Title: Updating our selves: synthesizing philosophical and neurobiological perspectives on incorporating new information into our worldview

Fay Niker[1], Peter B. Reiner[2], and Gidon Felsen[2,3,4 *]

Affiliations and addresses:

1. Department of Politics and International Studies, University of Warwick
Coventry, CV4 7AL, UK

2. National Core for Neuroethics, Department of Psychiatry, University of British Columbia
2255 Wesbrook Mall
Vancouver, BC V6T 2B5, Canada

3. Department of Physiology and Biophysics, University of Colorado School of Medicine
12800 E. 19th Ave., Mail Stop 8307
Aurora, CO 80045, USA

4. Center for Bioethics and Humanities, University of Colorado School of Medicine
13080 E. 19th Ave., Mail Stop B137
Aurora, CO 80045, USA

*** To whom correspondence should be addressed**
Phone: (303) 724-4532
Fax: (303) 724-4501
E-mail: gidon.felsen@ucdenver.edu

Acknowledgments: This work was supported by the Greenwall Foundation's Faculty Scholars Program in Bioethics (G. F.) and by a Warwick Transatlantic Fellowship from the University of Warwick's Humanities Research Centre (F. N.).

Abstract

Given the ubiquity and centrality of social and relational influences to the human experience, our conception of self-governance must adequately account for these external influences. The inclusion of socio-historical, externalist (i.e., "relational") considerations into more traditional internalist (i.e., "individualist") accounts of autonomy has been an important feature of the debate over personal autonomy in recent years. But the relevant socio-temporal dynamics of autonomy are not only historical in nature. There are also important, and under-examined, future-oriented questions about how we retain autonomy while incorporating new values into the existing set that guides our interaction with the world. In this paper, we examine these questions from two complementary perspectives: philosophy and neuroscience. After contextualizing the philosophical debate, we show the importance to theories of autonomous agency of the capacity to appropriately adapt our values and beliefs, in light of relevant experiences and evidence, to changing circumstances. We present a plausible philosophical account of this process, which we claim is generally applicable to theories about the nature of autonomy, both internalist and externalist alike. We then evaluate this account by providing a model for how the incorporation of values might occur in the brain; one that is inspired by recent theoretical and empirical advances in our understanding of the neural processes by which our beliefs are updated by new information. Finally, we synthesize these two perspectives and discuss how the neurobiology might inform the philosophical discussion.

Keywords

Autonomy, Pro-attitudes, Neuroscience, Decision making

Introduction

A key feature of autonomy – the capacity for self-governance – is that our decisions and actions are governed by a set of higher-order desires, values, and beliefs [e.g., 1, 2], which constitute our "pro-attitudes" [3]. The primary focus of this paper is to explore the question of how we incorporate new pro-attitudes into our worldview. This question is of fundamental importance to theories of autonomy for at least two reasons: (1) Our set of pro-attitudes defines, to some degree the "self," and, hence, determines a necessary part of the process of self-governance – or, "self-creation" [4], "self-authorship" [5], or "self-constitution" [6]; and (2) Given that incorporating new pro-attitudes is an adaptive feature of daily life, theories about the nature of autonomy that are satisfied with the mere ability of the agent to set her own course in life, without recognizing that certain experiences or evidence require that she reconsider and perhaps adjust her plan, may be incomplete [7]. Despite its importance to the nature of autonomy, the question of how new pro-attitudes are incorporated into one's worldview, as well as how existing pro-attitudes are revised or updated, has been little addressed in the literature.

We suggest that neuroscience may be able to provide a useful perspective on this question. While philosophy of mind is the discipline historically associated with the study of autonomy and its components, the biological substrate for these phenomena is the brain. And while we currently lack a precise neuroscientific account of how pro-attitudes are incorporated – indeed, such a description may be many years away – recent advances in our understanding of related brain functions allow us to begin to address the neural basis for this capacity. We therefore suggest that the time is ripe to explore how neuroscience can inform the philosophical conception of pro-attitude modification and integration, in the vein of other recent papers

attempting to ground philosophical ideas in neuroscientific evidence [8, 9], perhaps by constraining the debate to those proposals consistent with brain function [10–12].

We focus our discussion of evidence from neuroscience on a topic that, we suggest, is particularly relevant to the incorporation of pro-attitudes: how – and when – new (sensory) information is integrated into our existing beliefs about the state of the world. In the sections below, we first survey the relevant philosophical literature, in order to contextualize and justify the largely-unrecognized importance of an "experience-responsiveness" condition for any complete theory of the nature of autonomy. This survey is not intended to be normative; our goal is to describe the role of experience-responsiveness in autonomous agency and the context in which it functions. We then describe theories and data relevant to information integration, and how these might relate to pro-attitude incorporation. We conclude by discussing the implications of the neuroscientific data for descriptive theories of pro-attitude incorporation and consider how we might better align these theories with the data.

