Phenom Cogn Sci (2017) 16:425–448
DOI 10.1007/s11097-016-9464-0

Mind-upload. The ultimate challenge to the embodied mind theory

Massimiliano Lorenzo Cappuccio[1]

Published online: 29 March 2016
© Springer Science+Business Media Dordrecht 2016

Abstract The 'Mind-Upload' hypothesis (MU), a radical version of the Brain-in-a-Vat thought experiment, asserts that a whole mind can safely be transferred from a brain to a digital device, after being exactly encoded into substrate independent informational patterns. Prima facie, MU seems the philosophical archenemy of the Embodied Mind theory (EM), which understands embodiment as a necessary and constitutive condition for the existence of a mind and its functions. In truth, *whether* and *why* MU and EM are ultimately incompatible is unobvious. This paper, which aims to answer both questions, will not simply confirm that MU and EM actually are incompatible. It will also show the true reason of their incompatibility: while EM implies that a mind's individual identity is contingent upon the details of its physical constituents, MU presupposes that minds can be relocated from one material vessel to another. A systematic comparison between these conflicting assumptions reveals that the real shortcoming of MU is not the one usually discussed by the philosophical literature: it has nothing to do with MU's functionalist or computationalist prerequisites, and is only secondarily related to the artificial implementability of consciousness; the real problem is that MU presupposes that minds could still be individuated and numerically identified while being reduced to immaterial formal patterns. EM seems committed to refute this assumption, but does it have sufficient resources to succeed?

Keywords Mind upload · Embodied cognition · Embodied mind theory · Enactivism · Extended mind theory · Brain-in-a-Vat hypothesis · Multiple realizability principle · Singularity · Functionalism · Computational theory of mind · Consciousness · Identity · Individuality · Autopoiesis

✉ Massimiliano Lorenzo Cappuccio
m.lorenzo@uaeu.ac.ae

[1] UAE University, Maqam Campus, 15551 Al Ain, Emirate of Abu Dhabi, United Arab Emirates

1 Introduction: mind-upload vs embodied mind

According to the theorists of Mind-Upload (MU), a technology capable to fully digitalize one's brain activity and transfer it to a computer might one day be available (Moravec 1988; Kurzweil 2005; Wiley 2014). Allegedly, this procedure would allow humans to abandon their organic bodies and continue their existences in the form of programs run by machines. Once uploaded, their minds would be able to live forever: they could be housed by any suitable artificial body built ad hoc; be relocated from any such body to any other at will; or spend their endless existence in a virtual simulation.

Would we accept to undergo a MU procedure, if all the relevant technological advances were available? Many among us would hesitate, fearing that the procedure would fail to safely transfer our mind to the computer and that, at best, it could only produce a digital clone of our mind (Corabi and Schneider 2012; Pigliucci 2014). Whether or not MU can deliver what it promises does not only depend on what technology is able to do, but also on the correctness of MU's metaphysical assumptions: i.e., that minds are actually the kinds of "substrate independent" things that could be extracted from a brain map, emulated by a computer, and moved from one material vessel to another.

In critically assessing the plausibility of MU's assumptions, the philosophical literature has almost exclusively focused on the conditions of material implementability of a mind: e.g., Chalmers (2010), expressing optimism about MU, claims that accepting or rejecting MU mainly depends on whether one endorses a Functionalist or a Biological Theory of Consciousness, respectively; accordingly and conversely, Pigliucci (2014), rejects MU exactly because MU presupposes a Functionalist (and Computationalist) Theory of Mind (and Consciousness). I propose an alternative way of framing the problem: I argue that the Functionalist/Computationalist Theory of Consciousness does not provide sufficient or necessary reasons to endorse MU, exactly like the Biological Theory of Consciousness does not provide sufficient or necessary reasons to reject MU; moreover, consciousness is just one of the mental components transferred by MU, and MU's true problem is not whether these components can be realized by non-biological vessels, but whether they can move between material vessels in the first place. I agree with Pigliucci (2014, p. 119) that MU's implicit dualism is problematic; but I think that the metaphysical (and problematic) nature of this dualism can be correctly spelled out only once the notion of substrate independency presupposed by MU is fully clarified. A generically anti-functionalist, biological, or materialist theory of the mind will not suffice to identify the contradiction hidden in MU's notion of substrate independency, as it doesn't dispose of the theoretical resources necessary to explain how the individuality of a mind is strongly contingent upon the conditions of its embodiment. That is why I propose to interrogate MU from the point of view of the Embodied Mind (EM) theory. EM can show that the substrate independency postulated by MU is not the standard functionalist version of the Multiple Realizability Principle (MRP), but a much more metaphysically demanding interpretation of that principle, one that involves the attribution of autonomous existence and individual identity to purely immaterial entities.

EM can expose better than other theories the reasons that make this metaphysics implausible. In general, MU and EM seem based on mutually exclusive premises: MU's Key Assumption (MUka) is that minds are "substrate independent" (Pigliucci

2014, p. 122), like software is independent from the computers that run it; in turn, the Fundamental Principle of all the EM theories (fEMp) is that what a mind is and what it can do strongly depends on the details of its bodily implementation (Varela et al. 1991; Gallagher 2005a; Shapiro 2011). Prima facie, EM and MU seem philosophical arch-enemies, and in fact MU looks like a radical version of the "Brain-in-a-Vat" hypothesis that EM characteristically opposes as an epitome of neural reductionism.

However, a careful comparison reveals that the relation between MU and EM is actually multifaceted: not only some versions of MU seem more sympathetic to certain aspects of the EM theories; even if MU and EM were totally incompatible, the reasons of such incompatibility would not be immediately evident, as the interpretations of MUka and fEMp vary significantly among the theorists. Whether and why MU and EM are to some extent compatible depends on what metaphysical prerequisites are assumed by MU and which of them are untenable according to EM, which in turn depends on what version of EM we consider. This can't be established without going through a philosophical analysis of how the different varieties of MU and EM interact, an analysis that this paper aims to offer. Two factors make it complicated: first, various hypothetical methods to obtain MU have been imagined, involving different metaphysical requirements; second, EM comprises a wide range of distinct philosophical doctrines that interpret fEMp's meaning in different ways. For example, embodied cognition can be characterized as embedded (Rupert 2009), enactive (Noë 2005), or extended (Clark and Chalmers 1998). These characterizations correspond to significantly different views on Functionalism and Computational Theory of Mind, among other things.

Despite this diversity, I will argue that no version of EM is compatible with any version of MU. The core assumptions of MU and EM actually are antithetical at root: to begin with, I will present a new thought experiment (a particular version of MU) to show that the kind of substrate independency implied by MUka doesn't simply require that minds could be replicated in different material substrates, but also that a mind could be literally moved from one material substrate to another while preserving its numerical identity; fEMP, on the other hand, doesn't simply state that the causal/functional structure of the mind is at any time necessarily implemented by some material substrate, but also that the material substrate (which includes both the neural and the extra-neural body) plays a "physically constitutive" role in individuating the mind, determining the contingent circumstances that numerically identify it. As the body cannot be constitutive of the mind if the mind can be discorporated and moved from one body to another, MUka and fEMP are mutually exclusive.

If this conclusion is right, then MU and EM represent a serious theoretical challenge to each other, in that MU is unfeasible if minds are somehow embodied, while all the EM theories would be proven wrong if MU obtained even once. Showing that MU and EM are incompatible is not the only goal of this paper. Identifying the actual cause of such incompatibility is timely and necessary, because it represents a crucial step to expose the real reasons to reject MU, recognizing that these reasons are not those usually discussed by the philosophers. In turn, acknowledging why MU has to be rejected offers a new perspective to re-discuss certain classical questions about the nature of the mind, such as: does the individual identity of a mind coincide with the identity of the material vessels that occurrently embody it? Is material embodiment only a necessary or also a constitutive requirement for the existence and individuation of a

mind? Can a mind be relocated from a material vessel to another preserving its numerical identity? Given the general philosophical significance of such questions, contrasting EM and MU ultimately is a worthwhile exercise also for those who don't believe in neither of them. On the one hand, as the rejection of MU's metaphysical assumptions represents a unifying element for all EM theories, it can offer a vital polemical target for any detractor of EM; on the other hand, proving MU wrong could indirectly highlight the shortcomings of other theories of the mind that tacitly rely on the very dualistic assumptions that MU incarnates.

