Routledge Advances in Behavioural Economics and Finance
Edited by Roger Frantz

Traditionally, economists have based their analysis of financial markets and corporate finance on the assumption that agents are fully rational, emotionless, self-interested maximizers of expected utility. However, behavioural economists are increasingly recognizing that financial decision makers may be subject to psychological biases, and the effects of emotions. Examples of this include the effects on investors' and managers' decision-making of such biases as excessive optimism, overconfidence, confirmation bias, and illusion of control. At a practical level, the current state of the financial markets suggests that trust between investors and managers is of paramount importance.

Routledge Advances in Behavioural Economics and Finance presents innovative and cutting-edge research in this fast-paced and rapidly growing area, and will be of great interest to academics, practitioners, and policy-makers alike.

All proposals for new books in the series can be sent to the series editor, Roger Frantz, at rabeandf@gmail.com.

Behavioural Economics and Business Ethics
Interrelations and Applications
Alexander Rajko

Bounded Rationality and Behavioural Economics
Graham Mallard

Behavioural Approaches to Corporate Governance
Cameron Elliott Gordon

Trusting Nudges
Toward a Bill of Rights for Nudging
Cass R. Sunstein and Lucia A. Reisch

Social Neuroeconomics
Mechanistic Integration of the Neurosciences and the Social Sciences
Edited by Jens Harbecke and Carsten Herrmann-Pillath

For more information about this series, please visit: www.routledge.com/Routledge-Advances-in-Behavioural-Economics-and-Finance/book-series/RABEF

Social Neuroeconomics

Mechanistic Integration of the Neurosciences and the Social Sciences

Edited by Jens Harbecke and Carsten Herrmann-Pillath

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

11 Neuroscience of autonomy and paternalistic policies

Jaakko Kuorikoski, Samuli Reijula and Susanne Uusitalo

Abstract: Policies designed either to keep us from doing harm to ourselves or to others, or to improve our wellbeing by correcting for predictable biases in our behaviour, are justified by the idea that there is something wrong with our agency. The legitimacy of influencing our behaviour, often against our momentary, and sometimes even long-term, preferences, is based on the assumption that we would choose to act in the way intended by the paternalist, if only we were in full control of our actions and faculties. Early proponents of neuroeconomics have suggested that neuroeconomics can shed light on the neural basis of valuation and decision-making in a way which could be relevant for assessing the well-being and, by implication, autonomy of decision-makers. Can neuroscience inform us about the conditions under which we are not in sufficient control of ourselves?

In this chapter we discuss neuroscience research relevant to understanding the neural basis of self-control, motivation, and choice. Our focus is on cases of less-than-ideal agency, which do not yet amount to clearly defined pathologies but might warrant external intervention. We focus especially on the neuroscience of addiction and neuroeconomics of choice. We problematize the appropriate notion of control in this context and argue that the normative concept of personal autonomy relevant to the justification of paternalistic and coercive policies cannot be reduced to abnormalities in the neural mechanisms regulating our behaviour. We appeal to the literature on relational autonomy in arguing that autonomy is partly socially constituted and irreducibly normative.

Introduction

In the early days of neuroeconomics, the power of the neuroscientific methods was occasionally argued to have far-ranging normative implications in the future. For example, Park and Zak argued in 2007 that neuroeconomics had already made substantial progress in answering questions such as "How much regulation is optimal?" and even "How to achieve happiness?" Camerer, Loewenstein and Prelec (2005) also express enthusiasm about the potential of neuroscientific methods to shed light on not only how we make decisions, but also how we *should* make decisions and, by implication, when we could improve upon the decisions actually made. At the same time, social psychology has

taught us that we are often quite oblivious to what actually drives our actions. Perhaps looking inside our skulls can teach us something important about who we are and what is good for us?

Consider a not totally science fiction thought experiment: We image the functioning of the neural decision-making network for agent A in a choice context C and observe that as it comes to A's brain functions, everything seems to be going well. Can we conclude that the choice made was autonomous? Now consider a purely science fiction thought experiment: We have mapped agent A's ventromedial prefrontal cortex (vmPFC) activity in multiple choice situations and used a machine learning algorithm to reconstruct their preferences from the data. We therefore "read" the agent's subjective utility from the brain measurements and observe that their choice was made according to their utilities. Can we now conclude that the choice made was autonomous and in A's own best interests?

In what follows, we will clarify what we mean by the normative nature of making right choices, i.e., autonomy and its relation to wellbeing. We will argue that a view of autonomy that ignores the constitutive role of the social context falls short of capturing normatively adequate autonomous agency. In order for the neuroscientific evidence to qualify as decisive in the assessment of individual agency, the agential conditions need to accommodate this insight. Conceptualizing "the ingredients of autonomy", namely competence and authenticity, as categorical states intrinsic to the individual runs the risk of reducing agency from its normative dimension. A further consideration is that policies almost invariably target populations, not individuals, which may further facilitate the reductionist approach to assessing individuals. Nevertheless, the potential that neuroscience holds for understanding individual decision-making should not be wasted. If paternalistic policies require justification based on assessment of individual autonomy, then neuroscientific evidence could, and indeed can, cast light on some of the relevant sub-personal capacities and flag circumstances which can undermine them. We accept, for the sake of argument, that any policies affecting individual choices have to respect individual autonomy (barring considerations of harm to others). Our conclusion is that in order for neuroeconomics to achieve policy relevance, it has to more substantially include considerations of the social scaffolding of agency and autonomy, i.e., to become social neuroeconomics.