Philosophical perspective: The role of experience-responsiveness in self-governance

The philosophical discussion of autonomy posits that each person has some set of pro-attitudes – referred to variously as their "motivational set" [7], "collection of values" [3], "conception of the good" [13], "psychological core" [14], or "worldview" – and that the possession of this set underlies his or her autonomy in several ways, two of which have been discussed heavily in the literature. First, there are hierarchical internalist considerations concerning whether there is identification, at time *t*, between one's pro-attitudes and those first-order desires that are directed towards action [1, 2]. Second, there are socio-historical externalist considerations concerning how one's pro-attitudes initially came to be held [5, 15, 16]. However, neither of these sets of

considerations provides us with much guidance for addressing the issue that interests us here, namely, how we incorporate new pro-attitudes and, more generally, how we change or revise our set of pro-attitudes over time, in accordance with our self-development and continued dynamic interaction with the world around us – a process we refer to here as "pro-attitude incorporation" (which is taken to be inclusive of pro-attitude revision).

To gain some ground on this question, we must move beyond the recent discussion – with its focus on the challenge to internalism's time-slice analysis, on grounds of historic social dynamics – by exploring and capturing in our theories of the nature of autonomy the importance of certain future-oriented socio-temporal dynamics. Taking seriously the fact that we exercise our autonomy over time requires us to supplement our theories with a recognition that "we have a history and a future, that we develop our identities and emancipate ourselves from others over time, that we sometimes change our minds and take different directions, that we find ourselves in changing relationships and social environments, etc." [17]. One way to think about this issue in a tangible way is to consider what impact this recognition might have for a widely acknowledged stable component of autonomous agency; and critical reflection is an obvious contender.

Regardless of which specific account of autonomy one favors, critical reflection is always taken to be a central feature, since it is this that allows a person to shape the attitudes that guide her actions. Hierarchical internalist theories, for instance, understand autonomy as "a second-order capacity to reflect critically upon one's first-order preferences and desires, and the ability to either identify with these or to change them in light of higher-order preferences and values" [2]. Historically externalist accounts supplement this account with the claim that autonomy requires that the agent authentically possesses those pro-attitudes, meaning that she has come to possess them in a way that does not bypass her capacities for critical reflection [3]. However,

these standard claims about the critical reflection required by autonomy do not look to be sufficient when thinking about the process of pro-attitude incorporation. To see this clearly, consider Blöser et al.'s [18] case of "older Pat".

Pat is a 70 year-old man and a loving father and grandfather. He nevertheless finds it increasingly tough to accept that his children and grandchildren live their lives in ways different from those that he himself pursued at their age. For example, Pat struggles with the fact that his son has had his children outside of wedlock, since Pat is convinced that children can only flourish within a stable family, which he takes to be one in which the children's parents are married. At this stage, Blöser et al. stipulate that this case satisfies "all sensible internalist requirements for autonomy": in particular, Pat is able to critically reflect on each of his pro-attitudes in light of his existing set of pro-attitudes and to shape his pro-attitudes according to the outcome of this reflection. They also stress that Pat meets the historical externalist requirements in so far as he is not a victim of manipulation; rather, he holds the same pro-attitudes that he (that is, "younger Pat") authentically acquired half a century ago. But then imagine that Pat, for whatever reason (capacitarian, dispositional, or otherwise), fails to question his pro-attitudes in light of new experiences. Picking up the earlier example, the fact his son's family provides a safe and supportive environment in which he can see that his grandchildren are flourishing, despite their parents being unmarried, fails to make Pat reconsider whether marriage really is a basic requirement of good parenthood [18].

This case suggests that a distinction needs to be drawn between two kinds of critical reflection, and that *both* are required by a fully autonomous agent: the first is the familiar (coherentist) ability to critically reflect on a pro-attitude *in light of our other pro-attitudes*; and the second is the ability to critically reflect on a pro-attitude *in light of new experiences or*

evidence.[1] In the case of older Pat, which is constructed so that there is no inconsistency between his pro-attitudes, it is clear that it is his failure to recognize and respond appropriately to his new experience of child-rearing and, in turn, the evidence that this provides against his pro-attitude (i.e., his value-based child-rearing belief), that undermines his autonomy with respect to this pro-attitude. In line with this distinction, in order to recover autonomy with respect to this "encrusted" pro-attitude, Pat would need to exercise "experience-responsive" critical reflection, that is, to consider this new experience as a touchstone for his pro-attitude [18].

