bioethics

*Bioethics* ISSN 0269-9702 (print); 1467-8519 (online)
*Volume 23 Number 6 2009* pp 360–374

doi:10.1111/j.1467-8519.2009.01725.x

# AUTONOMY AND AUTHENTICITY OF ENHANCED PERSONALITY TRAITS

*JAN CHRISTOPH BUBLITZ AND REINHARD MERKEL*

**Keywords**
*autonomy,*
*authenticity,*
*neuroenhancement,*
*manipulation,*
*neuroethics,*
*Prozac Defence,*
*direct brain intervention*

***ABSTRACT***
*There is concern that the use of neuroenhancements to alter character traits undermines consumer's authenticity. But the meaning, scope and value of authenticity remain vague. However, the majority of contemporary autonomy accounts ground individual autonomy on a notion of authenticity. So if neuroenhancements diminish an agent's authenticity, they may undermine his autonomy. This paper clarifies the relation between autonomy, authenticity and possible threats by neuroenhancements. We present six neuroenhancement scenarios and analyse how autonomy accounts evaluate them. Some cases are considered differently by criminal courts; we demonstrate where academic autonomy theories and legal reasoning diverge and ascertain whether courts should reconsider their concept of autonomy. We argue that authenticity is not an appropriate condition for autonomy and that new enhancement technologies pose no unique threats to personal autonomy.*

This paper aims to clarify the relation between personal autonomy and authenticity and possible threats posed by neuroenhancements. Recently authenticity has emerged as a key notion in the debate on neuroenhancements. A widespread worry is that the use of neuroenhancements to improve cognitive functions or to alter one's emotional makeup with the help of pharmaceutical or other biotechnological means undermines an agent's authenticity.

Indeed, one of the most interesting questions about neuroenhancements is how they may affect a person's identity. It is important, however, to be clear which sense of identity is at stake. As David DeGrazia has convincingly pointed out, there is no need to worry about the enhancing procedure affecting a person's numerical identity or his diachronic persistence.[1]

A different sense of identity is involved when someone alters his personality to the extent that others might call him 'not the same anymore'. Profound personality transformations may excite suspicions of inauthenticity. Being authentic, roughly understood as being 'true to oneself', is an ideal shared by both proponents and opponents of enhancement technologies. Yet they disagree on its content.[2] As we will discuss below, the main tension among theories of authenticity is between essentialist views, in which authenticity is threatened by everything that makes

[1] Regardless of one's view of the contested issue of personal persistence over time, with today's enhancing methods and those of the foreseeable future, the numerical identity of a person is preserved. Cases that involve (inadvertent) complete amnesia would be a exception, see H. Markowitsch et al. A PET Study of Persistent Psychogenic Amnesia Covering the Whole Life Span. *Cognit Neuropsychiatry* 1997; 2: 135–158. For in-depth philosophical discussion of identity and biotechnology, see D. DeGrazia. 2005. *Human Identity and Bioethics*. Cambridge: Cambridge University Press; G. Gillett. 2008. *Subjectivity and Being Somebody: Human Identity and Neuroethics*. Exeter: Imprint.

[2] E. Parens. Authenticity and Ambivalence: Toward Understanding the Enhancement Debate. *Hastings Cent Rep* 2005; 35: 34–41; L. Bolt. True to Oneself? Broad and Narrow Ideas on Authenticity in the Enhancement Debate. *Theor Med Bioeth* 2007; 28: 285–300: 286.

Address for correspondence: Jan Christoph Bublitz, University of Hamburg – Legal Studies, Schlueterstrasse 28, Hamburg 20146, Germany. T: 00491771747182; Email: christoph.bublitz@uni-hamburg.de

people depart from who they truly are, and existentialist views, in which we create ourselves according to our own ideals, and an authentic personality consists of self-defined and self-established characteristics.

These two different interpretations of 'authenticity' lead to different answers to the question, 'Do neuroenhancements threaten authenticity?' Some critics fear that neuroenhancements separate us from 'who we really are' and what is 'most our own,'[3] while others praise neurotechnological tools for facilitating self-creation and self-fulfillment, enabling persons to become who they want to be. Settling this debate requires a firm understanding of the meaning and value of authenticity – without which it remains unclear whether it is truly a moral ideal, as Charles Taylor famously argued.[4]

While this paper will address those questions, it will focus on another concern involving authenticity and neuroenhancements: their relation to autonomy. Most contemporary theories of personal autonomy are at least implicitly based on an idea of authenticity. Some of them consider authenticity a necessary precondition of autonomy, ensuring that actions issue from an agent's *own* character. This implies that neuroenhancements might threaten personal autonomy by undermining authenticity. In this way neuroenhancements may impair 'our capacity to act freely and to consider ourselves responsible for the things we do.'[5] In order to assess this concern we will discuss several neuroenhancement scenarios and their evaluation according to contemporary theories of autonomy.[6] Interestingly, neuroenhancements may make some of the rather bizarre thought experiments in the autonomy debate come true. Moreover, legal cases involving a 'Prozac defence' already raise similar autonomy questions, and courts answer them in a way that is at odds with major theories of autonomy.

Drawing on several major theories of personal autonomy, we will put forward our own theory of autonomy, in which autonomous agents: possess the capacity for discerning right from wrong, are reason-responsive, have a minimal level of self-control, have a minimally proper understanding of the world around them, have not been manipulated (in a sense to be spelled out below), and identify with their traits (including their desires). Therefore, if agents who possess the minimal autonomy capacities identify with their enhanced personality traits and have not been not manipulated, there is no reason to deny them autonomy on the grounds that they are inauthentic.

## I. AUTONOMY

For clarity's sake, we are confining our argument to the notion of personal autonomy as *an agent's status of being an apt target for reactive attitudes such as praise and punishment*; on this view, autonomy is *a condition for moral accountability*.[7] Autonomous actions are to be respected by others and basically preclude paternalistic interventions.[8]

Autonomy requires certain minimal capacities. Agents must possess the capacity for discerning right from wrong and be reason-responsive in the sense that they would have acted otherwise if there were halfway plausible reasons to do so.[9] Furthermore, they need sufficient self-control to act according to their judgment (they must not be akratic). At least to some extent, they also need a proper understanding of the world and of the consequences of their actions.

Some extraordinary forms of neuroenhancements can impair these minimal capacities. When an artist consumes hallucinogens for inspirational purposes and curses someone for being a demon, he is probably not autonomous since he lacks minimal rationality, information and self-control. (Whether his hallucinogenic painting is praiseworthy depends – artistic taste aside – on the relation of the praiseworthiness of a product and the autonomy of its creator, an issue that we will not pursue here.) But cases like this – in which agents act under a capacity-diminishing influence – do not give rise to

[3] President's Council on Bioethics. 2003. *Beyond Therapy*. Washington, DC: U.S. Government Printing Office: 253.

[4] C. Taylor. 1991. *The Ethics of Authenticity*. Cambridge, MA: Harvard University Press.

[5] M. Sandel. 2003. *What's Wrong with Enhancement?* Background paper for the President's Council on Bioethics, *op. cit.* note 3. www.bioethics.gov/background [Accessed 15 Jan 2009].

[6] We only consider pharmaceutical enhancements, but our arguments can be generalized so as to apply to other means of brain intervention. For a comprehensive review of state-of-the-art techniques, see: R. Merkel et al. 2007. *Intervening in the Brain: Changing Psyche and Society*. Berlin: Springer; D. Repantis & I. Heuser. 2008. Antidepressants for Neuroenhancement in Healthy Individuals: A Systematic Review. *Poiesis & Praxis* [DOI: 10.1007/s10202-008-0060-4].

