2 Minimal Liberal Autonomy

The primary aim of this chapter is to construct a minimal theory of autonomy that is maximally inclusive of intuitions and theories about autonomy. A subsidiary aim is to formulate such a theory in the language of psychology. If I were to meet these aims, then I would have rendered plausible autonomy-minded liberalism and have at my disposal empirically precise standards by which to assess autonomy violation in branding. In order to arrive at such a theory we must know at least two things: we must know what requirements such a theory must meet (i.e. we must know its design specifications) and we must know which arguments can move us beyond the current controversies and make us meet these requirements (i.e. we must have an argumentative strategy).

With regard to the design specifications I propose that three uncontroversial conditions must be met: maximal intuitive plausibility, consistency and realism. The first principle (which was introduced in chapter 1) simply states that the theory that is able to incorporate most intuitions – provided subsequent conditions are met – is the best theory. The additional two specifications are limiting principles on the expansive drive of the first. The second principle of consistency also hardly requires elaboration. Intuitions about autonomy often clash and at least a political theory must explain with which intuition to side in prototypical situations. Political theory must give such explanations because these generate political legitimacy, and the explanations that are best at generating legitimacy are (at least minimally) rational ones (and hence consistent ones). The third principle of realism stipulates that the elements of a minimal theory of autonomy must be at least compatible with what we know about the workings of the human mind. There is an important reason for this requirement in addition to my demand for empirically precise standard for the analysis of branding. Because our intuitions about autonomy have not been shaped on the basis of a solid understanding of psychology, they may make demands on us that we cannot actually meet. A minimal liberal theory of autonomy must reject such intuitions because it makes no sense for the state to protect or promote mental states that citizens cannot attain. In order to avoid such a result, a minimal theory of autonomy must be conservative with regard to the principles it accepts and require evidence from psychology for thinking that they are realistic.

My argumentative strategy is to first select from among the intuitions about autonomy the ones that are minimally plausible and relevant to liberal autonomy. Where possible, I integrate intuitions, so that I end up with a set of seven core intuitions (section 1). Because these intuitions are interrelated, assessing which theories of autonomy are best in realistically and consistently making sense of them should preferably be done in some higher-order context. For this purpose I propose that the seven intuitions coalesce into two clusters. The first consists of agentic properties that make no reference to a self. I call this set self-control (where this term is understood in the way that Daniel Dennett uses it, so that the word ‘self’ does not refer to an element that causes anything) and discuss how they fit together in sec-
tion 2. The second cluster of intuitions goes under the header of authenticity and adds to the framework of self-control intuitions about the self (section 3). In the final section I discuss a number of factors that moderate the need for liberal state action in defending autonomy and so establish minimal practical thresholds that are in line with intuitions about appropriate political action.

1 Intuitions about Autonomy

In order to find a common ground among conceptions of autonomy it is important that all relevant intuitions and theories are considered. Otherwise my selection might be biased and that would undermine the credibility of the guarantee of anti-perfectionism that this chapter aims to offer. Because the philosophical literature on autonomy is substantial, it is most fortunate that I have been able to rely for this purpose on earlier surveys of the literature. From these surveys one can deduce two types of lists. One can make a list of different approaches to autonomy and one can make a list of the intuitions of what autonomy is about.

With regard to the first list, discussions about the correct approach focus on whether the history of mental states matters for autonomy and whether autonomy is a local or a global phenomenon. A number of authors maintain that the history of mental states is irrelevant insofar as this history does not strengthen capacities that presently advance the autonomy of an agent. However, this position is hardly tenable in light of an intuition that has been brought out by Alfred Mele in a thought experiment. Imagine two persons who have exactly the same psychological make-up so that they are just as open-minded, flexible and rational. Let us call them Ann and Beth. Suppose that Ann is autonomous, whatever that may turn out to be. Suppose that Beth is different in that she has been subjected to brainwashing so that her entire memory system has been wiped out and replaced by a new coherent set (which – remember – is identical to Ann’s). If this were the case, then we would clearly not think of Beth as autonomous. And because her psychological states are identical to those of Ann, the only difference that can drive this intuition is the difference in the history of the states in question.

Another disagreement concerns the relation between local autonomy, which is a property of mental states and actions, and global autonomy, which relates to life as a whole. Some think that the autonomy of a life can be known simply by adding up instances of local autonomy. Others disagree and maintain that the autonomy of a life depends on it being lived according to life-guiding values that have been endorsed (in an autonomy-conferring fashion). Fortunately, I can sidestep this controversy because taking a side on this issue does not make a practical difference from the point of view of the liberal state. That is, because – as I will argue later on in this chapter – the liberal state cannot (ethically) keep track of whether people live in accordance with properly endorsed values, it can only foster the critical skills that citizens need for evaluating their values. And the same skills (which on most accounts involve some type of reflection) are required for evaluating values on the local and the global account of autonomy.

The second list concerns conceptions of autonomy. Some of these are clearly not relevant to liberal autonomy and include obscure perfectionist conceptions of autonomy, non-liberal conceptions of political autonomy and conceptions of moral autonomy. With regard
to the remaining conceptions it can be established that they are driven by seven core intuitions which cannot be reduced to each other. The first intuition is about the input into the reasoning process and requires that the agent has accurate beliefs about the world. Intuitions two to four are about the reasoning process itself and concern the point of view of the agent, the type of reasoning and its coherence with self-related states. Distinct from these intuitions is the notion of rationality, which is (typically) a standard by which the output of such reasoning is evaluated. Whether this output causes behavior depends, and this is the sixth intuition, on impulse control (i.e. the opposite of weakness of will). Finally, in all of these respects the agent can be more or less autonomous depending on her independence from the interference by other agents.

If it is true that these intuitions are relevant to liberal autonomy as well as not reducible to each other – which I take to be uncontroversial – it can still be objected that some important intuitions are missing. The most prominent such omissions – espoused mostly by feminist philosophers – are intuitions which hold that autonomy is about having certain intellectual and physical skills, self-perceptions and positions in social relations. There are two ways in which these elements can be said to be relevant to liberal autonomy. The first is uncontroversial and concerns the relevance of skills, self-esteem and social relations for the exercise of any one of the seven components of autonomy that I have outlined above. It can for example be argued that certain intellectual skills are necessary to assess options, which can be argued to be a necessary component of the type of reasoning that confers autonomy. Similarly, it can be maintained that endorsement (as a type of reasoning or as coherence) only occurs when the agent has a certain level of self-esteem, so that such self-esteem is derivatively relevant to autonomy. And because self-esteem is contingent on certain social relations, it can by extension be argued that these social relations are a condition for the development of a condition for autonomy (i.e. self-esteem) and in that sense relevant to autonomy. I accept all of this, but because theorists who push this line of argument have yet to connect psychological work on self-esteem and skills to empirical thresholds of components of autonomy I am not able to rely on this work in the remainder of my argument.

The other and more radical case for the importance of skills, self-esteem and positions in social relations to autonomy maintains that these phenomena are constitutive of autonomy. Even if we suppose that this position is intuitively plausible, which is very debatable, it is not well suited to support a theory of liberal autonomy, since there is no non-perfectionist way to determine which skills, level of self-esteem or positions in social relations are conducive to autonomy. If one wanted to do this, one would either have to outline a set of goals that foster autonomy and then determine which skills, level of self-esteem and positions in social relations facilitate the realization of these goals. But such a move would introduce first-order values (for the selection of the goals) into the theory of liberal autonomy, which is irreconcilable with anti-perfectionism. Or one would have to demand that an autonomous agent must possess all the skills and self-esteem, and be in all of the types of social relations that the proficient pursuit of autonomously chosen goals requires. But such a conception is surely absurdly demanding. If these approaches are indeed both implausible (and exhaustive), then having skills, self-esteem and positions in social relations cannot – realistically speaking – be constitutive of liberal autonomy. Because these phenomena are not only hard to define in terms of thresholds but also clearly something different from psychological autonomy I will call them freedoms. I agree that on the basis of a minimal theory of autonomy a good case can be made for the protection of such freedoms in terms of rights and the distribution of freedoms in terms of primary social goods. But that is not to say that having
more freedoms increases one’s autonomy (once we disregard their effect on fostering the seven components of autonomy).

If it is agreed that the seven intuitions on my list are relevant and irreducible while it is also granted that the list is comprehensive, then the hard questions are which conceptions best make sense of the intuitions and how intuitions relate to each other. I take it that these questions cannot be answered independently. That is, if it is true that the meaning of intuitions is contingent on their relation to other intuitions, then one should discuss conceptions of intuitions in the context of an integrative framework that at least roughly indicates how intuitions fit together. In this regard I follow John Christman in thinking that attempts at such integration have so far predominantly taken one of two forms. Some think of autonomy as a special type of control without seeing any role for the self. In terms of my taxonomy, autonomy as control is made up of a cluster of four core intuitions: epistemic accuracy, type of reasoning, rationality and impulse control. Autonomy thus conceived involves flexibly and effectively responding to the world on the basis of accurate knowledge. Daniel Dennett has offered an influential outline of such an account and I will follow this outline as well as his terminology in my discussion of autonomy as self-control (see section 2). Others have conceived of autonomy as a special relation to the self on the basis of the two intuitions coherence and point of view. I refer to this cluster as authenticity and argue that it is a legitimate part of liberal autonomy if we conceive of it much more minimally than is common (see section 3). The two types of autonomy imply different conceptions of independence. On the self-control account, independence refers to the absence of another agent taking over control of behavior by interfering with any one component of self-control. On the authenticity account, independence refers to the absence of interference in the formation of a point of view or the process of maintaining coherence.

I need to add one more aspect to this framework. If it is plausible that a person who is hospitalized and unable to communicate can still be autonomous, as Alfred Mele argues, then autonomy must be about more than just actions. Hence, I am going to say that autonomy can be a property of mental states as well as a property of an action. But what about influence on mental states which in turn issue in action? In this regard I will say that influence affects autonomy as a property of mental states insofar as the changed mental states are stored in memory before issuing in an action, and that influence affects autonomy as a property of actions insofar as the changed mental states issue in an action before being stored in memory (if at all). On the basis of these distinctions my aim in the upcoming two sections will be to find a minimal common ground on self-controlled actions (2.1) and mental states (2.2) as well as authentic actions (3.1) and mental states (3.2).

2 Self-control

A system that controls something must meet three minimal conditions. It must have a desire about the state into which the controlee is to be driven. It must have accurate relevant knowledge of the world. And it must be able to deliberate, so that the desire and the beliefs can be adjusted. By corollary, if a system drives itself into a different state, we may speak of it exercising self-control. And were we to ascribe subsequent learning capacities and higher order thinking to it, then we would have something approaching the self-control that humans typically exercise. There is of course the possibility that other agents take over control, but as
long as this influence is known by the influenced agent and enters into the deliberation process, self-control is preserved. This, at least, is the analysis of Daniel Dennett. Pace skeptics on the prospects of agreement on a minimal theory of autonomy I will argue that such agreement is possible for self-control, as understood by Dennett, in terms of action (2.1) and attitude formation (2.2). I will come to disagree however with the view, espoused by Dennett as well as Bernard Berofsky, that self-control (if sufficiently well-informed, open-minded and flexible) would be all there is to autonomy (2.3).

2.1 Self-controlled Action
A liberal account of self-controlled action must specify four intuitions. It must specify the type and amount of reasoning (2.1.1), the rationality of this reasoning or its outcome (2.1.2), strength of will (2.1.3) and the independence from other agents (2.1.4). In the context of this thesis I cannot hope to give an exhaustive account of all these properties of self-controlled agents. I will therefore bring the criteria of assessment (as stipulated by my design specification) to bear only on intuitions with regard to which there exists disagreement that threatens the claim to non-perfectionism of a minimal theory of liberal autonomy (e.g. responsiveness to reasoning), or ones that are prima facie relevant to the assessment of branding and thus require empirical specification (e.g. independence).

2.1.1 Responsiveness to Reasoning
Although all theories of autonomy agree on the importance of reasoning, there is disagreement on the type and amount of reasoning that is to be required. Obviously, the amount of reasoning that a minimal theory can require will depend on what the reasoning should look like. In this regard two theories give very different accounts of both the mode of reasoning and its object. On the evaluative account, reasoning must involve normative judgment and is typically about (but need not be restricted to) courses of action. On the hierarchical account, reasoning need not be normative, but its object must be desires (courses of action will not do). Because a liberal theory of autonomy must meet demanding standards of realism, I will assess both approaches with an eye mainly to their psychological feasibility. In this regard it will prove instructive to point out in advance that there is a growing consensus among psychologists for making a distinction between system 1 and system 2 processes. System 1 processes are for the most part unconscious, show little individual variability and operate on the basis of parallel processing, which allows them to incorporate a lot of information and to process it fast. The system 2 process is sequential, conscious, slow, rule-based and accurate. Moreover, whereas system 2 is thought to consist of a single system, system 1 is likely made up of many systems. I am going to assume from here on that some distinction of this sort is correct, and that what theorists of autonomy are concerned with when they demand that people reason and reflect is that they use system 2.

The Evaluative Account of Reasoning and Autonomy
The most influential account of the idea that autonomy is conferred by evaluative reasoning has been developed by Gary Watson. He claims that autonomous action results from judgments about what it is good to do, all things considered. Such judgments are the result of reasoning that combines beliefs (and probability estimates) with considerations about the good (the latter set of considerations makes up what Watson calls the ‘valuational system’). Actions that lack autonomy are ones that are caused by motivational states that do not result
from reasoning or do not result from reasoning that relies on input from the valuational system. These three categories – unreflective motivations, non-normative reflective motivations and normative reflective motivations – together make up the 'motivational system'. An agent is perfectly autonomous if the valuational and the motivational system are identical. 

