Jan Romportl
Pavel Ircing
Eva Zackova
Michal Polak
Radek Schuster (eds.)

Beyond AI: Artificial Golem Intelligence

Proceedings of the International Conference
Beyond AI 2013
Pilsen, Czech Republic, November 12–14, 2013

ISBN 978-80-261-0275-5

Published by University of West Bohemia

Creating Free Will in Artificial Intelligence

Alžběta Krausová[1] and Hananel Hazan[2]

[1] Faculty of Law, Masaryk University, Brno, Czech Republic
betty.krausova@seznam.cz
[2] NeuroComputation Lab, Department of Computer Science
University of Haifa, Haifa, Israel
hhazan01@cs.haifa.ac.il

Abstract. The aim of this paper is to provide an answer to the question whether it is necessary to artificially construct free will in order to reach the ultimate goal of AGI to fully emulate human mental functioning or even exceed its average capacities. Firstly, the paper introduces various definitions of will based in the field of psychology and points out the importance of free will in human mental processing. Next, the paper analyzes specificities of incorporating will into AGI. It provides a list of general justifications for creating artificial free will and describes various approaches with their limitations. Finally, the paper proposes possible future approach inspired by current neurobiological research. The paper concludes that a mechanism of free will shall form a necessary part of AGI.

Keywords: artificial intelligence, artificial general intelligence, free will, volition, determinism, indeterminism, real random generator

1 Introduction

The highest goal of the science of Artificial Intelligence (AI) has been to create a being that can be considered as equal or even superior

to a human in the sphere of intelligence. This goal is made yet more difficult in a specific field called Artificial General Intelligence (AGI) that attempts to create "a software program that can solve a variety of complex problems in a variety of different domains, and that controls itself autonomously with its own thoughts, worries, feelings, strengths, weaknesses and predispositions" [1]. In other words, AGI aims at creating a being that not only resembles a human in the sphere of intelligence, i.e. "[the] ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought" [2] but also in all other aspects of human functioning.

Given this aim, the science of AGI needs to explore fields like neurosciences, psychology and philosophy in order to be able to emulate such degree of evolution. It has been proven that, unlike animals, human beings possess special processing capabilities resulting from a specific construction of their brain. Out of many important functions of a human brain one function is, however, probably the most outstanding, widely discussed, examined and doubted: the free will.

Existence of will as a specific quality in a person is recognized by modern psychology. Nevertheless, there remains an important and so far unresolved question: Is this will free, or is it deterministic? Moreover, does it matter if this will is free and do people need to consider themselves free anyway?

Since the concept of free will is so puzzling and still so characteristic for the species of *homo sapiens*, the purpose of this paper is to explore the problem of its construction in the context of creating the desired AGI being that fulfills the criteria of emulating human mental functioning or even exceeding its average capacities. Namely, the research question of this paper is whether it is necessary to artificially construct free will in order to reach the ultimate goal of AGI.

In order to formulate an answer at least two questions need to be addressed. The first question focuses on how will and its freedom are defined, and what are the limitations of this understanding. Given

the ultimate goal of AGI, the nature of will must be explored and, moreover, it needs to be proven that will has an important role in human mental processing.

The second question deals with specificities of incorporating an element of will into AGI and its usefulness. Assumed reasons for such incorporation will be summarized together with current approaches and identification of their constraints to answer this question. Moreover, main problems relating to creating indeterministic will shall be pointed out. Special focus will be put on creating real random generator and its role in creating artificial free will. Lastly, the latest neurobiological research shall be described in order to possibly suggest a way to inspire future attempts of AGI.

2 Definition of Free Will

Will or "volition" is the key concept of this paper. In order to proceed further with the examination of this concept, it is necessary to define at first, what is will as such, and later to explain how we understand free will which is, contrary to simple volition, distinctive with its specific characteristics. Finally, philosophical constraints of understanding freedom of will shall be briefly mentioned.

Will or volition itself can be defined in the simplest way as an "an act of making a choice or decision", as well as "the power of choosing or determining" [3].

However, there exist also different and more complex definitions of volition. Even in the field of psychology, opinions vary. For instance, a social psychologist Kurt Lewin considered volition to comprise two aspects: so called goal setting and goal striving. Goal setting represents the motivation of a person, her desires and needs, while goal striving means the particular ways in which a person then exercises her will in practice [4]. A more specific definition was later provided by Julius Kuhl that proposed so called action control theory. According to him, volition can be understood as a mechanism of action control that decides about which strategies out of

those available will be used and in which order so the goal would be achieved [4].

