國立陽明交通大學心智哲學研究所
Institute of Philosophy of Mind & Cognition
National Yang Ming Chiao Tung University

台北市北投區立農街 **2** 段**155** 號
TEL：**(02)2826-7000 ext. 65041**
FAX：**(02)2823-2920**

5 January 2024

Report on Przemysław Zawadzki's doctoral thesis:
"The Ethics of Neuromodulation of the Self: Personal Identity, Authenticity, Autonomy and Moral Responsibility in the Light of Neurointerventions"

It was a pleasure to be given the opportunity to review Mr Zawadski's doctoral thesis, even though, this being a thesis by publication, I was already familiar with some of its contents, having previously read several of the published articles featured in it. In the form of a series of nine such articles (four of which co-authored), the thesis discusses various ethical issues raised by the use – chiefly for therapeutic purposes – of direct interventions into the human brain, also known as "neuro-interventions". It pursues the commendable goal of combining a philosophical analysis of those issues with the latest knowledge from related empirical disciplines like psychology. I found the author's analysis throughout the thesis to be detailed, systematic, creative and thought-provoking.

Overview of the main ideas presented in the thesis

In the first part of the thesis (articles 1-4), the author looks to offer "a holistic and consensual explanation of the effects of DBS [Deep Brain Stimulation] on the selves of patients", and to help develop "a uniform procedure for evaluating the implications of DBS in individual clinical cases". The second part of the thesis (articles 5-9) focuses on philosophical issues raised by the advent of new methods for altering human memory, in particular optogenetics. These issues revolve around concepts like identity, authenticity, autonomy and moral responsibility.

The main claims and arguments put forward in the thesis include the following:

- The impact of DBS on the "selves" of mental patients is a major topic of discussion in the contemporary neuroethics literature. The relevant dimensions at risk from such an impact relate to personal identity, autonomy, and authenticity (articles 1 and 2).
- Despite being highly influential, the narrative approach to the self (especially in the form inspired by Marya Schechtman) is inadequate, as it is narrow and unidimensional, failing for instance to acknowledge that people with no episodic memory (and therefore no self-narrative) can still have a sense of self (art. 1). Furthermore, the approach neglects other aspects of the self and human personality, such as those posited by Dan McAdams's three-level model of personality, which provides a sounder empirical basis for philosophical reflection on this topic (art. 6).
- The pattern theory of the self, advocated by Dings and de Bruin following Gallagher, is on the right track when it comes to developing the comprehensive model of the self needed in neuroethics. Yet in its current state, it is still unsatisfactory. In particular, it should incorporate specifically moral aspects of the self, namely autonomy, authenticity, and moral responsibility. These moral aspects provide "the lens through which a self-pattern constituting a particular self may be evaluated" (art. 2).

- People's self-narratives are the channel through which other aspects of the self-pattern get expressed (art. 2). They also provide a unifying mechanism for various dimensions of personal agency (art. 3). Because the patient has unique access to the final product of the dynamical interactions of aspects that build his self-pattern, his own perspective on whether DBS altered any aspects of his self deserves some degree of priority over others (art. 2).
- The "Public Health Quarantine Model" (PHQ) is better suited than the retributivist approach to legal punishment to dealing with mental patients who commit crimes, but lack compatibilist capabilities (such as rational control, autonomy, and reason-responsiveness). These capabilities can help determine what measures are required to protect members of the public, and what forms of treatment are appropriate for the patient – but they should not be used to justify the attribution of basic desert moral responsibility (BDMR) and legal punishment (art. 4). Rather, the temporary restraint, and ultimate rehabilitation of such offenders are the solutions recommended by this approach.
- The new technology of optogenetics can be used to alter human memories, holding promise for trauma victims, as well as patients with depression and other disorders. However, it also raises ethical dilemmas that deserve greater attention from philosophers (articles 5 and 6). Some such dilemmas are shared with other memory modification technologies (MMTs). These include a patient's loss of the opportunity to recover from a trauma on their own, e.g. by constructing a self-narrative of redemption that promotes posttraumatic growth, higher well-being, and pro-social activities (articles 5, 6 and 7); threats to the patient's authenticity, particularly via optogenetics' impact on personality (art. 6); fostering maladaptive responses to harmful situations; and a loss of truthfulness (articles 5 and 8).
- Ethical concerns uniquely raised by optogenetics include: the risk of losing motivation to work for systemic change (articles 5 and 6); the risk of negative composition effects; the challenge of balancing respect for autonomy with the promotion of societal interests; and the risk of disrupting self-narratives (art. 5).
- Given that memories grounding factual and trait self-knowledge (i.e. semantic and implicit memories) need not be affected by the erasure of self-defining episodic memories, the loss of authenticity that might result from memory erasure using optogenetics need only be temporary. Also, other people than the patient could have a say in whether such erasure led to authenticity or not, and whether the erased memories should be reinstated (art. 6).
- The invasiveness of existing optogenetic procedures need not remain an insoluble problem in the future (art. 7). Also, some degree of speculation about future technological developments is appropriate to ensure that ethical reflection does not lag behind scientific innovation (articles 5 and 7).
- Memory modification can also negatively affect a patient's sense of personal agency (art. 8). This is problematic in that it can both lead to a failure of self-respect, and hinder improvements in the person's mental health. Yet in some cases, such interventions can also help enhance agency, autonomy, and well-being.
- Françoise Baylis's approach to narrative identity is preferable to those proposed by Schechtman and Hilde Lindemann, because Schechtman's reality constraint, and

