Refuting the unfolding-argument on the irrelevance of causal structure to consciousness

Marius Usher

Tel-Aviv University, Israel

ARTICLE INFO

Keywords:
Neural dynamics
Neural-networks
Consciousness
Temporal perturbations
Recurrent-networks
Cognitive-control
Unfolding-argument
Universal approximation theorem

ABSTRACT

The unfolding argument (UA) was advanced as a refutation of prominent theories, which posit that phenomenal experience is determined by patterns of neural activation in a recurrent (neural) network (RN) structure. The argument is based on the statement that any input–output function of an RN can be approximated by an “equivalent” feedforward-network (FFN). According to UA, if consciousness depends on causal structure, its presence is unfalsifiable (thus non-scientific), as an equivalent FFN structure is behaviorally indistinguishable with regards to any behavioral test. Here I refute UA by appealing to computational theory and cognitive-neuroscience. I argue that a robust functional equivalence between FFN and RN is not supported by the mathematical work on the Universal Approximator theorem, and is also unlikely to hold, as a conjecture, given data in cognitive neuroscience; I argue that an equivalence of RN and FFN can only apply to static functions between input/output layers and not to the temporal patterns or to the network’s reactions to structural perturbations. Finally, I review data indicating that consciousness has functional characteristics, such as a flexible control of behavior, and that cognitive/brain dynamics reveal interacting top-down and bottom-up processes, which are necessary for the mediation of such control processes.

While cognitive neuroscience has made much progress in accounting for aspects of conscious and unconscious processing and in determining their neural correlates (Block, 2011; Bronfman, Jacobson, & Usher, 2018; Dehaene & Naccache, 2001; Koch, 2012; Lamme, 2006; Noy et al., 2015; Zeki & Bartels, 1998), a genuine understanding of the neural nature of conscious phenomenal experience remains elusive. This challenge is often referred to as the hard-problem of consciousness (Chalmers, 1995a). In the attempt to answer the hard-problem, a number of prominent theories have proposed that conscious experience is an emergent property within a special type of neural processes that involves recurrent processing (RP; lateral as well as interacting top-down and bottom-up; Lamme, 2006). A special type of such theory is the Integrated-Information Theory (IIT). According to IIT, phenomenal experience

E-mail address: marius@post.tau.ac.il.

1 For a discussion of emergent conscious experience, in general, and in IIT in particular, see Chalmers (2017).

2 There is debate on whether a theory, such as IIT can answer the hard-problem (HP; Mindt, 2017). For example, Garrett Mindt has raised the objection that we can conceive of a zombie (physical duplicate of one of us), which is totally devoid of phenomenal experience. On the other hand, Chis-Ciure & Ellia (2021) have provided an analysis of the HP, which supports the position that IIT can answer it. It is not my aim here to argue that theories such as IIT can (or cannot) solve the HP (I refer interested readers to the papers above). Furthermore, Jylkka and Railo (2019) have argued that HP is unsolvable and merely “reflects the difference between a scientific model and the modeled concrete process”. There are some aspects of the IIT, which I do not endorse (for example, its acceptance of a potential dissociation between consciousness and functional organization). However, I believe, that like RP, the IIT is a legitimate scientific hypothesis and that the unfolding argument does not invalidate it.

https://doi.org/10.1016/j.concog.2021.103212

Received 1 June 2020; Received in revised form 15 September 2021; Accepted 16 September 2021
is identical with a maximally irreducible cause-effect structure, which can be realized in a physical structure and can mathematically be described by the integrated information formalism to account for the quality of experience, via a variable $\Phi$, which quantifies the irreducibility of the causal-effect structure and corresponds to the level of consciousness. Critically, $\Phi$ vanishes in FFN and thus according to IIT, FFNs have the wrong causal architecture or connectivity to mediate conscious experience (Oizumi, Albantakis, & Tononi, 2014; Tononi, 2015; Tononi, Boly, Massimini, & Koch, 2016). Thus, according to both the RP and the IIT theories, conscious experience is dependent on the causal (physical) brain structure, in particular on recurrent neural dynamics.

Recently, an influential argument has been advanced against any causal brain structure theory, including RP and IIT (Doerig, Schurger, Hess & Hertzog, 2019). According to the authors, any behavioral test, which a mechanism mediated by a recurrent (neural) network (RN) architecture can pass, can also be passed by a functionally equivalent feed-forward network (FFN) – this equivalent FFN is obtained via a process referred to as ‘unfolding the network’ (see Section 2). Consequently, according to the unfolding-argument (UA), if consciousness depends on the causal brain structure (e.g., on an RN structure), then by assumption (see premise 4 below), its presence is scientifically untestable (due to circularity) and unfalsifiable and thus, unscientific (see premise-5 below). This lead the authors to conclude that “only theories that abstract away implementation details and focus on explaining which kinds of functions are important for consciousness can avoid these challenges. To remain within the realm of science, consciousness must be described in terms of what it does, and not how it does it” (Doerig et al., 2019, p. 56; italics added). This argument has the following structure (see also Tsuchiya, Andrillon, & Haun, 2020):

1. In science we rely on physical measurements (based on subjective reports about consciousness).
2. For any recurrent system with a given input–output function, there exist feed-forward systems with the same input–output function (and visa-versa).
3. Two systems that have identical input–output functions cannot be distinguished by any experiment that relies on a physical measurement (other than a measurement of brain activity itself or of other internal workings of the system).
4. We cannot use measures of brain activity as apriori indicators of consciousness, because the brain basis of consciousness is what we are trying to understand in the first place.
5. Therefore, either causal structure theories are falsified (if they accept that unfolded, feed-forward networks can be conscious), or causal structure theories are outside the realm of scientific inquiry (if they maintain that unfolded feed-forward networks are not conscious despite being empirically indistinguishable from functionally equivalent recurrent networks).

I believe that Doerig et al. (2019) association of scientific theory with testability (premises 1, 4, 5) is too restrictive, however, I do not wish to press this point here; but see Tsuchiya et al. (2020) and Negro (2020) replies to UA for detailed discussions on these points. In this article I take a complementary strategy for refuting UA: I deny premise-2. While, I share Doerig et al. (2019) aspiration to determine functional characteristics of conscious processes, I believe that the dissociation between function and neural structure (premise-2 of the unfolding-argument) is premature. Like Tsuchiya et al. (2020) and Negro (2020), I also believe that the argument relies too heavily on a naive behaviorist baggage (the Turing test) and neglects much of what has been learned in cognitive neuroscience. Since much of the setup for the argument relies on implicit assumptions in philosophy of mind favoring some versions of input–output functionalism (Braddon-Mitchell & Jackson, 2007), I start with a philosophical discussion of previous attempts to deploy Turing-test criteria for the possession of mental states and consciousness (Block, 1981), which, to some extent, follows Negro (2020). I then consider the premise-2 of UA and I argue that the functional equivalence between FFN and RN is not supported by the mathematical work on the Universal Approximator theorem, and is also unlikely, as a conjecture, given data in cognitive neuroscience, showing interacting top-down and bottom-up processes, whose dynamics differ from those of FFNs. Finally, I focus on a central functional property of conscious processes, cognitive-control, which, I argue, to critically depend on recurrent (lateral and on interacting top-down and bottom-up) processing.

