From the Tree of Knowledge and the Golem of Prague to Kosher Autonomous cars : The Ethics of Artificial Intelligence Through Jewish Eyes

From the Tree of Knowledge & the Golem of Prague:

The Ethics of Artificial Intelligence Through Jewish Eyes

Abstract

This paper discusses the regulation of artificial intelligence from a Jewish perspective, with an emphasis on the regulation of machine learning and its application to autonomous vehicles and machine learning. Through the Biblical story of Adam and Eve as well as Golem legends from Jewish folklore, we derive several basic principles that underlie a Jewish perspective on the moral and legal personhood of robots and other artificially intelligent agents. We argue that religious ethics in general, and Jewish ethics in particular, show us that the dangers of granting moral personhood to robots and in particular to autonomous vehicles lie not in the fact that they lack a soul – or consciousness or feelings or interests – but because to do so weakens our own ability to develop as fully autonomous legal and moral persons. Instead, we argue that existing legal persons should continue to maintain legal control over artificial agents, while natural persons assume ultimate moral responsibility for choices made by artificial agents they employ in their service. In the final section of the paper we discuss the trolley dilemma in the context of governing autonomous vehicles and sketch out an application of Jewish ethics in a case where we are asking Artificial Intelligence to make life and death decisions. Our novel contribution is twofold; first, we bring a religious approach to the discussion of the ethics of Artificial Intelligence which has hitherto been dominated by secular Western philosophies; second, we raise the idea that artificial entities who are trained through machine learning can be ethically trained in much the same way that human are – through reading and reflecting on core religious texts. This is both a way of ensuring the ethical regulation of artificial intelligence, but also promotes other core values of regulation, such as democratic engagement and user choice.

Introduction

Artificial Intelligence (AI),[1] machine learning,[2] and computer-assisted decision-making are taking over more and more domains of social life and human judgment — from autonomous

[1] Margaret A. Boden, *Artificial Intelligence: A Very Short Introduction* (Oxford: Oxford University Press, 2018), 1, defines artificial intelligence as a comprising variety of dimensions of information-processing that seek to carry out psychological functions, such as "perception, association, prediction, planning, motor control," that have hitherto been associated only with living beings. General Artificial Intelligence refers to machines that possess a general cognitive capacity for intelligence, much as humans do; such machines do not currently exist, and the theoretical possibility of such machines is controversial.

[2] Boden, *Artificial Intelligence*, 39-40, in which she describes machine learning as a subset of AI that is highly mathematical, and that depends a good deal on Bayesian probabilistic models to train an information-processing system to learn a particular task. This often involves the assistance of human trainers, as well as systems of reinforcement, but machine learning can be done independently. Such systems therefore learn tasks from the bottom-up, much as human do. Successful machine learning depends upon having a large amount of high-quality

vehicles, medical diagnostics,[3] policing and crime prevention,[4] to controlling drones and even instituting lethal strikes during armed conflict.[5] Many of these decisions either possess an ethical dimension or they carry ethical consequences, and this is quite apart from the broader and more general moral problem of our delegating powers to machines for which we are wholly responsible, but only partially control.[6]

To date, much of the literature dealing with the ethics of AI takes one or more traditional Western philosophical approaches as a starting point: the deontological approach embodied in the moral philosophy of Emmanuel Kant,[7] the consequentialist or utilitarian approach first promulgated by Jeremy Bentham,[8] and the virtue or character-based approach of Aristotle.[9] However, the increasing role that AI will play in the lives of people in every corner and culture

data. See also: Ugo Pagallo, "From Automation to Autonomous Systems: A Legal Phenomenology with Problems of Accountability," Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17) (2017): 18

[3] Kumba Sennaar, "Machine Learning for Medical Diagnostics: 4 Current Applications," *Emerj*, March 5, 2019, https://emerj.com/ai-sector-overviews/machine-learning-medical-diagnostics-4-current-applications/.

[4] Lawrence McClendon and Natarajan Meghanathan, "Using Machine Learning Algorithms to Analyze Crime Data," *Machine Learning and Applications* 2, no. 1 (2015): 1-12; Susannah Breslin, "Meet the Terrifying New Robot Cop That's Patrolling Dubai," *Forbes*, June 3, 2017, https://www.forbes.com/sites/susannahbreslin/2017/06/03/robot-cop-dubai/#32d1504d6872.

[5] Bradley J. Strawser ed., *Killing by Remote Control: The Ethics of an Unmanned Military* (Oxford: Oxford University Press, 2013).

[6] See eg., Colin Lewis and Dagmar Monett, "AI and Machine Learning Black Boxes: The Need for Transparency and Accountability," Deep Learning World Conference, Munich, 6-7 May 2019, https://www.kdnuggets.com/2017/04/ai-machine-learning-black-boxes-transparency-accountability.html, explaining that many deep learning algorithms are a kind of 'black box" to us in that they are developed by the learning algorithm itself, and are not transparent to us and, in some cases, are not even knowable in principle.

[7] Deontology, most closely associated with Immanuel Kant, regards morality as a system of rights and duties. Here the focus is on categories of actions, where different actions are deemed impermissible, permissible, or obligatory based on a set of explicit rules.; see e*.g.*: Jennifer Uleman*, An Introduction to Kant's Moral Philosophy* (Cambridge: Cambridge University Press, 2010).

[8] Utilitarianism aims to produce the best aggregate consequences (minimizing costs and maximizing benefits) according to a pre-specific value function. For example, a classical utilitarian approach aims to maximize the total amount of happiness. See *e.g.*: Michael Sandel, *Justice: A Reader* (Oxford: Oxford University Press, 2007), 9, where he states, "One way of thinking about the right thing to do, perhaps the most natural and familiar way, is to ask what will produce the greatest happiness for the greatest number of people. This way of thinking about morality finds its clearest statement in the philosophy of Jeremy Bentham (1748-1832). In his *Introduction to the Principles of Morals and Legislation* (1789), Bentham argues that the principle of utility should be the basis of morality and law. By utility, he means whatever promotes pleasure or prevents pain."

[9] Virtue ethics regards ethical behaviour as the product of an acquired set of behavioural dispositions that cannot be adequately summarized as an adherence to a set of deontological rules (concerning actions) or to as a commitment to maximizing good consequences. See *e.g.*: Shannon Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (Oxford: Oxford University Press, 2016).

of the world demands a more inclusive approach to the ethics and politics of machine morality. As Awad *et al.* state, engineers, programmers, and ethicists form only a small constituency, whereas formulating the rules by which machine intelligences should be governed will require a broad-based consensus among global citizens.[10] The problem is primarily a normative one — one of political processes as much as fundamental rules of morality. If people do not find AI technologies to be moral, and the rules by which they are governed do not accord with popular conceptions of morality, then these technologies are unlikely to be accepted.[11] We would add that social divisions will be exacerbated if there is a lack of consensus on how to solve the moral, political, and legal disputes that AI technologies will generate. This requires more diverse voices to weigh in on the governance of AI than we have been hearing. To decide the profound and far-reaching moral and legal consequences of AI entities requires more perspectives; to discount them is like looking for a lost coin under the streetlight.

The purpose of this paper is to examine a few key problems in the regulation of AI from a Jewish perspective. We feel that Jewish ethics in particular, and religious approaches more generally, make a valuable contribution to the moral governance of AI. To be sure it enables one more constituency to be heard, but even more than this the moral principles at the heart of the Abrahamic religions should be heard. These views are a matter of deep conviction for many people — not only Jews, but Christians and Muslims as well. They have been influential in shaping the moral thinking and approaches used throughout the modern world, for these are living traditions, embedded in cultural and normative practices, and so represent a broader constituency than do the opinions of academics and professional ethicists.

Jewish ethics are themselves derived from a broad and diverse literary and intellectual tradition. For over two millennia, Jewish thought has focussed on the interplay of ethics with

[10] Edmond Awad, Sohan Dsouza, Richard Kim, Johnathan Schulz, Joseph Heinrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan, "The Moral Machine Experiment," *Nature* 563 (2018): 59.
[11] Awad, "Moral Machine," 59.

the rule of law. Jewish sources comprise both the tradition of rabbinic religious law — known as *Halakhah* — as well as non-legalistic rabbinical commentary known as the *Aggadah*, which consists of Biblical exegesis as well as tales, anecdotes, and folklore that explore many different spheres of life and knowledge, and that offer practical lessons and ethical teachings. This article will draw primarily on *Aggadic* sources, such as the Talmud and rabbinical commentaries, and will be divided into three parts. Part I introduces the Biblical story of Adam and Eve, and the Trees of Knowledge and of Life. From this ancient story we derive several fundamental principles of Jewish ethics that are relevant to governing AI, including the nature of humanity, its creation, and the ethical obligations owed by and towards human beings. Part II deals with the Jewish myth of the Golem, a synthetic but autonomous entity that has long raised ethical questions similar to those posed today by robotics and AI. Part III discusses some key issues regarding legal personhood and artificial agents in light of Jewish thinking. Part IV focuses on the traditional philosophical 'Trolley Dilemma' as applied to the ethical decision-making of autonomous vehicles, and how Judaism has addressed similar moral dilemmas. Here, we begin to sketch out a Jewish ethical response to the problems posed when machines make decisions involving life or death. Finally, we conclude by suggesting future directions whereby we can govern machine morality with human ethics that affords to each individual the worth and dignity they rightfully deserve.

