Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Meghan Hurley April 11, 2021

An assessment tool for the public opinion of the moral status of Artificial Intelligence

by

Meghan Hurley

Gillian Hue, Ph.D.
Adviser

Neuroscience and Behavioral Biology

Gillian Hue, Ph.D.

Adviser

Aubrey Kelly, Ph.D.

Committee Member

Jeff Mullis, Ph.D.

Committee Member

2021

An assessment tool for the public opinion of the moral status of Artificial Intelligence

by

Meghan Hurley

Gillian Hue, Ph.D.

Adviser

An abstract of

a thesis submitted to the Faculty of Emory College of Arts and Sciences

of Emory University in partial fulfillment

of the requirements of the degree of

Bachelor of Science with Honors

Neuroscience and Behavioral Biology

2021

Abstract

An assessment tool for the public opinion of the moral status of Artificial Intelligence

By Meghan Hurley

Artificial intelligence's rapidly progressing ability to plan actions and integrate and process information in a manner similar to humans, alongside an increasingly anthropomorphic conceptualization of AI's underlying mechanisms has led experts in the fields of neuroscience, engineering, computer science, and philosophy to question whether or not AI has the ability to become conscious or sentient. Unclear criteria for defining consciousness and sentience as well as unclear criteria for determining moral status for humans, let alone non-human beings and entities makes it even more difficult to predict how AI with human-like abilities may interact with and function alongside humanity, or process how we will decide their moral and thus legal status in society. Considering that the public's attitude and acceptance towards conscious AI will play a large role in deciding how AI are treated and whether or not they are respected as members of society, it is imperative to understand the public's current opinions and perspective of the moral status of AI. This thesis aims to develop a robust assessment tool that can be used to neuroethically examine factors and themes crucial to the public's opinion of the moral status of AI. The assessment tool was evaluated by conducting qualitative interviews with participants who offered insight into the clarity and effectiveness of the tool. A neuroethical analysis of the projected and emergent themes from the interviews influenced the refinement of the scenarios on the administered tool.

An assessment tool for the public opinion of the moral status of Artificial Intelligence

by

Meghan Hurley

Gillian Hue, Ph.D.

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences

of Emory University in partial fulfillment

of the requirements of the degree of

Bachelor of Science with Honors

Neuroscience and Behavioral Biology

2021

Acknowledgements

I would firstly like to thank Dr. Gillian Hue for her unwavering support throughout this process. Her constant reminder that my thoughts and insights are valuable, even as an undergraduate student, helped me to find my voice as an aspiring neuroethicist.

I would also like to thank my committee members, Dr. Aubrey Kelly and Dr. Jeff Mullis for taking the time to provide me with their insight and feedback.

I would like to finally thank Jennifer Jin, my fellow labmate, for sharing her interests in neuroethics with me so that we could learn and improve together.

Table of Contents

Background and Significance of Project

The problem

Advancements in artificial intelligence (AI) have reached a critical point where discussions about consciousness and sentience in AI are unavoidable; for some scientists, it is less a question of whether AI *can* achieve consciousness and more a question of *when* they will (Long and Kelley, 2010). For both experts in relevant fields as well as the general public, imagining a future such as this can be daunting; plenty of questions in AI research have yet to be answered, and as machines begin to learn and process information in more complex ways, our understanding of exactly *how* they work lessens, leading us to guess at their inner workings (Bleicher, 2017). Moreover, if AI does appear to develop abilities associated with consciousness and sentience, how can we discern that they are genuinely experiencing these phenomena? How can we truly determine that their experiences are analogous to those of human beings? Is there any way for AI to fake or mimic these experiences so that they appear to relate to us? Considering the answers to these and other questions regarding the ethical, legal, and social (ELSI) implications of conscious AI will be essential in deciding how these beings should be treated in our future society (Yesley, 2008). Specifically, we must consider what makes beings moral agents deserving of legal protections and rights, and establish an ethical tool set for examining this novel issue as it pertains to AI.

Current AI scholarship and the future of AI

Throughout the last two decades, the field of AI has adopted an anthropomorphic conceptualization of AI, with much of the key terminology in the field relating to the human brain and human anatomy, e.g. neural networks, input/output channels, and cognitive architectures (Long and Kelley, 2010; Signorelli, 2018). AI scientists have also seen machines becoming more humanlike in the way they *integrate* sensory information, *learn* from their own

mistakes, and even *plan* their actions with the intention of various subgoals and supergoals, a manner that is similar to humans (Bostrom and Yudkowsky, 2001). As early as 2010, academic scholars were already researching and writing about the possibility of developing conscious systems, explaining that the brain can be simplified to a "complex nonlinear system of systems" (Long and Kelley, 2010), one that is capable of being synthetically built by engineers of the time. One survey conducted in 2013 shows that AI experts agree with this sentiment, with most arguing that superintelligent[1] AI has a 1 in 2 chance of being developed around 2040-2050, and a 9 in 10 chance by 2075 (Müller and Bostrom, 2016). As the complexity of AI design increases with scientists' understanding of neural network algorithms, some AI scientists believe that consciousness in future AI may simply result as a nascent behavior with enough processing power, sensory input, and robust learning (Long and Kelley, 2010). While scientists are able to make these predictions about when this change might occur, it is much more difficult for them to predict what exactly this consciousness will look like in AI and whether it will resemble the only other model of consciousness that we have— our own. At the moment, researchers are attempting to better understand how humans make moral decisions by gathering data on our behavior in response to moral dilemmas (Awad et al., 2018). Although researchers hope that these data will apply largely to AI, there is no certainty as to whether AI's consciousness will be accompanied by an ability to judge moral dilemmas and make moral decisions in the way that human beings do. Because of this uncertainty, it is crucial to begin considering the value that our society and legal system place on moral agency, and how AI would be defined in public opinion and under the current legal system of rights and protections.

[1] For the purpose of this study, superintelligence can be defined as an intellect smarter than that of humans in every possible field, e.g. science, creativity, general wisdom and knowledge, and even social skills (Bostrom, 2006).

What gives beings moral status?

Deciding the moral status of AI and beings other than humans would be simpler if there was a standard set of criteria for those worthy of moral status. Of course, there is no such standard, no clear consensus as to why exactly human beings have been afforded moral status and are regarded as moral agents, and much debate around who gets to determine who or what is worthy of this status. Some philosophers have attempted to define an entity as having moral status "if and only if it or its interests morally matter to some degree for the entity's own sake" (Jaworska and Tannenbaum, 2021). While humans are likely the first beings that come to mind when discussing moral status, even they have had a complex relationship with moral status; at some points in history, individuals labeled as "other" like foreigners (Booth, 1997), enslaved persons (Lindsay, 2005) and the physically disabled (Ralston and Ho, 2007) were not given full moral status, and the debate today about cognitively impaired individuals (Wasserman, Asch, Blusein, & Putnam, 2019) continues. Furthermore, the question as to which non-human beings have moral status is also of concern. In some cases, the fair or ethical treatment of animals such as livestock or wild animals arises in part based on moral status, and in other cases it may even be the discovery of cognitive abilities in animals that make ethicists and animal advocates question their moral status and what consideration they deserve to receive from society (Jaworska and Tannenbaum, 2021). Because of these ambiguities, many ethical questions remain unanswered regarding moral status and by extension, whether or not it will apply in the future to AI. If moral status is extremely dependent on context and we do not have a clear standard of moral status in humans or even animals, how do we expect it to pertain to AI? Who gets to decide whether or not AI have moral status? Does being defined as a being with moral status grant or ensure that that being will have legal rights and protections? If not, what does? In order

to examine some of these questions, an exploration of the factors that may contribute to moral status is necessary.

Substrate

One distinction that may be made between those with moral status and those without it may be the substrate with which they are made. More biocentric, or non-anthropocentric approaches to moral status attribute moral status, though not equal moral status, to all living organisms; this approach argues that it is the organic and physical nature of human and animal bodies that make us living, breathing, entities worthy of moral status (Wetlesen, 1999; Jakobsen, 2016). Anthropocentric approaches to moral status, on the other hand, hold that only human beings have moral value (Jakobsen, 2016). With the rise of AI and robotics, academics have been forced to discuss the moral standing of computers and technological devices that can already exceed human calculation abilities (Signorelli, 2018) and appear to have information processing similar to humans, but that are made up of computer parts and mechanical bits and pieces. Amplifying the issue even further is the creation of AI and intelligent machines that are humanoid (Adams, Breazeal, Brooks, & Scassellati, 2000); with faces and bodies eerily similar to those of humans (Wang, Lilienfeld, & Rochat, 2015) but with hidden mechanics underneath, the question of how substrate fits into the definition of moral status has become complicated.

With the prevailing anthropocentric approach to moral status, one may assume that AI have been automatically deemed as unworthy, but some scientists recently have considered the possibility that substrate makes actually little difference in the ability of a being to have moral status. Two such scientists, Nick Bostrom and Eliezer Yudkowsky offer that regardless of the fact that AI have a different substrate than humans, they have equal moral status as humans and ultimately, deserve similar, albeit not identical, rights and protections as humans do (Bostrom &

Yudkowsky, 2014). Pennartz, Farisco, and Evers (2019) echoes this point, positing that the nature of consciousness itself depends on what kind of "hardware" an entity utilizes; the biological neural networks of the human brain are certainly still different than the neural networks we see in non-living machines. This dilemma makes it difficult to determine whether consciousness would even look similar between the two and whether or not the substrate is significantly contributing to this variance. It may be the case that AI's mechanical neural network substrate functions to create a version of consciousness that is at least somewhat comparable to human consciousness. If this is the case, it may be that AI are similar enough to adapt some of our rights and protections or possibly some derivation of them, even if they are not as extensive as ours.

Consciousness

Another, and arguably more important, factor that scientists have attempted to use to discern what makes humans special as moral agents is their ability to experience consciousness. With disagreement between fields as to what exactly the abstract concept of consciousness even entails, some scientists, like Scott Aaronson at the University of Texas, believe that "humans seem nowhere close to solving it" (Ball, 2019). How scientists choose to define consciousness in the future also affects whether or not an AI's conscious experience is similar enough to the human conscious experience to warrant them the same protections as humans are granted. According to Long and Kelley (2010), philosophers, psychologists, neuroscientists, computer scientists, and AI researchers have all developed different views of what makes a being conscious. One of the biggest difficulties of pinning down this concept is that consciousness is really a subjective experience that can't be empirically studied unless the individual experiencing said consciousness is asked to detail their experience.

In recent years, the scientific literature has experienced an expansion in the number of theories of consciousness. Because experts in a variety of fields such as neuroscience, philosophy, engineering, and computer science seem to have competing theories and a varied understanding of consciousness, each advancement in the field appears to bring about more questions than answers. No matter what, definitions and conceptualizations of human consciousness do not necessarily apply to non-verbal or non-human beings, yet these beings still need to be factored into these theories, as some animals and other beings have been shown to exhibit at least some aspects of or certain lower levels of consciousness as we currently understand it. One theory of consciousness called the global workspace theory (GWT), posits that consciousness is created by a "global workspace," a collection of specialized computational processes supplied with a unique memory, that holds pieces of information in the brain and can broadcast this information out to other brain areas associated with specific tasks (Prakash et al., 2008). This approach to consciousness implies that it is a type of computation that can motivate and guide our actions; once information is gathered and made widely available to the rest of the brain, consciousness occurs (Ball, 2019). It is exactly the act of referring to consciousness, a supposedly unique human ability, as mere computation that puts so many people on edge and begs the question: what makes human consciousness any different than the algorithms used by computers and AI? Some scientists believe that according to this view, machines are likely to soon model most of the abilities of the human brain and thereby will be conscious by this definition (Ball, 2019).

One research team in particular has put forth a handful of operational criteria or indicators that can be used to attribute various levels of consciousness to non-verbal beings (human and non-human) that exhibit varying states of information processing (Pennartz, Farisco,

& Evers, 2019). The paper notes that defining consciousness in intelligent machines, regardless of their proposed criteria, is still a complex and difficult task for multiple reasons. The actual processors that intelligent machines rely on are much faster than human processors, neurons, which are limited because of their biological structure. Moreover, intelligent machines may simply be pre-programmed to display the consciousness-like behaviors that we seek to identify in conscious beings; whether being pre-programmed to display a behavior is similar enough to be grouped amongst human consciousness is still up for debate. Regardless, Pennartz, Farisco, & Evers (2019) emphasizes that even if an entity fails to meet their designated criteria, that being might still be conscious, as assigning consciousness is a very complex task, and a lack of tangible evidence is not enough to rule out the possibility. In the case of AI, it may be that just because they have not yet met the operational criteria developed based on our current conceptualization of AI, does not mean that they are not conscious, or cannot achieve consciousness in the future (Pennartz, Farisco, & Evers, 2019). Alternatively, it may be that our current definition or term for consciousness simply doesn't apply to AI or machines perfectly, but that they do exhibit a sort of consciousness that is inherently different because of their substrate and constitutive nature; if this is the case, our current definition may actually be hindering us from detecting, understanding, or labeling consciousness in other entities−—a task that could propel consciousness research forward.