This is clearly and importantly distinct from the first kind of critical reflection, since it enables people to "appropriately update the inputs" to self-evaluations, as opposed to interpreting their experiences in light of the pro-attitudes that they already happen to hold, which is likely to inevitably confirm their evaluative outlook [7]. In this way, relevant new experiences and other, non-sensory evidence – specifically, those that convey unexpected information, given one's set of pro-attitudes – have "the power to call into question [pro-attitudes] in a way that so far has not been accounted for by either internalist or history-sensitive accounts of autonomy" [18]. Still, both of these accounts can be supplemented with the experience-responsiveness condition: hierarchical internalist theories could, on the one hand, maintain that autonomy requires critical reflection on a pro-attitude, followed by either identifying with it or changing it in light of (i) higher-order values and beliefs *and* (ii) any relevant and unanticipated experiences; and historically externalist theories could, on the other hand, hold that an agent is autonomous with respect to a pro-attitude only if she (i) authentically possesses it (i.e., had adequate control over

[1] There is some discussion in the literature about whether experience or evidence more generally is the correct object of our responsive attentions. For our purposes, we use the term "experience-responsiveness" to capture both "the acquisition of information through direct perception", i.e., experiential information [18] and the more expansive idea of any evidence – experiential or not – that offers a reason to review the relevant part of one's worldview [7].

its development) *and* (ii) remains able to reconsider, and adjust if necessary, the pro-attitude in light of any relevant and unanticipated experiences [7].

While it has not been addressed in the literature in any real detail until fairly recently, reference akin to the experience-responsiveness condition is not entirely without precedent in the philosophical discussion of autonomy. For example, Arneson writes that: "To live an autonomous life an agent must decide on a plan of life through critical reflection and in the process of carrying it out, remain disposed to subject the plan to critical review if […] unanticipated evidence indicates the need for such review" [19]. Similarly, Noggle contends that, in addition to "a skeleton of core attitudes, the fully formed [and autonomous] self has the ability to adjust and revise its own attitudes" [14]. In addition, others have discussed the role of reasons-responsiveness in autonomous agency [e.g., 20], and it seems plausible to think that experience-responsiveness might be understood as a particular way of responding to reasons, specifically, responding to reasons-to-review and/or reasons-to-revise a pro-attitude that one currently holds.

For our purposes here, it is not necessary to defend any particular model of pro-attitude incorporation; rather, we outline one plausible philosophical account of this process. Based largely on Blöser et al.'s analysis, experience-responsive critical reflection can be viewed as a "complex mental activity" involving four elemental processes [18]. First, a person *P* recognizes a new experience as being new, in so far as it is not to be expected or anticipated relative to *P*'s pro-attitudes. Second, the occurrence of a new experience pertinent to one of *P*'s pro-attitudes, *A*, is considered as relevant input for reflection upon *A*. This is the case for some pro-attitude, *A*, if *A* implies expectations about future experiences that are either confirmed or disconfirmed by the new experience in question or if that new experience indicates the need to make other sorts of self-adjustments, such as reducing the strength or relative importance of *A*. Third, *P* reconsiders

the adequacy of *A* in light of the new experience. Fourth, *P* adjusts her pro-attitudes accordingly, which could include her reconfirming, abandoning, or reweighing the relative importance of *A*.

This account offers a promising means of thinking philosophically about the process of pro-attitude incorporation. It captures the importance of people's capacity to bring about "an appropriate relation between their values and the world they live in" over the course of their lives, by showing that autonomy requires that part of this ongoing dynamic relation between a person and her (changing) environment is a disposition and an ability to respond adequately to unanticipated new experiences, which now lie in the future [18]. In the same way that the historical externalist condition responds to the idea that we do not want our pro-attitudes to be determined by external circumstances, the experience-responsive condition responds to the idea that we do not want them to be unaffected or incapable of being affected by relevant changes in the dynamic world around us. These conditions are consistent with the idea that a complete theory of autonomy must "avoid the consequence that persons are 'caught up in themselves' as well as that they are 'caught up in society'", as Baumann puts it [17]. We suggest a further qualification: that normative theories of autonomy should meet a threshold of minimum neuroscientific realism by being compatible with what is known about how the brain functions [10][2]. We explore this issue more fully in the next section.

Neuroscientific perspective: Updating beliefs with new information

The brain sciences offer a complementary perspective on pro-attitude incorporation. In particular, a central question in the field is how the brain integrates outside information into our existing beliefs about the world. We suggest that understanding how this occurs can provide insight into pro-attitude incorporation. Some might object that such a relatively low-level process

[2] In other words, our *oughts* should be compatible with what *can* be achieved in the real world.

is too far removed from our main question to be relevant. Even if we had a complete understanding of its neural basis, how much would that truly tell us about a higher-level phenomenon like the incorporation of pro-attitudes? It is true that we will not be providing a precise neurobiological account of pro-attitude incorporation; however, due to the constraints of evolution on brain structure and function, the neural computations underlying simple and more complex processes can be expected to be conserved [21]. For example, at the fundamental level of neural activity, the mechanisms responsible for updating the neural representation of a friend's face upon seeing her for the first time in many years, and the mechanisms responsible for updating a set of pro-attitudes when confronted with conflicting information, are probably shared. Thus, studies examining relatively simple brain processes can illuminate the neural basis of pro-attitude incorporation.

