The Despotic Fragment:
A critique of the traditional account of the relationship between the different internal principles of action and the various parts of the soul

Julian Urrutia

It is the conceptions of ourselves that are most important to us that give rise to unconditional obligations. For, to violate them is to lose your integrity and so your identity, and to no longer be who you are. That is, it is to no longer be able to think of yourself under the description under which you value yourself and find your life to be worth living and your actions to be worth undertaking. It is to be for all practical purposes dead or worse than dead.

-Christine Korsgaard [1]

The path to self-discovery is one which today's youth are very concerned to travel. The tropical paradises of the world are populated by beach-bums who 'found themselves' on their sandy shores. But my grandfather, and many of his generation, never quite understood how it could be possible that one should discover, or 'find', oneself (or rather, how it could be possible that one should *not* do so). And it must be conceded that the manner in which the topic is usually discussed begs the question. My grandfather never accepted the notion, which now has become common knowledge, that one needs to actively discover one's 'self'. How could one possibly loose one's 'self' such that one should have to find it? How could one have reached the age of reason without a clear idea of who one is?

My grandfather was born before the psychological revolution that began with Freud's theory of the subconscious, which might explain why he never empathized with these concerns. However, the idea that we have repressed desires, impulses and inclinations, whose existence we ignore, and that influence how we act and make decisions, certainly legitimizes much of the prevalent preoccupation over the character of our agential identities. Unfortunately, exploring Freud's neatly dichotomized mind might prove more disconcerting than one might hope. We often do not identify with that urge we discovered,

1. Korsgaard, Christine. *The Sources of Normativity*. Cambridge: Cambridge University Press, 2004.

or with that desire we didn't know we had—they seem external to us; we do not feel they are expressive of who we really are. So we might come to the conclusion that the process of finding who we *are* necessarily involves a process of deciding who we *are not*—for through our introspection we might discover more than our self. The problem thus becomes one not only of self-discovery, but also of self-formation. We are faced with the problem of having to decide which of the many internal principles of action that motivate our behavior are part of what we are as agents, and which are not—and this is no small problem.

But the problem is more than just an academic one. It is not just a matter of wanting to understand ourselves because it a meaningful goal in itself—it is also a matter of much more profound significance. Too many important aspects of our lives—ranging from our professional reputation, and our position in our networks of friends and family, to our religious and cultural affiliations—are irrevocably and unequivocally linked to our agential identities. In *The Sources of Normativity* (1996), Korsgaard gives us an analytical account of the enormous implications that our practical identities carry with them. She argues that our practical identities are the source of our normative commitments.

The conception of one's identity in question here is not a theoretical one, a view about what as a matter of inescapable scientific fact you are. It is better understood as a description under which you value yourself, a description under which you find your life to be worth living and your actions to be worth undertaking. You are a human being, a woman or a man, an adherent of a certain religion, a member of an ethnic group; someone's lover or friend, and so on. And all of these identities give rise to reasons and obligations. Your reasons express your identity, your nature; your obligations spring from what that identity forbids.

As such, my grandfather's position regarding the process of self-identification does not seem like it should be counseled. Instead, it would perhaps be advisable that one should take pains to form a self with which one is comfortable and satisfied, since our normative commitments are inseparable from our practical identity as an agent. So the question

becomes about how it is that we navigate this self-formation that results in an identity which is, according to Korsgaard, "the source of normativity."[2]

Harry Frankfurt believes that our capacity for introspection, by which we are able to confront and evaluate our motives and desires—which is what gives rise to the problem in the first place—also provides the solution. The reflexive structure of our mind, by virtue of which we are self-conscious, not only forces us to confront and evaluate those principles by which we are moved to action; it further allows us to be active with respect to those very principles. In so far as we are the kinds of beings who can reflect on our appetites, passions, reasons and desires—we are also the kinds of beings who can decide which of those motives truly constitute what we are; we can decide whether any particular disposition in question is a characteristic with which we identify, and thus actively incorporate into ourselves, or whether we reject it and exclude it from our agential identity. Frankfurt would therefore have us believe that it is through a conscious exercise of our will that we come to create our agential identity. Korsgaard seems to share a very similar view. There is a strong affinity between her strategy, which involves a process of 'reflective endorsement' of our motives, and Frankfurt's treatment of the question at hand.

I do not believe Frankfurt and Korsgaard's theory of identification is correct. Their error is consistent with a long-standing but misguided view of the structure of the human soul. It results from assuming a fallacious relationship of authority between our conscious, deliberative faculties and the other internal principles of action. I reject the notion that reflection stands in a privileged position of prerogative over the other internal principles of action as the most legitimate source of our agency. I maintain that the philosopher's habit of deferring to conscious deliberation before and above any other of our internal springs is unfounded. There is a mounting body of empirical evidence that suggests that no amount of reflexivity has the capacity to unify our agency such that we can create 'self' through reflexive self-evaluation in the way Harry Frankfurt envisions. Rather, I believe the evidence suggests that any self-conception at which we arrive through reflection will fail to

2. Korsgaard, *The Sources of Normativity*.

be representative of the entire quality of our agency, and further, it might estrange us from aspects of our agency which we might fully endorse in different contexts or under different circumstances.