Autonomy and self-control

Individual autonomy is a foundational value in modern societies. Self-government is considered an important capacity as it provides people the chance to make meaningful choices in their lives in a controlled manner, and as they make them, they can be held responsible for such choices. This is a core ideal in individual lives as well as in democratic societies. But sometimes people clearly choose in a way that goes against their own (and others') wellbeing. What should we make of such cases? Are such agents really autonomous and appropriately in control over their actions? Or is there something wrong in their decision-making, and is it the case that were their decision-making abilities to function appropriately, they would choose the right option, live the right way?

Individual autonomy is usually considered as an intrinsically good thing. Policies promoting the realization of autonomy are therefore desirable in virtue of this very reason, and any policies infringing on individual autonomy require strong justification. In this chapter we focus mainly on the latter kind of relevance, the ability of neuroscience to reveal deficiencies in autonomy.

The first step in our argument is the distinction between self-control and autonomy. We use "self-control" to refer to the internal processes of behaviour regulation, such as executive and impulse control, attention, and memory. Although in the research literature there have been various ways of using the notion, self-control (or regulation) generally refers to the capacity of the organism (and the processes involved) to control its thoughts, responses, or behaviour, typically in ways that are guided by its goals and purposes (Grouzet et al. 2013, pp. 3–4). Psychological and neurocognitive theories of self-control usually model oppositions between top-down and bottom-up processes, cold and hot processes, and long-term and short-term motivations or goals. Much of the top-down, cold, long-term processing is taken to happen in prefrontal regions, whereas the bottom-up, hot, short-term resides in subcortical areas (Kelley et al. 2014; cf. Hommel and Wiers 2017).

Internal self-control capacities can fail in a number of distinct ways. For example, Kotabe and Hofmann (2015) provide a taxonomy of self-control failures, according to which such failures can be broadly divided into conflict-based and control-effort-based. The former encompasses different problematic outcomes of failures to resolve conflicts between immediate desires and higher-order goals, e.g., when strong momentary desires override higher-order long-term goals, and when cognitive load hinders the processing of higher-order goals. The latter covers cases in which desires are enacted due to insufficient control motivation or insufficient control capacity, i.e., top-down control is for some reason or another "under-resourced" or the relevant executive capacities are, for some reason, defective. We do not take a stance on whether this or some other model of self-control failure is the correct one. In fact, Hommel and Wiers (2017) have questioned whether the usual juxtapositions between endogenous and exogenous, intentional and habitual, are a fruitful starting point for modelling action control to begin with. What is important here is that all of these processes, and failures therein, are internally realized by individual psychological mechanisms.

By autonomy we refer to the normative status of being self-*governing* – to, roughly, possessing and being able to competently pursue one's *own* (authentic, endorsed and potentially reflected, perhaps reasonable) interests in an environment that allows this kind of agency. We suggest that self-control must be distinguished from self-governance and, hence, autonomy. Deficiencies in self-control do not automatically translate into failures of autonomy or vice versa: what is required for such competence in pursuing goals, and what makes it the case that a goal or desire is aligned with the agent's authentic interests involve irreducibly normative considerations. Furthermore, many of the factors constitutive of autonomy are socially extended and distributed outside the agent's head. As John Doris (2015) has argued, the reasons in light of which our actions and projects unfold are socially negotiated, mined from jointly woven

life-narratives. And, as Tadeusz Zawidzki (2013) has argued, the function of the normative stories we tell each other about the causes of our behaviour is to facilitate coordination of social behaviour (see Chapter 9).

Failures of self-control are common and often momentary – everyone occasionally makes errors in judgement and suffers from weakness of will. Failures of autonomy, in contrast, are more fundamental and (usually) evaluated from a longer temporal perspective. A failure of self-control is an error in choice (in the internal decision-making machinery), whereas a failure of autonomy is a shortcoming or breach in agency. Self-control problems begin to impair autonomy when they become systematic and begin to hinder the pursuit of meaningful and self-endorsed projects of the agent. Most importantly, self-control processes are *sub-personal*, whereas autonomy is an agent-level normative status. Many of the agent-level capabilities constitutive of autonomy are doubtlessly realized by sub-personal self-control mechanisms (executive and impulse control, working and episodic memory, different attentional mechanisms, etc.), but there need not exist any simple mapping between personal autonomy and sub-personal cognitive and affective mechanisms.

Conversely, breakdowns in self-control processes are not necessary for the failure of autonomy. Severe indoctrination and oppression are, at least arguably, ways in which the agent's actions can be alienated from their authentic values and goals. In fact, such non-autonomous actions often require exceptionally strong sub-personal self-control capacities (think of physically or psychologically demanding rituals or extreme acts of violence).