[7] There are a great many theories of autonomy that sometimes share the same name but not the same subject matter. It is impossible to find an all-encompassing notion of autonomy. Often it is understood as an ideal with rather strong conditions, while our narrow and technical understanding relates only to the functional sense of moral accountability, which is much less demanding. Cf. N. Aparly. 2004. Which Autonomy? *Freedom and Determinism*. J. Campbell et al., eds. Cambridge, MA: MIT Press: 173–187.

[8] Notwithstanding certain exceptions such as the legal prohibition of 'killing on demand.'

[9] Note that our argument does not presuppose the metaphysical possibility of acting otherwise – as contended and contested in the free will debate.

particularly novel normative questions, for jurisdictions deal with issues like 'temporary insanity' as routinely as they deal with intoxicated criminals.

This paper will look instead at neuroenhancements that arguably threaten autonomy in a different way. We will look at agents who possess minimal autonomy capacities but have so drastically transformed their personality traits through neuroenhancements that their newly formed traits may be regarded as inauthentic. When agents' behaviour originates from neuroenhanced traits, theories of autonomy disagree on agential autonomy.

We should point out that actions are not in the strict sense *caused* by neuroenhancements. Neuroenhancements do not directly initiate actions on the neuronal level through activation of the motor cortex. If they did, the resulting bodily movement would not qualify as actions proper, i.e. as 'belonging' to the actor.

The influence of neuroenhancements on actions we are concerned with is less immediate. Rather than making a person behave in a certain way, neuroenhancements may modify a person's motives or general disposition to undertake certain actions. Depending on the consistency and scope of this modification, one may for instance distinguish changes in an agent's mood or character traits. For lack of a better term, we will identify such behavioural modifications by saying that a person's *pro-attitudes* toward certain actions have been changed. This broad term encompasses the agent's mood, character traits, motives and general disposition to undertake certain actions.

In what follows, we introduce six enhancement scenarios that differ in terms of voluntariness in the administration of the enhancement, foreseeability of the resulting effect, and post-enhancement identification with the new trait. All agents are assumed to possess minimal autonomy capacities.

### (1) Voluntary intervention, foreseeable result, identification

*Marc is an autonomous agent and aspiring philosopher; he has just finished his PhD on 'The Importance of the Categorical Imperative in Postmodernism' and strictly adheres to Kant's teachings. He is shy, low in self-confidence, and feels something lacking in his life. Marc autonomously decides (d1) to take somafinil, a substance raising his dopamine level. He changes drastically, discovers new sides of his personality, and generates interest in things inconceivable to him before. He becomes more outgoing and has several dates a week. At first he is surprised about his change, but he quickly adapts his self-conceptions and pro-attitudes. He decides (d2) to liberate himself from Kant's grip, reads Epicurus, and lives by hedonistic ideals from then on. The dean of his faculty calls his transformation into question, remarking that he is 'someone else now' – someone whom he rather dislikes. His colleagues charge him with inauthenticity, saying that 'it's not him, it's the drug.' Is Marc's decision d2 autonomous?*

Marc's successful transformation in (1) is the basic neuroenhancement situation: someone is discontented with a character trait and intentionally modifies it. Afterward he is satisfied with what he has become. However, the reactions of his social environment, suspecting his hedonistic lifestyle to be inauthentic, are comprehensible. In order to examine whether they may be justified in questioning Marc's autonomy, let us first review how his case is assessed by structural theories of autonomy.

## II. STRUCTURAL THEORIES OF AUTONOMY

According to structural or hierarchical theories, an agent's pro-attitude is autonomous when it fits into the wider structure of the rest of his pro-attitudes. In Harry Frankfurt's prominent account,[10] pro-attitudes are hierarchically structured, so that *an agent is autonomous with respect to her effective first-order desire to φ if she both had a second-order desire to have the first-order desire and she also wanted her desire to φ to move her to act (a second-order volition)*. Higher-order desires make the lower-order desire autonomous. This harmony between lower and higher-order desires is constitutive for autonomy in hierarchical theories and has been termed 'authenticity' by Gerald Dworkin.[11] Frankfurt subsequently refined his proposal by requiring the relation between higher and lower-order desires to be of a special quality: an agent has to 'decisively' or 'wholeheartedly identify' with his first-order desire. (In what follows, we shall disregard whether the first-order entity is a desire and shall look more broadly at individuals' identification with any of their pro-attitudes – for example, their moods or character traits). This kind of identification process confirms a pro-attitude's authenticity and ensures that it derives from an agent's 'essential character of will,' not from inauthentic factors such as addiction or neurosis. The notion of identification is blurred in detail and subject to a lively debate; for Frankfurt it is not necessarily a rational approval but

[10] H. Frankfurt. Freedom of the Will and the Concept of a Person. *J Philos* 1971; 67: 5–20.

[11] G. Dworkin. 1988. *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press: 25.

an 'altogether neutral attitude of acceptance'[12] or 'incorporation'[13] of one's desires. Identification is related to satisfaction; it 'entails an absence of restlessness or resistance ... A satisfied person has no interest in bringing about a change, being better is not interesting or important to him.'[14] The antipode of identification is alienation, roughly understood as the inability to accommodate pro-attitudes that conflict with one's self-concept. If someone is alienated from his first-order pro-attitudes, the resulting actions are not autonomous. This applies to the following variation of Marc's case:

### (2) Voluntary intervention, foreseeable result, alienation

*Kant's grip is too strong. Although somafinil works in the way Marc imagined, he is unable to integrate the new traits into his self-conception. He dislikes his hedonistic inclinations and feels alienated from his 'true self.'*

Whereas in (1) Marc identifies with his new desires, he rejects them in (2). Thus, according to structural theories of autonomy, he is autonomous in (1) but is not in (2).[15] As we can see in the first case, neuroenhancements do not necessarily undermine autonomy and authenticity, according to structural theories. (This said, one sees an important difference between colloquial and specialized meanings of authenticity. According to structural theories, authenticity depends only on the agent's self-evaluation, whereas colloquially it also depends on the view of others.) Thus, authenticity as identification is not necessarily undermined by neuroenhancements.

Another concern about identification and enhancements is worth noting: The availability of enhancements may lead people to be less inclined or even reluctant to identify with their existing traits. When a non-disposable trait becomes modifiable, persons may seek to alter it by any means necessary instead of just trying to accept it. In some respects modern societies burden people with perfectionist expectations. People may find themselves under legitimizing pressure to be a certain way as soon as being that way evolves from 'chance to choice.' This lesson can be learned from cosmetic surgery. In the days before plastic surgery became widely available, were people so concerned with such trivial facts of bodily appearance as the angle of the tip of their nose? We suspect not. Without a surgical option, people are probably more inclined to identify with the way they are. Obsessive striving to become better is *not* the way to a good and satisfactory life but rather facilitates discontent and, in autonomy contexts, alienation. Hence, through their availability alone, new technologies promising a better life might indeed be the source of mass-scale dissatisfaction. This indirect relationship between enhancements and identification should be taken into account by policymakers, but it does not undermine the autonomy of someone who identifies with his enhanced traits.

## III. HISTORICAL THEORIES OF AUTONOMY

Structural theories are challenged on various grounds, especially because they are insensitive to the agent's social relations. The people of Aldous Huxley's *Brave New World* are manipulated by an oppressive society with the help of neuroenhancements. All their troubles – and likewise all their ambitions – are drugged away by 'soma,' which keeps everybody in a shallow state of contentment and complacency. Strikingly, structural theories disregard these background conditions. The people in *Brave New World* have only a few desires, but they identify with them. Hence, though limited in their reach, their actions would be autonomous according to structural theories.[16] The same shortcoming of structural theories can be found in manipulative two-person cases: if an agent's identification is brought about by heteronomous intervention, he is not autonomous. To address such manipulations, historical theories of autonomy have been proposed. They are of special interest for the enhancement debate. Consider a manipulation case:[17]

[12] H. Frankfurt. 2002. Reply to Gary Watson. *Contours of Agency: Essays on Themes from Harry Frankfurt*. S. Buss & L. Overton, eds. Cambridge, MA: MIT Press: 161.