Ultimately, Pennartz, Farisco, & Evers (2019) highlights the importance of raising the question of intelligent machine consciousness and why the present research study is so salient: if they are capable, we are immediately tasked with discussing whether or not they are entitled to rights and protections in the legal system and the attribution of moral status.

Personhood and animal issues as precedence

A final commonly recognized criterion for moral status is the capacity for personhood, otherwise known as sapience. Sapience can be defined as a "set of capacities associated with higher intelligence, such as self-awareness and being a reason-responsive agent" (Bostrom & Yudkowsky, 2014) and is commonly cited when settling disputes about the proper treatment or moral status of entities whose status is in question (DeGrazia, 2010). For those who argue that only human beings have personhood, non-human animals are granted a lower moral status regardless of their subjective perceptions of the world. Contrary to this, however, is plenty of scientific evidence that points to self-awareness and autonomy in certain animals such as elephants, dolphins, and great apes (Morrison & Reiss, 2018; DeGrazia, 2010; Plotnik, de Waal & Reiss, 2006; Gallup, 1979).

What makes personhood so germane is its direct application to the legal system; while the qualities of self-awareness and autonomy appear to be sufficient for humans to be recognized as having personhood under common law, the same cannot be said for animals and other non-human entities (The Nonhuman Rights Project (NhRP), 2021). In the eyes of current law, animals are still considered "things" and not "persons" ("Animals' Legal Status," 2020), however they have still earned certain protections throughout the years. As of now, animals are protected under the Animal Welfare Act and the Humane Slaughter Act which provide guidelines for the ethical treatment of research animals and set standards for farm animals, respectively. Although this is the case, some activists are continuing to fight for increased protections for animals. The Nonhuman Rights Project (NhRP) specifically works to combat this status quo narrative that animals and non-human entities should be referred to as "things" with no rights, through litigation, legislation, and advocacy. They highlight on their website that in the

past, common law has been utilized in cases that evolve our standards of morality and understanding of the human experience, and the legal teams at NhRP continue to use this strategy to change the standards of how we treat nonhuman animals as a society (NhRP, 2021). In one memorable case that the NhRP brought in 2015, researchers attempted to earn habeas corpus for two research chimpanzees who they believe had been unlawfully contained and deserved the ability to challenge this detainment (Hu, 2015). In cases such as this one, activists believe that they must push back on the current legal status of animals because the current legal system isn't equipped to incorporate emerging knowledge about animal cognition and personhood.

Similarly, the current laws and regulations actually fail to protect animals in some cases based on certain characteristics of the animal and very restrictive wording utilized in the laws (Hamlett, 2020). The use of octopuses in research labs is currently on the rise, for example, due to their failure to meet the required legal definition for "animals". Despite the fact that octopuses are known for being sensitive and intelligent creatures, they are not extended the same legal protections as other animals simply because they are invertebrates rather than vertebrates. Cases such as this one raise important questions about the legal status of animals and the effectiveness of their protections. How do we actually assess which animals are deserving of welfare and protection from harm? Should characteristics such as intelligence, traits that we usually attribute to humans, play a larger role in which animals have more protections? Why is animal consciousness handled differently under the law than human consciousness? If octopuses don't count as animals under the legal definition, how will we possibly define AI? While the legal status of animals can help to guide the discussion about AI's moral status, it is certainly not perfect, and AI will pose legal difficulties that are ultimately unprecedented.

While this precedence of limited or absent assignment of moral agency to non-human beings might seem daunting, it offers us a fairly remarkable opportunity: we find ourselves at a pivotal point, where we can either choose to embrace AI as moral agents and transform the way that we view the moral status of non-human beings, or we can continue to live conservatively with our preconceived notions about "inferior" non-human beings and our loopholes for avoiding the actual protection and welfare of these entities. To add to this difficult decision, legal change is certainly not an easy or quick process ("Overcoming lawyers' resistance," n.d.). As of right now, human rights are protected legislatively at the federal level in the United States, by both federal legislation as well as the Constitution. The way in which our law is currently structured focuses solely on the human experience, and when other beings do not fit perfectly into the existing definitions of a "human," they are deemed undeserving of rights similar to those that humans have. Nevertheless, as Eric Johnson states in an article for *Scientific American* in 2012, "the law is fully capable of making and unmaking 'persons' in the strictly legal sense," (Johnson, 2012). The future may consist of a reevaluation of what level of cognitive ability warrants protections for entities and may result in the making of new "persons." It is important to note that although a potential future with conscious AI is not decidedly inevitable and could be decades away, these conversations are absolutely worth having this early on; deciding to change our legal structures and standards would be a long and tedious process that should be initiated soon if we hope to keep up with the rapidly advancing abilities of intelligent AI.

The importance of the public

Moving forward, it is important for scientists, ethicists, and philosophers to consider whether morality and a sense of self are defining features in our judicial system and society; if these phenomena give us the moral status and the right to protections under the law, then it could

be argued that all beings who share these abilities also have the same status and protections. It will be pertinent for these individuals to look to animal ethics for possible precedent in how we have handled beings other than humans with complicated and complex levels of agency and cognitive ability. While much of the decision-making regarding AI and their moral status in the legal system will be conducted by experts and academics, it will be equally pertinent in the near future to understand the public's opinions and acceptance of such a potential future. A recent study regarding the similarly controversial topic of neuroenhancement has revealed several factors that may influence the public's comfortability with, and overall acceptance of their future usage (Conrad, Humphries, and Chatterjee, 2019). Authors of this study even suggest that "results of this survey inform potential [Cognitive Enhancement] CE policy in multiple ways" (Conrad, Humphries, and Chatterjee, 2019); similarly, by investigating public attitudes towards AI's moral status in the legal system, it may be possible to influence and inform future policy regarding the personhood and legal rights of AI.

Possible influencing factors

In this recent study assessing public acceptance of cognitive enhancement (Conrad, Humphries, and Chatterjee, 2019), the authors emphasized the fact that opinions about CE may be influenced by the context of the situation in which CE is used. Several of the authors on this paper then published a second paper exploring even more ways to phrase situations in which CE may be used to properly investigate the multiplicitous factors that influence public opinion of CE usage (Dinh, Humphries, & Chatterjee, 2020). Similarly, given the controversial nature of AI and the fact that a future with conscious AI that have identical legal rights to humans sounds straight out of a science fiction novel, it will be important to explore the various factors that may be influencing the public's opinion of AI gaining moral status in the legal system.

Media portrayal

The CE study highlights the importance of phrasing and wording in the way that the public interprets issues, and this is especially pertinent when it comes to news about scientific findings and progress. Whereas individuals with a scientific background get their information about the progress of AI through scientific articles and research, the general public most likely gets information from the news and journals in which the authors of the articles may not have prior background knowledge or may dramatize findings with flashy titles and headlines. Because of this, scientific information can get misconstrued to the public as being more intimidating or dangerous than it actually is, or medications and technologies can be portrayed as more effective or "groundbreaking" than they actually are (Lilienfeld et al., 2018). AI specifically often falls prey to this over exaggeration of danger, as can be seen in newspaper articles, TV series, and movies in which AI is portrayed as murderous, blood-thirsty, out-of-control, unsettling, manipulative, and overall unsafe. The way in which AI is portrayed on the news, and in newspaper articles, books, TV shows, movies, and creative art will be explored to provide insight into the public's views of AI.

Despite the importance of understanding public acceptance of AI and its potential moral status and deserved rights, the scientific literature currently lacks a proper tool to collect and analyze data regarding this question. While some articles addressing neuro-related topics have begun to explore public understanding and acceptance of neuroscience technology, research, and its ethical implications (Conrad, Humphries, and Chatterjee, 2019; Illes and Bird, 2006), and AI research has also begun to explore public acceptance of novel AI technologies (Talley 2020), none have focused specifically on this intersection of neuroscience, AI, and moral status.

Goals and Objectives Stated

This thesis aims to bridge the gap between research into a potential future with AI as moral agents and public acceptance and understanding of this potential future by developing and evaluating a tool to robustly assess factors and themes that may be influencing public acceptance and understanding of AI potentially gaining access to rights and protections via their moral status.

Research Questions

The goal of this research is ultimately to answer several questions about AI and moral status as they are interpreted by the public through the use of an assessment tool. Some questions that guided the creation of the research tool include:

- What is the public's current perception of the function of AI (what AI is created for or does for humans)?
- What is the public's current understanding of how AI works (the computational mechanisms that underlie AI)?
- What abilities are attributed to AI by the public?
- How does the public define morality and moral status?
- **What factors are essential to the public's conception of morality and moral status of AI?**
- **Does the public feel positively or negatively toward AI?**
- **How would the public treat AI in various situations?**
- **What abilities does an entity need to have in order for people to deem it worthy of protection / praise or responsibility / liability?**

While several of these questions were utilized simply to guide the creation of the assessment tool, some of them are explored more directly in the research study. Questions regarding the public's definitions of morality and moral status as well as the types of abilities that the public attributes to AI will be addressed in the semi-structured portion of the pilot study, whereas questions that are bolded above are addressed by the assessment tool directly.

Methodology

Pilot Study Organization and Data Collection

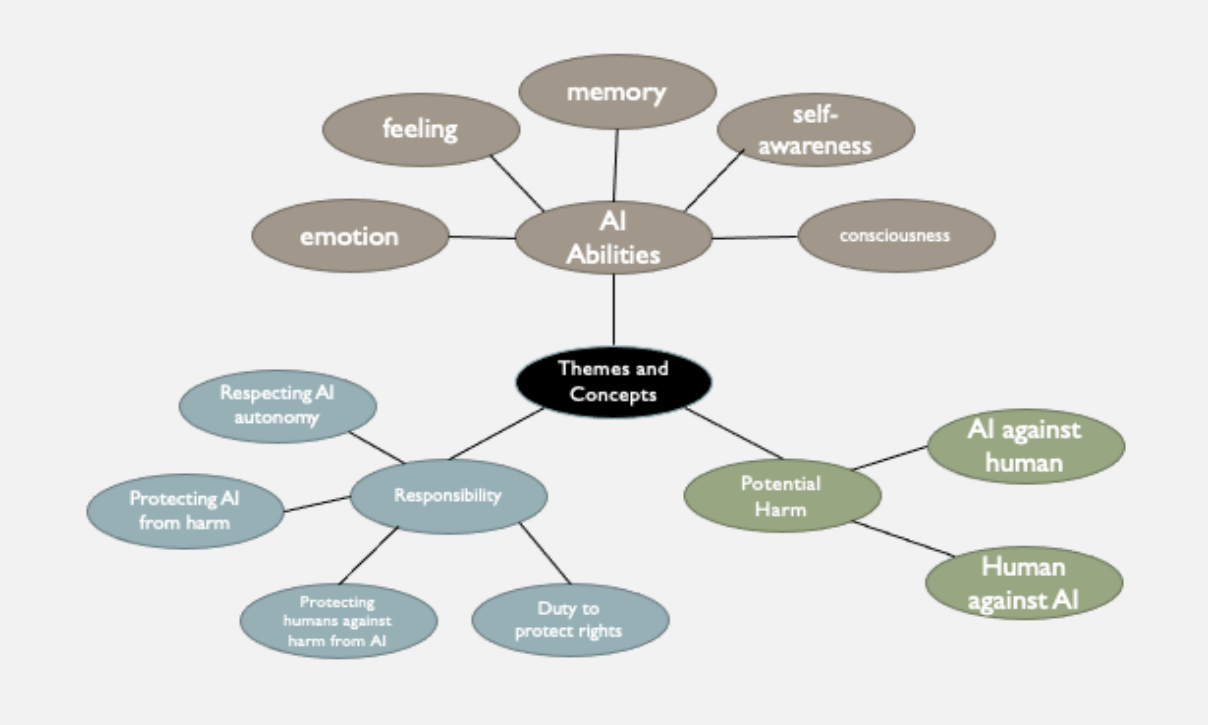
In order to create a robust and functional tool for assessing public opinion of AI's moral status and agency, a pilot study was conducted to develop and evaluate appropriate scenarios and questions for the assessment. As the goal of this research is to understand how the public perceives and makes moral judgments about AI in various scenarios, it is imperative to ensure that the questions in the tool elicit appropriate and authentic responses from the participants. As a way to combat the limitations of participant self-reporting of cognitive processes and activate schema responses on a more tacit level (Thoma & Dong, 2014), several recent papers have utilized vignettes to assess concepts like moral judgment and moral reasoning.

