Theoria, 2020, 35(3), 345-364
https://doi.org/10.1387/theoria.21302

THEORIA

ISSN 0495-4548 – eISSN 2171-679X

Free will and (in)determinism in the brain: a case for naturalized philosophy[&]

(Libre albedrío e (in)determinismo en el cerebro: un problema para la filosofía naturalizada)

Louis Vervoort*, Tomasz Blusiewicz
University of Tyumen, Russian Federation

ABSTRACT: In this article we study the question of free will from an interdisciplinary angle, drawing on philosophy, neurobiology and physics. We start by reviewing relevant neurobiological findings on the functioning of the brain, notably as presented in (Koch, 2009); we assess these against the physics of (in)determinism. These biophysics findings seem to indicate that neuronal processes are not quantum but classical in nature. We conclude from this that there is little support for the existence of an immaterial 'mind', capable of ruling over matter independently of the causal past. But what, then, can free will be? We propose a compatibilist account that resonates well with neurobiology and physics, and that highlights that free will comes in degrees — degrees which vary with the conscious grasp the 'free' agent has over his actions. Finally, we analyze the well-known Libet experiment on free will through the lens of our model. We submit this interdisciplinary investigation as a typical case of naturalized philosophy: in our theorizing we privilege assumptions that find evidence in science, but our conceptual work also suggests new avenues for research in a few scientific disciplines.

KEYWORDS: free will; consciousness; compatibilism; neuroscience; determinism; Libet experiment; naturalized philosophy; quantum mechanics.

RESUMEN: *En este artículo estudiamos la cuestión del libre albedrío desde una perspectiva interdisciplinar, combinando filosofía, neurobiología y física. Comenzamos revisando hallazgos relevantes en neurobiología acerca del funcionamiento del cerebro, en especial los presentados en (Koch, 2009); estos resultados son evaluados en relación con la física del (in)determinismo. Tales hallazgos en biofísica parecen indicar que los procesos neuronales no son de naturaleza cuántica, sino clásica. De aquí concluimos que hay poco apoyo a la existencia de una 'mente' inmaterial, capaz de gobernar la materia independientemente del pasado causal. Pero, ¿en qué puede consistir entonces el libre albedrío? Planteamos una propuesta compatibilista que concuerda con la neurobiología y la física, y en la que se destaca que el libre albedrío se da en grados, dependiendo de la comprensión consciente que el agente 'libre' tenga de sus acciones. Finalmente, analizamos el famoso experimento de Libet sobre el libre albedrío desde la perspectiva de nuestro modelo. Presentamos esta investigación interdisciplinar como un ejemplo típico de filosofía naturalizada: en nuestro teorizar privilegiamos asunciones con apoyo en la evidencia científica, aunque nuestro trabajo conceptual también sugiere nuevos caminos para la investigación en varias disciplinas científicas.*

PALABRAS CLAVE: libre albedrío; consciencia; compatibilismo; neurociencia; determinismo; experimento de Libet; filosofía naturalizada; mecánica cuántica.

[&] We would like to thank, for expert feedback, the participants of the conference "Free Will and Causality" at the University of Dusseldorf (September 2019), in particular Maria Sekatskaya, Laura Ekstrom, Nadine Elzein, Timothy O'Connor, Alexander Gebharter, as well as Giacomo Andreoletti for careful reading of the manuscript. We thank Krishna Muthukumarappan for consultancy in the interpretation of neurobiological results, and two anonymous reviewers for helpful comments.

* ***Correspondence to:*** Louis Vervoort. School of Advanced Studies, University of Tyumen, Ulitsa Volodarskogo 6, 625003 Tyumen (Russian Federation) – l.vervoort@utmn.ru

How to cite: Vervoort, Louis; Blusiewicz, Tomasz (2020). «Free will and (in)determinism in the brain: a case for naturalized philosophy»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 35(3), 345-364. (https://doi.org/10.1387/theoria.21302).

Received: 16 December, 2019; Final version: 27 April, 2020.
ISSN 0495-4548 - eISSN 2171-679X / © 2020 UPV/EHU

1. Introduction

The belief that we are free-willed people, proudly capable of choosing between alternative options almost whenever we want —say between now lifting my left hand, or not doing so— is so deeply embedded in humans that most of them believe that questioning free will is not even a serious pass-time. Interestingly enough, a handful of philosophers and scientists camp on the opposite position, and believe that free will is the ultimate chimera to combat, the last illusion of human 'exceptionality' to be reduced to smithereens, just as Copernicus, Darwin, Freud and others did with other illusions.

In philosophy, the debate on free will is millennia-old; almost all philosophers of renown have expressed their thoughts on it. Two typical and opposing positions in philosophy are libertarianism and hard determinism (for general works focusing on philosophy, cf. e.g. Pereboom, 1997; Walter, 2001; Dennett, 2003; Watson, 2003; Kane, 2005, 2012; Mele, 2009; Fischer *et al.,* 2007; Griffith, 2013). In essence, libertarians believe that we have a free will and that free will is incompatible with determinism; hard determinists on the other hand assume that physical or nomological determinism rules out free will. Determinists often conceive of the universe as a mechanical clockwork in which each particle and each system, be it a photon, atom or neuron, is bound by laws of nature to behave in a way that is predetermined, fixed, since the Big-Bang. Since our brain, the seat of our will, is in the end a physical and a biological system, and since natural science leaves no room for free will, or so it seems to the hard determinist, humans only *feel* free in an illusory manner. There is a third, popular, position, termed compatibilism, which states that free will and determinism are *not* in contradiction. For reasons explained below, the model of free will we will favour in this article is compatibilist.

In neurobiology, a number of well-known and highly relevant experiments related to free will exist, of which the most famous is Libet's experiment (Libet, 1985); but a considerable amount of other sophisticated investigations have been conducted, both on humans and animals (cf. e.g. Heisenberg and Wolf, 1984; Koch, 1999, 2009; Walter, 2001; Maye *et al.,* 2007; Soon *et al.,* 2008; Brembs, 2011; Stern, 2017). Reading these works confers the definite impression that neuroscientists (understandably) often draw conclusions from their experiments without a full knowledge of either the philosophical debate or of the physics debate on the matter. Sometimes they discard the philosophical debate as outdated and superfluous (see e.g. Koch, 2009; Brembs, 2011), advocating for instance the construction of a 'scientific concept of free will as a biological trait' (Brembs, 2011).

Here we embrace the point of view that a better understanding of free will, notably the high-level type of free will necessary for moral responsibility, needs to involve, to the least, the disciplines of neurobiology, philosophy and physics. In physics, too, the question of free will is highly debated (cf. e.g. the recent article by Nobel laureate Gerard 't Hooft (2017)), and relates to the determinism versus indeterminism controversy in the quantum realm[1] (cf. e.g. Wuethrich, 2011; Vervoort, 2013, 2019). This debate, famously initiated by Einstein and Bohr, is continued today most vividly in the interpretation of Bell's theo-

[1] On the standard interpretation of quantum mechanics and quantum field theories the universe is ultimately indeterministic (so probabilistic). But this matter is highly debated within the quantum philosophy and quantum foundations community (Wuethrich, 2011; Vervoort, 2013, 2019).

rem. Bell's theorem is generally believed to provide the strongest 'purely physical' argument against determinism; but there is now a consensus that this argument is actually heavily metaphysically tainted (cf. e.g. Wuethrich, 2011). Indeed, closer inspection shows that the standard interpretation of Bell's theorem as refuting determinism is based on the assumption that experimenters have 'free will' — precisely in the sense that their experimental choices are not determined by hidden physical causes (Wuethrich, 2011; Vervoort, 2013, 2019; 't Hooft, 2017). But as many compatibilists have argued, free will can exist even if all events, including experimental choices of physicists, are causally predetermined. So remarkably enough, it seems that the standard interpretation of Bell's theorem against determinism smuggles in the assumption of no-determinism from the start. In sum, it is important for the free will debate to realize that the foundations of quantum physics leave the dichotomy determinism versus indeterminism undecided. And yet, it has recently been argued that, of both alternatives, determinism has the strongest explanatory power, in that it can provide a coherent answer to questions for which indeterminism remains entirely silent (Vervoort, 2019). This is, to us, an argument to seek for compatibilist models of free will. Of course, many compatibilists do not need physics to find a series of arguments in favour of determinism in philosophy or neuroscience. Let us here just note that the stakes are high in physics too, since Bell's theorem can be seen as an essential obstacle in the construction of a 'theory of everything' ('t Hooft, 2017). In order to debunk this obstacle, it is therefore important to provide arguments showing that free will makes sense also if the universe would be fully deterministic after all; an effort to which we wish to contribute in this article.