I will first review the significance and the fundamental characteristics of Frankfurt's theory of identification, including Michael Bratman's expansion on it, in order to obtain a picture of the kind of agents they would have us believe we are. Then, in an effort to produce a comprehensive argument, I will begin my own exposition with a systematic account of what it means to be one kind of agent as opposed to another—of what I believe are the correct criteria by which we arrive at these distinctions. From there, I will proceed to review a body of empirical evidence which suggests a picture of human agency that discords with Frankfurt and Korsgaard's perspective. This will lead us to my view of the structure of our cognitive architecture, which will allow us to make sense of that odd body of empirical evidence regarding the actual character of our agency.

Identification, Freedom and Normative Agency

Appetites, passions, affections, and the principle of reflection, considered merely as the several parts of our inward nature, do not at all give us an idea of the system or constitution of this nature, because the constitution is formed by somewhat not yet taken into consideration, namely, by the relations which these several parts have to each other; the chief of which is the authority of reflection or conscience.

-Bishop Joseph Butler[3]

In his sermons, Bishop Butler articulates a view of human nature that can be traced at least back to Aristotle, and that has been—and still is—the most prevalent view of our motivational structure. I will call it 'the traditional view' and it is the view which I intend to challenge. This view of the 'soul', if you will, is biased in favor of one of our internal principles of action, viz. the 'principle of reflection', which is given a prerogative over all the

3. Butler, Joseph. *Five Sermons*. Indianapolis: Hackett Publishing Company, 1983.

others as if the legitimacy of its authority was completely obvious. Although Frankfurt's and Korsgaard's views are, in my opinion, more aligned with the facts than the traditional view, they nonetheless beg the question by granting our capacity for reflexive self-evaluation an unwarrantedly privileged position. And the problem with this assumption regarding the status of our capacity for self-evaluation is that it allows us to arrive at fallacious solutions to several important problems, such as those regarding the freedom of our wills or the sources of normativity. Thus, a challenge to the traditional view amounts to a challenge to the solutions for those problems that follow from it.

Freedom of the Will and Frankfurt's View

Let's begin with Frankfurt's take on the subject. He begins with the reality that people in general care about what kind of persons they are. And the truth of the matter is that we do, generally, care very deeply about what moves us to action—about what motives, desires, inclinations, or reasons (etc.) underlie our behavior. To use Frankfurt's language, "It matters greatly to us whether the desires by which we are moved to act as we do motivate us because we want them to be effective in moving us or whether they move us regardless of ourselves or even despite ourselves."[4] According to Frankfurt, it is as if we were not in control of our own behavior—as if our actions were being determined by something that is not a part of ourselves such that we become passive observers of our own behavior. Thus, he argues that the problem of who we are can also be construed as problem of free will. Frankfurt reminds us that "according to one familiar philosophical tradition, being free is fundamentally a matter of doing what one wants to do."[5] But under this account, it is impossible to distinguish between an agent who *acts* freely, and one whose *will* is free. As such, it is important to draw a distinction between a free action and free will. Frankfurt

4. Frankfurt, H. "Identification and Wholeheartedness." In: *The Importance of What we Care About*, by Harry G. Frankfurt. Cambridge: Cambridge University Press, 1988.
5. Frankfurt, H. "Freedom of the Will and the Concept of a Person." In: *Free Will*, edited by Gary Watson. Oxford: Oxford University Press, 2003.

defines the will as that desire which effectively moves a person to action.

He therefore states:

> Freedom of action is (roughly) the freedom to do what one wants to do. Analogously, then…freedom of the will means (also roughly) that [the agent] is free to want what he wants to want. More precisely, it means that he is free to will what he wants to will, or to have the will he wants. Just as the question about the freedom of an agent's action has to do with whether it is the action he wants to perform, so the question about the freedom of his will has to do with whether it is the will he wants to have.[6]

As such, the free will problem is not one that all agents have to face. Only those who can have desires regarding their desires—those who can have second-order desires—face this problem. Only those agents who can, as it were, reflect on their desire to act this way or that, and evaluate them from a distance, face the problem of finding that they are moved to act by forces that they do not endorse. In other words, from Frankfurt's perspective, the problem of free will presupposes an identity with which some desires can conflict. And as we shall see, so does the problem of normativity. Thus, it becomes apparent that the path to self-discovery takes us far beyond the beaches of Tahiti into the very depths of what it means to be, qua agents, human. It takes us down to those fundamental qualities that distinguish us from any other type of agents.

Frankfurt's Theory of Identification

It is these acts of ordering and of rejection—integration and separation—that create a self out of the raw materials of inner life. They define the intrapsychic constraints and boundaries with respect to which a person's autonomy may be threatened even by his own desires.