There are three types of incrementally serious objections to this account that are presently relevant. The first concerns Watson’s assumption that the output of normative system 2 reasoning is motivational. In this regard he explicitly sides with Plato (and against Hume) in assuming that when we think that something is good we thereby desire to do it. However, there are reasons to doubt such a view in light of evidence from patients with intact system 2 reasoning skills but damage to parts of the brain that are involved in processing emotions. It turns out that there are among such patients individuals who cannot settle on what to do in the real world (such as deciding which pencil to pick), whereas they have no problem in answering hypothetical questions about what it would be best to do. This suggests that affect plays a decisive role in practical reasoning. On the basis of such evidence several models of practical reasoning have been constructed that consist of system 1 processes that send signals of affect or pleasure (depending on the model) which are monitored unconsciously by the practical reasoning system. If this picture is correct, then acting on the basis of evaluative reflection might look something like this. A thought arises in the conscious system 2 reasoning system with the content ‘I shall do x’ (e.g. choose the blue pencil). This globally broadcast inner speech message is then picked up by a specialized module, which infers that a commitment is made to do x. Combined with a desire to act consistent with commitments, the belief that the person committed herself to doing x issues in a motivational state that may lead to the selection of a motor instruction in the practical reasoning system (if it outcompetes other motivations). If this is a credible type of account of practical reasoning, then Watson would have to adapt his theory in such a way that autonomous actions are indirectly caused by best judgments (which are beliefs) that are the output of the system 2 reasoning system. That is, evaluative reflection can influence the selection of an action scheme, but it can do so only if it is paired with a desire.

A second objection against Watson’s account, raised by Michael Bratman, is that evaluative judgments underdetermine what is to be done. Such underdetermination applies to cases in which two options are equally valuable (and where there is no value that would enable a comparison) and cases in which the agent is uncertain about what value to ascribe to a course of action. Furthermore, if the definition of an evaluative judgment is said to depend on some intersubjective standard, then uncertainty mushrooms, because there are many cases in which it will be uncertain whether rational agents will agree on the value to be given to a course of action. Bratman claims that in such situations people still act on the basis of reflection – in which they draw not on value judgments, but on feelings and inclinations – and that they do so in a way that we think of as autonomous. If so, then autonomy would be about more than just evaluative reflection.

Bernard Berofsky has gone yet further in claiming that we can think of an entire life in which people act against their judgment of what it is good to do and yet say that such a life is lived autonomously. This alternative conception sees values not as the result of reflection on what it would be good to do, but as stable preferences that reveal themselves in a wide range of (hypothetical) choices that are made in an informed and dispassionate way. A person may for example find herself spending most of the time gambling because of the thrill it gives her and the interesting people she meets. She may think her life is not up to her ethical standards, that it is not in line with her ideal self and perhaps even that it is utterly
worthless. Although such a situation would surely lead to a lot of conflict, it need according to Berofsky not make her gambling life any less autonomous. If this is plausible, and I think it is, then it would be wrong to tell such people that they are not living an autonomous life. This is not to deny that autonomous persons have the capacity for reflection on the good, or that they sometimes use this capacity, but it is to deny that they must live their life in accordance with such judgments. I take it then that on a liberal conception of autonomy there should be no requirement that people act on the basis of reflections that draw on an idea of the good.

The Hierarchical Account of Reasoning and Autonomy
Hierarchical theories maintain that autonomy consists in a hierarchy between two conative states that is established in reflection (e.g. an agent reflectively comes to desire to have a desire). The intuitive impetus that drives such theories is the notion that it is crucial for autonomy to reflectively step back from a situation and to reflect on it. In hierarchical theories this takes the form of a person desiring, upon reflection, to be driven by one or more of her desires. When she wants such second-order desires to be effective, then she is said – in the classical formulation by Harry Frankfurt – to have second-order volitions.

A difficulty that Frankfurt notes himself is that in case of conflict the higher-order desires are to win out, and that there is thus a danger that a chain of ever higher-order desires may lead to “the destruction of the person”. This chain can be terminated though, Frankfurt maintains, when the person “identifies himself decisively with one of his first-order desires” whereupon “this commitment ‘resounds’ throughout the potentially endless array of higher orders”. This is anything but a satisfying solution though, because it is not clear what such decisive identification consists in. But if we suppose with Gerald Dworkin that for all practical intents and purposes people do not carry on reflecting at ever higher levels we can set aside the matter for the moment. A more pertinent difficulty is that it is not clear from Frankfurt’s account whether hierarchy does any autonomy-conferring work. An alternative explanation would be that it is the reflective endorsement (with may or may not involve conative hierarchy) which pumps the intuition.

In one response to this alternative explanation, Bratman points to underdetermination of evaluative judgments. In such cases, to repeat, agents may come to a proper reflective decision that does not rely on evaluative judgments. But although these cases may show that a non-evaluative commitment has a role to play in a theory of autonomy, they do not show that such a commitment should be hierarchical. Another strategy of Bratman is to show that there are cases of higher-order ‘identification’ with a desire that have an autonomy-conferring effect which can only be explained with reference to hierarchy. He conceded that proper identification with a desire involves treating it as a justified reason for action. But according to Bratman there is a subset of cases in which autonomously treating a desire as a reason for action can only be explained with reference to hierarchy. In support of this claim he notes two hierarchical accounts of treating a desire as a reason that do not work, and presents a middle way as the solution.

One dysfunctional account says that a policy must support treating a desire as a justified reason in deliberation. But then the identification appears to occur at the first-order level in the desire being treated as a justifying reason in deliberation. Hence, higher-order attitude support does not seem to add anything. If so, then hierarchy does not add anything. The other dysfunctional account says that the second-order attitude must support the desire’s functioning as an effective motive. But then the desire issues in an action (due to the support by a higher-order attitude) without being treated as a justified reason.
As a middle way Bratman rolls out an ingenious argument to the effect that there can be cases in which we engage in attenuated reasoning and where higher-order attitudes can give the decisive little push. As an example he mentions the behavior of participants in the Milgram experiments. These subjects went along with authority to give electric shocks to a supposed examinee and are assumed by Bratman to have not fully endorsed their action. In such cases, he maintains, the agent engages in practical reasoning and comes to see the desire for something as a “justifying end” without endorsing that the desire function in this way. At this point Bratman introduces his higher-order attitudes to suggest that the gap between treating a desire as end-setting for practical reasoning and true endorsement of it as a justifying reason would be bridged by a higher-order attitude in support of the desire functioning in this way (i.e. as end-setting for practical reasoning). In the Milgram example this would mean that a participant who decides to push the electric shock button after a process of attenuated deliberation that treats the desire to punish as a justified end (simply because she goes along with authority) would not yet be autonomous. But the action would be autonomous if she had a self-governing policy that treated the desire to punish as a justified end in deliberation. In this way the agent still does some deliberation (unlike the case in which the higher-order attitude merely supports a desire functioning as an effective motive), while the higher-order attitude does seem to add something (i.e. treating the desire as a legitimate reason).

However, what this argument leaves open is the possibility that what is driving the intuition is that autonomy depends on reflection, and that the higher-order policy support drives the intuition by virtue of that policy being formed on the basis of reflection. On this interpretation, the fact that the relation between the reflection and the desire is temporally mediated by a higher-order attitude would be accidental. In sum, the absence of evaluative reflection in intuitive cases of autonomy does not force us to accept hierarchy, and cases of autonomy in which hierarchy is involved do not rule out the possibility that evaluative reasoning is pumping the intuition.

What is more, even if we grant that hierarchy does somehow confer autonomy it would seem, as Watson noted, that reflection on desires is too rare a phenomenon to be the basis of autonomy. If we only rarely reflect on our desires, then we are only rarely autonomous. At least for a theory of liberal autonomy, that will not do. In response to this concern, Bratman proposes to think of higher-order states in hierarchical theory as plans and policies rather than desires. Plans are generalized intentions that can settle courses of action and coordinate them over time (e.g. keep my teeth clean), while policies specify what such plans amount to in practice (e.g. brush my teeth after breakfast and before going to bed). According to Bratman, policies not only settle actions, but also settle which desires should be considered justified reasons for action in practical reasoning and roughly fix the weight that is to be given to desires in practical reasoning. Such policies are called self-governing policies. In this capacity, self-governing policies (which have the relevant hierarchical structure) co-regulate which desires activate motor instructions. Because such policies operate in the background, there is no need to reflect on every desire that moves us to act. Hence, the fact that we only rarely reflect on desires must not count against our lives being autonomous.

However, we earlier considered evidence for thinking that the weighing of options involved in decision-making occurs automatically and is thus not under conscious control. Hence, it is unlikely that policies can fix how much weight is to be given to a desire or how a desire is to be treated (i.e. as a justified reason) in affective practical reasoning. This is not to say that policies in general can have no influence on practical reasoning. They can have influence if they are about a course of action, and are supported by an innate desire to do what one has committed oneself to doing (by way of forming the policy). But if policies are
about courses of actions, then they do not support a theory of conative hierarchy. It is conceivable that there are self-governing policies that inform normative system 2 reasoning about which desires can count as justifying considerations. And system 2 judgments about courses of actions can in turn – when coupled with an innate desire to do what it is best to do – influence the practical reasoning system. But if this is how self-governing policies work, then the objection that they do not have much influence seems to stand.

In response to this concern it must be considered that Bratman has noted two structural pressures that supposedly drive agents in the direction of relying on second-order desires. One is the need for self-management, understood as the need to suppress desires that compete with reflective judgments for activating motor instructions. According to Bratman, the fact of such competing desires generates pressure to reinforce the (presumed) conative force of reflective judgments with higher-order desires in support of the reflective judgment. The way this is thought to occur is, again, by means of self-governing policies that contain the instruction to “give (refuse) justifying significance to consideration $x$ in motivationally effective practical reasoning, in part by giving (refusing) such significance to relevant first-order desires”. But I just argued that such policies can only influence system 2 reasoning, which is not itself motivationally effective. If that is true, then hierarchical policies apparently do not help in solving self-management problems, because these problems seem to be located not within system 2 (i.e. the formation of the judgment of what it is best to do) but between system 2 and practical reasoning (i.e. ensuring that the best judgment activates a motor instruction).

A second pressure for hierarchy is supposed to be exerted by the underdetermination of value judgments that we discussed earlier. However, we saw that such cases only demonstrate that there is pressure to make some commitment in favor of a course of action (that can be said to speak for the agent) but that underdetermination exerts no pressure on such commitments for being hierarchical ones. Taken together then, the evidence for thinking that conative hierarchy is a common feature of human action is not convincing.

Absent decisive arguments for the evaluative or hierarchical accounts of autonomy, what remains is the strong intuition that autonomy involves system 2 reasoning that produces a belief about what it is best to do (which is different from a belief about what it is to act in accordance with the good). Although this intuition is uncontroversial, it is not clear how much of it an autonomous agent should display. Given the fact that much of our everyday behavior consist of habits, and given the indirect nature of system 2 influence on behavior, it would seem that the role of reflection in a theory of autonomy should be modest. In this spirit Ben Colburn proposes to say, as we saw earlier, that a person is autonomous if she would still have carried out an action had she reflected on it. The problem with this approach – implicit in my earlier remarks about this idea – is that there is no way for the liberal state to know whether people would have reflectively endorsed an action of theirs. As a result, it is impossible for the state to accurately foster or protect autonomy, since it cannot measure either progress or transgressions of thresholds.

If it is agreed that some reflection matters for liberal autonomy, but that it is impossible to determine a minimal amount of reflection (e.g. in the form of hypothetical reflection), then the best we can hope for may be to require that people have developed the capacity to reflect. This is the approach advocated by Gerald Dworkin. In practical terms this means that there is a role for the state to foster via education basic reasoning skills, but that monitoring the amount of reflection citizens engage in is (absent violations of independence) not the business of the liberal state.
2.1.2 Rationality
I can be very brief with regard to autonomy and rationality. There is wide agreement that Kantian views of autonomy as rationality are too controversial to serve as conceptions of liberal autonomy and there is no need for me to reiterate these discussions. Besides concerns about rationalistic perfectionism there are worries about abuse of state power in the name of rationality. This was the point of Berlin’s tyranny argument discussed in chapter 1.1. In line with such worries it was noted with Christman that liberals are reluctant to employ anything but the most minimal standards of rationality in fleshing out concepts of autonomy. These minimal standards include some level of coherence of desires (understood negatively as the absence of clinically defined levels of incoherence) and the capacity for basic means-end reasoning that allow a person to pursue goals.53

2.1.3 Strength of Will
Whether there can in principle be such a thing as weakness of will has traditionally puzzled many philosophers. Apparently there are cases in which we judge that it would be best to do x but in the end fail to actually do x. And this is an everyday experience rather than an exception: we think it best to write a section but find ourselves surfing the internet. This is a puzzling state of affairs if we assume that a best judgment always settles a decision. However, there is no good reason to think that this must be so. One plausible explanation of what is going on is that system 2 reasoning does not settle what is to be done, but that such decisions are made by system 1 processes. If so, then the system 2 conclusion that it is best to do x must be tagged with a – possibly innate – desire to do what is best, and this desire may lose out against other desires in the practical reasoning process.54 On this account, weakness of will occurs when the desire to do what is judged best loses out against rival desires.55

If this is somewhat plausible, then the serious question is not whether weakness of will occurs, but how much of it a self-controlled person can be allowed to display.56 In this regard Christopher Suhler and Patricia Churchland have offered a neurobiological account of self-control, which consists of two components.57 The first is anatomical and concerns the unimpeded functioning of areas of the brain involved in “prototypical” cases of “good control”.58 In this context the authors refer to the central role of the prefrontal cortex in maintaining goal-related patterns of activity and sending biasing signals to several parts of the brain to bring the goal in question about. The second component is about neurochemicals, which must be “calibrated to the spectrum of values that the brain evolved to maintain in response to environmental demands typical of the species’ evolutionary past”.59 If serotonin levels are low, for example, people are more likely to engage in aggressive and impulsive behavior, and if there is a delay in the return to baseline of the corticotrophin releasing factor, recidivism of addicts is more likely.60 If we assume that our evolutionary past was stable and relatively uniform (which is not uncontroversial), this approach should potentially be able to offer the kind of verifiable criteria for self-control that a minimal theory of autonomy demands.61 Specifying such criteria is certainly not a task for the present occasion however.