1. Turing-type tests fail as a criterion for the possession of mental states: The ‘how’ matters

The flavor of the unfolding argument debate is reminiscent of previous debates between functionalists and logical behaviorists on the nature of mental states (for a textbook discussion see Braddon-Mitchell & Jackson, 2007). At the center of the UA is the behaviorist/verificationist position that any scientific investigation of consciousness should be based on behavioral experimental tests that distinguish between the presence and absence of consciousness, and this can only be carried out by posing to the subject a series of behavioral tests (e.g., “do you feel something right now?”). Consequently, if two organisms answer the same in all such tests (which involve sensory stimuli and motor or verbal responses), they must be equivalent in terms of their conscious experience (at least as far as science can reveal).

In a recent reply to the UA, Tsuchiya et al. (2020) have objected to the exclusively behavioral operationalization of consciousness data, which the argument relies on. According to Tsuchiya et al. (2020), consciousness researchers also rely “on the link - known to

---

3 Another prominent consciousness theory is the Global-Workspace Theory (GWT; Dehaene & Naccache, 2001), which assumes that conscious experience is associated with a frontal-parietal network, which is bilaterally connected to sensory and motor areas. As such, GWT is also a recurrent model, but it is described at a functional level (to make conscious contents globally available to all brain systems and to report) and the recurrence is less of essence then in RP and IIT and Doerig et al. (2019) argue it can be implemented in an FFN architecture. I am skeptic of this assertion, at least under physical brain constraints (see section 3).
each of us by first-hand experience - between conscious experience and behavioral reports” (p. 2). This type of attack on the UA, was extended by Negro, who showed that the argument also fails with regards to premise-5 (this requires a more careful examination of falsibility in philosophy of science theories) and with regards to premise-4 (see below). However, while I believe this is a justified approach, it is important to remember that behavioral criteria of mental states and processes, was what helped the functionalists to avoid the danger of a type of brain-chauvinism, which faced the rival identity-theory approach, according to which we would exclude organisms, whose brain are different from ours (bats, octopuses or potential intelligent aliens), as possessors of mental states. Taking this argument from the broader characterization of mental states to consciousness, may lead us to conclude that we can accept (consciousness) in other humans (but not in aliens; although in both cases we have no direct access to their mental life; with humans (but not with aliens) we can reason that, as our brains are similar enough to ours, if consciousness depends on causal structure, they are likely to be conscious too.

The functionalist proposal was that mental states could be multiply realized (in aliens or robots) so that insisting on a common brain structure would be too restrictive (for having mental states, like beliefs or desires), and thus a behavioral functional organization for the possession of mental states was proposed as more liberally adequate (Putnam, 1967). It is possible that this strategy becomes too ‘liberal’ in the case of consciousness (it may deem functional analogs of us, such as Searle’s Chinese room robot (Searle, 1980), or Block’s China brain (1978), as subjects of phenomenal experience. Whereas, I believe that functionalism is a productive project in cognition, I think it faces difficulties in addressing the phenomenal experience5. I think, however, that there is a strong motivation for accepting a functional characterization of consciousness (additional to the characterization based on causal structure). Thus, while I tend to accept that the causal/physical structure is relevant for consciousness (more than it is for beliefs/desires/intentions), I propose that the ideal situation (for a satisfactory explanation) is to have convergence of the functional and causal-structure characterization of consciousness. Because of this, I suggest that a more fruitful way to defend causal-structure theories of consciousness is to refute the premise-2 of the argument: the assertion of a robust functional equivalence between organisms whose brains differ drastically in their causal structure (FFN/RN). As I will illustrate in Section 2, this is a plausible position to take, given the fact the FFN and RN have fundamentally different dynamical properties, and I assert that – systems with fundamentally different dynamical properties cannot be functionally equivalent.

Ned Block (1981) has already made the case that organisms who are behaviorally identical (in answering queries in a Turing type test, which can include experimental tests of consciousness), are not necessarily identical in their possession of mental states. This objection was directed to the possession of mental capacities (such as intelligence and understanding), but its structure can be applied to the possession of consciousness. This argument was also the starting point in the discussion of the unfolding-argument by Negro (2020), who I will follow to some extent here.

In his article, known colloquially as the “Blockhead argument” (Block, 1981), Ned Block makes an important distinction between the possession of actual (and even potential) behavioral characteristics of mental/intelligent systems, and the actual possession of mentality or intelligence.

“Let psychologism be the doctrine that whether behavior is intelligent behavior depends on the character of the internal information processing that produces it. More specifically, I mean psychologism to involve the doctrine that two systems could have actual and potential behavior typical of familiar intelligent beings, that the two systems could be exactly alike in their actual and potential behavior, and in their behavioral dispositions and capacities and counterfactual behavioral properties (i.e., what behaviors, behavioral dispositions, and behavioral capacities they would have exhibited had their stimuli differed) – the two systems could be alike in all these ways, yet there could be a difference in the information processing that mediates their stimuli and responses that determines that one is not at all intelligent while the other is fully intelligent” (1981, p. 1).

To support psychologism, Block imagines a machine (which we may label Blockhead) that responds to inputs (say, questions posed to it in a Turing test) by examining a very large lookup table, which contains the outputs that would be provided in the same situation by an intelligent organism (say, a specific person) and follows them. As Block observes, while the original human has intelligence (and mental life), as her behavior is mediated by information processing (i.e., memory search, inferences, decision-making based on goals, etc.), the Blockhead machine “has the intelligence of a toaster”. All it knows to do is to follow rules. Thus at least, in principle, behavioral equivalence does not guarantee equivalence with regards to intelligence; while the person experience involves mental

---

4 There is debate on the validity of "multiple-realization" based on cases such as bats or octopuses (see e.g., Polger and Shapiro (2016); reply by Cough, 2018 and further review in Bickle (2020). For example, it is possible that different realizations of a function are not relevant to it, or that mental states of octopuses have a biochemical type that is identical to that of humans and this grounds mental states like beliefs, desires or pains. This strategy would still exclude potential sentient aliens or robots as bearers of mentality. Alternatively, it is possible that the RN and the FFN systems are not truly equivalent. This is the strategy we take here to answer the UA.