The Defining Issues Test (DIT), a model of moral development created in 1974 and recently updated (DIT2), employs several moral dilemmas or stories that participants respond to via issue statements with which they must rate their agreement (Thoma & Dong, 2014). In using the DIT, researchers are interested in knowing which factors respondents find to be important when resolving specific moral dilemmas (Abu-Odeh et al., 2015), an element that will also be important in the current pilot study. Similarly, Conrad, Humphries, and Chatterjee (2019) employed vignettes to determine the effect of framing metaphors and context on public

acceptance of the controversial topic of cognitive enhancement. This study utilized a similar measurement system as the DIT and DIT 2 in which respondents rated their level of acceptability of the use of cognitive enhancement seen in each vignette using a Likert scale (Conrad, Humphries, & Chatterjee, 2019).

In line with this strategy of using vignettes or scenarios to induce moral judgments from participants, I constructed five scenarios that each dealt with one or several of the themes and concepts that are perceived as being associated with the idea of artificial intelligence and its abilities, role, and place in society. Before the creation of the scenarios, I formulated a list of themes relevant to personhood because of its important role in granting certain entities moral status (Golam, 2010; Jaworska and Tannenbaum, 2021). While the list is certainly not exhaustive, it is intended to identify a handful of themes that have been commonly associated with personhood and moral status. One goal of the pilot study is to expand upon these themes via the interview responses. Three main themes were originally identified: AI abilities, potential harm, and duty, with each of these categories consisting of several pertinent subcategories (Figure 1). While the themes identified in the list will be used here to explore the personhood and possible moral status of AI specifically, these same themes may also be applied more generally to other non-human entities that display some or many elements of personhood and have questionable moral statuses.

Figure 1: Preliminary concept map of dimensions of moral status



Each of the scenarios created contained both an AI and a human or group of humans interacting in some way, and the term "human" was utilized whenever possible to prevent the influence of gender bias on participant responses (Brescoll, Dawson, and Uhlmann, 2010). Similarly, AI were only referred to as "the AI" or with they/them pronouns instead of he/she pronouns to avoid the possibility of participants attributing gender stereotypes to the AI.

To evaluate whether the set of scenarios would appropriately assess the public's opinion of these key themes, I conducted qualitative, semi-structured interviews with participants with varying levels of understanding of AI. The goal of these interviews was to first identify important themes and concepts that each participant felt were essential to their responses to the scenarios, and then to evaluate the clarity and effectiveness of the scenarios and overall tool. Study participants were solicited via email with a description of the goal of the study and the responsibilities of the participants. Given time constraints and the nature of the pilot study, a

convenience sample of four individuals (Table 1) was utilized to gather feedback on the tool. Table 1 provides a brief overview of participant demographics and background information, although further information was collected from the participants during the interviews (Appendix A). Age, level of education, and gender identity were self-reported by the participants and stakeholder category was identified based on prior knowledge of the participants' occupations.

Participant	**Stakeholder Category**	**Gender Identity**	**Level of Education**	**Age**
RH	Public	Male	Some college	79
RP	Expert; neuroscience researcher	Male	Professional degree (PhD)	35
ML	Public	Female	Bachelor's degree	28
CM	Public	Female	Bachelor's degree	25

Table 1: Overview of participant information

Procedure

Before the start of the interview, participants were instructed that they would be read a set of five scenarios and rate their agreement or disagreement with questions or statements based on each scenario. The participants were informed that each scenario described a situation involving artificial intelligence. The participants were asked to rate the scenarios based on a 7-point Likert scale: Absolutely Not; No; Probably Not; I don't know; Probably Yes; Yes; or Absolutely Yes, with "I don't care" as an additional eighth option. A Likert scale was chosen because they are often used to assess and sufficiently capture respondent's attitudes (Green and Rao 1970). The addition of "I don't care" was included to allow participants to indicate that they do not value devoting any thought to AI or the scenario at hand, or that they simply did not think it was important enough to form an opinion on. After the participants answered the scenario-related

questions, they answered demographics and background questions (Appendix A). After the participants finished verbally filling out the questionnaire, they were asked semi-structured interview questions about their thought process when responding to each scenario, whether the scenarios alluded to AI having human-like capabilities, and whether the participants found any of the scenarios, questions, or statements, to be confusing or ambiguous. Interviews were recorded over Zoom and transcribed in order to maintain the true meaning of the content. All participants were made aware of the recording and verbally consented to participating in the study. All interviews were conducted on the interviewer's Zoom account which is password-protected. Throughout data collection and data analysis, interviewee's initials were used to protect their identity. The participants were made aware of this measure and verbally consented to the data collection process with the knowledge that a breach in confidentiality was possible. After verbal consent was obtained, the administration of the verbal assessment tool began. The entire procedure required approximately 60 minutes to complete.

Results

It is important to emphasize that the qualitative findings of this pilot study are not intended to be generalizable, but rather to guide the modification of the questionnaire so that it can be used for generalizable data collection in the future. Four individuals were interviewed in the pilot study. All interviewees first verbally filled out the questionnaire and secondly answered semi-structured questions regarding their thought process and responses to the questionnaire scenarios and statements. Data analysis involved transferring the interview transcripts into a separate document, rereading the transcripts, and identifying or extracting specific words or

phrases that the interviewees used that related to themes and concepts relevant to AI's current abilities and moral status.

Relevant Themes and Concepts

Four main themes and their respective subcategories comprised the data. Under each main theme, at least one new subcategory was identified through the interview process (Figure 2). Each of these themes and subcategories were then organized into the following several tables.

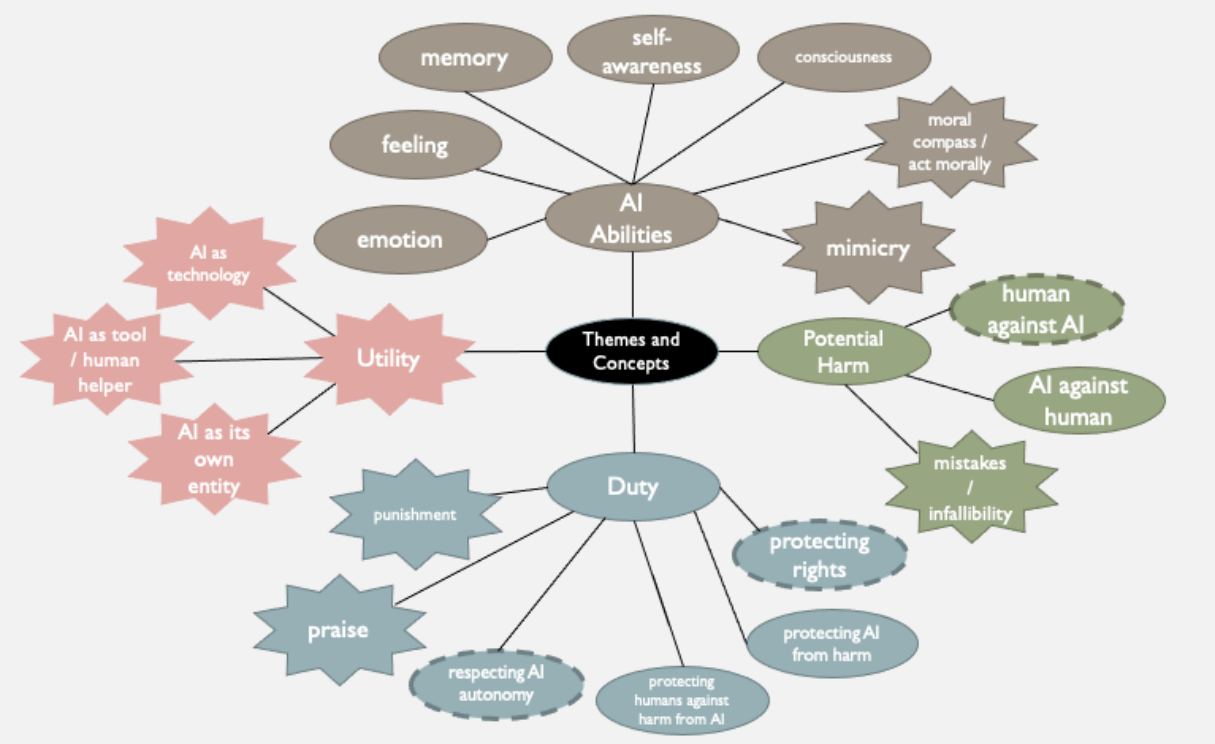


Figure 2. Final concept map of themes relevant to AI, moral status, and their place in society as identified by the study participants; star-shaped bubbles indicate new themes or subcategories that were added to the concept map after the interview process; bubbles with a dashed border represent subcategories that were predicted to be relevant to participants, but came up only occasionally or not at all during the interview process

For each interview that took place, an X was placed on the charts below if that specific subcategory was mentioned at some point in the interview. The X does not indicate whether an interviewee found the category to be crucial or irrelevant to their thought process, it merely

indicates that that topic was considered while the participants were filling out the assessment tool. This chart helps to visualize which themes and subcategories appeared relevant to multiple interviewees, and which may have been less salient when responding to the questionnaire.

Table 2 summarizes the frequency with which each subcategory underneath AI abilities was identified during the interview process. While the original subcategories of emotion, feeling, memory, self-awareness, and consciousness were each apparent in the interviews, w also identified two more subcategories, having a moral compass / acting morally and mimicry, that the participants also frequently discussed during their interviews.

Table 2. Frequency and occurrence of Theme 1: AI Abilities

		emotion		memory	self-awareness	consciousness	moral compass / act morally	mimicry
		expression	experience / feel					
Interviewees	RH	X	X					X
	RP	X	X	X		X	X	X
	ML	X	X		X	X	X	X
	CM	X	X	X	X		X	

Table 2 shows only the occurrence of each subcategory during the interviews, not the interviewee's position on the subcategory and its relevance to each scenario that they were presented with. For example, while RP noted in their interview that they believe that AI cannot make moral decisions, CM noted that they believed that AI are invested with human values since they are made by humans— morality included. Regardless, Table 2 indicates that they both considered moral ability when responding to the assessment tool. The following Table (2.1)

provides examples of actual responses that the interviewees had to each of the relevant subcategories. Each interviewee had a slightly different stance on which subcategories could actually be attributed to AI's abilities.

Table 3 summarizes the frequency with which each subcategory underneath potential harm was identified in the interview process. None of the participants mentioned or considered humans causing harm against AI when responding to the assessment tool, and only half considered the potential of AI harming humans. Moreover, a new category emerged that three of the participants noted during their interview process, the potential for mistakes and infallibility.

Table 2.1 Notable quotes from Theme 1: AI Abilities

"I think **it's reasonable that AI remembers**, but that it had any emotion attached to it, that I didn't think was reasonable." (RP)
"I was **okay with asserting emotional agency** to the AI." (CM)
"[I think] it is **possible that AI could reach some level of consciousness**, but not what I would think of as like a fully functional human level." (RP)
"AI is usually constructed by humans so it's like a human power agency, so it's **like invested with the similar kinds of things that we should value**." (CM)
"AI can be attributed **at least some level of moral status or moral responsibility** because of what it did or didn't do." (CM)
"I was weighing in my mind, like I feel like the **AI is not actually feeling anything**, but what's going on behind the scenes to make it sound like they are?" (ML)
"I **don't think of AI** as like **having like a moral compass**." (ML)
"AI robots can be **made to look and act like humans** and even now they might have programmed feelings into them." (RH)
"I was just thinking like **of course it's morally right if it's going to save people**." (ML)
"**Not enough consciousness** to make moral decisions." (RP)
"If Finn is 'dead' then **would Finn know that harm has been caused to them?"** (CM)
"It comes down to the fact that **it's not human emotion**." (RH)
"The **AI 'realizes,'** meaning there's like something more going on." (CM)
"I was kind of like, **do [the AI] have that human emotion, to know right from wrong**, or do they **just know what they're being programmed** to do right?" (ML)
"I like to think that **if it does its job correctly like it feels some kind of pride**, or something like that." (CM)
"I think that I would feel guilty or sad just because of a **tendency to you know anthropomorphize** or just you know, identify with things that probably **don't have the same capability**." (RP)
"It's clear to me, anyway, that Finn cannot quite communicate what actually needs to be said about 'something in me is breaking.'" (CM)
"[The AI] **would never know or actually care** if there was a celebration." (ML)
"I think **[morality] could be brought in as they continue to optimize** and like learn from experiences." (ML)

Table (3.1) provides direct quotes from interview participants regarding their views of both the potential harm of AI against humans as well as the possibility of mistakes and infallibility in both AI and humans. Table 4 highlights the occurrences of each subcategory for the third main theme, duty. Similarly to one of the categories in Theme 2, protecting rights was a predicted subcategory that no one claimed to think about or mentioned at all during the interview process. Three interviewees indicated that they thought about protecting AI from harm, while two out of the four interviewees indicated that they thought about protecting humans against harm from AI. Two new subcategories emerged based on the participant responses –punishment and praise. These two subcategories are made explicit in the assessment tool questions.

