Gonzalo Génova & Ignacio Quintanilla Navarro (2018) Are human beings humean robots?, Journal of Experimental & Theoretical Artificial Intelligence, 30:1, 177-186

DOI:10.1080/0952813X.2017.1409279

Journal of Experimental & Theoretical Artificial Intelligence 30(1):177–186, January 2018
(Published online 01 Dec 2017)
The final publication is available at Taylor & Francis via:
http://dx.doi.org/10.1080/0952813X.2017.1409279

Are Human Beings Humean Robots?

Gonzalo Génova
Departamento de Informática, Universidad Carlos III de Madrid.
Avda. Universidad 30, 28911 Leganés (Madrid), Spain
ggenova@inf.uc3m.es

Ignacio Quintanilla Navarro
Departamento de Teoría e Historia de la Educación, Universidad Complutense de Madrid
Avda. Rector Royo Villanova s/n, 28040 Madrid, Spain
ignacioq@ucm.es

We live in an age in which we are proud of thinking machines,
but distrustful of people that try to think.
Juan Génova, in memoriam

Keywords: human nature; free will; self-determination; algorithm; computational machine; Turing Test.

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 beings, nor computational machines. We examine also some objections to the Turing Test as a model to understand human intelligence.

1. Introduction

Throughout history, human beings have always reflected on themselves, initially comparing with something not human, e.g. divinity, animals or nature. The instance that today fulfills that function is mainly artificial machines and, in particular, artificial intelligent machines: robots are the most human of our machines. Artificial intelligence plays thus a key role in our current understanding of ourselves. We assume, then, that asking about the nature and limits of the intelligence we produce today is an appropriate framework to investigate the essence of the human condition.

In this paper we want to show the connection between Hume's conception of human nature and the modern conception of robots as they are in reality (robots in science fiction pose different problems, which rather reinforce our thesis). 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. 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 18th century 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 does not necessarily mean that the ends of human behavior are completely fixed; it only means that their selection and prioritization is not *rational*.)

Our purpose here is to challenge the suitability of a research program inspired in Hume's conception of human nature, and we think the current development of artificial intelligence supports our thesis.

2. 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" (Hume, 1739). Hume's philosophical plan was to apply to the science of man the method of 'experimental philosophy' (i.e. natural philosophy, or physics), and to extend to philosophy in general the fruitful methodological self-limitations of Newtonian physics (Copleston, 1999). 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 the aid of a stronger and opposite passion, much in the same way as physical forces operate on bodies. Passions, in this sense, must be understood as drives or impulses, leaving aside the category of emotions. 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*. Apart from the precise historical meaning of the assertion that "reason is the slave of the passions", we think Hume proposes a suggestive account of instrumental reason that anticipates and prepares a modern algorithmic model of intelligence.

Indeed, if we understand ourselves as machines, it is almost inevitable –in light of the current state of the art in artificial intelligence– to identify Humean passions or drives with predetermined objectives or final states in our 'natural programming', and Humean reason with calculations aimed at optimizing the achievement of these objectives.

As we will argue later on, we think this is a very restrictive and debatable notion of human intelligence. However, it seems interesting to initially accept Hume's suggestion in order to explore some current approaches to the relationship between algorithms and human minds. If we assume, as Hume himself does, that human passions and desires admit a quasi-algorithmic or mechanical formalization –by a homeostatic model, for instance– we can enable a common theoretical model to explain the means and ends of all observable human behavior.

In Hume's sentimentalist approach to ethics, happiness is achieved when passions are satisfied. Passions are attractive forces which, in analogy with Newtonian forces, can be resisted only if there is another passion attracting in the opposite direction. Since there are multiple passions (pride and humility, love and hatred, etc.), and they point in different directions, the resulting behavior is a balance of forces that tends to preserve inner peace and tranquility (*homeostasis*). 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 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 their attracting force; passions and objectives are pre-rational or meta-rational.