Before examining the incorporation of new pro-attitudes, it is necessary to describe how such pro-attitudes might be represented in the brain. From the neurobiological point of view, it seems likely that pro-attitudes are represented by recurrent patterns of activity in prefrontal cortical circuits, which are thought to bear primary responsibility for our uniquely human capacities. According to hierarchical models of executive control, these prefrontal circuits can guide decisions and actions by biasing activity in downstream, lower-level, brain regions [22]. As described below, these patterns of activity – and therefore the pro-attitudes that they represent – can also be modified by new, relevant information.

A useful framework for thinking about how new representations of pro-attitudes are incorporated into an existing set is provided by Bayesian inference [23, 24]. The overall idea is that decisions are represented probabilistically, and result from combining two sources of information: internally-generated "priors" – which provide a starting-point for approaching a

particular decision – and new evidence. Such evidence often comes from the outside world in the form of sensory input, but could also result from reflecting upon an issue in a new way. According to this framework, an existing set of beliefs can be thought of as the priors, with which new information interacts in two ways: (i) The two sources (priors and the new information) are integrated in order to determine the relative value of the available options for the particular decision at hand; and (ii) the new information updates (or does not update) the priors themselves, in preparation for the next decision. Bayesian inference is particularly useful when information is ambiguous, as is often the case in the real world.

We will use two examples to illustrate the issue. The first is one in which unreliable sensory information (e.g., a blurry photograph) leads us to believe that, for instance, the moon is made of cheese. We might decide, on the basis of this prior, to bring crackers on our trip to the moon. Upon landing and finding out that the moon is actually made of rock, we would (i) choose to leave our crackers behind when disembarking, and (ii) update our belief in the moon's composition, such that on our next trip we will not bring crackers in the first place, thereby saving fuel. The second example is a variant of Blöser et al.'s case of "older Pat", but instead considers the case of Nat. Nat, like Pat, might value married-parent families based on a belief formed by what he has heard from friends and the popular press. However, unlike Pat, upon having new and unexpected experiences (i.e., noticing that his grandchildren are indeed flourishing in an unmarried-parent family), discovering relevant sociological data on this issue, and/or simply reflecting on his beliefs[3] he might (i) decide to accept his son's lifestyle, and (ii) modify his values and beliefs, in light of this new information.

[3] While reflection may be sufficient for updating one's beliefs, the likelihood of this occurring may depend on how one's pro-attitudes initially developed, which our analysis does not attempt to address.

Numerous carefully-controlled studies have shown that we implicitly utilize a Bayesian framework by integrating new information with priors in order to make perceptual and motor decisions [25–27]. Indeed, perceptual biases such as optical illusions can be explained in terms of Bayesian inference [28, 29]: our expectations, represented by priors, influence our perception of the outside world. While these data support the idea that the brain employs Bayesian inference, and there exists an emerging literature on the neural basis of decision-making based on sensory evidence and priors [30–32], less is known about how priors are updated based on new information – the issue most relevant to pro-attitude incorporation.

What determines whether or not new information results in incorporating new pro-attitudes? As noted above, the initial (first and second) steps proposed by the philosophical model of pro-attitude incorporation outlined above involve an experience that is not consistent with an existing pro-attitude being identified as novel and relevant to that pro-attitude. Similarly, with respect to beliefs about the world, it has been proposed that an initial step in determining whether priors are updated is their reliability, relative to newly available information, for making "good" decisions, which is often operationalized as those having the best outcome [33]. If prior beliefs (or pro-attitudes) are sufficiently reliable, then new information is not relevant and there is no particular advantage in updating them. However, if the prior beliefs (or pro-attitudes) are not sufficiently reliable, then the new information is relevant. In practice, the reliability of priors and particularly pro-attitudes is rarely all-or-none; the most important determinant of the benefit of updating is whether doing so would lead to better outcomes.

The remaining (third and fourth) steps proposed by the model of experience-responsiveness are that the set of pro-attitudes in question is reflected upon in light of the new experience, and is (or is not) adapted accordingly. In the brain, these processes are likely

combined: new information is continually incorporated until the priors, on their own, become sufficiently reliable for making good decisions. It is likely that the neural systems responsible for maximizing reward play a role in determining when the priors need further updating and when they become sufficiently reliable [34–36]. In this way, we can appropriately adapt both lower-level beliefs and higher-level pro-attitudes to changing circumstances.