Lindemann's credibility criteria, are not adequate to their task (art. 9). Baylis's "equilibrium" condition is a more promising candidate.

- We should distinguish between *internal* and *external* identity-related autonomy. The fact that memory modification may promote autonomy of the internal kind does not mean that it will also boost external autonomy. Prospective users of MMTs should be informed of the risks that these technologies can present to the latter kind of autonomy.

I now turn to an assessment of how successful the thesis is as a research project. The next section thus critically examines some of the ideas presented in it, as well as the consistency among its component parts (introduction, conclusion and published articles).

Evaluation of the thesis's content

This thesis is an impressive intellectual effort. It addresses important ethical questions raised by new developments in brain science, in an era when the social impact of neurodegenerative diseases and mental disorders is of greater relevance than ever before. The thesis is ambitious in its stated objective of developing a systematic approach that can allow us to assess the ethical pros and cons of neuro-interventions like DBS or MMTs in any particular case. It is also ground-breaking in the way it looks into the future, even though it thereby invites the charge of being too speculative. It should be noted, however, that part of the thesis material is precisely aimed at answering worries about undue speculation. (I will just mention that article 7 strikes me as making substantial concessions to skeptics about the feasibility of selective memory erasure, concessions I believe the authors need not have felt compelled to make.) The nine articles constituting the bulk of the thesis show undeniable thematic unity, centered around neuromodulation therapies and philosophical notions like the self, identity, authenticity and autonomy.

The author demonstrates genuine creativity in the scenarios he envisages (such as the cases of Nietzcheana and Tarana), and in how he applies existing philosophical and psychological theories (sometimes in further refined form) to the neuroethical issues under discussion. Some of his examples, like that of Tarana in article 9, might be seen as somewhat provocative (I suppose some might get offended by the suggestion that an activist like Tarana Burke could turn into a go-getting corporate lawyer following the erasure of even significant autobiographical memories), but they are certainly thought-provoking. He (and his co-author) consider(s) ethical concerns from a variety of perspectives, first making a detailed case for a particular concern, such as the loss of authenticity in Elizabeth's case, and then providing considerations that assuage it. This is the hallmark of good analytical philosophy, even though there are passages in which one feels the author's ultimate conclusion could be stated more clearly.

The thesis also has the merit of being informed by a large body of scientific literature pertaining to the interventions discussed (in impressive detail) by the author, most prominently in the cases of DBS and optogenetics. As illustrated by the fact that all nine articles constituting the body of the thesis have previously been published in peer-reviewed

journals, and by the series of peer commentaries on article 6, the author's work has already made an impact on the relevant debates in neuroethics.

There is also much to like when it comes to the substance of the author's arguments. For instance, I believe he is fully correct to suggest that a Schechtman-style narrative approach to the self is inadequate, as the nature of the self (or of individual identity) is arguably complex, and cannot either be reduced to, or be fully captured by, a person's self-narrative. In this context, the incorporation of McAdams's three-level personality framework in article 6 represents a laudable attempt at combining ethical analysis with the best theories currently on offer in psychology. The author also deserves credit for highlighting, unlike many discussions of memory modification in neuroethics, the fact that erasing episodic memories need not automatically deprive the person of any recollection of the relevant event, given the possibility of preserving or re-acquiring semantic memories of that event. Moreover, he helpfully points out that emerging MMTs like optogenetics refute the standard assumption according to which the erasure of even an episodic memory must be irreversible, and he draws some of the practical implications of this new development.