When there are several opposing passions in contention, and it is not possible to satisfy them all (for example, the wish to steal something and the fear to be imprisoned), it is nevertheless possible to define a multi-objective function, a kind of weighted average satisfaction of all their objectives (or some other kind of aggregate function, not necessarily weighted average). This function (to be maximized) is capable to unify in a single objective the plurality of attractions, and thus is able to bring order to the individual passions. The key point is that the form of this function (its weights, so to speak) is also pre-rational or meta-rational. As a consequence of its being the slave of the passions, *reason does not impose any hierarchy among objectives*. The multi-objective function to be maximized, which expresses the order and hierarchy of passions, is also imposed upon reason.

In this scheme, where Reason is integrated into the realm of passions as an algorithmic calculation, Will is reduced to a sort of psychological motivation as well. What we usually call '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.

3. Human intelligence, robotic intelligence

Since the 17th century, Western culture has developed an epistemological program where we can truly understand only what we are able to replicate or produce, even though in ideal conditions. Therefore, understanding human *natural* intelligence requires, or at least is improved by, producing first *artificial* intelligence. However, some philosophical caveats must be put forward before we proceed with our argumentation:

1. Artificial does not necessarily mean non-organic.

2. Intelligence is not necessarily a quality exclusive of human beings; moreover, perhaps human intelligence is not the archetype of intelligence.

3. We cannot assume that intelligence is the key defining element of the human condition, even from a cognitive perspective.

4. We know and understand artificial intelligence a lot better than human natural intelligence, because we have produced the former, whilst the latter has been given to us.

5. Research in artificial intelligence encompasses more aspects than performing algorithms in a computational machine, such as: having emotions, perceiving the world as a totality, having awareness of oneself, having personal consciousness, having one's own desires, having the capacity of choosing between good and evil, and so on.

6. We do not know exactly what it means being intelligent, not even in the restricted human sense; therefore we do not know whether this sort of intelligence can be properly expressed in algorithmic terms.

7. There is not an undisputed definition of algorithm.

Even being conscious of these difficulties, it seems undeniable that the possibility of an inorganic artefact to emulate all definable mental processes that can be observed in a human being possesses a great theoretical importance, both for our notion of humanity and for the development of artificial intelligence.

4. 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. In this paper we are not concerned with the physical aspect of the robot, i.e. whether it resembles or not a human being (male android or female gynoid); nor with the fact that the robot can be made of inorganic materials, organic materials, or mixed. We are solely concerned with the fact that a robot is controlled by an algorithm, or set of algorithms; 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 (Turing, 1936), or, as Turing himself calls it, a Logical Computing Machine (Turing, 1948). This is the substance of the Church-Turing thesis (Copeland, 2002).

However, and perhaps surprisingly, there is a lack of satisfactory consensus on the definition of algorithm (Vardi, 2012). A recent study by Hill (Hill, 2015) 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" (Knuth, 1997), to more formal ones.

After some analysis, Hill offers a preliminary definition: "An algorithm is a finite, abstract, effective, compound control structure, imperatively given." The author argues that this definition is incomplete because it does not account for *what the algorithm actually does*: "there is more to an algorithm than its procedure". Therefore, the definition is further refined as follows (her addition in italics): "An algorithm is a finite, abstract, effective, compound control structure, imperatively given, *accomplishing a given purpose under given provisions*." The reason she gives for this addition –with which we fully agree– is the necessity to manifest the *intentionality* of algorithms. An algorithm is not simply something that happens, but something that happens *with a purpose*, i.e. it does something *for somebody*.

This sense of utility or purposefulness is shared by machines in general, in contrast with other kinds of artefacts that may not have a useful purpose, like works of art. As it has been extensively dealt with in the philosophy of technology (Kroes, 2010), a machine has a dual nature that encompasses both its physical *structure* and the *function* it has to accomplish. It is the success or failure in accomplishing its function that permits us to

tell whether the machine works properly or not. Therefore, *a machine cannot be defined and accounted for without reference to its purpose*.