5 There are a wide series of arguments such as, the 'Inverted-Spectrum' (Shoemaker, 1982) or the 'Inverted-Earth' (Block, 1990), which articulate dissociations between functional and phenomenal content. Most of these dissociations are related to the fact that while functionalism is a relational theory, phenomenal experience is intrinsic rather than relational (see Braddon-Mitchell & Jackson, 2007). It is likely, therefore that a brain-identity approach could be more relevant for grounding phenomenal experience (Shoemaker, 1982).

6 See for example, Chalmers, ’Absent/Fading/Dancing’ qualia argument (Chalmers, 1995b) for a defence of the principle of organizational invariance, according to which organisms that have the same organization (functional characterization) of conscious experiences share the same experiences. One can deny the functional-organizational invariance (in favour of a more strict causal invariance) principle and accept that we could be subject to epistemic uncertainty about our experiences in the Chalmers scenarios, however, this principle also allows a simple answer to the textbook challenge of why would consciousness evolve if it carried no behavioral benefit (see Braddon-Mitchell & Jackson, 2007), as it allows for the position that there are specific functional characteristics of consciousness (see Section-4).
process, such as memory search and inferences, the Blockhead-machine experience is drastically different and possibly lacking altogether (Blockhead is not endowed with a proper internal causal organization, rooted in a history of causal dependencies, that constitute the information processing of real thinkers).

Note that an important aspect of Block’s argument is the asymmetry between the person and the lookup table machine. While the former processes information, the latter does not (in the sense described above) and this is what accounts for our intuition of a difference in intelligence and in understanding. In his discussion of this paper, Negro (2020) suggests that this assertion is not circular (as the unfolding argument would hold; premise-4 on internal structure involving circularity), but rather this is a pre-theoretical intuition of what intelligence is not: “This entails the possibility of having a grasp of a notion, without having a theoretical definition of that notion: what drives the Blockhead argument is thus essentially a pre-theoretical intuition” (Negro, 2020, p. 985). Negro concludes that “The argument is therefore not circular, but grounded on a pre-theoretical foundation made of common-sense intuitions and widely shared beliefs.” Moreover, the structure of the Blockhead argument shares many features with the UA against IIT. In the UA we are also asked to consider of a functional duplicate of a person (obtained by replacing her RNs with functionally equivalent FFNs), and we are asked to conclude that they cannot differ in a mental property (now consciousness, instead of intelligence). The fact that this argument failed with Blockhead for mental states, should provide us a warning, it may fail again. While here we do not have such a clear pre-theoretical intuition as with Blockhead (the difference between the RN and the FFN duplicates does not evoke the same intuitive reaction as the one between Blockhead and the real person), one can argue that the association between RN and consciousness follows from the IIT “axiomatic basis that are derived from a phenomenological basis of the essential features of conscious experience, and thus similar to Blockhead, the difference between a conscious RN person and her FFN duplicate is grounded on a pre-theoretical phenomenological investigation” (Negro, 2020, p. 985). In other words, if IIT is uniquely determined by properties of phenomenal experience that are evident in 1st person introspection, the theory is grounded on pre-theoretical intuitions). Negro (2020) concludes that this is also not circular, refuting premise-4 of UA.

While I find this defense of IIT from the UA appealing, it can be subject to a number of objections. As discussed by Negro (2020), it has been argued that “IIT is undetermined, in the sense that even universally accepted axioms could be translated into different postulates, leading to different versions of IIT (Bayne, 2018; Barrett & Mediano, 2019). In turn, different versions of the theory could bring about different results. How should we choose which version is the correct one? Let us assume that one version attributes high \( \Phi \text{Max} \) to a system, while another version attributes low \( \Phi \text{Max} \). If behavioral functions cannot be employed to establish which version is correct, it seems we can appeal to aprioristic considerations only, and this could be seen as a form of circularity, or question-begging” (Negro, 2020, p. 986). Negro is optimistic that the IIT project can overcome these problems without having to rely on circular assumptions, and I do not wish to dispute this assessment. However, I believe it would be helpful to refute UA by attacking its central assumption of functional equivalence between RN and FFN in premise-2. Demasking the overstatement of equivalence contained in the second premise can benefit the field of consciousness study by making researchers wary of quick solutions to such long-standing conundrums.

In the next section I examine the assumptions of the UA in relation to the functional equivalence between FFN and RN architectures (premise-2). Following, I examine data from neuroscience which indicates strong signatures of recurrent neural dynamics, and I focus on a functional property, cognitive control, which I argue to depend on recurrent (interacting, top-down and bottom up) processing.

2. Are RN and FFN functional equivalent in a robust manner?

The heart of UA is a mathematical theorem on Universal function approximators, which Doerig et al. (2019), explain to mean that “any input–output function can be approximated to any degree of accuracy.” They then cite research to have shown that “RN, as well as FFN are both universal approximator functions (Hornik, Stinchcombe, & White, 1989; Schäfer & Zimmermann, 2006)” and they conclude that both RN and FFN “can be used to generate any desired input–output function to any degree of accuracy using a finite number of neurons”. Therefore, they conclude that “for any recurrent network with a given input output behavior, there are corresponding feedforward networks with the same characteristics (although feedforward networks often need many more neurons than their recurrent counterparts).” (p. 51). Premise-2 of the unfolding argument builds on this to conclude that “Anything that can be done by recurrent networks can also be done in a feedforward manner”, since “for any given input–output function we can find both feedforward and recurrent networks that realize the same function in different ways” (p, 52).

While I do not contest the mathematical Universal Approximator theorem (UAT), I believe that the way the UAT was used in the unfolding argument is not precise and that the conclusions drawn from it have been expanded beyond their validity domain and thus the UAT does not secure a robust-equivalence between RN and their “functional unfolded” FFN. However, even if premise-2 has not been proved, it may be advanced as a conjecture. To provide a proof that this conjecture is false is beyond the scope of this paper. In order to not make this paper too technical I focus below on qualitative aspects that indicate why I believe this conjecture to be false (but see some technical details here\(^\text{7}\)).

