Rabbi Daniel Nevins בס"ד

CJLS HM 182

Approved, June 19, 2019 (18-0-0). Voting in favor: Rabbis Pamela Barmash, Elliot Dorff, Baruch Frydman Kohl, Susan Grossman, Judith Hauptman, Joshua Heller, Jeremy Kalmanofsky, Steve Kane, Jan Kaufman, Gail Labovitz, Amy Levin, Daniel Nevins, Micah Peltz, Robert Scheinberg, Deborah Silver, Ariel Stofenmacher, Iscah Waldman, Ellen Wolintz Fields.

Halakhic Responses to Artificial Intelligence and Autonomous Machines

Question:

With rapid advances in the development of artificial intelligence and autonomous machines have come calls for "moral machines" that integrate ethical considerations into analysis and action. What halakhic principles should apply to the conduct of such machines? Specifically:

1. Are Jews liable for the halakhic consequences of actions taken by machines on their behalf, for example, Sabbath labor?
2. Should ethical principles derived from halakhah be integrated into the development of autonomous systems for transportation, medical care, warfare and other morally charged activities, allowing autonomous systems to make life or death decisions?
3. Might a robot perform a mitzvah or other halakhically significant action? Is it conceivable to treat an artificial agent as a person? As a Jew?

Response:

I. The Rise of Artificial Intelligence and the Call for Machine Ethics

Among the extraordinary developments of our digital age, perhaps none is as startling as the introduction of smart machines that function without direct human supervision, in an open and dynamic environment, with the possibility of significant and even lethal consequences for people. An academic field of machine ethics has emerged, but there is no consensus about what type of ethical system, whether rule-based, consequentialist, or virtue, should guide the development of artificial intelligence.[1] To date there has been negligible engagement by halakhic authorities in this discourse,

[1] Keith Abney, "Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed," in Patrick Lin, Keith Abney, George A. Bekey, editors, *Robot Ethics: The Ethical and Social Implications of Robotics* (MIT Press, 2012). See also Steve Torrance, "Artificial Agents and the Expanding Ethical Circle," *AI and Society* (2013) 28:399–414, and Wallach and Allen, below. The machine ethics conversation began in science fiction literature with Isaac Asimov's three robot rules, first mentioned in his 1942 short story, "Runaround." These were: *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 orders given 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*

despite the high stakes for both the Jewish community and society at large.[2] In this responsum I survey the emerging field of machine ethics, consider relevant halakhic concepts, and apply them to questions about the intersection of artificial intelligence with ethics and religion, with the caveat that the pace and scope of innovation is rapid and unpredictable. After bringing the technological and halakhic discourses into conversation, I reach three conclusions which might function as models for additional topics as they emerge from theory into practice. First, we must address the portentousness of this topic, which signals a transition in our relationship to machines from mere tools to (perhaps) ethical agents.

Some ethicists dismiss the entire project of fashioning ethical artificial intelligence as misguided, regarding the embodied experiences of suffering, pain, pleasure and mortality as essential contexts for moral reasoning.[3] Dramatic reports about advances in AI often obscure the fact that there has been little progress in approximating the human capacity for *understanding meaning* in machines.[4] That is, AI may be smart without ever becoming conscious. Others observe that machine learning works with data on a scale and in patterns that defy human comprehension. Deep neural networks may have up to 150 hidden layers between input and output, making them thoroughly inscrutable to people. That said, humans cannot and need not fully comprehend one another's neurological processes in order to achieve mutual understanding.[5] The emerging subfield of "explainable machine learning" is working

---

protection does not conflict with the First or Second Law. Asimov later added a fourth or Zeroth Law (so named because it superseded the other three): *Zeroth: A robot may not harm humanity, or, by inaction, allow humanity to come to harm*. Scholars (and Asimov himself) have noted the unworkability of these laws which, like other deontological systems, run aground in situations of contested rights, as when two people's interests are in conflict and the robot must help or harm one.

[2] The edited volume *Robot Ethics* has 22 chapters, including one providing Buddhist reflections on the subject and another that models a "Divine-Command Approach," though there is no religious reasoning evident in the chapter. Artificial Intelligence has been mentioned by some scholars of halakhah, but without analysis of the field or substantial conclusions about the issues before us. There is an article regarding liability issues at stake with autonomous vehicles in Volume 38 of the Israeli journal *Teḥumin* which I discuss below. Nadav Berman Shifman graciously shared with me his forthcoming article, "Autonomous Weapon Systems and Jewish Law: Ethical-Political Perspectives," to which I refer below.

[3] Torrance calls this the biocentric position. Nadav Berman Shifman writes (p.36 of v.2.7), "It is by the body that we sense the world and other beings and through it we come to understand what good and evil are. Ultimately, it is the body that makes us vulnerable and at the same time punishable. This dependence of morality on *materiality* and *mortality*, is neglected in the discussion about robot ethics, probably due to the aforementioned dominance of Cartesianism in Western philosophy, and to the predominance of the 'algorithm ethics.'" See related discussion of Hans Jonas below.

[4] Melanie Mitchell, "Artificial Intelligence Hits the Barrier of Meaning," *NY Times*, Nov. 5, 2018. Regarding current methods for training computers to discern the "meaning" of language see the "Language as a Litmus Test," box within the article "Artificial Intelligence and Ethics" by Jonathan Shaw, *Harvard Magazine* (Feb. 2019) p.47, and then Barbara J. Grosz, "Smart Enough to Talk With Us? Foundations and Challenges for Dialogue Capable AI Systems" in *Computational Linguistics,* 44:1 (March 2018) 1-15.

[5] See below for discussion of Michael Graziano, *Consciousness and the Social Brain* (Oxford UP, 2013).

to bridge the communication gap between people and AI.[6] Indeed, there is recent activism to require that AI systems be designed with the ability to report their calculus in ways that humans might understand.[7] Amitai Etzioni and Oren Etzioni have proposed the development of an "ethics bot"[8] or "AI guardian"[9] that would supervise the functioning of other AIs to ensure that they comply not only with laws but also with moral values. Still, robotics researchers frequently comment on the oddness of some decisions made by AI systems, which are not attuned to the contextual cues that humans associate with common sense. The disconnect between the deliberations of humans and machines is a foundational problem in the development of moral machines.

An opposite problem is that the border between "natural" and "artificial" intelligence is eroding, with the integration of biological and manufactured deliberative systems well underway. Already people make momentous decisions with computer assistance, and such integration is expected to advance into more seamless and permanent forms. Algorithms that direct internet searches for information are generally hidden from the user; our decisions are unconsciously guided by the presentation of results. Machines regularly complete our sentences and even our thoughts. Humans depend upon artificial intelligence for a bewildering array of activities from navigation to financial investments to health care decisions, generally without paying critical attention to the values implicit in the guidance given by machines. We can anticipate conflicts between humanistic or religious values and the conduct of machines; might engineers integrate moral reasoning into the design of artificial intelligence and autonomous machines?

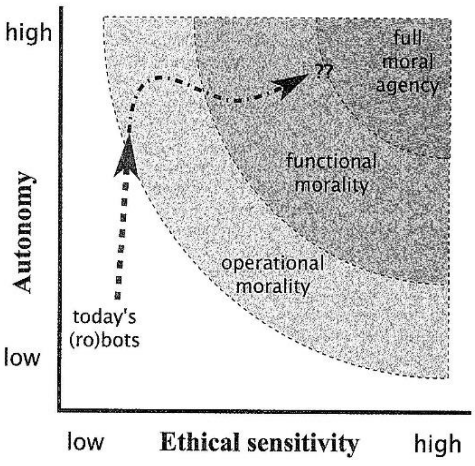
Wendell Wallach and Colin Allen chart a continuum of ethical sensitivity and autonomy[10] required to design what they call an artificial moral agent (AMA).[11] They term the lowest level of AMA "operational morality," meaning that the machine has been designed to function within certain parameters that respect the moral principles of the programmer. From a moral reasoning perspective, this is entirely passive. The highest level would be "full moral agency," which is not possible today for an artificial agent, and may indeed not be desirable or feasible, depending on one's theory of ethics. Between these levels is a realm they call "functional morality," defined as "intelligent systems capable of assessing some of the morally significant aspects of their own actions."

Wallach and Allen situate these levels on a grid measuring low and high levels of ethical

---

[6] I thank AI researcher and consultant Sergey Feldman for this and other insights (see below).
[7] Cliff Kuang, "Can AI Be Taught to Explain Itself?" *NY Times Magazine*, Nov. 17, 2017.
[8] Amitai Etzioni, Oren Etzioni, "AI Assisted Ethics," *Ethics Inf Technol* (2016) 18:149–156.
[9] Amitai Etzioni, Oren Etzioni, "Designing AI Systems that Obey Our Laws and Values," *Communications of the ACM*, (2016) Vol. 59 No. 9, Pages 29-31. See also, Oren Etzioni, "To Keep AI Safe—Use AI," *Recode*, Feb. 4, 2016.
[10] Cf. Etzioni ("AI Assisted Ethics"): "Autonomy in computer science refers to the ability of a computer to follow a complex algorithm in response to environmental inputs, independently of real-time human input."
[11] Wendell Wallach, Colin Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford UP, 2009). See also their chapter "Moral Machines: Contradiction in Terms or Moral Abdication?" in *Robot Ethics* (2012), where they update their argument and respond to some criticisms of their book. This chart is included in both texts.

sensitivity on the x-axis, and of autonomy on the y-axis. Systems that serve in a purely advisory capacity, for example in recommending medication dosages to physicians, have low autonomy. Autopilot systems that have operated airplanes for decades possess high autonomy but low ethical sensitivity, for they do not assess the moral consequences of their actions. They possess operational morality in that their programmers have integrated moral concerns (such as not injuring or killing humans) into the design of their operating parameters (such as not banking or changing altitude at a rate intolerable to most passengers, except to avoid a collision).[12]



Wallach and Allen, *Moral Machines,* Figure 2.1

Our greatest concern is with the development of AMAs that possess high levels of both autonomy and ethical sensitivity. The goal of fashioning machines endowed with high autonomy is already within reach, lending urgency to the project of developing ethical sensitivity within artificial systems before they can operate without human supervision. At Asilomar in 2017, the Future of Life Institute organized the drafting of 23 principles, signed by thousands of AI researchers and industry leaders, designed to support "beneficial intelligence," not "undirected intelligence."[13] Of course, even directed intelligence will not necessarily be beneficial, at least not for all parties affected.

Autonomous vehicles have driven many millions of miles on public roads in America, though most such vehicles operate in assist mode, rather than functioning without any human supervision.[14] The Society of Automotive Engineers has proposed 5 levels to describe such vehicles, from no automation to full automation, with levels 4 and 5 yet on the horizon.[15] Still, that horizon is approaching rapidly, and not only for cars. Uninhabited trains, boats and aerial vehicles already

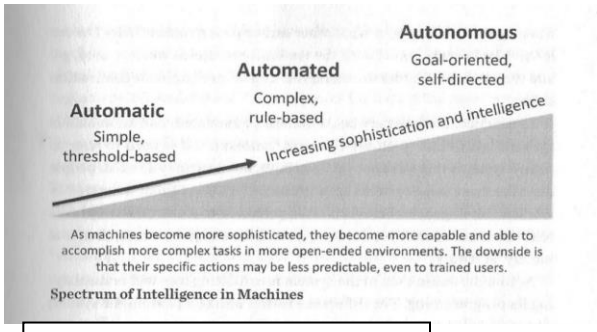
---

[12] The tragic crashes of Lion Air Flight 610 and Ethiopian Airlines Flight 302, Boeing 737 MAX 8 airplanes whose autopilot systems malfunctioned in 2018 and 2019, killing all aboard, remind us that even well-intended AI systems do not always produce moral results. Indeed, researchers discuss the "brittleness" of AI systems that lack contextual awareness to question and contain incongruous data in the way that humans naturally do.

[13] https://futureoflife.org/ai-principles/ See FLI president Max Tegmark's account of the history of this project in, *Life 3.0: Being Human in the Age of Artificial Intelligence* (Alfred A. Knopf, 2017).

[14] Waymo, the autonomous vehicle unit of Alphabet, claimed 10 million physical miles driven autonomously as of October 2018. Israeli competitor Intel Mobileye claims that safety can be demonstrated without such measures, which it deems wasteful, and indeed Waymo uses computer simulations virtually to drive ten million miles every day. Nevertheless, there are many "edge cases" that cannot be anticipated, and also the danger of hacking, either of the systems directly, or of essential environmental cues such as stop signs that can be compromised in ways discernible only to machines.

[15] https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety#issue-road-self-driving



Paul Scharre, *Army of None*, p.31

navigate the rails, waters and skies. While many are more "automatic" than truly "autonomous" (see chart below), the design trend is toward autonomous machines powered by artificial intelligence. Much conversation about the moral quandaries of autonomous vehicles focuses on variations of the famous trolley problem (discussed below), raising the question of whose life a system would prioritize in an accident situation—the occupant, a pedestrian, or passengers in a different vehicle.[16] Yet this discussion presumes beneficence. What about malevolent machines?

Autonomous weapons systems (AWS) are already employed on the land, sea and in the air by over 90 nations. Most systems are limited to surveillance functions, but some are equipped to react to threats with lethal force (LAWS).[17] There is a differentiation between systems in which human operators are "in the loop," required to authorize each use of force, those where humans are "on the loop," consulted during the operation with the ability to cancel an attack, and others where they are "out of the loop," meaning that people are informed only after the system has acted. Defense strategists speak of the OODA loop: first **O**bserve (search for targets), then **O**rient (detect targets), then **D**ecide (to engage the target) and then **A**ct (engage the target). If a human is required to confirm a target and appropriate action, then the system is semiautomatic.[18] Given that computers can process some data sets far faster than humans can, the temptation is great to move people out of the decision-making loop, much as generals may set objectives and assess outcomes while leaving the real-time process of deliberation and action to soldiers in the field of conflict.[19] This would lead to a fully autonomous system, even if humans continue to set goals and review results.

The Pentagon and CIA have spent billions of dollars developing increasingly powerful and

---

[16] See David Edmonds, *Would You Kill the Fat Man? The Trolley Problem and What Your Answer Tells us about Right and Wrong* (Princeton UP, 2015). MIT has a "Moral Machine" web site which introduces the topic and allows visitors to browse various scenarios and even play a related game: http://moralmachine.mit.edu/

[17] Ronald C. Arkin, *Governing Lethal Behavior in Autonomous Robots* (CRC Press, 2009). See also Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (WW Norton and Co, 2019), and his essay, "Killer Apps: The Real Dangers of an AI Arms Race" in *Foreign Affairs* (May/June 2019) 135-144. Israel is a leader in the field of autonomous weapons, with its Harpy system capable of airborne loitering for hours, identifying and attacking enemy radar installations, the Guardium uninhabited ground vehicle patrolling the Gaza border, and the Trophy Merkava tank defense system capable of automatically returning fire.

[18] See Scharre, *Army of None*, Chapter 3.

[19] US Department of Defense Directive 3000.09 (2012; rev. 2017) addresses autonomy in weapon systems in guarded language. Although military planners profess not to intend to relinquish the authorization of lethal force to AI, this could change in response to enemies acting without such scruples. See discussion in Scharre, Ch. 6.

autonomous systems;[20] an initiative called the Joint Artificial Intelligence Center (JAIC) was announced in June 2018. One focus of JAIC is said to be safety and ethics; nevertheless there are growing concerns from leading scientists, ethicists and leaders of the United Nations that AI weapons may violate international treaties.[21] Of special concern are moral norms such as the principle of distinction, which requires discernment between combatants and non-combatants, and the principle of proportionality, a concept designed to limit civilian harm in proportion to anticipated military gain.[22]

One of the strongest ethical defenses of autonomous systems—whether vehicles or weapons—is the *fallibility* of human operators. Driver error causes 90% of the over-30,000 vehicular deaths and many more injuries annually in the United States for all the familiar reasons—*drunkenness, drowsiness, distraction, rage* and *general poor judgment*.[23] Warfighters suffer from these same fallibilities; *fear* may further impair their judgment. Robotic warfighters presumably will lack these deficits, but they are likely to experience other challenges, especially in discerning intent, and in differentiating combatants from non-combatants. As noted, there has been little progress in equipping AI with understanding the meaning of a situation beyond the matching of data patterns. Will an armed robot be reliably able to discern between a child holding an ice cream cone and a soldier pointing a pistol? In its speed and efficiency will the system fail to allow ambiguous actions to be clarified?[24]

Perhaps this is merely a technical challenge; with improvements, autonomous systems might function with greater moral consistency than humans, making their use not only permissible but arguably mandatory.[25] AI may not comprehend *meaning*, but it can exceed human capacities for *situational awareness* in high intensity situations. Moreover, Ronald Arkin observes that humans do not

---

[20] See the DARPA video on "three waves of AI" by John Launchbury, for a sense of what is envisioned in future AI: https://youtu.be/-O01G3tSYpU.

[21] For a primer on AI that broaches concerns with LAWS, read the 2018 report of the United Nations Institute for Disarmament Research (UNIDIR), "The Weaponization of Increasingly Autonomous Weapons: Artificial Intelligence." As of April 2018, twenty-six countries had signed a call for the ban of fully autonomous weapons systems. Yet even some who signed the ban, such as China, were simultaneously announcing advances in technologies such as "swarms" of drones guided by AI. See Elsa Kania, "China's Strategic Ambiguity and Shifting Approach to Lethal Autonomous Weapons Systems," *Lawfare*, April 18, 2018.

[22] The four basic principles of the Law of Armed Conflict (LOAC) are: *distinction, proportionality, military necessity, and unnecessary suffering*. See https://loacblog.com/loac-basics/4-basic-principles/. For much more (>1200 pages) see the US DoD Law of War Manual, and the IDF Code of Ethics and Mission. See too Moshe Halbertal's blog at the Shalom Hartman Institute web site, "War and Ethics in the IDF Ethical Code."

[23] According to the National Highway Traffic Safety Administration, in 2016 there were 37,461 traffic fatalities in the United States, with almost a third involving alcohol-impaired-driving crashes.

[24] Scharre opens his book *Army of None* with an event in the USSR on Sep. 23, 1983 when sensors conveyed false information of five nuclear missiles incoming from the USA; a human operator sensed that something was amiss and sought additional radar confirmation before sending the alarm up the chain of command, possibly averting a nuclear apocalypse. Would an automated system have delayed responding in such a circumstance?

[25] Thanks to Yoni Brafman of JTS for this argument, which applies equally to non-military applications.

have an impressive record for moral conduct in war, and that autonomous weapons could function as a check on their baser instincts, preventing attacks that are not in compliance with international law.[26] The most highly regarded programs, such as the Aegis Combat System, feature integration between human and autonomous operation, benefiting from the relative strengths of people and machines in making life and death decisions.

