Forthcoming in M. Blitz & J.C. Bublitz (eds.), *Neuroscience, Autonomy, and Individual Rights*, London: Palgrave Macmillan. Please cite published version.

Autonomy, Evidence-Responsiveness, and the Ethics of Influence

Fay Niker, Gidon Felsen, Saskia K. Nagel & Peter B. Reiner

Fay Niker is Lecturer in Philosophy at the University of Stirling.

Gidon Felsen is Associate Professor in the Department of Physiology and Biophysics and the Center for Bioethics and Humanities at the University of Colorado School of Medicine.

Saskia Nagel is Professor for Applied Ethics in the Department of Society, Technology, and Human Factors at RWTH Aachen University.

Peter Reiner is Professor in the Department of Psychiatry and founder of The Neuroethics Collective at the University of British Columbia.

I. Introduction

It is uncontroversial that the rise of the cognitive sciences, broadly construed, has had a significant impact on how we understand how humans think and behave. Robust sets of neurobiological and psychological findings concerning human cognitive processes have both challenged orthodox positions in, and raised new questions for, the disciplines of economics, philosophy, politics, and beyond.

To give a brief example: Findings relating to the automaticity and context-dependency of our rational processing and the dual-process theories of cognition that purport to explain them (Stanovich and West, 2000; Kahneman, 2011) have challenged our traditional views on rationality and suggest that *situated* conceptions of reason may be more appropriate (Hurley, 2011). In economics, this body of empirical research has served to establish behavioural economics as a distinct way of modelling human behaviour (Simon, 1972). In psychology, these findings were instrumental in directing attention towards the emotions and their role in practical and moral reasoning (Bagnoli, 2011; May and Kumar, 2018), precipitating debates over the viability of virtue ethics as a metaethical enterprise (Doris, 2002) and setting the contours for revisionary theories of moral responsibility (Doris, 2015). In public policy and political theory, empirical research (Tversky and Kahneman, 1981) informed the shift towards the use of "nudges" as a public policy lever (Thaler and Sunstein, 2009), and reopened philosophical debates about the nature and wrongness of both paternalism and manipulation (Coons and Weber, 2013; 2014).

Within this broad tradition of inquiry, there are questions that can be raised about the relationship between empirical work in the cognitive sciences and the concept of autonomy. Specifically, one might ask whether empirical insights from fields such as neuroscience, psychology, and experimental philosophy can enrich our understanding of the nature of personal autonomy.[1] This is the focus of the current work.

[1] This is distinct from the metaphysical question asking whether neuroscientific experiments have shown that free will is an illusion. For those who understand "autonomy" as within the family of metaphysical freedom terms (e.g., Mele, 1995; 2012), this metaphysical question is the same as asking whether neuroscience has shown that there are no autonomous human beings. There has been a lively debate over this issue (Lavazza, 2016). Yet, it is more common to make a distinction between personal autonomy and freedom, and we take this route. Freedom concerns the ability to act (and on some conceptions, "having sufficient resources and power to make one's desires effective"); whereas autonomy concerns "the independence and authenticity of the desires (values, emotions, etc.) that move one to act in the first place" (Christman, 2015).

Broadly understood, to be autonomous "is to be one's own person, to be directed by considerations, desires, conditions, and characteristics that are not simply imposed externally upon one, but are part of what can somehow be considered one's authentic self" (Christman, 2005). An agent who exercises this capacity to direct herself thus is said to be self-governing. The ability of individuals to exercise this self-government over their lives is a central value of many (though certainly not all) cultures and political systems, and it plays a weighty role in moral and political theorizing. As fundamental as this concept has been in the development of liberal thought, the task of specifying more precisely the conditions for autonomy has proven controversial.

A central distinction within this debate is between *internalist* and *externalist* conditions for autonomy. Some accounts are purely internalist in so far as they hold that whether an agent, or one of her decisions, can be described as autonomous or not depends entirely on features concerning her mental states. On Frankfurt's view (Frankfurt, 1971), for instance, a decision is autonomous if the first-order desire that motivates it coheres with the person's higher-order attitude on the matter. This is deemed to be the central factor relevant to autonomy ascription, regardless of how this higher-order attitude came about. This makes such coherentist accounts doubly internalist: autonomy ascription depends on *neither* how we came to have the relevant higher-order attitude or make the decision at hand – "a fact that is prior to (and in this sense external to) the action itself" – *nor* how our beliefs and attitudes relate to reality – "a fact that is independent of (and in this sense external to) the beliefs and attitudes themselves" (Buss and Westlund, 2018). This reveals two ways in which external factors may be relevant to autonomy ascription. First, we might be concerned with various ways in which the internal conditions of autonomy, such as the quality of our rational deliberation, are affected by external factors, such as socialization or manipulation. For this reason, many autonomy theorists place a procedural constraint on such internal conditions: what matters is that an individual's preferences and values have (or could have) survived the right kind of critical reflection (Dworkin, 1988; Friedman, 2003; Christman, 2010). Second, we might move beyond the effects of external factors on internal, subjective criteria, and instead hold that there are external conditions for autonomy concerning, e.g., how our beliefs and attitudes relate to reality.

We aim to show in this chapter that empirical research can provide some insights into the nature of autonomy, in particular, the internalist and externalist character of two broadly consensus conditions of autonomy. We'll be assuming an account that requires: (1) critical reflection on one's pro-attitudes; and (2) that one's decisions are not subject to undue external influence. We explore some ways in which empirical work interacts with this philosophical view, so as to bring additional nuance to the way in which the internalist–externalist distinction plays out with respect to these two conditions. More specifically, we explore an overlooked aspect of the relation between critical reflection and autonomy, which adds to the existing externalist concern about the historical formation of a person's higher-order attitudes another concern about how her beliefs and values relate to reality (section II). We then explore a novel, internalist dimension of the way in which a person's decision making is influenced by a range of external factors and actors, and consider the complex relationship between what we have termed 'pre-authorization' and autonomy (section III). We do so with particular reference to research that we have conducted on this topic in recent years (Felsen and Reiner, 2011; Nagel and Reiner, 2013; Felsen and Reiner, 2015; Niker et al., 2016; 2018a; 2018b), and with the aim of integrating this with other relevant work in philosophy and neuroscience. We then apply our analysis to practical situations in which infringement of autonomy is a concern – specifically, with respect to public policy nudges and the design of persuasive technologies – in order to draw out some of the implications of our theoretical discussion (section IV).

II. Critical Reflection and "Evidence-Responsiveness"

Making autonomous decisions requires certain *competencies*, such as capacities for internal self-reflection and for forming and revising one's beliefs and values. Moreover, these autonomy competencies must be exercised in ways that ensure that the resulting decision is *authentic* to the person in question – that it is her own decision in the relevant sense (Christman and Anderson, 2005). Classically, models of authenticity ensure this by claiming that autonomy requires critically reflecting upon and endorsing (or rejecting) one's motives. Critical reflection is generally considered to be the principal internalist dimension of autonomy, so-called because the process occurs entirely within the bounds of the mind. Through this process, a person shapes the attitudes that guide her decisions and actions. It is therefore both an important competency for autonomous decision-making, as well as a key part of the story about how the authenticity that is required for autonomy is achieved.

As noted above, this process of critical reflection is often thought to aim at bringing our first-order desires into coherence with our more reflective, higher-order desires (Frankfurt, 1971), thereby ensuring that a person identifies with or endorses her motives (Dworkin, 1988).[2] Other autonomy theorists maintain that there is more to the capacity for self-reflection than the capacity to hold higher-order attitudes. For instance, when we endorse our motives, we also implicitly make claims about which motives have the support of our practical reasoning (Buss and Westlund, 2018). This understanding of critical reflection has two important implications, both of which take us beyond coherentist accounts. The first is that it captures the intuition that someone who has been unduly influenced with respect to the development of their higher-order attitudes (e.g., indoctrinated or oppressively socialized), or whose practical reasoning has been manipulated in some other sense, would not be properly self-governing. We discuss the idea of undue external influence in more detail in the next section. The second is that, when we take account of practical reasoning, we see that autonomy requires that someone can change her mind when she discovers good reason to do so. We consider this feature of critical reflection in more detail in this section. We present a philosophical innovation, and then assess whether this garners empirical support from a neurobiological perspective.