\(^\text{7}\) While the UAT studies cited, show that both FFN and RN are universal approximators, they are approximators for different entities. FFN are approximators for Borel-measurable functions on a compact domain (Hornik et al., 1989), whereas, RN are approximators for open dynamical systems (Schäfer & Zimmermann, 2006). From the fact that FN and RN are approximator of different entities it does not follow that they approximate each other. Moreover, Schäfer and Zimmermann (2006) are discussing and citing research showing that RNs compared to FFNs have several computational advantages (REF-4 in Schäfer & Zimmermann, 2006). In the text below, I advance some arguments for why FFN are not likely to robustly approximate RN.
It is important to note first, that while the notion of input and output layers are clearly distinct for FFNs they are less so for RNs, in which most units are both inputs and outputs. However, we can stipulate that the RN input layer would correspond to units that receive sensory stimulation, while the output units are those that correspond to vocal or motor responses (even if they provide feedback (thus input) to other units). There are two important considerations, critical to the argument. The first one is that while FFNs can formally be characterized as functions (from input to output layers) and RN can approximate functions (from input to output layers), RN cannot formally be characterized as functions of external inputs. What this means is that (even in the absence of noise or stochastic processes) the activations RN units are not expressible as a function of the inputs the networks receives. This is a well-known phenomenon for dynamical interacting (i.e., recurrent) systems (not only neural) called “hysteresis”. I illustrate this in Fig. 1, for a magnetic system. In a magnet, the magnetization variable, $M$, depends on the interaction between the magnetic spins of individual atoms in a grid (the recurrent component), which are also affected by an external electric field (the input). While the input (the external electric field) affects the magnetization (the magnet response), there is no function that can uniquely express this dependency. This is because the magnetization (D, in Fig. 1) is not only a function of the input (E) but also of the state of the magnet itself, which is subject to history (see Fig. 1).

Thus, strictly speaking, as the activations in an RN are not expressible as functions of external inputs (the Turing test queries), the UAT cannot approximate them via an FFN, which are function approximators based on external inputs applied to it. However, one may try approximate the activation in a specific layer of a RN as a more complex function of the inputs from all previous times (this brings in the history, as in the hysteresis illustration). While this may be achievable under some circumstances (especially if the input dominates over the recurrence), such approximations will fail under more general conditions, for example, when the recurrent input dominates (as seems to be the case in the neural system; Douglas & Martin, 2007). To consider an extreme case, we can have self-sustained dynamical states in an RN, based on attractors or oscillatory (limit cycle) patterns (Amit, 1989), even in the total absence of any external input (as in the case in REM dreaming); obviously, approximating the activity of the RN (from its null input) must fail.

One may still try to generate an FFN approximation of an RN under more limited conditions, which correspond to a subject tested in lab-experiment (the type of Turing test that the unfolding arguments relies on). Let’s further assume that during the test, the sensory input dominates (or at least that there are no intervals the subject is left without such inputs (i.e., day-dreaming states are excluded). Another problem here is that the FFN version of the UAT (Hornik et al., 1989) applies to static maps, from a state in the input space to a state of the output space. There are indeed, many cognitive functions, such as, classification problems (e.g., object recognition from visual input), in which the same static classification can be generated either by RN or by FFN networks. Indeed, recent work with deep convolution networks (which are feedforward; LeCun, Bengio, & Hinton, 2015) have shown that such deep FFNs are able to generate very powerful classifications functions, somewhat similar to those achieved by humans (but see Van Bergen & Kriegeskorte, 2020 for an insightful comparison of the efficiency of FFN and RN based on recent research). The UAT, however, does not imply that this functional equivalence extends from static maps to full dynamical trajectories (i.e., to temporally changing inputs and outputs) or to arbitrary perturbations (dynamic or structural), which can be applied to such a networks (as needed for a robust equivalence). I expand on these two cases below.

Before doing so, I want to clarify a potential confusion on the type of network architectures I believe that this discussion should focus on.

### 2.1. Network constraints

Although the neocortex has a fundamentally recurrent connectivity (Douglas & Martin, 2007), there is a trivial sense in which any RN is functionally equivalent to a FFN in a perfect way (so there is no need for an approximator, at least if we assume discrete time steps). The procedure to derive this type of equivalence is often called, ‘unrolling the network’ (from a time to a space dimension), which is used as a mathematical trick in RN learning algorithms (Van Bergen & Kriegeskorte, 2020). As illustrated in Fig. 2, this involves replicating the whole RN for each time step (that is small enough to approximate the neural dynamics in numerical-integration) and remapping the connections from within a network to connections between the relevant networks at successive steps.

Note that in order to maintain the functional equivalence the inputs and outputs also need to be dynamically shifted. For example, while in the original RN, the input $x(t)$ is provided on a fixed input layer, and the motor output $h(t)$ are received also from a fixed output layer, in the unrolled network, inputs and outputs need to be shifted to new layers at every time-step. If the time-step is small enough, such an unrolled network is an FFN-equivalent of the original RN, because each state of activation (at time-step $t$) is precisely replicated on the $t$-layer of the unrolled FFN. If we have a task (say, object recognition), which is carried out by a person within a short time (~1 sec), the unrolled network, could involve about 100 layers to produce the state of the RN at every 10-ms steps (1000 layers would be needed to improve precision to 1-ms steps). In FFN terms this would require a FF-depth of 1000 (or 1000).

---

8 Note that the time step needs to be small for the approximation to work. This is because we approximated a process that takes place in continuous time in the RN, to a sequence of discrete layers. In principle, there will always remain some discrepancy to the discretization (unless we believe that time is discrete too). There are some modern physics theories, such as quantum-gravity, which positis that time (and also space) are quantized, and thus subject to discrete steps. This should bring little comfort to the UA proponents, as the quantum gravity time step is of the order of Plank-scale, which is extremely small (order of $10^{-43}$ s)

9 Such a FFN unfolding of a RN is discussed by proponents of the IIT theory (Oizumi et al., 2014; Fig 22) and cited in the UA (Doerig et al., 2019). While Oizumi et al., 2014, who built a 5 (binary) units x 5 (time steps) FFN-equivalent to a RN, see this as a dissociation between function and consciousness. Doerig et al, see this distinction as untenable and unscientific.
In the context of the consciousness debate, this depth needs to be much larger. First, we are interested in the functional equivalence on more complex tasks that span longer intervals (hours/days corresponding to a Turing type test). Second, if the RN includes chaotic dynamics (Hendin, Horn, & Usher, 1991; Landau & Sompolinsky, 2018; Skarda & Freeman, 1987), for any finite time-step the approximation will break down, because any deviation (see footnote-8) in the starting points grows exponentially (https://en.wikipedia.org/wiki/Chaos_theory). Thus, the depth of the FFN needed to unroll the recurrent network of a conscious person, who is being examined in a behavioral consciousness test, is obviously not realistic in terms of allowing it to operate under constraints that brains operate in (see discussion in Van Bergen & Krigeskorte, 2020). Even if we neglect such realism constraints, I believe there are philosophical arguments to rule out such deep networks (unrolled networks of a full brain) as distinct from the original RN brain. In the following, I adopt the following realism-constrain. I only consider FFN, which are not unrolled networks of an RN, i.e., I consider networks that have a reasonable depth (like those used in deep-FFN today), and which may approximate an RN via the universal approximator theorem. In the next two sections, I will argue that for such networks, the functional equivalence (achieved via universal approximator) does not extend from static maps to full dynamic trajectories or to arbitrary perturbations (dynamic or structural), which are needed to establish robust equivalence.

