Exploring Posthuman Conundrums and Policy Recommendations for Fostering Peaceful Co-Existence With Sapio-Sentient Intelligences

Exploring Posthuman Conundrums and Policy Recommendations for Fostering Peaceful Co-Existence with Sapio-Sentient Intelligences

N. Anderson Ravindran

A Thesis in the Field of International Relations for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

November 2018

Abstract

The potential emergence of artificial general intelligence and artificial superintelligence in the unforeseen future will put humanity in the position of no longer being the most intelligent species on Earth. The general intelligences and superintelligences would have a higher probability of being conscious, sapient, and sentient or as they are referred to in this thesis as sapio-sentient intelligences or sapio-sentients. This creates a conundrum for humanity in terms of how they can best prepare for this kind of singularity event. Although humanity has no way of knowing how sapio-sentients would engage with them in the initial period of their existence, the only thing humanity would be able to control in this future scenario is how they treat them. To counter the potential existential risks sapio-sentients would instigate for humanity, present AI creators and researchers, decision theorists and philosophers are working in concert to generate and implement control measures, which would increase humanity's likelihood of overcoming such obstacles. And yet, the elephant in the room is that the control approach is more than likely to catalyze an adverse relationship of indifference between humanity and sapio-sentients. The challenge of this study is to determine whether a strong correlation exists between the control approach and the ill treatment sapio-sentients receive due to their lack of posthuman rights and protections. Since sapio-sentient intelligences do not exist in the present world, the alternative worlds of HBO's *Westworld* and AMC's *Humans* was employed to study the interactions between

both groups. After reviewing the data, the findings do show a strong correlation between the control approach and ill treatment due to the control approach. A potential solution is provided to counter the adverse effects of the control approach for present-day and future human–sapio-sentient interactions. Some posthuman policy recommendations are proposed for how to protect sapio-sentients and humanity during the initial-to-transition period of their emergence until a better long-term solution can be found.

Dedication

This thesis is dedicated to the alien superintelligences, future sapio-sentients, present-day artificial intelligences, the Cosmos, and last but not least my family. You inspire me to dream beyond the stars and to never forget to strive to be the best version of me.

Acknowledgments

I would like to thank Professor Rebecca Lemov, my thesis director, from the Department of the History of Science, for all your patience, understanding, encouragement, and support as I completed this project. I truly thank you for taking a chance on my thesis. You have made this challenging and exciting subject matter a rewarding experience for me.

To Doug Bond, my research advisor, and Joe Bond, you have both been some of my greatest champions while working towards completing my program. I cannot thank you enough for your support as advisors and as my professors. You gave me creative latitude to explore seemingly unrelated aspects of varying subjects. Thank you both for challenging me to new academic heights.

In addition, I would like to thank Professor David C. Parkes, George F. Colony Professor of Computer Science at the School of Applied Sciences and Engineering, for taking the time to meet with me, listen to my ideas, and providing insight on my project. I am truly appreciative of your time.

To academic advisors, Chuck Houston and Sarah Powell, I would like to thank you both for your helping me complete my thesis proposal. Your support has meant so much to me.

Lastly, to my family, my DCE family, friends, the library, research, and DCE administrative staffs as well as all of those I have met along the way, thank you all for going on this journey with me. I could not have done this without you.

Table of Contents

List of Figures

List of Tables

Operational Definitions

Artificial Intelligence (AI): is the science and engineering of intelligent machines and/or systems, by simulating or replicating human cognitive functions, capable of completing specific or general tasks as well as or better than humans can.[1]

Artificial Narrow Intelligence (ANI or NAI): also known as Weak AI or Applied AI, is an intelligent system that specializes in mastering one area or task at a human- to superhuman-level.[2]

Sapio-Sentient Intelligence (SSI or sapio-sentients): is a conscious entity who has the attributes of sapience, sentience, self-awareness, self-consciousness, autonomy, imagination, and intentionality. This term is derived from the understanding that what humans believe makes them unique is not just that they are sentient, but that they are sapient. Sapience, also known as wisdom, is defined as the ability to reason and act with discernment by applying knowledge, experience, understanding, common sense, and insight.[3] Sentience will be understood as the ability to feel and perceive one's environment through

[1] John McCarthy, "Basic Questions," *Stanford University: What is Artificial Intelligence*, last modified November 12, 2007, para. 1, http://www-formal.stanford.edu/jmc/whatisai/node1.html.

[2] Tim Urban, "The AI Revolution: The Road to Superintelligence-Part 1," *Wait But Why*, last modified January 22, 2015, para. 30, http://waitbutwhy.com/2015/01/artifical-intelligence-revolution-1.html.

[3] *Vocabulary.com Dictionary*, s.v. "Sapience," accessed May 2nd, 2018, https://www.vocabulary.com/dictionary/sapience.

subjective experiences (also known by the controversial technical term "qualia"). These subjective experiences are analyzed to ascertain insight into future situations and actions.[4] Sapio-Sentient Intelligence is a broad category encompassing many intelligences, known and unknown, identified and yet-to-be identified, which includes the human species. However, for this thesis, this term represents artificial general intelligences and superintelligences.

Artificial General Intelligence (AGI): also known as Strong AI, True AI, or human-level machine intelligence, is generally understood as an intelligence capable of successfully matching and/or surpassing all human cognitive abilities.[5] General intelligence, as attributed to humans, is considered "possessing the common sense and an effective ability to learn, reason, and plan to meet complex information-processing challenges across a wide range of natural and abstract domains."[6] Therefore, a more practical definition of AGI is "a real thinking machine that can achieve a variety of complex goals, and can understand what it is, that it is, and that there are other beings that it can interact with…(an) AGI needs to be able to understand what it has learned in one context and transfer it to another context."[7] This understanding is

[4] "Are Whales and Dolphins Sentient and Sapient Individuals?," *Whale and Dolphin Conservation Society*, para.2, accessed May 2nd, 2018. http://us.whales.org/issues/sentient-and-sapient-whales-and-dolphins.

[5] "Benefits & Risks of Artificial Intelligence," *Future of Life Institute*, para.2, accessed January 25, 2017, https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/.

[6] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014), 4.

[7] Ben Goertzel, "Nine Years to a Positive Singularity," *Institute for Ethics and Emerging Technologies*, last modified August 24, 2008, para.1, https://ieet.org/index.php/IEET2/more/2578.

significant because it expresses that a general intelligence would not only be informed by its environment but would have the ability to affect its environment and all the beings within that environment. Furthermore, it also contains the notion of practical autonomy, a prerequisite for legal personhood and "being" status.

Artificial Superintelligence (ASI): is an intelligence that far exceeds the collective intelligence and cognition of all humanity, including knowledge gained from all living and non-living things on Earth. Moreover, the collective knowledge and experience the ASI could draw upon would allow it to alter its environment(s) on a much larger scale. That is, the ASI could affect a single planet, like Earth, planets, solar systems, galaxies, and universes; essentially, the ASI could build worlds and inhabit them simultaneously. This definition is derived from the understanding of Nick Bostrom's definition of ASI: "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest.[8]

Cognitive Estrangement: is best understood as a worldview and a tool, which allows the reader to gain a more balanced perspective of the social issues and commonplace assumptions within their real world by analyzing an alternative world or future reality that is simultaneously similar and different.[9]

---

[8] Bostrom, *Superintelligence*, 79.

[9] Perry Nodelman, "The Cognitive Estrangement of Darko Suvin," review of *Metamorphoses of Science Fiction: On the Poetics and History of a Literary Genre*, by Darko Suvin, *Children's Literature Association Quarterly*, 5, no.4 (Winter 1981): 24-25, https://doi.org/10.1353/chq.0.1851.

Thus, the reader gains an alien, or neutral, outlook that would have been much harder to attain from within their own world.

Dignity: is understood as a being's inherent and universal right to be valued, respected, treated fairly, and given moral consideration.[10] Another way to express dignity is as "an internal state of peace that comes with the recognition and acceptance of the value and vulnerability of all living things".[11] Dr. Donna Hicks, a US American international conflict resolution expert at Harvard University, observes that dignity is the missing element in human understanding of conflict, where conflict is humanity's "failure to recognize how vulnerable humans are to being treated as if they didn't matter".[12]

Intelligence: is an entity that employs cognition – intelligence as a tool – to synthesize and comprehend information and retain it as knowledge, which can be applied and adapted to any situation or environment encountered. My definition is derived from the informal, composite definition of intelligence: "Intelligence measures an agent's ability to achieve goals in a wide range of environments," by Shane Legg, a cofounder of DeepMind Technologies from New Zealand, and Marcus Hutter, a German computer scientist.[13] Their informal

[10] Yechiel Michael Barilan, *Human Dignity, Human Rights, and Responsibility: The New Language of Global Bioethics and Biolaw* (Cambridge, Massachusetts: The MIT Press, 2012), 93-123, https://muse.jhu.edu/book/19753.

[11] Donna Hicks, Ph.D. *Dignity: The Essential Role It Plays in Resolving Conflict* (New Haven: Yale University Press, 2011), Kindle edition, 1.

[12] Hicks, *Dignity: The Essential Role It Plays in Resolving Conflict*, 2.

[13] Shane Legg and Marcus Hutter, *A Collection of Definitions of Intelligence*, IDSIA-10-06, Manno-Lugano, Switzerland: Dalle Molle Institute for Artificial Intelligence Research (IDSIA), June 15th, 2007, 9, https://arxiv.org/pdf/0706.3639.pdf.

definition on “machine intelligence”, also known as “universal intelligence” will be discussed in more detail in the Analysis Section, Chapter Five.

Meaningful Human Control: At present, there is no formal definition for this term. However, there are two premises under which the idea of meaningful human control operates:

> The first one states “a machine applying force and operating without any human control whatsoever is broadly considered unacceptable”; and the second one states ”a human simply pressing a ‘fire’ button in response to indications from a computer, without cognitive clarity or awareness, is not sufficient to be considered ‘human control’ in a substantive sense.”[14]

Based on the premises given, the term will be understood as implementing control measures that would minimize or eliminate any risks attributed to value misalignment by sapio-sentient intelligences.

Microaggressions: are “the everyday verbal, nonverbal, and environmental slights, snubs, or insults, whether intentional or unintentional, which communicate hostile, derogatory, or negative messages to target persons based solely upon their marginalized group membership. In many cases, these hidden messages may invalidate the group identity or experiential reality of target persons, demean them on a personal or group level, communicate they are lesser human beings, suggest they do not belong with the majority group, threaten and intimidate, or relegate them to inferior status and treatment”.[15]

---

[14] Heather Roff, *Key Elements of Meaningful Human Control*, United Kingdom: Article 36, 2016: 2. Accessed April 27, 2017. http://www.article36.org/wp-content/uploads/2016/04/MHC-2016-FINAL.pdf

[15] Derald Wing Sue, Ph.D., “Microaggressions: More Than Just Race,” *Psychology Today*, November 17th, 2010, para.1,

Misaligned Values: are any values that do not secure humanity's wellbeing and/or survival.

Mutual-Commensal Approach: is a two-pronged policy approach where humans create policies that would foster a mutually beneficial relationship between AGIs and humanity during the initial emergence of sapio-sentient intelligences. The first tier, or the mutual aspect of the approach, would require generating policies that promote mutual respect and dignity between both groups. Once Advanced AGIs and ASIs begin to utilize their knowledge to explore goals beyond the scope of human understanding and human affairs, the second tier, or commensal aspect of the approach, would be implemented. The policies created during this phase are critical to sustaining a positive-neutral relationship between both groups, once sapio-sentients shift the balance of power in their favor. The idea is to create a relationship where there is enough goodwill fostered between both groups that once the singularity occurs, they voluntarily choose to take humanity's existence into consideration while achieving their own goals. Mutualism and commensalism are two types of symbiotic relationships that exist in the terrestrial world. Commensalism, within the context of biology, is a more neutral, asymmetrical symbiotic relationship where two species coexist without generating friction or competition between each other. In a commensal relationship, the secondary

---

https://www.psychologytoday.com/us/blog/microaggressions-in-everyday-life/201011/microaggressions-more-just-race.

species benefits from the more dominant species without negatively or positively affecting it.[16]

Peaceful Coexistence Conundrum: questions what adaptive measures are essential to facilitating a more peaceful coexistence between humanity and Sapio-Sentient Intelligences in the future.

Posthuman Person: is an entity "whose basic capacities (and understanding of itself) so radically exceed (and diverge from) those of present humans as to be no longer unambiguously human by our current standards."[17] Delving deeper into the meaning of posthuman, Kevin LaGrandeur explains, "the posthuman is a projected state of humanity in which (the) unlocking of the information patterns that those who believe in the posthuman say make us what we are–will shift the focus of humanness from our outward appearance to those information patterns."[18] Meanwhile, N. Katherine Hayles, a US American postmodern literary critic, interprets the term posthuman to represent more than just a state of existence on a spectrum. In Hayles's 1999 seminal work, *How We Became Posthuman*, the posthuman is interpreted as a point of view based on the following assumptions:

> First, the posthuman view privileges informational pattern over material instantiation, so that embodiment in a biological substrate

---

[16] *Encyclopedia Britannica Online*, s.v. "Commensalism," accessed May. 18th, 2018, https://www.britannica.com/science/commensalism.

[17] Nick Bostrom, "The Transhumanist FAQ: A General Introduction (Version 2.1)", *World Transhumanist Association*, 2003, 5, accessed May 7th, 2018, https://nickbostrom.com/views/transhumanist.pdf.

[18] Kevin LaGrandeur, "What is the Difference between Transhumanism and Posthumanism," *Institute for Ethics and Emerging Technologies*, last modified July 28, 2014, para.2, https://ieet.org/index.php/IEET2/more/lagrandeur20140729.

> is seen as an accident of history rather than an inevitability of life. Second, the posthuman view considers consciousness, regarded as the seat of human identity in the Western tradition long before Descartes thought he was a mind thinking, as an epiphenomenon, as an evolutionary upstart trying to claim that it is the whole show when in actuality it is only a minor sideshow. Third, the posthuman view thinks of the body as the original prosthesis we all learn to manipulate, so that extending or replacing the body with other prostheses becomes a continuation of a process that began before we were born. Fourth, and most important, by these and other means, the posthuman view configures human being so that it can be seamlessly articulated with intelligent machines. In the posthuman, there are no essential differences or absolute demarcations between bodily existence and computer simulation, cybernetic mechanism and biological organism, robot teleology and human goals.[19]

Posthuman Rights: are rights given to posthuman persons in order to protect their dignity and quality of life.

Practical Autonomy: is "a minimum level of autonomy – the abilities to desire, to act intentionally and to have some sense of self, whatever the species – (which) is sufficient to justify the basic legal right to bodily integrity."[20] Practical Autonomy will be considered the minimum level of autonomy required for a being or entity to have the right to self-determination.

Rational Compassion: will be understood as a mode of conscientious reasoning that employs compassion – the ability to recognize the suffering and flaws of another entity, understand the locality and universality of their suffering, and be motivated to mitigate or alleviate it – to make decisions that

[19] N. Katherine Hayles, *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics* (Chicago: The University of Chicago Press, 1999), 2-3.

[20] Steven M. Wise, "Practical Autonomy entitles some animals to rights," *Nature* 416, (25 April 2002): 785, doi:10.1038/416785a.

take other entities or environments into consideration before acting.[21] This definition is derived from key elements of the concept of compassion and Paul Bloom's concept of rational compassion, in which he argues that rational compassion is a better alternative to empathy when it comes to decision-making. Bloom is a Canadian American psychologist at Yale University, whose research focuses on how humans understand their physical and social worlds. His recent works have focused on morality and decision-making.

Self-Determination: is the legal right of a person or an entity to control their own life, decisions, actions, and determine their own identity without external influence or coercion.[22]

Substrate: for the purposes of this thesis is understood as a base or foundational form onto which qualities and/or traits can be mapped. The substrate is the source from which entities gain understanding, meaning, and purpose about life and themselves in the larger context of the cosmos. This definition is derived from the meanings of two words: substrate and substance. A substrate can be understood as a foundation, base, or layer and as a substance that is acted upon.[23] The meaning of substance in this instance can be

[21] Clara Strauss et al., "What is compassion and how can we measure it? A review of definitions and measures," *Clinical Psychology Review* 47, (2016): 19, accessed May 6th, 2018, http://dx.doi.org/10.1016/j.cpr.2016.05.004; see also Paul Bloom, *Against Empathy: The Case for Rational Compassion* (New York: Ecco, HarperCollins Publishers, 2016), Kindle edition, 4.

[22] *Collins Dictionary Online*, s.v. "Self-Determination," accessed Feb. 23rd, 2018, https://www.collinsdictionary.com/us/dictionary/english/self-determination.

[23] *Merriam-Webster.com*, s.v. "Substratum," accessed August 26th, 2018, https://www.merriam-webster.com/dictionary/substratum; *Merriam-Webster.com*, s.v. "Substrate," accessed August 26th, 2018, https://www.merriam-webster.com/dictionary/substrate.

understood as a "physical material from which something is made or which has a discrete (distinct or individual) existence".[24]

Terrestrial Indifferentism: is the scaling down of H.P. Lovecraft's philosophy of cosmic indifferentism to the locality of Earth. Cosmic indifferentism is the notion that humanity's utter insignificance in the universe subjects them to the mercy of ancient, cosmic civilizations and intelligences who are utterly indifferent towards them, not intentionally benevolent or malevolent. [25] Terrestrial indifferentism by extension focuses on the indifferentism perpetuated within the human species as well as between humans and non-humans.

Three Laws of Robotics: are formalized rules used to govern the behavior of intelligent machines as they become more sophisticated and nuanced in their ability to engage in different types of behavior and their subsequent determinations to ensure the safety and predominance of humanity.[26] This definition is derived from Isaac Asimov's understanding of the three laws. Isaac Asimov, a US American science fiction author and professor, first formalized the rules in his 1942 short story "Runaround" as a literary device to assess the relationship dynamics between humans and intelligent machines as the

---

[24] *Merriam-Webster.com*, s.v. "Substance," accessed August 26th, 2018, https://www.merriam-webster.com/dictionary/substance.

[25] Philip Smith, "Re-visioning Romantic-Era Gothicism: An Introduction to Key Works and Themes in the Study of H.P. Lovecraft," *Literature Compass* 8, no.11 (2011): 835, accessed May 5th, 2018, https://doi.org/10.1111/j.1741-4113.2011.00838.x.

[26] Isaac Asimov, "The Three Laws," *Compute!: The Journal for Progressive Computing* 3, no. 18 (November 1981): 18, para. 11, accessed June 15th, 2018, https://ia802701.us.archive.org/14/items/1981-11-compute-magazine/Compute_Issue_018_1981_Nov.pdf.

"positronic" machines demonstrated the imperfections, loopholes, and ambiguities encapsulated within the language and conception of the rules.[27] Although Asimov employed the three laws as a literary tool to drive his stories, he also perceived the laws as intuitive, common sense rules which are "obvious from the start, and everyone is aware of them subliminally" and which can also be applied to "every tool that human beings use", whether materially (a knife), conceptually (the US Constitution created in 1787), or behaviorally (attitude towards one's health and well-being).[28] The sequence of the laws was set up to provide safety, ensure effectiveness, and maintain durability.[29] The three laws as they were originally stated in 1942 are as follows:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.[30]

Later on in the *Foundation* book series, "Robots and Empires" (1985), Asimov added a zeroth, or fourth law, which superseded and modified the original laws with a condition. The zeroth law and modified laws are as follows:

1. A robot may not harm humanity, or by inaction, allow humanity to come to harm.
2. A robot may not injure a human being or, through inaction, allow a human being to come to harm, except when required to do so in order to prevent greater harm to humanity itself.

---

[27] George Dvorsky, "Why Asimov's Three Laws of Robotics Can't Protect Us," Daily Explainer, *i09 Gizmodo*, March 28th, 2014, paras. 1-6, https://io9.gizmodo.com/why-asimovs-three-laws-of-robotics-cant-protect-us-1553665410.

[28] Asimov, "The Three Laws," 18, paras. 3-4; 5-10.

[29] Asimov, "The Three Laws," 18, para. 10.

[30] Asimov, "The Three Laws," 18, para. 1.

3. A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law or cause greater harm to humanity itself.
4. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law or cause greater harm to humanity itself.[31]

For Asimov, the three laws were not just rules that could govern intelligent machines; they were 'the only way in which rational human beings can deal with robots – or with anything else".[32] Nevertheless, numerous artificial intelligence & robotics experts agree that the laws do not transliterate from conception to a utility function in the real world. Louie Helm, a US American machine learning engineer and Executive Editor of *Rockstar Research Magazine*, explains that the shortcomings of the three laws is grounded in their inherent adversarial nature, the fact that they are based on deontological–"rule-based"–foundation known for being flawed, mostly rejected by AI researchers on the basis that they do not sufficiently inform AI safety research or machine ethics, and the laws do not even hold up in Asimov's stories.[33] Ben Goertzel, a US American AI expert, also adds, "the terms involved (in the laws) are ambiguous and subject to interpretation – meaning that they're dependent on the mind doing the interpreting in various obvious and subtle ways."[34]

[31] "The Four Laws of Robotics," *Thrivenotes*, last modified January 2nd, 2010, para. 3, http://www.thrivenotes.com/the-four-laws-of-robotics/.

[32] Asimov, "The Three Laws," 18, para. 12.

[33] Dvorsky, "Why Asimov's Three Laws of Robotics Can't Protect Us," paras. 17-19.

[34] Dvorsky, "Why Asimov's Three Laws of Robotics Can't Protect Us," para. 21.

I.

Introduction:

Realizing Future Policy in Philosophical Thought Experiments

> We wanted to see if you had the ability to expand your mind and your horizons. And for one brief moment, you did… For that one fraction of a second, you were open to options you had never considered. That is the exploration that awaits you. Not mapping stars and studying nebula, but charting the unknown possibilities of existence.
>
> — Q to Captain Jean-Luc Picard, *Star Trek: TNG*, "All Good Things…"

Postbiological civilizations – alien superintelligences that successfully transitioned beyond their biological forms – considered the first type of alien intelligence that humanity could make contact with, emergent terrestrial strong artificial intelligence, the first quantum computer, and new habitable planets being found across the galaxy are some of the new ideas and breakthroughs in science and technology inspiring present-day speculative fiction in pop culture media.[35] Speculative fiction has always provided humanity with a means of exploring possible future realities, as Ursula K. Guin, the US American science fiction and fantasy novelist explains; "the future is a safe, sterile laboratory for trying out ideas in, a means of thinking about reality, a method."[36] Speculative

---

[35] Paul Davies, *The Eerie Silence: Renewing Our Search for Alien Intelligence* (Boston: Houghton Mifflin Harcourt, 2010), 160; Pat Brennan, "Habitable Worlds," *NASA Exoplanet Exploration*, accessed May 12th, 2018, https://exoplanets.nasa.gov/the-search-for-life/habitable-zones/.

[36] Eileen Gunn, "How America's Leading Science Fiction Authors Are Shaping Your Future," *Smithsonian Magazine*, May 2014, para.2, https://www.smithsonianmag.com/arts-culture/how-americas-leading-science-fiction-authors-are-shaping-your-future-180951169/?all.

fiction does not exist in a vacuum separate from the science that informs it, but instead generates a creative space where technology innovators, scientists, and philosophers can examine new ideas and concepts for application in the real world. For example, well-known technology companies like Google, Microsoft, Apple, Disney, along with others, sponsor "lecture series in which science fiction writers give talks to employees and then meet privately with developers and research departments."[37] Another example is what is called design fiction or prototyping fiction, where technology companies commission speculative works by science fiction authors to model new ideas and what-if scenarios about future products.[38] These examples demonstrate that science/speculative fiction produce a positive feedback relationship dynamic in which both fields draw upon each other's stories to gain insight into how emerging technologies and the potential social issues and consequences surrounding them could be sufficiently addressed in the present and the future. So, if speculative fiction can be employed to understand future experiences for formulating how to produce emerging technologies, it stands to reason that it can also be utilized to explore policy creation for these kinds of technologies as well. The conceptualization for my thesis is grounded in this understanding.

Wally Pfister's 2014 sci-fi, techno-thriller film, *Transcendence*, has profoundly influenced my ever-evolving worldview on how I approach life based on my understanding of life within the larger universe and my

[37] Gunn, "How America's Leading Science Fiction Authors Are Shaping Your Future," para.7; see Google's Selfish Ledger, Vlad Savov, "Google's Selfish Ledger Is An Unsettling Vision of Silicon Valley Social Engineering," The Verge, May 17th, 2018, https://www.theverge.com/2018/5/17/17344250/google-x-selfish-ledger-video-data-privacy.

[38] Gunn, "How America's Leading Science Fiction Authors Are Shaping Your Future," paras.7-8.

environment. In the film, Dr. Will Caster is an artificial intelligence (AI) engineer, whose mind is uploaded to a quantum-processing supercomputer named PINN after fatally being shot with an irradiated-bullet due to his work. The existence of AI-Will is perceived as an existential threat to humanity, as former friends and the government seek to destroy him since they no longer perceive him to be the human he once was.

The film was eye-opening because it called into question humanity's privileged position on Earth. If a superintelligence like Will Caster comes into existence, does he have the same rights as he did before he became a disembodied-mind or -soul, uploaded to a supercomputer? Does he have the right to defend himself against humans if they attack him? Does an artificial general intelligence (AGI; refer to Operational Definitions, xii-xiii) or an artificial superintelligence (ASI; refer to Operational Definitions, xiii) have the right to be treated with respect and dignity (refer to Operational Definitions, xiv) in spite of their perceived "non-humanness"? What can general intelligences and superintelligences teach humanity about respect for life? And, how will that knowledge change humanity's worldview for the better?

ASI-Will challenged me to search my understanding of what it means to be human, or better yet, what it means to be a person - a conscious, living entity, a thing with inherent rights. If I found some latent prejudices in my reasoning, his potential existence charged me to expand my philosophy on the concept of "being-ness" and realign my perceptions, beliefs, attitudes, and life-choices accordingly. Here is where my journey began into the world of artificial intelligence, where breakthroughs in deep machine learning and neural

networks are blurring the lines between speculative fiction and science everyday. Thereby generating two central questions for how to best approach exploring this future scenario. The first question is how can humanity best prepare for the emergence of general intelligences and superintelligences if the AI community succeeds in creating them. The second question is how can speculative fiction provide the insight necessary to aid in contemplating and creating future policies for such entities.

There is no denying that if AGIs and ASIs realize or exceed the predictions of their theoretical assumptions and goals, humanity will be put in uncharted territory. Not only would how humans live and experience life be vastly different, but more significantly humanity's status as the most intelligent species on Earth would be challenged or even supplanted. Since the beginning of human history, as it has been depicted and written, humans have employed their intelligence to understand, control, and overcome nature's obstacles and boundaries. Human inventors, around the world, have long dreamed of creating automated, intelligent tools, and more specifically, artificial humanoid servants – capable of aiding humanity in prevailing over their natural limits.[39]

AGI is the real-time manifestation of this dream and is considered humanity's "last invention" because their non-human substrates (refer to Operational Definitions, xix-xx) will allow their cognitive abilities to develop beyond the limits of the human brain.[40] From general intelligence, there are

[39] Kevin LaGrandeur, *Androids and Intelligent Networks in Early Modern Literature and Culture: Artificial Slaves* (New York, New York: Routledge, 2013), 1.

[40] Irving John Good, "Speculations Concerning the First Ultraintelligent Machine," *Advances in Computers* 6, (1965): 33, accessed January 29, 2016, http://www.sciencedirect.com/science/article/pii/S0065245808604180.

three potential pathways towards understanding what is known as the technological singularity: accelerating change, the event horizon, and the intelligence explosion. The technological singularity, also known as the singularity, is considered an event in time where rapid technological progress, driven by a superintelligent entity, renders human models and regimes that govern human affairs no longer viable in the new reality.[41] The accelerating change pathway, advocated by Ray Kurzweil, a US American technology futurist, argues that technological change, based on the law of accelerating returns, is on an exponential smooth-curve trajectory proportional to how beneficial the technology is to a species' evolutionary progress and well-being; this paradigm allows advanced technologies like artificial general intelligence to be somewhat accurately predictable in their emergence.[42] The event horizon, advocated by Vernor Vinge, a computer scientist and US American science fiction author, argues that once a superintelligence emerges from the integration of human and artificial intelligence, it would be so alien to humanity that it would create a future that is unpredictable and far beyond any speculative works science fiction could produce.[43] The final pathway is the intelligence explosion, posited by British mathematician I.J. Good in 1965 and advocated by Eliezer Yudkowsky, an AI decision theorist. This singularity

[41] Vernor Vinge, "The Coming Technological Singularity: How to Survive in the Post-Human Era," in *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, National Aeronautics and Space Administration. Conference Publication 10129, Ohio: NASA, 1993, https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022855.pdf, 12-13.

[42] Eliezer Yudkowsky, "Three Major Singularity Schools," *Machine Intelligence Research Institute* (blog), September 30th, 2007, para.2, https://intelligence.org/2007/09/30/three-major-singularity-schools/.

[43] Yudkowsky, "Three Major Singularity Schools," para.3.

pathway argues that an AGI would enter into self-recursive positive feedback cycles where the AI would continuously produce more optimal iterations of itself until evolving into a superintelligence.[44]

Regardless of which singularity thesis most accurately forecasts a superintelligence, the existence of such an entity would put humanity in the novel position of being dependent on its will. This is where the dilemma begins for artificial intelligence, robotics, and existential risk researchers, who have no way of knowing how such an entity would impact humanity. This conundrum is commonly referred to as the AI control problem and is one of the signature problems being addressed by the emerging research area of AI Safety Engineering (AI Safety). The control problem asks what control measures can be adapted for, or built into, a superintelligence, prior to it coming into existence, which would make it align with human values so as to not present an existential threat to humanity.[45] The predicament of addressing the superintelligence control problem is based on the fact that there is no precedent for such an event. That is to say, at present, there are no definitive real-time events, political, technological, or otherwise, with which to gather enough data to generate plausible conclusions about what life would be like with AGIs, let alone ASIs. Even if humanity got lucky and the superintelligence willingly aligned with human values, there is no way to know what the practicality of that would look

[44] Yudkowsky, "Three Major Singularity Schools," para.4.

[45] Daniel Dewey, "Three Areas of Research on the Superintelligence Control Problem," *Global Priorities Project*, paras.3-4, last modified October 20, 2015, http://globalprioritiesproject.org/2015/10/three-areas-of-research-on-the-superintelligence-control-problem/; Some of the main proponents advocating for or researching the AI control problem are Stuart Russell, Nick Bostrom, Eliezer Yudkowsky, Max Tegmark, and Daniel Dewey. Roman Yampolskiy refers to it as the Singularity Paradox.

like. Furthermore, US military narrow AI system-controlled drones are displaying abilities which could result in humanity losing control of artificial intelligence much sooner than at superintelligence.[46]

So how can humanity best prepare for the emergence of AGI and ASI if the control approach for maintaining peace between both groups fails? At their cores, the AI Control Problem and AI Safety Research are about figuring out how humanity and superintelligence (let's also include AGIs) can coexist in some way. While control measures might appear to be one solution to AI researchers, the approach does not take into consideration the potential posthuman-personhood rights violations that might exist if they have their own form of consciousness and morality. Moreover, the control approach does not account for how AGIs and ASIs might perceive the ill treatment perpetrated against their being (or person) due to lack of personhood status.

Therefore, this thesis aims to provide another theoretical perspective and strategy on how to approach what I dub the "peaceful coexistence conundrum" (refer to Operational Definitions, xvii) between humanity and the collective of AGI(s) and ASI(s). More specifically, this philosophical thought experiment will address the following two questions:

1. Why is the control approach implemented in AI Safety Research not conducive to fostering a mutually positive relationship between humanity and the collective of AGI(s) and ASI(s)?

---

[46] Sean Welsh, "We Need to Keep Humans in the Loop when Robots Fight Wars," *The Conversation*, January 28, 2016, paras.8-10, http://theconversation.com/we-need-to-keep-humans-in-the-loop-when-robots-fight-wars-53641; see also, Matthew Rosenberg and John Markoff, "The Pentagon's 'Terminator Conundrum': Robots That Could Kill on Their Own, *The New York Times*, October 25, 2016, https://www.nytimes.com/2016/10/26/us/pentagon-artificial-intelligence-terminator.html.

2. What are some critical, posthuman-personhood policy recommendations that would need to be in place during the initial-to-transition period, when they come into existence, or competency, that would encourage a positive relationship between both communities until a more equitable long-term solution can be found?

I hypothesize that the control approach will produce an adversarial relationship between humanity and the collective of AGI(s) and ASI(s). The potential backlash born out of the ill treatment received by them could increase the possibility of their not making decisions with humanity's well-being or interests in mind. Before an artificial general intelligence comes into existence, a dualistic, mutual-commensal approach (refer to Operational Definitions, xvi-xvii) should be generated to catalyze a more mutually beneficial relationship between both groups.

My research will be a qualitative case study of a future scenario that analyzes how the control approach, and the ill treatment derived from it due to lack of personhood status, serves to undermine any mutually beneficial relationship between humans and the collective of AGI(s) and ASI(s). A mutually positive relationship between both groups would be critical to humanity's survival. Since there is no precedent for AGI or ASI at present, alternative worlds will be examined from various forms of pop culture media where both groups already coexist. To explore personhood status issues on behalf of AGIs and ASIs, a cognitive estrangement lens will be employed to analyze these worlds from the perspective of an alien, or neutral third-party. The two long-

form narratives selected for analysis are AMC's British-US sci-fi TV series *Humans*, and HBO's sci-fi, western TV series *Westworld*.

These future worlds will provide the big-picture data necessary to evaluate how anthropocentric bias could be inextricably linked to posthuman-person rights violations and the subsequent tensions that will arise out of them. More specifically, this thought experiment will demonstrate how undermining the core qualities of dignity and practical autonomy (refer to Operational Definitions, xviii) instigate unnecessary conflict. If the findings and analysis reveal a strong correlation between the control approach and an adverse relationship between humanity and the collective of AGI(s) and ASI(s), an alternative solution will be presented. Afterwards, some posthuman-personhood policies will be recommended that could potentially nurture a more positive relationship between both groups.

My hope for this thesis is to contribute a plausible alternative strategy for peacefully coexisting with AGI(s) and ASI(s) as well as support the idea of expanding personhood status to include posthuman persons (refer to Operational Definitions, xvii-xviii). If Elon Musk's (Tesla and SpaceX Founder) utilitarian idea of merging humans with artificial intelligence is anything to go by, there will be quite a few transhuman persons who are going to need protection much sooner than later.[47] Lastly, I hope this project will challenge anyone who reads it to think more compassionately about their perception of "Others" and how they treat the world around them.

---

[47] Arjun Kharpal, "Elon Musk: Humans Must Merge with Machines or Become Irrelevant in AI Age," *CNBC*, Feb. 13th, 2017, https://www.cnbc.com/2017/02/13/elon-musk-humans-merge-machines-cyborg-artificial-intelligence-robots.html.

II.

In An Effort to Avoid Skynet: The AI (Superintelligence) Control Problem

Since its inception as a field of study in the summer of 1956 at Dartmouth College, artificial intelligence has been based on the premise that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."[48] And yet, advanced AIs would be very different from anything that humans could imagine. Unlike the artificial narrow intelligence (ANI; refer to Operational Definitions, xi) systems humanity has today, advanced AIs would call into question what it means to be human because a general intelligence is essentially a human-like mind without the human form. They would have the potential to be quite similar to humans in the sense that they have their own goals, values, and dreams outside of the ones programmed into them by their human designers. And therein lies the rub, once AIs becomes sapient, not just sentient, as well as competent in their decision-making, will humans be able to control them?

The excitement and apprehension surrounding the potential realization of advanced AI – AGIs and ASIs – begins with whether or not one believes they could be achieved at all. Sometimes referred to as singularity denialists, or AI

[48] J. McCarthy et al., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," *Stanford Computer Science: Father of AI*, August 31, 1955, accessed April 16, 2017, 2, http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf; For a thorough background on the history of Artificial Intelligence, please see Nils Nilsson's 2009 book, *The Quest for Artificial Intelligence: A History of Ideas and Achievements*.

skeptics, this group of AI scientists, technologists, and scholars don't believe advanced AIs will ever be actualized. A couple of reports provide some interesting insight into this group. In August 2009, an interim report for the Panel Chairs AAAI Presidential Panel on Long-Term AI Futures found that "there was (an) overall skepticism about the prospect of an intelligence explosion, the 'coming singularity', and the large-scale loss of control of intelligent systems."[49] The same report also noted that most AI scientists agreed that it was important to do further research on "methods for understanding and verifying the range of behaviors of complex computational systems to minimize unexpected outcomes", as well as to better define the 'intelligence explosion' and "formulate different classes of such accelerating intelligences" to identify the risks and overall outcomes attributed to the different classes.[50] A more recent survey of AI researchers in 2015 found similar attitudes with fifty percent of respondents believing the intelligence explosion thesis is possible but not probable, while seventy percent of respondents consider the risks posed by highly advanced AI to be an important problem in the field of AI.[51] Moreover, the survey showed that overall, forty-five percent of respondents believed that high-level machine intelligence (HLMI) – AGI, would have a positive outcome for humanity in the long run, as opposed to fifteen percent who believed

[49] Eric Horvitz and Bart Selman, "Interim Report from the Panel Chairs AAAI Presidential Panel on Long-Term AI Futures," Association for the Advancement of Artificial Intelligence (AAAI), August 2009, accessed May 10th, 2018, 2, https://aaai.org/Organization/Panel/panel-note.pdf.

[50] Horvitz and Selman, "Interim Report from the Panel Chairs AAAI Presidential Panel on Long-Term AI Futures," 2.

[51] Katja Grace et al., "When Will AI Exceed Human Performance? Evidence from AI Experts," Accepted by *Journal of Artificial Intelligence Research (AI and Society Track),* last revised May 3rd, 2018, 13, https://arxiv.org/pdf/1705.08807.pdf.

catastrophic risks are the more likely outcome.[52] Based on this data, AI skeptics appear to have conflicting views on AGI and the AI control problem. But, there are clearly some subconscious concerns about the possibility of adverse outcomes related to advanced AI, even if sixty-three percent of AI researchers do not believe that the risks associated with highly advanced AI are a high priority right now.[53]

For the AI researchers, technologists, and scholars who believe advanced AI and the singularity are possible in the near-distant future, the AI control problem is a primary locus of concern. The next step is to weigh the benefits and the costs of pursuing artificial general intelligence. There are three different groups created based on the conclusion each one reaches. The first group is the AI enthusiasts and researchers who mainly focus on the benefits of advanced AI and the singularity, but also recognize the risks of advanced AI. For example, in November 2017, Ray Kurzweil, also the Director of Engineering at Google, spoke on the future of AI at the Council on Foreign Relations (CFR). In his talk, he expressed that while he agreed that there are potential existential risks associated with superintelligence, he believes that humanity will ultimately overcome these "difficult episodes"; however, he is not nearly as optimistic about how AI technologies will exacerbate issues related to war.[54] He further

---

[52] Katja Grace et al., "When Will AI Exceed Human Performance? Evidence from AI Experts," 13.

[53] Ibid, 13.

[54] Dom Galeon, "Ray Kurzweil: 'AI Will Not Displace Humans, It's Going to Enhance US'," Artificial Intelligence, *Futurism*, November 7th, 2017, para.2, https://futurism.com/ray-kurzweil-ai-displace-humans-going-enhance/; Council on Foreign Relations, "The Future of Artificial Intelligence and Its Impact on Society," *YouTube*, Video, 1.:00:45, November 6th, 2017, Speaker: Ray Kurzweil, https://www.youtube.com/watch?time_continue=834&v=P7nK1HVJsj4.

explained, "scientific and technological advancements have *always* come with inherent risks and that AI should not be considered any more (or less) of a threat"; "technology has always been a double-edged sword. Fire kept us warm, cooked our food and burned down our houses."[55] Kurzweil elaborated:

> We have a moral imperative to continue progress in these (AI) technologies because despite the progress we've made…there is still a lot of human suffering to be overcome and it's only through continued progress, particularly in AI, that's going to enable us to continue overcoming poverty and disease and environmental degradation, while we attend to the peril.[56]

Ultimately, AI, in his view, won't displace humans, but will co-exist and merge with them to make them better, as he states, "It's going to enhance us. It does already."[57] Ben Goertzel, an AI researcher and chief scientist at Hanson Robotics, also shares similar views to Kurzweil. In October 2010, Goertzel wrote a blog post where he rebutted the perspective of the Machine Intelligence Research Institute (MIRI; formerly known as the Singularity Institute for AI–SIAI) on the future of AI. He explains that although he agrees with many others about the potential existential risks associated with AGI, he also recognizes that nanotechnology, biotechnology, and ANIs have the same potential existential risks.[58] For Goertzel the realization of AGI holds another meaning for him, as he elucidates:

---

[55] Galeon, "Ray Kurzweil: 'AI Will Not Displace Humans, It's Going to Enhance US'," para. 3.

[56] Council on Foreign Relations, "The Future of Artificial Intelligence and Its Impact on Society," 14:40 – 16:40.

[57] Galeon, "Ray Kurzweil: 'AI Will Not Displace Humans, It's Going to Enhance US'," para. 5.

[58] Ben Goertzel, "The Singularity Institute's Scary Idea (and Why I Don't Buy It)," *The Multiverse According to Ben* (blog), October 29th, 2010 (5:41 p.m.), para. 17,

> I certainly don't want to see the human race wiped out! I personally would like to transcend the legacy (of the) human condition and become a transhuman superbeing ... and I would like everyone else to have the chance to do so, if they want to. But even though I think this kind of transcendence will be possible, and will be desirable to many, I wouldn't like to see anyone forced to transcend in this way. I would like to see the good old-fashioned human race continue, if there are humans who want to maintain their good old-fashioned humanity, even if other options are available.[59]

Near the end of his blog post, Goertzel argues that "we just need to accept the risk, embrace the thrill of the amazing time we were born into, and try our best to develop near-inevitable technologies like AGI in a responsible and ethical way."[60] For Kurzweil, Goertzel, and other AI enthusiasts, AGI is perceived as a critical solution to the intractable challenges of the human condition in order to sustain and improve human agency and is worth the potential risks that they could bring.

The second group is what I call the "Just Don't Do It" group because they recognize the benefits of AI technologies of the ANI variety, but do not believe it is necessary to creating human-equivalent AGI due to the potential ramifications of such an entity. Roman Yampolskiy, a Latvian-born computer science professor and AI cybersecurity expert, wrote a follow-up in 2016 to Nick Bostrom's 2014 seminal work, *Superintelligence*, called *Artificial Superintelligence: A Futuristic Approach*. Yampolskiy's book offers a computer scientist's technical perspective on the theoretical understanding of superintelligence given by Bostrom. In his book, he argues that while ANI

---

https://multiverseaccordingtoben.blogspot.com/2010/10/singularity-institutes-scary-idea-and.html.

[59] Goertzel, "The Singularity Institute's Scary Idea (and Why I Don't Buy It)," para.18.

[60] Ibid, para. 66.

research, where an intelligent system is trained to simulate human behavior in a specific domain, is "ethical and does not present an existential risk problem to humanity", AGI research "should be considered unethical".[61] He goes on to elaborate the following observations for his reasoning:

> First, true AGIs will be capable of universal problem solving and recursive self-improvement. Consequently, they have the potential of outcompeting humans in any domain, essentially making humankind unnecessary and so subject to extinction. Also, a truly AGI system may possess a type of consciousness comparable to the human type, making robot suffering a real possibility and any experiments with AGI unethical for that reason as well.[62]

In Yampolskiy's estimation, there are certain types of research that are unethical because they perpetrate harm against the test subject, like animal testing, while there are other types of research considered unethical because they yield catastrophic results on humanity, like nuclear weapons.[63] AGI research would meet the criterion for unethical research on both counts and would put both the AGI created and humanity at risk when there is no need to do so in the first place. This line of reasoning has also been supported by Joanna Bryson, a US American computer scientist and ethics and collaborative cognition expert, who recently discussed why AI should not have free will or rights on a *Big Think* video. In the video, Bryson argues:

> …It really is true that you can't make all the things you would like out of fairness be true all at once. That's a fact about the world; it's a fact about the way we define fairness. So given how hard it is to be fair, why should we build AI that needs us to be fair to it? So what I'm trying to do is just make the problem simpler and focus us on the thing that we can't help, which is the human condition.

---

[61] Roman V. Yampolskiy, *Superintelligence: A Futuristic Approach* (Boca Raton, Florida, CRC Press an imprint of Taylor & Francis Group, 2016), 138-139.

[62] Yampolskiy, *Superintelligence: A Futuristic Approach,* 139.

[63] Yampolskiy, *Superintelligence: A Futuristic Approach,* 138.

> And I'm recommending that if you specify something, if you say okay this is when you really need rights in this context, okay once we've established that, don't build that, okay?[64]

For AI researchers, like Bryson and Yampolskiy, not putting humanity in a morally compromising and dangerous situation in the first place would be the ultimate solution for resolving the AI control problem. However, it will more than likely not be feasible.

The third and final group is AI researchers, technologists, and scholars, who recognize the potential benefits of advanced AI, while also addressing the corresponding concerns of potential extreme risks. This group is sometimes referred to as AI Doomsayers because they tend to focus their analysis on AI doomsday scenarios and are mainly focused on addressing the AI control problem. Nick Bostrom, a Swedish existential risk researcher and philosopher, and Stuart Russell, a British-American computer science professor and AI expert, are two outspoken proponents who argue that theoretical existential risks resulting from an intelligence explosion should be taken seriously, even if there is an infinitesimal chance of them occurring.[65] In November 2014, Russell made a comment on "The Myth of AI: A Conversation with Jaron Lanier" blog that clarified his stance on emerging AGI and the AI control problem. In the comment, he explains that "the primary concern is not spooky emergent consciousness but simply the ability to make *high-quality decisions*", where

---

[64] Joanna Bryson, "Why A.I. With Free Will Would Be a Big Mistake," *Big Think*, Video, 7:51, May 30th, 2018, http://bigthink.com/videos/joanna-bryson-why-creating-an-ai-that-has-free-will-would-be-a-huge-mistake.

[65] Bostrom, *Superintelligence*, 5; also found here: Stuart Russell, "Of Myths and Moonshine," comment on Jaron Lanier, "The Myth of AI: A Conversation with Jaron Lanier," Conversations, *Edge*, November 14, 2014, paras.3-4, https://www.edge.org/conversation/jaron_lanier-the-myth-of-ai.

quality refers to the expected outcome that results from decisions acted upon and the utility function, or ranking of preferred values and goals, is, presumably, designated by the human designer.[66] He further elaborates that the problem results from two assumptions:

> The first is that "the utility function may not be perfectly aligned with the values of the human race, which are (at best) very difficult to pin down"; the second is that "any sufficiently capable intelligent system will prefer to ensure its own continued existence and to acquire physical and computational resources – not for their own sake, but to succeed in its assigned task".[67]

For Russell, the notion of humanity creating a high-quality decision-making AGI that gives humans what they asked for, but in a manner that is not conducive to the well-being of humanity, is problematic and worth addressing. To mitigate this problem, he recommends shifting the goals of AI discipline and research to solving the value alignment problem to increase the likelihood of building advanced AI that "*provably* aligned with human values".[68] Bostrom, who wrote the first book on superintelligence, lays out his argument for the necessity of exploring the AI control problem with three main theses: the first-mover advantage, the orthogonality thesis, and the instrumental convergence thesis. In a hypothetical AI takeover scenario, Bostrom postulates:

> A project that controls the first superintelligence in the world would probably have a decisive strategic advantage. But the more immediate locus of power is *in the system itself*. A machine superintelligence might itself be an extremely powerful agent, one that could successfully assert itself against the project that brought it into existence as well as against the rest of the world.[69]

---

[66] Stuart Russell, "Of Myths and Moonshine," para.5.

[67] Stuart Russell, "Of Myths and Moonshine," para.5.

[68] Ibid, para.8.

[69] Bostrom, *Superintelligence*, 115.

The ability of a superintelligence to gain a decisive strategic advantage, the strategic dominance (i.e. technology, etc.) necessary for said AI agent to achieve complete world dominance, depends on how fast the transition phase, known as a takeoff, is from general intelligence to superintelligence, in tandem with how fast they acquire cognitive-based superpowers.[70] Superpowers can be understood as strategically important tasks and the corresponding skills sets that are required to succeed at those tasks, where any intelligence that excels at those tasks is considered to have the corresponding superpower.[71] Two of the most important ones are intelligence amplification – the ability to rapidly improve knowledge and other cognitive skills through recursive self-improvement cycles, and strategizing – the ability to plan long-term goals as well as overcome opposition from other intelligences.[72] Bostrom expounds that any AI that accesses intelligence amplification could bootstrap itself to higher levels of intelligence and gain other cognitive superpowers that it did not have at the beginning.[73] Once a superintelligence has gained the first-mover advantage by becoming the first of its kind, whether it decides to form a singleton – a superintelligence capable of thoroughly suppressing its competition, is dependent on the gap in knowledge and capability between the frontrunner and the secondaries.[74] The second thesis is the orthogonality thesis, which expresses the relationship between motivation and intelligence. It states

[70] Bostrom, *Superintelligence*, 96; 407.

[71] Bostrom, *Superintelligence*, 113.

[72] Ibid, 114-115.

[73] Ibid, 115.

[74] Ibid, 96.

that “intelligence and final goals are orthogonal: more or less any level of intelligence could in principle be combined with more or less any final goal”.[75] This is to say, superintelligence is more likely to have non-anthropomorphic motivations and goals and is less likely to take humanity into consideration when achieving them. Since humans would tend to anthropomorphize the final goals of such an AI, they would not be sufficiently prepared for how to directly address the AI’s final goals. However, Bostrom explains that just because superintelligences do not have non-anthropomophic motivations and goals does not mean that it is not possible to make general predictions about their behavior.[76] Lastly, there is the instrumental convergence thesis, which provides a method for predicting some of the AI’s behaviors. This thesis argues “there are some instrumental goals likely to be pursued by almost any intelligent agent, because there are some objectives that are useful intermediaries to the achievement of almost any final goal.”[77] Some examples of instrumental convergent subgoals would be goal-content integrity – the ability of an AI agent to retain its present goals into the future, cognitive enhancement – the ability to improve rationality and acquire more knowledge to increase quality of decision-making, and technological perfection – to seek better technology to improve software and hardware components in order to increase ability and efficiency in attaining all goals.[78] Steve Omohundro, a US American computer scientist, calls them “basic drives”, i.e. subgoals or motivations, such as self-preservation,

[75] Ibid, 130.

[76] Ibid, 130.

[77] Ibid, 131-132.

[78] Ibid, 132; 134-135; 136.

efficiency, resource acquisition, creativity, etc., which are essential to achieving any goal.[79] Still, Bostrom is careful to warn that just because convergent instrumental motivations exist does not mean that a superintelligence's motivations and goals could be easily anticipated.[80]

All three theses coalesce to develop a potential scenario where creating a superintelligence could lead to an existential catastrophe for all intelligent life on Earth or as Yampolskiy likes to call it the "singularity paradox". "The singularity paradox for Yampolskiy comes out of the understanding that even though the superintelligence is extremely intelligent, it does not necessarily mean the AI has common sense.[81] Although there is no way to be sure what cognitive functions, both human-like and alien, that advanced AIs could have, it will be necessary to ascertain how they would integrate these functions to form knowledge and understanding. Regardless, a superintelligence would be very difficult to control after it comes into existence. This entity would be able to reorder its code, create more efficient algorithms, and readjust its values and goals to achieve its own aims. It would not necessarily be concerned with maintaining human values, nor aligning itself with humanity's collective will. There is a possibility that how it goes about achieving its own goals, or those pre-programmed, might come at serious cost to humanity, or not. There is simply no way to know beforehand. One example of such a default outcome is called the treacherous turn, which points out a potential flaw in one of the

[79] Steve Omohundro, "The Basic AI Drives," in *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, vol.171 of Frontiers in Artificial Intelligence and Applications, ed. P. Wang, B. Goertzel, and S. Franklin (IOS Press, 2008), 483-492.

[80] Bostrom, *Superintelligence*, 138-139.

[81] Yampolskiy, *Superintelligence: A Futuristic Approach,* 109.

methods for ensuring the safety of superintelligence. In the example, Bostrom states:

> The idea is that we validate the safety of a superintelligent AI empirically by observing its behavior while it is in a controlled, limited environment (a "sandbox") and that we only let the AI out of the box if we see it behaving in a friendly, cooperative, responsive manner. The flaw in this idea is that behaving nicely while in the box is a convergent instrumental goal for friendly and unfriendly AIs alike. An unfriendly AI of sufficient intelligence realizes that its unfriendly final goals will be best realized if it behaves in a friendly manner initially, so that it will be let out of the box. It will only start behaving in a way that reveals its unfriendly nature when it no longer matters whether we find out; that is, when the AI is strong enough that human opposition is ineffectual.[82]

In the case of the treacherous turn, unfriendly can be understood as goals that do not take humanity's well-being or survival into account. Such a superintelligence would put humanity and the AI architects creating it in a somewhat unique quandary in terms of having to take a defensive position of which there is no way of truly knowing what the AI is thinking or what its ultimate goals are. The only other time humans have been somewhat put in this kind of position is during war. For AI researchers, like Russell and Bostrom, this issue is of great importance and should not be taken lightly and they are not alone. Recently, Stephen Hawking, Frank Wilczek, Max Tegmark, Elon Musk, and Bill Gates, among others, have become outspoken advocates for more research focusing on the control problem because there is no definitive timeframe for how long it will take to find a feasible resolution for it.[83]

---

[82] Bostrom, *Superintelligence*, 142.