A person's set of *pro-attitudes* – a term we use as shorthand for higher-order desires, preferences, values, beliefs, etc.[3] – underlies her autonomy in important ways. Debate on pro-attitudes in this context has focused either on coherence (i.e., between these attitudes and lower-order desires, as referenced above) or on history (i.e., how pro-attitudes were initially formed). But, a complete account of autonomy requires deeper consideration of the fact that we exercise and maintain our autonomy competencies *over time*. As experience of the world continues throughout life, our pro-attitudes may need to change in order to accommodate relevant new information – a process we can call *pro-attitude revision*. A thought experiment developed by Blöser et al. illustrates the matter:

[2] In our earliest studies of the relationship between the cognitive sciences and the concept of autonomy (Felsen and Reiner, 2011), we found that this philosophically-defined hierarchical schema broadly aligns with our understanding of the fundamental neurobiology of the brain – in particular with executive control theory in which the prefrontal cortex exerts a top-down influence over other brain regions (Miller and Cohen, 2001).

[3] Elsewhere within the philosophical debate over the nature of autonomy, what we are here labelling as a person's set of pro-attitudes are referred to variously as her "motivational set" (Weimer, 2013), "psychological core"(Noggle, 2005), or "collection of values" (Mele, 1995).

> "Pat is a 70-year-old man and a loving father and grandfather. He nevertheless finds it difficult to accept that his children and grandchildren live their lives in ways different from those that he himself pursued at their age. For example, his son has had his children out of wedlock, and Pat is convinced that children can only flourish within a stable family, which he believes to be one in which the children's parents are married. In accordance with the procedural account of autonomy, Pat is able to critically reflect on these issues in light of his existing pro-attitudes. But he holds the same pro-attitudes that he (that is, "younger Pat") authentically acquired half a century ago. What 'old Pat' struggles with is questioning his pro-attitudes in light of new experiences. Although his son's family provides a stable environment in which his grandchildren are flourishing, Pat is unable to reconsider whether marriage really is a basic requirement of good parenthood." (Blöser et al., 2010)

This thought experiment has been constructed to show that something important remains for a complete account of autonomy, even when the standard internalist requirements (i.e., those relating to Pat's capacity to reflect upon and endorse his pro-attitudes) and historical externalist requirements (i.e., that Pat's pro-attitude did not come about via any problematic interference) are met. This remainder relates to Pat's ability (or, more precisely, lack thereof) to reconsider his pro-attitudes in the light of new experiences or evidence – or, as we put it above, in light of the reality of the situation.

It would appear, then, that we can draw a distinction between two kinds of critical reflection that are relevant to autonomy: (i) critically reflecting on a pro-attitude in light of our other pro-attitudes, and (ii) critically reflecting on a pro-attitude in light of new experiences or evidence (Niker et al., 2018b). The problem with respect to old Pat's autonomy does not have to do with (i), because there is no inconsistency between his pro-attitudes. Rather, the problem arises from the fact that his value-based childrearing belief is "encrusted" (Blöser et al., 2010). He does not reflect upon this pro-attitude in light of his new experience of and evidence about childrearing as it applies to his grandchildren; it is his failure to exercise the critical reflection as in (ii) which undermines his autonomy with respect to this pro-attitude. In other words, the intransigence of Pat's previously acquired pro-attitude prevents him from (skillfully) adapting to new situations that merit re-evaluation of his existing pro-attitudes.[4]

[4] Similar views can be found, more implicitly, in earlier accounts of autonomy. One example is Richard Arneson's view, demonstrated by his claim that, "To live an autonomous life an agent must decide on a plan of life through

If this is correct, then a robust view of autonomy requires that we have the ability to critically reflect upon and to modify our existing pro-attitudes when our experiences or evidence call them into question. Elsewhere, we have described this in terms of the process of "updating our selves", by appropriately revising our pro-attitudes over time (Niker et al., 2018b). Blöser et al. (2010) label this capability *experience-responsiveness*, while Weimer (Weimer, 2013; 2017) refers to the same condition as *evidence-responsiveness*.[5] Here, we use the latter term, as this captures both the information acquired via a person's own direct perception as well as information garnered through the testimony of others. If we accept that evidence-responsive critical reflection has a place within a complete account of autonomy, we can see the externalist character of critical reflection *itself*, which goes beyond the weaker externalist character of protecting a person's internal critical reflection process from being unduly shaped by external influences.

To what degree does neurobiological data align with this this philosophical innovation? This is a complex issue; here, we outline a set of observations relating to how pro-attitudes might be represented in the brain which suggest the beginnings of a neurobiological framework for evidence-responsive critical reflection. We begin from the claim that, given that pro-attitudes represent a distributed set of desires, beliefs, values, and so on, they are less likely to be instantiated as discrete memories than as widely dispersed networks of information, consistent with modern theories of information storage in the brain (Dehaene et al., 1998; Squire, 2004). These distributed networks then represent the neurobiological correlates of our pro-attitudes.

Arguably the best candidate for a plausible mechanistic explanation of the process of critical reflection is the phenomenon of Bayesian inference (Knill and Pouget, 2004). In this schema, decisions are represented probabilistically and result from combining two sources of information: internally generated "priors" – our pro-attitudes – and the new information that is associated with a particular decision. The two are provisionally integrated in the brain, generating a new statistical inference. The relative value of this new inference, as well as a measure of confidence in this evaluation, is then determined (Meyniel and Dehaene, 2017). Such evaluation

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" (Arneson, 1994).

[5] There is a strand of autonomy theory which defines autonomous decision-making in terms of reasons-responsiveness. Without endorsing this theory, here we simply point out that evidence-responsiveness might plausibly be understood as a specific way of responding to reasons, namely, responding to reasons-to-review or reasons-to-revise a pro-attitude that one currently holds (Niker et al., 2018b).

is the essence of critical reflection – appraising the likelihood that the new inference provides a more or less useful strategy for moving forward. When this process makes space for incorporation of new information, it qualifies as evidence-responsive. In addition to influencing specific decisions, new information – if it provides sufficiently compelling evidence – can also be used to update the priors themselves, which will then be applied to subsequent decisions. To return to our example of Pat: if he were capable of revising his pro-attitude that family stability requires marriage, based on the strong evidence provided by his flourishing grandchildren, he would be able to autonomously accept his son's – and even others' – decisions to have children out of wedlock.

The diffusion-to-bound model, a formal model of perceptual decision-making (Ratcliff and Rouder, 2016), helps to illuminate how this might work (Bitzer et al., 2014). The model proposes that one's options are represented as bounds, and a "decision variable" evolves in a multi-dimensional bounded space as we integrate information relevant to the decision with our priors. When the decision variable reaches one of the bounds, a decision is made which corresponds to selecting that option. This model can explain a range of behavioral phenomena and is consistent with extant neurophysiological data (Smith and Ratcliff, 2004; Gold and Shadlen, 2007); it also provides a useful framework for evaluating external influences on decision making (Bode et al., 2014) and how they affect autonomy of choice (Felsen and Reiner, 2015). While this model is consistent with executive control theory (Miller and Cohen, 2001), and represents an explicit, top-down process of evaluation, it is also possible to incorporate new information *below* the level of conscious awareness (Dehaene and Changeux, 2011). This does not preclude the possibility of top-down reflection, but it is consistent with the idea that non-conscious processing of inputs such as emotions can provide a useful heuristic for efficient decision making (Gigerenzer and Gaissmaier, 2011).