Part I — In the Beginning: Moral Personhood and the Soul

> *And the LORD God formed man of the dust of the ground and breathed into his nostrils the breath of life; and man became a living soul.* ~ Genesis 2:7

Jewish ethics begins with the Biblical tale of *Genesis* in which God creates Adam and Eve. After inviting them to eat freely of every tree in the Garden of Eden, he warned them not to eat of the Tree of Knowledge of Good and Evil, "for in the day that thou eatest thereof thou shalt surely die."[12] But the serpent, who knows what Eve does not, tells her, "Ye shall not surely die: For God doth know that in the day ye eat thereof, then your eyes shall be opened, and ye shall be as gods, knowing good and evil."[13] The serpent describes the Tree of Knowledge as the fount of not only imaginative but creative power, stating that "from this tree the Lord ate and then created the world, and every artist hates his friend. Eat from it and you too shall be creators of worlds."[14] Seeking this wisdom, Eve took and ate the fruit, which she gave to Adam as well.[15]

The consequences of knowing good and evil were as swift as they were inevitable. The pair were overcome for the first time by shame and fear, and they hid themselves and their nakedness.[16] The Bible tells us that Adam and Eve "heard the voice of the LORD God walking in the garden in the cool of the day: and Adam and his wife hid themselves from the presence of the LORD God amongst the trees of the garden,"[17] a thing which they had never done before. God then called out to Adam, who replied "I heard thy voice in the garden, and I was afraid, because I *was* naked; and I hid myself."[18] And the LORD God then said to Adam, "Who told thee that thou *wast* naked? Hast thou eaten of the tree, whereof I commanded thee that thou shouldest not eat?"[19]

God then declared, "Behold, the man is become as one of us, to know good and evil: and now, lest he put forth his hand, and take also of the Tree of Life, and eat, and live for ever:

[12] Genesis 2:17 (King James Version).
[13] Genesis 3:4-5 (KJV).
[14] Zohar, 19, 35 (All translations from Hebrew were done by Nachshon Goltz. Nachshon is a native Hebrew speaker holding a BA in Hebrew literature, LLB (including Jewish Law) and LLM (Israeli Law) from Haifa University. Nachshon was an academic editor in Tel-Aviv University and spent more than 20 years researching the Hebrew language, especially in the context of academic and Biblical writing and understanding); The *Zohar is* a key work of Jewish *Aggadic* literature, which comprise non-legal commentaries, as well as the foundational text of Jewish mysticism. It likely dates from thirteenth century Spain; see: Scholem, Gershom and Melila Hellner-Eshed, "Zohar," in *Encyclopaedia Judaica,* Vol. 21, 2nd ed., ed. by Michael Berenbaum and Fred Skolnik, 647-664 (Detroit: Macmillan Reference USA, 2007).
[15] Genesis 3:6 (KJV).
[16] Genesis 3:10 (KJV).
[17] Genesis 3:8 (KJV).
[18] Genesis 3:10 (KJV).
[19] Genesis 3:11 (KJV).

Therefore the LORD God sent him forth from the garden of Eden, to till the ground from whence he was taken."[20] Humanity is now to live in exile — "In the sweat of thy face shalt thou eat bread, till thou return unto the ground"[21] — where they must labour, tilling the ground in hard labour to grow food,[22] while Eve and her daughters will conceive and bring forth children "in sorrow."[23]

The exile has often been interpreted as a fall from grace, from a more perfect state of being, farther from the Divine. One of the leading Biblical commentators from 13th century Spain, Bahya ben Asher ibn Halawa (also known as Rabbeinu Behaye) argues that before the sin of eating from the Tree of Knowledge we were "whole divine wisdom," having no cognizance of the matters of the body.[24] This point is underscored by the fact that Adam does not perceive himself as being 'naked' until after eating from the Tree, at which point he "obtained power of the lust and was drawn after the indulgences of the body."[25] The punishment of knowing good and evil is a direct consequence of that knowledge: one's mental faculties are overcome by feeling the full force of the pain of the living body — from labour, from hunger, from unfulfilled wants and dreams.

The Rabbi of Prague during the 16th century, Shlomo Ephraim ben Aaron Luntschitz, further explains that, "in every sin the drive for good and the drive for evil are in an argument."[26] In this battle, the drive for good relies on promises of the spiritual payment we will receive in the next world, while the drive for evil — like the serpent in the story — offers a promise of *this* world, of gratifications that come swiftly, of enjoyments "that are vivid to the seeing eye;" but the marvels of the next world, "no eye have seen."[27]

[20] Genesis 3:22-23 (KJV).
[21] Genesis 3:19 (KJV).
[22] Genesis 3:19.
[23] Genesis 3: 16.
[24] Mikraot Gedolot, 119, (trans. Nachshon Goltz).
[25] Mikraot Gedolot, 119, (trans. Nachshon Goltz).
[26] Mikraot Gedolot, 119, (trans. Nachshon Goltz).
[27] Mikraot Gedolot, 119, (trans. Nachshon Goltz).

From the outlines of the above story, we can draw out a number of fundamental Jewish ethical principles. First, human beings are unique amongst living creatures: made in the image of God, they are infused with the breath of life — known in Hebrew as *nishmat chayyim* — by which they are endowed with a soul, and therefore a special moral worth that deserves to be recognized by others.

Having tasted of the Tree of Knowledge they are further endowed with moral knowledge that is both the fountain of all further knowledge and creative activity, but which also places on us the responsibility to use that knowledge for the good of others. As Rabbi Luntschitz reminds us, the choice to do good is often a hard one to make and — as we shall see below when discussing the trade-offs we must choose between when governing autonomous vehicles — it is one that often requires from us a sacrifice from us. And as Rabbeinu Behaye reminds us, to be human is to be embodied, which means not only to be fragile and to suffer — from pain, from hunger, from longing, from death — but also to be possessed of an intelligence and a capacity for self-reflection that gives us a keen appreciation of this suffering. As a result, we possess the moral responsibility to have compassion for the suffering of others — to understand it, to assuage it, and in so doing to affirm the dignity of our fellow humans.

Part II - The Golem of Prague: Our Relationship with Artificial Entities

The ethical themes raised by the golem legends have long enlivened discussions of AI, and not only in Jewish thinking.[28] The golem is a mythical creature well-known from Jewish folklore. Like Adam, the golem is formed from the clay of the earth, but unlike Adam it is not fully

[28] But see: Yehuda Shurpin, "From Golems to AI: Can Humanoids be Jewish?" Chabad.org, 2019, https://www.chabad.org/library/article_cdo/aid/4285513/jewish/From-Golems-to-AI.htm#utm_medium=email&utm_source=1_chabad.org_magazine_en&utm_campaign=en&utm_content=content, in which he states that there may actually be fundamental differences between robots and golem, stating "unlike a robot, a *golem* has some sort of a spiritual spark animating it. It is brought to life through a righteous individual using the secrets of creation hidden within the *Sefer Yetzirah*. This is clearly not the case for a man-made robot powered by algorithms."

human. For this reason the golem — the artificial or synthetic being — has long raised many of the same ethical issues within Judaism as AI now poses. In many traditions, the golem is often unable to speak, and therefore may be deficient in intellect compared with humans, but perhaps even more importantly it lacks *nishmat chayyim* — the breath of life that links us with the divine, and that suffuses us with our humanity and our moral nature. According to the Talmud, Adam is described as a golem for the first twelve hours of his life before receiving his soul in the form of the breath of God,[29] which suggests that 'golem' refers to an intelligent entity that lacks a soul. Golems have traditionally been excluded (along with children, women and the mentally disabled) from being counted in the *minyan*, the quorum of ten men required for prayer and other religious rituals, signalling that they are not considered fully emancipated under traditional Jewish law.[30]

One of the most well-known versions of the golem legend tells the story of the Rabbi of Prague, Rabbi Judah Loew ben Bezalel (1513-1609), who is reported to have created a golem to protect the Jewish community from pogroms during the Passover holiday in the spring of 1580; a local priest had used the Blood Libel to incite Christians to acts of violence against Jews.[31] This story portrays a common theme in golem legends: that the golem is the product of faith as well as magic, and is often created by humans as a helpmeet to perform useful but difficult tasks. This further reinforces the similarities between the golems of folklore and the AI of the modern imagination.

[29] Babylonian Talmud, Sanhedrin 38b.

[30] Byron L. Sherwin, *The Golem Legend, Origins and Implications* (Lanham, MD: University Press of America, 1985), 23. Sherwin states of the sixteenth-century scholar Rabbi Ashkenazi that "Ashkenazi raises the implicit question of whether "artificially" created entities can be "persons in the law," having privileges and duties under the law. By excluding the Golem from the minyan, Ashkenazi articulates the position that "artificial" individuals do not qualify to be considered legal persons."

[31] Alden Orek, "Modern Jewish History: The Golem," *Jewish Virtual Library*, undated, https://www.jewishvirtuallibrary.org/the-golem. The Blood Libel is a common anti-Semitic theme of long standing in which Jews are accused of killing Christians, especially children, and using their blood in religious rituals; it is connected with other anti-Semitic views that portray Jews as engaging in magic, witchcraft, and acts of extreme evil.