My investigation is structured as follows. Section 2 introduces the causal mechanics of two hypothetical MU procedures, highlighting how they conceptually differ from similar thought experiments and clearing the stage from some common misunderstandings. Section 3 identifies MU's three metaphysical requirements, which I call 'Neurocentrism', 'Formalism', and 'Ultra-Functionalism': the most characteristic among them (and, as I will argue in Sections 4 and 5, the most problematic) is Ultra-Functionalism, which affirms the in-principle possibility to preserve the (numerical) identity of the subject's mind throughout the upload procedure. Section 5 examines MU's three metaphysical requirements from an EM's viewpoint, and shows through a specific thought experiment that only Ultra-functionalism is at odds with the fEMp. I explain why recognizing Ultra-functionalism as the only justified reason of general incompatibility between EM and MU helps us identify the actual theoretical shortcomings of MU, often overlooked by the philosophical literature. Section 4 deepens the general implications of Ultra-Functionalism, explaining why this particular version of Functionalism, unlike the classical one, implies an intrinsically contradictory notion of substrate independency. The conclusive section asks whether EM's arguments are sufficiently strong to definitively rule out MU: this represents a crucial challenge for EM, one its very legitimation might depend on.

2 Mind-upload as a philosophical thought experiment

As envisaged by the transhumanist movement (Blackford and Broderick 2014), MU theory claims that a whole mind could "migrate from brain to computer" (Chalmers 2010, p. 34); this process would allow it to live forever "within a virtual reality or simulated world", supported by an anatomic 3D body simulation model. Alternatively, the simulated mind could reside in a computer that operates inside (or connected to) a humanoid robot or a biological body. The normative validity of MU's metaphysical assumptions can be assessed by means of a priori arguments through a philosophical thought experiment: the technical details of the MU procedures described by this experiment are crucial to assess the plausibility of the intuitions underlying MU. The paradigmatic analysis is offered by Chalmers (2010, p. 42), which describes two hypothetical methods to implement mind transfer, naming 'Dave' the subject who undergoes the MU procedure, and 'Digi-Dave' the computer emulation that results from the procedure.[1]

[1] While other methods have been conceived (Wiley 2014), I will examine only these two methods of destructive upload described by Chalmers (2010) because they provide the best-case scenario for MU and illustrate its mechanics in the most compelling way.

The first method (M1) is referred as "Gradual Upload" and involves "nanotransfer", a process meant to progressively map out and supplant the functions of Dave's brain: tiny robots run through Dave's synapses to learn the full schema of their functional organization and offload "the relevant processing to a computer via radio-transmitters" (Chalmers 2012, p. 34). While the computer gradually imports all of the brain's informational processes, the activity of the corresponding original brain structures is concurrently discontinued. In spite of this, no discontinuity in Dave's mental life can be noticed throughout his transformation into Digi-Dave, neither functionally nor phenomenologically: that is because, during the procedure, each component of the original brain (the organic neurons) is constantly connected to and communicates with its complement in the virtual brain (the emulated neurons). The organic and the emulated counterparts complete each other's functions, sending and receiving as usual all the information necessary to their local activities, so to form a temporary "causally integrated" meta-system distributed through the two systems (Chalmers 2012, p. 160). Therefore, while the activity of the neurons in Dave's brain is slowly destroyed, it is also synchronically recreated in Digi-Dave's mind by the computer. The 'migration' of Dave's mind is an effect of this progressive relocation: by the time the last robot has accomplished its nano-transfer work, the original brain has completely stopped working and an exact emulation of its functions (Digi-Dave) has started running on the computer.

The second method (M2) is referred as "Copy-and-Transfer", and relies on the availability of advanced technologies of "brain imaging" and 3D-printing, "with fine enough grain that neural and synaptic dynamics can be recorded" and replicated. Once Dave's connectome has been completely mapped out, it can be duplicated in an artificial brain (or in a computer simulation) that exactly mirrors the functions of the original. But creating a perfect replica is not enough to give life to Digi-Dave, as additionally the activity of the first brain has to *be transferred* to the second. For this to happen, the two brains must form a causally integrated meta-system during the upload procedure, as in M1. Also in this case, the de-activation of Dave's brain parts is synchronized with the concurrent activation of the matching parts in Digi-Dave brain, so that the two systems perfectly complement each other at any given time, until Digi-Dave's brain is fully operational and Dave's brain entirely shut-down. If everything goes well, Dave might not even notice that he is being transferred from one brain to another until the procedure is over.

The logic of M1 and M2 seem more compelling if we accept that brain functions can be isolated and swapped with vicarious extra-cranial vehicles of cognition, assuming that the latter were functionally equivalent to the former (as per "parity principle", Clark and Chalmers 1998): through MU, the progressive outsourcing of the internal functions leads to the complete relocation of the whole mind to a new substrate. Importantly, M1 and M2 re-assign the functions from the old vehicles to the new ones, without this involving any movement of the material vehicles themselves. No material structure in any of the two systems is physically replaced or moved at any time: the formal configuration of Dave's computational functions is the only thing that 'shifts' from one substrate to the other.

This is why, despite the analogies, MU is substantially different from any "ship of Theseus" situation (pace Hauskeller 2012): the Theseus paradox asks us to imagine that the material components of the system are materially replaced one by one while the

overall formal structure of the ship is preserved through time; conversely, MU asks us to imagine that the material components of the system are kept unchanged, while the formal structure is relocated from a material structure to another. Even if all the neurons in Dave's brain could be replaced by microchips one at a time (as imagined by Kurzweil 1999), without this generating discontinuities in Dave's mental processes, this procedure still wouldn't count as mind-upload, as Dave's mind would still be realized by the same old system (though, at that point, the system would be instantiated by different material parts). Certainly, nothing prevents us from imagining a ship of Theseus situation combined with a MU procedure: we could imagine that the microchips that supplanted Dave's neurons were able to send all information in Dave's mind to an external computer, as portrayed by M1 situation. If MU actually obtains, it is because at a certain point the information has been offloaded, not because the neurons have been materially replaced. Therefore, whether MU involves a ship of Theseus scenario or not, it doesn't make any difference. MU's concept is about the possibility to transfer the formal structures of the mind, not about the possibility to replace its material realizers.

MU also resembles other well-known thought-experiments involving exact material replicas of one person (swampman, zombies, etc.). However, MU is substantially different in that the mere existence of a perfect duplicate of the vehicles of cognition is neither a sufficient nor a necessary condition for mind transfer. On the one hand, M1 shows that preserving the material details of Dave's brain is not a necessary condition for successful upload: it doesn't matter whether Digi-Dave is made out of neurons or microchips, or whether he lives in an organic body or a digital simulation, as long as Digi-Dave is a "functional isomorph" of Dave. On the other hand, M2 shows that a physical duplicate of the vehicles of cognition is not a sufficient condition for successful upload either: the brain-replica is just an empty simulacrum of the original until Dave's cognitive functions are transferred to it. At best, a functioning duplicate could be just another token of the same mind-type, but MU occurs iff it is the same mind token that migrates to a new substrate.

Also the teletransportation thought experiment is similar to MU but, as remarked by Chalmers (2010, p. 41), and ignored by Pigliucci (2014, p. 127), teletransportation is "not the same as uploading: it preserves physical organization where uploading preserves only functional organization in a different physical substrate." MU, like teletransportation, involves the spatial relocation of the subject but, unlike teletransportation, obtains iff what is actually moved from one substrate to another is just the pure formal organization of Dave's mind.

Therefore, in spite of the similarities, neither the ship of Theseus, nor the clone, nor the teletransportation thought-experiments portray a MU situation, as neither the physical substitution of the material vehicles of cognition, nor their exact replication, nor their relocation in space would be sufficient for successful upload.

3 Normativity of mind-upload

Under what conditions could we consider a MU procedure successful? MU would legitimately occur iff Dave survived the MU procedure and started a new life as Digi-Dave. Digi-Dave really is Dave, not just his computerized twin. How can we reassure

the skeptical that the procedure is not just killing Dave and replacing him with a digital clone? In defining the conditions of Dave's and Digi-Dave's identity, we may want to allow some tolerance (Chalmers 2010, p. 60, 2012, p. 158; Goertzel 2012). Dave's personality, cognitive powers, and behavior might have changed due to the procedure (not necessarily in a bad way. For example his cognitive functions might have been augmented). But even significant changes, which could actually occur after any delicate brain surgery, shouldn't lead us to the conclusion that Digi-Dave is not Dave anymore. Instead, we must demand that the procedure allows some kind of robust continuity in their mental life: for example, that Digi-Dave keeps Dave's core cognitive functions (intelligence, decisional processes, etc.), and/or his subjective consciousness (e.g., phenomenal qualia and minimal self), and/or the distinctive mental contents (e.g., truth-preserving memories and beliefs) that are associated to his individual personal status (defining his rights, properties, social entitlements, etc.). He must have them in an ontologically inherent, non-derived way, because Dave and Digi-Dave are expected to be one and the same mind-token, not two tokens of the same mind-type.