Critically, the distance between the starting point of the decision variable and the decision bound determines the degree of evidence required to select the option represented by the bound: the further the bound, the more evidence is required. Thus, by setting the bound corresponding to the choice consistent with priors closer to the starting point of the decision variable, that option is more likely to be selected, without precluding the selection of alternatives given sufficient countervailing evidence. Bound setting is under top-down cortical control (Mulder et al., 2012), providing a mechanism for the influence of priors on decisions and for updating the priors themselves. To return again to Pat: Given his pro-attitude that marriage is required for family stability, the variable representing his decision about his son's choice to raise children out of wedlock effectively begins at the "reject" bound. To have any chance of the variable reaching the "accept" bound, the reject bound must be shifted away from the decision variable's starting point in response to the new evidence that Pat's grandchildren are flourishing despite their parents being unmarried.

We hope to have provided a philosophical account of evidence responsiveness and a sketch of how this process might occur in the brain. While much work remains to link our philosophical and neurobiological explanations (Niker et al., 2018b), we hope that our preliminary work can provide a framework for future studies examining pro-attitudes in terms of priors, how the neural representations of priors are updated by new evidence and the extent to which decisions based on these updated priors are perceived as autonomous[6].

[6] A second stream of neurobiological observations, specifically developed to account for long-term memory formation but likely also relevant to the incorporation of the distributed set of desires, beliefs, values, and so on that represent our pro-attitudes, provides a plausible mechanism for this process. The key finding is that memories are not static but subject to a cycle of deconsolidation and reconsolidation (Nader, 2015). To best understand how this works, think for a moment of a teacher that you had when you were in elementary school. The first salient observation is that you have been able to maintain a memory of that teacher for all these years – for some readers that would be several decades. This is the way we normally think of memory – as a stable feature of normal brain physiology. But while memories are indeed stable for years, the very act of recalling them transforms them from stable to labile. At this very moment, your memory of your elementary school teacher is not protected in the same way that it has been during the years that it lay dormant, but rather is available to develop a new set of associations. These associations, which likely arise via the process of Bayesian inference discussed above, are then stabilized by a process known as reconsolidation, most likely during a subsequent night's sleep (Tononi and Cirelli, 2014; Klinzing et al., 2019). Critically, when the existing memories and the new information are reconsolidated, they are linked; in our example the array of memories about your years in elementary school would be linked to this particular discourse on memory consolidation. Weeks from now, perhaps at a dinner party, you may share this thought experiment with a group of friends. When you do so, you will be drawing upon the association of these two memories to recall how the experiment works. As you delight your friends with your new insight, these memories will once again be labile in our brain, slated for reconsolidation when you return home for a good night's sleep.

Together, this set of observations provides a neurobiological framework for evidence-responsive critical reflection. Bayesian inference draws together extant and new information, providing a mechanism for critical reflection, and then the iterative process of deconsolidation and reconsolidation provides a mechanism for incorporating external information into our existing pro-attitudes – the essence of evidence-responsiveness. This process repeats itself throughout our lives, and we suggest that ability to engage in evidence-responsive critical reflection represents an important part of this key condition of autonomy.

There is evidence to suggest that older brains are less agile in this regard. While substantial plasticity occurs in the aging brain (Gutchess, 2014), a wealth of data supports the view that fluid cognitive abilities such as working memory, attention and executive control decline with age, while crystallized cognitive abilities are preserved (Samanez-Larkin and Knutson, 2015). Because fluid cognitive abilities are precisely those that are required to nimbly manage new information, those who are best endowed with these traits will naturally be in the strongest position to utilize them in a process of evidence-responsive reflection. It is for this reason that Blöser et al.'s choice of an elderly person in the example of 'old Pat' is so plausible: it is certainly not the case that *all* elderly people have strongly fixed pro-attitudes, A second stream of neurobiological observations, specifically developed to account for long-term memory formation but likely also relevant to the incorporation of the distributed set of desires, beliefs, values, and so on that represent our pro-attitudes, provides a plausible mechanism for this process. The key finding is that memories are not static but subject to a cycle of deconsolidation and reconsolidation (Nader, 2015). To best understand how this works, think for a moment of a teacher that you had when you were in elementary school. The first salient observation is that you have been able to maintain a memory of that teacher for all these years – for some readers that would be several decades. This is the way we normally think of memory – as a stable feature of normal brain physiology. But while memories are indeed stable for years, the very act of recalling them transforms them from stable to labile. At this very moment, your memory of your elementary school teacher is not protected in the same way that it has been during the years that it lay dormant, but rather is available to develop a new set of associations. These associations, which likely arise via the process of Bayesian inference discussed above, are then stabilized by a process known as reconsolidation, most likely during a subsequent night's sleep (Tononi and Cirelli, 2014; Klinzing et al., 2019). Critically, when the existing memories and the new information are reconsolidated, they are linked; in our example the array of memories about your years in elementary school would be linked to this particular discourse on memory consolidation. Weeks from now, perhaps at a dinner party, you may share this thought experiment with a group of friends. When you do so, you will be drawing upon the association of these two memories to recall how the experiment works. As you delight your friends with your new insight, these memories will once again be labile in our brain, slated for reconsolidation when you return home for a good night's sleep.

Together, this set of observations provides a neurobiological framework for evidence-responsive critical reflection. Bayesian inference draws together extant and new information, providing a mechanism for critical reflection, and then the iterative process of deconsolidation and reconsolidation provides a mechanism for incorporating external information into our existing pro-attitudes – the essence of evidence-responsiveness. This process repeats itself throughout our

but it is common to encounter older people who cling to their previously acquired pro-attitudes, and this impairs his ability to fully engage in evidence-responsive reflection.

lives, and we suggest that ability to engage in evidence-responsive critical reflection represents an important part of this key condition of autonomy.

There is evidence to suggest that older brains are less agile in this regard. While substantial plasticity occurs in the aging brain (Gutchess, 2014), a wealth of data supports the view that fluid cognitive abilities such as working memory, attention and executive control decline with age, while crystallized cognitive abilities are preserved (Samanez-Larkin and Knutson, 2015). Because fluid cognitive abilities are precisely those that are required to nimbly manage new information, those who are best endowed with these traits will naturally be in the strongest position to utilize them in a process of evidence-responsive reflection. It is for this reason that Blöser et al.'s choice of an elderly person in the example of 'old Pat' is so plausible: it is certainly not the case that *all* elderly people have strongly fixed pro-attitudes, but it is common to encounter older people who cling to their previously acquired pro-attitudes, and this impairs his ability to fully engage in evidence-responsive reflection.

III. External Influence and "Pre-authorization"

The second condition of autonomy that we're assuming in our inquiry is that, for a person's decision to be autonomous, it must not be the result of undue external influence. It must be "hers" in the appropriate sense. It is relatively simple to agree that certain forms of influence are undue, in so far as they present an obvious threat to a person's autonomy. This is especially the case when it comes to heavy-handed forms of external influence such as brainwashing or coercion (Chen-Wishart, 2006). But in the course of our day-to-day lives, we continuously encounter a range of external influences that run the spectrum from overt to subtle, and on to imperceptible to those whom they affect. Determining which of these various influences are to be considered "undue" is a complicated matter.

As mentioned in the introduction, research in the cognitive sciences has shed light upon the extent to which our decisions are influenced by seemingly-irrelevant situational factors, and has sought to explain how and why this often happens below the level of our conscious awareness. The robustness of this empirical research has laid the foundations for important shifts in philosophy of mind, including moves towards understanding cognition as embedded in and extended into our external environments. On such situated conceptions, decisions result from an interaction between mind and environment; decisions are, as a matter of fact, always influenced by external factors to some extent. We might worry about this from the perspective of autonomy, perhaps because it makes it more difficult to discern which influences are permissible (in so far as they respect autonomy's authenticity conditions) and which are not; but we might also think that this situated conception of cognition provides some insight into the concept of autonomy itself.