-Harry Frankfurt[7]

6. Frankfurt, "Freedom of the Will and the Concept of a Person."
7. Frankfurt, "Identification and Wholeheartedness."

In "Freedom of the Will and the Concept of a Person" (1971), Frankfurt claims that,

> Many animals appear to have the capacity for…'first-order desires', which are simply desires to do or not to do one thing or another. No animal other than man, however, appears to have the capacity for reflective self-evaluation that is manifested in the formation of second-order desires.[8]

In his terminology, a second-order desire is an intentional state about a first-order desire; it is, for instance, a desire to have a desire. Under Frankfurt's account, the relevant kind of identification requires more than just second-order desires—it is not enough to simply want to have a certain desire; we must want a certain desire to be our will. He calls these kinds of second-order desires 'second-order volitions'. It is not enough for an agent who cares about her will to want to have certain desires;, she needs to want that those desires to be effective. Frankfurt argues that we identify with a desire by endorsing it through a second-order volition that it be our will. This second-order volition can be understood as a commitment to include that desire as one of the internal principles of action that characterizes our agency. Unfortunately, the matter is not so simple; we are motivationally more complex that this sketchy picture of our soul suggests. Just as we are many times conflicted or ambivalent regarding our first-order desires, we can be similarly ambivalent regarding our second-order desires. Just as we are not always sure about which desires we should act on, we are not always sure about which desires we would like to have—we are not always sure about which of our desires we endorse, which ones we tolerate, and which ones we reject.

So what happens in the case of conflicting second-order desires? Must we now move to a higher level and adopt a volition of the third order regarding our second-order desires? And what if there is similar tension at this level? How can we avoid the threat of infinite regress without just cutting it off arbitrarily? In the 1971 essay, Frankfurt tries to deal with this concern by stating that a person can identify herself 'decisively' with one of

8. Frankfurt, "Freedom of the Will and the Concept of a Person."

her first-order desires such that her commitment to it 'resounds' throughout, and hence terminates, the potential regress. But as Gary Watson points out,

> We wanted to know what prevents wantonness with regard to one's higher order volitions. What gives these volitions any special relation to 'oneself'? It is unhelpful to answer that one makes a 'decisive commitment', where this just means that an interminable ascent to higher orders is not going to be permitted. This *is* arbitrary.[9]

Thus, in "Identification and Wholeheartedness" (1988), Frankfurt gives a more complete account of how this problem can be dealt with by appealing to an example of an individual engaged in an arithmetic calculation. Having arrived at an answer, this person verifies it with a second calculation, but it is possible that both calculations are faulty. This person faces a similar series of calculations that might extend *ad infinitum*. One way to resolve the situation is to simply quit and allow the result of the last calculation to serve as the answer. In this case the agent is passive with regard to the resolution of the conflict; in fact, the conflict has not really been resolved but merely abandoned or ignored. A more conclusive way to solve the situation is to make an active decision to adopt that solution as the answer. This is possible if, for example, the agent is fully confident that she has arrived at the correct answer. In this case, the future is clear to her and her decision to adopt this result as the answer resounds throughout the threatening regress because her confidence allows her to "anticipate the outcome of an indefinite number of possible future calculations."[10] But even in the case where she is not fully confident about her results, she will be capable of deciding in favor of her answer, without being arbitrary, if she feels she has reason to believe that any future results will not conflict with those at which she has already arrived. In other words, absent complete confidence in the correctness of her results, if the agent nonetheless does not find a disturbing conflict between those results, and between those results and any results at which she might reasonably expect to arrive given further calculations—she

9. Watson, Gary. "Free Agency." In: *Free Will*, edited by Gary Watson. Oxford: Oxford University Press, 2003.
10. Frankfurt, "Freedom of the Will and the Concept of a Person."

is left without a reason to engage in any such further calculations. Thus, her decision to commit to her answer is not arbitrary by any means. Similarly, a person reflecting on her desires—either because they conflict with each other or because a general lack of confidence casts doubt on her satisfaction with them—can avoid the regress by making a decisive commitment to endorse one of them. If she finds there is no conflict between the results already obtained by her evaluation, or between these results and any she might reasonably expect to obtain through further evaluations, the agent is left without a reason to continue evaluating her desires and can decide to endorse one or the other. Terminating the evaluative regress at this point can hardly be deemed arbitrary. And through this decision to cut off the process of forming desires of increasingly higher orders, our agent determines what she really wants because it allows her to separate the desires she endorses from those she does not—the decision makes that desire which she has chosen to endorse fully her own, it makes her *identify* with that desire.

Frankfurt and Korsgaard's Agent

A lower animal's attention is fixed on the world. Its perceptions are its beliefs and its desires are its will. It is engaged in conscious activities, but it is not conscious of them. That is, they are not the objects of its attention. But we human animals turn our attention on to our perceptions and desires themselves, on to our own mental activities and we are conscious of them. That is why we can think about them. And this sets us a problem no other animal has. It is the problem of the normative.