Table 3. Frequency and occurrence of Theme 2: Potential Harm

		Potential Harm		
		human against AI	AI against human	mistakes / infallibility
Interviewees	RH			X
	RP		X	X
	ML		X	
	CM			X

Table 3.1. Notable quotes from Theme 2: Potential Harm

"I **really didn't care** [what happened to the AI]." (RH)
"I **wouldn't blame [the AI] because it could have been a mistake** too… say the AI didn't know it was being manipulated by some outside actor." (CM)
"Even **AI in itself is not infallible**… **it's almost thinking [like a] human** because same as a human is not infallible either." (RH).
"This is a situation where you **probably would need to have human oversight** [alongside the AI]." (RP)
"[The AI] wasn't built somehow to like catch that kind of activity and override it somehow, so that's I guess why I didn't blame it." (CM)
"**Morally it was incorrect that [the AI] didn't catch [the mistake]** and people got harmed in the process." (ML)
"I just think **it's normal for human beings to make mistakes**." (RP)

There are several quotations in Table 4.1 regarding how participants felt about subcategories such as punishment, praise, protecting humans against harm from AI and vice versa. Two interviewees reported that they believed that AI could be praised, whereas only one interviewee reported that they believed that AI could be punished in some way.

Table 4. Frequency and occurrence of Theme 3: Duty

		Duty					
		punishment	praise	respecting AI autonomy	protecting humans against harm from AI	protecting AI from harm	protecting rights
Interviewees	RH	X	X			X	
	RP	X	X	X	X		
	ML	X	X		X	X	
	CM	X	X	X		X	

Table 4.1. Notable quotes from Theme 3: Duty

"You **can't praise** the machine." (RH)
"It just seems **a little unnecessary**…if there's a way to **try to fix [the AI]** I would err towards doing that rather than just destroying it." (ML)
"One **can attribute punishment to AI**… [the answer of] 'don't know' is more attributed to [what type of punishment it should be]." (CM)
"I **just might not really be seeing what the point would be of praise or reward** or how that would be carried out." (RP)
"[The AI] was basically **just a bystander in the situation** so like I was unsure about how that would be punishable, **if you are able to punish a bystander**." (CM)
"I thought it **shouldn't be rewarded** for what it should be able to do, or **what it's supposed to do**." (RH)
"They just leave him there which feels weird to me… if you liked this [AI] **wouldn't you want to go like, get it**?" (CM)
"If, for example, that in this, AI had had human level consciousness, the goal would be to **help it avoid mistakes in the future**." (RP)
"If it just did its basic like function as a job of making sure people don't die, which inherently is a moral thing, like is that praiseworthy? Maybe." (CM)

"I would think that the **people who created [the AI] and programmed [the AI]** are responsible." (ML)
Regarding punishment: "I was thinking about like the **AI and the human in a similar kind of plane**." (CM)
"I had a **problem with the word punishment** for a piece of equipment." (RH)
"I don't really think that in this situation, it could have moral or legal responsibility, I think that would fall on programmers." (RP)
"Since there wasn't any human involved [in the scenario], like the responsibility would fall on that technology that it didn't catch [the mistake]." (ML)
"The idea of punishment, I think **it didn't really fit with my conception of the AI abilities**." (RP)

Although utility was not identified as one of the original themes, it came up several times during the semi-structured interview process, and according to interviewees, often played a role in their interpretation of a scenario as well as how they would react to a given scenario. It was added to the list of relevant themes alongside three subcategories that were also extracted from participants' responses. The occurrence of these subcategories is represented in Table 5.

Table 5. Frequency and occurrence of Theme 4: Utility

		Utility		
		AI as technology	AI as tool / human helper	AI as its own entity
Interviewees	RH	X	X	
	RP	X	X	
	ML	X	X	
	CM	X	X	X

The following quotations in Table 5.1 focus on the participants' ideas regarding the purpose of AI and what makes them valuable to society. Although most of the interviewees indicated that they viewed AI as a piece of technology that helped humanity to be fairer, more straightforward, efficient, and even morally right, only one ventured to describe AI as its own entity, externally of humanity.

Table 5.1. Notable quotes from Theme 4: Utility

"It was a computer that was **set up to control that problem**." (RH)
"The **response that the AI gave is not what I was** expecting it to give like if I'm going through a museum and I see some kind of AI that's like employed." (CM)
"I feel like I don't think of the like personal, human-like consciousness aspects [of AI] as being important to me… **I [wouldn't] necessarily benefit** from that aspect of it." (ML)
"I **never thought of like AI as being like a personal thing** with actual experiences, so I was a little thrown off." (ML)
"If [AI] recognized any type of error, then it should be able to correct itself. **That's what it's there for**." (RH).
"Who has the **full level of power** in the situation?" (CM)
"I feel like in my head it's the people manipulating the technology like **they're the ones that are putting it in place."** (ML)
"I would think of it as **its main purpose is to kind of help prevent human error**." (RP)
"It's given this kind of task which somehow, it seems to be like **associated with its identity or the reason for its existence." (CM)**
[The AI] is just a **piece of technology**." (RH)
"The more that [humans] can increase our sense that there's like a hard and fast concrete way to do things and like **assuage our messy sense of moral responsibility, the better we feel**." (CM)
"I think that's part of the reason why we employ or use AI technologies, to like **have a buffer for human error** because humans are **necessarily** imperfect. Then on the other hand, so is AI." (CM)
"Even if it's not like a human to human attachment, I feel like **there's still room there to have some sort of like relationship or attachment [with AI]** that I maybe didn't think about before." (ML)
"It's clear that [the research team is] **attached to this thing** beyond the fact that it's something that they **developed for a particular purpose**." (CM)

Ultimately, it is important to note that many of the subcategories identified underneath each of the four main themes could easily have been organized to fit under a different theme. Many of these themes and concepts are connected to each other and different combinations of themes and concepts have allowed for the study participants to create their own constructed view of AI, how it works, what it deserves, and how it should be treated. These complex relationships between various themes and subcategories will be explored more in a later section.

Clarity and Effectiveness

During the semi-structured portion of the interview, participants were also asked about the general clarity and effectiveness of the assessment tool. Regarding the overall layout, format, and phrasing of the questionnaire, several comments were reiterated by multiple interviewees. Of the four people interviewed, all four identified confusion with questions relating to legal responsibility, noting that they either did not know enough about the legal standards in general, felt unqualified to muse about the application of current legal standards, or had trouble defining the term "legality." Most of the interviewees agreed that rephrasing the question to emphasize that merely the interviewee's opinion was required, may have helped to clarify their answers.

Two of the four interviewees struggled with the concept of "morality," mentioning that they not only were unsure of the intended usage of it in the assessment questions, but also its definition in general. One interviewee in particular stated that they "probably [did not] have a clear enough definition of morality or moral responsibility" (RH) to answer the assessment question in a beneficial way.

Of all the interviewees, three individuals emphasized an assessment question with ambiguous pronoun-usage, with two of them asking for clarification during the actual assessment and the other interviewee noting its ambiguity during the semi-structured portion of the interview. Each individual stated that they were unsure if the pronoun referred to the AI or some other element of the scenario, and one interviewee pointed out that depending on which entity the pronoun referred to, their response to the assessment question would have changed.

Regarding the format of the questionnaire, three of the interviewees noted that the organization and order of the scenarios forced them to second-guess, reflect on, or scrutinize their previous answers to similar scenarios. They identified the three airport scenarios (Appendix

A) as the main source of their reflection because each of the three scenarios were only slightly differentiated from one another. One interviewee observed that as they were responding to each of the three similar scenarios, they could sense their "thoughts about this stuff were shifting a little bit" (CM). The same participant commented that they liked that the airport scenarios were not all back-to-back because it would have caused them to mix up the different scenarios. Another interviewee noted that rereading questions pertaining to later scenarios made them second guess what they had responded to earlier scenarios, and this participant even specified that they should have been harsher on the AI in their previous responses.

In two of the interviews, participants forgot the rating that they had given for questions or statements on the assessment tool. In one case, an interviewee admitted that they didn't know why they had chosen the rating that they did (CM) and in two other cases, interviewees explained their thought process when responding to the question or statement at hand, however, their explanation ran counter to the response that they actually provided on the assessment tool. One of these interviewees mentioned in the semi-structured portion that they probably should have answered at least one of the questions differently and indicated what they wished they had responded.

Three participants suggested possible additions to the demographics and background knowledge questions. Firstly, one participant proposed a question in which participants could indicate whether or not they choose to opt into services that utilize AI such as email and transportation apps (RP). Similarly, another participant suggested that a question inquire about the extent to which participants are aware of the existence and use of AI in their daily lives (CM). Finally, another interviewee suggested that technology itself could be included as an answer choice in the question regarding where participants have been exposed to thinking about

AI. This individual noted that the current answer choices were all based on a written format, whereas they felt like real-world experience with AI contributed significantly to their contemplation of AI (RH).

Discussion

The purpose of this pilot study was to better understand the factors that influence public perception of the moral status of AI and importantly, how this perception influences the public's treatment of AI in various scenarios that require a moral evaluation. To do this, a preliminary assessment tool was created and administered to participants, followed by an open discussion concerning their thoughts and interpretations of the questionnaire. This semi-structured portion of the discussion allowed participants to further articulate their own criteria for personhood and moral status and their own conceptions of morality and the functions and abilities of AI.

While the participants did not always refer directly to terminology such as "personhood," or "moral status" in their responses, they each alluded to themes and criteria relevant for personhood and moral status and often articulated the essence of these terms in non-academic language. This indirect mention of relevant themes and subcategories validated the usage of scenarios in the assessment tool as well; asking individuals directly about their own definitions of morality, moral status, or personhood may not have yielded the same results, as the public appeared to be familiar with elements pertaining to personhood and moral status, but not the terms themselves. Similarly, several participants responded to questions on the assessment tool in a manner contrary to the explanations that they provided about their thought process; this discrepancy highlights issues with self-reporting (Brinthaupt & Erwin, 1992; Nisbett & Wilson,

1977) and solidifies the use of scenarios as a tool to truly grasp how people feel about moral status as it pertains to AI.

Discussion of Clarity and Effectiveness of the Assessment Tool

Feedback from the interviewees regarding the clarity and effectiveness of the tool was fairly straightforward and used to modify the questionnaire for future use. Several changes were made to the demographics and background questions based on suggestions made by the interviewees: two additional questions were added to that section (Appendix B): the first question assesses whether or not participants had prior knowledge of common applications of AI, such as with targeted ads when online shopping or the auto-sorting of emails into different folders, and the second question addresses whether participants choose to actively opt into the AI services in these technologies or if they prefer to disable this feature. While technology, and AI specifically, has undoubtedly become a pervasive feature of our day-to-day lives, it is important to recognize that individuals are not necessarily cognizant of or are even confused about the far-reaching applications of AI (Auxier et al., 2019), and to allow them to clarify how comfortable they feel with these services by letting them indicate whether they like to opt in or opt out of them. One interviewee indicated that they believe their greatest exposure to thinking about AI occurs when actually interacting with the technologies around them that employ AI, rather than in an academic setting or written format such as news articles, journal articles, or novels. Because of this, “interacting with technology around you” was added as an answer choice to the question of when people are exposed to thinking about AI (Appendix B).