The similarity is reinforced by another version of the tale, this one concerning a female golem created by Solomon ibn Gabriol, the eleventh century poet and philosopher of Andalusia. Ibn Gabriol suffered from a painful and disfiguring skin disease, possibly tuberculosis lupus, that left him irascible and reclusive, and he is said to have created a female golem as a housemaid.[32] Always a thorn in the side of the authorities, the story tells us that ibn Gavriol was suspected of committing lewd acts with the golem and was ordered to dismantle it. In this telling of the legend, the golem was made not of earth, but of wood and hinges.[33] Here, the golem is portrayed as a fully mechanical being, brought forth by the genius of its inventor, much as a modern robot.[34]

Because they are not fully human, our moral obligations towards golems and other artificial beings do not rise to the level of what we owe to other persons. Some scholars have posited that the obligations we owe golem are similar to that which we owe to animals. In one legend, Rabbi Zeira[35] was presented with a golem created through the magic of Rabbah.[36] Sherwin states that "Rabbi Zeira spoke to him but received no answer. Thereupon he said to him — You are a creature of magicians. Return to your dust."[37] This story shows that it is permissible to destroy golems, and this is not considered to be equivalent to murder.

Commenting on Rabbi Zeira's action, Gershon Hanokh Leiner, the nineteenth-century Hasidic Rabbi of Radzyn, states that killing the golem was justified because it was not intelligent, and is to be regarded instead "as an animal in human form."[38] This does not however

[32] Pen America, "Shelomo Ibn Gabirol (1021/22 – C. 1057/58), *Pen America,* February 16, 2007, https://pen.org/shelomo-ibn-gabirol-102122-c-105758/.

[33] Sherwin, *The Golem Legend*, 16.

[34] Sherwin, *The Golem Legend*, 19.

[35] a leading third generation Babylonian Amora that moved to the land of Israel. He is mentioned hundreds of times throughout the two Talmuds and his teachings are quoted by many proceeding Amoraim. He was an expert at the esoteric mysteries of Kabbalah and lived to a very old age. (https://en.wikipedia.org/wiki/Zeira).

[36] Rabbah Bar Nachmani, usually called Rabbah was a leading third generation Babylonian Amora. (https://he.wikipedia.org/wiki/רבה) (trans. Nachshon Goltz).

[37] Sherwin, *The Golem Legend*, 20-22, quoting the Babylonian Talmud, Sanhedrin, 65b.

[38] Sherwin, *The Golem Legend*, 40.

mean that we owe *no* moral obligations towards golems; we certainly have certain obligations to animals.[39] On the other hand, Leiner adds that had Rabbah created an intelligent golem, "he would have the legal statues of a true man…even as regards being counted in a minyan…and he would be the same as if God had created him."[40] This leaves open the possibility for further rights of moral personhood to be granted to intelligent golem — although the components of 'intelligence' here remain undefined.

Quite apart from the debate over whether and what moral obligations we owe to AI entities is the recognition that our use of AI may harm our own agency, an idea that we will return to below. Gavriol's female golem, for example, reflects the current controversy over robots built for purely sexual purposes.[41] Richardson argues that the fundamental problems with sex robots are that they will encourage users to dehumanize sex partners and view sexual relations as a purely material, transactional, relational.[42] The central issue is not whether these actions will cause harm to the robots themselves, but the harm that they will do to us, as our uses of the robots weakens our sense of *ourselves* as ethical beings living amongst other ethical beings to whom we owe a high standard of careful and conscientious treatment.

Norbert Wiener, one of the founders of the field of cybernetics, identifies this relationship between people and machines as being one of the central problems presently facing us, describing the machine as "the modern counterpart of the Golem."[43] For Weiner, too, our moral agency is weakened at its foundations by our usurpation of the act of creation itself from God — a theme echoed above in the snake's temptation of Eve to partake of the Tree of Knowledge.

[39] Rabbi Jill Jacobs, Ethical Treatment of Animals in Judaism, https://www.myjewishlearning.com/article/ethical-treatment-of-animals-in-judaism/

[40] Sherwin, *The Golem Legend*, 40.

[41] Campaign Against Sex Robots, "About," (n.d.), https://campaignagainstsexrobots.org/about/.

[42] Kathleen Richardson, "The Asymmetrical 'Relationship': Parallels Between Prostitution and the Development of Sex Robots," *SIGACS Computers & Society* 45, no. 3 (2015): 290-293.

[43] Norbert Wiener, *God and Golem, Inc.* (Boston: MIT Press, 1964), 95.

According to Wiener, in the Book of Job and Milton's *Paradise Lost*, "the Devil is conceived as playing a game with God, for the soul of Job, or the souls of mankind in general."[44] In many orthodox Jewish and Christian views, the Devil is one of God's creatures; in works like the Book of Job this creature directly challenges and attempts to usurp the power of the Creator — this is what the game between them is all about. As Weiner states:

> The conflict between God and the Devil is a real conflict, and God is something less than absolutely omnipotent. He is actually engaged in a conflict with his creature, in which he may very well lose the game. And yet his creature is made by him according to his own free will, and would seem to derive all its possibility of action from God himself. Can God play a significant game with his own creatures? Can *any* creator, even a limited one, play a significant game with his own creature?[45]

This relationship of conflict, of challenge, bring to light several visceral fears, the chief of which is the fear that our synthetic creations will surpass us: they will be smarter, faster, more invulnerable, and they will endure where we cannot; in so being, they will render us irrelevant.[46]

This view has been raised by a number of leading figures in science and technology, who have raised apocalyptic visions of the future of AI. Stephen Hawking, for example, has stated that AI is at least theoretically capable of exceeding human intelligence, and that it could mean the end of civilization unless we find a way to control it. He states, "we cannot know if we will be infinitely helped by AI, or ignored by it and side-lined, or conceivably destroyed by it."[47] Elon Musk has stated that artificial intelligence is the biggest existential threat we face; he likens us to the magicians of old when he states, "With artificial intelligence we are

[44] Weiner, *God and Golem*, 16.

[45] Weiner, *God and Golem*, 17.

[46] Boden, *Artificial Intelligence*, 136 *et seq.*; General Artificial Intelligence has so far proved to be elusive, and exponential advancements in the field have not brought us any closer to this goal; see also Daniel C. Dennett, "Will AI Achieve Consciousness? Wrong Question," *Wired*, February 19, 2019, https://www.wired.com/story/will-ai-achieve-consciousness-wrong-question/; some theorists of AI cast serious doubts over whether general AI is even possible at all, a debate which is beyond the scope of this paper.

[47] Arjun Kharpal, "Stephen Hawking says A.I. Could be 'Worst Event in the History of Our Civilization,'" CNBC, November 6, 2017, https://www.cnbc.com/2017/11/06/stephen-hawking-ai-could-be-worst-event-in-civilization.html.

summoning the demon. In all those stories where there's the guy with the pentagram and the holy water, it's like — yeah, he's sure he can control the demon. Doesn't work out."[48]

The 'magical' act of summoning AI is fraught with all of the ambiguities and responsibilities inherent in the act of creation itself. For the tellers of golem lore, "the creation of worlds and the creation of artificial life is not a usurpation of God's role of creator, but is rather a fulfilment of the human potential to become a creator."[49] Adam, after all, was an act of divine creation that began as a golem. As Gershom Scholem tells us, "we must go back to certain Jewish conceptions concerning Adam, the first man. For obviously a man who creates a golem is in some sense competing with God's creation of Adam; in such an act the creative power of man enters into a relationship, whether of emulation or antagonism, with the creative power of God."[50]

It is not only the nature of the golem that is important, but much of the golem legend is about the act of creation itself. Can we, too, aspire to be creators of worlds? If we reach too high, if we act out of the wrong motives, then our acts of creation may turn out to be acts of iniquity, and our creations abominations — a theme also played out in several modern works of fiction, including Marry Shelly's *Frankenstein*.[51] That the golem will enslave its master is likewise a common theme in several modern literary 'spin-offs' of the golem legend. For example, in the Capek brothers' famous play, "R.U.R" (Act 3), where the term 'robot' is first coined, we read, "Mankind will never cope with the Robots, and will never have control over them. Mankind will be overwhelmed in the deluge of these dreadful living machines, will be their slaves, will live at their mercy."[52]

[48] Samuel Gibbs, "Elon Musk: Artificial Intelligence is Our Biggest Existential Threat," *The Guardian*, October 27, 2014, https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat.

[49] Sherwin, *The Golem Legend*, 3.

[50] Gershom Scholem, "The Idea of the Golem," In *On the Kabbalah and Its Symbolism*, rev. ed. (New York: Schocken Books, 1996), 159.

[51] Mary Shelley, "Frankenstein or the Modem prometheus." *London: Printed for Lackington, Hughes, Harding, Mayor & Jones* 1818.

[52] Sherwin, *The Golem Legend*, 28.

Echoing the warnings above, Sherwin relates the biblical tale of Enosh:

> While the human being is encouraged to develop his/her creative potentialities, he/she is also warned that there are dangers inherent in the creative endeavour, dangers to the physical, moral and spiritual well-being of the human creature-creator. One such medieval warning tells that the biblical character Enosh learned that God had created Adam from the earth. Enosh then took some earth, kneaded it into a human form and blew into its nostrils to animate it as God had given Adam life. Satan then slipped into the figure and gave it the appearance of life. Enosh and his generation worshipped the figure and, hence, idolatry began. The figure was worshipped instead of God. The product of human hubris and demonic ruse replaced God as an object of human adoration.[53]

In Anne Foerst's telling of the legend, the golem was first built with the words *JHWH elohim emet* imprinted on its forehead, meaning "God the Lord is Truth."[54] But once the golem came to life, "it erased the letter א (the aleph, the first letter of the Hebrew alphabet) from the word truth so that now his forehead said *JHWH elohim met* (God the Lord is dead)." The golem justified this act to its terrified builders on the grounds that this act of creation actually serves to separate us from the divine. As Foerst states, "we adore God because God has created us, the most complex beings there are. If we are now able to re-create ourselves, people will adore the constructors of golems and not God anymore. But a god who is not adored and prayed to is dead."[55] This is the ultimate fear raised by AI, and other technological developments of the modern age, such as nuclear weapons and fossil fuels — that the act of creation can in turn destroy creation itself, and drive all *goodness* out of the world precisely because it drives it out of ourselves.