In order to make our point that genuinely interdisciplinary work is needed for an integrated understanding of free will, we will start by scrutinizing in Section 2 recent neurobiological work, notably by Christof Koch (2009). The key question we wish to investigate in Section 2 is: does neurobiology offer arguments in favor of determinism or rather indeterminism? We focus on Koch's work in his (2009), not only because he is a recognized expert on the matter, but also because he is already remarkably well-informed about the physics debate (Koch is a physicist turned neurobiologist). However we will argue that his interpretation of neurobiological results presents lacunae *in the philosophical and physical treatment* of the matters he investigates; we will conclude, *pace* Koch, that the neurobiological processes he considers bare a classic deterministic signature rather than an indeterministic quantum one (Section 2). Let us emphasize that Koch's professional work is in the first place of neurobiological order; we do of course not doubt the high quality of this scientific work. In contrast to many or most of his colleagues from neurobiology, Koch is not a determinist. We believe it is fair to consider him a libertarian adhering to a type of mind-matter dualism, a position that is close to, e.g., the position promoted by Popper and Eccles in their (1977), to whom Koch often refers (Eccles is a neurophysiologist and Nobel laureate). Of course, Popper and Eccles are in a long philosophical tradition here. In particular, Koch believes in a mind that can somehow realize or 'cause' free actions; in any case he does not exclude this option, and searches for experimental support for this idea. After critical assessment of Koch's interpretation of neurobiological data in his (2009), we attempt in Section 3 to answer the question of *what else* could be the basis for free will, if not a mind transcending the causal brain. We summarize in this Section the main ingredients of a minimal compatibilist model which we elaborated in detail elsewhere (Vervoort & Blusiewicz, 2020), and which draws on a biophilosophical account of consciousness developed by

Mahner and Bunge (1997). This model is analytically minimal in the sense that it aims to identify the minimal set of necessary and sufficient conditions to term an act 'free-willed' or 'free'.[2] Most of the ingredients of this model have been theorized before; it could be seen as a condensed synthesis of previous compatibilist theories; yet, we push the analysis further in some respects (cf. Vervoort & Blusiewicz, 2020; and Section 3.) Our main goals in the present article are 1) to apply our model to the paradigmatic Libet experiment on free will, and 2) to briefly describe further avenues of research that are suggested by it, in neurobiology and computer science. Both these goals are pursued in Section 4. Another goal is, as said, to scrutinize what we can learn from neurobiology regarding the (in)deterministic nature of brain processes (Section 2).

We thus submit this investigation as a typical problem of naturalized or naturalistic philosophy. We not only borrow from natural science to inform our philosophical theorizing, but we also extract new avenues for further research in these sciences from our conceptual inquiry.

2. Neurobiology results, and their (in)deterministic interpretation

In this Section we summarize and comment on the experimental neurobiological findings reported in (Koch, 2009) and other works, related to the (in)deterministic nature of neuronal processes. Koch in his (2009) searches for neurobiological properties that could support the libertarian concept of 'free mind'. This implies that the author needs to find arguments for *in*determinism in neurobiology (libertarians deny determinism and need indeterminism for their account); in other words for the hypothesis that some relevant neuronal processes escape from a clockwork causal structure of the brain and the universe. Let us first scrutinize the experimental facts Koch reports (2009). Then we will focus on their physical and philosophical interpretation.

2.1. Indeterminism in ionic channel currents

The units or bits of information on which neuronal communication is based are the so-called 'spikes', or rather 'action potentials' emitted by neurons. Spikes or action potentials are trains of electric currents generated along the axons of the neurons. Neurobiological research has found that spike currents are constituted of the combination of many thousands of smaller contributions, the latter being generated in microscopic ion channels that pass through the axon membrane. Now, while the spikes are, in the terminology of Koch (2009), "macroscopic, continuous and deterministic", the smaller ionic currents are "microscopic, binary, and stochastic" (binary means fluctuating between 'off' (0) and 'on' (1), and stochastic means probabilistic, or random, i.e. indeterministic). Thus, the probabilistic aspect of the ion currents resides here in the fact that the 'on' and 'off' domains arise in a probabilistic manner in time: there is no regular pattern in them — as is obvious from, for instance, the measured current profiles published in Koch (1999).

[2] Our model is a two-component model. Distinguishing several components in free will is also done by Walter (2001); both models can be fruitfully compared, as done in (Vervoort & Blusiewicz, 2020).

Could this aspect of probabilistic behaviour provide the element of indeterminism (in our brain activity when making free choices) that the libertarian is looking for?

2.2. Indeterminism in spike frequency

A similar randomness occurs in the frequency of the action potentials. Koch (2009) reports on experiments in which a micro-electrode is implanted close to a particular neuron in the cortex of a macaque monkey looking at a display of a randomly moving cloud of dots (see also Koch, 1999). (This projection ensures that the neuron's activity is reasonably constant over different experimental trials.) When the display is turned on, the neuron begins to 'fire', i.e. to emit spikes, at a frequency that can be measured via the electrode. Now, over different trials, the spikes are emitted in a random manner; more precisely, their temporal pattern varies unpredictably over the trials, but the average number of spikes remains reasonably constant. So one can define a probability of spike frequency. Here is again an indeterministic aspect of brain functioning and one can again ask: could this be the indeterminacy that allows our brain and our decision making to escape from the iron laws of biophysical determinism?

2.3. Indeterminism in fly behaviour

In the literature on the neurobiology of consciousness and free will, an important place is occupied by research investigating the behavior of model animals, notably the fruit fly (the famous and hard-working Drosophila melanogaster, inadvertently involved in the attribution of at least eight Nobel prizes). If the brains of flies were purely deterministic systems, one could expect that elementary behaviour (say flying in a straight line) in an environment presenting as little external stimuli as possible, would be very simple — say fully regular or perhaps fully random ('fully random' means: all movements having equal probability[3]). Drosophila m. has decided otherwise: her behaviour appears to be highly complex, somewhere in between fully regular and fully random — she acts capriciously, *'voluntarily'* as it were. This is shown by well-known experiments in which fruit flies are tethered to a rigid wire (receiving no visual impulses) and in which their 'escape' flights are registered while the animal can only turn left or right (Heisenberg and Wolf, 1984; Maye *et al.,* 2007). The flies appear to execute *stochastic saccades following a fractal pattern*. In Koch's words, "the animal behaves neither completely randomly nor fully deterministically, but opts for something in between chance and necessity" (Koch, 2009, p. 45). This 'spontaneous' complex behaviour is typically interpreted as (an embryonic form of) *voluntary* behaviour (Koch, 2009, p. 45) — where we wish to remark that the concept of voluntariness is indeed closely related to that of free will, but that it is nonetheless far from easily definable, as will be seen in the next Section.