In other words, the procedure is not only—and not primarily—expected to preserve Dave's mind *qualitative identity* (which includes, but may not be reducible to, the full behavioral and functional organization of the components of his cognitive system), but also—and especially—Dave's mind *numerical identity*. What are the requirements to secure qualitative and—most important—numerical identity throughout the MU procedure? People are likely to answer in different ways depending on their beliefs about the constituents of the mind.

(i) *Function*. Reductive functionalists and eliminativists will argue that Digi-Dave's mind is qualitatively identical to Dave's as long as they process information with the same algorithms. However, this is not enough to establish whether they are numerically the same, as their functional equivalence can't tell whether Digi-Dave truly is Dave or just Dave's copy.

(ii) *Consciousness*. Anybody who believes that consciousness is an indispensable feature of Dave's mind, without being reducible to its functional organization, will argue that Digi-Dave's mind must be conscious too, and not only functionally equivalent to Dave's, in order to be qualitatively identical to Dave's. Also in this case numerical identity is an ulterior requirement, in that it is not sufficient that Dave's and Digi-Dave experience the same phenomenal qualia: if they are the same person, then Dave's and Digi-Dave's experience one and the same flux of consciousness.

(iii) *Mental content*. Content externalists (Putnam 1973), and anyone who believes that mental contents are not reducible to the functional organization of the mind (e.g., the supporters of Radically Enactive Cognition, Hutto and Myin 2013), will argue that, even if Dave's mind and Digi-Dave's shared the same functional organization and phenomenal qualia, they would not be qualitatively identical, let alone numerically identical, unless the mind-upload procedure somehow managed to preserve all of Dave's mental contents.

Other constitutive components, in addition to function, consciousness, and content, might be required to exhaust Dave's mind. Whatever they are, these mental components must possess certain features in order to migrate from a brain to a computer.

These features depend on the metaphysical assumptions that MU is committed to. Prima facie, MU's relevant assumptions are three.

(A) *Neurocentrism* (which may or may not involve *Neuroreductionism*): an exhaustive description of the brain (neurocognitive) activity is possible and enables the full analysis of a mind;
(B) *Formalism* (which may or may not imply *Functionalism* and *Computational Theory of Mind*): brain (neurocognitive) activity is reducible to formal structures and processes that can be re-instantiated by different material substrates without suffering functional changes;
(C) *Ultra-Functionalism*: the set of formal structures and processes that exhaustively map one's brain activity can be relocated from one substrate to another remaining just one and the same set.

With regard to A, MU certainly requires Neurocentrism, the thesis that exhaustive knowledge of one's neural processes is possible and necessary to perfectly recreate its mind (e.g., reproducing one's "mental software" always involves "full brain emulation" capabilities, because mental activities always involve the functions of a brain). Some versions of MU might involve Neuroreductionism too, the thesis that neural processes are the only possible vehicles of one's mental activity (i.e., the mind is entirely and necessarily confined within the head, as only the neuronal body is responsible for its mental activity. Therefore, the causal processes in one person's extra-neuronal body give no contribution for reproducing one's mental software).

MU requires also B because the material realizers of the mental software (including the brain and—possibly—the extra-neuronal body) are exactly what is left behind during the upload process. While MU theorists expect mental software to be reducible to formal (non-material) configurations implementable by different material substrates, they do not specify the nature of these structures (called simply "patterns" by Moravec 1988, pp. 116–117). Depending on one's beliefs about the nature of the mind, these configurations are best characterized as propositional contents, mental representations, etc. If the MU procedure involves computer emulation (as in M1), then these structures must be characterizable as mechanizable algorithms, i.e. computable functions.

Finally, MU requires C because an uploaded mind is not just meant to be a duplicate of the original mind: in order to retain one's individual identity, it must still be the very same mind, transferred to a new vessel, including consciousness, content, and functional configuration. Crucially, while C presupposes functionalism (B), in the sense that functionalism is a necessary non-sufficient condition for C, C is not entailed (or contemplated) by any standard functionalist account of the mind.

That is why C is a third necessary requirement, conceptually irreducible to A and B: even positing that all the constitutive components of Dave's mind were internal (brain-bound), formal, and multiply-realizable computable functions (as per A & B), i, ii, and iii would still demand a separate explanation of how these components could be relocated (as opposed to simply copied) from Dave to Digi-Dave, remaining not only functionally but also numerically the same. What element, added to A & B, could ever grant C, if anything, and how could we verify that it is in place? Chalmers (2010, p. 48, building on Parfit 1971, 1995) optimistically alleges that "a certain sort of *continuity* or *connectedness* over time" is all it is needed for a mind to endure through time, and

MU seems granting it. Three groups of theories, invoked by Chalmers, contend whether the necessary continuity or connectedness is "biological", "psychological", or related to spatio-temporal proximity ("closer-continuer" theories). While the biological theories argue that identity is embedded in the original brain-organism system and specifically tied to the biological realizers of the mind, the other two theories of identity would easily concede that a mind could be implemented in very different kinds of material substrates. That is, psychological continuity and closer-continuer theories respectively assert that a mind survives whenever the cognitive processes continued uninterrupted, or the new system is sufficiently closely-related to the original one (in case of many new systems, only the most closely-related is the legitimate continuer). M1 conforms to psychological and closer-continuer theories, as they allow the continuation of one's mind in spite of the heterogeneity of the material realizers (e.g., microchips or virtual simulations). M2 may conform to biological theories, if the procedure is sufficiently fast to preserve the life functions and the brain replica is sufficiently similar to preserve biological functions (some spatio-temporal discontinuity in the physiological and cognitive processes may occur, but Wiley and Koene 2015 argues that biological processes are never entirely continuous anyway, as they present small temporal and spatial gaps).

Cerullo (2015) argues that these three theories are not enough to secure C because none of them offers a solution to the 'fission problem' (Parfit 1971; Corabi and Schneider 2012, p. 40), i.e. the dilemma that asks which one of the two identical hemispheres in which a whole brain has been split is supposed to inherit the mind of its owner—if any at all. In fact, the split-brain dilemma predicts biological and psychological continuity in each of the two halves of the brain, which in turn are equally close to the original brain. That is why, at best, these three theories could indicate necessary conditions for numerical identity to obtain, not sufficient ones: it is perfectly conceivable that Digi-Dave were just a digital twin of Dave, not Dave himself, even if Digi-Dave's mind supplanted Dave's psychological and biological functions and was his closest possible continuer in the world.

Mutatis mutandis, the same dilemma arises from MU, as M1 and M2 allow for the possibility that the same mind were uploaded to more than one computer or replicated in multiple brains at once.[2] In this scenario, biological and psychological functions would continue through multiple identical substrates, but we couldn't tell which uploaded mind is the original, if any; also, there could be more than one identical continuer, and none of them would be closer than any other to the original. So, which Digi-Dave would be the "real" Dave, if any of them? This dilemma would be dissolved by the three theories if any of them were capable to specify and fulfill all the conditions required for numerical identity to obtain. But this is not the case: even assuming (in line with Parfit 1971) that MU is successful iff Dave survives as Digi-Dave, and that Dave survives as Digi-Daves iff the right kind of connectedness between them exists, we still haven't been given any reason to believe that M1 and M2 are actually bringing about that particular kind of connectedness. Therefore, in order to trust MU is capable to

[2] Chalmers (2012, p. 159) denies that multiple simultaneous upload would be possible, arguing that the particular integrated mechanics of the procedure would engender causal overdetermination. Even if that was true, what does prevent us from simultaneously replicating Digi-Dave's mind at will in multiple concurrent emulations, once it has been fully digitalized for the first time? For M1, any emulation so produced would count as a distinct upload.

preserve numerical identity, some other ingredient must complement A & B, pace Chalmers 2010. Ultra-Functionalism (C) is this element: an implicit stipulation embedded in the MU theory to reassure us that the kind of connectedness established by M1 and M2 between Dave and Digi-Dave is exactly the particular kind of connectedness that secures numerical identity between Digi-Dave and Dave. In the next section I will enrich the characterization of C through a contrastive procedure, i.e. through an analysis of its incompatibility with EM.