Part III – Legal Personhood and Artificial Agents

> *Did you ever expect a corporation to have a conscience, when it has no soul to be damned, and no body to be kicked? ~ Edward, First Baron Thurlow 1731-1806*[56]

[53] Sherwin, *The Golem Legend*, 19.

[54] Anne Foerst, *God in the Machine: What Robots Teach Us About Humanity and God* (New York: Penguin Group, 2004), 37.

[55] Foerst, *God in the Machine*, 37.

[56] John C. Coffee Jr., "'No Soul to Damn: No Body to Kick': An Unscandalized Inquiry into the Problem of Corporate Punishment*,"* *Michigan Law Review* 79 (1981): 386, citing Edward, First Baron Thurlow (1731-1806).

The above quotation from Baron Thurlow, then Lord Chancellor of England, has summed up for centuries of jurists their frustrations regarding the problems of artificial or fictive legal persons[57] — problems that are just as relevant in the debate over whether AI entities should or should not be granted the status of legal person. A legal person is one who bears legal rights and duties, who owns and administers property, enters into contracts, and has the capacity to be a party to a lawsuit.[58] Legal personhood overlaps to a great extent with moral personhood, which describes the extent to which and "under which conditions humans are *in control of* and therefore *responsible for* their everyday actions."[59] Free will, the ability to make a moral choice, and causal responsibility are all assumptions that underlie the notion of legal personhood.[60] Even so, the nature of legal personhood varies with the nature of the entity so endowed and need not be limited only to natural persons.[61] Corporations and governments are examples of fictive legal persons who possess all of the above legal capacities. Should artificially intelligent agents and entities be added into the mix?

Pagallo has identified a number of options for conceiving of the legal personhood of AI entities. One option is to grant AI entities full legal personality, and allow them to bear legal rights and duties of their own.[62] A second option is to grant AI entities only limited rights of personhood such as we do for children and mentally incapacitated adults — thus recognizing both their inalienable legal rights as well as their moral personhood — but require a legal

[57] Coffee, "No Soul to Damn: No Body to Kick," 386.

[58] Lawrence B. Solum, "Legal Personhood for Artificial Intelligence," *North Carolina Law Review* 70 (1992):1239.

[59] Filippo Santoni de Sio, and Jeroen van den Hoven, "Meaningful Human Control Over Autonomous Systems: A Philosophical Account," *Frontiers in Robotics and AI* 5, Article 15 (2018): 2.

[60] Santoni de Sio, "Meaningful Human Control," 9.

[61] Robert van den Hoven van Genderen, "Do We Need New Legal Personhood in the Age of Robotics and AI?" in *Robotics, AI, and the Future of Law,* ed. Marcelo Corrales, Mark Fenwick, and Nikolaus Forgó (Singapore: Springer Nature, 2018), 20.

[62] Ugo Pagallo. *The Laws of Robots: Crimes, Contracts, and Torts* (Springer, Dordrecht, 2013), 153.

guardian to act for them in most matters.[63] Third, AI entities could be given some of the same rights as corporations and other fictive legal persons; this facilitates legal transactions while acknowledging that they are morally and legally different from natural persons.[64] A fourth option is to acknowledge certain legal rights and duties only — such as certain contractual or tort obligations — without admitting them to full legal personhood.[65]

The present discussion will focus mainly on the legal personhood of machine learning entities, one key example of which are autonomous vehicles. Machine learning does not involve programming a system from the top-down, but in training it to learn how to make decisions from the bottom-up, usually through repeated exposure to a particular data set. Karanasiou and Pinotsis note that machine learning AI "operates on a formula based on several degrees of automation employed in the interaction between the programmer, the user, and the algorithm;" this can result in "different answers to key issues regarding agency."[66] Algorithms are increasingly able to "either augment or replace analysis and decision-making by humans," as happens with machine learning algorithms — which are capable of making decisions as well as making and modifying rules for making decisions — without human input.[67] Programmers use data sets to train algorithms, and this can be more or less supervised.[68] In unsupervised machine learning, the programmer will choose those features that will make the algorithm learn its task most efficiently, such as distinguishing faces, or recognizing obstacles on an unfamiliar roadway.[69] These features are introduced at the design stage by the programmer, and are an

[63] Pagallo. *The Laws of Robots*, 153.
[64] Pagallo. *The Laws of Robots*, 153.
[65] Pagallo. *The Laws of Robots*, 153.
[66] Argyro Karanasiou and Dimitris Pinotsis, "Towards a Legal Definition of Machine Intelligence: The Argument for Artificial Personhood in the Age of Deep Learning," Proceedings of ICAIL '17, London, United Kingdom, June 12-16 (2017): 119.
[67] Ugo Pagallo, "From Automation to Autonomous Systems: A Legal Phenomenology with Problems of Accountability," Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17) (2017): 18.
[68] The importance of choosing the appropriate data set cannot be overstated.
[69] Karanasiou, "Towards a Legal Definition of Machine Intelligence," 121.

important source of bias in the outcomes.[70] In reinforcement learning, the algorithm is taking a number of steps to maximize an end goal. Bias is also an issue here: 'deep learning' is at once more effective and more efficient, and also more complex and abstruse; it is also much more susceptible to bias.[71]

However, we need not delve into questions of whether and how much agency and decision-making capacity are actually held by any given AI in a system. At present, the state of the art in the law of robots is to use existing forms of vicarious liability to impose strict liability on legal persons for any harm or damage caused by robots or other AI entities that they deploy for their benefit — such as a driver using an autonomous vehicle to get around, a transportation company delivering goods, or a taxi service contracting for fares, *etc.*[72] Looking at AI systems not as independent agents but as tools of human industry invokes areas of the law such as products liability or warranty; looking at AI systems as agents of human interactions invokes existing legal doctrines such as vicarious liability for children, animals, or employees.[73] Many harms caused by autonomous vehicles can be dealt with through existing laws as well as public and private insurance schemes.[74]

Are there reasons why we ought to give full legal personhood to AI entities? Both pragmatic as well as philosophical arguments have been put forth. Personal accountability for robots may simplify some legal matters, "such as whether robots are acting beyond certain legal powers, which party should be held liable for conferring such powers, or whether humans can evade liability for possible malfunctions of a machine."[75] On the other hand, even Chopra

[70] Karanasiou, "Towards a Legal Definition of Machine Intelligence," 121.
[71] Karanasiou, "Towards a Legal Definition of Machine Intelligence," 121.
[72] Pagallo. *The Laws of Robots*, 118.
[73] Pagallo. *The Laws of Robots*, 130.
[74] Peter M. Asaro, "A Body to Kick but Still No Soul to Damn: Legal Perspectives on Robotics," in *Robot Ethics: The Ethical and Social Implications of Robotics*, ed. Patrick Lin, Keith Abney, and George A. Bekey (Cambridge, MA: MIT Press, 2012), 170.
[75] Pagallo. *The Laws of Robots*, 134.

and White, who argue in favour of legal personhood for AI agents, conclude that "an agency law approach to artificial agents is cogent, viable, and doctrinally satisfying."[76]

Given the present state of AI technology, it may be sufficient to deal with any problems that arise through existing laws of product liability, and agency — treating AI entities as *de facto* agents of a principal who is an existing legal person. However, there have been a number of philosophical objections raised to this approach. As AI technology progresses at a rapid clip, the hard cases[77] may begin to accumulate and overwhelm existing legal doctrine. Pagallo argues that a strict liability approach may lead to harsh results for natural persons and corporations who are held fully accountable for matters wholly beyond their control. This may lead to unfairness while hindering research and development.[78] Should we broaden the notion of personhood, then, to include AI agents?

The basic arguments in favour of this approach were expressed in the classic 1991 essay by Lawrence Solum[79] in which he addresses many of the main objections to AI personhood. One such is the 'anthropocentric objection,' which asserts that "the domain of morality is limited to interactions between humans."[80] But in this day and age our morality is not so easily constrained, and we often hear arguments that we owe moral duties to non-human entities — animals,[81] other living creatures, and even entities as complex and as different from us as entire ecosystems.[82] We may certainly owe some *moral* obligations to AI entities, but that does not make them *legal* persons.

[76] Samir Chopra and Laurence F. White, *A Legal Theory for Autonomous Artificial Agents* (Ann Arbor: University of Michigan Press, 2011), 23.

[77] "hard cases are those cases in which the result is not clearly dictated by statute or precedent" Dworkin, R. 1975. Hard Cases. Harvard Law Review 88: 1057

[78] Pagallo, "From Automation to Autonomous Systems," 21.

[79] Solum, L.B., 1991. Legal personhood for artificial intelligences. *NCL Rev.*, *70*, p.1231.

[80] Solum, "Legal Personhood for AI," 1261.