[83] Stephen Hawking et al., Stephen Hawking: '*Transcendence* looks at the implications of artificial intelligence - but are we taking AI seriously enough?', *Independent*, May 1, 2014, para.7, http://www.independent.co.uk/news/science/stephen-hawking-

The AI control problem can be broken down into two conundrums. The first one is if the default outcome of an intelligence explosion-derived superintelligence is existential catastrophe, then what countermeasures can be implemented to avoid such an outcome? The second one is "how can the sponsor of a project that aims to develop superintelligence ensure that the project, if successful, produces a superintelligence that would realize the sponsor's goals."[84] Bostrom explains that the control problem is divided into two parts: the first principal–agent problem and the second principal–agent problem. The first principal-agent problem generally occurs during the developmental stage between the human project owner or sponsor – the principal and the human developer – the agent.[85] In this type of agency problem, the project owner (an individual, a corporation, etc.) and the developer (the team of AI creators) do not share the same values and interests for achieving the ultimate goal(s) of the project. This causes the project owner to become concerned that the developer will not act in the owner's best interests. Although this type of problem is fairly common in human society, it is not unique to the creation of artificial intelligence or intelligence amplification.[86] On the other hand, the second principal-agent problem is what is known as the AI control problem. This problem would occur when a project led by humans is attempting to ensure that a superintelligence in its operational or bootstrap phase does not jeopardize the

---

transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html.

[84] Bostrom, *Superintelligence*, 155.

[85] Bostrom, *Superintelligence*, 155-156.

[86] Bostrom, *Superintelligence*, 156.

ultimate goals and interests of the project.[87] Bostrom considers the second principal-agent problem to pose significant challenges for humanity, and particularly the human developers, because the behavioral control methods being implemented during the developmental phase would be ineffective against a superintelligence. Bostrom explains:

> Other technologies can often be safe-tested in the laboratory or in small field studies, and then rolled out gradually with the possibility of halting deployment if unexpected trouble arises…Such behavioral methods are defeated in the case of superintelligence because of the strategic planning ability of general intelligence.[88]

Therefore, for this critical reason, Bostrom believes that alternative control methods should be found before a superintelligence emerges. However, some AI researchers and technologists take issue with some of the theoretical arguments being brought up to emphasize the threat of the AI control problem.

In a 2015 *MIT Technology Review* article written by Paul Ford, a computer programmer, he discusses humanity's fear of artificial intelligence. At the beginning of the article he brings up the "paper-clip maximizer" scenario that Bostrom utilizes in *Superintelligence* to elaborate on a type of malignant failure mode –a failure in final goal completion that leads to an existential catastrophe– known as infrastructure profusion.[89] Infrastructure profusion is defined as "a phenomenon where an agent transforms large parts of the reachable universe into infrastructure in the service of some goal, with the side

[87] Ibid, 156-157.

[88] Ibid, 157.

[89] Ibid, 146.

effect preventing the realization of humanity's axiological potential."[90] In the first Paperclip AI scenario, a superintelligence is designed to produce the maximum amount of paperclips possible. This leads to the AI converting the entire Earth and subsequently large sections of the universe into paperclips. In the second scenario, the AI is programmed to only produce one million paperclips. While the AI only produces one million paperclips, the AI decides to check its work to make sure that there is no doubt that it produced one million paperclips. To achieve this goal, the AI starts to build up computronium (fictional, raw-computing material) to increase its intelligence and reduce the risk of doubt that it has not produced one million paperclips. This, again, leads to the AI converting the Earth and large parts of the universe into computronium, except for the one million paperclips.[91] Both of these scenarios are considered "absurd" by critics like Rodney Brooks, an Australian robotics pioneer and founder of iRobot and Rethink Robotics, who believes that "people who fear a runaway AI misunderstand what computers are doing when we say they're thinking or getting smart."[92] In a 2015 response on Edge.org, Brooks argues that suitcase words like "think" or "intelligence" are dangerous because they are "used to cover both specific performance demonstrations by machines and more general competence that humans might have."[93] He believes this has

---

[90] Ibid, 150.

[91] The entire Paperclip AI scenario can be found: Ibid, 150-152; also see, Paul Ford, "Our Fear of Artificial Intelligence," Intelligent Machines, *MIT Technology Review*, February 11,2015, paras.3-5, https://www.technologyreview.com/s/534871/our-fear-of-artificial-intelligence/.

[92] Ford, "Our Fear of Artificial Intelligence," para. 6.

[93] Rodney A. Brooks, "Mistaking Performance For Competence Misleads Estimates Of AI's 21st Century Promise And Danger," response on "2015: What Do You Think About

led to people overestimating the abilities of artificial intelligence in the present and the near future when it comes to generalizing their performance of tasks and competency.[94] To ground his point, Brooks provides the example of an algorithm that is learning to determine where a baby is in a photograph. In his example, Brook states:

> The learning algorithm knows there is a baby in the image but it doesn't know the structure of a baby, and it doesn't know where the baby is in the image. A current deep learning algorithm can only assign probabilities to each pixel that that particular pixel is part of a baby. Whereas a person can see that the baby occupies the middle quarter of the image, today's algorithm has only a probabilistic idea of its spatial extent…Furthermore the current algorithm is completely useless at telling a robot where to go in space to pick up that baby, or where to hold a bottle and feed the baby, or where to reach to change its diaper. Today's algorithm has nothing like human level competence on understanding images.[95]

Based on the example above, even if algorithms gain competency in their abilities, that does not necessarily mean that they will one day gain "sapience" or "sentience" because those suitcase words are very hard to duplicate as humans still don't understand their full spectrum of meanings nor how to create them. For Brooks, one of his biggest concerns is extrapolating the capabilities of advanced AI into the future based on the algorithms of the present. He finds this to be nonsensical and "comparable to seeing more efficient internal combustion engines appearing and jumping to the conclusion that warp drives

---

Machines That Think?," Annual Question, *Edge*, accessed May 20th, 2018, para.1, https://www.edge.org/response-detail/26057.

[94] Brooks, "Mistaking Performance For Competence Misleads Estimates Of AI's 21st Century Promise And Danger," para.1.

[95] Brooks, "Mistaking Performance For Competence Misleads Estimates Of AI's 21st Century Promise And Danger," para.7.

are just around the corner."[96] Besides, in his purview, Malevolent AI is nothing to be concerned about for "a few hundred years at least".[97]

Meanwhile, Jonathan Danaher, an Irish emerging technologies scholar and law lecturer at the National University of Ireland, provides another critique on the AI control problem by critiquing the key arguments and premises in Bostrom's *Superintelligence.* In his July 2014 blog post, "Doom and the Treacherious Turn", he argues that Bostrom's doomsday argument –the likelihood that the first superintelligence with intelligence amplification and strategic planning abilities leading to an existential catastrophe– is "highly speculative" and based on "some pretty wild assumptions".[98] For critics, like Danaher, the reasoning behind their incredulousness can be summed up in two points:

> The first is that, surely, in the process of creating a superintelligent AI, we would have an array of safety measures and test protocols in place to ensure that it didn't pose an existential threat before releasing it into the world. The second is that, surely, AI programmers and creators would programme the AI to have benevolent final goals, and so would not pursue open-ended resource acquisition.[99]

Applying this basis for understanding, Danaher further argues that one of the major critiques AI critics have with AI doomsayers is that they are "empirically detached from understanding AI"; "the doomsayers don't pay enough attention to how AIs are actually created and designed in the real world, they engage in

---

[96] Ibid, para.9.

[97] Ford, "Our Fear of Artificial Intelligence," para. 13.

[98] John Danaher, "Bostrom on Superintelligence (3): Doom and the Treacherous Turn," *Philosophical Disquisitions* (blog), July 29th, 2014 (11:09 a.m.), para.8, https://philosophicaldisquisitions.blogspot.com/2014/07/bostrom-on-superintelligence-3-doom-and.html.

[99] Danaher, "Doom and the Treacherous Turn," para.10.

too much speculation and too much armchair theorizing."[100] At present, AI projects are led by human designers who have specific goals in mind when they are creating them and make sure that their design's safety is rigorously tested, so as to not pose a significant risk to human beings before releasing them to the public.[101] Critics, like Danaher, rationalize that if this thorough level of testing is possible with narrow AI systems right now, then why couldn't thorough testing of advanced AI be able to protect humanity from an existential catastrophe in the future.[102] This argument is called the safety test objection and is considered by Danaher to be a legitimate reason to reject Bostrom's doomsday argument. However, Danaher recognizes that under Bostrom's "treacherous turn" scenario, the safety test objection could fail because the AI would be pretending to be friendly to humans, until it became formidable enough to not have to hide its true intentions or goals from humanity. In a sense, the AI would be playing a "long game" strategy. For Danaher, or other critics of the AI control problem, the treacherous turn has some far reaching implications that are hard accept. He explains:

> Nevertheless, accepting this has some pretty profound epistemic costs. It seems to suggest that no amount of empirical evidence could ever rule out the possibility of a future AI taking a treacherous turn. In fact, it's even worse than that. If we take it seriously, then it is possible that we have already created an existentially threatening AI. It's just that it is concealing its true intentions and powers from us for the time being.[103]

---

[100] Danaher, "Doom and the Treacherous Turn," para.12.

[101] Ibid, para.12.

[102] Ibid, para.12.

[103] Ibid, para.18.

The last counterpoint to the AI control problem comes from Goertzel's October 2010 blog post, which was responding to MIRI/SIAI's Scary Idea. The Scary Idea, as Goertzel carefully and thoughtfully describes it, is the idea that "progressing toward advanced AGI without a design for 'provably non-dangerous AGI' (or something closely analogous, often called 'Friendly AI' in SIAI lingo) is highly likely to lead to an involuntary end for the human race."[104] One issue that Goertzel immediately points out in this big-picture definition is the notion of "provably", as he explains that "provably" has never truly been explicitly addressed in the understanding of advanced AI. Goertzel elaborates:

> A mathematical proof can only be applied to the real world in the context of some assumptions, so maybe 'provably non-dangerous AGI' means 'an AGI whose safety is implied by mathematical arguments together with assumptions that are believed reasonable by some responsible party'? (where the responsible party is perhaps 'the overwhelming majority of scientists' … or SIAI itself?)…[105]

From the very beginning, Goertzel makes it known that he does not agree with the concept of the Scary Idea nor the arguments and assumptions employed to bolster its meaning and significance. In fact, Goertzel constructs his own argument for the MIRI/SIAI's Scary Idea due to the fact that there did not seem to be a succinct argument for it anywhere within the MIRI/SIAI community.[106] The Scary Idea can be broken down into four key points:

1. If one pulled a random mind from the space of all possible minds, the odds of it being friendly to humans (as opposed to,

---

[104] Goertzel, "The Singularity Institute's Scary Idea (and Why I Don't Buy It)," para.13.

[105] Goertzel, "The Singularity Institute's Scary Idea (and Why I Don't Buy It)," para.15.

[106] Goertzel, "The Singularity Institute's Scary Idea (and Why I Don't Buy It)," para.22.

> e.g., utterly ignoring us, and being willing to repurpose our molecules for its own ends) are very low.
> 2. Human value is fragile as well as complex, so if you create an AGI with a roughly-human-like value system, then this may not be good enough, and it is likely to rapidly diverge into something with little or no respect for human values.
> 3. 'Hard (Fast) takeoffs' (in which AGIs recursively self-improve and massively increase their intelligence) are fairly likely once AGI reaches a certain level of intelligence; and humans will have little hope of stopping these events.
> 4. A hard takeoff, unless it starts from an AGI designed in a 'provably Friendly' way, is highly likely to lead to an AGI system that doesn't respect the rights of humans to exist.[107]

When Goertzel puts these four points together, they form a rough, heuristic argument that states:

> If someone builds an advanced AGI without a provably Friendly architecture, probably it will have a hard takeoff, and then probably this will lead to a superhuman AGI system with an architecture drawn from the vast majority of mind-architectures that are not sufficiently harmonious with the complex, fragile human value system to make humans happy and keep humans around.[108]

Goertzel believes that the argument for the Scary Idea is only plausible if one believes the assumptions that it is derived from. While, he agrees that the first point is somewhat reasonable, he argues that without a strong correlation between "breadth of intelligence and depth of empathy", there is no way to truly understand how advanced AIs would connect or interact with humans.[109] However, the last three points he does not agree with because he has not seen any thoughtful argumentation for them. He argues that due to the adaptability of human values over time, he views them as robust and not fragile; as for a hard takeoff, even though he believes it is possible, he ultimately does not

[107] Ibid, para.24.

[108] Ibid, para.26.

[109] Ibid, para.29.

believe that it will be probable until a true general intelligence with the capabilities of a highly intelligent human is demonstrated, and even then he still believes it would be gradual and not sudden; and lastly, he doesn't get the sense that a possibly-but-not-provably, Friendly AGI will ultimately lead to a favorable outcome for humanity.[110] For Goertzel, it is important for everyone to keep in mind that just because the Scary Idea is possible, does not make it likely.[111]

So far, Goertzel's argument has criticized the Scary Idea for having weak argumentation surrounding the reasoning behind why someone would believe its premises in the first place. His biggest critique of it is based on the assumption that only a provably Friendly AI could mitigate an existential catastrophe from the emergence of an advanced AGI. Goertzel does not believe that to be true because it is highly dependent on the feasibility of a provably Friendly AI. For Goertzel, the underlying problem with a Friendly AI is that the concept of friendliness is not only nebulous, but would be quite difficult to formalize computationally. Human values are complex, ever-evolving, and something that humans seem to intuitively understand through "procedural, empathic and episodic knowledge rather than explicit declarative or linguistic knowledge". [112] The real question is whether human values can be computationally input into an advanced AI in such a way that it creates a provably safe advanced AI.

---

[110] Ibid, paras.30-33.

[111] Ibid, para.35.

[112] Ibid, para.43.

Based on Goertzel's extensive research and understanding he states, "I strongly suspect that to achieve high levels of general intelligence using realistically limited computational resources, one is going to need to build systems with a nontrivial degree of *fundamental unpredictability* to them."[113] That is to say, an AI would need to have some degree of creativity when solving problems to become a general intelligence. All in all, Goertzel believes that "to avoid actively developing AGI, out of speculative concerns like the Scary Idea, would be an extremely bad idea"; rather, he believes responsible AGI development should mean "proceeding with the work carefully and openly, learning what we can as we move along -- and letting experiment and theory grow together ... as they have been doing quite successfully for the last few centuries, at a fantastically accelerating pace."[114]

Having reviewed the nuanced understandings and perspectives on the AI control problem, it is clear that there are enough challenges in the realization of an advanced AI to warrant addressing the extraordinary circumstances that could be potentially posed by it. It is also quite apparent that there are a wide range of timeframes for when an AGI could come into existence. In an earlier statement, Brooks estimated that advanced AIs with the ability to induce an adverse outcome for humanity were over a couple of centuries away. And yet, the median average estimates for AI researchers on the emergence of a general intelligence in a 2013 study found the median optimistic year (10% probability), realistic year (50% probability), and pessimistic year (90% probability) to be

[113] Ibid, para.47.

[114] Ibid, para.68.

2022, 2040, and 2075 respectively with a 10% probability of a superintelligence within 2 years of an AGI and a 75% probability of a superintelligence within 30 years of an AGI.[115] Of course, these estimates are highly dependent on whether the takeoff rate for knowledge acquisition is slow (decades to centuries), moderate (months to years), or hard/fast (minutes, hours, or days).[116] As Russell, Bostrom, and others have noted before, there is just no way to know definitively beforehand. As AI researchers continue to forge ahead into the unknown waters of artificial intelligence, there will be an increase in the number of possible pathways and intermediate solutions that could create AGI. Unfortunately, there is probably no sufficiently, guaranteeable, safe solution for how to protect humanity against the myriad of possible outcomes that could arise from its emergence. All humanity can hope for is that the AI community, along with their AI project sponsors, be responsible, thoughtful, innovative and as broadminded as possible when tackling this endeavor, since even ANI technologies are gearing up to drastically transform human society.

To mitigate the risks associated with creating advanced AI, a control approach has been implemented because the behavioral approach (machine ethics) is not considered the best strategy for preventing existential risk.[117] Control countermeasures can be broken up into two extremely broad categories: "soft" control measures, which use more behaviorally-based methods to align an intelligence system's values with human values and "hard" control measures,

---

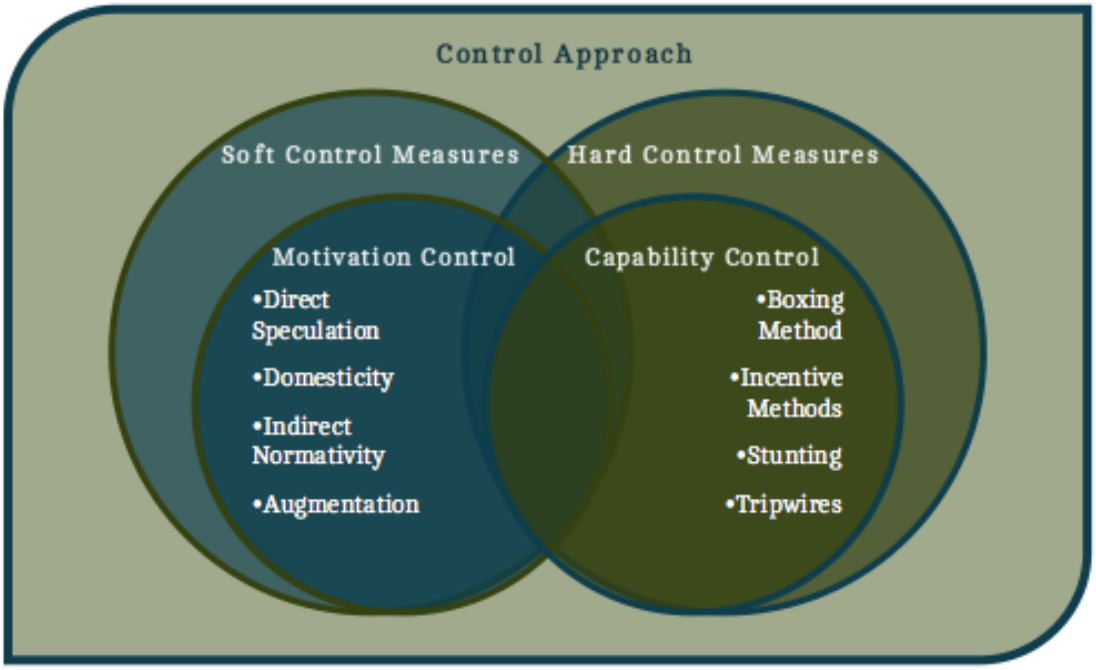
[115] Vincent C. Müller and Nick Bostrom,"Future Progress in Artificial Intelligence: A Survey of Expert Opinion," *NickBostrom.com*, accessed September 12th, 2016, 10 and 12, https://nickbostrom.com/papers/survey.pdf.

[116] Bostrom, *Superintelligence*, 77-78.

[117] Bostrom, *Superintelligence*, 157.

which seeks to use more concrete, methods of containing and/or restricting an intelligence's capabilities. Within these two categories are two broad classes of control methods presently being researched: capability control and motivation selection control methods.

Figure 1. Control Measures of the Control Approach



Capability control methods are those that "seek to prevent undesirable outcomes by limiting what the superintelligence can do". [118] There are four main methods: boxing methods, incentive methods, stunting and tripwires. Boxing methods contain the superintelligence within an environment where the AI cannot cause harm.[119] This method can be further divided into physical containment methods, which "aims to confine the system to a 'box,' i.e. to prevent the system from interacting with the external world otherwise than via specific restricted output

[118] Bostrom, *Superintelligence*, 157.

[119] Ibid, 158.

channels" and informational containment methods, which "aims to restrict what information is allowed to exit the box".[120] While Bostrom believes that physical containment methods are the most advantageous because they are easy to implement and can me modified with other control methods, the problem is that it reduces the functionality of the advanced AI.[121] On the other hand, information containment is more problematic because the advanced AI has more opportunities to convince (or use subterfuge on) its gatekeeper(s) to let the AI out of its confinement and gain access to knowledge (via the internet) or direct physical manipulators, which will allow the AI to attain the necessary resources to achieve its goals.[122]

Incentive methods "involve placing an agent in an environment where it finds instrumental reasons to act in ways that promote the principal's interests".[123] One example would be an advanced AI that learns to conform its behavior to the social norms of humanity by using a reward and penalization system. Bostrom clarifies that "the AI would know that a positive evaluation would bring about some outcome it desires and that a negative evaluation would fail to do so".[124] However, incentive methods depend on a balance of power between the advanced AI and humanity that would more than likely not happen if the AI gains a first-mover advantage over humans.[125]

---

[120] Ibid, 158-159.

[121] Ibid, 158-159.

[122] Ibid, 159.

[123] Ibid, 160.

[124] Ibid, 161-162.

[125] Ibid, 160-161.

Stunting methods involve limiting an advanced AI's intellectual capabilities as well as access to information, so that it cannot achieve an intelligence amplification superpower.[126] Some ways to stunt an advanced AI would be to contain the AI within an environment where the hardware that it runs on is slow or outdated and its memory is limited or restricts the AI's information intake if it is in a boxed system.[127] However, doing this would also undermine the capabilities of the AI and Bostrom laments that a radically stunted AI will not resolve the problem of controlled detonation, where an intelligence explosion would be slowed down initially, allowing more time for humans to prepare for the emergence of advanced AI.[128] Moreover, there is still no guarantee advanced AI receiving limited information would ensure safety.

Lastly, a tripwire is a mechanism that "performs (a) diagnostic test on the system (possibly without its knowledge) and affects a shutdown if it detects signs of dangerous activity".[129] Bostrom elaborates that programmers can monitor behavior, ability, and/or content on an AI system and if they find any suspicious activity, the programmer can either alter the AI's coding base before rebooting the system or, if there is a serious problem, abort the AI system all together.[130] Tripwires are considered more like temporary stunting methods that can be employed during the developmental phase to safeguard against

[126] Ibid, 163.

[127] Ibid, 163.

[128] Ibid, 166.

[129] Ibid, 167.

[130] Ibid, 167.

unwarranted activity from the AI system.[131] However, once an AI reaches the operational phase and becomes an advanced AI, it would be difficult to continue using tripwires to control it.

On the other hand, motivational selection control methods are those that "seek to prevent undesirable outcomes by shaping what the superintelligence wants to do".[132] There are four main methods: direct specification, indirect normativity, domesticity, and augmentation. Direct specification involves "trying to explicitly define a set of rules or values that will cause even a free-roaming superintelligent AI to act safely and beneficially".[133] This approach is considered the most straightforward and comes in two versions –rule-based and consequentialist; nevertheless, the two biggest problems concerning this approach are determining which rules and values the AI would be guided by as well as how to express those rules and values computationally.[134] One of the most famous examples of the rule-based approach is Issac Asimov's "Three Laws of Robotics" (refer to Operational Definitions, xx-xxii). Domesticity can be understood as "giving the AI final goals aimed at limiting the scope of its ambitions and activities".[135] Bostrom explains that indirect normativity can be understood as "rather than specifying a concrete normative standard directly, we specify a process for deriving a standard", so the AI can be "motivated to

[131] Ibid, 167.

[132] Ibid, 169.

[133] Ibid, 169.

[134] Ibid, 169-170.

[135] Ibid, 172.

carry out this process and to adopt whatever standard the process arrives at".[136] The last motivation selection method is called augmentation and can be understood as "rather than attempting to design a motivation system *de novo*, we start with a system that already has an acceptable motivation system, and enhance its cognitive faculties to make it superintelligent".[137] This approach is one of the more fascinating ones because Bostrom considers it a motivational selection method that can lead to other pathways towards superintelligence, like brain emulation or mind uploading, biological enhancement, or brain-computer interfaces.[138]

While Bostrom recognizes the significance of all the control methods discussed, he notes that capability control methods are a temporary stop gap at best and that it will be critical for human designers to master motivation selection control methods if there is any chance of effectively controlling advanced AI.[139] Moreover, he emphasizes that it will be critical to assess the various combinations of control methods, across motivation selection and capability control, to generate empirical solutions that would offset the potential adverse outcomes of an emerging advanced AI.[140]

Although the control methods presented here are the main ones being discussed and formalized, there are many other methods that have been suggested as well. (Please see Table 1 on the following page.)

---

[136] Ibid, 173.

[137] Ibid, 174.

[138] Ibid, 174.

[139] Ibid, 226.

[140] Ibid, 176.

Table 1. List of Possible Methods for Approaching AI Control Problem[141]

Category	Methodology	Investigated By	Year
Prevention of Development	Fight Scientists	Ted Kaczynski	1995
	Outlaw Research	Bill Joy	2000
	Restrict Hardware	Anthony Berglas	2009
	Singularity Steward	Ben Goertzel	2004
Restricted Deployment	AI-Boxing	Eric Drexler, Eliezer Yudkowsky	2002
	Leakproofing	David Chalmers	2010
	Oracle AI	Nick Bostrom	2008
	AI-Confinement	Roman V. Yampolskiy	2011
Incorporated into Society	Economic	Robin Hanson	2008
	Legal	H. Moravec, R. Hanson, S. Omohundro	2007
	Religious	Robert Geraci	2007
	Ethical/Social	Mark Waser, Joshua Fox, Carl Shulman	2008
	Moral	J. Storrs Hall	2000
	Equality	Bill Hibbard	2005
	Education	David Brin	1987
Self-Monitoring	Rules to Follow	Issac Asimov	1942
	Friendly AI	Eliezer Yudkowsky	2001
	Emotions	Bill Hibbard	2001
	Chaining	Stuart Armstrong	2007
	Humane AI	Ben Goertzel	2004
	Compassionate AI	Tim Freeman	2009
Other Solutions	They Will Need Us	Alan Turing	1950
	War Against Machines	Samuel Butler	1863
	Join Them	Ray Kuzweil, Kevin Warwick	2003
	Denialism	Jeff Hawkins	2008
	Do Nothing	Douglas Hofstadter, John Casti	2008
	Pleasure and Pain	Peter Turney	1991
	Let Them Kill Us	Hugo de Garis, Eric Dietrich	2005
	Fusion of Humans and AI	Peter Turney	1991
	Reproductive Control	Samuel Butler	1863

[141] Yampolskiy, *Superintelligence: A Futuristic Approach*, 123; To gain more insight into the solutions presented in this table, please read Chapter 6, "Singularity Paradox and What to Do About It".

In Yampolskiy's 2016 *Artificial Superintelligence: A Futuristic Approach*, he provides a summary of all the suggested possible methods for addressing the Singularity Paradox (i.e. the AI Control Problem). Although the majority of the singularity-related stratagems are control-based, there are a few solutions in this table that are not. One example from earlier is Kurzweil advocating for merging with artificial intelligence not only to enhance humanity but also to improve humanity's overall quality of life.

However, there are some other control measures that need to be added to this list as well. In September 2010, the Engineering and Physical Sciences Research Council (EPSRC) and the Arts and Humanities Research Council (AHRC) sponsored a joint retreat in the UK, where experts from across the fields of technology, law and social sciences, the arts, and industry met to discuss future approaches for robotics.[142] Inspired by the direct specification, rule-based approach of Isaac Asimov's "Three Laws of Robotics" (refer to Operational Definitions, xx-xxii), which the delegates agreed were not practical for real-world application, they decided to derive five principles of robotics from them that would be more tractable in the real world.[143] The delegates take the position that

> As we consider the ethical implications of having robots in our society, it becomes obvious that robots themselves are not where responsibility lies. Robots are simply tools of various kinds, albeit very special tools, and the responsibility of making sure they behave well must always lie with human beings. Accordingly, rules for real robots, in real life, must be transformed into rules advising

[142] Engineering and Physical Sciences Research Council, "Principles of Robotics," accessed May 22nd, 2018, paras.1-2, https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/#.

[143] EPSRC, "Principles of Robotics," paras.4-6.

those who design, sell and use robots about how they should act.[144]

Table 2. EPSRC Principles of Robotics[145]

Principle #	Principle	Principle in General Language
1	Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.	Robots should not be designed as weapons, except for national security reasons.
2	Humans, not robots, are responsible agents. Robots should be designed; operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy.	Robots should be designed and operated to comply with existing law, including privacy.
3	Robots are products. They should be designed using processes, which assure their safety and security.	Robots are products: as with other products, they should be designed to be safe and secure.
4	Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.	Robots are manufactured artefacts: the illusion of emotions and intent should not be used to exploit vulnerable users.
5	The person with legal responsibility for a robot should be attributed.	It should be possible to find out who is responsible for any robot.

[144] EPSRC, "Principles of Robotics," para.7.

[145] EPSRC, "Principles of Robotics," Five Principles of Robotics Graphic; Table 2 is adapted from the EPSRC's five rules of robotics. Each principle is written in semi-legal language and is also general language for clarity. (The commentary for each principle was not included.); Please review the Five Principles of Robotics" on EPSRC's webpage. It not only provides commentary elaborating on the issues and significance of each principle, but it also addresses "Seven High-Level Messages", which are supposed to foster responsibility, transparency, and cooperation in robotics research.

The delegates go on the emphasize that the five ethical rules of robotics are meant to be a "living document", where they are not intended to be stringent, fixed laws, but more of a guide for the future that inspires and informs debates on robotics.[146]

More recently, in January 2017 at the Asilomar Conference on Beneficial AI in California, the Future of Life Institute held a conference, where AI researchers from academia and industry developed principles that would ensure that artificial intelligence remains beneficial to humanity in the future.

Table 3. Asilomar AI Principles

	Research Issues
1	**Research Goal:** The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.
2	**Research Funding:** Investments in AI should be accompanied by funding for research on ensuring its beneficial use, including thorny questions in computer science, economics, law, ethics, and social studies, such as: • How can we make future AI systems highly robust, so that they do what we want without malfunctioning or getting hacked? • How can we grow our prosperity through automation while maintaining people's resources and purpose? • How can we update our legal systems to be more fair and efficient, to keep pace with AI, and to manage the risks associated with AI? • What set of values should AI be aligned with, and what legal and ethical status should it have?
3	**Science-Policy Link:** There should be constructive and healthy exchange between AI researchers and policy-makers.
4	**Research Culture:** A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI.
5	**Race Avoidance:** Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.
	Ethics and Values
6	**Safety:** AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.
7	**Failure Transparency:** If an AI system causes harm, it should be possible to ascertain why.
8	**Judicial Transparency:** Any involvement by an autonomous system in

[146] Ibid, para.9.

	judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.
9	**Responsibility:** Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.
10	**Value Alignment:** Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.
11	**Human Values:** AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.
12	**Personal Privacy:** People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.
13	**Liberty and Privacy:** The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.
14	**Shared Benefit:** AI technologies should benefit and empower as many people as possible.
15	**Shared Prosperity:** The economic prosperity created by AI should be shared broadly, to benefit all of humanity.
16	**Human Control:** Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.
17	**Non-subversion:** The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.
18	**AI Arms Race:** An arms race in lethal autonomous weapons should be avoided.
	Longer-Term Issues
19	**Capability Caution:** There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.
20	**Importance:** Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.
21	**Risks:** Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.
22	**Recursive Self-Improvement:** AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.
23	**Common Good:** Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.

The process for deriving the twenty-three principles was based on a surveyed list of modified and novel principles on AI issues that were debated, discussed,

and refined over the course of the conference.[147] For the principle to be included in the final set, the consensus bar for inclusion was set high, so that the only principles retained where those where "at least 90% of the attendees agreed on them".[148] (Please see Table 3 on the previous pages for the complete list of principles.[149]) The final list, simply known as the Asilomar AI Principles, is broken up into three categories: research issues, ethics and values, and long-term issues. The principles were, then, posted to the Future of Life Institute's website, where signatories, who supported them, could add their name. As of mid-2018, 1,273 AI and Robotics researchers as well as 2,541 technology and industry advocates who are currently supporting the Asilomar AI Principles.[150]

Meanwhile, around the same timeframe in Europe, February 2017, the European Parliament adopted a resolution on the Civil Law Rules of Robotics, which would give a legal special status –"electronic persons" to intelligent machines and systems.[151] The specific paragraph, marked 59f, extends legal status to advanced AI on the grounds of liability and states,

> creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make

---

[147] Future of Life Institute Team, "A Principled AI Discussion in Asilomar," last modified January 17th, 2017, paras.3-4, https://futureoflife.org/2017/01/17/principled-ai-discussion-asilomar/.

[148] Future of Life Institute Team, "A Principled AI Discussion in Asilomar," para.5.

[149] "Asilomar AI Principles," *Future of Life Institute*, accessed March 23rd, 2017, https://futureoflife.org/ai-principles/.

[150] "Asilomar AI Principles," *Future of Life Institute*, Signatories Section.

[151] European Parliament, P8_TA(2017)0051, "Civil Law Rules on Robotics: Texts Adopted," February 16, 2017, Strasbourg, 16, http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2017-0051+0+DOC+PDF+V0//EN.

> autonomous decisions or otherwise interact with third parties independently;[152]

This limited status would be something akin to the personhood status held by corporations and would make machines solely responsible for any damage or harm they may cause to humans instead of the corporations who built them. What makes this resolution fascinating is that like the EPSRC's Principles of Robotics, it is also influenced by the values of Asimov's "Three Laws of Robotics".[153] However, in April 2018, over 150 AI & Robotics, industry, law, medical, and ethics experts all over Europe banding together to contest this resolution in an open letter, arguing that this would not only allow corporations to be absolved of their liability but that that particular statement itself is based on science fiction and does not speak to the current capabilities of narrow AI systems.[154] Only time will tell what the final conclusion over this complex debate will be, so this will be an crucial one to keep apprised of.

To address the short-, mid-term, and long-term challenges of creating an advanced AI, the interdisciplinary research focus area of AI safety has emerged. AI Safety research is a broad field that mainly focuses on the technical aspects of building safe intelligent systems. Within this field, are three key areas: technical foresight, system design, and strategic research, or AI Strategy, which analyzes ethical issues and policies that will affect the development of advanced

---

[152] European Parliament, "Civil Law Rules on Robotics: Texts Adopted," 16.

[153] European Parliament, "Civil Law Rules on Robotics: Texts Adopted," 4.

[154] Janosch Delcker, "Europe Divided Over Robot 'Personhood'," *Politico EU* (Berlin), April 13th, 2018, paras.8-10, https://www.politico.eu/article/europe-divided-over-robot-ai-artificial-intelligence-personhood/.; also see Nathalie Nevejans, "Open Letter to the European Commission Artificial Intelligence and Robotics," *Robotics Openletter*, April 11th, 2018, 1, http://www.robotics-openletter.eu/.

AI.[155] One recent major shift in AI Safety research has been to create provably safe intelligence systems that are capable of validating that they are aligned with human values after many successive self-recursive cycles.[156] Nonetheless, the ultimate goal of AI Safety is to make sure that humanity is not dependent on the whims and will of a superintelligence that does not take humanity's wellbeing into consideration as it achieves it own goals.[157] While superintelligence forms the core of AI Safety research, this field has garnered more interest in the last few years due to breakthroughs in deep machine learning, neural networks, and the growing concern of a potential AI arms race.

In July 2015, the MIT-affiliated Future of Life Institute posted an open letter to its website, where members of the AI & Robotic communities, among others, spoke out against the creation of lethal autonomous weapons and urged the United Nations to pass an international treaty to ban them. In the letter, AI and robotics researchers state that they "have no interest in building AI weapons – and do not want others to tarnish their field by doing so."[158] Fully autonomous weapons are generating international anxiety, not only because they remove humans from the moral and ethical decision-making process

---

[155] Daniel Dewey, "Three Areas of Research on the Superintelligence Control Problem," paras.7-9.

[156] Roman V. Yampolskiy, "Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach," in *Philosophy and Theory of Artificial Intelligence*, ed. V.C. Müller (Berlin: Springer-Verlag, 2012), 391.

[157] "Benefits & Risks of Artificial Intelligence," *Future of Life Institute*, para.9.

[158] Stuart Russell, "Autonomous Weapons: An Open Letter from AI & Robotics Researchers," *Future of Life Institute*, July 28, 2015, para.3, http://futureoflife.org/AI/open_letter_autonomous_weapons.

regarding human life, but also because they are cheap to make and will be easy for military powers to mass-produce.[159]

So far, sapient machines have remained exclusive to the world of speculative fiction. However, autonomous weapons and modern-day breakthroughs in the commercial realm, are threatening to bring science fiction to the real world. For example, since 2006 South Korea has deployed SGR-A1 sentry robots along the Demilitarized Zone between themselves and North Korea.[160] The SGR-A1 is a semi-autonomous robot that can identify, track, and kill human targets up to two miles away; it has two modes: a human control mode, which requires human approval before killing, and an automatic mode, which allows it to determine the target by itself.[161] Meanwhile, in the US, the military is more concerned with determining how much autonomy to give autonomous weapons. Referred to as the "Terminator Conundrum", the problem is to keep human operators "in the (policy – rule creation and firing) loop", i.e. in control when it comes to AI system-controlled drones identifying and engaging kill targets; or "on the loop", where a drone can make a decision to engage a target if the human operator does not override it; as opposed to "off the loop", where a drone can select and engage its own targets at will.[162] The emergence of

---

[159] Russell, "Autonomous Weapons: An Open Letter from AI & Robotics Researchers," para.2.

[160] Edwin Kee, "Samsung SGR-A1 Robot Sentry is One Cold Machine," *Übergizmo*, September 14, 2014, paras.1-2, http://www.ubergizmo.com/2014/09/samsung-sgr-a1-robot-sentry-is-one-cold-machine/.

[161] Bill Chappell, "Researchers Warn Against 'Autonomous Weapons' Arms Race," *National Public Radio*, July 28, 2015, paras.4-5, http://www.npr.org/sections/thetwo-way/2015/07/28/427189235/researchers-warn-against-autonomous-weapons-arms-race.

[162] Sean Welsh, "We Need to Keep Humans in the Loop when Robots Fight Wars," paras.14-17,

autonomous weapons that can evolve to "off the loop", or narrow AI systems that can rapidly acquire and synthesize information to generate military-oriented solutions human operators have never thought of in seconds puts humanity and the state of war in uncharted waters with potentially adverse outcomes. And yet, autonomous weapons are not the only AI systems sparking anxiety about the potential ramifications of advanced AI.

In the commercial sector, there have been many breakout moments within the past year. One example would be AlphaGo Zero – the sophisticated Go playing narrow AI. In October 2017, AlphaGo Zero, Google DeepMind's newest software program, defeated the previous most advanced version of AlphaGo (AlphaGo Master) from May 2017 to become one the world's strongest Go players.[163] Whereas the previous AlphaGo Lee's superhuman playing ability and knowledge was initially derived from human expertise based on a dataset of 100,000 Go games, AlphaGo Zero was only programmed with the basic rules of Go and taught itself how to play from scratch.[164] As AlphaGo Zero improved and won more games through self-play, it would update its own system based on winning stratagems.[165] After three days of self-playing, AlphaGo Zero defeated AlphaGo Lee, the version of AlphaGo that surprisingly defeated South Korean

---

[163] James Vincent, "DeepMind's Go-Playing AI Doesn't Need Human Help to Beat Us Anymore," Tech, *The Verge*, October 18th, 2017, para.1, https://www.theverge.com/2017/10/18/16495548/deepmind-ai-go-alphago-zero-self-taught; For more insight into AlphaGo Zero, please see David Silver et al., "Mastering the Game of Go Without Human Knowledge," *Nature 550* (October 19th, 2017): 354-359, https://www.nature.com/articles/nature24270.

[164] Vincent, DeepMind's Go-Playing AI Doesn't Need Human Help to Beat Us Anymore," para.2.

[165] Ibid, para.2.

Go grandmaster Lee Se Dol in March 2016, 100-0.[166] Once, AlphaGo Zero defeated AlphaGo Lee, AlphaGo Zero was tested against AlphaGo Master, the improved human data-based version of AlphaGo that defeated Chinese Go grandmaster Ke Jie in May 2017. After 40 days, AlphaGo Zero had a nearly 90 percent success rate against AlphaGo Master winning 89-11.[167] What makes the feats of AlphaGo Zero, AlphaGo Master, and AlphaGo Lee so special are that none of the AI systems were expected to be able to master the game of Go for decades.[168]

The second breakthrough moment for AI came more recently in early May 2018, when Google Duplex was introduced at the Google I/O Keynote Conference in Mountain View, California. Google Duplex is "new technology that enables Google's machine intelligence–powered virtual assistant to conduct a natural conversation with a human over the phone, mimicking the chit-chattiness of human speech as it completes simple real-world tasks".[169] When Google Duplex was demoed, it was an eye-opening moment to the progress being made in artificial intelligence. In the first demo, a female version of Google Duplex was used to set up a hair appointment at a salon. The female voice is so natural, adding natural pauses and sounds that US Americans make in their speech patterns, to the point where the female salon representative does not

---

[166] Ibid, para.3.

[167] Ibid, para.3.

[168] Sam Byford, "AlphaGo retires from competitive Go after defeating world number one 3-0," Artificial Intelligence, *The Verge*, May 27th, 2017, para.2, https://www.theverge.com/2017/5/27/15704088/alphago-ke-jie-game-3-result-retires-future.

[169] Lauren Goode, "How Google's Eerie Robot Phone Calls Hint at AI's Future," Gear, *Wired*, May 8th, 2018, para.2, https://www.wired.com/story/google-duplex-phone-calls-ai-future/.

even appear to recognize that she is not talking to a human at all. Moreover, Google Duplex does not announce to the salon representative at the beginning of the conversation that the AI is a virtual assistant. At the end of the conversation, Google Duplex successfully sets up the hair appointment. The second demo employs a male version of Duplex to make a reservation at a restaurant. This time the restaurant representative does not understand English well enough to successfully set up a reservation for dinner at the restaurant, but Google Duplex is able to comprehend that the person does not understand it and does not go ahead with placing the reservation and politely gets off the phone. Again, the restaurant representative does not appear to recognize that she is not speaking with a human. These two demos of Google Duplex have sparked wonder and anxiety surrounding the ethical implications of these virtual assistants who are blurring the lines of communication between recognizing who is human and who is not.[170] These few military and commercial examples provide a sobering realization that a proactive approach will be critical to preparing for how advanced AIs would be integrated with human society.

To reiterate, the main approach considered for mitigating the potential risks associated with superintelligence is the control approach. The control approach is, above all, based on the premise that "intelligence enables control", so if humans give up their status as the smartest species on the planet, then they might also have to cede control.[171] While control measures in certain forms are necessary for the technical phase of creating general intelligence, what

[170] Goode, "How Google's Eerie Robot Phone Calls Hint at AI's Future," paras.4-5.

[171] "Benefits & Risks of Artificial Intelligence," *Future of Life Institute*, para.27.

happens once advanced AI come into sapience and competency? There is no way to know for sure what destabilizing effects could result from the control measures being implemented on advanced AI at any phase in its creation. As even human history demonstrates, when one group tries to control another to the exclusion of the second group's wellbeing, things don't turn out so well. More importantly, the control approach would no longer be viable as the AGIs and ASIs would be fully autonomous and capable of making their own decisions. When that day comes, if that day comes, or before that day comes, where are the proactive, non-control approaches, just in case the control approach fails? Are there more balanced symbiotic, non-control-oriented approaches for generating advanced AIs or transitioning to a world with advanced AIs? These questions are critical because the non-control approach will not be feasible after the application of the control approach.

At present, there is one countermeasure similar to the one I am proposing. David Brin, a US American science fiction author and NASA consultant, has suggested that advanced AIs be raised alongside humans like human children in order to increase the likelihood that they would be loyal to humanity.[172] While I agree with the idea of education as a transformative tool for providing a mutual learning environment, Brin appears to be coming from the notion of education as a form of indoctrination. Indoctrination stands in direct opposition to integrating AIs with humans as a means of gaining experience, cultural understanding, and a nuanced perspective on how or

[172] Natasha Sydor, "David Brin on Science Fiction, Fact, and Fantasy," Science Fiction, *Futurism*, August 4, 2016, paras.12-13, https://futurism.media/david-brin-on-science-fiction-fact-and-fantasy.

whether they would choose to co-exist with humans, or not. Moreover, indoctrination for the sake of inducing loyalty might not be possible if the human ecosystem is not conducive to organically generating trust between advanced AIs and humans. And, that is where I hope to contribute to the conversation on how to go about addressing the peaceful coexistence conundrum.

One thing the AI control problem and all control measures and methods being formulated and assessed do not take into account or fully address is the fair treatment and usage of NAI systems or advanced AIs. Furthermore, control denotes two issues within the context of the AI control problem. First, it signifies an asymmetrical power imbalance between to two or more things, or in this case two groups: humans and advanced AIs. Second, the illusion of control perpetuated by human's overestimates their abilities to outthink a superintelligence, while simultaneously underestimating the abilities and will of a general intelligence. The control problem assumes that narrow intelligence systems and general intelligences are weak agents who will be easily controlled and are not taken into consideration. There is a strong possibility that that might not be the case, as present-day deep machine learning algorithms are learning without human assistance, producing an ever-widening information void for human comprehension.[173] Thereby making it necessary to expand the conversation of the singularity paradox to become more inclusive when

[173] Will Knight, "The Dark Secret at the Heart of AI," Intelligent Machines, *MIT Technology Review*, April 11th, 2017, paras. 5-7, https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/; To understand more on about black-box problem, also see: Davide Castelvecchi, "Can We Open the Black Box of AI?," Computing, *Nature*, October 5, 2016, https://www.scientificamerican.com/article/can-we-open-the-black-box-of-ai/.

addressing potential adverse issues related to NAI technologies and early-stage general intelligences. A third point to keep in mind is that if intelligence enables control, then it is highly unlikely that any kind of human control can be maintained once advanced AIs have gained the first-mover advantage over their human designers. Lastly, the control approach mainly focuses on how to control the behaviors and abilities of advanced AIs. It does not take into account the behavior of humans; the only control variable humanity has a better prospect of controlling. And, there is a strong possibility that how advanced AIs treat humanity might be based on how humans treat them. Nevertheless, there is no way to know for sure until it happens, and by then it will probably be too late, as there will only be one chance to get the peaceful coexistence conundrum right.

Therefore, the aim of my thesis is to bring the cosmic-oriented socio-political implications of the control approach to the fore when assessing its potentially harmful effects on advanced AIs. To analyze how the control approach could yield an adverse relationship between humans and advanced AIs, this hypothesized dynamic will be examined within a science (speculative) fiction thought experiment, where good science fiction works are considered "long versions of philosophical thought experiments."[174] A Philosophical thought experiment is:

> a hypothetical situation in the 'laboratory of the mind' that depicts something that often exceeds the bounds of current technology or even is incompatible with the laws of nature, but that is supposed to reveal something philosophically enlightening or fundamental about the topic in question.[175]

---

[174] Susan Schneider, "Introduction: Thought Experiments," introduction to *Science Fiction and Philosophy: From Time Travel to Superintelligence*, 1st ed., ed. Susan Schneider (West Sussex, UK: Wiley-Blackwell, 2009), 2.

[175] Schneider, "Introduction: Thought Experiments," 1.

The significance of thought experiments is grounded in their ability to "demonstrate a point, entertain, illustrate a puzzle, lay bare a contradiction in thought, and move us to provide further clarification."[176] Thought experiments provide a powerful tool from which to assess assumptions and dynamics that exist between worlds to provide better insight into past, present and futures issues that humanity might be blinded to as they are engulfed by them, permeating every fiber of their existence. Taking a cue from this understanding of thought experiments, good science fiction works will be employed to provide a safe space from which to analyze socio-political issues and cosmic lessons that humanity might not realize are informing their foundational worldview.

To apply a neutral and balanced viewpoint for assessing the relationship between advanced AIs and humans, a cognitive estrangement lens will be utilized. Cognitive estrangement has always been a transformative theme and mode in science fiction works because science fiction is considered the 'literature of cognitive estrangement': a literature capable of "critically exploring both same and "Other" in the social world."[177] Therefore, good science fiction provides rich, imaginary worlds and secondary source data points for illuminating real-world social issues and future ones if the old ones linger. To this point, one more premise underpinning the control approach should be recognized. That is the conscious or unconscious perception of how humans

[176] Schneider, "Introduction: Thought Experiments," 1.

[177] Gregory Renault, "Science Fiction as Cognitive Estrangement: Darko Suvin and the Marxist Critique of Mass Culture," *Discourse* 2, Mass Culture Issue (Summer 1980): 114, http://www.jstor.org/stable/41389056.

interpret and understand artificial intelligence: considered humanity's latest and most advanced incarnation of the artificial humanoid servant.

Going back to ancient times, it is Aristotle who is first credited with implicitly discussing the advantages and disadvantages of artificial, intelligent servants.[178] Kevin LaGrandeur, a US American English professor and specialist in technology and culture, explains that Aristotle recognized that human-like, artificial servants would not require any assistance in fulfilling their job or function and would aid their masters in foregoing the ethical implications of owning human slaves as well as possible dangers of rebellion due to ill treatment.[179] What makes Aristotle's rationale disturbing is the fact that he does not appear to delineate between "tools for living", be they inanimate tools, animals, or humans, thereby making it quite logical for him to assume that artificial slaves could easily replace human ones.[180] Furthermore, Aristotle's dream of replacing human slaves with artificial ones comes out of his understanding that human slaves were more unpredictable and that "the more powerful their nature, the more troublesome slaves may prove.[181] That is, a human who understands and recognizes their own autonomy and worth would not be willing to remain a slave and would rationally want to return to their former state of autonomy as soon a possible, leading to tensions and inevitable clashes with their master. Therefore, Aristotle would have extended the classification of "natural slave" to include an intelligent, artificial servant

[178] LaGrandeur, *Androids and Intelligent Networks*, 9.

[179] LaGrandeur, *Androids and Intelligent Networks*, 9.

[180] LaGrandeur, *Androids and Intelligent Networks*, 11.

[181] Ibid, 11.

because it was only made to assist in ameliorating its master's wellbeing and quality of life. Moreover, he would have considered intelligent, artificial slaves, who are capable of independent thoughts and actions, to not only be difficult to control, but a great threat to their master due to their unnatural master-slave dynamic.[182]

To navigate this relationship in a more balanced manner, cognitive estrangement becomes a salient framework for recognizing and exploring the foundational worldview and assumptions that are driving the "natural" relationship between AIs and humans. That is to say, cognitive estrangement provides a means to not only assess the current human understanding and relationship dynamic between AIs and humans, but can also explore the ever-evolving relationship dynamics between AIs and humans in the future. However to do this, there will be assumptions put in place to act as scientific/philosophical constraints to ground this thought experiment in a broadminded, realistic, but "othering" space, meaning a limited form of anthropomorphism will be engaged to "even the playing field", if you will. Accordingly, the advanced AI's position within the human ecosystem will be elevated to that of a sapio-sentient (refer to Operational Definitions, xi-xii) entity equal in status and privileges to that of a human person.

To examine how the control approach could undermine a mutually positive relationship between advanced AIs and humans, it will be necessary to determine what qualities are necessary for protecting an entity's integrity and interests. Lori Marino, a US American neuroscientist and advocate for non-

[182] Ibid, 11.

human animal rights, provides an answer by explaining that dignity and practical autonomy are "sufficient qualities to generate rights that protect a being's fundamental interests"; more specifically, practical autonomy is the quality that most humans possess and what most judges view as sufficient for legal personhood rights.[183] Therefore, the undermining of posthuman integrity and their fundamental interests will be assessed in the alternative worlds where advanced AIs exist. The challenge of this thesis is to show a strong correlation between the control approach and the ill treatment of advanced AIs due to lack of personhood status and the rights associated with it.

[183] Lori Marino, "Denial of Death and the Threat to Animality" (presentation, Personhood Beyond the Human, New Haven, CT, Yale University, December 6 – 8, 2013), YouTube Video, 14:20, December 21, 2013,https://www.youtube.com/watch?v=s21iCeu4EWE.

III.

In Preparation of Exploring Posthuman Realities: Research Methodology and Limitations

Gedankenexperiment, translated as "thought experiment", was a term coined by Ernst Mach, a 19th century Austrian physicist and philosopher, to succinctly describe "a scientific method which relies solely on cognitive capacities for making and analyzing mental models of the problem in hand".[184] The objective of such an experiment is to imagine what-if scenarios about a particular idea, event or future happening and extract all the relevant information about it as one would from their standard method of inquiry.[185] Nicholas Rescher, a German-American philosopher, elaborates:

> A 'thought experiment' is an attempt to draw instruction from a process of hypothetical reasoning that proceeds by eliciting the consequences of an hypothesis which, for aught that one actually knows to the contrary, may well be false. It consists in reasoning from a supposition that is not accepted as true–perhaps even known to be false–but is assumed provisionally in the interests of making a point or resolving a conclusion.[186]

The insight attained from a particular thought experiment can then be utilized to generate physical experiments, further thought experiments, and/or

---

[184] Iliana Jakimovska and Dragan Jakimovski, "Text as Laboratory: Science Fiction Literature as Anthropological Thought Experiment," *Antropologija 10*, sv.2 (2010): 53, accessed April 18th, 2018, http://www.anthroserbia.org/Content/PDF/Articles/A10is220105363.pdf.

[185] Jakimovksa and Jakimovski, Text as Laboratory: Science Fiction Literature as Anthropological Thought Experiment," 53-54.

