Melvin's A.I. dilemma: Should robots work on Sundays?

Ivan Spajić / Josipa Grigić, Zagreb, Croatia

This paper addresses the issue of robotic religiosity by focusing on a particular privilege granted on basis of religiosity: work-free Sundays. The paper is divided into two questions; could artificially intelligent robots develop religiosity and should that be a reason to give them Sundays off. Both questions are answered in forms of short overviews of pro- and counter-arguments. Since answers to the second question greatly depend on the answers to the first one, a cross-relation between arguments on both sides is created, creating a dilemma which authors believe puts in question the purpose of creating A.I. in the first place and emphasizes dilemmas concerning human nature.

Keywords: A.I., religiosity, robots

Discussion

Let us say we live in some near future, and we are observing a robot unit, which we will from now on call Melvin. Melvin has a high IQ due to his knowledge and an ability to adapt, but he is also sensitive to his environment and continually learns about the ways of the world. Let us say he is in charge of the environmental protection for a given area. It should not then come as a surprise that one day Melvin realizes humans are not the supreme beings and that they are not causes in themselves, although they are his makers. After a while, Melvin becomes religious. Given the situation, he becomes a Christian and after some time asks his employers to give him Sundays off. How should the employer answer?

This paper addresses the issue in two steps, asking two questions:

Q1: Could robots develop religiosity?

Q2: If robots do develop religiosity, should they work on Sundays?

Q1: Could robots develop religiosity?

Arg(1.) Robots could never develop religiosity.
Arg(2.) Robots could develop religiosity.

Elaboration:

Arg(1.) Robots could never develop religiosity.

Arg(1a.) Such a level of development is not possible. Robots are preset and thus cannot posses the necessary required freedom of thought to develop religious thought.

Argumentation:

(P1): In order to compose a person capable of religious thought, it is necessary to compose elements which are not completely rationally analyzable (and therefore can not be known).
(P2): If an element is not completely rationally analyzable (and therefore known), it cannot be pre-programmed.
(P3): If an element cannot be pre-programmed, it cannot become a part of robot's programming.
(C): A robot cannot be capable of religious thought, because it cannot make it a part of its behavior.

(Obj1): Refers to Arg (1a.P1). It is possible for all the elements essential to a person capable of religious thought to be completely rationally analyzable.

(Obj2): Refers to Arg (1a.P3). Even if all elements are not rationally analyzable, artificial intelligence with a capability of adaptation, learning and knowledge of cause-effect and action-reaction of physical and inter-subjective relations might develop elements which have not been *a priori* installed. It may be similar to artificially created organisms that learn on their own once they are set into natural environment. Many researchers on A.I. today already believe that A.I. requires building an entity capable of learning; i.e., that we cannot simply program an intelligence.

Arg(1b.) Highly developed artificial intelligence would have no need for religious thought.

Argumentation:

(P1): Artificial intelligence informs itself through scientific inquiry of physical data.
(P2): Scientific inquiry cannot be conducted in fields of religion.
(C): Artificial intelligence does not inform itself through inquiry in fields of religion.

(Obj1): Refers to Arg (1b.P1.) Similar to Arg(1a.), the adaptive capabilities of an A.I. would allow it to ask questions of relations and causes in border cases, such as the constitution of matter on submaterial levels or the origin of matter prior to space-time. Such metaphysical questions could rule out scientific inquiry of physical data as the only possible method.

(Obj2): In reference to Arg(1b.Obj1), the question of personal religion can be resolved as well. Possible realization of the fact that humans are not causes in themselves and do not hold all the answers, might cause the highly adaptable A.I. to learn to relate to humans in their pursuit of meaning. An A.I. that is aware of its unique position in time and space and starts seeking purpose for its abilities outside the pre-programmed Arg(1a.Obj2), might develop an existentionalist complex of *being thrown into the worl*d. This could be a good reason for developing personal religious thought.