The author's wish to find a middle-ground between an overly static "essentialist" conception of the self and an unrealistically flexible "existentialist" one, evidenced by his endorsement of the dual-basis framework of authenticity, is eminently reasonable, although the consistency of the author's commitment to that framework across different articles is not so clear, a point I shall return to. On the issue of the moral responsibility and liability to punishment of offenders undergoing neuro-interventions like DBS, the author defends a bold position, based on strong philosophical foundations. Finally, he offers some helpful remarks about the need to consider the possible impacts of memory modification on a patient's value system, which could include serious and sometimes irreversible disruptions to their social ties.

Having highlighted these various merits of the thesis, I now turn to a critical examination of some aspects of the author's analysis, offered in the spirit of constructive philosophical dialogue.

My first comment has to do with the primacy that the author seems willing to grant to the patient's own perspective on themselves in article 2, in response to the contrary view expressed by Gilbert and colleagues with regards to Patient 4. I see two difficulties with it: first, the rationale given for that claim does not strike me as persuasive, and secondly, it does not seem to fit well with what the author argues elsewhere in the thesis (especially in article 9). On the former issue, the author defends Patient 4's perspective, expressed in his statement "I don't feel different at all", by arguing that the dimensions of the self that DBS may have disrupted need not have been subjectively "weighty" enough for the patient to view himself as changed, even though his relatives may have viewed things differently. The author adds that "it is the patient whose epistemic access to disruptions of the self-aspects is the most informative". Yet it is not clear to me why we should take this claim to hold in this particular case, at least if we interpret Patient 4's report of not feeling any different as implying that he *was* no different on DBS. Such a statement seems to contradict the reality of the clear personality changes reported by the patient's family following the intervention (and which, as I understand, the author does not deny).

Two possible ways of trying to rescue Patient 4 from the charge of self-deception (or self-blindness) would be to either interpret his statement as expressing his own subjective sense of authenticity, with no further claim about the "objective" lack of any changes in him; or as indicating that even if he were to acknowledge the relevant personality changes, he would still not view them as altering who he fundamentally was, because he did not regard the features in question as having much relevance in that context. The first approach cannot refute Gilbert and colleagues, but rather makes the author's disagreement with them disappear, as their analysis seems focused on the objective evolution of the patient's self, not on his subjective sense of authenticity. As for the second interpretation, it still fails to provide compelling grounds for thinking that the patient's own view of what constitutes their identity carries special authority, to those who do not find that premise intuitively appealing.

This leads me to the issue of coherence among the different parts of the thesis. Despite highlighting real shortcomings of Schechtman's narrative approach to identity in other articles, the author still seems to want to retain Schechtman's prioritization of the first-person perspective in article 2. However, I note that in article 9, the author endorses Baylis's narrative relational account of identity, which states that who someone is emerges from a state of "equilibrium" between one's self-narrative and the stories that others tell about the person. It is not clear that Baylis's equilibrium constraint is compatible with any form of prioritization of the first-person perspective when it comes to identifying a person's defining characteristics.

The same can be said of the author's emphasis, again in article 9, on the evidence suggesting that humans have a widespread propensity to confabulation and self-deception. Did he change his mind on this particular issue between articles 2 and 9? If so, this is certainly legitimate (and I happen to think that it is also the correct move to make), yet I think it ought to be made clearer in the introduction. Especially given the author's stated goal of helping develop a uniform framework for the ethical assessment of neuromodulation therapies, it is important to explicitly acknowledge any tensions between different parts of the thesis, whether these be linked to an evolution in his views, or perhaps to the fact that some but not all of the articles featured in the thesis were co-authored.