Nevertheless, autonomy should not be conceptualized as a unitary, stable, context-independent, and binary status (Mackenzie 2014). As the bioethical and legal debates over patients' capacity to make informed decisions about their treatment have demonstrated, people can at the same time competently pursue their interests in one domain (e.g., medical), while still being severely limited in their capacities to understand and manage their life in others (e.g., legal and financial) (Hooper and Chiong 2017). The model of the assessment of decision-making capacities of the elderly by Moye et al. (2013) also suggests differentiating capacities with respect to different kinds of decisions, such as those involving low and high risk. Autonomy-relevant capacities may also change over time and, as we will stress later, changes may ensue from changing psychological, social, and material conditions external to the agent.[1]

Conceptions of autonomy can be divided into two main camps, procedural and substantial, and most real-world normative applications involve a mixture of the two. A procedural view defines autonomy purely in terms of the properties of the decision-making process (at the agent level): an agent is procedurally autonomous if they make decisions in "the right way", regardless of the content of their preferences. Usually procedural autonomy also demands at least some level of stability of preferences over time: a fully self-governing agent must be able to formulate plans and control their actions so as to achieve long-term goals in the face of incompatible short-term urges – at least to some extent. In contrast, according to substantial views of autonomy it matters not only *how* people choose but also

what they choose, i.e., that they also prefer *the right* things (Mackenzie and Stoljar 2000). For example, an orthodox Kantian view would be that an agent cannot be fully self-governing unless they act according to the categorical imperative.

We do not dwell deeper on the more subtle philosophical distinctions with regard to autonomy, as the distinction between procedural and substantive autonomy is the most relevant for paternalism and policy purposes more generally. Furthermore, practically all conceptions of autonomy imply two broad classes of conditions of failure. As autonomy broadly means the competence to pursue one's own interests (in an environment that allows this), it can fail if either (a) the interests are not pursued competently (*competency* conditions), or (b) that the pursued goals are not really the agent's own (*authenticity* conditions). Failures of autonomy are often due to problems in an individual's internal self-control, but they are not reducible to them, nor does there need to exist any simple mapping between failures of autonomy and failures of self-control.

Protecting individual wellbeing

Paternalism is "the intentional overriding of one person's known preferences or actions by another person, where the person who overrides justifies the action by the goal of benefitting or avoiding harm to the person whose preferences or actions are overridden" (Beauchamp and Childress 2001, 178, df. in biomedical ethics) – i.e., an action or policy which violates agent's autonomy in the name of the good of the agent. A strongly paternalistic policy ignores the agent's preferences altogether (the policy maker knows best), whereas a softly paternalistic policy influences the agent's decision-making with the goal of helping the agent to competently pursue their authentic preferences (Thaler and Sunstein 2008), e.g., by providing them with relevant information.

Antipaternalists oppose (at least strong) paternalistic interventions. This is because, so the argument goes, such interventions violate individual rights and restrict the individual's free choice in an undue manner. Proponents of antipaternalism can also argue that paternalistic standards are too broad; paternalism would authorize and institutionalize too much intervention if made the basis of policy (Beauchamp and Childress 2001, 182). However, there are many cases of lack of autonomy, in which even staunch antipaternalists would not necessarily see harm in strongly paternalist interventions (such as intervening on the behaviour of small children) (Beauchamp and Childress 2001, 183).

Justification of paternalism focuses on different aspects of decision-making depending on the kind of criteria we accept for self-government. In light of the distinction discussed earlier, regardless of the specific theory of autonomy, we can divide such aspects to be either about the competence conditions or about the authenticity conditions. In cases of serious failure of either, a strong paternalist would allow for influencing (usually limiting) the possible actions of the agent, regardless of their current preferences. A soft paternalist would allow for influencing the choice of the agent only in cases in which a failure in either competence or authenticity leads the agent to act against their authentic preferences.

Now consider the role of neuroscience in producing evidence of (in) authentic preferences. Both the soft and strong *neuropaternalist* have to be able to provide an answer to what could be called the *so-what challenge*: does not behavioural evidence (including verbal behaviour) tell us everything we need to know about the practical rationality, reason-responsiveness, and, possibly, ends and values, relevant to the evaluation of autonomy? The neuropaternalist can adopt either a strong or weak stance on this issue of the relevance of neuroscience. According to a strong thesis, there are cases in which neuroscientific evidence remains relevant for the evaluation of autonomy even after all the (possible) behavioural evidence is in. A weaker thesis maintains that there are cases in which neuroscientific evidence is relevant for evaluating autonomous agency, because not all (possible) behavioural evidence is available.

Furthermore, neuroscientific evidence could be relevant for paternalism in two ways: it could provide evidence of a failure of either competence or authenticity (i.e., in the identification of impaired agency), or it could provide evidence about choice contexts which are liable to lead otherwise autonomous agents to suffer serious failures of either competence or authenticity. We discuss these possibilities in turn.