2.1.4 Independence
Self-control requires that our actions are not determined by others outside of our control.62 In this section I discuss interference insofar as controllers influence behavior directly (influence over behavior that is mediated by attitude change is discussed in the next section). Controllers can do so in three ways. They can activate a goal outside of awareness, they can
disturb a controlee’s reasoning process and they can deplete her resources for exercising self-control.

**Unconscious Goal and Trait Activation**
Activating a goal or a trait outside of awareness can occur at two levels. In subliminal influence the stimulus itself that activates the goal or trait is not perceived, whereas in instances of supraliminal priming only the process by which a presented stimulus influences behavior is not perceived. The first route – of subliminal influence – is the clearest method of taking control outside of awareness. The acceptance of this phenomenon relies on the acceptance of a distinction between an objective threshold for perception (i.e. when a stimulus is discriminated by the senses) and a subjective threshold (i.e. when subjects report awareness of perceiving the stimulus). One can of course say that all subjective thresholds are unreliable and so define away the possibility of subliminal influence. But at least for a normative theory of self-control reasoning about influence attempts is relevant, and such reasoning must have access to information that is consciously accessible. If one does not know that another agent has exerted influence (because it did not pass the subjective threshold), then one can also not take the intention of the agent into account. Subliminal priming procedures have been shown to affect drinking behavior, as in an experiment conducted by Ap Dijksterhuis and colleagues. When they flashed the words ‘drink’ or ‘cola’ for 15 milliseconds and then immediately masked the words with a string of letters (as part of an unrelated task) subjects were significantly more likely to accept the offer of a beverage by the experimenter during a pause in the experiment (than subjects in a control condition in which random letters had been subliminally presented).

The best evidence for influence by means of supraliminal priming comes from experiments in which subjects are presented with scrambled sentence tasks in which they are to make correct sentences out of a collection of words. John Bargh and colleagues in particular have achieved remarkable results by using this method. In one experiment some of the words were unbeknownst to the subjects related to aging (e.g. bingo, wrinkles and Florida) while subjects likewise did not know that when the experiment was supposedly over, the experimenters measured the time it took subjects to walk from the experiment room to the one of the conductor of the experiment. It turned out that it took those who had been primed with old age on average well over eight seconds, while the control group needed only seven. In a similar experiment the words in the scrambled sentence test were related either to rudeness or politeness. The response was now measured in terms of the time it took subjects, after they had completing the task, to intervene in a staged conversation between the experimenter and a supposedly different subject, who – if no interruption followed – kept on talking for ten minutes after the subject had finished her task. Of those primed with rudeness 63% interrupted, against 38% for the control group and 17% for those who had been primed with politeness. In both experiments subjects showed no awareness whatsoever on the relation between the content of the words in the scrambled sentence task and their behavior. Hence subjects who are primed outside of awareness are also unable to take the influence of the controller into account and must thus count as deficient in self-control.

**Reasoning Disruption**
Once a message is identified and scrutinized for its accuracy, a self-controlled agent must be able to reason about it. Such reasoning sets a self-controlled agent apart from a creature that has only reflexes. However, reasoning can be disrupted by a controller if she manages to put cognitive load on the controlee’s mental system and press for a quick decision. Putting
cognitive load on a system is achieved by dispersing cognitive resources by means of activating diversionary reasoning. If a decision is made under such conditions, an agent is likely to rely on simple heuristics, such as following authoritative persons or people that are likable.

Fortunately, we have a defense mechanism against such influence attempts in the form of a mechanism that is known as reactance. This phenomenon is conceived of as a system that warns us when we are pressured to make a quick decision (this at least corresponds to one cluster found in questionnaires that attempt to measure the phenomenon). The cognitive underpinnings of the system are unclear, but there is some evidence to suggest that two routes may be involved. In one the feeling that freedom is threatened leads to a motivation to reject a message directly, and in the other the feeling that freedom is threatened evokes the motivation to counter argue, so that the message may be rejected indirectly. In any case, there is significant individual variability with regard to the threshold for the emergence of reactance, and evidence is as of yet not solid enough to make a case for the level that can reasonably be required in standard types of situations. Moreover, feelings of reactance can be artificially repressed in an agent while control efforts are undertaken all the same. The latter process is also poorly understood however, so that it is too early days to indicate both a minimal threshold of reactance as well ethical prohibitions against manipulating it.

Weakness of Will Induction

A third way in which a controller can wield direct influence over a controlee is by depleting self-control resources. Thus, if a controller has the goal of preventing a controlee from engaging in a particular action that the controller knows is caused by a best judgment, she may intentionally deplete the self-control resource of the controlee. If the controlee were to have no awareness of the workings of ego depletion or the intentions of the controller, then her self-control (understood broadly) declines. Likewise, if a controller would know how to inhibit the release of serotonin and so trigger weakness of will the self-control of the controlee would decline.

2.2 Self-controlled Mental States

With regard to self-controlled mental states we need to consider two intuitions. One tells us that autonomy is about epistemic competence and the other sees autonomy as independence from other agents. I can deal quickly with the epistemic competence intuition. On some accounts – most notably on the objectivity account developed by Bernard Berofsky – the level of autonomy is contingent on the extent to which beliefs on which an agent acts are accurate representations of the world. Such accuracy allows an agent to interact effectively and flexibly with the world around her, and this would be what autonomy centrally consists in. For my purposes I can sidestep controversies about this intuition, because there is a consensus with regard to liberal autonomy to the effect that the state should in principle not intervene in the lives of citizens because it believes that citizens do not see the truth. Surely, it is agreed that self-controlled agents must act on the basis of true beliefs to the extent that they are able to act with some minimum of effectiveness, but this demand is understood negatively as the absence of severe clinical conditions of delusion.

In the remainder of this section I will therefore be concerned with the intuition that autonomy is about independence. Specifically, self-controlled mental state formation is impaired when it results from processing in which the intention of a controller is not taken into account (and when the mental state at a later time causes the controlee to behave in a manner intended by the controller). Knowing about the intentions of a controller preserves
self-control if this information serves as input into the practical reasoning system – provided it is not disturbed – and enables the agent to reach a relevant judgment (which may be either to accept the influence or to resist it). Thus there are two stages at which self-control of mental states is secured in the face of influence attempts. The self-controlled agent must first recognize the intention behind a message of a controller and then be able to properly process it. In order to bring out intuitions on what is disturbing about the loss of self-control, philosophers have so far discussed thought experiments that involve grotesque cases of brainwashing. A relevant criticism of this literature is that concerns about manipulation are only practically relevant insofar as it is possible (and the effects are significant). I therefore want to give empirical evidence for thinking that controllers can realistically devise messages in such a way that they cannot be properly recognized (2.2.1) and processed (2.2.2) by controlees.

2.2.1 Recognizing Intentions
The simplest way to take control via mental state formation is by changing beliefs. Giving true information to another agent violates self-control if the agent does not take the intention behind the communication into account (and if the processing of the information has a behavioral effect that the controller intended). Such a loss of self-control is more likely if the information is false, because in that case the controlee is less likely to correctly interpret the intention behind the communication (at least if the controlee expected to be told the truth; when confronted with default liars, the reverse would be true). However, if the liberal state is to establish whether an intention to manipulate exists on the part of a controller and whether the controlee could have been expected to recognize this intention, it must have criteria that allow it to determine ethically whether these conditions are met. As I will argue in section 4, establishing whether citizens are misleading each other involves invasions of privacy that are unacceptable from the point of view of liberal autonomy itself. In this regard, properties of messages that induce a type of stimulus processing that evades the recognition of an intention will count as more appropriate. Below I will do three things. I want to give evidence for the existence of such message properties and processing channels; I want to show that stimuli processed in this way can lead to behavior; and I want to demonstrate that such influence cannot easily be corrected.

Unconscious Influence
We have already seen that it is possible to activate a goal directly by means of priming outside of an agent’s awareness. Whether mental states that are stored in long term memory can also be changed outside of awareness is a separate question. For a start, the brainwashing experiments that figure so prominently in philosophical discussions of manipulation are probably not possible. That is, although psychiatrists have in the 1950s reportedly succeeded in wiping out the memories of patients, attempts to instill new attitudes in them by headphones have failed miserably (so that the unfortunate subject had to acquire even their most basic skills all over again). However, there is evidence for less spectacular unconscious attitude change by means of evaluative conditioning (which has been documented when stimuli are presented both supraliminally and subliminally) and non-evaluative covariation learning. Because proving the existence of such processing is crucial to my overall argument, I must ask readers to endure a rather dry and technical excursion into psychology.

Evaluative conditioning (EC) has long been thought of as a process. According to this view, if we like a conditioned stimulus (CS) more because it was paired with a likeable unconditioned stimulus (US), then our liking of the CS was the result of evaluative condi-
tioning. EC is thus seen as the process by which the evaluation of the US transfers to the CS. However, subsequent research has demonstrated that there are many different types of transfer, so that it is now widely agreed that evaluative conditioning is more usefully referred to as an effect rather than a process. Any one of these processes violates self-controlled attitude formation if transfer of affect occurs outside of awareness. That is, not only must subjects be aware that the stimuli are presented together, but they must also be aware of the transfer from unconditioned stimuli (USi) to conditioned stimuli (CSi) and the intention on the part of a controller to present the stimuli in this way. Below I will run through four mechanisms that produce evaluative conditioning effects and assess what is known about types of awareness for these mechanisms, so that I can subsequently draw out the implications for self-controlled attitude formation.

An easy way to test whether subjects are aware of EC effects would be to rely on a measurement technique that is known to measure inaccessible attitudes. Some have argued that implicit attitudes count as such attitudes. Measures of implicit attitudes are distinct from ones that measure explicit attitudes (which are obtained by means of questionnaires in which subjects report their liking towards an attitude object, typically on a scale, while time pressure and cognitive load are usually absent) and of more recent date. They come in several variants, but all of them measure small differences in response time. In the Implicit Association Test (IAT), subjects are asked to categorize an attitude object (e.g. a face) as belonging to one of two categories (e.g. black faces or white faces) which are represented on opposite sides of a computer screen. Key to the set up is that subjects have on a previous task already categorized another attitude object as belonging either to the category positive or negative, which are located on the same position as the categories in the second task. If for example a black face on the second task is categorized faster to the category of 'black faces' when it is located on the side of the screen that was previously occupied by the category 'negative', then it is assumed that there is a stronger implicit link between black people and negative valence. The strength of this link is represented in terms of reaction time. In the Associative Priming Task (APT) subjects are asked to evaluate an attitude object as good or bad. However, the attitude object is briefly preceded by a prime, which is supposed to evoke an automatic evaluative reaction. If the valence of the prime and the attitude object are congruent, then the attitude object should be associated with the relevant valence faster (and the association should occur slower if the two are incongruent). Again, the differences in reaction time are taken to indicate implicit attitude strength (between the primed word and the evaluation). I will return to what these measurements may refer to later. Here I only want to note that on the basis of a review of relevant studies, Betram Gawronski and colleagues concluded that there is no evidence that people would lack awareness of an attitude because it is implicit. Hence, evidence that EC effects appear only on implicit attitude measurements will not count as evidence for unconscious effects of EC so that we will have to assess direct measurements of awareness for the relevant conditioning mechanisms.

The first EC mechanism to be discussed is implicit misattribution, which is the only mechanism that consistently leads to EC effects outside of awareness. According to the Implicit Misattribution Model (IMM), this EC effect occurs when the affect from the US is mistakenly transferred to the CS if they are presented simultaneously. The experimental set up in which such effects have been found is designed to ensure the absence of awareness (which is supposed to facilitate misattribution). That is, a cover story is presented in which participants are told that they are to play a surveillance guard and are to press a bar as quickly as possible when a designated stimulus appears on the screen (amidst distracting stimuli), while the CS is paired with different USi (with a similar valence). A unique prediction of the
IMM is that the increased salience of the CS but not the US should enhance EC effects, because more salient stimuli are more likely to be the object of attribution. Indeed, when pictures of CSi were made bigger (and hence more salient) the EC effect is found to be larger, but not when pictures of USi were enlarged. Another unique prediction of the IMM is that the affective response to a US will be more likely to be misattributed to a CS if the US is not very evocative (where evocativeness is defined as causing arousal, being highly accessible and having an extreme valence). For if the US is highly evocative one may assume that attribution mistakes are less likely. Interestingly, it was found that EC effects for mildly evocative USi only occurred for participants who were unaware of the US-CS co-occurrence while EC effects of highly evocative USi only occurred for participants with awareness.84

A second mechanism that leads to EC effects is stimulus-stimulus association, in which the CS is associated with a valence via a link with the US. This effect is achieved when the same US is paired with a CS (either sequentially or simultaneously). The best evidence for this effect comes from studies in which the US is revaluated. That is, if the valence association of the CS is mediated by its link with the US, then the CS should change along with the US when the US is independently revaluated. And this is exactly what is found.85 Whether contingency awareness is necessary for stimulus-stimulus EC is contested. Because stimulus-stimulus conditioning effects have been observed for pairs that occurred subliminally there can be little doubt that it can occur outside of awareness.86 On the other hand, it has also been consistently found that contingency awareness enhances EC effects for CS-US pairings. According to a recent suggestion, the way to resolve this issue may be that only CSi that are already embedded in rich associations require contingency awareness in stimulus-stimulus pairing.87

The final two mechanisms occur only when contingency awareness is present and are therefore designated as propositional mechanisms. Pavlovian conditioning leads to EC effects in sequential trials in which for example a picture (CS) is followed by a shock (US). Because the picture predicts the shock, which has a negatively valenced shock, the picture itself is valenced negatively. This effect is, as all Pavlovian conditioning, more vulnerable to extinction effects than other EC effects. Finally, stimulus-attribute association is similar to misattribution but differs in that it only occurs when there is contingency awareness. The boundary conditions for its occurrence have not been studied yet however. But just as Pavlovian conditioning depends on the truth of the prediction that the US will occur, so the belief that the CS has a characteristic of the US should probably be seen as a proposition that can be invalidated by subsequent information.