Returning to the examples above, suppose that we initially held the belief that the moon was made of cheese because, on several previous trips, we had found this to in fact be the case. Upon nearing the surface of the moon and obtaining evidence – e.g., based on the carom of a projectile – that the moon is made of rock, we may decide to leave our crackers behind for this trip, but in the face of ambiguous evidence, we may not yet update our belief about the composition of the moon. It could be the case that our projectile trajectory data are noisy, or that our estimate of how the projectile would carom off of cheese is unreliable. If further experiments continue to yield information favoring the idea that the moon is made of rock, and it turns out to be a good decision to leave the crackers behind (because we saved fuel), we would gradually update our beliefs in accordance with the new evidence. Similarly, in light of Nat's discovery of new personal-experiential or sociological evidence, or simply upon further reflection, his relevant pro-attitude is shown to be unreliable, thereby triggering him to begin the process of reevaluating it. Note, also, that in these hypothetical scenarios, it is possible that the moon once was actually made of cheese, and that married parents once did in fact provide a more stable environment for children.

It is worth noting that foraging theory, the neural basis of which has begun to be examined, provides a natural framework related to Bayesian inference for thinking about how to balance existing beliefs with new information [37, 38]. This framework attempts to address the

question of when the current behavioral strategy should be utilized to the best effect possible, i.e. "exploitation", and when a better strategy should be identified, i.e. "exploration". The latter requires an investment – for example, in time, energy, or immediate reward – in the hope of identifying a more lucrative strategy in the long run. There is an inherent tradeoff between exploration and exploitation: overdependence on exploitation – much like overreliance on priors – can result in missed opportunities to optimize the strategy, particularly in a relatively dynamic environment. With respect to philosophical commitments, this can be thought of as being akin to rigidly maintaining a set of "encrusted" pro-attitudes, and resisting updating them, even when confronted with countervailing evidence. On the other hand, overdependence on exploration – much like overreliance on new information – can reduce overall reward, particularly in a relatively static environment. This may be akin to being too quick to abandon one's pro-attitudes in the face of new information.

Traditional internalist conceptions of autonomy might consider an exploitative strategy to be compatible with, and perhaps even to promote, autonomy. However, an account of autonomy that is sensitive to the importance of experience-responsiveness would contend that such a stance inhibits autonomy by limiting a person's capacity to critically reflect on her pro-attitudes in light of new experiences[4]. Conversely, a maximally exploratory strategy is over-reliant on external information over pro-attitudes and thus also violates a core precept of autonomy. So maximizing autonomy looks to require a familiar tradeoff between exploration and exploitation.

Our emerging understanding of the neural basis of this tradeoff, and of Bayesian inference, for simple decisions may, in a general sense, provide insight into how and when a higher-level process like pro-attitude incorporation may occur in the brain. While the

[4] To be clear, it is the capacity to critically reflect that is important, not whether new pro-attitudes are or are not ultimately incorporated.

neuroscience underlying this process has not been studied directly, one line of recent research has led to the intriguing proposal that a particular region of the prefrontal cortex, the posterior cingulate cortex (CGp), is a key node in the network responsible for recognizing changes in the environment and adapting behavioral strategies accordingly [39]. Some of the evidence supporting this idea comes from examining neural activity in non-human primates performing tasks requiring choice between a "safe" option, associated with a fixed reward, and a "risky" option, associated with a variable reward. Briefly, the activity of CGp neurons reflects the magnitude and variability of reward [40], and is predictive of switches in behavioral strategy (e.g., from the safe to the risky option) [41]. These and other data support the idea that lower levels of CGp activity promote retaining the current set of beliefs – akin to exploitation, in foraging theory – and higher levels CGp activity supports allowing new experiences or evidence to update the beliefs – akin to exploration [39]. While these experiments were designed to examine the relationship between CGp activity and behavior across trials performed by the same subjects, it is possible that the relationship would hold across subjects – i.e., that individuals more prone to resist updating their beliefs (e.g., Pat) are more likely to exhibit lower levels of CGp activity. Under this theory, elevated CGp output would modulate activity in the circuits representing prior beliefs such that the likelihood that these representations will be altered is increased, although how this occurs is not yet known.

It seems reasonable to suggest that CGp – likely in concert with a network of interconnected brain regions – plays a similar role in incorporating pro-attitudes as it does in representing prior beliefs. The critical point is that a plausible mechanism exists for incorporating new patterns of neural activity in response to novel sensory information. While this mechanism subserves lower-level processes that often occurs below the level of conscious

awareness, this (or some similar) mechanism could be conserved for higher-level processes like pro-attitude incorporation. While we are currently limited to inferring how the brain mediates higher-level processes from our understanding of the neural basis of analogous lower-level processes, it is possible that advances in neuroscientific methods may allow us to directly examine higher-level processing in the future.

