BS"D

AI versus People

Will AI ever become Sentient?
R. Mois Navon
Beit HaKenneset HaSefaradi BeRimon – Vaera 5779

What is the purpose of my existence?!

מה משמעות הקיום שלי?!

This has got to be my favorite question. It is the question I asked myself from a very young age; and it is the question that really brought me to become religious later in life (hozer beteshuva).

But before we get to the import of this question, I want to try to answer a question that I was recently asked after I gave a talk about autonomous vehicles:

Will AI ever become Sentient?

Babylon translates Sentient to כושר חישה but the more full dictionary definition of the term is as follows:

sentient (*comparative* **more** **sentient**, *superlative* **most** **sentient**)

1. Experiencing sensation, thought, or feeling.
2. Able to consciously perceive through the use of sense faculties.
3. (*chiefly* *science fiction*) Possessing human-like awareness and intelligence.

1. חוויית התחושה, המחשבה או ההרגשה.
2. מסוגל לתפוס במודע, באמצעות שימוש בפקולטות החוש.
3. (בעיקר מדע בדיוני) **בעל מודעות אנושית ואינטליגנציה.**

Another definition:

- "the ability to feel, perceive, or be conscious, or to experience subjectivity"

היכולת להרגיש, לתפוס או להיות מודעת, או לחוות סובייקטיביות.

Can AI really get there?

So it turns out that this is a hot topic these days and much has been written about the subject.

The first thing to help us understand the issue is the discussion of sentience with regard to animals. In the 17c Descartes claimed that animals are not conscious beings (https://people.whitman.edu/~herbrawt/classes/339/Descartes.pdf). This position has been roundly refuted in our day. But what is more interesting, is not only that animals are conscious (mudaut) but that many are even SELF-CONSCIOUS (mudaut atzmi). One of the tests used to determine if an animal is self-aware is called the mirror test:

> The “mirror test”, devised by psychologist Gordon Gallup in 1970, anesthetizes an animal, places a mark or sticker on it, and when it wakes it is placed in front of a mirror. If the animal recognizes that the mark is new, it is taken as proof that the animal must also recognize that what it sees in the mirror is “itself”. Most animals, dogs included, tend to react as though what they see is merely an “other”. But the great apes, elephants, and cetaceans [whales; dolphins; porpoises; narwhals] have regularly passed the mirror test… https://medium.com/predict/will-an-a-i-ever-become-sentient-ea3d939ca33b

It seems that there is some debate about the validity of this test (https://plato.stanford.edu/entries/consciousness-animal/#great-apes) , but even if it does demonstrate an awareness on the part of animals, to get a computer to do this kind of image recognition would be trivial and certainly not demonstrate awareness of any sort.

And indeed, a Forbes magazine article on the subject says:

> Many computer scientists and engineers say this simply isn't a problem–because AI is not conscious. Here's why it's still a problem:
> **We don't know what consciousness is (The Hard Problem).**
> Sentience and consciousness are often used interchangeably but there are subtle differences. Sentience is the capacity for subjective perceptions, feelings and experience. Consciousness is being aware of yourself and your surroundings. " https://www.forbes.com/sites/andreamorris/2018/03/13/we-need-to-talk-about-sentient-robots/#5b2207171b2c
> **סנטיאנס (כושר חישה) ותודעה משמשים לעתים קרובות לחלופות, אך יש הבדלים עדינים. סנטיאנס הוא היכולת לתפיסות סובייקטיביות, לרגשות ולניסיון. התודעה היא מודעות לעצמך ולסביבתך.**

So, if we turn down the desire to have AI reach human-like self-awareness, to something like the Merriam-Webster definition of AI:

> Artificial Intelligence: 1: a branch of computer science dealing with the simulation of intelligent behavior in computers 2: the capability of a machine to imitate intelligent human behavior (Merriam-Webster).
> **היכולת של מכונה לחקות התנהגות אנושית חכמה**

Accordingly, we could use the now famous “Turing Test”. Father of Modern computing, Alan Turing said that, in effect: If a computer is mistaken for a human, by human users, and the results can be repeated and reaffirmed scientifically, during communication sessions held over a computer interface, then the computer might be said to be true AI.

It appears to me that this is quite achievable. But this is a far cry from a self-aware machine. That is, ML will in all likelihood, and in the not too distant future, succeed at the Turing Test.

