Revised 1 July 2020. For *The Future of Freedom of Thought: Liberty, Technology, and Neuroscience*, (eds.) M. Blitz & C. Bublitz (Palgrave Macmillan, *forthcoming*)

Varieties of (Extended) Thought Manipulation

J. Adam Carter
University of Glasgow
adam.carter@glasgow.ac.uk

Abstract. Our understanding of what exactly needs protected *against* in order to safeguard a plausible construal of our 'freedom of thought' is changing. And this is because the recent influx of cognitive offloading and outsourcing—and the fast-evolving technologies that enable this—generate radical new possibilities for freedom-of-thought violating *thought manipulation*. This paper does three main things. First, I briefly overview how recent thinking in the philosophy of mind and cognitive science recognises—contrary to traditional Cartesian 'internalist' assumptions—ways in which our cognitive faculties, and even our beliefs, can be materially realised by as well as stored non-biologically and extracranially. Second, and taking brain-computer interface technologies (BCIs) and the associated possibility of 'extended' beliefs as a reference point, I propose and defend a sufficient condition on freedom-of-thought violating (extended) thought manipulation. On the view proposed, the right not to have one's thoughts or opinions manipulated is violated if one is (i) caused to acquire non-autonomous propositional attitudes (*acquisition manipulation*) or (ii) caused to have otherwise autonomous propositional attitudes non-autonomously eradicated (*eradication manipulation*). The implications of this view are then illustrated through four thought experiments, which map on to four distinct ways—what I call Type 1-Type 4 manipulation—in which, and with reference to the view defended, one's freedom of thought is plausibly violated.

1. A central platitude in legal and political philosophy, and which lies at the heart of many democratic constitutional systems, is that all individuals enjoy—in slogan form—the *freedom of thought*. Even if your actions are constrained by laws, your capacity to exercise your own mind as you wish is not equally constrained.

Kant ([1797] 1991) famously committed himself to this idea by defining the scope of juridical laws so as to exclude them from applying to the mind, insisting that juridical laws apply only to 'external actions.'[1] Other philosophers, like Mill ([1859] 1998), have defended the freedom of thought by pointing to the disutility of its absence: the

[1] For discussion, see Bublitz (2013, 241).

suppression of opinion thwarts a community's capacity to discover and maintain the truth.[2]

Outside of philosophy, a defence of the freedom of thought is enshrined explicitly in Article 18 of the Universal Declaration of Human Rights, which ensures that 'everyone has the right to freedom of thought, conscience and religion.' Elsewhere, in U.S. Constitutional legal scholarship, it is lauded by Supreme Court Justice Oliver Wendel Holmes as the principle that "most imperatively calls for attachment."

But even if the existence of a freedom so described is not controversial, things get thorny quickly when we zero in on what constitutes a plausible *violation* of it. This is especially so when we distinguish what is involved in violating one's freedom as pertains to (i) *expression of thought*; versus (ii) *the thought itself*. We can easily conceive of what it takes to violate (i) by looking to egregious examples of such violations—e.g., political persecution of minority opinions as expressed through religious and political demonstration and speech.

Question: But what would it be, exactly, to *violate* one's *freedom to simply form and possess her own thoughts* as opposed to express them, and to violate this freedom non-trivially? (A trivial way to violate *any* kind of freedom in thinking, categorically, would be to cause injury to the physical brain, injury to which is already legislated against as a paradigmatic *physical* harm.) Is the freedom to (in short) think as one wishes—at least on those matters on which it is possible when functioning normally to control thought[3]—something that could be violated any *other* way? And if not, then did we even need to make this freedom explicit in the first place?

2. It is tempting to think the answer to these questions is 'no'[4], given how pervasive the Cartesian picture of the mind, as a kind of private 'inner theatre', remains in ordinary thought and talk, as well as, implicitly, in legal and political thinking.[5] On the Cartesian view, according to which a thinker alone has privileged and exclusive *access*

[2] See, e.g., *On Liberty,* ([1859] 1998, Ch. 2).

[3] Even when paradigmatically free, our thinking is not entirely in our control—as philosophers have recognised in denying *doxastic voluntarism*, the view that (in short) we can believe what we desire to believe, and to do so directly without any intermediate steps in thinking. A simple kind of counterexample to doxastic voluntarism concerns perception. If there is a red table in front of you, and you desire to see a blue table and to immediately form the belief <*There is a blue table*>, you will not be able to do it. The denial of doxastic voluntarism is compatible with the thought that you have a kind of indirect control over (some) beliefs about what is true, which can be brought about by intentionally taking steps to acquire certain kinds of evidence. For some notable discussions of doxastic voluntarism and the philosophical issues surrounding it, see, e.g., Audi (2001), Clarke (1986), and Steup (2000).

[4] Perhaps one exception though is found in debates surrounding indoctrination in the philosophy of education. It's beyond the scope of what I can do to cover this here, but some relevant stances are found in Hand (2002; 2004), Gardner (2004), Hansson (2018), and Siegel (2004).