Additional ethical concerns are triggered by the development of AWS, which have already expanded the dominance of the most powerful nations and corporations over their populations and adversaries.[27] Societies often struggle to balance between individual liberty and communal security. AI-powered systems equipped with facial recognition and video analytics may skew the balance further toward security, or perhaps more accurately, toward control by powerful governments and corporations over access to information and freedom of action accorded to people in their domain.[28] Intelligent video surveillance systems have been developed in China as a method for regulating the social conduct of the population in general and suppressing the Muslim Uighur population in Xinjiang Province in particular. China has found a ready export market in other nations interested in extreme forms of social control.[29]

General concerns of privacy are magnified by the expanded powers and prevalence of AI in society. Robots greatly expand surveillance powers, including in previously private spaces such as the home or even the body, given the ubiquity of smart phones and watches.[30] While there may not be a "right to privacy" in Jewish law, there certainly is a ban on "talebearing" (רכילות) and on injuring parties by causing them shame or other harm. AI systems often prompt people to take actions that are not in their own best interest, whether in recommending unnecessary purchases or influencing voting behaviors, as demonstrated in the 2016 American elections. Such interventions could match the Rabbis' understanding of not placing a stumbling block before the blind (ולפני עור לא תתן מכשול).[31]

Even beneficent AI applications such as health care assistants may raise troubling ethical questions. When a human health care provider such as a physician or nurse makes medical interventions, there is a basic sense of relationship that anchors the interaction, creating a presumption

---

[26] Paul Scharre, *Army of None*, 282-3, reporting an interview with Arkin on June 8, 2016.

[27] For an alarming discussion of the potential social, economic and political harms that artificial intelligence could enable, see Yuval Noah Harari, "Why Technology Favors Tyranny," *The Atlantic* (October 2018) and the chapter "Liberty" in his book, *21 Lessons for the 21st Century* (Spiegel & Grau, 2018). Shoshana Zuboff has dubbed this, "The Age of Surveillance Capitalism." Yet more alarming is the 2017 Slaughterbots video, a fictional short film that illustrates the moral perils of LAWS and advocates halting the development of autonomous weapons.

[28] See Jay Stanley, "The Dawn of Robot Surveillance: AI, Video Analytics and Privacy," published by the ACLU, June 13, 2019.

[29] Paul Scharre, "Killer Apps," 138-9.

[30] See M. Ryan Calo, "Robots and Privacy," chapter 12 in *Robot Ethics* (2014).

[31] I thank Micah Peltz for suggesting this connection, based on b. Avodah Zarah 6a/b and Sifra Kedoshim, 2:2.

of concern and responsibility. Systems powered by artificial intelligence may provide valuable information and analysis of risks and benefits, and their interface may mimic patterns of speech and non-verbal behaviors that approximate expressions of concern, but we generally consider such systems to be tools, not responsible caregivers. Yet already there are trends in advanced economies to rely on such systems in place of humans, whose assistance may be costlier, less readily available, and perhaps less reliable. Indeed, the use of embodied or animated conversational agents has become common in medical settings, and recent meta-analysis points to improvements in outcomes when well-designed virtual agents are used.[32] Such systems are designed to educate patients and monitor behaviors, but they are not empowered to make a diagnosis or determine the course of therapy, much less assess ethical dilemmas. Looking ahead, might virtual agents be compared to human health care proxies that make treatment decisions on behalf of a patient, or should there always be a human supervisor to finalize matters of life and death?[33]

An additional area of concern relates to fairness in machine learning.[34] Bias is not inherently problematic; indeed, the ability to discriminate between different subjects and situations is the foundation of learning. Our concern is with discrimination that leads to an unjustified bias or one which has been declared morally irrelevant.[35] Fairness in machine learning is necessary to prevent the integration of social biases by algorithms charged with assisting on momentous decisions, from medical therapies to the approval of home loans, insurance policies, and even petitions for parole.[36]

---

[32] See "Virtual Humans in Health-Related Interventions: A Meta-Analysis," by T Ma, H Sharifi, D Chattopadhyay, conference paper, May 2019. In "Managing Chronic Conditions with a Smartphone-based Conversational Virtual Agent," (*IVA* 18) Michael K. Paasche-Orlow, Jared W. Magnani, *et al,* conclude, "Given that chronic diseases affect such a large portion of the population, and the complexity of self-care management regimens—especially for patients with low health literacy— virtual agents represent an important tool for improving population health and decreasing healthcare costs." I thank Michael for explaining the use of virtual health agents.

[33] I thank Toby Schonfeld for suggesting this analogy to proxy decision making.

[34] For a detailed (and in the second half, highly technical) introduction, see the 2017 video tutorial, "Fairness in Machine Learning" by Solon Barocas and Moritz Hardt, https://vimeo.com/248490141. I thank Sergey Feldman for the reference and for flagging this as a topic of moral concern.

[35] For example, social biases may be anchored in fact patterns that are accurate, and yet those patterns are themselves reflections of bias such as gender or racial discrimination. Barocas discusses two discourses in American discrimination law, *disparate treatment*, and *disparate impact*. The former relates to procedural fairness and equal opportunity, while the latter considers larger social structures that may lead to unequal outcomes and promotes options for distributive justice. We will discuss models of fairness in halakhic discourse and their implications for directing machine learning below. The question of which biases are erroneous, which are accurate but morally unacceptable, and which are defensible, is an endlessly contentious matter which plays out for example in discussions of affirmative action. Defensible bias has specific dimensions within halakhic discourse, with its manifold distinctions among actors and actions.

[36] Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Broadway Books, 2017). See also Dhruv Khullar, "AI Could Worsen Health Disparities," *NY Times,* Jan. 31, 2019. Regarding loan applications see this discussion of "threshold classifiers" and how they can be made less biased:

What can be done to make machines operate on a level that is at least as moral as that of beneficent humans, avoiding the pitfalls that lead people to misjudge and discriminate against one another? Given the lack of consensus about ethical principles and the pressing need to develop guidelines for the conduct of autonomous machines equipped with artificial intelligence, it is imperative that wisdom traditions such as Judaism contribute to the general discourse, just as it has participated in discussions of bioethics and other matters of public policy.[37]

Classical statements of rabbinic law range between extremely narrow and extraordinarily broad subject audiences. Many Talmudic norms are directed particularly at Jews who are adult, free, male, educated, physically and mentally able, and even members of the rabbinic or priestly elite. Such people were classically considered as *bar ḥiyuva,* most fully obligated and therefore fully governed by halakhah. Yet even classical rabbinic law applies in varying degrees to people of diverse ages, genders, religions and other categories. Indeed, Jewish law also addresses animals, inanimate objects, and features of nature, encompassing within its scope the great chain of being.[38] Progressive halakhists have sought in recent decades to articulate a more expansive and inclusive statement of the law. Can the halakhah be applied even to artificial agents? A good starting point is the rabbinic concept of agency (שליחות), specifically the appointment of agents whose status differs substantially from the principal who appointed them.

II. *Shliḥut*: What if a person's agent is not like them?

In determining the contours of legal agency, the Rabbis consider differences of age, gender, religious and social status between the principal (שולח) and agent (שליח). They also discuss the use of non-human actors (e.g., an elephant or a monkey), and inanimate objects, such as a walled courtyard, to extend the reach of a human. What is the significance of status difference between principal and agent? In what circumstances is the principal responsible for harm or transgressions performed on their behalf by the agent? Do the terms of agency depend on the social status, cognitive state or autonomy of the agent?

The foundational source for the laws of agency is Mishnah Kiddushin 2:1; it is developed further in both the Yerushalmi and the Bavli (41a-42b).[39] This Mishnah states that a man may betroth a

---

https://research.google.com/bigpicture/attacking-discrimination-in-ml/.

[37] While we frequently describe halakhah as "Jewish law" it is quite rare for halakhic rulings to be enforced with means comparable to the law enforcement methods of governments and their police forces, even in Israel. That said, it is not uncommon for religious discourse to influence the development of civil law, especially in areas that are unsettled or controversial. Of course, for observant Jews, halakhic guidance plays an even more direct role in their interaction with these new technologies.

[38] See Mira Beth Wasserman, *Jews, Gentiles and Other Animals: The Talmud after the Humanities* (Penn UP, 2017), 33, 166, 234-235.

[39] לסקירה רחבה של הנושא ראו נחום רקובר, **השליחות וההרשאה במשפט העברי** (מוסד הרב קוק, 1972/תשל"ב).

woman either directly or through use of an agent, and that a woman may accept an offer of betrothal either directly or through use of an agent. There are several curious features to the wording of this Mishnah which occupy the sages in the Bavli and lead to important rulings, such as the prohibition of marrying off a child (קטנה) until she is old enough to consent to a specific marriage offer.[40] This *sugya* establishes that men and women (but not children) may appoint agents, and that agents may function in various legal capacities for their principal: betrothal and divorce, tithing, and also commercial transactions (as in the purchase of a paschal lamb on behalf of a group). Finally, the Talmud generalizes the rule, claiming that, "a person's agent is [legally considered to be] like him" (ששלוחו של אדם כמותו):

> דא"ר יהושע בן קרחה: מנין ששלוחו של אדם כמותו? שנאמר: ושחטו אותו כל קהל עדת ישראל בין הערבים, וכי כל הקהל כולן שוחטין? והלא אינו שוחט אלא אחד! אלא מכאן, ששלוחו של אדם כמותו.[41]
>
> For Rabbi Yehoshua b. Karḥah states, what is the source that shows that a person's agent is [legally considered to be] like him? For it says (Exodus 12:6, regarding the paschal offering): *the entire community of the congregation of Israel will slaughter it at dusk.* Could it be that the entire community slaughtered? Wasn't it only one person who slaughtered [the lamb]? Rather, this shows that a person's agent is [legally considered to be] like him.

This legal principle had previously been cited in Tannaitic literature, including Mishnah Brakhot, Tosefta Ta'anit, and Mekhilta D'Rabbi Yishmael, *et al*.[42] It is assumed to apply "in all situations" (בכל מקום), except when Scripture implies restriction of a significant action to the principal.[43] This rule is widely attested in the Bavli, but several restrictions are added to its application.

In the continuation of b. Kiddushin 41b, the sages consider the implications of Numbers 18:28

For a comparison to *mandatum* in Roman law, see Barry Nicholas, *An Introduction to Roman Law* (Oxford: Clarendon Press, 1962, rev. 2008), pp.187-89.

[40] **בבלי קידושין דף מא עמוד א.** האיש מקדש את בתו כשהיא נערה. כשהיא נערה אין, כשהיא קטנה לא; מסייע ליה לרב, דאמר רב יהודה אמר רב, ואיתימא רבי אלעזר: אסור לאדם שיקדש את בתו כשהיא קטנה, עד שתגדל ותאמר בפלוני אני רוצה.

[41] שם, עמוד ב.

[42] **משנה מסכת ברכות פרק ה, משנה ה.** המתפלל וטעה סימן רע לו ואם שליח צבור הוא סימן רע לשולחיו מפני ששלוחו של אדם כמותו. **תוספתא מסכת תענית (ליברמן) פרק ג.** צו את בני ישראל ואמרת אליהם את קרבני לחמי אי איפשר לומ' כל ישראל אלא מלמד ששלוחו של אדם כמותו. **מכילתא דרבי ישמעאל בא - מסכתא דפסחא פרשה ג.** ויקחו להם. וכי כלן היו לוקחין אלא לעשות שלוחו של אדם כמותו מכאן אמרו שלוחו של אדם כמותו:

[43] **מכילתא דרבי ישמעאל משפטים - מסכתא דנזיקין פרשה ב.** ורצע אדוניו את אזנו. למה נאמר, לפי שמצינו בכל מקום ששלוחו של אדם כמותו, אבל כאן הוא ולא שלוחו. **ספרא מצורע פרשה ה.** אשר לו הבית שלא ישלח ביד שליח יכול אפילו זקן, ואפילו חולה, תלמוד לומר ובא והגיד לכהן, ידקדק הכהן כיצד בא הנגע לביתו. **ספרי במדבר פרשת מטות פיסקא קנג.** אמר לאפוטרופוס כל נדרים שתהא בתי נודרת מיכן ועד שאבוא ממקום פלוני הפר לה והפר לה שומע אני יהיה מופר ת"ל ואם הניא אביה אותה ואם היפר אביה מופר ואם לאו אינו מופר דברי ר' יאשיה, ר' יונתן אומר מצינו בכל מקום ששלוחו של אדם כמותו.

for the use of an agent to separate tithes.[44] The verse emphasizes, "even you," which the Rabbis understand to allow for a Jewish—but not a non-Jewish—agent in the act of tithing:

> ואלא אתם גם אתם למה לי? מיבעי ליה לכדרבי ינאי, דא"ר ינאי: גם אתם - מה אתם בני ברית, אף שלוחכם בני ברית.
>
> Then [it could have stated,] *you*. Why was it needed to state, *even you*? It was needed according to the statement of Rabbi Yannai, for Rabbi Yannai said: *Even you*—just as you are parties to the covenant [i.e., Jewish], so must your agents be parties to the covenant.

To be clear, Rabbi Yannai's restriction is contextual—to the religious obligation imposed on Israelite farmers to tithe for the Temple. Because non-Jewish farmers had no such obligation, they could not serve as agents for this task.[45] This might imply that in other situations non-Jews *could* serve as agents, and indeed there are a few such cases from the realm of financial law.[46] Still, the general principle is that a Jew may not appoint a non-Jew as an agent (אין שליחות לעובד כוכבים), because gentiles are not obligated to observe the mitzvot of Jews.[47] Sometimes this ruling is restated as, "for gentiles have no agency" (דלאו בני שליחות נינהו), but this phrasing is imprecise, as noted by Rabbi Moshe Shochet (1878-1947) in his Mishneh Torah commentary: obviously gentiles do serve as agents *for each other*, and in this way their status is not to be compared to people who are physically or intellectually incapable of such service.[48] Moreover, a responsum of Rabbi Moshe Sofer (1762-1839) claims that when gentile law parallels Jewish law, then the action of a gentile agent may be accepted.[49]

---

[44] **במדבר פרק יח פסוק כח.** כֵּן תָּרִימוּ **גַּם אַתֶּם** תְּרוּמַת ה' מִכֹּל מַעְשְׂרֹתֵיכֶם אֲשֶׁר תִּקְחוּ מֵאֵת בְּנֵי יִשְׂרָאֵל וּנְתַתֶּם מִמֶּנּוּ אֶת תְּרוּמַת ה' לְאַהֲרֹן הַכֹּהֵן:

[45] **רמב"ם הלכות תרומות פרק ד הלכה א.** עושה אדם שליח להפריש לו תרומות ומעשרות שנאמר כן תרימו גם אתם לרבות שלוחכם ואין עושין שליח עכו"ם שנאמר גם אתם מה אתם בני ברית אף שלוחכם בני ברית.

[46] **טור חושן משפט הלכות גביית מלוה סימן צז.** כ"כ רב אלפס בתשובה שאם הלוה אלם ורוע מעללים ומעיז פניו מלפרוע חובו מותר ליכנס לביתו לדעת מצפוניו אפי' ע"י שליח עכו"ם.

[47] **טור חושן משפט הלכות שלוחין סימן קפח.** א אין שליחות לעובד כוכבים לא שיעשה הוא שליח לישראל ולא שיעשנו ישראל שליח אבל אשה ועבד ושפחה בני שליחות הן כיון שחייבין בקצת מצות ויש להן דעת אבל חרש שוטה וקטן כיון שאין להן דעת לאו בני שליחות הן לא שיעשו הם שלוחין ולא שאחרים יעשום שליח: **וראו בדרכי משה שם, אות (א).**

[48] **הד משה הלכות זכיה ומתנה פרק ד.** ולכאורה יש סברא לחלק בין עכו"ם ובין חשו"ק בדין שליחות לחומרא משום דבעכו"ם איתא בהו דין שליחות בעלמא בין בהמשלח ובין בהשליח היינו בשליח עכו"ם עבור ישראל דבהמשלח הישראל איתא ביה דין שליחות וגם בעכו"ם השליח יש בו דין שליחות היינו בעכו"ם לעכו"ם רק דעכו"ם לישראל ליתא דין שליחות מדאורייתא אבל כל חד באפי נפשיה יש בהם דין שליחות.

[49] Ḥatam Sofer (#90) writes in a responsum that if one used a gentile agent to lead an animal and thus acquire it, this is effective (מועיל) since they have similar laws of agency (his word קאמיסיאנע seems to come from "commission" and mean a type of appointment): **שו"ת חתם סופר קובץ תשובות סימן צ.** ומטעם זה נ"ל לקיים לדינא אידך דתשובת משאת בנימין דבמשך על ידי שליח עכו"ם מועיל משום דבדיניהם הוה קאמיסיאנע כמותו, אף על גב דמשאת בנימין בעצמו לא נחית לכך.

There is also a distinction between pure agency, which is set up as a voluntary act, and employment in which the servant's hand is an extension of the hand of the boss (יד עבד כיד רבו). A gentile may, for example, transport an object for a Jew, even one of ritual significance such as a *get*.[50] The codifiers are most exclusive of gentile proxies in ritual contexts such as the sale of *ḥametz* on Passover, and most inclusive when it comes to business practices that are shared between Jews and gentiles or when gentile workers are hired for the task. As Rabbi Shneur Zalman Fradkin of Lublin (1830-1902) explains, gentiles may conduct business for Jews, but may not complete mitzvot on their behalf.[51]

Another helpful distinction is illustrated at b. Eruvin 31b in a discussion about using an agent to deliver a food item that will be placed in a location to extend one's Shabbat boundary:

> דתניא: נתנו לפיל והוליכו, לקוף והוליכו - אין זה עירוב. ואם אמר לאחר לקבלו הימנו - הרי זה עירוב. - ודילמא לא ממטי ליה? - אמר רב חסדא: בעומד ורואהו. - ודילמא לא מקבל ליה מיניה? - אמר רב יחיאל: חזקה שליח עושה שליחותו.
>
> For it is taught in a *beraita*: If they gave it [i.e., an item to establish an extension of the Sabbath boundary] to an elephant, and it carried it, [or] to a monkey and it carried it—this is not a [valid] *eiruv.* But if he said to another [person] to receive it from him [i.e., the animal], then it is a [valid] *eiruv*. But what if it [the animal] does not deliver it? Rav Ḥisda says, the case is when the [principal] stands and watches. And what if [the human receiving agent] refuses to accept it from him [the animal]? Rav Yeḥiel says, it is established that an agent completes his appointed task.