Such an insight, we think, would be related to a philosophical innovation in the debate over autonomy in recent years. This has centered not on empirical work on cognition, but rather on theoretical work on conceptions of the self; both kinds of work, though, are connected by the fundamental role that they give to social embeddedness. Constructive critiques from feminist philosophers have led to a reconceptualization of autonomy in light of appropriate appreciation being given to the fact that we are relational beings – beings who are not only continually subject to external influences, but who require them in order to develop and exercise our autonomy competencies (Meyers, 1989). Often collectively termed *relational autonomy* (Mackenzie and Stoljar, 2000), the twofold motivation of such accounts is to show, on the one hand, that "rational autonomous capacities are made possible by the support of numerous surrounding agents who enable careful reflection and judgment" and on the other, that "individuals' autonomous capacities can be disabled or oppressed by the withholding of this contextual support" (Specker Sullivan and Niker, 2018). This reconceptualization offers rich opportunities to delve deeper into the question of when and why an influence is considered undue.

Much philosophical attention has been given to determining which types of influences are morally problematic – *how* a decision is influenced, and the ethical character of these various types, has been debated in detail. There are, for instance, distinct and in some cases, burgeoning philosophical literatures on the nature and (political) morality of coercion (Anderson, 2010; Wertheimer, 2014), manipulation (Coons and Weber, 2013), persuasion (McKenna, n.d.), upbringing and socialization (Clayton, 2006), and nudging (Sunstein, 2016; Niker, 2018). Interestingly, though, there has been much less discussion of a different feature that may be relevant to the "dueness" of external influence, namely, *who* is exerting the influence and how the person who is subject to it understands their relationship to this influencing actor. We intuitively allow some people, institutions, and so on to have a greater influence upon our decision making than others. To put it another way, information from certain actors is viewed as a welcome input into our decision making, but this is not so when the very same information comes from other actors. In recent work, we have sought to offer a conceptual tool for better understanding this selective process regarding the source of external influences and to examine how this relates to (relational) autonomy.

It is plausible to think that whether information is regarded as welcomed or not by a given person depends not only upon its relevance to the decision at hand, but also upon that person's perception of the reliability of the source of that information. We have termed the latter sort of consideration *pre-authorization* (Niker et al., 2016). We operationally define pre-authorization as an evaluative stance by which an individual gives certain agents preferential access to influencing her decision-making processes (Niker et al., 2018a). Several reasons can be put forward for pre-authorizing an agent. One prominent example occurs when we perceive that the agent has values, commitments, and goals that are similar to ours – that is, that in some meaningful way they share our worldview. Another common situation is one in which the agent has some relevant knowledge or expertise that we do not have and which we can trust, for example when we consult with a physician or a lawyer. The result, in both cases, is that we feel comfortable incorporating information from these agents into our decision-making. More specifically, the evaluative stance taken by an individual towards some agents means that an influence from a pre-authorized agent is incorporated in relevant future interactions without necessarily needing to be consciously evaluated, and without impacting the individual's perception of the control that she has over, and the authenticity of, her resultant decision (Niker et al., 2018a). We have suggested that the extent to which the source of an influence is pre-authorized contributes to our perception that we are making an autonomous decision. A person's actual autonomy and her perceived autonomy can be distinct – for example, while in practice an intervention does not impact on a person's decision-making capacity, she might perceive that it does, or vice versa. Yet, if pre-authorization can be shown to play a role in what we might call the "folk" conception of autonomy, this would justify consideration of its relation to the autonomy competencies, as understood on a relational account of autonomy.

To further explore whether the *concept* of pre-authorization has some basis in the way that people view influences upon their decision making, we carried out a set of empirical studies. We particularly examined how people perceive of everyday socio-relational influences on their decisions, such as a news clip on a social media platform, a friend's comment or suggestion, a notification from an app, and so on. The data, derived from carefully balanced contrastive vignettes, demonstrated that the influence of pre-authorized agents with whom we share a worldview – be they individuals or institutions – was judged to be significantly less undue than when that same influence derived from non-preauthorized agents. One might imagine that this was secondary to our familiarity with the agent, because in the normal course of events, we are usually better acquainted with those with whom we share a worldview than those with whom we do not. Yet these effects persisted even after controlling for the familiarity of the agent. Thus, we found that the public's conception of when an influence is welcome or not is indeed dependent upon the source of the influence, providing initial support for the validity of the concept of preauthorization (Niker et al., 2018a).

Another way of saying this is that we evaluate not just the content of information that arrives at our doorstep but also its pedigree. Does it come from a trusted source? Is it from someone who shares our worldview? Is it from someone who has expert knowledge on the topic? These questions define our attitude towards the source, and that in and of itself affects the degree to which we allow it to have an influence over our decision-making processes. From an empirical perspective, we have hypothesized that our brains have something akin to a *skeptical filter*, and that our evaluation of the pedigree of the information determines the stringency of the skeptical filter we apply to it. When it comes from a pre-authorized source, the skeptical filter is loosened, making it easier for that information to "get through" and influence the decision at hand. When it comes from a source that is not pre-authorized, our evaluation of the information is more rigorous, calling for further cognitive work. We suggest that an important autonomy competency is the ongoing maintenance of this skeptical filter, using it as a means of authorizing external influences that are consistent with one's goals, values, desires, convictions, and life plan.[7] There is a modicum of evidence in support of the existence of this filter. For example, people use more stringent criteria to evaluate others' arguments than when they produce arguments themselves (Bode et al., 2014; Felsen and Reiner, 2015). Moreover, the concept is consistent with neurobiological descriptions of decision making that account for the incorporation of external influences (Shadlen and Roskies, 2012; Bode et al., 2014). Nonetheless, the precise neural circuitry that undergirds this phenomenon is currently unknown.

[7] This maintenance may include engaging in evidence-responsive critical reflection in order to update the stringencies attached to way the filter functions, as and when appropriate, so that they don't become "encrusted" in the way discussed in section II.

How does this relate to autonomy? The answer is not entirely straightforward. As noted above, our studies grounding the concept of pre-authorization test a person's *perceptions* about whether an external influence is welcome or not. But whether an influence is considered welcome by a person for the purposes of her decision making is not the same thing as it being a "due" or morally permitted influence; it acts merely as a proxy. While often overlapping, a person's autonomy and her perceived autonomy aren't the same thing. Insofar as they overlap, we might say that the phenomenon of pre-authorization is one particular way in which we can capture the role and value of interpersonal relations in supporting autonomy – or more specifically, in supporting a person's ability to make autonomous decisions under real-world conditions, where information is both abundant and costly and where time is often limited. Pre-authorization provides us with a possible mechanism by which we exercise autonomy relationally.

We might think, then, that pre-authorization fits into the framework of *autonomy support* (Nagel and Reiner, 2013; Nagel, 2015), which acknowledges the social and relational ties that bind and support individuals in their making of decisions throughout their life (and is discussed in more detail in the section IV). On this view, autonomy is an intersubjective phenomenon that is not only developed socially but is also constantly reflected, maintained, and advanced in relational contexts. What is interesting about the concept of pre-authorization is that it posits an empirically plausible (though unverified) means by which an individual can exert control over the differential impact of external sources of influence on her decision-making processes, as determined by how these sources relate to her own beliefs, values, life plan, etc. This is interesting from the perspective of the framework set up by the chapter because, if accepted, pre-authorization highlights a novel *internalist* feature of this externalist condition (i.e., of not being subject to undue external influence). Together with the conclusion of the previous section, this further problematizes any clear distinction between internalist and externalist conditions for personal autonomy.