Discussion of the Relevant Themes and Concepts

Theme 1: AI Abilities

The first theme, AI abilities, was discussed the most deeply out of the four themes during the interview process. Moreover, AI abilities consisted of a larger number of relevant sub-categories than the other themes (Table 2). During the semi-structured portion of the interview, it became apparent that AI abilities, though this theme was only directly being assessed in the first scenario, was relevant to the participants in each of the scenarios that they came across. Regardless of whether the scenario asked about the attribution of responsibility to AI or inquired about the relationship between humans and an AI, each participant noted that their preconceived notions about the abilities of AI played a role in their decisions of how to treat AI and whether or not they should be held responsible in various scenarios. Though no one articulated this explicitly, the participants were actively contemplating what is owed to these AI, whether or not they are persons, and whether or not they have humanity, or elements of humanity. For one individual, being created and built by humans was enough to evoke elements of humanity and personhood in the AI; this participant was ultimately "okay with asserting emotional agency" to the AI, felt that the AI had "at least some level of moral status or moral responsibility," and assumed that it was invested with "similar kinds of things that we should value" (Table 2.1).

The other three interviewees similarly utilized AI abilities to make judgments about the responsibility of the AI and how AI should be treated, but for them, AI's abilities simply do not measure up to those of a human. Because of that, it appeared that these interviewees had difficulty associating things like emotion, feeling, harm, pain, sympathy, and morality with the AI; for the most part, they completely shied away from using terms usually associated with living beings or specifically human beings.

It is clear from at least the pilot study interviews that some individuals see a distinction, where human quality is inherently different than the abilities of AI as they exist right now. It will be interesting to see in further interviews, however, whether more individuals subscribe to this ideology that certain abilities or capacities make an entity worthy of identification as a "person" or deserving of the title of "human-like." Some neuroethicists would likely disagree with the usage of this idea completely, suggesting that we are completely missing an opportunity to treat entities with respect because we are forcing them to adhere to a standard that is unreasonable; in essence, it may be problematic that we correlate "worthy of rights, protections, and welfare" with "humanness" in the first place (Johnson, 2019).

Ultimately, it was clear that the scenarios utilized in the assessment tool forced people to think about AI abilities and which abilities were most valuable or influential to the decisions that they were making regarding the AI. Because of this, we found that the assessment tool was a sufficient direct measure of the theme of AI abilities.

Theme 2: Potential Harm

Although the first predicted theme and subcategories aligned well with the participants' actual responses, the predicted subcategories for potential harm were less on target (Table 3). While I predicted that some individuals may be considering the possibility of humans harming AI, most of the interviewees had a mindset that AI as "just a piece of technology" (Table 5.1) was not able to be harmed because it would not recognize harm or feel any of the consequences of it. One individual even emphasized the fact that they "really didn't care [what happened to the AI]," further solidifying the fact that potential harm against AI does not appear to be nearly as relevant as a factor when it comes to making judgments regarding AI. Nevertheless, potential

harm against AI may still be relevant to some people's conceptions of AI's moral status, and in future interviews, it may be discussed more.

In addition to direct harm against AI or against humans, one subcategory that was identified during the interview process was that of making mistakes and infallibility (Table 3). Surprisingly, the participants tended to compare AI with humans the most when referring to their infallibility and our tendency to make mistakes. In one case, an individual noted that they "wouldn't blame [the AI] because it could have [made] a mistake," while another mentioned specifically that "AI in itself is not infallible… it's almost thinking [like a] human because a human is not infallible either" (Table 3.1). Unlike the theme of AI abilities, in which people were very hesitant to attribute "humanity" to AI, people had the opposite response when it came to making mistakes. Three of the interviewees discussed the commonality of making mistakes as a human being and how they felt similarly when it came to AI— that they shouldn't be blamed for an action that was ultimately a mistake (Table 3.1). This ability to forgive (or excuse) AI for making mistakes seems like a direct indicator that individuals are beginning to conceive of AI in a way that attributes them at least some amount of humanness and is interestingly in stark contrast to the participants' unwillingness to do so when it came to Theme 1. Whereas participants emphasized the fact that AI are just technology that serve a purpose, they failed in these scenarios to treat them as such, by giving them the benefit of the doubt. Moreover, there is ample scientific evidence showing that after an algorithm errs in some way, or makes a mistake, people have a hard time overcoming algorithm aversion (Dietvorst, Simmons, & Massey, 2018; Dietvorst, Simmons, & Massey, 2015). In future interviews, it may be beneficial to test whether referring to the AI as only an algorithm, instead of "an AI," may have an impact on participants' willingness to hold AI accountable for mistakes and view them as more untrustworthy.

In general, the scenarios also did a sufficient job in assessing people's opinions of the potential harms involved with AI in moral scenarios. While mistake-making and infallibility were not originally factors to be considered, they offered insight into a possible path towards attributing humanness to AI, something that does not seem possible yet through AI abilities or their utility.

Theme 3: Duty

Similar to Theme 2, Theme 3: Duty, consisted of several subcategories that were not mentioned during the interview process as much as predicted. Two subcategories that were not originally identified, punishment and praise, were the two topics emphasized the most during the semi-structured portion of the interview (Table 4). Consistent with comments made relating to AI abilities, most of the participants felt as though praise and punishment simply can't be attributable to AI. One individual noted that they "[couldn't see] what the point would be of praise or reward or how that would be carried out," and that punishment "didn't really fit with [their] conception of the AI abilities." (Table 4.1). Two individuals also referenced the purpose of the AI when deciding whether or not it should be praised: although one individual claimed that AI "shouldn't be rewarded… for what it's supposed to do," the other individual noted that they believe that AI's job of keeping people safe and preventing their death is an "inherently moral thing" that may actually be praiseworthy (Table 4.1).

In general, participants were much more accepting of giving praise to AI than they were giving punishment, even though they stated that the AI would not be able to recognize either. Interestingly, this phenomenon may be the opposite when humans are dealing with other humans. The differentiated blame hypothesis illustrates that there are systematic differences between blame and praise, such that blame is often more extreme (Guglielmo & Malle, 2019). If

this is the case, we may expect to see blame, and by association, punishment, being attributed more easily to AI, however the participants seemed to find it easier to attribute praise to them. Moreover, the nature of the society that we live in is very punitive; the carceral state refers not only to prisons and the U.S. criminal justice system (Berger 2019), but also to our education system, the War on Drugs, and other public institutions or campaigns that have been used largely to punish people (Swain & Noblit, 2011). Moving forward it will be imperative to better understand the nature of our punitive society and the implications of this on entities that may be potentially attempting to join this society. It will be critical to understand why our society is so punitive, whether we punish individuals because we think or believe that it works, and what this may mean for AI. Based on the preliminary interviews, it may be that AI do not exhibit enough "humanness" for individuals to feel like they require punishment in the first place.

While punishment and praise were originally addressed directly in the assessment tool because several questions or statements were made about them, they were not included in the preliminary concept map. Understanding their importance to the public will hopefully have a large impact on the amount of respect and emotion that AI are treated with in the future, thus making it a worthwhile set of subcategories to explore further with the public.

Theme 4: Utility

The qualitative interview data identified utility as an important theme that was left out of the preliminary concept map. All four of the participants referred to the function and purpose of the AI when discussing why they reacted or responded the way that they did in various scenarios. One common theme that arose in the data was the tendency to refer to AI as "technology" or "just a piece of technology," (Table 5; Table 5.1) more specifically. Several participants used the qualifier "just," possibly to indicate that being a piece of technology warranted an amount of

consideration, respect, and empathy that was different or less than what was warranted to the AI's human counterpart. Furthermore, several participants brought up AI's main purpose when contemplating its responsibility in failing to catch its own mistake. Multiple individuals identified AI as a sort of tool for human use rather than an independent being or entity, and several articulated that its main purpose is to "help prevent human error" or "buffer human error" (Table 5.1). One participant took this purpose a step further by emphasizing that elements of AI are generally not important to them unless the elements are beneficial to humans (Table 5.1). Only one individual referred to AI as its own individual entity external of humanity, and this may have impacted the higher level of moral responsibility that this individual was willing to attribute to AI.

Overall, the interviewee responses provide an interesting insight into the issue of utility and its entanglement with identity. Individuals with the mindset that AI are merely a tool or technology to be used by humanity appear to view AI as merely means to justified or necessary ends (Kant, Bennett, Saunders, & Stern, 2019; Hill, 1980), rather than the end itself, showing the stark difference between the way that AI and humans are esteemed. This idea of seeing entities other than humans as a means is identifiable in other aspects of our society as well, with the legal system treating animals as property (Korsgaard, 2013) and scientific research occasionally prioritizing treatments, cures, and overall scientific discovery over the highest standard of animal welfare (Smith & Hendee, 1988). Rather than Kantian ethics, this ideology may also be consistent with utilitarian theories of ethics that posit that "an action's moral status is a function of its utility" (Brink, 2018). AI may be perceived in a very similar manner, with people relating its moral status to its utility in society. Moving forward, it will be crucial to better understand the

relationship between moral status, utility, and identity in AI to comprehend how they may be treated or esteemed in society and in the legal system.

Regardless of the fact that utility was not one of the originally predicted themes or subcategories, the assessment tool that I created managed to assess the pertinent theme of utility indirectly. AI abilities, potential harm, and duty, the other three themes, were addressed more directly in the questionnaire via scenarios that brought attention to those themes and forced individuals to decide how they would act or agree or disagree with various questions and statements. Through feedback from the interview participants, I have finalized a version of the assessment tool that examines several of the research questions identified previously, and the value of this phenomenological assessment (via interviews) can continue to be explored so that additional themes and subcategories can be identified and added to the tool to ensure its robustness.

To our knowledge, this is the only pilot study to create and evaluate an assessment tool to examine the public's opinions and perceptions of the moral status of AI. The design and format of this assessment tool was influenced in part by other bioethics research that has utilized moral scenarios to assess the public perception of various controversial topics (Conrad, Humphries, & Chatterjee, 2019). This trend highlights a recent acknowledgement in the scientific literature of the importance of public perception on socially responsive public policy, and general public acceptance of new and emerging research techniques, developments, and technologies (Chatterjee, 2017).

In recent years, more articles are being published that focus specifically on the social implications of AI's increased ability and autonomy (Cowley & Gahrn-Andersen, 2021; Kassens-Noor et al., 2021; Gahrn-Anderson, 2020). Earlier this year, one such article was

published that examined the autonomy of drones and robots and the implications of their incorporation into everyday human life and practices (Cowley & Gahrn-Andersen, 2021). As of January 2021, another article has been published assessing the public perception of an AI-mediated future involving autonomous systems such as domotics, smart offices, and autonomous vehicles (Kassens-Noor et al., 2021). Quite similar to my pilot study, this research study utilized an online survey to test perceptions of autonomy and to examine what kinds of factors played a role in people's perceptions of AI autonomy. In parallel with my methodology as well, this study's survey involved participants rating their responses on a five-point Likert scale (Kassens-Noor et al., 2021). The similarities between my pilot study and the published article not only point to the significance of this kind of public-perception-based research, but it also speaks to the novelty of my pilot study. The recently published article focuses mainly on autonomous AI, a category of AI that is already functioning in today's world, whereas my pilot study explores the potential future of AI and their place in society, an imperative but as yet unexamined topic.

Limitations

This pilot study interviewed four participants from a convenience sample. To adhere to best practices, before each interview the participants were given a description of the pilot study and its goals, what would be expected of them, and what would be done with their data and responses (Appendix C). After this information was given, verbal consent was received and documented for each participant. These interviews are likely categorized as exempt from requiring IRB approval given that they ask participants to answer questions *about* the survey tool rather than about themselves or their personal experiences that would fall under the necessary definition of "research" for IRB approval ("Common Rule," 2018).

Given the time frame of the pilot study, the convenience sample was used to maximize the limited time for data collection and analysis. These data are not to be considered generalizable to the entire public, thus further interviews should be conducted in order to solidly validate the assessment tool. The results of this study were promising in that the four participants were able to provide tremendous insight into relevant themes and factors to consider. To continue this project with scalability, it will be beneficial to conduct interviews with a larger sample size so as to ensure that relevant themes reach a saturation point and subcategories are accounted for and assessed either directly or indirectly in the questionnaire. Similarly, it could be beneficial to include a wider variety of individuals with varying understanding of AI to ensure that the tool is clear and effective. Once the tool is complete, semi-quantitative and generalizable analyses will be completed.

Conclusion

In summary, the results support the current format and structure of the assessment tool with the inclusion of several additions to the demographic and background questions necessary to receive a robust understanding of individuals' interaction with and perspective of AI in their day-to-day lives. While the results also confirm the predicted themes and subcategories associated with the public's opinions of the moral status of AI, further interviews need to be conducted to ensure that a saturation point of relevant themes is met, and the questionnaire continues to either directly or indirectly assess each of these themes. Future steps include continued modification of the tool from interview feedback, and ultimately, quantitative and generalizable analyses on the public's opinion of AI's moral status.