On a related note, while the author's rejection of Schechtman's version of the narrative approach is, in principle, certainly compatible with his endorsement of Baylis's in article 9, I nevertheless wonder whether some of the strongest criticisms he levels at Schechtman's analysis do not equally apply to Baylis's view. In article 6, he thus plausibly argues that while narratives do define one level of personality, namely the third one distinguished by McAdams, they nevertheless do not represent "a constitutive condition of the person's self…as there are other levels of personality". Yet given that Baylis's relational narrative account still defines the self in terms of the stories people tell about a person (including that person's own inner story), it is not clear that it can reliably incorporate those other personality levels, since in real life such stories are often marked by ignorance and error – and Baylis's constraints about equilibrium and absence of oppression do not seem to help in that regard.

I have similar questions about the relationship between articles 6 and 9. The former article puts the issue of authenticity front and centre in an ethical assessment of optogenetic memory

modification, and endorses the coherentist approach of Pugh and colleagues. In the latter, by contrast, the author declares agreeing with Mackenzie and Walker's that "we should focus in the neuroethical analyses on the concepts of identity and autonomy rather than authenticity". As before, if this indicates an evolution in the author's thinking, it is perfectly acceptable, but ought to be clarified, for the benefit of the reader, in the introduction (and perhaps also conclusion) to the thesis. I also wasn't sure, upon reading that article, whether the author had changed his view of the ethical significance of a loss of motivation to work for systemic change (which characterizes at least some variants of Tarana's case), or whether he simply wanted to bracket out this issue in that particular article, to focus instead on the question of autonomy and social relationships. (I would personally hope that the latter is correct.)

Another question is whether the distinction between authenticity and autonomy collapses if we follow the author's view, since the coherentist, dual-basis account of Pugh and colleagues can plausibly be characterized as analysing authenticity as a form of autonomy. Perhaps, in light of what the author says in article 9, he would view Pugh and colleagues' account as an inadequate analysis of autonomy, because insufficiently relational. Some clarifications would again be welcome on these points.

Still on this issue, I would offer the suggestion that a move away from the *language* of authenticity could indeed be desirable in neuroethics debates, given the ambiguous nature of this term, which even in philosophical contexts can be used to refer to distinct values, such as autonomy, accurate self-presentation, and reliable contact with reality. At least, disambiguating the term when one uses it should be required. Yet this does not mean that the *values* themselves to which the term can refer, beyond autonomy, can be safely ignored. Arguably, the fact that a neuro-intervention distorts a person's grip on reality, and/or their self-presentation, can count against it, even if it does not also undermine their autonomy.

Finally, there are several points mentioned in the thesis which I hope the author will consider elaborating on in future work. Article 5, for instance, mentions the possibility that optogenetics could serve as a more precise alternative to DBS, or at least that DBS could be refined to achieve optogenetic-like precision. One wonders what the author expects the impact of such cutting-edge interventions to be, if applied for *other* purposes than memory modification, from the perspective of identity, authenticity or autonomy. Another relevant point is how exactly "significant others" are to be understood in the context of article 9. One question concerns the degree to which someone like Tarana is free to decide whom she wishes to regard as having the "special authority" allowing her, according to the author, to participate in the co-creation of her identity. The greater her freedom to determine this, the less acute one might think the challenge to the assertion of her new identity will be, even if most of society happens to strongly disapprove of her latest self-narrative.

Overall assessment

Mr Zawadzki's thesis represents a solid, rich and original contribution to major contemporary debates in neuroethics. Given that he is either the sole author or the main contributor to eight of the nine articles constituting the thesis, I believe he has proven himself worthy of the degree of Doctor in Philosophy, and also of being considered an important voice in the field.

國立陽明交通大學心智哲學研究所 台北市北投區立農街 2 段155 號
Institute of Philosophy of Mind & Cognition **TEL**：**(02)2826-7000 ext. 65041**
National Yang Ming Chiao Tung University **FAX**：**(02)2823-2920**

While I have expressed reservations regarding some aspects of his philosophical analysis, this is only a normal phenomenon in our discipline, and should certainly not be taken to imply that the thesis is not an excellent piece of work. My only request would be for the author to add some clarifications in the introduction (and if deemed appropriate, in the conclusion) regarding the possible evolution of his views across articles, on the issues of authenticity, autonomy and the priority of the first-person perspective on the self, as per my comments above. I understand from my communications with the Department of Philosophy at Jagellonian University that I ought to abstain from giving a final recommendation about Mr Zawadski's dissertation until such revisions have been completed. I will therefore await your further input before doing so.

Dr Alexandre Erler
DPhil, Oxon