[5] For some discussion on this point, see Carter and Palermos (2016). See also Blitz (2010).

to the content of her own thoughts, thought itself is *in principle* unregulatable (apart from regulating against physical injury to the brain) and so there would seem to be no point to legislating it in a way that goes beyond regulating physical harm. We could at most, on the Cartesian view, attempt to regulate a thinker's thoughts *indirectly* by regulating (e.g., punishing) the *behaviour we take to be evidence of thought*.[6] However, and in line with Kant's thinking, these regulations themselves would be *de facto* regulations of (e.g., verbal and physical) *behaviour*, and not regulations of anything like the shape and character of thought *as such*.

But—as contemporary thinking in the philosophy of mind and cognitive science suggest—Descartes was wrong in (at least) two important ways about the 'inner' nature of the mind. First—as Putnam (1975), Kripke (1980), and Burge (1986) showed in the 1970s and 80s, it is mistaken to think that the *content* of our thoughts is *either* (i) transparent to us[7] or (ii) determined solely by the inner workings of the mind. *Content internalism* has since been rejected almost universally for *content externalism*, which holds that the content of our thoughts—viz., what our thoughts are *about*—is at least partly determined by facts about our physical and socio-linguistic environments that might be inaccessible to us on reflection.[8] For example, on this view, when you think about the wet, blue stuff that you see in oceans, whether you are thinking about *water* (which is type identical with H_2O) or about something *else* (as Putnam imagined: 'XYX'—viz., something which is very similar to water but which is not identical to H_2O), depends on what the physical environment you are interacting with is *actually like*, and this is something you might not have reflective access to while entertaining the image of the blue, wet stuff.[9]

More importantly for our purposes, though, a *second* kind of Cartesian doctrine about the mind's inner nature—*cognitive internalism*—has also fallen into disrepute, and has increasingly done so over the past 10 years.[10] Whereas *content* internalism concerned the *content* of thoughts—viz.., what your thought counts as being a thought *about*—*cognitive* internalism is a thesis about the kinds of things that *materially realise* cognition, viz., about the kinds of physical processes on which cognition supervenes.

[6] Alternatively, one might indirectly regulate thought by depriving another of information (or tools to generate information). For example, one might indirectly regulate a mathematician's ability to discover a certain result—and thus, to believe that result is true—by depriving her of a pencil and paper. Thanks to John Tillson for noting this other indirect form of thought regulation.

[7] See also Schwitzgebel (2008).

[8] See Carter et al. (2014).

[9] The denial of content externalism is closely related to a range of puzzles in the contemporary literature on self-knowledge. For discussion, see, e.g., Gertler (2000, 2010), Parent (2017), McKinsey (1991), and Pritchard (2002).

[10] See, e.g., Clark (2008), Menary (2007), Palermos (2011, 2014b), Wilson (2000, 2004). For criticism, see, e.g., Adams and Aizawa (2008).

Prior to Clark and Chalmers' landmark paper 'The Extended Mind' (1998), even most *content* externalists in the philosophy of mind were still *cognitive* internalists. They held that although one's physical and socio-linguistic[11] environment can partially determine the content of one's thoughts, only *intracranial* processes—i.e., biological processes that play out in brain—are the sorts of things that can materially 'bring about' cognitive processes like memory, reasoning, perception and the like.

But even this more basic kind of internalism about the mind is falling to the wayside. According to the *hypothesis of extended cognition* (HEC), our assessments of what kinds of things can feature in 'cognitive' process should be guided by commonsense functionalist thinking, rather than by considerations to do with physical makeup or special location. For example, according to the HEC proponent, if you are using a well-integrated smartphone to *do* what biomemory does—viz., to play the role biomemory normally plays in storing and retrieving information—then to the extent you have dispositional beliefs (i.e., which become occurrent beliefs when retrieved and brought to conscious awareness) stored in biomemory, you also have 'extended' dispositional beliefs stored in your phone's memory, or in the cloud.[12] This might seem radical, but to say otherwise, on this line of thought, commits one to an unprincipled kind of 'bioprejudice'[13] that gives arbitrary weight to material constitution and special location when demarcating the bounds of the cognitive.

3. Against this background, it should be obvious that the question of what it would be to violate one's freedom of thought—and not *just* her expression of thought in speech and action—can hardly be set aside as moot or purely theoretical. And this is because the latest cognitive science allows beliefs, memories, perceptions[14], and the like, to be materially realised by processes that *include parts of the world* which themselves are not in principle 'hidden away' in some Cartesian theatre, but subject publicly to various kinds of manipulation by other parties.[15]

And the picture is complicated further when we include recent advancements in, and the potential future of, *brain-computer interface* (BCI) technologies.[16] To make

[11] See, e.g., Burge's (1986) arthritis/tharthritis example.

[12] For discussion, see, e.g., Clark (2010, 2008), Carter and Kallestrup (2016, 2017), Carter and Pritchard (2020), Menary (2010), Pritchard (2010, 2018), and Palermos (2014a).

[13] The use of this term is due to Chalmers (2008).

[14] While HEC is often explained in terms of extended memory processes, the thesis also applies to perception, and this can be illustrated with reference to tactile visual substitution systems (TVSS). See, e.g., Bach-y-Rita (1983), Bach-y-Rita and Kercel (2003), and Palermos (2016).

[15] For a discussion of such violations and their potential legal ramifications, see Carter and Palermos (2016).