Soft neuropaternalism for impaired agents

Let us first consider whether neuroscience can be informative for identifying *impaired agents* – agents with chronic anomalies in their decision-making and self-control machinery serious enough to undermine their autonomy. In health care ethics, one of the issues that measure a person's autonomy-relevant competencies is decision-making capacity (see, e.g., Appelbaum 2007; Charland 2015). Assessing the competence of older adults with age-related cognitive decline to manage their affairs also involves making judgements about decision-making capacity. In the health care context, decision-making capacity requires not only the ability to understand the issue at hand (the diagnosis and possible courses of treatment) and what follows from it (possible outcomes and associated risks), but also to appreciate what it means in relation to one's own life. This also concerns one's preferences, and decision-making capacity is therefore also argued to include an essential component related to the values of the agent (Charland 2015; Peterson 2018). The assessment of the authenticity of those preferences and values, however, can be problematic. Arthur Caplan (2006), for instance, argues that individuals with addiction fail to appreciate abstinence and thus fail to seek the option of recovery in addiction. He draws an analogy to individuals who have experienced severe trauma. The initial reactions of such individuals to refuse treatment are indeed likely to go against their authentic preferences. In this light, even Caplan's suggestion of the legitimate coercive treatment would be along the lines of soft paternalism.

Serious mental illnesses or disabilities are usually considered as candidate cases of clearly pathological agency – few would deny justification for stopping a psychotic person from harming themself or others – and some neuroanatomical correlates

for several psychopathologies are relatively well established (although not in broad diagnostic use). Patients with vmPFC damage may be unable to motivate themselves to act according to their own long-term interests and may fail to suppress sudden urges, while discursively acknowledging that they really ought to act otherwise. Studies of patients with lateral prefrontal cortex damage report problems in planning, maintaining, and coordinating among complex goals (Kelley et al. 2015).

Our primary aim in this chapter is not to add to the literature on the criteria and the assessment tools concerning clinically impaired agents, but to look into the possibilities of using neuroscience to shed light on instances of less-than-ideal agency. What we mean by this is the domain-specific autonomy of non-pathological populations possibly relevant for social neuroeconomics. But what kind of evidence could warrant soft neuropaternalism in less clear-cut cases of compromised autonomy? There are studies which suggest that neuroscientifically measurable individual differences in self-control in laboratory conditions predict differences in behaviour outside the lab. For example, Demos et al. (2012) demonstrated that individual differences in ventral striatal responses to food cues predicted subsequent weight gain in a six-month follow-up, and in a study by Lopez and colleagues (2014), increased activity in the left inferior frontal gyrus (IFG) during a standard go/no-go self-regulation task predicted successful restraint regarding food temptations outside the laboratory. Furthermore, Lopez at al. (2016) demonstrated that a measure of relative imaged activation of executive control and reward related areas predicted differences in eating behaviour outside the laboratory. The Demos et al. study also showed that neural responses to images of erotic scenes predicted individual differences in sexual interest. Casey et al. (2011) suggest that individual differences in abilities to resist temptation and delay gratification remain stable across time (decades) and are associated with differences in frontostriatal activation related to motivational and control processes. Nevertheless, although people certainly are different, only a few would be ready to admit that modest quantitative differences in appetitive desires and resources of self-control would amount to serious failures of either competence or authenticity.

Addictions are an important and contested case of possibly impaired autonomy. Whether people with serious substance use disorders have impaired decision-making capacities, and whether their stated preferences reflect, in some sense, their authentic agency, has direct implications to policies of treatment and substance control. The neuroscience-based brain disease model of addiction stresses the importance of long-term physiological and functional changes in the brain caused by long-term substance use. Such functionally and anatomically congruent changes include sensitized reward response, stronger stress-reactivity, impaired executive and impulse control, reduced ability for reflection and insight, and changes in the direction of attention (Koob and Volkow 2016).

Much ink has been spilled in relation to whether these changes imply changes in the authentic preferences of individuals with addiction. For instance, Charland (2002) argues that addiction involves a change in one's values, whereas Heyman (2009, 145) discusses the toxic nature of addiction in which the

substance use "poisons" other options, thus making the addictive reward preferable. Metaphorically, addiction hijacks one's values and harnesses the individual's preferences to its own benefit. This has been contested, as some instead characterize addiction as acting in accordance with one's long-term authentic preferences and still resulting in suffering.

Are paternalistic treatments of addiction ever justified? We take a reasonable operationalization of (at least procedural) autonomy in this context to be that of the ability to give informed consent for treatment. The soft paternalist would therefore regard such paternalistic treatments as justified if, counterfactually, it would have been possible to obtain the person's informed consent for the paternalistic intervention. Beauchamp and Childress (2001, 80) provide the standard categorization of elements of informed consent in health care. The first category is threshold elements that are the preconditions for consent. These include competence (to understand and decide) and voluntariness (in deciding). The second category contains the information elements: consent requires disclosure of material information and a recommendation of a plan, and that the individual demonstrably understands these two issues. This may be of limited relevance to our case. The third category involves the actual consent, and it has two elements: the decision in favour of the plan, and the authorization of the plan.