-Korsgaard[11]

In "Identification, Decision, and Treating as a Reason" (1999), Michael Bratman fills in the possible gaps Frankfurt leaves by giving a systematic account of what it means to make a decision by which we identify with our desires. Bratman argues that we come to identify with a desire when we decide to treat it as reason-giving. It involves a decision

11. Korsgaard, *The Sources of Normativity.*

about whether or not to count a desire as a reason in our deliberation. Thus, the decision by which we identify with a desire involves deciding to treat the desire as reason-giving, and being satisfied with that decision in the sense of it not being in conflict with other standing decisions about which desires to treat as reason giving.[12] In doing so, we endorse that desire and make it part of who we are. A full description of ourselves would have to include a commitment to pursue the satisfaction of that desire, since we have decided to treat it as reason-giving.

According to Bratman, 'reasons' give us ends that we feel we are justified in pursuing; 'reasons' justify the relevant means we take towards an end.[13] But they play more than simply a justificatory role; as Korsgaard points out, the word 'reason' is a normative word. As such, we are more than simply justified in pursuing an end which we believe we have reason to pursue—we are committed to its pursuit. Reasons commit us in the sense that not behaving in the way we have reason to behave would make our actions unintelligible both to ourselves and to third parties. A chess master playing to win a tournament has a reason to make a certain move if he realizes it will put his opponent in checkmate. That fact not only justifies his making the move—it makes it so that any other move would be impossible to explain; it would be incoherent and unintelligible. Thus, through the same process by which we come to identify with our desires, we find that we become obligated to them. So by the same process through which we create a practical identify for ourselves, we come to create commitments for ourselves—we come to obligate ourselves.

Thus, the same faculty which according to Frankfurt makes us into the kinds of agents who *can* have a free will, for whom freedom of the will can be a problem by virtue of being able to make our desires the objects of our attention, also makes us into the kinds of agents who impose duties and obligations on ourselves. The process of identifying with our desires through reflexive self-evaluation also makes us commit to the satisfaction of those desires because, as Korsgaard puts it, our "identities give rise to reasons and obligations. [Our]

12. Bratman, M. "Identification, Decision, and Treating as a Reason." In: *Faces of Intention: Selected Essays on Intention and Agency*, by Michael Bratman. Cambridge: Cambridge University Press, 1999.
13. Bratman, "Identification, Decision, and Treating as a Reason."

reasons express [our] identity, [our] nature; [our] obligations spring from what that identity forbids."[14] Therefore, our capacity for reflexive self-evaluation makes "obligation in general a reality of human life."[15]

The argument is certainly compelling. As Frankfurt himself points out, his account of identification allows for a theory concerning freedom of the will that meets all the essential conditions which such a theory should meet: it accounts for our disinclination to adjudicate this freedom to any other species than our own, and it makes it apparent why this kind of freedom is desirable. According to Frankfurt, a person who is free to do what he wants to do, and to want what he wants to want, enjoys "all the freedom it is possible to desire or to conceive."[16] It also allows Korsgaard to ground the elusive source of normativity in a way that is consistent with that freedom. It allows her to provide a view of human agency that is simultaneously autonomous and normatively constrained. In other words, Frankfurt and Korsgaard's appeal to our self-consciousness allows them to simultaneously establish the source of the characteristics which we believe are epitomic of human agency: freedom, autonomy, and morality.

Unfortunately, if our capacity for reflexive self-evaluation really *is* the source of our freedom, our autonomy, and our morality, it is not at all clear to me that we really *are* free, autonomous and moral agents—nor that we would want to so be. I am of the opinion that we are much less self-conscious than we would like to be, and that any identity we might create through a process of reflexive endorsement will necessarily leave out aspects of our agency that will eventually be effective in governing our behavior, and that we would endorse if we could. However, I do not believe the problem is simply one of epistemic limitations. My claim is much stronger than that: my claim is that even if we did have conscious access to all the aspects of our agency, a process of reflexive endorsement would necessarily estrange aspects of our agency that will, again, eventually govern our behavior, and that we would fully endorse given the appropriate circumstances or context. As such,

14. Korsgaard, *The Sources of Normativity.*
15. Korsgaard, *The Sources of Normativity.*
16. Frankfurt, "Freedom of the Will and the Concept of a Person."

not only will we not be free much of the time, nor will we be autonomous or moral—we will not want to be any of these things because that identity which *would* be the source of our freedom, autonomy, and morality will always be more a product of circumstance than of our volition, and it will therefore necessarily estrange aspects of agency which, *mutatis mutandis*, we would fully endorse as being a part of ourselves, and which we would desire that they be effective in governing our behavior given the appropriate circumstances or the appropriate context.

To establish my claim, I will rely mostly on empirical evidence that shows how remarkably ineffective, inconsistent, and self-deluded our conscious agency really is. I will use this evidence to show that we are *not* the kind of agents Frankfurt and Korsgaard would have us believe we are. I will then give an account of my view of the structure of our cognitive architecture, which I will use to argue why we *cannot* be that kind of agents. Why we would not *want* to be those kinds of agents will be self-evident given that my view of the structure of our cognitive architecture is indeed accurate. However, it is worthwhile to indulge in a quick digression about the criterion by which we distinguish between different kinds of agents so as to avoid any potential misunderstandings about what I mean when I talk about 'different kind of agents or agencies.'