Apart from the above mentioned definitions, there are many specific theories exploring the notion of volition. However, in general a will can be understood as a special mechanism indicating intention or purpose and governing mental processing of information in order to achieve the indicated purpose.

Free will, on the other hand, is a concept that is enriched with specific features that are not present in a simple will defined above. The reason is that a simple will might be programmed or conditioned to function in a certain rigid and predictable manner.

The notion of free will has been explored mostly with regard to humans. This is due to the fact that experiencing a freedom to choose and then act upon such choice has a very private and unique nature. Each person most probably perceives this experience in other way. Free will is simply defined as "a freedom of humans to make choices that are not determined by prior causes or by divine intervention" [5]. However, this concept has been understood differently by various philosophers; for instance as an ability to choose deliberatively based on desires and values, self-mastery (i.e., trained freedom from desires), an ability to identify true personal goals higher than basic need and to act upon those goals, or as so called "ultimate origination", i.e. an ability to act otherwise [6].

Psychologist Chris Firth mentions important characteristics of free will: "the origin of the behavior lies in the behaving individual rather than in the environment. The behavior is self-generated or endogenous ... a response with no eliciting stimulus" [7]. However, with regard to the philosophical notions we deem the definition to be broader.

Free will can be defined as an independent force that is able to determine own purpose, create own intentions and change them deliberately and unpredictably, form respective goals, pick strategies based on recommendation from an intelligence unit, and give orders to perform particular chosen actions. Such will is free to ignore external stimuli or past experience that may predict future outcomes.

With regard to this definition, will plays an important role in human mental processing. To speak metaphorically, free will can be compared to a driver of a car who has a freedom to change the route at any time according to her feelings and desires that may not be logical. Within this metaphor the driver can also choose to turn away from a route that others normally follow, leave prints on previously untouched ground and originally influence the outer world.

Since freedom of will lies in the ability to act contradictory to logical reasoning from past experience, i.e. unpredictably, employ emotions instead of cognition and for example decide randomly in situations when two different courses of action have a completely same probability to reach a desired goal, a respective subject characteristic with free will is enabled to develop own cognition, innovate and form better survival strategies [8].

Deploying volition and self-control in humans leads to activation of other specific processes; for instance attempts to conserve own resources [9]. Moreover, perception of own free will, or on the other hand perception of its absence, has an impact on formation of own identity and approach of an individual to solving problems. For instance, it has been proven that people tend to give up responsibilities and start to cheat when they are exposed to deterministic arguments [10]. In general, perception of being autonomous influences behavior in various domains [11, 12].

After illustrating the importance of free will and its perception in human mental processing, it is necessary to make at least a short note on its existence. Although the question of existence of free will belongs to one of the most significant problems in philosophy, it has not yet been possible to scientifically prove it. This crucial question deals with problem of mental causation, i.e. how pure thoughts or mental acts can influence the matter. A precise mechanism is not known yet. Monistic approach solves the question of causality by stating that any mental state is caused by organization of matter, therefore thoughts are mere products of matter and not a force influencing the matter [13]. Dualistic approach on the other hand presumes existence of a separate mental force that influences and

changes the matter. This, however, makes a scientific approach impossible since it considers spiritual to be unexplainable [13].

The absence of scientific proof of free will represents the most serious limitation of its understanding. However, for the purpose of our paper we consider that it is not important to solve this fundamental philosophical question at this point. What psychologists now call free will is undoubtedly an important element in governing mental functioning and, therefore, needs to be reflected in AGI as truly as possible. Within the course of construction of such will researchers may then come with new ideas that may contribute to the solution of the argument between determinists and indeterminists, as well as materialists and idealists.

3 Incorporation of Free Will into AGI

3.1 Justification of the Effort to Create Artificial Free Will

With regard to the research question of whether it is necessary to construct free will in order for AGI to reach its goal of emulating human mental functioning or even exceeding its average capacities, it is necessary to ask at first whether, given the high complexity and at the same time uncertainty of the concept, there is any meaning in attempting to create free will in AGI and whether the highest goal of AGI is justifiable at all. Maybe the humanity would benefit enough from a highly intelligent being that functions only in a deterministic way as we understand it now.