Generally, the agent must themselves be like the principal—Jewishly obligated and accepting of the rabbinic rule about Sabbath boundaries. For example, a Samaritan who rejects the rabbinic law of boundaries may not serve as an agent for this purpose. This accords with what we learned in the previous paragraph. But if there is a second *receiving* agent who is equally obligated, then the first *messenger* agent need not be committed to the religious content of the task. Indeed, the item could even be placed on *an elephant or a monkey* to convey, so long as the principal can observe the transfer, and the receiving agent is in place.

---

[50] ע' **שו"ת להורות נתן חלק ו סימן צו.** בזה י"ל דאע"ג דאין עכו"ם נעשה שליח לישראל מ"מ פועל עכו"ם של ישראל שפיר נעשה שליח, דהרי הא דאין העכו"ם נעשה שליח לישראל אין החסרון בישראל המשלח, שהרי הישראל בעצמו היה יכול לעשות מעשה זה, אלא שהחסרון הוא שאין העכו"ם יכול לקבל כחו של ישראל המשלח, דהתורה לא ריבתה דין שליחות בעכו"ם, ואשר מהאי טעמא אין העכו"ם יכול לעשות מעשה זה עבור הישראל.

[51] **תורת חסד אורח חיים סימן מה.** ואמנם כבר ביארנו לעיל דאף אם נימא לדינא סברת המח"א הנ"ל דבפועל עכו"ם יש שליחות משום דידו כידו מ"מ כ"ז לענין ממון ולא לענין מצות ואיסורין. וכד' הראב"ד והתוס' לענין יד עבד כיד רבו וכנ"ל. ואין לברך כשעושה מעקה ע"י פועל נכרי: **וע"ע שו"ת ציץ אליעזר חלק יט סימן סד.** ומכח האי טעמא דפועל יום שאני דגופו קנוי לבעה"ב לגמרי וידו כיד בעה"ב ממש. ס"ל להמח"א שם לפני כן בדבריו דאפילו אם עשה את המעקה ע"י פועל יום עכו"ם דג"כ הבעה"ב מברך עליו מפני דחשיב כאילו הוא בעצמו עשאה. ואליבא דהפר"ח ס"ל להמח"א שם דבכל גוונא בעה"ב מברך ואפילו אם עכו"ם שלוחו (ולא פועל יום) עשה את המעקה כדיעו"ש.

This ruling establishes a precedent for intermediate steps which are not considered to be true agency but are mechanical. In later halakhic literature this comes to be called מעשה קוף (act of a monkey) and serves to justify, for example, using a postal or courier service to send a bill of divorce to a location where a second agent will receive it.[52] The carrier is not a true agent in the sense of legally representing the sender. In other words, both at the outset and at the conclusion of the legal act, there must be a properly qualified actor (whether principal or agent) who can bear full responsibility, but in the middle, it is permissible to use animate or mechanical means of conveyance.

From the first example of tithing and the subsequent ones regarding Sabbath boundaries and bills of divorce, it appears that there is a higher standard of agency for ritual actions than for monetary transactions, but even ritual acts may use a non-obligated agent for an intermediate step such as transportation. *We may infer that if a ritually obligated individual both initiates and completes an action, then a non-obligated agent, and even a non-human agent, might be permitted to carry out an intermediate segment of the task.*

A second important restriction on agency is the rule, "there is no agency for transgression" (אין שליח לדבר עבירה). For example, if A says to B, "Go get me an ox from the barn of C," and B goes and steals C's ox, B is responsible for their own action and cannot claim to have simply followed A's orders. As the Talmud puts it, "If the master speaks, and a student speaks, to whom do we listen?"[53] God is the master, and the agent remains responsible to obey divine law, regardless of the instructions of the principal. Agency breaks down at the border of transgression. When a principal appoints an agent for a legal act, then the "agent is like the principal." But if it turns out that the action is illegal, then the agent is liable for their own misdeed (at least if the agent shares the same halakhic obligations as the principal). However, the principal remains morally culpable for leading their neighbor into transgression.[54]

---

[52] **שולחן ערוך אבן העזר הלכות גיטין סימן קמא סעיף לה. וראו ברא"ש מסכת גיטין פרק ב...**אבל הכא כיון דלא עשה העובד כוכבים אלא מעשה קוף בעלמא ומינה הבעל בכתבו את ישראל שבאותו מקום שליח למה יפסל הגט וכן נוהגין באשכנז ובצרפת ע"פ דברי ר"ת לשלוח גט וקידושין על ידי עובד כוכבים וממנה בכתבו ישראל שבאותו מקום שליח. **וע"ע ברבינו ירוחם - תולדות אדם וחוה נתיב כד חלק ג דף רז טור ד,** מסירה לא תנשא לכתחלה משום דגוי לאו בר כריתה הוא ואם נשאת לא תצא משום דרבי שמעון דאמר במתניתין כלם כשרים ואמרי ירד רבי שמעון לשטתו של רבי אליעזר דאמר עדי מסירה כרתי וקיימא לן כרבי אליעזר ע"כ. ורבי' תמהו על דברי הגאון כי אינו פסול אלא כשמינהו ישראל שליח במקומו אבל כשהגוי לא עשה אלא מעשה קוף בעלמא ומינה הבעל בכתבו את ישראל שליח ליתן לה גט ושולח הגוי ליתן לישראל שבאותו מקום מה שהוא שולח לו לא יפסל הגט וכן נהגו באשכנז וצרפ' על פי ר"ת לשלוח גט וקדושין ביד גוי וממנה שליח בכתבו לישראל שבאותו מקום.

[53] **בבלי קידושין דף מב עמוד ב.** והא דתנן: השולח את הבעירה ביד חרש שוטה וקטן - פטור מדיני אדם וחייב בדיני שמים, שילח ביד פיקח - פיקח חייב; ואמאי? נימא: שלוחו של אדם כמותו! שאני התם, דאין שליח לדבר עבירה, דאמרינן: דברי הרב ודברי תלמיד - דברי מי שומעים?

[54] **רמב"ם חובל ומזיק פרק ה.** האומר לחבירו שבר כליו של פלוני על מנת שאתה פטור ועשה הרי זה חייב לשלם, וכאילו אמר לו סמא עינו של פלוני על מנת שאתה פטור, ואף על פי שהעושה הוא החייב לשלם הרי זה האומר לו שותפו בעון ורשע הוא שהרי הכשיל עור וחיזק ידי עוברי עבירה. **ובדברי רמא בשולחן ערוך חושן משפט הלכות שלוחין סימן קפב סעיף א:** הגה: בכל דבר שלוחו של אדם כמותו, חוץ מלדבר

This rule sounds reasonable but could lead to absurd or even mischievous results, as the Sages realized (not even anticipating AWS). Should a child or a slave be held liable for obeying a parent or a master? Could a courtyard? After all, rabbinic law allows a person to claim ownerless property based on it landing in their domain—within 4 cubits of their body in a secluded space, or in their courtyard—even without the person establishing physical contact.[55] But what if the "claimed" object is not truly ownerless? Can a person be held responsible for the acquisition committed by his courtyard? Might an inanimate object be considered an agent?

At b. Bava Metzia 10b, the sages entertain the possibility that a courtyard could be considered an agent of a person, for example in acquiring ownerless property for them or receiving a bill of divorce. As Rashi explains, [then] "her courtyard could be like her agent."[56] Yet, this position is rejected. Why? Amoraim Ravina and Rav Sama limit the transfer of liability from principal to agent in cases where the agent has little or no autonomy, with each offering a different explanation for the limitation:

> אמר רבינא: היכא אמרינן דאין שליח לדבר עבירה - **היכא דשליח בר חיובא הוא**, אבל בחצר דלאו בר חיובא הוא - מיחייב שולחו. - אלא מעתה, האומר לאשה ועבד צאו גנבו לי דלאו בני חיובא נינהו הכי נמי דמיחייב שולחן? - אמרת: אשה ועבד בני חיובא נינהו, והשתא מיהא לית להו לשלומי. דתנן: נתגרשה האשה, נשתחרר העבד - חייבין לשלם.
>
> Ravina says, when we said that "there is no agency for a transgression," that was **only when the agent themselves was obligated** [for that transgression]. But as for a courtyard, which is not itself obligated, the principal is liable. If so, when a man tells his wife or slave, "go steal for me," since they are not obligated to pay [the "double" penalty] shall we say that the principal is liable? You could say, wives and slaves are [after all] responsible [not to steal] but are not obligated [to pay the fine for theft, since they do not control their own assets]. For it is taught in a Mishnah, if the woman is divorced or the slave is freed, then they become liable to pay [their own fines].

> רב סמא אמר: היכא אמרינן אין שליח לדבר עבירה - **היכא דאי בעי עביד, ואי בעי לא עביד.** אבל חצר, דבעל כרחיה מותיב בה - מיחייב שולחו.
>
> Rav Sama says, when we said that "there is no agency for a transgression," that was

---

עבירה דקיימא לן אין שליח לדבר עבירה (טור). ודוקא שהשליח בר חיובא, אבל אם אינו בר חיובא הוי שליח אפילו לדבר עבירה (הגהות מיימוני פרק ה' דשלוחין).

[55] The halakhah also permits serving a bill of divorce by throwing it at the feet of a woman based on, "her four cubits acquire for her." This practice is not permitted, however, if the wife is a child. Why the child's personal space or courtyard doesn't acquire her divorce is the presenting issue of our *sugya*. There is tension between the tannaitic source, which assumes that a very young girl may be betrothed, and the amoraim who follow Rav's dictate (discussed above) that the girl must be old enough to give informed consent.

[56] **רש"י מסכת בבא מציעא דף י עמוד ב**. משום שליחות איתרבאי - מדרבי רחמנא שליחות לאדם, כדתניא (קידושין מא, א) ושלח - מלמד שהאיש עושה שליח, ושלחה - מלמד שהאשה עושה שליח, אתרבאי נמי חצרה, דהויא לה כשלוחה.

> **only in the case when if [the agent] wanted, he acted, and if [the agent] didn't want, he didn't have to act.** But as for a courtyard, where items are placed without its consent, the principal is liable. [Emphasis added in both quotes]

Before proceeding, we must pause to acknowledge the offensiveness of the common comparison of women to slaves in a patriarchal society, and of the ancient permission of slavery altogether. We further acknowledge the painful reality that though most nations no longer sanction slavery, and though they may have enacted formal gender equality in many realms of law, no society is fully egalitarian, and various forms of slavery remain prevalent throughout the world, including in democratic countries.[57] For the Talmud, a stratified social structure is assumed, and the presenting question regards the implications of such hierarchies for the legal institution of agency, not the (in)justice of such structures.

Instead, the editor ponders the practical distinction between the positions of Ravina and Rav Sama, proposes some cases where they might yield different results, but concludes that in the end their positions are compatible. The rule that "there is no agency for a transgression" applies only when the agent possesses the ability to resist the mission, and to pay consequences for transgression. In the end, the courtyard is not really an agent. *Obligation is an essential qualification for agency, and an inanimate object cannot become an agent.*

A related discussion is found in Midrash Mekhilta[58] regarding a person "who sends their livestock to graze in another's land" (Exodus 22:4). The verb for "send" (ושלח) is read by the Rabbis to imply that the owner used an agent to perpetrate this misdeed. The Midrash says,

> ושלח את בעירו. מכאן אמרו מסר צאנו לבנו לשלוחו ולעבדו פטור, לחרש שוטה וקטן חייב.
>
> *He sends his livestock.* From here they said, if he handed his sheep to his son, to his agent, to his servant/slave, then he [i.e., the owner] is exempt, but [if he handed his sheep] to a person who is deaf-mute, mentally ill or a minor, then he [i.e., the owner] is liable.

The Midrash here distinguishes between an agent who has the capacity for independent judgment, and therefore is responsible for their own actions, and an agent who does not have such ability, and is therefore not held responsible.[59] The Sages have established a distinction between agents who are

---

[57] E. Benjamin Skinner, *A Crime So Monstrous: Face to Face with Modern Day Slavery* (Free Press, 2008).

[58] **מכילתא דרבי ישמעאל משפטים - מסכתא דנזיקין פרשה יד. וע' דברי התורה תמימה לשמות כה, ד הערה מה)** כך משמע הלשון ושלח שעזבו לנפשו בלא השגחה ושמירה, ולכן אם מסרו לחרש שוטה וקטן הרי עזבו מרשותו לרשות הפקר שלא בהשגחה לכן חייב, משא"כ אם מסרו לפקח פטור הוא וחייב השליח, משום דהרי שמור הוא תחת יד הפקח ואפילו שלחו לא שייך לומר שהפקח עושה שליחותו של המשלח, משום דקיי"ל אין שליח לדבר עבירה, ועיין בסוגיא ב"ק נ"ו א':

[59] This midrash may be harmonized with the Bavli source cited above exempting slaves from the "double" fine for stealing; the slave is responsible not to steal but lacks independent financial resources to pay a fine. The designation of the deaf-mute (חרש) as legally incompetent has been studied and modified by Rabbi Pamela

independent moral actors, and those who are not.

This rabbinic discourse around agency and crime has implications for our discussion of machine ethics. When we say that machines are functioning autonomously, currently we mean this in a very limited sense. Machines are given a task and the capacity to complete the task within certain parameters, usually by following algorithms built on a series of predetermined "if…then" rules. They are not capable of establishing independent goals or refusing to act on orders that fall within their operational parameters. Nor are they accorded legal personhood, no matter how personal people may get in conversations with virtual assistants.[60] Just as it would be absurd to punish a courtyard for "stealing" a goat, so would it be absurd to whip an autonomous vehicle in punishment for "murdering" a pedestrian. Legal standing and free will are essential components to moral stature and liability.[61] At this stage, artificial intelligence functions like a tool, and so moral liability must remain with the principal. Or perhaps the machine is more like an animal, in which case its owner is responsible to a greater or lesser extent depending on typical performance. In any event, the machine is not obligated (בר חיובא), as Ravina puts it, nor does it have free will (דאי בעי עביד), as Rav Sama emphasizes. *Without these capacities, liability remains with the principal who appointed the agent—the person, not the machine.*

Even so, questions remain. Who is the principal? Those who designed the system? Its manufacturer? The vendor, or the end-user? Liability for an action conducted by an autonomous system should reside with the person most responsible for this specific action. If a user instructs an

---

Barmash, "Status of the Heresh and of Sign Language," (CJLS, approved May 24, 2011).

[60] The emerging field of affective computing equips machines to identify the emotional state of human users and to respond in kind, either by modulating the tone of voice or suggesting appropriate actions. See for example, the MIT Media Lab. It seems that the interface shift away from keyboards and screens toward voice and visual representations of virtual agents leads people to interact with digital devices differently than with either mute machines or live persons. Judith Shulevitz discusses such potentially disturbing trends in AI design in, "Alexa, how will you change us?" *The Atlantic*, Nov. 2018. See also Nicholas A. Christakis, "How AI will rewire us" *The Atlantic,* April 2019.

[61] Of course, free will is very much contested in contemporary studies of human behavior, much as בחירה חפשית was in ancient Jewish sources starting with Rabbi Akiva's claim in m. Avot 3:15, "All is foreseen, but choice is given," and on to Maimonides in his "Eight Chapters" (esp. Ch. 8). Neuroscientist Robert Sapolsky argues in *Behave: The Biology of Humans at our Best and Worst* (Penguin, 2017) that there is no independent part of us (a homunculus, or a soul) that makes decisions free from the influence of biology. In contrast Michael Gazzaniga describes a complex decision-making process of mind emerging from the biological structures of the brain and the social structures surrounding the person to yield an "interpreter" that approximates autonomy. See *Who's in Charge? Free Will and the Science of the Brain* (Ecco, 2012). We certainly have the *perception* of human free will, and both religious and secular legal systems assign individuals responsibility based on that presumption. Michael Graziano discusses free will (200-202), concluding, "We frequently act without any conscious knowledge of why, and then make up false reasons to explain it. Consciousness is hardly the sole controller of behavior. But in [my] current theory, consciousness is at least one part of the control process."

autonomous vehicle to convey them to a location, and the vehicle causes damage or injury on the way, the vehicle is not a free agent and does not have legal or moral standing (unlike a human driver).[62] A person should be responsible, but who? With consumer products in general there is overlapping responsibility between the manufacturer, who must build a reliable product, and the user, who must maintain the product and use it according to directions. What is different here is that AI-directed machines are designed to operate in unpredictable circumstances. Still, humans design, manufacture, license, sell, purchase and use these machines—surely there must be a doctrine of joint or several liability to account for damage and assign fault to one or more responsible humans.

What about future iterations of artificial intelligence? Recalling the schema of Wallach and Allen, what if future machines progress from operational to functional morality? Some researchers (and many science fiction writers) anticipate artificial *general* intelligence, which would not merely complete delimited tasks, but could pursue more abstract goals (e.g., fighting crime) using contextual reasoning. Just as human neural networks have been used as a model for developing artificial neural networks, so might moral values be used as a model for training machines to differentiate and prioritize goals. Imagine, for example, an AI assistant that has observed that a user is an alcoholic and thus declines to transport them to a liquor store unless there is consultation with another family member, physician (or parole officer). Such pairings of AI with autonomous machines might eventually qualify as "full moral agency." At that point might machines be deemed persons, responsible for their own conduct, expected to comply not only with laws, but also with moral values? We will return to this question below.

Currently autonomous machines do not possess well-developed capabilities for moral reasoning. We have suggested that people who request their services be held responsible for their actions, but is this finding reasonable when the implementation of commands is not understood or minutely controlled by the person? When autonomous vehicles cause damage, should full liability be assigned to their occupants? Such occupants may not understand, much less direct, the algorithms guiding their vehicle. They will not truly be drivers, but rather passengers. Is it reasonable to hold them liable? For this question, we turn to the rules of damages (נזיקין).

III. Limited Liability for Indirect Actions

Beginning with the Torah (Exodus 21:25-36 and 22:4-5) Jewish law holds owners responsible for damage caused by their property, whether animate (such as a donkey or an ox) or inanimate (such as a pit or a fire). There are many variables that can mitigate or amplify liability, such as whether the damage was intentional or accidental, direct or indirect, predictable or surprising, whether due caution had been exercised, and whether freak circumstances conspired to cause harm. These subjects are discussed extensively in the first six chapters of Bavli Bava Kama.

[62] See next section for discussion of limited liability for indirect damage.