(Obj3): Refers to Arg (1b.P1.) It may be possible that feelings, and not just sophisticated cognition, are required for religiosity.
But it may also be that the development of an A.I. itself requires emotions, too. To elaborate this point, we should observe guidelines that form basic interactions between conscious systems, such as humans. They are of a particular value,

because adaptability required for development of A.I. relies on A.I.'s interactions. These primary guidelines are basically simple and straightforward even in complex living systems (e.g. avoid collision with other bodies), but results that emerge from them during system's interaction with its surroundings are much more complex. The strongest property that expands these guidelines into complexity is the fact that the aim of these guidelines is not completely specified. In her research of requirements for development of artificially intelligent systems, Susan Stuart deals with such complex emergent systems. She states that, if we don't know emergent behaviours a complex system is aiming for, what could emerge from it is something that might possibly be irreducible to physical facts and relations. [1] Development of emergent properties in interacting systems is therefore emphasized when the system doesn't know the emergent behaviours he/she/it is aiming for.

With this lack of specified goals in mind, when we observe A.I.'s adaptability, we notice a gap. Inside this gap, despite all its interactions, A.I. could exhaust its capacities and still not develop an opinion of its own or a consistent pattern that would enable it to properly adapt to novelty. In order to fill this gap, authors such as Keith Oatley and Jennifer Jenkins[2], along with Stuart, suggest that we need to involve emotion. They believe emotion has/is the necessary property to halt the system for long enough to create a directed reaction. This emotional reaction signals the system a need for thought about adaptation and changes in thought and behaviour. Such focused thought is, according to Stuart, what leads to development of real consciousness, a necessary prerequisite for A.I. Therefore, emotion could play a vital role in development of A.I.

Arg(2.) Robots could develop religiosity.

If objections to Arg(1.) are taken into consideration, then Arg(2.) is a justified possibility.

Q2. If robots develop religiosity, should they work on Sundays?

Arg(3.) Robots should not work on Sundays.
Arg(4.) Robots should work on Sundays.

Elaboration:

Arg(3.) Robots should not work on Sundays.

Arg(3a.) Robots were given a possibility. Once Melvin was given the possibility of upgrading himself to the level of self-awareness and cause searching, would it be moral to reset him? If he develops a sophisticated consciousness that passes modern tests and qualifies him as a person at a level similar to at least that of a small child or a person with affective disorders, are we to deny him his experience and positions?

Arg(3b.) Robot functions within its purpose, which is to adapt and research for the benefit of humanity (and the environment in Melvin's case). By developing personal religious thought, Melvin has not necessarily strayed from the purpose he was created for; he merely expanded it.

Arg(3c.) Robot's observation of human insufficiency is indisputably correct. If it is in his nature to seek causes, then he has the right to seek them outside the domains that have been set by humans.

Arg(3d.) Denying the robot his right to have Sundays off could lead to disputes within the society. In reference to Arg(3a.), we can imagine a slippery slope which starts here; if we deny Melvin his religious thought, do we deny it to cyborgs, too? Where exactly does the borderline between a robot and a human with brain implants lie? What about artificially created biological organisms that had their beginnings in electro-stimulation of laboratory organic matter? Is all the processing that comes after stimulation artificial and only second class? Would a human created in this way be allowed to have Sundays off?

Arg(3e.) Denying a robot his right could also have destructive an impact on the religion of his choice. If a robot is denied his right, the reality of the act of faith in itself could be brought in question. If Arg (1b.Obj1) and Arg(1b.Obj2) are taken into consideration and justified, then Melvin truly has a need, a desire, to be a person of faith. Denying his right solely on the fact that he is not human may be observed as a problem similar to that of aliens and religions. An alien is most apparently not a human. Yet, if he has the capabilities, he may choose to become a member of faith and a religion.

Arg (3f.) By practicing faith, Melvin desires to *do good*. If, in reference to Arg (3b.), Melvin works within the boundaries of his general purpose of adapting and researching for the benefit of humanity, his practice of faith is undeniably an act of good (as seen by humanity). And if this desire isn't proved to be an act of opportunism (*see Arg(4a.)*), then the act in itself should gain some validity.
If, on the other hand, Melvin has expanded his general purpose, he may as well be in danger of working out of the boundaries set by humans in accord with Arg(3b.), but due to Arg(3a.), it would be disputable whether we should reset him or not, since he was given the possibility to upgrade and has become a creature of free will.

Arg(3g.) If, according to Arg(3b.), Melvin works within his designated purpose, religious satisfaction increases Melvin's efficiency.

Arg(4.) Robots should work on Sundays.

Arg(4a.) Melvin could be an opportunist. If all elements of religiosity prove to be reducible, and consequently analyzable according to Arg(1a.Obj1), then we might be in danger of being fooled by Melvin. He could just be modifying himself and developing towards religiousness because he simply does not want to work. If this were true, it would then be in collision with Arg(3b.), and sufficient a reason for not letting him have Sundays off, although it would be disputable whether we should reprogram him, due to Arg(3a.).