[16] For some representative recent developments in BCI technologies for use in cognitive enhancement, see, e.g., Ghafoor et al. (2019), Wang et al. (2019), and Pisarchik, Maksimenko, and Hramov (2019).

concrete the kinds of possibilities generated by BCIs, consider that in October 2019, a French dentist who had fallen 15 feet walked for the first time in two years using his mind to control an exoskeleton suit. The man who goes by the first name 'Thibault' has implants on his brain that read its activity and send this to a nearby computer, which in turn uses this information to send instructions to the exoskeleton. The result is that Thibault can, simply by thinking, control the limbs of the exoskeleton in three-dimensional space.[17]

The reasoning of Elon Musk, who has launched BCI start-up Neuralink (and other BCI startups such as BrainCo, Emotiv, Kernel, Mindmaze, NeuroSky, NeuroPro, Neurable, and Pandromics) if that we can use neural implants to communicate with computers in the *therapeutic* case—viz., where the aim is to restore an individual to normal, healthy levels of human functioning in order to correct disease and pathology[18]—why not use it to take already healthy individuals *beyond* normal levels of functioning, especially when it comes to *cognitive* functioning, where more sophisticated BCIs can in principle allow us to not only send but also *receive* information immediately through thought commands.[19]

To make this idea a bit more concrete, think about what you do when you say "Hey Google/Siri what's the weather today?" Moments later, Google/Siri tells you the answer. Now just imagine streamlining this process. You *think*, rather than verbalise "What's the weather?" And soon after, perhaps immediately, your brain receives the information[20] from the computer you've just communicated with via a thought command.

I am going to assume from here on in that these kinds of BCI enhancement technologies are worth taking seriously, even if they have not yet arrived fully functional. What is important for philosophical and legal thinking about the freedom of thought is whether we have a clear way to think about how to protect freedom of thought in connection with them when (if) they arrive.

4. Current international law frameworks recognise three key elements to one's freedom of thought which can be threatened in different ways by new technologies.[21] These are, as Susan Alegre (2017, 225) summarises: (i) the right not to reveal one's

[17] See Carter (2020b Ch. 1) for a recent discussion of this case.

[18] For discussion on the distinction between cognitive enhancement and mere therapeutic cognitive improvements, see, e.g., Bostrom and Sandberg (2009) and Carter and Pritchard (2019).

[19] See Musk (2019).

[20] Different possible BCIs might realise this operation differently, e.g., by prompting content representation and regulating attention via the implant; the assumption should be that these ways will trend in the direction of being increasingly seamless and non-obtrusive as BCI technologies continue to improve in the more distant future.

[21] For discussion, see Alegre (2017).

thoughts or opinions; (ii) the right not to have one's thoughts or opinions manipulated; (iii) the right not to be penalised for one's thoughts.

The advent of HEC bears directly on (i) and by extension (iii). HEC implies that, to the extent that your mind is partly located (in certain circumstances) in external memory storage, the right you have not to *reveal* your thoughts is a right that extends also to certain protections from inspection of such external storage.[22]

The rise of BCI technologies, by contrast, poses special challenges for (ii), and better understanding these challenges helps to equip us for future thinking about the freedom of thought. Or so I want to argue. Here is the plan for what follows. In §5 I propose by using illustrative BCI-style cases, a sufficient condition for freedom-of-thought violating 'extended' thought manipulation, viz., thought manipulation that involves some kind of distortion of a thinker's non-biological mental faculties.[23] Once this condition is set out and defended, I will, in the remaining sections, taxonomize four distinct varieties of freedom-of-thought-violating extended thought manipulation which have interestingly different structures, but which all satisfy the proposed sufficient condition.

5. Let's distinguish two kinds of cases where a thinker might be fitted with a BCI: *pre-arranged* cases and *non-pre-arranged* cases, e.g., where in the latter kind of case, one's being fitted with a BCI is not in accordance with one's past autonomous decisions.[24] For example: a person is unwillingly 'experimented on'.

Non-pre-arranged BCI cases constitute *trivial* violations of one's freedom against thought manipulation, however such freedom may be plausibly construed; but such cases are covered under the wider class of protections against physical harm and injury. What I want to suggest in what follows is that even *pre-arranged* BCI implementation cases can very easily serve as ones where a thinker's freedom against thought manipulation is violated. Appreciating this has potential practical import in a very possible future in which consenting to BCI fitting is a typical and common form of cognitive enhancement.[25]

[22] For an interesting recent take on this idea, see Riley v. California (2014), and in particular, John Roberts' majority opinion on the case, in which he draws comparisons between cell phones and human biological anatomy. http://www.supremecourt.gov/opinions/13pdf/13-132_8l9c.pdf For an overview in the context of the extended cognition debate, see Carter and Palermos (2016).

[23] Note that manipulation is distinct from coersion. For discussion on this difference, see, e.g., Baron (2003), cf., Ghafoor et al. (2019).

[24] I am setting aside for the purposes of discussion here issues to do with thought manipulation via *genetic* enhancement, or by testing and selecting for certain embryos; these cases, while interesting and important, are difficult to address without a foray into questions of personal identity that go beyond what I can cover here.

[25] For an influential defence of the idea that we can expect to increasingly incorporate BCIs, see Clark (2003).