How can neuroscience of addiction answer the so-what challenge in the case of addiction, i.e. tell us something about counterfactual informed consent to the intervention that goes beyond behavioural evidence? Assessments of competence carrying well-defined legal (custodial) implications are usually made with the help of standardized capacity assessment tools, such as the MacArthur competence assessment tool (MacCAT-CR). These are sets of questions to be used by a care professional in a semi-structured interview. It has been argued that, at least in cases in which communication with the patient is difficult or impossible, neuroscientific evidence could offer supportive evidence to such standardized tests (Peterson 2018). After reviewing the widely accepted neuroscientific results, Carter and Hall (2011) come to the conclusion that despite the empirically well-established long-term changes in functionality and anatomy brought about by many addictive substances, when asked for informed consent for treatment (when not currently intoxicated or under acute withdrawal), most individuals with even serious substance use problems retain adequate decision competence for autonomous choice – because they are demonstrably competent according to behavioural evidence and standardized measures. For example, a study by Morán-Sánchez and associates (2016) measures the decision-making capacity of individuals with serious substance use problems using the MacCAT – CR tool and clinical interviews. They found that the majority of the subjects exhibited adequate competence for informed consent, although roughly a third had such serious impairments (especially in understanding the relevance of the considered treatment to their own lives, perhaps reflecting problems both in competence as well as authenticity) so as to be judged lacking in decision-making capacity.

As autonomy is partly constituted by the reason-responsiveness of the agent, it is hard to see how neuroscientific evidence could ever override such overt behavioural evidence: if an individual was demonstrably able to answer relevant questions, make correct inferences about the consequences of available information, and be correctly motivated to act in the light of these consequences, it would be hard to see how a brain scan or functional imaging data could prove such performances somehow illusory. However, although behavioural evidence strongly suggests that people with addictions should not be seen as lacking in autonomy as a default, a significant portion of the subjects in the study by Morán-Sánchez et al. did exhibit significant impairments in decision-making capacity. This is congruent with the neuroscientific picture of addiction and, at a general level, neuroscience could therefore be seen as providing important *corroborating* evidence for *the potential* of addictive substance use to impair autonomy.

At the individual level, even though behavioural evidence would in principle be exhaustive concerning authenticity and competence, all such relevant evidence is never available in practice. There is no conceptual reason why a set of neuromarkers could not outpredict a limited set of behavioural data relevant for the assessment of, say, individual decision-making capacities. In an ambitious and programmatic proposal for complementing current diagnostic practice in relation to addiction, Kwako et al. (2016) propose the use of imaging data in a cue reactivity task and a monetary incentive delay task to assess possibly problematic features in incentive salience, and imaging data in facial emotion matching task to assess negative emotionality. The predictive power of such neuromarkers with regard to deficiencies in sub-personal decision-making and control mechanisms is obviously a purely empirical question. Nevertheless, as such deficiencies in sub-personal processes are neither sufficient nor even strictly necessary, for failures in agent-level autonomy, neuromarkers can only flag possible problems in autonomy. They should never be taken as decisive.

Soft neuropaternalism for pathological situations

Consider now the question whether neuroscience can help in diagnosing situations in which otherwise competent agents are liable to act against their better judgement. For example, emotional and social distress have been shown to increase activation in brain areas related to assessing reward cues, and this may be a central neural mechanism linking social and emotional stress with problems in self-control (Wagner et al. 2012).

An important normative foundation for soft paternalism is the conception of authentic preference: influencing the choice of an agent is justified if the intervention increases the probability that the agent will choose according to the preferences that they truly identify with or at least reflectively accepts. If the neuropaternalist could reliably detect situations in which an otherwise competent person's neural decision-making machinery predictably fails, "nudges" correcting for this failure would be autonomy preserving, not undermining.

It is well known that soft paternalists face the daunting problem of providing an account of what makes some set of preferences truly authentic (preference identification problem) – let alone how we could come to know such things

(Infante et al. 2016; Reijula and Hertwig 2020). The concept of revealed preference often used in economic welfare assessments is obviously inadequate, as observed behaviour is more likely a mix of several sets of competing preference orderings. As has been argued by, for example, Ainslie (2000), there is little normative reason to automatically equate long-term preferences with the will of the true "self". Furthermore, Guala and Mittone (2015) argue that, in the welfarist framework, the preference identification problem is simply intractable.

Here we might find a possible answer to the so-what challenge: since we simply cannot assume that people's preferences are the same or even similar to those of others subjected to the same intervention, and an individual's behaviour can be caused by several competing preference orderings, perhaps behavioural evidence is simply insufficient for justifying an intervention. Could neuroscience evidence help uncover people's true preferences (internal states actually motivating observed choices), or even authentic preferences (the preferences the agent identifies with)?

According to a widely accepted hypothesis, a population of neurons in the orbitofrontal cortex (OFC) and vmPFC integrates decision-relevant inputs from multiple stimulus types and bottom-up and top-down processes and hence computes a single subjective value representation (Padoa-Schioppa 2011). Some neuroeconomists have taken quite literally the idea that measured OFC, vmPFC or striatal activity could be taken to be informative of preferences. For example, Krajbich et al. (2017) propose the idea of neurometrically informed mechanism design:

> A fundamental assumption behind the classic impossibility results in mechanism design is that the only way the planner can gain information about individual preferences is by eliciting them behaviourally through a cleverly constructed mechanism. Although this has been a valid assumption for the last 30 years, modern neurometric technologies are now making it possible to obtain direct, but noisy, signals of subjects' preferences.