But, as insinuated above, there is much more to say about the relationship between pre-authorization and autonomy. Our notion of the skeptical filter provides some insight into one of the pitfalls of pre-authorization. As Onora O'Neill has pointed out, trusting others to provide us with information is only valuable if the individual or institution is in fact trustworthy (O'Neill, 2018). Thus, if we pre-authorize an agent and they lead us astray by convincing us of incorrect information that they sincerely believe, or worse, by using our confidence in them to manipulate us, we are in a very bad situation indeed, as the loosening of the skeptical filter causes us to less rigorously assess the veracity of their claims. In this way we see that the heuristic nature of pre-authorization – a quick and efficient but nonetheless imperfect solution to evaluating external information – can lead to situations in which our autonomy may be subverted.

The ideal version of pre-authorization is one in which a person has reflected upon the issue and intentionally decides to preauthorize another agent (these days it is probably wise to include algorithmic agents in the mix). But in practice, this is not what normally happens. The canonical example is a friendship that develops over time. Initially, both parties might be open to each others' ideas but still a bit skeptical. Over time, as they get to know and trust each other, they begin to pre-authorize each other to influence their thinking on certain matters. But they are unlikely to stop and say something like, "Wow, my friend Judy seems like a really good person to take advice from. I think I will do so from here on in." Rather, the pre-authorized relationship develops in an implicit manner. Indeed, one may not even explicitly realize it has happened, unless prompted to reflect on the issue. What we don't know is how, from a mechanistic point of view, this process plays out. What we do know is that over time, we come to rely upon some individuals more than others, and past experience is one factor that plays into the process. All of this is to say that our vision of the concept of preauthorization holds less in common with a legally binding grant of power than the sort of power exchange that occurs informally amongst parties with everyday social interaction.

Another interesting dimension of the relationship between pre-authorization and autonomy comes from the former's inverse. Although we have not specifically tested the hypothesis, it seems plausible that actors may not only preauthorize but also *anti-preauthorize* other agents. This has become a common trope in modern life in which the partisan nature of political positions and the structure of our informational landscape allows us to ignore information that derives, e.g., from news sources that do not align with our worldview, irrespective of the comparative factual quality of the different outlets (Bessi et al., 2016; Del Vicario et al., 2016). Thus, while pre-authorization may be a useful heuristic in so far as it allows us to more easily integrate information from trusted kith and kin, its inverse, anti-preauthorization, may be a factor that negatively affects our capacity to make informed decisions and to engage in the evidence-responsive critical reflection discussed in the previous section.

IV. Implications for the Ethics of Influence

In previous sections, we have considered some of the issues involved in two consensus conditions of autonomous decision making – critical reflection and not being subject to undue external influence – from the perspective of both philosophy and neurobiology. We turn our attention now to exploring the practical relevance and potential implications of our theoretical discussion for real-world scenarios about which there is concern over autonomy. We focus in particular on the phenomenon of *nudging*, both as it functions as a public policy lever and the role it plays in the design of persuasive technologies (Thaler and Sunstein, 2009). It makes an illustrative case because: (i) the ethical debate over nudging has centered on autonomy; and (ii) we think that both the issues of critical reflection and evidence-responsiveness (section II) and pre-authorizing selected sources of external influence (section III) have interesting implications for the debate over the ethics of nudging. Indeed, our analysis shows that these two aspects of our theoretical discussion are heavily interrelated in the practical case of nudging.

Nudging involves intentionally modifying a person's choice environment in order to predictably, yet non-coercively, influence her decision making towards a specified end. Introduced as a public policy lever aimed at promoting individual and social welfare (Thaler and Sunstein, 2009), this form of influence was provocatively termed *libertarian paternalism*.[8] When motivated in this way, a nudge is paternalistic because the "choice architect" intervenes with the best interests of the nudged person in mind. But this welfare-promoting aim is reined in by liberal values, it is thought, because the nudged person is not forced to decide in accordance with the nudge; for it to count as a nudge, she needs to be free to opt out with relative ease. Such interventions find their rationale and operational mechanisms in the empirical research grounding situated conceptions of rational agency. This research has shown that environmental settings have a deep impact on the decisions people make, such that "seemingly trivial changes in the way information is conveyed, choices are arranged, or default rules are set" can affect the decisions they make (Moles 2015). For instance, whether an in-work pension scheme or organ donation registration scheme has an opt-in or an opt-out default makes a considerable difference to the uptake of both. Knowledge of the various ways in which cognitive heuristics and biases affect our decisions makes it possible to

[8] Despite the initial equation of nudges with a form of paternalism, it is now well-established that nudging is a type of influence that can be used in service of different ends. While we may be motivated to nudge for paternalistic reasons, we might also use nudges for the purpose of promoting justice, utility, commercial profit, or so on.

design choice architecture in a way that steers, or "nudges," people in a particular direction. Several governments have, in recent years, changed the default of these two schemes, with the explicit aim of, for example, producing higher rates of savings and cadaveric organ donation. Several governments have now adopted nudging as a policymaking technique, but this move has not been without its critics (Goodwin, 2012; Yeung, 2012; Waldron, 2014).

Much of this critical engagement has examined nudging's relationship to autonomy (Grune-Yanoff, 2012; Felsen et al., 2013; Wilkinson, 2013; Engelen and Nys, 2019). One under-theorized critique of nudging from this direction is that it may infringe upon the *development* of autonomy competencies. Blöser et al. (2010) emphasize that one must recognize an experience as being new and relevant in some manner as a pre-condition for evidence-responsive critical reflection; nudges may diminish the opportunity to engage in such reflection. Consider an adolescent who, rather than finding their own way in the world by 'learning from their mistakes', has parents who remove obstacles from their path – a situation commonly known as "snowplow parenting". In essence, these adolescents live in an environment that is designed by choice architects (their parents, in this case) to make the best decisions most likely. They may end up with decisions that are welfare-promoting, or even ideal in some sense, but there is less opportunity for them to develop the fundamental skills involved in decision making. The worry is that a similar sort of diminishment of human decision-making competencies is going on in a world structured by public policy nudges. This is especially so if we agree with critics that nudges work by bypassing our deliberative capacities (e.g., Grüne-Yanoff, 2012); operating in this way would threaten the development and exercise of several autonomy competencies, not only evidence-responsive critical reflection.

But, as Neil Levy has recently argued, there is at least a certain kind of nudge – which he calls *nudges to reason* – which might have an important role to play in helping us to become *more responsive* to genuine evidence (Levy, 2017). In recent years, much attention has been directed towards issues relating to evidence-responsiveness in a so-called "post-truth" world. This has been bolstered by findings such as the "backfire effect" – which describes the phenomenon that occurs when those who are motivated to resist and reject (some kinds of) evidence become more entrenched in their false beliefs after being presented with arguments citing such evidence (Nyhan and Reifler, 2013). This related set of issues clearly pose a threat to the flourishing of democratic systems (e.g., the possibility of having a well-informed electorate), to public health (e.g., the case of anti-vaxxers), and to climate justice (e.g., the case of climate change deniers). Levy suggests that nudges to reason may offer an effective and ethically permissible means of addressing such false beliefs by increasing responsiveness (or, at least, reducing *perverse* responsiveness) to evidence. He accepts the critic's claim that interventions into decision-making and belief formation threaten a person's autonomy when they bypass her capacities for deliberation; but nudges to reason, he argues, address themselves to capacities that are partially constitutive of a person's reasoning (Levy, 2017). In so doing, these interventions do not offend against autonomous decision-making and, in fact, they may support autonomy by enabling people to engage in evidence-responsive critical reflection. How might they do this?

