Computational machines, free will and human reason

Gonzalo Génova
Departamento de Informática
Universidad Carlos III de Madrid, Spain
ggenova@inf.uc3m.es

Ignacio Quintanilla Navarro
Departamento de Teoría e Historia de la Educación
Universidad Complutense de Madrid, Spain
ignacioq@ucm.es

***Abstract*—David Hume, the Scottish philosopher, conceives reason as the slave of the passions, which implies that human reason has predetermined objectives it cannot question. An essential element of an algorithm running on a computational machine (or Logical Computing Machine, as Alan Turing calls it) is its having a predetermined purpose: an algorithm cannot question its purpose, because it would cease to be an algorithm. Therefore, if self-determination is essential to human intelligence, then human beings are neither Humean robots, nor computational machines.**

Keywords—human nature; free will; self-determination; algorithm; computational machine; goal and strategy selection

I. Introduction

In this paper we want to show the connection between Hume's conception of human nature and the modern conception of robots. Even if, quite possibly, the concept of 'robot' would have proved deeply strange to Hume, the truth is that his conception of reason as 'the slave of the passions' anticipated the modern concept of computing machine: we call his conception a *Humean robot*, that is, an instrumental intelligence at the service of predetermined objectives, or passions. In fact, if for us humans of the 21st century, it is tempting to consider ourselves complicated biological robots, it is only because we have previously accepted the Humean paradigm of reason as the slave of the passions. We are prone to believe that we are robots, because we have first accepted that reason neither chooses nor prioritizes its ends.

This paper is a summary of the one published in the Journal of Experimental & Theoretical Artificial Intelligence (Jan 2018) with the title *Are Human Beings Humean Robots?* [2].

II. David Hume: reason is the slave of the passions

David Hume (1711-1776) wrote in *A Treatise of Human Nature*, under the section devoted to the influencing motives of the will, that "reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them" [6]. Hume wanted to understand the human mind as Isaac Newton had understood the cosmos, by adopting a mechanistic approach to human intelligence. Human beings are attracted by passions, and moving towards a concrete passion can be resisted only with aid of a stronger and opposite passion, much in the same way as physical forces operate on bodies. In this conception of human nature, *the role of reason is to elaborate a strategy* to best fulfill the set of passions; but reason neither questions nor chooses the passions it has to serve. We think Hume proposes a suggestive account of instrumental reason that anticipates and prepares a modern algorithmic model of intelligence, aimed at optimizing the achievement of its predetermined objectives (i.e. passions).

In Hume's sentimentalist approach to ethics, happiness is achieved when passions are satisfied. In this model, *reason is understood primarily as an optimization tool* (technical or instrumental reason, therefore), used to calculate the behavior that better satisfies the passions involved and that demands less effort from the subject. *Reason is the slave of the passions*: it does not question those passions that are irresistibly imposed upon it, nor their objectives, nor the strength of their attracting force; passions and objectives are pre-rational or meta-rational.

If Reason is integrated into the realm of passions, or goals, as an algorithmic calculation, Will cannot be anything but automatic: once the optimal path is known, all that is left is to start out, give the order, but not properly 'decide'. What is implied here is *a radical denial of human freedom* in the usual sense of the term, to which we will refer later.

III. Alan Turing: what is a computational machine

A robot is usually defined as a mechanical device that is controlled by a computer running a program; a robot is, in this sense, an algorithmic or computational machine. An algorithm can be preliminary defined as a rule-based procedure that obtains a desired result in a finite number of steps. Alan Turing laid the foundations of the modern notion of algorithm, establishing that a computation method is *effective* (a.k.a. mechanical) if it can be carried out by a Turing Machine [8], or, as Turing himself calls it, a Logical Computing Machine [9]. This is the substance of the Church-Turing thesis [1].

However, and perhaps surprisingly, there is a lack of satisfactory consensus on the definition of algorithm [10]. A recent study by Hill [4] examines existing approaches to the notion of algorithm, from semi-formal definitions like the one by Donald Knuth, "an algorithm is a finite set of rules that gives a sequence of operations for solving a specific type of problem" [7], to more formal ones.

After some analysis, Hill offers a definition: "An algorithm is a finite, abstract, effective, compound control structure, imperatively given, *accomplishing a given purpose* under given provisions." The italics manifest the *intentionality* of algorithms. This sense of utility or purposefulness is shared by machines in general. It is the success or failure in accomplishing its function that permits us to tell whether the machine works properly or not. Thus, *a machine cannot be defined and accounted for without reference to its purpose* [3].

Take for example a game playing machine, designed to play against a human. Initially, the machine has the goal to win the game. If the game is very simple (like Tic-Tac-Toe), designing a strategy (an algorithm) to win, or at least not to lose, is rather easy. In the case of chess, the complexity of the game has not permitted, until now, an infallible strategy, even though, with current technology, most of human players will lose against a rather common artificial chess player.