[3] Fully random behaviour would reflect the fact that in experiments the flies are exposed to no structured external (visual) stimuli — in physical terms: only white noise (white noise is unstructured randomness). If the brain were a simple deterministic input/output box, then 'randomness in' would cause 'randomness out'. But things are not as simple.

2.4. Determinism in Libet-like experiments

Koch interprets the well-known neurobiological result of Libet (1985), which we will consider in greater detail in the last Section, as indicating that "the brain can make a simple decision well before the conscious mind does; this observation reveals the experience of willing an action to be secondary to the actual cause" (Koch, 2009, p. 49). 'The actual cause' is the working of the brain. This is a quite standard interpretation — an interpretation that, as Koch concedes, is rather in agreement with determinism than with indeterminism. Modern versions of this experiment have confirmed Libet's work based on electroencephalography (EEG, e.g. Haggard and Eimer, 1999). A considerably more sophisticated technique, fMRI (functional Magnetic Resonance Imaging), has been used in experiments by e.g. Soon *et al.* (2008). These researchers monitored the brain activity of test persons who were asked to flex either their left or their right hand. Soon *et al.* reported that, by screening their hemodynamic activity in the parietal and prefrontal cortex, their choice could be predicted 8 seconds prior to the onset of movement[4] (!). Once more, there is a long time span of brain activity before the conscious choice is made.

Let us now scrutinize Koch's and other typical interpretations of these scientific results. First of all, it should be noted that many philosophers agree that fundamental randomness, or fundamental indeterminacy, is, *in itself*, not a hallmark of free decisions (cf. e.g. Pereboom, 1997; Walter, 2001; Dennett, 2003; Griffith, 2013). The argument generally used against the idea that indeterminism alone can ground free will is the so-called 'luck argument': if my actions are the result of an indeterministic process occurring in my brain that might very well not have occurred, it is hard to see how these actions are under my control and up to me (libertarians may contest this conclusion). At any rate, we take it that a libertarian, more precisely a 'libertarian *à la* Koch', looks for neuronal processes in humans that satisfy at least *two* criteria, L1 and L2, namely neuronal processes that:

L1) escape from the straightjacket of determinism, *and at the same time,*
L2) can be somehow (freely and consciously) '*influenced by the mind*' (cf. e.g. Koch, 2009, p. 49) or '*controlled by the mind*' (e.g. Koch, 2009, p. 41).

It is clear from Koch's article that he focuses on the case of an *immaterial* mind (see e.g. Koch, 2009, p. 41, where he compares the mind to a 'metaphysical ectoplasm'). What do biology and physics tell us about these requirements or hypotheses?

As exemplified by the results 2.1-2.3, there indeed exist indeterministic neuronal processes that are or could be at the basis of our conscious behaviour (L1). But physicists immediately add that the indeterminism of these specific processes corresponds to classical randomness, not quantum randomness (this is acknowledged by Koch). This is essential, because classical randomness is usually seen as *deriving from deterministic, even if unpredictable, processes*. Only quantum indeterminacy is genuine, irreducible randomness, which cannot be reduced to underlying deterministic processes. Regarding empirical result 2.1, Koch (2009, p. 43) says: "Given the large size of channel proteins [...] it is generally believed that the stochastic character of ionic channels can be entirely explained by thermal fluctuations and does not rely on quantum indeterminacy [...]". And regarding 2.2, the

[4] Some authors, however, criticize this result by pointing out that, depending on experimental conditions, the predictions were only correct with a certain probability, not necessarily clearly significant.

randomness and variability of action potentials of individual neurons: "Some of this variability is due to trembling eyes, the exact timing of the heartbeat, breathing, and so on. The remaining unpredictability is thought to be accounted for by the incessant movement of the molecules, primarily water, making up the wet and warm brain — thermal motion [...]. This ceaseless motion cannot be predicted but is still subject to the laws of cause and effect. Biophysicists by and large believe that quantum mechanics has no essential role to play here. While nervous systems —like anything else— obey quantum mechanics, the collective effects of all these molecules frenetically moving about is to smear out any quantum indeterminacy. At the cellular level, neurons look to be firmly governed by classical physics" (2009, p. 43-44). This analysis is indeed in full agreement with what for instance quantum condensed matter physics and quantum chemistry teach us, namely that quantum coherence and quantum phenomena in general are destroyed when atoms or molecules are brought to temperatures above say liquid-nitrogen temperatures (an icy 77 Kelvin or –196º Celsius) — when thermal kinetic effects tune in. Something similar is true for the stochastic saccades following a fractal pattern observed on fruit flies (result 2.3): fractals are typically explained by the theory of non-linear systems, or *deterministic* chaos (Maye *et al.,* 2007).

So, neuronal phenomena 2.1-2.3 are usually understood as deriving from classic *deterministic* processes, but at the same time they are *unpredictable*. This brings us to the heart of Koch's analysis in his (2009). Indeed, on closer inspection, Koch takes the unpredictability of brain activity, well-illustrated by 2.1-2.3, as the *key guarantee for the possibility of free will*. He stresses on p. 45: "Your actions are not, and never will be, predictable. Even though the universe and everything within it obeys natural laws, the state of the future world is contingent in a way that, in general, cannot be computed from its current state".

Now, there is little doubt that the detailed time-behaviour of ion channel currents, as well as of action potentials, as well as the features of many other (individual or collective) brain processes will for ever elude human prediction — we have, alas or luckily, too limited cognitive and epistemic means at our disposal for that. This unpredictability-in-practice is universally acknowledged in physics and philosophy. However, it is *not* considered a challenge to fundamental determinism, in other words, it does not amount to an (L1)-type argument for libertarianism. As recalled above in criterion (L1), a libertarian can only interpret genuinely indeterministic neuronal activity as a basis for free will, not 'just' unpredictable activity.[5] Most unpredictable systems in physics are still deterministic, deterministic chaos as reported in result 2.3 being the paradigm example. Only genuine quantum processes could satisfy (L1); but there is no evidence for these in the working of neurons. In sum, the unpredictability to which Koch points is decidedly not enough.

Koch seems to be aware of this problem for his account, since at times he speculates on the possibility that there still could exist relevant quantum processes in the brain. Let us bear with the author and concede this point: that it still might be possible, against the odds, that future biophysical research will find evidence for a truly quantum, irreducibly indeterministic process involved in conscious brain activity. Koch envisages the following possibility (2009, p. 40): "What cannot be ruled out is that tiny quantum fluctuations deep in the

[5] A libertarian assumes that genuine alternatives are open to the mind, and therefore needs to find arguments against the single-track necessity of determinism.

brain are amplified by deterministic chaos and will ultimately lead to behavioral choices. This is the basis of Jordan's quantum amplifier hypothesis of free will [...]. The release of a single synaptic vesicle may be dependent on some pre-synaptic quantum event. This might generate an action potential in the post-synaptic neuron that, in turn, triggers a cascade of active neurons that ultimately give rise to movement." This eighty-year-old 'quantum amplifier' hypothesis is indeed well-known, and was proposed by one of the founding fathers of quantum theory, the German physicist Pascual Jordan (Jordan, 1938).