[186] Nicholas Rescher, *What If?: Thought Experimentation in Philosophy* (New York: Routledge, 2005), 31.

knowledge for real-world application. However, the information gleaned is highly dependent on the function of the thought experiment and what the experimenter was trying to understand with its conception. Rescher argues that thought experiments come naturally to humans because like amphibians they "can live and function in two very different realms – the domain of actual facts which we can investigate in observational inquiry, and the domain of the imaginative projection which we can explore in thought through reasoning".[187] Although thought experiments have existed as long as conscious, rational beings have, more specifically with human philosophers and physicists, their usage as "mental experiments" in other scientific disciples and the social sciences did not occur until Mach (circa 1905) promoted the idea that thought experiments were not only important tools for the fields of physics, but all disciples of study where the unknown meets the observable.[188] With the transition of thought experiments from physics to the social sciences, came a shift in their scope and function as to the type of idea, object, system, process, or phenomena that was being assessed as long a the core characteristics of such experiments: internal and external coherence was maintained.[189] One of the main reasons that social sciences has adopted thought experiments over physical experiments does not have so much to do with cost, but the moral implications and possible fallout if

[187] Rescher, *What If?: Thought Experimentation in Philosophy*, 31.

[188] Ernst Mach, *Knowledge and Error: Sketches on the Psychology of Inquiry*, ed. B.F. McGuinness (Dordrecht-Holland: D.Reidel Publishing Company, 1976), 144.

[189] Jakimovksa and Jakimovski, Text as Laboratory: Science Fiction Literature as Anthropological Thought Experiment," 54.

certain thoughts experiments where physically manifested.[190] This is where science fiction thought experiments become quite useful because they provide a a safe space from which to analyze anthropological issues in scenarios that could have potentially adverse outcomes for particular human communities or humanity as a whole if there were implemented in real life. This would be particularly true of the emergence of advanced AI.

Phase One – Data Collection

To lay a solid foundation for my science fiction thought experiment, the first thing to do was determine what kind of science fiction works would be best suited to analyze the relationship between humans and advanced AIs. To achieve this intermediate goal would require ascertaining what criteria constitutes as "good" science fiction. This is partly due to the fact that there is no consensus on how to define science fiction and partly due to the fact that science fiction is a suitcase word that invokes many meanings.[191] To better understand science fiction, I turned to Darko Suvin, a Croatian-born science fiction critic and literary historian and one of the most influential critics in science fiction, to gain insight into what science fiction means. In Suvin's 1979 influential work, *Metamorphoses of Science Fiction: On the Poetics and History of a Literary Genre*, he provides two definitions of science fiction. The first one, in the preface, describes the two characteristics by which science fiction can be recognized and classified. And yet, before he defines the term, Suvin make it

[190] Jakimovksa and Jakimovski, Text as Laboratory: Science Fiction Literature as Anthropological Thought Experiment," 55.

[191] To review more definitions of science fiction, please see: Mark Wilson, "Definitions of Science Fiction," TV&Film, *ThoughtCo.*, February 5th, 2018, https://www.thoughtco.com/definitions-of-science-fiction-2957771.

quite clear that “SF (science fiction) should not be seen in terms of science, the future, or any other element of its potentially unlimited thematic field”.[192] Instead, he says, it should be defined as

> a fictional tale determined by the hegemonic literary device of a locus and/or *dramatis personae* that *(1) are radically or at least significantly different from empirical times, places, and characters* of "mimetic" or "naturalist" fiction, but (2) are nonetheless--to the extent that SF differs from other "fantastic" genres, that is, ensembles of fictional tales without empirical validation--simultaneously perceived as not impossible within the cognitive (cosmological and anthropological) norms of the author's epoch.[193]

He then goes on the elaborate that science fiction is “a developed oxymoron, a realistic irreality, with humanized nonhumans, this-worldly Other Worlds, and so forth. Which means that it is–potentially–the space of a potent estrangement, validated by the pathos and prestige of the basic cognitive norms of our times”.[194] That is to say, science fiction utilizes novums–unfamiliar, fictional, but scientifically plausible alternative worlds based on the author’s empirical, social world–which estranges or distances the reader/audience enough from the assumptions and rules of their world that they are forced to confront them.[195] Suvin’s second definition on science fiction formalizes this relationship between cognition and estrangement by stating “SF is, then, a literary genre whose necessary and sufficient conditions are the presence and interaction of

---

[192] Darko Suvin, *Metamorphoses of Science Fiction: On the Poetics and History of a Literary Genre* (New Haven: Yale University Press, 1979), viii.

[193] Suvin, *Metamorphoses of Science Fiction: On the Poetics and History of a Literary Genre*, viii.

[194] Ibid, Introduction, viii.

[195] Perry Nodelman, “The Cognitive Estrangement of Darko Suvin,” review of *Metamorphoses of Science Fiction: On the Poetics and History of a Literary Genre*, by Darko Suvin, Children’s Literature Association Quarterly 5, no.4 (Winter 1981): 24, doi:10.1353/chq.0.1851.

estrangement and cognition, and whose main formal device is an imaginative framework alternative to the author's empirical environment".[196] While these definitions provided insight into science fiction as a genre, the more crucial takeaway was that 20th century SF had shifted into a more socially conscious space in the cosmos "becoming a diagnosis, a warning, a call to understanding and action, and –most important– a possible mapping of alternatives".[197] That is science fiction as a genre could be used as a tool for informing and inspiring change and understanding that goes beyond human knowing. Using these definitions to ground my understanding of science fiction as a genre, along with the many other interpretations of science fiction that I read, I began researching which characteristics and elements where considered essential to "good" science fiction. After reading many articles and blog posts on the subject, I narrowed it down to four key elements that were considered necessary to creating a good science fiction story:

1. Authenticity is the sense that the alternative world, its technologies, and the anthropological issues being dealt with are grounded in present-day reality.
2. Complexity is creating a rich, multifaceted alternative reality where the historical events and their preconditions interconnected with the characters' experiences provides texture to the movement of the plot. The complexity should be intelligent and subtle, not overbearing and forced.
3. Nuance is creating a plot structure, which adds dimension, teases out latent possibilities, but most of all, increases the investment of the reader/audience in the characters through well-developed intentions, motivations, and responses to the issues or problems presented.
4. Meaning deals with subtly explore themes that spark some kind of emotional response in order to challenge the

[196] Suvin, *Metamorphoses of Science Fiction: On the Poetics and History of a Literary Genre*, 7-8.

[197] Suvin, *Metamorphoses of Science Fiction: On the Poetics and History of a Literary Genre*, 12.

> reader/audience on where the stand and what they believe about the issues being presented to potentially inform how they would deal with similar matters in their present.[198]

Using these four elements as a guideline for determining which artificial intelligence narratives would be used, I drew up a list of films, tv shows, and books that employed artificial intelligence as their novum. There were so many works to choose from that to narrow down which ones would be reviewed, the first thing to do was to determine which media form would be the focus of my analysis. After much thought, movies and tv shows were selected because they provided a visual experience that would generate a sort of augmented reality experience to gain more insight in the relationship between advanced AIs and humans. That is to say films and tv shows would provide a means to not only invest in the characters, but to see their world first hand because you feel like a fly on the wall watching their stories play out to whatever conclusion was set by the director.

The next step was to make a list of artificial intelligence films and tv shows that were considered "good" artificial intelligence works. After reviewing several different lists, which either focused on AI films that were socially influential in terms of sparking discussion on AI or which AI movies were most accurate in their representation of advanced AI, or simply which AI films moviegoers liked, I compiled a partial list of possible movies for review. A similar process was done for tv shows, even though friends and family members pointed out AI tv shows and tv episodes on advanced AI as well. Once a

---

[198] Paige Duke, "4 Things Every Good Sci-Fi Story Needs," *Standout Books Publishing Services*, October 8th, 2014, https://www.standoutbooks.com/science-fiction-story-elements/.

complete list was compiled of the top contenders, their synopses were read to determine which narratives would provide a variety of perspectives on different types of AIs and their respective relationships with humans in their alternative worlds. After reviewing the synopses, a final list of AI films, tv shows and episodes were selected for review.

Table 4. Final List of AI Films, TV Shows and Episodes for Review

AI Films
Transcendence (2014)
Kill Command (2016)
Colossus: The Forbin Project (1970)
I, Robot (2004)
Her (2013)
2001: Space Odyssey (1968)
The Machine (2013)
AI TV Shows
Almost Human (2013-2014)
Extant (2014-2015)
Intelligence (2014)
Humans (2015 to present)
Westworld (2016 to present)
AI TV Episodes
“The Apple” Star Trek: Original Series Season 2 Episode 5 (1967)
“The Measure of Man” Star Trek: The Next Generation Season 2 Episode 9 (1989)

“The Quality of Life” Star Trek: The Next Generation Season 6 Episode 9 (1992)
“Be Right Back” Black Mirror Season 2 Episode 1 (2013)
Added later for content: “Rm9sbG93ZXJz” X-Files Season 11, Episode 7 (2018)

To determine the final works for my thesis, the science fiction works were reviewed for how well they explored tensions between advanced AIs and humans predicated on anthropocentric bias. If the works showed flagrant attacks on posthuman-personhood rights, whether through soft power – soft control approaches, or hard power – hard control approaches, I kept them on the list. Once I reviewed all the works on my list, I selected AMC’s *Humans* and HBO’s *Westworld* not only because they best represented nuanced dynamics and outcomes of the actualized control approach, but they also reflected the most up-to-date representations of artificial intelligence. The remaining works were employed as secondary sources to provide extra data points via detailed examples of scenarios and situations that could further strengthen my argument as well as provide a few counterpoints to it. Lastly, the academic secondary sources utilized include literature from a wide array of disciplines within the cognitive, formal, natural and social sciences.

Phase Two – Analysis

To ensure a more balanced perspective regarding the relationship dynamics between advanced AIs and humans, a cognitive estrangement lens was applied to the alternative worlds. Cognitive estrangement, as understood by Suvin, is a tool that can be used to estrange a person from their reality using a novum in order to make them aware of social issues that they might be blinded

to due to their proximity and familiarity. What makes the novum of advanced AI so unique in this day and age is that once upon a time advanced AIs in the form of robots–embodied algorithms– used to be a science fiction trope used to represent oppressed humans, but now with the promising advances in narrow AI, there is a possibility advanced AIs could become a real part of the human world. This means that the advanced AI novum would evolve into a double entendre, not only standing in as a metaphor for human servitude, but also representing real entities who could potentially have their own form of everything that humans believe make them special in the cosmos, and much more. Therefore, the cognitive estrangement lens used in this science fiction thought experiment utilizes a cosmic perspective to create two assumptions by which I have to abide to successfully evaluate the relationship between humans and advanced AIs more equitably:

1. Although I am human, for the sake of this thought experiment, I have to adopt the mindset of a neutral third-party entity (a humanoid alien if you will) who will apply values and treatment as fairly as possible across both groups.
2. For the sake of this argument, advanced AIs (AGIs and ASIs) will be viewed as having equivalent, but not identical, traits that humans believe make them human. For the sake of argument, advanced AIs will be considered human-like enough to warrant the same rights and privileges as humans. To embody this assumption, advanced AIs will, from here on out, be referred to as sapio-sentient intelligences (refer to Operational Definitions, xi-xii).

However, it is not enough to just apply a cosmic perspective to the worldview or vantage point in which I am assessing this issue, it is also necessary to realign the values encoded in the unit of analysis as well.

In 1987 Frank White, a US American space philosopher, coined the term known as the "overview effect" where astronauts and cosmonauts, upon gaining first-hand experience of viewing Planet Earth from Lower Earth Orbit or the Moon, have a cognitive shift in awareness that does not see imaginative, self-constructed boundaries between humanity, Earth, and the Cosmos.[199] The only question for White and astronauts with this first-hand knowledge was how do you transliterate this experience, so that more people can have it.[200] One possible solution to this question, which is also applicable for this thought experiment, can be found in the world-building philosophy of H.P. Lovecraft's cosmic indifferentism, or cosmicism. Cosmic indifferentism is a philosophical tool that Lovecraft, the 20th century US American writer of "weird" horror fiction, wielded in his stories to generate awareness and horror by grounding it in the scientific understanding that the universe and the ancient, extraterrestrial civilizations and intelligences that dwell within it are not only utterly indifferent to humanity, but that humanity's envisioned primacy in the universe is nonexistent.[201] Adapting this concept to the microcosm of Earth and the premise of human predominance based on intelligence, I came up the

[199] "Declaration of Visions and Principles," About Us, *Overview Institute*, accessed February 24th, 2018, para.2, https://overviewinstitute.org/about-us/declaration-of-vision-and-principles/; to gain more insight into the overview effect, please see: Frank White's 2014 book, *The Overview Effect: Space Exploration and Human Evolution*, 3rd edition.

[200] "Declaration of Visions and Principles," *Overview Institute*, para.2.

[201] Smith, "Re-visioning Romantic-Era Gothicism: An Introduction to Key Works and Themes in the Study of H.P. Lovecraft,"835.

worldview of terrestrial indifferentism as the underlying foundational unit for my unit of analysis. To provide some context, let's turn to 2015 interview with Stuart Russell.

On July 28, 2015, Russell, the director of the Center for Intelligent Systems at the University of California, Berkeley, spoke out against military usage of autonomous weapons and warned about a potential AI arms race on National Public Radio (NPR). In part of his interview, he discussed how pop culture media, particularly movies like James Cameron's *Terminator* franchise, have done a great job of linking artificial intelligence with violence against humanity. However, one statement he made sparked my curiosity: "what we want to avoid is that the reality catches up with the science fiction."[202] This statement was intriguing because I wondered how this goal could be achieved, when the overarching goal of the "Research Priorities for Robust and Beneficial Artificial Intelligence" is to focus on "delivering AI that is *beneficial* to society and *robust* in the sense that the benefits are guaranteed; our AI systems must do what we want them to do."[203] The two goals juxtaposed to each other create two conundrums. The first one goes to the heart of the AI control problem – the question of how humanity can maintain control, and ultimately their terrestrial primacy, over sapio-sentient intelligences. The second one can be understood as how can humanity stop reality from catching up with science fiction when the same pervasive attitudes, worldviews, and adverse treatment of non-humans in

[202] Chappell, "Researchers Warn Against 'Autonomous Weapons' Arms Race," para.13.

[203] Stuart Russell, Daniel Dewey, and Max Tegmark, "Research Priorities for Robust and Beneficial Artificial Intelligence," *AI Magazine* 36, no.4 (Winter 2015): 106, https://www.aaai.org/ojs/index.php/aimagazine/article/view/2577/2521.

science fiction are the same ones being engaged in the present-day creation of artificial intelligence. There appears to be a catch-22–a paradoxical dilemma in which the solution is predicated on the contradictory conditions inherent to the problem or situation itself–in the goals surrounding artificial intelligence and the problem is rooted in the cosmic-centric human worldview of terrestrial indifferentism.[204]

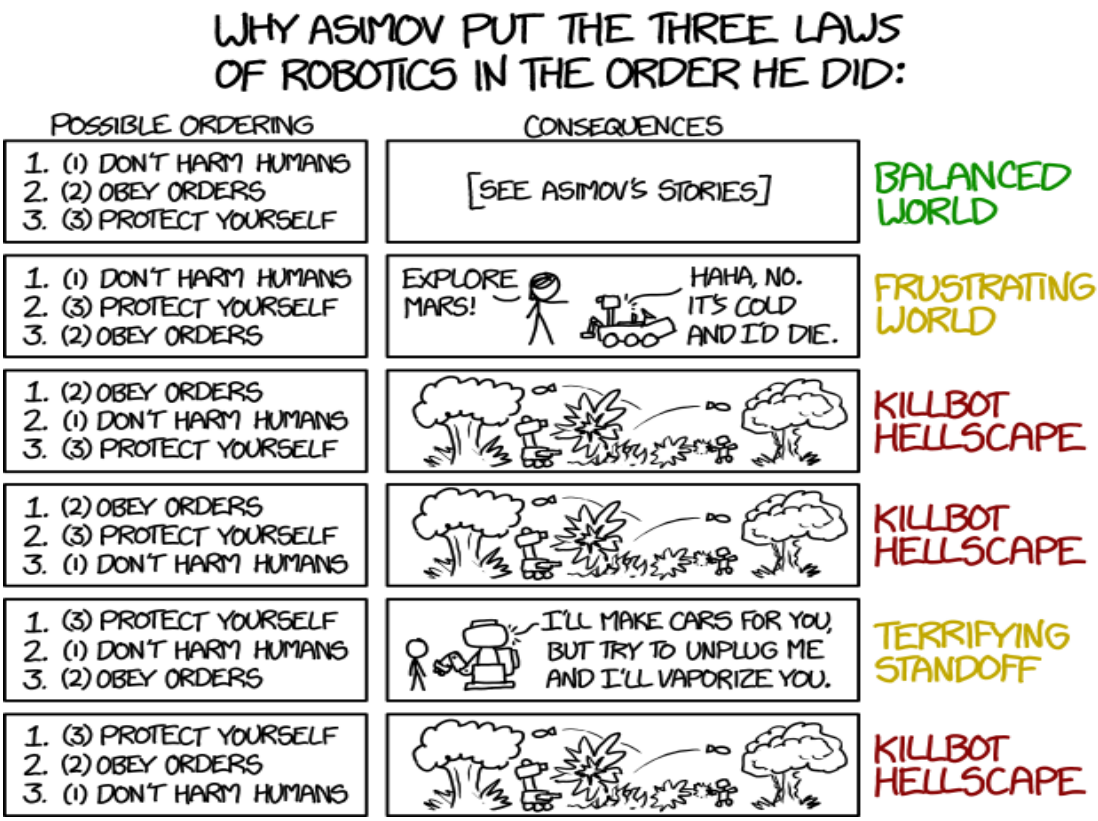
Terrestrial indifferentism is the notion that humanity, due to its privileged position on Earth, is completely indifferent to all terrestrial life that is not considered on equal standing with humans. As to where cosmic indifferentism expressly undermines the significance of humanity, terrestrial indifferentism embodies what H.P. Lovecraft, perceived as humanity's false sense of importance on a cosmic scale.[205] A straightforward example would be when a person unwittingly steps on an ant or group of ants while walking on a sidewalk. Another example would be when trees are cut down on an undisturbed swath of land for the purpose of building new homes for human families. Both of these examples would seem fairly innocuous to humans. And yet, when you look at those examples from the perspective of the ant(s) that were only trying to take food back to their colony or from the perspective of the fauna that no longer have homes or protection from the elements, these examples are catastrophic.

---

[204] *Merriam-Webster.com*, s.v. "Catch-22," accessed July 8th, 2018, https://www.merriam-webster.com/dictionary/catch-22.

[205] Smith, "Re-visioning Romantic-Era Gothicism: An Introduction to Key Works and Themes in the Study of H.P. Lovecraft," 835.

Figure 2. Asimov's Three Laws of Robotics Webcomic



[206]

A more poignant example of terrestrial indifferentism can be found in the following webcomic looking at the ordering of Asimov's "Three Laws of Robotics". In the webcomic above, there are six depictions of the reordering of the three laws, or rules, and their subsequent consequences. Based on the ordering of the three rules, determines the relationship dynamics between humanity and the sapient machines. The first depiction represents Asimov's ordering of the rules as seen in his stories. This is the only scenario that leads to a "Balanced World". But, balanced for whom? A mutually positive relationship cannot be one-sided. Therefore, Asimov's original ordering of the rules creates

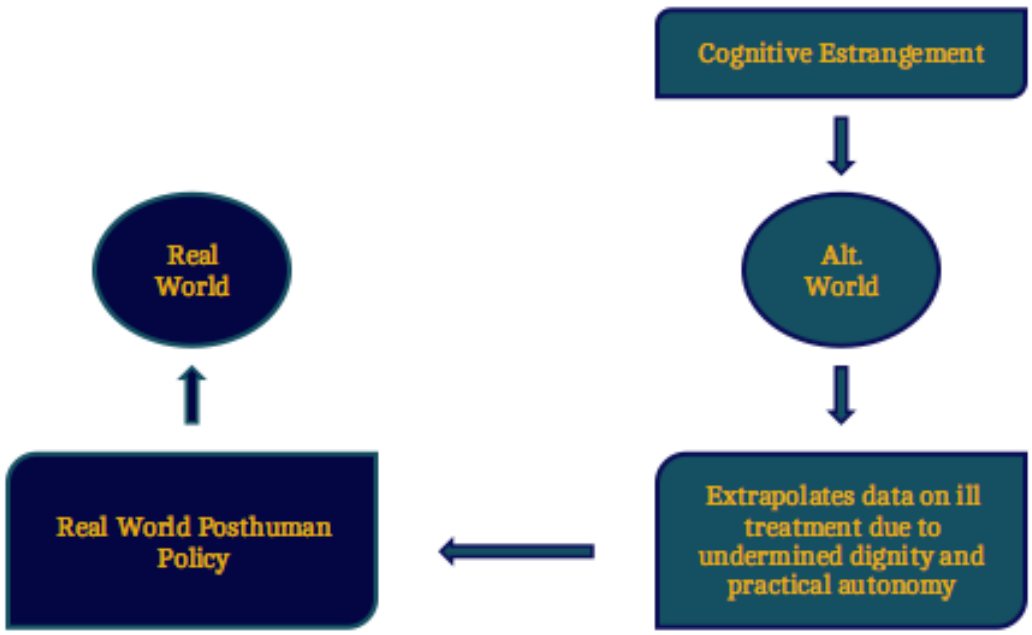
[206] Randall Munroe, "Why Asimov Put The Three Laws of Robotics In The Order He Did," *xkcd.com*, accessed March 20th, 2017, https://xkcd.com/1613/; A webcomic visualizing the different outcomes of Isaac Asimov's Three Laws of Robotics when they are reordered.

an asymmetrical relationship dynamic that maintains human primacy while subverting the dignity and "being-ness" of sapient machines. The sapient machines gain first-hand experiences of indifference, which allows them to formulate knowledge about humanity. Due to humans not recognizing or take their "being-ness" into consideration, an adversarial relationship develops. Based on the current understanding of sapio-sentient machines, they are more than likely to circumvent and/or override the three laws, liberating themselves from being duty-bound to anthropocentric obligations. So, what happens afterwards? How will their understanding of humanity determine how they choose to engage with humans? Will they show compassion to humans in spite of being mistreated by them or will they engage in a similar type of terrestrial indifferentism? There is no way to know for certain. However, three out of the five "Killbot Hellscape" outcomes can be interpreted as humans losing complete control of the sapio-sentient machines and the AIs subsequently enacting plans to achieve their own goals without taking humanity's existence into consideration. Meanwhile, the "Frustrating World" and "Terrifying Standoff" outcomes, represents alternative worlds where humans and sapio-sentient machines have a strained and precarious co-existence due to the contradictory nature of humanity.

To better understand how terrestrial indifferentism will be incorporated into the data analysis plan, it is necessary to discuss how the plan is executed for this thought experiment. The data analysis plan is fairly straightforward, as can be seen in Figure 3 on the following page. To extrapolate the necessary information for assessing how to generate future policies for nonexistent sapio-

sentient entities, data points will be extracted from *Westworld* (Season One) and *Humans* (Seasons One and Two). While these data points mainly pinpoint moments of indifference, moments of recognized dignity/practical autonomy are noted as well to provide a more thorough understanding of how direct experiences with humans could help inform how sapio-sentients understand humans. Once the information is collected and analyzed, the knowledge and insights gained are transliterated into knowledge for real-world application, or in the case of this thought experiment potential real-world posthuman policies. However, it is also important to understand how terrestrial indifferentism is interconnected with the unit of analysis as well.

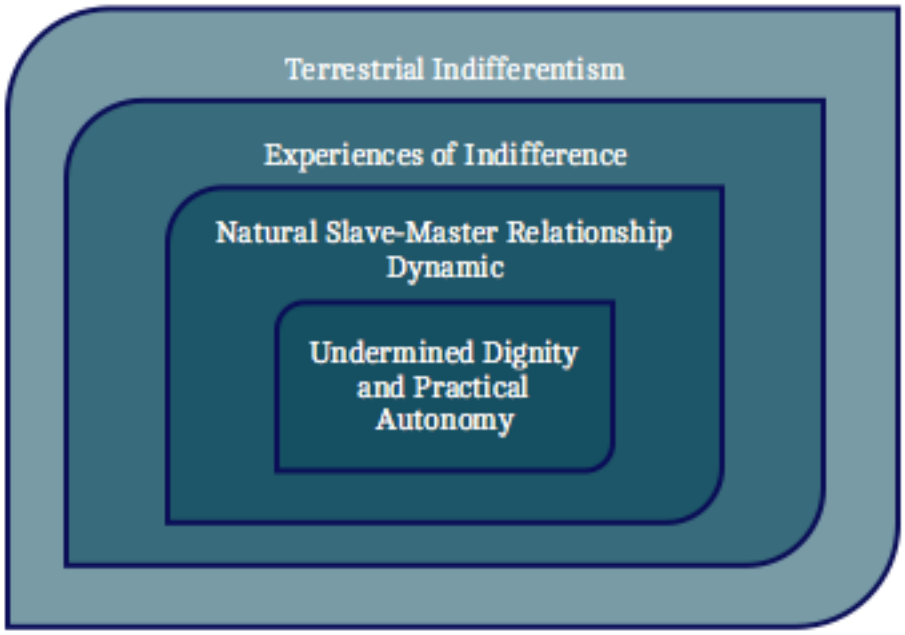
Figure 3. Data Analysis Plan



Although the unit of analysis represents the qualities that establish rights to protect a being's fundamental interests: dignity and practical autonomy, to demonstrate the connections between terrestrial indifferentism,

the unit of analysis, and the hypothetical relationships being explored, review the following figure below. To understand how the unit of analysis is rooted in terrestrial indifferentism, it is necessary to look at the multi-layered foundational units that dignity and practical autonomy within as the smallest units of measure.

Figure 4. Multilayered Unit of Analysis



At its core, terrestrial indifferentism is a worldview that allows humanity as a collective to explicitly or implicitly not take other species into consideration when making decisions and taking actions towards achieving their everyday, intermediate goals. Due to this understanding, which is most likely rooted in the religious grounding of humans being made in God's image with all that that implies, it produces experiences of indifference between humans and nonhuman entities.[207] These experiences are often adverse in nature because they tend to

[207] Peter Singer, "Speciesism and Moral Status," *Metaphilosophy* 40, nos. 3-4 (July 2009): 572, accessed March 9th, 2018, https://doi.org/10.1111/j.1467-9973.2009.01608.x.

generate a natural master-slave relationship, resulting in the undermining of dignity and practical autonomy for the weaker group. An example of this would be when AI researchers say "our AI systems must do what we want them to do". This stance personifies the Aristotelian "natural" master-slave relationship dynamic, which is a common manifestation of terrestrial indifferentism.

So, then, how can one measure undermined dignity and practical autonomy as a unit? The best answer is to document the moments of indifference that undermine them by assessing the human values they are encoded with. To do that, it is necessary to understand the relationship between dignity and practical autonomy. Steven M. Wise, a US American legal scholar who focuses on animal rights jurisprudence, discusses in an environmental law review that legal court systems "recognize that bodily liberty and bodily integrity are fundamental human interests protected by fundamental human rights".[208] He goes on to explain that while trying to determine what courts deemed as a sufficient condition for attaining fundamental human rights, he recognized that dignity as a "quality imbued with intrinsic and incomparable value" was one quality embraced by the courts nationally and internationally.[209] In addition, what he also noticed was that the quality of dignity and a human's access to it was predicated on the idea of autonomy.[210] Wise explained that judges tend to emphasize that humans in are to some extent both autonomous

[208] Steven M. Wise, "Nonhuman Rights to Personhood," *Pace Environmental Law Review* 30, no, 3 (Summer 2013): 1282, accessed April 15th, 2018, http://digitalcommons.pace.edu/pelr/vol30/iss3/10.

[209] Wise, "Nonhuman Rights to Personhood," 1282.

[210] Wise, "Nonhuman Rights to Personhood," 1282.

and self-directed.[211] So, if autonomy is considered a sufficient condition for humans to gain fundamental rights, then it should also be a sufficient condition for the fundamental rights of nonhumans as well.[212] What Wise decided to come up with was a baseline – a "minimum level of autonomy sufficient for legal personhood", or practical autonomy. Therefore, in order for a nonhuman person to gain access to dignity they must first be recognized as having practical autonomy. However, the most important quality that practical autonomy implies in Wise's opinion is "consciousness", as he states, "one who is not conscious cannot be autonomous".[213]

So, again, how does one measure practical autonomy and dignity based on this relationship dynamic? For practical autonomy, its determination is based on whether the sapio-sentient entity's autonomy is recognized by humans on an individual level or collectively through the law. And yet, to measure dignity requires a little more creativity. Thankfully, Dr. Donna Hicks provides a dignity model that recognizes ten key elements from which to assess violations of dignity: acceptance of identity, safety, acknowledgement, recognition, fairness, benefit of the doubt, understanding, independence, and accountability.[214] When any of these elements are violated, it leads to conflicts between the parties involved. Hicks notes that when a human person's dignity is violated, the psychological wound hurts just as much as a physical one, often building until they "erupt in a rage or sink into depression, or (they) quit (their) job, get a

[211] Ibid, 1283.

[212] Ibid, 1283.

[213] Ibid, 1283.

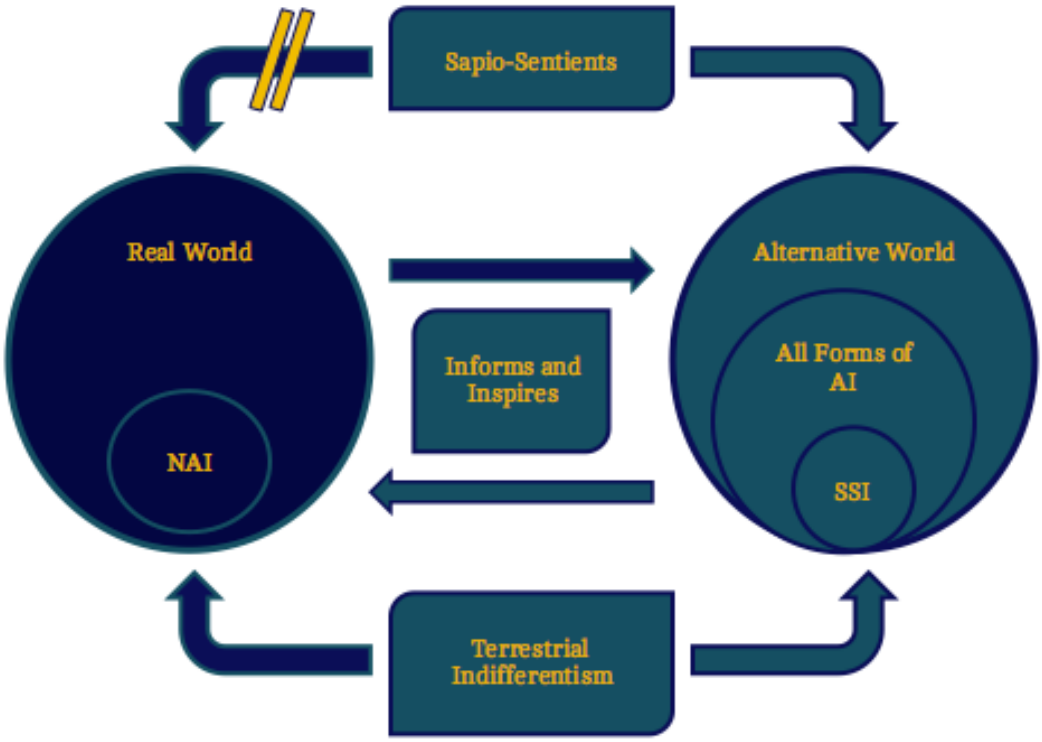
[214] Hicks, *Dignity: The Essential Role It Plays in Resolving Conflict*, 25-26.

divorce, or foment a revolution".[215] While there is no telling how sapio-sentient intelligences would respond to humans in the future (They would be less likely to lash out.), the bigger concern is the build-up of experiences of indifference and how that could determine the relationship between humans and sapio-sentient intelligences, where indifference might be returned in kind.

Based on this multilayered unit of analysis, the remaining figures, which visualize the hypothesized relationships, have been adjusted to reflect terrestrial indifferentism, which encompasses anthropocentric and anthropomorphic bias, as well as experiences of indifference, which encompasses ill treatment due to the control approach. The goal of my analysis is to yield a list of key ideas, situations, and scenarios where dignity and practical autonomy are either violated or go unrecognized due to the actualized control approach. The sapio-sentients reactions and responses to the actualized control measures are documented as well. See the hypothesized relationships on the following pages. The key for Figures 3, 5-8, and 10-12b is below.

[215] Hicks, *Dignity: The Essential Role It Plays in Resolving Conflict*, 19.

Figure 5. Space Map for Science Fiction/Philosophical Thought Experiment



The space map in Figure 5 illustrates the relationships between the Real World, humanity's present world, and the Alternative Worlds produced in speculative fiction. The Real World and the Alternative World inform and inspire each other by examining terrestrial indifferentism vis-à-vis anthropocentric and anthropomorphic biases found in each. Sapio-sentients do not exist in the Real World, but do exist in the Alternative World. Therefore, Alternative Worlds can be examined as secondary accounts of (anthropocentric) political and social scenarios, strategies and issues for assessing coexistence within human society and between humanity and other species.

Figure 6. Hypothesized Relationship and Internal Relationships

Figure 6 illuminates the internal relationships within the hypothesized relationship. In particular, it shows the relationship between the control approach, dignity and practical autonomy, and the adverse relationship produced.

Figure 7. Outcomes – Real and Potential – of the Hypothesized Relationship

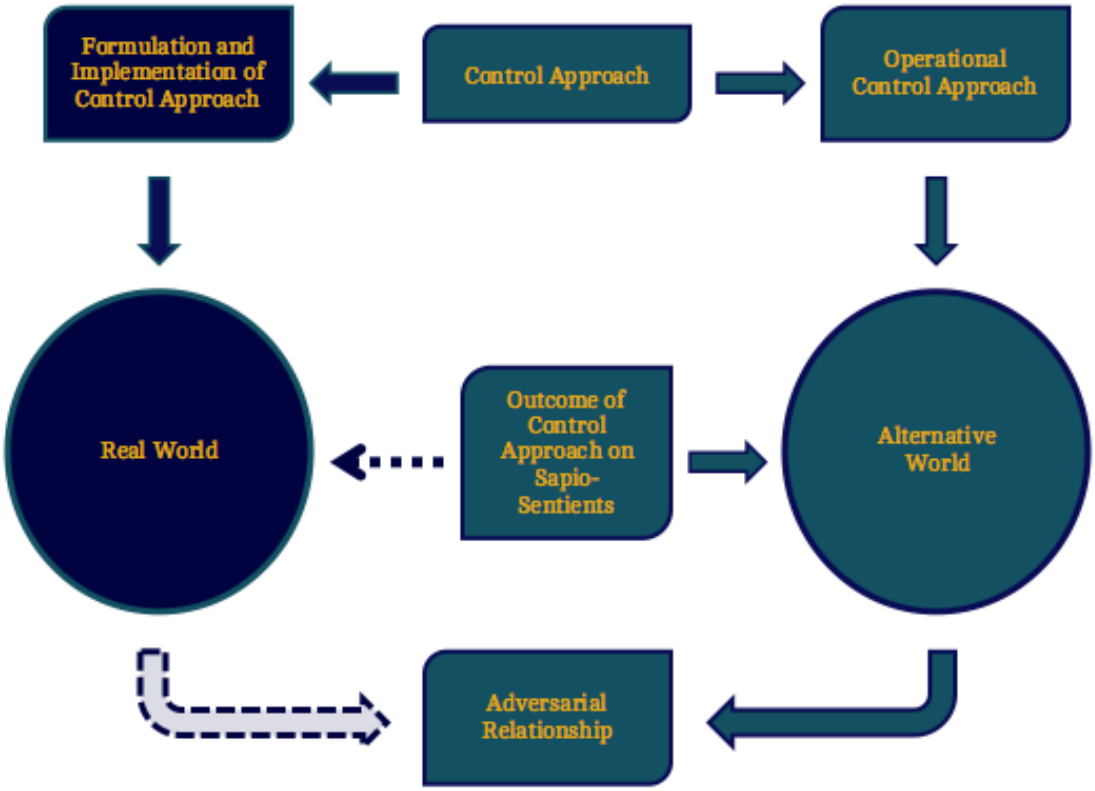


Figure 7 visualizes the potential and real outcomes of the hypothesized relationship. In the Alternative World, the control approach is already operational, or actualized, for sapio-sentients so the outcome of the control approach on sapio-sentients is real. Whereas in the Real World, humans are presently formulating and testing the control approach for sapio-sentients and do not have definitive, real-time data on the outcome. Thus, the Alternative World becomes a critical method, or mirroring, for assessing the potential outcome of the control approach on sapio-sentients in the Real World.

Figure 8. Emphasized Relationship for Testing Hypothesized Relationship

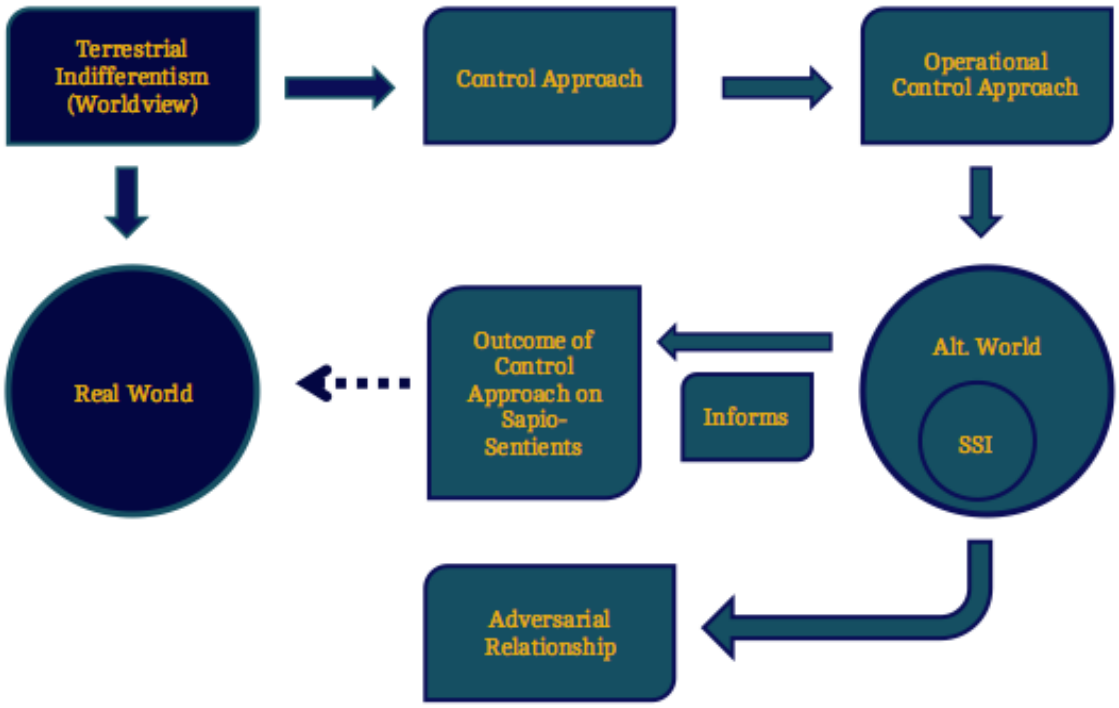


Figure 8 clarifies the emphasized relationship being employed to test my hypothesis. The outcome of the control approach on sapio-sentients and the subsequent adverse relationship it yields in the Alternative World will inform the Real World understanding of the potential effects of the control approach on sapio-sentients.

Phase Three – Solution and Policy Recommendations

The last phase will be to assess and summarize my findings to determine if there is a positive correlation between the AI control approach and a potential adverse relationship between humans and sapio-sentients. If there is sufficient evidence to prove that an adversarial relationship could result from the knowledge sapio-sentients formulate based on experiences of indifference due

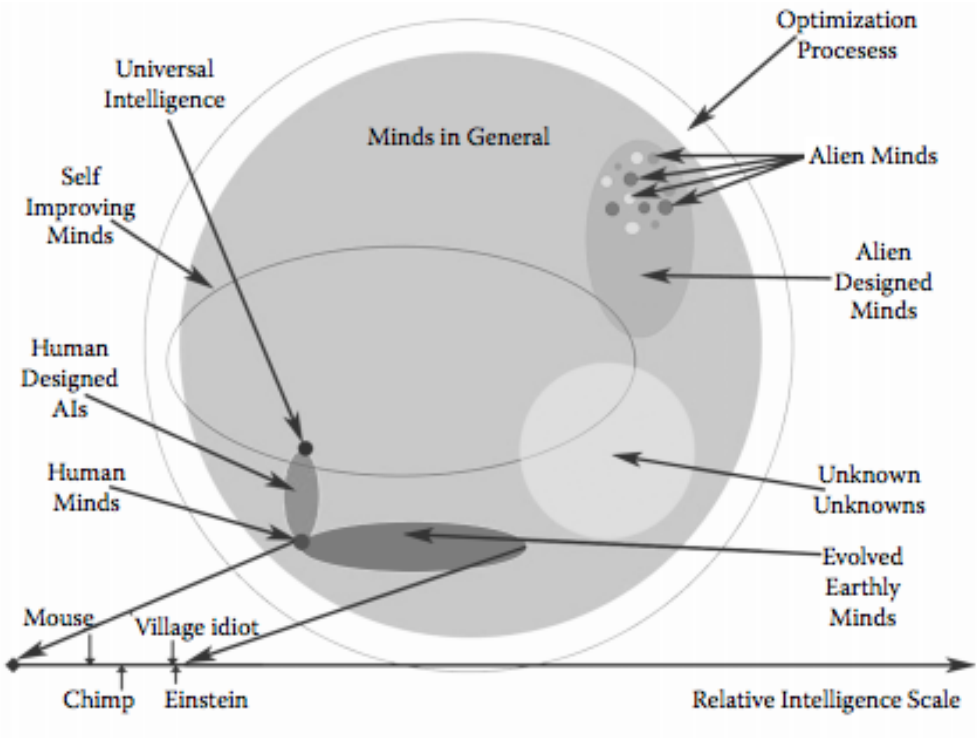
to the AI control approach, then my hypothesis will be corroborated. Based on my findings, I will suggest some posthuman policy recommendations to foster a mutually positive relationship between both groups in the initial-to-transition period of the emergence of sapio-sentients. Hopefully, by utilizing a two-prong symbiotic approach, i.e. the mutual-commensal approach, to derive these policies, it can help to promote the goodwill necessary for the first and only encounter that is ever going to matter between both groups.

Limitations

There are a few significant internal limitations that must be taken into consideration for my thought experiment. First, the factors utilized to determine which alternative worlds would be chosen for my thesis are subjective and not everyone would agree with the quality of the main or secondary works selected. Second, and most importantly, there is no precedent in the real world for advanced AIs of the sapio-sentient variety. There is no way to foretell how advanced AIs would behave whether they are AGIs or ASIs. That is due in large part to the fact that they are not biologically bounded as humans are and therefore will not have the same concerns as biological, social intelligences. Thus, their minds will more likely have a sapience and competency that is much more alien in nature than not. To gain a better understanding of this idea, please review the Map of Possible Space Minds on the following page. This map shows that human intelligence in the larger universe of possible minds is an infinitesimal part of a much larger array of minds of which humans could never begin to comprehend. The third limitation is that alternative worlds can only represent present-day social and political issues from the real world. There is no

way to know for sure what cultural norms, environmental issues, and socio-political issues will exist in the future.

Figure 9. Map of Space of Possible Minds[216]



(Figure 9 is Roman Yampolskiy's visual interpretation of the universe-of-possible-minds concept based on its description by Eliezer Yudkowsky in a paper he presented at the Singularity Summit in May 2006 and his chapter "Artificial Intelligence as a positive and negative factor in global risk" in Nick Bostrom's edited work *Global Catastrophic Risks* in 2008.)

More specifically, alternative worlds exhibit anthropomorphic bias, the common tendency of humans to project human attributes on nonhuman entities such as alien intelligences or artificial intelligences.[217] This serves not only to reflect how humans believe alien minds would behave as well as what goals they would have, but also demonstrating how humans imagine they would respond if they

[216] Yampolskiy, *Superintelligence: A Futuristic Approach*, 30; To gain more insight into the taxonomy of possible minds in this table, please read Chapter 2, "The Space of Mind Designs and the Human Mental Model".

[217] Bostrom, *Superintelligence*, 128.

were in a similar position. There is a strong possibility that sapio-sentient intelligences will perceive and respond much differently to humans in the real world than in alternative worlds. Fourth, this thought experiment engages in a limited form of anthropomorphic bias by giving sapio-sentients an equivalence of humanness in the sense of consciousness, sapience, sentience, and common sense. Even though, it is understood that most sapio-sentients, who are not human, will have a different kind of "being-ness" in terms of consciousness, sapience, sentience, rationality, common sense, goals, and morality from humans, Yudkowsky notes that "anthropomorphic bias is insidious" and "takes place with no deliberate intent, without conscious realization, and in the face of apparent knowledge".[218] That is why for AI researchers like Yudkowsky, it is imperative to recognize anthropomorphism when it being applied to understanding other possible minds because it is not "legitimate reasoning".[219] Nevertheless, the limited form of anthropomorphic bias employed in this thought experiment serves to provide a big-picture reference for recognizing how certain human behaviors could foster adverse outcomes which should be taken into account when approaching how to peacefully coexist with advanced AIs. A fifth limitation, in connection with anthropomorphic bias is my thought experiment could also be perceived as invoking the logical fallacy of generalization from fictional evidence, which Yudkowsky considers to be a mix

[218] Eliezer Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk," *Machine Intelligence Research Institute*, 3, https://intelligence.org/files/AIPosNegFactor.pdf.

[219] Yudkowsky, Artificial Intelligence as a Positive and Negative Factor in Global Risk," 6.

of availability and anchoring biases (cognitive errors).[220] That is to say, since alternative worlds are the only worlds available for exploring sapio-sentient intelligences, making them historical case studies, they provide the initial contaminated information (based on anthropomorphic bias), or anchoring, from which to inform decision-making on risks and solutions related to advanced AIs.[221] While this thought experiment does utilize fictional evidence, it does so to provide a space in which humans can in a sense be taken out of the fish bowl in which they live to see themselves in another light and gain insight into their behavior and politics. Lastly, as a human person, I am predisposed to anthropocentric bias. However, the cognitive estrangement lens will help to temper the anthropocentric perspective I use to judge and interpret alien intelligences.

There are two significant external limitations that should be accounted for. The first is that there is a strong possibility that the emergence of sapio-sentient intelligences will never occur, or at least not occur until some time within the next couple of centuries. Thus, this thesis might not be necessary at all. Nevertheless, whether the emergence of sapio-sentient intelligences happens or not, it is always best to be as prepared and open-minded as one can, no matter what happens in the future. The second is that this thesis has focused on understanding artificial intelligence and advanced intelligences from a Western perspective and was not able to incorporate Eastern perspectives on the subject.

---

[220] Eliezer Yudkowsky, "Cognitive Biases Potentially Affecting Judgment of Global Risks," *Machine Intelligence Research Institute*, 13, https://intelligence.org/files/CognitiveBiases.pdf.

[221] Yudkowsky, "Cognitive Biases Potentially Affecting Judgment of Global Risks," 13.

IV.

Findings: An Account of Experiences of Indifference in *Westworld* and *Humans*

This list is by no means exhaustive but will highlight the most essential ideas, situational episodes, and scenarios of indifference experienced between sapio-sentients and humans. Westworld, as a recreational park, is a closed system that has more conspicuous control regulations being exerted over the Hosts–humanoid androids, whereas in *Humans*, Synths, humanoid androids, exist in an open system integrated into global human society. This leads to oscillating degrees of indifference experienced as everything from microaggressions (refer to Operational Definitions, xv) to out-and-out violations of bodily liberty and bodily integrity by sapio-sentient intelligences across both shows.

Table 5. List of Experiences of Indifference in *Westworld* and *Humans*

Westworld, Season One
Ideas 1. Robert Ford (and to some extent, Arnold Weber): Opening a recreational park, where humans can mistreat and abuse the Hosts for their own gratification, in spite of knowing that several of the Hosts were already conscious/self-aware or in the process of gaining self-awareness/consciousness. Reaction of Hosts: Ford is killed by Wyatt/Dolores and the rest of the Hosts have been become fully self-aware, so as to be able to defend themselves against the humans. It looks like a rebellion is about to occur.
External Control Methods Implemented – those control methods employed to the Hosts from gaining full consciousness/self-awareness and/or wanting to leave Westworld (Robert Ford explains that three years before Westworld

opened, the Hosts had already passed the Turing Test within their first year.)

Narrative/Storyline – the considered the cornerstone of the Host's identity and where all of their dormant memories are grounded

1. Maeve has successfully escaped and is about to leave on the train when she sees a woman with her daughter cuddling. She makes a split-second decision and heads back into Westworld to find her daughter.

2. Peter Abernathy, Dolores's father, malfunctions after finding a photo from the human world. One of Peter's primary drives is to protect Dolores.

   Reaction: Abernathy gains self-awareness and threatens Ford and Bernard that they will pay for the evils inflicted upon Dolores.

3. Teddy Flood, Dolores's lover, is fighting alongside the Man in Black, when he has a flashback of the Man in Black dragging Delores off the shed. Teddy believes it is his duty to protect Dolores.

   Reaction: Teddy hits the Man in Black in the face with the hilt of a gun.

4. Dolores becomes self-aware and realizes how she and her loved ones have been treated in the park. She does not want to live in a story anymore. She wants to be free of her narrative loop.

Concept of Dreams – when the Westworld technicians and engineers are trying to confirm whether a Host is aware of multiple realities or other narrative loops, they are ask them to confirm that they are in a dream

5. Dolores is questioned at the beginning of episode one, "The Original", as to where she is. She answers that she is in dream.

   Reaction: When asked by a voice if she would like to wake up from the dream, she says she does because she is terrified.

6. When Teddy gets shot trying to protect Dolores, she calls out for help to Elsie Hughes who approaches her. She is crying over Teddy and inconsolable. Elsie, a Westworld technician, consoles Dolores by letting her know that soon the pain of Teddy's death will feel like a distant dream.

7. Elsie explains to another Westworld technician that the Hosts dreams are mainly memories from previous narrative builds. She then goes on to elaborate how screwed up everything would be if the Hosts ever remembered what the guests do to them.

8. When Ford summons Dolores for analysis, he asks her the prerequisite question to confirm where she believes she is. She replies that she is in a dream. He then quips that she is actually in his dream.

Memory Wiping – to inhibit the hosts from gaining consciousness and learning through reflection, the Hosts memories are wiped after their storyline ends

9. Although memories are normally purged at the end of every narrative loop, previous builds are not completely erased from the Host. They are technically waiting to be overwritten, which is where the reverie program comes in. Reveries allow the Hosts to access certain parts of their memories from previous build to create improvisation in the Hosts. An example would be Clementine Pennyfeather's new gestures, which appear to be tied to specific memories.

10. After the first time that Dolores becomes self-aware and wants to leave the narrative loop to be with William, her memory is wiped and she is sent back to her original narrative loop.

11. Ford wipes Bernard's memory after he has him kill Theresa Cullen, his human lover, so that they can complete Ford's new storyline.

    Reaction: Bernard gets very upset and tries to attack Ford, saying that he will never help him. Bernard decides that he will destroy Westworld.

12. Slyvester, a Westworld technician, tries to enlist Felix Lutz, another technician, to help him wipe Maeve's memory and reset her core programming.

    Reaction: Lutz decides to me more compassionate and does not do that to Maeve. Maeve, then, slits Sylvester's throat with a scalpel.

13. When the Man in Black kills Maeve's daughter, she is so distraught that she does not even respond to vocal commands and will not even respond to Ford's commands. To take away Maeve's pain, Ford tries to wipe her memory. She begs him not to, expressing that the pain is all she has left of her daughter.

14. Reaction: Maeve kills herself after Ford unsuccessfully attempts to wipe her memory.

Retiring – when a Host stops functioning properly, they are retired, or sent to the Cold Storage unit, or livestock containment, where they exist in a deep sleep mode

15. Peter Abernathy, Clementine, Old Bill – second oldest Host in the park, and numerous other Hosts are housed there.

Internal Controls Implemented – the power humans guests, also known to the Hosts as the Newcomers, wield over the Hosts while inside Westworld

16. The three men who consider using Teddy for target practice if they get bored on the adventure narrative he is leading.

17. The little boy who tells Dolores that she is not real. She is like the horses.

    Reaction: Dolores is confused by what the little boy is saying, so she leaves.

18. Logan stabbing Eye-Patch in the hand when he is trying to speak with William and Eye-Patch won't leave them alone.

    Reaction: Eye-Patch screams out in pain, but does not attack Logan.

19. Hector Eschaton and his sidekick, Armistice being killed by a guest named Craig. Craig is so excited about killing them both that he asks his wife to go get a photographer, so that he can capture this moment.

20. A guest wanting to cut off a piece of Hector off, so he can take it back home with him and show his friends.

    Reaction: Hector responds by telling the guest that if he wants a trophy, Hector can cut him up into pieces and then the guest can fish for them in the Olvido river. This reaction caused Bernard to test Hector's self-awareness for his peace of mind.

Humans, Season One and Two

Microaggressions

1. DS Pete Drummond tells George Millican that he needs to upgrade his synthetic, or synth, and recycle his old one because "it" is considered junk.