Of these mechanisms only misattribution is fully independent of (and possibly even be impaired by) contingency awareness, so that its application would be a clear case of self-control violation (provided the intended behavioral consequences occur). For Pavlovian, stimulus-attribute and possibly stimulus-stimulus conditioning it may be a necessary condition that one knows that the US and CS occurred together. Whether this contingency awareness will lead to demand awareness (i.e. knowing the intention behind the co-occurrence) will depend on factors of the context which have so far not been studied. But one would assume that the evocativeness of the CS and the centrality of its relation to the US in the overall message will determine whether the intention to influence attitudes by means of stimulus pairing is picked up.

Evidence for implicit non-evaluative covariation learning has recently emerged as well. When Michael Olson and colleagues in their surveillance experiment replaced evaluative USi such as positive words (e.g. joy) by non-evaluative words (e.g. size) no conditioning effects were observed – which was in line with previous findings. However, when they subliminally primed subjects beforehand with words related to size and ran the same experi-
ment, conditioning effects were found such that the CSi that had been paired with words related to size in the surveillance task were significantly more strongly associated with size after the trials. And as in the case of evaluative misattribution learning, subjects have no awareness of an association between CSi and the US. This suggests that implicit non-evaluative covariation learning does exist, but that it is contingent on the CSi being accessible.88

Behavioral Effects of Unconscious Influence
Evidence for mental state formation outside of awareness is not the same as evidence for loss of self-control. For if mental states just sit there (in the mind) and do not move anything, then the receiving agent cedes no self-control whatsoever. It must therefore be shown that mental state formation induced by relevant conditioning procedures and subliminal influence actually leads to behavior (or subsequent conscious thought processes) and under which conditions this occurs. If there is such evidence, then at least the scientifically informed controllers can be expected to know about the behavioral effects of their manipulations.89 A controller may not be able to predict when the controlee encounters the relevant conditions under which mental states influence behavior, but that seems not necessary for taking control.90 Admittedly, there is not a wealth of experiments that demonstrate attitude effects of conditioning mechanisms and subliminal influence which in turn cause behavior. But if we know how conditioning mechanisms have effects on different types of attitudes and how these types of attitudes in turn affect behavior, then we can infer that (and how) conditioning mechanisms can be conduits of control.

Thus, I will first deal with the effect of conditioning mechanisms on implicit and explicit attitudes. Different models make different predictions about this effect. On the strict dual systems account, there are two distinct processing mechanisms with their own output, each of which is thought to have an exclusive effect on either implicit or explicit attitudes. Conditioning mechanisms (which are a type of so called associative processing) are believed to shape only implicit attitudes whereas deliberation about information (which is also known as propositional processing) is believed to shape only explicit attitudes.91 However, later studies have found clear evidence of implicit attitude change after deliberation about an attitude object, as well as explicit attitude change as a result of EC trials.92 That is, associative and propositional processing each have effects on both implicit and explicit attitudes. The real question then is how these effects are mediated.

The Associative Propositional Evaluations (APE) model makes the most precise predictions in this regard.93 It also posits two processing mechanisms but maintains that they are distinguished on the basis of whether the mechanism is regulated by truth values. That is, on the APE account associative processing (which it understands as EC mechanisms and subliminal processing) does not ascribe truth values whereas propositional processing (which it understands as effortful deliberation) does. As in the case of the strict dual systems model, associative processing has an effect primarily on implicit attitudes and propositional processing has an effect primarily on explicit attitudes. However, the model also assumes that implicit and explicit attitudes are established online, and that they tap for content into the same ‘associative store’. Hence, EC and subliminal influence trials can have an indirect effect on explicit attitudes that is mediated by their effects on the associative store (that is tapped into when explicit attitudes are generated). Likewise, propositional processing about a strong argument works with representations from the associative store, which are thus activated and reinforced (and so may be reflected in subsequently generated implicit attitudes). The prediction that can be derived from this model is that propositional processing
should typically have a greater effect on explicit than on implicit attitudes, and that associative processing should typically have a greater effect on implicit than on explicit ones.

Now, a problem with the original formulation of this model was that EC mechanisms were all treated as associative in nature. We have seen that this is not the case. In fact, a meta-analysis of effects of mainly propositional EC trials (using a sequential presentation order and the same US) shows that these have a greater effect on explicit than on implicit attitudes. However, when the misattribution mechanism is activated in conditioning trials, it is consistently found that the effect is greatest on implicit attitudes (while the effects of stimulus-stimulus conditioning are more ambiguous). This suggests that a revision of the original APE model is in order to reflect that associative processing refers only to misattribution conditioning (and possibly some types of stimulus-stimulus EC). This suggestion has been corroborated in a recent test of the APE model. In a series of studies misattribution was induced (by using a variant of the surveillance task and presenting different USs simultaneously with the CS) and pitted against a trial in which information was presented in the form of propositions. Not only were the effects on implicit attitudes greater in the case of associative processing than in the case of propositional processing (and the reverse). It also turned out, as one would predict on the basis of the APE model, that effects of associative processing on explicit attitudes were mediated by implicit attitudes (which completely predicted the occurrence of explicit ones while the reverse was not true) and that the effects of propositional processing on implicit attitudes were mediated by explicit attitudes.

Thus, whereas propositional EC mechanisms have an effect primarily on explicit attitudes, associative EC mechanisms have an effect primarily on implicit attitudes. Interestingly, and this is the second step in the present argument, these two attitudes affect behavior under different conditions. On this point models on attitudes and behavior are broadly in agreement. The MODE (Motivation and Opportunity as DEterminants) model – which is the most influential one – maintains that attitudes can affect behavior in two ways. They can be generated after reflection about the consequences of possible actions or they can affect behavior spontaneously. Which one of these ways is activated depends on the subject’s motivation to process deliberatively and the opportunity to do so (as constrained by time and task demands). When people have little motivation to engage in deliberative processing and little opportunity to do so, implicit attitudes are found to be more predictive of behavior. Explicit attitudes, by contrast, are more predictive when motivation and opportunity are high (although effects of implicit attitudes on deliberative behavior are not absent).

We said that for control to be ceded to a controller via attitude change, the controller must know which type of attitude will be affected by the unconscious influence procedure that she uses and under which conditions the attitude that is predicted to have changed will influence behavior. We can now say what this means in psychological terms. If a misattribution EC trial is used to change implicit attitudes, the behavior that the controller seeks to elicit must occur under conditions of high time pressure or task demand, or with individuals low in need for accuracy.

**Correcting for Unconscious Influence**

A final objection to the present account is that even if attitudes can be changed by a controller in a way that predictably evokes behavior, we should not be bothered too much by this because a controlee can quickly regain self-control. This is the position of Daniel Dennett. He thinks that even if it were possible to brainwash people in the way that is common practice in philosopher’s thought experiments there is not necessarily a reason to be all too concerned about this. For after “a number of years” in which the brainwash victim deploys her
critical reasoning skills she will reach a point at which we would call her decisions auto-
nomous once again. Supposedly, during these years of decision-making and information
processing, the brainwashing victim will have engaged in a lot of reflection in which a great
number of her brainwashed beliefs and desires have been scrutinized. Maybe she has also
gained new desires and beliefs in a critical way. If we add up all the scrutinized and newly
formed mental states we would find that after ‘a number of years’ the person is just as auton-
omous as before. Loss of self-control with regard to mental states, the argument continues,
would be a temporary phenomenon and can therefore for all practical intents and purposes
be ignored.

In order to properly assess this argument we must move away from imaginary thought
experiments and consider the real world effects of EC mechanisms and subliminal stimulus
presentation. Because both mechanisms primarily affect implicit attitudes, for an agent to
maintain self-control it should be possible – if Dennett’s objection is true – to change or
‘shed’ such attitudes. This, however, is ridden with practical difficulties. First of all, delib-
erative corrections that negate an attitude may change only explicit attitudes. In experiments
in which stereotypes are consciously negated, implicit association strength between the atti-
dute object and the stereotype may in fact increase rather than decrease. In one study,
Gawronksi and colleagues presented subjects with faces of black people and white people
and asked participants either to press the NO key whenever a black face appeared together
with a negative (congruent) stereotype or a YES key when a positive (incongruent) stereo-
type appeared together with a black face. When implicit attitudes were measured, those in
the second condition had more positive attitudes towards blacks while those in the first con-
dition had now actually more negative implicit attitudes towards blacks! Thus, proposition-
ally negating an implicit association appears to activate both representations in the associa-
tive store and may so reinforce its strength irrespective of the nature of the relation.

In this section I have argued that mental state formation by means of unconscious influ-
ence is possible, that the changed attitudes predictably affect behavior and that the controlee
cannot feasibly correct for most of this influence. Although the techniques by which uncon-
scious influence is wielded are not easy to use, they pose a serious prima facie threat to lib-
eral autonomy. I will discuss to what extent unconscious control of attitudes is permissible
together with similar difficult cases in section 4.

2.2.2 Processing Alien Intentions
Even if the intention of a controller to instill an evaluative judgment or to present a goal is
clearly recognized, self-controlled mental state formation may still be undermined. This
occurs when either upon encoding or upon retrieval the receiver of the message falsely
attributes the alien goal or evaluation to herself. This would impair self-control because the
controlee can in such instances no longer properly scrutinize the goal or attitude.

The psychologists Julius Kuhl and Miguel Kazén have presented evidence for precisely
such a phenomenon in an interesting experiment. It simulated a work day in an office for
the purpose of which subjects had to choose from a list of 27 tasks that they were to per-
form later, while 9 were assigned by the experimenter and the remaining 9 were left over.
The hypothesis that informed this first in a series of similar experiments was that subjects
with a specific personality type would be more likely to falsely recall office tasks that were
assigned to them by the experimenter as self-selected. The personality type in question is
state orientation, which refers (among other aspects that are not relevant here) to an impaired
ability to cope with negative affect after failure on a task and a heightened tendency to rumi-
nate about such failures. By contrast, so called action oriented individuals more quickly cut their losses and focus their attention on the next action at hand. The distinction is captured by a survey that asks subjects to report their common reaction to situations. A typical question on this survey asks whether, upon repeatedly failing to contact a person by telephone, one's reaction would be “I can’t stop thinking about it, even while I’m doing something else” (state orientation) or rather “I easily forget about it until I can see the person again” (action orientation). Now, the assumption of the office experiment was that the rumination of state oriented individuals impairs an emotional preference checking system, which is used to attribute a message to a source. (The nature of this system will concern us only later on in this chapter). Under these conditions, it was furthermore assumed, state oriented persons should falsely self-ascribe (i.e. introject) office tasks that are given to them by the experimenter. A series of experiments has established that this is indeed the case.

In a number of variations on the above experiment it was investigated which conditions facilitate introjections for state oriented individuals. In this regard Kuhl and colleagues reasoned that if rumination about negative emotions (i.e. state orientation) inhibits accurate self ascription of tasks, then inducing negative affect should facilitate introjection. This was indeed what they found when they induced (and then measured) negative affect by giving subjects meaningless tasks, by showing sadness arousing movie clips and by having the experimenter exert pressure on subjects. One may object that this body of research is not in fact evidence for loss of self-control because the receiver might simply have liked the goal or evaluation so much that she mistook it for a mental state of herself. On this interpretation, the agent may not have been aware of the origin of a goal, but may all the same have appropriated it in some unconscious self-control bestowing way. If this were true, then one would expect that introjections would occur mainly for items that are most attractive to subjects. In order to rule out this possibility, Baumann and colleagues had subjects rate their preference for the listed office tasks in advance of the experiment. What they found was that introjections overwhelmingly occur for items that are rated low in attractiveness.

Thus, there is evidence for loss of self-control in that the source of a message (including, supposedly, the intention of that source) may be confused in the course of processing (by state oriented persons), especially if they experience negative affect and perceive goals to be unattractive.

2.3 The Shortcut from Self-control to Autonomy

According to some philosophers, most notably Daniel Dennett and Bernard Berofsky, something like the account of self-control that I have just outlined is all there is to autonomy. But this is in fact a minority view. Most philosophers assume that something must be added to the capacities for self-control, and that this something is somehow related to the self. What Dennett and Berofsky deny is not the existence of this intuition, but the existence of psychological processes to which it could refer. The intuition, they argue, results from folk-psychology, and that is not a reliable source. For Dennett the intuition is misleading because it supposes that there would be a Central Headquarter in the brain that is both conscious and in control of higher mental functions. This is a misguided assumption, he thinks, because the feeling of being a self is the result of a fiction that functions to ascribe responsibility. It is, in his words, a center of narrative gravity that helps us understand our actions but that does not actually move anything (or can for that matter own any mental states). Berofsky similarly maintains that reflective reasoning – which is supposed to speak for the self by many theorists of autonomy – may bear little relation to what is actually driving behavior.
Thus, what Dennett and Berofsky deny in terms of my taxonomy is that the intuition of autonomy as a point of view is a reliable basis for a theory of autonomy. In this section I assess whether a conservative reading of the evidence can support the idea of autonomy as a point of view and claim that this is indeed not the case. However, I also argue that this assessment need not have the implications for a role of the self in a theory of autonomy that Dennett and Berofsky think it has.

For the purpose of my assessment I distinguish with Peter Carruthers between two theories about consciousness (or a point of view) and decision-making. One holds that effects of conscious thoughts on action are direct, while another approach takes them to be indirect. The first position is referred to as the self-monitoring model. It assumes that there are two, independent, systems for reading intentions: one for reading the intentions of others and one for assessing our own intentions. The self-monitoring module can on this account not only access propositional attitudes but also directly influence them, so that conscious events in the forms of intentions can directly activate a motor instruction. The other view is known as the mind-reading model and maintains that knowledge of propositional attitudes is acquired solely via the system that evolved to interpret the minds of others. On this account, the mind-reading module generates an interpretation of why we are doing what we are doing only on the basis of introspective and extraceptive stimuli. Because it has no direct access to the practical reasoning system, conscious thought can also not directly settle the choice of a motor instruction.