[13] H. Frankfurt. 1988. Identification and Wholeheartedness. *The Importance of What We Care About*. Cambridge: Cambridge University Press: 172.

[14] H. Frankfurt. 1999. The Faintest Passion. *Necessity, Volition, Love*. Cambridge: Cambridge University Press: 105.

[15] Note that it is a different question whether M bears responsibility for his hedonistic lifestyle if it were a foreseeable consequence of the autonomous decision d1. If a reprehensible act can be traced back to an autonomous decision and was foreseeable at that time, there is room for a charge of negligently causing the second nonautonomous act d2. However, charges of negligence are weaker than those of intent, and events occurring after a long stretch of time may not have been foreseeable at d1.

[16] Analogously, feminists challenge the individualistic conceptions of autonomy that disregard personal interdependence and cultural and societal influences. An interesting collection of such views is C. Mackenzie & N. Stoljar, eds. 2000. *Relational Autonomy: Feminist Perspectives on Autonomy, Agency and the Social Self*. Oxford: Oxford University Press.

[17] This is a modified version of a case developed by A. Mele. 1995. *Autonomous Agents: From Self-Control to Autonomy*. Oxford: Oxford University Press.

## (3) Involuntary intervention, identification

*Beth is an autonomous agent and an aspiring philosopher. She has just received her PhD on Kant, whose categorical imperative she strictly adheres to. Her misanthropic and postmodernist colleague Ann is envious of Beth's success and her coherent philosophical belief system. Ann starts to stir things up by giving Beth work that deals with Schopenhauer and Nietzsche. With the aid of a nefarious neuroscientist, Ann discovers that Beth's neurological structure is prone to transformation with malafinil. Ann therewith implants in Beth a Charles Manson-like hierarchy of values and eradicates all competing values. Beth is now, in a relevant sense, a psychological twin of Manson. Beth realizes her change; the categorical imperative affects her no more. Yet she traces her change to the fact that her old life in accord with the categorical imperative has emerged as less satisfying than once imagined. Beth reflectively endorses her Manson-like desires. She commits M, a Manson-like act.*

## (4) Involuntary intervention, alienated by effect

*Beth realizes her change and is alienated by her Manson-like desires.*

For structural theories, Beth's autonomy depends solely on higher-order identification. In (4) Beth is alienated by her new desires – she is not autonomous. The challenging case is (3). Ann brainwashes Beth so thoroughly that Beth identifies with her instilled pro-attitudes. According to structural theories, Beth is therefore autonomous. Frankfurt says that when someone succeeds 'in providing a person ... with a new character, [t]hat person is then morally responsible for the choices and the conduct to which having this character leads.'[18]

There is ample reason to object to this result. Intuitively, manipulated agents are paradigmatic examples of heteronomy. Hence, historical theories stipulate a further condition: autonomous pro-attitudes have to be acquired in an appropriate process – they need to have an autonomy-conferring etiology (causal history). Fischer and Ravizza (hereafter F&R) have developed one of the most comprehensive and influential history-sensitive theories of autonomy. Some of their remarks are significant for personality transformations through neuroenhancements. According to them, numerous personality traits must be evaluated historically. For instance, virtue is a historical concept:

> Virtues depend on certain *processes of acquisition*. Virtues are not simply propensities or dispositions to behave in certain ways ... they are these dispositions only provided that *they have been acquired through certain appropriate processes of education and habituation*. It is impossible in a strong sense that there be 'virtue pills,' pills that one could take that could induce dispositions that would count as virtues. Whereas these pills might induce the pertinent propensities, these would *not count as virtues* insofar as they were not acquired in the relevant fashion ... It's a conceptual and metaphysical impossibility that a person have the relevant virtue without having acquired it in the specified manner.[19]

If F&R are right, the value of any virtuous personality trait sufficiently influenced by neuroenhancements is severely undermined. Actions originating from such enhanced traits may not have the right etiology and may even be nonautonomous. F&R's central idea is that autonomy requires guidance control. Agents exhibit guidance control when actions issue from their own moderate reason-responsive mechanisms.[20] Generally, persons who use neuroenhancements do not lose their reason-responsiveness. In (3) Beth is reason-responsive; unfortunately, she acts (as Manson did) for the wrong reasons. What undermines Beth's autonomy is that her actions were not caused by her *own mechanisms*: 'Agents who perform actions produced by ... potent drugs and certain sorts of direct manipulation of the brain are not reasonably to be held responsible for their actions insofar as they lack the relevant sort of control.'[21] In those cases, the 'behavior does not issue from *one's own mechanism*'.[22]

A plausible result in brainwashing cases. But stated in this general way, agents who use potent drugs or direct brain interventions *never* act autonomously, even in the basic case (1), where Marc changed his personality from Kantianism to hedonism with the help of somafinil. Thus the interesting point in F&R's theory is the suggestion that mechanisms influenced by *direct* brain interventions (neuroenhancements such as pills and psychosurgery) are not the agent's own since their history relevantly differs from mechanisms produced by traditional or *indirect* brain interventions. Of course we cannot do justice to

[18] H. Frankfurt. 2002. Reply to J. M. Fischer. In *op. cit.* note 12, p. 27.

[19] J. Fischer & M. Ravizza. 1998. *Responsibility and Control*. Cambridge: Cambridge University Press: 182. Of course, in their view autonomy is also a historical concept.

[20] J. Fischer. Responsibility and Manipulation. *J of Ethics* 2004; 145–177: 146.

[21] Fischer & Ravizza, *op. cit*. note 19, p. 35.

[22] J. Fischer. Responsibility, History and Manipulation. *J of Ethics* 2000; 385–391: 391.

F&R's theory of mechanism ownership here, but surely their remarks strike an important note in the enhancement debate.

Is there a way to incorporate the attractive idea that the causal history of a pro-attitude affects autonomy? Yes there is, but we must first address (and then ultimately dismiss) arguments that conclusions about agential autonomy can be inferred from historical facts about whether an intervention was direct or indirect. In order to assess these arguments, we will first isolate the relevant historical difference between pro-attitudes acquired or transformed through ordinary, indirect interventions and pro-attitudes acquired or transformed through direct interventions. Once we have correctly described this relevant historical difference, we will then show that it does not map onto our normative intuitions regarding the attribution of autonomy.

## IV. DIRECT VS. INDIRECT BRAIN INTERVENTIONS

How do direct interventions differ from indirect ones? Neil Levy attempts to capture the peculiarity of direct interventions. He observes 'a widespread *presumption* in favor of the traditional way of changing minds, other things being equal ... If we *can* use the traditional means, we should, or so many people believe.'[23] Levy makes the reasoning of these people explicit in the following manner:

> [D]irect manipulation of the brain differs from indirect in an extremely significant way: whereas the presentation of evidence and argument manipulates the brain via the rational capacities of the mind, direct manipulation bypasses the agent's rational capacities altogether. It works directly on the neurons or on the larger structures of the brain.[24]
>
> Anti-depressants, psychosurgery and the other technologies of direct manipulation introduce an alien element into the equation: After treatment with these technologies, I am no longer the person I was ... [P]sychotherapy is preferable to direct manipulation. Psychotherapy explores *my* self, my inner depths. It seeks coherence and equilibrium between my inner states, and between my inner states and the world. But direct manipulation simply imposes itself over my self.[25]

Among others, Levy presents three characteristics:[26] direct brain interventions bypass rational capacities, they introduce an alien element that undermines authenticity, and they impose themselves over my self.