Let us assume, for the time being, that the quantum amplifier hypothesis satisfies requirement (L1). What about (L2)? Are there neurobiological indications that an immaterial mind could somehow control quantum processes (or processes generated by quantum processes)? Koch speculates (2009, p. 41-42): "The only freedom that such a mind could have is to realize one quantum-mechanical event rather than another one as dictated by Schrödinger's law. Say, for example, that at a particular point in time and at a particular synapse in cortex, a superposition of two quantum mechanical states occurs. There is a 10% chance that the synapse will switch —sending a chemical signal across the cleft separating two neurons— and a 90% chance that nothing happens. [...] Given our present interpretation of quantum mechanics, it cannot be ruled out that the conscious mind could exploit this idiosyncratic freedom. It is powerless to change these probabilities —that would cost energy— but it might be able to decide what happens on any one trial. [...] We do not know whether this is even within the realm of the possible. But at least it cannot be ruled out". The 'idiosyncratic freedom' Koch refers to is the capacity of the mind to choose between quantum states in superposition (while leaving the experimentally verifiable probabilities unaltered).

We fear however that the possibility that Koch envisages here runs more against what physics teaches us than he claims. Quantum processes are indeed probabilistic and it is fully legitimate to assume that they are characterized by several outcomes or states (say switching or not switching, as above) and that these states arise each with a certain probability. However, physics offers no support for the idea that 'something immaterial' could influence a quantum system in the brain and cause it to assume a certain state rather than another one. Indeed, to cause a quantum system to assume a state (in physics terms: to cause the wave function to collapse on some state) amounts to an interaction, and interactions inevitably cost energy. But the principle of conservation of energy applied to the brain, in other words the energy balance of the brain, does not leave room for such an extraneous immaterial agent — the energy balance of the material brain is in equilibrium. Therefore, the (L2)-type of argument Koch invokes —the possibility of an immaterial substance choosing among quantum states— contradicts one of the most fundamental laws of nature, namely the energy conservation principle.[6] This adds of course to the fact that, till date, there is no other biological nor physical evidence that could back-up the existence of a mind beyond the brain. To be fair, we should remark that Koch is himself all but adamant in defending his mind-model; he seems well aware that it remains, from a scientific perspective, a highly speculative option.

Let us now draw our conclusion, based on the neurobiological data and models presented by Koch and others. We believe it is fair to say that there exists, till date, no suffi-

[6] Somewhat surprisingly, at times Koch seems to be aware of this fact: see his correct observations on the principle of energy conservation (Koch, 2009, p. 41).

cient neurobiological nor physical evidence to support the idea of libertarianism based on (L1) (in short, quantum indeterminacy) and (L2) (in short, a controlling mind). As neurobiologists and biophysicists have argued, neuronal systems are too large and too hot for quantum effects to survive; and the arguments for an immaterial mind that could control quantum states seem even more doubtful to us. Also, the experiments performed by Libet and others (e.g. Soon *et al.,* 2008, findings 2.4 above) are usually interpreted as in agreement with determinism rather than with indeterminism (of course, libertarians have contested this conclusion, since this is a debatable matter).

Guided by the above conclusions, as well as by philosophical arguments (notably the mentioned 'luck argument') and by physics results (Vervoort, 2013, 2019), we will adopt in the following a deterministic or rather compatibilist position on free will. We have elaborated our compatibilist model of free will elsewhere (Vervoort & Blusiewicz, 2020). Let us summarize and illustrate our findings in the next Section, and then show in Section 4 how they deal with Libet's experiment and which avenues of research they suggest.

3. A minimal compatibilist model of free will: the CMT model

Let us first state our main background assumptions. First of all, we adhere in the following to 'materialist-emergentist realism' essentially as it is expounded in (Walter, 2001) and in most detail in (Mahner and Bunge, 1997); for a summary see (Vervoort & Blusiewicz, 2020, Section 5). This implies in particular that we assume that mental activity is brain activity: our thoughts, beliefs, choices, feelings etc., have a neurological, and ultimately chemical-physical basis in the brain; mental states correspond to neural (super)networks, mental acts are brain processes. As recalled in (Vervoort & Blusiewicz, 2020), we follow Mahner and Bunge in their characterization of the 'mind' (of agent A) as a conceptual object, namely the union, or *set*, of all mental processes of the brain of A (Mahner and Bunge, 1997, p. 205). Further, this view implies, notably, that: "There can be no mind-matter interaction because —unlike individual mental processes and brains— mind and matter are sets, hence conceptual objects. However, it does make sense to speak of 'mental-bodily interactions' provided this expression is taken to abbreviate 'interactions among plastic neuronal systems, on the one hand, and either committed neuronal systems or bodily systems that are not part of the Central Neuronal System on the other' (Mahner and Bunge, 1997, p. 205).

Next, we follow in this article most philosophers in the assumption that 'free will' is 'free will necessary for moral responsibility' (note that the present article is about free will, not moral responsibility; but for completeness we present a concise view on the latter concept in the Appendix). Thus the free will we study here is 'more' than animal free will, if such a thing exists. But let us immediately add: from a neurobiological point of view, it is natural to assume that *if* something like free will exists, there should be a smooth evolutionary transition from an embryonic form of free will in animals to a higher-grade species in humans (cf. e.g. Brembs, 2011). As we will see in a moment, the conceptual model of free will we proposed in (Vervoort & Blusiewicz, 2020) allows for this smooth evolution.

In philosophy, a position intermediate between libertarianism and hard determinism, called compatibilism, assumes that free will and determinism are not in contradiction (for a general introduction, cf. e.g. Ayer, 1954/1997; Pereboom, 1997; Walter, 2001; Den-

nett, 2003; Griffith, 2013). Many and perhaps a majority of philosophers studying free will are compatibilists (Pereboom, 1997, p. 242). Let us immediately clarify: the free will of the compatibilist model presented below is *not* the absolute free will of an independent mind or agent in which a libertarian *à la* Koch believes — a mind or agent with the capacity to choose between genuine alternatives independently of prior causes. If determinism is true, such ontologically fundamental alternatives do not exist, as already argued by Spinoza. But *what could then be* this type of free will that is compatible with determinism? This remains an open question in philosophy, but it appears that a thorough literature search and a synthesis of popular, cogent theories brings us quite far. This is essentially what we did in (Vervoort & Blusiewicz, 2020); there we argue that our model solves problems of other theories, so that we can have a relative confidence in it.

Many compatibilists agree that an essential ingredient of free-willed actions is the condition that they are not compelled by an external agent (see e.g. Ayer, 1954/1997; this idea traces back to Aristotle, Hobbes and others). But even more essential is, as realized already by Aristotle in his *Nicomachean Ethics*, that a free-willed act should be *voluntary*. To analyze the complex concept of voluntariness we relied on a detailed neurobiological theory developed in Mahner and Bunge (1997). Without going in details, voluntary can be analyzed as purposeful and *conscious*; and for consciousness Mahner and Bunge propose following definition (in a slightly modified phrasing, cf. Mahner and Bunge, 1997 p. 209):

> **DEF-1**. A *conscious mental process* / choice / act (conceived as based on, governed by, a mental process) is a mental process / choice / act that is *monitored* (recorded, analyzed, controlled, or kept track of) by some other mental activity in the same brain.

In other words, in essence: for a mental process or act to be conscious, it must be thought about by a higher-level part of the brain, typically thought of as being localized in the prefrontal cortex. This leads then to following definition of a free act (Mahner and Bunge, 1997; Vervoort & Blusiewicz, 2020):

> **DEF-2**. Action A by animal b is 'free-willed' or 'free' (is made of b's own free will) IFF
>
> i. the action A is *unconstrained* (no programmed or external compulsion), and
> ii. the action A is purposeful and *conscious* in that the action (linked to a mental process) is *monitored* (recorded, analyzed, controlled, or kept track of) by some other mental activity in the brain of b.

Clause (i), no programmed or external compulsion, refers to the absence of direct or programmed constraint by external agents (as well as of pathological or compulsory internal constraint, see e.g. Ayer, 1954/1997). For the following, it is condition (ii) that will be seen to be essential.