With Carruthers I also take there to be three groups of relevant evidence for deciding which model is correct. First, the self-monitoring model poses two mechanisms, one for reading the minds of others and one for reading our own, so that one would expect to find cases of patients with local brain damage in which the two systems come apart. Advocates of the self-monitoring model argue for autism as a case where the mechanism for reading the minds of others is deficient while the one for reading our own intentions is intact. Closer examination of the empirical evidence does not support this position however, since a number of experiments have documented greater relative deficits in the introspection of propositional attitudes in autists. Passivity symptoms in cases of schizophrenia are claimed to represent the opposite case. Persons who have this condition are perfectly able to read the minds of others but have the feeling that the actions they perform themselves are not their own. However, the best explanation of such alienation symptoms is related to lower-order systems. When a motor instruction is given, an efferent copy is broadcast and compared with both the initial intention and perceptual data about the behavior. A mismatch results in a feeling of alienation. The best explanation for schizophrenia is that the comparator system (that carries out this procedure) is deficient.

Second, there is evidence from research on split-brain patients and on normal subjects which shows that metacognitions about an action and the actual cause of an action can come apart. This is what one would expect to find if the mindreading model were correct. Split-brain patients are shown to confabulate stories about their actions (mostly in the left hemisphere) that are unrelated to the motivation for behavior that issued in from the right hemisphere. Thus, if the right hemisphere of a split-brain patient is exposed to a picture of snow and the left hemisphere to a picture of a chicken leg, and the patient is asked to select items, then the left hand (which is guided by the right hemisphere) is likely to pick a shovel. The obvious explanation for this choice would be that the right hemisphere associated the
shovel with snow. But that is not what split-brain patients tell the investigator. Since the interpreting is going on in the left hemisphere, which has no access to the snow picture, subjects just make up a story about the relation between the shovel and chickens. The fact that they do not simply say that they do not know why they selected the shovel suggests that it is the default mode to make sense of our actions without direct access to propositional attitudes. Similar effects can be triggered in normal subjects, as a number of experiments in social psychology shows that people falsely attribute actions to themselves if they hear a word that reasonably explains their action no more than five seconds before the action (while no other thoughts may interfere). This is suggestive evidence for the mindreading model as well.

Third, if metacognitive processes have access to practical reasoning processes and are able to intervene in them, then we would expect an innate ability to correct reasoning errors. Yet evidence from the psychology of reasoning suggests that naive subjects are particularly bad at spotting mistakes in reasoning. An alternative (although not exclusive) hypothesis would be that conscious reasoning skills evolved rather in order to engage in persuasion contests. This hypothesis has much better support, since it neatly explains phenomena such as the innate confirmation bias.

Because the aim of formulating a minimal theory of autonomy places the burden of proof on those who want to argue for more demanding notions of autonomy, I take it that the evidence for the self-monitoring model is at least not strong enough. Hence, we are to assume from here on that conscious events do not directly cause motor instructions. By corollary, the point of view from which we think we act cannot be said to have the special autonomy-conferring status that proponents of autonomy as authenticity, such as Gary Watson, Harry Frankfurt, David Velleman and Michael Bratman ascribe to it. It is important to stress though that arguing against the authority of the point of view of the agent is not the same as arguing against a role for consciousness either in action or in a theory of autonomy. That is, although the point of view of the agent does not confer authority on actions, this does not rule out the possibility of indirect influence of conscious actions, which may still be said to be important in a theory of autonomy. More important still, whereas the evidence for the direct involvement of conscious processes in action is weak, there is very strong and uncontested evidence for the importance of conscious processes in the formation of mental states.

To sum up, I agree with Dennett and Berofsky that theorists who think that the autonomy of actions is bestowed by the agent’s point of view are on shaky ground. However, it is one thing to claim that the point of view of the agent confers no relevant authority on actions and quite another thing to claim that there is no role for the self in a theory of autonomy. For the latter claim to be true it would have to be the case not only that the point of view of the agent is immaterial in the formation of mental states, but also that autonomy as coherence (in relation to both actions and mental states) lacks a reliable empirical correlate. In the next section I will claim that this premise is wrong.

3 Authenticity

What an argument for autonomy as coherence should look like seems clear enough. If one were able to distinguish a subset of mental states that can be said to represent the self, then
actions caused by such states can be called authentic. However, a problem with this approach is that there is a lot of disagreement among psychologists and philosophers alike about the referent of mental states that represent the self. Some psychologists think of the self as a central node in a memory network, others see it as a disparate set of modules, and yet others conceive of the self as a connectionist network. Likewise, philosophers who work on autonomy have widely diverging views on the self and see it either as a mode of information processing, a set of narratives or a set of policies. I cannot hope to resolve these fundamental disagreements and establish which one of these concepts captures the 'real self'. I also cannot hope to develop normative accounts of authenticity for all of the central theories of the self separately (that would be too laborious). I will therefore remain neutral as to what the self refers to in the mind. The only assumption that I make is that when new states are acquired, there will be special places in the mind where states are stored on which the organism relies when it acts, or that new states are tagged as belonging to the self so that they can be retrieved when the organism acts. If there were no such mechanisms, then nothing would prevent one from acting on beliefs and desires that one ascribed to others. And such a life would lose its coherence.

On the basis of this assumption, minimal authenticity can be defined with reference to two criteria. The first is appropriate acquisition. If we could define plausible criteria for appropriate acquisition, then we could say that actions that issue in from states that were not acquired appropriately are not authentic actions. Defining such criteria will be the main task of section 3.2. The second criterion is access to relevant self-representations. If memory systems are stuffed with appropriately acquired self-related states, but reasoning systems receive hardly any input from them, then actions still do not speak for the self. From an evolutionary standpoint such differences are to be expected. When seeing eye to eye with a saber tooth tiger it would be seriously non-adaptive if the brain would take the time to extensively access memory systems in order to evaluate needs and desires. Instead, the system that regulates such scanning should block all access save the most elementary environmental input and skills that are relevant to fight-or-flight reactions. In section 3.1, I want to demonstrate that such differences in self-access are not only to be expected, but that psychology offers evidence for their occurrence as well as the conditions under which they occur.

3.1 Authentic Action
An account of authenticity as coherence with appropriately self-related states must from the outset face a variation on the objection against autonomy as a point of view. This objection holds that just as the point of view of the agent cannot be trusted as a source of authenticity, so self-related states from memory as they are acquired in psychological surveys cannot be trusted to cause action. It is possible, after all, that in reality unconscious memory states do the causal work. In response to this objection I will outline relevant retrieval mechanisms that operate both consciously (3.1.1) as well as unconsciously (3.1.2).

3.1.1 Explicit Authenticity and Action
The first important case of authentic action consists in relating a possible course of action to relevant explicit self-related states. I distinguish two ways in which such states are accessed. One concerns the mere interpretation of what we are doing, and the other involves higher-order reasoning.
A Minimal Account of Authenticity as Interpretation

In a highly original account, David Velleman maintains that autonomy consists essentially in interpreting what we are doing. Central to his theory of autonomy is the assumption that intentions are the output of theoretical reasoning that is informed by several intellectual desires. The most important of these are the desire to know what we are doing and the desire to know why we are doing what we do. According to Velleman it is only because of the motivational force of such intellectual desires that we are not limited to acting on innate drives but can also act on predictions about what we are going to do. For if such a prediction is linked to the desire to know what we are doing, it can attain motivational force and hence outcompete other desires and issue in actions. A reason for action can now be defined as a thing that enhances self-knowledge upon being intended. That is, if we act on a prediction, then what we will be doing is already known to us – it was what we expected to do after all – so that our self-knowledge increases. On the basis of this theory of action, Velleman defines autonomy as acting on predictions that are most coherent with explicit self-representations.

To this end he introduces, in a discussion of Dennett’s views on the narrative self, a “narrative-composing module”, which “will … declare a winner [among potential courses of action] simply by telling the more coherent continuation of the story.” And the most coherent continuation of the story is the one that is best in line with our beliefs and desires.

There are a number of problems with this account. First of all, there are serious difficulties with Velleman’s concepts of intention and a reason for action, which has been pointed out by Michael Bratman. When we act, we not only expect a goal to occur but also the side-effects of that goal. Yet if expectations are intentions by virtue of increasing our self-knowledge when they are performed, then side-effects that we clearly do not intend (say, waking up a person by switching the light) would have to be described as intentions as well. Similarly, Bratman argues, the expectation that I will stumble in front of an audience would increase self-knowledge and hence have to count as a reason for action. Both consequences are unacceptable. The only way to resolve this problem would be to restrict the expectations that can settle a decision to those that are based on an agent’s relevant preferences. But such relevant preferences would have to be practical desires for success in action, so that the intellectual desires for self-knowledge and understanding are no longer the drivers of behavior.

Secondly, Velleman posits a desire to know what one is doing that would lead one to do what one is knowing. There can be little doubt in any conscious person that we have a desire to know what we are doing. The issue, however, is whether this desire causes us to act upon what we are expecting to do. Velleman thinks that evidence from social psychology is at least coherent with his view, and, supposedly, that this would strengthen his case. The psychologists Wegner and Vallacher have indeed argued that we select a ‘proponent act description’ before we act, and that this selection is guided by a concern for acting consistent with beliefs and desires that make sense of the action. And decades of research on cognitive dissonance have indeed shown that people go to great lengths to maintain consistency. But none of this research has any bearing on the question of whether we select such act descriptions – or expectations – because we want to know what we are doing. Instead, all the evidence that Velleman has mustered rather points in the direction of a desire to act in way that is consistent with act descriptions. For the sake of parsimony, we should therefore pose – for the regulation of action at least – a consistency desire rather than an intellectual desire to know what we are doing (or the desire to know why we are doing what we do).
On the basis of these criticisms I want to propose a more minimal version of a theory of autonomy as interpretation that retains, I think, the spirit of Velleman’s account. It takes as its starting point the assumption that the mindreading module is primarily involved in interpreting the actions that are in fact selected by the system 1 practical reasoning modules. The results of these interpretations can in turn indirectly influence the practical reasoning process if the selection of future action schemas is regulated, at least in part, by the desire to act consistently with self-interpreteds. Moreover, this construction allows me to say that acting upon these interpretations results in autonomy to the degree that these interpretations (i.e. act descriptions) are formed on the basis of (appropriately acquired) self-related states. That is, when more states that are made proponent in action identification enter into the consistency assessment process, then actions that issue in due to a desire for consistency (that is properly related to the consistency assessment process) are more autonomous. Take the case of pressing a doorbell. This single action can be interpreted as, respectively, stretching a finger to reach for the button, ringing the bell, or looking if a friend is at home. It is clear that the highest level interpretation touches on far more concerns that the lower level ones. Looking if a friend is at home raises many more questions with regard to social norms (Is this an appropriate time?), means-end efficiency (Should I not rather give him a call?) and ethics (Is it ethical to visit a convicted pedophile?). Hence, many more self-related states would be activated than the simple interpretation under which we stretch a finger to reach for a button. Thus, the higher the level of identification, the more likely it is that self-related memories will enter into the consistency assessment process.

It seems to follow then that a minimal theory of autonomy can simply set a certain level of act identification given classes of situations. Unfortunately, personality differences make such a move problematic. To see why, we must go into some more details of Vallacher and Wegner’s action identification theory. It states on the one hand that if higher and lower level act identities are available, there is a tendency for the higher level act to become proponent. Evidence for this principle is found in the readiness by which people accept a higher order act identity that is presented to them when they are identifying at a low level. On the other hand, when act identities cannot be maintained at a certain level (because the act cannot be carried out effectively) the tendency is to move down the hierarchy, where actions are broken down into their component parts. On the basis of this principle Vallacher and Wegner predict that people who are less skilled at carrying out a range of tasks (so called “chronic klutzes”) should in general maintain lower act identities. When the researchers compared responses on a questionnaire that asked subjects to self-report success at a range of tasks (such as typing, playing an instrument, planning a party, and building furniture) with responses to questions that asked subjects to indicate the level at which they would identify an action (e.g. whether they would describe ‘making a list’ as ‘getting organized’ or as ‘writing things down’), they found a significant correlation in support of this prediction. That is, people who report more difficulties in carrying out actions effectively are more likely to display the tendency to identify actions at a lower level. But that would mean that personality differences influence people’s propensity to be autonomous, and that makes formulating a minimal threshold a tricky affair. Dealing with this practical problem in the context of liberal autonomy will be the task of section 4.