A somewhat different kind of chess machine might include a certain degree of randomness in its 'decisions', or it might be able to self-limit the effectiveness of its strategy in order to configure an affordable level of difficulty, so that the human player still enjoys the game and does not throw in the towel too soon. These two kinds of chess machines have slightly different objectives: either winning the game, or else having the human player learn how to play better and enjoy the learning process. Nevertheless, in each case the machine has a well determined purpose or function that defines it. What we do not expect from a chess machine of the first kind (i.e. designed to win) is that it chooses to lose the game… *It can fail to achieve its goal, but it cannot change its goal*. Of course, there can be algorithms with different levels of goal selection and prioritization. However, those *dynamic goal-selection algorithms* are in fact obeying higher-order goals (meta-goals) to select convenient sub-goals and strategies.

We think Turing himself acknowledged this *lack of freedom* was essential in his conception of a computational machine, even if implemented by humans performing calculations: "A man provided with paper, pencil, and rubber, and *subject to strict discipline*, is in effect a universal machine" [9] (our italics). Notably, it happened exactly in this way in the internal organization of Bletchley Park labor groups set up by Turing and others to decipher German codes during World War Two [5]. Being 'subject to strict discipline' means not questioning at all the rules and purposes of the procedure, i.e. the computation.

IV. Determination, indetermination, self-determination

Mechanism in philosophy is the view that all beings, whether lifeless or alive, are like complicated machines. Mechanism is closely linked to determinism, since the scientific and technological revolution of the 17th century made some philosophers –Hume among them– believe that all phenomena could eventually be explained in terms of 'mechanical laws', i.e. natural laws governing the motion and collision of matter under the influence of physical forces. Modern mechanistic views of living beings, including humans, comprise mechanical *information processing* as an essential element of the 'living machine', for example in behaviorist stimulus-response theories. We distinguish three ways of relationship between mechanistic determination and the behavior of humans and computational machines.

1. **Hetero-determination.** The behavior is fully determined by the received stimuli and the computational or neurological processing these stimuli undergo to produce a response, according to more or less complex programs and evaluation systems.
2. **Indetermination.** This view complements the previous one by adding a certain degree of uncertainty. However, indeterminism does not add anything essentially different to the Humean conception of human nature. In fact, these two views, hetero-determination and indetermination, agree in their radical negation of human freedom.
3. **Self-determination.** In this position the previous two are rejected. If human freedom, in its usual sense, is not an illusion, then it is not true that human behavior is determined (even only statistically) just by the material aspects of the body and the phenomena that occur in it. On the contrary, being truly free means that human beings self-determine in their actions.

V. Summary of the argument

David Hume conceives reason as the slave of the passions, which implies that human reason has predetermined objectives it cannot question. On the other hand, an essential element of an algorithm running on a computational machine is its predefined purpose: an algorithm cannot question its purpose, because it would cease to be an algorithm. We have reached a critical point for the Humean-computational view of human beings, since self-determination is not an algorithmically programmable function: *purpose is the prerequisite of an algorithm, not its result*. Therefore, if self-determination is the true essence of human freedom, then human beings are neither Humean robots, nor algorithmic machines.

We have *not* demonstrated that human beings are truly free (self-determined). We have only demonstrated that *if* humans are free, *then* they cannot be algorithmic machines; *then* human intelligence cannot be properly defined as an algorithmic process; and *then* human behavior cannot be perfectly emulated by algorithmic robots. Whether we, in some uncertain future, can produce in our laboratories a kind of non-algorithmic robots that can be properly called free, and whether they still can be called 'robots', will be the subject of further research.

References

[1] Copeland, B.J. (2002). The Church-Turing Thesis. *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition), Edward N. Zalta (ed.). http://plato.stanford.edu/archives/sum2015/entries/church-turing.

[2] Génova, G., Quintanilla Navarro, I. Are Human Beings Humean Robots? *J. Exp. & Theor. Art. Int.* 30(1):177–186, Jan 2018.

[3] Génova, G., Quintanilla Navarro, I. Discovering the principle of finality in computational machines. *Found. of Science*. Online 13 Feb 2018.

[4] Hill, R.K. (2015). What an algorithm is. *Phil. & Techn.* 29(1), 35–59.

[5] Hinsley, F. H., Stripp, A., eds. (1993). *Codebreakers: The inside story of Bletchley Park*. Oxford: Oxford University Press.

[6] Hume. D. (1739). *A Treatise of Human Nature*, II-iii-3. London: John Noon. Text available online at http://www.gutenberg.org/ebooks/4705 and http://www.davidhume.org/texts/thn.html.

[7] Knuth, D.E. (1997). *The art of computer programming (vol. 1): fundamental algorithms*. Redwood City: Addison Wesley.

[8] Turing, A. (1936). On computable numbers, with an application to the Entscheidungsproblem, *Proceedings of the London Mathematical Society* 2(42):230–265.

[9] Turing, A. (1948). Intelligent Machinery. *National Physical Laboratory Report*. (Facsimile at http://www.AlanTuring.net/intelligent_machinery.)

[10] Vardi, M. (2012). What is an algorithm? *Comm. of the ACM* 55(3):5–5.