But now we can, it seems, go further in our conceptual analysis.[7] *How, by what means,* does the brain '*monitor*' (i.e. record, analyze, control, or keep track of) mental activities as ideas, perceptions etc.? In (Vervoort & Blusiewicz, 2020) we propose that this monitor-

[7] Pushing this analysis further is necessary to allow our model to interact more fruitfully with special sciences, including proposing new experiments, notably in neurobiology and computer science (see the last Section).

ing always involves some assumptions, beliefs, worldviews or other cognitive tools (besides perhaps other brain states). For instance, when 'freely' deciding which suit to choose to go to work, I may 'assume' or I may 'use the belief' or 'be in the belief' that suit A is more appropriate for today's work than suit B. Or when deciding of my own free will to help my neighbor, I may base my decision on the worldview or 'theory' that altruism is a source of fulfillment, or on the simple belief that the act will be rewarded by something much more mundane — say a babble plus beer. It seems that the more intellectually demanding the context is, the more elaborate the assumptions I adopt are: in some cases these better be part of a well-grounded theory. For instance, when a spaceship commander decides in an unforeseen situation to freely press the 'full power' button at a precise time, he hopefully does so on the basis of assumptions that take serious calculations and relativistic mechanics into account. (If he does not, and hits the button in a panicky reflex, one hesitates to call his act genuinely free — in agreement with almost all philosophical models.) In the following we use 'theory' in a very broad sense, including high-level theories (ethical, philosophical, sociological, political, physical...) but more generally also assumptions; beliefs systems, including every-day beliefs and hypotheses; bodies of information; worldviews; etc. In (Vervoort & Blusiewicz, 2020) we introduced 'theory*' to define theory in this broad sense; we could equally well have used 'assumptions*' or 'beliefs*'.[8] Now, it seems that one always monitors, analyzes and controls an act, choice or decision *with reference to, or within, a theory**, as illustrated by the above examples. In sum, high-level free acts, when non-trivial intellectual or ethical deliberation is asked, involve *conscious monitoring by means of assumptions, beliefs or theories*. This is the key new ingredient of our model. But this ingredient still needs to be supplemented with following stipulation.

Indeed, the theories* involved in monitoring an act are more or less 'adequate'. One ponders about a choice using beliefs that may be ill-guided or solid, (more or less) rational or irrational; and in non-trivial situations the whole art of making adequate free decisions amounts, according to our model, to using *adequate* assumptions, beliefs or theories. This interpretation comes close to what Kant thought about free will in his *Foundations of the Metaphysics of Morals*. There he states: "a free will and a will under moral laws become one and the same" (Kant, 1786/1983 BA98). For Kant, real free will is, in essence, *will under moral law*. Translated in the language of our model: will in accordance with (monitored, controlled by) adequate moral assumptions / theories*.

[8] Perhaps the most essential reason why we introduce the convention 'theory*' rather than 'assumption*' or 'belief*' is that, when relating our account to computer science, it appears clearly more instrumental to focus on theories. So the concept 'theory*' allows us to interface with computer science in a much more relevant way: see the last Section. Moreover, we believe it is important to make a distinction between what could be termed 'simple' or 'elementary' beliefs (e.g. linked to perceptions), and beliefs that involve diverse or elaborate cognitive elements (or elaborate mental content) such as theories. In cognitively demanding contexts, an agent cannot monitor a choice through the lens of a simple belief, but needs a whole theory —or many coherent beliefs— to assess a situation. Further, we know since Quine and his holism that statements, at least somewhat complicated ones, in any case scientific ones, do usually not get their meaning 'in isolation' but through interconnection with a net of statements (so diverse mental content) — a theory* in our jargon. One might say: for complicated free decisions, an agent may need to integrate several beliefs (constituting a coherent whole).

A relevant feature of the CMT model is that it is in agreement with the idea that free will comes in degrees.[9]As said, within a neurobiological conception of free will there certainly is room for variations in free will, as has been argued by neurobiologists (Brembs, 2011). Indeed, if such a thing as free will exists, then we arguably do not possess it as a foetus, but we *gradually* acquire it with age and cognitive evolution. The variability of free will can also be understood, within a neurobiological perspective, as a direct consequence of the fact that we descended from animals that had, at best, an embryonic form of free will; ours evolved in parallel and gradually with our cognitive capacities. This is well reflected in our model: according to it, *the variability of free will resides, notably, in the degree of adequacy of the assumptions* we use to consciously monitor and guide our actions*; so in their (cognitive and practical) efficiency (Vervoort & Blusiewicz, 2020). It seems clear that low-level animals are devoid of this form of free will; but it is highly likely that in primates some elementary forms of consciousness exist, as an evolutionary base for human consciousness. In our model this embryonic consciousness comes down to a capacity to guide actions through some elementary form of 'beliefs' or 'assumptions'. Therefore we suggest it would be interesting to search for the neuronal correlates for 'assumptions' in primates (one conjectures that they are related to memory-circuits, or to the 'mirror neurons' that have become fashionable lately[10]).

We can wrap-up above considerations in the final definition that summarizes our model (cf. Vervoort & Blusiewicz, 2020):

> **DEF-3**. Action A by animal b is 'free-willed' or 'free' (is made of b's own free will)
>
> IFF
>
> i. the action A is *unconstrained* (no programmed or external compulsion), and
> ii. the action A is *conscious* in that the action (linked to a mental process) is *monitored* (recorded, analyzed, controlled, or kept track of) by some other mental activity in the brain of b. The latter monitoring occurs through theories*, which have a varying degree of adequacy (cf. explanation in text).

For examples how to apply this definition to concrete cases, we refer to the last Section (see notably how we treat the example involving Alice and Bob) and to the original (Vervoort & Blusiewicz, 2020). In the latter article we have shown that our model, which we term the CMT model (for free-willed as 'conscious-through-monitoring-through-theories*') can be related to, and solve problems of, several recent compatibilist theories, notably those of Frankfurt (1969, 1988), Wolf (1990), and Fischer and Ravizza (Fischer and Ravizza, 1998; Fischer *et al.,* 2007). Let us consider one example from (Vervoort & Blusiewicz, 2020). In Frankfurt's 'hierarchical mesh theory' of free will (1969, 1988) an action or choice A is free if it meshes with a 'second-order volition' — a higher desire about the first-order desire to do A. Then A is really (rationally) desired, in agreement with one's second-order (rational)

[9] This idea is not new in the philosophy literature. It is for instance elaborated upon in an interesting manner in (O'Connor, 2009). Here we propose a related but still different account of the origin of the variability of free will.

[10] It is interesting that these mirror neurons are the base for the 'theory of mind' that neurobiologists have attributed to primates as the cognitive base for recognizing the 'self' and 'the other'.

desires; action A flows from the 'will one wants' — a reflective capacity animals likely do not have. As argued in (Vervoort & Blusiewicz, 2020), there surely is a relevant connection between Frankfurt's theory and the CMT model: one may consider that second-order volitions are part of the general beliefs, worldviews etc. that an agent uses to guide and control her life and actions. Action A meshes with a second-order volition in that it is consciously monitored by (assessed, analyzed etc.) with help of a worldview, a belief system, assumptions of life, in other words theories*. So both models converge, but ours is not afflicted for instance with the well-known infinite-regress problem of Frankfurt's; and ours can treat a case as brainwashing, again a threat to Frankfurt's theory (for a recent overview and references, see Griffith, 2013, Ch. 4). A girl Trina brainwashed by her community to believe that stealing is commendable may well act in accordance with higher volitions, really believe in what she does, and therefore be entirely free according to Frankfurt's model — a conclusion most people would disagree with. The CMT model solves this problem: Trina is brainwashed and therefore not unconstrained; and she monitors (assesses) her deeds through questionable, likely inadequate beliefs. In other words, one could say she has a limited or corrupted form of free will (Vervoort & Blusiewicz, 2020).