Human Agency

The Constraint-Conforming Approach to Agency

To be an intentional system, and therefore to qualify as 'minded' in some minimal sense, is, on standard approaches, to be a system that is well-behaved in representational related respects. The well-behaved system represents things as they appear within the constraints of its perceptual and cognitive organization. And it acts in ways that further its desires…in the light of those representations or beliefs. An intentional system may not be perfectly behaved in these action-

related and evidence-related ways, but it will have to attain a certain threshold of rational performance—and perhaps do so as a result of a certain history or organization—if it is to seem like it is minded at all.

-Victoria McGeer and Philip Pettit[17]

According to the Stanford Encyclopedia of Philosophy, an agent is one who performs activity that is directed at a goal. Thus, agency can be understood as the capacity for goal-directed activity. This definition is perhaps too broad if one intends to address human agency, but I want to start with the most general understanding of what constitutes agency and build my way up to a precise understanding of what constitutes human agency—what distinguishes human beings from other agents. However, I am willing to make one pre-analytical claim about human agency that I don't believe will be disagreeable: human beings are very sophisticated agents. Compared to a cow, a pigeon, or even Deep Blue (the computer that beat Kasparov at chess), human beings are much more complex, refined and multifarious agents. But it is important to distinguish between the faculties that support a certain kind of agency, the capacity to exercise that agency, and the actual character of the agency that is exercised.

As I mentioned very briefly in the introduction, I believe that the mind, which I will argue is embodied in our cognitive faculties, is the very source of our agency. A good understanding of the structure and the functioning of our mind is therefore requisite in any discussion about agency. I began this section with a quote from McGeer and Pettit's "The Self Regulating Mind," where they articulate what they call 'the constraint-conforming approach to the mind,' which I have borrowed to suit my purposes. We ordinarily think of the difference between human beings and other animals in positive terms. We think of the things that we humans can do that other agents cannot—we think of the difference in terms of being differently able. Another way of approaching the subject would be to think of the difference in negative terms—in terms of being differently constrained. I think both

17. McGeer, Victoria; Pettit, Philip. "The Self Regulating Mind." In: *Language & Communication*, 22 (2002) 281-299.

approaches are ultimately equivalent, but it will suit my purposes better to use the latter approach. I will argue that we should distinguish between different kinds of agents on the basis of the constraints that limit the manner and the extent to which they exercise their agency.

It is almost trivial to state that there are constraints to agency. An agent locked in a small room has her capacity for activity effectively constrained. She might be similarly constrained by non-physical forces, like strong social pressures or her family's expectations. However, most constraints are not exogenous to the agent in this way. Agency can also be constrained by physical, perceptual, epistemic, and motivational limitations (to name only a few) that are themselves endogenous to the agent. As McGeer and Pettit point out in the quote above, some of the most significant constraints on our agency are representational. The more accurately and extensively an agent can represent his environment, the more nuanced, precise, and sophisticated his agency. Of course, different agents face different constraints, and similar constraints might limit different agents to varying degrees. In other words, the constraints that limit the extent to which agency can be exerted vary in degree as well as in kind. A very myopic individual faces a constraint to her agency which, *ceteris paribus*, someone who is only moderately afflicted by the same condition does not face; the difference between the constraints these individuals face is one of degree. In contrast, an arachnophobe faces a constraint which, *ceteris paribus*, a dendrophobe does not; these constraints are different in kind. Though it is perhaps an obvious distinction, I believe it has profound implications on our discussion.

The very myopic individual is less able to discriminate and attend to the relevant stimuli in her environment; she is less *sensitive* to the relevant stimuli than the other individual. As such, the scope of her agency is more limited because she is less able to respond precisely to the relevant constraints. On the other hand, the arachnophobe might be just as sensitive to spiders as the dendrophobe is to trees; they are simply sensitive to different kinds of stimuli as constraints. As such, given the right proportions of spiders and trees, the scope of their

agency could be equally limited, but each identifies very different limiting constraints; they each find very different stimuli to be relevant. This is significant because, far and wide, our phobics will face different constraints in any given circumstance; the degree to which they are limited will depend on considerations of how and when. There is nothing we can know about the extent to which they will be limited without specifications about the particular situation or context. The case of our myopes is different; we know that one will be more limited than the other under any set of conditions. This is significant in that it allows us to distinguish between different types of agents. Agents that face similar kinds of constraints might be said to be the same type of agents, even if they can be further differentiated by the degree to which those constraints limit the extent to which they can engage in goal-directed activity. Human agency is such that we face many kinds of constraints, and are very sensitive to individual instances of each kind of stimuli. We are the type of agents who are most sensitive to the biggest range of constraints.

