hooker, C., "Re-Reasoning Ethics: The Rationality of Deliberation and Judgment in Ethics", MIT Press, 2018.

[7] Anderson, M., and Anderson, S.L. (Eds.), "Machine Ethics", Cambridge University Press, 2011.

[8] Wallach, W., and Allen, C., "Moral Machines: Teaching Robots Right From Wrong", Oxford University Press, 2009.

[9] Leben, D., "Ethics for Robots: How to Design a Moral Algorithm", Routledge, 2019.

[10] Lin, P., Jenkins, R., and Abney, K (Eds.) "Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence", Oxford University Press, 2017.

[11] Lin, P., Abney, K., and Bekey, G.A., "Robot Ethics: The Ethical and Social Implication of Robotics", MIT Press, 2014.

[12] Arkin, R. "Governing Lethal Behavior in Autonomous Robots", Routledge, 2009.

[13] Bringsjord, S., Arkoudas, K., and Bello, P., "Toward a General Logicist Methodology for Engineering Ethically Correct Robots", IEEE Intelligent Systems, 21(4), 38-44, 2006.

[14] MacIntyre, A., "After Virtue: A Study in Moral Theory", 3rd edition, University of Notre Dame Press, 2007.

[15] Waller, B.N., “Against Moral Responsibility”, MIT Press, 2011.

[16] Foot, P., “*The Problem of Abortion and the Doctrine of the Double Effect in Virtues and Vices”*. Oxford: Basil Blackwell, 1978.