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 predetermined 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 different levels of goal selection. There are in fact algorithms that can dynamically change their goals, prioritize them, etc. So they are able to perform some kind of meta-reasoning in relation to the goals to be achieved. However, those dynamic goal-selection algorithms in turn do not analyze themselves and change their goals. They are in fact obeying higher-order goals (meta-goals) to select convenient sub-goals. They cannot decide to stop behaving as goal-selection algorithms. Therefore, this objection does not affect our argument.

Summing up, an essential element of an algorithm running on a computational machine is its predetermined purpose: *an algorithm cannot question its purpose, because it would cease to be an algorithm*. Thus, in a certain sense, an algorithmic robot is the intelligent slave of its purpose… it is a *Humean robot*. Conversely, if human reason can question the passions (i.e. goals and meta-goals) it is supposed to serve, then human reason is not the slave of those passions, that is, human reason is not algorithmic.

We think Turing himself acknowledged that *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" (Turing, 1948; our italics). Notably, it happened exactly in this way in the internal organization of Bletchley Park labor groups set up by Turing to decipher German codes during World War Two (Hinsley & Stripp, 1993). Being 'subject to strict discipline' means not questioning at all the *rules and purposes* of the procedure, i.e. the computation.

5. 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.

Computers are a special kind of machines, namely algorithmic machines, where this information processing aspect lies at the core; therefore we want to explore the role of mechanistic determination in our model of intelligence, whether human or robotic.

We distinguish three –or rather four– ways of relationship between determinism and the behavior of humans and computational machines.

1. **Hetero-determination.** The behavior is fully determined by the received stimuli and by the computational or neurological processing these stimuli undergo to produce a response, according to more or less complex programs and evaluation systems. In the case of computers, the programmer wrote these programs and evaluation systems. In the case of humans, they are grounded on genetics, or have been acquired from education and influence of the environment. This paradigm is a natural evolution of the Humean conception of human nature, where humans are no more than complex biological robots: *we are the obedient slaves of our passions, our biology, our education or cultural inheritance*. The paradoxes of deterministic behavior have been illustrated long ago with the Burindan's Ass dilemma, based on the writings of the French philosopher Jean Buridan (c. 1295–1363), and more recently in science fiction stories such as *Runaround* (1942) by Isaac Asimov. The problem received a mathematical foundation from computer scientist Leslie Lamport, who calls it the Buridan's Principle (Lamport, 2012): "A discrete decision based upon an input having a continuous range of values cannot be made within a bounded length of time". In the paper, originally written in 1984 though not published until 2012, the author also stresses the consequences of ignoring the principle when designing engineering devices.

2. **Indetermination.** This view complements the previous one by adding a certain degree of uncertainty due to physical causes, either in the evaluative subsystem (decision by a factor of randomness) or in the executive subsystem (which in fact means the physical system does not behave exactly as commanded). The lack of determinism in the outcome of an algorithm is no surprise for computer scientists, who have been using randomness in algorithms for decades (e.g. genetic algorithms and other biologically inspired computing techniques). It pragmatically solves Buridan's Ass dilemma and makes behavior relatively unpredictable (though statistically predictable). However, indeterminism does not add anything essentially different to the Humean conception of human nature. In fact, these two views, hetero-determinism and indeterminism, agree in their radical negation of human freedom, which we can observe also in the interpretation of the neuroscientific experiments by Benjamin Libet and others (Libet et al., 1983; Soon et al., 2008): *freedom is an illusion*, since voluntary acts are unconsciously initiated in the brain before the subject is aware of them. Even if there are philosophers (Hume included, in his 1748 *Enquiry*, though not

in this 1739 *Treatise*) that consider freedom is compatible with hetero-determinism, we think the usual sense of freedom is precisely what their interpretation of experimental results expose: if behavior is hetero-determined or indetermined, then freedom is an illusion. Modern authors like Daniel Dennett (Dennett, 1991) also hold a compatibilist approach, but in our view their arguments in fact confirm that they think freedom is not something real that can influence on human behavior, but an illusion produced by the brain, be it deterministic or indeterministic.