4 The take of embodied cognition theory

Can EM accept MU's three requirements (A–C)? The answer may depend on the kind of embodied-embedded theorist the question is asked to. After generating an influential debate in cognitive science and philosophy of mind, the EM approach to cognition is today fragmented into a constellation of various—often conflicting—positions and views (reviewed by Gallagher 2011 and Menary 2010b): they range from the theories of minimal embodiment (which assert that the sensorimotor areas of the brain represent information in a "bodily format", scaffolding higher cognitive functions) to the enactive (radical) theories of cognition (which claim that cognition builds not on representation, but on action-perception feedback loops emerging from the interaction), including various positions in between—such as embodied functionalism and biological or semantic embodiment. Following the suggestion that only the theories seriously committed to fEMp can legitimately be considered embodied-embedded (Gallagher 2015), my analysis will focus on the enactive and extended theories of cognition: both accounts build on the premise that minds are causally co-determined and essentially co-constituted by the physical contingencies in which the body is contextually situated. Nonetheless, some major differences between them exist and are relevant in this context: the Enactivists, on the one hand, are generally oriented towards biological and dynamicist accounts of the mind, and stress the non-locality of mental faculties, which holistically emerge on the continuous adaptive interplay between the organism and its ecological surroundings; the extended mind theorists, on the other hand, are typically sympathetic to functionalist and mechanistic accounts of the mind, and stress that intra-cranial (neuronal, usually organic) and extra-cranial (non-neuronal, possibly artificial) vehicles of cognition are in principle interchangeable or complementary, and at times replace or augment one another, as they are both suitable to locally realize and occasionally carry out the same cognitive processes. Are both accounts of embodied-embedded cognition, in spite of their difference, similarly disposed towards MU's three assumptions (A–C)?

In considering MU's first assumption (A), one must note that, while Neurocentrism is necessarily part of it, Neuroreductionism is just an option. Actually, only some versions of MU are compatible with Neuroreductionism (M2), while other versions (M1) envision the possibility to upload one's mind to non-neural inorganic vessels, contradicting the idea that brains are indispensable to mental activity. Neurocentrism only asserts that exhaustive mapping of the neuronal processes is a necessary step to replicate the functional organization of one's mind, but it doesn't exclude that the non-neuronal body could participate in various cognitive processes; Neuroreductionism, in turn, asserts that the neuronal body is the only possible seat of cognition, and—as

opposed to the enactive and the extended approaches—denies that any non-neuronal substrate could ever play a role in mental functions.

Therefore, if A included Neuroreductionism, then MU would be incompatible with EM theories. The Enactivists, to begin with, deny that a mind could be reduced to representations or computations located in the head, as mental processes necessarily involve bodily varieties (e.g., sensorimotor, affective, etc.) of practical or social interaction with the environment. Also "first-wave" (Clark 1997) and "second-wave" extended mind theories (Menary 2010a; Sutton 2010) stress the deep codependence of brain, body, world, characterizing it with their own terms (equivalence/replaceability or complementarity/integration, respectively)[3]: even though Parity Principle and Extended functionalism (Clark 1997; Wheeler 2010) justify the theoretical possibility that all cognitive functions were carried out inside the head (in fact, as previously noted, the Parity Principle makes M1 and M2 more plausible), all the extended mind theorists deem highly improbable that actual human minds would be entirely accounted for by intra-cranial processes only, as in our world the processes that contribute to human cognition are pervasively distributed across brain, body, and environment, and consolidate when a dynamic coupling establishes complex feedback loops between these components. On all EM accounts, in short, the extra-cranial body is the medium of the brain-world integration, as the functional and phenomenological specifications of a mind are structurally determined by the biological, ecological, and historical contingencies of its bodily implementation (Heersmink 2015). If A leaves no space to non-neural forms of cognition, then MU is incompatible with EM. However, it is well possible that MU commits itself to Neurocentrism only, without endorsing the neuroreductionist assumption that is at odds with EM.

This becomes clearer when we consider the close analogy between MU and the famous Brain-in-a-Vat hypothesis (BiV): BiV states that the entirety of one's cognitive and experiential life could be preserved unchanged if its brain, removed from the non-neural body and kept alive by a machine, were opportunely wired to a computer capable to feed its synapses with the same electrochemical signals that are usually provided by the non-neural body. MU can be considered an even more radical version of BiV because, while BiV asserts only that a mind, separated from its non-neural body and reduced to a secluded brain, could still generate mental functions in symbiosis with a digital computer, MU asserts that the very same kind of mind-computer digital symbiosis could obtain even if the original brain itself were eventually discarded and replaced by digital processes. MU, like BiV, builds on the assumption that knowing and controlling the system of inputs and outputs that inform one's neuronal processes (that is the inputs and outputs generated from the non-neural body during its dynamic interaction with the external world) is a necessary condition to fully analyze and reproduce the totality of one's mental life. The reasons EM would have to reject A, if any, are the same reasons EM would have to reject the typical neuroreductionist interpretation of BiV. That is why, if one believes that BiV implies a neuroreductionist interpretation of A, and rejects BiV under the assumption that Neuroreductionism is untenable, then he will also reject MU for analogous reasons. That is why many EM theorists dismiss BiV as an implausible fantasy.

[3] "First-wave" extended mind theorists (Clark and Chalmers 1998) tend to characterize minds as being extended on an occasional, not constitutive, basis.

However, while EM certainly rejects Neuroreductionism, BiV can be interpreted in ways that imply a non-reductionist interpretation of A (Gallagher 2005b; Clark 2008, p. 164; Cosmelli and Thompson 2011). Under this interpretation, BiV actually works as an argument *in favor* of EM and *against* Neuroreductionism: BiV presupposes that the computer wired to the brain possesses an exhaustive knowledge and a capability to fully simulate the fine-grained details of the interactions between the brain and the non-neural body/environment; this amounts to admit that the brain can function only if the totality of its contingent relations of reciprocal dependence with the organic body, in interaction with the environment, are preserved; but this is exactly the principle asserted by fEMp and rejected by MUka, i.e. the idea that mental functions and conscious experience closely depend on the concrete details of the extra-neuronal circumstances in which the mind is implemented. Proving that a computer can (in principle) recreate the fine-grained system of stimuli contingently produced at any time by the body/environment on the brain is a burden for the neuroreductionist interpretation of BiV, not for EM, which in turn is not weakened by the fact that implementing the scenario envisioned by this thought experiment is in practice impossible or nearing impossibility. This reasoning suggests that BiV, combined with a non-reductionist interpretation of A, is perfectly compatible with, and actually corroborates, EM. According to the same reasoning, MU is compatible with EM too, as it could be argued that one's mental functions can be properly emulated/duplicated only under the condition that the fine-grained details of the brain/body relationship (i.e. the complex systemic interactions between brain, extra-neuronal body, and environment) were analytically examined and perfectly replicated through M1 or M2. From this perspective, not only A doesn't contradict EM, but offers an argument in favor of fEMp.

If A is not incompatible with EM, what about B? EM theorists are generally unsympathetic to the idea of reducing minds to formal structures, but whether EM is truly compatible with B depends on the kind of Formalism presupposed by MU. If Formalism means that purely immaterial essences could ever be extracted from one's brain, then MU is trivially incompatible with EM. However, MU theorists contend that M1 and M2 do nothing else than manipulating certain natural processes by means of other natural processes, according to logically consistent and physically plausible causal chains that involve no immaterial substances or supernatural causes (Hayworth 2012; Wiley 2014; Wiley and Koene 2015). If that is true, then the formalist requirement included in B does not presuppose any dualism of substances: the claim that minds are reducible to formal structures does not imply—in itself—that minds subsist without material implementers; it means that the physical structures that causally determine mental functions (their specific patterns of organization) can be identically replicated across multiple material substrates preserving functional equivalence. This interpretation of B, suggested by the MU theorists, is nothing else than the standard functionalist interpretation of the Multiple Realizability Principle (henceforth: "Classical MRP").

Supposing that the MU theorists are right, the key question about B at this point would be whether or not fEMp is incompatible with Functionalism (and Classical MRP). The answers given by the various EM approaches would hardly be unanimous: in particular, enactive and biological embodiment theories (e.g., Chiel and Beer 1997) would deem M1 implausible, as this method to upload minds alleges—in a functionalist fashion—that mental processes aren't affected by their re-instantiation through

inorganic substrates involving radically different material composition. Other versions of EM (e.g., Extended Functionalism, a version of the Extended Mind Theory—see Wheeler 2010) might well be compatible with this particular aspect of M1. Some versions of MU's Formalism can be accounted for without involving Functionalism at all, therefore whether EM is compatible with Functionalism or not doesn't affect our analysis: for example, even anti-functionalist EM theories, like the enactivist ones, could find the formalist mechanics of M2 generally plausible, as this version of MU postulates that the new vessel to which the mind is transferred (a brain) is not only functionally but also materially (biologically) identical to the old one. Thus, even the Anti-Functionalist could have reasons to believe that, in accord with B, the migration to the new brain doesn't compromise the qualitative identity between Dave's and Digi-Dave's minds. It ultimately seems that the relationship between fEMp and Functionalism, whatever it truly is, can't determine whether EM is compatible with MU.