Let's begin with a closer look at the last characteristic: the idea that direct brain interventions impose themselves over my self. Sidestepping metaphysical minefields surrounding notions of a 'self,' the 'imposition over my self' seems a valid description only if specific mechanisms are replaced by others (e.g., an implanted chip takes over brain functions). But pharmaceuticals work differently. They make use of the existing biological framework: neurotransmitters relay, modulate or amplify signals. Raising serotonin levels that alter mood is not an imposition or replacement of an alien mechanism over an authentic one but rather the modification or reconfiguration of a system. Although one can identify functionally different areas in the brain, it seems to be a large unified and densely interconnected network. It is hard to conceive of an increase or decrease of neurotransmitters available in synaptic clefts throughout the brain as distinct mechanisms; and it is patently implausible to consider traits originating from specific neurotransmitters, say serotonin, as *not* an agent's own. Any approach – particularly F&R's – that depends on the individuation of specific mechanisms is problematic, as it presupposes a reasonable way to identify distinct mechanisms and to ground the attribution of autonomy thereupon.

In a different sense, however, pharmaceuticals may introduce an alien element into the neuronal system. This hints at another distinction often appealed to in the enhancement debate: *natural vs. artificial*. The plausibility of an appeal to the natural is, of course, closely interwoven with the debate over what constitutes human nature in an increasingly technical world. We cannot pursue this here, but would like to note that anyone drawing on this distinction would have to further support the rather futile claim that only the natural conveys autonomy.

Most interesting is the first characteristic: the suggestion that direct interventions bypass rational capacities. Here the means of intervention closely relate to autonomy considerations: 'Certain kinds of manipulation that bypass or somehow supercede the human capacity for practical reasoning are salient examples of

[23] N. Levy. 2007. *Neuroethics: Challenges for the 21st Century*. Cambridge: Cambridge University Press: 71.
[24] Ibid: 70.
[25] Ibid: 75.
[26] We hasten to note that Levy does not endorse the claims he depicts. In fact, he refutes several presumptions against direct interventions. Ibid.

responsibility undermining factors'.[27] But how do direct interventions supposedly bypass rational capacities?

To begin with, the choice of pharmaceutical over traditional means of self-transformation can be perfectly rational. Persons usually have reasons for seeking personality changes through neuroenhancements, and these reasons may be as good or bad as those of people utilizing 'indirect' means. Rather, if direct interventions bypass rationality in some way, this apparently pertains to the transformation process itself. What is the peculiarity of a transformation process by direct interventions?

A transformative process via pharmaceuticals is not *initiated* by rational mechanisms (but by physiological mechanisms), nor does it *operate* rationally (but on brain chemistry). If rationality is the appropriate criterion by which to identify concerns about enhancements, ethically dubious transformative processes apparently have to be initiated by or operate on rational capacities.

Here, however, caution is in order to avoid the conceptual traps of the mind-brain–problem. Supervenience theories (requiring no strong ontological commitments) maintain that every mental change correlates with or is dependent on physical (neuronal) changes. Accordingly, one can cautiously say that direct interventions work primarily on the physical (neuronal) level, whereas indirect interventions aim at the mental level but cannot achieve any effect without acting physically as well. Thus, on the physical side, every transformation – direct or indirect – 'works directly on neuronal or larger structures of the brain'; and such processes cannot be rational in any strict sense because the concept of rationality is not applicable to electro-chemical occurrences. Rationality is a property of persons.

Interventions may differ in the way decisions are *executed*. Ordinary decisions for self-transformation do not necessarily induce the intended change. Oftentimes decisions are short of 'executive power' to induce changes on psychological and neuronal levels. Then one can resort to other means, e.g., changing something in the world and hoping for coinciding change 'in the mind' or by deploying mental techniques ('pulling oneself together,' imagination, repetitive thinking, etc.) to 'strengthen' the desire. A large portion of the problems people have with themselves is due to the fact that they cannot get themselves to live up to their (reasonably formed) resolutions. Direct (neuronal) interventions have the advantage of not relying on 'willpower.' At times they may even enable rationality by helping a person overcome 'weakness of will.'

[27] J. Fischer, *op. cit*. note 20, p. 145.

Closely related is another difference. Indirect transformations – and this is a tentative approach to the problem – may involve mechanisms that possibly function in a way resembling a system of 'checks and balances' which prompts us to reexamine evidence or reconsider a decision if transformations change the personality in undesirable ways. Furthermore, the effectiveness of traditional interventions might be restricted by opposing desires or beliefs. Normally people do not change randomly but in accordance with an existing personality structure. Acquiring 'maverick' traits is a hard task; an optimistic person may simply fail to become melancholic or pessimistic, and the depressed may prove incapable of becoming hopeful and easy-going unless they undergo drastic changes in their overall personality structure.

Such restrictions do not apply to pharmaceutical interventions. At least to some extent they can simply *override* the status quo chemically, irrespective of individual desires and beliefs. Direct interventions have an immediate impact on neuronal functioning, whereas traditional interventions change personality structures slowly and more holistically. Thus neuroenhancements may bypass the 'checks and balances' of an existing personality structure, and perhaps this is why enhanced traits are susceptible to being deemed inauthentic. But how do neuroenhancements supposedly bypass rational capacities?

We think they normally don't. One can speak of rational capacities being bypassed in only two exceptional cases: One is that *other* persons can avail themselves of the effectiveness for manipulative purposes. The (perhaps unknowing) consumer's personality can be transformed without his having adequate control over the change occurring in himself; we elaborate on manipulation cases below. The other case concerns *unintended side-effects*. In both cases, pharmaceutical interventions are capable of initiating transformations *without* a preceding (rational) decision, and in this sense they bypass rational capacities.

The majority of indirect personality transformations, however, are not initiated by rational decisions and are accompanied not by reasoning but by the psyche's mysterious, irrational, inexplicable and oftentimes unconscious forces. In fact, only few personality traits are the result of a rational process of self-creation, and not every decision is subject to rational scrutiny. An abundance of neuroscientific and psychological evidence points to many non-rational factors that shape our moods and affect pro-attitudes and decision-making. Most prominently, Antonio Damasio's somatic markers hypothesis claims that without certain physiological affective states, decision-making processes are severely

impaired.[28] Several forms of psychotherapy work with conditioning procedures rather than by reflecting on unconscious conditions. Rationality and reflective deliberation – which are oft-invoked ideals – might be exceptions rather than the rule in personality change. A great share of the background factors in pro-attitude formation seems to be beyond rational control.

What does this mean for autonomy? Agents are nonautonomous insofar as neuroenhancements produce transformations against their better judgment and agents do not identify with the new traits.[29] However, the troublesome cases we are interested in are those in which agents *do* identify with their transformed traits. Unless one declares as nonautonomous any pro-attitude originating from processes that bypass rationality, a claim we will deal with below, rationality seems an inappropriate criterion by which to identify the peculiarities of changes by neuroenhancements. Any principled distinction drawn between ethically suspect direct interventions and nondubious indirect ones would seem to come at the cost of challenging autonomy per se.

Descriptively, the best candidate for a criterion to mark the contrast between direct and indirect brain interventions is that the latter are *cognitively mediated*. Transformations caused by cognitively mediated interventions need not be initiated by a rational decision but may well originate in subconscious and irrational processes. They differ from pharmaceuticals and brain surgery insofar as these do not involve or require any cognitive component. Think of subliminal advertising and hypnosis. They are only effective if perceived yet do not rise to the level of consciousness; both bypass rationality, but are still different from pharmaceuticals.