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 hetero-determined (even only statistically) just by the material body and the phenomena that occur in it. On the contrary, being truly free means that human beings self-determine in their actions. Self-determination admits two versions:

    a. **Self-determination *towards* the ends.** In this version, human beings are free to pursue a certain final end, and to choose among different behaviors in order to achieve it. But the final end, as such, is given. This entails a rather modest affirmation of freedom, consisting only in the (maybe computable) choice among various means to reach a given end, together with the more substantial free option to pursue that end, or not.

    b. **Self-determination *of* the ends.** In this more radical version, the human being not only self-determines to an end, but self-determines the ends: ends are not given. It is claimed that humans not only *have* a destiny (even less a tragic fate), but they *forge* their own destiny. A human being not only chooses *how* to become something, but *what* he or she wants to become. And this is precisely what makes it so difficult to make certain choices (Chang, 2013).

Self-determination poses two difficult problems we do not intend to solve here. First, the metaphysical mind-body problem, i.e. the interplay between the immaterial and the material (in any case, we do not think Cartesian dualism is a valid solution). Second, the moral problem of arbitrariness in the self-determined choice of ends: Does it matter whether one chooses this or that end for one's life? Are certain ends recognizably better than others?

Leaving apart these two problems, we have reached a critical point for the Humean-computational view of human beings, since *self-determination is not an algorithmically programmable function* (an algorithm cannot self-determine its purpose). In other words, free behavior is not computable. Therefore, if self-determination is the true essence of human freedom, then human beings are not Humean robots.

6. The objections of panpsychism and panfreedom

Two philosophical stances have been revived in recent times, which could seem to refute our argument. *Panpsychism* states that everything in the universe is conscious in some way or other; in other words, consciousness is a fundamental feature of the Universe, as David Chalmers puts it in *The Conscious Mind* (Chalmers, 1996).

Panfreedom is a version of panpsychism, where it is stated that everything in the universe has a certain degree of freedom, even the elementary particles of physics. This is the alleged result of the *Free Will Theorem* by Conway and Kochen (Conway & Kochen, 2006): under certain assumptions, if human beings have a free will, in the sense that their behavior is not a computable function of the past, then so must elementary particles. Apparently, one could conclude that, if elementary particles are free, then very likely so are computers; or that, if everything is conscious, then a computer is conscious, though in a different way as a human being.

Indeed, these two stances are controversial themselves, with their defenders and opponents, like everything that you can imagine that can be said about freedom and human nature (including our own position). For the sake of concision of our argument, we do not intend to refute them. Instead, we are satisfied with the demonstration that, even if panpsychism or panfreedom were true properties of the universe, our argument would be still valid.

It is as follows. The essence of an algorithmic computer (recall Turing's definition of effective or mechanical computation method) is to carry out a computation, i.e. to obtain a desired result, to reach a predefined goal. If the computation fails to achieve its goal, it is not because *it is* a computation, but because *it is not*. In other words, it is a bad implementation of the ideal computation, or else it is a computation that is not capable to overcome the resistance of hardware to obey software. If the computer does not compute as expected because it cannot harness the disobedient indeterminism of elementary particles (be it due to their true random indeterminism, or to their alleged uncontrollable free will), then it is not a true computer, it is a failure.

Therefore, even if we assumed that elementary particles are free, we should not conclude that computers are, because the whole purpose of a computer is to neutralize and control the free (or at least indeterministic) behavior of its components, in order to achieve its goal. In fact, we can argue in a similar way when the computer components are commonly recognized as free beings. Remember the computer made of humans that worked in Bletchley Park, running programs designed by Turing and others, as quoted above: "A man provided with paper, pencil, and rubber, and *subject to strict discipline*, is in effect a universal machine" (Turing, 1948). Bletchley Park machinery worked as a computer only because individual initiative was completely suppressed. If it had not been so, if a member of the staff had failed (or had not wanted) to accomplish its subtask, then Bletchley Park would have not operated like an algorithmic computer any longer.