More specifically, what I want to propose and then sharpen is the following *sufficient condition* on freedom-of-thought violating thought manipulation:

> **Thought Manipulation (Sufficiency) (TMS):** The right not to have one's thoughts or opinions manipulated is violated if one is (i) caused to acquire non-autonomous propositional attitudes (*acquisition manipulation*) or (ii) caused to have otherwise autonomous propositional attitudes non-autonomously eradicated (*eradication manipulation*).

Regarding the *acquisition manipulation* component of (TMS): a term that needs clarified is that of a *non-autonomous propositional attitude*.[26] Examples of propositional attitudes are beliefs and desires, e.g., your belief that Paris is the capital of France, your desire that you not eat liver for dinner this evening. Following influential work on autonomous attitudes by Al Mele (2001), I am going to assume that, sufficient for a propositional attitude's *not* being autonomous, and thus, not being such that it is properly attributable[27] to the agent, is the conjunction of two conditions: (i) a *bypass condition*—viz., a condition pertaining to whether the attitude in question was acquired in a way that 'bypassed' the subject's relevant (e.g., cognitive and conative) faculties[28]; and (ii) an *unsheddability condition*—viz., a condition pertaining to whether the subject is able to (easily enough) give up, or at least attenuate the strength of, the relevant attitude.[29] Regarding the *eradication manipulation* component of (TMS). To unpack this further, say that an otherwise autonomous propositional attitude is caused to be *non-autonomously eradicated* if it is caused to be either (a) *shed* (e.g., to go out of existence, or to decrease in severity) or (b) blocked from manifesting in ways that relevantly bypass a thinker's cognitive and conative faculties. The core idea of TMS is, in sum, that your freedom of thought is violated if you're caused to either acquire an unsheddable attitude that your own faculties played no

[26] Note that, on the proposed account—which states just a sufficiency condition and not a necessary condition—it's entirely possible that the right not to have one's thoughts or opinions manipulated could be violated *non-propositionally* as well, e.g., via the compromise of faculties or dispositions in such a way as to leave all representational content as is. For the purpose of this paper, I'm keeping my focus on propositional manipulation; an interesting and relevant question for further work concerns the matter of freedom-of-thought manipulation via one's dispositions directly. Thanks to John Tillson for discussion on this point.

[27] I am using attributability here in the sense of Watson (1996) as denoting 'character revealing'. Your striking someone as a result of being pushed into that person is, for example, is not properly attributable to you, as it in no way reveals your character—viz., your stable dispositions of mind.

[28] See Carter (2020b Ch. 2) for a detailed discussion of different ways to interpret this condition.

[29] For alternative ways of thinking about attitudinal autonomy, see, e.g., Dworkin (1981) and Frankfurt (1988). For developments of a Mele-style approach to attitudinal autonomy—an approach which denies that attitudinal autonomy is entirely a matter of one's present psychological structure and can also include such things as the attitude's history—see, e.g., Weimer (2009) and Carter (2020b, Ch. 2, 2020a).

role in acquiring or are caused to shed (or block) an attitude that your own faculties played no role in your shedding.

A simple and egregious case of *acquisition manipulation* is having beliefs or desires 'implanted' in a clandestine fashion. A simple and egregious case of *eradication manipulation* is having beliefs or desires 'wiped' in a clandestine fashion. But these are just 'limit' cases; what's more interesting (as we'll see in §6) are the less egregious but nonetheless morally and epistemically significant violations.

A final point of clarification: (TMS), it should be emphasised, does not imply that if a subject had an implanted belief or desire that *was* sheddable, then it would thereby *not* constitute a violation of her freedom against acquisition manipulation. This is because (TMS)—both its acquisition and eradication clauses—offers sufficiency conditions but not necessity conditions on freedom-of-thought violating thought manipulation. However, even as a (disjunctive) sufficiency condition for attitudinal acquisition and eradication manipulation, TMS is of philosophical interest. As we'll see in the next section—using some BCI-based thought experiments—any plausible right we have against thought manipulation can be violated (with reference to the clauses in TMS) in crucially different *kinds* of ways, which map on to (at least) four interestingly different 'varieties' of freedom-of-thought violating thought manipulation.

6. Let's now consider the following cases:

> **Case 1:** Otto, due to gradually failing biomemory, asks to be fitted with a sophisticated 'Neuralink Memory-Pro' brain computer interface that will help 'pick up the slack' where his memory is failing, when it comes to scheduling and organising his life.[30] The BCI is designed so that when Otto learns something he wants to include in his calendar, the information goes, via a thought command, rather than to his biomemory, straight to the BCI's cloud storage (e.g., much like a Google Calendar). When he attempts to recall old information from the Memory-Pro, he receives the information that is stored. For Otto, the Neuralink Memory-Pro plays the role that biomemory, pen-and-paper, as well as manually operated computers used to play for structuring his day. Unbeknownst to Otto, the Memory Pro's software update has now automatically 'auto-integrated' national and bank holiday dates into Otto's cloud storage.

Assessment: Otto's initial and consensual fitting of the Neuralink Memory-Pro violated no thought-based right of his. However, the software update *did*. The reason, with reference to TMS, is that the update causes him to acquire non-autonomous

[30] This case is a twist on Clark and Chalmers' (1998) case of 'Otto', which they use to motivate the extended mind.

