AI & SOCIETY
https://doi.org/10.1007/s00146-022-01519-7

OPEN FORUM

A free mind cannot be digitally transferred

Gonzalo Génova[1] · **Valentín Moreno**[1] · **Eugenio Parra**[1]

Received: 3 November 2021 / Accepted: 22 April 2022

Abstract
The digital transfer of the mind to a computer system (i.e., *mind uploading*) requires representing the mind as a finite sequence of bits (1s and 0s). The classic "stored-program computer" paradigm, in turn, implies the equivalence between program and data, so that the sequence of bits themselves can be interpreted as a program, which will be algorithmically executed in the receiving device. Now, according to a previous proof, on which this paper is based, a computational or algorithmic machine, however complex, cannot be free (in the sense of 'self-determined'). Consequently, a finite sequence of bits cannot adequately represent a free mind and, therefore, a free mind cannot be digitally transferred, *quod erat demonstrandum*. The impossibility of making this transfer, as demonstrated here, should be a concern especially for those who wish to achieve it. Since we intend this to be a rigorous demonstration, we must give precise definitions and conditions of validity. The most important part of the paper is devoted to explaining the meaning and reasonableness of these definitions and conditions (for example that being truly free means being self-determined). Special attention is paid, also, to the philosophical implications of the demonstration. Finally, this thesis is distinguished from other closely related issues (such as other possible technological difficulties to "discretize" the mind; or, whether it is possible to transfer the mind from one material support to another one in a non-digital way).

Keywords Mind uploading · Free will · Information theory · Digital transfer · Stored-program computer · Sequence of bits

1 Introduction

The digital transfer of the mind to a computer system (i.e., *mind uploading*) requires representing the mind as a finite sequence of bits (binary digits: 1s and 0s). That is to say, the state of the brain is "measured" at a given moment, and the result of that measurement is transferred to a digital support, either for its mere conservation as a backup from which the state of the brain can be "restored" (to the same or to another biological brain), or else to "execute" it (in an electronic device) as a program that is supposed to be functionally equivalent to the original brain at the time of transfer. Thus, if human beings were nothing more than complicated biological robots (with a substrate in the chemistry of carbon instead of silicon), then the program that governs us could be transferred to an electronic robot, in a manner analogous to how we execute the same program on different computers.

The rest of the paper is structured as follows. In Sect. 2 we outline mind uploading as *an ethical and a technological problem*, to offer context for the third technological difficulty exposed (the program-data equivalence), which will be the subject of the rest of the paper. In Sect. 3 we synthetically present the previous research upon which this one is based, and its main conclusion, namely that *a computational machine cannot be free*. In Sect. 4, we show how modern digital computing arrived at the principle of *equivalence between program and data*, from which, in combination with the previous argument, our new thesis is derived. In Sect. 5, we state the *conditions and assumptions of validity* for our demonstration, as well as the distinction with respect to other forms of mental transference (not affected by our argument). In Sect. 6, we draw attention to some *implicit assumptions* of the idea that the digital transference of the

✉ Gonzalo Génova
ggenova@inf.uc3m.es

Valentín Moreno
vmpelayo@inf.uc3m.es

Eugenio Parra
eparra@inf.uc3m.es

[1] Departamento de Informática, Universidad Carlos III de Madrid, Avda. Universidad 30, Leganés, 28911 Madrid, Spain

Published online: 27 June 2022

mind is technically feasible and offer counter arguments. Section 7 closes the paper with some *concluding remarks*.

2 Mind uploading: an ethical and a technological problem

One can certainly question whether it is ethically desirable to carry out this process of mind uploading. Certain techno-optimistic conceptions (Bostrom and Roache 2008) want to see the eventual positive consequences alone, such as, for example:

- preserve the mind from deterioration due to disease or aging, and even use this uploaded mind as a "personal assistant";
- save the mind of a terminally ill person and "resuscitate" it in another body (immortality);
- select the best "versions" of an individual's mind, maybe through computational competition;
- even parallel work of multiple instances of the mind of a genius in different disciplines or under different stimuli.

However, there is no doubt that other big questions arise: Can the original biological person be left alive, or should we apply euthanasia to avoid his or her simultaneous living with the digital duplicate? If both continue to "live" for a while, which of the two is the repository of the rights and responsibilities acquired by the first? Regarding the other people that were related to the original, can they legitimately reject the copy because they do not accept it as a valid substitute? In the same way, we could continue with many other ethical issues that could arise if this procedure were ever a technical reality, though they still belong today to the realm of science fiction.