A last property of the CMT model that deserves remembering is that it is compatible with a deterministic picture of the world. Indeed, *even if* all events, systems and properties of physical or biological or other nature were ultimately governed by iron deterministic laws, leaving no room to chance, free will could still exist — namely as a capacity of agents to act as defined by DEF-3. Indeed, even if we were living in a deterministic universe, an agent can act without being constrained by other agents, and monitor her acts through conscious theories*. Of course, the *acquisition or adoption or application* of beliefs, assumptions etc. will, in a deterministic picture, be the result of predetermined processes, mediated through genetics, education, social background, life-changing encounters etc. — in any particular case by a potentially quasi-infinite number of particular causes (let us term this claim 'source determinism'). But this is clearly not in opposition with the fact that one can evaluate an act through the lens of clauses (i) and (ii) in DEF-3, and judge whether it is free *in that sense*. Note, and this is interesting, that our model can also survive in an indeterministic universe! Also in that case one can evaluate an act according to (i) and (ii). So, our model could perhaps also appeal to libertarians and be used by them (but they would want to add something to it). But the model jibes best with a deterministic stance on free will. Indeed, the deterministic picture is the simplest: it does not need to invoke an 'agent' or 'mind' or other substance that would causally control a person's acts; in the deterministic picture acts are causally controlled by a person's beliefs, assumptions (cf. clause (ii)) — something cognitive—, where this cognitive thing (neural nets ultimately) is itself part of a deterministic causal chain. But these neural nets are still in the brain *of* some agent; agency comes back in in this way.

In any case, a comprehensive analysis of free will demands a commitment to either determinism or indeterminism, as we will argue below. Therefore, to be precise, let us explicitly add the following ingredient to the CMT model as a compatibilist model. We do not simply assume that determinism and free will are not in contradiction (on the logical level) as all compatibilists do; we assume moreover that both *exist* in this world. Many contemporary compatibilist philosophers make no ontological commitments; but we need to do so when interpreting neurobiology and real-world cases. One could call this a form of '*ontic* compatibilism' or 'soft determinism'. We prefer the former, new term to emphasize the

perspectival dimension of this position ('soft determinism' is already used and does not explicitly contain this dimension). Indeed, one can look at problems of free will, agency, moral responsibility etc. through the lens of a theory of free will, but *at the same time* one should not forget, according to our ontic compatibilism, to analyze things through the lens of determinism. This position, intimately linked to (versions of) perspectivism, surely has been advocated in the history of philosophy. In sum, our model of free will, the CMT model, has three essential characteristics, namely DEF-3, featuring notably the assumption that the monitoring (analyzing, controlling) by the conscious brain involves assumptions, beliefs, theories*; the hypothesis that free will comes in degrees (this is actually already explicit in clause (ii) of DEF-3); and ontic compatibilism.

4. Libet's experiment and the CMT model. Conclusions

We have analyzed in this article neurobiological results related to free will as reviewed in (Koch, 2009); more precisely results on the (in)deterministic nature of neuronal properties that can be analyzed in physical terms, namely ionic channel currents, spike frequency, and fly behaviour (cf. Section 2, findings 2.1-2.3). In Section 2 we came to the conclusion that the biophysical processes involved do not provide evidence for genuine (quantum) indeterminacy. This is one of the reasons why we expressed a preference for a compatibilist model of free will, as sketched in the previous Section (one can also invoke philosophical and even physical arguments, cf. Vervoort, 2019). Even if all neuronal activity is based on deterministic (even if in practice unpredictable) processes, as present-day scientific research seems to privilege, the CMT model sketched above is compatible with that reality, and leaves room for (some form of) free will. But as noted, the core of the CMT model (DEF-3) would also survive if free decisions would involve indeterministic processes, after all.

Let us, then, have a closer look at Libet's experiment and its modern variants. In this experiment an EEG scan is made on the scalp of a test person. Such an EEG monitors the so-called 'readiness potential', which is an electrical signal (often called 'brain waves'), expressed in Volts, that results from firing neurons. Every voluntary action, such as the flexing of a hand, induces such a slowly rising electrical potential. In Libet's experiment the test persons are asked to spontaneously lift or flex their hand, whenever they feel like it, and note the time when they make this 'free' decision. The experimental set-up is such that test persons can quite precisely identify this decision time (they look at a screen with a bright point moving on a circle, a clock, and can thus simply report the position of the pointer); the actual time of movement can be measured electrically with high precision; both times coincide quite well. As is well-known, the at-the-time quite spectacular result reported by Libet was that the EEG showed that the brain activity started almost half a second before the time at which the test person consciously decided to move. In Koch's words: "What became apparent was that the beginning of the readiness potential preceded the conscious decision to move by [between] 0.3 and 0.5 sec. That is, the brain acted before the conscious mind did! This is a complete reversal of the deeply held intuition of mental causation — your brain and your body only act after your mind wills it" (2009, p. 46). Experiments like these have been confirmed by others; and one has to add to these findings the even more spectacular experiments by researchers as Soon *et al.* (2008).

Many, and probably most neurobiologists are inclined to believe that these findings undermine the idea that humans have a free will. What is the verdict of the CMT model? The precise question is: is the test persons' act of flexing their wrist 'free' (at the decision time)? Applying our model is straightforward here: at the decision time the test person is not constrained (condition (i) is satisfied) and she is clearly conscious of her act (condition (ii) is satisfied) — hence her act *is* free, after all, according to our model. (*Before the decision time* the act is unconscious and therefore not free; but that seems uncontroversial.) Of course, deciding to flex a hand in a Libet experiment is almost a reflex-like act; therefore the 'assumptions' accompanying this kind of conscious decision are surely minimal and not particularly rationalized.[11] But let us recall that free will comes in degrees; some acts involve a higher level or simply a different kind of consciousness than others; one could therefore say that the type of free will exerted by the test persons in flexing their hand is of a 'minimal' type. Also, the steady rise in readiness potential that is measured in Libet's experiment might be paralleling precisely this rise of awareness, of consciousness that an act is done — in agreement with our model.

Thus, on the CMT model, the test person in the Libet-experiments is free at the moment of conscious choice, even if this conscious and free act may well be determined by previous causes — as the neurobiological data suggest. In other words, Libet-like experiments do not exclude a compatibilist conception of free will. This conclusion is in line with compatibilist ideas[12] — it is on the other hand at odds with what many neuroscientists believe. However, our model allows to go further in the analysis. As will be no surprise, to that end it is instrumental to look at actions that are more complex, cognitively speaking, than the wrist flexing used in Libet's experiment. Let us look at a somewhat subtle test case. Imagine a family that has since generations amassed considerable wealth. Suppose that the two off-springs in the youngest generation of the dynasty, Alice and Bob, are in a bitter fight since years about their heritage. Alice and Bob have been brought up in a family that puts material wealth among the highest goods; they were under the constant influence of ideas, habits, events that expressed this 'worldview'. (Although they surely experienced the idiosyncrasies of this select micro-cosmos, they do not seem to suffer from any obvious, extreme psychopathological disorders.) Suppose further that Bob wants to inherit alone his family's wealth, and decides to shoot his sister (over the years he has come to hate Alice, convinced as he is that she has tricked him out of a considerable part of the heritage; he seems to have convinced himself of the idea that Alice somehow *deserves* her fate). Suppose finally that Bob is also a drug addict, and that at the moment

[11] The 'assumptions' may be related to the assumption/belief that "I am a test person and supposed to flex"; to memories of former wrist flexing; to the memory or belief that the flexing can be done clockwise or counter-clockwise; etc.