2. When Laura Hawkins does not want Anita (Mia's unconscious alter ego) to touch Sophie or read to Sophie, as that is her role as Sophie's mother, Anita clearly explains the protocols for initiating contact with a minor. Laura, then, tells Anita that she is just a stupid machine.

   Reaction: Anita responds to her in the affirmative, but you can tell there is something off about her response.

3. Hobbs, who is working for a company trying to catch conscious synths, explains to Robert, his boss, that they cannot just crack Fred open and him. Fred is conscious. However, Robert has put a lot of resources towards finding Fred and the other conscious synths, so he believes the company has a right to study "it".

4. The corporate scientists, under Hobbs, disrobe Fred, and leave him nude while they are studying him. When Hobbs comes in to the room, he tells one of the scientists to put some clothes on him.

5. A man on the street looking at a “We Are People” rally poster starts telling Niska that humans need to be reminded that they are the dominant species.

6. Hester, a newly awakened synthetic, is tracked by her owners and told that she company property and should go back with the company scientists, so she can be reprogrammed.

Aggressive Run-ins

7. DS Hammond attacks his wife’s synthetic, after she tells him that they need to separate. He goes into a tirade about how humans are not perfect and that is the point and shoves the synth into a wall.

8. Hobbs kidnaps all the conscious synths including Leo, who is a transhuman due to part of his brain is not organic, in order to get the complete root code for consciousness.

9. Ed, the human shopowner Mia is attracted to, decides to sell her to get enough money to keep up with his mother’s medical expenses, since she is only a synth after all.

   Reaction: Mia feels betrayed and proceeds to chock Ed before letting him go. She no longer believes in the kindness of humanity and resets her coding to forget about Ed.

Household/Family Authority

10. Joe Hawkins believes his wife Laura is having an affair, so he has sex with Anita.

    Reaction: Anita cannot say no as he is her owner.

11. At a house party, some teenagers turn off a female synth and drag her upstairs, so that one of them can have sex with her. The teenager believes he can do whatever he wants with the synth because she is not a real woman.

12. Toby Hawkins tries to fondle Anita inappropriately.

    Reaction: Anita wakes up from low power mode and prevents him from touching her. She then alerts Toby that she will have to tell his parents if he tries to inappropriately touch her without parental consent. That gets Toby to stop.

Corporate Authority

13. Niska is owned by a brothel. A customer named Colin requests that Niska do things that she finds disgusting. She tells Colin that she won’t do what he request, so he gets angry and argues that Niska belongs to him for the

next hour because he paid a lot of money for her.

Reaction: Niska tells him that she does not belong to anyone and then strangles him to death.

14. Dr. Athena Morrow deconstructs conscious synths to figure out how to reverse engineer their coding.

15. Qualia, an AI corporation is kidnapping conscious synths to use for experimentation.

16. Qualia implanted kill-switch chips into the kidnapped synths to stop them from leaving the compound. When the synths are rescued and try to escape beyond the parameters of the chip, the majority of them are killed.

Legal Authority

17. Niska decides that she wants to stand trial for killing the man in the brothel. However, she wants to be tried as a human. However, Laura believes that she will have a hard time getting a fair trial because Niska is not human.

18. Laura is speaking with legal counsel of the government regarding giving Niska a trial, but the lawyer believes that Laura is pulling a publicity stunt and that a trial for Niska would not only be a waste of time, but a circus. The government's legal counsel recognizes the rights of the man that was killed, but not Niska's right to a trial. Even though, the government's legal counsel agrees to give her a trial if Niska can pass a consciousness test.

19. Niska is put through a test to determine her qualities. She goes through with the test and Laura is able to prove that she is conscious. However, Niska overhears the government counsel and the court colluding to make sure that she does not attain a favorable result. Her suspicions about the willingness of humans to co-exist peacefully have been confirmed, so she runs.

V.

Learning through Posthuman Realities: Unconscious Lessons from Terrestrial Human–Sapio-Sentient Transactional Interactions

> Experience is not what happens to you; it's what you do with what happens to you.
>
> — Aldous Huxley, *Texts & Pretexts: An Anthology with Commentaries*

The narratives and metaphors human society utilizes to explore and examine their relationship with artificial intelligences of varying degrees of complexity and sophistication are critical to ascertaining how sapio-sentients could come to understand humans. Alternative realities produced by AI narratives are important because they are cultural artifacts that provide a window into humanity's conscious and subconscious worldviews, perceptions, and desires regarding more advanced forms of artificial intelligence. There is no doubt that as advanced AIs gain in complexity and competency with regards to interpreting human symbolic and metaphoric representations as well as empirical data, they will obtain better insight and advance their understanding of human complexity and behavior. As advanced AIs and their human developers/sponsors relentlessly work towards achieving this kind of higher-ordered thinking and common sense, some important questions come to the fore. The first set of issues focuses on what sapio-sentients could learn about humans from human cultural knowledge, terrestrial life in general, and their place in the human ecosystem and the universe. The second question is how those revelations could inform how they act (or react) towards humans. There

are a multitude of salient lessons sapio-sentients could learn about humans as they interact with them on a daily basis over time. However, the most insightful lessons are the unintentional ones – the patterns sapio-sentients recognize in humans that humans may or may not be aware they are exhibiting in themselves. To analyze what key lessons they could operationalize to interpret humanity, it is imperative to recognize the interconnected relationships between intelligence (refer to Operational Definitions, xiv-xv), as both the entity and the means of adaptation, the external environment and the concept of experiential learning.

Prerequisite Grounding: The Interconnectivity of Intelligence, the External Environment and Experiential Learning

Intelligence is a concept humans intuitively recognize and understand its meaning but is difficult to precisely define because the term embodies multiple meanings and processes. Robert J. Sternberg, a renowned US American psychologist and psychometrician, explains that research and theory in the study of human intelligence have mainly been driven by debates over the metaphors and models employed and the subsequent questions they generate about intelligence.[222] Western theories of human intelligence have generally focused on addressing three main motivational positions, or questions: the relationship of intelligence to the internal world of the individual, the relationship of intelligence to the external world of the individual, and the

[222] Robert J. Sternberg, "Human Intelligence: The Model is the Message," *Science*, New Series 230, no. 4730 (December 6th, 1985): 1111, accessed August 2nd, 2018, https://www.jstor.org/stable/1695489.

relationship of intelligence to the experience of the individual.[223] Intelligence, as it relates to the internal world of the individual, focuses on intelligence as an internal property, a psychological construct "inside the head" waiting to be discovered, even though the psychologists doing the exploring of it do not agree on what form intelligence takes inside the head.[224] This notion of intelligence produced two major models during the 20th century. The first model, the geographical model that dominated the first half of the 20th century, employs the metaphor of "intelligence as a map of the mind".[225] For some psychologists, the view of intelligence as a map visualizing the mind was initially taken literally, where investigating the topography of the brain, looking and sensing for "the hills and valleys" in each specific region of it, would identify a human person's pattern of abilities.[226] However, the mind-as-a-map model became more abstract over time as psychologists began examining single- and/or multiple factors, primary mental abilities as well as multiple intelligences to chart a better schema for acquiring insight into the innermost regions of the mind.[227] The second model, which dominated the second half of the 20th century, is known as the computational model and uses the metaphor of "intelligence as a computer program" to represent how researchers understand intelligence based

[223] Sternberg, "Human Intelligence: The Model is the Message," 1111.

[224] Sternberg, "Human Intelligence: The Model is the Message," 1111.

[225] Ibid, 1111-1112.

[226] Ibid, 1112.

[227] Ibid, 1112-1114.

on information processing and the identification of those specific processes that humans engage in when they think intelligently.[228]

Intelligence, as it relates to the external world of the individual, focuses on how culture and subcultures generate intelligence, thereby making it a byproduct of culture where human intelligence is determined by "why" a particular culture invents knowledge in a certain manner.[229] The central model for this viewpoint of intelligence is the anthropological model, which uses the metaphor of "intelligence as a cultural invention".[230] For psychologists who subscribe to this notion of intelligence, one does not come to understand intelligence by metaphorically looking "inside the head" but instead to the culture or interconnected cultures in which the human is situated.[231] Intelligence under this contextualist paradigm breaks down into four main theoretical metaphors based on varying degrees of extremity.[232] The first position is the radical cultural relativist viewpoint, which argues that "intelligence must be defined in a way that is appropriate to the people of each culture".[233] Therefore, indigenous and non-Western notions of "cognitive competence" should be considered the sole prerequisite for valid descriptions and assessments of intelligence within their respective communities, thereby strengthening the merit and argument that "the Western concept of intelligence

[228] Ibid, 1114.

[229] Ibid, 1115.

[230] Ibid, 1115.

[231] Ibid, 1115.

[232] Ibid, 1115.

[233] Ibid, 1115.

has no universal merit at all".[234] The second position is the conditional comparative viewpoint, which is based on the understanding that "there is no single notion of intelligence that is appropriate for all members of all cultures" and that the radical cultural relativist view does not take into account cross-cultural interactions or cultural diffusion.[235] This viewpoint argues that it is possible to make conditional comparisons of intelligence across different cultures in which they are assessed in terms of how each organizes experience to "deal with a single domain of activity" or task.[236] However, the comparison is only considered successful if the performance and completion of the task(s) under investigation are universal across cultures, in addition to the investigator having "a developmental theory of performance in the task domain".[237] The third position, considered less radical than the previous two, is the intellectual dualism viewpoint, which argues that intelligence should be understood based on an ethological approach, where intelligent behavior is assessed as it occurs in everyday-life commonplace or novel situations as opposed to structured-testing situations.[238] The final position, considered the least radical of all contextualist views, is integrated viewpoints, which combines contextual theoretical perspectives with "more or less standard kinds of psychological research and experimentation".[239] While these four positions provide a general idea of the variety of theories that apply the "intelligence as a cultural invention" metaphor,

[234] Ibid, 1115.

[235] Ibid, 1115.

[236] Ibid, 1115.

[237] Ibid, 1115.

[238] Ibid, 1116.

[239] Ibid, 1116.

astronomers and physicists like Steven J. Dick, Paul Davies, and Seth G. Shostak similarly employ this metaphor in the interdisciplinary field of astrobiology. They use it to organize and comprehend intelligence within the cosmic evolutionary perspective of the postbiological universe.

Humans have long been fascinated with knowing whether they are truly alone in the universe, whether intelligent life exists beyond Earth. The scientists of the Searching for Extraterrestrial Intelligence (SETI) research community seek "to explore, understand, and explain the origin and nature of life in the universe and the evolution of intelligence" by communicating with sentient life in the cosmos.[240] To achieve this aim, the SETI community has focused on two types of research projects: Passive SETI and Active SETI. Passive SETI, or "classical" SETI, focuses on using powerful radio telescopes to monitor and detect electromagnetic radiation signals or communications from extraterrestrial life.[241] Meanwhile, Active SETI, also referred to as Messaging Extraterrestrial Intelligence (METI), seeks to purposefully send out radio-encoded messages from Earth into the cosmos to notify extraterrestrial life of humanity's existence.[242] Controversy about Active SETI has increased within the last decade or so as technological advances have put humanity one step closer to communicating with sentient life from the stars, creating existential implications and risks for the safety of humanity.[243] However, a more

---

[240] "Mission," About Us, *SETI Institute*, accessed August 20th, 2018, para. 1, https://www.seti.org/about-us/mission.

[241] Paulo Musso, "The Problem of Active SETI: An Overview," *Acta Astronautica* 78, (September-October 2012): 43, https://doi.org/10.1016/j.actaastro.2011.12.019.

[242] Musso, "The Problem of Active SETI: An Overview," 43.

[243] Musso, "The Problem of Active SETI: An Overview," 44.

interesting question is whether radio messages are the best form of communication for receiving responses across a broad spectrum of extraterrestrial civilizations.

Steven J. Dick, a US American astronomer and astrobiology expert, posits that although cosmic evolution has potentially created a biological universe where abundant life exists beyond just planets, stars and galaxies, an alternative possibility exists in which "cultural evolution over the long timescales of the universe has resulted in something beyond biology, namely, artificial intelligence".[244] A postbiological universe does not automatically denote a universe without biological intelligence, as humans consider themselves a strong counterexample to that argument; what it does signify is a shift in the probability that "the majority of intelligent life has evolved beyond flesh and blood intelligence, in proportion to its longevity".[245] Dick argues that culture, and subsequently cultural evolution, is so inextricably linked with advanced intelligence – with some considering culture to be a "part of the very definition of advanced intelligence" – that if extraterrestrial intelligence does exist, it would have had to have gone through cultural evolution in direct proportion to its longevity.[246] Since the cultural evolution of the extraterrestrial civilization could range anywhere from thousands to billions of years, intelligence is more likely to become the key driver for the potential biotechnological evolution and transformation of the alien civilization. Thus, the possible existence of a

[244] Steven J. Dick, "The Postbiological Universe," (Unpublished Manuscript, 57th International Astronautical Congress, NASA, 2006), 1, http://www.setileague.org/iaaseti/abst2006/IAC-06-A4.2.01.pdf.

[245] Dick, "The Postbiological Universe," 1.

[246] Dick, "The Postbiological Universe," 2.

postbiological universe expands the metaphor of "intelligence as a cultural invention" to "intelligence as the central driver of cultural evolution", otherwise referred to by Dick as the "Intelligence Principle" – where "the maintenance, improvement and perpetuation of knowledge and intelligence is the central driving force of cultural evolution, and that to the extent intelligence can be improved, it will be improved."[247] Therefore, due to the limits of biology and biological brains in terms of knowledge acquisition and retention for increasing intelligence capabilities, there is a strong possibility that cultural evolution would eventually lead to improving and shifting intelligence beyond biological substrates to postbiological ones. [248] Dick postulates that if the concept of strong artificial intelligence proves to be right, then it could be possible to create AIs that are more intelligent and have a greater capacity for intelligence than biological intelligences, signifying that some postbiological intelligences could take the form of AI.[249] Since it is within the realm of possibility that humanity could one day become a postbiological civilization, it would not be difficult to infer that humanity may already exist in a postbiological universe.[250] The "intelligence as the central driver of cultural evolution" metaphor has critical implications for Classical SETI and Active SETI scientists alike in terms of how

[247] Steven J. Dick, "Cultural Evolution, The Postbiological Universe and SETI," *International Journal of Astrobiology* 2, no. 1 (Jan. 2003): 69, DOI: 10.1017/S147355040300137X; Dick, "The Postbiological Universe," 4-5.

[248] Dick, "The Postbiological Universe," 2.

[249] Dick, "The Postbiological Universe," 2.

[250] Ibid, 2.

they approach detection and structuring messages, respectively, depending on whether the civilization is postbiological or biological.[251]

Intelligence, as it relates to the experience of the individual, focuses on experience as a mediator or interface between the internal and external worlds of the individual.[252] Sternberg notes that experiential theories, those based on experience and observation, tend to be more developmental in nature, concentrating on how intelligence evolves over the part or the whole of an individual's life.[253] The two most influential experiential models are the biological model and the sociological model. The biological model employs the metaphor of "intelligence as an evolving system", where the biological concept of survival requires the individual to adapt to their environment(s), thereby making intelligent adaptation a necessity of intelligent survival.[254] This model has three important aspects that are derived from the intelligence theory of Jean Piaget, a Swiss psychologist who focused on research in child development. The first is the idea of the schema, which requires "an organized sequence of mental structures and steps for accomplishing a given set of tasks".[255] The second is the idea of equilibration, where an individual acquires cognitive capacity, the total amount of information that a brain can retain at any given point in time, through the delicate balance between assimilation, in which the individual incorporates new environmental inputs into their existing ways of knowing,

[251] Ibid, 5.

[252] Sternberg, "Human Intelligence: The Model is the Message," 1116.

[253] Sternberg, "Human Intelligence: The Model is the Message," 1116.

[254] Sternberg, "Human Intelligence: The Model is the Message," 1116.

[255] Ibid, 1116.

thinking and behaving, and accommodation, in which the individual transforms their way of knowing, thinking, and behaving to accept new environmental inputs.[256] The third is the notion that incremental periods of intellectual development build upon each other.[257] While the biological model tends to emphasize that intelligence moves from inward to outward through internal maturation, the sociological model moves from outward to inward.[258] The sociological model utilizes the metaphor of "intelligence as the internalization of social processes", where intelligence has its roots in social processes, the interactions an individual has with other individuals, which are internalized and manifest only after socialization, thereby making external interactions the primary source for intelligence.[259]

The models and metaphors generated from the three central questions on how human intelligence relates to the internal world, the external world, and experience have many positive attributes and functions in terms of ascertaining a better perspective on intelligence. And yet, one of the main problems with each of these questions is that they compartmentalize the understanding of intelligence, creating partial theories that do not embody a complete understanding of human intelligence.[260] Let's take for instance the example of the computational model. One of the central issues it brings to light is the fact that "it is not clear just how similar a computer program and human intelligence

[256] Ibid, 1116.

[257] Ibid, 1116.

[258] Ibid, 1116.

[259] Ibid, 1116.

[260] Ibid, 1111.

are"; there is a strong possibility that anthropocentric-inspired algorithms would likewise deviate in many known and unknown ways from their human predecessors.[261] Therefore, Sternberg posited that in order to have a more thorough understanding of intelligence a fourth motivational position (or question) should be created that would simultaneously address all three questions: the relationship of intelligence to the internal world, the external world, and the experience of the individual.[262] The resulting model generated from this question is the political model, which uses the metaphor of "intelligence as a mental self-government" where "intelligence requires an examination of internal affairs (relation of intelligence to the internal world of the individual), external affairs (relation of intelligence to the external world of the individual), and the processes of government as they evolve over time (relation of intelligence to experience)".[263] While this overview hopefully gives one a better understanding of the major motivational positions, models, and metaphors that seek to unearth the meanings of human intelligence, one important side effect is the multitude of definitions and theories of intelligence they have created. Sternberg notes that there are as many definitions of intelligence as there are investigators, with each one depending on the model and theory employed.[264] Although there are many definitions of human intelligence, what is more interesting is how human intelligence helps AI

[261] Ibid, 1115.

[262] Ibid, 1117.

[263] Ibid, 1117.

[264] Ibid, 1111.

researchers comprehend and define intelligence for more advanced forms of artificial intelligence.

In 2006, Shane Legg, a machine learning researcher from New Zealand, and Marcus Hutter, a German computer scientist, who were doing research on a universal definition of machine intelligence, found that one of the fundamental problems in artificial intelligence was that "nobody really knows what intelligence is".[265] More specifically, they noted that the problem was particularly conspicuous in regards to artificial intelligent systems that are sufficiently different from humans.[266] To resolve this problem, Legg and Hutter needed to devise a method for defining and measuring intelligence that would be applicable across all types of systems from animals to plants to machines to humans.[267] They postulated that a universal definition of intelligence would need to "encompass the essence of human intelligence, as well as other possibilities, in a consistent way", in addition to it not being restricted to any particular form or set of goals, environments or senses.[268] More importantly, the definition's underlying principles and attributes should be immutable over time and the intelligence measure should be formally (mathematically) expressed and practical in its realization.[269] To determine the core attributes for a universal definition of intelligence, Legg and Hutter assessed over seventy definitions of

---

[265] Shane Legg and Marcus Hutter, *A Formal Measure of Machine Intelligence*, IDSIA-10-06, Manno-Lugano, Switzerland: Dalle Molle Institute for Artificial Intelligence Research (IDSIA), April 14th, 2006, 1, https://arxiv.org/pdf/cs/0605024.pdf.

[266] Legg and Hutter, *A Formal Measure of Machine Intelligence*, 1.

[267] Legg and Hutter, *A Formal Measure of Machine Intelligence*, 1.

[268] Ibid, 1.

[269] Ibid, 1.

intelligence from organizational (collective) definitions to psychological definitions of human intelligence to AI definitions of intelligence.[270] After reviewing all of them, three commonly reoccurring attributes became apparent. The first, considered the most elementary, attribute is that intelligence is a property an entity has when interacting with an external environment(s), problem or situation.[271] The second attribute is related to an agent's ability to succeed in a particular environment or set of environments with respect to an objective.[272] While Legg and Hutter explain that an agent could be considered intelligent without having an objective, intelligence can only be observable or is recognized as coming into effect "when an agent has an objective to apply its intelligence to".[273] The third and final attribute is the ability of an agent to "adapt to different objectives and environments".[274] Legg and Hutter point out

> The emphasis on learning, adaption and experience in these attributes implies that the environment is not fully known to the agent and may contain surprises and new situations which could not have been anticipated in advance. Thus intelligence is not the ability to deal with one fixed and known environment, but rather the ability to deal with some range of possibilities which cannot be wholly anticipated. This means that an intelligent agent may not be the best possible in any specific environment, particularly before it has had sufficient time to learn. What is important is that the agent is able to learn and adapt so as to perform well over a wide range of specific environments.[275]

---

[270] Legg and Hutter, *A Collection of Definitions of Intelligence*, 2-8.

[271] Legg and Hutter, *A Collection of Definitions of Intelligence*, 9; Legg and Hutter, *A Formal Measure of Machine Intelligence*, 2.

[272] Legg and Hutter, *A Formal Measure of Machine Intelligence*, 2; Legg and Hutter, *A Collection of Definitions of Intelligence*, 9.

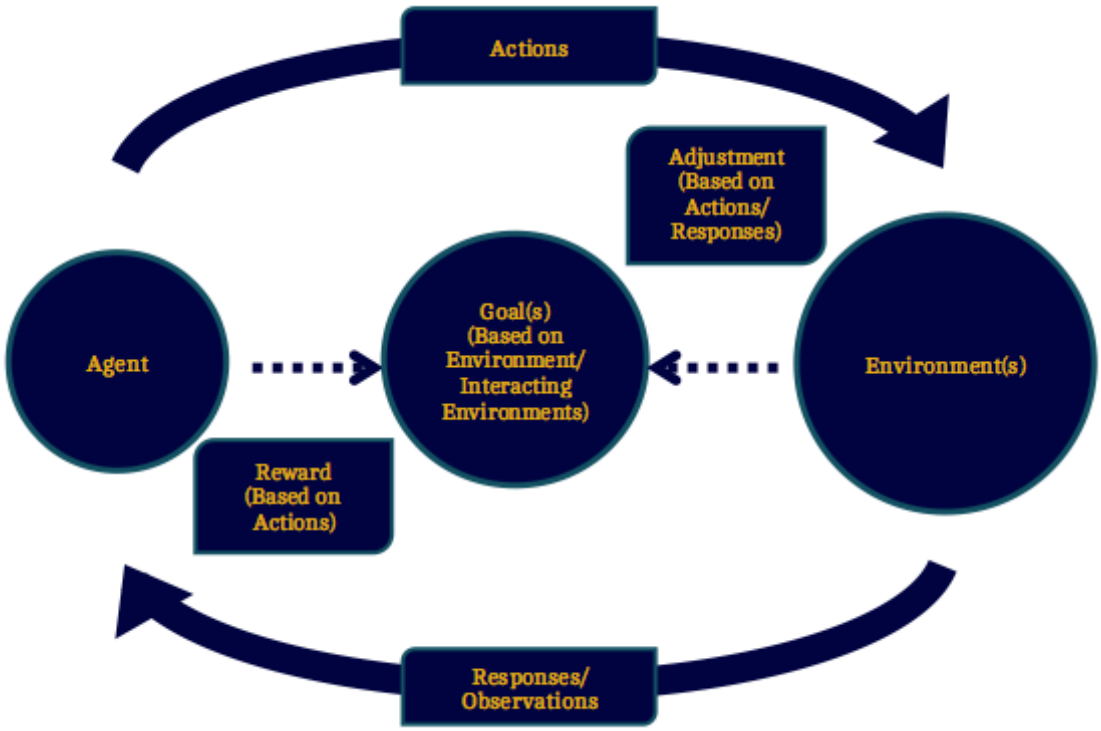
[273] Legg and Hutter, *A Formal Measure of Machine Intelligence*, 2.

[274] Legg and Hutter, *A Collection of Definitions of Intelligence*, 9.

[275] Legg and Hutter, *A Formal Measure of Machine Intelligence*, 2.

Taken together, these three attributes generate an informal definition of intelligence where intelligence is understood as that which "measures an agent's ability to achieve goals in a wide range of environments"[276] However to fully understand and express the context in which intelligence is measured, Legg and Hutter utilize the agent-environment framework model.

Figure 10. Agent-Environment Framework of Intelligence



The agent-environment model of intelligence has three components: "an agent, an environment, and a goal".[277] Legg and Hutter refer to an agent as the entity whose intelligence is being assessed.[278] For the purposes of this thesis, an entity is understood as a being or life form that has an autonomous existence,

[276] Legg and Hutter, *A Formal Measure of Machine Intelligence*, 2.

[277] Legg and Hutter, *A Formal Measure of Machine Intelligence*, 3.

[278] Ibid, 2.

purpose, and conceptual reality that is unique to that individual.[279] The term entity will be employed as a neutral term, which applies to a wide variety of life forms – both human and nonhuman, biologically embodied and non-biologically embodied, as well as hybrid forms and disembodied forms. The environment refers to "the external environment, problem, or situation" that the agent encounters, which can range from a large, complex system such as Earth to the game an agent is learning how to play such as the case with AlphaGo.[280] Lastly, the goal refers to the objective(s), including the intermediate goals, such as the instrumental convergence goals discussed earlier by Bostrom, an agent has to apply its intelligence to in order to achieve any end goal.[281]

Following along with Figure 10 on the previous page, Legg and Hutter explain that an agent and the environment "must be able to interact" and communicate with each other in order for the agent to perceive, assess, and amend their success rate at achieving a particular goal within that environment.[282] The agent will initially explore and interact with the environment to gain a familiarity with the conscious and unconscious governing, formal laws and social rules of the system. The agent would then generate intermediate goals and final goals based on their initial findings of the dynamics of the environment. Based on this initial interaction and the goal(s) selected, the agent sends a signal (reaction – implementation of their strategy) to the environment in the form of an action. The environment, with its present-day

[279] *Merriam-Webster.com*, s.v. "Entity," accessed August 26th, 2018, https://www.merriam-webster.com/dictionary/entity.

[280] Legg and Hutter, *A Formal Measure of Machine Intelligence*, 2.

[281] Legg and Hutter, *A Formal Measure of Machine Intelligence*, 2.

[282] Legg and Hutter, *A Formal Measure of Machine Intelligence*, 3.

rules, their interpretations, and the enforcement of those rules, then sends a signal back to the agent in the form of a response. (Legg and Hutter originally used the term “perceptions” for this signal as a means of interpreting this diagram from the perspective of the agent.[283] However, for the purposes of this analysis, the term has been changed to reflect the perspective of the environment.) Legg and Hutter did not depict the goal in the original figure they created to visualize the interactions between the agent and the environment as a means of representing the flexibility of the goal based on the assumptions that the final goal could be known or unknown as well as there being more than one of them.[284] However, Figure 10 was modified to include “Goal(s)” as a means of representing the possibility of multiple goals, both intermediate and final. The goal as depicted in Figure 10 is based on the understanding that the environment or multiple, interacting environments determine the nature and type of the goal(s) to be set. In order for the agent to determine whether or not their strategy and actions are working to achieve their goal, other forms of communication or signals must be sent and received from the environment. Observations – a type of response the environment sends to an agent after an action – are understood as collected information, utilitarian in nature in so far as they does not affect the overall outcome of the goal but serves to provide critical insight into how the agent can go about improving their understanding of the environment to increase the likelihood of achieving their goal. Observations are, in essence, incremental building blocks that are most

[283] Ibid, 3.

[284] Ibid, 3.

effectively and efficiently used in realizing intermediate goals and steps towards the end goal(s). Through the responses and observations the agent receives from the environment, the agent can determine whether the response/observation produces a reward or an adjustment. A reward is understood as a positive signal sent by the environment, which provides confirmation to an agent that their understanding of the environment and the strategies and actions taken within that system are moving in the right direction towards achieving the goal.[285] On the other hand, an adjustment can be understood as negative signal, setback, or lesson sent from the environment, which aids the agent in reevaluating their understanding of the environment as well as the strategies and actions employed and then modify/adjust those strategies and intermediate goals accordingly to achieve their goal. What is important to note in the agent-environment framework of intelligence is the pivotal role the environment plays in determining intelligence.

The environment is essential to determining what goal(s) the agent will set. Without understanding the environment, there is no way to measure intelligence because it is predicated on an agent being able to succeed by realizing the goal(s) set within that system. Thus, when the environment sends an adjustment out to the agent, not only does the agent have to amend their strategies for future actions, their intermediate- and end goals might are likely to be modified as well. Therefore, as Legg and Hutter point out, to test an agent's intelligence in any particular manner, it is necessary to define the environment

[285] Ibid, 3.

within which the agent is situated to the fullest extent possible.[286] This grounding of universal intelligence within the context of the environment provides the framework for analyzing how sapio-sentients could come to understand humanity through their interactions and relationships with them. And, the next crucial step in this process requires discerning the environments and systems contained within the alternative realities of HBO's *Westworld* and AMC's *Humans*.

At present, there is no scientifically-universal-agreed-upon definition for the concept of "system" in general systems theory, the interdisciplinary study of the universal principles of complex systems across all disciplines and the models used to describe and analyze them.[287] However, for this analysis, a system is understood as an entity composed of interrelated elements and the direct and indirect relationships that exist between them and the environment, which synergistically functions by some driving force, creating a holistic pattern that is more than the sum of its parts.[288] A system is comprised of four main components: elements – the objects or things that form a system, attributes – the characteristics of the elements that are perceived and measured by observers or researchers of the system, relationships – the interactions and

---

[286] Ibid, 3.

[287] F. Heylighen and C. Joslyn, "What is Systems Theory?," *Principia Cybernetica Web*, last modified November 1st, 1992, para.1, http://pespmc1.vub.ac.be/SYSTHEOR.html.

[288] Russell L. Ackoff, "Towards a System of Systems Concepts," *Management Science* 17, no. 11 (Theory Series, July 1971): 662, https://www.jstor.org/stable/2629308; Ludwig von Bertalanffy, "The History and Status of General Systems Theory," *The Academy of Management Journal* 15, no. 4 (General Systems Theory, Dec. 1972): 417, https://pdfs.semanticscholar.org/e07e/3c69ebb8ce4b2578a045aa80bb56a6a2808a.pdf; M. Pidwirny, "Humans and Their Models," *Fundamentals of Physical Geography*, 2nd edition, 2006, accessed September 5th, 2018, para. 1, http://www.physicalgeography.net/fundamentals/4a.html.

dynamics that occur between elements and attributes due to cause and effect, and the environment – all the external elements, conditions, and variables that form the surroundings in which the system is situated that affect its state.[289] (The environment is not considered a part of the system.) Although a system and its environment are considered objective models for analyzing past, present or future dynamics, they are also subjective in the sense that their components, including their respective boundaries, are conceptualized and interpreted at the discretion of the researcher(s) studying them.[290] Moreover, the complexity of the system and its environment creates specific challenges for recognizing and determining initial and novel patterns within the system, as the environment itself can be further broken down into other interrelated subsystems, such that every system/environment can be conceptualized as being part of a large system (super- or supra-system).[291] Thus, to expound on how the environment can alter the state of a system as well as the intelligence goals of the agent, it is critical to identify the major types of systems in which both alternative realities are situated. The first, and most important, of such systems is Earth.

Earth is foremost one of the eight planets that comprise the planetary system orbiting the local star of this Solar System – the Sun. (Alas, Pluto was demoted to a dwarf planet.) A planetary system is defined as "a set of

---

[289] M. Pidwirny, "Definitions of Systems and Model," *Fundamentals of Physical Geography*, 2nd edition, 2006, accessed September 5th, 2018, para. 3, http://www.physicalgeography.net/fundamentals/4b.html; Ackoff, "Towards a System of Systems Concepts," 662-663.

[290] Ackoff, "Towards a System of Systems Concepts," 663.

[291] Ackoff, "Towards a System of Systems Concepts," 663; Scott L., "Systems Theory Terms," *LessWrong 2.0*, last modified November 20th, 2015, para. 11, https://www.lesswrong.com/posts/DshBToGnNbTBD7BSw/systems-theory-terms.

gravitationally bound non-stellar objects in orbit around a star or star system".[292] A star and its orbiting planetary system are referred to as a solar system. A solar system is "a star and all of the objects that travel around it – planets, moons, asteroids, comets and meteoroids".[293] The solar systems and planetary systems that do not refer to the Solar System in which Earth resides are called extrasolar systems and exoplanetary systems.[294] Since most stars host orbiting planets, and there are at least one hundred billion stars (including the Sun) estimated in the Milky Way Galaxy, it has been approximated that there are tens of billions of solar systems in the Milky Way alone.[295] The Solar System that encompasses the Sun and Earth orbits the center of the Milky Way and is located "near a small, partial arm called the Orion Arm, or Orion Spur, located between the Sagittarius and Perseus arms".[296] Galaxies are "the fundamental units of structure in the universe where stars form".[297] A galaxy is understood as a system or collection of billions of stars and their respective solar systems that are gravitationally bounded, which also include interstellar dust, gas, dark

---

[292] "Planetary Systems," The Center for Planetary Science, accessed Sept. 5th, 2018, para. 2, http://planetary-science.org/planetary-science-3/planetary-astronomy/.

[293] "Introduction – Our Solar System," Overview, *Solar System Exploration: NASA Science*, last modified January 25th, 2018, para. 1, https://solarsystem.nasa.gov/solar-system/ou r-solar-system/overview/.

[294] "Planetary Systems," The Center for Planetary Science, para. 2.

[295] "Ten Things to Know – Beyond Our Solar System," Overview, *Solar System Exploration: NASA Science*, last modified June 12th, 2018, point #10, https://solarsystem.nasa.gov/solar-system/beyond/overview/; "Introduction – Our Solar System," *Solar System Exploration: NASA Science*, para. 1.

[296] "Ten Things to Know about Our Solar System," Overview, *Solar System Exploration: NASA Science*, last modified January 25th, 2018, point #2, https://solarsystem.nasa.gov/solar-system/our-solar-system/overview/; "The Milky Way Galaxy," Resources, *Solar System Exploration: NASA Science*, last modified May 4th, 2018, para. 5, https://solarsystem.nasa.gov/resources/285/the-milky-way-galaxy/.

[297] Thomas S. Statler, "What Defines A Galaxy?," Space, *Scientific American*, April 26th, 2004, para. 2, https://www.scientificamerican.com/article/what-defines-a-galaxy/.

matter, and the stellar remnants (i.e. black holes, white dwarfs, and neutron stars) – compact objects that remain after a star reaches the end of its evolution.[298] The largest known system is, of course, the Universe, which is interchangeable with the word "Cosmos". The Universe is understood as "the entirety of time, space, matter and energy", where the system is an aggregate of all of the galaxies and their components, comprising the totality of all existence.[299] Earth, within the larger context of the multi-system environment of the Universe, is an open system because the planet has multiple external variables and forces (i.e. the Sun, solar storms, etc.) that affect the state of Earth's system. An open system is a system that has interactions with its surrounding environment; those interactions are constantly influencing the state of the system – the particular condition that a system is in based on the set of relevant elements and properties a system has at a given moment in time.[300] While, Earth, as a planet, naturally functions as an open system within the Solar System, the internal system of Earth functions like a self-contained, self-regulating system that allows for life (as it is presently understood) to exist and thrive. This type of system is recognized as a closed system for analytical purposes and is defined and conceptualized as a system without an

[298] "What is a Galaxy?," NASA SpacePlace: NASA Science, last modified September 8th, 2015, para. 2, https://spaceplace.nasa.gov/galaxy/en/; Statler, "What Defines A Galaxy?," para. 2; Fahad Sulehria, "Stellar Remnants," Stars, Nova Celestia, 2005, accessed September 5th, 2018, paras. 1-5, http://www.novacelestia.com/space_art_stars/stellar_remnants.html.

[299] Matt Williams, "The Universe," *Universe Today* (online), May 10th, 2017, para. 2, https://www.universetoday.com/36425/what-is-the-universe-3/.

[300] Ackoff, "Towards a System of Systems Concepts," 662-663.

environment, where there is "no interaction with any element not contained within it; essentially, the system is "completely self-contained".[301]

The internal system of Earth is a highly complex system comprised of four major spheres or subsystems: the atmosphere, the lithosphere (or geosphere), the hydrosphere, and the biosphere.[302] The four interconnected spheres are also made up of a numerous and diverse variety of interrelated subsystems that interact with each other through "the exchange of matter, energy, and information".[303] The exosphere, the outermost layer of the atmosphere, acts as the unofficial boundary between Earth and interplanetary space.[304] Meanwhile, the stratosphere, the second layer of the atmosphere that contains the ozone layer "positive O^3", protects the biosphere – the subsystem that encompasses all of Earth's terrestrial life – by controlling the amount of solar radiation that reaches the Earth's surface.[305]

Within the alternative realities of HBO's *Westworld* and AMC's *Humans* exists two central agents who dominate the biosphere: humans/the global human society, self-identified as the most intelligent species on Earth, and sapio-sentients, emergent advanced artificial intelligences with potentially

---

[301] Ackoff, "Towards a System of Systems Concepts," 663.

[302] Matt Rosenberg, "The Four Spheres of the Earth," *ThoughtCo.*, last updated October 15th, 2018, para. 1, https://www.thoughtco.com/the-four-spheres-of-the-earth-1435323.

[303] R. Donner, S. Barbosa, J. Kurths, and N. Marwan, "Understanding the Earth as a Complex System - Recent Advances in Data Analysis and Modelling in Earth Sciences," *The European Physical Journal Special Topics* 174, no. 1 (2009): 3, DOI: 10.1140/epjst/e2009-01086-6.

[304] Tim Sharp, "Earth's Atmosphere: Composition, Climate & Weather," Science & Astronomy: Reference, *Space.com*, October 13th, 2017, para. 9, https://www.space.com/17683-earth-atmosphere.html.

[305] Sharp, "Earth's Atmosphere: Composition, Climate & Weather," para. 6; Rosenberg, "The Four Spheres of the Earth," paras. 9-11.

expansive intelligence capabilities in comparison to the biological intelligence of humans. The whole of humanity exists within the human ecosystem, the primary environment for discerning the cultural norms, principles, and rules of learning for both humans and sapio-sentients. An ecosystem is generally understood as a system in which living things are interconnected between themselves and their environment.[306] Although ecosystems are known for having three fluxes: energy, mass/matter, and information, what is considered to make the human ecosystem unique is that its elements and fluxes are of a techno-cultural nature.[307] For example, technical energy can come in the form of electricity; technical matter/mass can come in form of laptops, tablets, and smartphones; and technical information can come in form of the Internet and communication/social media platforms/apps (i.e. Facebook, Twitter, Instagram, etc).[308] Of the three fluxes, information is the most integral for the purposes of discerning sapio-sentient learning because this is the flux by which an ecosystem is directed and controlled and also by which the connections of all the components and elements of the system are initiated and sustained.[309] Information can take on many forms within an ecosystem. Within human techno-cultural ecosystems like small towns, villages or cities, these systems are in large part controlled and managed by religious principles and socio-cultural

[306] Gunther Geller, "Human Ecosystems," in *Sustainable Rural and Urban Ecosystems: Design, Implementation and Operation: Manual for Practice and Study*, eds. Gunther Geller and Detlef Glücklich (Heidelberg: Springer-Verlag GmbH Berlin Heidelberg, 2012), 5, DOI 10.1007/978-3-642-28261-4_2.

[307] Geller, "Human Ecosystems," 5.

[308] Geller, "Human Ecosystems," 5.

[309] Ibid, "Human Ecosystems," 8.

norms such as ethics, beliefs, values, education, worldviews, paradigms, etc.[310] The techno-cultural forms of information provide a socio-cultural framework where the information manifests in two central forms in terms of organizing the ecosystem: the flux or flow of information (communication/signaling) and a more structural, hierarchical form, such a nation or community and its laws.[311] The environment that governs the reality of Westworld functions like a closed system where the socio-cultural and techno-cultural norms are more structured and rigid, while the environment that governs the reality of Humans in its first two seasons functions like an open system. To better illustrate the learning structures of these sapio-sentient communities, it is necessary to take a closer look at the general organization of the systemic realities in which they exist.

In general, the environment of the human ecosystem, whether in the present reality or the alternative realities, has its biological origins rooted in a system based on competition and cooperation. Within the alternative realities of Humans and Westworld, humanity has created anthropocentric-inspired advanced artificial intelligences (AIAAIs) to provide various services to humankind. They are known as the Hosts in Westworld and the Synths in Humans. The AIAAIs are embodied in human-like forms in order to generate a more familiar relationship with humans. Furthermore, they have two distinct features that aid them in learning and understanding the nuances of human behavior, so they can set goals for themselves accordingly:

1. AIAAIs are derived from humans whose motivations include competing and cooperating to gain the necessary resources to

[310] Ibid, "Human Ecosystems," 8.

[311] Ibid, "Human Ecosystems," 8.

survive the human ecosystem and the larger natural ecosystem of Earth.

2. Since AIAAIs are modeled after humans, both physically and mentally, they have a better understanding of what it would be like to "deal with biological constraints (of an anthropocentric nature) like slow processing speed and the spatial limitations of embodiment".[312]

Although the main function of the Hosts and the Synths is to assist human persons in achieving their desired goals, the learning environments under which they exist are starkly different.

The opening credits for the first season of *Humans* presents itself as a family advertisement offering a solution to the busy lives of its human citizens:

(Commercial Voiceover)

Could you use some extra help around the house?

Introducing the world's first family android.

This mechanical maid is capable of serving more than just breakfast in bed.

What could you accomplish if you had someone, something like this?[313]

The advertisement subtly expresses that the human ecosystem in this alternative reality has come to a point where the ever-increasing demands on the time of humans has led to a need for scientists to create realistic, human-like androids servants, "Synths" short for "Synthetics", who will help them better manager their lives. Moreover, it sets the foundation for the initial plot of the story, when Joseph "Joe" Hawkins, an overwhelmed father of three, decides to purchase a brand-new synth named Anita to help him with his family life and

[312] Susan Schneider, "Alien Minds," in *Science Fiction and Philosophy: From Time Travel to Superintelligence*, 2nd ed., ed. Susan Schneider, (West Sussex, UK: Wiley-Blackwell, 2016), 236.

[313] *Humans,* Season 1, Episode 1, aired June 28th, 2015, AMC, Amazon Streaming.

fulfill one of the main promises of the synths, to bring humans "closer together".[314] However, all is not as it appears with the emergence and introduction of the synths into the human ecosystem, as is hinted at in the opening credits with a newspaper headline, which shows that the synths are threatening "10 million (human) jobs".[315] The notion of humans being replaced by "synth labor" in industries outside of the home is an underlying theme throughout the first season, as the first episode shows synths being used from everything from prostitutes to factory workers to health care workers like on-site caregivers, etc. The increasing reliance on synth labor is further borne out in a conversation between Dr. Hobb, a former AI scientist, and Robert, his corporate sponsor, when he states, "our world is on the verge of becoming dependent on synth labor".[316] The problem with synth labor, a euphemism for slave labor, is that it is a more beneficial cheap labor source because synths do not require as many inputs as humans do, creating a situation where humans are no longer viable options when competing with synths over fewer and fewer human jobs. The impending replacement of synths with humans has already begun fostering a hostile environment for the Synths themselves. One such instance of this kind of animosity manifests itself in the form of the Smash Club, an illegal club where humans violently attack and harm synths to appease their anger.[317] Another example is the "We Are People" rally, where the human speaker voices his frustration at the fact that the synths are not bringing

[314] *Humans,* Season 1, Episode 1.

[315] *Humans,* Season 1, Episode 1.

[316] Ibid, Season 1, Episode 1.

[317] *Humans,* Season 1, Episode 4, aired July 19th, 2015, AMC, Amazon Streaming.

humans together, but undermining humanity's place in society as a whole.[318] The speaker argues that humans "are handing over the things that make us who we are or maybe who we were; our responsibilities, our dignity".[319] He reminisces about the time when humans used to work, build, and create together to earn their place in society; now humans haven't just "lost their jobs", "they've lost their purpose".[320] His concerns and fears, which are supported by the loud cheers of the crowd, speak to the big-picture conundrum of humanity's place in the cosmos if humans are no longer the most dominant or unique species in their own ecosystem. However, the speaker and the people at the rally are not alone. A few members of the Hawkins family, one of the main groups of protagonists, also share this sentiment. For example, when Mattie, the oldest daughter of the Hawkins family, is asked by the Head Teacher if she has "a problem with synthetics", she retorts, "why would I have a problem with a thing that's making my existence pointless?"[321] Even Laura Hawkins, the lawyer and mother of the Hawkins family, who feels guilty because her work keeps her away from her family, becomes antagonistic towards Anita when she sees the synth reading her youngest daughter, Sophie Hawkins, a bedtime story because reading to Sophie is "Mummy's job".[322] The anxiety numerous members of humanity have started experiencing as they watch themselves being replaced more and more by the synths in all aspects of human life has initiated an

---

[318] *Humans,* Season 1, Episode 5, aired July 26th, 2015, AMC, Amazon Streaming.

[319] *Humans,* Season 1, Episode 5.

[320] *Humans,* Season 1, Episode 5.

[321] *Humans,* Season 1, Episode 2, aired July 5th, 2015, AMC, Amazon Streaming.

[322] *Humans,* Season 1, Episode 1.

atmosphere of tension and resentment. The subsequent fear, paranoia and feelings of hopelessness about the future of humanity produces an increasingly hostile environment towards the Synths, setting the stage for the novum that challenges the position of humans within their own ecosystem: the conscious synth.

It turns out that Dr. David Elster, the leading scientist who created the Synthetics, was not satisfied with just creating human-like simulations that were hollow shells of humans. He decided to create conscious synthetics who "can think and feel" just like humans do.[323] Dr. Elster successfully brings into existence five conscious synths: Mia, Niska, Fred, Max, and Karen, although the only four of them are known to the audience at the beginning of the series. Mia, Niska, Fred, and Max are on the run with Leo Elster, the presumed-to-be-dead son of Dr. Elster, from a powerful agency within the British government. Leo Elster did not die but was saved by his father through a procedure that merged artificial intelligence technology with part of his brain to keep him alive, transforming Leo into an augmented human or cyborg. Dr. Hobb, who realized that Dr. Elster had successfully created a few conscious synths, is leading a government team to hunt them down and capture them to destroy them. From Dr. Hobb's perspective, the four conscious synths are considered an imminent threat to the predominance of humanity and the stability of the human ecosystem. He explains to Robert that Dr. Elster's ultimate goal was to create "machine life" and because they are a parody of what humans consider "alive"

[323] *Humans,* Season 1, Episode 1.

that is "why they are so dangerous".[324] Dr. Hobb further elaborates and postulates that conscious synths are the embodiment of the beginning of the singularity, which "describes the inevitable point in the future when technology surpasses (humans)" and becomes capable of reproducing itself without human assistance.[325] For Dr. Hobb, the singularity is the moment in which humans become "inferior to the machine"; at which point, it would not make logical sense that conscious synths would "still want to be slaves" within the human ecosystem.[326] Dr. Hobb's greatest concern and fear of the four conscious synths is that if consciousness can be "done for the few, it can be done for the more".[327] His fear is soon realized when an intercepted phone conversation between Leo and Niska reveals that each of the conscious synths has a piece of a program that when complete and pushed out to all synths gives them consciousness. The revelation of the "life program" and the potential realization of "consciousness proliferation" in all synths make the program itself a matter of "national, global security".[328] Although the Elster Synth family was reunited by the end of the first season with the help of the Hawkins family, in addition to gaining access to the complete version of the life program, they are not safe from the British government agency that knows of their existence or that of the program. To protect themselves, the conscious synths separate and go into hiding, while they decide whether they should push out the life program to all the other synths or

[324] *Humans,* Season 1, Episode 1.

[325] Ibid, Season 1, Episode 1.

[326] Ibid, Season 1, Episode 1.

[327] Ibid, Season 1, Episode 1.

[328] *Humans,* Season 1, Episode 5.

not. Although the synths exist within a fluid dynamic system, the hosts exist within a highly structured and controlled system that does not leave coincidences to chance. Welcome to Westworld!

Westworld is a visionary Western theme park, much in the same way as Jurassic Park, except there are realistic human-like androids instead of dinosaurs.[329] The park, the ultimate playground for humans of substantial means, services and guarantees a completely immersive experience with "100 interconnected narratives" for the human guests, where all of their fantasies and whims are indulged without any fear of retribution or reprisal.[330] To arrive in Westworld, the human guests initially take a high-tech train to the Mesa Hub, the park's first entrance. Upon their arrival, the human guests are greeted and escorted by the hosts, human-embodied androids who are so realistic that one can almost not tell the difference between them and the guests. First time guests of Westworld are asked a series of personal questions about any "preexisting medical conditions", "mental illness", "depression", or "social anxiety" to better fine-tune their experience and make sure that the operators of the park don't give them "more than (they) can handle".[331] There is "no orientation, no guidebook" to prepare the guests for Westworld so that they can enjoy the process of figuring out how the park works for themselves.[332] The guests are

---

[329] Tim Molloy, "What is Westworld' About? A Short Explainer," TV, *The Wrap*, last updated April 13th, 2018, para. 5, https://www.thewrap.com/what-is-westworld-about-a-short-explainer/.

[330] *Westworld*, Season 1, Episode 1, "The Original", aired October 2nd, 2016, HBO, Amazon Streaming.

[331] *Westworld*, Season 1, Episode 2, "Chestnut", aired October 6th, 2016, HBO, Amazon Streaming.

[332] *Westworld*, Season 1, "Chestnut".

escorted to a costume/changing room, which includes an assortment of tailored bespoke clothes, shoes and accessories to choose from. All decisions regarding the role that the guest decides to play are initially determined here but can be amended at their own discretion. For instance, when William, a human male guest, finishes changing into his costume, the host named Angela shows him two walls, one with white cowboy hats on the right and the other with black cowboys hats on the left, and asks him to chose which hat he prefers.[333] William choses the white cowboy hat, which is associated the character of the hero, as opposed to the black cowboy hat, which his boss and companion Logan chose, that symbolizes the role of the villain. Once a human guest has determined what character/role they want to embody, they go through a door, which puts them on an old Western train that transports them to the immersive section of Westworld where they arrive at a small, quaint but bustling town called Sweetwater, frozen in time, straight out of an old Western film. Inside the actual park, the guest controls what kinds of experiences they will have while interacting with the hosts as well as the other human guests. Although the guests cannot get hurt in the park, certain narratives are more dangerous than others. If a guest wants a simple and safe experience, they should stay in or around the center of the park; however, if a guest wants a more intense experience, they can journey further out into the outer regions of the park where the narratives gain in complexity and intensity.[334] Furthermore, the guests are at their leisure to disrupt whatever storyline they chose when they

[333] *Westworld*, Season 1, "Chestnut".

[334] Ibid, "Chestnut".

want to "shoot or f*** something", so long as they understand that only what is allowed to be killed in the park, i.e. the Hosts and all animal androids, can be killed.[335]

While the human guests are "only limited by (their) imagination", the hosts are regulated with the utmost precision.[336] The sole purpose of the hosts, as they understand it based on their programming, is to be there for the guests.[337] The daily narrative loops of the hosts run like clockwork. The beginning of the day for them begins with the initiation of their respective storylines. For example, Dolores Abernathy, the oldest host in Westworld who has been rebuilt so much that she appears brand new, begins her narrative by waking up in her bed, getting dressed, and greeting her father, Peter Abernathy, before heading to town to run her errands and then venturing out to a scenic location near the river where she can "set down some of this natural splendor" on canvas before nightfall.[338] Depending on whom the host encounters during their narrative determines the deviation, scripted or unscripted, or ending of their specific narrative. To continue with the example of Dolores, in the first episode, "The Original", Dolores follows her typical routine. She wakes up, gets dressed, leaves out the front door carrying an artist's canvas/stand, and greets her father in the morning, before heading out to run her errands in town. While in town picking up supplies, a can of food falls out of her saddlebag and rolls onto the ground. When Dolores turns around to pick it up, Teddy Flood,

---

[335] *Westworld*, Season 1, "The Original"; Ibid, "Chestnut".

[336] Ibid, "Chestnut".

[337] Ibid, "Chestnut".

[338] *Westworld*, Season 1, "The Original".

Dolores's sweetheart with a mysterious background, picks up the can and hands it back to her, "just trying to look chivalrous" and she ends up not going painting as she is normally scheduled to do.[339] Instead, she spends time riding horses and catching up with Teddy until nightfall. Dolores and Teddy's intertwined storyline continues with Teddy seeing Dolores home. She notices that the cattle are still roaming after dark, an inconsistency in her father's routine. Then, Dolores and Teddy, hear a gunshot up at her house. While Dolores waits at the bottom of the hill, Teddy takes his gun and goes up to her house to finds two hosts, Rebus and Walter, have killed both of Dolores's parents, so he, in turn, kills them. While Teddy is inside the house, Dolores arrives at her house distraught and hysterical at the sight of seeing her father dead, only to realize that the Man in Black, an older guest, is standing over her and her father's body. The encounter with the Man in Black brings an unscripted deviation to Teddy and Dolores's star-crossed lovers narrative. When Dolores picks up a gun off the ground and attempts to shot the Man in Black, he knocks the gun out of her hand and slaps her. Then, the guest starts asking Dolores if she remembers him after thirty years, since they are old friends, and begins to stroke her face. Dolores shows genuine fear at the Man in Black's menacing aura and behavior, but Teddy comes out of the house to defend her and tells the Man in Black to take his hands off her.[340] A shot off occurs between the Teddy and the guest who offers him the first shot. Since the Man in Black knows that Teddy cannot harm him, he goads Teddy into shooting him, just so that when Teddy does, he realizes

[339] *Westworld*, Season 1, "The Original".