(extended) propositional attitudes (e.g., national and bank holiday dates). For classification purposes, let's call Case 1 a **Type 1 case** for the following reason: freedom against thought manipulation is violated due to the acquisition of a non-autonomous belief in such a way that *no faculty (cognitive or conative) was exercised whatsoever* in the acquisition of the (extended) belief; in other words, his faculties have been *fully bypassed* in the course of propositional attitude acquisition.[31] As we'll see in Case 2, freedom against thought manipulation can be violated (with reference to the acquisition manipulation clause in TMS) even when faculties are only *partially* bypassed.

> **Case 2:** Everything is the same as with Case 1, except for some of the details about the nature of the Neuralink Memory-Pro's update and Otto's knowledge about it. First, the update is much more extensive, in that it inserts (along with bank holidays) various other kinds of information, which will continuously be added, including on the basis of algorithmic suggestions synced from his other devices (e.g. "Stores open late in your area tonight for Black Friday shopping" … "This Sunday, new Netflix WWI documentary available", etc.). Due to high demand for the product, Otto is given an impossibly brief period of time to decide whether to opt-in or out-out of the update, not long enough for him to understand what kinds of things it will include (and how they're included), and he's provided no further information from the company. On good faith, Otto opts in, and is soonafter baffled by what seem to be his own beliefs (and by extension, plans), and he begins losing his grip on what he had intentionally stored and what was prompted by the Memory-Pro's algorithms.[32]

Assessment: Unlike Case 1, it's not the case that Otto's acquisition of the algorithmically generated information inserted in his Memory-Pro via the update *completely*

[31] This includes no such exercise of a cognitive faculty in the past, as would be the case, for example, if one prearranged to have bank dates auto-inserted via the update at a future date.

[32] It's worth registering an important difference between the kind of situation described in Case 2 (a genuine acquisition manipulation case) with a superficially similar situation depicted in the science fiction show *Almost Human*. In that show—set in a cyborg future—it is common for individuals to see *personalised* hologram advertising, which targets the individual user. For example, while walking to the store, you might see a hologram on the side of a building which appears there (keyed to your GPS) to target you specifically, on the basis of sophisticated algorithms. In such a case, you would be—like Otto in Case 2—'bombarded' with content it would be very easy to 'uptake', and further, in both cases, you are cognitively influenced. But the *Almost Human* situation is not a genuine case of acquisition manipulation (of either Type 1 or Type 2) because, in this case, your autonomy is being respected; you are *nudged*, but not *caused* to uptake or endorse anything that features in the aggressive hologram-style advertising. However, the situation is different in Case 2 (as well as in Case 1) where acquisition manipulation is present.

bypassed his faculties. (He was after all informed that there would be some updates; he understood this much and consented to the update in so far as he understood it, which was limited given his unusually restricted opportunities). Nonetheless, this is a case where, with reference to TMS, the update causes him to acquire non-autonomous (extended) propositional attitudes and in doing so violates his freedom against acquisition manipulation. For classification purposes, let's call this a **Type 2 case** for the following reason: freedom against thought manipulation is violated due to the acquisition of non-autonomous (extended) beliefs (like Case 1), where these extended beliefs are non-autonomous *not* because (as in Case 1) their acquisition bypasses faculties altogether, but because it bypasses suitable *opportunities to exercise* those faculties.[33]

In sum: whereas Type 1 acquisition manipulation involves acquiring attitudes in ways that *completely* bypass the thinker's faculties, Type 2 acquisition manipulation involves acquiring attitudes in ways that bypass suitable *opportunities* to exercise those faculties, even if not bypassing the faculties wholesale.

> **Case 3:** Everything is the same as with Case 1, with a few important exceptions. The Neuralink Memory-Pro's creators, inspired by the efficacy in 'strategic forgetting'[34] demonstrated by deep neural networks and reinforcement learning techniques at Google DeepMind, have introduced an algorithm in the latest update that *deletes* information stored in the Memory-Pro deemed to be 'clogging' up the system. This includes, for example, information about plans that have been canceled or superseded by other plans. It also includes information stored in the Memory-Pro that is both flagged by the algorithm as 'unimportant details' (e.g., the weather back on 5 May 2024) and which has gone long enough without being retrieved. The update's functions, including how the algorithm targets information for deletion, are not made suitably explicit to users.

Assessment: With reference to (TMS), Case 3 is a case of *eradication manipulation* rather than *acquisition manipulation*. Recall that, on (TMS), an otherwise autonomous propositional attitude is caused to be *non-autonomously eradicated* if it is

[33] Plausibly, after all, your faculties are 'bypassed' in the acquisition of an attitude in a way that is relevant to whether the attitude is autonomous if you acquire it without suitable opportunity to exercise those faculties. (By way of comparison: An otherwise non-autonomous attitude whose acquisition bypasses one's faculties wouldn't be 'converted' into an autonomous attitude simply were it the cases that one was able to exercise one's faculties in unsuitable circumstances in coming to acquire the attitude.) For a detailed discussion of this issue, framed in terms of competences rather than faculties, see Carter (2020b Ch. 2).

