2023 Annual Conference & Exposition
Baltimore Convention Center, MD | June 25 - 28, 2023

The Harbor of Engineering
Education for 130 Years

ASEE

Paper ID #36947

# Personhood at the Extremes

**Dr. Suzanne Keilson, Loyola University, Maryland**

Suzanne Keilson is a faculty member at Loyola University Maryland. Her background and degrees are in Applied Physics and her research interests include signal processing, biomedical and materials engineering, design, STEM education and assistive technologies.. She has served in the Mid-Atlantic section of ASEE for a number of years and is active in ASME and IEEE activities.

# Personhood at the Extremes

## Abstract

This paper investigates the implications of competing definitions of 'personhood' for technology, specifically artificial intelligence (AI) agents and ways in which their legal and moral status may evolve over time. This exploration was the initial basis for a course in a liberal studies program. The basic structure of that course will be presented, including readings. An important starting point for the course and discussion was to look at historical, philosophical, and religious definitions of a person. One of the more natural points of comparison was and continues to be how we regard the status and rights of animals. The questions become one of setting boundaries for categorization. For example, is the boundary about intellectual capacity? How would that be defined? What are the defining hallmarks of cognition? Is it language or logic or something else? And what is the role and importance of physically embedded sensation and perception? What of these features do AI agents possess or are likely to possess? The animal rights movements and legal protections for pets and animals may serve as a template for exploring what may be eventually likely for such artificial agents. The ability to feel, both positive and negative, pleasure and pain, has been brought into arguments about regulating our relationship with the living world and how far ownership and domination may extend. It is also useful to remember earlier understanding of rights, humanity, and personhood of women, children, and slaves and the ways in which that understanding has evolved in Western thought and legal systems. Certainly, the personhood of artificial lifeforms has been a staple of science fiction books, television, and movies since Frankenstein, but the import of such moral thought experiments is often dismissed as irrelevant when discussing the status of artificial agents and ways in which moral guidance will be instilled into such semi-autonomous beings. Isaac Asimov's laws of robotics are marginally used as a starting point for such discussions, however seeing what is missing in his statement of the problem can be productive. Generally, in Western thought it seems that primacy is given to individual interaction and decision making and little emphasis is placed on the expanding circles of obligation from family, kin, tribe, nation, humanity as a whole. The issues raised here are not just a sterile intellectual exercise but have real consequences as we wrestle with programming decision making in such agents as autonomous cars and prioritizing associated legal and moral goods and virtues.

## General Introduction and Background

The point of view this paper presents is that a quest for general AI is situated in our moral and ethical imagination in the ways in which we understand our relations with animals and other humans. A course in a liberal studies program at the post bachelor's level at Loyola University Maryland explored the question of where personhood is situated and what are its key distinguishing characteristics. Course discussions and readings revolved around questions such as what are the obligations of society towards persons as well as what are the responsibilities of those recognized as persons towards society? What is the mutual set of rights and responsibilities that flow in multiple directions? These questions may seem to be the provenance of science fiction, speculation, philosophy, or theology, but as our machines, AI software and approaches to general AI become more sophisticated these are questions that need to be explored by a wide range of constituents, including, philosophers, students, engineers, businesspeople, and the public.

The extremes that were considered in the course were the animal and the mechanical. As advances in AI have progressed the definition of cognition, sentience, and personhood have also been adjusted and refined. It is often said that whenever we approach a new achievement in machine capability, we move the goalpost of the requirements for intelligence. Such changes can have real impacts in various legal arenas, as well as the debates surrounding AI and our machines.

Historically, there has been a focus on cognition and rationality as being the hallmarks of intelligence and personhood with its accompanying legal rights. It was ‘understood’ that animals, pets, and even children, women, slaves, and others did not have the same degree of capability for rational thought. Even as legal standards to protect and empower animals, pets, and children have been codified, part of the legal underpinning includes an assumption of diminished capacity and a need for special protections. Children are understood to represent potential but are not held responsible for actions in the same way as adults. Animal rights and protections have been expanding, but again, they are not generally held responsible for their actions. If an animal is considered dangerous to the public, it can suffer the consequence of euthanasia. Clearly, we have expanded and improved our understanding of persons and living things deserving of respect, but what about our machines? A course was developed in 2017 to investigate this.