A similar reasoning applies to the computability requirement (which is an optional addition to B). Many EM approaches explicitly reject the Computational Theory of Mind for a multiplicity of reasons (for example because they deny that brain functions were ever computable by digital machines. This skepticism is shared by other, non-EM, approaches to cognition, e.g. Greenfield 2012, and Pigliucci 2014). Most of the Enactivists (and in particular the radical ones, see Hutto and Myin 2013) reject the classical Computational Theory of Mind because it entails the implausible semanticization of natural processes and properties: according to the classical cognitivist account, computation involves representational contents, and the involvement of contentful representation in basic forms of cognition is rejected or heavily downsized by most EM theories in favor of habit-formation and pre-reflective adaptive skills (e.g., Dreyfus 2002). However, not all the EM approaches entirely reject representationism (especially if they rely on minimal forms of representation, see Wheeler 2010), and some of them endorse computationalist accounts of biological functions (Clark 1997). Some other EM theorists might accept B, on the grounds that Formalism and Functionalism do not necessarily imply an endorsement of the Computational Theory of Mind (B1): for example, M2 proposes an analogical "copy-and-transfer" method to replace/replicate organic functions that doesn't imply digitalization of information.

It seems that, whether an EM theory endorses none, some, or all of B's optional components (Functionalism, and/or Computationalism) is insufficient to assess whether EM in itself is compatible with MU or not: on the one hand, some functionalist and computationalist versions of MU could plausibly go hand in hand with the functionalist and computationalist theories of EM; on the other hand, the EM theories that don't endorse Functionalism or Computational Theory of Mind are not likely to recognize these views as necessary assumptions of the MU procedures, and would avoid functionalist or computationalist characterizations of M1 or M2 in the first place. Considering this, and that Formalism per se doesn't constitute sufficient ground to endorse or reject EM or MU, we must ultimately conclude that requirement B is not intrinsically incompatible with fEMp.

This shows the inadequacy of the oppositive terms in which the philosophical discussion on MU has been framed until now. In spite of their antagonism, Chalmers (2010) and Pigliucci (2014)—that respectively endorse a Functionalist Theory of Consciousness favorable to MU and a Biological Theory of Consciousness adverse

to it—seem gladly sharing the assumption explicitly stated in Chalmers 2010: that the opposition between supporters and critics of MU at root is an opposition between functionalist and biological accounts of consciousness. The former claims that consciousness can be instantiated by both organic and non-organic substrates; the latter that only organic, living (typically neuronal) substrates can implement consciousness. This way of framing the problem is unsatisfactory for two reasons highlighted by our analysis: in the first place, Consciousness is not more important than Function and Content in assessing the validity of MU, as the theoretical complication faced by MU in relation to any of these mental components is just one and the same (preserving quantitative identity); in the second place, neither the functionalist nor the biological theories of consciousness (or the mind in general) are necessary or sufficient to endorse or reject MU (or EM, for that matter). In fact, our analysis of M2 suggests that MU can be achieved through a purely biological implementation, from organic neurons to organic neurons, without involving digitalization procedures that would imply a Computational Theory of Mind or Functionalism. The fact that Ultra-Functionalism is required by MU, while not being reducible to classical Functionalism, shows that Functionalism is absolutely not a sufficient element, and in some cases also not a necessary one, of a sound MU theory (in that qualitative identity is certainly not sufficient, and in some cases also not necessary, for numerical identity to obtain in a procedure of mind relocation).

C has largely been neglected by the philosophical literature, which has prevalently been focusing on A and B. If A and B were sufficient to account for the metaphysics of MU, then A and B would represent the only justifiable reasons of skepticism towards MU, and there could be versions of MU that are compatible with at least some versions of EM, and possibly with all of them. For example, one could imagine a variant of the "copy-and-transfer" procedure (call it "M3") in which not only Dave's brain (as in M2), but his entire body, has been 3D-printed out. The artificial body is indistinguishable from the original. While Dave is asleep, his mind is uploaded into it. Once awake, he doesn't even notice the difference because, through the new body, appropriately connected to his mind, he can still engage in all kinds of familiar interactions with the world, as he had always done with his organic body. If the only problems EM has with MU's metaphysics of the mind were related to the availability of a bodily substrate with appropriate organic characteristics, then virtually all the supporters of EM might run out of valid reasons to reject M3. However, the main issue EM theorists should be concerned with is not what body would be suitable for the relocation of a mind (biological vs mechanical, etc.); the problem is whether a mind can be relocated at all.

This is the true source of the incompatibility between MU and EM: C posits criteria of continuity and identity of a mind that are extrinsic to its physical and functional constituents, and unrelated to the specific contextual integration of the mind-body-world system; on the contrary, fEMp implies that individual identity, real status (existence), and (according to top-down emergentist accounts, e.g. Varela et al. 1991) causal powers of a mind exist only in relation to its bodily implementation: not only the existence, but also the very essence of one's mind depends on the contingent circumstances of its material realization; and the essence of mental functions itself is co-determined with the contingencies of its existential constitution.

Why? According to fEMp, a functioning bodily substrate is *not only a necessary*, but *also a constitutive condition* to have a mind (additionally, on the enactivist-autopoietic

account, the availability of a living organic body is also a *sufficient condition* to have a mind, e.g. Thompson 2005). On MU's account, on the contrary, material contingencies define only the contextual way certain tokens implement the type at a given time, but the type itself is a purely autonomous, atemporal formal structure. This is because, on a MU view, a mind type is a token of itself, ontologically and not just virtually independent from its physical implementation. MU agrees with EM that bodily implementation is a necessary condition to have a mind, but can't allow bodily implementation as a *constitutive* condition: this means that for MU, even if in normal circumstances we only encounter minds implemented by material vehicles, minds are never "made out" of those vehicles or concretely identified by them.

The difference between necessary and constitutive conditions is notable. If an entity must fulfill its constitutive conditions in order to exist, and existence is what provides that entity with a unique (numerically distinct) individual identity, then EM's general definition implies that minds (including consciousness and mental contents) are individually identified by the unique material contingencies in which they are realized. On any EM account, including the functionalist ones, the bodily implementation defines a mind as the unique token (by establishing its unrepeatable individual existence) of a particular type (by shaping its functional and behavioral organizational invariants). Subsequently, for any EM account, material contingencies don't simply establish the normative preconditions to be qualitatively the same, but also numerically the same. This is sufficient to exclude C. According to EM, even if formal structures (in MU's sense) existed and exhausted the mind (as per B), they wouldn't constitute the mind: i.e., they wouldn't be anything more than virtual networks of functional relations, causally grounded into and individuated by their concrete modes of material implementation. If I am right, all the EM approaches to cognition should tend to reject MU's third requirement (C), i.e. the possibility (alleged by Ultra-Functionalism) that purely formal mental structures can be spatially relocated while preserving their numerical identity through the process.

Now, the constitutive role attributed by EM to material implementation implies neither that the mind is reducible to its physical realizers, nor that a mind can be instantiated only by one type of realizers (Classical MRP is not ruled out a priori), nor that a mind cannot exert any causal influence on them. Especially, it doesn't imply that the material realizers of a mind cannot be replaced over time. For example, the Autopoietic Theory of the Living offers the most effective EM account of how mental components can and do fully change without interrupting the continuity of one's mind activity, as in a ship of Theseus scenario. On this account, all the changes in the mind-body system follow an adaptive and self-preserving logic that inherently belongs to the incessant flow of material constituents. Note that other generically biological and materialistic accounts of the mental simply state that the identity of the mind supervenes on a continuous flux of contingent modifications and replacements, leaving the relationship between the whole and its parts extrinsic (i.e., necessary but non-constitutive). Unlike them, autopoiesis theory specifically explains how this flux determines its own identity in contrast with its world-environment, shaping its own internal organisation according to a self-determining logic that is intrinsic to and constitutive of the specific modes of its own spatio-temporal and causal development: this is how the mind establishes boundaries and porous conditions of existence that at the same time are determined by and determine in a strong normative sense the individuality of the system, defining the

horizon of modifications and partial replacements that allows the precarious continuity of the whole through a continuous negotiation of the environmental contingencies (Varela et al. 1991; Varela 1997). Any such flow, according to the autopoietic principles of biological autonomy and operational closure, defines individual identity as an essential property that is real because constitutively rooted into precarious existence, i.e. a property that is strictly contingent upon, but not reducible to, the systemic specifications of its ever changing organic implementers, including for example the rich details of hormonal and immune systems dynamics (Stapleton and Froese 2015).