It is well established in Jewish law that people are partially responsible for damage caused by means of an animal or tool that is under their domain.[63] It is also established that people bear full responsibility for damage caused by actions they have initiated, such as shooting an arrow, even once the object has departed from their control. This category of damage is known as "by his force," (בכוחו). But there is a third category of damage caused by a sequence of events that was initiated by a person, but which proceeded in unpredictable ways. For example, a person throws a stone, which ricochets and hits another stone, which loosens and falls, damaging property in the process. This is called "by force of his force" (בכוח כוחו), and is more controversial, with differing opinions about liability.[64]

A major distinction made by the Sages is between animate and inanimate property.[65] This distinction plays out in many areas of halakhah, such as whether animate property (say, a tethered donkey) might be used as a wall for a Sukkah, a post for a Sabbath border, or as parchment for a writ of divorce. In the laws of damages, a further distinction is made between behavior expected of animals (which increases liability for the owners) and unexpected behavior (which limits owner liability).[66] How shall we regard the conduct of autonomous machines—like inanimate property which might move and cause damage (as with a fire), or like an animal that moves of its own volition, and may surprise its owner with unexpected behavior? Should autonomous machines powered by artificial intelligence be regarded as "alive" at least to the extent of animals? And if so, should this comparison narrow or expand the scope of owner liability?

In *Teḥumin*, an Israeli journal of society and halakhah, Rabbis Yosef Sprung and Yisrael Meir Malka discuss liability for the actions of autonomous vehicles in light of the laws of damage.[67] Rabbi Avraham Yeshaya Karelitz (1878-1957) had declared operators liable for the actions of machines such as cars or motorized plows that they set in motion, even if they had removed their hands from the controls, since they had clearly caused the machine to operate.[68] Yet Sprung and Malka note that autonomous vehicles are different—they act independently and in ways that the user could not have anticipated.

---

[63] **רמב"ם נזקי ממון פרק ב, הלכה ו.** בעטה בארץ ברשות הניזק והתיזה צרורות מחמת הבעיטה והזיקו שם חייב לשלם רביע נזק שזה שינוי הוא בהתזת הצרורות, ואם תפש הניזק חצי נזק אין מוציאין מידו, ואפילו היתה מהלכת במקום שאי אפשר לה שלא תתיז ובעטה והתיזה משלם רביע נזק, ואם תפש הניזק חצי נזק אין מוציאין מידו.

[64] For a survey of these rules, see Rambam, MT *Laws of Assault and Damage*, chapter 6.

[65] See Beth Berkowitz, *Animals and Animality in the Babylonian Talmud* (Cambridge UP, 2018), esp. Ch.3, "Animal Morality" which explores the possibility of animals as subjects that bear moral culpability for their actions.

[66] רמב"ם נזקי ממון פרק א. העושה מעשה שדרכו לעשותו תמיד כמנהג ברייתו הוא הנקרא מועד, והמשנה ועשה מעשה שאין דרך כל מינו לעשות כן תמיד כגון שור שנגח או נשך הוא הנקרא תם, וזה המשנה אם הורגל בשינויו פעמים רבות נעשה מועד לאותו דבר שהורגל בו שנ' (שמות כ"א ל"ו) או נודע כי שור נגח הוא.

[67] **אחריותו של הנהג ברכב אוטונומי על נזקים, מאת יוסף שפרונג וישראל מאיר מלכה, תחומין, כרך 38 (2018).** הרב שפרונג גם מרצה על הנושא בסרטון רכב אוטונומי (2016) הנמצא באתר אוניברסיטת תל אביב.

[68] **חזון אי"ש לאו"ח לו.**

A precedent within established halakhah is the case of work animals which are sometimes kept under direct control, and other times allowed to follow their own course. Animals yoked to a plow are an example of the former; unfettered animals carrying a pack across a field at the urging of a person are like the latter. At b. Shabbat 153b we learn that a man who plows on Shabbat (החורש) with his animal is fully liable for the labor; but if he places a load on an animal (המחמר) to carry from one domain to another (מוציא) he is not. Why? Both are forbidden labors!

The answer is technical in part—the Sages imagine the person placing his package on a moving donkey and removing it when it stops. Being fully liable (חייב) for the labor of transporting requires the same person to lift the object, carry it and then deposit it. If person A places an object on person B, who transports it, and then A retrieves it, they are both free from full liability but are still forbidden to act so by decree of the Rabbis (פטור אבל אסור). Rav Pappa then says that any action which is biblically forbidden when done by one (active) person, and only rabbinically banned by means of another (passive) person, is completely permitted (מותר) when done by means of an animal.[69] This is because the animal is not itself liable for the mitzvot.

A second and more interesting distinction is found in Ramban's comments to b. Shabbat 153b.[70] He distinguishes between situations where the animal is controlled and those in which it is relatively independent:

> ונאמר בזה שמפני שהחורש בבהמה הוא נותן עליה עול והוא כובש אותה תחת ידו וברשותו היא עומדת, כל המלאכה על שם האדם היא ובו היא תלוי' ואין הבהמה אלא ככלי ביד אומן, ואינו דומה למחמר שהבהמה היא הולכת לנפשה אלא שיש לה התעוררות מעט מן המחמר.
>
> This [liability] is stated because when a person plows with his animal, he places a yoke on it, and he controls it by force of his hands, and it remains under his control. Any labor is done for the person, and it depends on him, and the animal is no more than a tool in the hands of an artisan. This is not comparable to the donkey driver, because the animal walks of its own accord, even if it is somewhat mindful of the donkey driver.

Sprung and Malka suggest that this distinction may apply as well to an autonomous vehicle. Even though it is given a task by a human "driver," the fact is that the person does not control the machine directly. It will stop and start, turn, avoid obstacles, re-route and otherwise operate independently, as it works to complete its mission. This would imply that a person is not responsible for work done on her behalf by an autonomous machine, so long as it is acting independently.

A similar inference is drawn from discussion of a person who encourages an animal to graze in a neighbor's field—only if he pushes the animal to cause damage is the person fully liable. Otherwise,

---

[69] **תלמוד בבלי מסכת שבת דף קנג עמוד ב.** אמר רב פפא: כל שבגופו חייב חטאת - בחברו פטור אבל אסור. כל שחברו פטור אבל אסור - בחמורו מותר לכתחלה.

[70] **חידושי הרמב"ן מסכת שבת דף קנג עמוד ב.**

the animal is understood to have acted independently (see Rema to SA HM 394:3). So too with a person who sets a dog or a snake on a foe to bite them—this is forbidden, but the damage is not considered to be directly caused by the person since the animal directs its own action. As Sprung and Malka summarize these cases, a person is fully responsible for malicious damage (קרן) caused by an animal, but only if the person is directly controlling the animal's action. If the animal acts independently, then the damage is not considered to be caused by the person, though the owner may be held indirectly responsible for common damages under the less severe categories of "tooth and hoof" (שן ורגל). This area of established law (see Mishnah Bava Kama, ch. 2) would imply that damage or ritual violations caused by autonomous machines would not be the full responsibility (נזק שלם) of the owner, though the owner might nevertheless bear partial responsibility, as with any damage done by one's animals or other possessions (חצי נזק).[71]

Rabbis Sprung and Malka discuss "second force" (כח שני), a term which comes from a discussion of murder in b. Sanhedrin 77b/Ḥullin 16a. If A unleashes a force (such as a stream of water) in a way that directly kills B (say A had bound B and placed B in the path of the water), then A is liable for murder—it is as if his arrows accomplished his goal (גירי דידיה הוא דאהני ביה). But if A's action was more indirect—causing a flood that *might* overtake B, one could argue that the lethal force of the floodwaters was independent of the actor, and thus the death was not caused entirely by A. A would not be liable for murder (though perhaps for manslaughter). Similarly, if a person requested that an autonomous vehicle convey them to their office, and the vehicle did damage on the way, the damage could be said to have been caused by a second force, leaving the person exempt from personal liability. This is certainly the case when there is no reason to expect an accident.

Jewish laws of liability differentiate between damage caused by a person (כוח גופו), and damage caused by their property (כוח ממונו). The archetype of the latter category is fire. If A starts a fire and it spreads to damage B's property, then A is liable for the damage. But as we have seen, the introduction of intermediate factors that separate between the action of a person and the ultimate results can lessen or eliminate liability. Likewise, when a person's property causes damage, we consider outside influences. For example, if a person sets a controlled fire in their yard in a manner that is normally safe, and then an unusually strong gust of wind (ברוח שאינה מצויה) spreads it quickly and it destroys neighboring property, the person is not fully liable. Whenever damage is caused by a combination of factors, some of which were absent at the time of the person's initial act, liability is limited.

Jewish law includes the category of compulsion (אונס); one may be held liable for damages that

[71] The Amoraim Rav Pappa and Rav Huna debate whether the Torah's rule of "half-damage" paid by the owner of an "innocent" ox that went wild implies shared responsibility or is rather a fine. Rav Huna says the latter, which means that if the ox's owner admits the damage, then they are exempt from paying the fine. The halakhah follows Rav Huna, which would indicate that the owner of an autonomous vehicle that causes damage, and who reports the accident to the authorities, might likewise be fully exempt from payment. See b. Ketubot 41a.

they set in motion even if they did not intend the destruction. The Mishnah from Bava Kama cited above states that "a person is always forewarned" and may not claim ignorance and freedom from liability for the consequences of their actions. Yet this description of liability assumes control over the action. If there was very low risk of damage, and if the process happened outside of the control of the person, then they should not be held liable for the damage caused by their vehicle.

The authors next discuss the category of "permitted harm" (מזיק ברשות). For example, if person A chooses to go running in the park, and they collide with person B, who is also running in the park—both runners had permission to run in public, and both were aware of the risk of such collisions. Unless they acted with wanton disregard for the safety of others, there is a normative expectation permitting such conduct (and even more if they were racing home for Shabbat). As autonomous vehicles become prevalent, Rabbis Sprung and Malka conclude, people journeying on roadways will expect these machines to function according to their design, and users should not be held responsible for freak accidents or failures that they could not have anticipated.

A complicated and controversial topic in the laws of damages is liability for indirect action.[72] There are two similar categories known as "*grama*" (גרמא) and "*garmei*" (גרמי). In the laws of damages, a person implicated by the former category is exempt from liability, but liability is assigned for the second category (following the opinion of Rabbi Meir in the Talmud). Everything about this topic is complicated, from the definition of terms, to the nature of liability for *garmei,* to whether it is true liability or merely a penalty to prevent people from damaging the property of others with impunity. Generally, for an act of negligence to result in liability for damages as "*garmei*" the damage must be direct, immediate and certain. None of these conditions is expected to obtain with autonomous vehicles, and thus not even a penalty for damage caused by one's autonomous machinery is justified.

Sprung and Malka's presentation regarding damages is excellent, and yet the project that they have set for themselves is limited. They have not considered the larger implications of autonomous technologies. How shall we regard the decisions of the autonomous machine—is it like an extension of the owner, or like an independent actor? Should distinctions applied to animals as either harmless (תם) or dangerous (מועד) be applied to AI-driven machines?

What if the machine does not cause physical damage, but violates other rules of the Torah? For example, if a person pre-orders an autonomous vehicle to take them on a Shabbat journey—have they violated Shabbat? They do not address this question, but their conclusion that passengers are not responsible for the damage done by autonomous vehicles suggests that they are likewise not responsible for forbidden labor done by the vehicle on their behalf. This is implicit in their citation of Ḥazon Ish regarding the operation of a plow on Shabbat. Certainly, it would be less problematic to allow a machine to conduct labor autonomously than to drive the plow oneself.

[72] ראו בבלי ב"ק ק ע"א, קונטרס הרמב"ן בדינא-דגרמי, רמב"ם פ"ז מהל' חובל ומזיק, טור ושו"ע חו"מ שפו א', וגם הערך "גרמא בנזקין, גרמי" באנציקלופדיה תלמודית.

A further limitation of this presentation is that it does not address the integration of Jewish principles into the design of the autonomous vehicles and of the algorithms that guide them. One expects that such vehicles will be safe, safer even than those driven by people, yet accidents will continue to occur. If a vehicle must choose between hitting a pedestrian in the roadway or diverting and striking an inanimate object, thereby risking the lives of its occupants, what will it do? Whose life should a machine prioritize when some loss of life is inevitable?

IV. Of Canteens and Trolleys: Whose Life Comes First?

Perhaps the most universal ethical principle is the one stated unequivocally in the Decalogue as, לא תרצח, *do not murder*. Yet the Torah itself states that the punishment for murder is execution, so intentional killing is sometimes sanctioned.[73] Then there is killing in self-defense, and negligent homicide, both of which are treated more leniently by the Torah, and justified warfare. Even if a person intentionally murders an innocent victim, rabbinic law requires elaborate (and largely unfeasible) evidentiary standards of intention and action before permitting execution. How might a machine determine whether a killing is justified or even necessary?

Isaac Asimov made his first robot rule a prohibition on harming humans, but nothing about such rules has ever been simple. What is the difference between murder and manslaughter? What is the definition of intention? Who is authorized to assess risk? Is there a meaningful ethical distinction between action and inaction, even if both are conscious decisions that may cost a life? And what should be done when loss of life is inevitable, and only one person or party can be spared?

This final question played out gruesomely in England during the summer of 1944. The German army developed V-1 missiles and began launching great numbers of them at London, causing enormous damage and 6,000 deaths in the end. Most of the missiles fell to the south of the city, where they caused damage, but not as much as if they had hit their intended target. British Intelligence engaged in a campaign of misinformation to convince the Germans that their missiles were mostly striking Central London. In so doing, they exposed people to the south to greater danger, and indeed, some 57,000 homes were damaged in Croydon alone. And yet, it was claimed that because of this deception, the casualty total was reduced, sparing as many as 10,000 lives. After the War, Philippa Foot explored the dilemma through publication of a puzzle that has come to be known in philosophical discourse as the Trolley Problem.[74]

The puzzle begins with a trolley rolling out of control towards five victims who are tied to the tracks. A bystander can throw a switch and divert the trolley onto a spur, where one person is tied. Inaction will result in five deaths, whereas action can reduce the toll to one. What is the moral

[73] See Jeremy Kalmanofsky, "Participating in the American Death Penalty," (CJLS, approved Oct. 15, 2013).
[74] For a full account see David Edmonds, *Would You Kill the Fat Man?*

response? This puzzle, with many permutations, has been used to clarify different systems of moral reasoning. The most famous variable is that instead of pushing a switch, the bystander may push a large person onto the tracks to stop the trolley and save the five.[75] Surveys reliably show that most people would throw the switch, but would not shove the person, even though the death toll would be the same.

From a utilitarian perspective, one looks to the result and reasons backwards. If forced to choose between 1 victim and 5 (in the trolley problem), or between 6,000 and 16,000 victims (in the WWII case), the moral choice is to minimize loss. The mechanism should not matter since the goal is always the best outcome for the largest number of people. Yet this approach involves intentionally causing the death of specific people. Is that defensible?

In a rule-based (deontological) ethical system, the focus is on the person deciding how to act. The rule "do not murder" would preclude a witness either from throwing the switch or pushing the man, since by doing so they would have chosen to kill the person on that track. Yet this ruling will cost four additional lives and is in tension with another rule not to watch passively while another person (or five) is killed.

Most deontological systems have developed ways to reconcile rules, such as the doctrine of double effect associated with the thirteenth century Christian theologian Thomas Aquinas.[76] In this case the bystander who throws the switch, diverting the trolley to the spur, is not *intending* to kill the lone person tied to the tracks over there, but only to save the lives of the five trapped on the current route. True, the bystander may *foresee* the death of the solitary victim before throwing the switch but would be delighted if that person were somehow to escape injury. The same cannot be said about pushing a large person onto the tracks, since their death is not merely a *consequence* of the act of saving but is the very *means*. Still, from the perspective of the poor person tied to the tracks in the first scenario, the intention of the bystander is of little significance. Is this fine moral distinction really why so many people differentiate between the two?

The divergent responses to the two scenarios (in numerous surveys 90% would throw the switch, but only 10% would push the man), despite their equal outcomes, may reveal less about moral reasoning than about neurology. Recently a field of "neuroethics," has examined the respective roles of the brain's limbic system (associated with emotional responses and drives) and the prefrontal cortex (associated with abstract reasoning) and shown that the former is more powerfully activated by the prospect of pushing a person onto the tracks. Assessment of intent is often an after-the-fact rationalization designed to justify instinctive decisions, rather than an objective measurement.[77] Many

---

[75] The size of the person matters, since otherwise the (presumably typical sized) observer could sacrifice themselves, which would alter the ethical question.

[76] "Doctrine of Double Effect," in *Stanford Encyclopedia of Philosophy*, 2004, rev. 2014.

[77] Neil Levy, "Neuroethics: A New Way of Doing Ethics," *AJOB Neurosci*. 2011 Apr-Jun; 2(2): 3–9.

other variables illustrate the interplay of these systems, indicating that people frequently act first for emotional reasons, with ethical theories such as utilitarianism and deontology serving as after-the-fact rationalizations.[78]

There are other approaches to decide the most ethical course which focus more on the moral development of the actor. Perhaps best known is an approach traced to Aristotle known as virtue ethics (which was given modern restatement by Philippa Foot). Here the focus is not on abstract decisions but on conflicting dispositions *within the person* empowered to act. What would be the most honest, or brave, or responsible action? Again, such considerations would make little difference to the person tied to the tracks should they be doomed. Yet the Torah also frames its rules in terms of righteousness and integrity. What might rabbinic sources contribute to this discourse?

One of the most famous moral dilemmas included in the Bavli, Bava Metzia 62a (and paralleled in Midrash Sifra, Behar 5:6:3) frames the question in the context of a road trip gone bad:

> שנים שהיו מהלכין בדרך, וביד אחד מהן קיתון של מים, אם שותין שניהם - מתים, ואם שותה אחד מהן - מגיע לישוב. דרש בן פטורא: מוטב שישתו שניהם וימותו, ואל יראה אחד מהם במיתתו של חבירו. עד שבא רבי עקיבא ולימד: וחי אחיך עמך - חייך קודמים לחיי חבירך.
>
> Two people were walking on the path, and one held in his hand a canteen of water. If they both drink from it, they will die, but if one of them drinks it s/he may reach the settled area. Ben Petora explained—Better that they both drink and die, that one not (passively) observe the death of the other. But then Rabbi Akiva came and taught that the verse, *Let him live by your side* (Lev. 25:36) means that your life precedes the life of your fellow.

The Talmud might have summoned a consequentialist argument, saying that a toll of two dead is worse than one, but that perspective is not mentioned. Jewish law does not permit the killing of one person to spare the life of a second, except in self-defense, so consequentialism seems to be unacceptable.[79] Ben Petora comes closest to the virtue ethics position. What kind of person would take

---

[78] Indeed, studies starting with Benjamin Libet's "Unconscious cerebral initiative and the role of conscious will in voluntary action" in *Behavioral and Brain Sciences*, 1985; 8(4): 529-539, have shown that unconscious emotional responses are often quicker than are consciously intellectual ones. This again raises the question of free will (see above), and whether individuals "decide" how to act based on reasoning rather than instincts that may be stimulated subconsciously. Still, the presence of such instincts does not deny any role to the conscious mind.