One of the ways in which psychologists have found we can become more responsive to evidence relates to recent insights into how we respond to testimony. As Levy explains,

> "Children and adults must learn from others: there is a great deal that we cannot check for ourselves, and a great deal more that it would be too time-consuming or otherwise costly to check. In the contemporary world, we rely on medical specialists to diagnose our ills, technology specialists to fix our computers, accountants to manage funds for our retirement and meteorologists to advise us when to hold a picnic. But this reliance on specialists […] is a feature of traditional societies too. Canoe making, for instance, is a specialised skill, and not everyone has the time to acquire it. Moreover, skill acquisition is itself dependent on the acceptance of testimony: children often cannot discover essential techniques for survival themselves, and must be taught them. […] For all these reasons, we are often forced to learn from others in the absence of a capacity directly to gauge how reliable they are. We are therefore forced to use cues to reliability; cues which reliably enough correlate with being a good source of testimony." (Levy, 2017)

This relates directly to the concept of pre-authorization discussed above; in essence, a person uses cues of reliability and benevolence to help her to determine which information to take account of in their belief-formation and decision-making processes. In the case of correcting false beliefs, it has been shown that a person's sensitivity to these cues plays a role in explaining why some corrections are successful, while others are not. For instance, Nyhan and Reifler (2013) found that the source of the information made a significant difference to whether corrections of myths about President Obama's policies were successful for conservatives or not. In fact, there were two source-based considerations that produced this effect: both the perceived ideological leaning of the media outlet that reported the debunking claim, and the source of the claim (i.e., whether it was attributed to a liberal, non-partisan, or conservative think tank) (Nyhan and Reifler, 2013; Levy, 2017).

This evidence opens up the possibility of counteracting public ignorance and misconceptions by designing interventions that present evidence in certain ways. The most relevant case for our purposes concerns intentionally selecting the source(s) of the evidence so as to increase the likelihood that (a certain set of) people will respond to it as they rationally ought to. But there are also other techniques such as “moral reframing”, which works by framing a position that an individual would normally not support in a way that is consistent with her values and so thus positively affects the credence she gives to it (Feinberg and Willer, 2019), in line with the rational significance of genuine evidence. Should these nudge interventions – and, in particular, the testimonial version that is of particular interest to us – be regarded with the same sort of suspicion as other nudges? And if not, why not?

According to Levy, these testimonial nudges count as nudges to reason because, rather than modify a person’s behaviour directly, they do so by seeking to alter her beliefs through the process of making her more responsive to evidence. Nevertheless, critics may accept this while remaining worried about how these nudges affect this change of mind, where the concern is just a variant of the standard worry that such interventions operate by bypassing our deliberative faculties. The real reason explaining why we changed our mind, it might be thought, has to do with the selection of a source that has been intentionally chosen to avoid the backfire effect; and so, “by bypassing our deliberative capacities, [such nudges] may threaten the substantive freedom of our choices even if they succeed in making us more responsive to the evidence” (Levy, 2017). There are different responses available; but the more interesting, from our perspective, is to deny that nudges to reason do in fact bypass an individual’s deliberative faculties. Instead, such interventions are “designed to be processed by filters that are partially *constitutive* of reasoning in normal functioning agents” (ibid.). In Levy’s terms,

> “[a] process is a proper part of reasoning […] when it regularly and reliably supports better deliberation (either in a domain-general or a domain-specific manner)… Appeals to the mechanisms that weigh testimony by reference to their source are very plausibly appeals to mechanisms that are partially constitutive of rationality, because we likely have such mechanisms in virtue of the role they played in enabling better decision-making. [T]hese mechanisms are sensitive to the previous track record of the source. That is, very obviously, sensitivity to a property that is truth-conducive. We

> should put less weight on the testimony of those who are frequently wrong than those who have better records. Similarly, sensitivity to the ideological orientation of the source is also truth-conducive. We should be wary of the claims of people who lack benevolence towards us, because they may be motivated to exploit us. We also should put more stock in testimony from agents who have an incentive to reject the claim they affirm… Sensitivity to these properties is sensitivity to considerations that are relevant to the credence we should place on testimony. Appealing to them is appealing to capacities that have as their proper function the assessment of reasons for belief – a function that is obviously partially constitutive of reasoning – in their role as reasoning mechanisms." (Levy, 2017)

If this argument is correct, nudges to reason may permissibly be used to counteract false beliefs held by the public. By presenting evidence via a source that is more likely to be pre-authorized, and hence more likely to make it through the skeptical filter, these nudges support a person's capacity for evidence-responsiveness and for evidence-responsive critical reflection.[9] Given our analysis, then, it is plausible that nudges to reason support the exercise of autonomy competencies, especially when autonomy is conceptualized in relational terms. Of course, not all nudges are nudges to reason; indeed, most would not be categorized as such, so our conclusion applies only to a subset of nudges.

In a sense, Levy's nudges to reason can be viewed as an example of *autonomy support* – a strategy introduced in the previous section that aims to help individuals arrive at decisions that are aligned with their values, needs, preferences, and desires. Originally developed as a means of supporting individuals in developing autonomy competencies, particularly in the domains of education and the workplace (Reeve, 1998; Ryan and Deci, 2000), the concept of autonomy support can be thought of as a set of strategies that assist people in developing and executing autonomy competencies throughout their life course (Nagel, 2015). Unlike classical nudges that are designed to make it more likely that an individual arrives at a decision that the choice architect has deemed to be in their best interests, the external influences that comprise autonomy support give extra weight to respect for the person, devoting effort to consider how one might empower individuals to arrive at decisions that are in their own best interests.

[9] Neurobiologically, this could be represented as shifting the starting point of the drift-diffusion process closer to one of the bounds (Felsen and Reiner, 2015). Often, as with encrusted values, bounds are set by internal biases. By changing the relative distances to bounds, nudges can be seen to counteract such internal biases in ways that are (more) consistent with the agent's pro-attitudes.

But there is another sense in which pre-authorization seems to be a useful concept for understanding another phenomenon associated with public policy nudging. Namely, pre-authorization may be one of the factors that explain why certain nudges are perceived as more or less welcome. There is empirical data showing that certain contextual factors make a difference to whether any given nudge is perceived by the public as infringing upon or respecting their autonomous decision-making (Castelo et al., 2012; Felsen et al., 2013; Jung and Mellers, 2016). In an era in which trust in institutions is weakening, this has substantial implications for public policy initiatives which employ nudges to alter citizens' behavior. Indeed, these data may go some way towards explaining the phenomenon of *partisan nudge bias*, whereby attitudes toward particular policy goals or policymakers – i.e., whether they align with the actor's goals and commitments – affect attitudes about the moral permissibility of the nudge policy itself (Tannenbaum et al., 2017).

We move now to another example within the ethics of influence that draws together the concepts of nudging, pre-authorization, and autonomy support, namely, the ethical dimensions of persuasive technologies. In the modern world, influence over our decision-making is increasingly exerted not by other humans but rather via software on our algorithmic devices, colloquially known as 'apps'. It is well-established that by monitoring our digital footprints, software can predict a great deal about us, from Big Five personality traits to our political views and more (Kosinski et al., 2013; Matz et al., 2017). This information can then be used to micro-target individuals in an effort to persuade – or nudge – them to follow one or another course of action (Calo, 2014; Frischmann and Selinger, 2018; Susser et al., 2018). Karen Yeung calls this "hypernudging", because these Big Data analytic nudges are much more potent than their standard public policy counterparts on account of "their networked, continuously updated, dynamic and pervasive nature" (Yeung, 2016).

The potency and personalization of persuasive technologies make them novel; but so does the fact that, through repeated use, we accept our algorithmic devices – exemplified most obviously by the smartphone – as extensions of our minds (Clark, 2008). As we do so, we increasingly rely upon them as a trusted source of information, social interaction and approval, and a means of offloading cognitive work (Fitz and Reiner, 2016; Reiner and Nagel, 2017). If, as seems to be the case, we treat apps as pre-authorized agents (Niker et al., 2018a), we allow them to have an outsized influence upon our decision-making. Although there have already been several substantial efforts to explore these issues (Yeung, 2016; Susser et al., 2018; Williams, 2018), there is much work still to be done in this area of applied ethics.