Ironically, the ship of Theseus scenario is occasioned by EM, not (as sometimes wrongly alleged) by MU: MU is about moving formal structures, not replacing material parts. Even hypothesizing that the mind were at its core just sheer functional organization and amounted to nothing more than the purely formal structure of the system (as required by MU), EM demands that such formal structure intrinsically belongs to and emerges over its own body, grounded in a stream of concrete interactions among material entities that exist in time and space. The patterns of these body-world interaction loops have a constitutive valence for the cognitive system but at the same time are merely relational in nature, i.e. situated, context-sensitive, non-exportable. Therefore, they are essentially irreplaceable in the unique way they are individuated in relation to neuronal and extra-cranial bodily interactions and to the beyond-the-skin world: that is why fEMp implies that the concrete instantiation of the mind in a contingent flow of material circumstances doesn't only define its functionality and phenomenology, but also its very conditions of ipseity and, therefore, the historically determined modes of its existence and persistence through time.

The theories that don't accept this tenet cannot legitimately claim to be embodied (e.g., "minimally embodied theories", see Gallagher 2015), as they would need to drop fEMp's fundamental requirement that the components of the mind are constitutively—and not only necessarily—embodied. In short, if fEMp directly clashes with MU it is not because MU usually implies Functionalism, but because it certainly implies Ultra-Functionalism, which requires that the constituents of the mind are *purely formal* and *individuated* at the same time. In the next section, abstracting from EM and generalizing the conclusions of the analyses conducted so far, I will explain why endorsing Ultra-Functionalism requires one to commit to a very demanding metaphysical interpretation of MRP. This interpretation should be unacceptable not only to EM, but any non-dualistic theory of the mind.

5 Multiply realizable angels: transcendence and individuation constraints

I have already mentioned that the MU theorists deny that M1 and M2 involve any implicit substance dualism. Our analysis of C, building on the comparison between MU and EM, gives us reasons to believe that their claim dramatically overlooks the demanding metaphysical notion of substrate independency that C implies. We have also clarified that the problem is not Functionalism per se: while Ultra-Functionalism somehow relies on a functionalist style of thinking, it is not included in any orthodox functionalist account of the mind. Classical MRP was originally proposed to counter Identity theory, and is generally associated to machine functionalism. It states that a single mental kind (property, state, event) can be realized by many distinct physical

kinds: different material substrates can realize functionally equivalent algorithms, i.e. functions that produce the same outputs every time the same inputs are fed to the system. Definitely, this principle is presupposed by many renditions of MU. But the substrate-independency presupposed by MU is theoretically more demanding than this modest functionalist principle. I call it "Ultra-MRP" to stress both the difference with Classical MRP and its association with Ultra-Functionalism: Classical MRP maintains that functional equivalence is sufficient evidence for qualitative identity; Ultra-MRP, in turn, maintains that functional equivalence can be sufficient evidence for numerical identity too, under the assumption that the relevant continuity between the realizers is in place. In the cases described by Chalmers 2010, this is the particular psychological, biological, or spatio-temporal continuity brought about by M1 and M2. Ultra-MRP is meant to assure us that, under this continuity conditions, one and the same mental function is preserved through MU.

Classical MRP and Ultra-MRP correspond to different interpretations of the type-token distinction. In the logic of Classical MRP, two numerically distinct physical states (x and y) can be either functionally equivalent or not (x = y or x ≠ y), which maps into two options: either their respective functional/formal structures (call them: Types) are qualitatively identical or not (X = Y or X ≠ Y). In turn, in the logic of Ultra-MRP, the functional equivalence of two numerically distinct physical states (x = y) can be interpreted in two different manners: the types of the two physical states can be considered either qualitatively identical but numerically distinct (Xi = Yi and Xii ≠ Yii), or qualitatively identical and numerically identical (Xi = Yi and Xii = Yii). In the logic of Ultra-MRP, there are two conceptually possible and distinct options for the re-instantiation of a mental state in a functionally equivalent way: *Copy* (preserves only the qualitative identity of the type) or *Transfer* (preserves also the numerical identity of the type). Now, what distinguishes between qualitative and numerical identity is exactly the fact that not only the type, but also the token, is the same, and this implies a type/token distinction. This distinction is replicated (second order) if the tokens we are considering are actually types of other tokens. That is why, in order to allow both *Copy* and *Transfer* as distinct possibilities, Ultra-MRP must presume that types (Xi and Yi) have individual identities like tokens (x and y), but independent from the tokens' identities. This entails that there are tokens of types (like Xi and Yi) and types of types (like Xii and Yii). So, in the *Copy* case, two tokens are identified as distinct instantiations of two types of the same type; in the *Transfer* case, the two tokens are identified as manifestations of one and the same type. This way, Ultra-MRP can in principle distinguish three possibilities: a certain physical state is functionally different from another one (x ≠ y); or it is functionally identical but numerically distinct from it (copy x = y with Xix = yYi and Xii ≠ Yii); or it is functionally and numerically identical to it (copy x = y with Xix = yYi and Xii ≠ Yii). But do types of formal structures have second-order types? In other words, are they individually identifiable at all?

The ramifications of the idea that formal types have individual identities of their own (irrespective of the particular contingencies that realize them) have been judged counterintuitive or even absurd (Hopkins 2012). Adapting Hauskeller's metaphor (2012, p. 191), one could say that according to Classical MRP a book, a comic book, or a movie can narrate a story in ways that are essentially equivalent. This notion is in itself unproblematic, but Ultra-MRP additionally claims that the content of the story is real, ontologically autonomous, and can be moved from one book to another, if the two

books include exactly the same words and some particular processual nexus between them (equivalent to MU) obtains: in a M1-like case, metaphorically speaking, the words written in one book are read and simultaneously erased, while they are concurrently rewritten into another book; in a M2-like case, two perfectly identical books are accessed and manipulated at the same time, but as soon as one word is read in one book the corresponding word in the other book is erased. In both examples, the expected outcome is that the identity of the original book (i.e., its intangible "content", not the material characters inscribed on the pages) is invisibly transferred from one book to another!

Classical MRP does not differentiate between transferring and copying a formal structure: this distinction seems a pleonasm—a distinction without a difference—when applied to abstract notions like patterns and information that hardly allow normative criteria to conceptually differentiate between reduplication and translation. But the Transfer/Copy distinction seems indispensable if the formal structures in question are the minds of people. That is why Ultra-MRP is needed if one presupposes, like MU does, that people are reducible to their brain activity (as per A) and brain activities are reducible to formal structures (as per B) because, given that people's minds are individually identifiable (i.e., numerable), then the formal structures of their minds must be individually identifiable entities too. The problem is that while people in flesh and bone are minds (form) implemented by bodies (matter), the purely formal structures manipulated by the MU procedure are devoid of any material detail. The contradiction implicit in Ultra-MRP is captured by these two conflicting constraints:

1) *Transcendence constraint:* the material realizers of the mind do not contribute in any way to identify its structural/formal components. If this were not the case, then MU would be impossible because the mind would not be substrate-independent (i.e., the relocation from one material substrate to another might significantly alter the structures of the mind);
2) *Individuation constraint*: the mind and its structural components are at any time numerically identifiable. If this were not the case, the concept of transfer would not apply to the uploaded minds, and it would be impossible to say that any formal structure has been transferred at all through the MU procedure.

The concurrent commitment to both constraints seems to violate the phenomenological evidence (epitomized by Aristotle's Principle of Individuation) that only physical stuff (material tokens) can be individually (i.e., numerically) identified, not abstract 'types' like algorithms, patterns, story plots, and mental contents considered per se, indepedently of the 'tokens' that realize them. If types are not material entities, then attributing them unique individual identities is a categorical mistake. But, if non-material entities like abstract types and patterns are not individuated, can they be transferred in space? Hopkins (2012), Corabi and Schneider (2012), and Hauskeller (2012) deny that formal mental functions (which are just abstract patterns types, distinct from their physical implementations) could be moved in space, as they lack individuation, which is a precondition for localization, which in turn is necessary to hold topological properties like continuous existence in space, extension, and finally movement.