## Course Background and Description

The course catalogue description was as follows:

> Humans have persisted in thinking of themselves as a species apart, but what makes humankind unique, both individually and as a species, remains unclear. Advances in neuroscience and computer science, as well as ethics, generate questions about the nature of intelligence, consciousness, and personhood and the rights and protections associated with being human. In this course students tackle classic readings from Descartes to modern ruminations on artificial intelligence, examine our relation to our creations and pets, and the way our various identities affect how our personhood is perceived and protected.

Some basic information from the course syllabus is described below.

## Broad Topics covered

1. Basic neuroscience
2. Distinctions between humans and nonhumans
3. Emotional connection and dependencies between humans and nonhumans
4. Definition and developments in artificial intelligence

5. Historical philosophical arguments about unique elements of personhood
6. Moral and legal implications of various visions and definitions of personhood

**Conduct of the Course:**

> In order to achieve the education objectives of the course you must come to class prepared to participate. This involves, regular attendance, regular reading, regular writing, regular participation. The quality of the class depends upon your engagement. If you will be absent, please let me know ahead of time. If you are absent more than one class meeting your grade will be affected.

Reading, writing, and conversation were the focus of the course and so the following breakdown of activities and grading were used. The aim was to not put undue emphasis on a single activity or mode of learning and expressing achievement.

- Research Paper: draft and final version (200 points =28% )
- Oral Presentation based on paper topic (100 points = 14%)
- Discussion leader of an article you choose (100 points = 14%)
- Class Participation (100 points = 14%)
- Written responses/Posts (100 points = 14% )
- Reading Quizzes (100 points = 14%)
- Total Points: 700 points

To give a sense of the breadth of topics that might be addressed in such a course and area of intellectual inquiry the following list of suggested topics for a research paper was presented to the class:

Suggested Broad Research Paper Topics that the idea of "Personhood" touches upon:

1. Corporate status as a legal person
2. Nonhuman animal status
3. Nonhuman companion rights
4. Fetal status
5. Women's movement and personhood
6. Capital Punishment and the moral dimension of personhood
7. Slavery and personhood
8. Race and the "alien" person
9. Artificial intelligence status as persons
10. Status of disabled
11. Status of children
12. End-of-life bioethics and personhood
13. Property, personhood, and environmental protection
14. What do inanimate objects teach us about personhood

Table One presents a sampling of authors and texts possible for a course that explores these issues. The selections with the asterisks were primary texts for the course, of which there were four. We started with some background in Neuroscience and Michael Gazzaniga's book, "Human: The Science Behind What Makes Us Unique". Next, we read the early philosophical foundations of Peter Singer in "Animal Liberation". We then considered a more full-blown exploration in the theological and philosophical history of understandings of personhood with the broad survey work by Joseph Torchia. Finally, we discussed the realms of speculative fiction and the hopes and concerns expressed by computer scientists and others in the compilation of "What to Think About Machines that Think". Numerous additional readings are easy to find in general circulation magazines such as *The New Yorker* or *The Atlantic*. There is more than ample material for reading, discussion, reflections, and student self-directed projects. A small sampling of those articles is provided in Table Two and it can easily be filled in with more current work.

The course was only offered once and with a small enrollment, so it is difficult to provide much in the way of assessment data or even suggestions for the next course offering as the graduate program was closed. One student in the course did take the course paper and expand it into a master's thesis topic looking at the role of altruism and its motivations. He conducted qualitative research with interviews and analyses of motivations for alumni giving in higher education and considered what of those drives might be significant for future general AI.

Table One: Initial Course Bibliography

| | |
|---|---|
| Citizen Canine | David Grimm |
| Some we love, some we hate, some we eat | Hal Herzog |
| The Emotion Machine | Marvin Minsky |
| Animals Make Us Human | Temple Grandin |
| Human: The Science Behind What Makes Us Unique | Michael Gazzaniga* |
| What Makes Us Human | Charles Pasternak |
| Exploring Personhood: Introduction to the Philosophy of Human Nature | Joseph Torchia* |
| In Search of Self | J. Wentzel Van Huyssteen |
| How to Create A Mind | Ray Kurzweil |
| Our Final Invention | James Barratt |
| The Emergence of Personhood | Malcolm Jeeves |
| Artificial Intelligence Simplified | Binto George |
| Ethics Into Action | Peter Singer |
| Animal Liberation | Peter Singer* |
| What to Think About Machines that Think | John Brockman* |

Table Two: Illustrative List of Additional Articles

| Title | Author | Journal/Magazine |
|---|---|---|
| Daniel Dennett's Science of Soul | Joshua Rothman | The New Yorker |

| | | |
|---|---|---|
| Get Smart: How will we know when machines are more intelligent than we are? | Adam Gopnik | The New Yorker |
| How Smart is an Octopus? | Olivia Judson | The Atlantic |
| If Animals Have Rights, Should Robots? | Nathan Heller | The New Yorker |

**Moral Machines**

The questions of goals, motivations, and intent are central to our understanding of independent and responsible agents. As we contemplate the possible consequential scenarios of devices such as self-driving cars, the questions of legal responsibility are important. Since such tools have a repertoire of actions, they exhibit decision making capabilities. To what extent are such actions truly autonomous? Who in the chain of system design, engineering, and programming, if anyone, is responsible for the outcomes of those actions?