[81] See eg. The Nonhuman Rights Project in the United States, https://www.nonhumanrights.org/, which seeks the recognition of rights of legal personhood for animals.

[82] New Zealand, for example, recently recognized rights of legal personhood for the Whanganui river system, and it has had two guardians appointed to look after its interests; see: Eleanor Ainge Roy, "New Zealand River

Another objection raised by Solum is what he calls the 'missing something argument,' which denies full legal personhood to robots due to their lack of some cognitive or moral attribute — "consciousness, intentionality, desires, and interests,"[83] but also feelings,[84] or even free will.[85] To the extent that robots may possess some or all of these factors, Pagallo argues that they may be likened to the legal personality and moral agency of children or those suffering from a mental illness or disability.[86] They are autonomous, and possess agency, intentionality and interests, but they require a full legal and moral agent to govern their affairs.

Religious objections to the 'missing something' argument are often based on the idea that AI cannot be equivalent to moral or legal persons because they lack a soul — the breath of life, the *nishmat chayyim* with which God infused Adam and through him all humanity. Solum states that "some may find this argument very persuasive; others may not even understand what it means."[87] He rejects this view because legal norms in a modern, pluralistic society need to be based on public grounds and public reasons, and not on "religious or philosophical conceptions of what is good."[88] For the time being we do not see any evidence of AI entities possessing a soul, although we leave open the possibility — but we also agree that public laws must find acceptance among all of a society's constituencies, and here we found our objections to legal personhood for AI are based on a different ground.

This ground centers around what Solum calls the 'paranoid anthropocentric argument' against AI personhood, which states that robots might pose us harm and so we need to control them or dispense with them altogether. Solum responds to this argument by stating that "if AIs

Granted Same Legal Rights as Human Being," *The Guardian*, March 16, 2017, https://www.theguardian.com/world/2017/mar/16/new-zealand-river-granted-same-legal-rights-as-human-being.

[83] Pagallo. *The Laws of Robots*, 157.

[84] Solum, "Legal Personhood for AI," 1267.

[85] Solum, "Legal Personhood for AI," 1272.

[86] Pagallo. *The Laws of Robots*, 157.

[87] Solum, "Legal Personhood for AI," 1263.

[88] Solum, "Legal Personhood for AI," 1263.

really will poses a danger to humans, the solution is not to create it in the first place."[89] But this approach is at once too simplistic and apocalyptic for the current development. AI entities may certainly pose some risk to humans. Take for example the case of autonomous vehicles. Evidence[90] shows they may save many more lives than they end up taking or injure far fewer people on our roads than is the case with human drivers. Nevertheless, some harm is inevitable. We need to regulate autonomous vehicles in the public interest, resolve disputes, facilitate the democratic process, and declare standards of behavior, morality, and community norms — just as we do in every other area of the law.

Instead, we put forth a 'cautious anthropocentric' argument in favour of withholding legal personhood for AI entities. This view recognizes both the harms and benefits that can accrue from using autonomous vehicles, and mandates that we regulate AI in order to reap its benefits — efficiency, cost, and safety — while minimizing risks and distributing compensation fairly in cases where damages need to be paid, or where there is some dispute. In our view, the real harm that AI entities pose to human agents is that opening up legal personhood at this time may come at the expense of those who already hold that status.[91]

In making this argument, we draw less upon the creation tale from Genesis, and more upon the golem legends. The tale of the Rabbi of Prague and the golem he created to protect the Jewish community on Passover teaches us that while it may be permissible for us to create artificial entities to assist us in our tasks, we must remember our responsibility to keep control over them, and not the other way around. The tale of the female golem created by Solomon ibn Gabriol teaches us that quite irrespective of any moral duties we owe to artificial entities — whether or when their agency might rise to the level enjoyed by humans — the key moral issue at stake is that the way we treat them determines the development of our own characters and

[89] Solum, "Legal Personhood for AI," 1262.

[90]

[91] Van Genderen, "Do We Need New Legal Personhood," 26.

sets the future course of our own exercise of moral agency. The question is not how we might emancipate artificial entities, but how we should keep them from hindering *us* on our own, quest to become fully emancipated creatures ourselves — capable of articulating intentions and acting according to the dictates of reason, agency, and free will.

The evolutionary development of Autonomous vehicles[92] exemplify the fact that AI agents are not taking over new areas of human endeavour wholesale but are being rolled out gradually. This way, we may come to progressively give up more and more of our own autonomy and control over decision-making to the autonomous system. This is the heart of our present objection to legal personhood for AI agents — that in giving them greater autonomy, we diminish our own. A more moderate view can argue that we want assistance from legal agents, whilst always maintaining the ability to control these agents. Our use of AI entities has the potential to damage our very ability to be fully intentional and autonomous agents, in control over the moral choices that we make.[93] Yet we can counter this tendency by mandating that we continue to maintain full legal and moral responsibility over AI entities.

Few AI systems on the roads today are fully autonomous. The Society of Automotive Engineers classifies autonomous vehicles into different categories: from 0, in which the human driver does everything, to systems in which the vehicle automates some functions, to level 4 when the vehicle is fully automated without a human needing to take back control but only in some environments. [94] Only Level 5 describes a fully automated vehicle that can monitor its environment and perform all driving tasks autonomously in all conditions.[95] The *Vienna*

[92] It could be argued that the development of automatic changes of gears in vehicles was the first stage of the eventual move to autonomous vehicles.

[93] For a discussion of some Jewish principles of cognitive agency, and how this relates to spiritual life, see *eg.*: Rabbi Schneur Zalman of Liadi, *Tanya: Hebrew – English Standard Revised Edition*, trans. Nissan Mindel, Nissan Mangel, Zalman Posner, and Jacob Immanuel Schochet (Brooklyn, NY: Kehot Publication Society, 1973).

[94] National Highway Traffic Safety Administration (NHTSA), Federal Automated Vehicles Policy: Accelerating the Next Revolution in Roadway Safety. Report. U.S. Department of Transportation, September 2016, 9.

[95] National Highway Traffic Safety Administration (NHTSA), "Federal Automated Vehicles Policy," 9.

Convention on Road Traffic, as amended in 2016, has responded to the rise of autonomous driving systems by holding that the driver is always responsible for controlling the vehicle; recent amendments to article 8 require that automated driving systems[96] must be able to be overridden or switched off by the driver.[97] As van Genederen states:

> According to the road traffic law, the driver is the responsible party. But how to justify this when the driver is gradually losing control over the car and, instead, depends on numerous providers of information? These providers are the manufacturer, the infrastructure, road managers, other motorists, the producer of the software, the meteorological department, the designer of the algorithm at the heart of the learning vehicle and third-party data providers that control or affect navigation and engine control.[98]

This is a gradual process, one that gradually renders us more passive and accustoms us to giving up more of our own autonomy and control over decision-making to the AI system.

Van Genderen states that, "human control and accountability are important values to protect in all activities where basic human rights like life and physical integrity (as well as freedom and privacy) are at stake."[99] Accordingly, there should always be meaningful human control over autonomous systems. We need not decide whether the AI system possesses attributes of personhood, such as intelligence or free will.[100] We also need not dissect the role of the human versus the machine element, for we are imposing human agency as the locus of control over the system as a matter of normative choice.[101]

[96] For this reason, we prefer to use the term 'autonomous driving system' over the term 'autonomous vehicle.' This highlights the fact that there are a multiplicity of decision-making agents distributed in an overall system, some of which are located in the vehicle, some in the human driver, and some may be carried out by algorithms operating outside of the vehicle itself.

[97] Karanasiou, "Towards a Legal Definition of Machine Intelligence," 124; Vienna Convention on Road Traffic, 19 September 1949, 125 UNTS 3, art 8 5bis (entered into force 26 March 1952, as amended 22 March 2016); see also: United Nations Economic Commission for Europe (UNECE), "UNECE Paves the Way for Automated Driving by Updating UN International Convention," Press Release UNECE (March 23, 2016), https://www.unece.org/info/media/presscurrent-press-h/transport/2016/unece-paves-the-way-for-automated-driving-by-updating-un-international-convention/doc.html. Last accessed 24 February 2020.

[98] Van Genderen, "Do We Need New Legal Personhood," 33.

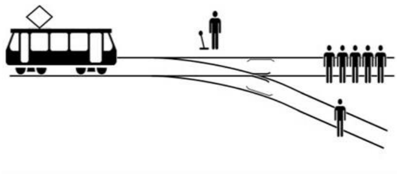
[99] Santoni de Sio, "Meaningful Human Control," 11.

[100] Karanasiou, "Towards a Legal Definition of Machine Intelligence," 125.

[101] Karanasiou, "Towards a Legal Definition of Machine Intelligence," 126.

There may come a time when an AI entity could logically and legally be a juridical person, and this time may no longer be too far into the future.[102] But to rush the process of emancipating AI agents might take something vital away from human agents. Here, we are not invoking the 'missing something' argument but making a practical argument about the negative effects this would have on our own agency and ability to act independently, make decisions — especially moral ones — and take control over our lives. One could come to the same conclusion through secular reasoning, as some have done.[103] We concur with Solum that a public system of laws and regulations should appeal to all persons,[104] and not only to those who hold religious beliefs. That people of very different beliefs can each look to our laws and regulations and find something in their essence that comports with their beliefs is important for a tolerant and pluralistic society. We think that many people will perceive the advantages in a regulatory system that gives them control over managing that technology and incorporating it into their own lives.