But rather than simply critiquing persuasive technologies, it is perhaps apropos to highlight how our relationship with persuasive technologies might be constituted such that it is supportive of our autonomy competencies. Consider the app *Moment* which helps people manage their smartphone usage. It resides on the device and, after you grant it sufficient privileges, it monitors most of what you do on your phone during the day. It doesn't prevent you from using your phone (unless you ask it to), but from time to time it gives you for feedback on how much you have used your phone, and even includes a reminder of what your goal for phone usage might be. In this way, the *Moment* app causes you to critically reflect upon your phone usage by presenting you with evidence of your current usage. This, we suggest, is an existing example of an algorithmic nudge to reason (Levy, 2017). By regularly prompting you to reflect on your choices of phone usage, the app helps you to make an autonomous decision to keep your phone usage at a level that you wish it to be (Specker Sullivan and Reiner, 2019). This represents a plausible example in which nudges can be harnessed to support autonomy, at once helping humans make better decisions and become better decision makers.[10]

V. Conclusion

Autonomy, with its implications for moral, political, and philosophical thought, is a well-studied concept in Western intellectual thought. Nonetheless, there remain opportunities to advance our knowledge in this realm, and this chapter represents our attempt to explore recent progress in our understanding of two consensus conditions of autonomy – critical reflection and not being subject to undue influence. Our consideration of these matters has attempted to integrate conceptual work with empirical research in the cognitive sciences. In both cases, our analysis has put pressure on the idea that we can draw any clear distinction between internalist and externalist conditions for personal autonomy.

Critical reflection upon one's pro-attitudes is a fundamental internalist condition of autonomy. We have suggested that the critical reflection required for autonomy includes critically

[10] We do recognize, though, that most of the worries about nudges to reason are diminished in the case of *Momentum* (vis-à-vis public policy nudges to reason) by the fact that a person has intentionally granted permission to the app to influence her decision-making in this way.

reflecting in direct response to new experiences and genuine evidence, in order to assess how our beliefs and values relate to reality. This evidence-responsive critical reflection requires that we consider and revise our pro-attitudes wherever these are found to be called into question by relevant external factors, such as reliable evidence garnered by first-personal experience or from trustworthy third-party experts. By exploring what is known about relevant neurobiology, we have been able to suggest a neurobiological framework for evidence-responsive critical reflection. We have also deepened our understanding of the concept of undue influence, in particular in the realm of the sorts of everyday influences that we experience in virtue of being socially embedded. As part of this exploration we have developed the concept of pre-authorization, which suggests that the pedigree of information that might influence us has some bearing upon how we view such information – admitting it with relatively little skepticism or examining it more carefully. Not being subject to undue external influence on our decision-making processes tends to be viewed as an externalist condition for autonomy; but pre-authorization, with its role in determining who counts as the external actors whose influence is welcomed in our decision-making, represents a novel internalist aspect that is relevant to understanding when this condition has and has not met.

We brought both sets of insights together to analyse the ethics of influence, with a particular focus on nudging carried out by governments and by our increasingly technologically enriched environment. Taken together, these investigations add to the existing body of knowledge about autonomy and its discontents, recognizing our desire for control over our own decisions as well as helping us to better understand how we might preserve autonomy as socially embedded beings.

References

Anderson, Scott (2010). “The Enforcement Approach to Coercion,” *Journal of Ethics and Social Philosophy*, 5: 1–31.

Arneson R (1994) Autonomy and preference formation In Harm's Way: Essays in Honor of Joel Feinberg. Feinberg J, Coleman JL, Buchanan AE (eds). Cambridge University Press. pp. 42–75.

Bagnoli C1 (2011) Morality and the emotions. New York ; Oxford : Oxford University Press.

Bessi A, Zollo F, Del Vicario M, Puliga M, Scala A, Caldarelli G, Uzzi B, Quattrociocchi W (2016) Users Polarization on Facebook and Youtube Preis T, ed. PLoS ONE 11:e0159641.

Blöser C, Schöpf A, Willaschek M (2010) Autonomy, experience, and reflection. On a neglected aspect of personal autonomy. Ethic Theory Moral Prac 13:239–253.

Bitzer S, Hame P, Felix B, Stefan K (2014) Perceptual decision making: drift-diffusion model is equivalent to a Bayesian model. Frontiers in Human Neuroscience 8.

Bode S, Murawski C, Soon CS, Bode P, Stahl J, Smith PL (2014) Demystifying "free will": the role of contextual information and evidence accumulation for predictive brain activity. Neuroscience and Biobehavioral Reviews 47:636–645.

Buss S, Westlund AC (2018) Personal Autonomy. Stanford University:1–56 Available at: https://plato.stanford.edu/entries/personal-autonomy/.

Calo R (2014) Digital market manipulation. Stanford Technology Law Review 82:995–1051.

Castelo N, Reiner PB, Felsen G (2012) Balancing autonomy and decisional enhancement: an evidence-based approach. Am J Bioeth 12:30–31.

Chen-Wishart M (2006) Undue influence: vindicating relationships of influence. Current Legal Problems.

Christman J (2005) Autonomy, self-knowledge, and liberal legitimacy. Autonomy and the challenges to liberalism: new essays:330–358.

Christman J (2010) The politics of persons: individual autonomy and socio-historical selves. Cambridge: Cambridge University Press.

Christman J (2015) Autonomy in moral and political philosophy. Stanford University.

Christman J, Anderson J (2005) Autonomy and the challenges to liberalism: New essays (Christman J, Anderson J, eds). Cambridge University Press.

Clark A (2008) Supersizing the mind: embodiment, action, and cognitive extension. Oxford; New York : Oxford University Press.

Clayton M1 (2006) Justice and legitimacy in upbringing. Oxford; New York : Oxford University Press.

Coons C, Weber M (2013) Paternalism: Theory and practice. Cambridge University Press.

Coons C, Weber M (2014) Manipulation: theory and practice. Oxford University Press.

Dehaene S, Changeux J-P (2011) Experimental and theoretical approaches to conscious processing. Neuron 70:200–227.

Dehaene S, Kerszberg M, Changeux JP (1998) A neuronal model of a global workspace in effortful cognitive tasks. Proceedings of the National Academy of Sciences 95:14529–14534.

Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, Stanley HE, Quattrociocchi W (2016) The spreading of misinformation online. Proc Natl Acad Sci USA 113:554–559.

Doris JM (2002) Lack of Character. Cambridge University Press.

Doris JM (2015) Talking to Our Selves. OUP Oxford.

Dworkin G (1988) The Theory and Practice of Autonomy. Cambridge University Press.

Engelen B, Nys T (2019) Nudging and Autonomy: Analyzing and Alleviating the Worries. Rev Phil Psych 16:676.

Feinberg M, Willer R (2019) Moral reframing: A technique for effective and persuasive communication across political divides. Social and Personality Psychology Compass 13:29.

Felsen G, Castelo N, Reiner PB (2013) Decisional enhancement and autonomy: public attitudes towards overt and covert nudges. Judgment and Decision Making 8:202–213.

Felsen G, Reiner PB (2011) How the neuroscience of decision making informs our conception of autonomy. AJOB Neuroscience 2:3–14.

Felsen G, Reiner PB (2015) What can Neuroscience Contribute to the Debate Over Nudging? Rev Phil Psych 6:469–479.

Fitz NS, Reiner PB (2016) Perspective: Time to expand the mind. Nature 531:S9–S9.

Frankfurt HG (1971) Freedom of the Will and the Concept of a Person. The Journal of Philosophy 68:5–20.

Friedman M (2003) Autonomy and social relationships: Rethinking the feminist critique. In: *Autonomy, Gender, Politics* (Friedman M, ed), pp 81–97. Oxford: Oxford University Press.

Frischmann BM, Selinger E (2018) Re-engineering humanity. Cambridge University Press.