Advocates of creating free will in AGI mention important benefits. First of all, scientists believe that construction of free will in an artificial agent would enable us to understand better human nature and learn about it [14]. Secondly, we consider it as a fair presumption that free will would enable artificial beings to develop their intelligence to much higher level and, therefore serve people better. A deterministic agent or intelligent system that simply creates own rules upon existing rules without being able to deviate from them or

to make random decisions is prevented from being able to gain own individual and original understanding. In this sense, artificial agents could be prevented from gaining wisdom, i.e. knowledge how to use knowledge. Finally, some consider as probably the greatest benefit to the humanity having an equal that would give us an opportunity to define ourselves as humans in relationship to the new species.

As opposed to the mentioned positive side of artificial free will, there arise also concerns about social implications. Current legal systems are based on presumption of existence of free will. Humans are the only subjects who are entitled to enter relations protected by state and enforceable by state power. Law would then need to solve the dilemma of who is to be protected in case a new entity comparable with humans would come into existence. Should there be species neutrality as long as the species have the same abilities and awareness? Should these beings be protected at least like animals given the condition that they can feel suffering? Moreover, another issue rises with a possibility that we may later not like what we would have created. At the same time the scientists would then face an ethical question whether these artificially created beings could be destroyed. All these questions are for now highly theoretic. Since we have not yet experienced the particular problems which cannot be all precisely predicted, we can unfortunately only guess. But even these guesses are important. For instance one of the classic arguments against creating a being equal to a human is a fear of machines becoming more powerful than people and possessing the same desire to control the outer environment such as people strive for. This fear although currently unreal may be prevented in the future by taking appropriate steps during research.

The answer to the question of the very purpose of AGI seems to be based on balancing possible pros and cons. Since the construction of free will in AGI is not an easy quest, it is presumable that there would be constant monitoring of progress in development and advantages and disadvantages of creating and incorporating such new beings into the society would be evaluated simultaneously together with assessment of new risks. A possibility of learning more about

us provides an extra advantage to the human kind and a reason why to continue persuading the goal of the development of the ultimate AGI.

3.2 Models of Artificial Free Will

As it has been mentioned earlier, philosophers and scientists have not yet agreed on whether there exists free will in humans. Both sides come with strong arguments. Determinists refer to causality as the basic principle ruling the existence while indeterminists claim that while causality is valid, the outcome cannot be predicted with absolute certainty. Some advocates of free will postulate that free will represents an original cause itself.

There have been various approaches by computer scientists aiming at reflecting volition or even free will in artificial intelligence. Approaches vary from creation of deterministic will that is called "free" to proposals to emulate free will resembling human mental functioning.

In this chapter at first a deterministic model will be described and assessed from the AGI's point of view. Next, an approach to human-like free will shall be presented. Finally, one of intermediate stages will be mentioned as well.

In 1988 John McCarthy proclaimed that with regard to free will "the robots we plan to build are entirely deterministic systems" [15]. Later in 2002 – 2005, he proposed a model of *Simple deterministic free will* [16] in which he reduced the notion of free will to (1) computing possible actions and their consequences, and (2) deciding about most preferable action. As an essential element he considers knowledge of choices. This approach refuses complexity of a system to exhibit free will.

Although this proposed model seems to be effective for the existing AI, it seems that such notion is not suitable for AGI purposes and emulation of human mental functioning since it is too simplistic. From psychological point of view, human mental processing is claimed to be based on three cooperating elements: volition,

cognition and emotions [17]. Avoiding or reducing impacts of these elements in the processing then prevents making unpredictable solutions of which humans seem to be capable. Although some experiments have been made to disprove existence of free will (namely Libet's experiment), results of these experiments have been widely criticized and not fully accepted [8]. Moreover, unpredictability of human decisions is highly presumable with regard to the very biological nature of a human brain ("The brain is warm and wet, unpredictable, unstable, and inhomogeneous.") and principles of quantum physics [18]. According to those principles it is possible to determine only probabilities of future behavior but not exact future behavior [19].

An argument against existence of deterministic free will based on causality was also made by Perlovsky. He claims that causality reflected in logic is prominent in consciousness, but consciousness does not represent "a fundamental mechanism of mind" [13]. According to him in computer science dynamic logic is necessary to overcome the issue of complexity of mind that has own hierarchy. Moreover, conditions for existence of free will that can be formalized were already proposed and based on physical theories. These are said to be based on pairwise interactions of particles. Research shows that free will can in principle exist in case of interaction between three or more particles [19].