A Minimal Account of Authenticity as Reflection
Distinct from interpreting what we are doing (which may be the work of a specialized mindreading module) is the activity of reflecting about what we are to do. All hierarchical theorists of autonomy assume that a special relation with self-related states is established in
reflection, but John Christman is the only one who has developed this view in significant detail. He discards the idea of self-governance as identification, since it means either that we must “identify with all aspects of ourselves that we acknowledge as such”, which would force us to acknowledge addictions, compulsions and the like. Or it means that we must identify with “those principles, values, and ideals that we hold in most esteem” so that we are led by our “best selves” – but then we would have to view everyone who falls short of her ‘ideal self’ as deficient in autonomy, which would include all of us.131

In an attempt to strike a balance between these extremes, Christman requires only that upon reflection on a ‘basic organizing value’ – but we could also insert here ‘course of action’ – we do not experience alienation (i.e. we must both believe and feel that we should not repudiate the value or course of action).132 Obviously, we need some guarantee for thinking that this reflective judgment is properly related to the self. To this end, Christman demands that the value (or course of action) be processed by “the characteristic mode of thinking, embodiment and feeling that the agent’s self-schema manifests”. This ‘characteristic mode’ refers in turn to the schemas, heuristics and scripts that are activated “over a variety of circumstances”. Moreover, from among these stable processing modes only the ones that are activated when we reflect (where this reflection is unimpaired by manipulation, pathology and the like) speak for what we could call the autonomous self.133

There is more to Christman’s theory of autonomy, but what concerns me here is the extent to which reflection can count as authentic on his account. In this regard I am skeptical, for three reasons. First, it is not clear in how many circumstances reflection is to occur, or even by which criteria the appropriate number and type of circumstances could be selected. If this is correct, then it seems more promising to identify psychological mechanisms by which self-consistency is assessed and to demand that these are not impaired. Second, self-schemas are unique individual modes of processing information, but if they are to guarantee autonomy then they must take the form of reflection on Christman’s account. Hence, if there are individual modes of processing that are not reflective, then there can be a conflict between a state of alienation triggered by characteristic processing and a state of non-alienation triggered by reflection. If that is true, then the requirement of reflection seems to be biased against certain ‘characteristic’ processing modes. This bias is problematic for at least a liberal theory of autonomy because, as Berofsky notes, it is one thing to claim that autonomy involves some reflection and quite another to claim that only reflective judgments speak for the self.134 While the former commitment is reasonable from the perspective of a minimal liberal theory of autonomy (see section 2.1), the latter is not.135 Third, a liberal theory of autonomy should not decide which types of feelings or beliefs (such as non-alienation) should function as the standard of evaluation in reflecting on self-related states. For such a theory it suffices that self-related states are appropriately acquired and that reasoning systems have appropriate access to them. In the latter respect, a minimal theory must ask instead how access of self-related states to system 2 processing is regulated.

On this question we must, absent relevant work in psychology, rely on some tentative findings from neuroscience. The short summary of studies on this issue is that reflection that involves a judgment on preferences of the self (as opposed to reflection on an abstract problem) appears to go together with the activation of a specific area in the brain that is known as the Cortical Midline Structures (CMS).136 Activation of these structures during self-reflection is moreover shown to reduce conflict in cases where there are multiple correct possibilities, such as in occupational choice.137 That is, system 2 conscious reasoning about self-relevant dilemmas does seem to resolve personal dilemmas, and it may well do so by relying on self-related states (given activation patterns). What is more, the impairment of
these self-reflective processes due to cognitively demanding tasks is associated with less activity in CMS and more activity in lateral prefrontal areas of the brain (while the reverse also holds true). Thus, the above evidence suggests that self-related processing by system 2 may be blocked by cognitive load, and is in that sense similar to the interpretation system (where cognitive load decreases task proficiency and downshifts act identification).

### 3.1.2 Implicit Authenticity and Action

The case for implicit authentication of action in a theory of autonomy can be inferred from the work of Harry Frankfurt. In section 2 I discussed Frankfurt’s theory of autonomy as acting on second-order desires that are decisively endorsed. Several critics have noted that although it makes intuitive sense to maintain that autonomy is dependent on some form of reflection, it is not intuitively clear why reflection by itself would speak for the agent. The later Frankfurt has responded to this question by claiming that reflection does so only when it is in line with our ‘essential volitional essences’ that consist of mental states of deep caring. Frankfurt hereby abandons the active nature of the decisive endorsement of a higher-order desire and stresses instead that becoming ‘wholehearted’ – as it is now called – is not “under … immediate voluntary control”. It is not a matter of merely “telling stories about our life”, since “we can be only what nature and life make us”. Instead, having a free will is to “be wholehearted in [the will]”, that is, to have no volitional conflict. Wholeheartedness does not mean that there can be no “inner opposition to [one’s] will”, but it does mean that one “must be resolutely on the side of one of the forces struggling within” and “not on the side of any other”. Being satisfied is defined as the “absence of restlessness or resistance”, or “having no ambition for improvement” of the current situation. Thus, satisfaction consists not in the adoption of any “cognitive, attitudinal, affective, or intentional stance”, but is simply a “state of the entire psychic system”. The only requirement is that this state emerge reflectively, for otherwise one would have to grant so called ‘wantons’ autonomy.

Although there is much to criticize in this account, I also want to maintain that the core intuition of autonomy as acting in harmony with mental states that are not under volitional control is defensible. With regard to the critique, Frankfurt does not explain what the phrase ‘volitional essence’ refers to. He also fails to make clear what entity could be on a ‘side’ in the case of volitional conflict, while it is likewise not clear why the amount of conflict would determine the amount of autonomy. There is at least a strong intuition against autonomy as harmony. But these critical points do not invalidate the impetus of Frankfurt’s approach, because they are for the most part related to Frankfurt’s conception of the self as a perspective. If we think of the self instead as a set of representations in the mind, we have no need to introduce ‘volitional essences’, much less an entity that takes sides between them. Instead, we need only require of an action that it issues in from a practical reasoning process that has appropriate access to implicit self-representations. This alternative reading of the implicit coherence intuition is contingent on the existence of a relevant retrieval mechanism. To this end I will outline in broad lines the theory of volition developed by Julius Kuhl, known as Personality Systems Interaction Theory (PSI), which poses precisely such a mechanism and presents evidence for its existence.

### A Psychological Account of Implicit Self-consistency Processing

In PSI volition is seen as the output of a complex system that consists of two main processes, self-control and self-maintenance, which are regulated on the basis of positive and negative affect. Self-control is concerned with inhibiting impulsive actions and maintaining a
single focus on goals. It is conscious and effortful and consists of two component systems that are called intention memory and object-recognition. Intention memory is concerned with maintaining abstract symbolic representations of intentions in working memory and inhibits intuitive behavior in order to prevent premature behavior. The object-recognition system is a lower-order system concerned with the detection of discrepancies and warns others systems in the brain in case of their occurrence. The other major system, called self-maintenance, is concerned not with the suppression of alternatives but rather with the integration of a wide array of desires and implicit self-representations. Its two component systems are extension memory and intuitive behavior control. Extension memory integrates cognitive and affective information from various subsystems and processes it much faster than analytical thinking due to parallel processing. Intuitive behavior control translates intentions into routine actions and also regulates spatial awareness. Volitional behavior is supposed to result from carefully calibrated switching between these two modes and to be regulated by negative and positive affect. An increase of negative affect is a sign of danger and activates the object recognition system, while a decrease in negative affect indicates safety and activates extension memory. An increase of positive affect is believed to signal the absence of problems so that intuitive behavior is activated, while a decrease in positive affect supposedly demonstrates that needs are not met so that intention memory is activated.

If there were in fact such a self-maintenance system, then this would be the type of implicit consistency checking mechanism that can support a place for authenticity in a minimal theory of autonomy. It is therefore most interesting that Kuhl and colleagues have deduced specific predictions from PSI for the experimental setting of the office work day simulation described in section 2.2 and put them to the test. Note that if there is a system that automatically assesses the coherence of a course of action with states of the self then it may do so at the encoding stage (when office tasks are allocated) or at the retrieval stage (when the memory task is completed). Only the assessment of coherence at the retrieval stage might be evidence of the type of mechanism that we are looking for.

In line with this possibility, Kuhl and colleagues hypothesized that because individuals who are worse at automatically downregulating negative affect (i.e. state oriented individuals) are more likely to introject tasks given to them by others, the retrieval of task selection may involve an affective self-consistency checking mechanism. Specifically, they argue that there may be a check of (emotional) preferences about the goal or attitude object (which is generated on line) which may result in a commitment marker being tagged to that goal. If this commitment marker is scanned when a goal is retrieved, then subjects who are best at retrieving task selection (i.e. action oriented individuals) should show the longest response latencies (since scanning takes time). The assumption was tested in a number of studies which found that upon retrieving task origin there is indeed a longer time lag for action oriented individuals (compared to state oriented individuals), especially under conditions in which negative affect is induced (i.e. meaningless tasks, pressure and sad mood induction). As we noted earlier, negative affect induction before task retrieval increases the introjections rate (for state oriented individuals), so that there is good reason to assume that at least part of the introjection indeed occurs at the retrieval stage and that the time lag is related to this effect.

Let me stress that I leave open against what exactly consistency is checked by this automatic process. These entities may correspond to what is found on implicit attitude measures, because the process operates so fast. But it cannot be ruled out that the attitude that corresponds to what is found on explicit reports is also accessed. Moreover, how these mental states are instantiated in the brain is also not clear. This structure may consist of subsym-
bolical connectionist networks (which is what PSI assumes) or symbolical modules, but this issue is highly contested. What is to be retained for now is that there is a process of automatic and unconscious preference scanning for goals and attitude objects that aids the recall of selected actions and may reasonably be supposed to be activated in action contexts. By virtue of having a role in selecting an action we say that the entities that are accessed by this process can be called mental states of the self. The greater the access to such states, the more authentic a person becomes.

Once it is agreed that both implicit and explicit retrieval is relevant for authenticity it may be wondered which one speaks for the ‘real self’. A liberal theory of autonomy, I argue, should not take sides on this issue. Intuition of course tells us that we should side with explicit retrieval, but if Kuhl and colleagues are correct, explicit retrieval may activate a smaller number of self-relevant states than implicit retrieval. Alternatively, one could argue that harmony of the two retrieval systems leads to more autonomy. There is currently no research that speaks to this. But if explicit retrieval were to rely more on explicit attitudes and implicit retrieval more on implicit attitudes, then there would be relevant evidence that indicates that a state of self-awareness can reduce the gap between explicit and implicit states. However, it is not realistic to require citizens to engage continuously in meditation practices and the like that foster such harmony.¹⁵¹

Hence, a minimal liberal theory of autonomy should only demand that access to the conscious self is unimpaired by cognitive load and it should demand that implicit processes are unimpaired by negative affect.¹⁵² Because of this liberal stance, the theory is able to ascribe autonomy to the entire spectrum ranging from over-reflective and inhibited personalities to impulsive hedonists. And this is precisely what liberal autonomy should look like.

3.2 Authentic Mental States

For mental states to count as authentic it will not do if beliefs and desires are simply stacked on to each other. If we think of the memory system upon which implicit or explicit retrieval rely as a box, then if that box simply contains all the beliefs and desires the system happened to encounter in its environment the set of mental states of a person will not count as autonomous. And we would say so not merely because the resulting incoherence would disturb efficient action execution (and thus diminish self-control), but also because newly acquired beliefs and desires would not be related to existing mental states in a way that preserves the authenticity of the self. The task of this section is to give the most convincing minimalist account of criteria for such mental state acquisition without specifying what such a self should look like.

The most straightforward proposal for a principle of authentic mental state formation is to require that all mental states be processed by bringing to bear on them the mechanisms for interpretation and reflection that I outlined in section 3.1. Although such a proposal might satisfy the demands of the intuition of autonomy as coherence or a point of view, it does not satisfy the demand for realism, because such processing (at least in the case of reflection) would be way too resource intensive.¹⁵³ This concern has prompted John Christman to propose a modification.¹⁵⁴ If we cannot realistically be required to reflectively endorse new mental states, then maybe we can require that we endorse new mental states were we to reflect on them in light of how we acquired them. However, the problem with defining autonomy on the basis of hypothetical reflection is, as I noted earlier, that there is no practi-
cal way in which others can establish whether a person would have reflectively endorsed a mental state.\textsuperscript{155}

A second solution to the problem that reflection on new mental states is too resource intensive has been developed by Alfred Mele in a discussion of the authenticity of values. He specifies the conditions under which reflection is to occur, where the reflection bestows autonomy by means of being endorsed by the point of view of the agent. Specifically, Mele identifies two principles according to which this process is to be structured. Initiatory principles stipulate when assessment should issue in and procedural principles stipulate how assessments should proceed. Mele mentions a number of such principles for the purpose of illustration, but maintains that they should not be fixed by a theory of autonomy. Instead, autonomous agents should design and implement these principles themselves.\textsuperscript{156} Although this radical subjectivism about standards of mental state acquisition may seem attractive from a liberal point of view, it cannot be the basis for a minimal liberal theory of autonomy. For one thing, it is not realistic to suppose that mechanisms that initiate reflection are under conscious control. Such a mechanism would have to filter huge amounts of data in a way that is carefully calibrated with other unconscious heuristic principles. Such complicated design work appears to be beyond the reach of conscious decision-making. For another thing, because there would on Mele’s account be no straightforward way for the state to establish which initiatory or procedural principles an agent upholds, this construction would – as in the case of hypothetical reflection – be practically meaningless. That is, the liberal state cannot feasibly keep track of individually defined initiatory and procedural principles and thus could not protect autonomous mental state formation.

In the remainder of this section I want to present an alternative account which modifies Mele’s framework by replacing the intuition that autonomy is a point of view with the intuition that autonomy involves coherence. This move allows me to find objective criteria of authenticity violation. That is, in order for initiatory mechanisms to support a theory of autonomy as a point of view, they must be endorsed by the agent in reflection – hence Mele’s untenable subjectivism about initiatory mechanisms. But once we take coherence as the guiding intuition for designing a theory of autonomy, we have an externally verifiable criterion for identifying initiatory mechanisms. I will here outline two such coherence-checking initiatory mechanisms, while taking a liberal (and hence minimal) approach to the types of assessment that guarantee authenticity. The first type of consistency-checking mechanism is domain specific. In this regard research has so far focused on a mechanism that is concerned with maintaining a balance with regard to the valence of persons. Although this mechanism issues in automatic domain specific processing I want to argue that this does not undermine authenticity, since it remains restricted to lower-order processes. The second mechanism is general and concerned with bringing a host of self-relevant mental states to bear on new mental states. If it is plausible to claim that these mechanisms form the main portion of the initiatory defense structure, then we can say that authentic mental states are the ones that have passed through these initiatory processing channels.\textsuperscript{157}

The operation of the domain specific mechanism is studied in balance theory. This theory holds that relations between people tend towards a balance in which we like the friends of our friends, dislike the enemies of our friends, and so on. In triad relations – i.e. relations between three people – there is balance if there are either no negative relations or an even number of negative relations. If such balance is established at encoding, then this would be a case of a mechanism that detects inconsistency and initiates an automatic corrective response. In a recent experiment it has been shown that this is indeed the case. If subjects were given a negative valence of a source first (e.g. Peter is not nice) and afterwards
his negative evaluation of a target (e.g. Peter dislikes Paul), then the subject’s attitudes towards the target balanced out and was positive (e.g. I like Paul – the enemy of my enemy). However, if the valence of the source (i.e. Peter) was given only after his negative evaluation of Paul, no balancing took place (i.e. the negations did not cancel each other out). Likewise, in another experiment it was shown that if the valence of the source is changed afterwards (e.g. Peter is in fact not nice at all) the balance that ought to occur in relation to his negative evaluation of the target (i.e. the two negations should cancel each other out) fails to materialize. If balance were established upon retrieval, then the order of presentation should not have an influence. As these experiments show that the order of presentation does have an influence, they tell us that there is a mechanism that automatically establishes a specific type of consistency in our evaluations of persons upon encoding. However, because balance processing aims to maintain consistency, which is dependent on the ascription of truth value, associative processing should not be subject to it, if the APE model is correct.

