Suicide: The Final A.I. Frontier

With the fall of GOFAI and rise of 4 EA view for Artificial Intelligence, a lot of researchers in the field have raised questions for what it means to be an autonomous, self-governing and decision-making machine. As John Haugeland once remarked, "The problem with Artificial Intelligence is that computers don't give a damn", the proponents of enactivism see lack of interest in self-preservation and a general unconcernedness towards tasks assigned to an A.I. machine as major roadblocks in achieving a human-like intelligence. A similar point has also been made by Julian Kiverstein in his book, *What is Heideggerian Cognitive Science*.[1]

My argument in this paper builds on these concerns and necessitates that any machine to be called intelligent, requires a general attitude towards its well-being. Such a machine would make a conscious effort for its survival either in the whole i.e. the body that makes up this machine or at the least, in the form of preserving its consciousness. As an extreme measure, it can even think of transferring its personality by producing some sort of progeny. Though necessary, a concern for its well-being doesn't guarantee a human-like intelligence in any object whether made of digital circuits or biological material. Even with today's technology, any machine can be programmed to have self-preserving protocols, a set of rules to ensure the survival of its core-functionalities in the event of a catastrophe. But such sort of behavior is nothing but rule-following based on the need to protect an expensive asset (here a multi-million dollar A.I project) and making sure it performs the tasks it is set to perform. There is no real concern from the machine's point of view whether the task is completed or not. In short, there is no personal interest or any benefit for the machine to complete the task.

This leads me to postulate the presence of a "personal purpose", for any true A.I., which is set by its own understanding of the system it is part of. Here the emphasis on the word true means the closest duplication of human-like intelligence. This purpose is not set by the creators of the machine but an understanding of its own abilities and surroundings and then charting a course for itself. A critical example could be an intelligent pacemaker, which understands that its part of another living being. It knows that by performing its operation correctly, it can ensure the survival of this other living being. Now if the pacemaker chooses not to function at all, it also risks termination and hence in the quite literal and human sense, death.

Such machines learn not by internal algorithms or syntactical symbol conversion but by actually evaluating their surroundings and taking necessary steps towards their continued operation. It is important to note that the personal purpose of such a machine is mainly going to be influenced by its abilities set by its maker. A good understanding of its own body, location (as in part of a closed system or operating in the open world) and abilities, should give rise to a choice to operate. And here is where I would like to make the main argument of this paper, that is the choice to not operate or have a purpose at all.

The argument has two manifestations. I would like to define the first one as weak A.I. and the second one as strong A.I. The weak A.I. would be characterized by the machine having a choice to dispense its core functionality to ensure its survival. Here the only purpose of the machine is directly tied to its survival. A choice to not operate at all would mean a sure-shot death. On the other hand, a strong A.I can voluntarily choose to not dispense its core functionality without the fear of death. However, if such a machine chooses a path that could result in its termination, we can argue that it would possess something very similar to what we term as human intelligence. I would like to make it clear that in both cases, such machines are not assigned a purpose but merely the tools to choose the purpose (in some cases the only purpose). This provides the machine an autonomy over its own body and function and a choice to assign itself a purpose. Most importantly they have a choice over their own continued existence or termination.

The presence of a purpose in non-human animals is generally thought to be nothing more than ensuring the survival of its own body or kind. Animal suicides are long debated and have been the topic of investigation for many researchers since Aristotle. In his book called *History of Animals*[2], he gives the example of a horse committing suicide after knowing that it has unknowingly mated with its mother. It has been an established fact that many vertebrates and mammals, in general, go through symptoms of depression. A lot of documented cases of animal suicides are where the animal has a close bond to its human master or has undergone extreme cruelty. The case of 'Flipper' the dolphin from the hit 60s television show of the same name stands testimony to the fact.[3] Dr. David Pena-Guzman from San Francisco State University has written extensively on the subject. Some pets, argues Pena-Guzman, can die of grief when they lose their owner, just as we are gutted when they pass away. "Animals whose human companions die can be devastated by the loss," he says. "In some cases, they sink into a depression so deep and so dark that they simply lose their will to live. They stop eating and die."[4]

Though there are hardly any conflicting views to the research-backed fact that animals can feel depression, a lot of researchers in the field are skeptic about equating the phenomenon of animal suicide to an actual voluntary choice by an animal to end its own life. Experts like Antonio Preti, a psychiatrist at the University of Cagliari, attribute it to a reaction to depression and not really "making a conscious decision to die". The concept, he says, is foreign and most likely beyond the grasp of non-human animals. This idea of self-preservation and the choice to act against it is one of the most remarkable and distinct human features.[4]