But perhaps the question of how to distinguish between different types of agents is not so simple. It is also true that agents will differ in how they stand in relation to the same constraints. For example, an agent might be capable of exerting its agency adaptively given the constraints that it faces, without being aware of the constraints themselves. In contrast, an agent might be explicitly aware of the limits to its agency, of what constraints it faces, and as a result be capable of responding to them deliberately. An agent that is explicitly aware of the constraints it faces might not only be able to exercise its agency more deliberately with relation to those constraints, it might further be able to exercise its agency in a way that deliberately alters those constraints; it might be able to affect the limits to its own agency.[18] One could imagine agents that face the same constraints but differ in how they stand in relation to them in this way. Human beings are the kinds of agents that can affect the constraints that affect their own agency.

I believe the first approach is most correct, but perhaps someone might object on the grounds that this method for distinguishing types of agents is not sufficiently refined to

18 . McGeer & Pettit, "The Self Regulating Mind."

allow for distinctions of more than academic relevance. The critic might argue that agents might face all the same constraints, and yet differ in how they stand in relation to those constraints. If the only criteria by which we distinguish agents are the kinds of constraints they face, then there is an important sense in which we might not be able to distinguish between Gary Kasparov and IBM's Deep Blue. But there *is* a very significant difference between the two Chess World Champions, and this difference lies exactly in how they stand in relation to the constraints they face. An agent who can attend to the constraints that it faces can act deliberately *in* response to the constraints themselves, while an agent who cannot attend to those constraints, who is unaware of what those constraints are, can only act *as* a response to them. An agent who can attend to its constraints can direct its actions at the constraints themselves—it can deliberately direct its actions at the constraints themselves. An agent who is unaware of what constraints it faces—one who cannot identify those constraints *as* constraints, cannot deliberately make those constraints the object of its intentional action. It cannot intentionally alter what constraints it faces. Thus an agent that can intentionally alter its constraints, such as Kasparov, can potentially change both the degree to which the constraints limit the scope of its agency, as well as the kind of constraints it faces. Deep Blue, on the other hand, does not have this option; it will always act in response to the chess-playing algorithms that constitute the source of its agency, and *having that option* is an important difference in the kinds of agency which Kasparov and Deep Blue can exercise.

But this admittedly intuitive argument is fallacious. Let's go back to our myopes and our phobics: by identifying that the quality of her vision limits the scope of her agency, our myope can choose to buy glasses with the intention of changing the degree to which she is constrained my her myopia. Similarly, our dendrophobe can choose to begin some kind of therapy to eliminate his irrational fear of trees. But notice that their capacity to attend to the constraints by itself does not necessarily change the kind of agents these persons are. It perhaps changes the kind of agent the person *can* be, but the capacity itself doesn't change

the kind of constraints she faces; it does not change whether any particular kind of stimuli actually constrain her agency. If the myope doesn't buy glasses, and the dendrophobe doesn't go to a shrink, they will continue to be the kind of agents that are respectively nearsighted or fearful of trees. Similarly, though Kasparov can choose not to be limited by the same constraints the limit Deep Blue's agency, if he doesn't choose to exercise that capacity but instead chooses to spend the rest of his life playing chess against Deep Blue, he will effectively be no different from the computer.

So it is not the case, as the critic believed, that agents should be classified according to how they stand in relation to their constraints. Instead, it is more in line with our current intuitions to classify them according to the kinds of constraints which effectively limit their agency. Our capacity to attend to our constraints—including our motivational, passional, appetitive, or otherwise intentional constraints—allows us to distinguish ourselves in a theoretical sense, but it is not sufficient to distinguish us in the relevant, practical sense. So what Frankfurt and Korsgaard have articulated really refers to the kinds of agents human beings can be—not to the kind of agents that we *are*. Though they are both correct when they assert that our capacity for reflexive self-evaluation presents us with the problems of freedom of the will, autonomy, and normativity—it is not at all clear that it is also true that this capacity allows us to resolve the problem.

Part of the appeal of Frankfurt, and Korsgaard's account is that it seems to encompass the most fundamental and epitomic aspects of human agency: freedom, autonomy and normartivity. It is by virtue of our capacity for reflective self-evaluation, through which we create the identity that characterizes us, that we are the kind of agents that are free, autonomous and moral. But it is worth wondering whether we are ever truly characterized by that identity at which we arrive through the process of reflective endorsement. Whether this identity, which Frankfurt and Korsgaard would have us believe characterizes our selves, is ever an accurate description of the kind of agents that we are. I believe the question regarding the true character of our agency remains open, but under the constraint-

conforming approach to agency, the question becomes an empirical one. So let's turn to the evidence so that we may consider whether we can ever be the kind of agents that, upon reflection, we would like to be. Let's explore whether human beings generally behave in a way that is consistent with the idea that we are characteristically self-conscious beings

The Pre-Conscious Aspect of Our Agency

The difference between the instrument called philosopher and the instrument called clavichord [is that] the philosopher-instrument is sensitive, being at one and the same time player and instrument.