Such neurometrically informed market design or choice architectures are, as of yet, speculative social science fiction. The suggestion points, however, to an important and thus far-unresolved philosophical question about the relationship between the state of the valuation system and our preferences (cf. Fumagalli 2013). When the early neuromarketing results claimed that people in general really "liked" Pepsi more than Coca-Cola, the results were easy to dispute on the grounds that the methodology was ill-equipped to discriminate between motivation- and reasoning-relevant brain activation and other brain responses (the difference being caused by Pepsi simply having more sugar in it). With the current improved understanding of the neural basis of subjective valuation, this response is not as convincing. The subjective value function integrates all decision-relevant inputs and is therefore, in a sense, a neural-level preference ordering – a comparative all-things-considered valuation (cf. Hausman 2011). It is just that such an all-things consideration is a sub-personal, not an agent-level, process. The value appears to incorporate the momentary inputs from top-down and bottom-up processes and various sensory modalities. Of course, the top-down processes may have long-term "content", but the integration happens here and now. There are therefore no guarantees that this brain-level preference ordering corresponds to authentic preferences of the agent – or any other way we wish to define the normative, autonomy-relevant preferences corresponding to the agent's genuine interests. In fact, we have every reason to expect that this brain-level preference ordering is highly unstable across time. Even if neuroscience could distinguish between "true" preferences, something that actually motivates action in a given choice situation, from "mere" impulses, drives, and the like, it could not tell us which of these true preferences were authentic (reflectively endorsed and in line with the agent's practical identity).

Finally, we do not necessarily have to engage in such speculation in searching for a plausible role of neuroscience in informing when a choice situation, instead of any specific agent, could be considered "pathological". Neuroscientific evidence of the brain mechanisms of gambling has been used as an argument for considering pathological gambling on a par with substance addictions, with similar implications for policy. The arguments for this are that gambling activates the same areas and pathways as the use of addictive substances and, crucially, results in some similar long-term anatomical and functional changes.[2] One of the most politically important consequences of these arguments was the inclusion of gambling disorder under the substance-related and addictive disorders in DSM-5. Insofar as serious substance addictions can impair the autonomy of people suffering from them, neuroscience has provided (at least corroborating) evidence for regarding some forms of gambling as just as harmful.

Relational autonomy and socially scaffolded agency

Human agency by nature is filled with bias and partiality. Yet, we would argue, in a social context we need to be able to (and typically do) distinguish autonomous individuals from non-autonomous ones. Likewise, we are able to distinguish between different kinds of actions these agents perform. Such competences allow us not only to judge questions of responsibility, praise, and blameworthiness of the action, but also in certain situations to guide and facilitate individuals' agency and action. Substantive accounts of autonomy have been criticized for assuming certain universal values to be constitutive of autonomy, leaving no room for variation in personal, cultural, and social preferences. Modern societies are to an increasing extent pluralist, and it is clear that not all individuals conform to and prioritize the same values and goals. Even requirements for "rationality" in the abstract result in various different realizations and goals. Procedural accounts of autonomy, in turn, allow variation in the authentic preferences but face challenges when viewed in toxic situations such as oppressive political regimes in which oppressive values are internalized. Partly due to these reasons, the philosophical focus has turned to relational aspects of agency and autonomy, bringing social and interpersonal dynamics more to the fore.

The context in which we assess agents and their actions is always embedded in social and political practices that contribute to the normative framework for that assessment. Hutchinson et al. (2018) stress that the authenticity condition for autonomy must take into account the social constitution of practical identity of an

individual, and Vargas (2018) suggests that responsible agency is partly constituted by social feedback continuously shaping the acquisition and maintenance of our values and dispositions. These insights do not as such facilitate the identification of authentic values from other kinds of values, but remind us to look beyond the individual in the assessment of individual agency. People who have less-than-ideal capacities of self-regulation and control, e.g. people with addictive tendencies, may well hold values that suit their circumstances and the environment they live in. Their values are shaped by the circumstances and resources they have at their disposal. It is far from evident that informing them about the abnormalities in their neural mechanisms would result in different kinds of choices and actions.

More concretely, internal self-control processes are not determinative of self-governance even at the level of individual action. Experimental studies have shown that strong impulse control is not necessarily predictive of effective self-governance, as skills and strategies in avoiding temptation and prospectively adjusting the relative costs of tempting courses of action play a larger role in real-life self-governance than the disputed internal "mental muscle" of self-control (Levy 2017; see Duckworth et al. 2018). Many such strategies involve other people. Public announcement of long-term goals (and sometimes the metaphorical burning of bridges) are well-known examples of socially constituted strategies of self-governance. In the case of substance addictions, social support and the characteristics of social networks have been shown to have a big impact on recovery (e.g. Stevens et al. 2015). Having friends who tell you when to stop drinking is a more effective strategy for preserving self-governance than having a prescription for antabus – or building up your mental muscle by subjecting yourself to temptations. Similarly, being embedded in a social network of substance users, in which the relative social costs of substance use and abstinence work against the maintenance of other long-term goals, has a stifling effect on self-governance, regardless of changes in brain chemistry.