With regard to these facts it is obvious that a concept of free will should not be dismissed in AGI as inherently impossible or useless. It is, therefore, necessary to look at other, more complex models of free will emulation or their proposals. Much more favorable approach to artificial (mechanical) free will was taken by Manzotti. He claims that "free will stems out of very complex causal processes akin to those exploited by human beings. However, it is plain that simple deterministic devices are not up to the task" [20]. He states that freedom of an agent lies in capability of making real choices, i.e. choices that are not random but also not resulting only from external causes. He mentions a concept of gradual freedom in which freedom of an agent depends on its complexity and a degree to which

individuality of an agent is expressed [20]. A degree of freedom in decision is also related to the degree of involved causal structures in an agent. An action resembling an instinct is considered to be much less free than an action involving own causal structure formed by individual history.

The presented technical approach is much more complex than simple deterministic free will. However, it does not provide any particular solutions. Only conceptual guidelines are outlined. Moreover, many constraints and problematic concepts to be solved are mentioned: temporal integration in an agent, polytropism, or automatic and conscious responses [20].

The two models, one of simple deterministic free will and the second of human-like free will, represent two ends on a scale of AGI development. It is obvious that any development is gradual (various approaches were briefly summarized by McCarthy and Hayes [21]); therefore, one needs to presume stages in which technology will improve over time. It has been shown that free will is rather difficult concept and includes many components. One of its main characteristics is unpredictability. As it has already been argued, the very unpredictability is caused by the biological structure of the brain [18]. Randomness represents its inherent feature. Therefore, this component should also be included in order to reach the next step in developing artificial free will.

In terms of computer science, however, creation of real random generator has been quite a difficult task to accomplish. The question is how can a deterministic model produce indeterministic results while it is working based on laws of logic? Software-generated randomness can be computed and is not then truly random. Fortunately, new research shows paths how to create real randomness. Some recent random generators are based on physical phenomena and use noise sources such as chaotic semiconductor lasers [22]. The most promising research is though in the area of quantum physics. Quantum randomness was proven incomputable and "is not exactly reproducible by any algorithm" [23]. The next step in developing artificial free will would then be incorporating and testing quantum

random generator in order to provide AGI with a mechanism that can at any time provide it with a possibility to decide in a completely illogical way.

3.3 Possible Future Approach

Previous chapters outlined the current knowledge about volition, free will and attempts so reflect this characteristic in artificial intelligence. This last chapter should focus on the future and other possible steps or sources of inspiration for creating free will in AGI. One of the promising fields is neuroscience that studies neural systems and mechanisms that underline psychological functions.

With regard to biological basis of volition, very interesting research has been done by prof. Peter Ulric Tse who studied activity of neurons. Based on the results of his research he claims that free will has a neurological basis. According to his theory neurons react only in case some particular and predefined criteria are fulfilled. Decision of a person and her will are conditioned by the current structure and definitions. However, freedom of a person and her will lies in rapid neuronal plasticity. After a person made a decision, the neurons can reprogram themselves and define new criteria for future decision-making [24].

These findings are in fact in line with previous findings and specified psychological characteristics of free will. A person bases her decisions on previous experience. However, in case of employing complex cognitive processes, reaction can be changed for future cases. There is also delay in performing decisions by humans so it is presumable that before acting in a decided way, the particular person can quickly reconsider the action and act otherwise. To other humans such action seems instant and, therefore, free.

The comprehensive description of neural functioning by prof. Tse provides a great starting point for computer scientists to try to emulate similar functioning in the sphere of artificial intelligence. It seems to be the most feasible to use neural networks in order to achieve the same manner of functioning.

However, a serious limitation still persists even in this approach. Even when the activity of organic neurons would be perfectly emulated, it would be a mere presumption that as of this moment an artificial being has a free will. The problem with the free will is, as already mentioned, that this quality is dubious due to its first person perspective experience and cannot yet be even confirmed in animals. Further research in this field is necessary.

4 Conclusion

The aim of this paper was to frame the problem of free will in the context of AGI. In order to answer the question whether it is necessary to artificially construct free will to reach the ultimate goal of AGI two main problems were explored: (1) the nature and importance of free will for human mental functioning, (2) usefulness and technical possibility of its creation and incorporation into AGI.

It has been shown that free will as such significantly influences mental processing and overall behavior of a human. Characteristics associated with free will are considered to be uniquely human and contributing to development of intelligence.

In the field of AGI incorporation of such an element is presumed to bring significant improvement for agents situated in complex environments. Although there are many limitations and constraints yet to be solved, the possibility of creating free will seems to be viable and in case of continuous risk assessments also beneficial to the society.