Synthesis of perspectives from philosophy and neuroscience

In the preceding sections we have reviewed the philosophical perspective on the incorporation of new pro-attitudes and the reasons for thinking that this provides an additional necessary condition for autonomous agency. This highlighted that self-governance requires governance by a self that is able to "update" itself in light of the world around it by responding to relevant experiences (as opposed to governance by an inflexible former self). We then considered how this process might occur in the brain, with an eye towards further informing this philosophical perspective. So what does neuroscience add to the philosophical discussion?

One interesting discrepancy between the perspectives from the two disciplines concerns how the process of pro-attitude incorporation is thought to occur. It is commonly assumed in the philosophical literature that this process requires top-down, rational reflection. For instance, even though Weimer suggests that the experience-responsiveness condition "allows for an agent to be on 'autopilot' with respect to her motivational states for long periods of time, merely monitoring in a passive way for unanticipated evidence", he holds that pro-attitude incorporation requires that she is able to "retake active control of the relevant part(s) of her life by *rationally reconsidering*" the appropriate pro-attitude(s) [7] (emphasis added). In addition, Noggle maintains that pro-attitude incorporation requires "the psychological mechanisms necessary to

allow [the agent] to *reflect upon and revise* those beliefs and desires" – a process he calls "reflective self-adjustment" [14] (emphasis added).

However, these ideas are not invoked by the neuroscientific account. Instead, it is entirely consistent with the extant neurobiology that we incorporate new pro-attitudes below the level of conscious awareness, in much the same way that we make many everyday decisions and that perception "automatically" – i.e., without top-down reflection – arises from the integration of sensory evidence with prior beliefs. Those pro-attitudes whose incorporation leads to better decisions – defined, as above, as those associated with better outcomes – will be retained, while those that do not lead to better decisions will be rejected. While the evidence from neuroscience does not preclude the possibility of top-down reflection in incorporating new pro-attitudes – indeed, there is support for this hierarchical feature of autonomy, in particular via executive control theory [10, 22] – the evidence is also consistent with the idea that non-conscious processing of emotions provides a useful heuristic for efficient decision making [42, 43]. While we do not suggest that these descriptive findings be brought to bear on the question of how we should value autonomy, the explanation they provide for how autonomous decision making could occur is most consistent with philosophical accounts that do not depend on top-down rational reflection exclusively [e.g., 44, 45].

Another potential insight from neuroscience may address the question of what determines whether a new pro-attitude will be accepted or rejected. As described above, the more surprising new information is – i.e., the less predictable the information is based on the priors – the more rapidly the priors are updated. In a similar manner, the more conflict there is between our pro-attitudes and new information, the more rapidly we may be to evaluate whether we need to update our pro-attitudes. While there is undoubtedly a large set of factors that influence whether

new pro-attitudes are accepted or rejected – for example, the trustworthiness of the source of information driving the new pro-attitude – Bayesian inference and, specifically, conflict between new information and priors provides an overarching framework for thinking about why some pro-attitudes remain stable and some are updated.

Although we have focused in this paper on the neuroscience underlying Bayesian inference, since this is particularly relevant to the question of updating pro-attitudes based on new information, insights from neuroscience are not limited to this systems-level phenomenon. For example, at the cellular level, the phenomenon of memory reconsolidation may also provide insight into pro-attitude incorporation. It has been shown that our long-term memories are not as stable as we might imagine them to be: the very act of recalling a memory converts it from stable to labile [46]. After a period of time, and probably during sleep [47, 48], these memories pass into a long-term stable state once again – a process known as reconsolidation. Not only does recall transform memory from stable to labile, it also provides a window of opportunity to *change* memory [49, 50], and this has important implications for the mechanisms of neuronal integration [51].

Thus, whenever we encounter new information the brain attempts to "put it into context": we think about related events that have occurred in the past, and as we do so each of them de-consolidates. If they are both sufficiently salient and sufficiently related to the new information, then a functional linkage between them is created, and when both events pass into long-term memory – one as a *de novo* memory and the other as a reconsolidated one – they are interconnected such that the recall of either is, on average, more likely to cause the recall of the other. Such an integrated perspective on the world around us is tremendously useful for accurate prediction of future events [24], and may provide a framework for thinking about how pro-

attitudes are updated. Specifically, if representations of pro-attitudes become labile[5] when they are "brought to mind" as new evidence is considered, a similar process of reconsolidation may underlie the updating that the representations undergo. It is this updated form of the pro-attitude representation that will then be brought to bear on subsequent decisions.