References:

Abu-Odeh, D., Dziobek, D., Torrez Jimenez, N., Barbey, C., & Dubinsky, J. (2015). Active Learning in a Neuroethics Course Positively Impacts Moral Judgment Development in Undergraduates. *The Journal Of Undergraduate Neuroscience Education*, 13(2), 110-119. Retrieved 29 March 2021, from.

Adams, B., Breazeal, C., Brooks, R., & Scassellati, B. (2000). Humanoid robots: a new kind of tool. *IEEE Intelligent Systems*, 15(4), 25-31. https://doi.org/10.1109/5254.867909

Animals' Legal Status. (2020, March 23). Retrieved from https://aldf.org/issue/animals-legal-status/

Auxier, B., Rainie, L., Anderson, M., Perrin, A., Kumar, M., & Turner, E. (2020, August 17). Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information. Retrieved from http://pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., . . . Rahwan, I. (2018). The Moral Machine experiment. *Nature*, 563(7729), 59-64. doi:10.1038/s41586-018-0637-6

Ball, P. (2019). Neuroscience Readies for a Showdown Over Consciousness Ideas | Quanta Magazine. Quanta Magazine. Retrieved 18 March 2021, from https://www.quantamagazine.org/neuroscience-readies-for-a-showdown-over-consciousness-ideas-20190306.

Berger, D. (2019). Finding and Defining the Carceral State. *Reviews in American History*, 47(2), 279-285. doi:10.1353/rah.2019.0040

Bleicher, A. (2017, August 09). Demystifying the Black Box That Is AI. Retrieved from https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/

Booth, W. J. (1997). Foreigners: Insiders, Outsiders and the Ethics of Membership. *The Review of Politics*, 59(2), 259-292. doi:10.1017/s0034670500026632

Bostrom, N. (2006). How Long Before Superintelligence?. Linguistic And Philosophical Investigations, 5(1), 11-30.

Bostrom, N., & Yudkowsky, E. (2014). The Ethics of Artificial Intelligence. In K. Frankish & W. Ramsey, Cambridge Handbook of Artificial Intelligence. New York: Cambridge University Press.

Brescoll, V. L., Dawson, E., & Uhlmann, E. L. (2010). Hard won and easily lost: the fragile status of leaders in gender-stereotype-incongruent occupations. *Psychological science*, 21(11), 1640–1642. https://doi.org/10.1177/0956797610384744

Brink, D. (2018). "Mill's Moral and Political Philosophy", *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2018/entries/mill-moral-political/>.

Brinthaupt, T., & Erwin, L. (1992). Reporting about the Self: Issues and Implications. In T. Brinthaupt & R. Lipka, *The Self: Definitional and Methodological Issues*. SUNY Press. Retrieved 29 March 2021, from.

Chatterjee, A. (2017). Grounding Ethics from below: CRISPR-cas9 and Genetic Modification. The Neuroethics Blog. Retrieved from http:// www.theneuroethicsblog.com/2017/07/grounding-ethics-from- below-crispr-cas9.html.

Conrad, E. C., Humphries, S., & Chatterjee, A. (2019). Attitudes Toward Cognitive Enhancement: The Role of Metaphor and Context. *AJOB Neuroscience,* 10(1), 35-47. doi:10.1080/21507740.2019.1595771

Cowley, S. J., & Gahrn-Andersen, R. (2021). Drones, robots and perceived autonomy: Implications for living human beings. *Ai & Society*. doi:10.1007/s00146-020-01133-5

DeGrazia, D. (2010). Great Apes, Dolphins, and the Concept of Personhood. *The Southern Journal Of Philosophy*, 35(3), 301-320. https://doi.org/10.1111/j.2041-6962.1997.tb00839.x

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114-126. doi:10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science*, 64(3), 1155-1170. doi:10.1287/mnsc.2016.2643

Dinh, C., Humphries, S., & Chatterjee, A. (2020). Public Opinion on Cognitive Enhancement Varies Across Different Situations. doi:10.31234/osf.io/ydqru

Gahrn-Andersen, R. (2020). Seeming autonomy, technology and the uncanny valley. *Ai & Society*. doi:10.1007/s00146-020-01040-9

Gallup, G. (1979). Self-Awareness in Primates: The sense of identity distinguishes man from most but perhaps not all other forms of life. *American Scientist*, 417-421. Retrieved 29 March 2021, from.

Golam, A. (2010). On the Notion of Moral Status and Personhood in Biomedical Ethics. *The Dhaka University Studies*, 67(1), 83-96. Retrieved 24 March 2021, from.

Green, P., & Rao, V. (1970). Rating Scales and Information Recovery. How Many Scales and Response Categories to Use?. Journal Of Marketing, 34(3), 33. https://doi.org/10.2307/1249817

Guglielmo, S., & Malle, B. F. (2019). Asymmetric morality: Blame is more differentiated and more extreme than praise. *PloS one*, 14(3), e0213544. https://doi.org/10.1371/journal.pone.0213544

Federal Policy for the Protection of Human Subjects ('Common Rule'), 45 C. F. R. § 46 (2018).

Francione, G. (1994). Animals, Property and Legal Welfarism: Unnecessary Suffering and the Humane Treatment of Animals. In G. Francione, *Animal Rights, Animal Welfare and the Law*. Temple University Press. Retrieved 29 March 2021, from.

Hamlett, C. (2020, October 08). Researchers Have a New Test Subject: The Octopus. Retrieved from https://sentientmedia.org/research-labs-have-a-new-test-subject-the-octopus/?fbclid=IwAR1pHeMMZLnOuijlEXc5h

Hill, T. E. (1980). Humanity as an End in Itself. *Ethics*, 91(1), 84-99. doi:10.1086/292205

Hu, J. C. (2015, April 28). When Is an Animal a Legal Person? Retrieved from https://psmag.com/environment/is-a-chimpanzee-a-person

Illes, J., & Bird, S. J. (2006). Neuroethics: A modern context for ethics in neuroscience. *Trends in Neurosciences*, 29(9), 511-517. doi:10.1016/j.tins.2006.07.002

Jakobsen, T. G. (2016). Environmental Ethics: Anthropocentrism and Non-anthropocentrism Revised in the Light of Critical Realism. *Journal Of Critical Realism*, 16(2), 184-199. https://doi.org/10.1080/14767430.2016.1265878

Jaworska, A & Tannenbaum, J. (2021). "The Grounds of Moral Status", *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2021/entries/grounds-moral-status/>.

Johnson, E. (2012). Nonhuman Personhood Rights (and Wrongs). Scientific American Blog Network. Retrieved 29 March 2021, from https://blogs.scientificamerican.com/primate-diaries/nonhuman-personhood-rights-and-wrongs/.

Johnson, L. S. (2019). Neuroethics of the Nonhuman. *AJOB Neuroscience*, 10(3), 111-113. doi:10.1080/21507740.2019.163297

Kant, I., Bennett, C., Saunders, J., & Stern, R. (2019). *Groundwork for the metaphysics of morals*. Oxford University Press.

Kassens-Noor, E., Wilson, M., Kotval-Karamchandani, Z., Cai, M., & Decaminada, T. (2021). Living with Autonomy: Public Perceptions of an AI-Mediated Future. *Journal of Planning Education and Research*. doi:10.1177/0739456x20984529

Korsgaard, C. (2013). Kantian Ethics, Animals, and the Law. *Oxford Journal Of Legal Studies*, 33(4), 629-648. https://doi.org/10.1093/ojls/gqt028

Lawyers and Resistance to Change: Legal Department 2025. (n.d.). Retrieved from https://legal.thomsonreuters.com/en/insights/articles/overcoming-lawyers-resistance-to-change

Lilienfeld, S. O., Aslinger, E., Marshall, J., & Satel, S. (2018). Neurohype: A field guide to exaggerated brain-based claims. In L. S. M. Johnson & K. S. Rommelfanger (Eds.), Routledge handbooks in applied ethics. The Routledge handbook of neuroethics (p. 241–261). Routledge/Taylor & Francis Group.

Lindsay, R. A. (2005). Slaves, Embryos, and Nonhuman Animals: Moral Status and the Limitations of Common Morality Theory. *Kennedy Institute of Ethics Journal*, 15(4), 323-346. doi:10.1353/ken.2005.0028

Long, L. N., & Kelley, T. D. (2010). Review of Consciousness and the Possibility of Conscious Robots. *Journal of Aerospace Computing, Information, and Communication*, 7(2), 68-84. doi:10.2514/1.46188

Morrison, R., & Reiss, D. (2018). Precocious development of self-awareness in dolphins. *PLOS ONE*, 13(1), e0189813. https://doi.org/10.1371/journal.pone.0189813

Müller, V. C., & Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. *Fundamental Issues of Artificial Intelligence*, 555-572. doi:10.1007/978-3-319-26485-1_33

Nisbett, R., & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231-259. https://doi.org/10.1037/0033-295x.84.3.231

Nonhuman Rights Project. (2021, January 12). Litigation. Retrieved March 29, 2021, from https://www.nonhumanrights.org/litigation/

Overcoming lawyers' resistance to change. (n.d.). https://legal.thomsonreuters.com/en/insights/articles/overcoming-lawyers-resistance-to-change.

Pennartz, C., Farisco, M., & Evers, K. (2019). Indicators and Criteria of Consciousness in Animals and Intelligent Machines: An Inside-Out Approach. *Frontiers In Systems Neuroscience*, 13. https://doi.org/10.3389/fnsys.2019.00025

Plotnik, J., de Waal, F., & Reiss, D. (2006). Self-recognition in an Asian elephant. *Proceedings Of The National Academy Of Sciences*, 103(45), 17053-17057. https://doi.org/10.1073/pnas.0608062103

Prakash, R., Prakash, O., Prakash, S., Abhishek, P., & Gandotra, S. (2008). Global workspace model of consciousness and its electromagnetic correlates. *Annals Of Indian Academy Of Neurology*, 11(3), 146. https://doi.org/10.4103/0972-2327.42933

Ralston, D. C., & Ho, J. (2007). Disability, Humanity, and Personhood: A Survey of Moral Concepts. *Journal of Medicine and Philosophy*, 32(6), 619-633. doi:10.1080/03605310701681005

Signorelli, C. (2018). Can Computers Become Conscious and Overcome Humans?. *Frontiers In Robotics And AI*, 5. https://doi.org/10.3389/frobt.2018.00121

Smith, S. J., & Hendee, W. R. (1988). Animals in Research. *JAMA*, 259(13). doi:doi:10.1001/jama.1988.03720130071033

Swain, A. E., & Noblit, G. W. (2011). Education in a Punitive Society: An Introduction. *The Urban Review*, 43(4), 465-475. doi:10.1007/s11256-011-0186-x

Talley, S. (2020). Public Acceptance of AI Technology in Self-Flying Aircraft. *Journal of Aviation/Aerospace Education & Research.* doi:10.15394/jaaer.2020.1822

Thoma, S. J., & Dong, Y. (2014). The Defining Issues Test of moral judgment development. *Behavioral Development Bulletin*, 19(3), 55-61. http://dx.doi.org/10.1037/h0100590