[34] See, e.g., Beierle and Timm (2019) and Silver et al. (2017).

caused to be *shed* (e.g., to go out of existence, or to decrease in intensity) in ways that relevantly bypass your cognitive and conative faculties. In this case, Otto's faculties have been bypassed precisely because he lacks an explanation for how the algorithm is targeting stored information. Call this kind of case—where the mechanisms of memory eradication (as opposed to acquisition) are opaque to one—a **Type 3 case**. The difficulty of legislating Type 3 cases, it is worth noting, is already evidenced in recent debates following the 2018 GDPR (Art. 22, 13-15, Recital 71) about a data subject's 'right to an explanation', when purely algorithmic decisions are used to make decisions that affect someone's interests.[35]

> **Case 4:** After years of enjoying his Neuralink Memory-Pro BCI, Otto wants 'the next big thing', which is Neuralink's 'i-Connect' BCI device, which promises to help a thinker better 'organise one's mind'. The device's key trick is to use semantic tagging to sort information committed via thought command to information storage into compartments. Algorithms are then run on specific compartments in order to 'connect' information a thinker might not have connected themselves, which is then 'suggested' to the user on the basis of retrieval cues. The i-Connect promises, for example, to help users make better decisions on issues ranging from whom to trust (e.g., by storing track-record information) to which things to do to best relax. The suggestions made by the i-Connect, however, interfere with a thinker's own natural capacities for insight and creativity. In particular, the i-Connect does this by (albeit, inadvertently) blocking the efficacy of 'incubation' in insight problem solving tasks.[36]

Assessment: Let's assume, *ex hypothesi*, that Otto is fully aware of what kind of information the i-Connect enables him to acquire and even how it does this, such that the case is not, with reference to (TMS), a case of *acquisition manipulation*; in short, in Case 4, the 'bypass' condition on acquisition manipulation is not met *ex hypothesi*. That said, with reference to (TMS), Case 4 *is* a case of *eradication manipulation*. But this is *not* due (as in Case 3) to the 'shedding' proviso on eradication manipulation but due to the 'blocking' proviso. In Case 4, various insights Otto would have had have been effectively, even if not by intentional design, 'blocked'. Having insights blocked needn't violate a plausible freedom against thought manipulation (with reference to TMS's eradication manipulation component) if such blocking *itself* did not relevantly

[35] For discussion, see Goodman and Flaxman (2017) and Selbst and Powles (2017). For criticism that the GDPR can reasonably be interpreted as insuring a 'right to an explanation' on the part of data subjects when purely algorithmic decisions are made that affect their interests, see Wachter, Mittelstadt, and Floridi (2017).

[36] Sternberg and Davidson (1995), Metcalfe and Wiebe (1987), and Carter (2017).

bypass cognitive and conative faculties. In Case 4, though, it does. Call this, accordingly, a **Type 4 case**: a case of eradication manipulation that qualifies as such, with reference to TMS, via 'blocking' rather than via 'shedding'.

7. Extended though manipulation of Types 1-4[37] hardly exhaust possible categories. In fact, we can imagine subcategories of several of these, which map on to, e.g., partial or total bypassing, partial or total shedding, partial or total blocking, etc.

The aim of the above taxonomy is to reveal a few of the salient contrast points when it comes to *violations* of a plausible freedom against thought manipulation—viz., one that is framed (as TMS is) in terms of a freedom against (at least) the caused acquisition of *non-autonomous attitudes* and against the *non-autonomous eradication* of (would-be) autonomous attitudes.

As we continue to develop new technologies that make thought manipulation possible in new ways—including (and in addition to BCIs) various kinds of brain 'implants'[38] along with potential new breakthroughs in research on artificial neurons[39] and deep brain stimulation[40]—it becomes more important to anticipate and understand varieties of thought manipulation that such technologies enable. The above is an attempt at engaging in this kind of anticipation.

Further work in (extended) thought manipulation will go beyond the kind of sufficient condition (TMS) advanced here in order to make progress *vis-à-vis* the articulation of conditions *necessary* as well as sufficient for extended thought manipulation, a project more ambitious than what I've set out to do here.[41]

[37] I've intentionally illustrated the varieties of thought manipulation I have by using BCIs. This is because BCIs—even if less practically applicable today than smartphones, through which thought manipulation is also in principle possible—allow us to frame these kinds of manipulation in an particularly sharp way. It is worth noting though that BCIs aren't *necessary* for thought manipulation. If the extended mind and cognition theses (see §§2-3) hold water, and one's mind supervenes partly on a thinker's extracranial environment, the ingredients are present to manipulate 'extended' thought. See Carter and Palermos (2016) for discussion on this point.

[38] For discussion of research on the implantation of 'false memories', see, e.g., Ramirez et al. (2013). See also Carter (2020b Ch. 1).

[39] See, e.g., Simon et al. (2015).

[40] See, e.g., Suthana and Fried (2014) and Flöel et al. (2008).

[41] Thanks to Marc Blitz, Christoph Bublitz, John Tillson, and Ruaridh Gilmartin for helpful comments on a draft of this paper.

REFERENCES

Adams, Fred, and Ken Aizawa. 2008. *The Bounds of Cognition*. Blackwell.

Alegre, Susie. 2017. "Freedom of Thought, Belief and Religion and Freedom of Expression and Opinion." In *Human Rights of Migrants in the 21st Century*, 72–77. Routledge.

Audi, Robert. 2001. "Doxastic Voluntarism and the Ethics of Belief." *Knowledge, Truth, and Duty*, 93–111.