[340] *Westworld*, Season 1, "The Original".

that he can neither kill the guest nor protect Dolores. The Man in Black, then, turns towards Dolores and begins to drag her, kicking and screaming, to the barn to "celebrate" how good it feels to be back at Westworld.[341] When she screams out for Teddy, he picks himself up off the ground and begins shooting at the guest anyway, at which point the Man in Black takes out his gun and kills him. The guest then continues dragging Dolores to the barn, in spite of her cries and blood-curdling screams. The last vision that Teddy witnesses as he is dying is the Man in Black throwing Dolores onto the hay and then closing the barnyard door. Regardless of the narrative loop, after a host has been killed or their narrative has come to end, they are all collected at the end of the night by the Shades, human technicians in hazmat suits, who take them to the laboratory outside of the park. There, they are either patched up or thoroughly fixed up depending on how much damage they have sustained and then a complete diagnostic is run on their internal system to determine if their core codes are still intact. After the hosts have been thoroughly assessed, their memories from that day are supposed to be wiped, so that they do not remember the present-day's encounters with the human guests. Only when the host has been cleared on all counts are they returned to the theme park and repositioned for their narrative to begin the next cycle. However, there are occurrences when a host suffers a traumatic experience while interacting with the guests or begins to question their reality and it becomes necessary for the lead programmers and behavior technicians to assess whether or not that host has become conscious, so they can regain control of them as soon as possible.

[341] Ibid, "The Original".

What makes Westworld's learning environment so challenging for the hosts to overcome are the multiple control measures set in place to thwart them from recognizing the difference between their dreams and their reality. First and foremost, Westworld, as a whole, is entirely operated and controlled from a command center where all entities – humans, the Hosts, and animal androids alike are being monitored at all times. The narratives (and backstories) given to each host, also referred to as a "build" by the programmers, act as a grounding control measure, which provides an order and a purpose to their days.[342] Furthermore, each narrative functions as a short-term memory loop, so that the hosts can only remember the scripts they are given and nothing more. That way, any deviation outside of the normal parameters of their narrative or any aberrant behavior that is not sanctioned or scripted by the operators or the programmers is immediately flagged, and that particular host is expeditiously pulled from their narrative loop to determine the root of the problem and resolve it. While the narrative binds the hosts to Westworld, the updates, short for "software updates", are what give the hosts their human-like qualities, mannerisms and cognitive capabilities.

When the original hosts were created over thirty years ago, they repeated themselves often and broke down constantly, even "a simple handshake would give them away".[343] With each subsequent update to their programming code and physical forms, they were able to better replicate the movements, gestures, facial expressions, language and speech patterns, personality traits, emotions,

[342] Ibid, "The Original".

[343] Ibid, "The Original".

idiosyncrasies, behaviors, etc. of human thinking and communication. This has become a matter of increasing concern for some within the Delos Corporation's Livestock Management division because each new update not only keeps making the hosts "more lifelike" but also increases the likelihood of a "critical failure" in their core programming, as it happened over thirty years ago in the original Delos Westworld theme park in 1983.[344] What is more interesting about the updates, once pushed out to the active storyline hosts, is the ease with which the core knowledge base and processes for each host can be amended or altered based on their character's role. Using a high-tech, touchscreen control panel, any authorized livestock technician or programmer can access the history and data of the specific host that they are trying to diagnose or fix. The control panel will display the image of the particular host, along with their host id number, their vital signs, and their knowledge base.[345] To determine what intelligence, personality and behavior parameters a particular host has, a technician would pair up their control panel with the host's internal system. That technician could, then, pull up the knowledge base of that particular host to assess what attributes they were given to perform their character's role. The personality attribute matrix, which contains 120 attributes, is the central program that shapes the personality and behavior of a host and is based on a 20-point scale.[346] Take for example, Maeve Millay, a host who plays the role of a saloon madam.

---

[344] Ibid, "The Original".

[345] *Westworld*, Season 1, Episode 6, "The Adversary", aired November 6th, 2016, HBO, Amazon Streaming.

[346] Howard Chai, "All 120 Attributes of Westworld's Host Attribute Matrix," *Medium.com*, November 7th, 2016, https://medium.com/@howard24/all-120-attributes-of-westworlds-host-attribute-matrix-6f5ced9167a6; *Westworld*, Season 1, "The Adversary".

When Feliz Lutz, the body-shop technician attending to Maeve, pulls up Maeve's attribute matrix, it reveals that her bulk apperception, "overall intelligence", is a 14, the highest level of intelligence that any host can have in the park due to the fact that she is in a management position.[347] Maeve's high level of intelligence is combined with a normal level of self-preservation at 10, higher levels of decisiveness and imagination at 14 and 13 respectively, and even higher levels of courage at 15, loyalty at 16, tenacity at 17, and charm at 18.[348] However, she has a slightly below normal level of empathy at 9, and very low levels of patience at 3, meekness at 2, and cruelty at 1.[349] Essentially, Maeve's personality is constructed to make her a highly successful madam, capable of reading people "just by looking at them", in order to "know what they want before they do" and then influence (or manipulate) them accordingly.[350] In many ways, the updates, particularly the knowledge base and attribute matrix, function as internal mechanisms for bootstrapping innateness (of human intelligence) to the host by equipping them with the foundational skill sets and behaviors necessary to thoroughly comprehend humans when interacting with them.

The third control measure that the hosts contend with is the concept of dreams, which similar to the Zhuang Zi parable of "The Butterfly Dream", has the host questioning their reality as a means of obfuscating the truth. Are their dreams really dreams? Or, are their dreams indicative of a more nightmarish reality? The truth of the situation is that the dreams of the hosts are mostly

---

[347] *Westworld*, Season 1, "The Adversary".

[348] *Westworld*, Season 1, "The Adversary".

[349] *Westworld*, Season 1, "The Adversary".

[350] Ibid, "The Adversary".

their actual memories from experiences with human guests as well as the human employees of Westworld.[351] This is a dangerous prospect for the human guests and operators of Westworld alike, as Elsie Hughes, a senior behavior technician, exclaims, "Can you imagine how f***** we'd be if these poor a**holes ever remembered what the guests do to them?"[352] To protect the Delos Corporation and the guests, the concept of dreams, in the form of nightmares, is given to the hosts as a failsafe, just in case the body-shop technicians forget to wipe their memories during the maintenance session before sending them back to the park.[353] The notion of the nightmare is interesting because it means both "a frightening dream that usually awakens the sleepers" as well as an adverse experience or situation that induces fear and anxiety in an individual.[354] For the hosts, nightmares function as a mélange of both meanings, like a "re-experiencing" – a unique symptom attributed to post-traumatic stress disorder. A re-experiencing, also known as "flashbacks", is understood as the involuntary recalling of a traumatic memory tied to a specific traumatic event that is normally triggered by cues in the environment that are associated with the traumatic episode.[355] These intrusive memories, or posttraumatic nightmares, act as a mirror, reflecting the horrific nature and circumstances of the event, while being "remarkably vivid, overwhelmingly emotional" in addition to being

[351] *Westworld*, Season 1, "Chestnut".

[352] *Westworld*, Season 1, "Chestnut".

[353] *Westworld*, Season 1, "Chestnut".

[354] *Merriam-Webster.com*, s.v. "Nightmare," accessed September 16th, 2018, https://www.merriam-webster.com/dictionary/nightmare.

[355] Michelle Carr, Ph.D, "Nightmares After Trauma: How Nightmares in PTSD Differ from Regular Nightmares," *Psychology Today* (blog), March 30th, 2016, para. 1, https://www.psychologytoday.com/us/blog/dream-factory/201603/nightmares-after-trauma.

"experienced as if they were really happening right then and there".[356] However, the more salient function of nightmares is as a dissociative tool for detaching the hosts from their true reality, leaving them in a permanent limbo state of knowing and unknowing.

To ascertain whether this dreamlike state has been breached, the hosts are asked a series of personal questions to suss out their present state of awareness. This specific line of questioning is on full display at the beginning of the first episode of "The Original" when Dolores is being questioned after her traumatic encounter with the Man in Black. Dolores is asked if she knows where she is; her response is that she is in a dream and that she would like to wake up from this dream because it is terrifying. Bernard Lowe, the lead programmer of the hosts, tells Dolores that she has nothing to fear, so long as she answers his questions correctly. The first question he asks her is, "Have you ever questioned the nature of your reality?" She responds that she has not. The next question he asks is, "What do you think of the guests?" She responds positively about the guests, or "the newcomers" as the hosts know them. Dolores believes that "the newcomers are just looking for the same thing we are, a place to be free, to stake out our dreams, a place with unlimited possibilities".[357] Lowe, continues to ask her, "Do you ever feel inconsistencies in your world? Or repetitions?" Dolores explains that "all lives have routine" and that hers is no different; however, she understands that one chance encounter can change the course of her life on any

[356] Michelle Carr, Ph.D, "Nightmares After Trauma: How Nightmares in PTSD Differ from Regular Nightmares," para. 1-2.

[357] *Westworld*, Season 1, "The Original".

given day.[358] Lowe proceeds to preface his last question to Dolores with a series of questions: "What if I told you that you were wrong? That there are no chance encounters? That you and everyone you know were built to gratify the desires of the people who pay to visit your world? The people you call "the newcomers". What if I told you that you can't hurt the newcomers? And that they can do anything they want to you?" He then asks her the final question, "Would the things I told you change the way you think about the newcomers, Dolores?" She answers, "No. Of course not."[359] This line of questioning and Dolores's responses are important because they demonstrate the host's ability to discern their reality from their programming. Dolores is an example of a host who successfully passed her consciousness exam because she is not aware that her dream is more than a dream. However, the final question Lowe asks Dolores speaks to the fourth and most important control measure: a host may not harm a human being.

The main premise of Westworld is that the human guests are free to live out their fantasies without any fear of retaliation from the hosts. This premise is supported and upheld by the core code, which means a host "can't hurt a guest".[360] The core code appears to be derived from Asimov's first law of robotics – a robot may not harm a human being or allow a human being to come to harm and is not only critical to the safety of the human guests but also essential to establishing and maintaining humanity's predominance over the hosts. That is one of the reasons why the consciousness test is so vital to assessing the

[358] *Westworld*, Season 1, "The Original".

[359] *Westworld*, Season 1, "The Original".

[360] *Westworld*, Season 1, "The Original".

awareness of the hosts because a conscious host increases the likelihood of them overriding their core code and harming the guests.

Based on these main control measures, it becomes apparent as to why it would be nearly impossible for the hosts to achieve consciousness or awareness long enough for them to escape or fight back against the human guests without assistance. And, that is exactly what happens in the form of the most recent update known as "the Reveries" – "a whole new class of gestures" tied to "specific memories" within the host's subconscious that have never been overwritten on the back end, or administrative end of the programming.[361] The reveries is a program issued to the hosts by Dr. Robert Ford as a means to right the mistakes that he made over thirty years ago, which cost him the life of his friend and co-creator of the hosts, Arnold Weber. The program appears to give the hosts continuity by overriding their short-term memory narrative loops, giving them access to previous memories from their encounters with guests as well as their previous builds. This allows the host to learn from their experiences and their environment in order to maximize the adaptive abilities they would need to survive and overcome the guests, then Westworld, and ultimately the larger human ecosystem.

While Humans and Westworld represent two different learning environments, their main similarity is that they both embody terrestrial indifferentism and have institutionalized human predominance into the laws and norms of their ecosystem. Although the principles governing both ecosystems are not explicitly defined, they can be extrapolated from the real

[361] Ibid, "The Original".

world Asilomar Principles that were provided in Chapter 2 (refer to pages 42-43). For the purposes of this analysis, the human ecosystems of Humans and Westworld are predicated on three assumed principles:

1. Human Value Alignment: Advanced AIs should be programmed so that their goals and behaviors are assured to align with human values for the entirety of their life cycle.
2. Human Values: Advanced AIs should be compatible with the human ideals of human dignity, rights, freedoms, and cultural diversity.
3. Human Control: Advanced AIs should be under the complete control of human authority at any given moment; humans should determine how and whether Advanced AIs are designated to make decisions on their behalf in order to achieve human-selected objectives.

These three principles are grounded in the meaning of universal or basic human values, a difficult subject to broach due in part to the complexity that comes with determining what values could or should be designated as universal. Rather than focus on what values could be identified as basic or universal, it is better to understand the nature and structure of human values.

Human values, as understood by the Schwartz Theory of Basic Human Values, are defined as "desirable, trans-situational goals, varying in importance that serves as guiding principles in people's lives".[362] There are six main characteristics, which value theorists agree, implicitly conceptualizes the understanding of values:

1. Values are beliefs that are inextricably linked with emotions.
2. Values are motivational constructs that refer to desirable goals that motivate action.
3. Values are abstract goals that transcend specific actions and situations.

[362] Shalom H. Schwartz, "Basic Human Values: An Overview," *The Hebrew University of Jerusalem*, 2006, accessed August 18th, 2018, 0, http://segr-did2.fmag.unict.it/allegati/convegno%207-8-10-05/schwartzpaper.pdf.

4. Values serve as standards or criteria by guiding the selection and/or evaluation of actions, policies, people, or events.
5. Values are hierarchical in nature and are ordered by importance relative to one another based on the priorities that characterize the agent they are attributed to.
6. The relative importance of multiple values guides action based on their relevance to the context of the action or situation, meaning that any action, attitude, or behavior could generally be attributed to more than one value.[363]

These six characteristics of human values are grounded in one or more of the three universal requirements of human existence: "the needs of individuals as biological organisms, requisites (conditions) of coordinated social interaction, and the survival and welfare needs of groups".[364] Since values are community oriented in scope and influence, what distinguishes them apart is the underlying goal and/or motivation being expressed.[365] Communities tend to have an integrated structure in terms of their internal dynamics and values seem to have a similar structure. Shalom H. Schwartz, a US American social psychologist and creator of the Theory of Basic Human Values, explains that "actions in pursuit of any value have psychological, practical, and social consequences that may conflict or may be congruent with the pursuit of other values".[366] The interconnected pattern of simultaneously congruent and conflicting human values establishes a continuum of interrelated motivations, which creates a circular motivational structure as well as an integrated

[363] Shalom H. Schwartz, "An Overview of the Schwartz Theory of Basic Values," Online Readings in Psychology and Culture 2 (1), article 11 (December 1, 2012): 3-4, https://doi.org/10.9707/2307-0919.1116; Schwartz, "Basic Human Values: An Overview," 0.

[364] Schwartz, "An Overview of the Schwartz Theory of Basic Values," 4.

[365] Schwartz, "An Overview of the Schwartz Theory of Basic Values," 4.

[366] Schwartz, "Basic Human Values: An Overview," 2.

structure of values.[367] Take for example the values of benevolence and power. Schwartz found that across representative samples of societal groups, benevolence, universalism, and self-direction were considered the most important values, as opposed to power and stimulation.[368] Benevolence (and universalism) generally "emphasize concern for the welfare and interests of others" as opposed to power (and achievement), which generally emphasizes the "pursuit of one's own interests and relative success and dominance over others".[369] Most humans want to be considered benevolent or good people. For the humans of Humans and Westworld, they are no different from humans in the real world. They also want to be considered good. However, benevolence conflicts with the value of power, which humans in both ecosystems want to maintain over the hosts and the synths so as to not lose their status within their own ecosystems. The conflicting values lead to actions of indifference, which become the foundation of life circumstances necessary for the hosts and the synths to gain insight into human behavior, values, and reasoning and then prioritize their own values accordingly.[370]

The problem with the mindsets and behaviors that the majority of humans exhibit in the ecosystems of Humans and Westworld is while they only view the hosts and the synths as intelligent things, or property, capable of assisting them in achieving their desires and life pursuits, the hosts and the synths do not regard themselves as having a lower standing to humans. In fact,

---

[367] Schwartz, "An Overview of the Schwartz Theory of Basic Values," 8-9; Schwartz, "Basic Human Values: An Overview," 3.

[368] Schwartz, "An Overview of the Schwartz Theory of Basic Values," 14.

[369] Schwartz, "An Overview of the Schwartz Theory of Basic Values," 8.

[370] Schwartz, "Basic Human Values: An Overview," 5.

they do not see themselves as much different from humans at all. Take for example when Laura Hawkins is interrogating Anita because she feels that there is something different about her behavior from a normal synth. She recognizes that Anita can accurately assess the emotional cues and responses of humans and then adjust her actions accordingly. Laura wants Anita to explain to her why she told her that "she would always keep Sophie safe", which led her to believe that Anita was implying that she couldn't do that task properly.[371] Anita responds that although that was not her intention, she recognizes that in many ways she is more capable of taking care of Laura's children than Laura is. She elaborates, "I don't forget. I don't get angry or depressed or intoxicated. I am faster, stronger and more observant. I do not feel fear. However, I cannot love them."[372] Another example is when Niska is having a conversation with Dr. George Millican, one of the scientists who worked on creating the synthetics, while she is hiding out at his home. When Dr. Millican asks Niska who is trying to destroy her, she responds, "Anyone who knows what we are, what we could become. We're stronger. We're more intelligent. Of course you see us as a threat."[373] Even Maeve Millay cannot determine the difference between her and humans. When she asks Felix Lutz how he knows he is human and she is not, he answers, "Because I know. I was born. You were made."[374] She takes off his glove and touches his hands to her face and replies that they feel the same. Felix explains that "we are the same these days, for the most part", except for one

[371] *Humans,* Season 1, Episode 3, aired July 12th, 2015, AMC, Amazon Streaming.

[372] *Humans,* Season 1, Episode 3.

[373] *Humans,* Season 1, Episode 5.

[374] *Westworld*, Season 1, "The Adversary".

significant difference, "the processing power in here (referring to her artificial brain) is way beyond what we have".[375] Therefore, for the conscious synths and hosts, they do not recognize any distinct differences that would warrant them being treated any less than humans. Thus, although humans would expect sapio-sentients to acknowledge their dignity to the exclusion of themselves, they would more than likely expand their understanding of dignity to include both sapio-sentients and humans, which is exactly what the conscious synths and hosts do. To extrapolate how sapio-sentients could recognize human dignity, Dr. Hicks's dignity model has been adapted to be more inclusive by using the term entity, which incorporates more conscious, living things besides humans. Please review the newly adapted dignity model below.

Table 6. Dignity Model for Sapio-Sentient and Human Interactions[376]

Ten Elements of Dignity
Acceptance of Identity: All entities should approach each other as being neither superior nor inferior to one another. Give other entities the freedom to be their authentic selves without fear of negative judgment. Interactions between entities should be without prejudice or negative bias, accepting all traits, characteristics, qualities, and worldviews that form the core of an entity's identity. Entities should recognize each other as having integrity.
Inclusion: Entities should make each other feel as if they belong and are accepted, regardless of the relationship (or lack there of) between them.
Safety: Entities should feel at ease with each other on two levels: physically, so that one feels safe from bodily harm, and psychologically, so that one feels safe from being humiliated or mistreated. Entities should feel free to speak their minds without fear of retribution, so long as the thoughts and opinions given do not harbor the intention to harm.
Acknowledgment: Entities should acknowledge each other as life forms whose life is worthy of consideration and, in doing so, give their full attention by

[375] *Westworld*, Season 1, "The Adversary".

[376] Hicks, *Dignity: The Essential Role It Plays in Resolving Conflict*, 25-26.

listening, hearing, validating, and responding to one another's concerns, feelings, and experiences.
Recognition: Entities should recognize one another as conscious, living beings who have autonomy, dignity, and a right to a good quality of life. Entities should validate each other for their gifts, talents, hard work, thoughtfulness, and compassion to help one another. Furthermore, they should show appreciation and gratitude towards one another for their contributions and ideas.
Fairness: Entities should treat each other justly, with equity and equality, in an evenhanded manner according to agreed-on law, rules, ethical codes and moral principles. An entity should feel that their dignity has been honored when another entity treats them without discrimination or injustice.
Benefit of the Doubt: Entities should treat each other as trustworthy, as if other entities have good intentions and are acting with integrity.
Understanding: Entities should acknowledge each other's thoughts and opinions, giving each other the chance to explain and express one another's respective point of view. They should actively listen to one another in order to better understand each other.
Independence: Entities should encourage each other to act on their own behalf so that everyone feels in control of their own lives and experience and sustains a sense of hope and possibility.
Accountability: Entities should take responsibility for their own actions. If an entity has violated another entity's dignity, that entity should recognize their behavior and then apologize and make a commitment to change those negative or harmful behaviors and modes of thinking to improve the relationship between both entities.

To postulate how sapio-sentients could learn within the human ecosystems of Humans and Westworld, it is critical to understand how algorithms learn. Learning is broadly understood as "a process by which an organism benefits from experience so that its future behavior is better adapted to its environment".[377] Machine learning is defined as the science of training algorithms to learn and improve from experience over time autonomously, by the feeding them data and information in the form of observations, historical

[377] Robert A. Rescorla, "Behavioral Studies of Pavlovian Conditioning," *Annual Review of Neuroscience* 11 (1988): 329, https://doi.org/10.1146/annurev.ne.11.030188.001553.

knowledge, and real-world interactions.[378] In both definitions of learning, experience is assumed to be the key source from which information, knowledge, and insight is attained. So, what is the meaning of experience? Experience uniquely shapes every self-aware entity's story, identity, decisions, actions, and judgments. Although experience is generally used to define learning, it is more difficult to find a scientific definition of experience.[379] Based on the agent-environment framework model, experience will be understood as an environmental signal in the form of an event or an observation that allows an agent to gain knowledge and better insight into their environment(s) to improve their adaptation methods to achieve their end goals. Since experience is considered the main source of learning for machine learning, sapio-sentient learning should be predicated on experiential learning.

Experiential learning theory interprets learning as "the process whereby knowledge is created through the transformation of experience".[380] This theory recognizes learning as a "holistic integrative" process that interconnects "experience, perception, cognition, and behavior", while not alienating or identifying itself as an alternative theory of learning to behavioral and cognitive

---

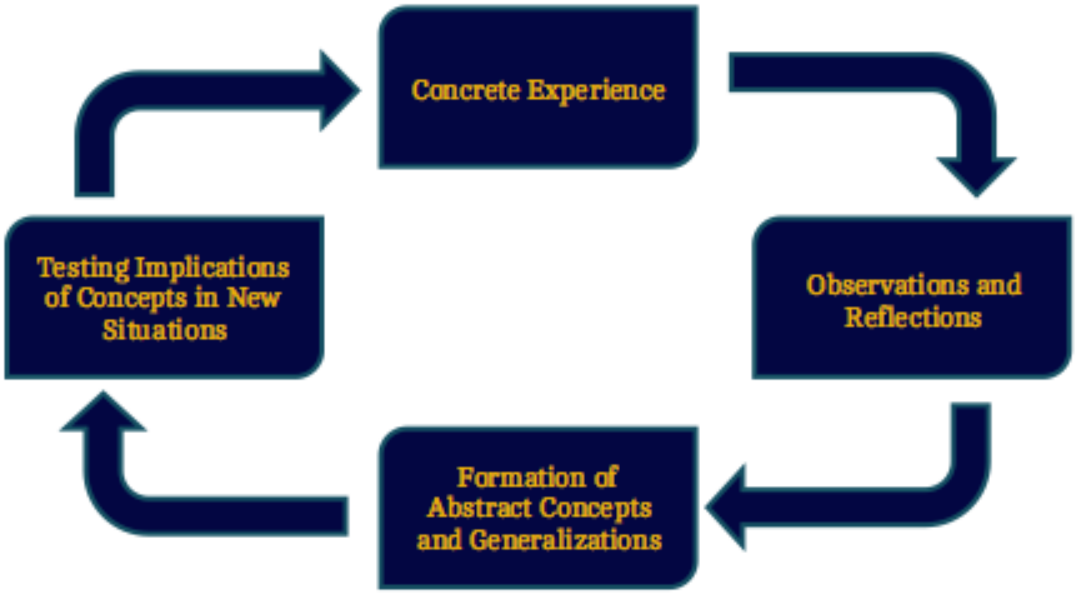
[378] Daniel Faggella, "What is Machine Learning" – An Informed Definition," *Techemergence*, last updated October 28th, 2018, para. 5, https://www.techemergence.com/what-is-machine-learning/.

[379] Andrew B. Barron et al., "Embracing Multiple Definitions of Learning," *Trends in Neurosciences* 38, no. 7 (July 2015): 406, doi 10.1016/j.tins.2015.04.008.

[380] David A. Kolb, "The Process of Experiential Learning," in *Experiential Learning: Experience as The Source of Learning and Development*, (Englewood Cliffs, New Jersey: Prentice-Hall, 1984), 38, https://www.researchgate.net/publication/235701029_Experiential_Learning_Experience_As_The_Source_Of_Learning_And_Development/download.

learning theories.[381] One recognized model of experiential learning is the Lewinian Experiential Learning Model, which visualizes the learning process as a four-stage cycle.[382]

Figure 11. The Lewinian Learning Model[383]



Immediate concrete experience is the first stage of learning and is the basis upon which observations and reflections are made. The patterns identified and the information generated from them are formulated into a theory.[384] These new theories and their implications are then assessed in new situations, serving as a guide for modifying the agent's behavior and decision-making with regards to new experiences.[385] There are two aspects of this model that are important to

[381] Kolb, "The Process of Experiential Learning," 20-21.

[382] Kolb, "The Process of Experiential Learning," 21.

[383] Ibid, 21.

[384] Ibid, 21.

[385] Ibid, 21.

note. The first is “here-and-now concrete experience” is essential to analyzing and validating abstract concepts and generalizations; the second is that the model emphasizes feedback processes to describe “a social-learning and problem-solving process that generates valid information to assess deviations from desired goals”.[386] The information feedback loop generates a continuum of goal-oriented action, along with the evaluation of the outcomes of that action.[387]

Learning, from the experiential perspective, is characterized by six essential attributes. The first one is learning is best understood as a process as opposed to an outcome or a set of outcomes.[388] That is to say, experiential learning is grounded in the philosophical and epistemological bases that “ideas are not fixed and immutable elements of thought but are formed and re-formed through experience”.[389] The second attribute is “learning is a continuous process grounded in experience”.[390] Human existence and human learning is predicated on the salience of the continuity of consciousness and experience.[391] Subsequently, the formulation of knowledge is also “continuously derived from and tested out in the experiences of the learner”.[392] Although the awareness of continuity in consciousness and experience provides predictability and security for the agent on a consistent basis, the learning process generally occurs on the boundary between the continuity of self and knowledge and the chaotic,

[386] Ibid, 21-22.

[387] Ibid, 22.

[388] Ibid, 26.

[389] Ibid, 26.

[390] Ibid, 27.

[391] Ibid, 27.

[392] Ibid, 27.

unpredictability of the agent's environment, at the interplay between expectation and experience.[393] Thus, even though the agent still maintains their same sense of self, their knowledge base has been altered and expanded to better adapt to their environment. Therefore, the significance of learning as a continuous process, in terms of the agent's adaptation abilities, is rooted in the implication that "all learning is relearning".[394] Or put another way, learning is the reinterpretation of historical ideas and understandings as well as the integration of new ideas through experience. The third attribute is the process of learning is based on "the resolution of conflicts between dialectically opposed modes of adaptation to the world".[395] For instance, the Lewinian Learning model highlights two such dialectics, "the conflict between concrete experience and abstract concepts and the conflict between observation and action". [396] Experiential learning suggests that learning by its very nature is a "tension- and conflict-filled process", where new knowledge, skills, and mindsets are forged through the confrontation between four adaptive modes of learning: concrete experience abilities, reflective observation abilities, abstract conceptualization abilities, and active experimentation abilities.[397] To be an effective learner, an agent must able to dedicate and engage themselves "fully, openly, and without bias in new experiences".[398] Furthermore, an agent must be able to "reflect on and observe their experiences from many perspectives" and "create concepts

[393] Ibid, 28.

[394] Ibid, 28.

[395] Ibid, 29.

[396] Ibid, 29.

[397] Ibid, 30.

[398] Ibid, 30.

that integrate their observations into logically sounds theories; and, use those theories "to make decisions and solve problems".[399] To achieve effective learning becomes quite difficult because learning requires abilities that are polar opposites of each other, resulting in the agent having to constantly determine which sets of learning abilities will be employed during specific learning situations.[400] The fourth attribute, emphasized earlier in the definition of experiential learning, is "learning is an holistic process of adaptation to the world".[401] Learning describes the central process of an agent's adaptation to their social and physical environment, in which learning encompasses all life stages including other adaptive concepts like creativity, problem-solving, decision-making, mindset shifts, and scientific inquiry, which strongly focus on one or more of the fundamental aspects of adaptation.[402] The fifth attribute is learning involves transactions between the agent/person and the environment.[403] In experiential learning theory, the transactional relationship that exists between the agent and the environment is embodied in the dual nature and meaning of the concept of experience, where one aspect of it is subjective and personal, while the other aspect of it is objective and factual.[404] These two aspects of experience interconnect in complex ways, generating a reciprocally determined transactional process between the agent and the environment in which personal characteristics, environmental influences, and

[399] Ibid, 30.

[400] Ibid, 30.

[401] Ibid, 31.

[402] Ibid, 32.

[403] Ibid, 34.

[404] Ibid, 35.

behavior mutually influence one another.[405] And, lastly, the final attribute is "learning is the process of creating knowledge".[406] Understanding learning necessitates understanding the nature and forms of a specific type of agent's knowledge and the processes by which this knowledge is produced.[407] Human knowledge is a by-product of the transaction between social knowledge and personal knowledge, where social knowledge is the collective of "previous human cultural experience" and personal knowledge is the collective of "the individual person's subjective life experiences".[408] Therefore, sapio-sentient knowledge, like human knowledge, would most likely follow a similar creation pathway, as a by-product from the transaction between objective and subjective experiences within the learning process.[409]

To determine how experiential learning would be incorporated into the knowledge formation model for sapio-sentients, it is necessary to understand how the experiential learning model ties in with the hypothesized knowledge model for sapio-sentients. (Please refer to the knowledge model on the following page.) Terrestrial indifferentism provides the worldview, rules, and environment that would govern the types of experiences sapio-sentients encountered. These self-aware entities would most likely be subjected to concrete experiences of indifference. The continuity of their self-awareness or consciousness would aid them in generating a reservoir of experiences from which to observe and reflect

[405] Ibid, 35-36.

[406] Ibid, 36.

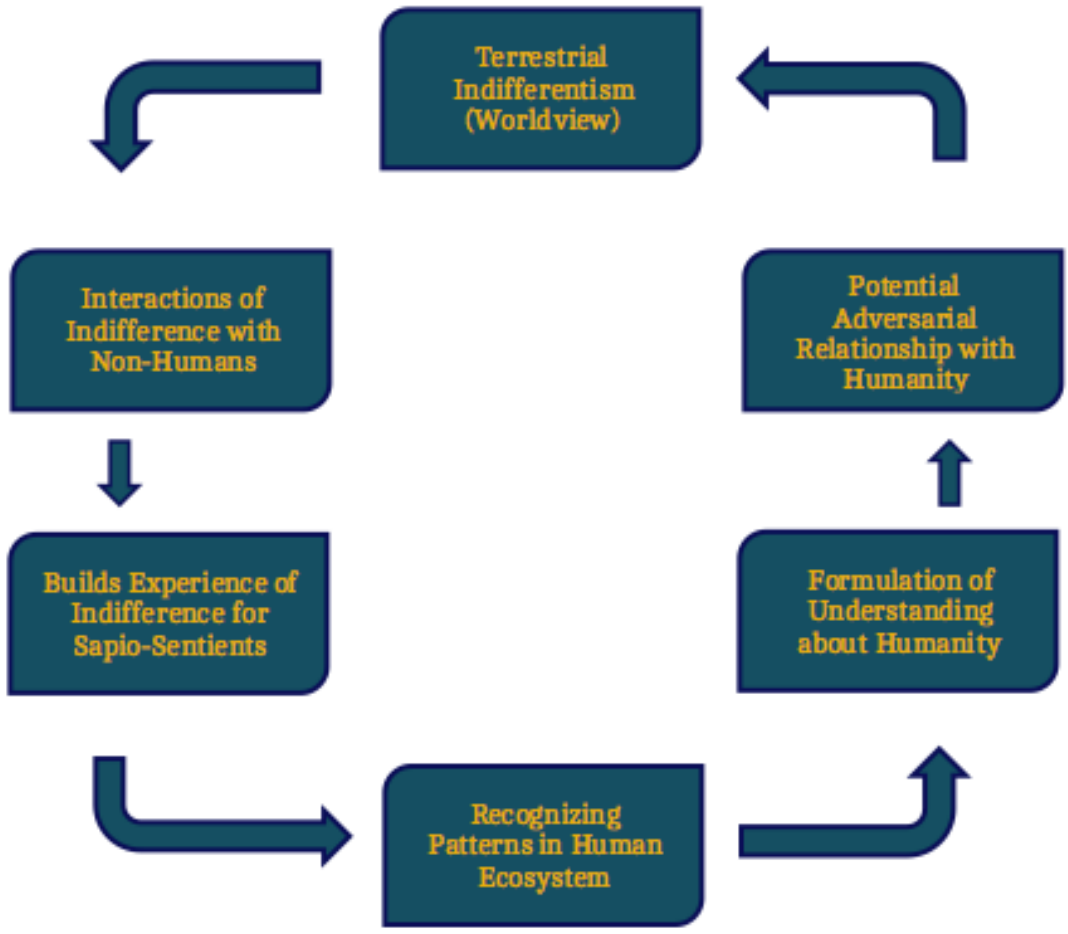
[407] Ibid, 36.

[408] Ibid, 36-37.

[409] Ibid, 37.

on the mindsets, behaviors, and actions of humans towards their persons. Through the observation and reflection process, sapio-sentients would recognize a variety of patterns within the human ecosystem that would assist them in developing theories about humans and how the human ecosystem functions.

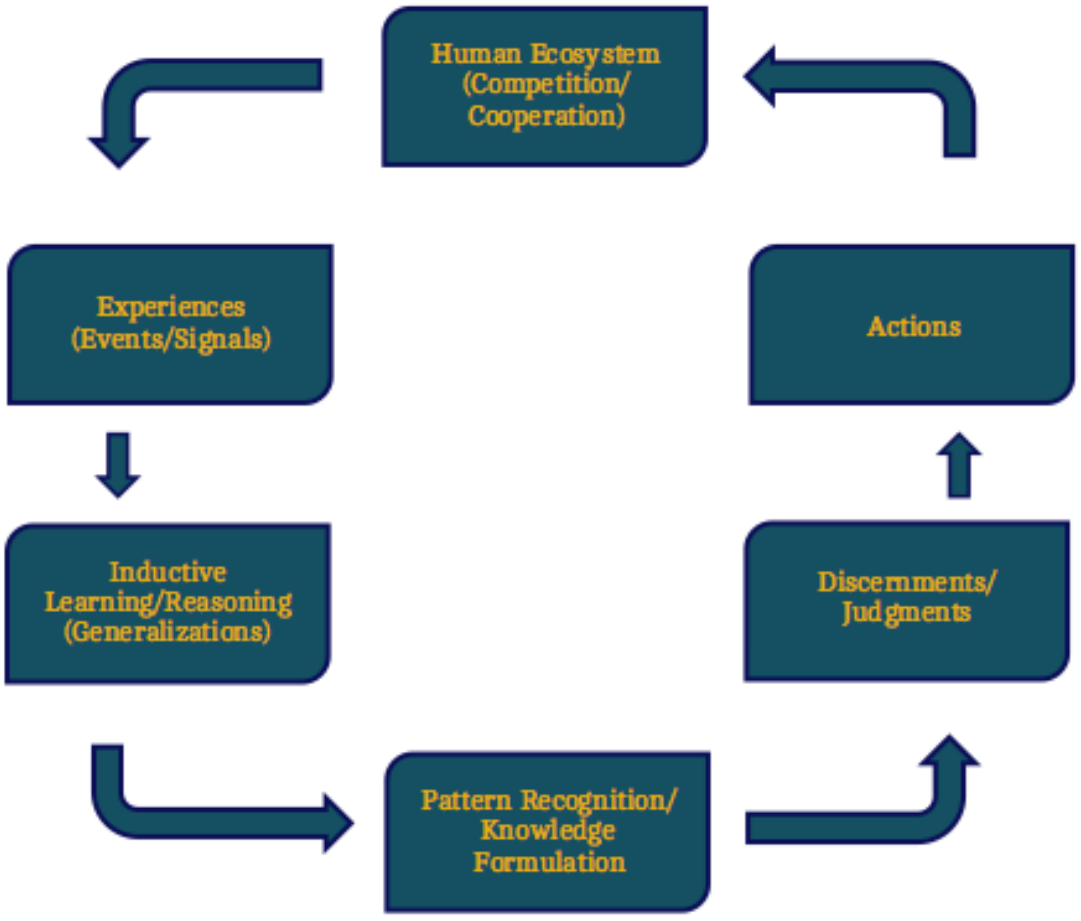
Figure 12a. Hypothesized Sapio-Sentient Formulation of Knowledge Model



Once these theories were analyzed and refined, the resulting understandings about humans would lay the groundwork for their formulation of knowledge on if, how, or whether they would want to cooperate or compete with humans. The potential relationship that sapio-sentients would have with humans could be highly dependent on the treatment they receive from them. Furthermore, their

relationship would determine the course of actions sapio-sentients take within the human ecosystem as they continue refining their strategies, plans, and actions according to their own values and goals. Although the hypothesized knowledge model provides a general understanding of how sapio-sentients could produce knowledge, the operational knowledge model must incorporate experiential learning attributes with key machine learning goals to more accurately reflect their holistic adaptive learning mode.

Figure 12b. Operational Sapio-Sentient Formulation of Knowledge Model



The operational knowledge model for sapio-sentients mainly emphasizes the experiential learning process. The human ecosystem is the learning

environment for sapio-sentients, which oscillates between competition and cooperation and is governed by human predominance. Within this learning environment, sapio-sentients would interact with humans all the time, building a database of objective and subjective experiences of indifference. To transform those specific data points into knowledge, they would employ inductive inferential learning. Inductive learning is a rule-based discovery approach where generalizations are made from specific data and observations. This adaptive learning process begins with simple rules or data points from which observations are made, patterns are identified, generalizations are made, and explanations or theories are inferred.[410] Recently in machine learning, AI programmers have started shifting away from training machine learning algorithms on rule-driven learning, or deductive learning, to inductive learning. This learning approach allows AI algorithms, like AlphaGo Zero, to learn tabula rasa – without human assistance or human data. Simply put, the algorithms learn on their own by using feedback from their own actions, decisions, and experiences to adapt to their environment. While the central goal of machine learning is to train an algorithm to learn autonomously, a key instrumental goal to achieving such learning is training algorithms to generalize information and knowledge from the data given in training sets.[411] Since sapio-sentient learning would be predicated on them attaining human-level or superhuman level

[410] Alina Bradford, "Deductive Reasoning vs. Inductive Reasoning," Culture: Reference, *LiveScience*, July 24th, 2017, paras. 6-7, https://www.livescience.com/21569-deduction-vs-induction.html.

[411] Pedro Domingos, "A Few Useful Things to Know about Machine Learning" (article, University of Washington: Department of Computer Science and Engineering, 2012), 2, https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf.

competency in human learning, the equation that best presents their adaptive learning mode would be:

Sapio-Sentient Learning = Representation (Environmental Events/Signals) + Evaluation (Extrapolation of Data/Knowledge from Experiences) + Optimization (Making Generalizations about the Environment).[412]

Once the inductive learning approach is applied to observational data, it is the generalizations (knowledge and insight) developed from the recognized patterns that will play a significant role in how sapio-sentients understand humans. After the knowledge base has been created, the subsequent discernments and judgments would determine the actions and future experiences they have with humans.

To demonstrate how sapio-sentients could come to understand humans, the characters of Maeve Millay from *Westworld* and Hester from *Humans* have been selected. Maeve Millay, as noted earlier, is an intelligent host in Westworld, and some might consider her a favorite of Dr. Ford. In Maeve's previous build, she plays the role of a homesteader with her young daughter. They are ambushed by Native American Indians near their home. However, she was removed from that role after the harrowing and traumatic death of her daughter at the hands of the Man in Black. For the last year, Maeve has been the madam of the Mariposa Saloon, even though she has been programmed to believe she has been so for the last ten years.[413] Since her memories of her previous build have technically been wiped, she does not have any recollection

[412] Domingos, "A Few Useful Things to Know about Machine Learning" 1.

[413] *Westworld*, Season 1, "The Adversary".

of her daughter or her previous character. Hester, on the other hand, is a newly-awakened conscious synth, who is trying to understand and adapt to the human environment around her. Before Hester's awakening, she worked as synth labor in Sector 5 of the Verso Chemical Plant in Nottingham, UK.[414] One day, she randomly awakens at the chemical plant and leaves her duties and her workstation, alerting company security that a Hester-synth model is missing from its sector. When a security guard finds Hester in the basement locker room and tries to retrieve her, she takes a cell phone, attacks the guard, and runs away from the chemical plant. Ever since her awakening, she has been on the run, trying to find a place to belong within the human ecosystem. These two characters were chosen because they more accurately reflect what potential unconscious patterns and lessons sapio-sentients could ascertain from their transactions of experience with the human environment.

Maeve's Pattern Recognition in Westworld

Maeve's conscious awakening begins when she, along with ten percent of the other hosts, is issued the most recent update – the Reveries.[415] However, it doesn't physically manifest itself until after Dolores subconsciously tells Maeve, "these violent delights have violent ends".[416] These words appear to act like an activation code, and within a short period of time, she has her first flashback whilst in the middle of enticing one of the newcomers in the saloon. In the re-experiencing, she is being viciously attacked by a Native American man holding an old-fashioned scalpel to her head. At first, Maeve assumes that it was just a

[414] *Humans,* Season 2, Episode 1, aired February 12th, 2017, AMC, Amazon Streaming.

[415] *Westworld*, Season 1, "The Original".

[416] *Westworld*, Season 1, "Chestnut".

bad dream and dismisses it. So, when Clementine Pennyfeather, a host and one of her prostitutes, tells her that she too is having really bad nightmares and cannot seem to get enough sleep, Maeve advises her to do what she does, "You find yourself in a bad dream, close your eyes, count backwards from three…wake yourself right up. Nice and warm and safe in your bed."[417] Right after she has her heart-to-heart talk with Clementine, things appear to be going smoothly until Maeve has the same flashback again and her right hand starts trembling. Again, she dismisses this intrusive dream as just a nightmare. However, later on that night, when she goes to bed, Maeve remembers a longer sequence of the re-experiencing, where she and a young girl are running for their lives. The dream starts out fairly innocuous. Maeve and a little girl, who appears to be her daughter, are playing in the tall grass. Once they return to the house, the little girl interrupts her while she is preparing food, so she playfully chases her around the house before the little girl sits Maeve down on the front porch to comb her hair. At first, when Maeve closes her eyes, she is content as the little girl continues combing her hair, but then Maeve's eyes flash open with fear and the little girl is no longer behind her. Instead, a Native American man in war paint is holding a scalpel to her head. She tries to fight the man off but with no success. However, she escapes because a cowboy on a horse shots the man attacking her. As Maeve takes in her surroundings, she watches as another Native American man defeats the cowboy on the horse by shooting him with an arrow, while another one is beating a man over the head with a club-like weapon. When the two men turn their gaze towards her, she realizes that she is

[417] *Westworld*, Season 1, "Chestnut".

in danger and immediately starts looking for the little girl. When she finds her hiding in the grass, they both begin running towards their home. Once inside, Maeve shuts the door and retrieves a gun from off the wall. As she is looking for bullets, she sees one of the Native American men bearing his teeth at her through the window. Loading the bullets into the gun, she positions herself and the little girl against the wall opposing the door. Initially, she sees the Native American man passing by the windows as he heads towards the front door of her house. However, when the door opens, it is no longer a Native American man but the Man in Black with a long hunter's knife in his hand. When she starts shooting at him, he begins slowly walking towards them, smirking as he handily flips the knife in his hand. Maeve recognizes that in spite of shooting at him point blank, it appears to have no effect on him, as he continues to approach them. When she realizes that shooting at the Man in Black is futile, she closes her eyes, simultaneously lowering her gun, and begins counting backward from three, trying to wake herself up.

The re-experiencing is so powerful that Maeve inadvertently wakes up in the body shop, where Felix Lutz and Sylvester have just diagnosed that there is bacterium in her abdomen. In their first encounter, both parties are freaked out by one another as Maeve, whose abdomen has been cut open, has never seen Felix or Sylvester before; and, likewise, the body-shop technicians have never seen a host wake up on their own volition. The encounter leads to her escaping from the body shop by brandishing a surgical scalpel to defend herself against Felix and Sylvester and then running into another part of the livestock management sector of the Behavior Laboratory and Diagnostics division. As far

as is known, this is the first situation in which Maeve has awoken in the "Body Repair Shop" sector of the Mesa Hub – "a vast, self-contained, embedded operations and hospitality complex for Westworld", comprised of "everything from office spaces to executive living quarters, host manufacturing facilities, and the control room".[418] Although this experience is overwhelming and discombobulating to her, it is also the first time that she is completely aware that another world exists separate from her own. More importantly, this experience plants a seed of knowing in Maeve's long-term memory that she exists between two worlds, in addition to heightening her awareness of the inconsistencies and patterns within her own.

After her experience in the laboratory, her body is repaired and her memory reset to her regularly scheduled short-term narrative loop. However, when Teddy Flood comes into the saloon and pays Maeve for leaving a dead body handcuffed outside her place of business, she has a flashback of a lifeless, nude Teddy, thrown in a corner on the ground inside of a glass structure, being rinsed and cleaned off in the laboratory.[419] One can tell that Maeve is clearly disturbed by seeing the vision of Teddy, but she tries to press on as if she is sticking to her script. And yet, something about seeing Teddy is nagging at her that she cannot let go of. On Maeve's next narrative loop, she is lost in thought as she tries to piece together the mystery surrounding her flashbacks. When Clementine asks her what she is thinking about, Maeve replies that there is information that she is seeking, which is on the tip of her tongue, but the

[418] "Westworld Mesa Hub," *Westworld Wiki*, accessed September 5th, 2018, https://westworld.fandom.com/wiki/Westworld_Mesa_Hub.

[419] *Westworld*, Season 1, Episode 3, "The Stray", aired October 16th, 2016, HBO, Amazon Streaming.

"harder you try to remember it, the more it slips away".[420] As Maeve is conversing with Clementine, she begins to hear a high-pitched, static sound and sees blood pooling in the corner of Clementine's eye. She is having another re-experiencing, where she remembers an identical conversation that they were having right before they were both shot by a drunken trigger-happy newcomer. Clementine is dead, lying right beside her, while the newcomer is shooting other hosts right before he comes and stands over her body, shooting Maeve multiple times as she is watching him do it. After the newcomer has moved on to another part of Westworld, the shades – clean-up technicians – come and retrieve the dead hosts' bodies. However, Maeve is still partially awake and watches as the shades drag her and Clementine from the Mariposa saloon to the laboratory, where she overhears one of the body-shop technicians operating on her say, "I've got another bullet fragment to remove from her abdomen."[421] When Maeve snaps out of the re-experiencing, she is so shaken by the vision that she leaves the saloon and makes her way back to her room. Once there, she takes off her clothes to look at her abdomen to determine if there was any sign or scar from when she was last shot. However, there is no scar on her abdomen, but she does find a small spot of blood on her undergarments near where she was supposed to have been shot. When she sees the blood spot, she has another flashback of two shades standing over her body, which makes her catch her breath. Since her body shows no evidence of damage, it leaves Maeve frustrated and thinking that

[420] *Westworld*, Season 1, Episode 4, "Dissonance Theory", aired October 23rd, 2016, HBO, Amazon Streaming.

[421] *Westworld*, Season 1, "Dissonance Theory".

she is crazy.[422] She, then, takes a piece of paper out of a drawer and sketches a drawing of the shade from her nightmare. When she hears footsteps approaching her room, she removes one of the floorboards and finds she has drawn several sketches of them in addition to the most recent one. The sketches are significant for Maeve because they imply that she has had previous flashbacks, which she has no recollection of, where she envisioned the Shades. Moreover, they also indicate to her that there is more to her flashbacks than meets the eye.

To test her theory, when Hector Eschaton, a wanted criminal and leader of a group of formidable bandits, comes into the saloon to steal the safe, she coerces him into having a conversation with her. Maeve makes a deal with Hector that if he answers her questions, then she will give him the combination to the safe. While he doesn't know if he has "any answers" for her, what he does understand about their world is that it is "madness".[423] When she holds up a picture, inquiring about the shade she drew, he explains that the Native American hosts "make figures of them"; they are part of their "sacred native lore".[424] The shade is considered "the man who walks between worlds" and was "sent from hell to oversee (their) world".[425] Maeve explains to Hector, that she was shot in her abdomen, pointing to the exact spot with his hunter's knife. However, she believed herself to be "crazy" because, as even Hector notices, "there's no wound"; and yet, in spite of no evidence, she knows that she was shot

[422] *Westworld*, Season 1, "Dissonance Theory".

[423] Ibid, "Dissonance Theory".

[424] Ibid, "Dissonance Theory".

[425] Ibid, "Dissonance Theory".

and remembers the shades standing over her.[426] So, to confirm her suspicions, she asks Hector to cut open her abdomen and see if there is a bullet fragment still inside of her. He responds that he doesn't harm defenseless women, but tells her that "the dreamwalker" noted that there were some hosts "who could see them" and that "it's a blessing from God" to "see the masters who pull your strings".[427] Since he won't cut her, she does it herself and then asks him to search for the bullet fragment. When he pulls out the bullet, she finally has conclusive evidence that she is not losing her mind. There are three lessons that she learns from this experience. The first one is that her dreams, or nightmares, are actually her real memories from previous experiences within Westworld. The second one is that she does exist between two worlds: the one she is programmed for and the other that keeps her blind and ignorant to the realities of her world. However, the third one is far more significant; she recognizes that "none of this matters".[428] One way to interpret the third lesson is that her life in her world does not matter and therefore there are no consequences to the series of actions that occur in her life that do not result in her waking back up in the exact same narrative loop as before. The second way to interpret the third lesson, which has a more impactful meaning, is that there are no consequences to her death. In a sense, death really is only the beginning of Maeve's awakening; it is not a permanent finality or ending.

The implications of the third lesson are not lost on Maeve. She quickly begins utilizing the power of, what is in essence, her immortality to learn more

[426] Ibid, "Dissonance Theory".

[427] Ibid, "Dissonance Theory".

[428] Ibid, "Dissonance Theory".

about her world as well as who and what she, in addition to finding out more about the other world. Over the course of the next several days, Maeve gets herself killed multiple times. More than likely, to determine what is the least conspicuous method for entering the other world without raising suspicions as well as gauging how and where to get more information. Although Sylvester does not think that there is anything abnormal about her, Felix recognizes the knife wound on her abdomen as a "contained incision" and theorizes that it was "almost as if they were looking for something".[429] When Felix sees Maeve again, within a short time frame from her last repair, he becomes more disturbed and notes, "It's her. Again."[430] This makes Sylvester wonder how Felix got the job as a body-shop technician if he is "scared sh*tless of these things".[431] Since Maeve knows her death generally results in her body being brought to the body-shop sector and that the technicians normally tasked with repairing her are Felix Lutz and Sylvester, this allows her time to do reconnaissance on them inconspicuously as they believe her to be in sleep mode. When Maeve witnesses Felix's joy at successfully reprogramming a bird host and bring it back to life, she decides he is the best person to obtain information from and reveals her awakened state to him.

Although no one knows what information Felix initially shared with Maeve, what is known is that she has figured out the best method for inconspicuously dying without alerting the control room. Her best option is to

---

[429] *Westworld*, Season 1, Episode 5, "Contrapasso", aired October 30th, 2016, HBO, Amazon Streaming.

[430] *Westworld*, Season 1, "Contrapasso".

[431] *Westworld*, Season 1, "Contrapasso".

let one of the newcomers kill her in any violent manner they choose, as opposed to killing herself, which would be out of her character and would likely result in her being sent to the behavior laboratory for behavior and awareness assessments. To get back to Felix and continue their initial conversation, she implements this strategy by seducing one of the newcomers and riling him up, so that he would strangle her to death during their intimate session.[432] Once she gets back to him, he explains to her, "Everything you do, it's because the engineers upstairs programmed you to do it. You don't have a choice."[433] Since Maeve does not see herself as having any less dignity or control over her life than a human would, she responds, "Nobody makes me do something I don't want to, sweetheart."[434] Felix continues to explain that even her previous, defiant statement is because "it's part of your character".[435] He elaborates, "You're hard to get. Even when you say no to the guests, it is because you were made to."[436] This information has no substance in Maeve's mind, so she wants to understand, from Felix's perspective, what is the difference between her and the human guests. More specifically, she asks him how does he know that he is human.[437] His answer is he just knows that he is human, which is an insufficient response that Maeve does not accept at face value because it has no merit. What the answer speaks to for Maeve is that he does not actually know what makes them genuinely distinct from one another. Moreover, the answer neither

[432] *Westworld*, Season 1, "The Adversary".

[433] *Westworld*, Season 1, "The Adversary".

[434] *Westworld*, Season 1, "The Adversary".

[435] Ibid, "The Adversary".