A challenge for neuroscientific evidence is that, at least in these kinds of non-pathological contexts, it cannot provide clear-cut thresholds for when features of the sub-personal are in danger of causing problems for autonomy. In "safe" environments, even less than ideal cognitive capacities and mechanisms can be sufficient for effective self-governance and, as was discussed above, some environments (such as Australian casinos equipped with modern neuroscientifically informed slot machines) will corrupt even well functioning brains. In order for neuroeconomics to fulfil its early promise of providing insight into the design of interventions and institutions for the betterment of our welfare, it has to acknowledge the irreducibly social constitution of autonomy – to become truly social neuroeconomics.

Notes

1 In the same vein, Manuel Vargas emphasizes that the capacities constituting morally responsible agency (autonomy) may be multiply realized across and even within agents, and that what our folk-theorizing about human agency takes as a unitary and general capacity of reason-responsiveness "is really a cluster of more specific, ecologically limited capacities indexed to particular circumstances" (Vargas 2013, 205).

2 Neuroscience has not only been a factor in identifying such autonomy-impairing situations, but it has been harnessed in creating them. The latest generations of slot machines have been specifically designed so as to utilize our brains weaknesses and compromise autonomy-relevant functionalities (cf. Schüll 2014).

References

Ainslie, G. 2000. A Research-based Theory of Addictive Motivation. *Law and Philosophy*, 19(1), 77–115.

Appelbaum. P.S. 2007. Assessment of Patient's Competence to Consent to Treatment. *New England Journal of Medicine*, 357(18), 1834–1840.

Beauchamp, T.L., and Childress, J.F. 2001. *Principles of Biomedical Ethics* (5th ed.). Oxford: Oxford University Press.

Camerer, C., Loewenstein, G., and Prelec, D. 2005. Neuroeconomics: How Neuroscience Can Inform Economics. *Journal of Economic Literature*, 43(1), 9–64.

Caplan, A.L. 2006. Ethical Issues Surrounding Forced, Mandated, or Coerced Treatment *Journal of Substance Abuse Treatment*, 31, 117–120.

Carter, A., and Hall, W. 2011. *Addiction Neuroethics: The Promises and Perils of Neuroscience Research on Addiction*. Cambridge: Cambridge University Press.

Casey, B.J., Somerville, L.H., Gotlib, I.H., Ayduk, O., Franklin, N.T., Askren, M.K., Jonides, J., Berman, M.G., Wilson, N.L., Teslovich, T., and Glover, G. 2011. Behavioral and Neural Correlates of Delay of Gratification 40 Years Later. *Proceedings of the National Academy of Sciences*, 108(36), 14998–15003.

Charland, L.C. 2002. Cynthia's Dilemma: Consenting to Heroin Prescription. *American Journal of Bioethics*, 2(2), 37–47.

Charland, L.C. 2015. Decision-Making Capacity. In *The Stanford Encyclopedia of Philosophy* (Fall 2015 ed.), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/fall2015/entries/decision-capacity/.

Demos, K.E., Heatherton, T.F., and Kelley, W.M. 2012. Individual Differences in Nucleus Accumbens Activity to Food and Sexual Images Predict Weight Gain and Sexual Behavior. *Journal of Neuroscience*, 32(16), 5549–5552.

Doris, J.M. 2015. *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford: Oxford University Press.

Duckworth, A.L., Milkman, K.L., and Laibson, D. 2018. Beyond Willpower: Strategies for Reducing Failures of Self-control. *Psychological Science in the Public Interest*, 19(3), 102–129.

Fumagalli, R. 2013. The Futile Search for True Utility. *Economics & Philosophy*, 29(3), 325–347.

Grouzet, F.M., Sokol, B.W., and Müller, U. 2013. Self-regulation and Autonomy: An Introduction. In F.M. Grouzet, B.W. Sokol, and U. Müller (eds.), *Self-regulation and Autonomy: Social and Developmental Dimensions of Human Conduct*. New York: Cambridge University Press, 1–16.

Guala, F., and Mittone, L. 2015. A Political Justification of Nudging. *Review of Philosophy and Psychology*, 6(3), 385–395.

Hausman, D. 2011. *Preference, Value, Choice, and Welfare*. Cambridge and New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139058537

Heyman, G.M. 2009. *Addiction: A Disorder of Choice*. Cambridge, MA: Harvard University Press.

Hommel, B., and Wiers, R.W. 2017. Towards a Unitary Approach to Human Action Control. *Trends in Cognitive Sciences*, 21(12), 940–949.

Hooper, S.M., and Chiong, W. 2017. Decision Making Capacity and Frontal Lobe Dysfunction. In B. Miller and J. Cummings (eds.), *The Human Frontal Lobes: Functions and Disorders* (3rd ed.). New York: The Guilford Press, 184–199.