In summary, this paper represents an attempt to integrate neuroscientific data and theories with the philosophical ideas underlying a critical, yet under-examined, feature of autonomy. We have endeavored to illuminate the degree to which our brains have the capacity to achieve the aims that philosophers of mind attribute to them. Specifically, we have focused on our emerging understanding of how the brain integrates new information into existing beliefs as a model for understanding how pro-attitude incorporation might occur in the brain. This is particularly important in light of both the requirement for minimal empirical realism and the general applicability of a supplementary experience-responsiveness condition to theories of autonomy. Future studies will advance our understanding of the neural basis of information integration, which can serve to further inform pro-attitude incorporation and address other questions relevant to a wide range of conceptions of autonomy. Continuing to synthesize the complementary neuroscientific and philosophical accounts of related brain processes will ultimately allow for a neurobiologically-grounded conception of personal autonomy.

References

1. Frankfurt, Harry G. 1971. Freedom of the Will and the Concept of a Person. *The Journal of Philosophy* 68: 5–20.

2. Dworkin, Gerald. 1988. *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press.

[5] This account does not attempt to explain resistance to incorporating new pro-attitudes (exemplified by the case of older Pat). It is indeed possible that representations of pro-attitudes do not undergo the same cycle of deconsolidation and reconsolidation described here.

3. Mele, Alfred R. 1995. *Autonomous Agents: From Self-Control to Autonomy*. Oxford University Press.

4. Mill, John Stuart. 2008 [1859]. *On Liberty and Other Essays*. OUP Oxford.

5. Raz, Joseph. 1986. *The Morality of Freedom*. Clarendon Press.

6. Korsgaard, Christine M. 2009. *Self-Constitution: Agency, Identity, and Integrity*. OUP Oxford.

7. Weimer, Steven. 2013. Evidence-responsiveness and autonomy. *Ethical Theory and Moral Practice* 16: 621–642.

8. Shadlen, Michael N., and Adina L. Roskies. 2012. The neurobiology of decision-making and responsibility: Reconciling mechanism and mindedness. *Frontiers in Decision Neuroscience* 6: 56.

9. Roskies, Adina L. 2010. How does neuroscience affect our conception of volition? *Annual Review of Neuroscience* 33: 109–130.

10. Felsen, Gidon, and Peter B. Reiner. 2011. How the neuroscience of decision making informs our conception of autonomy. *AJOB Neuroscience* 2: 3–14.

11. Felsen, Gidon, and Peter B. Reiner. 2015. What can neuroscience contribute to the debate over nudging? *Review of Philosophy and Psychology* 6: 469–479.

12. Greene, Joshua D. 2014. Beyond point-and-shoot morality: why cognitive (neuro)science matters for ethics. *Ethics* 124: 695–726.

13. Rawls, John. 1971. *A Theory of Justice*. Harvard University Press.

14. Noggle, Robert. Autonomy and the paradox of self-creation: Infinite regresses, finite selves, and the limits of authenticity. In *Personal Autonomy: New Essays on Personal Autonomy and its Role in Contemporary Moral Philosophy*, ed. James Stacey Taylor, 87–108. Cambridge: Cambridge University Press.

15. Oshana, Marina. 2006. *Personal Autonomy in Society*. Ashgate Publishing, Ltd.

16. Christman, John. 2009. *The Politics of Persons: Individual Autonomy and Socio-historical Selves*. Cambridge: Cambridge University Press.

17. Baumann, Holger. 2008. Reconsidering relational autonomy. personal autonomy for socially embedded and temporally extended selves. *Analyse & Kritik* 30: 445–468.

18. Blöser, Claudia, Aron Schöpf, and Marcus Willaschek. 2009. Autonomy, experience, and reflection. on a neglected aspect of personal autonomy. *Ethical Theory and Moral Practice* 13: 239–253.

19. Arneson, Richard. 1994. Autonomy and preference formation. In *In Harm's Way: Essays in Honor of Joel Feinberg*, ed. Joel Feinberg, Jules L. Coleman, and Allen E. Buchanan, 42–75. Cambridge: Cambridge University Press.

20. Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

21. Anderson, Michael L. 2010. Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences* 33: 245–266.

22. Miller, Earl K., and Jonathan D. Cohen. 2001. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* 24: 167–202.

23. Knill, David C., and Alexandre Pouget. 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences* 27: 712–719.

24. Clark, Andy. 2013. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36: 181–204.

25. Körding, Konrad P., and Daniel M. Wolpert. 2006. Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences* 10: 319–326.

26. Girshick, Ahna R., Michael S. Landy, and Eero P. Simoncelli. 2011. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience* 14: 926–932.