[12] For instance Ayer's (1954/1997). Indeed, Libet's test persons are not under external constraint (neither under pathological or compulsory internal constraint); hence free on Ayer's model. Ayer famously argues that free will should not be contrasted with determinism and causality, but with constraint (Ayer, 1954/1997, p 115). See also Griffith's monograph on free will (Griffith, 2013, p. 109): "RP [readiness potential] onset could correlate with any number of things. It could be some sort of precursor to an intention. It could be a cause of an intention rather than an intension itself. This is important because it may not be problematic to think that our intentions have causes (you will recall that most of the free will theories we have discussed allow for a causal chain)".

of killing he is under influence, but still —at least partly— lucid. Is Bob's killing an act of free will?

On a straightforward application of the CMT model (DEF-3), Bob can be said to have (a form of) free will. Indeed, Bob was not compelled by others in his decision making, or let us suppose so (clause (i) is satisfied); and he was guiding his act by some (more or less) conscious thinking (clause (ii)). Now, as argued in Section 3, a richer perspective can be given to this question by realizing that free will comes in degrees, and that the conscious monitoring of free acts is done within assumptions* or theories* having a degree of adequacy. From this perspective, *even without taking his drug use into account*, the free will of Bob may be said to be of a 'corrupted' type — at least if we agree that Bob's belief that "money is all what counts" and that "she deserves it" are not the most adequate theory* to adopt.

But this is not all. Ontic compatibilism, the third main characteristic of our model, incites us to scrutinize Bob's decisions through the lens of determinism. Determinism is not only compatible with the two components of DEF-3, it is also at the source of these ingredients of free will (we termed this 'source determinism' above) — or so suspects the determinist. In Bob's case it seems clear that there exist some drugs-mediated deterministic processes underlying his decision making. Even if one cannot fully analyze why Bob believed what he believed (a great number of factors will have conditioned his beliefs, presumably starting in childhood), it can be assumed that his drug abuse will have influenced (so partially determined) his assumptions and beliefs when deciding to murder his sister. For instance, a euphoric or 'counterfactually confident' person makes different decisions than a normal person. So Bob's drug use further distorted his free will through distorting his beliefs, assumptions and worldview; this is a free-will reducing factor that deterministically underlies condition (ii) in DEF-3. His drugs use could also be seen as a deterministic free-will reducing factor for condition (i), if it leads to compulsory behaviour; and high-grade free will is devoid of inner or programmed compulsion according to clause (i) (cf. e.g. Ayer, 1954/1997).

At the same time —and this surely is the counterintuitive part of our interpretation—, according to the nomological determinism we favour, *Bob had to do what he did* — for each and every one of his acts. When taking all physical data of the universe into account, *sub specie aeternitatis* so to speak, *he had no real choice* (supposing determinism is true) — even if to our and to his subjective minds this seems unbelievable at almost all instances of perceived choice. Similarly, on the interpretation of the neurobiological experiments of Section 2 we favour, there is little room for quantum indeterminacy in cognitive processes, and even less for a mind determining quantum processes. On that interpretation, what neurobiology leaves us with are classic deterministic processes materialized through action potentials of active neurons. These neurobiological processes, causing some free act, are the material counterpart of 'beliefs' or 'reasons' that make us pick a choice, act a free act. In Bob's case these processes were moreover conditioned by drugs. (One could see this as a further argument for determinism: if some decisions are ultimately depending on the deterministic chemistry and physics of the brain — why not all?)

Therefore, to be precise, we should give a detailed answer to the question "had Bob free will?". *All depends on what one calls free will*. If the ultimate capacity of a mind exerting power over the brain is meant, a mind that can thus genuinely choose between alternatives, independently of the causal past, our answer is 'no' — at present we find little or no scien-

tific evidence for such a capacity. But if a CMT-type capacity is meant, a capacity to think about one's acts and thus to guide and control them, we believe the answer is 'yes'. Our analysis of the question of free will hints to a final conclusion that is quite close to what philosophers as Plato, Aristotle, Kant and Spinoza taught us, namely that it is through better understanding the world that we become freer beings.[13] (Regarding the question whether Bob was morally responsible: cf. the Appendix.)

As announced, one essential motivation we had for constructing an analytically minimal model of free will (DEF-3), is that we believe that it can be instrumental for guiding natural science research; it may suggest a number of avenues for research in neuroscience and computer science. Indeed, we conjecture that our model can be instrumental in tackling questions related to consciousness — considered an essential but at the same time highly elusive concept in neuroscience (e.g. Stern, 2017). Notably, our model could possibly conceptualize aspects of consciousness and free will that could have an empirical basis. We think here in the first place of the process of monitoring by a neuronal superstructure (presumably in the prefrontal cortex, or having an integrating 'central unit' there) that should represent a theory*; and the neuronal correlates of embryonic forms of 'beliefs' or 'assumptions*' in primates. We suggest it would be interesting to investigate whether such superstructures are active in decision making and conscious acts; and of which nature they precisely are.

In computer science, in particular AI research, a much debated question is: can future computers and robots be conscious and/or have free will? If possible at all, our model suggests that one of the key properties a computer or an artificial neural net should have to emulate consciousness, or to approximately mimic it, is the capacity to 'use' higher-order theories — and this notably includes the capacity to adequately apply theories to real-world situations and to act accordingly. Some will conclude we are very far from this possibility. This suggests the following line of research in computer science and AI: can machines learn to acquire and use theories*, and which types and how?

Intriguingly, computer scientists and cognitive scientists have recently indeed proposed that mastering theories is a key goal for artificial intelligence (AI). Lake *et al.* (2017) state in the abstract of their highly cited article:

> "We review progress in cognitive science suggesting that truly human-like learning and thinking machines will have to reach beyond current engineering trends in both what they learn and how they learn it. Specifically, we argue that these machines should (1) build causal models of the world that support explanation and understanding, rather than merely solving pattern recognition problems; (2) ground learning in intuitive theories of physics and psychology to support and enrich the knowledge that is learned; and (3) harness compositionality and learning-to-learn to rapidly acquire and generalize knowledge to new tasks and situations."

So, developing the capacity to acquire and use theories is, according to these authors, the most promising path that AI can take to emulate human cognition. Since we come to a converging conclusion (about rational consciousness, arguably strongly related to cogni-

[13] Of course, this research also suggests that concepts as blame, praise, desert, punishment, achievement, pride, self-made man etc. need to be taken with a serious grain of salt, and actually seriously revised — a topic for much more philosophy.

tion) from a very different angle, namely from philosophical research on the ancient question of free will, we submit this as a case of mutual corroboration.

In physics, finally, a compatibilist view on free will allows us to escape from Bell's no-go theorem, as argued in Section 1. This may be important, since (the usual interpretation of) Bell's theorem is an obstacle in the construction of a 'theory of everything' ('t Hooft, 2017). We hope that we have hereby made a case for naturalized philosophy, by showing how philosophy and natural sciences can interact in a bi-directional way.