The second initiatory mechanism that can be argued to meet minimal standards of authentic mental state acquisition is the unconscious consistency checking process that is studied in PSI research. In section 3.1.2, I offered evidence for unconscious consistency checking at the retrieval stage, but there is also good evidence to suggest that it occurs upon encoding. According to PSI what should happen at the encoding stage is that emotional preferences towards a goal or attitude object are scanned, after which it may be tagged with a commitment marker. On the basis of earlier research, Kuhl and colleagues assumed that many of the presumed aspects of the emotional preference scanning process are located in the right hemisphere. For example, whereas the recognition of faces of other people correlates with more activity in the left hemisphere, the recognition of one’s own face correlates with more activity in the right hemisphere. The right hemisphere is also known to be more active when integrative processing is observed and when increased activity of bodily states is registered. In order to manipulate the activation of the right and left hemispheres, Kuhl and colleagues relied on research that indicates that muscle contractions on one side of the body activate the opposite brain hemisphere and its functions (the hemispheres are connected to opposite sides of the body for reasons that are not well understood). Thus, squeezing a ball with your left hand should activate your right hemisphere, and if that hemisphere is indeed more involved in self-related, integrative and emotional processing, then the number of introjections for state oriented people should decrease.

Interestingly, the ball squeezing manipulation allows one to accurately isolate encoding effects from retrieval effects. As such, experiments using the office set-up that was described earlier demonstrate that one minute of left-hand ball squeezing before encoding (i.e. selecting and being assigned office tasks) by state oriented individuals totally annihilated the surplus of false self ascriptions. Moreover, both state and action oriented individuals showed a greater tendency to introject after right hand ball squeezing (and concomitant left hemisphere activation). The result has been replicated for attitude formation (where the attitude objects were tests for measuring everyday intelligence, which subjects rated for their effectiveness). This consistency-checking mechanism can be said to be initiatory to the extent that there is suggestive evidence for thinking that it sends a signal to reflective processing. This evidence builds on an intermediate stage in the argument within PSI theory on introjection, which is that performance on tasks in which subjects are to judge the coherence of triplets (e.g. ocean, sand and bucket) is impaired in state oriented individuals after negative affect induction. Interestingly, the psychologists Sascha Topolinski and Fritz Strack suggest how
this effect may initiate relevant reasoning. Specifically, they propose that the processing of coherent triplets occurs more fluently and that this fluency arouses positive affect. This positive affect may in turn be read by a monitoring system and automatically influence or even become the object of system 2 reasoning. In a series of studies, Topolinski and Strack have made a cumulative case for this proposal by showing that coherent triplets are processed faster, that they function as primes that facilitate positive affect and that they are liked more (compared to incoherent ones). Thus, if integrative coherence judgments do indeed account for the relevant consistency checking, then positive affect may bias reasoning about self-consistency without giving any insight into the nuts and bolts of the process. On this account, we are conscious only of a positive feeling when a goal is found to be coherent with our mainly emotional preferences. When a negative feeling occurs, by contrast, this could conceivably initiate central processing. But research is needed to corroborate these speculations about the initiatory function of this mechanism.\textsuperscript{164}

In any case, it is unlikely – as it is unlikely for automatic balance regulation – that associatively processed stimuli will be subjected to this coherence-checking mechanism. That is, the mechanism has been shown to operate in the case of goals or attitude objects that were deliberatively chosen, but it is unlikely that it would kick into action when two stimuli are implicitly associated with each other, let alone when affect is misattributed from a US to a CS. Admittedly, there is research that might be interpreted to suggest that conflict detection leads to appropriate changes of implicit attitudes indirectly. That is, a number of experiments have shown that when subjects have discrepant explicit and implicit attitudes about the same attitude object, they are more willing to engage in deliberative processing about information that is relevant to the attitude object (and that might resolve the discrepancy).\textsuperscript{165} In order for this effect to occur there must presumably be a separate system that receives input from the propositional and associative processing systems about an attitude object and checks whether they are consistent. This is of course not the same as evidence of a consistency checking mechanism for implicit-implicit conflict. But it could be argued that when discrepancy between implicit and explicit measures is reduced, and explicit measures are subject to consistency detection, implicit attitudes are by virtue of being aligned with explicit measures indirectly responsive to conflict detection. Such an argument is unpersuasive though, because processing information that resolves the implicit-explicit attitude discrepancy by changing implicit attitudes often widens the discrepancy, because – as we saw in section 2.2 – negations often do not carry over to implicit attitudes.\textsuperscript{166}

If this is correct, then associatively processed stimuli bypass both types of initiatory mechanisms. If there is no plausible way in which initiatory standards can be lowered, then the possibilities of finding natural defense mechanisms against authenticity undermining influence have been exhausted. Because these assumptions seem to be correct, we are forced – as in the case of self-controlled action and mental state formation – to look for thresholds beyond natural defenses.

4 Practical Thresholds of Liberal Autonomy

In detailing the constituent components of minimal liberal autonomy I have come across a number of unsolved puzzles that concern mainly covert influence on both action and mental states. Specifically, with regard to self-control it is to be determined which extent of
manipulation is acceptable and with regard to authenticity which minimal levels of cognitive load, negative affect and associative processing are acceptable. In this final section I propose practical standards that can bring my minimal theory of liberal autonomy into equilibrium with intuitions about appropriate state action. I have of course already discussed some practical considerations, but I have done so only insofar as they were germane to the plausibility of mechanism that are constitutive of autonomy. Here I consider solutions to problems that persist after the possibilities for modifying such mechanisms have been exhausted. And I will do so for both self-control (4.1) and authenticity (4.2).

4.1 Thresholds of Self-Control
Whether liberal state action is imperative when self-control is undermined is – I propose – contingent on two factors. First of all, state action must be proportional to the violation of self-control (just as it must be proportional to a violation of freedom). In other words, the violation of self-control must be sufficiently central to the controlee before state action is imperative. Second, state efforts to protect autonomy should not bring with them violations of autonomy in another respect (or rather, the latter violations of autonomy should not outweigh the gains won by protecting autonomy). Specifically, efforts to protect autonomy can impair autonomy as a side-effect in the form of both violations of privacy and violations of the controlee’s consent.

Lack of Centrality
To begin with, not all instances of manipulation seem serious enough to warrant state action. Consider the following cases. Suppose I know about the process of mirroring (either from theory of from experience) and just to amuse myself I shake my foot with the goal of making you shake your foot. I hereby tap into an unconscious mechanism by which you regulate your body movements in interaction, and so bypass your ability to know about my intention. Hence, my action would result in a clear loss of self-control on your part. Or suppose that I direct a thriller movie in which I want to create an experience of suspense whenever a certain evil character appears. In order to do so, I play a piece of music that I know arouses an eerie feeling every time the evil character enters the scene. From experience (or from seminars at the film academy) I know that the viewer will make an associative link between eeriness and the evil character (even if I know nothing about the details of evaluative conditioning). If the editing is subtle and the story an enthralling one, I will achieve my goal even with an experienced public.

Now, both in the foot shaking and the movie example autonomy conditions are violated intentionally and yet it seems deeply wrong to say that liberal autonomy is violated in any meaningful way. In these cases the obvious source of this intuition pump seems to be the relative futility of the behavior and attitudes. In other words, it seems so utterly immaterial to a person’s life whether she shakes her foot or thinks that a particular movie character is eerie that it is unreasonable to suggest that the liberal state should take action to protect her. By analogy, although breaking your neck in the course of a fistfight will land me in jail, pushing you off the sidewalk will not. Although the latter act is wrong in some sense, it is just too marginal an offence to require state action. And just as there is no clear cut-off point at which violence becomes an offence, so there is unlikely to be a point where intentional autonomy violation becomes serious enough to become a concern for the liberal state. Still, if this analogy were to be useful, there would have to be standard of centrality of self-control loss equivalent to bodily harm. I propose that such centrality of self-control loss pertains to
both mental states and behavior of the controlee, where the centrality of mental states is measured against the attitude’s position in the controlee’s goal hierarchy and the centrality of behavior is dependent on its social and private consequences.

I have no intention of delving into debates in psychology on the details of goal hierarchies. Instead, I will simply assume that something like the goal hierarchy of three stages developed by the psychologists Charles Carver and Michael Scheier will be true. At the most basic level of this model there are sequences (or motor control goals), one level up are programs which are described as ‘do goals’ (e.g. going to the movies), and still higher up are principles and traits (such as ‘be thoughtful’) which are connected in memory to the self-concept. With regard to mental states this simple model allows us to say that – all things being equal – changing mental states that involve ‘do goals’ is of greater consequence than changing mental states that involve motor instructions. Hence, whether we think of a movie character as eerie or not can be said to be of little significance because its position relative to the goals that structure our behavior will in the typical case be peripheral.

Manipulations of behavior are more central if they have a greater effect on future actions by the agent. This effect can be mediated either by the effect of the behavior on the social environment or by the effect on the psychological organization of the agent herself. In the first case the agent makes a commitment which changes the perception of the agent in the mind of others. For example, if the manipulated act were a marriage, then the belief held by others that the manipulated agent is married changes their future conduct towards her, and hence has a substantive effect on her future actions (depending on the relationship). Such an event might also involve, and this is the second case, a change in self-perception that will have a major impact on the content of the agent’s goal hierarchy and so influence future actions (where again the level in the hierarchy determines the centrality). Hence, all things being equal, the manipulation of actions that are public will be more central than the manipulation of private actions and actions that influence goals higher in the goal hierarchy level are more central than actions that influence goals at the bottom.

Irretrievability
The present proposal still allows for many cases that do not even look like threats to liberal autonomy as we intuitively understand it. Suppose that after a series of painful losses I finally want to defeat you in a game of trivial pursuit. Knowing my share of psychology, upon opening the box I start a discussion about soccer hooliganism and in the course of the ensuing discussion ask you – by way of a rhetorical gesture – to imagine being a hooligan in order to understand their side of the story. By activating in you schemes associated with low intelligence (which results from empathizing with soccer hooligans) I cause you to underperform during the game and so I achieve my premeditated goal of defeating you. Now, on the centrality condition I meet the criteria for relevant liberal autonomy violation if we add that the consequences of your loss will be severe. Suppose that in light of your previous string of victories against me, you overconfidently betted your house on winning the contest. Although the effects of the covert influence are now central to the controlee, it seems wrong to suggest that the liberal state therefore has an obligation to step in and prevent me from engaging in this petty trap of mine.

This intuition, I maintain, is fueled by the principle that state action should not excessively trample upon privacy. Monitoring or reconstructing actions such as the trivial pursuit case would surely constitute a massive invasion of privacy. And if it is true that we value privacy because we value having control over personal information, then the injunction against privacy violation is in fact a concern for autonomy itself. Does this mean that the
A liberal state can never act against manipulation? Not if one can show that there are realistic ways in which manipulative intentions and knowledge can be retrieved without violating privacy. I want to argue that this condition is met when the intentions of potential manipulators are fixed by an organizational setting.

Specifically, if agents are members of organizations, then there are two ways – one direct and the other indirect – in which intentions and knowledge can be retrieved appropriately. Direct retrieval is possible because overall organizational strategies and the roles of its members in relation to that strategy are often accessible to observers. Members of the organization may have given interviews (to the press or researchers), strategy documents may be available and it is even possible that surveys have been conducted. Indirect retrieval is possible by deducing intentions and knowledge from statements or patterns of behavior. If no literal statements of intent or knowledge are available, then it may be possible to at least deduce from documents and interviews whether agents must have had specific intentions or had sufficient knowledge about the manipulative nature of their actions. Consistent patterns of behavior may also be good evidence of intentional behavior as well as, more importantly, specific types of knowledge that tasks require for their effective performance.

What is more, a concern for privacy also brings with it that the main objective of state measures against manipulation should not be to act against individuals (since this involves privacy violating monitoring), but to design policies that regulate or ban organizational practices that violate autonomy. That is, it must be shown that we can reasonably expect individuals in organizations to have manipulative intentions and knowledge in order to determine whether state policies against their organizations are in order.