Note that in order to apply the UAT to approximate and RN via FFN we need to designate what are the inputs and outputs that the FFN will have to approximate. The issue is not only to pick up input and output units, but also a time-delay. One could, for example, find a FFN that approximate the function that connects input into the network (with some history) to the outputs at some time-delay.

For lack of space I will not be able to expand on this here, but here is a sketch. For such a network (a recurrent brain that is replicated for each time step, as in Fig. 2), one can reinterpret the situation as one involving a regular (recurrent) brain that moves to the next brain replica location in the unrolled FFN as each time step (remember the inputs are also shifted with time). We can then ask “what is the fact of the matter that would distinguish the FFN in Fig. 2, from an RN that moves rightwards (jumping by one layer at each time step)”. If the unrolled FFN has consciousness the RN theorist can then interpret this as residing from the moving (rolled) RNN. Thus only an equivalence between RN and FN that cannot be rolled back into a simple moving RNN can threaten causal structure theories that require recurrent neural structure.

Fig. 1. **Hysteresis in a magnetic system.** Electric displacement field $D$ of a ferroelectric material, as the electric field $E$ (external input) is first decreased, then increased. The curves form a hysteresis loop. For example, when the electric field ($E$) is zero, the $D$ variable can be either positive or negative depending on the history: when starting with a positive magnetization, one needs to apply a negative electric field ($E_C$) to bring the magnetization to zero, while after the magnet has a negative magnetization, the electric field needed to bring the magnetization to zero is positive. There is thus no unique function that can express the relation of the magnetization (output) and the electric field (the input). From (https://en.wikipedia.org/wiki/Hysteresis).

Fig. 2. Unfolding a recurrent network by replicating for each (digital) time-step. Within each A-box, one may insert the corresponding RN (this could involve a simple RN like those in Section 2.3, but they could also involve the whole brain). Connections $W_{ij}$ are re-mapped so that if $X_i$ connected to $X_j$ via a weight $W_{ij}$ in the original RN ($A_0$), now $W_{ij}$ connects $X_i$ in $A_0$ to $X_j$ in $A_1$, and so on.

In the context of the consciousness debate, this depth needs to be much larger. First, we are interested in the functional equivalence on more complex tasks that span longer intervals (hours/days corresponding to a Turing type test). Second, if the RN includes chaotic dynamics (Hendin, Horn, & Usher, 1991; Landau & Sompolinsky, 2018; Skarda & Freeman, 1987), for any finite time-step the approximation will break down, because any deviation (see footnote-8) in the starting points grows exponentially (https://en.wikipedia.org/wiki/Chaos_theory). Thus, the depth of the FFN needed to unroll the recurrent network of a conscious person, who is being examined in a behavioral consciousness test, is obviously not realistic in terms of allowing it to operate under constraints that brains operate in (see discussion in Van Bergen & Krigeskorte, 2020). Even if we neglect such realism constraints, I believe there are philosophical arguments to rule out such deep networks (unrolled networks of a full brain) as distinct from the original RN brain. In the following, I adopt the following realism-constrain. I only consider FFN, which are not unrolled networks of an RN, i.e., I consider networks that have a reasonable depth (like those used in deep-FFN today), and which may approximate an RN via the universal approximator theorem. In the next two sections, I will argue that for such networks, the functional equivalence (achieved via universal approximator) does not extend from static maps to full dynamic trajectories or to arbitrary perturbations (dynamic or structural), which are needed to establish robust equivalence.

Note that in order to apply the UAT to approximate and RN via FFN we need to designate what are the inputs and outputs that the FFN will have to approximate. The issue is not only to pick up input and output units, but also a time-delay. One could, for example, find a FFN that approximate the function that connects input into the network (with some history) to the outputs at some time-delay.
2.2. Dynamics, transients and perturbations

The FFN version of the UAT applies to static maps from input to output spaces. While those are of interest to many practical applications, in neural and cognitive systems, the temporal dynamics are critical. First, neural systems exhibit, self-generated, dynamic activity patterns, that are not generated by any external input, like in REM-sleep, and recent research on the default network indicate that this type of self-generated activation is the dominant neural process in the brain (Buckner, Andrews Hanna, & Schacter, 2008; Raichle et al., 2001). Moreover, the default network activation and spontaneous resting-state brain fluctuations have been implicated in subtle cognitive functions, such as creativity (Beaty et al., 2014; Broday-Dvir & Malach, 2020) and have been reported to reflect personality characteristics and cognitive traits of individuals (Harmelech & Malach, 2013). Unlike, an FFN, which only reacts to inputs, the neural process of humans and animals is dominated by self-generated activation, which only interacts with the sensory input (Douglas & Martin, 2007; Lamme, 2006). I believe that this alone, should make a faithful imitation of the behavior of a normal person (possessing an RN causal structure) by an FFN-equivalent, even for a short temporal interval of few hours, quite unlikely.

Second, even ignoring self-generated activation, the responses of FFN and of RN typically differ in their dynamic trajectory, even when the static (or asymptotic) input-output is the same (i.e., in the conditions for which the Universal Approximation theorem applies). To illustrate this, we can examine two simple neural models that were proposed to account for behavioral and neural data in perceptual choice. I consider here to the simplest, and paradigmatic classification task, in which a subject has to respond (left/right) to noisy evidence shown to her (dots moving left/right with some probability; Shadlen & Newsome, 2001). The two choice models have been widely discussed in the literature: i) feed-forward inhibition (FFI; Mazurek, Roitman, Ditterich, & Shadlen, 2003, and ii) lateral (recurrent) inhibition, as implemented in the Leaky-Competing-Accumulator (LCA, Usher & McClelland, 2001; Teodorescu & Usher, 2013; see also the attractor model (Wang, 2009), which has similar characteristics).

In both cases, the networks carry out the same map from the input space – motion vectors (I_L, I_R) – to a binary decision variable (L, R). The map is a simple, one: L = 1, R = 0, if I_L < I_R, and L = 0, R = 1, if I_L > I_R. In other words, those networks generate functionally equivalent static maps between inputs and outputs. As shown in Fig. 3, however, the functional equivalence is limited to the static (asymptotic) behavior. As soon as we examine the activation trajectories (i.e., the transients), we observe profound differences. While the activation trajectories of the FFN are monotonic, those of the RN are not (red upward arrow in Fig. 3): when the input is applied, both response units show an initial increase in activation, but while the unit that receives the higher input keeps growing up in its activation, the unit receiving weaker input is suppressed to zero. Such non-monotonic trajectories are a signature of RN, as illustrated bellow (Fig. 4) in the classical Interactive-Activation model (McClelland & Rumelhart, 1981; see response of the R-unit, in Fig. 4; see red leftward arrow in the right panel), which accounts for context effects in word/letter recognition. The computational strength of such recurrent networks is that they allow incompatible hypotheses to compete in the representation, while (at the same time) allowing compatible hypotheses to strengthen each other; this makes the ‘work’-unit win the competition and top-down help the disambiguation of the ‘k’-hypothesis (Fig. 4).