An interesting online experiment [1] [2] that was not part of the course, asks the question of how one might program the decision making of a self-driving car. The results were culturally dependent in interesting ways. The question that was asked was whether a car should swerve to save pedestrians which increases the risk to the driver. The survey, called Moral Machines, asks about several different scenarios which vary in the number of people involved, their socioeconomic status and age. The results broke out into three geographic and cultural groups where one group preferred inaction on the part of the self-driving car while another group preferred pedestrians and the lawful and the third group prioritized females and high-status individuals. The two important points to this are that these kinds of programming questions are real, necessary and in our future and that there is no universal algorithm or values to be applied to such questions, as they are culturally dependent. One can imagine a future where machines are programmed based on such cultural differences and multinational corporations will perhaps have to implement different algorithms depending on the customers or country's preferences. Or perhaps you get to choose different modes for your car. Or go elsewhere to buy it. Perhaps you want to tweak the thresholds of decision making between the value of your life and that of pedestrians or other motorists. Underlying all of this are our assumptions about the nature of persons and humanity and which lives are valued.

One way to think about the issues raised here is to focus on the question of autonomy of decision-making and action. Such a focus can help clarify the various categories of 'persons' that exist legally and in our imaginations. Children (and women) might have been seen to have rational thought and cognition or the capacity for that, but they did not have the ability to act or put in practice many areas of their decision-making, intents, and desires. That understanding has clearly changed in Western legal thought and society generally.

The status of pets and animals has also been shifting. One reason that has occurred is because of our changing knowledge about animal cognition and emotional life, but also a recognition that animals have independent intent and desire. Peter Singer [3] and others have turned a focus upon animals' capacity for emotion and, especially, for pain and discomfort in equal measure to a human. Singer argued that such capacity elevated the moral standing of all non-human animals. This approach, while widening the category of moral or legal standing seems to obscure important differences. It certainly makes ethical decision making in the context of two competing goods more complex by raising the value threshold for nonhuman life and perhaps diminishing the value assigned to certain kinds of human life.

Measuring personhood solely by affective or cognitive domain requirements by themselves does not answer our questions or scenarios. The extremes of cognition (AI) without affect or of affect with minimal communication of cognition does not satisfactorily address our intuitive understanding of what makes a person. Back in 2017 an article [4]looked at what a global consumer base trusted AI to do, from a high of 79% approval for AI providing reminders about medications to a low of 20% approval for AI engaging in childcare. Even as these percentages may shift, this probably correlates well to our sense of AI being superior in the cognitive domain and certainly for repetitive sequential tasks, but not being preferred with tasks that require emotional intelligence [5]. At the other end of the spectrum, however much emotional intelligence the family pet may have, it still would not be entrusted with childcare duties, because of a recognized lack of cognition that engages fully with long-range future consequences.

## Implications for Engineering

I propose that the capacity for decision-making and the ability to implement intent are key determinants for helping us to differentiate various systems and scenarios involving engineered machines and software. Historically an engineered machine does not have any decision-making capability. This was obviously true for tools such as hammers or axes and was equally true for steam engines and all its associated applications. Early control systems and mechanical automata amazed but were still understood to be devoid of independent goal setting.

With the advent of general programmable digital electronic machines many problems could be solved that could not be approached before. As speed, memory and efficient algorithms improved so did the space of difficult problems that could be tackled. None of this implies anything about decision-making. Alan Turing envisioned a general programmable machine and suggested that the key to assigning it sentience would be its ability to conduct a natural language conversation to such an extent that the human partner could not distinguish the replies from the machine or from a person. It might be wondered whether or not a fast enough look-up table (dictionary) could fool such a Turing test and that some element of awareness of another mind that 'understands' the language is hard to define but understood to be missing. This dilemma is often referred to as John Searle's Chinese Room Problem [6]. [7]This remains true with our interactions with large scale natural language AI such as the ChatGPT that has generated so much current buzz. In this case it is not a brute force lookup but a prior machine learning from prior examples of natural language. That does not mean that these language AI can 'understand'. It could be that aspects of language such as humor or lying or subtext generally are the frontier where Chatbots fail [7].