The ultimate goal of AGI is to create a system that resembles or exceeds human capabilities in all areas including cognition and emotions. Since free will contributes to intelligence development, emotional control and possibly also self-awareness, and it seems to be construable, AGI needs to create this element to resemble human capabilities. Future attempts not only need to include real random generator that will be incorporated into the decision mechanism but also learn from neuroscience and get inspiration from mechanical functioning of the brain.

Last remark we wish to make concerns constructing a system that exceeds human capabilities. It needs to be noted that "exceeding human capabilities" is a very vague term. Since AGI aims at first to resemble a human, free will seems to be necessary. However, this will may also enable an AGI system to experience dilemmas, contradictions and human states in which it is sometimes difficult to make any decision. It is questionable which role free will plays in this drama. It can be at the same time the cause of all these problems as well as their solution.

References

1. Goertzel, B., Pennachin, C., eds.: Contemporary Approaches to Artificial General Intelligence. In: Artificial General Intelligence. Springer (2007)
2. Neisser, U., Boodoo, G., Bouchard Jr, T.J., Boykin, A.W., Brody, N., Ceci, S.J., Halpern, D.F., Loehlin, J.C., Perloff, R., Sternberg, R.J., et al.: Intelligence: Knowns and unknowns. American psychologist **51**(2) (1996) 77
3. Volition. `http://www.merriam-webster.com/dictionary/volition`
4. Kazdin, A.E., ed.: Volition. In: Encyclopedia of Psychology. Oxford University Press (2000)
5. Free will. `http://www.merriam-webster.com/dictionary/free%20will`
6. O'Connor, T.: Free will. In Zalta, E.N., ed.: The Stanford Encyclopedia of Philosophy. Spring 2013 edn. (2013)
7. Frith, C.: The psychology of volition. Experimental Brain Research **229**(3) (2013) 289–299
8. Haggard, P.: Human volition: towards a neuroscience of will. Nature Reviews Neuroscience **9**(12) (2008) 934–946
9. Baumeister, R.F., Muraven, M., Tice, D.M.: Ego depletion: A resource model of volition, self-regulation, and controlled processing. Social Cognition **18**(2) (2000) 130–150
10. Vohs, K.D., Schooler, J.W.: The value of believing in free will. encouraging a belief in determinism increases cheating. Psychological science **19**(1) (2008) 49–54

11. Ryan, R.M., Deci, E.L.: Self-regulation and the problem of human autonomy: Does psychology need choice, self-determination, and will? Journal of personality **74**(6) (2006) 1557–1586
12. Hong, F.T.: On microscopic irreversibility and non-deterministic chaos: Resolving the conflict between determinism and free will. In: Integral Biomathics. Springer (2012) 227–243
13. Perlovsky, L.: Free will and advances in cognitive science. ArXiv e-prints (2010)
14. Fisher, M.: A note on free will and artificial intelligence. Philosophia **13**(1-2) (1983) 75–80
15. McCarthy, J.: Mathematical logic in artificial intelligence. Daedalus **117**(1) (1988) 297–311
16. McCarthy, J.: Simple deterministic free will. `http://www-formal.stanford.edu/jmc/freewill2/`
17. Corno, L.: The best-laid plans modern conceptions of volition and educational research. Educational researcher **22**(2) (1993) 14–22
18. Donald, M.J.: Neural unpredictability, the interpretation of quantum theory, and the mind-body problem. arXiv preprint quant-ph/0208033 (2002)
19. Urenda, J.C., Kosheleva, O.: How to reconcile physical theories with the idea of free will: from analysis of a simple model to interval and fuzzy approaches. In: Fuzzy Systems, 2008. FUZZ-IEEE 2008.(IEEE World Congress on Computational Intelligence). IEEE International Conference, IEEE (2008) 1024–1029
20. Manzotti, R.: Machine free will: Is free will a necessary ingredient of machine consciousness? In: From Brains to Systems. Springer (2011) 181–191
21. McCarthy, J., Hayes, P.: Some philosophical problems from the standpoint of artificial intelligence. Stanford University (1968)
22. Kanter, I., Aviad, Y., Reidler, I., Cohen, E., Rosenbluh, M.: An optical ultrafast random bit generator. Nature Photonics **4**(1) (2009) 58–61
23. Calude, C.S., Dinneen, M.J., Dumitrescu, M., Svozil, K.: Experimental evidence of quantum randomness incomputability. Physical Review A **82** (2010) 022102
24. Tse, P.U.: The Neural Basis of Free Will. Criterial Causation. The MIT Press (2013)