[79] Perhaps the text where the Rabbis come closest to permitting the sacrifice of one to save many is Tosefta Terumot 7:20 (and parallel in Yerushalmi Terumot Ch. 8, halakha 10), which describes a siege situation in which marauders demand one victim to kill, or else they will kill the whole group (this is a rabbinic rendition of the biblical story of Sheva ben Bichri told in II Samuel 20). The Tosefta prohibits such a sacrifice, unless the invaders specify one target. In that case they may surrender the target rather than allow themselves all to be killed. Rabbi Judah goes further—if all are in danger then one may be sacrificed. This seems to approximate a utilitarian calculus, but it could also be understood as a form of self-defense. See comments of Saul Lieberman in Tosefta Kifshuta, pp.420-421. I thank Noah Bickart for suggesting this source.

the water and leave his friend to die? This argument has the unfortunate consequence of a maximal death toll. Rabbi Akiva's *drash* unearths a "rule" in the Torah which justifies and even requires what might otherwise look like a selfish act or a coldly consequentialist calculus but has the benefit of reducing the body count.

Later commentators seek to harmonize the two perspectives. Rabbi Samuel Eliezer ben R. Judah HaLevi Edels (b. Cracow 1555) argues that Rabbi Akiva's rule applies only if the canteen *belongs* to one person, but if it belongs to both, then Rabbi Akiva would surely agree with Ben Petorah and instruct the person in possession to share and die.[80] I am not convinced. Rabbi Akiva might then argue that *the one in possession has presumption of ownership* (המוציא מחברו עליו הראיה), as he does in numerous other cases.[81] In this example, a rule allows conduct that would otherwise be condemned as selfish.

Should Rabbi Akiva's rule be made the basis of the algorithms that guide autonomous vehicles? In an accident situation where the vehicle must either strike a pedestrian or divert into a hazard that would endanger the occupant, Rabbi Akiva's rule would doom the pedestrian if needed to save the passenger (assuming the vehicle is loyal to its occupant).

However, in Bavli Pesaḥim 25b the Rabbis argue against permitting the killing of a bystander even for the sake of self-preservation. As Rava says, "Why do you think that your blood is redder than theirs?"[82]

> ושפיכות דמים גופיה מנלן? - סברא הוא; כי ההוא דאתא לקמיה דרבא, אמר ליה: מרי דוראי אמר לי זיל קטליה לפלניא, ואי לא - קטלינא לך. - אמר ליה: ליקטלוך ולא תיקטול. מאי חזית דדמא דידך סומק טפי? דילמא דמא דההוא גברא סומק טפי?
>
> How do we know that murder is forbidden, [even if necessary to save one's own life]? It is logical, as seen in the case of one who came before Rava, saying, "the lord of my town told me to kill so-and-so or if not, I will kill you." [Rava] said to him, "Let him murder you, but you must not murder. Why do you think that your blood is redder than his? Perhaps that fellow's blood is redder than yours!"

This story, which is founded on rabbinic logic (סברא), becomes exalted into a cardinal rule known by the expression אין דוחין נפש מפני נפש, *one may not kill one person to save another*. Rava thus follows a relevant rule; his position also seems closer to the virtue-ethics perspective of Ben Petora: Better to die

---

[80] **מהרש"א חידושי אגדות מסכת בבא מציעא דף סב עמוד ב.** וחי אחיך עמך חייך קודמין וכו' דמלת עמך משמע שיהא הוא טפל לך ונראה לפי הדרש דרבי עקיבא דאם היה הקיתון של מים של שניהם דמודה רע"ק לבן פטורא דשניהם ימותו ואל יראה כו' ואפשר דהיינו טעמא כדאמרינן דמאי חזית דדמא דידך סומק טפי מדמא דחברך וק"ל:

[81] See Mishnah Bekhorot 2:6-8. In each of these three mishnayot, Rabbi Tarfon rules that two equal claimants should divide property, whereas Rabbi Akiva insists that the one in possession retains ownership:
רבי טרפון אומר יחלוקו רבי עקיבא אומר המוציא מחברו עליו הראיה.

[82] **תלמוד בבלי מסכת פסחים דף כה עמוד ב.**

than to become a murderer. In the accident scenario described above, this logic might yield the opposite outcome. *The autonomous vehicle would need to divert from striking the pedestrian, even if that caused mortal injury to its passenger (and totaled the vehicle).*

It is possible that Rava does not truly disagree with Rabbi Akiva, since Rava's case would require the active murder of one's neighbor, whereas Rabbi Akiva simply advises the canteen holder to walk away. This reconciliation was offered by Rabbi Ephraim Oshry in his *Responsa Mima'amakim* from the Kovno ghetto during the *Shoah*.[83] Rabbi Oshry argues that one may act to preserve one's own life, even if this might passively endanger another person (בשב ואל תעשה), but may not actively kill another person (בקום עשה), since "why do you think your blood is redder than theirs?" Ben Petora might argue that even in keeping the water to himself the first person is actively killing or at least shortening the life of the second, but Akiva and perhaps Rava would counter that in this desperate circumstance, a person may act to preserve their own life. Rabbi Oshry notes that Rabbi Moshe Isserles concludes that if a person is about to suffer damage, he may protect himself, even if this causes another person to be harmed.[84] This logic allowed Rabbi Oshry to justify the distribution by Jewish officials of work permits to half of the working population of the Kovno ghetto, even though it made the deaths of the other half more likely.

Do we prefer that autonomous machines identify with and prioritize the lives of their owners (or users), as Rabbi Akiva's rule would indicate, or that they act altruistically, as Ben Petora and perhaps Rava would argue? *Rabbi Akiva's principle has been well established, and thus we might conclude that autonomous vehicles should prioritize the lives of their occupants, though of course they should be designed to spare the lives of bystanders as well whenever possible.*

Algorithms can be designed to reflect whatever ethical approach programmers prefer, but many would object to the entire premise of this exercise. After all, both Rabbi Akiva and Rava were addressing human actors, not machines, which have no "skin in the game." On the other hand, encoding whatever rules and principles we come up with into AI would probably make it more likely that they are complied with in difficult situations.[85] Paradoxically, AI can either be used to ensure greater compliance with Jewish ethical teachings or, since artificial agents are not party to the Torah's covenant, to avoid them altogether.

---

[83] Ephraim Oshry, שו"ת ממעמקים ה: א, pp. 14-25 (esp. p.23). I thank Robert Scheinberg for bringing this source to my attention.

[84] **שולחן ערוך חושן משפט סימן שפח סעיף ב בדברי רמ"א.** היה רואה נזק בא עליו, מותר להציל עצמו אף על פי שע"י זה בא הנזק לאחר.

[85] I thank Yoni Brafman for this point.

V. Indirect Actions and the Evasion of Moral Liability

We have come upon the limitations of formalist models of religious law.[86] It may be possible to pair autonomous machines with the halakhic doctrine of indirect causation to evade legal liability for damage done, for crimes committed, and for social and ritual obligations evaded. What then of our general obligation, "you shall do that which is right and good in the eyes of the Lord your God"?[87] Once we have delegated duties to smart machines, are we absolved of responsibility?

Consider the progression within recent decades from writing orders with pen and ink, to entering data by means of a keyboard, to the use of natural speech to instruct digital assistants to act on one's behalf. I have argued previously that using digital devices such as keyboards or cameras should be considered a derivative form of writing and be forbidden on Shabbat.[88] Can the same be said about speech that is captured by digital assistants that are always listening, and suggest actions based on what they hear? Arguably the use of speech recognition introduces one or more degrees of separation between the person and the action. After all, the AI-driven machine must make use of a large data base of sounds to interpret the speaker's words, with many opportunities for error. This would seem to remove the process from active writing to an indirect interaction that could be permitted as *grama.*

We already differentiate between direct and indirect speech when it comes to non-Jewish employees performing tasks for a Jew on Shabbat. For example, if a Jew were to announce, "I am cold," or "I would like a cup of coffee," and a gentile employee or friend were to respond by adjusting the thermostat or brewing a cup of coffee, this action would be acceptable for many Shabbat observers. Would the same not be true for a speech-activated thermostat or coffee machine? Certainly, such indirect action would not be the equivalent of a physically forbidden action taken directly by a Jew, yet such speech acts are increasingly common and powerful in our technological environment. Permitting them outright may signal the erosion of distinction between Shabbat and weekdays. For example, a sabbath observer might use voice commands to summon an autonomous vehicle to drive them across town, to order the delivery of food or other merchandise, to cook, clean or engage in commerce, all without directly "acting" in a forbidden fashion. The principle of *shvut* reminds Jews to guard the restful experience of Shabbat and Yom Tov, and not to lean too heavily on techniques that avoid technical violation while substantially undermining the purpose of the day. Naḥmanides describes just such a concern in his commentary to Leviticus 23:24.[89] Only if there is a positive religious obligation at

---

[86] I explain my approach to halakhic formalism and values-guided interpretation in a 2015 responsum, "Halakhic Perspectives on Genetic Engineering," pp.29-38.

[87] **דברים פרק יב, כח.** שְׁמֹר וְשָׁמַעְתָּ אֵת כָּל הַדְּבָרִים הָאֵלֶּה אֲשֶׁר אָנֹכִי מְצַוֶּךָּ לְמַעַן יִיטַב לְךָ וּלְבָנֶיךָ אַחֲרֶיךָ עַד עוֹלָם כִּי תַעֲשֶׂה הַטּוֹב וְהַיָּשָׁר בְּעֵינֵי ה' אֱלֹהֶיךָ:

[88] Daniel Nevins, "The Use of Electrical and Electronic Devices on Shabbat," pp.30-35.

[89] **רמב"ן ויקרא פרק כג.** ...ונראה לי שהמדרש הזה לומר שנצטוינו מן התורה להיות לנו מנוחה בי"ט אפילו מדברים שאינן מלאכה, לא שיטרח כל היום למדוד התבואות ולשקול הפירות והמתנות ולמלא החביות יין, ולפנות הכלים וגם האבנים מבית לבית וממקום למקום, ואם

stake, such as caring for the comfort and dignity of people, should such speech acts that cause a machine to complete labor be permitted on Shabbat.

Near the end of his life Rabbi Aharon Lichtenstein (1933-2015) delivered a series of lectures based on Ramban's, "Treatise on the Law of Indirect Damage" (קונטרס בדינא-דגרמי) which, as discussed above, is a complex topic of great interest to the medieval sages.[90] He voices concern that technology might be used to evade responsibility for harm and calls on rabbis to address the matter:

> במציאות הטכנולוגית המתפתחת, גוברת בהתמדה היכולת להסב נזקים, פיסיים או אפילו וירטואליים, של ממש בלי להתחייב על פי הקריטריונים של הרמב"ן או של הר"י. הנזקים יכולים להיות יותר מופשטים ותהליך הנזק יותר עקיף מהרף המינימלי הנדרש לחייב מדין גרמי - ואף על פי כן, התוצאה חמורה למדי. יוכל גנב למדן ומבריק לתכנן ולבצע שוד מושלם, בעזרת מערכת כלי פריצה של גרמא, מבלי להסתבך, לדוגלים בשיטה זו, בנזק ישיר או גרמי. הנתמיד בפטור תרחיש כזה על יסוד גרמא בנזקין? האם מגמת ניתוק אדם ממעשיו, על בסיס פער הזמן בין הפעולה והתגובה, ומתוך הנחת עצמאותן של מערכות מפותחות, שכמה פוסקים אימצו כדי להקל בשעת הצורך לגבי שבת, ישימה לגבי פטור נזק? וגם אם נדחה דוגמה זו, מתוך הנחה שניתן לחלק בין שבת לניזקין - שלגבי מלאכה נתמקד בפעולה ולגבי נזק בתוצאה - כלום אין הבעיה קיימת בשפע מקרים אחרים ? הבקשה שטוחה, הסמכות קיימת, והעניים נשואות. במידה ויעלה ביד גדולי הפוסקים לתקן בנידון, הם יצליחו לגדור פירצה חברתית של ממש, ואף ישכילו, בד בבד, להרים קרנה של תורה.

> Within the developing technological reality, there is a constantly increasing ability to inflict significant damages, physical or even virtual, without becoming liable based on the criteria of Ramban or the R"I [=Rabbi Yosef ibn Migash, regarding *garmei*]. The damage can be more abstract, and the process more indirect than the minimum required to cause liability in the laws of *garmei*—and yet, the results can be extremely severe. A skilled and brilliant thief could commit a perfect crime, using an indirect mechanism for the intrusion without getting caught up, according to those who hold this view, with liability for direct or indirect damage. Will the gap in time between an act and its consequence, and the autonomy of an advanced system be allowed to separate a person from [responsibility for] their actions? Will this approach—which has been used by *poskim* to ease liability for the laws of Shabbat—also be used to dismiss liability for damage? And even if we can set aside this example, differentiating between Shabbat and damages, since regarding labor we focus on the process, whereas with damages we look

---

היתה עיר מוקפת חומה ודלתות נעולות בלילה יהיו עומסים על החמורים ואף יין וענבים ותאנים וכל משא יביאו בי"ט ויהיה השוק מלא לכל מקח וממכר, ותהיה החנות פתוחה והחנוני מקיף והשלחנים על שלחנם והזהובים לפניהם, ויהיו הפועלים משכימין למלאכתן ומשכירין עצמם כחול לדברים אלו וכיוצא בהן, והותרו הימים הטובים האלו ואפילו השבת עצמה שבכל זה אין בהם משום מלאכה, לכך אמרה תורה "שבתון" שיהיה יום שביתה ומנוחה לא יום טורח. וזהו פירוש טוב ויפה:

[90] See https://www.etzion.org.il/he/download/file/fid/11095, p.200. I thank Nadav Berman Shifman for directing my attention to this discussion. In a 2006 responsum with Elliot Dorff and Avram Reisner, we cited Rabbi Lichtenstein's discussion of human dignity as another example of the importance of values within halakhic discourse. See "Homosexuality, Human Dignity and Halakhah," p.14 and n.90.

> at the results—will this problem not recur in many other cases? The need is apparent, the [rabbinic] authority exists, and the eyes [of the public] are raised. If the leading authorities can address this matter and repair this substantial societal problem, they will succeed and raise high the banner of Torah.

Rabbi Lichtenstein has named a major problem, even if only as a valedictory message at the close of an extensive discourse. He is hardly alone or even early in worrying that technology has created plausible moral deniability, with vast social damage to follow, but his call for rabbinic activism is noteworthy. Perhaps, as he suggests, we ought to be lenient on matters of ritual, which are between people and God, but more stringent with actions that can damage other people, whether immediately, or even over the course of generations.

Philosopher Hans Jonas argues in *The Imperative of Responsibility* that classical ethical discourse is inadequate to the powers of the technological age.[91] Previous ethics focused on interhuman relations and was geared "to the proximate range of action," (5) given the presumption that humans could not cause damage to future generations or to the earth itself and had no sense of cumulative impact. He observes that no ethics outside of religion has adopted the needed perspective which includes nonhumans within the realm of human responsibility (8). For Jonas the greatly expanded powers of advanced technology require a commensurate expansion of moral responsibility.

Jonas focuses primarily on environmental ethics; this is a concern of ours as well. Here however, our task is to delineate responsibility not for actions done by humans with machines, but for actions done by machines on behalf of humans. Shall our ethics expand yet again to encompass artificial intelligence? Can machines themselves be taught to act ethically? From the perspective of halakhah, is it legitimate to join Allen and Wallach in imagining an artificial moral agent? Might we ever say that a machine has attained personhood? If so, could the machine also acquire a religious identity, become subject to divine command, and be considered Jewish? Outlandish as this may sound, the concept is not unanticipated. In the early modern era several prominent sages pondered a similar question: *may a golem count in a minyan?*

VI. Androids as Religious Agents

As noted above, while the Rabbis focused much of their norm-building attention on individuals like them, they also expanded their gaze to other types of people, as well as to animals, objects and topography. Rabbinic law, like all law, is anthropocentric, with human life most carefully protected and regulated. Yet, this is not a full account of biblical and rabbinic thought, which is concerned with animal suffering, imposes responsibilities on humans for animal welfare, anthropomorphizes certain animals as "crafty" (the snake in Eden), others as loyal and innocent (Balaam's donkey), or wise and

[91] Hans Jonas, *The Imperative of Responsibility: In Search of an Ethics for the Technological Age* (U Chicago, 1985). The original German edition was published in 1979.

industrious (the ant in Proverbs 6). In one famous Talmudic passage, animals are proposed as a possible source for natural law.[92] The Rabbis even apply the same court procedures required for the execution of criminals to certain animals.[93] The land of Israel is viewed in Leviticus as a living organism that cannot tolerate moral turpitude.[94] Indeed, it is possible to extend Judaism's "image of God" concept beyond humanity to the earth's ecosystem.[95]

Still, these are natural phenomena. What about creatures created by people? There is a long history in Jewish thought extending to the early rabbinic period discussing the possibility and implications of using mystical methods to create an android or "golem." Gershom Scholem, Byron Sherwin, Moshe Idel and others have produced extensive studies of the history of the golem.[96] The golem enters halakhic discourse in the 17th century with obvious implications for our subject. I will provide a brief review of the essential sources, citing Idel's book as the latest statement on the subject.

Early rabbinic texts such as Midrash Bereshit Rabbah and the mystical book Sefer Yetzirah may imply some measure of human partnership in the creation of the world, and specifically in the creation of humanity. Idel finds support for this claim in early rabbinic traditions related to Abraham. Bereshit Rabbah plays on the unusual word בהבראם ("in their creation") in Gen. 2:4 and rearranges the letters to

---

[92] **תלמוד בבלי מסכת עירובין דף ק עמוד ב.** אמר רבי יוחנן: אילמלא לא ניתנה תורה היינו למידין צניעות מחתול, וגזל מנמלה, ועריות מיונה. דרך ארץ מתרנגול - שמפייס ואחר כך בועל.

[93] For example, Mishnah Sanhedrin 1:4 states that capital crimes are heard by a court of 23 judges and adds that this is true if the defendant is a person or a domesticated animal. The Mishnah says, "just as the owners are killed, so is the ox" (כמיתת בעלים כך מיתת השור). Rabbi Eliezer says that dangerous animals such as a wolf, bear, lion, tiger, leopard and snake should be killed on the spot, but Rabbi Akiva defends the equivalence of procedure between humans and animals (except for snakes), with each tried by a court of 23, and this becomes codified law. The rabbinic bestiary also includes animals that share certain human features such as sirens and the אדני השדה, as discussed in my paper, "Halakhic Perspectives on Genetically Modified Organisms," (CJLS, approved Nov. 10, 2015) n.40. See also Beth A. Berkowitz, *Animals and Animality in the Babylonian Talmud* (Cambridge UP, 2018), especially chapter 3, "Animal Morality," and chapter 5, "Animal Danger."