Critically, although the static map of inputs to outputs for the FFI and the LCA are identical, their diverging transients (Fig. 3b,d) have important consequences on the classification, when the inputs are temporally perturbed, for example, by shifting the input pulses to L and to R with a small relative delay, as indicated by their different temporal-weighing (see Bronfman, Brezis, & Usher, 2016; Tsetsos, Gao, McClelland, & Usher, 2012). Furthermore, other recent work has shown that due to their dynamic properties, RN networks can carry out processing of temporal information, in a way that exploits laws of visual (or lexical) inference to optimally combine previous with present observations in order to estimate a current state, for example via recurrent network implementation of Kalman filters (Deneve, Duhamel, & Pouget, 2007). While a feedforward approximation (of Kalman filters) can also be generated, they will have a finite temporal window of memory, while recurrent networks can integrate information over arbitrary periods (Van Bergen & Krigeskorte, 2020).

2.3. Structural perturbations

Even if we could generate an FFN approximator function, which is an equivalent-RN, in the sense that it produces the same (classification) output for any set of inputs given to it, it is unlikely that this will be robust to structural (lesion) or physiological (TMS) perturbations (such as eliminating 5% of the units or applying a magnetic pulse to a targeted area). There is no reason to expect that the

I chose lateral inhibition as the simplest form of recurrent connectivity. Other types of recurrence (e.g., top down connectivity, will only increase the dynamic differences)
Fig. 3. Illustrations of the drift–diffusion and LCA models and their dynamics. a) The drift–diffusion model (DDM), implemented as two accumulators with feed-forward inhibition, an upper absorbing boundary and a zero activation reflecting boundary; b) The LCA model with two accumulators, mutual-inhibition, leak and zero-activation reflective boundary; c,d) representative example activations trajectories (with Gaussian noise) for the two models (reproduced from Bronfman et al., 2016). The red-arrow points to a critical difference in the dynamic trajectories of the two networks, which have the same asymptotic outputs. Note that in the LCA, the activation of the loosing unit is not monotonic. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 4. Illustration of the interactive-activation model of word-perception (McCelland & Rumelhart, 1981). Left: model architecture involves lateral, bottom-up and top-down connections. In particular, all connections between layers are bi-directional, so while word-units receive bottom-up input from letter-units they also project top-down input to letter units, allowing them to disambiguate evidence based on context. Right: activation of word and letter units (for the last letter location) in response to a corrupted input pattern (“work”). Initially, several letter units are activated (K, R) are initially activated in response to the missing sensory input. As a result of bottom up (recurrent) input from the letter layer, the unit K wins the competition, while the unit-R is suppressed (red-arrow). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
effect of such perturbation will be the same when applied on the FFN approximator. After all, as we are not speaking of the (totally equivalent) unrolled-RN. Rather, the FFN approximator achieves the map in a different way, and thus the effect of the lesion will have different outcomes. Note that lesion methodologies are central to modern accounts of cognitive functions, like Connectionism, which not only aspire to account for actual behavior but for its pattern of degradation, upon structural perturbations (McClelland, McNaughton, O’Reilly, 1995; Patterson et al., 2006). Moreover, there is nothing circular, with testing the functional behavior of a network under additional constraints (drugs or TMS protocols). If a FFN responds like the RN only under control conditions but not under drugs or TMS stimulation, then the two networks are not robustly equivalent.

The reason why the unfolding argument is not sensitive to such perturbations is its restrictive reliance on a purely behavioral test (as specified by inputs and outputs). Obviously, such tests are not exhaustive of scientific experiments, which can include direct intervention on the network itself by the administration of drugs, alcohol or of TMS protocols. The latter are particularly relevant to the debate, since targeted consciousness research has shown that carefully timed interventions, which are specifically timed to disrupt the arrival of re-entrant information into a visual area, disrupt conscious perception (Di Lollo, Enns, & Rensink, 2000; Fahrenfort, Scholte, & Lamme, 2007; Lamme, Zipser & Spekreijse, 2000). Thus once we expand the definition of ‘input’ to include electrical/psychoactive inputs, the outputs will be different in the two networks. Note that the Universal Approximator theorem assumes a separation between input and output units, and therefore a move in which all units can be either inputs or outputs conflicts with its premises.

To conclude, I have argued that both temporal and structural/physiological perturbations, provide efficient and scientific ways to functionally distinguish between what may appear as functionally equivalent (RN and FNN) networks, at the level of static maps. While I do not consider those cases proofs of non-existence of an equivalence, given the doubts I stated over the validity of the UAT to establish an FFN equivalence for RN with dynamical trajectories, I challenge the UA proponents to show us a FFN equivalent for any of them. In the next section I will focus on what I believe to be the most central functional characteristic of consciousness: cognitive control.

3. A function of conscious processing: Flexible cognitive-control

I join Doerig et al. (2019) in their aspiration to describe consciousness in terms of what it does and not only of how it does it. This suggests that there is a function that conscious processes achieve, which cannot be achieved by unconscious processes. This is a hotly debated issue, with various cognitive functions proposed to uniquely characterize consciousness and with some cognitive scientists arguing that all cognitive functions can be also achieved by unconscious processes (Goldstein & Hassin, 2017). In this section I will review evidence for one functional dissociation between conscious and unconscious processes: cognitive control. Focusing on such dissociation may provide us some cues of the type of brain processes, in which conscious is typically involved.

I believe that there is evidence that consciousness is necessary for certain types of flexible control operations (DeLange, Van Gaal, Lamme & Dehaene, 2011; Jack & Shallice, 2001; Umiltà, 1988; see van Gaal, deLange, & Cohen, 2012; Kunde, Reuss, & Kiesel, 2012, for recent reviews)12. For example, conscious processes are often distinguished from automatic processes, which can be performed on autopilot. A simple way to illustrate this is by considering a common life situation, like walking home, after having moved apartment to a neighboring street. In such a case, one can suddenly find her/himself walking to the old apartment (the routine and automated action), until the awareness of the mismatch (between the action and current goal) is detected, and the behavior is controlled towards the goal-relevant action (change direction to the new home). The idea that attentional control is mediated by recurrent, top-down connections is consistent with much research in cognitive neuroscience (Corbetta, Patel & Shulman, 2008; Desimone & Duncan, 1995; Maunsell & Treue, 2006). In this section I review a computational framework, which illustrates the essential role of recurrent (top-down) connectivity in control processes, and then data that supports the statement that consciousness is necessary for some types of control to be exerted.