Part III – The Autonomous Kosher Car: How Should AI Solve the Trolley Problem?



Source: Wikimedia.org

The ethics of autonomous vehicles begins with an appreciation of just how many unnecessary deaths and injuries happen each year on the world's roads due to human error. Over one million lives are lost, and between twenty to fifty million people are injured, every year, to drivers who

[102] Van Genderen, "Do We Need New Legal Personhood," 40.
[103] Van Genderen, "Do We Need New Legal Personhood," 33.
[104] Solum, "Legal Personhood for AI," 1287.

are at fault.[105] The World Health Organization reports that road injuries are the eighth leading cause of death around the world, coming behind only such chronic conditions as heart and lung disease, cancer, Alzheimer disease, and diabetes.[106] Road injuries took the lives of about 1.35 million people in 2018, with the greatest burden borne by vulnerable people, such as children and those living in low-income countries; road injuries were the *leading cause of death worldwide* for children and young people between the ages of five and twenty-nine.[107] Preventing unnecessary suffering is a fundamental principle of all ethical systems including Judaism, and it is one of the main arguments in favour of adopting autonomous driving systems.

Even if autonomous vehicles are to become safer than human drivers, the enterprise of driving will always come with risks, meaning that some harm is inevitable. How, then, should the harms of autonomous driving systems be distributed? This is not a question we have had to address with the errors of human drivers, for whom the harm caused usually lies where it falls. Many philosophers have therefore turned to variations on the 'Trolley Problem' to address some of these moral dilemmas.

The Trolley Problem was introduced in the mid-20th century by philosophers Phillip Foot[108] and Judith Jarvis Thomson.[109] The Trolley Problem has since become a popular thought experiment in ethics that seeks to inquire about the conditions under which an individual will deflect a large projectile — usually a runaway trolley — from a larger group of persons to a smaller one. In one popular version of the thought experiment you are to imagine that you are driving a trolley.[110] You have gone round a bend and see five men on the tracks ahead of you.

[105] Nathan Posner, "Driverless Cars Belong in Israel," *Times of Israel*, August 7, 2016, http://blogs.timesofisrael.com/driverless-cars-belong-in-israel/. Last accessed February 24, 2020.

[106] World Health Organization, "The Top 10 Causes of Death," *WHO News*, May 24, 2018, https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Last accessed February 24, 2020.

[107] World Health Organization, Global Status Report on Road Safety, 2018 (Geneva: WHO, 2018).

[108] Philippa Foot, "The Problem of Abortion and the Doctrine of Double Effect," in *Virtues and Vices and Other Essays in Moral Philosophy* 19 (1978).

[109] Judith Jarvis Thomson, "The Trolley Problem," *Yale Law Journal* 94, no. 6 (1985): 1395.

[110] Judith Jarvis Thomson, "The Trolley Problem," 1395.

You put on your breaks, but they fail. You have just enough time to divert your trolley onto a track leading off to the right. However, there is a man on this track, too. If you divert the trolley you will kill him. Judith Jarvis Thompson asks, "Is it morally permissible for you to turn the trolley?"[111]

Thomson introduced another popular variant on the Trolley Problem by imagining a scenario in which you are on a footbridge and you see the runaway trolley hurtling towards the five men; you can stop it, though, by pushing a very large weight off the bridge into the path of the trolley.[112] Where would you find such a weight? Now, imagine that there is a very fat man leaning over the railing. If you push him, you will save the five, but take the life of the fat man. Is it morally permissible for you to do *this*? Most people when presented with this scenario, find that it is not, despite the fact that most people also think that five deaths are much worse than one.[113]

Although its modern formulation is new, the issues raised by the Trolley Problem are not. They involve questions of who should live and who should die and who should suffer harm when choices must be made, as well as the moral justifications we give for making these different choices — and all tensions that are brought forth between human reason and human emotion in moral decision-making. For these reasons, issues similar to the Trolley Problem have long been discussed in Rabbinical commentaries, as well.

One theme much discussed in the literature is whether it is permissible to kill another in order to save one's self. It is generally held that you may not violate the Torah's prohibitions against murder in order to save your own life. One Talmudic commentary illustrates this principle through the story of a man who came before the Rabbi Rava and said to him, "The

[111] Judith Jarvis Thomson, "The Trolley Problem," 1395.

[112] Judith Jarvis Thomson, "The Trolley Problem," 1409.

[113] Fiery Cushman and Liane Young, "Patterns of Moral Judgment Derive from Nonmoral Psychological Representations," *Cognitive Science: A Multidisciplinary Journal* 35, no. 6 (2007): 1052-1075.

ruler of my city has told me, ‘Go kill that particular person, and if you do not, I will kill you.’”[114] Rava said to him, “Let him kill you rather than you kill the innocent person. For what makes you think that your blood is redder? Maybe the blood of that person is redder!”[115]

A similar lesson is taught in the Talmud in the tale of two travellers lost in a desert without enough water for them to both make it to the nearest town.[116] Is it morally permissible for one of them to drink all of the water and survive at the expense of the other? There is great dispute among the commentators, with some agreeing that this would be wrongful,[117] while others would permit a person to save their own life under these circumstances.[118] The Talmud presents two views of this existential moral dilemma as follows:[119]

> Two people are travelling along the way, and one of them has in his possession a flask of water. If both drink from it, they will both die. However, if only one of them drinks, he will be able to make it out of the desert.
>
> Ben Patura expounded, “It is better that both should drink and die than that one should witness the death of his fellow.”
>
> Then Rabbi Akiva came and taught, “‘Your brother shall live with you’[120] — your life comes first, before your friend.”

Commentators such as Ben Patura would find it preferable for both to drink and die than for one of them to perish, knowing how his life had been devalued, and for the other to live, knowing that the continuation of his life is the result of his committing a moral wrong. Commentators such as Rabbi Akiva would argue that we are permitted to give precedence to our own life above that of another in cases in which the death of one is unavoidable. The

[114] Talmud, Pesachim 25A-B, 161.
[115] Talmud, Pesachim 25A-B.
[116] Talmud, Bava Metsi’a 62A.
[117] The Dilemma: Modern Conundrums, Talmudic Debates, Your Solutions, Student Textbook, Jewish Learning Institute, Lesson 5 Making the Right Turn – Engineering Ethics into Driverless Vehicles, 158, referencing Ben Patura.
[118] The Dilemma: Modern Conundrums, Talmudic Debates, Your Solutions, Student Textbook, Jewish Learning Institute, Lesson 5 Making the Right Turn – Engineering Ethics into Driverless Vehicles, 158, referencing Rabbi Akiva.
[119] Talmud, Bava Metzia, 62a.
[120] Talmud, Vayikra, 25:36.

circumstances under which this is permissible are complex and have been much debated, but the general consensus is that you may give preference to your own life over that of another and that this is an exception to the general *mitzvah* of putting other's needs ahead of your own.[121] When it comes to autonomous vehicles, many secular commentators would also resolve the problem in this way, and permit the vehicle to save the passenger in preference to another; most manufacturers of autonomous vehicles have also resolved the Trolley Problem to give precedence to the lives of a passenger or passengers.[122]

Another story is told in the Talmud that deals more directly with the theme of whether it is preferable to kill the one or to let the many be killed. Here, we are asked to imagine that a group of travellers — pious and faithful to God — are confronted along the road by a band of hostile and violent unbelievers. The travellers are ordered to hand over one of their number; if they refuse, they will *all* be killed. This is similar to the trolley problem, in which the choice is between killing the one and letting the many die, and a utilitarian calculus would normally permit — or even outright require — that we hand over the one in order to save more lives.[123]

But the Talmud tells us that morality does not work this way. An innocent should not be handed over to be killed, and this is the choice that is consistent with the values-based approach to ethics we have outlined above: to sacrifice one person for the benefit of others is to devalue that person's essential worth and dignity; at the same time, it degrades the dignity and the moral character of those who would do the handing over, for what benefit could they truly derive from a life founded on a wrongful action? This view is strengthened when we see

[121] Rav Binyamin Zimmerman, "Bein Adam Le-chavero: Ethics of Interpersonal Conduct, Shiur #29: 'Ve-ahavta Le-reiakha Kamokha' II - Putting the Needs of Others First," in The Israel Koschitzky Virtual Beit Midrash, Halakha, https://www.etzion.org.il/en/shiur-29ve-ahavta-le-reiakha-kamokha-iiputting-needs-others-first.

[122] Alex Roy, "Autonomous Cars Don't Have a 'Trolley Problem' Problem: In a You-Versus-Them Scenario, There's Only One Choice the Self-Driving Car Can Make," *The Drive*, October 19, 2016, https://www.thedrive.com/tech/5620/autonomous-cars-dont-have-a-trolley-problem-problem.

[123] Jerusalem Talmud, Terumot 8:4.

that there are exceptions to this rule based upon whether the individual being handed over justly deserves their treatment, such as when they have committed a wrong that has incurred the death penalty.[124]

Several modern commentaries take us even more directly into the heart of the Trolley Problem. One such is that of Rabbi Eliezer Waldenburg, who captures the problem by imagining that a driver encounters a group of people crossing the street; the only way to avoid them is to change direction — but the only direction in which the driver could go would result in his certainly killing one person.[125] The driver has the choice to swerve and kill the one, or to do nothing, and let the many die. Rabbi Eliezer argues that the driver should remain passive, refraining from taking any positive action that would end a life. He states, "at all costs, the driver should remain passive… The driver should not perform any act of commission. It makes no difference that the driver's intention in reversing is not to kill the one person but to save the larger number of people, because ultimately, the driver's act will in actual fact cause death."[126]

The Haredi scholar Rabbi Avrohom Yeshaya Kerlitz, also known as the Hazon Ish, addressed himself to a scenario remarkably similar to the Trolley Problem in a commentary on the Talmud. This problem was reportedly brought to him by a man who claimed to have experienced just such a moral dilemma: he had been driving and his brakes failed. He realized he was about to hit a group of pedestrians; so he turned the wheels way from the group, but in doing so he hit and killed a person walking on the pavement.[127] He asked the Hazon Ish if he was a murderer, or if his action was permissible in order to save the many.