But this distinction between 'cognitively mediated' and 'non-cognitively mediated' does not map onto our normative intuitions regarding the attribution of autonomy. We are not in principle suspicious of some non-cognitively mediated interventions: if a (non-pathologically) depressed and stressed-out philosophy professor does physical exercise on a daily basis, thus increasing his ephedrine production, elevating his mood and acquiring new pro-attitudes, no cognitive factors are involved, yet his autonomy is beyond doubt. Conversely, we are uncomfortable with hypnosis and subliminal advertising in a way that we are not with exercise – although physiologically the latter's functioning is much more akin to pharmaceutical intervention than the former's. Moreover, some direct interventions such as pharmaceutical treatment of mental illnesses certainly *restore* autonomy and may even *enhance* capacities for rational thinking, whereas cognitively mediated persuasion thwarts it. These examples demonstrate that the apt description of the two intervention techniques does not correspond to normative autonomy considerations.

## V. AUTONOMY AND MANIPULATION

Thus there is no unambiguous way of inferring agential autonomy from the means of intervention or the mechanisms issuing in action. The relation is more intricate and depends on the notion and function of autonomy. Let's take manipulation cases like (3) for a start. Recall that in these cases, the direct interventions are involuntary, but the nonautonomous agents identify with the results. How and why does the direct brain intervention make these agents nonautonomous? Well, they lack autonomy not because the history of their pro-attitudes involved direct brain interventions but because they were *manipulated*. Persons can be manipulated through various means, from the presentation of false evidence, hypnosis and advertisements, through to pharmaceutical interventions. It is not the means that render them nonautonomous but the fact that someone else *illegitimately infringed upon their rights*. Under certain circumstances, the violation of rights leads to a shift of responsibility from the manipulated agent to the manipulator, thus exempting the manipulated. The manipulator bears primary responsibility for the actions of the manipulated. In such cases we consider the manipulated *heteronomous* and not the proper addressee of reactive attitudes. This may sound obvious, but it is significant for appreciating the difference in reasoning between exclusively internal theories of autonomy (like structural theories) – which try to locate autonomy in an agent's inner states alone – and those that include external factors such as social relations among agents. The reason for the nonautonomy of the manipulated is not to be found in his brain but in the fact that someone else illegitimately interfered with his brain states. Thus, an agent is responsible if he possesses minimal autonomy capacities and, additionally, if normative principles for the attribution of autonomy do not shift responsibility for his action to someone else. Hence, *self-induced* brain interventions *never* thwart autonomy regardless of the means of intervention. Marc

[28] A. Damasio. 1995. *Descartes' Error. Emotion, Reason, and the Human Brain*. New York: Avon Books.

[29] Although more powerful, changes by direct interventions are not necessarily irreversible. Especially if pharmaceuticals require repeated intake, the agent has opting-out opportunities to correct changes against his better judgment. Still, as a caveat, the rational decision to employ highly effective neuroenhancements with hard-to-reverse effects must be well considered. A way to ensure this would be to restrict access to such neuroenhancements to qualified consumers.

is autonomous in the basic case (1) but Beth is not in (3). This theory is contrary to several prevalent theories and we will deal with a few consequences that some may find counter-intuitive in Prozac cases (below). But first, let us briefly discuss illegitimate influence.

## VI. ILLEGITIMATE INFLUENCE

The guiding principle is agential control. As far as we understand the structure and functioning of consciousness, consciously available information enables the best decision-making. For instance, functional conceptions of consciousness such as the global workspace model claim that only consciously available information is accessible to several sub-modules involved in decision-making.[30] Therefore an agent has most control when he makes conscious decisions, which presuppose the agent being consciously aware of relevant decision-making factors. Based on this premise, we can incorporate certain insights of the previously outlined distinctions among interventions and propose a framework of illegitimate influence by others:

1. Interventions involving conscious and uncoerced processes do not thwart agential control and are *prima facie* legitimate, leaving the agent autonomous. Arguments and rational consideration involve the highest cognitive functioning, but even interventions that make use of less rational but still conscious mechanisms (e.g. the presentation of goods in favorable lighting in supermarkets) are legitimate unless other normative considerations apply (e.g. the fact that goods are rotten is masked by infrared light).
2. Interventions that are cognitively mediated but never come to conscious awareness, such as hypnosis and subliminal advertising, take away conscious control but still involve the 'checks and balances' of subconscious processes. Thus they infringe less than direct interventions but more than conscious ones. In these cases, the key question is a normative one: is the intervener *entitled* to change the agent's mental states by such interventions? It's fine for a shrink to perform hypnosis and undertake subliminal behaviour-reprogramming with prior consent, but it is impermissible for a company attempting to make people buy whisky to invisibly insert pictures of nude women into every twentieth frame of a TV commercial. The guideline here is that we ought to treat each other as beings who act for reasons and who respect each other's capacities for self-control, which implies the duty of not intentionally undermining control capacities. This said, communication in daily life always involves non-conscious factors. Some salesmen, for instance, may strongly influence decision-making through their charm, gestures and appearance – all of which are subconsciously appreciated. They may even train to deploy these qualities intentionally, yet these are largely unavoidable and *socially accepted* subconscious interventions, and therefore legitimate.
3. In contrast, non-cognitively mediated, direct interventions by their very nature do not involve any of these processes and therefore bypass control capacities. But most non-cognitive interventions by others are always illegitimate for a different reason. Body and brain are protected against outside interference without consent; every jurisdiction penalizes the infliction of physical injury. (We note that 'physical harm' does not traditionally comprise minor detriment to bodily substance below a certain threshold of significance, and in manipulation scenarios there may not even be any damage to the brain's substance if, for example, neurotransmitters are increased or a certain area in the brain is stimulated. Thus legislations will have to deal with the question of whether a toxic-free reconfiguration of neuronal states with primarily mental outcomes constitutes bodily harm.) For our purposes, we can revert to the well-founded principle that no one has to endure or accept any bodily interference; hence direct interventions by others are illegitimate unless warranted for special reasons (e.g. pharmaceutical treatment of mental illness) or by informed consent.

Although this framework needs further refinement, we think that illegitimacy is the only feasible way to distinguish which interventions preserve autonomy and which do not.[31]

Now, finally, we can turn to the challenging 'Prozac defence' cases.

### (5) Voluntary intake, unforeseen side-effects, identification (Prozac case)

*Hedonism leads Beth away from philosophy to more financially rewarding enterprises. She conducts international stock market transactions and has to stay alert and focused*

[30] B. Baars. 1997. *In the Theater of Consciousness: The Workspace of the Mind*. Oxford: Oxford University Press.

[31] Further refinements need to incorporate empirical findings on the effectiveness of the influence and to define a normative threshold for the transfer of responsibility.

*for several hours at night, when Asian markets are open. In order to remain fit for the demands placed on her, she autonomously takes promafinil, a substance that significantly raises cognitive capacities and endurance. An unknown side-effect of promafinil is that it influences pro-attitudes for violent behaviour (while promoting a positive self-image through stimulation of the 'reward system'). Over a stretch of time, Beth takes on new pro-attitudes and identifies with them. One night she commits M, a Manson-like murder, because of the newly acquired pro-attitudes. Did she act autonomously?*

Unlike brainwashing scenarios, this case is far from being merely of theoretical interest. An alleged relation between SSRI anti-depressants and suicide and homicide has been presented to courts worldwide in so called 'Prozac defences', defendants claiming that the unknown side-effects of SSRIs caused them to commit crimes. When such a defence was successful in courts, it was because defendants were held to be 'chemically insane' or 'temporarily insane'.

Legal insanity is established by a test of either cognitive or control capacities. The former tries to establish whether the defendant was unable to know, appreciate, or understand the nature of his conduct or that it was morally wrong or legally prohibited. Control tests seek to explore whether the defendant was unable to exert sufficient control over his conduct or conform his conduct to the requirements of the law.[32] Agents in a state of legal insanity lack the minimal capacities we stipulated above.