7. The Turing Test revisited

Alan Turing proposed in 1950 his famous test to determine whether a machine can or cannot think, or rather, as a means to define what a thinking machine could be (Turing, 1950):

> *I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms 'machine' and 'think.'*

The test is conceived as a methodical procedure, an 'experiment' to tell in a verifiable way whether or not a machine can think. The test does not require that the role of the

interrogator be performed by a human, a group of humans, or even a machine. Indeed, a CAPTCHA (acronym for Completely Automated Public Turing test to tell Computers and Humans Apart) is a Turing Test performed by a machine. The essence of the test is its being a methodical, repeatable procedure, independent of who carries it out. In this sense, it is automatic, algorithmic. However, we know that we cannot algorithmically distinguish whether a sequence of events (i.e. the behavior of the subject under examination) has or has not a purpose. Gregory Chaitin demonstrated (Chaitin, 2005), as a derivation to Turing's Halting Problem, that there is no algorithm that can unequivocally tell whether a given sequence of numbers is deterministic or random (i.e. with or without a purpose).

Let's suppose a human being (assuming truly free, i.e. self-determined) is subjected to the Turing Test. What will happen if that human being self proposes to mimicry a machine and deceive the interrogator? How can the interrogator (human or mechanical, but in any case methodical) defend itself from deception? We do not know what a radically free being subjected to the test will do. Since it has not a predefined objective, it can decide to fail the test. And if it decides to strictly follow the interrogator's instructions, then it is not a true representative of humanity from that moment on. Even if there is not a clear will to deceive, if the imitator receives instructions to 'behave like a human', *what does it mean, to behave like a human?*

On the other side, a machine that is built to pass the test, i.e. that perfectly mimicries human behavior (the imitation game), has the objective to imitate a being… who has no *apriori* objective! How can that be? *What can possibly be the specification for doing an unspecified task?* Of course, we can build a machine that mimicries the *typical* human behavior: somewhat erratic, somewhat purposeful, and so on. This has been already achieved to a certain degree with current technology. But the point here is that *we cannot methodically (algorithmically) distinguish between purposeful and purposeless behavior*, even less between beings that behave according to given instructions and beings that behave according to instructions they choose themselves (i.e. with their own purposes). The typical behavior (i.e. the average behavior) is not necessarily the behavior of an individual. *Free (self-determined) behavior is not computable.*

In other words, we can build a machine (the imitator) that behaves like typical humans, and we can build a machine (the tester) that distinguishes between typical and weird human behavior. But this leaves out the essential question of whether that being is or is not truly free, in the sense that it self proposes its own objectives. We can build robots that disobey their human owners (Briggs & Scheutz, 2015), but in doing that, those robots do nothing more than obeying their human programmers. *Creativity, i.e. the capacity to create new projects and go beyond predetermined objectives, is not a verifiable property.* If human beings are, as Hume wanted, 'the slaves of the passions', then there is no difficulty in principle to build an imitator machine, because humans are machines in the end; and also there is no difficulty to build a tester machine, once those passions are known (the test specification would be: 'human' is that being that pursues those known passions). But if self-determination is a genuine property of humans, then the road towards humanness is closed for computational machines. In this sense, our position is a double objection (from both the imitator and tester viewpoints) to the Turing Test as a model to understand human intelligence, because it leaves out the non-computational aspects of freedom and rational choice (of course, if human freedom is denied, then our objection is void). Our objection, however, does not diminish the

importance of the Turing Test as a research program for the development of AI or as a criterion for defining what would be artificial thinking.

8. 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, we have also established that, by construction, an essential element of an algorithm running on a computational machine (or Logical Computing Machine, as Alan Turing calls it) is its predetermined purpose: an algorithm cannot question its purpose, because it would cease to be an algorithm.