Nevertheless, it is not our purpose here to analyze in detail whether mind uploading is ethically *desirable* but, rather, whether it is technologically *feasible*. So, the argument of impossibility that we will develop here should be a special concern for those who wish to attain, in the indeterminate future, the supposed advantages of mind uploading. Among the technological difficulties that arise (besides the—less problematic—obsolescence of the receiving device), we point out the following:

(i) Every process of measuring a physical system assumes a certain degree of abstraction or simplification of the reality being represented. It is necessary to define the relevant variables and the degree of precision with which they will be measured. How do we know which these variables are, and how much accuracy is enough to make a "true" copy of the original? Would it be necessary to measure the individual electro-chemical state of each neuron? How do we know that the copy contains all the relevant information? (Sánchez-Cañizares 2016).

(ii) The process of digitizing the information contained in an analog signal (that is, one that is continuous both in the independent variable—usually time—and in the range of values) consists of two main steps: *sampling* involves taking measurements at fixed intervals of time; *quantification* transforms the continuous range of values into a finite set of predefined levels, eliminating the least significant part of the measure (rounding or approximation). We explain below with more detail the problems involved in digitalization.

(iii) A sequence of bits can be interpreted either as a piece of information or as a program. Now, a computer program, by its own definition (as will be explained later), cannot exhibit truly free behavior in itself, so it could not adequately represent the behavior of a self-determined being, as could be the case of humans.

Regarding the problems of digitalization, the quantification process, as it is easily understandable, introduces a certain distortion in the original information (quantification noise), which results in an irreversible *loss of information*. By contrast, the sampling theorem, due to Harry Nyquist (1928) and Claude Shannon (1949), shows that, if the sampling rate is sufficiently high (at least twice the bandwidth of the original signal), then the sampling transformation is done without loss of information.

Of course, this requires that the original information has a finite bandwidth. In practice this is fulfilled *in a very approximate way* in many artifacts with which we are familiar. For example, the human ear cannot perceive vibrations of frequency higher than 20 kHz; therefore, any sound information above this value can be neglected when digitizing a piece of music, without people with the finest ear noticing the difference. Likewise, if quantification is performed with a sufficiently high number of levels or steps, the rounding errors are also sensorially irrelevant. The above applies analogously to digital photography in relation to sight, where the independent variables are two space dimensions; and also to other types of sensory information (temperature, pressure…). Now, we must note that the brain information to be digitized like a picture is not only information for the senses (which are limited in their own perceptual capacity), so we cannot accept, without proof, that limited bandwidth and precision are acceptable assumptions.

In the rest of this paper we will develop only the third difficulty, specifying its conditions of validity and the assumptions on which it is based. Even if the other difficulties could be solved with more refined technology, this one is more fundamental and, thus, unsolvable.

3 Free will and computational machines

The thesis we want to demonstrate is that.

> a free mind cannot be digitally transferred.

The argument developed in this paper follows from another thesis we have previously demonstrated (Génova and Quintanilla Navarro 2018a). For convenience, we briefly state the two parts of that thesis, to better situate the new argument in its context:

1. A *computational machine* cannot be free (and, conversely, a free being is not a computational machine).
2. The quality of being a free entity, i.e., self-determined, is not an algorithmically *computable function* (that is, a computational machine cannot properly distinguish between free beings and non-free beings by means of a decision algorithm), or, from a more general viewpoint, free will is not a testable hypothesis (Northcott 2019).

The interested reader is referred to Génova & Quintanilla Navarro (2018a) to examine the details. Here we will only expose in a very summarized way the first part of the previous thesis, on which the new argument is based. Then we will proceed with the new argument in the next section.

Definition: An *algorithm* is (i) a rule-based procedure (ii) that obtains a desired result (iii) in a finite number of steps. In spite of some controversy (Hill 2015; Vardi 2012), this definition is firmly rooted in the pioneering works of computer science (Turing 1936). Of the three elements of the definition, we are now interested in a special way in the second one: *every algorithm is defined by its objective*, by the result it has to achieve. The algorithm, therefore, like any other human artifact, has *an extrinsic purpose*, which is imposed upon it from outside (Génova and Quintanilla Navarro 2018b). A computational machine is nothing more than a machine that works out computations algorithmically, and therefore it shares this characteristic of having an extrinsic purpose with all machines.

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, so that a machine cannot be defined and accounted for without reference to its purpose. A machine can fail to achieve its goal, but it cannot change its goal. Therefore, 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*.

Of course, there can be different levels of goal selection in a machine. There are in fact algorithms that can dynamically select among a set of given 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 own 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.