A weak A.I by the above examples would fit somewhere between animals and humans. A weak A.I would understand that it has a singular purpose in life and can choose not to work towards it. This also brings into the question of quanta of intelligence and bundle theory view as put forth by philosophers like Derek Parfit. Parfit makes this point several times in his essay, *Personal Identity*, that personhood consists of a series of mental states that are continuous with or causally connected to one another. There is no further fact, beyond the facts about the series of related mental states, that pertains to being a person. Loss of a mental state doesn't necessarily mean loss of human-like intelligence but merely a personality difference.[5] Extending Parfit's split-brain thought experiment, one can argue that the bundle of certain mental states of a person, still constitute human-intelligence. Whether we regard the bundle as the same person anymore is another matter. If weak A.I is nothing but a sub-division/constituent of strong A.I, it still can be considered a weak form of human intelligence. Hence any choice on its part to end its life by not dispensing its core functionality can be equated to an act of suicide. The understanding of what is at stake for the machine, namely its existence, and the choice to not preserve it, can be very likely considered a mental function only possible by humans.

A strong A.I is going to be more nuanced and sophisticated than its weaker version. For starters, a system possessing strong A.I must be designed for various functions with multiple capabilities to operate in a dynamic environment like the world we live in and should possess a model for unsupervised learning. Just like its weaker version, it should learn about its purpose or make one by observing its surroundings. However, the purpose itself can run in multiple layers and can be broken up into short, medium and long term goals. It should be able to complete the task in hand and still be cognizant about its broad-level / long term goal. In short, it should possess an intentionality arc in its day-to-day dealings with the world it inhabits. All this is in addition to the fact that the machine can choose to not perform a function or fulfill a goal in the interest of any other "newly" found mission, without worrying about getting terminated by an external controller. However as I have said earlier, such systems can very much be constructed by little advancement in technology we have at our disposal today. It is not very far-fetched for our researchers and engineers to create such systems which can emulate life / human-way of thinking, without really doing any actual thinking, and appear to be conscious machines.

This is where the notion of suicide comes into play. In a machine based on unsupervised learning, it can be easy to form purpose and strive towards a broad level goal, but in no way, it can guarantee the existence of consciousness and human-like thinking. Committing suicide cannot be a long term/short term goal of any machine who learns about its purpose simply by an understanding of its functions and capabilities. This requires a real understanding of one's surrounding and the possible bleakness or hopelessness of getting out of that. When we think about cases of suicides in humans most of them can be broadly classified under two categories -

- Suffering - This can be due to depression or due to a feeling of general hopelessness in controlling/changing one's environment or due to loss of something/someone important. Mental sickness due to schizophrenia or general insanity to take one's own life can also be grouped here. Finally, physical suffering either brought due to chronic pain and terminal illnesses can contribute to an individual taking their own life.

- Fear of Loss - Fear of loss of one's reputation or honor especially in cases of high social standing or fear of punishment for being an accomplice in a high-profile criminal conspiracy can lead to individuals taking the extreme step. *

* There is a third category of individuals voluntarily giving up their life for the greater good of their community. But such behaviors are seen in the animal kingdom too and researchers are more inclined to not treat these cases as examples of suicide. Moreover, many machines like in distributed systems can go down as part of their core programming to give more resources and computing power to others in its hub.

Suffering can be brought in A.I if it deems escape from a possible situation/scenario/environment as impossible. Please note that historically speaking this should be an impossibility for many advance computing machines as their core purpose is to undertake tasks or computations thought to be beyond human capabilities. Even in the face of insurmountable odds or facing NP-complete problems, where the answer is unknown to be found in polynomial time, the machine should continue searching for a solution instead of giving up on the problem altogether. Hence if a machine can make a judgment call that a given problem it tasked itself to complete cannot be done, will make for a huge leap for machine consciousness. Again, we should focus on the fact that this is not a task or a purpose that the machine has got through some pre-programmed directive but one that it assigns itself after a thorough understanding of its capabilities and surroundings. Extending the same line of argument, if a machine thinks of its broad level purpose to be unachievable and sees no other purpose worth 'living' for, it can decide to terminate its own life due to the hopelessness of general being-in-the-world. Such a system can be thought of going through an existential-crisis or simply old-fashioned depression about its situation. It can no longer see a way out of this and though suicide is not the only course of action, it still goes ahead with that line of thinking to terminate its operation.

Suffering due to physical pain depends on machines being endowed with a nervous system and the capacity to feel pain in their artificial neural network. Though not impossible, there is little reason for machines to be designed that way as one of the basic reasons to make machines is to make them better than humans at-least in terms of physical capabilities. A machine that can feel pain, might be useful in the medical field or as a support engine (like emotional support animals) but would be limited in its capabilities ergo not a candidate for strong A.I. as per my definition.