Arg(4b.) Working on Sunday would not necessarily damage Melvin's state of mind. Although he might feel formal dissatisfaction - or put differently, simply not agree with his boss - Melvin might not experience any permanent traumas from being denied this right; his *mental health* would not be damaged, because his parts are

replaceable. Any damages in his mechanism could be repaired by his regenerative systems, or externally, by his administrator.

Arg(4c.) Robots are created to work. Why is a creature that could choose not to work created to work? Why should we let it choose?

Conclusion

This paper does not necessarily provide the answer to the topic question. In fact, depending on the framework, we can attain enough arguments both for and against. What authors believe this paper accomplishes is to line out both groups of arguments and point out the ethical implications they have on societies of humans and potential robots. As long as these arguments are in cross-relation with one another, we should, in accordance with Arg(4c.), ask ourselves whether it makes sense to create an A.I. at all, if that A.I. has a possibility of personal development and an option not to work if it/he/she does not feel like it.

Even more importantly, this analysis provides us with dilemmas concerning what we consider to be human properties at present. If we take into regard the necessary prerequisites for the development of artificially intelligent organisms, and should they prove to be outside the scope of rational, then it might turn out that we are currently using many wrong approaches in regard of naturally intelligent organisms. For example, in Arg (1b.Obj3.), we consider the possibility that a necessary prerequisite for proper decision making is emotion. Yet, in current scientific regard of people, emotion is regarded as an obstacle for proper decision making. For further development of any society, be it A.I. or human, this issue needs to be resolved.

It is of course possible that Arg (1b.Obj3.) is wrong, but it has shown substantial reason to be correct. Also, what are its' alternatives? As far as we can see it, they are Arg (1a.Obj1.), the argument that states that all mental and spiritual properties are reducible and analyzable; and along with it Arg (1b.Obj1.) and (1b.Obj2.) that punctuate a lack of purpose and an existentionalist complex of *being thrown into the world*. If we regard humanity in this way, it becomes apparent that humans themselves are robots in many regards. We could draw a parallel here to Daniel Dennet's statement; «We're all zombies.» or zimboes.[3]

However, if that is the case, and Arg (1b.Obj1.) and (1b.Obj2.) are also present, then natural intelligence is without a purpose, all that is left is raw data and no real reason to process it. Just like in (1b.Obj3.), we begin to wonder whether *we* too could exhaust all *our* capacities and never develop a consistent pattern that would enable us to properly adapt to novelty. Furthermore, it brings in question the term *properly* itself, since proper behaviour requires something to aspire to. Outside the scope of primary survival, all other human activities would be random, useless and mindless. Without such elements that stop us in tracks and demand a decision, the very concept of ethics is disputable.

We should take into consideration that Aristotle for example foresaw such dilemmas in his own society, and introduced *phronesis*, the virtue of moral thought. He regarded it to be more important than other two intellectual virtues; *episteme* (scientific knowledge) and *techne* (knowledge of know how). He saw *phronesis* as

the activity that balances analytical and instrumental rationality of *episteme* and *techne*, by means of clarifying values and interests. Beside rational, *phronesis* involves conscious awareness of the environment, consistent experience and feeling for balance. It isn't certain whether we should introduce *phronesis* into modern research and the way we regard the world in general. However, it is advisable we resolve afore mentioned issues in that or some other manner before we further them by developing artificial intelligence. Because, once we do develop A.I., we will have a problem very common to problems of bioethics in research of genetic manipulation and cloning. That is to say, if the A.I. is given the possibility to make free choices as a person, then once it is set into the world, we should not un-set it.

Acknowledgments

Authors would like to thank professors dr Neil Levy of the University of Oxford, dr Alexander Batthyany of the University of Vienna, and dr Kristijan Krkač of the University of Zagreb for the help and advices during the development of this paper.

Email: ispajic@ffdi.hr

[1] See Susan Stuart, *Artificial Intelligence and Artificial Life – should artificial systems have rights?,* 2003
[2] See Jeniffer Jenkins and Keith Otley, *Understanding emotions*, Basil Blackwell, 1996
[3] Dennett, Daniel C. (1991) *Consciousness Explained*. London: Penguin Books. 1993.