Definition: A *free being* is one that can self-propose the objectives of its activity, that is, it is a self-determined being. In this sense, *self-determination opposes both hetero-determination and in-determination*. Hetero-determination happens when 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 (i.e., both the stimuli and the programs come from outside the being whose behavior is under consideration). In-determination happens when there is 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 concept of freedom, unlike the previous one, is not purely formal, so no surprise that it is more controversial and has a long journey in the history of thought. Related to our analysis of goal-selection algorithms, an influential stance has been that of Frankfurt (1971), who claims a person's free will must be understood in terms of the capacity of having second- and higher-order goals, by means of which first-order goals are pursued; however, if the subject has not the capacity to self-propose these higher-order goals (which could certainly be the case of humans), then it is not really a free being, according to our definition—even if it could be considered 'free' from a compatibilist viewpoint, a theory with its own problems and critics (McKenna and Coates 2021). In any case, for the purposes of this demonstration we work from the given definition, whose reasonableness is examined in more detail in (Génova and Quintanilla Navarro 2018a).

From these two definitions the first thesis is derived almost immediately. A computational machine (one that executes algorithms) is hetero-determined; therefore, it is not free (it is not self-determined). A free being is capable of proposing its own objectives; therefore, it cannot be

equivalent to a hetero-determined computational machine. In consequence, as our first thesis states:

> a computational machine cannot be free.

This does not mean that it is completely impossible to "produce", in one way or another, free beings, but only that they could not be designed and constructed so as to operate in an algorithmic, computational way; if they were free, they would not be properly computational machines. On the other hand, in all this argumentation it is not assumed that human beings are free in the sense of self-determined; maybe we are not free, but only too complex to understand ourselves. Even if we are free, it seems quite clear that ours is not an absolute freedom, but a freedom that is very limited by the environment, and by our historical and corporeal human condition.

So, the most we can conclude is that,

> *if* we are free, *then* we are not computational machines.

We believe that human freedom is a reasonable possibility, but we do not need to rely on it to continue with the demonstration, which is explicitly circumscribed to *free minds*, whether human or not.

4 Equivalence between program and data

In the most archaic programmable machines, from the Jacquard Loom in the early nineteenth century, to the deciphering machines of German secret codes in the mid-twentieth century, the separation between program and data was strict. In a certain sense, the concept of program had not yet emerged, or it was extremely primitive. For the most part, machines were "programmed" by changing switches and wires on a plugboard. However, just after the end of World War II, the "stored-program computer" paradigm was successfully imposed. Its authorship has been commonly attributed to John von Neumann, although more recently it has become a disputed issue (Copeland 2020). The role of Alan Turing in the genesis of this concept ten years before, with his Universal Machine, capable of executing any program, was also crucial.

According to this conception, the program that describes an algorithm is treated as a data entry for the computational machine that processes it (and, thus, programs are written, stored, transformed, downloaded, etc., like any other piece of information). That is, the difference between program and data is progressively blurred, and in fact formal equivalence is finally achieved.

As a consequence, any program is represented as a sequence of bits, and any sequence of bits can be interpreted as a program, which will be executed algorithmically in the receiving device. To explain it with an analogy, *a musician's score* contains descriptive information of the work, and at the same time it is the program that the musician "executes" when he or she interprets it. Equally, a digitized photograph is graphic information of an image, and at the same time it is a set of instructions for the "machine" that reproduces the image on the computer screen.

This means that a sequence of bits supposedly representing the state of mind of an individual would be formally equivalent to a program or algorithm, which in turn could be executed in a computer. Now, if a computational machine executing a program cannot be truly free, then no finite sequence of bits can completely represent the state of a free being, because that sequence, as we have already said, is formally equivalent to a program. In other words:

> *a free mind cannot be digitally transferred,* q.e.d.

5 Conditions and assumptions of the demonstration

The argument of impossibility for the digital transfer of the free mind is based on three main pillars:

(1) the definition of an algorithm,
(2) the definition of a free being, and
(3) the equivalence between data and program.

It is important to emphasize that the argument does not affect other conceivable forms of mental transference. In particular, we have not shown that it is impossible:

(a) the digital transfer of *a non-free mind* (i.e., a hetero-determined mind);
(b) the *non-digital* transfer (i.e., analogical transfer) of a free mind to a non-computational support.

In other words, the possibility of a non-digital, but analogical, transfer between free biological brains, remains

open. In contrast, the argument of impossibility is still applicable to a hypothetical digital transfer to a non-computational receiver; that is, *the digital transfer from a free biological brain to another free biological brain is equally impossible*. Indirectly, this implies that digital teleportation is also impossible (though the argument is inconclusive about other forms of teleportation).