But, as mentioned, in those interesting cases where agents *do* possess minimal autonomy capacities, there still remains a somewhat vague but insistent intuition that – in cases of unanticipated side-effects – the agents should not be deemed autonomous. For the sake of argument let's assume that promafinil contributes to violent behaviour while still preserving agential capacities of reason-responsiveness and self-control.[33] It only adds a violent quality to the agent's thoughts and pro-attitudes combined with an increased activation that enables their realization. Beth is still reason–responsive, as she would have refrained from M'ing had some plausible reason presented itself (say, the victim had cried), and she has self-control since she acts the way she wishes. She is not irresistibly urged to M; rather, she identifies with M'ing. Promafinil only contributed significantly to the formation of the respective pro-attitude.[34]

How do – or, at any rate, *should* – courts decide such cases? If defendants pass the insanity tests – which Beth would – criminal courts disregard the etiology of the pro-attitude. They deem it irrelevant whether the tendency for violent behaviour was caused by a pill, by nature, or by 'traditional ways' – as a chain of unlucky incidents, say, from miserable family conditions to social exclusion that furnished the defendant's psychology with similar propensities.[35]

Perhaps one is tempted to disagree with this legal reasoning. Leaving theoretical considerations aside for a moment, we read stories of real people who lost their jobs or were experiencing personal turmoil and sought relief through pharmaceuticals. They then underwent drastic personality changes due to side-effects and committed violent acts.[36] Having been law-abiding citizens throughout their lives, it seems unfair and unjust to hold them responsible in the same way that we do persons who become criminal in the ordinary ways. Rather, one is inclined to blame what happened on the drug. Apparently, a notion of authenticity underlies these intuitions: persons who have committed pharmaceutically induced violence are, in a diffuse sense, not really the evil persons their deeds indicate. Somehow the essential core of their personality is good and only darkened by factors alien to their character; hence we hesitate to hold them responsible. On this view, inauthenticity renders the agents nonautonomous.

At this point, autonomy considerations are closely related to notions of authenticity in the enhancement debate, so we must look more closely at authenticity.

[32] The insanity defences date back to the famous M'Naghten case in 1843. M'Naghten had the delusional belief that there was a conspiratorial Tory plot to kill him, so he concocted a preemptive plan to kill the Tory Prime Minister, Robert Peel, but only ended up shooting Peel's secretary. For a comprehensive and critical discussion of insanity tests and recent developments in United States law, see: S. Morse & B. Hoffman. 2007. The Uneasy Entente Between Insanity and Mens Rea: Beyond Clark v. Arizona. *U Penn Public Law and Legal Theory Research Paper Series*. http://papers.ssrn.com/abstract=962945 [Accessed 15 Jan 2009].

[33] Scientifically, the relation between SSRIs and suicide or violent behaviour is a highly contested issue on which we cannot comment here. In 2007 the FDA directed the use of black-boxed warning labels stating an increased risk of suicide.

[34] Here our theoretical case possibly differs from real cases: defendants report that they had no self-control during the incidents, as they felt like spectators to everything their body did.

[35] Of course we oversimplify the way courts deal with Prozac cases. In the US, insanity defences vary from state to state. Globally the picture is even more diverse, as concepts of criminal responsibility in common law and in continental systems differ fundamentally. Moreover, insanity defences and their continental equivalents are subject to a lively debate growing proportionally to increasing understanding of the neuronal underpinnings of criminal behaviour. Nevertheless, our portrait of legal reasoning can be deemed the classic way in which courts deal with such cases. Cf. a special issue on: Responsibility and Mental Impairment. *Int J Law Psychiatry*. 2004: 395–503.

[36] For some of such stories, see D. Healy. 2004. *Let Them Eat Prozac*. New York: New York University Press.

Theories of authenticity can be roughly placed on a continuum between two poles: one pole is rather essentialist and considers authenticity as threatened by everything that makes people depart from who they truly are. This tacitly presupposes that there is a way someone *truly* is – an essentialist self. Authenticity then means to connect one's present person-stage to such a pre-given, rather static self through an introspective journey of self-discovery, and to live accordingly. This conception is often combined with a postulate of gratitude toward the given, which renders altering such traits by neuroenhancements dubious or even impermissible.[37] The other pole is marked by existentialist beliefs. Thrown into this world without any preordained essence, each of us must create himself according to his own ideals. An authentic personality consists of self-defined and self–established characteristics, for which neuroenhancements provide a powerful tool.[38]

These irreconcilable poles underlie – and potentially obstruct – the discussion about authenticity in the enhancement context; however, both poles agree on Beth's inauthenticity in the Prozac case, since the promafinil-induced transformations make her depart from her true self and are not under her control. It thus appears to be a good case to use as a springboard into a discussion of whether inauthentic agents can nevertheless be autonomous.

Essentialist and existentialist ideas are both widely adopted by historical theories of autonomy. Some historicists subscribe to essentialist beliefs. To them autonomy is the preservation and unimpeded development of a self solely through internal resources. Thus agents are only autonomous if a transformation can be traced back to a preceding autonomous decision. In this way the essential self is transmitted over time and conveys autonomy to the respective follow-up stages of its personality.

However, if one starts tracing back an agent's decision to former autonomous decisions, one ends by regressing back to one's birth. Even if there exists a chain unbroken by any external influences that thwart one's autonomy, it makes no sense to consider humans as autonomous at birth. On the contrary, everyone starts life as heteronomous and dependent, and the process of growing up by which one's personality is shaped is primarily defined by nonautonomous events, from parental education through to schooling and many other like influences. How then can someone emerge 'autonomous' solely through reliance on his internal mechanisms? We suspect that a fundamental misconception stems from the translation of autonomy as '*self*-rule.' When the *self* indicates something other than rule by *others*, we are verging on essentialist thinking. Again, the idea is plausible only if there exists an essentialist self, an inner core that is intrinsically good, autonomous, and which can be transmitted throughout a person's life. This is unlikely. Rather than an autonomy-conferring entity within us that needs to be preserved and shielded from outside corrupting influence, we have dispositions that develop by interacting with the world and others. Heteronomous influence is a normal part of life and 'some people literally have to be kicked into autonomy.'[39] Essentialism does not help to solve the problem as to how someone can be autonomous if there is no autonomy to be found in the earliest, 'most historical and most authentic' starting conditions and why agents should be – gratefully – bound by their essence with regard to autonomy. We therefore reject the essentialist view of authenticity.

Other theories of autonomy are influenced by existentialist beliefs demanding acts of self-creation. Agents are autonomous if they are in control of all transformations. As we have shown, neuroenhancements may spark transformations without any decision preceding; according to this conception, then, agents are not responsible. This tendency can be found in Mele:

> There is also a negative historical constraint on autonomy which I have called authenticity ... A necessary condition of an agent authentically possessing a pro-attitude P ... is that it be false that having P ... is, as I will say, compelled* – where compulsion* is compulsion not arranged by S ... [Sometimes] agents come to possess pro-attitudes in ways that *bypass* their control capacities over their mental lives ... Bypassing is sufficient for compulsion ... provided that the bypassing was not itself arranged or performed by the manipulated person.[40]

Thus agents *authentically* possess pro-attitudes acquired through direct interventions that bypass control capacities if, and only if, the agents have arranged for the interventions themselves and foreseen the results. In case (5), promafinil bypassed control capacities and contributed crucially to the formation of Beth's pro-attitude to M. Although Beth took the pill herself, she didn't foresee the

[37] Especially M. Sandel draws on the notion of giftedness in his case against enhancements. 2007. *The Case Against Perfection*. Cambridge, MA: Harvard University Press.

[38] The inconsistencies of an in-between-position that denies a pre-given self and understands authenticity as being true to the self *as it is* are pointed out by Levy in *op. cit.* note 23.

[39] B. Berofsky. 1995. *Liberation from Self: A Theory of Personal Autonomy*. Cambridge: Cambridge University Press.

[40] Mele, *op. cit.* note 17, pp. 166 f.

resulting change and didn't arrange for the change herself; hence her violent desires are inauthentic. According to Mele, she is nonautonomous. Might this sense of authenticity render Beth nonautonomous in (5)?