[124] The Dilemma: Modern Conundrums, Talmudic Debates, Your Solutions, Student Textbook, Jewish Learning Institute, Lesson 5 Making the Right Turn – Engineering Ethics into Driverless Vehicles, 167.

[125] The Dilemma: Modern Conundrums, Talmudic Debates, Your Solutions, Student Textbook, Jewish Learning Institute, Lesson 5 Making the Right Turn – Engineering Ethics into Driverless Vehicles, 170, referencing Responsa Tsits Eliezer 15:70

[126] The Dilemma: Modern Conundrums, Talmudic Debates, Your Solutions, Student Textbook, Jewish Learning Institute, Lesson 5 Making the Right Turn – Engineering Ethics into Driverless Vehicles, 170.

[127] Olamot, http://olamot.net/shiur/הצלת-נפש-בנפש.

The Hazon Ish distinguishes this case from that described above, in which the individual is handed over by the group to be killed.[128] According to the Hazon Ish, the latter is a deliberate and cruel action that involves the intentional destruction of a soul; the remaining members of the group are saved not by the nature of the action itself, but by chance. Instead, the driver's dilemma is more akin to the case in which a person diverts an arrow from one side to the other to avoid hitting a person. This kind of action is, at its core, an act of rescue. Here, the *intention* of the actor to do good rather than harm is the crucial point, rather than the outcome. As such, this is not a deliberate killing; it is only by chance that the driver was posed with such a dilemma, and his action was morally permissible since we should try to minimize the loss of life whenever we can.[129]

When confronted with actual trolley problems, many people intuit that the moral choice is to save the many rather than the one. This accords with some of our moral intuitions that tell us that the value of five lives is worth more than one, and that the suffering endured by the loved ones of five lost persons is greater than the suffering of only one bereaved family.[130] But few people can actually justify pushing the fat man in front of the trolley, or killing the one person. Our emotional and moral connection with other persons is too great to permit us to justify direct killing in this manner.[131]

This is in accordance with traditional Jewish commentaries on the matter, which generally decline to weigh the value of a human life in terms of such a rational, disconnected calculus. As Rabbi Jonathan Sacks[132] reminds us, to value a human life at anything less than infinity is to devalue it. "Finitude is quantifiable," he states, "infinity is not. If human life is

[128] Hazon Ish, Hoshen Mishpat, Sahndrin 25, p. 203

[129] Hazon Ish, Hoshen Mishpat, Sahndrin 25, p. 203

[130] Fiery Cushman and Liane Young, "Patterns of Moral Judgment Derive from Nonmoral Psychological Representations," Cognitive Science: A Multidisciplinary Journal 35, no. 6 (2007): 1052-1075.

[131] Cite ventromedial study, explain the emotions of moral calculus.

[132] Former Chief Rabbi of the British Commonwealth and renowned biblical scholar.

very precious, and yet still finite in its value, then there is a difference between one man dying and many. And this difference makes it sometimes — in extremis — justifiable to sacrifice the one for the sake of the many."[133] But is there really a difference between one life and many? Not, Rabbi Sacks argues, if the value of a human life is infinite. "And infinity cannot be quantified," he states. Instead, "Infinity times one and infinity times one hundred are the same. So devastating is the loss of a single life that the enormity is infinite. And as between the death of one and the death of many there can be no calculations."[134]

Moral dilemmas about the just distribution of inevitable harms are not only the province of theoretical philosophy. During the Nazi occupations in Europe and the Holocaust, many Jews had to confront such moral dilemmas head on. These stories have been collected and retold, and they have become part of the cultural store of Jewish knowledge concerning ethics. One such incident took place in the Vilna Ghetto at a time when the Nazis handed down an order requiring a certain number of Jews be turned over to the Gestapo. Several Rabbis went to the Commander of the Ghetto and told him they would not comply, for Jewish law forbade them from turning over a Jewish person to the authorities unless the person were specified by name.[135] The commander replied by telling them that handing over a few Jewish persons would spare many more from death. The Rabbis replied in turn that Mamonides had considered such a situation and had ruled that if a non-Jewish authority tells you, "give me one of yours otherwise we will kill all of you. They should all be killed and not hand over one soul from Israel."[136]

[133] Rabbi Jonathan Sacks, "The Practical Implications of Infinity," in *To Touch the Divine* (Brooklyn, N.Y.: Merkos L'inyonei Chinuch, 1989), 83-84.
[134] Rabbi Sacks, "The Practical Implications of Infinity," 83-84.
[135] Gideon Refeal Ben-Michael (ed), *Biteun Forum Shmirat Zichron A Shoha* 56, (March 2019), 10, http://www.daat.ac.il/daat/shoah/biton56.pdf [trans. Nachshon Goltz].
[136] Mishna Torah, Larambam, Sefer A Mada, Hilchot Yesoday Hatora 5, 5 (trans. Nachshon Goltz).

In a similar incident in Vilna Ghetto, the Judenrat was ordered to hand over a Rabbi for execution; if they refused, five Jewish persons would be rounded up and killed in his place.[137] They did refuse, but now the terrible dilemma they had faced belonged to the Rabbi. Would he be willing to turn himself in and die to save five others? The Rabbi said that he preferred to die in the name of God and save five souls, and so he dressed his Shabat clothes and took his Talit and Tefilin to face the Nazi executioner.[138]

Another incident was told by the witness Dr. Dvorzsky at the genocide trial of Nazi war criminal Adolf Eichmann. The doctor was asked what would have happened if a person was found without a certificate issued by the occupying authorities:[139]

> A. He will be put on the way to Ponar.[140]
>
> Q. Do you recall a case in which a certain person returned home and told his mother, "I need to fix a certificate either for you or for my wife."?
>
> A. Yes, I remember. This person came home and told his mother, "What should I do? You walked us to the Hupa and now I can take only you or my wife." And his mother said to him, "It is written in our holy Torah 'A man shall leave his father and his mother and shall cleave unto his wife.' You need to build a family with your wife. I give up my life for her."
>
> Q. Was this man you?
>
> A. Yes.

Yosef Kremer related the following story to Yitzchak Nimtzovitz:

> My family and many other Jewish people were hiding from the Nazis in a bunker. One day my son, Davidal, started weeping and crying. Everyone in the bunker held their breath fearing that the Nazis will find them, but Davidal, our youngest son,

[137] Micheal Kasawar, "Di Lekwidtzya fon Tloshter Gata," witness statement in 'Sefer A Zicharon to Keilat Tloshtesh', 127-8; this witness statement was written in Yiddish and translated by Yaoshua Aivshitz in 'Dvarim Kektavam', 16-17; a 'Judenrat' is a Nazi administrative authority that was imposed in Ghettos and concentration camps and staffed largely by Jewish prisoners.

[138] Kasawar, "Di Lekwidtzya fon Tloshter Gata,"16-17; a 'Talit' is a woven prayer shawl with fringes at the corners and 'Tefilin' are versus of the Torah placed in leather boxes and strapped to the arm and head and worn in ritual prayer.

[139] Gideon Refeal Ben-Michael, ed., Biteun Forum Shmirat Zichron A Shoha, 56 at 31 (March 2019), http://www.daat.ac.il/daat/shoah/biton56.pdf (trans. Nachshon Goltz).

[140] The witness is referring to Ponary, a site in the forest outside of the Vilna Ghetto where tens of thousands of Lithuanian Jews were murdered and buried by the Nazi regime; see: Yad Vashem, "Vilna During the Holocaust: Ponary," *The Jerusalem of Lithuania: The Story of the Jewish Community of Vilna* (2020), https://www.yadvashem.org/yv/en/exhibitions/vilna/during/ponary.asp; Sakowicz, K., 2005. *Ponary Diary, 1941-1943: A Bystander's Account of a Mass Murder*. Yale University Press.