Whether programming is sequential, parallel, object-oriented or any other variant, the branching of decision-making is constrained by what the programmer envisioned. There is no adaptation to events or the environment except by the conditionals that the human programmer had developed prior. The search tree and valuations are fixed in advance and not contingent upon circumstances. In industrial robotics, machines were developed that could use complicated control algorithms as they interacted more thoroughly with their environment using sensors and actuators. If accidents or liabilities occurred, they were still seen as the responsibility of the human programmer and manufacturer. The robot was not seen to have any independent sphere of action based on intent established in the moment.

It is the advent of machine learning and all its variations that now bring that into question. Even in cases of current early prototypes of self-driving or autonomous cars, we generally assign ultimate responsibility to either the programmer, manufacturer, or driver, not to the machine.

What is qualitatively different about machine learning is that the approach to optimizing an algorithm or approach to decision-making is left to either supervised or boot-strapped learning. In either case, the programmers often say that they do not know or understand why the machine came up with the algorithm, categorization, or decision that it implemented. Therefore, a certain unknowable, or independent element has entered the decision-making process. Although approval and actual implementation may yet remain in human hands, this raises several questions for engineers engaged in the design, development, and deployment of such machines. The first and foremost of these considerations is failsafe design that does not allow for solutions or actions deemed totally off-limits to the machines sphere of action.

When considering how to program values one approach might be to look at expected value calculations (EV). The expected value, positive or negative, to an event or decision is equal to the probability of the occurrence of the event multiplied by the impact of the event. Therefore, along each decision branch, one can calculate the overall EV of that decision based upon the probabilities of possible outcomes and their impact. For example, with machine or Bayesian learning, a machine might update its prior probabilities, on the fly with the advent of new information, but if one possible outcome has a disastrous consequence or impact it could disallow the entire decision regardless of probabilities.

**Engineering Safety Factors**

Every machine design or computer program considers safety. That is simply best engineering practice. It may involve various kinds of constraints or safety interlocks. If a user chooses to disable the safety features, that must be a conscious decision on their part. There is acknowledge responsibility on both the part of the designer and the part of the user. There is no reason to believe that autonomous machines or AI would be designed any differently. The scenarios beloved of moral and science speculation generally have a simple solution. The concern of machines running amok, having infinite loops like ‘the sorcerer’s apprentice’ is almost laughable. We know how to prevent infinite loops or have software quality and reliability test for those conditions. Concerns of machines with outsized egos or disabling their own safety protocols should be easy to answer by competent programmers. Similarly, if one is concerned about a machine taking a destructive approach to others, an appropriate tweak of the various expected values for outcomes along different decision tree paths could resolve those concerns. Often, industry thinkers seem to project that machines will have goals and motivations towards power, but it is not clear why that would be the most likely outcome of general AI or that AI could not be taught to value autonomy or other things that are culturally situated.

The question of how to socialize AI and give it emotional-social education is something that has been and continues to be an area of active research by researchers, including with reviews of the field [8].

### Machine Learning and Training Sets

There are a few more novel concerns to be studied. Among them are the appropriate nature, size, and diversity of any sampling used in machine learning. We already have evidence of 'the coded gaze' where skewed sampling has resulted in statistically significant differential error rates in facial recognition. While we expect correct identification rates of 90% or greater, that is only true for white males, whereas black females are either not recognized or misgendered at rates approaching 40% for commercial software [9]. This can have real and disturbing consequences. Similar problems exist in samples for voice recognition programs which are not that flexible regarding accent or speech production disabilities. Other natural language or more general AI programs have been taken out of 'circulation in the wild' because the training data consisted of uncurated or vetted text from the internet which led to concerns about misinformation, offensive and disturbing interactions [10]. Technical personnel will need to have a better understanding of these concerns as they develop the software and machines. Safety of machine interactions now must include concerns about psychological and social distress and impacts. One of the more difficult questions facing us is who will be making decisions that place those guardrails on AI. These problems are as real and considerable as those involving decision-making in autonomous driving.

### Future Considerations

Culture, social convention, education, religion, and law act to constrain human action generally. Similar training and education, updated Asimov's Laws of Robotics [11] say, can make autonomous machines partners and not rogues in our society. In the original formulation individual human life was sacrosanct above all. Those so-called laws as originally stated will have to be expanded to consider valuation of many types of lives more carefully, 'persons', social goods and interactions among social circles of increasing size and weaker connections.