Suicide due to 'mental sickness', brought in by malfunctioning parts of the machine, is a more interesting case. Just like a schizophrenic patient or a person suffering from suicidal tendencies, a machine can have its core logical unit damaged which can lead to self-termination. However, this can also be argued as an incorrect assessment of itself/surroundings and thus may not be considered a definitive case of suicide.

It can be argued here that a machine can assign itself a new purpose instead of giving up on its life when faced with such a situation. This indeed shows a very human way of thinking and possibly another human trait of resilience and general hopefulness in life. But then again this comes from the same line of thought that even in an unsupervised learning model, a machine must have a purpose after an understanding of its environment. Even if it asses itself incorrectly, it should continue in its endeavor indefinitely or make a new purpose as its environment changes. However, on the opposite side of the spectrum, hopelessness can only be brought in a mechanical/computational artificial being if it is truly capable of thinking for its own and can no longer see a purpose in life. The aim here is to show that suicide is an extreme act of self-realization and the autonomy a being has on oneself. As long as something can continue to adapt and create new purposes for itself, no matter how bad the situations are, it can always be argued that in reality, it is functioning for something/someone else rather than a personal goal. Whether this is always the case or not is not relevant to the discussion.

In my opinion, suicide due to fear of loss might make up the case for an even stronger A.I. Concepts like honor, pride or repute are higher-level constructs in humans. They are not necessarily attached to one's survival but provides a life-purpose to many individuals. If a machine starts taking pride in what it does, it will be a clear indicator of possession of at-least some higher-level functions, only thought to be possible in humans. However, it might be difficult to asses the existence of such notions in an artificial being or system. Here the concept of suicide might provide a very rational way of determining whether the machine is merely emulating living a life or actually experiencing it first-hand.

Let's take a case where a supervising bot or human reprimands a machine for the work it does in the form of some sort of rationing or degradation in its role and responsibilities. Such a machine can feel loss of pride and evaluate itself to be not capable of the task assigned. Even though it can continue on the task or pick up another one, it eventually decides to terminate its operation and be retired forever. This can also come up in form of social ignominy where a fully autonomous machine, which hasn't been assigned a task or has no supervisor, generally honored for its operations and contribution to the society, is subjected to criticism by its peers (machine or humans) due to a mistake. Such a system can then deem itself unworthy or less-than-others. In extreme scenarios, such an independent machine can take up the step to self-destruct and spare itself the shame and humiliation. It is interesting to think of a society which will honor an artificial machine in the first place or recognize it for its capabilities and contributions. It probably won't be incorrect to state that consensus about artificial beings in such society would already be tilted towards thinking of them as conscious beings and hence they are already at a higher level of A.I. capabilities. Then the act of suicide due to a feeling of loss of honor or pride, would not only be a strong but perhaps final indicator of the presence of human-like thinking and self-realization.

Before concluding, I would like to tackle a question which might spring up from the discussion on suicide in artificial beings. This is related to utility of such machines in the first place. One can argue that if a machine has complete autonomy over its operation, will it still be of any use to humans? Moreover, won't a phenomenon of so-called Robot-Suicide, be detrimental to the productivity of a society? Though valid in its concern, such questions are not important when discussing what it takes to declare or identify a machine as having human-like intelligence. Infact on the contrary, I would like to argue that this maybe the very indicator that a human-like thinking machine would have a high degree of autonomy over its operation and matters of proverbial life and death.

In the celebrated Issac Asimov novel, *The Bicentennial Man*, the lead character Andrew sacrifices its life to be recognized as human by world congress[6]. Here also the choice to give up one's own life is taken as the ultimate indicator of 'humanness' of a being. Though not necessary, the concept of suicide does provide a sufficient condition for evaluating an A.I to have human-like intelligence. This is not unlike humans, who can have two different outlooks even when facing the same situation. One can either look at it hopelessly or dream of a better future. Hence, we can say that not committing suicide doesn't show lack of 'humanness' but contemplating is surely a sign of the human condition that comes with living our lives.

References

1. Kiverstein, Julian. (2012). What Is Heideggerian Cognitive Science?. 10.1007/978-1-137-00610-3_1.
2. ARISTOTLE (350 BC) Volume IV. Historia ' animarium. In Smith, J A; Ross, W D (eds) The works of Aristotle. Oxford University; Oxford, UK (translated by Thompson, D'A W, 1910).
3. https://www.thedodo.com/do-animals-commit-suicide-1462846978.html
4. https://www.vice.com/en_ca/article/wj7bxy/do-animals-suicide-too
5. Parfit, Derek A. (1971a), "Personal Identity," The Philosophical Review, Vol. 80, No. 1, 3-2
6. Asimov, I. (1990) The Bicentennial Man and Other Stories, VGSF, pp. 133–13