3.1. Computational mechanisms of cognitive control

Control tasks have been intensively studied in cognitive psychology, using a variety of laboratory tasks, such as the Stroop, the flanker, or the anti-saccade tasks (Eriksen & Eriksen, 1974; Hallett, 1978; MacLeod, 1992). To illustrate how control is achieved in computational cognitive models, I start by considering the computational framework that was developed by Jonathan Cohen and colleagues (Botvinick et al., 2001; Cohen, Dunbar, & McClelland (1990); Cohen, Servan-Schreiber, & McClelland, 1992).

The idea is that to achieve control one needs to apply top-down modulation (from frontal areas that encode goals) to feed-forward pathways from sensory to output areas, in order to bias these automatized pathways, when their (bottom-up) output diverges from the task relevant one (like when walking to the old apartment). In the Stroop task, for example, one needs, when presented with an incongruent stimuli (word ‘red’ written in blue), to respond with its color (blue) despite the automatic tendency to respond ‘red’ (reading is automatized for adult persons). A second type of control involves trial-by-trial (or block-frequency) adaptation in the degree of top-down control deployed (Gratton, Coles, & Donchin, 1992; Tzelgov, Henik, & Berger, 1992). For example, after experiencing conflict in a previous trial, people increase the control they deploy (and reduce interference) in the next trial.

This is illustrated in the Botvinick et al. (2001), model at two levels (Fig. 5). First, we have the top-down projection from the task-demand units (assumed to be located in the dorsolateral PFC) towards the sensory layers, in order to allow a correct response (blue) in

---

12 There are other functions that have been uniquely associated with Consciousness, such as “the integration of information: the ability to combine different signals into a coherent, unified one” (Mudrik, Faire, & Koch, 2014; Hirschhorn et al., 2020). The discussion here is focused on flexible control, but I do not exclude additional unique functions of consciousness.
the case above. Note also that the task demand units are connected in a recurrent loop with sensory layers; see Kalantronoff, Davelaar, Henik, Goldfarb, and Usher (2017), for specific behavioral consequences that involve the phenomenon of reverse-facilitation in the Stroop task, which is accounted for by these recurrent connections). Second, we have a second loop of conflict modulation (top unit), which goes from the response-conflict to the task-demand units (this upper loop is assumed to be mediated by the ACC (anterior cingulate cortex) brain area, and which accounts for trial-by-trial (or block) modulations in the amount of control deployed.

Control models of this type (based on RNs), have been highly successful in accounting for much behavioral and neural data in a variety of studies (Braver, 2012; Gratton et al., 1992; Kalanthroff & Henik, 2013; Kalantronoff et al., 2017; Tzelgov et al., 1992). We challenge the FFN theorists to provide a parsimonious mechanism of cognitive control that can, both, modulate task-relevant behavior and account for the experimental data.

One qualification of this conclusion is needed. An FFN-equivalent of such a control network may perhaps be achieved, but this will require a huge network, which may exceed the neural constraints of the organism. If this is the case, this could imply that cognitive control is achieved via RN in humans, but this fact is contingent on the neural constraints that humans and animals are subject to. I do not deny the viability of such a move, but the point remains that the main functional characteristic of conscious processes, is one which under the constraints of our environment, requires RN. (We do not know that under less restrictive environments, consciousness and control would still be (uniquely) associated). One possible narrative, consistent with RP and IIT, is that consciousness has developed in response to the pressure to develop a flexible control under the constraints of our environment (that limit the brain-size). As the association between control and consciousness (and dissociation between control and unconscious processes) is central to this argument and since there is debate on this issue, in the last section, I review empirical data that provides support for this association (and dissociation between control and unconscious processes). Additional to the argument about the relation of consciousness to RN, this data may be important to another central question that Doerig and colleagues, discussed. Is there a functional characteristic to conscious processes?

3.2. Does control dissociate between conscious and unconscious processes?

Although recurrent processing may be necessary for cognitive control, one may question the dissociation I made between conscious and unconscious processes with regards to their ability to deploy control. I have started with an anecdotal support for this dissociation (the routine/automatic walk to the old home), however, support for this dissociation was also obtained in experimental studies. Whereas there are some types of control (e.g., response inhibition in a stop-signal paradigm) that can be carried out (to a lower degree) unconsciously (i.e., based on subliminal cues; van Gaal, Ridderinkof, van den Wildenberg & Lamme, 2009), more flexible types of control appear to require consciousness.

One such case takes place when setting the conscious/controlled and the unconscious/automatic processes in opposition. For example, Merikle and Joordens (1997) presented a prime word (RED or GREEN, in gray color), which was either masked or not, and was followed by a color stimulus, which the participants had to categorize as red or green. Critically, the prime was negatively correlated with the target (in 75% a RED prime was followed by a green-target, and conversely, thus they had 75% incongruent and 25% congruent trials). The result was a cross-over interaction between the consciousness of the prime and its effect on the target. While for conscious primes, the participants were able to (inhibit the automatic response) and expect the alternative target, resulting in faster
RT for incongruent trials, with unconscious primes the RT was faster for congruent trials (see also Daza, Ortells, & Fox, 2002 for similar results). Below I briefly describe a similar study, reported in our lab (Bauer, Cheadle, Parton, Mueller, & Usher, 2009), which relies on a similar opposition method\textsuperscript{13}, to establish a qualitative dissociation between consciousness and control.

The experimental framework is a target detection task, in which observers monitor three Gabor stimuli and have to detect a subtle temporal change (spatial frequency) in one of them (Fig. 6A). Unknown to the observers, before the change, one of the Gabors (either the target or another of the three) flickers at a rate of either 30/50 Hz. Although the observers are at chance in detecting the 50-Hz flicker, the presence of this subliminal cue, makes them faster in detecting the target (the change in spatial-frequency) when the flicker-cue is congruent with the target (Fig. 6B). This indicates that 50 Hz flicker (but not the 30 Hz flicker) can act as a subliminal cue that attracts attention. In a second experiment, the observers are presented with a similar target detection with only two Gabors (left/right of fixation), one of which flickers in either a visible (25 Hz) or a subliminal (50 Hz) way. Critically, the flicker cue and the target are now placed in opposition: in 80% of trials the target appears on the opposite side of the flicker-cue (in 20% the target appears on the flicker-side). Participants are also informed about this validity manipulation, and thus they have the incentive to shift the attention away from the flicker cue in order to detect the target in most trials. As we can see in Fig. 6c, while a strong validity effect takes place for consciously seen cues (25 Hz), there is a smaller but significantly negative validity effect with subliminal cues. This shows that the presence of the cue is registered in the visual system but, because it does not reach conscious access, it cannot generate the control needed to divert attention towards the opposite side.