Gigerenzer G, Gaissmaier W (2011) Heuristic decision making. Annu Rev Psychol 62:451–482.

Gold JI, Shadlen MN (2007) The neural basis of decision making. Annu Rev Neurosci 30:535–574.

Goodwin T (2012) Why We Should Reject "Nudge." Politics 32:85–92.

Grune-Yanoff T (2012) Old wine in new casks: libertarian paternalism still violates liberal principles. Soc Choice Welf 38:635–645.

Gutchess A (2014) Plasticity of the aging brain: New directions in cognitive neuroscience. Science 346:579–582.

Hurley S (2011) The Public Ecology of Responsibility. In: Responsibility and Distributive Justice (Knight C, Stemplowska Z, eds), pp 187–217.

Jung JY, Mellers BA (2016) American attitudes toward nudges. Judgment and Decision Making 11:62–74.

Kahneman D (2011) Thinking, Fast and Slow. Farrar Straus Giroux.

Klinzing JG, Niethard N, Born J (2019) Mechanisms of systems memory consolidation during sleep. Nat Neurosci 35:1–13.

Knill DC, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci 27:712–719.

Kosinski M, Stillwell D, Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. Proc Natl Acad Sci USA 110:5802–5805.

Lavazza A (2016) Free Will and Neuroscience: From Explaining Freedom Away to New Ways of Operationalizing and Measuring It. Frontiers in Human Neuroscience 10:439.

Levy N (2017) Nudges in a post-truth world. Journal of Medical Ethics 43:495–500.

Mackenzie C, Stoljar N (2000) Relational Autonomy. Oxford University Press.

Matz SC, Kosinski M, Nave G, Stillwell DJ (2017) Psychological targeting as an effective approach to digital mass persuasion. Proceedings of the National Academy of Sciences 114:12714–12719.

May J, Kumar V (2018) Moral Reasoning and Emotion. In: The Routledge Handbook of Moral Epistemology, pp 139–156.

McKenna R (forthcoming) Persuasion and Epistemic Paternalism. In Guy Axtell & Amiel Bernal (eds.), *Epistemic Paternalism: Conceptions, Justifications, and Implications*. Rowman & Littlefield

Mele AR (1995) Autonomous Agents. Oxford University Press.

Mele AR (2012) Another scientific threat to free will? The Monist 95:422–440.

Meyers DT (1989) Self, society, and personal choice. New York : Columbia University Press.

Meyniel F, Dehaene S (2017) Brain networks for confidence weighting and hierarchical inference during probabilistic learning. Proc Natl Acad Sci USA 71:201615773.

Miller EK, Cohen J (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202.

Mulder MJ, Wagenmakers EJ, Ratcliff R, Beokel W, Forstmann BU (2012) Bias in the Brain: A Diffusion Model Analysis of Prior Probability and Potential Payoff. Journal of Neuroscience 32:2335-2343.

Nader K (2015) Reconsolidation and the Dynamic Nature of Memory. Cold Spring Harb Perspect Biol 7:1–16.

Nagel SK (2015) When Aid Is a Good Thing: Trusting Relationships as Autonomy Support in Health Care Settings. The American Journal of Bioethics 15:49–51.

Nagel SK, Reiner PB (2013) Autonomy support to foster individuals' flourishing. Am J Bioeth 13:36–37.

Niker F (2018) Policy-led virtue cultivation. In: The Theory and Practice of Virtue Education, 1st ed., pp 153–167. Routledge.

Niker F, Reiner PB, Felsen G (2016) Pre-Authorization: A Novel Decision-Making Heuristic That May Promote Autonomy. Am J Bioeth 16:27–29.

Niker F, Reiner PB, Felsen G (2018a) Perceptions of Undue Influence Shed Light on the Folk Conception of Autonomy. Front Psychology 9:57–11.

Niker F, Reiner PB, Felsen G (2018b) Updating our Selves: Synthesizing Philosophical and Neurobiological Perspectives on Incorporating New Information into our Worldview. Neuroethics 11:273–282.

Noggle R (2005) Autonomy and the paradox of self-creation: Infinite regresses, finite selves, and the limits of authenticity. In: New Essays on Personal Autonomy and its Role in Contemporary Moral Philosophy. Edited by James Stacey Taylor. Cambridge University Press.

Nyhan B, Reifler J (2013) Which Corrections Work? Research results and practice recommendations. New America Foundation Media Policy Initiative ….

Moles A (2015), "Nudging for Liberals", Social Theory and Practice, Vol. 41, No. 4 (October 2015): 644-667.

O'Neill O (2018) Linking Trust to Trustworthiness. International Journal of Philosophical Studies 26:1–8.

Ratcliff R, Rouder JN (2016) Modeling Response Times for Two-Choice Decisions. Psychological science : a journal of the American Psychological Society / APS 9:347–356.

Reeve J (1998) Autonomy Support as an Interpersonal Motivating Style: Is It Teachable? Contemporary Educational Psychology 23:312–330.

Reiner PB, Nagel SK (2017) Technologies of the Extended Mind: Defining the Issues. In: *Neuroethics: Anticipating the Future* (Illes J, ed), pp 111–126. Oxford University Press.

Ryan RM, Deci EL (2000) Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. Contemporary Educational Psychology 25:54–67.

Samanez-Larkin GR, Knutson B (2015) Decision making in the ageing brain: changes in affective and motivational circuits. Nat Rev Neurosci 16:278–289.

Shadlen MN, Roskies AL (2012) The neurobiology of decision-making and responsibility: reconciling mechanism and mindedness. Front Neurosci 6:56.

Simon H (1972) Theories of Bounded Rationality. In: Decision and Organization (McGuire CB, Radner R, eds). Amsterdam: North Holland.

Smith PL, Ratcliff R (2004) Psychology and neurobiology of simple decisions. Trends Neurosci 27:161–168.

Specker Sullivan L, Niker F (2018) Relational Autonomy, Paternalism, and Maternalism. Ethic Theory Moral Prac 25:1–19.

Specker Sullivan L, Reiner PB (2019) Digital Wellness and Persuasive Technologies. Philos Technol 37:76.

Squire LR (2004) Memory systems of the brain: a brief history and current perspective. Neurobiology of Learning and Memory 82:171–177.

Stanovich KE, West RF (2000) Individual differences in reasoning: Implications for the rationality debate? Behav Brain Sci 23:645–665.

Sunstein CR (2016) The ethics of influence: Government in the age of behavioral science.

Susser D, Roessler B, Nissenbaum HF (2018) Online Manipulation: Hidden Influences in a Digital World. Mich Telecomm & Tech L Rev.

Tannenbaum D, Fox CR, Rogers T (2017) On the misplaced politics of behavioural policy interventions. Nature Human Behaviour 1:0130.

Thaler RH, Sunstein CR (2009) Nudge. Penguin.

Tononi G, Cirelli C (2014) Sleep and the Price of Plasticity: From Synaptic and Cellular Homeostasis to Memory Consolidation and Integration. Neuron 81:12–34.

Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. Science 211:453–458.

Waldron J (2014) It’s All for Your Own Good. New York Review of Books:1–5.

Weimer S (2013) Evidence-Responsiveness and Autonomy. Ethic Theory Moral Practice 16:621-642.

Weimer S (2017) Evidence-Responsiveness and the Ongoing Autonomy of Treatment Preferences. HEC Forum 13:1–23.

Wertheimer A (2014) Coercion. Princeton University Press.

Wilkinson TM (2013) Nudging and Manipulation. Political Studies 61:341–355.

Williams J (2018) Stand Out of Our Light. https://wwwcambridgeorg/core/product/3F8D7BA2C0FE3A7126A4D9B73A89415D:1–152.

Yeung K (2012) Nudge as Fudge. The Modern Law Review 75:122–148.

Yeung K (2016) “Hypernudge”: Big Data as a mode of regulation by design. The Information Society 20:118–136.