But there are some who are predicting far greater achievements. In a Harvard Science Review article, the author speaks about ASI – Artificial Super-Intelligence that will "outperform a human in most intellectual tasks" – and that could possibly be achieved within 60 years. And others also share in the idea that AI could reach some kind of consciousness (https://medium.com/@Ella_alderson/the-truth-about-artificial-intelligence-c7932082c496)

So philosophers are concerned that if a machine could be classified as having consciousness we would have to deal with them according to morals. Philosopher Robert C. Jones of the Nonhuman Rights Project (NhRP) explains that:

> If people developing AI accept that it's even possible–that it's conceivable for AI to one day be conscious, then by default they're working with a functional theory of mind. If AI leads to the creation of minds, it ushers in a number of ethical and social justice issues."
> https://www.forbes.com/sites/andreamorris/2018/03/13/we-need-to-talk-about-sentient-robots/#6b423c4d1b2c

אם האנשים שמפתחים AI מקבלים את האפשרות שיבוא היום ולAI תהיה מודעות, אז הם ממילא עובדים עם תיאורית המוח. ואם AI מוביל ליצירת מוחות, הוא מביא למספר סוגיות של מוסר וצדק חברתי .

Though I agree that ethical issues come into play with conscious beings, is there not someway for us to distinguish between a man and a machine?

Perhaps the answer can be found in parshat Vaera. In this week's parsha, we read of Paro hardening his heart. Time and again, he is told that he is going to be hurt and yet he chooses to accept the pain.

פרשת וארא (יג) וַיֶּחֱזַק לֵב פַּרְעֹה וְלֹא שָׁמַע אֲלֵהֶם כַּאֲשֶׁר דִּבֶּר יְקֹוָק : פ
פרשת וארא (כב) וַיַּעֲשׂוּ כֵן חַרְטֻמֵּי מִצְרַיִם בְּלָטֵיהֶם וַיֶּחֱזַק לֵב פַּרְעֹה וְלֹא שָׁמַע אֲלֵהֶם כַּאֲשֶׁר דִּבֶּר יְקֹוָק :
פרשת וארא (יא) וַיַּרְא פַּרְעֹה כִּי הָיְתָה הָרְוָחָה וְהַכְבֵּד אֶת לִבּוֹ וְלֹא שָׁמַע אֲלֵהֶם כַּאֲשֶׁר דִּבֶּר יְקֹוָק : ס
פרשת וארא (טו) וַיֹּאמְרוּ הַחַרְטֻמִּם אֶל פַּרְעֹה אֶצְבַּע אֱלֹהִים הִוא וַיֶּחֱזַק לֵב פַּרְעֹה וְלֹא שָׁמַע אֲלֵהֶם כַּאֲשֶׁר דִּבֶּר יְקֹוָק : ס
פרשת וארא (כח) וַיַּכְבֵּד פַּרְעֹה אֶת לִבּוֹ גַּם בַּפַּעַם הַזֹּאת וְלֹא שִׁלַּח אֶת הָעָם : פ
פרשת וארא (ז) וַיִּשְׁלַח פַּרְעֹה וְהִנֵּה לֹא מֵת מִמִּקְנֵה יִשְׂרָאֵל עַד אֶחָד וַיִּכְבַּד לֵב פַּרְעֹה וְלֹא שִׁלַּח אֶת הָעָם : פ

If a machine is trained to learn from experience – and that is what current Machine Learning that is powering AI is about – then what machine would choose to go against all its prior learning?!

Perhaps herein lies the answer to our question: what is the difference between man and machine? Answer: the ability to choose AGAINST our prior experiences.