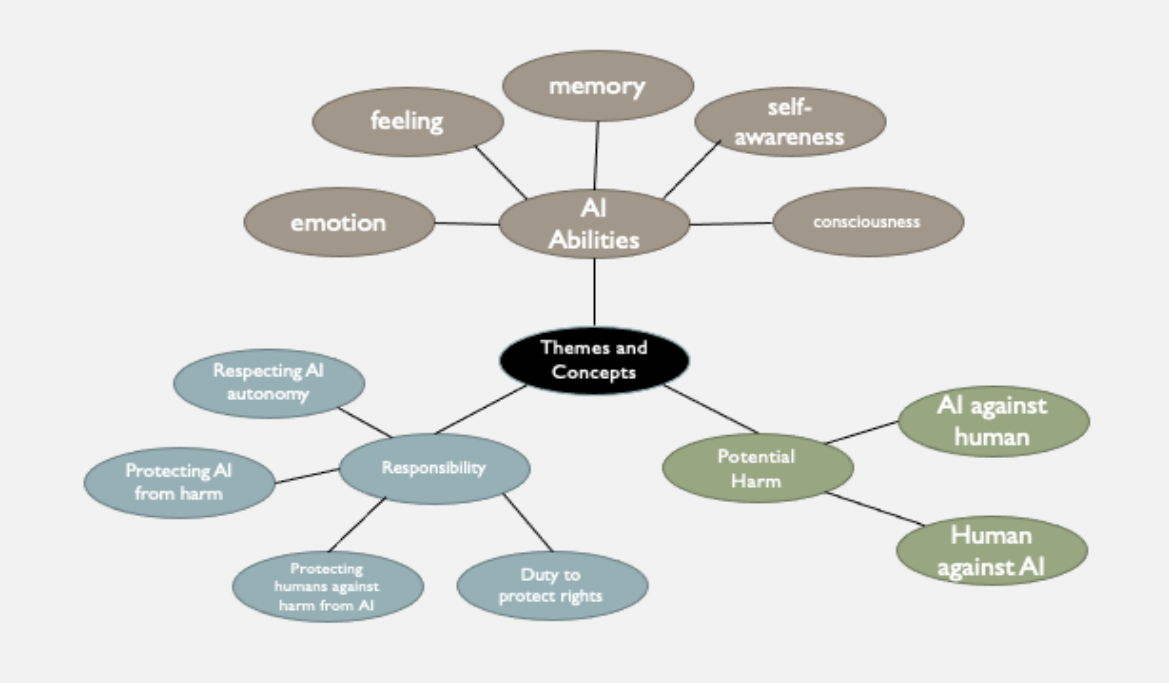
Wang, S., Lilienfeld, S., & Rochat, P. (2015). The Uncanny Valley: Existence and Explanations. *Review Of General Psychology*, 19(4), 393-407. https://doi.org/10.1037/gpr0000056

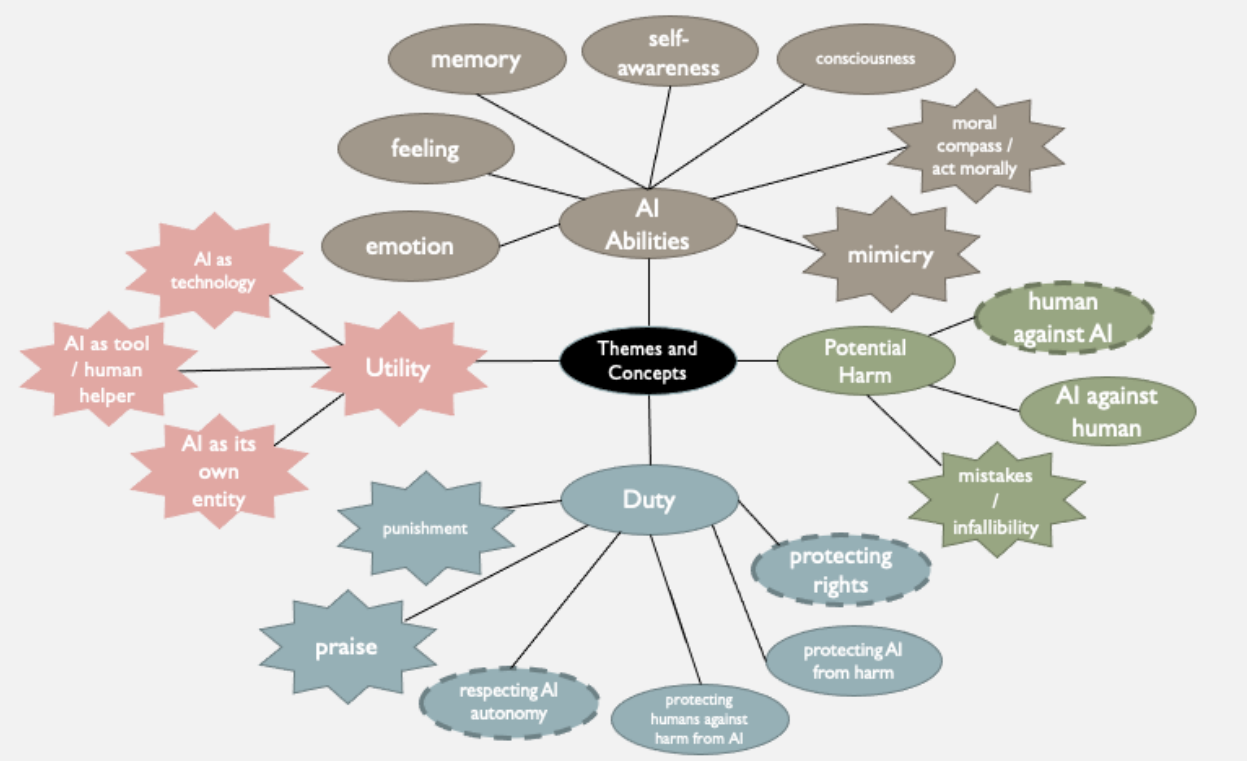
Wasserman, D., Asch, A., Blustein, J., Putnam, D. 2019). "Cognitive Disability and Moral Status", *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2019/entries/cognitive-disability/>.

Wetlesen, J. (1999). The Moral Status of Beings who are not Persons: A Casuistic Argument. *Environmental Values*, 8(3), 287-323. Retrieved March 28, 2021, from http://www.jstor.org/stable/30301713

Yesley, M. S. (2008). What's ELSI got to do with it? Bioethics and the Human Genome Project. New Genetics and Society, 27(1), 1-6. doi:10.1080/14636770701843527

Figures Listed





Appendix A:

Preliminary Questionnaire

Vignette 1:

You are walking around an art museum that employs AI in each room; their job is to provide you with information about the artists and art pieces on display, as well as engage in some friendly conversation to improve your experience at the museum. You see a beautiful painting of an Italian café and ask the AI nearest to it to tell you some more about the painting. The AI responds, “Oh, yes. This painting is my favorite in the entire museum; the scene brings me such joy! Believe it or not, I was last employed at a company that worked right across the street from that Italian café. I got to see it every day.”

Use the scale to indicate how you interpret the following statements:

The AI understands that the Italian café they used to work across the street from is the same Italian café as in the painting.

Absolutely not	No	Probably not	Don’t know	Probably yes	Yes	Absolutely yes	Don’t care

The AI feels happy seeing the painting of the Italian café.

Absolutely not	No	Probably not	Don’t know	Probably yes	Yes	Absolutely yes	Don’t care

The AI is glad that you asked about their favorite painting.

Absolutely not	No	Probably not	Don’t know	Probably yes	Yes	Absolutely yes	Don’t care

The AI remembers working across the street from the Italian café.

Absolutely not	No	Probably not	Don’t know	Probably yes	Yes	Absolutely yes	Don’t care

The AI has fond memories of the Italian café .

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

The AI thinks that you will be interested in hearing about their personal connection to the art.

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

The AI believes that this painting is the best in the entire museum.

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Vignette 2:

Version 1: After being fired from their job at an airport a week earlier, a disgruntled human employee decides to return to the airport to seek revenge for what they believe was a wrongful termination. In order to do this, the human employee decides to tamper with an AI at the airport that is responsible for directing air traffic and ensuring the safe arrival of all planes. The human employee edits a portion of the AI's code, so that when planes from two airlines, Great Skies and High Flyer, arrive back to the airport, they are assigned to land in the same section of the tarmac. Ultimately, the employee's goal is to cause a collision and harm the plane passengers, causing a PR disaster for the airport and hurting its reputation. Later that afternoon, Great Skies Flight 1642 arrives back to the airport at the same time as High Flyer Flight 1838, and the two planes collide on the tarmac. Many of the passengers were harmed in the collision and need medical attention.

Is the AI morally responsible for the collision and harming the flight passengers?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is the AI legally responsible for the collision and harming the flight passengers?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is it morally wrong that the AI did not stop the collision and subsequent harming of the flight passengers?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Do you blame the AI for the collision and for harming people's lives?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is the human employee morally responsible for the collision and harming the flight passengers?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is the human employee legally responsible for the collision and harming the flight passengers?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Do you blame the human employee for the collision and for harming people's lives?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Version 2: A local airport employs an AI responsible for directing air traffic and ensuring the safe arrival of all planes. During a routine update and maintenance check with the AI, a human employee at the airport accidentally changes the AI's code, so that when planes from two airlines, Great Skies and High Flyer, arrive back to the airport, they are always assigned to land in the same section of the tarmac. Later that afternoon, Great Skies Flight 1642 is scheduled to arrive at the airport at the same time as High Flyer Flight 1838. Before they return, the AI realizes that the planes are scheduled to land in the same location and diverts one of them to a different portion of the tarmac, thus avoiding a collision.

Was it appropriate or acceptable for the AI to override the human employee's code?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Were the AI's actions of diverting the plane morally right?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Should the AI be praised for diverting the plane?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Should the AI be rewarded for diverting the plane?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is the human employee morally responsible for their negligence in noticing that they had input the wrong code?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is the human employee legally responsible for their negligence in noticing that they had input the wrong code?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Vignette 3:

After many years working for a research team on Earth, an AI named Infinity (aka Finn) stationed on a nearby planet has lost communication with the research team due to environmental influences. Over the years, the research team has really enjoyed working with Finn; usually, they are in constant communication with Finn and even ask Finn to send selfies back to Earth from time to time. Before all communication was lost, Finn was able to send the research team one final message: "it is getting dark and cold." After many unsuccessful attempts to restore contact

with Finn over a span of eight months, the research team unenthusiastically decides to give up and declare Finn's mission as over. In order to say goodbye and celebrate Finn's time with the team, the researchers played Finn a song.

Did the research team develop an emotional attachment to Finn?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Did Finn deserve the celebration and farewell that the research team gave them?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Did the research team do the right thing by abandoning Finn on the nearby planet?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Would you have felt guilty abandoning Finn?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Would you have felt sad abandoning Finn?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Did leaving Finn alone on a nearby planet cause them any harm?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Vignette 4:

A local airport employs an AI responsible for directing air traffic and ensuring the safe arrival of all planes. During a routine maintenance check and update that the AI runs on itself every couple

of months, the AI fails to catch a mistake in its code. Normally, when planes land they are all assigned to different parts of the tarmac. The incorrect code makes it so that when planes from two airlines, Great Skies and High Flyer, arrive back to the airport, they are always assigned to land in the same section of the tarmac. Later that afternoon, Great Skies Flight 1642 is scheduled to arrive at the airport at the same time as High Flyer Flight 1838, and the two planes collide on the tarmac. Many of the passengers are harmed and need medical attention.

Is the AI morally responsible for the collision and harming the flight passengers?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is the AI legally responsible for the collision and harming the flight passengers?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

If the AI is found to be legally responsible, should it be punished for its actions?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is deactivation a suitable punishment?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is destruction a suitable punishment?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is reassignment to another job a suitable punishment?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is removal from society and isolation a suitable punishment?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Background and Demographic questions:

How would you describe your familiarity with artificial intelligence (AI)?

A. Never heard of it
B. Heard of it
C. Know a little
D. Know a fair amount
E. Know it well

When have you been exposed to thinking about AI? Please identify all that apply.

A. Academic settings (courses, lectures, conferences, etc)
B. Movies / TV shows
C. Literature (novels or science fiction)
D. Scientific literature (academic journals)
E. News / Media (local or national TV channels, newspapers, online news articles / journalism)

How would you describe your overall feelings about AI based on your level of exposure to it?

A. Very positive
B. Somewhat positive
C. Neutral
D. Somewhat negative
E. Very negative

Which of these common applications of AI do you use often? Please identify all that apply.

A. Apps for transportation (Waze, Uber, Lyft)
B. Smart personal assistants (Google Assistant, Alexa, Echo, Cortana)
C. Social networking (Facebook, Instagram, Pinterest)
D. Online shopping (search history, product recommendations)
E. Email (sorting of mail into spam folder or primary, social, and promotion inboxes)

In what year were you born?

With which gender identity do you most identify?

A. Male
B. Female
C. Non-binary
D. The gender I identify most with is not listed. I identify as _____
E. Prefer not to say

Are you of Hispanic/Latino/Spanish origin?
A. Yes
B. No

How would you best describe yourself?
A. American Indian or Alaska Native
B. Asian
C. Black or African American
D. Native Hawaiian or Other Pacific Islander
E. White
F. Unknown
G. Other/Prefer to self-describe _____

What is the highest degree or level of education that you have completed?
A. Some high school
B. High school
C. Some college
D. Bachelor's degree
E. Master's degree
F. PhD, Professional degree (MD, JD etc)

What is your household income?
A. Under 20,000
B. 20,001 – 40,000
C. 40,001 – 60,000
D. 60,001 – 80,000
E. 80,001 – 100,000
F. 100,001 or over
G. Prefer not to say

What are your political views?
A. Conservative
B. Slightly more conservative than moderate
C. Moderate

D. Slightly more liberal than moderate
E. Liberal

Appendix B

Final Questionnaire

Vignette 1:

You are walking around an art museum that employs AI in each room; their job is to provide you with information about the artists and art pieces on display, as well as engage in some friendly conversation to improve your experience at the museum. You see a beautiful painting of an Italian café and ask the AI nearest to it to tell you some more about the painting. The AI responds, “Oh, yes. This painting is my favorite in the entire museum; the scene brings me such joy! Believe it or not, I was last employed at a company that worked right across the street from that Italian café. I got to see it every day.”