Hutchinson, K., Mackenzie, C., and Oshana, M. 2018. Introduction. In K. Hutchinson, C. Mackenzie, and M. Oshana (eds.), *Social Dimensions of Moral Responsibility*. Oxford: Oxford University Press, 1–37.

Infante, G., Lecouteux, G., and Sugden, R. 2016. Preference Purification and the Inner Rational Agent: A Critique of the Conventional Wisdom of Behavioural Welfare Economics. *Journal of Economic Methodology*, 23(1), 1–25. https://doi.org/10.1080/1350178X.2015.1070527

Kelley, W., Wagner, D., and Heatherton, T. 2014. In Search of a Human Self-Regulation System. *Annual Review of Neuroscience 2015*, 38(1), 389–411.

Koob, G.F., and Volkow, N.D. 2016. Neurobiology of Addiction: A Neurocircuitry Analysis. *The Lancet Psychiatry*, 3(8), 760–773. doi:10.1016/S2215-0366(16)00104-8

Kotabe, H.P., Hofmann, W. 2015. On Integrating the Components of Self-Control. *Perspectives on Psychological Science*, 10(5), 618–638.

Krajbich, I., Camerer, C., and Rangel, A. 2017. Exploring the Scope of Neurometrically Informed Mechanism Design. *Games and Economic Behavior*, 101, 49–62.

Kwako, L.E., Momenan, R., Litten, R.Z., Koob, G.F., and Goldman, D. 2016. Addictions Neuroclinical Assessment: A Neuroscience-Based Framework for Addictive Disorders. *Biological Psychiatry*, 80(3), 179–189. doi:10.1016/j.biopsych.2015.10.024

Levy, N. 2017. Of Marshmallows and Moderation. In S.-A. Walter and C.B. Miller (eds.), *Moral Psychology: Virtue and Character*. Cambridge, MA and London: MIT Press, 197–214. www.jstor.org/stable/j.ctt1n2tvzm.17.

Lopez, R.B., Hofmann, W., Wagner, D.D., Kelley, W.M., and Heatherton, T.F. 2014. Neural Predictors of Giving in to Temptation in Daily Life. *Psychological Science*, 25(7), 1337–1344.

Lopez, R.B., Milyavskaya, M., Hofmann, W., and Heatherton, T.F. 2016. Motivational and Neural Correlates of Self-control of Eating: A Combined Neuroimaging and Experience Sampling Study in Dieting Female College Students. *Appetite*, 103, 192–199.

Mackenzie, C. 2014. Three Dimensions of Autonomy: A Relational Analysis. In Veltman and Piper (eds.), *Autonomy, Oppression and Gender*. Oxford: Oxford University Press, 15–41.

Mackenzie, C., and Stoljar, N. 2000. Introduction: Autonomy Reconfigured. In Mackenzie and Stoljar (eds.), *Relational Autonomy. Feminist Perspectives on Autonomy, Agency, and the Social Self*. Oxford: Oxford University Press, 3–31.

Morán-Sánchez, I., Luna, A., Sánchez-Muñoz, M., Aguilera-Alcaraz, B., and Pérez-Cárceles, M.D. 2016. Decision-making Capacity for Research Participation among Addicted People: A Cross-sectional Study. *BMC Medical Ethics*, 17(1), 3.

Moye, J., Marson, D.C., and Edelstein, B. 2013. Assessment of Capacity in an Aging Society. *American Psychologist*, 68(3), 158.

Padoa-Schioppa, C. 2011. Neurobiology of Economic Choice: A Good-based Model. *Annual Review of Neuroscience*, 34, 333–359.

Park, J.W., and Zak, P.J. 2007. Neuroeconomics Studies. *Analyse & Kritik*, 29(1), 47–59.

Peterson, A. 2018. Should Neuroscience Inform Judgements of Decision-Making Capacity? *Neuroethics*, 1–19.

Reijula, S., and Hertwig, R. 2020. Self-nudging and the Citizen Choice Architect. *Behavioural Public Policy*, 1–31. doi:10.1017/bpp.2020.5

Schüll, N.D. 2014. *Addiction by Design: Machine Gambling in Las Vegas*. Princeton, NJ: Princeton University Press.

Stevens, Ed, Jason, L.A., Ram, D., and Light, J. 2015. Investigating Social Support and Network Relationships in Substance Use Disorder Recovery. *Substance Abuse*, 36(4), 396–399. doi:10.1080/08897077.2014.965870

Thaler, R.H., and Sunstein, C. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.

Vargas, M. 2013. *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.

Vargas, M. 2018. The Social Constitution of Agency and Responsibility – Oppression, Politics, and Moral Ecology. In K. Hutchinson, C. Mackenzie, and M. Oshana (eds.), *Social Dimensions of Moral Responsibility*. Oxford: Oxford University Press, 110–136.

Wagner, D.D., Boswell, R.G., Kelley, W.M., and Heatherton, T.F. 2012. Inducing Negative Affect Increases the Reward Value of Appetizing Foods in Dieters. *Journal of Cognitive Neuroscience*, 24(7), 1625–1633.

Zawidzki, T.W. 2013. *Mindshaping: A New Framework for Understanding Human Social Cognition*. Cambridge, MA: MIT Press.