The utilitarian approach to ethics favored by philosopher Jeremy Bentham [12] has an appeal to American culture but is not the only way to understand such decisions. Even taking such an approach, we tend to deeply discount future events and generations and the weighting or importance that is given to certain improbably adverse scenarios. Such calculations for overall good, a kind of calculus of decision-making under uncertainty may need to significantly increase the weight given to future generations and the probabilities of improbable, but dire, outcomes should not be conveniently set to zero. This can leverage the work of Kahneman and Tversky as to how humans tend to judge risk and probability under uncertainty, called Prospect Theory [13].

Kenneth Feinberg, the attorney appointed by John Ashcroft, wrote a book about the difficulties of determining what a life is worth in the context of victim compensation for the September 11, 2001families [14]. Although the circumstances were far removed from AI, the question of human, animal, and machine valuation is still very relevant. It is not clear that the utilitarian approach will work or be accepted. Engineers will need to understand these issues just as they are educated in systems engineering about business cases, simulations, and sensitivity analyses.

For humans and machines to interact in positive and useful ways, considerations of intent and ultimate goals (on the part of governments, businesses, individuals, culturally diverse ways of seeing and organizing reality) and of affective interactions (emotional robotics) will be crucial areas for the education of engineering students. At some point, it may well be, that like with women, children, slaves, workers, animals, pets, some weight will have to be given to the value of the independent actions and decision-making abilities of our machines with a recognition of their personhood, legally and morally. They can then be both protected and culpable in their actions. They also would not be multi-million-dollar devices to be disposed of heedlessly as we have done in the past with human lives and potential. The challenge is in what ways will they compete economically with all of us and either disrupt or enhance our self-understanding and flourishing. Their expense, care and abilities may well devalue the human or can equally enhance our abilities, not just for work and productivity, but for more creative actions. Will the machines want to explore their own creative inner spaces as well? It is not inconceivable, and they may well develop a moral standing to be taken into consideration. It will be up to a partnership of broadly educated engineers and others to help navigate possible solutions and scenarios.

## References

[1] MIT, "moralmachine.net," [Online]. Available: https://www.moralmachine.net/. [Accessed 1 May 2023].

[2] G. Marcus, "Moral Machines," *The New Yorker,* 2012.

[3] P. Singer, Animal Liberation, New York: Avon Books, 1975-1079.

[4] J. Winters, "Tasks Trusted by Consumers for AI Performance," *Mechanical Engineering,* p. 27, January 2017.

[5] D. Gursoy, O. H. Chi, L. Lu and R. Nunkoo, "Consumers acceptance of artificially intelligent (AI) device use in service delivery," *International Journal of Information Management,* vol. 49, pp. 157-169, 2019.

[6] J. Searle, "The Chinese Room Revisited," *Behavorial and Brain Sciences,* 1982.

[7] M. DeWitte, "Stanford News: How will ChatGPT change the way we think and work? Stanford scholar examines," February 2023. [Online]. Available: https://news.stanford.edu/2023/02/13/will-chatgpt-change-way-think-work/. [Accessed April 2023].

[8] M. Spezialetti, G. Placidi and S. Rossi, "Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives," *Frontiers in Robotics and AI,* vol. 7, pp. 1-11, 2020.

[9] L. Hooberman, "The Coded Gaze: algorithmic bias, facial recognition and beyond: How research can change the law and influence people," in *WebSci '21: 13th ACM Web Science Conference 2021*, 2021.

[10] G. Neff and P. Nagy, "Talking to Bots: Symbiotic Agency and the Case of Tay," *International Journal of Communication,* vol. 10, pp. 4915-4931, 2016.

[11] R. Murphy and D. Woods, "Beyond Asimov: The Three Laws of Responsible Robotics," *Intelligent Systems, IEEE,* vol. 24, no. 4, pp. 14-20, 2009.

[12] J. Bentham, The collected works of Jeremy Bentham: Deontology together with a table of the springs of action and the article on utilitarianism, clarendon press, 1983.

[13] D. Kahneman and A. Tversky, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica,* vol. 47, no. 2, pp. 263-292, 1979.

[14] K. Feinberg, What Is Life Worth?: The Inside Story of the 9/11 Fund and Its Effort to Compensate the Victims of September 11th, PublicAffairs, 2005.