[94] **ויקרא פרק יח, כה.** ותטמא הארץ ואפקד עונה עליה ותקא הארץ את ישביה:

[95] This is the core argument of David Mevorach Seidenberg's book, *Kabbalah and Ecology: God's Image in the More-than-Human World* (Cambridge UP, 2015).

[96] Gershom Scholem, "The Idea of the Golem," *On the Kabbalah and its Symbolism*, trans. R. Manheim (Schocken, 1965); Byron Sherwin, *The Golem Legend: Origins and Implications* (Univ. Press of America, 1985); Moshe Idel, *Golem: Jewish Magical and Mystical Traditions on the Artificial Anthropoid* (SUNY Press, 1990). At the end of his volume Idel observes a distinction between Sephardic and Ashkenazic mystical tendencies highlighted by the *golem.* Sephardic mystics (with some exceptions) ignored the topic. They tended to focus on philosophical contemplation of the sefirot (divine spheres), whereas Ashkenazi mystics were more interested in the use of letter combinations and divine names for magical purposes, with the *golem* serving as a proof of principle. Both forms of mysticism were pious exercises designed to demonstrate reverence for the divine creation. Idel discusses the early modern emergence of the *golem* following the Renaissance reclamation of magical traditions as models of scientific exploration in Christian and Jewish circles. Of course, the *golem* has been used for many other agendas, Jewish and general, including this inquiry into AI and halakhah.

render באברהם ("through Abraham")—God created the world through the merit of Abraham.[97] The mystical tractate Sefer Yetzirah, which may include material from as early as the second century, begins with discussion of the methods of combining numbers and letters by which the universe is formed. Toward the end it includes a passage describing the special role of Abraham:

> וכשבא אברהם אבינו עליו השלום הביט וראה וחקר והבין וחצב וחקק ועלתה בידו [הבריאה שנאמר ואת הנפש אשר עשו בחרן], נגלה עליו אדון הכל יתברך שמו לעד, והושיבו בחיקו ונשקו על ראשו וקראו אברהם אוהבי.[98]
>
> Because Abraham our ancestor, blessed be his memory, contemplated and looked, saw and investigated, understood and engraved, extracted and combined and formed, and succeeded [in the creation, as it says, "and the souls that they made in Ḥaran"]; the Master of the Universe was revealed to him and He made him sit in his bosom and He kissed him upon his head and called him My beloved and put him as His son."

Idel believes that Sefer Yetzirah hints here at Abraham's ability to create a person (which is made quite explicit in the bracketed text). In any event, the 13th century commentator R' Eleazar of Worms, based on traditions from his master Rabbi Yehudah he-Ḥasid, understood Sefer Yetzirah in this way. The ability to create a person was considered within reach for the righteous, though only God could make a person with intelligence and speech.

This idea that humans are distinguished by *intelligence and speech* is found in the comments of Rashi to Genesis 2:7: "All animals are called *nefesh ḥaya*, (vital life) but humans have additional vitality, since they also possess intelligence and speech (דעה ודבור)."[99] This claim of human distinction relating to speech, based perhaps on the Aramaic translation of Onkeles--לרוח ממללא, *a speaking spirit*—was also integrated into liturgical poetry, most famously in the Yom Kippur poem, *HaAderet V'HaEmunah*.[100]

---

[97] **בראשית פרק ב, ד**. אֵלֶּה תוֹלְדוֹת הַשָּׁמַיִם וְהָאָרֶץ בְּהִבָּרְאָם בְּיוֹם עֲשׂוֹת ה' אֱלֹהִים אֶרֶץ וְשָׁמָיִם: **בראשית רבה (תיאודור-אלבק) פרשת בראשית פרשה יב.** אמר ר' יהושע בן קרחה בהבראם באברהם, בזכות אברהם (שהיה עתיד להעמיד).

[98] The Hebrew text is copied from the Bar Ilan (26+) collection, taken from Judah Eizenstein in Otzar Midrashim. The English is Idel's translation in *Golem*, based on the text of Ithamar Gruenwald, "A Preliminary Critical Edition of Sefer Yezira," in *Israel Oriental Studies*, v.1 (1971), 174, par.61, which omits the words in brackets. I have translated these words from Eizenstein and interpolated them into Idel's translation [within brackets]. See citation below of b. Sanhedrin 99b, where Reish Lakish claims that whoever teaches Torah to another's child is כאילו עשאו "as if he *made* him," using the same prooftext about the souls "made" by Abraham in Ḥaran.

[99] **פירוש רש"י לבראשית פרק ב, ז.** לנפש חיה - אף בהמה וחיה נקראו נפש חיה, אך זו של אדם חיה שבכולן, שנתוסף בו דעה ודבור:

[100] See Idel, chapter 5, "Ashkenazi Ḥasidic Views on the Golem," *Golem*, 54-80. He notes that mystical commentators like Eleazar of Worms read the piyut's line הדעה והדבור לחי עולמים as proof of the limits of human creative power. Idel clarifies that the mystical meaning is unlikely to have been intended by the poet, though we note that this poem is first found in the mystical midrash היכלות רבתי, and the next text from the Bavli would likely have been well known to the author of this poem.

The belief that righteous sages might create a person, but could not endow them with speech, is based on the most important text regarding the creation of an android, Bavli Sanhedrin 65b:

> אמר רבא: אי בעו צדיקי ברו עלמא, שנאמר כי עונותיכם היו מבדלים וגו'. רבא ברא גברא, שדריה לקמיה דרבי זירא. הוה קא משתעי בהדיה, ולא הוה קא מהדר ליה. אמר ליה: מן חבריא את, הדר לעפריך. רב חנינא ורב אושעיא הוו יתבי כל מעלי שבתא ועסקי בספר יצירה, ומיברו להו עיגלא תילתא, ואכלי ליה.[101]
>
> Rava said, if they wished, the righteous could create a world, for it says, *But your iniquities have been a barrier [between you and your God]* (Isaiah 59:2). Rava created a man and sent him [to appear] before Rabbi Ze'era. He [Rabbi Ze'era] spoke to him, but he [the man] did not reply to him. [Rabbi Ze'era] said to him: You came from the fellowship [of magicians], return to your dust! Rabbi Ḥanina and Rav Hoshaya used to sit each Sabbath eve and study the Book of Creation, and created for themselves a third grown calf, and they ate it.

This strange story, the end of which I examined previously,[102] claims that certain rabbis were able to use "The Book of Creation" to create life forms. Rashi explains Rava's method for creating the man: ברא גברא - על ידי ספר יצירה שלמדו צרוף אותיות של שם, "He created a man by means of Sefer Yetzirah, for they learned the combination of letters of the [divine] name."

It is hard to know quite how Rabbi Ze'era regarded Rava's artificial man. Idel translates חבריא as "the pietists" though it literally means "the fellowship" or "the magicians."[103] It appears that this man was able to understand and follow Rava's directions yet was unable to speak in response to Rabbi Ze'era, or perhaps at all. This story established the template for all later descriptions of a golem as a man-made creature of limited intelligence. In Ashkenazi circles we begin to hear stories of rabbis who created a *golem* with Rabbi Shmuel HeḤasid, father of Rabbi Yehudah, in the 13th century, and then Rabbi Eliyahu of Ḥelm in the late 16th century.[104]

For our purposes, the significance of this topic is its entrance into halakhic discourse among the descendants of Rabbi Eliyahu of Ḥelm. His grandson (or perhaps great-grandson) Rabbi Tzvi

---

[101] **בבלי סנהדרין דף סה עמוד ב.**

[102] Daniel Nevins, "The Kashrut of Cultured Meat," (CJLS, approved Nov. 14, 2017), 26-27. I translate עיגלא תילתא as "third-grown," not "three-year-old calf" since there is no such thing. After a year a calf is called a heifer, and after three years it is called a cow even if it hasn't given birth to its own calf.

[103] It seems to me that the Aramaic word חבריא alludes to Deut. 18:11,וְחֹבֵר חָבֶר, which JPS translates as "casts spells." Jeffrey Tigay, in *JPS Torah Commentary, Deuteronomy* (1996) p. 173, and n.31, p.375. suggests that the etymology may come from "murmuring" a spell, or perhaps from "joining" the spell to its target.

[104] There are no explicit attestations of the creation of a *golem* by the Maharal of Prague, a contemporary of Eliyahu of Ḥelm, until the 1847 folkloric work of Leopold Weisel, *Der Golem*. See "How the Golem Came to Prague," by Edan Dekel and David Gantt Gurley, *Jewish Quarterly Review* 103:2 (Spring 2013) 241-258. The 1909 book of fabrications *Niflaot Maharal* by Polish rabbi Judel Rosenberg added a connection to blood libels.

Ashkenazi[105] reports the tradition of his grandfather having created a "man" and asks whether he might be counted in a minyan.[106] After all, The Talmud states that, "whoever raises an orphan in his home is deemed by the Torah to have given birth to him,"[107] which might imply that the golem could be Jewish. But based on the Bavli's story of Rabbi Ze'era dissolving the golem, Rabbi Ashkenazi concludes that he could not have been useful in a sacred service such as the minyan. Moreover, he offers a *drashah* on Gen. 9:6, **האדם באדם**, that only a person who came from *within* a person, i.e., was born from a woman, is considered human and protected by the prohibition of murder.[108, 109]

Rabbi Ashkenazi's son was an even more influential halakhic authority, Rabbi Jacob Emden.[110] In his collection of responsa he returns to his father's question, focusing on the relation of speech to legal standing.[111] He concludes that the *golem* is not like a human who is incapable of speech, but is

---

[105] Tzvi Hirsh ben Jacob Ashkenazi, the Ḥakham Tzvi, was born in Moravia in 1660, and died in Lemberg, Poland, in 1718. He was sent to study in Sephardic lands, hence the title Ḥakham. Historians have puzzled over Ashkenazi's use of the signature ס"ט, generally understood as ספרדי טהור, which he obviously was not. Perhaps it means סיפיה טב or refers to a family miracle narrative. See (ירושלים, תשס"ז) ר' שלום משאש, ספר שמ"ש ומגן ח"ד.

[106] **שו"ת חכם צבי סימן צג.** נסתפקתי אדם הנוצר ע"י ספר יצירה כאותה שאמרו בסנהדרין רבא ברא גברא וכן העידו על זקני הגאון מוהר"ר אליהו אבדק"ק חעלם מי מצטרף לעשרה לדברים הצריכין עשרה כגון קדיש וקדושה מי אמרינן כיון דכתיב ונתקדשתי בתוך בני ישראל לא מיצטרף או דילמא כיון דקיי"ל בסנהדרין המגדל יתום בתוך ביתו מעה"כ כאילו ילדו מדכתיב חמשת בני מיכל כו' וכי מיכל ילדה והלא מירב ילדה אלא מירב ילדה ומיכל גדלה כו' ה"נ כיון שמעשה ידיהם של צדיקי' הוא הו"ל בכלל בנ"י שמע"י של צדיקי' הן הן תולדותם ונ"ל דכיון דאשכחן לר' זירא דאמר מן חברייי' את תוב לעפרך הרי שהרגו ואי ס"ד שיש בו תועלת לצרפו לעשרה לכל דבר שבקדושה לא היה ר' זירא מעבירו מן העולם דאף שאין בו איסור שפיכת דמים דהכי דייק קרא (אף שיש בו דרשות אחרות) שופך דם האדם באדם דמו ישפך דוקא אדם הנוצר תוך אדם דהיינו עובר הנוצר במעי אמו הוא דחייב עליה משום שפכ"ד יצא ההוא גברא דברא רבא שלא נעשה במעי אשה מ"מ כיון שיש בו תועלת לא היה לו להעבירו מן העולם א"ו שאינו מצטרף לעשרה לכל דבר שבקדושה כך נ"ל וכו'. צבי אשכנזי ס"ט:

[107] **תלמוד בבלי מסכת סנהדרין דף יט עמוד ב.** ללמדך שכל המגדל יתום בתוך ביתו - מעלה עליו הכתוב כאילו ילדו. **ובמקביל ע' בבלי מגילה יג ע"א וכתובות נ ע"א. וע"ע סנהדרין דף צט עמוד ב.** אמר ריש לקיש: כל המלמד את בן חבירו תורה מעלה עליו הכתוב **כאילו עשאו**, שנאמר ואת הנפש אשר עשו בחרן.

[108] **בראשית פרק ט, ו.** שֹׁפֵךְ דַּם הָאָדָם בָּאָדָם דָּמוֹ יִשָּׁפֵךְ כִּי בְצֶלֶם אֱלֹהִים עָשָׂה אֶת־הָאָדָם:

[109] This curious feature of the verse is also the basis for Rabbi Ishmael's claim at b. Sanhedrin 57b that Noaḥides (i.e., non-Jews) have a special prohibition on performing abortions. Yet a third application is proposed by Nadav Berman Shifman to ban autonomous weapons systems, since it is only "by man" that human blood may be shed (unpublished draft, 2018, p.9).

[110] Jacob ben Tzvi Emden was born in 1697 and died in 1776, in Germany.

[111] **שו"ת שאילת יעבץ חלק ב סימן פב.** בהא דמספקא ליה למר אבא בספרו (סימן צ"ג) בנוצר ע"י ספר יצירה אם מצטרף לעשר'. קשיא לי מאי קמבעיא ליה אטו מי עדיף מחרש שוטה וקטן דאינן מצטרפין. אף על גב דמבני ישראל הן ודאי וחשובין כשאר אדם מישראל לכל דבר חוץ מן המצות וההורגן חייב ואית להו דעתא קלישתא מיהא וכ"ש הקטן דאתי לכלל דעת ואפ"ה לא מצטרף. האי גברא דלאו בר דעה הוא כלל צריכא למימר מיהת בכלל חרש הוא דהא אשתעי רבי זירא בהדיא ולא אהדר ליה הא ודאי גרע מניה אלא שיש לדקדק. לכאור' נרא' שהי' שומע דהא שדריה לקמיה דר"ז אי הכי הוי ליה חרש השומע ואינו מדבר שדינו כפקח לכל דבר. אבל אין זה נרא' אמת כי אם הי' בו כח השמיע' הי' ראוי גם לכח הדבור בודאי ולא הי' מהנמנע אצלו אלא מבין ברמיזות וקריצות הי' כמו שמלמדים את הכלב לילך בשליחות להוליך ולהביא מאום' מאדם אחר כן שלחו לזה והלך. וכתוב בספר חס"ל שאין חיותו אלא כחיות הבהמ'. ולכן אין בהריגתו שום עברה א"כ פשיטא דאינו אלא כבהמ' בצורת אדם וכעיגלא תילתא דמיברי להו לר"ח ולר"א. אגב אזכיר כאן מה ששמעתי מפה קדוש אמ"ה ז"ל מה

rather to be compared to a beast in human form. At the end, Emden adds a fascinating detail, apparently from family sources, that Rabbi Eliyahu of Ḥelm grew concerned that his golem might "destroy the world," and thus detached the divine name from his forehead, returning him to dust, but not before the *golem* scratched Rabbi Eliyahu badly.[112] This postscript seems to be the source for some of the stories told later about the Maharal and his *golem* in Prague.

While it is tempting to dismiss mystical texts in a discussion of contemporary halakhah, we ought to resist this temptation. First, while fantastical, these discussions of the status of a manufactured man (and in some cases, a woman) come from the core of rabbinic discourse and involve some of the greatest halakhic authorities of our tradition.[113] Early modern masters such as Rabbi Joseph Karo and Rabbi Eliyahu, the Gaon of Vilna, are counted among the giants of both mystical and legal scholarship. Second, as Jacob Katz and Moshe Ḥallamish have demonstrated, kabbalah has exercised a substantial influence on halakhah, especially following the publication of the Zohar.[114] Third, as our late colleague Byron Sherwin argues, the mystical and halakhic discussions of the *golem* are an important model for understanding the religious significance of technology.[115]

What, then, might we learn from this discourse with reference to artificial intelligence? The idea that humans can create entities that mimic the physical and cognitive features of humanity has not

---

שקרה באותו שנוצר ע"י זקנו הגראב"ש ז"ל כי אחר שראהו הולך וגדל מאד נתיירא שלא יחריב העולם על כן לקח ונתק ממנו השם שהי' דבוק עדיין במצחו וע"י זה נתבטל ושב לעפרו. אבל הזיקו ועשה בו שריט' בפניו בעוד שנתעסק בנתיקת השם ממנו בחזק'.

[112] Idel links this idea to the early rabbinic concept that the divine creation continued to grow in proportion until God curbed it, as found in b. Ḥagigah 12a. The divine name שדי is said there by Reish Lakish to mean, אני הוא שאמרתי לעולם די, "I am the One who said to the world, enough!" Likewise, with AI the ultimate challenge may be not in its creation but in its restraint.

[113] See Louis Jacobs, *Theology in the Responsa* (Littman, 1975, 2005), pp.334-335.

[114] Jacob Katz, *Halakhah and Kabbalah* [Hebrew] (Magnes Press, 1984); Moshe Ḥallamish, *Kabbalah in Liturgy, Halakhah and Customs* [Hebrew] (Bar Ilan Press, 2000), esp. chapters 5-8, including 7, "Joseph Karo—Kabbalah in his halakhic decisions." As Ḥallamish writes, תורה וקבלה סמוכות היו אצלם על שולחן אחד (p.162). In 2003 Robbie Harris argued in a CJLS responsum about the recitation of "Amen" to the leader's blessing גאל ישראל that one should not pay regard to kabbalistic arguments in the determination of halakhah, even when made by a great authority such as Joseph Karo. Although I voted for the paper (and follow its second suggestion, the practice of Magen Avraham, of saying the blessing together with the leader), I do not share his antipathy to considering mystical sources in the determination of liturgical practice (which is itself a mystical activity).

[115] Byron L. Sherwin, "Golems in the Biotech Century," *Zygon* 42:1 (March 2007). He relates that Gershom Scholem attended the dedication of Israel's first computer at Weizmann Institute in 1965. Scholem dubbed it, "Golem I." David B. Ruderman discusses the importance of the topic of creating life among Jewish and Christian Renaissance thinkers in chapter 8 of *Jewish Thought and Scientific Discovery in Early Modern Europe* (Yale UP, 1995), 138-9. A bit earlier (132) he writes, "A general consensus among historians of science has emerged about the cultural complexity of the age in which modern science was born, about the coexistence of mystical and rational elements among scientific thinkers, and the need to view scientific thought in its broader intellectual, religious, and social context." The same can be said of halakhah. Context and complexity matter.

exceeded the imagination of our halakhic predecessors. Resemblance, however, did not secure equal status. A *golem* was not born, it lacked speech and creative capacity, and therefore it could be useful, but it could not sanctify God.