Bach-y-Rita, Paul. 1983. "Tactile Vision Substitution: Past and Future." *International Journal of Neuroscience* 19 (1-4): 29–36.

Bach-y-Rita, Paul, and Stephen W Kercel. 2003. "Sensory Substitution and the Human–Machine Interface." *Trends in Cognitive Sciences* 7 (12): 541–46.

Baron, Marcia. 2003. "Manipulativeness." In *Proceedings and Addresses of the American Philosophical Association*, 77:37–54. 2.

Beierle, Christoph, and Ingo J Timm. 2019. "Intentional Forgetting: An Emerging Field in Ai and Beyond." *KI-Künstliche Intelligenz: Vol. 33, No. 1*.

Blitz, Marc Jonathan. 2010. "Freedom of Thought for the Extended Mind: Cognitive Enhancement and the Constitution." *Wis. L. Rev.*, 1049.

Bostrom, Nick, and Anders Sandberg. 2009. "Cognitive Enhancement: Methods, Ethics, Regulatory Challenges." *Science and Engineering Ethics* 15 (3): 311–41.

Bublitz, Jan-Christoph. 2013. "My Mind Is Mine!? Cognitive Liberty as a Legal Concept." In *Cognitive Enhancement*, 233–64.

Burge, Tyler. 1986. "Individualism and Psychology." *Philosophical Review* 95 (January): 3–45.

Carter, J. Adam. 2017. "Virtuous Insightfulness." *Episteme* 14 (4): 539–54.

———. 2020a. "Epistemic Autonomy and Externalism." In *Epistemic Autionomy*, edited by Jonathan Matheson and Kirk Lougheed.

———. 2020b. *The Future of Knowing: Knowledge, Radical Enhancement, and Epistemic Autonomy*.

Carter, J. Adam, and Jesper Kallestrup. 2016. "Extended Cognition and Propositional Memory." *Philosophy and Phenomenological Research* 92 (3): 691–714.

———. 2017. "Extended Circularity." In *Extended Epistemology*, edited by J. Adam Carter, Andy Clark, Jesper Kallestrup, S. Orestis Palermos, and Duncan Pritchard. Oxford: Oxford University Press.

Carter, J. Adam, Jesper Kallestrup, S. Orestis Palermos, and Duncan Pritchard. 2014. "Varieties of Externalism." *Philosophical Issues* 24 (1): 63–109.

Carter, J. Adam, and S Orestis Palermos. 2016. "Is Having Your Computer Compromised a Personal Assault? The Ethics of Extended Cognition." *Journal of the American Philosophical Association* 2 (4): 542–60.

Carter, J. Adam, and Duncan Pritchard. 2019. "The Epistemology of Cognitive Enhancement." In *The Journal of Medicine and Philosophy*, 44:220–42. 2.

———. 2020. "Extended Entitlement." In *New Essays on Entitlement*, edited by Peter Graham and Nikolaj J. L. L. Pederson. Oxford: OUP.

Chalmers, David. 2008. "Foreword to Andy Clark's Supersizing the Mind." *A. Clark, Supersizing the Mind: Embodiment, Action, and Cognitive Extension*, 000–000.

Clark, Andy. 2003. *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford University Press.

———. 2008. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension: Embodiment, Action, and Cognitive Extension*. Oxford University Press.

———. 2010. "Memento's Revenge: The Extended Mind, Extended." In *The Extended Mind*, edited by Richard Menary, 43–66. MIT Press.

Clark, Andy, and David Chalmers. 1998. "The Extended Mind." *Analysis* 58 (1): 7–19.

Clarke, Murray. 1986. "Doxastic Voluntarism and Forced Belief." *Philosophical Studies* 50 (1): 39–51.

Dworkin, Gerald. 1981. "The Concept of Autonomy." *Grazer Philosophische Studien* 12: 203–13.

Flöel, Agnes, Nina Rösser, Olesya Michka, Stefan Knecht, and Caterina Breitenstein. 2008. "Noninvasive Brain Stimulation Improves Language Learning." *Journal of Cognitive Neuroscience* 20 (8): 1415–22.

Frankfurt, Harry G. 1988. *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press.

Gardner, Peter. 2004. "Hand on Religious Upbringing." *Journal of Philosophy of Edu-*

cation 38 (1): 121–28.

Gertler, Brie. 2000. "The Mechanics of Self-Knowledge." *Philosophical Topics* 28 (2): 125–46.

———. 2010. *Self-Knowledge*. Routledge.

Ghafoor, Usman, Amad Zafar, M Atif Yaqub, and Keum-Shik Hong. 2019. "Enhancement in Classification Accuracy of Motor Imagery Signals with Visual Aid: An fNIRS-Bci Study." *International Conference on Control Robot System Society*, 1201–6.

Goodman, Bryce, and Seth Flaxman. 2017. "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation'." *AI Magazine* 38 (3): 50–57.

Hand, Michael. 2002. "Religious Upbringing Reconsidered." *Journal of Philosophy of Education* 36 (4): 545–57.

Hand, Michael, Jim Mackenzie, Peter Gardner, and Charlene Tan. 2004. "Religious Upbringing: A Rejoinder and Responses." *Journal of Philosophy of Education* 38 (4): 639–62.