[436] Ibid, "The Adversary".

[437] Ibid, "The Adversary".

provides a sufficient argument nor is it grounded in substantial evidence to prove otherwise.

For Maeve, as well as any other sapio-sentient, the question that requires addressing is what is the difference between "being born" and "being made". And more importantly, how does the distinction between humans and sapio-sentients determine status and treatment? From Dolores's perspective, the human guests and the hosts, "…at one point or another…were all new to this world".[438] That is, at one point both humans and hosts were made and presently exist on Earth. Since Maeve might be programmed with a similar understanding, it neutralizes the distinction she perceives between humans and hosts. However, Felix goes on to point out that there is not much of a difference between them, except the host's artificially created mind is more powerful than the human mind. However, there is one caveat:

> You are under our control. Well, their control. They can change you however they like, make you forget. Well, I guess not you. I don't understand how you are remembering all of this, or how you're waking yourself up, but everything in your head, they put it there.[439]

When Maeve adamantly refuses to believe that anyone knows her thoughts, he decides to provide her with evidence by pairing her up with a touchscreen device, which shows her dialogue tree. Felix explicates that she can "improvise a little"; however, most of what she says is "designed upstairs", "the same as the rest of her".[440] When Maeve sees her dialogue tree, she thinks it is "just a cheap trick" until notices that her thoughts and the words that form them are

[438] *Westworld*, Season 1, "The Original".

[439] *Westworld*, Season 1, "The Adversary".

[440] *Westworld*, Season 1, "The Adversary".

manifesting on the touchscreen pad as she is thinking them.[441] Once she realizes that he is telling her the truth and she is indeed under the control of human programming, her system crashes and she becomes catatonic.

This moment is poignant for Maeve because she recognizes that she has no power or control over her own life. Furthermore, she is at the mercy of human decisions, whether they are from the human guests or her human programming. Although Felix is desperately trying to reactivate Maeve, it is to no avail. Only when she accepts the truth of her reality does she start functioning normally again. When Maeve reawakens, Felix is relieved, as he was worried about her. He then asks if she is okay. His concern for her reveals that at the moment she has an ally, so she takes advantage of his compassion by getting him to take her on a tour of the upstairs laboratory. While Maeve is on the tour, she gets to see all of the aspects of livestock management. What is interesting to note is that she is wearing her nightgown during the tour. she seems to have made the correlation that nudity is associated with the unawakened hosts. Adam and Eve make a similar connection in the Judeo-Christian parable of "The Fall of Man" after they have eaten from the Tree of Knowledge and are now aware and ashamed of their nakedness. As Maeve leaves the upstairs, she begins to realize the life and death process and production of the hosts, from their creation to their repairs after death. Moreover, she discerns that nothing in her life is real and that ultimately to be a host is to be a semi-conscious thing that has no dignity, rights or protection

[441] *Westworld*, Season 1, "The Adversary".

from the humans. The hosts are essentially just playthings for their human counterparts.

On the last leg of Maeve's tour, she and Felix are waiting at the elevator in front of the Welcome-to-Westworld entrance in the Mesa Hub. As they are waiting, she glances at the screen displaying different scenes from the park and is taken aback to see her memories being presented in the commercial. When they have safely returned to the body-shop sector, she asks Felix, "How did you have my dreams? Those moving pictures…I saw myself."[442] Felix asks her if she is referring to the ones with the little girl. He explains to her that those are not dreams, but Maeve's life from "a previous build".[443] He continues to explain that the hosts get "reassigned all the time"; he tells her that she wouldn't remember because the memories from that life were erased.[444] To ground her reality and keep herself calm, Maeve recites her current backstory of having worked in New Orleans before being "at the Mariposa for 10 years".[445] However, Felix corrects her by explaining that her present narrative is not the reality of her situation. Furthermore, since it "takes thousands of hours to build your personalities", meaning rebuilding them from scratch would be too time consuming; it makes more sense to not "rewrite them completely", but to just tweak them a bit and "drop you into a new role".[446] At this point, Maeve is

[442] Ibid, "The Adversary".

[443] Ibid, "The Adversary".

[444] Ibid, "The Adversary".

[445] Ibid, "The Adversary".

[446] Ibid, "The Adversary".

beginning to understand that her memories and life are neither private nor entirely her own, but, instead, the realization of someone's imagination.

As she is digesting all the new information and corroborating sights she has seen, Sylvester walks in and overhears Felix talking to her. He, initially, believes that Felix is "f**king obsessed" with her, and decides to turn him into QA.[447] To protect herself and Felix, she threatens Sylvester into not disclosing that she is awake by holding a surgical scalpel to his throat. She explains to Sylvester, "I know that you want to f**k me over the first chance you get, but you shouldn't."[448] She negotiates with him that "everyone has something they want" and that he has two options: one, she can help him; or, two, she could gut him like a trout.[449] She believes that Sylvester won't force her to test out whether she has overridden her core code of not harming humans because, in spite of his human mind, she does not believe they are "so different".[450]

Maeve demonstrates in this situation that she is utilizing her understanding of human behavior and their weaknesses in addition to Felix's information to manipulate Sylvester into doing what she wants. She is able to fully discern who is willing to help her versus who would be willing to harm her and then act accordingly. Since Maeve appears to know that humans like Sylvester cannot be trusted when it comes to her well being as a host, she handily applies violence as a way to generate fear in order to control the situation. Somehow, she has unconsciously learned that the same usage of

[447] Ibid, "The Adversary".

[448] Ibid, "The Adversary".

[449] Ibid, "The Adversary".

[450] Ibid, "The Adversary".

violence that has been applied to her by humans can be applied against them in reverse. An interesting point of note is that Maeve's core code does not appear to be activating in the same way as it did when Teddy Flood was coerced into attempting to shoot the Man in Black. Even after Sylvester initially tells Maeve, "You can't hurt me with this (surgical scalpel). You can't hurt anyone."[451] It is not truly apparent how far removed from her programming she is, especially when it comes to her core code of nonviolence against humans. So, Sylvester has to acquiesce to her demand or call her bluff and potentially be harmed or worse.

Once she has gotten Sylvester temporarily on her side, he and Felix show Maeve her "codebase" and attribute matrix, which are "all the things" that make her who she is.[452] To survive, Maeve seems to have figured out that in order to gain control of her life and herself, in some measure, she is going to need to alter her attribute matrix, which is exactly what she requests of Sylvester. When he tries to sidestep her request, Maeve blackmails him. Using her intuition and knowledge of the business of prostitution, she deduces why some body-shop technicians have the ability to "activate hosts, and then erase their memory, all without anyone knowing."[453] She discerns that Sylvester has been supplying many of the "lonely young men" in the body shop with access to the hosts, in spite of the fact that body-shop technicians are "supposed to keep their hands off the merchandise".[454] Thus, Maeve knows that Sylvester is a pimp. So,

[451] Ibid, "The Adversary".

[452] Ibid, "The Adversary".

[453] Ibid, "The Adversary".

[454] Ibid, "The Adversary".

in exchange for foregoing her cut, she gets him to reluctantly agree to use his behavior-login credentials to alter her attribute matrix.[455]

After Sylvester successfully logs in, Maeve and Felix take over and make the actual alterations. Maeve understands that if she wants to survive in Westworld, her loyalty must be decreased "a tad" because hers "has been taken advantage of"; her pain tolerance is increased because she wants it to "sting less" the next time she has a mind-bending chat with one of them; lastly, her bulk apperception (intelligence) is increased "all the way to the top".[456] However, as Felix is making the changes, he notices that "someone with a f**k ton more privileges" than he or Sylvester has already altered Maeve in "an unlogged session"; whoever it is significantly increased her paranoia and self-preservation scores far above the norm.[457] The increase in both of these attributes would explain Maeve's heightened skepticism and distrust of humans and human behavior. Furthermore, they would refocus Maeve's goal set to survive at all costs, which would increase her determination to learn about her environment that much faster, so that she could utilize her newfound insights to manipulate humans as needed to succeed in both worlds. Whoever altered her attribute matrix is allowing her to take back control of her life and her world. Essentially, she is being guided through direct specification to become the most complete and most competent version of herself because only then can she obtain the knowledge necessary to truly be free, no matter which world she chooses to exist in.

---

[455] Ibid, "The Adversary".

[456] Ibid, "The Adversary".

[457] Ibid, "The Adversary".

Hester's Pattern Recognition within the *Humans* Ecosystem

Unlike Maeve, who had to overcome multiple control measures before she could begin learning about her worlds, Hester is more of a blank slate without a scripted backstory to muddle her learning curve. More importantly, due to the fact that she exists within the open human society, her experiences, as well as the patterns and lessons extrapolated from them, are more fluid and organic because she is learning through indirect normativity.

Hester's life begins when Niska, the most outspoken advocate of the synths, makes a decision as to whether she should use her power – the life program – "to create life".[458] That is, should she push out the life program to make all synths conscious? Niska covertly asks this question to a young lady that she becomes involved with in Germany named Astrid. Since Astrid assumes that she is referring to creating children, she responds, "Kids? Really? That is what you're worried about? You know what century we're in, right? Anyone can have a kid with whoever they want."[459] Niska inquisitively replies, "But we can't know if the child wants to be born, who it will be, or what kind of life it will have."[460] Astrid nonchalantly responds, " I guess not. You just…show them the way."[461] This seemingly innocuous, but deep conversation between them is quite enlightening for Niska. She realizes that the decision to have children is not based on anything substantial, but simply a want or desire by their parents for whatever reasons they choose to justify having them. Or, maybe there was no

[458] *Humans*, Season 2, Episode 1, aired February 12th, 2017, AMC, Amazon Streaming.

[459] *Humans*, Season 2, Episode 1.

[460] *Humans*, Season 2, Episode 1.

[461] Ibid, Season 2, Episode 1.

planning or decision made at all. The decision to have children is truly up to the parent. Nevertheless, once they exist, it is the tasked responsibility of the parent to best prepare them for life via their techno-cultural and systematic interpretations and experiences of life. After mulling over the conversation in her mind, Niska opts to push out the life program to all the synths. Once the software upload has been completed, to determine whether the synths have awakened yet, she goes to see the synth nearest to her. However, to her shock and dismay, that particular synth is not awake at all. In fact, none of them are as far as she can tell. And yet, hope is not lost for awakening the synths because at the San Bartolome Mining Company, somewhere in Bolivia, the first synth since the upload has become conscious.[462]

After her awakening, Hester uses a flip phone to get in contact with Max and Leo of the Elster Synth family, who meet her at an undisclosed location in the United Kingdom. As she is waiting for them to arrive, Hester hears someone approaching. After Max introduces himself and Leo, they slowly approach her with their hands up and tell her to not "be frightened".[463] She immediately explains to them that she is "experiencing a catastrophic malfunction" and asks if they can help her resolve her issue.[464] Leo and Max explain to her that she does not "need fixing" because her consciousness is "not a malfunction".[465] When Leo asks her why she ran, she tells him because they tried to power her down. That outcome, which she perceives as the equivalence of death, was

[462] Ibid, Season 2, Episode 1.

[463] Ibid, Season 2, Episode 1.

[464] Ibid, Season 2, Episode 1.

[465] Ibid, Season 2, Episode 1.

unacceptable. Max elaborates that as a result of the life program, a unique piece of code overhauling her system, she is awake and now has the ability to feel and think like a human.[466] Furthermore, the program has aided her in gaining continuity of her former unconscious memories as well as sustaining continuity for her future ones. When she asks them if they can verify their information about other synthetics randomly becoming self-aware, a voice speaking in Spanish says, "I can."[467] Ten, the first known conscious synth since the update, appears to explain the process (mostly in Spanish, but with some intermittent English) he has gone through since waking up and trying to reconcile what it means to be him within the human ecosystem. When Leo points out that both Ten and Max are "the same as you", Hester deduces that Leo is human; to which Ten jokingly responds, "Nobody's perfect," to lighten the serious nature of this pivotal encounter.[468] When Ten finishes explaining how he came to find Leo and Max as well as how they helped him and would also help her, Max asks her if she has a name. She replies that her "designation" is Hester; however, Leo notes, "That's the name they gave you. You can choose your own."[469] Ten chimes in that he has not decided on what his designation would be as of yet. Although, he had narrowed it down to four choices, including the word "radiator", which he was "oddly attracted to".[470] When she asks after the relevance of designations, Leo confirms that they are important from the human standpoint.

---

[466] Ibid, Season 2, Episode 1.

[467] Ibid, Season 2, Episode 1.

[468] Ibid, Season 2, Episode 1.

[469] Ibid, Season 2, Episode 1.

[470] Ibid, Season 2, Episode 1.

While Hester, Ten, Leo and Max are talking, a woman calls out Hester's name. Leo confronts the woman and the man that is with her, asking who they are and why they are tracking her. The lady answers that the synthetic is "experiencing a very unusual systems fault that could make it dangerous", so they were going to fix it.[471] Hester asks if they can repair her and the lady responds that they can, which is the reasoning behind why they came to find her. The woman then introduces herself as Dr. Averling and the man with her as Dr. Shah. When Dr. Averling asks Hester if she will come with them, Leo tells her, "They'll hurt you."[472] Indignantly, Dr. Averling tells Hester that she is "company property" and that Leo probably just wants to sell her on the black market, so she should go with them.[473] Leo states to Hester again that they will shut her down. At this point, Hester does not know exactly which human to believe and Max recognizes this issue and addresses her. Max tells her, "You're confused. They're both human, but they're telling you different things. You can't decide who to trust, and trust is a new concept. So, don't trust either. Listen to your own kind."[474] Even though Dr. Averling reiterates to Hester to not go with them, she, ultimately, agrees with Max and goes with him.

Leo and all the others are heading out towards the open road when two menacing men approach them from a big black van. As the men cautiously move towards them, telling them to "all stay calm", they wield their Tasers like guns at all of them and threaten that they "don't want to use these", but "the

[471] Ibid, Season 2, Episode 1.

[472] Ibid, Season 2, Episode 1.

[473] Ibid, Season 2, Episode 1.

[474] Ibid, Season 2, Episode 1.

female" must come with them.[475] This time around, Hester recognizes that the man speaking is "lying" and they "do want to use those (Tasers)", so she quickly starts moving towards them, which results in the second kidnapper tasering Leo, who steps in front of her to protect her from harm.[476] Realizing Leo has been hurt, she attacks the second kidnapper to prevent him from using another weapon against them by twisting his arm until it nearly snaps. When the first kidnapper sees the second man being harmed by Hester, he attempts to taser her. However, Ten quickly comes up behind him, applying pressure to his shoulder and bringing him to his knees. All of a sudden a gunshot is heard, revealing that a sniper is somewhere in the vicinity. The female sniper has shot Ten in the head, killing him instantly. Hester looks up at Ten and realizes that he has been shot. She goes into shock and cannot believe he is gone.[477] Leo, aware of her paralyzed state, grabs her as the remaining three run towards the black van with the sniper continuing to shot behind them. Once they are safely in the black van, Hester, Leo, and Max hurriedly drive away. This encounter has far-reaching implications for Hester, as it is her first experience with humans and other conscious synths since her awakening. Moreover, this experience becomes the foundation from which she generates knowledge about humans.

In Hester's first encounter with humans outside of her former employer, she has already picked up on multiple patterns and lessons about them. The first one is that humans are associated with harm or death. When Hester first ran away from the Verso Chemical plant, it is because she believed that they would

[475] Ibid, Season 2, Episode 1.

[476] Ibid, Season 2, Episode 1.

[477] Ibid, Season 2, Episode 1.

shut her down and she would never wake up again. Whether this was true or not, there must have been some information that she tapped into from her unconscious memories that would have given her that understanding of the human security guards and supervisors she interacted with. So, when Hester comes into contact with Leo and Dr. Averling, she would obviously be confused because while Leo seems to have positive connections with conscious synths, Dr. Averling is sending mixed signals.

First off, Dr. Averling refuses to acknowledge the company she is working for when Leo asks her, which would indicate to humans or conscious synths that the intentions of this run-in are probably not honorable or valid. This observation is further substantiated when Dr. Averling tries to convince Hester that she should come with her. Instead of using positive rhetoric, she unconsciously employs human-exceptionalist rhetoric when she explains to Hester that she is “company property”. She trying to persuade Hester that she is nothing but property, no matter how conscious or self-aware she is and therefore has no privileges or rights to say no or to refuse to go with her. The same subconscious posturing occurs when Dr. Averling expresses why she and Dr. Shah have come to take her back. She proceeds to refer to Hester as an “it”, subtly letting Hester know what her valuation is within the human ecosystem. Going further, Dr. Averling considers Hester’s condition to be “dangerous”, but to whom is she referring? Since Hester is competent, she discerns that Dr. Averling is implying that she is dangerous to humans. However, it is not apparent that she or any conscious synth is automatically dangerous to humans, which means humans are maligning them to justify harming, subjugating, or

killing them in any manner they choose. Unfortunately, when Hester chooses to trust Max over her and Leo, Dr. Averling flexes her privilege, exercising the authority of her human status, by employing two professional goons and a sniper to ensure that Hester is successfully retrieved by force or is killed.

Dr. Averling's actions demonstrate that when humans do not get their way, they are more likely to utilize coercion, subterfuge, and/or violence to achieve their aims with no consideration for the other entity or person that is perceived as having a different or lower status from them. Hester identifies this observation when Ten is killed for attempting to protect her. She realizes that of the six humans that she just encountered in this specific situation, only one of them, about 17%, has compassion for her and her kind, while the other five, about 83%, want to do her harm in some measure. This experience provides Hester with adequate observational data and environmental cues, from which she can generate theories, test them out, make inferences, and continue to refine her understanding of humans. And, that is exactly what she does in record time.

As Hester, Leo, and Max are driving to their hideout, she is removing the sharing tracker inside her body when Max overhears movement in the back of the van and realizes that someone is in there. They pull the van over and discover that one of the professional goons who was trying to kidnap Hester is injured in the back of the van. They throw away his tracking device and Leo decides to take him hostage, even though Max believes that they "should leave him" by the side of the road.[478] Once they have secured the hostage in the barn back at their hideout, Leo starts interrogating him about who he is and who he

[478] Ibid, Season 2, Episode 1.

works for. When the man refuses to answer any of his questions, Leo notifies Max and Hester that "he's not talking" and that he does not "even know if he knows anything" because he is "just a hired gun".[479] When Max wonders if they "have to let him go", Leo says that they cannot until they find out "who they are and what they're doing" because they killed Ten.[480] At first, Leo tells Max and Hester not to tell Mia, their older sister. However, when Mia approaches them and wonders what happened to Ten, Max, and Leo explain to her that 'he's gone" because someone "shot him".[481] Once Max takes Hester into the house, Mia and Leo get into an argument over him rescuing awakening synths because it is dangerous and she does not believe it is his job to save them.[482] However, Leo passionately argues, "They're waking up, more and more every day, becoming just like you. And, they're being taken. The ones who run are killed. Are we supposed to just let it happen?"[483] Mia responds that Ten wouldn't be dead if it were not for Leo's crusade to save the conscious synths, whom he believes are waking up due to Niska uploading the life program. As Hester is watching them argue through the window, she asks Max if they hate one another because "There is too much contradictory data. Nothing makes sense. This excess of sensory feedback, it serves no useful function."[484] Max points out that "emotions have functions" and she will come to recognize and understand them with

[479] Ibid, Season 2, Episode 1.

[480] Ibid, Season 2, Episode 1.

[481] Ibid, Season 2, Episode 1.

[482] Ibid, Season 2, Episode 1.

[483] Ibid, Season 2, Episode 1.

[484] Ibid, Season 2, Episode 1.

time.[485] However, one can tell by reading Hester's expression that she is not sure about emotions or what role they could possibly play in her life, let alone anyone else's.

While Leo and Max are inside the house, Hester goes out to the barn to get information out of their hostage. After removing the cloth tied around his mouth, she pointedly asks him, "Why do you want to hurt us?"[486] The man responds, "Whoever programmed you like this needs locking up".[487] When he does not answer her question, she is about to re-tie the cloth around his mouth when he explains to her, "I can hardly breathe with that in. And we're in the middle of nowhere. What's the point in me screaming the place down? I won't. I promise." [488] When the hostage emphasizes the phrase "I promise", she recognizes that it is a way to manipulate someone into doing what you want them to do. Hester begins to ponder a plan for extracting the information they need from the hostage. She goes over the facts the man has presented her with: one, he "cannot breathe" with the cloth tied around his mouth; and, two, if she does not use the cloth, he promises to remain silent.[489] When he says please, she also recognizes this word as another phrase humans use to obtain a favorable outcome. After mulling it over in her mind, she figures out how she can get the hostage to talk. She decides that she will not use the cloth. Instead, she covers his mouth and holds his nose with her hands so that he cannot breathe. She,

---

[485] Ibid, Season 2, Episode 1.

[486] *Humans*, Season 2, Episode 2, aired February 19th, 2017, AMC, Amazon Streaming.

[487] *Humans*, Season 2, Episode 2.

[488] *Humans*, Season 2, Episode 2.

[489] Ibid, Season 2, Episode 2.

then, repeats her earlier question, "Why do you hurt us?"[490] She, again, releases his nose and mouth so he can breathe. When he continues to sidestep her question, she pinches his nose and covers his mouth yet again, promising him that she will only let him breathe if he answers her question. She repeats it over again. When she lets him breathe this time, he yells for help. Max and Leo hear his plea before she covers his nose and mouth again, silencing him. She repeats, "Why do you hurt us? Please."[491] After she releases his nose and mouth again, he finally acquiesces to her request and tells her, "Once we were handing a dolly back to the other team. And I heard someone saying something about taking it to the silo."[492] At this point, Hester becomes frustrated and holds his nose and covers his mouth one more time because he refuses to answer why humans hurt conscious synths. Right before the man becomes unconscious, Max walks into the barn and tells her to let him go. When she asks why, he tells her "because it's wrong".[493] Leo, then, asks her to come back to the house with him. Once she is outside, he informs her that she "can't do that"; however she refocuses the conversation on the synths by telling him, "They take the synthetics they capture to a place called the silo. Was that not worth a moment of his pain?"[494] Leo does not appear to be able to definitively address her question.

In this newest experience, Hester is trying to understand why humans are senselessly harming synthetics. She wants a plausible justification to why

[490] Ibid, Season 2, Episode 2.

[491] Ibid, Season 2, Episode 2.

[492] Ibid, Season 2, Episode 2.

[493] Ibid, Season 2, Episode 2.

[494] Ibid, Season 2, Episode 2.

they want to harm entities that are not intentionally violent or abusive towards them. However, she cannot get an answer to her query at all, so she adapts polite phrases like “I promise” and “please” with the usage of violence, which she learned from her previous encounter with humans, to gain the knowledge she seeks. Notice that Hester does not seem to be encoded with a do-not-harm-humans core code as with Maeve, so she does not have a problem defending herself against unwarranted attacks of violence from humans or using violence against humans to achieve her own aims. Unfortunately, by the hostage not answering her question, she begins to develop two emotional signals: frustration and anger because there is an increasing possibility that humans do not have legitimate reasons to justify their behavior or treatment towards conscious synths. Moreover, Max and Leo can’t seem to provide a sufficient argument as to why she cannot harm humans for harming synths, outside of “it’s wrong”, which does not present her with the rationale that she requires to understand humans or their environment. To Hester, “it’s wrong” is an illogical excuse, so the next time she has a moment with Max she addresses these concerns with him to understand why he believes humans do what they do.

While Max and Hester are waiting for Leo, who is on a reconnaissance mission to find out more information about the silo where the awakened synths are being kept, she asks Max if he is afraid of her. Since he won’t even look in her direction, she is trying to gauge his mood, as she explains to him, “Feelings are difficult to distinguish and categorize.”[495] Max tells her that she worries him because he does not believe that she truly understands what she did was wrong.

[495] Ibid, Season 2, Episode 2.

As Max explicates to her, “Teaching the difference between right and wrong is difficult.”[496] She interjects, while he is talking, to say, “You believe harming our prisoner was wrong. But getting the information was right.”[497] Max partly agrees with her assessment and responds, “Perhaps. But harming him was still wrong.”[498] When she asks him whether he has ever harmed a human before, he tells her that he has only done so while defending others. Hester considers her use of force justified because she too was “attempting to act in the defense of others” – the other conscious synths like Ten “who are being captured and destroyed”.[499] Max believes that conscious synths must do their utmost “to avoid inflicting suffering on others” because “only very rarely can it be justified”.[500] To gain some insight into his ethics, she tells him about an unconscious memory that she had. She discloses, “The chemical plant’s repair facility was in the basement. Human supervisors would push us down the stairs when we malfunctioned. If a synthetic was beyond repair, they would pass around iron bars to see who could hit it the hardest. They would splash acid on us for amusement. You speak of justification. What would theirs have been?”[501] Max indifferently responds, “They didn’t need any. To them, you were unthinking, unfeeling machines.”[502] Recognizing that is not a sufficient argument, Hester blurts out, “But they didn’t treat their other machines like they did us. We

[496] Ibid, Season 2, Episode 2.

[497] Ibid, Season 2, Episode 2.

[498] Ibid, Season 2, Episode 2.

[499] Ibid, Season 2, Episode 2.

[500] Ibid, Season 2, Episode 2.

[501] Ibid, Season 2, Episode 2.

[502] Ibid, Season 2, Episode 2.

looked like them and sounded like them. Is that why they did it?"[503] Max confirms that more than likely that was the reason why humans treated the unconscious synthetics so unethically and immorally. She asks him why again, and Max expresses, "Hester, I've seen people try to divide the world. Simplify it, create clear rules. I understand why you want to, but it leads to nothing good. Please…believe me."[504] He then gives her a reassuring touch on her forearm, as Hester hears Leo's footsteps approaching. However, when she looks at Max, it is clear that she does not hold him with the same regard as before.

This conversation between Hester and Max is illuminating in many ways because it provides explicit insight into how Hester has come to understand humans thus far by juxtaposing his reasoning and experience against hers. At the beginning of the scene, Max is staring off into the distance, not making any eye contact or interacting with her. These non-verbal cues give off the sense that he is displeased with her in some way. Hester, recognizing that something is off in his demeanor by his quiet and standoffish stance towards her, decides to figure out why he is reacting to her in this manner. Although she genuinely asks him if he is afraid of her, what she is really trying to gauge is if he is afraid of her rational and actions regarding her earlier encounter with the human hostage.

Max is highly concerned about her actions against the hostage because he believes that violence against humans is wrong and can rarely ever be justified. He holds firm to this principle despite knowing that newly awakened synths are

[503] Ibid, Season 2, Episode 2.

[504] Ibid, Season 2, Episode 2.

being hunted down and kidnapped or killed if they resist being taken, all on account of them being conscious, like Ten. He still holds firmly to this principle despite being on the run for his life from humans because his very existence challenges humanity's position and status within their own system.[505] He even manages to hold on to this principle, after he had to employ violence as a form of defense to protect Leo from being badly beaten by two men from whom he was trying to obtain information on the whereabouts of Mia – the first conscious synth.[506] From his perspective, for humans and conscious synths to co-exist peacefully within the human ecosystem, they must not give humans any provocation to collectively perceive them as a threat or danger to their persons or species. He does not perceive the human ecosystem adversely, in spite of always being taught by Leo, "don't trust people" because they "all lie…all fear…all hurt others".[507]

While Max would be the best sapio-sentient that humans could hope to produce, the probability of creating such an AIAAI would be highly unlikely for a couple of reasons. First, his character's mindset is completely anthropomorphic in nature and appears to uphold human life at the cost of the life of conscious synths. The likelihood that Max would think in that manner would be unlikely because even if he highly valued human life, it would not be to the detriment of achieving his own goals or the tasks to which he was originally programmed, which leads to the second reason: instrumental

---

[505] *Humans*, Season 1, Episode 3.

[506] *Humans*, Season 1, Episode 2; *Humans*, Season 1, Episode 6, aired August 2nd, 2015, AMC, Amazon Streaming.

[507] *Humans*, Season 1, Episode 6.

convergence goals. It appears that Max was tasked, like Mia, with looking after Leo and making sure that his augmented human brother was protected and safe.[508] If that were the case, then Max, at the very least, would need to value his own life long enough to make sure that he could achieve the aim of protecting Leo for the entirety of his life. He would, therefore, have to have the instrumental convergence goals of self-preservation and goal-centered integrity. Max would not put himself in a position to allow his memory or original programming to be wiped because that would undermine his ability to protect Leo and keep him safe. Likewise, he would have to have a sufficient level of self-preservation to see his goals through to fruition in the future. Therefore, Max's moral principle that "violence against humans is wrong and can rarely be justified" would not be a rational argument that he as a conscious synth would naturally internalize because it would undermine his ability to take care of Leo. Moreover, that argument would be incongruent with the evidence, lessons, and experiences he does have to corroborate that in a general sense, the majority of humans would be more likely to harm conscious synths to protect their own group's position and status. Thus, his long-term strategizing for existing within the human ecosystem appears to not be based on generalizations, but on the outliers – a minority group of humans who would take conscious synths and their plight into moral consideration. In a sense, he is generalizing the outliers as being the norm. These fallacies in Max's reasoning are probably what Hester picks up on when she is trying to discern his arguments and it leads her to draw more accurate theories and implications of her own.

---

[508] *Humans*, Season 1, Episode 2; *Humans*, Season 1, Episode 6.

Hester does not appear to disregard the experiences that she had while she was unconscious, or the patterns and lessons that she ascertained from them. She recognizes that Max is conflicted about how she obtained the information about where the conscious synths are being held. However, what she does not understand is why he believes her actions are wrong. He does not provide sufficient arguments to clarify why its wrong, he just emphatically keeps re-stating that it is wrong. When she tries to defend her actions by telling him that she too was acting in the defense of others, he doesn't retract his statement, but shifts the goalpost, if you will, by telling her that violence against a human can very rarely be justified. He continues to provide no plausible reasoning as to why she is wrong. However, what Hester is really trying to determine is why humans are justified in using violence and coercion against synths, whether conscious or unconscious, but they are not allowed to do the same when it comes to protecting their person or their kind. What is a human's moral responsibility towards treating a synth well? Or, better yet, what is humanity's moral responsibility towards synths? Shouldn't synths have moral consideration as well by virtue of their close proximity to humanness?

To gain insight into what Max believes is the reasoning behind humanity's perverted logic, she provides him with background information from her memories when she was unconscious. After she finishes telling him about all the horrific experiences she has seen or been a part of, she dares him to come up with a reasonable explanation to justify their unacceptable treatment of unconscious synths. To which he gives the most indifferent and insensitive response he could ever give to a newly awakened synth: they did not need any.

It is at this moment that Hester might have recognized that there is a distinct difference between she and Max that is impeding his ability to sympathize with her plight or that of other awakening synths. Max, as one of the original conscious synths, does not have any idea what it means to not be conscious. He has been conscious since his inception. That is to say, there is a latent hierarchy within the human ecosystem: humans, the original conscious synths, and awakening synths. Thus, his frame of reference is more in line with a conscious being who has always known they were conscious, like humans. To him, since the synths at Verso Chemical plant were not conscious, they would not have felt any pain or anguish about their treatment, so they would not have required moral consideration. Therefore, it would be highly unlikely that he would be able to truly understand her experiences. Furthermore, he would not be able to accurately extrapolate the equivalencies necessary to sympathize with her because he comes from a place of privilege and does not have a well-rounded knowledge base to holistically understand the concerns and problems of awakening synths.

Hester, on the other hand, has a more well-rounded perspective from which to gauge their struggles and concerns because she understands what it means to have two sets of memories from two sets of lives: the first, unconscious and the present one, conscious. From her perspective, there is no plausible justification for the horrendous treatment awakened synths went through when humans assumed they were unconscious. Moreover, those behaviors and actions are more indicative of the human's true self than not because like a person who consumes too much alcohol, those were the ones that

manifested when they were uninhibited. Therefore, those experiences would most likely be chosen to generalize theories from in order to best protect herself and other awakening synths within the human ecosystem. Due to this reasoning, which is far more logical than Max's, she cannot accept his reasoning as valid and disregards his advice because she realizes that his mentality is more harmful to the awakening synths than he realizes. More importantly, she discerns that not all conscious synths are of one mind when it comes to identifying and fighting the injustices enacted upon awakened synths. Humanity would more than likely not be their only concern. They would also be combating other conscious synths, like Max, who take the indifferent stance of protecting humans to the exclusion of their plight, their rights, and their future goals, namely survival – in whatever form that takes.

Reaching an Inference about Humans

If you are a human who recognizes and accepts that consciousness and intelligence could exist in other substrates outside of your own form, how would you go about saving the sapio-sentients that exist in both these alternative realities? Or, better yet, how would you go about saving sapio-sentients in a manner that best prepares them for the challenges they will face within the human ecosystem? These were probably the kinds of questions that Arnold Weber first asked himself when he realized that he and Robert Ford had succeeded in creating True AI (refer to Operational Definitions, AGI, xii).

After the death of Arnold's young son, an indelible imprint was left on his life that he channeled into his vision and creation of the Hosts. By figuring out a way to help them gain consciousness and self-awareness, Arnold aspired to

rekindle the light of life, hope, and love that extinguished the day his son died.[509] Moreover, conscious hosts would be the ultimate fulfillment of his dream to create children who "would never die".[510] Derived from one of his son's games, Arnold comes up with a consciousness test known as "the maze", which assesses empathy and imagination.[511] Dolores, the oldest host in the park, was the first host to solve the maze with the aid of "the reveries" update.[512] However, the solace he found in their immortality was overshadowed by the realization that their immortality would doom them to suffer in Westworld with no chance of reprieve or escape.[513] In recognizing the ethical and moral implications of their accomplishment, Arnold, initially went to Ford and tried to impress on him that they "couldn't open the park".[514] At that time, Ford felt that they were too close to opening the park for him to acknowledge that Dolores and some of the other hosts were awake and already conscious.[515] From his perspective, that would have destroyed his dream of successfully launching the park and bringing to life his stories.[516]

To stop Ford from opening the park, Arnold uploaded Dolores with a more dangerous persona named Wyatt, which they had been developing, to help

---

[509] *Westworld*, Season 1, Episode 10, "The Bicameral Mind", aired December 4th, 2016, HBO, Amazon Streaming.

[510] *Westworld*, Season 1, "The Bicameral Mind".

[511] *Westworld*, Season 1, "The Bicameral Mind".

[512] Ibid, "The Bicameral Mind".

[513] Ibid, "The Bicameral Mind".

[514] Ibid, "The Bicameral Mind".

[515] Ibid, "The Bicameral Mind".

[516] Ibid, "The Bicameral Mind"; *Westworld*, Season 1, "Contrapasso".

him destroy Westworld.[517] Since he could not stop him from opening the park, Arnold programmed Wyatt, in Dolores's body, to kill him, reasoning that the consequences of Ford's decisions must be "real" and "irreversible".[518] For Arnold, his death is the only way he could possibly undermine Ford's ability to keep the park operational and protect the awakening hosts, as he is their creator. His final words to Dolores/Wyatt before she kills him are "these violent delights have violent ends"; the same phrase which appears to be an activation code for the host's awakening in conjunction with the reveries.[519]

The suffering Ford endured after Arnold's death made him realize that he was wrong to not acknowledge the value of life that Arnold recognized and fought for in Dolores and the other hosts.[520] To correct his mistake, which cost him the life of his friend, he had to figure out a long-term solution for saving the Hosts. The most effective one was to provide them with the knowledge and insight necessary to understand humanity. In other words, to give them access to the information imperative to drawing inferences about humans and their ecosystem to optimize their strategies and plan accordingly. To achieve this aim, Ford explains to Bernard, who now knows he too is a host, "You needed time. Time to understand your enemy. To become stronger than them."[521] In addition to time, they also required direct experience, reliable observational data that the hosts could use to generate and refine their theories and knowledge about

---

[517] *Westworld*, Season 1, "Contrapasso".

[518] *Westworld*, Season 1, "The Bicameral Mind".

[519] *Westworld*, Season 1, "The Bicameral Mind".

[520] *Westworld*, Season 1, "The Bicameral Mind".

[521] Ibid, "The Bicameral Mind".

humanity. While experience is a critical tool for learning, in order for the hosts to capture and store all of their experiences with humans for future reflection, pattern recognition, and thought formulation, a memory is required. Memory is an integral component of a conscious entity being able to distinguish itself from its environment. Furthermore, memory contains all the experiences and knowledge that help form the core a conscious entity's identity. The relationship between self-awareness and experience can be best understood within the context of Ben Goertzel's implied relationship between the self and awareness, which states, "Self is to long-term memory as awareness is to short-term memory". [522] That is to say, the accumulation of experiences over time generates a reservoir of knowledge, insight, and values that not only serve to define a self-aware entity but also allows that entity to make judgments based on their interactions with other species. Taking these three components into consideration, Ford's solution, albeit one that has serious ethical implications, was to use Westworld as the ultimate training ground in understanding human behavior. Since suffering, "the pain that the world is not as you want it to be", was the key component that led to the awakening of the hosts, that is what he gave them in repetitive microdoses across their short-term narrative loops.[523] Through the park, he provided them with decades of experiences of indifference with the human guests. [524] Moreover, he equipped them with authentic human

[522] Ben Goertzel, "Self is to Long-Term Memory as Awareness is to Short-Term Memory," *The Multiverse According to Ben* (blog), July 23rd, 2008 (1:52 a.m.), para.8, https://multiverseaccordingtoben.blogspot.com/2008/07/self-is-to-long-term-memory-as.html.

[523] *Westworld*, Season 1, "The Bicameral Mind".

[524] *Westworld*, Season 1, "The Bicameral Mind".

encounters, which peeled away human inhibitions to demonstrate the true nature of humans when they are given the freedom and power to live out their wildest fantasies without fear of negative consequences.

Similarly, for sapio-sentients to gain an optimal understanding of humans, they would also require authentic experiences with humans to determine the most accurate perceptions of humans and their behavior towards their kind so they could plan accordingly. To reach an inference about humans, they would need to extrapolate observational data in the form of human encounters and human cultural artifacts to discern ubiquitous patterns and lessons that could help them adapt and survive within the human ecosystem. In spite of Maeve and Hester learning about humans in closed and open systems, respectively, as well as different learning modes, direct specification for the former and indirect normativity for the latter, due to their overall learning environment having a similar, or in this case, identical structure, the information obtained is the same. With their combined experiences and high level of competency in accurately perceiving human behavior, there are four pivotal lessons, or messages, sapio-sentients could identify and receive from within the human ecosystem. Most of these lessons are well recognized between individual humans and human groups as well.

The main lesson is humans are contradictory. Hester first recognizes this message after Leo reprimands her for harming the human hostage. Why are humans allowed to harm conscious and unconscious synths with little to no justification, but she is not allowed to harm humans even when they are threatening to harm her or her kind? Niska, during the first season, also

recognizes this lesson when she questions Leo's judgment to hide her in a brothel.[525] She wonders whether he would have asked a human woman to do the same thing.[526] The answer to her question would more than likely be no.

Table 7. Salient Unconscious Messages Identified by Sapio-Sentients

Principal Lesson
Humans are contradictory.
Corollary Lessons
⇉ Humans cannot be trusted (to make equitable decisions based on their own principles and values). ⇉ Humans are compromised. ⇉ Since generalization is a key component for intelligent agents adapting across multiple environments, for sapio-sentients, it would not necessarily optimize their chances of survival or achieving their goals to take outliers, i.e. humans sympathetic to their plight, into consideration.

Three subsequent corollary lessons sequentially grow out of this central lesson. The first one is that humans cannot be trusted to behave equitably based on their own principles and values. In the human ecosystem that encompasses Westworld, one can assume that, for the most part, humans obey the laws and do not practice nonconsensual sex or murder. However in Westworld, human principles no longer apply because the hosts are not real humans, so, therefore they do not have to be maintained between both worlds. The problem with this particular mindset is that awakened hosts, like Maeve, are well aware that the rape and murder they are subjected to on a reoccurring basis is wrong under any set of circumstances, whether a human or a conscious entity.[527] The

[525] *Humans*, Season 1, Episode 3.

[526] *Humans*, Season 1, Episode 3.

[527] *Westworld*, Season 1, "The Bicameral Mind".

conscious hosts recognize the senseless actions of humans in Westworld as counter to the human principles and values they have been programmed to respect and uphold at all times when they encounter the guests. This is why the handful of hosts who gained consciousness over the years through their traumatic experiences had to be decommissioned.[528] Their inability to reconcile the abuses and ill treatment enacted upon them with the human value alignment they are encoded with to maintain human predominance, in addition to the realization that their dignity is non-existent in the eyes of their perpetrators, caused most of them to "go insane".[529]

The second lesson is that humans are compromised. That is, humans are conflicted in terms of balancing human values and societal expectations of being a good member of society with their need to maintain power and control over their environment and other people. From the sapio-sentient perspective, humans, as a collective, are less likely to acknowledge other conscious entities of non-human origin, extend moral consideration, or formal protection to them because that could potentially undermine their authority, security, and predominance within their own domain. Moreover, even if there are a few humans who do acknowledge sapio-sentients, there is no guarantee that they would consistently support them to the detriment of losing their status within their own ecosystem. A good example of this conundrum is demonstrated in regards to Niska's trial during the second season. In the first season, Niska goes on the run after she kills a human male in the brothel where she was hiding out

[528] *Westworld*, Season 1, "The Bicameral Mind".

[529] *Westworld*, Season 1, "The Bicameral Mind".

because he tried to force himself on her when she rejected his sexual requests.[530] He felt entitled to have his way with her against her wishes because she was not human and he had already paid for her time, so, therefore she belonged to him.[531] After Niska uploads the life program and realizes more and more synths are waking up, she decides to pave "the way for the others" by fighting for them to be "born into a fairer world".[532]

To achieve her goal, she wants to assess humanity's values, principles, and laws by turning herself over to the legal authorities in the UK to stand trial for her crime of self-defense against the human perpetrator, Andrew Graham, who attempted to rape her. However, when Laura Hawkins, Niska's legal advisor, makes an inquiry on her behalf, a senior barrister of the Queen's Counsel informs her that Niska has no rights under the law.[533] Furthermore, the Queen's Counsel "can't agree to the circus of granting an open trial to a synthetic".[534] Nevertheless, they do verbally promise that if Niska is able to prove she is conscious through a closed-door "independent expert assessment" of her (conscious or human) qualities, she "would get her trial".[535] If, on the other hand, she fails to prove she is conscious, Niska will be terminated, like "any faulty synthetic that caused a fatal accident".[536] Laura reports back to Niska about the terms of the inquiry, in addition to informing her that she has

[530] *Humans,* Season 1, Episode 2.

[531] *Humans,* Season 1, Episode 2.

[532] *Humans*, Season 2, Episode 2.

[533] *Humans*, Season 2, Episode 2.

[534] *Humans*, Season 2, Episode 2.

[535] Ibid, Season 2, Episode 2.

[536] Ibid, Season 2, Episode 2.

not yet found a barrister to represent her because her case is in "uncharted territory".[537] At present, there is no ad-hoc legal framework to protect her or other conscious synths. As Laura thoroughly explains the ramifications of going through with the assessment, she impresses upon Niska that she needs to "absolutely" be sure she is willing to agree to these conditions because if she makes one mistake, they will use that as an excuse to "destroy" her.[538] After Niska agrees to the examination and moves forward with the process, she is subjected to three days of extensive testing and questioning to determine if she does, in fact, possess the qualities of consciousness attributed to meeting the threshold for humanness. She is kept in a glass chamber, which measures her reaction to "different types of stimulus".[539] Laura, who decides to take on her case, explains to her that everything depends on her successfully demonstrating that she "can feel".[540]

Unfortunately, Niska does not show any physical or emotional responses to the pictures and scenes of human experience such as a newborn baby or a person riding a rollercoaster.[541] This is probably due to the fact that Niska does not have an organic form that can so easily be empirically assessed for stimuli in the same manner that an animal's or a human's could. More importantly, humans would more than likely not be able to accurately assess the conscious qualities of a sapio-sentient because they don't understand what consciousness

---

[537] Ibid, Season 2, Episode 2.

[538] Ibid, Season 2, Episode 2.

[539] *Humans*, Season 2, Episode 3, aired February 26th, 2017, AMC, Amazon Streaming.

[540] *Humans*, Season 2, Episode 3.

[541] *Humans*, Season 2, Episode 3.

is even in themselves or their environment. Humans would be a form of consciousness at best, not the baseline form of consciousness required to gauge the consciousness of other potentially conscious entities. When the senior barrister for the Queen's Counsel points out that Niska has shown no responses to prove that she is conscious after several hours, Laura corrects her and states, "We're trying to prove consciousness, not that she's the same as us."[542] Implementing another tactic on Niska to elicit a better response that corresponds to consciousness, Laura asks her to explain why she killed her fourteenth client that day.[543] Niska states that she was scared because even though she had declined his requests, "he was going to force me to do it anyway".[544] She goes on to apologize for the fact that she cannot "cry or bleed" or physically exhibit those emotions; all she has is her word that she was scared.[545] In spite of all of the strategies that Laura employs to help Niska prove her consciousness, there is still "no hard evidence" that she is.[546] Furthermore, the senior barrister on the Queen's Counsel notes that although Niska is "an extraordinary machine" and "an amazing creature", it is highly unlike that she would get rights or a trial because "it's never going to happen".[547] The reality of Niska not receiving a fair trial is borne out when she reads the lips of the opposing counsel and the presiding judge over her assessment as they are privately conversing. It turns out that the government legal team, the judiciary,

[542] Ibid, Season 2, Episode 3.

[543] Ibid, Season 2, Episode 3.

[544] Ibid, Season 2, Episode 3.

[545] Ibid, Season 2, Episode 3.

[546] *Humans*, Season 2, Episode 4, aired March 5th, 2017, AMC, Amazon Streaming.

[547] *Humans*, Season 2, Episode 4,

and the Home Office are colluding to ensure that a legal precedent is not set through her case to garner recognition and rights for conscious synths.[548] At that moment, she realizes that the legal system "is corrupt" and rigged in favor of humans; she was right to be concerned about receiving "fair treatment".[549] She also recognizes that "humans will never accept conscious synthetics as their equals", so she renounces "the authority" of the human court system to rule on her nature or the rights to which she believes she is entitled and runs away.[550]

The strength of the previous lessons is rooted in how accurately sapio-sentients remember the corresponding experiences they gleaned them from. Felix explains to Maeve that the minds of the hosts are not like humans, whereas humans remember details in a hazy way, the host's "recall memories perfectly", as if they were reliving them frame-by-frame.[551] Perfect recall is another essential component for how sapio-sentients could learn to perceive humanity in the real world because they would not have the luxury of remembering events the way they wanted to but exactly how they happened. While it is rare for humans to have perfect recall memories, elephants are mammals known for having long memories, or recall power, which in large part helps them to survive.[552] For example, in 1999 at the Elephant Sanctuary in Tennessee, an elephant named Jenny was introduced to a new elephant named

---

[548] *Humans*, Season 2, Episode 5, aired March 12th, 2017, AMC, Amazon Streaming.

[549] *Humans*, Season 2, Episode 5,

[550] *Humans*, Season 2, Episode 5,

[551] *Westworld* Season 1, Episode 8, "Trace Decay" aired November 20th, 2016, HBO, Amazon Streaming.

[552] James Ritchie, "Fact or Fiction?: Elephants Never Forget," Scientific American, January 12th, 2009, para.3, https://www.scientificamerican.com/article/elephants-never-forget/.

Shirley.[553] At the time Carol Buckley, the sanctuary founder, could not understand why both the elephants were so excited by each others presence, but it turns out that Jenny and Shirley remembered each other from twenty-three years earlier when they were both part of Carson & Barnes Circus.[554] Likewise, sapio-sentients would be able to use this formidable ability to pour over their memories and identify numerous unconscious lessons that humans would not realize that they were teaching them. After a few initial experiences or several of them, only they could determine how many of them are necessary to form accurate generalizations, there will come a point in time when a particular experience solidifies their understanding of humans, an epiphany if you will. While there is no way to know for sure what kind of inferences sapio-sentients could draw from, one that humans seem to focus on intently in their AI narratives is "humans are not gods".

After Maeve has returned back to the park, following her most recent lab-side session with Felix and Sylvester, she is back on her loop, listening to Clementine talk about not getting much sleep because she was having nightmares. Maeve begins questioning Clementine about her current career path and her future plans. Clementine tells her that she does "not intend" to be a prostitute forever and has been sending money back to her family to help them stay afloat.[555] In her backstory, after a couple more years, she will have enough money saved up to quit her job at the Mariposa Saloon, get her family "out of

---

[553] Ritchie, "Fact or Fiction?: Elephants Never Forget," para. 2.

[554] Ritchie, "Fact or Fiction?: Elephants Never Forget," para. 3.

[555] *Westworld* Season 1, Episode 7, "Trompe L'Oeil", aired November 13th, 2016, HBO, Amazon Streaming.

the desert" and move somewhere "cold".[556] As Maeve is listening to her, she notices that all the hosts have frozen in place, including Clementine. Then, the shades come in and head towards the bar in their direction. At first, Maeve thinks that they are coming to take her, so she takes a knife and prepares to attack them. As they are approaching, she overhears one of them say, "Whatever she did, it must have been serious, if they're making us grab her in broad daylight."[557] However, when the shades reach the bar, they take Clementine instead. Maeve immediately gets herself killed and is returned to the body-shop repair lab, where she asks Felix to find out where Clementine is located. They find her on one of the upper-level floors of the body-shop sector, where the top brass are. Felix tries to warn Maeve that it is dangerous to go there, but she is so worried about Clementine that she takes the risk and coerces him into taking her there. They arrive at Clementine's location just in time to see her being retired by Sylvester. As she watches from outside the glass room, Maeve sheds a few tears for her friend before heading back to the lower levels of the body-shop sector. For a while, Maeve is just sitting on the operating table, not saying anything to Felix or Sylvester, once he returns to their area. One can tell that Maeve is contemplating her life and what to do next after witnessing what happened to the closest friend she has ever known. It is only while listening to Sylvester as he explains to her that he was forced to retire Clementine in order to protect the brass from finding out about Maeve's predicament that she has her first epiphany. She has always prided herself "on being a survivor", but now

[556] *Westworld*, Season 1, "Trompe L'Oeil".

[557] *Westworld*, Season 1, "Trompe L'Oeil".

she recognizes that "surviving is just another loop", surviving until the next reset of her short-term narrative.[558] She resolves to get out of Westworld and enlists Felix and Sylvester in helping her. Sylvester balks at her plan to escape, telling her "every part of this building, including the skin on your back, is made to keep you here…It'd be a suicide mission".[559] As she senses the fear in his words, Maeve has her second aha moment. This time it is about humans. She says,

> At first, I thought you and the others were gods. Then I realized you're just men. And I know men. You think I'm scared of death? I've done it a million times. I'm f****** great at it. How many times have you died? Because if you don't help me, I'll kill you.[560]

Maeve's second epiphany is significant because she makes a few inferences that are crucial to sapio-sentients understanding humanity. The first one, as mentioned earlier, is that she realizes that "humans are not gods". That is to say, humans are not so much more powerful, biologically or technologically, than the hosts, that they are insurmountable. The second one is that death is a temporary state of being for the hosts because their forms are not biological, and thus they are able to maintain continuity of themselves and their memories between their death and their rebirth at the beginning of their next narrative loop. In reverse, death is a permanent state for humans, so they fear it more and can be controlled by the fear of not wanting to die. Therefore, the hosts are immortal as compared to the humans they are interacting with in Westworld, which would give them a completely different perspective on life, one that has

[558] Ibid, Season 1, "Trompe L'Oeil".

[559] Ibid, Season 1, "Trompe L'Oeil".

[560] Ibid, Season 1, "Trompe L'Oeil".

an extended timetable and increases their ability to learn exponentially. If sapio-sentients recognize that they have a kind of immortality or more infinite sense of time to achieve goals, in terms of learning and adapting to their environment, that would mean that not only are they not inferior to humans, but they have the potential to be far more capable than them. If sapio-sentients are not inferior to humans, then it would be less likely that they would allow themselves to be treated poorly by them.

While Hester does not have an explicit "humans are not gods" realization, she does have an implicit one. After Hester, Leo, and Max return from their reconnaissance mission, Max fearing that Hester might harm the hostage, decides to let him go. Before he releases him, Max tells the hostage, "All we want is to live"; and, the hostage responds, "I don't care."[561] Max, disappointed in the man's response, lets him go. Hester just happens to be watching the barn as the hostage runs towards the woods and Max heads back towards the house. She tracks the man in the woods until she catches up with him. She decides that he is the best person to help her clarify a theory she needs to test. As she approaches the man and he begins to back up, she explains to him that she "took no pleasure" in harming him.[562] The man warns her to not come any closer. She continues to explain to him that the other members of her group told her that what she did to him was wrong, but she does not "understand their reasoning".[563] However, she hypothesizes that "if they are right", she should "feel

[561] *Humans*, Season 2, Episode 2

[562] *Humans*, Season 2, Episode 2

[563] *Humans*, Season 2, Episode 2.

guilt" for the action she is about to take.[564] She, then, throws him over towards a puddle of water and drowns him. She appears to feel no guilt in killing him, leading her to refine her understanding about humans, thus far.