Let H denote the set of 'Humean beings', that is, beings whose behavior obeys predetermined objectives. And let A denote the set of 'Algorithmic beings', that is, beings whose behavior is algorithmically computed. Then A is a subset of H, since having a predefined objective is an essential property of the definition of A; therefore, being A implies being H, and not being H implies not being A.

$$A \Rightarrow H$$
$$\neg H \Rightarrow \neg A$$

The reverse is not true. There may exist a 'goal-driven entity' (H) whose behavior is not algorithmic or mechanical (in Turing's precise sense of 'mechanical'), but that is still physically determined (Copeland, 2002).

Let S denote the set of 'Self-determined beings', that is, beings that can freely choose to pursue or not a given end, or even can self-determine the ends they want to pursue. Then, it is clear that being S implies not being H, and therefore not being A.

$$S \Rightarrow \neg H$$
$$\neg H \Rightarrow \neg A$$
$$\therefore S \Rightarrow \neg A$$

Conversely, being A implies not being S.

$$A \Rightarrow H$$
$$H \Rightarrow \neg S$$
$$\therefore A \Rightarrow \neg S$$

Summing up, if self-determination is essential to human intelligence, then human beings are neither Humean beings, nor algorithmic machines.

It could happen, as Hume wanted, that humans always act in accordance with drives, wishes and desires they cannot rationally challenge; this could be easily imitated by machines. It is also conceivable a physical system that is somehow flexible enough to self-determine its own goals: we humans are indeed physical systems, and we are capable (at least in the opinion of many, who do not share the Humean view of human nature) of self-determination. Nonetheless, such physical systems, even if they were produced by humans, lie out of the computational paradigm as was classically defined by Turing and others: *they are physical systems, but they are not, properly speaking,*

algorithmic machines that obey a program (note that, as we already explained, algorithmic goal selection is still algorithmic, even if at a higher level). An entity capable to self-determine its own goals at the highest level, well, it is not algorithmic.

At this point we need to recapitulate our assumptions (i.e. those statements we take for granted without need to demonstrate or at least argue in favor of them), and very specially our non-assumptions, and distinguish both of them from our conclusions. In particular: we *do not* assume that machines cannot be free; and we *do not* assume that human beings are free. Instead: we *conclude* that machines cannot be free; and, consequently, we argue that, *if* humans are free, *then* humans are not machines.

On the other side, what we really assume are the definitions of computer and freedom. We define 'computer' as an algorithmic or computational machine (not a very strange definition, indeed, well rooted in the tradition of computer science started by Turing). An algorithmic machine has a predefined objective it has been designed to accomplish; we do not invent this property. The definition of a computer as an algorithmic machine includes, as an essential element and design principle, the fact that it has a predefined objective. Therefore (by the strength of the definition of the terms) a computer cannot change its objective, because then it will be no more a computer, i.e. something with a predefined objective. It would simply be a contradiction: no more, and no less.

There are many ways to define 'freedom'. We have defined it (also following well established traditions) as self-determination, i.e. the ability to choose one's own objectives, or to put hierarchy among given objectives.

We certainly claim that computational machines cannot possess free will. But this is not an *assumption*: it is a *consequence* of the definitions of computational machine and freedom, and the whole point of this work.

However, note that we *do not* claim that it is impossible to 'make', 'build' or whatever you want to call it, an artificial free being, be it organic, inorganic (maybe electronic), or mixed. We only claim that it won't be a robot, a computer, an algorithmic machine. This is not only an issue for 'current' computers. It is an issue for computers (i.e. computational machines with predefined objectives) of all ages. If an artificial being comes to be free (i.e. self-determined), it will not be because of its being a computer, but because of its being more than that.

If, sometime in the future, we get to make machines without predefined objectives, our argument will not apply, obviously. However, we still think that 'a machine without a predefined objective' is a contradictory concept: we should invent a new name for it. In a sense, human reproduction already produces beings without predefined objectives.

In some way, we admit that we could build artificial beings capable to alter their passions, their goals; instead, we do not admit that those artificial beings could be properly named 'computational machines'. If a washing machine can decide to stop washing, then it is not any more a washing machine. If a robot can select its own goals, then it is not any more a robot, even if it can still be an electronic device. Indeed, an electronic device, a physical being, but not a computer, not a computational machine.