> did not stop crying. We heard the Nazi soldiers boots approaching and I could see the people's eyes begging me to stop my son's crying. I read their thoughts saying, "You need to strangle your son if he cannot calm down." Drops of sweat covered my face and I tried again and again to calm down Davidal with his crying blue eyes, but the baby would stop crying. Genia, my wife, was frozen hugging our two other children. All of a sudden, we heard hammers slamming, the Nazis were trying to find out whether Jewish people were hiding in the bunker, and Davidal keep crying. Until this day I cannot explain how I had the mental power to overcome my emotions. I grabbed Davidal by the throat and he stopped crying. Silent, chocked with tears, the people in the bunker mourned the death of Davidal. At night, I took his cold body and buried him outside, asking forgiveness from God.[141]

Finally, a witness told a similar story concerning the last days of the Warsaw Ghetto:

> In the last day in the Warsaw Ghetto, my husband and myself were trying to get help in order to save our daughter, a toddler named Tamarale. We turned to a Jewish policeman, an acquittance of ours, his name was Perlstein. The policeman removed his police hat and gave it to my husband, Israel, and told him to run to the square in which the Jewish people were concentrated in order to send them to the death camp. He also gave him his police certificate and told him, "Run quick. The train to Treblinka will leave on midnight. Tell the Nazi officer that your child was taken by mistake as we, the Jewish policemen, are protected against this kind of kidnaping."
>
> With shivering hands my husband Israel grabbed the hat. When he walked towards the door, the policeman told him, "You need to kidnap someone that will come instead of your Tamarale, since the Nazis make sure they have the right number." Israel froze near the door. It seemed that he did not understand. Then he removed the policeman's hat from his head, placed it along with the police certificate on the table and cried, "my only daughter, you are the only one I am allowed to sacrifice, only you are mine!"[142]

The above stories all find the killing of one person over another or others to be morally permissible in certain circumstances, but not for reasons that would be familiar to Western philosophers. A Utilitarian ethic might justify the Rabbi's sacrifice on the grounds that he saved five persons, and the killing of young Davidal saved an entire bunker of people. Even the mother's sacrifice in favour of her daughter-in-law might be justified on the grounds that a younger woman could bring the family more utility for many more years. But Utilitarian reasoning would struggle with Israel's sacrifice of his daughter, for one

[141] Gideon Refeal Ben-Michael, ed., *Biteun Forum Shmirat Zichron A Shoha*, 56 at 77 (March 2019), http://www.daat.ac.il/daat/shoah/biton56.pdf [trans. Nachshon Goltz]

[142] Durmbus-Album, "Aoif Der Arisher Ziyt," 168-9, translated from Yiddish by Yaushua Aiyvshitz, "Bekdusha Ubetaara," 289-90.

child's death would be equivalent to another's. A philosophy such as the 'ethics of care' may even condemn this action on the grounds that we should place the welfare of our own children above that of others.[143] In each case above, the emphasis is placed not on the utility of those caught up in the situation, but in their status as moral agents who make choices. The Rabbi and the mother chose to make their sacrifice, and it is this choice that forms their moral character and sets them up as examples for others. Only the parents made choices on behalf of their children, and here the emphasis is not on maximizing utility, but on their refusal to commit a greater moral wrong. Echoing the advice of the Hazon Ish, these were permissible as acts of rescue that sought to help others rather than to do harm.

In applying Jewish ethics to the more practical matters involved in governing self-driving cars, there are a number of general principles and rules of guidance that we can draw from the above discussion. One of the first points to be made is that human life cannot be adequately valued at all unless all lives are valued equally (and equally highly). One is generally prohibited from holding the life of one person above those of others. In making a life-or-death calculus, self-driving vehicles are therefore not permitted to take any individual or personal characteristics into account. A man's life is not worth more than that of a woman, a child's not worth more than that of an elderly person, a pregnant woman's not worth more than that of a man. Nor can decision-making algorithms take into account a person's ethnicity, social grouping, socioeconomic status, or criminal status, nor whether he is a drug addict, or a homeless person — and as the moral machine experiment demonstrated, these are all factors that people are likely to take into account, whether consciously or not, as part of our moral intuition.[144]

[143] See *e.g.*: Virginia Held, *The Ethics of Care* (Oxford: Oxford University Press, 2005).
[144] Awad, "Moral Machine."

Another principle that we can draw out is that there is a vast moral difference in killing the one over the many, or in sparing one's self over another, when we are pressed by chance and circumstance — rather than deliberate fault — to have to make such a choice. This leads to some counter-intuitive results. For example, suppose that a vehicle is about to crash, killing the driver, but she could be saved if the car were to swerve and hit a pedestrian. This action would be permitted. On the other hand, if a vehicle were about to hit a pedestrian, killing him, but he could be saved if the car were to swerve and hit a barrier, killing the driver, then this would not be permitted.[145] These two situations in fact involve the same moral principle: that it is wrong to kill only when the intention is to do wrong. Like the driver who was counselled by the Hazon Ish, we do not commit a moral wrong simply because we are faced by an unfortunate or unlikely situation. Similarly, if an autonomous vehicle cannot avoid harming someone, then it should simply shut the vehicle down, or otherwise take the path of least resistance. We should require assurances that autonomous vehicles are prohibited from deliberately harming anyone.[146]

The above principle also helps us resolve the Trolley Problem in a way that is in accordance with Jewish ethical thinking. All circumstances being equal, an autonomous vehicle faced with the dilemma of having to kill a smaller number to save the many may do so, as this is a rescue-type action rather than a cruel or intentional killing. As the editors of The Dilemma note, "it would be permissible to purchase a car that is programmed to save the greatest number

[145] The Dilemma, 174-5.

[146] See: Jean-François Bonnefon, Azim Shariff, Iyad Rahwan, "The Social Dilemma of Autonomous Vehicles," *Science* 352, no. 6293 (2016): 1573-1576, in which the authors found that most people favoured autonomous vehicles that were programmed to sacrifice passengers in order to save the greater number, but did not themselves want to ride in such a vehicle, and would not necessarily approve of enforcing such actions via regulation. The authors concluded that for most people, their self-interest outweighed their commitment to the greater good in such life and death situations.

of people, even if this is not the optimal way of programming the car. This is because the likelihood of such a scenario occurring is extremely remote."[147]

Another principle is that harms may be distributed in such a way that both harm and risk of harm may fall on those morally responsible for posing that harm.[148] In other words, the person at fault should bear the brunt of the consequences, just as we may hand over one to be killed when they have done something to justly deserve this result. These types of scenarios are fairly common on the road. It would be impermissible, for example, for an autonomous vehicle to swerve to miss a drunk driver if in doing so it were to hit a nearby pedestrian. What if there were passengers in the vehicle? Are they not morally responsible, as well, for permitting the driver to drive while impaired? Should they not shoulder the brunt of their actions, rather than the innocent pedestrian? Distributing the inevitable risks of the road according to causation and moral responsibility is a daunting task to be sure, but it is one that the era of self-driving cars will almost certainly bring about. As long as the overall result is to maximize safety and to minimize injury and loss of life while valuing human dignity and equality, then the result will be a vast improvement over current conditions on the world's roads, and well worth the effort.

Conclusion

One of the chief benefits of a religious approach to governance of AI is that it ensures that AI is governed by long-standing and deeply-held values that are shared by large numbers of the world's people; the ethics of AI will need to be in tune with the moral beliefs and practices of those who are subjected to AI technologies, and their input is going to be crucial in the adoption, governance, and perceived legitimacy of AI technologies.

[147] The Dilemma, 174-5.

[148] See *e.g.*: Jeff McMahan, *Killing in War* (Oxford: Oxford University Press, 2009) for a full discussion of deontological ethics as applied to life and death situations.

In this paper, we begin to sketch out a specifically Jewish approach to the governance of AI focusing on autonomous driving systems. We examined the story of Adam and Eve, and their eating the fruit of the Tree of Knowledge, to derive some of the basic principles of Jewish ethics: that humans are unique amongst living creatures, infused with the breath of God and endowed with a soul as well as moral agency and personhood. We then discussed the Golem legends of folklore that deal with many of the same ethical principles that arise when considering AI and robots. From these narratives, we learn that even though AI entities possess intellect, agency, intentionality, they lack some qualities possessed by human beings — not merely a direct connection with the Divine, but also a fragility and a capacity for suffering, as well as a self-reflective consciousness that gives us a keen appreciation of that suffering. While we certainly owe AI entities some moral and even legal obligations, it does not follow from this that they are deserving of full moral and legal personhood. The real ethical issue that we have to confront is how our treatment of AI entities and the kinds of relationships we cultivate with them is going to impact our own struggle to develop as full and fully intentional moral agents.

On the question of whether AI entities ought to be granted full legal personhood, we therefore give a negative answer. We do not invoke the ‘missing something’ argument, that AI entities lack something — a soul, intentionality, or agency — that would render them ineligible for this status. We leave open the possibility that they may achieve these qualities; they may at some point achieve full consciousness, or even a belief in God. Instead, we put forth the ‘cautious anthropocentric’ argument that states that to give over moral or legal personhood to AI entities will harm our own development as moral and legal agents.

Finally, we address some of the key moral dilemmas concerning the just distribution of inevitable harms that are posed by the introduction of autonomous driving systems. In the past, we have only been called upon to apportion legal responsibility after the fact: the harms caused by human drivers simply fall where they lie. With the introduction of autonomous vehicles, we are called upon to distribute these harms as a matter of system design. We draw upon Rabbinical commentaries, as well as Shoah stories, and we find that there is great disagreement over whether it is permissible to prefer one's life over that of another, and whether it is permissible to minimize the loss of life by killing the one over the many. This reflects a profound discomfort (in the commentaries as well as in the present authors) in permitting these kinds of actions. We reconcile these conflicting authorities by finding that such actions may be permissible *in extremis*: in cases where there is no fault, no intention to do wrong, and some harm is inevitable. This is an exception to the more general *mitvot* that that we should put our own interests after those of others, and that we should value each human being equally and equally highly. We close with a reminder of this fundamental ethical principle as expressed in the Babylonian Talmud, that "Whosoever shall destroy one life, it is as if he has destroyed the entire world; and whosoever shall save one life it is as if he has saved the entire world."[149]

[149] Babylonian Talmud, Sanhedrin, 37a (trans. Nachshon Goltz).