## VII. AGAINST AUTHENTICITY

We agree with Mele that autonomy is about control rather than some diffuse essentialist sense of authenticity; however, we cannot endorse his high threshold for autonomy. Agents should, of course, have as much control over their personality and pro-attitudes as possible.[41] The problem arises when the formation of pro-attitudes bypass control capacities. Having self-arranged for all of these bypassing transformations is too demanding a condition. If we take that criterion seriously, then the majority of our pro-attitudes would have to be declared inauthentic and all the resulting actions nonautonomous. There is no self-creation *ex nihilo*. From one's sex and other bodily constitutions through to moods, core character traits, behavioural dispositions, social environments and natural endowments, there exist myriad influences on the formation of pro-attitudes that bypass rational control, depend on natural contingencies and are not self-arranged.[42] We should reject that strong existentialist view of authenticity and autonomy; we should face the fact that – to a significant extent – we are the product of external influences beyond our control.[43]

Legal institutions punish people whether or not they have controlled the development of all of their relevant pro-attitudes. Consider an 'authentic' criminal (in the colloquial, not Mele's sense) whom legal institutions do not hesitate to hold accountable. He acquired, so we must assume, a large share of his characteristics through control-bypassing processes as indicated above. However, unlike Beth in (5), he is deemed 'authentic' because he has always been that way. Yet in terms of control we are anything but sure as to whether there is indeed a relevant difference. The criterion of control casts doubt on his autonomy in the same way as it does on Beth's. Authenticity makes sense only insofar as the reason for holding people accountable is to be found in an agent's character. In this regard, the authentically bad character is a more apt target for reactive attitudes than the inauthentic. But people do not get punished for having a certain character – at least not in liberal constitutional states; they get punished for having committed a criminal act. Reprehensible acts bring about, as it were, objective states of injustice in a legal community that have to be remedied and compensated for, irrespective of the perpetrator's character.

Not blaming the authentic or the inauthentic agent if both lack control – a move probably favored by most autonomy theorists – leads to an understanding of autonomy that is considerably more restrictive than the concept employed by courts and legal systems. Taking into account the various sources of the argument, this is no surprise. Whereas compatibilist theories of autonomy are eager to identify ideal conditions in order to dispel sceptical views of human freedom, legal systems presume persons to be free.[44] In the legalistic view of courts, what renders persons autonomous is first and foremost that they can be expected to conform to society's norms.[45] In this regard the etiology and authenticity of an agent's pro-attitudes is irrelevant. Why should the broker on promafinil not be expected to abide by law to the same degree as everyone else?

Furthermore, from a legal perspective, autonomy and responsibility are intrinsically linked to the functioning of the normative system as a whole. The violation of a norm causes a disturbance in its stability and its claim to universal validity within the purview of the legal order to which it belongs, for such a violation is an obvious instance of that claim's failure in one particular setting, and has an erosive impact on the validity of the respective

[41] That is why we hold agential control to be the guiding principle when it comes to illegitimate influence. Of course, we must neglect details of Mele's complex account here. For a deeper analysis, see S. Cuypers (who comes to the same conclusion). The Trouble with Externalist Compatibilist Autonomy. *Philos Stud* 2006; 129: 171–196: 180; and Mele's reply, Manipulation, Compatibilism and Moral Responsibility. *J Ethics* 2008; 263–286: 272.

[42] An interesting discussion of autonomy, authenticity and social identity can be found in D. Meyers. Intersectional Identity and the Authentic Self: Opposites Attract! In Mackenzie & Stoljar *op. cit.* note 16, pp. 151–180.

[43] See G. Watson's description of the life-story of a prime example of an 'authentic murderer.' Responsibility and the Limits of Evil: Variations on a Strawsonian Theme. Repr. in *Free Will: Critical Concepts in Philosophy*. Vol. 1. J. Fischer, ed. 2005. London: Routledge: 106–135.

[44] It is almost impossible to talk about autonomy without discussing free will, so a short comment on the topic: we can remain uncommitted on the question of whether the process of forming pro-attitudes and the follow-up processes that lead to action are deterministic or indeterministic (or whether causal theories are even applicable on a psychological level) and which other factors may be involved (i.e. the uncaused causal power of Kant's transcendental 'noumenal will'). Although we deal with compatibilist theories of autonomy, libertarian autonomy – though rarely developed explicitly – faces similar questions. Cf. I. Haji & S. Cuypers. Libertarian Free Will and CNC Manipulation. *Dialectica* 2001; 221–238.

[45] Hence, Hans Kelsen's notorious statement: 'A person is subject to legal imputation not because he is free; rather, he is free because he is subject to imputation.' Kelsen. 1967. *Pure Theory of Law*. Second Edition. Translated by M. Knight. Berkeley: University of California Press: 98.

norm. If that impact were left intentionally unremedied by the guarantor of the legal order – the state – this would symbolically amount to an abandonment of the norm itself (by abandoning its universal applicability). And this in turn would convey the message to society that the norm was in the process of gradually perishing. When a norm's validity is shaken in the way just sketched, legal systems ascribe responsibility to agents for the purpose of re-stabilization.[46] Were such perpetrators let off the hook too easily, norms would, in the long run, be substantially undermined as to their functions and effectiveness. Practically speaking, how should Beth in (5) be treated? If we deny her autonomy because of inauthenticity, she has a *carte blanche* to violate norms as she pleases. But norms do not sustain uncompensated contravention, and legal systems cannot endure loopholes in the responsibility of people who, by and large, are 'normal' (not 'insane' in the legal sense). We concur with the above legal reasoning: in order to maintain norms, Beth should be considered autonomous – and held responsible.

We therefore propose the following relationship between neuroenhancement and autonomy: if agents who possess the minimal autonomy capacities self-initiate neuroenhancements and then identify with the results, they are autonomous. If they are manipulated or do not identify with the results, then they are not autonomous. Authenticity – in either the essentialist or existentialist understandings of that term – is not necessary for autonomy.

Without essentialist or existentialist authenticity conditions, autonomy (in nonmanipulation cases assuming minimal autonomy capacities) is therefore reduced to identification; in Prozac cases, this may prompt a counterargument to our position. The Prozac cases show that enhancements which modify moods or emotional propensities may have self-legitimizing effects: insofar as they generate positive emotions, increase well-being, promote self-worth and the feeling of 'really being oneself,' they foster identification with the new personality.[47] When *the process of identification itself* is enhanced, the agent lacks an independent vantage point from which to evaluate their new traits. Our conclusion – that the agent is nonetheless autonomous – may leave a residue of uneasiness.

To which we reply: why should agents consider new traits from their old perspective? Why should the former, presumably authentic personality have normative priority over the current one as long as the agent identifies with it? Will Beth be nonautonomous for the rest of her life? Can she not regain her autonomy by post-transformative identification?[48] Should paternalistic interventions aiming to restore her former 'authentic personality' (against her will) be permissible? We think not. As long as she identifies with her personality, she can be anything she wants. The price for this freedom is accountability.

## VIII. THE DIFFERENCE BETWEEN NATURE AND PERSONS

One last possible objection needs to be addressed – the objection that our view makes too much of the distinction between compulsion by nature and compulsion by other agents. To some it may appear inconsistent to conclude that agents manipulated by other people are nonautonomous (3), whereas agents suffering from unanticipated side-effects (5) are autonomous. From the agent's perspective, it indeed makes *no* difference whether his proattitudes are compelled by someone else or influenced by natural forces – both are influences beyond his control.