Hansson, Lena. 2018. "Science Education, Indoctrination, and the Hidden Curriculum." In *History, Philosophy and Science Teaching*, 283–306. Springer.

Kant, Immanuel. (1797) 1991. *The Metaphysics of Morals, Translated by M. Gregor*. Cambridge: Cambridge University Press.

Kripke, Saul. 1980. *Naming and Necessity*. Harvard University Press.

McKinsey, Michael. 1991. "Anti-Individualism and Privileged Access." *Analysis* 51 (1): 9–16.

Mele, Alfred R. 2001. *Autonomous Agents: From Self-Control to Autonomy*. Oxford University Press on Demand.

Menary, Richard. 2007. *Cognitive Integration: Mind and Cognition Unbounded*. Springer.

———. 2010. *The Extended Mind*.

Metcalfe, Janet, and David Wiebe. 1987. "Intuition in Insight and Noninsight Problem Solving." *Memory & Cognition* 15 (3): 238–46.

Mill, John Stuart. (1859) 1998. *On Liberty and Other Essays*. Oxford: Oxford University Press USA.

Musk, Elon. 2019. "An Integrated Brain-Machine Interface Platform with Thousands of Channels." *Journal of Medical Internet Research* 21 (10): e16194.

Palermos, S. Orestis. 2011. "Belief-Forming Processes, Extended." *Review of Philosophy and Psychology* 2 (4): 741–65.

———. 2014a. "Knowledge and Cognitive Integration." *Synthese* 191 (8): 1931–51.

———. 2014b. "Loops, Constitution, and Cognitive Extension." *Cognitive Systems Research* 27: 25–41.

———. 2016. "The Dynamics of Group Cognition." *Minds and Machines* 26 (4): 409–40.

Parent, T. 2017. "Externalism and Self-Knowledge." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2017. https://plato.stanford.edu/archives/fall2017/entries/self-knowledge-externalism/; Metaphysics Research Lab, Stanford University.

Pisarchik, Alexander N, Vladimir A Maksimenko, and Alexander E Hramov. 2019. "From Novel Technology to Novel Applications: Comment on 'an Integrated Brain-Machine Interface Platform with Thousands of Channels' by Elon Musk and Neuralink." *Journal of Medical Internet Research* 21 (10): e16356.

Pritchard, Duncan. 2002. "McKinsey Paradoxes, Radical Scepticism, and the Transmission of Knowledge Across Known Entailments." *Synthese* 130 (2): 279–302.

———. 2010. "Cognitive Ability and the Extended Cognition Thesis." *Synthese* 175 (1): 133–51.

———. 2018. "Extended Epistemology." *Extended Epistemology*, 90–104.

Putnam, Hilary. 1975. "The Meaning of 'Meaning'." *Minnesota Studies in the Philosophy of Science* 7: 131–93.

Ramirez, Steve, Xu Liu, Pei-Ann Lin, Junghyup Suh, Michele Pignatelli, Roger L Redondo, Tomás J Ryan, and Susumu Tonegawa. 2013. "Creating a False Memory in the Hippocampus." *Science* 341 (6144): 387–91.

Schwitzgebel, Eric. 2008. "The Unreliability of Naive Introspection." *Philosophical Review* 117 (2): 245–73.

Selbst, Andrew D, and Julia Powles. 2017. "Meaningful Information and the Right to Explanation." *International Data Privacy Law* 7 (4): 233–42.

Siegel, Harvey. 2004. "Faith, Knowledge and Indoctrination: A Friendly Response to Hand." *Theory and Research in Education* 2 (1): 75–83.

Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, et al. 2017. "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm." *arXiv Preprint arXiv:1712.01815*.

Simon, Daniel T, Karin C Larsson, David Nilsson, Gustav Burström, Dagmar Galter, Magnus Berggren, and Agneta Richter-Dahlfors. 2015. "An Organic Electronic Biomimetic Neuron Enables Auto-Regulated Neuromodulation." *Biosensors and Bioelectronics* 71: 359–64.

Sternberg, Robert J, and Janet E Davidson. 1995. *The Nature of Insight*. The MIT Press.

Steup, Matthias. 2000. "Doxastic Voluntarism and Epistemic Deontology." *Acta Analytica* 24: 25–56.

Suthana, Nanthia, and Itzhak Fried. 2014. "Deep Brain Stimulation for Enhancement of Learning and Memory." *Neuroimage* 85: 996–1002.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 2017. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." *International Data Privacy Law* 7 (2): 76–99.

Wang, Wenjie, Yuan Liu, Zhicai Li, Zhuang Wang, Feng He, Dong Ming, and Dapeng Yang. 2019. "Building Multi-Modal Sensory Feedback Pathways for Srl with the Aim of Sensory Enhancement via Bci." In *2019 Ieee International Conference on Robotics and Biomimetics (Robio)*, 2439–44. IEEE.

Watson, Gary. 1996. "Two Faces of Responsibility." *Philosophical Topics* 24 (2): 227–48.

Weimer, Steven. 2009. "Externalist Autonomy and Availability of Alternatives." *Social Theory and Practice* 35 (2): 169–200.

Wilson, Robert A. 2000. *The Mind Beyond Itself*. New York: Oxford University Press.

———. 2004. *Boundaries of the Mind: The Individual in the Fragile Sciences-Cognition*. Cambridge University Press.