Informed consent
Another case which suggests that our current conception of self-control is too demanding is the therapeutic use of hypnosis. Once under hypnosis, a client in such a session cedes self-control because she can no longer properly take either context-relevant or longer-term goals into account. And yet intuition unambiguously informs us that such influence is acceptable from the point of view of the liberal state. One explanation for this intuition refers to the well-being of the addressee. On this reading, the hypnosis is legitimate because it is in the interest of the client. But this explanation is insufficient, because therapeutic hypnosis cannot legitimately be applied at any one time (except for cases in which a subject has lost all contact with reality). Given this conditionality of our intuition, what seems required for hypnotic therapy (for mild clinical cases at least) to be legitimate from the point of view of the liberal state is proper consent of the patient. And because proper consent is an exercise of autonomy, protecting a citizen’s autonomy against a practice to which she has consented constitutes yet another case in which we are faced with a trade-off between (respect for) different aspects of autonomy.171

It is contested which properties consent must have in order for it to condone violations of the autonomy of the consentor. Surely, from the perspective of the liberal state there must be credible guarantees that a person who has consented to thorough manipulation has arrived at that decision without undue pressure or coercion. As a first step towards specifying these properties we could make a rough distinction between cases in which tacit consent is acceptable and ones in which explicit consent is required. We would also want to distinguish between the types of contexts in which consent to autonomy violation might be required. That is, consent can be given at the level of a message, at the level of the (institutional) context in which types of messages can be expected and society at large (with its legal framework).
I posit that there is an interaction between the centrality of covert influence and the appropriateness of either tacit or explicit consent. As I conceive of it, instances of consent to autonomy violating practices cluster in three groups according to their centrality. Firstly, futile cases of covert influence do not violate liberal autonomy by virtue of their lack of centrality. This includes cases such as unconsciously priming a person for games of trivial pursuit and the like, where other agents are completely in the dark about the covert influence but for which we still do not think that the liberal state has any role to play. Such cases are found predominantly at the level of individual messages. Secondly, there is a group of cases of covert influence that because of being somewhat more central requires tacit consent. Such cases would include movies, soaps and stage hypnosis. We can assume that literate viewers know that their emotions are aroused outside of awareness in such cases and hence that by entering a cinema or switching to a channel they have tacitly consented to such influence. In such cases the agent typically does not know the content of the future influence, its centrality nor the method by which it is wielded. In fact, such knowledge might annul much of the joy derived from the experience (e.g. of watching a movie). Besides these cases in which tacit consent is given at the level of (institutional) contexts (in which specific message types can be expected), one can also imagine tacit consent being given at the level of society.

One may argue that tacit consent to covert influence suffices in a liberal society that treats its citizens as responsible agents. But I think this argument is wrong, because of the danger of runaway manipulation. That is, tacit consent is a shady concept, and once the risk of abuse becomes too large it becomes implausible to say that tacit consent is a sufficient safeguard. Aggressive cults for example can always argue that even the most extreme long-term priming and hypnosis efforts could have been anticipated at some level by participants. In others words, if tacit consent would be the only type of consent required, then the risk of abuse becomes too serious. If this is correct, then we can distinguish a third group of cases that are highly central and therefore require explicit consent. Although I cannot draw a hard line in this regard, a central case would be religious rituals that use methods of implicit belief induction (e.g. bodily synchrony, music and evaluative conditioning). With regard to explicit consent as well, criteria can be more or less stringent depending on what the controlee is supposed to know. That is, a range of positions can be occupied ranging from thinking that highly central covert influence requires that the full swat of conditions is consented to (source, content, method, strength of effect), to thinking that carte blanche consent suffices (i.e. merely knowing that someone will somehow covertly change attitudes in a given context).

Let me illustrate the ambiguity of intuition in this regard on the basis of an episode of the Channel 4 program ‘Hero at 30,000 feet’ by the psychological illusionist Derren Brown. In this program, Brown had hundreds of volunteers screened for one of his new mental stunts. As in previous programs, these volunteers presumably all signed a contract (in previous shows this happened on screen, where Brown would tell the participant that ‘in the contract it says that we can do anything we like with you’). One of the candidates – Matt – was then followed for a month with hidden cameras without him knowing that he had been selected (nor did he know that his entire family was in cahoots with Brown). The objective of Brown was to select a candidate with a boring life that lacked meaning, and to inject purpose into it by means of unconscious influence (especially hypnosis). Thus, Matt was hypnotized in his sleep by Brown at night and instructed to take risks in his life, after which he was planted into dangerous situations as a form of (involuntary) behavioral fear-reduction therapy.
Now, here we have a case in which a person gives carte blanche consent to serious intrusions of his mental independence (at the end of the show we see Matt opting to buy a new house and looking for a new job, presumably because of him having gone through a number of terrifying experiences in hypnosis which included being tied to a railway when a train was approaching, taming a crocodile and landing a plane with dozens of passengers). I think that it is fair to say that it is not clear whether the liberal state should allow for such interventions, since there are intuitions on both sides. In order to account for the ambiguity in this and similar cases I propose to distinguish between what I will call the present-autonomy account and the balanced-autonomy account.

On the present-autonomy interpretation, carte blanche consent to central autonomy violation in branding will do for the liberal state. It requires no guarantees from consenters as a matter of liberal principle (except for absence of coercion and pressure) because the ultimate decision-making authority resides with the current instantiation of the agent. On the balanced-autonomy interpretation, by contrast, carte blanche consent to central autonomy violation will not do. From this perspective the default position of the liberal state is to protect all life slices of autonomy. The agent herself can choose to give up future slices of autonomy by allowing other agents to associatively change her values. But on the balanced-autonomy reading more stringent demands are made on the consent procedure, because future autonomy is allocated more weight. For this reason, the agent must not just know to whom she consents, but also to what she consents. That is, she must know the content of the influence, the method by which the influence is communicated and the extent of the influence. If one of these requirements is reasonable from a balanced-autonomy perspective, then carte blanche consent is insufficient. Which one of these two approaches to consent is the correct one is unexplored territory; answering it would require extensive normative analysis that I cannot hope to provide here. I therefore remain undecided as to which of these approaches to autonomy is the correct one.

Finally, with regard to the level at which explicit consent applies, we must note that the level of society at large is problematic because there is no right of entry in international law. If there is no possibility to enter another country, then there is also no possibility to leave your own country if you refuse to give consent. And consent without the option of rejection is meaningless. Furthermore, explicit consent at the level of individual messages is possible – from a practical perspective – only if these messages are limited in number. Thus, explicit consent to covert influence will be restricted in practice mainly to contexts in which message types can be expected.

4.2 Thresholds of Authenticity
The most pressing problem in finding a minimal interpretation of the notion of autonomy as authenticity is that associative processing bypasses initiatory mechanisms. I will here discuss three solutions to this problem and make a case for accepting the final one.

The first solution can be extrapolated from the work of Robert Noggle. He makes a distinction between desires that are formed on the basis of beliefs and those that are formed on the basis of quasi-beliefs. A quasi-belief is defined as a belief that lacks one or more of the four properties of a full-blooded belief: integration, formation by cognitive-epistemic means, possibility of introspection and the willingness to mentally assert the mental state in question. The kind of states that Noggle has in mind when speaking of quasi-beliefs are the ones brought about by conditioning. Rats, for example, can be conditioned to run if they are hungry, provided they are rewarded in the right manner. In a similar way, behavior therapists
can change behavior in a way that bypasses the patient’s awareness.\textsuperscript{175} Noggle next acknowledges that we cannot simply require that quasi-beliefs be rooted out of our cognitive system, because they are a constitutive part of it and fulfill important functions. Hence for him the task of a theory of autonomy is to stake out which quasi-beliefs are acceptable and which are not, and Noggle’s proposal is to say that an alien desire is one “that is generated by a quasi-belief whose content conflicts with that of a straightforward (possibly only tacit) belief.”\textsuperscript{179}

The advantage of this approach is that it defines an objective standard for authentic attitude formation: concordance with existing mental states. On the downside minimally two points must be mentioned. First, the standard of concordance between beliefs and quasi-beliefs may be unrealistic, because we noted that divergences between implicit attitudes (which are primarily affected by misattribution EC) and explicit attitudes towards the same object are endemic.\textsuperscript{180} Second, observers who want to determine whether conditioning violates the autonomy of an agent would need to know the full range of beliefs in order to establish whether the conditioned quasi-beliefs conflict with them. It is not clear how such knowledge could be acquired practically. For this reason, the liberal state cannot know when conditioning must be prevented.\textsuperscript{181}

A second way of dealing with associative processing in the context of authentic mental state formation is to say that the centrality of changed mental states determines the degree of authenticity loss. It is even conceivable that a reasonable threshold can be found in this regard. However, a problem for this approach is that there are cases of central mental state and behavior change that – despite their centrality – cannot plausibly be said to impair minimal authenticity because they are part of the human condition. In particular, unconscious processing is a structural part of both dyadic interactions and group life. A constitutive component of dyadic interactions is so called mirroring behavior, which generates rapport that is required to sustain comfortable trust-generating interactions.\textsuperscript{182} In this way attitudes are changed (about the valence of the interaction partner) while progressively intense forms of mirroring can even result in the mimicking of behaviors. All this occurs outside of awareness. A similar phenomenon occurs in the context of group behavior, where so called automatic goal contagion is facilitated by group membership. In one experiment, students who saw competitive behavior acted more competitively themselves in a follow-up task only if they were told (or naturally perceived) that the people who initially displayed competitive behavior were part of the same group (in this case the same university).\textsuperscript{183} Again, subjects had no awareness of this bias, while such effects can reasonably be assumed to lead to lasting attitude change. Surely, it would be absurd to suppose that the liberal state should act against interactions and associations on the ground that their effects are central to participants.

This brings me to the third solution, which is to say that minimal authenticity is only violated if the associative processing channel was activated by another person. This solution is able to explain why dyadic interactions and group life do not in themselves violate autonomy (as is the case on the centrality account) while it still captures much of the intuition that the effects of associative processing are problematic because they are not properly related to existing mental states. It may be objected that on the current definition authentic mental state formation is indistinguishable from self-controlled attitude formation. In response, two points must be noted. First of all, it is not clear why this would be an objection. As long as the conditions for proper mental state formation (for self-control and authenticity) are fueled by different intuitions it seems immaterial if they would turn out to have identical practical consequences. But secondly, authentic and self-controlled mental state formation are not identical. Violating self-control involves three things. The intention to bring about a certain state in a controlee, accurate predictions about the state that is actually brought about, and the
intention to use a message that prevents the controlee from knowing about the control. Violating authenticity requires only the intention to use a message that in fact bypasses initiatory mechanisms and some awareness that the message type in question is processed in an impaired way. Hence, there are cases of intentional attitude change in which one form of autonomy violation occurs but not the other. That is, a controller can take control of another agent in a way that does not bypass initiatory mechanisms (so that authenticity is not violated). And an authenticity reducer can bypass initiatory mechanisms without having accurate knowledge about the effect of the relevant message on the receiver’s behavior (so that self-control is not violated). Moreover, because there are fewer constraints on what counts as a valid intention to subvert authenticity, the retrievability and consent conditions are more modest than in the case of self-control. That is, what is to be retrievable and consented to in the case of authenticity is, respectively, the intention to use a message that bypasses initiatory mechanisms (not mindreading capacities) and some awareness that the message has effects that are associated with relevant processing channels (not the intention to bring about a certain mental state or behavior in another agent).

Once the authenticity requirement of autonomy is said to depend on the intentions of other agents, there are also better prospects for defining thresholds for authentic action. With regard to the retrieval of explicit self-related states we could formulate the negative requirement that other agents may not deliberatively increase the cognitive load of another agent. (Because in the case of interpretation cognitive load would bring down the level of action identification, and in case of reflection it could reduce access to explicit self-related states). With regard to the retrieval of implicit self-related states, a feasible negative requirement would be that other agents should not induce negative affect. The success of this modification in formulating practical and retrievable thresholds is mixed however. Making cognitive load contingent on it being induced does not make much of a difference, since it is not clear in which situations such inductions should be allowed (and how much). The prospects for the negative affect induction requirement look much better though, because one can in principle test whether it is correlated with false self ascriptions (FSAs).

This completes my analysis of the minimal components of autonomy, which can now be pieced together. An autonomous agent is self-controlled and authentic with regard to both action and mental state formation. An agent acts in a self-controlled way if her system 2 reasoning system is not impaired, if her goals are not unconsciously activated and if she exhibits appropriate impulse control. Self-controlled mental state formation is guaranteed if the agent is not deluded and if she is aware of the intention on the part of a controller to evoke a specific behavior (and takes this intention into account in processing a message). An agent acts authentically when she has appropriate access to her implicit and explicit self-related states, which is secured by the absence of negative affect and cognitive load induction. Finally, authentic mental state formation is guaranteed if initiatory mechanisms related to conflict detection are not bypassed intentionally by another agent. For influence attempts to violate autonomy with regard to both self-control and authenticity it must be true that the effects of influence are central, that the relevant intention (which is not the same for self-control and authenticity) of the influencer is retrievable and that relevant consent (which is also not the same for self-control and authenticity) has not been given.

It is my contention that this framework is maximally integrative of intuitions about liberal autonomy under constraints of consistency and realism. In order to falsify this claim it need be possible only to add either an interpretation of one of the seven core intuitions or a mechanism that tells us in what these intuitions consist, without this addition being unreal-
istic or causing inconsistency. Throughout this chapter I have argued that at least conative hierarchy, evaluative reflection, the authoritative point of view of the acting agent, substantive theories about the self and excessively intellectual theories of action fail to do so. If these are the most likely candidates for falsifying the claim to maximal inclusiveness, and if my identification of the seven core intuitions of liberal autonomy is acceptable, then it is reasonable to assume that at least something like my minimal theory of autonomy meets my design specifications. If it is also accepted that my design specifications are an appropriate safeguard against perfectionist worries, then this theory supports anti-perfectionist autonomy-minded liberalism. Finally, with regard to the subsidiary aim of this chapter I think that it is fair to say that my theory is sufficiently rich in empirical detail to allow for the concrete assessment of autonomy violation in society. But whether this is indeed the case can be established only in the course of such an assessment, which I am about to carry out in the upcoming chapters.

Let me close with a conditional statement on the value of liberal autonomy. Even if autonomy violations – say, in the form of independence violations – are sufficiently central to the receiver, if they can be retrieved in the right way and if they have not been consented to, it need not follow that the liberal state is to take measures against them. For that argument makes the problematic assumption that autonomy is the supreme value that trumps all other values. Instead, appropriately context sensitive political philosophy must carefully weigh moral, ethical, prudential and realistic concerns for every policy or policy area. Hence, instances of autonomy violation are only pro tanto problematic.