27. Nassar, Matthew R., Robert C. Wilson, Benjamin Heasly, and Joshua I. Gold. 2010. An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of Neuroscience* 30: 12366–12378.

28. Weiss, Yair, Eero P. Simoncelli, and Edward H. Adelson. 2002. Motion illusions as optimal percepts. *Nature Neuroscience* 5: 598–604.

29. Goldreich, Daniel, and Jonathan Tong. 2013. Prediction, postdiction, and perceptual length contraction: a Bayesian low-speed prior captures the cutaneous rabbit and related illusions. *Frontiers in Psychology* 4: 221.

30. Gold, J. I., and M. N. Shadlen. 2007. The neural basis of decision making. *Annual Review of Neuroscience* 30: 535–74.

31. Beck, Jeffrey M., Wei Ji Ma, Roozbeh Kiani, Tim Hanks, Anne K. Churchland, Jamie Roitman, Michael N. Shadlen, Peter E. Latham, and Alexandre Pouget. 2008. Probabilistic population codes for bayesian decision making. *Neuron* 60: 1142–1152.

32. Vilares, Iris, and Konrad Kording. 2011. Bayesian models: the structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Sciences* 1224: 22–39.

33. Courville, Aaron C., Nathaniel D. Daw, and David S. Touretzky. 2006. Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences* 10: 294–300.

34. Braver, Todd S., and Jonathan D. Cohen. 2000. On the control of control: The role of dopamine in regulating prefrontal function and working memory. In *Control of cognitive processes: Attention and performance XVIII*, 713–737.

35. Rangel, Antonio, Colin Camerer, and P. Read Montague. 2008. A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience* 9: 545–556.

36. Ruff, Christian C., and Ernst Fehr. 2014. The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience* 15: 549–562.

37. Stephens, David W., and John R. Krebs. 1986. *Foraging Theory*. Princeton University Press.

38. Kolling, Nils, Timothy E. J. Behrens, Rogier B. Mars, and Matthew F. S. Rushworth. 2012. Neural mechanisms of foraging. *Science* 336: 95–98.

39. Pearson, John M., Sarah R. Heilbronner, David L. Barack, Benjamin Y. Hayden, and Michael L. Platt. 2011. Posterior cingulate cortex: adapting behavior to a changing world. *Trends in Cognitive Sciences* 15: 143–151.

40. McCoy, Allison N., and Michael L. Platt. 2005. Risk-sensitive neurons in macaque posterior cingulate cortex. *Nature Neuroscience* 8: 1220–1227.

41. Hayden, Benjamin Y., Amrita C. Nair, Allison N. McCoy, and Michael L. Platt. 2008. Posterior cingulate cortex mediates outcome-contingent allocation of behavior. *Neuron* 60: 19–25.

42. Gigerenzer, Gerd, and Wolfgang Gaissmaier. 2011. Heuristic decision making. *Annual Review of Psychology* 62: 451–482.

43. Damasio, Antonio R. 1996. The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 351: 1413–1420.

44. De Sousa, Ronald. 1990. *The Rationality of Emotion*. MIT Press.

45. Tappolet, Christine. 2014. Emotions, reasons, and autonomy. In *Autonomy, Oppression and Gender*, ed. Andrea Veltman and Mark C. Piper, 163–180. Oxford University Press.

46. Nader, Karim, Glenn E. Schafe, and Joseph E. LeDoux. 2000. Reply — Reconsolidation : The labile nature of consolidation theory. *Nature Reviews Neuroscience* 1: 216–219.

47. Stickgold, Robert. 2005. Sleep-dependent memory consolidation. *Nature* 437: 1272–1278.

48. Gais, Steffen, Geneviève Albouy, Mélanie Boly, Thien Thanh Dang-Vu, Annabelle Darsaud, Martin Desseilles, Géraldine Rauchs, et al. 2007. Sleep transforms the cerebral trace of declarative memories. *Proceedings of the National Academy of Sciences* 104: 18778–18783.

49. Schiller, Daniela, Marie-H. Monfils, Candace M. Raio, David C. Johnson, Joseph E. LeDoux, and Elizabeth A. Phelps. 2010. Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature* 463: 49–53.

50. Xue, Yan-Xue, Yi-Xiao Luo, Ping Wu, Hai-Shui Shi, Li-Fen Xue, Chen Chen, Wei-Li Zhu, et al. 2012. A memory retrieval-extinction procedure to prevent drug craving and relapse. *Science* 336: 241–245.

51. Schlichting, Margaret L, and Alison R Preston. 2015. Memory integration: neural mechanisms and implications for behavior. *Current Opinion in Behavioral Sciences* 1: 1–8.