We reply that, from a normative perspective, there is a widely agreed difference. Consider two cases:[49]

> *Doctor D asks patient P for consent to remove his cancerous kidney – otherwise, it is certain that P will shortly die, say within the next month. P consents and the kidney is removed.*

This is significantly different from:

> *Being held at gunpoint, P is 'asked' to consent to the removal of his kidney in order to transplant it to the coercee's son; otherwise P will die with a bullet through his head. P consents and the kidney is removed.*

Although the psychological pressure on P – the threat of imminent death – is the same in both cases, and neither causal forces are under his control, there is no doubt that in the cancer case P's consent is valid, whereas it is invalid when given under coercion. Joel Feinberg argues in a similar vein:

[46] Cf. N. Luhmann. 2004. *Law as a Social System*. Oxford: Oxford University Press: 140–150; G. Jakobs. Imputation in Criminal Law and the Conditions for Norm Validity. *Buffalo Crim Law Rev* 2004; 491–511.

[47] For stories of antidepressants making consumers 'feel like themselves' see P. Kramer. 1993. *Listening to Prozac*. New York: Penguin.

[48] Bolt (*op. cit*. note 2, p. 294) criticizes DeGrazia (*op. cit*. note 1, p. 102, fn. 37) for favoring, as we do, retrospective identification. Bolt does not have in mind the normative implications that follow from her account – she is in pursuit of stronger notions of autonomy and authenticity. The same holds true for I. Hyun. Authentic Values and Individual Autonomy. *J Value Inq* 2001; 35: 195–208. We are in line here with the F&R idea of taking responsibility for a new trait by acting on and identifying with it. Fischer & Ravizza *op. cit.* note 19, p. 234.

[49] Merkel *op. cit.* note 3.

> If [a] threat comes from nature rather from any other person ... we would take that threat simply to be one of the background conditions ... rather than an intervening force rendering his decision involuntary. The mere grimness of all alternatives to the action is often nothing more than the 'legitimate inequalities of fortune' which all of us must inevitably confront.[50]

We are aware that this distinction has been criticized in the free will debate; nonetheless, the kidney case plausibly shows a relevant normative difference between nature and persons.[51] Nature may sometimes be harsh, even cruel, but never illegitimate.

Residual intuitions of unfairness can still be accommodated to a degree: authenticity can provide guidance for the appropriate reactive attitude. The difference between 'authentic' and 'inauthentic' criminals can be recognized by applying different sanctions. If there is a 'monocausal' explanation for a reprehensible act, the effect should be reversed whenever possible. Pharmaceutical side-effects may be eliminated or mitigated through withdrawal and discontinuation; for rehabilitation, this is better suited than traditional punishment. And in this regard neuroenhancements may prompt a radical rethinking: perhaps neuroenhancements can aid in the rehabilitation of even authentic offenders by enhancing morally valuable traits, such as empathy, or by inhibiting aggressive behaviour.[52] Nevertheless, punishment serves more and other purposes than rehabilitation. It has inherently retributive and norm-stabilizing functions. These burdens legally constituted societies must place on any autonomous actor's shoulders.

Lastly, a short comment on a final scenario:

### (6) Involuntary intervention, foreseeable result, identification

*Due to a financial crisis, there is great competition for jobs in the brokering business. Beth thinks she has to take powerfinil to keep up with her neuroenhanced colleagues; otherwise, the chances are high that she'll lose her job. Did she take powerfinil autonomously?*

Once again, the sheer weight of the psychological pressure is not the decisive criterion for autonomy. The question is whether the pressure is illegitimate and, additionally, whether a person can be expected to withstand it. Beth can be expected to withstand the workspace pressure. As far as autonomy is concerned, persons can even be expected to withstand illegitimate pressures (particularly in cases of pressures to harm others). Professional athletes are likely to be subject to enormous pressures to take performance-enhancing drugs; nevertheless we deem them appropriate targets for reproach and sanctioning when they contravene regulations. Normatively, we must expect the athletes to comply. Politically, it might be wise to regulate highly competitive fields in order to protect individuals from illegitimate compulsive forces and to serve the interests of social justice.

## CONCLUSION

Neuroenhancements can cause consumers to feel self-alienated and impede their identification with new traits, rendering them nonautonomous. But even when agents identify with their enhanced traits, their autonomy appears doubtful according to some theories of autonomy. Some claim that pro-attitudes transformed by direct brain interventions such as neuroenhancements derive from mechanisms that are not the agent's own; hence, the resulting actions are nonautonomous. This is plausible only insofar as agents are manipulated by other agents who then bear primary responsibility, thus exempting the manipulated agent. What counts as manipulation can be broadly assessed by our proposed framework of illegitimate influences. Interventions on the neuronal level against the will of the agents are illegitimate, and hence shift responsibility onto the manipulator. If agents self-initiate such transformations, the only reason to conceive of them as inauthentic or nonautonomous would be that they are not true to an essentialist self. Autonomy, however, cannot be grounded upon such a notion. When agents enhance their traits intentionally and identify with the results, they are autonomous. In the most challenging enhancement scenario (5), agential personality is transformed without any preliminary decision, yet afterward the agent identifies with the results. In such a case, the transformative process was not under agential control; still, we contend that it suffices for autonomy as long as the agent subsequently identifies with the newly acquired trait.

[50] J. Feinberg. 1987. *The Moral Limits of the Criminal Law 3: Harm to Self*. Oxford: Oxford University Press: 196.

[51] It has also been criticized by A. Fenton in his review of Merkel et al. *Neuroethics* 2008; 215. However, Fenton seems to misunderstand (or, at any rate, erroneously not accept) that legitimate pressures originating from the compulsive force of legal norms (of a largely legitimate order) are akin to natural forces but not to blackmailing forces exerted by other agents. Furthermore, the nature/person distinction does not entail that agents are unentitled to a defence of necessity when threatened by a natural calamity. This is a matter of weighing the harm inflicted against the harm avoided, and it pertains to the justification of an autonomous act; Prozac cases question the agent's autonomy in *unjust* acts.

[52] Cf. T. Douglas. Moral Enhancement. *J Appl Philos* 2008; 228–245.

In this paper, we initially set out to clarify the relation between autonomy and authenticity and possible threats posed by neuroenhancements. We claim that authenticity is not an adequate condition for autonomy and that neuroenhancements do not threaten autonomy in principle. Finally, although we have argued that authenticity is not a necessary component of autonomy, we firmly believe that authenticity has important value *beyond autonomy*: social relationships are built on stable and enduring conceptions of other people. And these may be threatened by pharmaceutically induced changes in personality. After all, would you marry someone who asked for your hand under the spell of lovafinil, a potent love potion?

**Acknowledgements**

This paper was written in the larger framework of an interdisciplinary research project entitled 'Potentials and Risks of Psychopharmaceutical Enhancement' at Europäische Akademie, Bad Neuenahr – Ahrweiler, Germany. The authors would like to thank two anonymous reviewers and the editors of Bioethics, the collaborators of the project and especially the project's coordinator, Dr. T. Galert, for their substantial contribution and critical comments in the preparation of the manuscript, and Mr. K. McAleer for invaluable help in improving the style of our text. The work was supported by the German Federal Ministry of Education and Research (BMBF), and for Reinhard Merkel made possible by the generosity of his current scientific host, the marvelous Wissenschaftskolleg zu Berlin (Institute for Advanced Study).

**Jan Christoph Bublitz** is a research fellow at the University of Hamburg, Institute for Criminal Law and member of the project group 'Potentials and Risks of Psychopharmaceutical Enhancement' at the Europaeische Akademie for Technology Assessment, Bad Neuenahr-Ahrweiler. He is currently finalizing his PhD Thesis on 'Legal Implications of Neuroscience'.

**Reinhard Merkel** is professor for criminal law and legal philosophy at the University of Hamburg. Currently he is a fellow at the Wissenschaftskolleg zu Berlin (Institute for Advanced Study). He recently published *Intervening in the Brain: Changing Psyche and Society* (ed. Springer, 2007).