-Denis Diderot[19]

I choose the term 'pre-conscious' not only to avoid the Freudian baggage associated with concepts like the 'unconscious', but also because I think it is more appropriate given the structure of our mind—but we will get to that later. In any case, I think it is best to approach this subject from the angle pre-conscious agency in general. I will therefore begin by considering the degree of sophistication of which an agent that is not conscious is capable of. By conscious, I mean of being capable of reflexive self-evaluation, and therefore self-aware in the sense of being aware of its own deliberation. I find that McGeer and Pettit's distinction between routinized and self-regulating minds is a good point of departure:

> A routinized mind, be it animal or robot, will produce actions in an intentional or voluntary way…But the system will not act intentionally with regard to how far it conforms to evidence-related and action-related constraints as such; it will not even have beliefs as to what those constraints are or require. Its conformity with the constraints will happen by courtesy of nature or nurture; it will not be intentionally achieved or intentionally reinforced.

In contrast, a self-regulating mind is one that can discriminate and attend its own intentional states as such; it is explicitly aware of the content of its intentional states, and

19. Diderot, D. "d'Alemberts Dream." London: Penguin Books, 1966.

therefore is in a position to evaluate them and to act with the intention of altering them. That this is simply a more formal account of the same claims we have been examining all along becomes evident when we recall Korsgaard's claim that,

> A lower animal's attention is fixed on the world. Its perceptions are its beliefs and its desires are its will. It is engaged in conscious activities, but it is not conscious of them. That is, they are not the objects of its attention. But we human animals turn our attention on to our perceptions and desires themselves, on to our own mental activities and we are conscious *of* them.

So given a system with relatively simple cognitive faculties, it will be able to adjust to incoming information in a relatively faithful manner, thereby generating a representation that allows it to behave in an adaptive manner in relation to the constraints it faces. The point is that all of this is possible without the system being reflexively aware of its own intentional states in general. It need not be aware of its own representations or beliefs; it need not be able to act intentionally with regard to how it conforms to those constraints; it does not even need to represent what those constraints require. As long as the system is endowed with cognitive faculties that allow for sufficiently faithful representations, the agent will be able to act in response to them and behave adaptively. But its behavior will be automatic and unreflective; it will be a routinized agent.

This description seems to fit the mind of relatively simple animals. Honeybees (*Apis mellifera*) for example, will invariably follow a vector derived from simple dead-reckoning that would have taken them from their hive to a feeding station, even if they are captured as they leave the hive and are displaced to a different location.[20] Their behavior is governed by a routinized mind that unreflectively conforms to the beliefs it has about the constraints

20. Menzel, R., Greggers, U., Smith, A., Berger, S., Brandt, R., Brunke, S., et al.. "Honey bees navigate according to a map-like spatial memory." Proceedings of the National Academy of Sciences 102 (2005), 3040-3045.; De Marco, R.J. & Menzel, R.. "Learning and memory in communication and navigation in insects." In: *Learning Theory and Behavior Vol. 1 of Learning and Memory: A Comprehensive Reference*, 4 vols. (J. Byrne Ed.). Oxford: Elsevier; (2008) p 477-98. ; Menzel, R., Greggers, U., Smith, A., Berger, S., Brandt, R., Brunke, S., et al.. "Honey bees navigate according to a map-like spatial memory." Proceedings of the National Academy of Sciences 102 (2005), 3040-3045.; Menzel, R., De Marco R. & Greggers, U. "Spatial memory, navigation and dance behaviour in Apis mellifera." In: *Journal of Comparative Physiology* A 192 (2006), 889-903.

it faces. It is so routinized that the bees do not even respond to counterfactual evidence about the fidelity of those representations available to them as they travel along the false vector. Interestingly enough, however, once the bees have travelled the length of the false vector, they reorient themselves and return to the hive (or the feeder) along a straight homing-vector which they had probably never travelled before, suggesting that they also have a Euclidian map-like representation available to them which they only deploy *after* their initial route memory proves erroneous.[21] These findings show a clear case of an animal that automatically represents its spatial location in the environment in two different ways, and deploys them hierarchically when navigating. This strongly suggests that these navigating systems are supported by largely independent cognitive processes; or rather, by one independent cognitive process which only allows the animal to navigate according to dead reckoning, and a second which generates a richer, map-like representation that allows the animal to be more flexible in its behavior. The system that allows for the dead-reckoning causes the agent to follow a rigid, memorized vector—the other allows the agent to navigate according to a Euclidian map-like representation, which requires robust interaction between landmark, map-like based information and dead reckoning information, thus generating a richer representation. This suggests that the bee's mind is fragmented into at least two modules, one of which builds on the other, but each of which is a separate source of a different kind of agency. I will return to this important point later.