We hope we have shed light on the relationship between computational machines, human reason and free will. We have assumed that true freedom, in its usual sense, requires self-determination, at least in its weaker form of 'self-determination *towards* the ends' (even if we prefer the stronger version of self-determination *of* the ends).

However, 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. Therefore human intelligence, so understood, cannot be properly defined as an algorithmic process, and human behavior cannot be 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

Briggs, G., Scheutz, M. (2015). *"Sorry, I can't do that:" Developing mechanisms to appropriately reject directives in human-robot interactions*. Proceedings of the 2015 AAAI Fall Symposium on AI and HRI. Available at http://www.aaai.org/ocs/index.php/FSS/FSS15/paper/download/11709/11522.

Chaitin, G. (2005). *Meta Math! The Quest for Omega*. New York: Vintage Books. ISBN 978-1-40-007797-7.

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press. ISBN 0-19-511789-1.

Chang, R. (2013). *Grounding practical normativity: going hybrid*. Philosophical Studies 164(1): 163-187. DOI: 10.1007/s11098-013-0092-z.

Conway, J., Kochen, S. (2006). *The Free Will Theorem*. Foundations of Physics 36(10): 1441–1473. DOI: 10.1007/s10701-006-9068-6.

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

Copleston, F. (1999). *A History of Philosophy*, vol. 6, pp. 405–406. London: A&C Black. ISBN 0-385-47043-6.

Dennett, D.C. (1991). *Consciousness Explained*. London: Penguin Books. ISBN 0-316-18065-3.

Hill, R.K. (2015). *What an algorithm is*. Philosophy & Technology 29(1): 35–59. DOI: 10.1007/s13347-014-0184-5.

Hinsley, F. H., Stripp, A., eds. (1993). *Codebreakers: The inside story of Bletchley Park*. Oxford: Oxford University Press. ISBN 978-0-19-280132-6.

Hume, D. (1739). *A Treatise of Human Nature: Being an Attempt to introduce the experimental Method of Reasoning into Moral Subjects*, II-iii-3, London: John Noon, 1739. It was later reworked and published in 1748 as *An Enquiry Concerning Human Understanding*. The text is available online at http://www.gutenberg.org/ebooks/4705 and http://www.davidhume.org/texts/thn.html.

Knuth, D.E. (1997). *The art of computer programming, volume 1 (3rd Ed.): fundamental algorithms.* Redwood City: Addison Wesley Longman Publishing Co. Inc. ISBN 0-201-89683-4.

Kroes, P. (2010). *Engineering and the dual nature of technical artefacts*. Cambridge Journal of Economics 34(1): 51-62. DOI: 10.1093/cje/bep019.

Lamport, L. (2012). *Buridan's Principle*. Foundations of Physics 42(8): 1056-1066. DOI: 10.1007/s10701-012-9647-7.

Libet, B., Gleason, C.A., Wright, E.W., Pearl, D.K. (1983). *Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential) - The Unconscious Initiation of a Freely Voluntary Act*. Brain 106(3): 623–642. DOI: 10.1093/brain/106.3.623.

Soon, C.S., Brass, M., Heinze, H.J., Haynes, J.D. (2008). *Unconscious determinants of free decisions in the human brain*. Nature Neuroscience 11(5): 543–545. DOI: 10.1038/nn.2112.

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

Turing, A.M. (1948). *Intelligent Machinery*. National Physical Laboratory Report. In Meltzer, B., Michie, D. (eds), *Machine Intelligence 5*. Edinburgh: Edinburgh University Press, 1969. Digital facsimile viewable at http://www.AlanTuring.net/intelligent_machinery.

Turing, A.M. (1950). *Computing machinery and intelligence*. Mind 59: 433-460.

Vardi, M. (2012). *What is an algorithm?* Communications of the ACM 55(3): 5–5. DOI: 10.1145/2093548.2093549.