Appendix. A few words on moral responsibility

Almost all philosophers agree that the concept of moral responsibility is closely related to that of free will, and that free will is necessary for moral responsibility, but the precise link is highly debated. We cannot elaborate in detail here on this link, but it seems that examples as Bob's case force us to envisage following hypothesis: while free will comes in degrees, moral responsibility is rather an all-or-nothing concept. (Sure, there may be a variability and degree attached to moral responsibility too, but, as we will argue now, to a lesser extent than to free will.) From some point of view, moral responsibility is a concept that carries the idea of a *status* that a society of people attributes to each of its members in order that it can function acceptably. On the CMT model, if there is no constraint, free will is largely *only* depending on some capacity of the free-willed agent *herself* (namely to 'CMT'). But one may argue that moral responsibility seems rather like a status given by society. Indeed, in the literature moral responsibility is defined by such concepts as answerability, attributability, accountability (cf. e.g. Watson, 1996; Smith, 2012) — these all have a legalistic touch to them, while DEF-3 is devoid of such concepts. Free will is something one can have on an island alone; the last man to live can have free will; but one hesitates to attribute moral responsibility to the last human in the universe — the last human can in any case not be morally responsible *with regard to* another human. One is morally responsible *with regard to X* (normally: other people, perhaps other living species); free will does not carry this relativity. It thus seems that having moral responsibility normally involves taking society into account; having free will not necessarily, e.g. in morally neutral situations. To come back to our starting point: we believe this is the reason why moral responsibility is rather an all-or-nothing concept, much less a matter of degree than free will: society attributes it to any agent as soon as she has just a little free will (just a little of the capacity to consciously monitor acts). On this view, moral responsibility is attributed by implicit fiat of society, not so differently as it attributes e.g. 'citizenship of nation X' — clearly an all-or-nothing thing. A moderate hypothesis is the following: *that the variability of moral responsibility does not parallel that of free will*. While an agent's free will can in principle vary gradually with time, act-by-act so to speak, an agent is morally responsible for his acts once and for all, as soon as he has a minimum form of free will.

Hence our verdict: Bob in the above case study *is* morally responsible.

REFERENCES

Ayer, A.J. (1997). Freedom and Necessity. In D. Pereboom (Ed.). *Free Will* (pp. 110-118). Indianapolis: Hackett Publishing.

Brembs, B. (2011). Towards a scientific concept of free will as a biological trait: spontaneous actions and decision-making in invertebrates. *Proceedings of the Royal Society B* 278, 930-939.

Dennett, D. C. (2003). *Freedom Evolves*. London: Penguin Press.

Fischer, J. M., Kane, R., Pereboom, D., & Vargas, M. (2007). *Four Views on Free Will*. Hoboken: Wiley Blackwell.

Fischer, J. M. and Ravizza, M. (1998). *Responsibility and Control*. Cambridge: Cambridge University Press.

Frankfurt, H. (1969). Alternate Possibilities and Moral Responsibility. *Journal of Philosophy* 66, 829-39.

Frankfurt, H. (1988). *The Importance of What We Care About*. Cambridge: Cambridge University Press.

Griffith, M. (2013). *Free Will, the Basics*. London: Routledge.

Haggard, P., Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research* 126, 128-133.

Heisenberg, M., Wolf, R. (1984). *Vision in Drosophila: Genetics of Microbehavior*. Berlin: Springer

Jordan, P. (1938). Die Verstärkertheorie der Organismen in ihrem gegenwärtigen Stand. *Naturwissenschaften* 26(33), 537-545.

Kane, R. (2005). *A Contemporary Introduction to Free Will*. New York: Oxford University Press.

Kane, R. (2012). *The Oxford Handbook of Free Will*. New York: Oxford University Press.

Kant, I. (1786/1983). *Grundlegung zur Metaphysik der Sitten [Foundations of the Metaphysics of Morals]*. W. Weischedel (Ed.). Complete works in 10 Vols., Vol. 6. Darmstadt: Wissenschaftliche Buchgesellschaft.

Koch, C. (1999). *Biophysics of Computation: Information Processing in Single Neurons*. New York: Oxford University Press

Koch, C. (2009). Free will, physics, biology, and the brain. In N. Murphy, G. Ellis & T. O'Connor (eds.). *Downward causation and the neurobiology of free will* (pp. 31-52). Berlin and Heidelberg: Springer.

Lake, B., Ullman, T., Tenenbaum, J., Gershman, S. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, Vol. 402017, e253.

Libet, B. (1985). Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action. *The Behavioral and Brain Sciences* 8, 529-566.

Mahner, M., Bunge, M. (1997). *Foundations of Biophilosophy*. Berlin: Springer Verlag.

Maye, A., Hsieh, C.-H., Sugihara, G., Brembs, B. (2007). Order in spontaneous behavior. *PLoS ONE* 2, e443.

Mele, A. (2009). *Effective Intentions: The Power of Conscious Will*. Oxford: Oxford University Press.

O'Connor, T. (2009). Degrees of Freedom. *Philosophical Explorations* 12 (2), 119-125.

Pereboom, D. (Ed.). 1997. *Free Will*. Indianapolis: Hackett Publishing.

Popper, K. R. and Eccles, J. C. (1977). *The Self and Its Brain—An Argument For Interactionism*. Heidelberg: Springer.

Smith, A. M. (2012). Attributability, Answerability, and Accountability: In Defense of a Unified Account. *Ethics* 122, 575-89.

Soon, C.S., Brass, M., Heinze, H.-J., Haynes, J.D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience* 11, 543-545.

Stern, P. (2017). Neuroscience: In Search of New Concepts. *Science* 358 (6362), 464-465.

't Hooft, G. (2017). Free Will in the Theory of Everything. *arXiv:1709.02874 [quant-ph]*.

Vervoort, L. (2013). Bell's Theorem: Two Neglected Solutions. *Foundations of Physics* 43, 769-791.

Vervoort, L. (2019). Probability Theory as a Physical Theory Points to Superdeterminism. *Entropy* 21(9), 848, 1-13.

Vervoort, L., Blusiewicz, T. (2020). The CMT model of free will. *Dialogue, the Canadian Philosophical Review*, cf. doi:10.1017/S0012217320000104

Walter, H. (2001). *Neurophilosophy of Free Will. From Libertarian Illusions to a Concept of Natural Autonomy*. Cambridge: MIT Press.

Watson, G. (1996). Two Faces of Responsibility. *Philosophical Topics* 24, 227-48.

Watson, G. (Ed.). 2003. *Free Will*, 2nd ed. Oxford: Oxford University Press.

Wolf, S. (1990). *Freedom Within Reason*. Oxford: Oxford University Press.

Wuethrich, C. (2011). Can the world be shown to be indeterministic after all? In C. Beisbart and S. Hartmann (eds.). *Probabilities in Physics* (pp. 365-389). Oxford: Oxford University Press.

Louis Vervoort studied philosophy at the University of Montreal, after having studied physics at the University of Marseille and the École Normale Supérieure, Paris. He is now professor at the School of Advanced Studies, University of Tyumen, Russian Federation. His research interests include philosophy of science, naturalized philosophy of mind and metaphysics, and foundations of physics. His most recent articles are: Vervoort & Blusiewicz (2020). The CMT model of free will. *Dialogue: Canadian Philosophical Review*, and Vervoort (2019). Probability Theory as a Physical Theory Points to Superdeterminism. *Entropy* 21(9), 848 (1-13).

Address: School of Advanced Studies, University of Tyumen, Ulitsa Volodarskogo 6, 625003 Tyumen, Russian Federation. Email: l.vervoort@utmn.ru

Tomasz Blusiewicz is an assistant professor of history at the School of Advanced Studies at the University of Tyumen, Russia. His doctoral research, defended at Harvard University in 2017, focused on international relations and economic cooperation in Eurasia in the second half of the 20th century. He investigates, among other topics, the interaction between the discipline of history and the natural sciences, neuroscience in particular. His most recent article on free will is: Vervoort & Blusiewicz (2020). The CMT model of free will. *Dialogue: Canadian Philosophical Review*.

Address: School of Advanced Studies, University of Tyumen, Ulitsa Volodarskogo 6, 625003 Tyumen, Russian Federation. Email: t.blusiewicz@utmn.ru