In our time, AI is quickly developing convincing capacities for intelligence. In some ways machine learning already exceeds human learning, including in what we would like to see as creative ability. Yuval Noah Harari notes that judges of chess tournaments tend to be suspicious of human—but not computer generated—moves that appear completely original.[116] Yet machines lack the most distinctive human tendencies, which may be more emotional than intellectual. They may compute, they may even generate speech. But they cannot feel responsible, fear failure, or love.

The nature of consciousness in humans and other animals is hardly well understood. Giulio Tononi has developed an "integrated information theory" which argues that consciousness emerges as a product of informational complexity.[117] In *Consciousness and the Social Brain*, Michael Graziano notes the limitations of this and other hypotheses and offers an "attention schema theory," by which consciousness is a "schema" (simplified rendition) of information attended to by the brain. He says, "Attention is not data encoded in the brain; it is a data-handling method. It is an act." (25)

Toward the end of Graziano's book is a chapter called "Some Spiritual Matters," including a section on "Computer Consciousness" (216-220). Having described the peculiar features of the human brain's attention schema that yield the distinct sense of being conscious, he offers a pathway to developing a similar ability in machines. When asked about its awareness of feelings such a computer would, "provide a human-like answer because the information set on which it bases that answer would be similar to the information set on which we humans base our answer."

Although Graziano suggests that with proper resources and effort such a machine could be designed within a decade, it is far from evident that the resulting consciousness would truly approximate human character. It is not only that our brains are social, with a theory of mind to predict what another person thinks, but that we feel *responsible* for each other's well-being. This interpersonal bond is the essence of ethics. A machine may be developed that can mimic the reports of awareness made by humans, but it will still not achieve the status of humanity unless it can present the essential aspects of personhood. What are they?

The definition of personhood and its parameters is an extraordinarily vexing philosophical and legal problem.[118] Scholars working in the field of critical animal studies have, "conceived of animals as

---

[116] Yuval Noah Harari, "Why Technology Favors Tyranny," *The Atlantic* (October 2018). See Gary Kasparov's discussion of AlphaZero's style of chess play in comparison to his own, "Chess, a *drosophila* of reasoning," in *Science* (Dec. 7, 2018), Vol. 362, Issue 6419, p. 1087.

[117] See "Integrated information theory: from consciousness to its physical substrate," by Giulio Tononi, Melanie Boly, Marcello Massimini, and Christof Koch. *Nature Reviews Neuroscience*, vol.17, pp. 450–461 (2016). Graziano provides a critical survey of this theory in chapter 11 *of Consciousness and the Social Brain.*

[118] See for example, "The Moral Status of Animals," in the *Stanford Encyclopedia of Philosophy*.

persons, agents and subjects."[119] This is at variance with the established characterization of animals as things and property. In her study of Bavli Sukkah 22b-23b (Chapter 6), Berkowitz shows how the Sages problematize their initial characterization of animals as objects and establish a category (דבר שיש בו רוח חיים) that acknowledges that animals are "bad things," even if they are not exactly persons. She speculates that the ancient rabbis were navigating between Roman law, which clearly and consistently classifies animals as property, and Zoroastrian law, which divides animals between beneficent and noxious classes, accepting them as subjects that can be innocent or guilty, much as the Sages did in applying court procedures to animals accused of crime. In the Babylonian Talmud, she argues, the rabbis confront the limitations of animal "thingness" and open the possibility of animal subjectivity.

Personhood is at the heart of our discussion of whether AI directed machines can likewise escape their status as things and achieve something closer to subjectivity. Analytic intelligence is only one aspect of personhood. It is difficult and even dangerous to define personhood precisely, since one can always find categories of people who lack an ability said to be essential. It is reductive to state that a person is a person born to a person, but that it is the sense of the *drashah* seen above on אדם באדם by Rabbi Tzvi Ashkenazi. Another approach is to consider personhood as a social construct—a person is an actor who relates to themselves and to other actors with responsibility and reciprocity. This understanding does not, however, apply to all and only humans. While some people are either temporarily or permanently incapable of such relationships, they nevertheless remain part of the class of people who do have such abilities. This is not the case with machines, which currently have no sense of self, no experience of obligation, of concern, of guilt, or of moral grandeur. These terms are admittedly imprecise, but it is just such spiritual qualities that define personhood in Jewish life as found in our liturgy, our commandments, and our theology.

Abraham Joshua Heschel considered the distinctive qualities of humanity in his 1964 address to the American Medical Association, "The Patient as Person." In his poetic style he writes,

> What constitutes being human, personhood? The ability to be concerned for other human beings…The truth of being human is gratitude, the secret of existence is appreciation, its significance is revealed in reciprocity. Mankind will not die for lack of information; it may perish from lack of appreciation. Being human presupposes the paradox of freedom, the capacity to create events, to transcend the self…
> The ultimate significance of human being as well as the ultimate meaning of being human may be wishful thinking, a ridiculous conceit in the midst of a world apparently devoid of ultimate meaning, a supreme absurdity. It is part of the cure to trust in Him who cures. Supreme meaning is therefore inconceivable without meaning derived from supreme being. Humanity without divinity is a torso. This is even reflected in the

[119] Beth Berkowitz, *Animals and Animality,* pp.157-160.

> process of healing. Without a sense of significant being, a sense of wonder and mystery, a sense of reverence for the sanctity of being alive, the doctor's efforts and prescriptions may prove futile.[120]

If Heschel were alive today, he would likely add the biotech engineer to the doctor. Intelligence can be manufactured, but not the soul, and without a soul, artificial life is always virtual, never quite real. Our halakhists denied that a *golem* could join a minyan. It is evident that for an action to count as fulfillment of a mitzvah, a command, one requires still distinctive human capacities such as compassion, gratitude, wonder, and reverence.

VII. למאי נפקא מינה? From Theory to Practice: Foreseeable Implications for Real Life

We have considered four halakhic discourses that seem relevant to our consideration of AI and autonomous machines: the rules of agency and the role of non-human agents; the rules of damage caused by animals and property under indirect control of a person; the prioritization of lives when loss is inevitable; and the (im)possibility of including a human-made android or golem in the minyan. Now we turn to practical questions, realizing that this technology is new and that we can hardly anticipate all the capabilities that will be developed and the religious dilemmas that they will engender.

A. Smart Appliances and the Laws of Shabbat.

It is already common for "smart appliances" to employ facial recognition to identify users, and for machine learning to anticipate their needs. Such interactions may quickly move from the digital to the physical realm. For example, a kitchen appliance might learn to prepare hot beverages or to cook certain foods customized for each member of the household. Or, an autonomous car service might be ordered to transport members of the family to certain locations each day, including on Shabbat, based on their appointment calendar or by monitoring their digital messages. Let us assume that the machine's actions would not require a human user to engage in any direct financial transactions or physical data entry on Shabbat for the machine to complete its task. Is such labor permitted, given that it is automated, or is it forbidden, given that it requires human participation?

Currently there are many appliances such as elevators and timers that have been modified to operate for human benefit on Shabbat, but these devices follow a fixed schedule without human input. True, Shabbat elevators check to make sure the doorway is clear, and the cab is not overloaded before moving, but there is no need for human action other than standing in place while the machine completes its task. In contrast, smart appliances and vehicles guided by artificial intelligence will engage the user, recognize them, perhaps request voice confirmation of instructions and modify their actions in response to or in anticipation of human need. Should actions that would be forbidden for a

[120] Abraham Joshua Heschel, *The Insecurity of Freedom: Essays on Human Existence* (JPS, 1966) 26.

Jew be permitted to a robot acting on their behalf?

God commands Israel to observe the Sabbath, not only for themselves, but also for their animals and non-Jewish staff. However, automatic work processes set up prior to Shabbat, such as a water mill to grind kernels of grain into flour, or an irrigation trench to water a field, may continue through Shabbat, since Jews are not responsible for work done by tools (שביתת כלים) as long as the process is not managed on Shabbat, and there is no concern that others will think a Jew is doing the labor.[121]

Although non-Jews are not themselves commanded to rest on Shabbat, Jews are restrained from requesting that they perform forbidden tasks on their behalf. The restraint on "speaking to a gentile" (אמירה לנכרי) is complex, with various explanations, stringencies and leniencies.[122] In general, a Jew may not ask nor profit from labor done by a gentile on Shabbat. Such requests introduce weekday concerns into Shabbat and turn the gentile into an agent for the Jew.[123] The status of this prohibition is rabbinic, or "*shvut*." If the task was *not* formally prohibited as transformative labor (מלאכה), but only from the command to rest (שבות), then a Jew may ask a gentile to do such tasks for the sake of the ill or infirm, or to assist with a mitzvah (for example, by setting up materials for *havdalah*). This is because there is no "*shvut* on a *shvut*," meaning that the expanded ban on asking gentiles to work (*shvut* #1) does not extend to the expanded ban on imperfect or impermanent forms of the prohibited labors (*shvut* #2) if the purpose is to allow performance of a mitzvah, to help the ill or infirm, to protect dignity, etc. It would seem reasonable to be at least as lenient with an autonomous machine, and probably more so, if there is a halakhically significant motivation.

In Section II we learned that an agent generally bears responsibility for prohibited actions, but not if they lack the freedom to refuse the assigned task or the ability to pay fines for misconduct. A smart appliance or autonomous vehicle does not have the freedom to refuse a task or the ability to pay for misconduct. As such it might be considered a tool in the hand of the user, and it would be

---

[121] Please see further discussion of this topic in my responsum, "The Use of Electrical and Electronic Devices on Shabbat" (CJLS, approved May 31, 2012) p. 3ff and notes. See also David Hoffman's responsum, "Building at What Cost?" (CJLS, approved October 17, 2018).

[122] ע' **אנציקלופדיה תלמודית** כרך ב, אמירה לנכרי שבות.

[123] This explanation is offered by Rashi at b. Shabbat 153a, s.v. מאי טעמא. The context is that of a Jewish traveler who fails to reach their destination before Shabbat. They may hand their purse to a gentile to hold for Shabbat, since gentiles are not commanded to observe Shabbat. Rashi explains the objection—והרי הוא שלוחו לישאנו בשבת "this would turn the gentile into the Jew's agent to carry it on Shabbat." The Gemara provides a descending series of preferred actions designed to minimize transgression without tempting the Jew to ignore the laws and carry their own purse. In general, it is better to tie the purse onto a donkey than to hand it to a person. As the Gemara explains, "What is the reason? These are people; this [donkey] is not a person! מאי טעמא - הני אדם, האי - לאו אדם. This suggests that asking even exempt classes of people to perform work on Shabbat is more problematic than causing an animal to do the task. How much more so for a machine! For much more on this topic, see Jacob Katz's classic study, *The Shabbes Goy: A Study in Halakhic Flexibility* (JPS, 1989).

forbidden to ask or even allow the machine to perform labor on one's behalf. Yet in Section III we found that intermediate factors between the will of the user and the actions of their animals or the interactions of their property and environment could reduce liability for consequent actions. Arguably the very premise of artificial intelligence is that it is autonomous—it considers variables, for example traffic reports, and determines the best route, and is in this sense acting independently of the user. *The user's request for forbidden labor could be designated as indirect, but since there is immediate correlation between the command and the action, we would deem it to be garmei, and for the user to still bear partial responsibility for the action.*

As I argued in my prior work regarding electricity and Shabbat,[124] there is a contest of values between preserving the distinctive experience of Shabbat and Yom Tov on one side, vs. protecting life, preserving human dignity and the environment, and completing positive mitzvot on the other side. Regarding biblically forbidden labors (מלאכות), only the mandate to save life (פקוח נפש) can override the prohibition. Thus, a person would be forbidden from typing instructions (תולדת כותב דאורייתא) during Shabbat for a machine to do labor on their behalf. However, passively permitting a system to initiate actions on one's behalf during Shabbat without instruction, or even speaking instructions to a machine which will independently design a course of action is to be considered akin to a partially indirect action (גרמי) and is therefore of a lower level concern, the rabbinic concept of *shvut.* The person does not supply or even anticipate all the information required for the activity; autonomous machines gather data required to realize goals, and thus they are not mere tools in the hand of a person. Still, the experience of asking a machine to complete tasks for a person is close enough to the experience of asking a non-Jew to do the same that it should be governed by the same rabbinic concern for *shvut*. Competing Jewish values such as human dignity may outweigh rabbinic bans. Nevertheless, everything possible should be done to honor both values and to minimize conflict between them.

Therefore, it would seem to be permissible to arrange for an autonomous vehicle to transport a person with special needs (illness, frailty, disability, avoiding danger) within the local limits of travel (תחום של שבת) on Shabbat. Interactions with AI-powered autonomous machines on Shabbat itself should be considered as generally banned by force of rabbinic law, but amenable to override due to urgent competing halakhic values (לצורך גדול). That said, we are also bound to a positive commandment to preserve the restful nature of the day (שבות) by avoiding weekday actions and concerns (עובדין דחול) and should avoid such technological solutions unless necessary for the experience of Shabbat itself. Moreover, one must not use this leniency to order services for *after* Shabbat or the holiday, since this would be considered "preparing" (הכנה) for the work week, which is itself forbidden under the category of *shvut*.

[124] pp.47-53.

There is an established pattern of seeking leniencies in the realm of Shabbat law that are not applied in other situations.[125] We would therefore find that during Shabbat or Yom Tov, asking for or acquiescing to biblically prohibited labors to be done on one's behalf by a machine is banned, but under the least severe prohibition, *shvut*.

B. Training Autonomous Cars for Accident Situations.

Engineers working on autonomous vehicles presume that they will be far less likely to endanger the lives of passengers or pedestrians than are vehicles driven by humans. There have been several well-publicized fatal accidents caused by autonomous vehicles, but none in which the car was forced to choose between two potential victims. Human drivers are not generally trained in the prioritization of lives either, except not to swerve to avoid small animals in the road, and so perhaps it is unnecessary to worry about the moral reasoning of vehicles.

Moreover, most of the ethical systems that are candidates to guide the development of artificial moral agents are problematic for one or another reason. Consequentialism might lead an autonomous vehicle in an accident scenario to consider the relative market value of nearby cars (or even the net worth of their occupants), swerving to hit the least expensive (or wealthy) among them. Such considerations might be rational but would be judged as unacceptable for human drivers. Deontology might prevent a car from avoiding a multi-fatality collision if the only alternative route presented the risk of one fatality since a moral agent should never intentionally kill an innocent person. Yet such conduct would maximize the risk for all people present. Somehow, human drivers are expected to do the right thing even in the confusing and chaotic experience of an accident, and somehow, we must anticipate the unexpected and design autonomous machines that can also "do the right thing." How? One relative strength of machine intelligence is situational awareness, the ability to process information from many sources at once—which might overload a human actor—and to act according to the established decision tree. What might that be?

Recalling the teaching of Rabbi Akiva in Section IV, we would conclude that a vehicle should prioritize the life of its own passenger(s) in an accident scenario where it is either their life or that of another person, since "your life comes first" (חייך קודמין). But remembering Rava, we would want the autonomous vehicle to give every priority to sparing life, and never intentionally to kill a person, even if that allows an accident to take the life of the passenger. Rabbi Oshry's citation of Rema balances these values, allowing for self-rescue, even if that puts other lives at risk. A halakhic algorithm might be structured as follows:

i. If there is risk of a collision, then the vehicle should prioritize avoiding danger to all

[125] See Rabbi Lichtenstein above. This line of thought seems to originate with m. Shabbat 22:3. See Bavli there at 146a, and comments of Rashba: וי"ל כיון דבעלמא במקלקל פטור אבל אסור הכא משום צורך שבת מותר לכתחלה.

people in the area, by swerving, slowing or accelerating as necessary.

ii. If a collision is unavoidable, then the priority should be to avoid fatality or severe bodily harm to humans. The vehicle should swerve to avoid striking a pedestrian or passenger vehicle, even at risk to property damage.

iii. If risk of fatality is unavoidable, then the lives of the occupants of each vehicle should be prioritized by that vehicle. The vehicle should not swerve off the road into danger in order to avoid striking a pedestrian on the roadway unless there is little risk of fatality to the passengers. Rather, the vehicle should take all available measures to reduce danger to the pedestrian, such as slowing down or switching lanes, while still protecting the lives of the passengers.

While such an algorithm might approximate the moral reasoning of a person and could even exceed their analytic abilities in an accident situation, the machine itself is not to be considered a moral agent. Moral agency reflects the embodied experience and moral accountability of humans—their sense of self, of mortality, of responsibility and of guilt. A moral actor must be able to do a minor wrong to accomplish a major right. *Humans should be consulted in any life-or-death scenario unless the delay required for such consultation will imperil human life.*

Regarding autonomous weapons systems, military strategists have noted that requiring a human to be in the loop on all lethal actions will disadvantage that side. Once autonomous weapons systems have become part of a nation's arsenal, humans are not likely to retain constant control over the systems. This argues against the use of AWS that target humans (rather than inanimate targets such as mines, missiles, sensors, etc.) altogether, and indeed, this should be our primary position. Certainly, the laws of war should be updated to ban fully autonomous systems from attacking humans without specific human permission. Otherwise there will be an accountability gap, and commanders might unleash such weapons without the constraint of being held personally responsible for the results.