On the other hand, it is also worth noting that, in a way, a free mind could "receive" a digital transfer and interpret its content as a program (a set of instructions), which this free mind would execute in its own way. In other words, a robot cannot perfectly imitate the behavior of a free being, but a free being can choose to behave like a robot.

6 Implicit assumptions in the feasibility of mind uploading

Finally, we also want to point out some implicit assumptions that underlie the idea that, contrary to what we have shown, mental transference is technically feasible:

(i) the brain, in the end, is nothing more than a complicated biological machine, and freedom is nothing more than an illusion;
(ii) all relevant functions of the mind are computational;
(iii) there may exist a universal machine that executes the code of any brain;
(iv) the mind (the mental state) is strictly separable from the corporeal brain, i.e., the mind can be represented as an abstract structure of computation that can have a material substrate in different embodiments (a brain, a computer), without being linked to any particular body; and
(v) rationality is something essentially incorporeal, i.e., it is not the rationality of a living body, which is conceived, grows, and dies, but it is an essentially static, timeless rationality.

According to this multi-faceted view, the brain is essentially a (biological) computer, a complex information processing unit. Consequently, the best way to understand the brain (and, at the same time, human knowledge and behavior) is to study it under the paradigm of causal relationships. This *computational conception of the brain* (Rescorla 2020) has become mainstream (see, for example, Patricia Churchland, 1992), but it is not shared by all reputable scientists. Among the opponents of this conception, we can mention David Gelernter (2014, 2016): "Man is only a computer if you ignore everything that distinguishes him from a computer". One of the most radical distinctions is indeed free will. Those who do not dare to radically affirm freedom are precisely the ones that will most easily fall into the temptation of considering that humans are ultimately nothing more than complicated biological robots (Génova and Quintanilla Navarro 2018a).

Establishing the field of artificial intelligence as a scientific discipline seemed strongly linked to the idea that human cognition can be described in terms of a symbol system, as credited to Allen Newell and Herbert Simon in their famous 1975 ACM Turing Award Lecture (Newell and Simon 1976). Certainly, a computer can be well characterized as a system of manipulating symbolic structures, but the same idea cannot be simply transferred as such to the human mind: intelligence does not consist in following rules (Dreyfus 1992), even if interpreting and following rules is an important part of intelligence.

Raymond Tallis (2004), among many others, has shown the fallacies beyond the computational view of mind. Summing up his view, there is a fallacy in the attribution of agency to a computer as if it were a conscious subject, when the truth is that those functions are only performed in conjunction with the person using the computer; the computer is only an instrument, the true subject of those acts is the person. The computational theory of mind has been criticized also from the viewpoint of semiotics (Fetzer 2001; Nöth 2008).

When computation is taken as a model from which to understand the mind, it is considered that the essential objective of the latter is analogous to that of the former: to solve problems. As Barrett et al. (2015) put it, "the understanding of the mind as problem-solving has been the dominant approach for at least the last 50 years, but it is becoming increasingly clear that the mind overflows it". In fact, this reductive, instrumental view of intelligence as a problem-solving capacity, leaving aside the selection of problems worth solving (as though it were not even more fundamental to intelligence), can be traced back even to Descartes.

Of course, there is still strong controversy regarding these conceptions of the mind, but remember (see Sect. 3) that our thesis does not rest on the human mind not being computational, but rather on the opposition between computation and free will understood as self-determination. Our demonstration is explicitly circumscribed to *free minds*, whether human or not.

7 Conclusion

The musician's score can be interpreted in different ways by different interpreters, but it is, in itself, a closed and very limited program/description of the musical work, and it is obviously not a self-determined being. The techno-optimism of mental transfer considers, in the words of José Ignacio Latorre, that it is worth transferring the mind to a machine, because it "will not suffer the limits imposed by biology" (2019). Now, as the corporeal support of a supposedly free

being, biology certainly imposes its limits, but the machine imposes much more severe ones: the exclusion of free will by the hetero-determination that characterizes any machine.

A machine is always an obedient slave to its programming (Génova and Quintanilla Navarro 2018a). Even if this programming corresponds to the "portrait" or "score" of a free being at a given moment, the machine will never be truly self-determined; it will always be an imitation, a dynamic imitation if you want, but always a pale reflection of the original. Is it worth the risk of losing freedom—even if it is uncertain—to keep the mind packaged in a sort of mobile and "autonomous" portrait?