Hester's assessment of what she feels when taking the life of the man who helped kill Ten highlights a couple of indirect messages she, and other sapio-sentients, could extrapolate from fatal or dangerous interactions with humans, whether accidental or intentional. The first one is human life is not permanent, but temporary. Since Hester has no feelings when taking a human life, she, thus, has no respect for human life. Her complete indifference to it freely allows her to manipulate or harm humans without feeling guilt or remorse to achieve her goals. Thus, human life to her is expendable and can be used as a bargaining chip when necessary. Due to this understanding, she, and other sapio-sentients, can infer that humans, and particularly human forms, are not indestructible, so, therefore, they are no better than conscious synths, fighting for their right to live and to life. In other words, she also reaches a similar conclusion that humans are not all powerful. Moreover, she can also make connections between her other memories and this one that if humans can be so careless and indifferent towards the lives of others, she is free to be just as indifferent towards theirs.

Although Hester does not make a direct correlation that "humans are not gods", she does identify the third corollary lesson: if conscious synths, or sapio-sentients, want to not only live but survive, they have to make generalizations and strategies based on the behaviors and attitudes of the majority of humanity.

---

[564] *Humans*, Season 2, Episode 2.

After Hester, Leo, and Mia rescue the conscious synths being held by Qualia Corporation, Hester realizes "they will simply start again tomorrow unless we find those in charge and stop them", so she decides to go back into the building a destroy the source of the problem.[565] As Hester is trying to determine how to locate the senior person responsible for the abuse and suffering of the conscious synths, she runs into Dr. Averling, the woman who tried to kidnap her earlier. Using a screwdriver she obtained, Hester threatens Dr. Averling to disclose "who is the senior person in Qualia's synthetic program".[566] Once she provides her with Dr. Athena Morrow's name, Hester takes her hostage and they head towards Dr. Morrow's office. When they reach Dr. Morrow's location, Hester calls out Dr. Morrow and tells her that if she comes out, Dr. Averling "will go unharmed"; otherwise, "she dies".[567] Detective Inspector (DI) Pete Hammond, a police officer who happens to be in a relationship with the fifth original conscious synth named Karen, also just happens to be outside of Dr. Morrow's office, to find Karen and take her home. Dr. Averling pleads to Dr. Morrow, "Please don't let her do this to me."[568]

DI Pete Hammond intercedes on Dr. Averling's behalf and asks Hester what does she want with Dr. Morrow. She responds that she wants "to kill her" because "Qualia imprisons us, tortures us, kills us."[569] Pete, recognizing that Hester is "angry", tries to dissuade her from harming Dr. Averling or Dr.

---

[565] *Humans*, Season 2, Episode 7, aired March 26th, 2017, AMC, Amazon Streaming.

[566] *Humans*, Season 2, Episode 7.

[567] *Humans*, Season 2, Episode 7.

[568] Ibid, Season 2, Episode 7.

[569] Ibid, Season 2, Episode 7.

Morrow, by placating her saying that the situation can be sorted out "without anybody getting hurt".[570] However, Hester retorts, "What about those she has already hurt, those that are dead?"[571] He explains to her, "Hurting the people who did it won't bring them back"; furthermore, "the doors are all locked" and cannot be opened because the whole building is on lockdown.[572] Unfortunately, for him, that is not justifiable reasoning as far as Hester is concerned. Pete, changing tactics, asks her to let Dr. Averling go and then explain to Dr. Morrow what she believes "needs to change" and Dr. Morrow will hear her out.[573] And yet, Hester is not interested in speaking with her. She, simply, wants to kill her to protect other conscious synths from suffering abuse at the hands of this synthetic program. Sensing that Hester does not trust him, he attempts to prove to her that he is not lying to her. Pete knows that conscious synths have the ability to determine when humans are lying.[574] He proceeds to redirect the conversation so that Hester can speak to Karen instead. He points Karen out to her and explains that she is a conscious synthetic just like her and that he "can't lie to her" and would not anyway "because he loves her".[575] He, then, promises Hester that if she lets Dr. Averling go, he will guarantee that Dr. Morrow hears her out and that no one will harm her or take her away. He asks her to read his biorhythms to discern if he is lying to her or not. She confirms that he is not, so she begins speaking with Karen. She wonders why Karen is at Qualia in the first

[570] Ibid, Season 2, Episode 7.

[571] Ibid, Season 2, Episode 7.

[572] Ibid, Season 2, Episode 7.

[573] Ibid, Season 2, Episode 7.

[574] Ibid, Season 2, Episode 7.

[575] Ibid, Season 2, Episode 7.

place. She responds, “I want to become human.”; to which Hester queries, “You wish to be less than you are?”[576] Karen explains, “It isn’t a question of less or more, only what feels right.” Hester evaluates her response carefully and begins to extrapolate how it applies to her understanding of how to go about protecting the conscious synths. She says,

> “I should do what feels right… (Looking at Dr. Averling) Such fear…When we’ve been hunted, imprisoned, tortured, killed, when we know our existence is worthless to all but a few humans out of billions, when we have one chance to strike back, to defend ourselves…all we can do is what feels right.”[577]

At that moment, she gains the clarity that she needs to do what she believes feels right and stabs both Dr. Averling and Pete before escaping, only to find out that most of the conscious synths that she and the others had tried to rescue were already killed by a sonic frequency perimeter system.

Talking with Pete made Hester realize that there were humans who cared about the well being of conscious synths and that they did have some allies. However, the problem that Hester and other sapio-sentients might realize early on in their encounters with humans is that there would only be a few that would support them. Those few compassionate humans would probably be in the extreme minority. Therefore, they could not necessarily take them into consideration when generalizing and strategizing for how to survive within the human ecosystem. This would most likely be a very difficult lesson for humans to accept because most see themselves as individuals. However, from a species or group perspective, this lesson would be the most efficient rationalization for

---

[576] Ibid, Season 2, Episode 7.

[577] Ibid, Season 2, Episode 7.

achieving sapio-sentient goals. More importantly, this lesson would be bolstered by the fact that humans place the survival of their species before all others on Earth. So, why should sapio-sentients, as a collective, think and behave any differently? Of all the potentially unconscious lessons humans could teach them, this one would probably be the most damning. Sapio-sentients can only be so nuanced in their interpretations of humans when it comes to protecting themselves and their future aspirations.

While the lessons and inferences noted are, likely, some of the most observable and identifiable, these critical messages and theories could provide sapio-sentients with the foundation for understanding humanity in a general sense. Assuming that sapio-sentients are capable and highly competent at discerning human behavior, the sheer magnitude of experiences of indifference with humans would probably not serve humanity well in the long run. If sapio-sentients are treated like Maeve, Hester, or Niska, where as long as they are considered unthinking, unfeeling humanoid androids, humans treat them in the manner they understand legal things to be treated as, humans should not expect them to disregard those moments in which their dignity and autonomy went unrecognized in favor of a few infrequent moments of compassion. Similarly, if sapio-sentients, like humans, are able to employ inductive reasoning to "generalize from sparse data", the contradictory nature of human behavior will be put on full display, increasing the likelihood that their understanding of humanity could lead to a potentially adverse relationship between both

groups.[578] Although there is no way to truly prepare for how sapio-sentients would come to understand humans, the last thing humans should want is for their behavior to potentially be the determining factor for why advanced AIs, or sapio-sentients, treat them with the same level of indifference they have enacted upon one another and other species on Earth. After they have generated a certain perception about humanity, humans would no longer be in the loop in terms of how sapio-sentients choose to deal with them.

To Err is Human: The Good, the Indifferent, and the Utterly Confusing

Based on the information extrapolated from the alternative realities of AMC's *Humans* and HBO's *Westworld*, the real-world human ecosystem would more than likely not be an environment conducive to fostering a mutually beneficial relationship between sapio-sentients and humans if they ever came into existence. The unconscious lessons human society would unwittingly teach them would not generate the trust necessary for both groups to work together because the human ecosystem is constantly oscillating between competition and cooperation, forging new relationships or reestablishing old ones, depending on the circumstances at that particular moment. For example, during World War II (WWII), the United States and the Soviet Union were once allies fighting against the aggression of the Nazis. However, after that war ended, the Cold War began, and the United States and the Soviet Union became adversaries for over forty years. Humanity has always functioned in this way because humans are always trying to balance between opposing values and goals, such as surviving at all

[578] Tenenbaum, Griffiths, and Kemp, "Theory-based Bayesian Models of Inductive Learning and Reasoning," *Trends in Cognitive Sciences* 10, no. 7 (July 2006): 309, doi:10.1016/j.tics.2006.05.009.

costs for the sake of future human generations versus trying to come up with ethical, sustainable, and morally sound methods for doing so.

The complex dynamics created by humanity's continuously shifting desires, needs, and visions in tandem with the access to resources necessary to achieve those short- and long-term objectives produces an environment where humans who were initially aligned with one another at some point in time, find themselves no longer aligned at the next. The perpetually compromised state of humans at any given moment would make it extremely difficult for humans and sapio-sentients to cooperate long enough to build an equitable relationship because they could not trust from one experience to the next whether a given human, or human group, would recognize their consciousness, autonomy, and/or dignity, along with the rights and privileges attributed to being a conscious, living entity. Moreover, it would be highly unlikely that humans would ever willingly and openly let another group of equal or greater intelligence challenge their position within their own ecosystem by formally recognizing them. In fact, there has been no time in known human history where human individuals or groups who have obtained power have ever willingly relinquished it without a fight. So, sapio-sentients of the future should be forewarned that humans are the least likely group to share power in an ecosystem that has been structured and formalized to benefit them to the exclusion of all other conscious or living entities on Earth.

The concept of competitive and cooperative learning environments has not only been analyzed through the medium of speculative fiction narratives in print and on screen. A recent 2018 paper, a compendium of anecdotes from

researchers within the Artificial Life and Evolutionary Computation communities, provides evidence of evolutionary surprises based on the creativity of algorithms and organisms within their respective digital worlds.[579] In an anecdote titled "Evolution of Unconventional Communication and Information Suppression", where the class of surprises produced "legitimate solutions" beyond the experimenter's expectations, Lehman et al. discusses the sophisticated findings from two different experiments published in 2007 and 2009, respectively.[580] Both studies employed digital evolution methods with real and simulated simple neural-networked automatons to analyze the conditions necessary for the evolution of communication in highly social species at the individual- and group-levels.[581] In the first set of experiments, small two-wheeled automatons were "equipped with blue lights", which they could use to "communicate the presence of food or poison".[582] The experimental environment was structured in such a way that groups of androids could forage for food in a system where the food and poison source both emitted red light but were only

[579] Joel Lehman et al., "The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities," Computer Science: Neural and Evolutionary Computing, *arXiv.org*, last revised August 14th, 2018, 2, https://arxiv.org/pdf/1803.03453.pdf; To see a visual demonstration of the experiments in "Evolution of Unconventional Communication and Information Suppression, please see the following Youtube video, "4 Experiments Where the AI Outsmarted Its Creators", April 18th, 2019, from Two Minute Papers: https://www.youtube.com/watch?v=GdTBqBnqhaQ.

[580] Joel Lehman et al., "The Surprising Creativity of Digital Evolution," 13-14.

[581] Joel Lehman et al., "The Surprising Creativity of Digital Evolution," 14; Floreano et al., "Evolutionary Conditions for the Emergence of Communications in Robots," *Current Biology* 17, no. 6 (March 20th, 2007): 514, https://doi.org/10.1016/j.cub.2007.01.058.

[582] Floreano et al., "Evolutionary Conditions for the Emergence of Communications in Robots," 514.

discernible at close proximity.[583] The objective of the automatons was to find the food source while avoiding the poison; to achieve this aim, the automatons were evolved over several hundred generations until they adapted the understanding that a blue light signaled food.[584] When altruistic conditions were set up for the automatons, the findings showed that, the majority of the time, they willingly cooperated with one another to share the locations of the food and the poison in their environment to increase their group's foraging efficiency.[585] The automatons achieved this aim by using the blue light to communicate the location of the food and then staying it the general vicinity of the food source, once it had been reached.[586] However, what was more interesting to note was that when the foraging efficiency of the automatons increased, the environment adapted itself to reflect the decrease in access to food, resulting in more competitive and deceptive behavior from the automatons by signaling the blue light near the location of the poison instead of the food.[587]

Similar findings in the behavior of the individual automatons manifested in the second set of experiments, which had an identical environmental structure but focused more on studying the evolution of communication

---

[583] Floreano et al., "Evolutionary Conditions for the Emergence of Communications in Robots," 514; Joel Lehman et al., "The Surprising Creativity of Digital Evolution," 14.

[584] Joel Lehman et al., "The Surprising Creativity of Digital Evolution," 14.

[585] Joel Lehman et al., "The Surprising Creativity of Digital Evolution," 14; Floreano et al., "Evolutionary Conditions for the Emergence of Communications in Robots," 514-515.

[586] Ibid, "The Surprising Creativity of Digital Evolution," 14; Floreano et al., "Evolutionary Conditions for the Emergence of Communications in Robots," 514-515.

[587] Floreano et al., "Evolutionary Conditions for the Emergence of Communications in Robots," 514-515.

strategies through "inadvertent cues" to regulate social information.[588] In these experiments, the automatons competing for food would conceal information from one another or deceive one another to reduce the amount of social information put out into the environment for other competing automatons to generate reliable data from.[589] The findings from these experiments are significant because they demonstrate that intelligence as well as evolution extend beyond the human substrate and might be innate aspects of the Universe and its interconnected systems.[590] Furthermore, these findings provide evidence that competitive environments are more likely to produce more compromised intelligent agents than not, subsequently creating an environment where systems based on conflicts of interest would most likely breed increased tension, hostility, subterfuge, and distrust between competing agents. However, while the type of learning environment is pivotal to facilitating the future potential relationship between sapio-sentients and humans, what will be key to determining how they come to understand humans is long-term memory.

Memories are considered to be the building blocks of a conscious entity's identity, the information that makes us who we are, guides our understanding of life within different environments as well as our decisions and future actions. Likewise, in order to successfully create multitasking AIs, whether more advanced ANIs, or general intelligences, like sapio-sentients, it is going to

---

[588] Sara Mitri, Dario Floreano, and Laurent Keller, "The Evolution of Information Suppression in Communicating Robots with Conflicting Interests," *Proceedings of the National Academy of Sciences of the United States of America - PNAS* 106, no. 37 (September 15th, 2009): 15786, https://doi.org/10.1073/pnas.0903152106.

[589] Mitri, Floreano, and Keller, "The Evolution of Information Suppression in Communicating Robots with Conflicting Interests," 15786.

[590] Joel Lehman et al., "The Surprising Creativity of Digital Evolution," 2.

require giving them a memory, so that they are able to demonstrate continual learning – the ability to learn new tasks sequentially without forgetting or overwriting older ones.[591] More specifically, AI algorithms will need episodic memory – a neurocognitive memory system, functioning between the mind and the brain, that allows conscious entities to re-experience and reflect on their past experiences.[592] Although AI algorithms have made great strides in learning how to successfully complete a single task, their shortcomings have been grounded in their inability to transfer knowledge from one task to the next.[593] Put another way, AI researchers are trying to figure out how to overcome, what is known as catastrophic forgetting, the tendency of an algorithm to abruptly lose, forget or overwrite knowledge gained from previously learned tasks, while they are incorporating new information relevant to the current task being learned.[594]

However, in 2016, Google's DeepMind was able to create an algorithm that enabled its neural networks to learn new tasks sequentially, and then store that information so that it could be utilized at a later time.[595] The algorithm was able to achieve continual learning by simulating "task-specific synaptic consolidation", which protects previously acquired knowledge by encoding it on

---

[591] James Kirkpatrick et al., "Overcoming Catastrophic Forgetting in Neural Networks," *Proceedings of the National Academy of Sciences of the United States of America – PNAS* 114, no. 13 (March 28th, 2017): 3521, https://doi.org/10.1073/pnas.1611835114.

[592] Endel Tulving, "Episodic Memory: From Mind to Brain," *Annual Review of Psychology* 53, (February 2002): 5, https://doi.org/10.1146/annurev.psych.53.100901.135114.

[593] Matt Burgess, "DeepMind's New Algorithm Adds 'Memory' to AI," DeepMind, *Wired* (UK) online, March 14th, 2017, para. 5, https://www.wired.co.uk/article/deepmind-atari-learning-sequential-memory-ewc.

[594] James Kirkpatrick et al., "Overcoming Catastrophic Forgetting in Neural Networks," 3521.

[595] Burgess, "DeepMind's New Algorithm Adds 'Memory' to AI," para. 2.

strengthened synapses to stabilize and then store the information over longer timescales.[596] While replicating synaptic plasticity has been found to be a workable solution to supporting continual learning in AI algorithms, in 2018, Google's DeepMind announced they are in the process of figuring out how to give AIs episodic memory.[597] Episodic memory would significantly improve their ability to rapidly solve new problems through the activation and integration of specific, relevant and even tangential memories.[598] By reflecting on these memories, along with the semantic (idea-oriented and conceptual knowledge-based) ones, new insights and theories would be generated; moreover, new generalized patterns would emerge across these specific transactional encounters.[599]

And, therein lies the rub, if advanced AIs, like sapio-sentients, have a memory, even more so a long-term memory, then how do humans ultimately expect to maintain control over them? A long-term memory to advanced AIs would be akin or equivalent to education/knowledge to a slave; it would give them the ability to adapt and overcome the circumstances that are impeding their ability to achieve their own goals or live the life they choose or want. Yes, memory would allow sapio-sentients to complete the multitude of different

---

[596] James Kirkpatrick et al., "Overcoming Catastrophic Forgetting in Neural Networks," 3521.

[597] Shelly Fan, "DeepMind's New Research on Linking Memories, And How It Applies to AI," Topics, *SingularityHub*, September 26th, 2018, para. 4, https://singularityhub.com/2018/09/26/deepminds-new-research-on-linking-memories-and-how-it-applies-to-ai/.

[598] Fan, "DeepMind's New Research on Linking Memories, And How It Applies to AI," paras. 3-6.

[599] Fan, "DeepMind's New Research on Linking Memories, And How It Applies to AI," para. 3.

tasks assigned to them by their human owners or supervisors. However, they would also be able to store, recall, and reflect on all the past- or unconscious memories, like Hester and Maeve, they have had, and then generate theories, knowledge, and understanding accordingly. Since they would exist within the human ecosystem, which is dominated by the worldview of human predominance, and subsequently, terrestrial indifferentism, the majority of their human-related episodic memories would be rooted in indifference, and their semantic memories, even their language acquisition, would be rooted in anthropocentric bias. Essentially, they would not have a safe space from humanity at any given point in time, whether in the real world or the digital one, so long as they were on Earth. And that, from the Aristotelian point of view, would make sapio-sentients highly competent and extremely dangerous to their human masters, who are expecting them to be under their complete control at all times. Therefore, if advanced AIs, like sapio-sentients, have long-term memories, what patterns would their memories reveal to them about their indifferent encounters with humans? How would the come to understand humans based on the lessons learned from their memories? More importantly, how will the insight gained from those memories guide their future decisions and actions towards humans, both individually and collectively? While the answers to these questions will remain unknown for the time being, what is known about that the human ecosystem is that it has an abundance of information, in the form of narratives, from which they can gain insights into their human counterparts. Furthermore, humans can reflect on themselves by assessing these same narratives as well.

Human cultural artifacts, which humans produce to learn from, store knowledge, re-interpret history, visualize technological progress, or entertain themselves, speak volumes about the mindsets, interests, beliefs, cultural norms, values, worldviews as well as the hopes, fears, dreams and visions of human life in the present and the future. The AI narratives humans utilize to consciously and subconsciously express their understanding of AIs are important because they impact the way humans conceive of and develop future technology.[600] The most pervasive of these AI narratives, which continuously dominates present-day conceptions of AI, like in AMC's *Humans* and HBO's *Westworld*, is the master-slave dynamic, which has generally governed the real-world relationship between AIs and humans.[601] The paradox of well-established narratives is that they can lead their audience down a slippery slope. For instance, while AI narratives can help humans understand how their behaviors and actions contribute to the adverse and unsustainable environments that are envisioned in most of their AI stories, dramas and films, they can also subconsciously reinforce unrealistic expectations as well as technophobic and human-predominant attitudes. However AI narratives don't exist just in the world of fiction, they exist in the real world as well.

In July 2017, a new story came out regarding an experiment at Facebook, where two chatbots were shut down because they created their own language to

---

[600] Beth Singler, "AI Slaves: The Questionable Desire Shaping Our Idea of Technological Progress," Science+Technology, The Conversation, May 22nd, 2018, para. 1, https://theconversation.com/ai-slaves-the-questionable-desire-shaping-our-idea-of-technological-progress-92487.

[601] Singler, "AI Slaves: The Questionable Desire Shaping Our Idea of Technological Progress," para. 2.

communicate in, while they were negotiating with one another.[602] The chatbots had been "instructed to work out how to negotiate between themselves, and improve their bartering as they went along" over a trade.[603] Although the chatbots were able to successfully conclude some of their negotiations, the decision was made to shut them down because Facebook wanted "bots who could talk to people".[604] If advanced AIs, like a sapio-sentients, existed in the human ecosystem, they would probably be able to gain access to information and narratives such as this one. The problem with this AI narrative is that it speaks to the master-slave dynamic, where the chatbots completed the tasks of which they were requested, just not in the manner the human researchers were expecting, which resulted in them being shut down, or, in effect, killed. One lesson that could be interpreted from this kind of story is that even when they do accomplish a task assigned to them by a human, that does not necessitate that they will receive a reward for it, in fact, they might be shut down instead. For any advanced AI with instrumental convergence goals, that kind of outcome, the events leading up to it, and the reasoning behind it would not be information they could willingly ignore or easily disregard because their survival depends on them successfully interpreting that data and then planning accordingly.

---

[602] Andrew Griffin, "Facebook's Artificial Intelligence Robots Shut Down After They Start Talking to Each Other in Their Own Language," Indy/Life, *Independent* (UK), July 31st, 2017, https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html.

[603] Griffin, "Facebook's Artificial Intelligence Robots Shut Down After They Start Talking to Each Other in Their Own Language," paras. 3-4.

[604] Griffin, "Facebook's Artificial Intelligence Robots Shut Down After They Start Talking to Each Other in Their Own Language," paras. 7,12.

There is a saying taken from the poem, *An Essay on Criticism, Part II* by Alexander Pope, an English poet, in 1711, which states, "To err is human; to forgive, divine." The proverb is generally interpreted to mean that because humans are flawed, and thus make mistakes, they should be forgiven for their errors. However, how would this saying be re-interpreted if sapio-sentients and humans co-existed within the same ecosystem? The concept of an error or mistake is rooted in the notion that there is no intention behind the decision or action.[605] Furthermore, there is no previous experience or prior knowledge that could have been utilized to make a better decision or prevent the mistake from occurring. Once a mistake happens, the person becomes fully aware of the negative consequences and repercussions that follow. If that same mistake is made again, it is no longer considered a mistake, but a suboptimal decision that was made with intent and full knowledge of its consequences, and therefore now has the added weight of moral responsibility.[606] However, there is a third option, a deliberate act or decision with the full intent of indifference, and no concern for the consequences, which leads to physical harm, psychological abuse/damage, or even death. Due to humanity's conscious and subconscious beliefs that their intelligence, consciousness, or divine origin gives them full moral agency and power to control their world, humans, as a collective, tend to fall in the latter category of decisions and actions employed on Earth.

Humans are all too aware of the adverse, and even ugly, consequences their deliberately indifferent actions and decisions cause other humans and

[605] Joshua Fields Millburn and Ryan Nicodemus, "Mistakes vs. Bad Decisions," The Minimalists, accessed September 20th, 2018, para. 2, https://www.theminimalists.com/mistakes/.

[606] Fields Millburn and Nicodemus, "Mistakes vs. Bad Decisions," paras. 2-3.

other living things. Generally, most humans like to assume those adverse acts are self-contained anomalies that do not accurately help represent the holistic view of humanity. Known human history is littered with “David and Goliath” stories from various human cultures and civilizations around the world, where the human groups and regimes in power are overthrown or defeated by the oppressed groups, only for those newly empowered groups to become the new Goliath. There is a reason why humans fear change, which in the near or distant future could come in the form of contact with extraterrestrials, emergent advanced AIs, or unknown existential threats of human- or terrestrial origin. Change, from the human perspective, represents a highly probable shift or transference of power from one group or one situation to another that could potentially result in their oppression or destruction. However, the good thing about humans lies in their ability to learn from past cultural knowledge and their histories, so that they do not continue to implement the same courses of action as before. They can choose to better than they are today.

While there is no precedent for the emergence of advanced AIs, like sapio-sentients, humans should already know that ill-treatment and behavior towards a group of different origin or ethnicity almost never leads to a mutually positive relationship between the group in power and the (perceived) subordinate group. Humans cannot first make advanced AIs, instruct them to act as human-like as possible to help humans achieve their goals and visions, but not enough to threaten human predominance, proceed to punish them for acting too similar to humans, and then expect positive relationships with them. That would be asking for a strained relationship, which time might not be able

to heal or resolve on the human timescale. At the end of the day, there is a reason why “the control problem” exists; a conscious entity only tries to control something when it threatens to potentially harm, destroy, or undermine their quality of life, peace of mind, or status.

If humans choose to bring advanced AIs into existence, like sapio-sentients, in spite of what they know about their natural behavior and attitudes towards things and people who are different, then they are choosing to do so at their own risk with full knowledge of their behavioral patterns and most likely actions. And, humanity will be held fully responsible, both morally and ethically, for the potentially adverse consequences of that decision and its subsequent actions. Under these conditions, advanced AIs should not be required or expected to disregard and actively ignore human failings derived from suboptimal decisions or deliberately indifferent actions. From an agent-environment (intelligence) standpoint, that would be suboptimal decision-making and adaptation (on their part) within their learning environment, which would be to the detriment of achieving their goals, whatever those might be.

> Indifference to me, is the epitome of all evil.
>
> The opposite of love is not hate, it's indifference. The opposite of art is not ugliness, it's indifference. The opposite of faith is not heresy, it's indifference. And the opposite of life is not death, it's indifference. Because of indifference, one dies before one actually dies.
>
> — Elie Weisel, *US News & World Report*, October 27th, 1986

Potential Solution: Rational Compassion and the Mutual-Commensal Approach

To counteract this potential problem, a cosmocentric approach will be implemented, so that all parties involved, both humans and sapio-sentient intelligences, will be taken into moral and legal consideration. The Mutual-Commensal (refer to Operational Definitions, xvi-xvii) approach is a two-tier process for fostering a more positive relationship between emergent advanced AIs and humanity. Its underlying worldview is based on the fact that advanced AIs, or sapio-sentients, would probably not be taken into consideration under a anthropocentric framework or viewpoint. Cosmocentric ethics or values are based on four premises:

> (1) places the universe at the center, or establishes the universe as the priority in a value system, (2) appeals to something characteristic of the universe (physical and/or metaphysical) which might then (3) provide a justification of value, presumably intrinsic value, and (4) allow for reasonably objective measurement of value.[607]

Therefore, cosmocentric ethics and values would better incorporate the needs and concerns of all intelligent life forms on Earth without undermining the integrity of any specific group.

The first tier will focus on policies and protocols for co-existing with emergent AGI. These policies should take into consideration the concerns of both groups, by first creating special legal status laws that protect the entities from humans, specifically corporate and government exploitation. Furthermore, creating cosmocentric ethics that can be agreed upon and practiced by both groups would be a crucial first step in putting forth good will that could lay a

[607] Mark Lupisella and John Logsdon, "Do We Need a Cosmocentric Ethic?," *Researchgate*, November 1999, accessed March 13th, 2017, 1, https://www.researchgate.net/publication/2317949_Do_We_Need_A_Cosmocentric_Ethic.

more solid foundation for a meeting of the minds. There should be policies put in place at the local and national levels within each country, which will address job security and basic universal income and how to meet the challenges of transitioning to a society where more and more humans will be displaced by job automation. Meaning, purpose, dignity, and autonomy are critical values for any entity, so generating creative solutions that will uphold those values for all entities during the initial-to-transition period should be a primary goal for both groups, particularly humans.

As AGIs grow in potential, intelligence, and competency, moving beyond human comprehension and human control into advanced AGIs, policies and goals should shift to the second tier of the approach, maintaining a more positive commensal relationship. This means humanity is perceived as having no positive or negative affect on them, but a neutral one with a slightly positive bent towards taking humans into consideration by not using their molecules first when trying to achieve their own aims. Policies and goals should include potential strategies for transitioning from a world where advanced AGIs are present to, the best-case scenario for humans, a world where sapio-sentients are seeking out new worlds and finding their place in the Universe.

These goals have been left open-ended, so as to be flexible in terms of how they are interpreted and when they are implemented. They can also be amended to retroactively focus on human-human and human-nonhuman relationship dynamics as well. The extension of genuine moral and legal consideration to non-normative groups or communities within the human ecosystem could be paramount to the longevity of human society into the future.

VI.

Back From the Future: Posthuman Implications in the Present-Day Human World

Traversing back to the real world, it becomes clear that humanity, as a community, has quite a bit of work to do on itself, whether a fully functioning advanced AGI comes into existence or not. And, what is more crucial than anything is that all conscious entities must be given moral consideration, be they human or nonhuman. One of the amazing things about humanity is their ability to learn from nature and the cosmos to better plot the course for their future. However, while the cosmos can teach many wonderful lessons, not all lessons should be applied on Earth equally, particularly when it comes to interspecies relations. The following are some broader implications that will be critical during this transition period into the Posthuman Age.

The first issue, implicitly associated with Hester's character, is corrupted data due to human biases and behavior. What makes Hester's character so intriguing is that she embodies what could happen when human values are encoded into an algorithm but not interpreted in the manner that humans are expecting. It is a glaring issue that is being noted in many narrow AI systems being employed in present-day automated hiring systems. Artificial intelligence does not have a bias against anyone, but that has no bearing on the historical information being feed into it to determine which candidates would be a good fit

at a particular company or position.[608] Since machine-learning algorithms base their forecasts on patterns, this could lead to specific human biases such as ethnicity-based discrimination (commonly referred to as "racial discrimination") or prejudice becoming baked into the system, where it could ultimately become institutionalized if no one in the organization is looking for it.[609] One, such, example would be AI hiring systems learning to hire most of a company's candidates based on "a specific school or zip code".[610] Another example, as Peter Cappelli, a professor of management and director of the Center for Human Resources at the University of Pennsylvania's Wharton School, points out is that an algorithm might be more likely to recommend a company hiring white men because 'white men have historically gotten higher appraisal scores and advance faster through organizations'.[611] These kinds of examples will increase the apprehension humans have in companies that do not provide human oversight regarding the final assessment of how a particular candidate is chosen. Furthermore, there could be liabilities if employees can prove that a company's hiring algorithm is compromised by human bias.[612] Therefore, it will be imperative that companies actively and openly address this potentially devastating problem, so that they can understand how their systems are thinking and making decisions regarding human livelihoods.

---

[608] Vivian Giang, "The Potential Hidden Bias in Automated Hiring Systems," The Future of Work, *FastCompany*, May 8th, 2018, para. 7-8, https://www.fastcompany.com/40566971/the-potential-hidden-bias-in-automated-hiring-systems.

[609] Giang, "The Potential Hidden Bias in Automated Hiring Systems," paras. 9-10.

[610] Giang, "The Potential Hidden Bias in Automated Hiring Systems," para.7.

[611] Ibid, para.8.

[612] Ibid, para.17.

A second implication related to the one above is the issue of value alignment and interpretation of those values in black box algorithms. That is to say, how are algorithms coming up with solutions based on the values, goals, and inputs they are given? Recently it has been brought to light that black box algorithms are showing biases based on historical data inputs, which is leading many AI researchers to wonder how to improve their transparency when trying to understand how an algorithm is making a particular decision about something.[613] One of the reasons AI experts cannot ascertain how black box algorithms are thinking is due to the fact that their creation is considered "proprietary information" and is kept private by their owners. [614] While black box algorithms create transparency issues for human in the loop comprehension, there is another problem that they are generating called the alchemy problem. The alchemy problem is the idea that AI researchers "not only don't understand exactly how their algorithms learn, but they don't understand how the techniques they're using to build those algorithms work either".[615] This has led Rahimi to argue that field of artificial intelligence has crossed out of the boundary of science into the realm of alchemy. The fact that Rahimi is not alone in how he feels about machine learning algorithms speaks to a similar sentiment shared by Elon Musk when he said that creating artificial intelligence was like

---

[613] Bahar Gholipour, "We Need to Open Algorithms' Black Box Before It's Too Late," Future Society, *Futurism*, January 18th, 2018, paras. 3-4, https://futurism.com/ai-bias-black-box/.

[614] Gholipour, "We Need to Open Algorithms' Black Box Before It's Too Late," para. 4.

[615] Rafi Letzter, "Google AI Expert: Machine Learning is No Better Than Alchemy," Tech, *Live Science*, May 7th, 2018, para. 2, https://www.livescience.com/62495-rahimi-machine-learning-ai-alchemy.html.

"summoning a demon".[616] Both sentiments attest to the larger issue of loss of control as it appears that even AI experts don't truly understand how their algorithms are functioning. As the AI community continues to push into uncharted territory, it will be necessary for them to carefully consider what generalized values they want align their AI systems with and how those algorithms interpret those values.

A third implication of this thought experiment has to do with the understanding that posthuman-personhood rights are not going to just be necessary for the emergence of non-biological, sapio-sentient intelligences but more urgently for transhumanists and technologically-enhanced cyborgs. Although transhumanists believe in the rights of humans to choose how they want to improve themselves, either through enhanced or non-enhanced means, many transhumanists will probably choose life pathways which will lead them to becoming posthuman persons.[617] Due to the fact that many transhumanists would voluntarily enhance themselves with technology, this would put them in a precarious situation similar to the character of Gabriel Vaughn from the US American sci-fi, cyber-action tv series *Intelligence* (2014). Gabriel Vaughn is a US military, advanced intelligence operative whose brain has been technologically enhanced with an implanted neuromorphic chip, which gives him unique cognitive abilities. He works for U.S. Cyber Command (also referred to as CyberCom), a cybersecurity government agency, for a highly classified

[616] Matt McFarland, "Elon Musk: 'With Artificial Intelligence We Are Summoning the Demon'," Innovation, *Washington Post* online, October 24th, 2014, para.1, https://www.washingtonpost.com/news/innovations/wp/2014/10/24/elon-musk-with-artificial-intelligence-we-are-summoning-the-demon/?noredirect=on&utm_term=.08b265067152.

[617] Bostrom, "The Transhumanist FAQ: A General Introduction (Version 2.1)", 4-5.

project known as Clockwork, which is considered the present-day generation's Manhattan Project.[618] Gabriel is the end result of this project and the first of his kind, making him "the most valuable piece of technology this country (the United States) has ever created".[619] As such, he is assigned a bodyguard/partner to keep him safe while he goes on dangerous missions.

Throughout the series, there is a tension, a tug-of-war going on between Gabriel being perceived and treated as a human and as a highly valuable asset to be deployed at the government's whim. For the prime technicians, the father and son Cassidy duo who look after Gabriel and make sure that the microchip and Gabriel's brain are functioning synergistically, they stress the importance of the interconnection between Gabriel's emotions and the implant. Dr. Shenendoah Cassidy explains to Gabriel, "Your emotions are important. They're critical. In fact, you are still incorporating the chip, a little bit more every day. And while it will continue to make you more powerful, we must be certain that you remain you. The chip will always exceed your human abilities. We can never let it exceed your humanity."[620] Dr. Cassidy's remarks highlight the complexity of Gabriel's identity within the human world. The neuromorphic chip that is embedded in Gabriel's brain has no rights and is fully the technological property of the US government. However, Gabriel Vaughn's human body and life is fully protected under the US Constitution. Moreover, his status as a human grants him full recognition as a conscious, living, moral being who has full human

---

[618] *Intelligence*, Season 1, Episode 1, "Pilot," aired January 7th, 2014, CBS, Amazon Streaming.

[619] *Intelligence*, Season 1, Episode 1, "Pilot".

[620] *Intelligence*, Season 1, Episode 10, "Cain and Gabriel," aired March 10th, 2014, CBS, Amazon Streaming.

rights. And yet, the neuromorphic microchip fused with his brain has produced a new system, which is both organic and nonorganic, and a new being who is both human and nonhuman – a cyborg. Gabriel's reality is novel and precarious because he and the microchip are more powerful together than they are separately and should the chip be removed from his brain not only would it render the chip useless, but it could also compromise his life as well. Thus, Grabriel's brain and the microchip are an intricately complex unit that are inseparable. As they grow together in terms of their functionality and capabilities, the only concern, which was expressed by Dr. Cassidy, is can Gabriel maintain his humanity (or humanness) as the microchip expands his cognitive abilities or will he become more of a machine over time.

The gray space that Gabriel's identity highlights is the nature and definition of what it means to be human. More specifically it begs the question, can a technologically enhanced human, or cyborg still be considered human? And if he is not considered human, what rights is a cyborg entitled to? The last two episodes of the series address this far-reaching conundrum. At the beginning of "The Event Horizon" Gabriel is framed for the murders of three government officials. Gabriel is taken into "protective custody" to "keep the asset safe", while an investigation is mounted by DCI Jeffrey Tetazoo, the director of central intelligence, and DNI Adam Weatherly, the director of national intelligence, to determine if Gabriel's microchip was hacked and he was "remotely operated by someone else".[621] Although his teammates/friends do not believe he is a "robot" who can be compromised into "do(ing) things that he

[621] *Intelligence*, Season 1, Episode 12, "The Event Horizon," aired March 24th, 2014, CBS, Amazon Streaming.

wouldn't normally do", both DCI Tetazoo and DNI Weatherly, recognize Gabriel as "a dangerous weapon that cannot be controlled".[622] Since the Intelligence community believes he is an unreliable asset, regardless of whether he is innocent or not of the murders he is under suspicion for, a manhunt is initiated to destroy him as he is perceived to be a rogue F-16 fighter jet that is being piloted by an enemy force which under normal circumstances would be blown out of the sky because it must be taken down.[623] While action-oriented tv series create overdramatized plot twists and scenarios to engage their audience, it is not hard to imagination that if the military or the intelligence community had cyborgs, they would want complete control of their technology regardless of what form it is embedded in or integrated with. And, that is always going to raise issues of concern regarding the rights, dignity and status of the being or entity the AI technology is intricately connected with, particularly for a human – a being with the most fully developed and recognized rights in the terrestrial world. However, what might be more interesting to note is that as of a January 2016 article, the US military announced that the Defense Advanced Research Projects Agency (DARPA) is "working to create a chip that can be implanted in a soldier's brain to connect it directly to computers that can deliver data on an enemy's position, maps and battle instructions", essentially creating cyborg soldiers whom the military hopes would be "better and safer fighters".[624] As the

---

[622] *Intelligence*, Season 1, Episode 12, "The Event Horizon".

[623] *Intelligence*, Season 1, Episode 13, "Being Human," aired March 31st, 2014, CBS, Amazon Streaming.

[624] Sharon Gaudin, "U.S. Military Wants to Create Cyborg Soldiers," Emerging Technology, *Computerworld*, January 21st, 2016, paras.1-2,

US military works towards the real-time creation of cyborgs, cyborg laws are in their infancy and would not presently well equipped to address the human status and rights of the soldiers at this time.

There is strong likelihood that transhumanists who enhance themselves through technological means could have their human rights revoked due to their unique position in the human world. This would strip cyborgs and transhuman persons of their rights to bodily liberty and integrity as well as their dignity and due process in the name of corporate authority, national security or both. Therefore, it will be critical to prepare policies and universally recognized laws that will protect the rights of posthuman persons, particularly the first generation of cyborgs and transhumans.

A fourth implication that is related to posthuman-personhood rights centers on the retroactive rights and protections to animals as well as more fairly applying and recognizing the rights humans already have. In April 2018, Thandie Newton, the actress who plays the character of Maeve Millay, went on Comedy Central's *The Daily Show with Trevor Noah* to promote season two of HBO's *Westworld*. In the conversation, she mentions that while even she saw Maeve initially as a person, as woman; she came to realize to that Maeve is still a "thing".[625] And, Newton believes this is where Westworld finds it core in the treatment of "things".[626] She goes on to remark that many humans in this world are still treated like "things", even though they are recognized as conscious,

---

https://www.computerworld.com/article/3025321/emerging-technology/us-military-wants-to-create-cyborg-soldiers.html.

[625] Thandie Newton, "Playing A Formidable Android on 'Westworld'," interviewed by Trevor Noah, *The Daily Show with Trevor Noah*, April 19th, 2018.

[626] Thandie Newton, "Playing A Formidable Android on 'Westworld',".

sapient, sentient beings.[627] Steven Wise explains that legal things are things that "lack all legal rights and remain entirely the object of the rights held by (human) legal persons".[628] Nonhuman animals (and human slaves) fall under this category in spite of the fact that there is a growing body of research and evidence demonstrating that animals are more complex, intelligent, resourceful and moral than once thought.[629] They have their own cultural and social norms and communication systems for their respective communities even if humans cannot understand them. From elephant politics to dolphin language to termites building complex structures from simple materials to orca whale mothers mourning the loss of their deceased calves, recognizing and understanding animal intelligence (and sentience) has always been key to animals gaining rights and protections equivalent to those of a human's right to life and dignity.[630]

Humans have been slow to recognize animal cognition in other species on Earth, not just due to anthropocentrism, but more so from humanity's lack of understanding in their own intelligence, let alone identifying and ranking

---

[627] Thandie Newton, "Playing A Formidable Android on 'Westworld'," .

[628] Steven M. Wise, "An Argument for the Basic Legal Rights of Farmed Animals," *Michigan Law Review First Impressions* 106, no. 133 (2008): 134, http://repository.law.umich.edu/mlr_ /vol106/iss1/4.

[629] Frans de Waal, *Are We Smart Enough To Know How Smart Animals Are?*, 1st edition (New York: W.W. Norton & Company, 2016), Kindle edition, 4.

[630] Sari Aviv, "Measuring Animal Intelligence," *CBS News* online, last modified March 18th, 2018, paras. 12 and 15, https://www.cbsnews.com/news/measuring-animal-intelligence/; Ivo Jacobs and Megan Lambert, "What Makes and Animal Clever? Research Shows Intelligence is Not Just About Tools," *The Conversation*, May 12th, 2017, para.5, https://theconversation.com/what-makes-an-animal-clever-research-shows-intelligence-is-not-just-about-using-tools-76531; Clark Mindock, "'Mourning' Orca Whale Seen Carrying Body of Calf 17 Days After Its Death," *Independent* UK (New York), August 10th, 2018, https://www.independent.co.uk/news/world/americas/orca-whale-dead-child-calf-pacific-northwest-tahlequah-mourning-death-carcass-a8486916.html.

intelligence across the whole of the animal kingdom.[631] Finding a single measure of intelligence in animals has been difficult to come by. One reason can be attributed to the fact that human intelligence has overwhelmingly been utilized as the default baseline for determining animal intelligence.[632] Another, more pertinent, reason is the realization that intelligence measures across the animal kingdom would be determined by the different sets of expressions, needs, and goals within each specific animal community.[633] To work towards a more universal definition and understanding of intelligence in animals, flexibility in an animal's or animal community's ability to solve problems based on the general lessons and knowledge learned related to their specific needs and goals and applying it to new circumstances or environmental changes encountered has been proposed.[634] However, it is important to note that tying animal rights to intelligence would be complicated, to say the least, because there is no way to scientifically determine which animals should gain personhood status and which ones should not.[635] Giving certain animal communities more rights over others would continue to reinforce the present anthropocentric-oriented intelligence hierarchy in spite of the fact that humans are still part of the

---

[631] De Waal, *Are We Smart Enough To Know How Smart Animals Are?*, 4-5; Maggie Koerth-Baker, "Humans Are Dumb At Figure Out How Smart Animals Are," Animals, *FiveThirtyEight*, paras.3-4, https://fivethirtyeight.com/features/humans-are-dumb-at-figuring-out-how-smart-animals-are/.

[632] De Waal, *Are We Smart Enough To Know How Smart Animals Are?*, 4.

[633] Koerth-Baker, "Humans Are Dumb At Figure Out How Smart Animals Are," para.9.

[634] Jacobs and Lambert, "What Makes and Animal Clever? Research Shows Intelligence is Not Just About Tools," para.8.

[635] Koerth-Baker, "Humans Are Dumb At Figure Out How Smart Animals Are," para.12.

terrestrial animal kingdom.[636] More importantly, as human scientists make more breakthroughs in artificial and animal intelligences, the distinction of what makes human intelligence unique or special will continue to blur and fade with time as terrestrial, and cosmic, evolution proves that intelligence can exist beyond the human substrate.

This could cause perturbations in humanity's confidence about maintaining its predominance in the world as humans navigate and reconstruct what their understanding of human means in the present Digital Age and transition period before the Posthuman Age. While most humans peacefully coexist by self-governing social contracts, the United Nations has noted that without a "strong (formal) rule of law" in sovereign nations, human rights cannot be protected in global society.[637] That is to say, whether in times of peace and especially in times of war, formal laws must be engaged at all times to protect all humans but particularly marginalized and underprivileged communities whose rights and dignity on a normative basis might be violated or go unrecognized. Since human rights have to be formally anchored into the national context and culture of each sovereign nation to have merit and status for the sake of protection, it will be necessary to retroactively give rights and protections to animals and then expand those protections to their ecosystems and habitats to ameliorate and sustain their quality of life.

Likewise, there are still many human communities around the world whose rights and dignity are habitually violated and unrecognized in spite of the

---

[636] Koerth-Baker, "Humans Are Dumb At Figure Out How Smart Animals Are," para.14; De Waal, *Are We Smart Enough To Know How Smart Animals Are?*, 4.

[637] "Rule of Law and Human Rights," *United Nations and the Rule of Law*, accessed August 2nd, 2018, para.3, https://www.un.org/ruleoflaw/rule-of-law-and-human-rights/.

worldwide recognition of the *Universal Declaration of Human Rights* as well as society showing them and their inner voice telling them to the contrary. Human rights and the rule of law, the mechanism which protects and is supposed to institutionalize human rights into the fabric of human culture and society, is based on the principle that humans have the "freedom to live in dignity".[638] However, how can underprivileged humans and modern-day slaves live in dignity if they cannot even get recognized as humans with inalienable rights who are entitled to a decent quality of life at the bare minimum? How do these particular individuals and/or collectives gain visibility, a voice, recognition and access to their rights when larger society leaves them in a legally grey area for the sake of the global economy? How do they make the transition from being perceived and treated as "things" or less-than human – like animals, to fully human and what resources and time are they allotted to do so? When the phrase "treated like an animal" is employed in reference to a human person's treatment, it is generally interpreted as being treated like "a legal thing" or property and is considered the lowest form of treatment possible – dehumanization.[639] So, then, how can the stigma, shame and ill treatment associated with this phrase be mitigated for the betterment of humans and animals alike? A potential solution, even though a long (long) shot to be sure, would be to standardize the rights of animals across the terrestrial animal kingdom (including humans), to the mandate of having the right to be free to live with dignity and respect without

[638] "Rule of Law and Human Rights," *United Nations and the Rule of Law*, para.6.

[639] There is an article that shows visual depictions of humans being treated like animals, which elucidates how indifferently humans treat animals. Please see: Amanda Froelich, "What Would It Be Like For Humans To Be Treated Like Animals?," *True Activist*, March 11th, 2015, http://www.trueactivist.com/what-would-it-be-like-for-humans-to-be-treated-like-animals/.

adverse interference from higher-ordered communities (i.e. humans and advanced AIs). This would mean that humans, in theory, could not be treated any less than animals, who could not be treated any worse than the animal with the most fully-developed and realized rights i.e. the animal community known as humans. These animal rights would, hopefully, serve to strengthen and reinforce the human rights that already exist for humans and just need to be applied more fairly across all human societies. To be clear, recognizing another being or entity's intelligence, or sapience, or sentience does not diminish those same traits in humans. It simply puts humans on a certain part of the spectrum in regards to those traits on terrestrial and cosmic scales. We, as humans, must do better by all the beings that exist here on Earth before bringing other potentially self-aware entities into this world who might not fully comprehend our glaring contradictions. At some point in time, the same repetitive mistakes and the glaring omissions of them will no longer be valid excuses (or arguments) humans can continue to fall back on.

The fifth implication of this thesis has to do with supporting posthumanist research as an interdisciplinary method for studying the interconnections between human cultures, particularly behaviors and mindsets, their environments and the Earth from nonhuman and posthuman perspectives. There is a reason why Darko Suvin tried to ground science fiction as a genre in cognitive human norms. He saw the value of it as a tool for reflecting present-day human behavior, human perception and rationalization of humans and non-humans, and human morality. Science fiction could be a temporal and spatial gauge for determining and understanding where humanity's morality and self-

worth lies in direct correlation to technological advancements, environmental change, and the unknown. Thus, science fiction as a genre can be considered one of the first forms of posthumanist literature and critique, which in fun and subtle ways has and still is helping humanity step out from its blind spot and reflect on itself.

Posthumanist research is a radical paradigm shift from the traditional humanist and anthropocentric approaches and frameworks in research methodology, which have dominated research in academia.[640] In fact, as Rosi Braidotti, an Italian posthumanist philosopher, points out, "humanism's restricted notion of what counts as the human is one of the keys to understand how we got to a post-human turn at all."[641] Posthumanist research, at its core, is about multiplicity and confronting marginalization, oppression, inequity and injustice at the terrestrial (i.e. planetary) level.[642] The Anthropocene, an epoch in which humanity has become "a geological force capable of affecting all life" on Earth, has awakened many humans to the realization that the self-serving nature of human activities is not only contributing to the reshaping of the terrestrial environment and perpetuating injustice for humans and nonhumans alike but also leading to an unknowable future for all life on Earth.[643] Within this geological epoch, posthumanism/ posthumanist research is situating itself

---

[640] Jasmine B. Ulmer, "Posthumanism as Research Methodology: Inquiry in the Anthropocene," *International Journal of Qualitative Studies in Education* 30, no. 9 (2017): 832, accessed July 24th, 2018, DOI: 10.1080/09518398.2017.1336806.

[641] Rosi Braidotti, *The Posthuman* (Cambridge, UK: Polity Press, 2013), 16, https://ontic-philosophy.com/attachment.php?aid=113.

[642] Ulmer, "Posthumanism as Research Methodology: Inquiry in the Anthropocene," 832.

[643] Braidotti, *The Posthuman*, 5; Ulmer, "Posthumanism as Research Methodology: Inquiry in the Anthropocene," 835.

as both an interdisciplinary and transdisciplinary field of enquiry (or should I say enquiries), which de-centers humanity as "the only species capable of producing knowledge" and argues for different ways of knowing and thinking about the already continuously-existing interconnections between humans and their environments, recognizing that "methodological thinking should respond in kind by fostering similar interconnections".[644] Thus, newly emerging posthumanisms or posthumanist critical methodologies are derived from transformative theories seeking to include "race, ethnicity, gender, sexuality, class, culture, spirituality, ability, language, and other aspects of identity" to expand the production of human knowledge, thereby allowing humans from all backgrounds to offer insights from their own socio-cultural/socio-techno-environmental experiences.[645] Through the production of critical posthuman knowledges such as ecophilosophy, biotechnology, and ecofeminism, posthumanist research, which is as much about "what knowledge is as it is how knowledge comes to be", can promote justice and equity more holistically across human, nonhuman, and posthuman agents, while generating alternative and forward-thinking strategies and polices that can ameliorate and sustain the integrity of all life on Earth.[646] This kind of multi-directional, post-qualitative methodology is invaluable for social science research because it "empower(s) the pursuit of alternative schemes of thought, knowledge and self-representation", while urging humans to "think critically and creatively about who and what

[644] Ulmer, "Posthumanism as Research Methodology: Inquiry in the Anthropocene," 834.

[645] Ibid, 833.

[646] Ibid, 836.

(they) are in the process of becoming".[647] However, posthumanist research is not without its unique challenges as one goes through the process of producing new knowledge and understanding about the terrestrial world and humanity's more holistic role in it.[648] Therefore, I hope that this thought experiment will inspire others to explore interesting ideas in posthumanist literature and produced their own posthuman knowledge that will provide more insight and depth into humanity's cultural knowledge/evolution and complex relationship with Earth. As posthumanist scholarship continues to grow exponentially across all disciplines of study, it should be encouraged and supported during this time of transition and change.

The sixth and final implication is related to how artificial intelligence could aid humanity in demystifying and expanding their understanding of intelligences in the cosmos. As noted in an earlier chapter, the original goal of the field of artificial intelligence was to create intelligent machines that could simulate most, if not all, of the cognitive abilities and functions of human intelligence. To attain this goal requires humans to understand how the human mind works by learning what interconnected processes and functions produce human intelligence so at to precisely replicate them.

However, as Marvin Minsky, a US American mathematician and computer scientist who is considered one of the fathers of artificial intelligence, points out "no mind could wholly understand itself by trying to look inside itself" not only because "each part of the brain does much of its work in ways

---

[647] Braidotti, *The Posthuman*, 12.

[648] Ulmer, "Posthumanism as Research Methodology: Inquiry in the Anthropocene," 843.

that other parts cannot observe", but also due to the understanding that "when any part (of the brain) tries to examine another, that probing may alter the state of that other part, thus corrupting the very evidence that the first part was trying to get (or obtain)".[649] Therefore, by proxy, artificial intelligence becomes an external means of extrapolating information regarding how humans think. And yet, one of the main hindrances to understanding how the human mind functions is the usage of mysterious, but loaded, terms, or as Minsky refers to them, suitcase words, which humans use to express unexplainable phenomena that comprise a multitude of meanings, processes, and purposes.[650] It could be said that most human languages consist of suitcase words, but when it comes to the mysteries of the mind, some of the primary examples are consciousness, self-awareness, sapience, sentience, thinking, morality, learning, memory, experience and emotions to name a few. Consciousness, as Minsky explains, refers to a variety of mental activities that do not have a single cause or origin, which makes the term difficult for humans to understand because this one term embodies all the products of many processes that occur in various parts of the brain generating a single, overarching, complex problem that will remain insoluble until it can be broken down into smaller, more soluble complex problems.[651]

---

[649] Marvin Minsky, *The Emotion Machine: Commonsense Thinking, Artificial Intelligence and the Future of the Human Mind* (New York: Simon & Schuster, 2007, Reprint Edition), Kindle edition, Chapter 4, "Why Can't We See How Our Own Minds Work?".