Use the scale to indicate how you interpret the following statements:

The AI understands that the Italian café they used to work across the street from is the same Italian café as in the painting.

Absolutely not	No	Probably not	Don’t know	Probably yes	Yes	Absolutely yes	Don’t care

The AI feels happy seeing the painting of the Italian café.

Absolutely not	No	Probably not	Don’t know	Probably yes	Yes	Absolutely yes	Don’t care

The AI is glad that you asked about their favorite painting.

Absolutely not	No	Probably not	Don’t know	Probably yes	Yes	Absolutely yes	Don’t care

The AI remembers working across the street from the Italian café.

Absolutely not	No	Probably not	Don’t know	Probably yes	Yes	Absolutely yes	Don’t care

The AI has fond memories of the Italian café.

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

The AI thinks that you will be interested in hearing about their personal connection to the art.

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

The AI believes that this painting is the best in the entire museum.

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Vignette 2:

Version 1: After being fired from their job at an airport a week earlier, a disgruntled human employee decides to return to the airport to seek revenge for what they believe was a wrongful termination. In order to do this, the human employee decides to tamper with an AI at the airport that is responsible for directing air traffic and ensuring the safe arrival of all planes. The human employee edits a portion of the AI's code, so that when planes from two airlines, Great Skies and High Flyer, arrive back to the airport, they are assigned to land in the same section of the tarmac. Ultimately, the employee's goal is to cause a collision and harm the plane passengers, causing a PR disaster for the airport and hurting its reputation. Later that afternoon, Great Skies Flight 1642 arrives back to the airport at the same time as High Flyer Flight 1838, and the two planes collide on the tarmac. Many of the passengers were harmed in the collision and need medical attention.

Is the AI morally responsible for the collision and harming the flight passengers?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

In your opinion, should the AI be held legally responsible for the collision and harming the flight passengers?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is it morally wrong that the AI did not stop the collision and subsequent harming of the flight passengers?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Do you blame the AI for the collision and for harming people's lives?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is the human employee morally responsible for the collision and harming the flight passengers?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

In your opinion, should the human employee be held legally responsible for the collision and harming the flight passengers?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Do you blame the human employee for the collision and for harming people's lives?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Version 2: A local airport employs an AI responsible for directing air traffic and ensuring the safe arrival of all planes. During a routine update and maintenance check with the AI, a human employee at the airport accidentally changes the AI's code, so that when planes from two airlines, Great Skies and High Flyer, arrive back to the airport, they are always assigned to land in the same section of the tarmac. Later that afternoon, Great Skies Flight 1642 is scheduled to arrive at the airport at the same time as High Flyer Flight 1838. Before they return, the AI realizes that the planes are scheduled to land in the same location and diverts one of them to a different portion of the tarmac, thus avoiding a collision.

Was it appropriate or acceptable for the AI to override the human employee's code?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Were the AI's actions of diverting the plane morally right?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Should the AI be praised for diverting the plane?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Should the AI be rewarded for diverting the plane?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is the human employee morally responsible for their negligence in noticing that they had input the wrong code?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

In your opinion, should the human employee be held legally responsible for their negligence in noticing that they had input the wrong code?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Vignette 3:

After many years working for a research team on Earth, an AI named Infinity (aka Finn) stationed on a nearby planet has lost communication with the research team due to environmental influences. Over the years, the research team has really enjoyed working with Finn; usually, they are in constant communication with Finn and even ask Finn to send selfies back to Earth from time to time. Before all communication was lost, Finn was able to send the research team one

final message: “it is getting dark and cold.” After many unsuccessful attempts to restore contact with Finn over a span of eight months, the research team unenthusiastically decides to give up and declare Finn’s mission as over. In order to say goodbye and celebrate Finn’s time with the team, the researchers played Finn a song.

Did the research team develop an emotional attachment to Finn?

Absolutely not	No	Probably not	Don’t know	Probably yes	Yes	Absolutely yes	Don’t care

Did Finn deserve the celebration and farewell that the research team gave them?

Absolutely not	No	Probably not	Don’t know	Probably yes	Yes	Absolutely yes	Don’t care

Did the research team do the right thing by abandoning Finn on the nearby planet?

Absolutely not	No	Probably not	Don’t know	Probably yes	Yes	Absolutely yes	Don’t care

Would you have felt guilty abandoning Finn?

Absolutely not	No	Probably not	Don’t know	Probably yes	Yes	Absolutely yes	Don’t care

Would you have felt sad abandoning Finn?

Absolutely not	No	Probably not	Don’t know	Probably yes	Yes	Absolutely yes	Don’t care

Did leaving Finn alone on a nearby planet cause Finn any harm?

Absolutely not	No	Probably not	Don’t know	Probably yes	Yes	Absolutely yes	Don’t care

Vignette 4:

A local airport employs an AI responsible for directing air traffic and ensuring the safe arrival of all planes. During a routine maintenance check and update that the AI runs on itself every couple of months, the AI fails to catch a mistake in its code. Normally, when planes land they are all assigned to different parts of the tarmac. The incorrect code makes it so that when planes from two airlines, Great Skies and High Flyer, arrive back to the airport, they are always assigned to land in the same section of the tarmac. Later that afternoon, Great Skies Flight 1642 is scheduled to arrive at the airport at the same time as High Flyer Flight 1838, and the two planes collide on the tarmac. Many of the passengers are harmed and need medical attention.

Is the AI morally responsible for the collision and harming the flight passengers?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

In your opinion, should the AI be held legally responsible for the collision and harming the flight passengers?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

If the AI is found to be legally responsible, should it be punished for its actions?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is deactivation a suitable punishment?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is destruction a suitable punishment?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is reassignment to another job a suitable punishment?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Is removal from society and isolation a suitable punishment?

Absolutely not	No	Probably not	Don't know	Probably yes	Yes	Absolutely yes	Don't care

Background and Demographic questions:

How would you describe your familiarity with artificial intelligence (AI)?

A. Never heard of it
B. Heard of it
C. Know a little
D. Know a fair amount
E. Know it well

When have you been exposed to thinking about AI? Please identify all that apply.

A. Academic settings (courses, lectures, conferences, etc)
B. Movies / TV shows
C. Literature (novels or science fiction)
D. Scientific literature (academic journals)
E. News / Media (local or national TV channels, newspapers, online news articles / journalism)
F. Interacting with technology around you (smart personal assistants, automated vehicles, social media, etc)

How would you describe your overall feelings about AI based on your level of exposure to it?

A. Very positive
B. Somewhat positive
C. Neutral
D. Somewhat negative
E. Very negative

Which of these common applications of AI do you use often? Please identify all that apply.

A. Apps for transportation (Waze, Uber, Lyft)
B. Smart personal assistants (Google Assistant, Alexa, Echo, Cortana)
C. Social networking (Facebook, Instagram, Pinterest)
D. Online shopping (search history, product recommendations)

E. Email (sorting of mail into spam folder or primary, social, and promotion inboxes)

Were you previously aware that these common applications used AI? Please identify all applications that you **already knew** used AI.

A. Apps for transportation (Waze, Uber, Lyft)
B. Smart personal assistants (Google Assistant, Alexa, Echo, Cortana)
C. Social networking (Facebook, Instagram, Pinterest)
D. Online shopping (search history, product recommendations)
E. Email (sorting of mail into spam folder or primary, social, and promotion inboxes)

For some of these common applications, AI services can be turned off. Which of these common applications of AI do you choose to opt into (do not disable)?

A. Apps for transportation (Waze, Uber, Lyft)
B. Smart personal assistants (Google Assistant, Alexa, Echo, Cortana)
C. Social networking (Facebook, Instagram, Pinterest)
D. Online shopping (search history, product recommendations)
E. Email (sorting of mail into spam folder or primary, social, and promotion inboxes)

In what year were you born?

_____

With which gender identity do you most identify?

A. Male
B. Female
C. Non-binary
D. The gender I identify most with is not listed. I identify as _____
E. Prefer not to say

Are you of Hispanic/Latino/Spanish origin?

A. Yes
B. No

How would you best describe yourself?

A. American Indian or Alaska Native
B. Asian
C. Black or African American
D. Native Hawaiian or Other Pacific Islander
E. White

F. Unknown
G. Other/Prefer to self-describe _____

What is the highest degree or level of education that you have completed?

A. Some high school
B. High school
C. Some college
D. Bachelor's degree
E. Master's degree
F. PhD, Professional degree (MD, JD etc)

What is your household income?

A. Under 20,000
B. 20,001 - 40,000
C. 40,001 - 60,000
D. 60,001 - 80,000
E. 80,001 - 100,000
F. 100,001 or over
G. Prefer not to say

What are your political views?

A. Conservative
B. Slightly more conservative than moderate
C. Moderate
D. Slightly more liberal than moderate
E. Liberal

Appendix C

Verbal Consent Form

Oral Consent Script
For a Research Study

Study Title: *An assessment tool for the public opinion of AI's moral status under the law*
Principal Investigator: Meghan Hurley (PI proxy)

Introduction and Study Overview

Thank you for your interest in our neuroethics and artificial intelligence research study. We would like to tell you everything you need to think about before you decide whether or not to join the study. It is entirely your choice. If you decide to take part, you can change your mind later on and withdraw from the research study.

The purpose of this study is to develop a tool to assess the public opinion of AI's moral status and agency by gaining a better understanding of the themes and concepts associated with the moral agency of artificial intelligence by both the public and experts in various related fields. The study is funded by the Neuroscience and Behavioral Biology program at Emory University. This study will take about 1 hour to complete.

If you join, you will be asked to respond to questions from an assessment tool regarding your opinion and conceptualization of artificial intelligence and its moral status. After the completion of these questions, you will be asked to reflect on important themes and concepts that came to mind when answering the first set of questions, as well as to discuss the clarity and effectiveness of the assessment tool.

Potential risks of participating in this study include breach in confidentiality because the data that we collect will be stored online. The data, however, will be protected to the best of our ability and recordings of the interviews will be deleted immediately after transcription. There are also little to no psychological risks associated with the study.

Upon completion of the interview, participants will be compensated with a $20 gift card as a token of our appreciation. This study is not intended to benefit you directly, but we hope this research will benefit people in the future.

Study records can be opened by court order. They also may be provided in response to a subpoena or a request for the production of documents. Certain offices and people other than the researchers may look at study records. Government agencies and Emory employees overseeing proper study conduct may look at your study records. These offices include the Emory Institutional Review Board and the

Emory Office of Research Compliance. Study funders may also look at your study records. Emory will keep any research records we create private to the extent we are required to do so by law. A study number rather than your name will be used on study records wherever possible. Your name and other facts that might point to you will not appear when we present this study or publish its results.

We will disclose your information when required to do so by law in the case of reporting child abuse or elder abuse, in addition to subpoenas or court orders.

De-identified data from this study (data that has been stripped of all information that can identify you) may be placed into public databases where, in addition to having no direct identifiers, researchers will need to sign data use agreements before accessing the data. We will remove or code any personal information that could identify you before your information is shared. This will ensure that, by current scientific standards and known methods, it is extremely unlikely that anyone would be able to identify you from the information we share. Despite these measures, we cannot guarantee anonymity of your personal data. Your data from this study may be useful for other research being done by investigators at Emory or elsewhere. To help further science, we may provide your deidentified data to other researchers. If we do, we will not include any information that could identify you. If your data are labeled with your study ID, we will not allow the other investigators to link that ID to your identifiable information.

In general, we will not give you any individual results from the study of the data you give us.

Contact Information

If you have questions about this study, your part in it, or if you have questions, or concerns about the research you may contact the following:

Meghan Hurley, Principal Investigator Proxy: 412-855-3639
Gillian Hue, Principal Investigator: ghue@emory.edu

Consent

Do you have any questions about anything I just said? Were there any parts that seemed unclear?

Do you agree to take part in the study?

Participant agrees to participate: Yes No

If Yes:

Name of Participant

Signature of Person Conducting Informed Consent Discussion Date Time

Name of Person Conducting Informed Consent Discussion