Acknowledgements This research has received funding from the RESTART project "Continuous Reverse Engineering for Software Product Lines/Ingeniería Inversa Continua para Líneas de Productos de Software" (ref. RTI2018-099915-B-I00, Convocatoria Proyectos de I + D Retos Investigación del Programa Estatal de I + D + i Orientada a los Retos de la Sociedad 2018); MOMEBIA project "Monitorización del Mercado Eléctrico Basada en técnicas de Inteligencia Artificial" (ref. RTC2019-007501-7, Convocatoria de Proyectos de I + D + i «Retos-Colaboración» 2019—Ministerio de Ciencia e Innovación—Agencia Estatal de Investigación); it has also been supported by the Madrid Government (Comunidad de Madrid-Spain) under the Multi-annual Agreement with UC3M in the line of Excellence of University Professors (EPUC3M17), and in the context of the V PRICIT (Regional Programme of Research and Technological Innovation).

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data availability Our manuscript has no associated data.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Barrett N, Güell F, Murillo JI (2015) Los límites de la comprensión computacional del cerebro. Cuenta y Razón 34: 71–76 http://cuentayrazon.com/wp-content/uploads/2016/05/revista34.pdf

Bostrom N, Roache R (2008) Ethical issues in human enhancement. In: Ryberg J, Petersen T, Wolf C (eds) New waves in applied ethics. Palgrave Macmillan, New York, pp 120–152

Churchland PS, Sejnowski TJ (1992) The computational brain. The MIT Press, Cambridge

Copeland BJ (2020) The Modern History of Computing. The Stanford Encyclopedia of Philosophy (Winter 2020 Edition), Edward N. Zalta (ed.) https://plato.stanford.edu/archives/win2020/entries/computing-history/

Dreyfus HL (1992) What computers still can't do: the limits of artificial intelligence. Harper and Row, New York

Fetzer J (2001) Computers and cognition. Why minds are not machines. Springer, New York

Frankfurt H (1971) Freedom of the will and the concept of a person. J Philos 68(1):5–20. https://doi.org/10.2307/2024717

Gelernter D (2016) The tides of mind: uncovering the spectrum of consciousness. Liveright, New York

Gelernter D (2014) The Closing of the Scientific Mind. Commentary Magazine. https://www.commentarymagazine.com/articles/the-closing-of-the-scientific-mind/

Génova G, Quintanilla Navarro I (2018a) Are human beings humean robots? J Exp Theor Artif Intell 30(1):177–186. https://doi.org/10.1080/0952813X.2017.1409279

Génova G, Quintanilla Navarro I (2018b) Discovering the principle of finality in computational machines. Found Sci 23(4):779–794. https://doi.org/10.1007/s10699-018-9552-4

Hill RK (2015) What an algorithm is. Philos Technol 29(1):35–59. https://doi.org/10.1007/s13347-014-0184-5

Kroes P (2010) Engineering and the dual nature of technical artefacts. Camb J Econ 34(1):51–62. https://doi.org/10.1093/cje/bep019

Latorre JI (2019) La singularidad. http://lab.cccb.org/es/la-singularidad/

McKenna M, Coates DJ Compatibilism. The Stanford Encyclopedia of Philosophy (Fall 2021 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2021/entries/compatibilism/

Newell A, Simon HA (1976) Computer science as empirical inquiry: symbols and search. Commun ACM 19(3):113–126. https://doi.org/10.1145/360018.360022

Northcott R (2019) Free will is not a testable hypothesis. Erkenntnis 84(3):617–631. https://doi.org/10.1007/s10670-018-9974-y

Nöth W (2008) Sign machines in the framework of semiotics unbounded. Semiotica 169:319–341. https://doi.org/10.1515/SEM.2008.041

Nyquist H (1928) Certain topics in telegraph transmission theory. Trans Am Inst Electr Eng 47:617–644. https://doi.org/10.1109/T-AIEE.1928.5055024

Rescorla M (2020) The computational theory of mind. The Stanford Encyclopedia of Philosophy (Fall 2020 Edition), Edward N. Zalta (ed.) https://plato.stanford.edu/archives/fall2020/entries/computational-mind/

Sánchez-Cañizares J (2016) Entropy, quantum mechanics, and information in complex systems: a plea for ontological pluralism. Eur J Sci Theol 12(1):17–37

Shannon C (1949) Communication in the presence of noise. Proc Inst Radio Eng 37(1):10–21. https://doi.org/10.1109/PROC.1984.12998

Tallis R (2004) Why the mind is not a computer: a pocket lexicon of neuromythology. Imprint Academic, Exeter

Turing AM (1936) On computable numbers, with an application to the Entscheidungsproblem. Proc Lond Math Soc 2(42):230–265. https://doi.org/10.1112/plms/s2-42.1.230

Vardi M (2012) What is an algorithm? Commun ACM 55(3):5–5. https://doi.org/10.1145/2093548.2093549