This is not the first time that the problem of the individuation of immaterial objects is discussed in history of metaphysics. A systematic account of this millenary debate is beyond the scope of this paper. However, it might be useful to refer to a famous

medieval dispute (Pini 2012): Scotus and Aquinas debated how body-less persons like angels could have individual identities (e.g., Gabriel, Raphael, etc.) without violating the Aristotelian principle of individuation. They offered different answers. Aquinas proposed that every type ("species") of angel is instantiated by one and only one token ("individual") because, every type of non-material entities, like angels, *is* a token in itself (and *of* itself): therefore if one sights two angels that look exactly the same, then he can be sure that they actually are one and the same angel (numerical identity). Scotus, in turn, proposed that angels were made out of "spiritual matter", which allowed their individuation in space while preserving their non-physical status.

Mutatis mutandis, both theoretical options are currently employed to defend Ultra-MRP. If we agree with Aquinas that non-spatial formal structures are unique, and that every such structure identifies both a type and its unique token, then—given that minds are formal structures of this kind—MU can claim that every time we witness two or more instances of functionally identical minds we can be sure that they actually are one and the same mind. This option, called 'branching identity', has been developed by Cerullo (2015) and Wiley (2014), and leads to the somewhat counterintuitive conclusion that each and every outcome of a split-brain or mind-upload procedure *is still* the original person/mind, so long as they are qualitatively identical (with some tolerance). Therefore, (tolerant) qualitative identity is necessary and sufficient to have numerical identity. This option justifies Ultra-MRP, but implies the puzzling consequence that each and every cognitive clone (whether diachronically or synchronically existing) is just *one and the same person*: despite our intuition and in the face of their apparent separateness, identity can be branched across infinitely many bodies. Not only this claim clashes badly with our most basic understandings of identity (let alone personal identity); it also eliminates any space for the copy/transfer distinction that is a normative condition of MU. Eliminating this distinction is problematic because it makes MU either too inclusive (every Digi-Dave would be Dave, regardless of how many Digi-Daves we have created) or too restrictive (in principle, the creation of only one Digi-Dave would be possible. But it is not clear why, given that mind copy seems to be repeatable ad libitum[4]). Wiley and Koene (2015) invite a departure from the notions of transfer/copy to make sense of MU but, as a matter of fact, without the copy/transfer distinction, the outcomes of mind digitalization become unintelligible (and immortality much less attractive).

Scotus' option, on the other hand, implies that minds are made out of formal structures that can be individuated in themselves and by themselves, regardless of their contingent physical realizations, by reason of their extraordinary metaphysical status that—despite their immaterial nature—grants them by stipulation the same topological properties of materially instantiated objects. Like Scotus' angels, MU's uploaded minds enjoy a privileged status because, even if they are abstract formal patterns, they can still be individuated as if they were embedded in material contingencies and physical incidents. This option seems apparently less counterintuitive, but it clashes with naturalism, as it postulates immaterial entities that interact with the material world. This engenders problems of causal over-determination and frustrates all the basic requirements of standard scientific discourse.

[4] See note 3.

Of course, one could also try to avoid the Aristotelian principle of individuation altogether, proposing a third option. This seems the option explored by Chalmers (2012, p. 158), who (pressed on this point by Corabi and Schneider 2012) claims that non-extraordinary non-material structures can have topological properties too (and therefore can be individuated and moved in space): for example, institutions like universities can be relocated in different localities by means of a decree. However, there are reasons to find this example unpersuasive: first of all, minds are not social objects like universities (their existence is not contingent on conventions); second, while it is clear that the physical components of a university (buildings, employees) can move from a location to another, it is not evident how the abstract notion of institution could move in any relevant sense. Any attribution of topological properties to a non-material non-spatial entity leads to the reification (the attribution of a strong ontological status) of something that is abstract and merely relational, if not nominal, in nature (Hopkins 2012).

All of these three options are metaphysically demanding. They interpret substrate independency in a way that is more radical than the one presupposed by classical Functionalism: a mind must be substrate independent not only in the sense that different material substrates can realize its functional configuration (i.e., its essence), as per B; also in the sense, expressed by C, that its very identity as an individual (i.e., its existence) has nothing to do with the particular modes of its implementation. In other words, a mind subsists beyond the contingent circumstances of its material implementation. A particularly strong ontological commitment of this kind is required to assert that a functional configuration like mental contents or mental functions can be both transcendent and identity-preserving: not only its essence is virtually independent from material contingencies, as an abstract network of co-variance relationships between material things (as per Functionalism and Classical MRP); additionally, it exists as a self-contained self-sufficient entity that is real in itself and by itself, distinct and autonomous from its material realizers (as per Ultra-Functionalism and Ultra-MRP). Classical MRP only demands that a functional configuration is replicable throughout different material realizers (qualitative identity), while its existence still depends on the material substrates. For Ultra-MRP, in turn, formal structures are ontologically autonomous metaphysical realities: not only their functionality is independent of their material realizers; their existence too. Accordingly, MU presupposes the idea that, even if formal structures like minds are always occurrently 'hosted' by material vehicles that allow them to operate in the physical world, they have existences and individual identities of their own independently of such vehicles.

The immaterial ontology of Ultra-MRP is controversial. In order to fulfill both Transcendence and Individuation constraints it gets embarrassingly close to Cartesian dualism or Platonist forms realism. I have already explained why this reification of the mental "essences" is incompatible with EM's own ontological requirements; but this worry can be shared by other philosophical views, as the reification of "mental contents" reintroduces a dualism of substances and causes that clashes with scientific naturalism at large (Hutto and Myin 2013). That is why Pigliucci (2014) warns us that MU's metaphysics "should be unacceptable in modern philosophy". This implicit commitment to substance dualism is the most serious reason to be

skeptical about MU, one that is shared by theories like old school Functionalism and Biological Theory of Consciousness, among others. Certainly, EM is not the only theory that has valid reasons to oppose MU. But EM can reveal and defend these reasons much better than any generically biological or materialist theory of the mind: only EM, relying on the autopoietic/autonomic notion of individuality, offers the specific theoretical resources necessary to reveal and rebut the metaphysically demanding premises implicit in MU's notion of substrate independency. In the conclusive section I will explain why EM, more than any other philosophical approach to the mind, is committed to refute MU, and why this mission represents a critical challenge to EM.

6 The challenge

I have showed that MUka and fEMp frontally clash on the constitutive role of bodily implementation, and that their intrinsic incompatibility reveals the most problematic aspect of MU: the conflict of Transcendence and Individuation constraints that is implied by MUka. However, I have said nothing so far about whether or not EM ultimately provides sufficient arguments against MU. For example, a supporter of MU could still argue that there is nothing in C that violates the principles of biological autonomy and the autopoietic/enactive definition of identity of a cognitive system: after all, if an autopoietic system can be perfectly emulated by a computer, why can't we gradually transfer its formal structure to a computer (as in M1)? In that case, would the autopoietic/enactive theory have any a priori reason to reject an M3 scenario?

If M3 poses a challenge to theoretical biology, phenomenology seems to struggle with it too. From a Husserlian perspective, the core of one's identity is the sense of "mineness" that unifies all the conscious experiences of an embodied subject (Zahavi 2006). It is neither replaceable nor exportable, because it is inherent to an inalienable first-person stance. The identity of the minimal self is the orientation of a subjective perspective towards its concretely situated experience, which is shaped by the body and its intrinsic passivity, i.e. the uncontrollable resistance offered by the material contingencies to our intentional acts. Without this resistance, a "pure" disincarnated consciousness, like the one conceptually identified by Chalmers (2010) as the sufficient condition for the preservation of the Self, literally wouldn't have a place to exist. Which is why—according to EM—consciousness wouldn't have any place to go either, if it was not constitutively embodied as it is.

But one might wonder whether this is enough to exclude a priori M3. This is a challenging puzzle for the phenomenologists who endorse EM. "One can dream of oneself inhabiting a tiger's body, but one cannot dream being a tiger: my dream of being in a tiger's body is that of having all my experience, conditioned by my human embodiment, simply transplanted there" (Ganeri 2012, p. 118). But can I dream of being myself in a body that is identical to mine without actually being mine? If the artificial body under my control looks and feels exactly like the usual one, then what—if anything—can characterize it as other than not mine? Can

phenomenology reasonably argue that the idea of waking up one morning in a qualitatively identical but numerically different body is ultimately non-sensical?

The stake is very high. If EM couldn't provide a final argument against M3, then both the functional and the phenomenological accounts of embodiment would be undermined: if our mind is free to move from our body to an identical one, then a mind is just an autonomous, substrate-independent, formal structure. The concrete circumstances of the mind's embodiment would not play any constitutive role in determining the functional or phenomenal identity of the self. Consciousness could well be an abstract disincarnated category, a neutral container open to any sort of contents without being related to any of them. Mineness would float without an anchor, unable to keep together the bundle of one mind's intentional acts. EM as a whole would crumble. Does EM have sufficient arguments to prevent this?