These opposition tasks are not the only ones in which control appears to require consciousness. Two other examples are trial-by-trial control adaptation effects (Gratton et al., 1992; see previous section), which take place only when the conflict is consciously detected (Heinemann, Kunde, & Kiesel, 2009), and the ability to carry out two-steps tasks (Sackur & Dehaene, 2009). While there is some debate in the literature regarding the types of control processes that depend on consciousness, one idea is that control operations (e.g., stop signals) that can be mediated by fast feedforward sweeps of information processing can be carried out without consciousness, while more flexible and complex control operations that are mediated by recurrent processing require conscious processing (Lamme, 2006; Dehaene & Changeux, 2011; van Gaal, de Lange & Cohen, 2012).

4. Discussion

The UA relies on the Universal Approximator theorem to argue that any theory, holding consciousness to depend on causal/physical structure, is untestable or unscientific (Doerig et al., 2019). In particular, this was concluded to be the case with regards to the RP (Lamme, 2006) and the IIT (Oizumi et al., 2014; Tononi et al., 2016) theories. While, a possible move in reply to this argument is to adopt a view of consciousness, which accepts consciousness to be totally dissociated from functional behavior (Oizumi et al., 2014), here I aimed to defend causal-structure theories, while maintaining a functional role (or functional signature) of conscious processes (see Chalmers’ Fading-Qualia argument for an example of conceptual problems facing the position that accepts a total dissociation between phenomenal experience and functional organization; Chalmers, 1995b).

In a recent reply to the UA, Tsuchiya et al. (2020) have objected to the exclusively behavioral operationalization of consciousness-data, on which the UA relies on. Accordingly, consciousness researchers also rely on the introspective conscious experience subjects have in the experimental conditions. In addition, Negro (2020) has shown that the charge of circularity (premise-4 of the UA) is

---

\textsuperscript{13}The idea of setting conscious/controlled and unconscious/automatic processes in opposition is also the basis from Larry Jacoby’s Process Dissociation Procedure (Jacoby & Kelley, 1992).
unfounded; for the case of Blockhead, one can see the theory for possessing mental states (or intelligence), as grounded on an intuitive pre-theoretical understanding of the phenomena. For the case, of IIT, this may be subject to some difficulties, such as demonstrating that the axioms uniquely determine the theory. Moreover, Negro (2020) also refuted premises-5 of the argument, showing that it is not backed up neither based on Popperian nor on Lakatosian criteria. It is possible that those are enough grounds to put the unfolding-argument to rest. I believe, however, that there is value to independently refute its central theoretical assertion: premise-2, which asserts that it is possible to generate full functional FFN equivalent for any RN. Doing so can, additional to removing potential concerns of uniqueness within the IIT and allowing us to defend RP (which does not have the benefit of phenomenological axioms like IIT), enable us to obtain a consciousness theory in which both causal structure and functional behavioral characteristics (in sync) are grounds for dissociating conscious and unconscious processes.

I have argued that premise-2 of the unfolding-argument is unwarranted and that there are several reasons that undermine the FFN-RN functional equivalence. First, the argument adopts a methodological behaviorist framework that leads to a too restrictive criterion of testability (too close to the Turing test; Block, 1981; see also Tsujiyta et al., 2020) about RN and FFN functional equivalence. Accordingly, equivalence is exclusively operationalized as an identity of input–output mappings. I have argued that, even if the RN/FFN mechanisms would be identical in all their possible (static) input output mappings, they can be distinguished with other experimental techniques, such as structural or physiological lesions or the administration of drugs. Such techniques, not only are scientific but have played a large role in the development of the cognitive neuroscience theory. Second, I have argued that UAT studies were incorrectly interpreted to argue that FFN are universal approximators of RN. Rather what UAT studies showed is that both FFN and RN are universal approximators, but of different entities, respectively (see footnote-7). Moreover, I have argued that premise-2 is an unlikely conjecture and that a robust equivalence between RN and FNN is not likely to be possible, in particular for cases that include dynamics and transients.

To support this, I have shown examples of FFN/RN mechanisms, which are identical in their static maps but vary in their transients by which the static maps are converged to. In particular, I have shown that RN mechanisms have a distinctive non–monotonic activation signature, and they differ from FFN in their temporal weights, thus resulting in divergent predictions to temporal perturbations. Moreover, I have argued that the concept of an input- output system, which is what defines a FFN is an outdated and inadequate one for understanding brain and cognitive processes, which are mostly self-generated and only modulated by sensory inputs. For example, research in neuroscience has identified self-generated activity patterns, associated with cognitive function (i.e., spontaneous resting state brain fluctuations, sometimes associated with the default network; Beatty et al., 2014; Broday-Dvir & Malach, 2020; Harmelech & Malach, 2013; Raichle et al., 2001), as well as the presence of neural-avalanches that reflect cognitive deficits (Shriki et al., 2013; Seshadri, Klaus, & Winkowski, 2018).

Finally, I have focused on a specific functional role that is associated with consciousness: flexible cognitive control. I have reviewed cognitive control models, in which the concept of top-down (and recurrent) feedback from higher-level (goal) stages to earlier/sensory stages of processing is crucial to their functioning. In addition, I reviewed experimental data supporting a qualitative (and not only a quantitative) dissociation between the presence/absence of consciousness and flexible cognitive control (Bauer et al., 2009; Daza et al., 2002; Heinemann et al., 2009; Merkle & Joordens, 1997; Sackur & Dehaene, 2009). This supports a long prevailing idea in cognition—that consciousness is necessary for overriding automatic/routine stimulus–response associations (Dehaene & Naccache, 2001; Jack & Shallice, 2001; Umiltà, 1988). If my argument that flexible cognitive control requires recurrent processing is correct this provides a counter-example to the RN-FFN equivalence, which also explains why flexible control depends on consciousness, at least under the constraints of our environment.

To conclude, I suggest that it is premature to rule out a role for causal structure theories, in the theory of consciousness. I wish to end with a warning (I join here Negro, 2020, and Tsujiyta et al., 2020) that restricting the scope of science to behavioral tests that are framed in terms of inputs/output mappings is a too restrictive method, which should have been abandoned when cognitive psychology (and then cognitive neuroscience) replaced behaviorism as the dominant theory of mental states.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Special thanks are due to Michael Herrmann and Amon Rosenman for extensive and helpful discussions on the UAT and on neural network theory, to Nikolaus Kriegeskorte for a helpful email-discussion on FFN/RN and for sharing his recent arXiv preprint. I also wish to thank Rafi Malach, Liad Mudrik and Hilla Jacobson for helpful comments on a previous version of the manuscript and Robert Chis-Ciure for very helpful editorial advice during the review.

References