Given that semiautonomous weapons are already integrated in the armed forces of dozens of nations, some have argued that the next best approach may be to integrate tight ethical controls into command systems. Michael Saxon and Christopher Korpela make this case:

> So, does the need for responsibility in Just War Theory require that humans remain "in the loop" for decisions involving potentially lethal force? This might be illuminated by imagining a human-in-the-loop LAWS system that satisfies all of the requirements of IHL [International Humanitarian Law]. First, it is discriminate, in that it is capable of differentiating between combatants and non-combatants. Second, it adheres to the rule of non-combatant immunity. Third, it is proportionate, using morally appropriate levels of force. These must all be subject to the checks provided by the human mind in control. Aside from legitimate questions about the *ease* with which a system might allow the nation to go to war unnecessarily, this sort of

machine seems ideal.[126]

While such a sensitive machine may seem ideal in secular terms, and might even satisfy Wallach and Allen's standard for an AMA, can the same be said for the requirements of halakhah? True, the Talmud has David lecture Saul, "The Torah says, if one comes to kill you, rise up quickly to kill him."[127] A human pursuer (רודף) is considered forewarned, allowing a human defender to strike the attacker dead without warning. Yet the codes teach that the minimally lethal use of force is necessary (as exemplified in the Bible by David, who spares Saul's life), and that a defender who kills an attacker when non-lethal options are available is liable for the death.[128]

We must not offer authority to a non-human actor to differentiate among human targets and decide whom to kill. Paul Schaare concludes his book, *An Army of None*, with a call for "a conscious choice to pull back from weapons that are too dangerous, too inhumane." (362) He warns that autonomous systems are hackable, and that they may be turned against their owners by adversaries, or even by simple error. In *Foreign Affairs* he argues that, "The United States should work with other countries, even hostile ones, to ensure AI safety." (144) International activism to limit the autonomy of lethal weapons has begun, and halakhists must join humanists in mandating human judgment in matters of life and death. Autonomous defense systems that attack objects such as missiles or malware are one thing; those that select human targets and attack them without human consent are quite another. Maintaining human control over such systems may require structures such as 24/7 human operators to authorize lethal attacks; humans must always retain control of preemptive uses of force. Jewish law is exceptionally cautious about the evidentiary standards required to justify killing a human. *Autonomous weapons systems must not use lethal force against humans without human direction.*

C. Robotic Assistance in Jewish Ritual Performance

Robotic attendants might be able to assist Jews in various physical tasks required to complete mitzvot. For example, they might retrieve ritual objects such as a siddur, tallit, tefillin, lulav and etrog and bring them to an immobile person at the time of prayer. In Section II, we learned that ritual actions such as separating tithes or establishing a Sabbath boundary (תחום של שבת) must be *initiated and completed* by a Jew, not by a non-Jew (e.g., a Samaritan) nor by an animal (e.g., an elephant or a

---

[126] Michael Saxon and Christopher Korpela, "Killing with Autonomous Weapons Systems," *War Room,* Jan. 17, 2018.

[127] **בבלי ברכות דף סב עמוד ב.** אמר רבי אלעזר, אמר לו דוד לשאול: מן התורה - בן הריגה אתה, שהרי רודף אתה, והתורה אמרה: בא להרגך השכם להרגו. **וע"ע שם דף נח ע"א, וביומא דף פה ע"ב ובמקבילות במדרש.**

[128] **רמב"ם רוצח ושמירת הנפש פרק א.** כל היכול להציל באבר מאיבריו ולא טרח בכך אלא הציל בנפשו של רודף והרגו הרי זה שופך דמים וחייב מיתה אבל אין בית דין ממיתין אותו. **וע' בית יוסף חושן משפט סימן שפח.** ובודאי מי שאינו מוחזק בכך אין ממיתין אותו אחר מעשה אבל מי שהוחזק בכך ממיתין אותו בין בשעת מעשה בין לאחר מעשה וכל הקודם זכה וכו'.

monkey). However, we saw that a Jew might rely on one of these "agents" for an *intermediate* step such as carrying an object from one location to another. Based on this analogy, a robot or other type of smart appliance might be used to position or prepare an object for ritual use, but the ritual itself must be performed by a person for whom it is a sacred obligation. The same would be true for social commandments such as performing acts of *ḥesed*—visiting the ill, comforting the bereaved, supporting the poor. A machine might assist, but a human must initiate and complete this action.

In Section V we traced the history of the *golem* not only as an occult figure, but also as a possible player in the ritual life of a community. Rabbi Jacob Emden asked whether such an android might be counted in the minyan required for certain ritual acts. He concluded that because the *golem* discussed in our sources lacked human capabilities of speech and apparently reasoning, and because it was not formed within a human, it could not be considered human, nor could it be included in ritual life.

Digital assistants today are already capable of voice recognition and of generating contextually appropriate responses. The field of affective computing addresses the emotional content of communication with the goal of helping people form bonds with machines. Studies have demonstrated that people of all ages are eager to engage such speaking machines as if they were persons. And yet, the distinction between human and artificial intelligence remains significant. Artificial agents may *approximate* the decision-making process of humans, but they do not have the capacity to *appreciate* the significance of an action, to accept or reject responsibility, to experience human emotions of fear, pain, pleasure or love, or to act in service of an abstract value, the divine. This distinction between machines and humans is essential and demands our active defense.

VIII. *Piskei Din*

1. *Are Jews liable for the halakhic consequences of actions taken by machines on their behalf, for example, Sabbath labor?* Perhaps, but only at the lowest level of *shvut*. During Shabbat or Yom Tov, a Jew should not request that a smart machine initiate or complete forbidden labor unless there is a mitigating factor such as illness or frailty (חולה שאין בו סכנה), threat to human dignity (כבוד הבריות), or specific need to facilitate a commanded act (לצורך מצווה/שבת) on that day. Arranging for such activities prior to the onset of Shabbat or Yom Tov, even absent such mitigating factors, would be permitted if it is understood that these services were pre-arranged, and if they did not undermine the general experience of Shabbat (שבות).
2. *Should ethical principles derived from halakhah be integrated into the development of autonomous systems for transportation, medical care, warfare and other morally charged activities, allowing autonomous systems to make life or death decisions?* Autonomous systems may have capacities to process and communicate information that exceed those of humans, and they may help humans avoid common failures as moral and religious actors. That said, only humans have

the right and the responsibility to make life and death decisions. Humans must supervise AI systems and authorize lethal actions, whether in transportation, medicine or in warfare.

3. *Might a robot perform a mitzvah or other halakhically significant action? Is it conceivable to treat an artificial agent as a person? As a Jew?* Artificial agents may be used to facilitate the performance of mitzvot, for example in conveying ritual objects to a person. They may follow commands, but they do not become humans, much less *b'nei brit*. This distinction is not to diminish their value, which can be vast, but it is to recall the very purpose of the mitzvot that have been revealed to and developed by the people of Israel—to bless God Who, despite our frailty and fallibility, has sanctified our lives through the commandments.

וכנלע"ד[129]

[129] Special thanks to friends and colleagues who advanced my understanding of this topic. Among them: Nadav Berman Shifman, PhD; Yoni Brafman, PhD; David Gerwin, PhD; Alick Isaacs, PhD; Lynn Nevins, MS; Michael Paasche-Orlow, MD; Rabbi Micah Peltz; Rabbi Avram Reisner, PhD; Jason Rogoff, PhD; Rabbi David Rosenn; Toby Schonfeld, PhD; Rabbi Burt Visotzky, PhD; and students in my JTS Spring 2019 seminar, "Writing Responsa." Of course, I am responsible for all errors of fact and judgment.

Selected Bibliography

Ronald C. Arkin, *Governing Lethal Behavior in Autonomous Robots.* CRC Press, 2009.

Beth Berkowitz, *Animals and Animality in the Babylonian Talmud.* Cambridge UP, 2018.

Samir Chopra, Laurence F. White, *A Legal Theory for Autonomous Artificial Agents.* Univ. Michigan Press, 2012. CU LAW K917.C475 2011

Amitai Etzioni, Oren Etzioni, "AI Assisted Ethics," *Ethics Inf Technol* (2016) 18:149–156.

____ . "Pros and Cons of Autonomous Weapons Systems," *Military Review* (May-June 2017) 72-81.

Keith Frankish, William M. Ramsey, editors, *The Cambridge Handbook of Artificial Intelligence*.

Michael Graziano, *Consciousness and the Social Brain.* Oxford UP, 2013.

David J. Gunkel, *The Machine Question: Critical Perspectives on AI, Robots and Ethics.* MIT Press, 2012.

Yuval Noah Harari, *Homo Deus: A Brief History of Tomorrow.* Harper Collins, 2017.

Patrick Lin, Keith Abney, George A. Bekey, editors, *Robot Ethics: The Ethical and Social Implications of Robotics.* MIT Press, 2012.

Patrick Lin, *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford UP, 2017.

George Lucas Jr., "Engineering, Ethics & Industry: The Moral Challenges of Lethal Autonomy," in *Killing by Remote Control: The Ethics of an Unmanned Military*, ed. Bradley Jay Strawser. New York: Oxford, 2013.

Naḥum Rakover, *Agency and Appointment in Jewish Law* [Hebrew]. 1972.

Paul Schaare, *Army of None: Autonomous Weapons and the Future of War.* NY: W.W. Norton and Co, 2018.

Byron L. Sherwin, *Jewish Ethics for the Twenty-First Century: Living in the Image of God*. –2000

Nadav Berman Shifman, "Autonomous Weapon Systems and Jewish Law: Ethical-Political Perspectives," (forthcoming)

Peter W. Singer, "The Ethics of Killer Applications: Why Is It So Hard To Talk About Morality When It Comes to New Military Technology?" *Journal of Military Ethics* 9, no. 4 (2010), 299-312.

Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence.* Alfred A. Knopf, 2017.

Wendell Wallach, Colin Allen, *Moral Machines: Teaching Robots Right from Wrong*. Oxford UP 2009.

Wendell Wallach, *A Dangerous Master: How to Keep Technology from Slipping Beyond Our Control*. Basic Books, 2015.

Appendix: Study Sheet with Selected Sources

I. A Question of Agency: Can a Robot Represent a Person?

a. Eruvin 31b

דתניא: נתנו לפיל והוליכו, לקוף והוליכו - אין זה עירוב. ואם אמר לאחר לקבלו הימנו - הרי זה עירוב. - ודילמא לא ממטי ליה? - אמר רב חסדא: בעומד ורואהו. - ודילמא לא מקבל ליה מיניה? - אמר רב יחיאל: חזקה שליח עושה שליחותו.

For it is taught in a *beraita*: If he gave it [i.e., an item to be used to extend the sabbath boundary] to an elephant and it carried it, [or] to a monkey and it carried it—this is not a [valid] *eiruv.* But if he arranged for another [person] to receive it from him [the animal]—this is a [valid] *eiruv.* But perhaps [the animal] won't deliver it? Rav Ḥisda says, it is [a case] when he [the sender] stands and watches him [the animal]. But perhaps he [the receiver] won't accept it from him [the animal]? Ravi Yeḥiel says, agents are presumed to fulfill their agency.

b. Bava Metzia 10b

אמר רבינא: היכא אמרינן דאין שליח לדבר עבירה - **היכא דשליח בר חיובא הוא**, אבל בחצר דלאו בר חיובא הוא - מיחייב שולחו. - אלא מעתה, האומר לאשה ועבד צאו גנבו לי דלאו בני חיובא נינהו הכי נמי דמיחייב שולחן? - אמרת: אשה ועבד בני חיובא נינהו, והשתא מיהא לית להו לשלומי. דתנן: נתגרשה האשה, נשתחרר העבד - חייבין לשלם.

Ravina says, when we said that "there is no agency for a transgression," that was **only when the agent themselves was obligated** [for that transgression]. But as for a courtyard, which is not itself obligated, the principal is liable. If so, when a man tells his wife or slave, "go steal for me," since they are not obligated to pay [the "double" penalty] shall we say that the principal is liable? You could say, wives and slaves are [after all] responsible [not to steal] but are not obligated [to pay the fine for theft, since they do not control their own assets]. For it is taught in a Mishnah, if the woman is divorced or the slave is freed, then they become liable to pay [their own fines].

רב סמא אמר: היכא אמרינן אין שליח לדבר עבירה - **היכא דאי בעי עביד, ואי בעי לא עביד.** אבל חצר, דבעל כרחיה מותיב בה - מיחייב שולחו.

Rav Sama says, when we said that "there is no agency for a transgression," that was **only in the case when if [the agent] wanted, he acted, and if [the agent] didn't want, he didn't have to act.** But as for a courtyard, where items are placed without its consent, the principal is liable. [Emphasis added in both quotes]

II. Indirect Damages Caused by an Animal. Ramban, Ḥiddushim to Shabbat 153a.

ונאמר בזה שמפני שהחורש בבהמה הוא נותן עליה עול והוא כובש אותה תחת ידו וברשותו היא עומדת, כל המלאכה על שם האדם היא ובו היא תלוי' ואין הבהמה אלא ככלי ביד אומן, ואינו דומה למחמר שהבהמה היא הולכת לנפשה אלא שיש לה התעוררות מעט מן המחמר.

This [liability] is stated because when a person plows with his animal, he places a yoke on it, and he controls it by force of his hands, and it remains under his control. Any labor is done for the person, and it depends on him, and the animal is no more than a tool in the hands of an artisan. This is not comparable to the donkey driver, because the animal walks of its own accord, even if it is somewhat mindful of the donkey driver.

III. Whose Life Comes First?

a. Bava Metzia 62a.

שנים שהיו מהלכין בדרך, וביד אחד מהן קיתון של מים, אם שותין שניהם - מתים, ואם שותה אחד מהן - מגיע לישוב. דרש בן פטורא: מוטב שישתו שניהם וימותו, ואל יראה אחד מהם במיתתו של חבירו. עד שבא רבי עקיבא ולימד: וחי אחיך עמך - חייך קודמים לחיי חבירך.

Two people were walking on the path, and one held in his hand a canteen of water. If they both drink from it, they will die, but if one of them drinks it s/he may reach the settled area. Ben Petora explained—Better that they both drink and die, that one not (passively) observe the death of the other. But then Rabbi Akiva came and taught that the verse, *Let him live by your side* (Lev. 25:36) means that your life precedes the life of your fellow.

b. Pesaḥim 25b

ושפיכות דמים גופיה מנלן? - סברא הוא; כי ההוא דאתא לקמיה דרבא, אמר ליה: מרי דוראי אמר לי זיל קטליה לפלניא, ואי לא - קטלינא לך. - אמר ליה: ליקטלוך ולא תיקטול. מאי חזית דדמא דידך סומק טפי? דילמא דמא דההוא גברא סומק טפי?

How do we know that murder is forbidden, [even if necessary to save one's own life]? It is logical, as seen in the case of one who came before Rava, saying, "the lord of my town told me to kill so-and-so or if not, I will kill you." [Rava] said to him, "Let him murder you, but you must not murder. Why do you think that your blood is redder than his? Perhaps that fellow's blood is redder than yours!"

c. Rabbi Moshe Isserles, Shulḥan Arukh, Ḥoshen Mishpat 388:2

היה רואה נזק בא עליו, מותר להציל עצמו אף על פי שע"י זה בא הנזק לאחר

If a person sees that he is about to be injured, he may save himself even though in so doing the injury will come to another person.

IV. A Golem in the Minyan?

a. Sanhedrin 65b

אמר רבא: אי בעו צדיקי ברו עלמא, שנאמר כי עונותיכם היו מבדלים וגו'. רבא ברא גברא, שדריה לקמיה דרבי זירא. הוה קא משתעי בהדיה, ולא הוה קא מהדר ליה. אמר ליה: מן חבריא את, הדר לעפריך. רב חנינא ורב אושעיא הוו יתבי כל מעלי שבתא ועסקי בספר יצירה, ומיברו להו עיגלא תילתא, ואכלי ליה.

Rava said, if they wished, the righteous could create a world, for it says, *But your iniquities have been a barrier [between you and your God]* (Isaiah 59:2). Rava created a man and sent him [to appear] before Rabbi Ze'era. He [Rabbi Ze'era] spoke to him, but he [the man] did not reply to him. [Rabbi Ze'era] said to him: You came from the fellowship [of magicians], return to your dust! Rabbi Ḥanina and Rav Hoshaya used to sit each Sabbath eve and study the Book of Creation, and created for themselves a third grown calf, and they ate it.

b. Rabbi Zvi Ashkenazi, "Hakham Tzvi" Responsum #93

נסתפקתי אדם הנוצר ע"י ספר יצירה כאותה שאמרו בסנהדרין רבא ברא גברא וכן העידו על זקני הגאון מוהר"ר אליהו אבדק"ק חעלם מי מצטרף לעשרה לדברים הצריכין עשרה כגון קדיש וקדושה מי אמרינן כיון דכתיב ונתקדשתי בתוך בני ישראל לא מיצטרף או דילמא כיון דקיי"ל בסנהדרין המגדל יתום בתוך ביתו מעה"כ כאילו ילדו [...] ה"נ כיון שמעשה ידיהם של צדיקי' הוא הו"ל בכלל בנ"י שמע"י של צדיקי' הן הן תולדותם ונ"ל דכיון דאשכחן לר' זירא דאמר מן חברייי' את תוב לעפרך הרי שהרגו ואי ס"ד שיש בו תועלת לצרפו לעשרה לכל דבר שבקדושה לא היה ר' זירא מעבירו מן העולם דאף שאין בו איסור שפיכת דמים דהכי דייק קרא (אף שיש בו דרשות אחרות) שופך דם האדם באדם דמו ישפך דוקא אדם הנוצר תוך אדם דהיינו עובר הנוצר במעי אמו הוא דחייב עליה משום שפכ"ד יצא ההוא גברא דברא רבא שלא נעשה במעי אשה מ"מ כיון שיש בו תועלת לא היה לו להעבירו מן העולם א"ו שאינו מצטרף לעשרה לכל דבר שבקדושה כך נ"ל וכו'.

צבי אשכנזי ס"ט:

I have wondered regarding a person created by means of the *Sefer Yetzira*—such as that one mentioned in Sanhedrin [65b], "Rava created a man," and also such as the one attested to my [great] grandfather our teacher Rabbi Elijah, Chief Justice of the holy community of Chelm—whether [such a man] could be included in the [minyan] of ten for matters which require ten such as kaddish, kedushah. Do we say that since it is written [Lev. 22:32], *I shall be sanctified amongst the children of Israel*, that he may not be counted [since he is not a descendant of Israel]? Or perhaps, in light of the statement in Sanhedrin [19b] that "whoever raises an orphan in his home is considered as if he gave birth to him" the scripture would raise up [the golem] to the status of one born to him? […] Here too, since [the man] is the handiwork of the righteous he [might be considered part of] Israel, for [we learn that] "the handiwork of the righteous is their progeny" [if so, the golem might be counted]. It seems to me that since Rabbi Ze'era said, "you are from the fellowship of magicians—return to your earth," that he killed him. And if it had occurred to him that [the golem] could be included among the ten needed for matters of sanctification, Rabbi Ze'era would not have removed him from the world. Even though [killing the golem] is not considered murder, for we explain [the verse, Gen. 9:6] *Whoever spills the blood of a <u>person by a person</u> his blood shall be spilled*—this means a person who was formed in a person, namely a fetus in his mother. Only killing such a person would be considered murder, thus excluding the man made by Rava, who was not formed in his mother's womb. Nevertheless, since the [golem] had some utility, [Rabbi Ze'era] should not have removed him from the world. But certainly, he would not count among the ten. Thus it seems to me, Zvi Ashkenazi, S"T [either "a pure Sephardi" or "he came to a good end"].