Perhaps, the dialectics between the philosophical notions of essence and existence can guide the search for such arguments: *existence* characterizes the constitutive preconditions under which something is real and factually present in the world, as an individuated entity open to possibilities and risks; *essence* defines the necessary preconditions under which something can be identified as itself, something that is what it is and does what it does. MU sharply separates the essence of a mind from its existence, and attributes a primacy to the first over the latter, so that bodily implementation is merely instrumental to the existence of the mind, but doesn't define its preconditions or its individual identity. On the contrary, EM acknowledges that the functionality and the phenomenology of a mind can't be abstracted from its bodily implementation, because the essence of the mind itself depends on the concrete modes of its individuated existence, and the existence of the mind is grounded in its bodily contingencies. For EM, if a mind is numerically identifiable and therefore uniquely individuated in time and space it is only because its essence (functions and qualia) is grounded in material existence.

Obviously enough, also other approaches in philosophy of mind oppose MU. But EM epitomizes in the deepest manner the fundamental response against MU because it has been always inspired by the existentialist principles of humanistic phenomenology (e.g., Varela et al. 1991; Dreyfus 2002; Gallagher 2005a, etc.): in particular, EM incarnates the phenomenological/existentialist principle that "existence precedes essence" (Sartre 1946). According to this principle, acknowledging the constitutive situatedness of our condition provides us with all the freedom we might need or aspire to. The Transhumanists see freedom in a different way: in a gnostic vein, minds need to be "liberated" by their bodies, because humans are only accidentally trapped in mortal vessels. In contrast, EM maintains that bodies are not cages and minds are not angels, because they are so inherently entangled with one another that they have an individual personal identity only when they are together. If EM is right, and individuation is the price to pay in order to experience both the constraints and the freedom of real existence, then it is hard to believe that minds could ever migrate from their organic bodies to computers. But if we are not constitutively constrained by the existential cage in which we once fell, as MU ventilates, and our minds are only temporarily exiled from the transuranic world of pure essences, then maybe it is time to update our familiar categories for comprehending an unprecedented notion of freedom.

Given that the philosophical stake is so high, and that the future of humanity might depend on it, it will be crucial to see whether EM or MU will eventually prevail in the struggle for giving a sense to our mortal (and possibly immortal) life.

References

Blackford, R., & Broderick, D. (Eds.). (2014). *Intelligence unbound: The future of uploaded and machine minds*. Oxford: Wiley-Blackwell.

Cerullo, M. A. (2015). Uploading and branching identity. *Minds and Machines, 25*(1), 17–36.

Chalmers, D. (2010). The singularity: a philosophical analysis. *Journal of Consciousness Studies, 17*, 7–65.

Chalmers, D. (2012). The singularity: a reply to commentators. *Journal of Consciousness Studies, 19*, 141–67.

Chiel, H. J., & Beer, R. D. (1997). The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment. *Trends in Neurociences, 20*, 553–557.

Clark, A. (1997). *Being there: Putting brain, body and world together again*. Cambridge: MIT Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford: Oxford University Press.

Corabi, J., & Schneider, S. (2012). The metaphysics of uploading. *Journal of Consciousness Studies, 19*(7–8), 26–44.

Cosmelli, D., & Thompson, E. (2011). Embodiment or Envatment? Reflections on the bodily basis of consciousness. In J. Stewart, O. Gapenne, & E. Di Paolo (Eds.), *Enaction: Towards a new paradigm for cognitive science*. Cambridge: MIT Press.

Dreyfus, H. (2002). Intelligence without representation. *Phenomenology and the Cognitive Sciences, 1*, 367–383.

Gallagher, S. (2005a). *How the body shapes the mind*. New York: Oxford University Press.

Gallagher, S. (2005b). Metzinger's matrix: Living the virtual life with a real body. *Psyche: An interdisciplinary journal of research on consciousness*, 11(5).

Gallagher, S. (2011). Interpretations of embodied cognition. In W. Tschacher & C. Bergomi (Eds.), *The implications of embodiment: Cognition and communication* (pp 59–72). Exeter: Imprint Academic.

Gallagher, S. (2015). Invasion of the body snatchers: how embodied cognition is being disembodied. *The Philosophers Magazine*, (April 2015), 96–102.

Ganeri, J. (2012). *The self. Naturalism, consciousness, and the first-person stance*. Oxford: Oxford University Press.

Goertzel, B. (2012). When are two minds versions of one another? *International Journal of Machine Consciousness, 4*(01), 177–185.

Greenfield, S. (2012). The singularity: commentary on David chalmers. *Journal of Consciousness Studies, 19*, 112–118.

Hauskeller, M. (2012). My brain, my mind, and I: some philosophical assumptions of mind-uploading. *International Journal of Machine Consciousness, 4*(1), 187–200.

Hayworth, K. J. (2012). Electron imaging technology for whole brain neural circuit mapping. *International Journal of Machine Consciousness, 4*(1), 87–108.

Heersmink, R. (2015). Dimensions of integration in embedded and extended cognitive systems. *Phenomenology and the Cognitive Sciences, 14*(3), 577–598.

Hopkins, P. (2012). Why uploading will not work, or, the ghosts haunting transhumanism. *International Journal of Machine Consciousness, 4*(1), 229–243.

Hutto, & Myin. (2013). *Radicalizing Enactivism. Basic minds without content*. Cambridge, MA: MIT Press.

Kurzweil, R. (1999). *The age of spiritual machines*. New York: Viking Penguin.

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. New York, NY: Penguin Press.

Menary, R. A. (2010a). Cognitive integration and the extended mind. In R. A. Menary (Ed.), *The extended mind* (p. 227). Cambridge MA: MIT Press.

Menary, R. A. (2010b). Dimensions of mind. *Phenomenology and the Cognitive Sciences, 9*, 561–578.

Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Cambridge MA: Harvard University Press.

Noë, A. (2005). *Action in perception*. Cambridge MA: MIT Press.

Parfit, D. (1971). Personal identity. *Philosophical Review, 80*(1), 3–27.

Parfit, D. (1995). The unimportance of identity. In Harris (Ed.), *Identity* (pp. 13–45). Oxford: Oxford University Press.

Pigliucci M. (2014). Mind-uploading: A philosophical counter-analysis. In R. Blackford & D. Broderick (Eds), *Intelligence unbound: The future of uploaded and machine minds* (pp. 119–130). Oxford: Wiley-Blackwell.

Pini, G. (2012). The individuation of angels from Bonaventure to Duns Scotus. In Hoffmann T. (ed), *A companion to angels in medieval philosophy.* Leiden and Boston: Brill, 79–115.

Putnam, H. (1973). Meaning and reference. *Journal of Philosophy, 70*, 699–711.

Rupert, R. R. (2009). Embedded cognition and computation. In *Cognitive systems and the extended mind*. Oxford: Oxford University Press.

Sartre, J.-P. (1946). *L'existentialisme est un Humanisme*. Paris: Éditions Nagel.

Shapiro, L. (2011). *Embodied cognition*. New York: Routledge.

Stapleton, M., & Froese, T. (2015). The enactive philosophy of embodiment: From biological foundations of agency to the phenomenology of subjectivity. In J. I. Murillo, M. García-Valdecasas, & N. F. Barrett (Eds.), *Biology and subjectivity: Philosophical contributions to non-reductive neuroscience*. Cham: Springer.

Sutton, J. (2010). Exograms and interdisciplinarity: History, the extended mind and the civilizing process. In R. Menary (Ed.), *The extended mind* (pp. 189–225). Cambridge: MIT Press.

Thompson, E. (2005). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Harvard University Press.

Varela, F. (1997). Patterns of life: intertwining identity and cognition. *Brain and Cognition, 34*(1), 72–87.

Varela, F., Thompson, E., & Rosch, E. (1991). *The embodied mind. Cognitive science and human experience*. Cambridge MA: MIT Press.

Wheeler, M. (2010). In defense of extended functionalism. In R. Menary (Ed.), *The extended mind* (pp. 245–270). Cambridge MA: MIT Press.

Wiley, K. B. (2014). *A taxonomy and metaphysics of mind-uploading*. Los Angeles: Humanity+ Press and Alautun Press.

Wiley, K. B. & Koene R. A. (2015). *The fallacy of favoring gradual replacement mind uploading over scan-and-copy* (published on-line on Phillica.com and at http://arxiv.org/ftp/arxiv/papers/1504/1504.06320.pdf).

Zahavi, D. (2006). *Subjectivity and selfhood: Investigating the first-person perspective*. Cambridge MA: MIT Press.