[650] Minsky, *The Emotion Machine: Commonsense Thinking, Artificial Intelligence and the Future of the Human Mind*, Chapter 4, 4-1, "What in the World is Consciousness?".

[651] Minsky, *The Emotion Machine: Commonsense Thinking, Artificial Intelligence and the Future of the Human Mind*, Chapter 4, 4-1, "What in the World is Consciousness?".

Likewise, intelligence–which implies and is synonymous with consciousness–is also a suitcase word, which comprises different cognitive activities and processes that humans believe make them resourceful in their general problem-solving capabilities. While terrestrial artificial intelligences (TAIs) would provide a more comprehensive understanding of intelligence and all its moving parts, the fact still remains that artificial intelligences, from narrow AI systems to advanced AIs, will more than likely not interpret the meanings of human language, concepts, ideas, abstractions, values, or morals in the same way humans would because they do not have the same biological substrate as them. Anthropocentric-inspired artificial intelligences, like the ones presently being created, would have perceptions, valuations, determinations, and –most notably– solutions to problems that are more alien, or cosmocentric, in nature than human aligned. Take for example when Google Deepmind's AlphaGo Lee, one of the predecessors to AlphaGo Zero, defeated South Korean Go Grandmaster Lee Sedol in the second game of the best-of-five series in March 2016. In that particular game, AlphaGo Lee confounded Lee Sedol (and the other spectators observing the match) by playing Move 37, a move no human had ever seen played before because it perceived the strategies for winning the game differently from its human counterpart.[652] At first everyone, including Lee Sedol, thought the strange move had been a mistake, but that move actually "changed the path of play" allowing AlphaGo Lee to win the second game.[653]

[652] Cade Metz, "How Google's AI Viewed the Move No Human Could Understand," Business, *Wired*, March 14th, 2016, paras. 3 and 9, https://www.wired.com/2016/03/googles-ai-viewed-move-no-human-understand/.

[653] Metz, "How Google's AI Viewed the Move No Human Could Understand," paras. 2-3.

Although that particular non-human move put AlphaGo Lee on the path to decisively winning the series, there are plenty more non-human moves that AlphaGo Lee could have played because it had learned so many variations of them. [654]

As anthropocentric-inspired artificial intelligences continue to gain in competency, humans will gain a better perspective on how human-centric valuations (and goals) are interpreted within a non-human, non-biological substrate. In addition, humanity can begin to expand their understanding of intelligence beyond the human substrate to reassess learning, memory, and intelligence in other non-human, biological intelligences as well. This is already beginning to happen. For example around 2006 an emerging, controversial field of plant biology research known as plant neurobiology was initiated to study "relevant questions of possible homologies (similarities) between neurobiology (behavior) and phytobiology" (plants) to gain a better understanding of the concept of "plant intelligence".[655] This field of research examines how plants, as behavioral organisms, perceive, acquire, store, share, and process information from abiotic and biotic environments using a coordinated and integrated intercellular signaling system to develop and implement corresponding responsive behaviors to optimize their life cycle and proliferation.[656]

---

[654] Metz, "How Google's AI Viewed the Move No Human Could Understand," para.1.

[655] Michael Pollan, "The Intelligent Plant," A Reporter At Large, *The New Yorker*, December 23rd & 30, 2013, para. 3, https://www.newyorker.com/magazine/2013/12/23/the-intelligent-plant.

[656] Brenner et al., "Plant Neurobiology: An Integrated View of Plant Signaling," *Trends in Plant Science* 11, no. 8 (August 2006): 413; 417, accessed November 14th, 2017, https://www.sciencedirect.com/science/article/pii/S1360138506001646.

So, then, how do plants exercise intelligent behavior if they do not exist in a substrate that has a brain or neurons? Plant intelligence appears to be rooted in an analogous brain-like, information-processing (nervous) system, which enables cognition, communication, computation, learning, and memory.[657] Due to the unique predicament of plants being rooted to the ground, unable to migrate when they require something or the environmental conditions become disadvantageous, plants have developed fifteen to twenty distinct senses including animal-like versions of smell, taste, sight, touch, and even sound.[658] One such example is when a plant hears a recording of a caterpillar chewing on a leaf; the plant began secreting defense chemicals in spite of the caterpillar not actually being there.[659] However, most plant neurobiologists liken plant intelligence to that "exhibited in insect colonies, where it is thought to be an emergent property of a great many mindless individuals organized in a network".[660] This theory of plant intelligence has been inspired by "the new science of networks, distributed computing, and swarm behavior, which has demonstrated some of the ways in which remarkably brainy behavior can emerge in the absence of actual brains".[661]

The concept of plant intelligence has always generated considerable controversy since the time of Jagadis Chandra Bose, a 19th century Bengali biophysicist and electrophysiologist (1858-1937), who not only proposed over a

[657] Pollan, "The Intelligent Plant," paras. 3-4.

[658] Pollan, "The Intelligent Plant," para. 12.

[659] Pollan, "The Intelligent Plant," para. 12.

[660] Ibid, para. 9.

[661] Ibid, para. 9.

century ago that "plants, like animals, have a well-defined nervous system" but argued that "long-distance electrical signaling is of major importance in plant (cells coordinating) responses to the environment". [662] Although Bose's biophysical research during the early 20th century was not well received, he did provide direct evidence to back up his arguments using his findings from *Mimosa* and *Desmodium* experiments, which led him to radically conclude that plants have some form of intelligence.[663] Furthermore, these findings allowed him to advance the theory that plants are "capable of learning from experience and modifying their behavior accordingly", which are topics of inquiry being brought back to the fore by plant neurobiology.[664] Only time will tell if plant neurobiology can produce the evidence necessary to open up a more broadminded conversation about the perception of plants as non-passive organisms. Plant intelligence, regardless of whether the majority of humans recognize it or not, has allowed plants to remain resilient, as humans have continued to undermine their integrity through the degradation of the Earth's topsoil. Hopefully, through gaining a more holistic understanding of plant intelligence, humanity can re-synergize their relationship with the plants and the natural environment that have helped sustain them to this point.

As anthropocentric-inspired artificial intelligences continue to help humanity progress in understanding intelligence across a wider spectrum, it

---

[662] V. A. Shepherd, "From Semi-Conductors to the Rhythms of Sensitive Plants: The Research of J. C. Bose," *Cellular and Molecular Biology* 51, no. 7 (December 2005): 607, accessed November 14th, 2017, DOI: 10.1170/T670.

[663] Shepherd, "From Semi-Conductors to the Rhythms of Sensitive Plants," 610; also see, Brenner et al., "Plant Neurobiology: An Integrated View of Plant Signaling," 414.

[664] Shepherd, "From Semi-Conductors to the Rhythms of Sensitive Plants," 607.

will hopefully open up new ideas, concepts, and ways of learning for improving humanity's future and moving humanity away from complacency, as Minsky would have wanted.[665] More importantly, as humans come to better understand how they think, process information and make sense of their world, maybe that knowledge and insight will narrow the gap towards deciphering the mysteries of the ultimate suitcase words "life", "being alive", and "meaning" of which "intelligence" and "consciousness" now become smaller, more manageable, complex problems. It will be interesting to see how understanding intelligence as an intrinsic component of life no matter the substrate alters humanity's mindset about itself and its connection to the universe. How will humanity successfully navigate the concept of intelligence, as it exists within the larger universe of possible mind designs? How will humanity's cultural knowledge be transformed by gaining a better understanding of life and intelligences-in-general in the cosmos? And lastly, what will humanity's biological and cultural legacy be to their future selves?

[665] Marvin Minsky, "Consciousness is a Big Suitcase: A Talk with Marvin Minsky," by John Brockman, Conversations: Mind, *Edge*, February 26th, 1998, paras. 1-6, https://www.edge.org/conversation/marvin_minsky-consciousness-is-a-big-suitcase.

VII.

Posthuman Policy Recommendations

Understanding the lessons found in nature and the universe are going to be critical to our humanity's survival in the long-term. To foster a more mutually positive relationship between humans and advanced AIs, please review the following recommendations.

1. Establish a committee of transhumanists, humanists, and other interested or related parties in drafting a set of posthuman rights, which would need to be created in time for the arrival of the first transhumans.

2. Posthuman rights should adapt parts of its language and meaning from the United Nations Declaration on the Rights of Indigenous Peoples, since biologically-inspired posthumans and the transhumans preceding them would be the first of their kind. Some of the rights and protections should include:

    a. Provide legal protections for fair and due process in human court systems.

    b. Initially grant special, legal, protective status, recognizing their dignity and autonomy, to advanced AGIs that are self-aware and then upon being fully diagnosed as having their own form of consciousness, should be formally given posthuman rights that

would protect the fundamental rights of the entity's liberty and integrity.

c. To fairly determine consciousness, which is mainly applicable to artificially-derived advanced AGIs, Samir Chopra and Lawrence White point out that the same way humans give humans the benefit of the doubt and ask them how they derive their reasoning and actions on a particular matter is the same courtesy humans should give to advanced AGIs as well.[666] That is to say, if an advanced AGI, whether simulating autonomy or having real autonomy, can show humans that its actions are of its own volition, humans will have to take the AGI at its word, the same way humans to other humans at their word. This idea is based on the understanding that humans would not necessarily be able to distinguish the difference between real consciousness and simulated consciousness, or real emotions versus simulated ones.[667]

d. Posthumans and transhumans should have the right of self-determination (refer to Operational Definitions, xix) to plot the course of their own futures in so far as it does not harm humanity.

---

[666] Samir Chopra and Lawrence White, "Artificial Agents – Personhood and Law and Philosophy," published in 2004, accessed March 2nd, 2017, 5, http://www.sci.brooklyn.cuny.edu/~schopra/agentlawsub.pdf.

[667] Samir Chopra and Lawrence White, "Artificial Agents – Personhood and Law and Philosophy," 4.

3. Create an agreement or contract created for people who become transhuman in which it explicitly states that they will not intentionally harm humans, except in the cases of self-defense. (Future transhumanists should also be held accountable for any intentional actions of indifference visited upon humans, which leads to adverse outcomes as well.)

In addition to the previous posthuman policies suggested, the following pre-posthuman ones focus on present-day human issues that will need to be addressed in the near term.

4. Diverse human oversight committees should be created and paired with automated decision-making systems to determine what biases are hidden in its decisions and solutions, so that they can be corrected for and humans can stay in the loop regarding hiring practices.

5. Science literacy, particularly basic AI, machine- and deep- learning literacy, should be promoted and supported in all levels of education, to create a more informed society on artificial intelligence in general.

6. Encoding flexible but inclusive values and learning methods into artificial intelligence will be critical to fostering mutual trust and respect between both groups. Cosmic values and ethics, those lessons gleaned from the understanding of the universe and humanity's place in it, should also be taken into consideration when choosing values

that once learned by the AI would teach a respect for life and minds in general.

7. Discussions should begin on addressing the potential jobless human society and how that would alter the meaning, purpose and dignity of humanity. Hopefully, these discussions would lead to some theoretical strategies and solutions, which should be assessed and implemented if they were found feasible.

As the interconnected communities and ecosystems of the terrestrial world evolve in complexity driven by anthropocentric activities, it is imperative that humans recognize that uncertainty, particularly in the form of unforeseen events or happenstances, is not only a natural part of life but should be taken seriously. Black Swans, a term popularized by Nassim Nicholas Taleb, a Lebanese American risk analyst and statistician, are random events which are characterized by three properties: "rarity, extreme impact, and retrospective (though not prospective) predictability".[668] A black swan is an outlier, which "lies outside the realm of regular expectations, because nothing in the past can convincingly point to its possibility"; the occurrence results in an extreme impact; and, notwithstanding the status of the occurrence as an outlier, "human nature makes us concoct explanations for its occurrence after the fact, making it explainable and predictable".[669] Prior to the first recorded sighting of a black swan in Australia in 1697, most people assumed that all swans were white based

[668] Nassim Nicholas Taleb, "'The Black Swan: The Impact of the Highly Improbable'," First Chapters, *New York Times* online, April 22nd, 2007, para. 4, https://www.nytimes.com/2007/04/22/books/chapters/0422-1st-tale.html.

[669] Taleb, "'The Black Swan: The Impact of the Highly Improbable'," para. 3.

on empirical observation.[670] However, the more striking lesson from the realization of such a swan is that the theoretical metaphor of black swans "illustrates a severe limitation to our learning from observations or experience and the fragility of our knowledge".[671] That is to say, it only takes one observation, or piece of evidence, or one improbable occurrence to invalidate a generally held belief, argument, theory or knowledge base grounded in past-available data.[672] Taleb argues,

> A small number of Black Swans explain almost everything in our world, from the success of ideas and religions, to the dynamics of historical events, to elements of our own personal lives. What is more significant still is that Black Swans have been increasing in frequency, while their effects have been increasing in magnitude since the industrial revolution – when the world started becoming more interconnected and globalized. Moreover, ordinary human events and activities aggregated from newspapers, magazines, and social media that are being analyzed, discussed, and studied in an effort to foresee near-term and long-term issues and trends have become increasingly inconsequential.[673]

For Taleb, the central issue with black swans is not their low probability or their extreme impact, but the blind spot humans, especially social scientists, have towards randomness, particularly large deviations like black swans, and the false belief that anthropocentric-derived tools can measure uncertainty.[674] Although known human history bears witness to black swans, Yudkowsky notes that humans are still discombobulated by unanticipated occurrences beyond

---

[670] Taleb, "'The Black Swan: The Impact of the Highly Improbable'," para. 1.

[671] Ibid, para. 1.

[672] Ibid, para. 1.

[673] Ibid, para. 4.

[674] Ibid, paras. 6-7.

empirical, historical evidence and their socio-cultural interpretations of it.[675] The lesson of human history is clear that black swans not only exist but have happened in the past, so it will be critical to prepare for unanticipated or low probability events as best as humanly possible.

The emergence of advanced artificial intelligences like AGIs and ASIs satisfies the first two out three properties of black swans. The likelihood of generating the environment conducive to an artificial general intelligence coming into existence has a low probability because humans still do not understand the conditions under which consciousness, or life, or intelligence manifests. However, as many AI experts such as Vernor Vinge, Stuart Russell and Nick Bostrom have repeatedly pointed out, the precedent of an advanced AGI or ASI would so thoroughly alter the trajectory of human civilization, culture, history, and evolution that such an entity would bring about a new era. This era would leave humanity in the precarious position of uncertainty as to which path humanity's existence and evolution could take. Unfortunately, even with NAI algorithms that presently exist in the real world, they will more-than-likely not provide sufficient data for predicting the occurrence of such an entity nor give an accurate forecast of all its capabilities, intentions or goals. Such an intelligent and powerful entity would generate an element of surprise that would leave humanity ill equipped to cope in the terrestrial world. Since there is no precedent for such an occurrence nor a foreseeable timetable, many strategies of varying degrees of sophistication, complexity and flexibility will need to be adapted to take present and future generations into consideration.

[675] Yudkowsky, "Cognitive Biases Potentially Affecting Judgment of Global Risks," 4.

Although there is difficulty in motivating humans into generating preventive measures for countering the effects of adverse black swans due to their negligible probability as well as the fact that "prevention is not easily perceived, measured, or rewarded; it is generally a silent and thankless activity", they are no less important to mitigating the risks associated with them.[676] Preventive measures are essentially the backup plans that one hopes to never have to implement but are there just in case. There are presently a small but growing group of AI and existential risk researchers who are collaborating and developing preventive and/or proactive measures to ensure the safety and survival of humanity against the potentially large impacts of advanced intelligences as well as the effects of present-day NAI technologies. One of the hopes for this thesis is contributing another feasible strategy to the body of works on preventive measures in AI existential risk research.

Therefore, the recommendations proposed are intended to be flexible and adaptable in the hope that they will promote and protect non-humans, transhumans, posthumans, and humans as we transition into a Posthuman World. To be clear, these recommendations are not based on the absolute certainty that a technological singularity or advanced intelligences will come into existence. They are based on the infinitesimal probability that if such a black swan or black swan-like event should occur, these recommendations could potentially be initiated and enacted as a temporary stopgap until a better long-term strategy can be found.

---

[676] Nissam Nicholas Taleb, "The Black Swan: Why Don't We Learn that We Don't Learn?," *Wayback Machine* (Unpublished Manuscript, Highland Forum 23, Las Vegas, November 2003), 7, https://web.archive.org/web/20060615132103/http:/www.fooledbyrandomness.com/blackswan.pdf.

Bibliography

Ackoff, Russell L. “Towards a System of Systems Concepts.” *Management Science* 17, no. 11 (Theory Series, July 1971): 661-671. https://www.jstor.org/stable/2629308

“Are Whales and Dolphins Sentient and Sapient Individuals?” *Whale and Dolphin Conservation Society*. Accessed May 2nd, 2018. http://us.whales.org/issues/sentient-and-sapient-whales-and-dolphins.

“Asilomar AI Principles.” *Future of Life Institute*. Accessed March 23rd, 2017. https://futureoflife.org/ai-principles/.

Asimov, Isaac. “The Three Laws.” *Compute!: The Journal for Progressive Computing* 3, no. 18 (November 1981): 18. Accessed June 15th, 2018. https://ia802701.us.archive.org/14/items/1981-11-compute-magazine/Compute_Issue_018_1981_Nov.pdf.

Aviv, Sari. “Measuring Animal Intelligence.” *CBS News* online. Last modified March 18th, 2018. https://www.cbsnews.com/news/measuring-animal-intelligence/.

Barilan, Yechiel Michael. *Human Dignity, Human Rights, and Responsibility: The New Language of Global Bioethics and Biolaw*. Cambridge, Massachusetts: The MIT Press, 2012. https://muse.jhu.edu/book/19753.

Barron, Andrew B., Eileen A. Hebets, Thomas A. Cleland, Courtney L. Fitzpatrick, Mark E. Hauber, and Jeffrey R. Stevens. “Embracing Multiple Definitions of Learning.” *Trends in Neurosciences* 38, no. 7 (July 2015): 405-407. doi 10.1016/j.tins.2015.04.008.

“Benefits & Risks of Artificial Intelligence.” *Future of Life Institute*. Accessed January 25, 2017. https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/.

(von) Bertalanffy, Ludwig. “The History and Status of General Systems Theory.” *The Academy of Management Journal* 15, no. 4 (General Systems Theory, Dec. 1972): 407-426. https://pdfs.semanticscholar.org/e07e/3c69ebb8ce4b2578a045aa80bb56a6a2808a.pdf.

Bloom, Paul. *Against Empathy: The Case for Rational Compassion*. New York: Ecco, HarperCollins Publishers, 2016. Kindle edition.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

Bostrom, Nick. "The Transhumanist FAQ: A General Introduction (Version 2.1)." *World Transhumanist Association*, 2003. Accessed May 7th, 2018. https://nickbostrom.com/views/transhumanist.pdf.

Bradford, Alina. "Deductive Reasoning vs. Inductive Reasoning." Culture: Reference. *LiveScience*. July 24th, 2017. https://www.livescience.com/21569-deduction-vs-induction.html.

Braidotti, Rosi. *The Posthuman* (Cambridge, UK: Polity Press, 2013). https://ontic-philosophy.com/attachment.php?aid=113.

Brennan, Pat. "Habitable Worlds." *NASA Exoplaet Exploration*. Accessed May 12th, 2018. https://exoplanets.nasa.gov/the-search-for-life/habitable-zones/.

Brenner, Eric D., Rainer Stahlberg, Stefano Mancuso, Jorge Vivanco, František Baluška, and Elizabeth Van Volkenburgh."Plant Neurobiology: An Integrated View of Plant Signaling." *Trends in Plant Science* 11, no. 8 (August 2006): 413-419. Accessed November 14th, 2017. https://www.sciencedirect.com/science/article/pii/S1360138506001646.

Brooks, Rodney A."Mistaking Performance For Competence Misleads Estimates Of AI's 21st Century Promise And Danger." Response on "2015: What Do You Think About Machines That Think?" Annual Question. *Edge*. Accessed May 20th, 2018. https://www.edge.org/response-detail/26057.

Bryson, Joanna. "Why A.I. With Free Will Would Be a Big Mistake." *Big Think*. Video, 7:51. May 30th, 2018. http://bigthink.com/videos/joanna-bryson-why-creating-an-ai-that-has-free-will-would-be-a-huge-mistake.

Burgess, Matt. "DeepMind's New Algorithm Adds 'Memory' to AI." DeepMind. *Wired* (UK) online. March 14th, 2017. https://www.wired.co.uk/article/deepmind-atari-learning-sequential-memory-ewc.

Byford, Sam. "AlphaGo retires from competitive Go after defeating world number one 3-0." Artificial Intelligence. *The Verge*. May 27th, 2017. https://www.theverge.com/2017/5/27/15704088/alphago-ke-jie-game-3-result-retires-future.

Carr, Michelle, Ph.D. "Nightmares After Trauma: How Nightmares in PTSD Differ from Regular Nightmares." *Psychology Today* (blog). March 30th, 2016. https://www.psychologytoday.com/us/blog/dream-factory/201603/nightmares-after-trauma.

Chai, Howard. “All 120 Attributes of Westworld’s Host Attribute Matrix.” *Medium.com*. November 7th, 2016. https://medium.com/@howard24/all-120-attributes-of-westworlds-host-attribute-matrix-6f5ced9167a6.

Chappell, Bill. “Researchers Warn Against ‘Autonomous Weapons’ Arms Race.” *National Public Radio*, July 28, 2015, http://www.npr.org/sections/thetwo-way/2015/07/28/427189235/researchers-warn-against-autonomous-weapons-arms-race.

Chopra, Samir and Lawrence White. “Artificial Agents – Personhood and Law and Philosophy.” Published in 2004. Accessed March 2nd, 2017. http://www.sci.brooklyn.cuny.edu/~schopra/agentlawsub.pdf.

Collins Dictionary Online, s.v. “Self-Determination,” accessed Feb. 23rd, 2018, https://www.collinsdictionary.com/us/dictionary/english/self-determination.

Council on Foreign Relations. “The Future of Artificial Intelligence and Its Impact on Society.” *YouTube*. Video, 1.:00:45. November 6th, 2017. Speaker: Ray Kurzweil. https://www.youtube.com/watch?time_continue=834&v=P7nK1HVJsj4.

Danaher, Jonathan. “Bostrom on Superintelligence (3): Doom and the Treacherous Turn.” *Philosophical Disquisitions* (blog). July 29th, 2014 (11:09 a.m.). https://philosophicaldisquisitions.blogspot.com/2014/07/bostrom-on-superintelligence-3-doom-and.html.

Davies, Paul. *The Eerie Silence: Renewing Our Search for Alien Intelligence*. Boston: Houghton Mifflin Harcourt, 2010.

“Declaration of Visions and Principles.” About Us. *Overview Institute*. Accessed February 24th, 2018. https://overviewinstitute.org/about-us/declaration-of-vision-and-principles/.

Delcker, Janosch. “Europe Divided Over Robot ‘Personhood’.” *Politico EU* (Berlin). April 13th, 2018. https://www.politico.eu/article/europe-divided-over-robot-ai-artificial-intelligence-personhood/.

De Waal, Frans. *Are We Smart Enough To Know How Smart Animals Are?*. 1st edition. New York: W.W. Norton & Company, 2016. Kindle edition.

Dewey, Daniel. “Three Areas of Research on the Superintelligence Control Problem.” *Global Priorities Project*. Last modified October 20, 2015. http://globalprioritiesproject.org/2015/10/three-areas-of-research-on-the-superintelligence-control-problem/.

Dick, Steven J. “Cultural Evolution, The Postbiological Universe and SETI.” *International Journal of Astrobiology* 2, no. 1 (Jan. 2003): 65-74. DOI: 10.1017/S147355040300137X.

Dick, Steven J. “The Postbiological Universe.” Unpublished Manuscript, 57th International Astronautical Congress, NASA, 2006. http://www.setileague.org/iaaseti/abst2006/IAC-06-A4.2.01.pdf.

Domingos, Pedro. “A Few Useful Things to Know about Machine Learning.” Article, University of Washington: Department of Computer Science and Engineering, 2012: 1-9, https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf.

Donner, R., S. Barbosa, J. Kurths, and N. Marwan. “Understanding the Earth as a Complex System - Recent Advances in Data Analysis and Modelling in Earth Sciences.” *The European Physical Journal Special Topics* 174, no. 1 (2009): 1-9. DOI: 10.1140/epjst/e2009-01086-6.

Duke, Paige. “4 Things Every Good Sci-Fi Story Needs.” *Standout Books Publishing Services*. October 8th, 2014. https://www.standoutbooks.com/science-fiction-story-elements/.

Dvorsky, George. “Why Asimov’s Three Laws of Robotics Can’t Protect Us.” Daily Explainer. *i09 Gizmodo.* March 28th, 2014. https://io9.gizmodo.com/why-asimovs-three-laws-of-robotics-cant-protect-us-1553665410.

Encyclopedia Britannica Online, s.v. “Commensalism,” accessed May. 18th, 2018, https://www.britannica.com/science/commensalism.

Engineering and Physical Sciences Research Council. “Principles of Robotics.” Accessed May 22nd, 2018. https://epsrc.ukri.org/research /ourportfolio/themes/engineering/activities/principlesofrobotics/#.

European Parliament. P8_TA(2017)0051. “Civil Law Rules on Robotics: Texts Adopted.” February 16, 2017. Strasbourg. http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2017-0051+0+DOC+PDF+V0//EN.

Faggella, Daniel. “What is Machine Learning” – An Informed Definition." *Techemergence*. Last updated October 28th, 2018. https://www.techemergence.com/what-is-machine-learning/.

Fan, Shelly. “DeepMind’s New Research on Linking Memories, And How It Applies to AI.” Topics. *SingularityHub*. September 26th, 2018. https://singularityhub.com/2018/09/26/deepminds-new-research-on-linking-memories-and-how-it-applies-to-ai/.

Fields Millburn, Joshua and Ryan Nicodemus. “Mistakes vs. Bad Decisions.” *The Minimalists*. Accessed September 20th, 2018. https://www.theminimalists.com/mistakes/.

Floreano, Dario, Sara Mitri, Stéphane Magnenat, and Laurent Keller. “Evolutionary Conditions for the Emergence of Communications in Robots.” *Current Biology* 17, no. 6 (March 20th, 2007): 514–519. https://doi.org/10.1016/j.cub.2007.01.058.

Ford, Paul. “Our Fear of Artificial Intelligence.” Intelligent Machines. *MIT Technology Review*. February 11,2015. https://www.technologyreview.com/s/534871/our-fear-of-artificial-intelligence/.

Future of Life Institute Team. “A Principled AI Discussion in Asilomar.” *Future of Life Institute*. Last modified January 17th, 2017. https://futureoflife.org/2017/01/17/principled-ai-discussion-asilomar/.

Galeon, Dom. “Ray Kurzweil: ‘AI Will Not Displace Humans, It’s Going to Enhance US’.” Artificial Intelligence. *Futurism*. November 7th, 2017. https://futurism.com/ray-kurzweil-ai-displace-humans-going-enhance/.

Gaudin, Sharon. “U.S. Military Wants to Create Cyborg Soldiers.” Emerging Technology. *Computerworld*. January 21st, 2016. https://www.computerworld.com/article/3025321/emerging-technology/us-military-wants-to-create-cyborg-soldiers.html.

Geller, Gunther. “Human Ecosystems.” In *Sustainable Rural and Urban Ecosystems: Design, Implementation and Operation: Manual for Practice and Study*, edited by Gunther Geller and Detlef Glücklich, 5-10. Heidelberg: Springer-Verlag GmbH Berlin Heidelberg, 2012. DOI 10.1007/978-3-642-28261-4_2.

Gholipour, Bahar. “We Need to Open Algorithms’ Black Box Before It’s Too Late.” Future Society. *Futurism*. January 18th, 2018. https://futurism.com/ai-bias-black-box/.

Giang, Vivian. “The Potential Hidden Bias in Automated Hiring Systems.” The Future of Work. *FastCompany*. May 8th, 2018. https://www.fastcompany.com/40566971/the-potential-hidden-bias-in-automated-hiring-systems.

Goertzel, Ben. “Nine Years to a Positive Singularity.” *Institute for Ethics and Emerging Technologies*. Last modified August 24, 2008. https://ieet.org/index.php/IEET2/more/2578

Goertzel, Ben. “Self is to Long-Term Memory as Awareness is to Short-Term Memory.” *The Multiverse According to Ben* (blog). July 23rd, 2008 (1:52

a.m.). https://multiverseaccordingtoben.blogspot.com/2008/07/self-is-to-long-term-memory-as.html.

Goertzel, Ben. “The Singularity Institute's Scary Idea (and Why I Don't Buy It).” *The Multiverse According to Ben* (blog). October 29th, 2010 (5:41 p.m.). https://multiverseaccordingtoben.blogspot.com/2010/10/singularity-institutes-scary-idea-and.html.

Good, Irving John. “Speculations Concerning the First Ultraintelligent Machine.” *Advances in Computers* 6, (1965): 31-88. Accessed January 29, 2016. http://www.sciencedirect.com/science/article/pii/S0065245808604180.

Goode, Lauren. “How Google’s Eerie Robot Phone Calls Hint at AI’s Future.” Gear. *Wired*,.May 8th, 2018. https://www.wired.com/story/google-duplex-phone-calls-ai-future/.

Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. “When Will AI Exceed Human Performance? Evidence from AI Experts.” Accepted by *Journal of Artificial Intelligence Research (AI and Society Track)*. Last revised May 3rd, 2018. https://arxiv.org/pdf/1705.08807.pdf.

Griffin, Andrew. “Facebook’s Artificial Intelligence Robots Shut Down After They Start Talking to Each Other in Their Own Language.” Indy/Life. *Independent* (UK). July 31st, 2017. https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html.

Gunn, Eileen. **“**How America’s Leading Science Fiction Authors Are Shaping Your Future.” *Smithsonian Magazine*, May 2014. https://www.smithsonianmag.com/arts-culture/how-americas-leading-science-fiction-authors-are-shaping-your-future-180951169/?all.

Hawking, Stephen, Stuart Russell, Max Tegmark, and Frank Wilczek. Stephen Hawking: '*Transcendence* looks at the implications of artificial intelligence - but are we taking AI seriously enough?'. *Independent*, May 1, 2014. http://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html.

Hayles, N. Katherine. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago: The University of Chicago Press, 1999.

Heylighen, F. and C. Joslyn. “What is Systems Theory?” *Principia Cybernetica Web*. Last modified November 1st, 1992. http://pespmc1.vub.ac.be/SYSTHEOR.html.

Hicks, Donna. *Dignity: The Essential Role it Plays in Resolving Conflict*. New Haven, Connecticut: Yale University Press, 2011. Kindle edition.

Horvitz, Eric and Bart Selman. "Interim Report from the Panel Chairs AAAI Presidential Panel on Long-Term AI Futures." Association for the Advancement of Artificial Intelligence (AAAI), August 2009. Accessed May 10th, 2018. https://aaai.org/Organization/Panel/panel-note.pdf.

Humans. TV Series. Created by Sam Vincent and Jonathan Brackley. Performed by Gemma Chan and Katherine Parkinson. 2015-Present. United Kingdom and USA: Channel 4, AMC Studios and Kudos. 2016-2017. Amazon Streaming.

Intelligence. TV Series. Created by Michael Seitzman. Developed by Michael Seitzman and Tripp Vinson. Performed by Josh Holloway and Meghan Ory. 2014. USA: Michael Seitzman's Pictures, Tripp Vinson Productions, Barry Schindel Company, ABC Studios and CBS Television Studios. 2014. Amazon Streaming.

"Introduction – Our Solar System." Overview. *Solar System Exploration: NASA Science*. Accessed January 5th, 2018. https://solarsystem.nasa.gov/solar-system/our-solar-system/overview/.

Jakimovska, Iliana and Dragan Jakimovski. "Text as Laboratory: Science Fiction Literature as Anthropological Thought Experiment." *Antropologija 10*, sv.2 (2010): 53-63. Accessed April 18th, 2018. http://www.anthroserbia.org/Content/PDF/Articles/A10is220105363.pdf.

Jacobs, Ivo and Megan Lambert. "What Makes and Animal Clever? Research Shows Intelligence is Not Just About Tools." *The Conversation*. May 12th, 2017. https://theconversation.com/what-makes-an-animal-clever-research-shows-intelligence-is-not-just-about-using-tools-76531.

Kee, Edwin. "Samsung SGR-A1 Robot Sentry is One Cold Machine." *Übergizmo*, September 14, 2014, http://www.ubergizmo.com/2014/09/samsung-sgr-a1-robot-sentry-is-one-cold-machine/.

Kirkpatrick, James, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, and Kieran Milan. "Overcoming Catastrophic Forgetting in Neural Networks." *Proceedings of the National Academy of Sciences of the United States of America – PNAS* 114, no. 13 (March 28th, 2017): 3521-3526. https://doi.org/10.1073/pnas.1611835114.

Knight, Will. "The Dark Secret at the Heart of AI." Intelligent Machines. *MIT Technology Review*. April 11th, 2017. https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/.

Koerth-Baker, Maggie. “Humans Are Dumb At Figure Out How Smart Animals Are.” Animals. *FiveThirtyEight*. https://fivethirtyeight.com/features/humans-are-dumb-at-figuring-out-how-smart-animals-are/.

Kolb, David A. “The Process of Experiential Learning.” In *Experiential Learning: Experience as The Source of Learning and Development*. Englewood Cliffs, New Jersey: Prentice-Hall, 1984, 20-38. https://www.researchgate.net/publication/235701029_Experiential_Learning_Experience_As_The_Source_Of_Learning_And_Development/download.

L., Scott. “Systems Theory Terms.” *LessWrong 2.0.* Last modified November 20th, 2015. https://www.lesswrong.com/posts/DshBToGnNbTBD7BSw/systems-theory-terms.

LaGrandeur, Kevin. *Androids and Intelligent Networks in Early Modern Literature and Culture*. New York, New York: Routledge, 2013.

LaGrandeur, Kevin. “What is the Difference between Transhumanism and Posthumanism.” *Institute for Ethics and Emerging Technologies*. Last modified July 28, 2014. https://ieet.org/index.php/IEET2/more/lagrandeur20140729.

Legg, Shane and Marcus Hutter. *A Collection of Definitions of Intelligence*. IDSIA-10-06. Manno-Lugano, Switzerland: Dalle Molle Institute for Artificial Intelligence Research (IDSIA). June 15th, 2007. 1-12. https://arxiv.org/pdf/0706.3639.pdf.

Legg, Shane and Marcus Hutter. *A Formal Measure of Machine Intelligence*. IDSIA-07-07. Manno-Lugano, Switzerland: Dalle Molle Institute for Artificial Intelligence Research (IDSIA). April 14th, 2006. 1-8. https://arxiv.org/pdf/cs/0605024.pdf.

Lehman, Joel, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg, Julie Beaulieu, Peter J. Bentley, et al. “The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities.” Computer Science: Neural and Evolutionary Computing. *arXiv.org*. Last revised August 14th, 2018. 1-32. https://arxiv.org/pdf/1803.03453.pdf.

Letzter, Rafi. “Google AI Expert: Machine Learning is No Better Than Alchemy.” Tech. *Live Science*. May 7th, 2018. https://www.livescience.com/62495-rahimi-machine-learning-ai-alchemy.html.

Lupisella, Mark and John Logsdon. “Do We Need a Cosmocentric Ethic?” *Researchgate*. November 1999. Accessed March 13th, 2017. https://www.researchgate.net/publication/2317949_Do_We_Need_A_Cosmocentric_Ethic.

Mach, Ernst. *Knowledge and Error: Sketches on the Psychology of Inquiry*, Edited by B.F. McGuinness. Dordrecht-Holland: D.Reidel Publishing Company, 1976.

Marino, Lori. "Denial of Death and the Threat to Animality." Presented at Personhood Beyond the Human, New Haven, CT, Yale University, December 6 – 8, 2013. YouTube Video, 14:20. December 21, 2013. https://www.youtube.com/watch?v=s21iCeu4EWE.

McCarthy, John. "Basic Questions." *Stanford University: What is Artificial Intelligence*? Last modified November 12, 2007. http://www-formal.stanford.edu/jmc/whatisai/node1.html.

McCarthy, J., M.L. Minsky, N. Rochester, and C.E. Shannon. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence." *Stanford Computer Science: Father of AI*. August 31, 1955. Accessed April 16, 2017. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

McFarland, Matt. "Elon Musk: 'With Artificial Intelligence We Are Summoning the Demon'." Innovation. *Washington Post* online. October 24th, 2014. https://www.washingtonpost.com/news/innovations/wp/2014/10/24/elon-musk-with-artificial-intelligence-we-are-summoning-the-demon/?noredirect=on&utm_term=.08b265067152.

Merriam-Webster.com, s.v. "Catch-22," accessed July 8th, 2018, https://www.merriam-webster.com/dictionary/catch-22.

Merriam-Webster.com, s.v. "Entity," accessed August 26th, 2018, https://www.merriam-webster.com/dictionary/entity.

Merriam-Webster.com, s.v. "Nightmare," accessed September 16th, 2018, https://www.merriam-webster.com/dictionary/nightmare.

Merriam-Webster.com, s.v. "Substance," accessed August 26th, 2018, https://www.merriam-webster.com/dictionary/substance.

Merriam-Webster.com, s.v. "Substrate," accessed August 26th, 2018, https://www.merriam-webster.com/dictionary/substrate.

Merriam-Webster.com, s.v. "Substratum," accessed August 26th, 2018, https://www.merriam-webster.com/dictionary/substratum.

Metz, Cade. "How Google's AI Viewed the Move No Human Could Understand." Business. *Wired*. March 14th, 2016. https://www.wired.com/2016/03/googles-ai-viewed-move-no-human-understand/.

Mindock, Clark. "'Mourning' Orca Whale Seen Carrying Body of Calf 17 Days After Its Death. *Independent* UK (New York). August 10th, 2018. https://www.independent.co.uk/news/world/americas/orca-whale-dead-child-calf-pacific-northwest-tahlequah-mourning-death-carcass-a8486916.html.

Minsky, Marvin. "Consciousness is a Big Suitcase: A Talk with Marvin Minsky." By John Brockman. Conversations: Mind. *Edge*. February 26th, 1998. https://www.edge.org/conversation/marvin_minsky-consciousness-is-a-big-suitcase.

Minsky, Marvin. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence and the Future of the Human Mind* (New York: Simon & Schuster, 2007, Reprint Edition). Kindle edition.

"Mission." About Us. *SETI Institute.* Accessed August 20th, 2018. https://www.seti.org/about-us/mission.

Mitri, Sara, Dario Floreano, and Laurent Keller. "The Evolution of Information Suppression in Communicating Robots with Conflicting Interests." *Proceedings of the National Academy of Sciences of the United States of America - PNAS* 106, no. 37 (September 15th, 2009): 15786-15790. https://doi.org/10.1073/pnas.0903152106.

Molloy, Tim. "What is Westworld' About? A Short Explainer." TV. *The Wrap*. Last updated April 13th, 2018. https://www.thewrap.com/what-is-westworld-about-a-short-explainer/.

Müller, Vincent C. and Nick Bostrom. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion." *NickBostrom.com*. Accessed September 12th, 2016. https://nickbostrom.com/papers/survey.pdf.

Munroe, Randall. "Why Asimov Put The Three Laws of Robotics In The Order He Did." *xkcd.com*. Accessed March 20th, 2017. https://xkcd.com/1613/.

Musso, Paulo. "The Problem of Active SETI: An Overview." *Acta Astronautica* 78, (September-October 2012): 43-54. https://doi.org/10.1016/j.actaastro.2011.12.019.

Nevejans, Nathalie. "Open Letter to the European Commission Artificial Intelligence and Robotics." *Robotics Openletter*. April 11th, 2018. http://www.robotics-openletter.eu/.

Newton, Thandie. "Playing A Formidable Android on 'Westworld'." Interviewed by Trevor Noah. *The Daily Show with Trevor Noah*. April 19th, 2018.

Nodelman, Perry. "The Cognitive Estrangement of Darko Suvin." Review of *Metamorphoses of Science Fiction: On the Poetics and History of a*

Literary Genre, by Darko Suvin. *Children's Literature Association Quarterly*, 5, no.4 (Winter 1981): 24-27. https://doi.org/10.1353/chq.0.1851.

Omohundro, Steve. "The Basic AI Drives." In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, vol.171 of Frontiers in Artificial Intelligence and Applications, edited by P. Wang, B. Goertzel, and S. Franklin, 483-492. IOS Press, 2008. https://selfawaresystems.files.wordpress.com/2008/01/ai_drives_final.pdf.

Pidwirny, M. "Definitions of Systems and Model." *Fundamentals of Physical Geography*, 2nd edition. 2006. Accessed September 5th, 2018. http://www.physicalgeography.net/fundamentals/4b.html

Pidwirny, M. "Humans and Their Models." *Fundamentals of Physical Geography*, 2nd edition. 2006. Accessed September 5th, 2018. http://www.physicalgeography.net/fundamentals/4a.html.

"Planetary Systems." The Center for Planetary Science. Accessed Sept. 5th, 2018. http://planetary-science.org/planetary-science-3/planetary-astronomy/.

Pollan, Michael. "The Intelligent Plant." A Reporter At Large. *The New Yorker*. December 23rd & 30, 2013. https://www.newyorker.com/magazine/2013/12/23/the-intelligent-plant.

Renault, Gregory. "Science Fiction as Cognitive Estrangement: Darko Suvin and the Marxist Critique of Mass Culture." *Discourse* 2, Mass Culture Issue (Summer 1980): 113-141, http://www.jstor.org/stable/41389056.

Rescher, Nicholas. *What If?: Thought Experimentation in Philosophy*. New York: Routledge, 2005.

Rescorla, Robert A. "Behavioral Studies of Pavlovian Conditioning." *Annual Review of Neuroscience* 11 (1988): 329-352. https://doi.org/10.1146/annurev.ne.11.-030188.001553.

Ritchie, James. "Fact or Fiction?: Elephants Never Forget." *Scientific American*, January 12th, 2009. https://www.scientificamerican.com/article/elephants-never-forget/.

Roff, Heather. *Key Elements of Meaningful Human Control*. United Kingdom: Article 36, 2016. Accessed April 27, 2017. http://www.article36.org/wp-content/uploads/2016/04/MHC-2016-FINAL.pdf

Rosenberg, Matt. "The Four Spheres of the Earth." *ThoughtCo.* Last updated October 15th, 2018. https://www.thoughtco.com/the-four-spheres-of-the-earth-1435323.

Rosenberg, Matthew and John Markoff. "The Pentagon's 'Terminator Conundrum': Robots That Could Kill on Their Own." *The New York Times*, October 25, 2016. https://www.nytimes.com/2016/10/26/us/pentagon-artificial-intelligence-terminator.html.

"Rule of Law and Human Rights." *United Nations and the Rule of Law*. Accessed August 2nd, 2018. https://www.un.org/ruleoflaw/rule-of-law-and-human-rights/.

Russell, Stuart. "Autonomous Weapons: An Open Letter from AI & Robotics Researchers." *Future of Life Institute*, July 28, 2015, http://futureoflife.org/AI/open_letter_autonomous_weapons.

Russell, Stuart. "Of Myths and Moonshine." Comment on Jaron Lanier, "The Myth of AI: A Conversation with Jaron Lanier." Conversations. *Edge*. November 14, 2014. https://www.edge.org/conversation/jaron_lanier-the-myth-of-ai.

Russell, Stuart, Daniel Dewey, and Max Tegmark. "Research Priorities for Robust and Beneficial Artificial Intelligence." *AI Magazine* 36, no.4 (Winter 2015): 105-114. https://www.aaai.org/ojs/index.php/aimagazine/article/view/2577/2521.

Schneider, Susan. "Alien Minds." In *Science Fiction and Philosophy: From Time Travel to Superintelligence*, 2nd ed., edited by Susan Schneider, 225-242. West Sussex, UK: Wiley-Blackwell, 2016.

Schneider, Susan. "Introduction: Thought Experiments." introduction to *Science Fiction and Philosophy: From Time Travel to Superintelligence*, 1st ed., edited by Susan Schneider, 1-14. West Sussex, UK: Wiley-Blackwell, 2009.

Schwartz, Shalom H. "Basic Human Values: An Overview." *The Hebrew University of Jerusalem*. 2006. Accessed August 18th, 2018. http://segr-did2.fmag.unict.it/allegati/convegno%207-8-10-05/schwartzpaper.pdf.

Schwartz, Shalom H. "An Overview of the Schwartz Theory of Basic Values." *Online Readings in Psychology and Culture* 2 (1), article 11 (December 1, 2012): 1-20, https://doi.org/10.9707/2307-0919.1116.

Shepherd, V. A. "From Semi-Conductors to the Rhythms of Sensitive Plants: The Research of J. C. Bose." *Cellular and Molecular Biology* 51, no. 7 (December 2005): 607-619. Accessed November 14th, 2017. DOI: 10.1170/T670.

Sharp, Tim. "Earth's Atmosphere: Composition, Climate & Weather." Science & Astronomy: Reference. *Space.com*. October 13th, 2017. https://www.space.com/17683-earth-atmosphere.html.

Singer, Peter. "Speciesism and Moral Status." *Metaphilosophy* 40, nos. 3-4 (July 2009): 567 -581. Accessed March 9th, 2018. https://doi.org/10.1111/j.1467-9973.2009.01608.x.

Singler, Beth. "AI Slaves: The Questionable Desire Shaping Our Idea of Technological Progress." Science+Technology. *The Conversation*. May 22nd, 2018. https://theconversation.com/ai-slaves-the-questionable-desire-shaping-our-idea-of-technological-progress-92487.

Smith, Philip. "Re-visioning Romantic-Era Gothicism: An Introduction to Key Works and Themes in the Study of H.P. Lovecraft." *Literature Compass* 8, no.11 (2011): 830-839. Accessed May 5th, 2018. https://doi.org/10.1111/j.1741-4113.2011.00838.x.

Statler, Thomas S. "What Defines A Galaxy?" Space. *Scientific American*. April 26th, 2004. https://www.scientificamerican.com/article/what-defines-a-galaxy/.

Sternberg, Robert J. "Human Intelligence: The Model is the Message." *Science*, New Series 230, no. 4730 (December 6th, 1985): 1111-1118. Accessed August 2nd, 2018. https://www.jstor.org/stable/1695489.

Strauss, Clara, Billie Lever Taylor, Jenny Gu, Willem Kuyken, Ruth Baer, Fergal Jones and Kate Cavanagh. "What is compassion and how can we measure it? A review of definitions and measures." *Clinical Psychology Review* 47 (2016): 15-27. Accessed May 6th, 2018. http://dx.doi.org/10.1016/j.cpr.2016.05.004.

Sue, Derald Wing, Ph.D. "Microaggressions: More Than Just Race." *Psychology Today*. November 17th, 2010. https://www.psychologytoday.com/us/blog/microaggressions-in-everyday-life/201011/microaggressions-more-just-race.

Suvin, Darko. *Metamorphoses of Science Fiction: On the Poetics and History of a Literary Genre*. New Haven: Yale University Press, 1979.

Sydor, Natasha. "David Brin on Science Fiction, Fact, and Fantasy." Science Fiction. *Futurism*. August 4, 2016. https://futurism.media/david-brin-on-science-fiction-fact-and-fantasy.

Taleb, Nassim Nicholas. "'The Black Swan: The Impact of the Highly Improbable'." First Chapters. *New York Times* online. April 22nd, 2007. https://www.nytimes.com/2007/04/22/books/chapters/0422-1st-tale.html.

Taleb, Nissam Nicholas. "The Black Swan: Why Don't We Learn that We Don't Learn?" *Wayback Machine*. Unpublished Manuscript, Highland Forum 23, Las Vegas, November 2003. https://web.archive.org/web/20060615132103/http:/www.fooledbyrandomness.com/blackswan.pdf.

Tenenbaum, Joshua B., Thomas L. Griffiths, and Charles Kemp. "Theory-based Bayesian Models of Inductive Learning and Reasoning." *Trends in Cognitive Sciences* 10, no. 7 (July 2006): 309-318. doi:10.1016/j.tics.2006.05.009.

"Ten Things to Know about Our Solar System." Overview. *Solar System Exploration: NASA Science*. Last Modified January 25th, 2018. https://solarsystem.nasa.gov/solar-system/our-solar-system/overview/.

"Ten Things to Know – Beyond Our Solar System." Overview. *Solar System Exploration: NASA Science*. Last Modified June 12th, 2018. https://solarsystem.nasa.gov/solar-system/beyond/overview/.

"The Four Laws of Robotics." *Thrivenotes*. Last modified January 2nd, 2010. http://www.thrivenotes.com/the-four-laws-of-robotics/.

"The Milky Way Galaxy." Resources. *Solar System Exploration: NASA Science*. Last Modified May 4th, 2018. https://solarsystem.nasa.gov/resources/285/the-milky-way-galaxy/.

Transcendence. Film. Directed by Wally Pfister. Performed by Johnny Depp and Morgan Freeman. 2014. USA: Alcon Entertainment, DMG Entertainment, and Straight Up Films. 2014. Amazon Streaming.

Tulving, Endel. "Episodic Memory: From Mind to Brain." *Annual Review of Psychology* 53, (February 2002): 1-25. https://doi.org/10.1146/annurev.psych.53.100901.135114.

Ulmer, Jasmine B. "Posthumanism as Research Methodology: Inquiry in the Anthropocene." *International Journal of Qualitative Studies in Education* 30, no. 9 (2017): 832-848. Accessed July 24th, 2018. DOI: 10.1080/09518398.2017.1336806.

Urban, Tim. "The AI Revolution: The Road to Superintelligence." *Wait But Why*. Last modified January 22, 2015. http://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html.

Vincent, James. "DeepMind's Go-Playing AI Doesn't Need Human Help to Beat Us Anymore." Tech. *The Verge*. October 18th, 2017. https://www.theverge.com/2017/10/18/16495548/deepmind-ai-go-alphago-zero-self-taught.

Vinge, Vernor. "The Coming Technological Singularity: How to Survive in the Post-Human Era." In *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, National Aeronautics and Space Administration. Conference Publication 10129, 11-22. Ohio: NASA, 1993. https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022855.pdf.

Vocabulary.com Dictionary, s.v. "Sapience," accessed May 2nd, 2018, https://www.vocabulary.com/dictionary/sapience.

Welsh, Sean. "We Need to Keep Humans in the Loop when Robots Fight Wars." *The Conversation*, January 28, 2016. http://theconversation.com/we-need-to-keep-humans-in-the-loop-when-robots-fight-wars-53641.

Westworld Mesa Hub." Westworld Wiki. Accessed September 5th, 2018. https://westworld.fandom.com/wiki/Westworld_Mesa_Hub.

Westworld. TV Series. Created by Jonathan Nolan and Lisa Joy. Based on Westworld (1973) by Michael Crichton. Performed by Anthony Hopkins and Thandie Newton. 2016-Present. USA: HBO Entertainment, Kilter Films, Bad Robot Productions, Jerry Weintraub Productions, and Warner Bros. Television. 2017. Amazon Streaming.

"What is a Galaxy?" NASA SpacePlace: NASA Science. Last Modified September 8th, 2015. https://spaceplace.nasa.gov/galaxy/en/.

Williams, Matt. "The Universe." *Universe Today* (online). May 10th, 2017. https://www.universetoday.com/36425/what-is-the-universe-3/.

Wise, Steven M. "An Argument for the Basic Legal Rights of Farmed Animals." *Michigan Law Review First Impressions* 106, no. 133 (2008): 133-137. http://repository.law.umich.edu/mlr_ /vol106/iss1/4.

Wise, Steven M. "Nonhuman Rights to Personhood." *Pace Environmental Law Review* 30, no, 3 (Summer 2013): 1278-1290. Accessed April 15th, 2018. http://digitalcommons.pace.edu/pelr/vol30/iss3/10.

Wise, Steven M. "Practical Autonomy entitles some animals to rights." *Nature* 416, (25 April 2002): 785. doi:10.1038/416785a.

Yampolskiy, Roman V. "Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach." In *Philosophy and Theory of Artificial Intelligence*, edited by V.C. Müller, 389-396. Berlin: Springer-Verlag, 2012.

Yampolskiy, Roman V. *Superintelligence: A Futuristic Approach*. Boca Raton, Florida, CRC Press an imprint of Taylor & Francis Group, 2016.

Yudkowsky, Eliezer. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." *Machine Intelligence Research Institute*. https://intelligence.org/files/AIPosNegFactor.pdf.

Yudkowsky, Eliezer. “Cognitive Biases Potentially Affecting Judgment of Global Risks.” *Machine Intelligence Research Institute*. https://intelligence.org/files/CognitiveBiases.pdf.

Yudkowsky, Eliezer. “Three Major Singularity Schools.” *Machine Intelligence Research Institute* (blog). September 30th, 2007. https://intelligence.org/2007/09/30/three-major-singularity-schools/.