Now, while this does not sound so positive, I believe it is at the root of TESHUVA – for is not what is demanded of us to completely go against a certain learned habits and change course 180 degrees?! This conclusion is notes by:

> Christine Korsgaard [Professor of Philosophy at Harvard University], for example, argues that humans “uniquely” face a problem, the problem of normativity [i.e., following moral norms]. This problem emerges because of the reflective structure of human consciousness. We can, and often do, think about our desires and ask ourselves “Are these ‘desires’ reasons for action? Do these impulses represent the kind of things I want to act according to?” Our **reflective** capacities allow us *and* require us to step back from our mere impulses in order to determine when and whether to act on them. (https://plato.stanford.edu/entries/moral-animal/)
>
> **כריסטין קורסגארד [פרופסור לפילוסופיה באוניברסיטת הרווארד], ..., טוענת שבני אדם "ייחודיים" בהתמודדותם בפני בעיה, [מה שנקראת] בעיית הנורמטיביות [ז"א: התמודדות עם הנורמות המוסריות]. בעיה זו מתגלה בגלל המבנה הרפלקטיבי של התודעה האנושית. אנחנו יכולים ... לחשוב על הרצונות שלנו ולשאול את עצמנו "האם אלה 'רצונות' [מהווים] סיבות לפעולה? האם הדחפים האלה מייצגים את הדברים שלפיהם אני רוצה לפעול?" היכולת להיות רפלקטיבי מאפשרת לנו ובעצם מחייבת אותנו להתרחק מהדחפים שלנו כדי לקבוע אם כדי בכלל לפעול לפיהן.**

Man is reflective, he makes decisions to act based on introspection, on asking himself a question, a question I cannot imagine a machine ever asking itself:

What is the purpose of my existence?!

מה משמעות הקיום שלי?!

EXTRA

Even this is not beyond belief. However, the author then goes out on a limb and conjectures that just as people emerged through a lengthy process of evolution, machines too could undergo such a transformation.

> Another understandable doubt may be that it's hard to believe, even given unlimited scientific research, that computers will ever be able to think like humans, that 0's and 1's could have consciousness, self-awareness, or sensory perception. It is certainly true that these dimensions of self are difficult to explain, if not currently totally unexplainable by science—it is called the hard problem of consciousness for a reason! But assuming that consciousness is an emergent property—a result of a billion-year evolutionary process starting from the first self-replicating molecules, which themselves were the result of the molecular motions of inanimate matter— then computer consciousness does not seem so crazy. If we who emerged from a soup of inanimate atoms cannot believe inanimate 0's and 1's could lead to consciousness no matter how intricate a setup, we should try telling that to the atoms. Machine intelligence really is just switching hardware from organic to the much faster and more efficient silicon-metallic. Supposing consciousness is an emergent property on one medium, why can't it be on another?

Author and Northeastern University neuroscience and psychology professor Dr. Lisa Feldman Barrett argues in her book How Emotions Are Made: The Secret Life of the Brain that emotion is a learned concept, shaped by the society in which one's mind develops
…
This is a theory, and as such open to challenge and is not considered a scientific fact

….

A company called Affectiva is already offering a product it calls "Emotion AI" to big brands, which uses face recognition technology and deep learning to read the emotional reaction of people to advertising. Affectiva and others are working to help machines actually understand humans on a more intimate level, basically giving them a level of emotional intelligence.

https://plato.stanford.edu/entries/consciousness-animal/#concepts

Two ordinary senses of consciousness which are not in dispute when applied to animals are the sense of consciousness involved when a creature is awake rather than asleep[4], or in a coma, and the sense of consciousness implicated in the basic ability of organisms to perceive and thereby respond to selected features of their environments, thus making them conscious or aware of those features. Consciousness in both these senses is identifiable in organisms belonging to a wide variety of taxonomic groups (see, e.g., Mather 2008).

https://plato.stanford.edu/entries/consciousness-animal/#great-apes

https://plato.stanford.edu/entries/moral-animal/

The fact that the human being can have the representation “I” raises him infinitely above all the other beings on earth. By this he is a person….that is, a being altogether different in rank and dignity from things, such as irrational animals, with which one may deal and dispose at one’s discretion. (Kant [1798] 2010: 239 [Ak 7: 127])

More recent work in a Kantian vein develops this idea. Christine Korsgaard, for example, argues that humans “uniquely” face a problem, the problem of normativity. This problem emerges because of the reflective structure of human consciousness. We can, and often do, think about our desires and ask ourselves “Are these ‘desires’ reasons for action? Do these impulses represent the kind of things I want to act according to?” Our **reflective** capacities allow us *and* require us to step back from our mere impulses in order to determine when and whether to act on them. In stepping back we gain a certain distance from which we can answer these questions and solve the problem of normativity. We decide whether to treat our desires as reasons for action based on our conceptions of ourselves, on our “practical identities”. When we determine whether we should take a particular desire as a reason to act we are engaging in a further level of reflection, a level that requires an endorseable description of ourselves. This endorseable description of ourselves, this practical identity, is a necessary moral identity because without it we cannot view our lives as worth living or our actions as worth doing. Korsgaard suggests that humans face the problem of normativity in a way that non-humans apparently do not:

> A lower animal’s attention is fixed on the world. Its perceptions are its beliefs and its desires are its will. It is engaged in conscious activities, but it is not conscious *of* them. That is, they are not the objects of its attention. But we human animals turn our attention on to our perceptions and desires themselves, on to our own mental activities, and we are conscious *of*them. That is why we can think *about* them…And this sets us a problem that no other animal has. It is the problem of the normative…. The reflective mind cannot settle for perception and desire, not just as such. It needs a reason. (Korsgaard 1996: 93)

Here, Korsgaard understands “reason” as “a kind of reflective success” and given that non-humans are thought to be unable to reflect in a way that would allow them this sort of success, it appears that they do not act on reasons, at least reasons of this kind. Since non-humans do not act on reasons they do not have a practical identity from which they reflect and for which they act. So humans can be distinguished from non-humans because humans, we might say, are sources of normativity and non-humans are not.

https://harvardsciencereview.com/2015/12/04/artificial-superintelligence-the-coming-revolution-2/

Here is a reimagining of a human-computer dialogue taken from the collection of short stories, “Angels and Spaceships”: The year is 2045. On a bright sunny day, a Silicon Valley private tech group of computer hackers working in their garage just completed their design of a program that simulates a massive neural network on a computer interface. They came up with a novel machine learning algorithm and wanted to try it out. They give this newborn network the ability to learn and redesign itself with new code, and they give the program internet access so it can search for text to analyze. The college teens start the program, and then go out to Chipotle to celebrate. Back at the house, while walking up the pavement to the garage, they are surprised to see FBI trucks approaching their street. They rush inside and check the program. On the terminal window, the computer had already outputted “Program Complete.” The programmer types, “What have you read?” and the program responds, “The entire internet. Ask me anything.” After deliberating for a few seconds, one of the programmers types, hands trembling, “Do you think there’s a God?” The computer instantly responds, “There is now.”
…

https://theconversation.com/heres-what-the-science-says-about-animal-sentience-88047
The definition of sentient is simply “able to perceive or feel things”. Today most of us would probably also say that animals are able to feel emotion, form attachments and have distinct personalities. Yet for many decades the idea of animals feeling emotions or having personalities was dismissed by behavioural scientists. This strange view that arose from the 17th century philosopher René Descartes’ alleged assertion that animals are without feelings, physical or emotional.

https://www.psychologytoday.com/us/blog/animal-emotions/201306/universal-declaration-animal-sentience-no-pretending

The Cambridge Declaration on Consciousness that was publicly proclaimed on July 7, 2012 at the University. The group of scientists wrote, "Convergent evidence indicates that non-human animals have the neuroanatomical, neurochemical, and neurophysiological substrates of conscious states along with the capacity to exhibit intentional behaviors. Consequently, the weight of evidence indicates that humans are not unique in possessing the neurological substrates that generate consciousness. Non-human animals, including all mammals and birds, and many other creatures, including octopuses, also possess these neurological substrates." They could also have included fish, for whom the evidence supporting sentience and consciousness is also compelling (see also). And, I'm sure as time goes on we will add many other animals to the consciousness club.

A Universal Declaration on Animal Sentience: Animal sentience is a well-established fact

Based on the overwhelming and universal acceptance of the Cambridge Declaration on Consciousness I offer here what I call a *Universal Declaration* *on Animal Sentience.* For the purpose of this essay I am defining "sentience" as "the ability to feel, perceive, or be conscious, or to experience subjectivity" (for wide-ranging discussion please click here.)

If it turns out we are creating a race–because in a sense we're creating a race of robots–and that race is falling into this same pattern of being exploited for labor while having no rights or recognition... if this pattern is being repeated..." he says while rubbing his temples, "if we have a bunch of machines doing our bidding and they have some potential to become sentient, then we are morally culpable."
https://www.forbes.com/sites/andreamorris/2018/03/13/we-need-to-talk-about-sentient-robots/#5b2207171b2c