For now it is enough that we realize the degree of sophistication that is exhibited by the bee when it navigates according to its map-like representation. Certainly nobody would claim that honeybees are conscious in the way we are, and the rigidity of the way in which they deploy the two navigating systems hierarchically certainly suggests that there is very little reflexive self-evaluation going on. Yet their routinized minds allow these animals to be able to learn the relative locations of different landmarks in their environment, and to compute new trajectories which they have probably never travelled before. As such, these animals do not need to be able to discriminate and attend to the contents of their

21. Ibid.

intentional states; they do not need to be capable of intentionally altering the kind of agency they exert. The quality of their agency is given to them by their routinized minds; what kinds of stimuli are identified as relevant constraints on their agency is determined by which cognitive process—either the dead-reckoning system or the map-like system—is dictating their behavior in any given context. And this is not the most striking example of behavior supported by what seem to be completely routnized processes. Instead, ethologists believe most animal behavior can, and should, be explained in exactly these terms. This includes everything from altruistic behavior in birds[22] and teaching in wild meerkats,[23] to inferences of social rank in fish.[24] All of this behavior can be accounted for by routnized, irreflexive cognitive processes that are sensitive to very specific stimuli and the constraints they embody.

This is not a very controversial claim, and it is probably unsurprising to most of us; indeed, what people commonly mean by 'instinct' would probably comes down to something very similar. But most people would probably be much more reluctant to accept that anything but the crudest of our own behavior is governed by similarly routinized processes. Even though Freud's idea of the subconscious, or something similar enough, has become thoroughly entrenched in popular culture, it is usually understood in terms of repressed urges and impulses that mysteriously affect the outcome of our deliberative process, or as if it secretly inclined us towards certain idiosyncratic behavior that is nevertheless ultimately made effective through the conscious exercise of our agency. In other words, the common view is that this subconscious part of the mind secretly influences conscious decision-making; few would be willing to acknowledge that they actually do not have conscious control over much of their own behavior; that their 'subconscious' is an independent source of very sophisticated agency.

But be that as it may, I think the latter view is much closer to the truth than the

22. Komdeur, J. & Edelaar, P. "Male Seychelles warblers use territory budding to maximize lifetime fitness in a saturated environment." In: *Behavioral Ecology* 12 (2002), 706-715.
23. Thornton, A. & McAuliffe, K. 2006. "Teaching in wild meerkats." In: *Science* 313, 227-229.
24. Grosenick, L., Clement, T.S. & Fernald, R.D. "Fish can infer social rank by observation alone." In: *Nature* 445 (2007), 429-413.

alternative. We make many decisions without being aware that we made them, and much of our behavior is dictated by pre-conscious, routinized cognitive processes. No one would claim that she consciously controls all of the operations of her own brain: nobody in their right mind would claim that she deliberately makes her heart beat, or that she is conscious of the process by which the waves of electromagnetic radiation that hit her retina are converted into visual images. We all know we have no conscious control over certain 'knee-jerk' behaviors (like the reflex which gives rise to that expression), which by itself establishes that we do not control all the operations of the mind that dictate our behavior. But these examples of non-conscious activity are different from the activities that concern us here. They are instrumental and necessary, but they are not the kind of actions which interest us in this discussion: they do not require a cognitive or associative 'mind' of any sort, they do not require intentional states, and they are not goal-directed or purposeful in the relevant sense. So let's go to an example of relevant, agential behavior, like our daily commute to work: we most probably do not deliberate consciously about the route that we will follow; instead, we simply put one foot in front of the other and we arrive at our destinations even though we might have been concentrating on a deeply engrossing phone conversation the whole way. Nonetheless, the skeptics will grant that we do this unreflectively; but they will reply that this is a bad example of truly pre-conscious agency. They will argue that we learned the route to work after we had travelled it conscientiously many times. It becomes a habit that we no longer have to think about, but it is nothing like reflex, or instinct, or any other exercise of agency that is independent from our conscious faculties.

But the skeptic's reply assumes that there is a fundamental difference between habit and instinct. I believe that this (usually mistaken and unwarranted) assumption is the result of the vagueness with which both concepts are used in common language. When asked about the distinction between the two, the most common answers appeal to things like 'innate as opposed to learned behavior', or 'natural as opposed to acquired inclinations', but none of this is really helpful. It is a fact that animals learn much of their 'instinctual'

behavior during their ontogeny, and it seems that our inclination to navigate through our environment is just as natural as that of a bee. And if we are not willing to grant that bees learn their map-like representations through conscious deliberation, it seems odd to assume that we necessarily do. Instead, I believe that we become sensitive to the stimuli which are relevant constraints on our agency in a given environment, learn how they relate with other stimuli, and adjust our behavior on the basis of these new representations—*without any conscious mediation of the process*.

This theory of pre-conscious learning which allows an agent to quickly and faithfully adapt to those constraints that are particular to its environment is neither new, nor very controversial in either ethology or social theory. It is also not restricted to relatively simple behavior like navigation; it has been used to account for very sophisticated exercises of agency, like very subtle and complex social interactions in human beings.

There is ample empirical evidence which suggests that we do not have conscious access to the cognitive processes which support much of our agency and which dictate much of our behavior;, nor does conscious deliberation suffice for autonomous, self-regulating agency. John C. Marshall and Peter W. Halligan report that "in a variety of neurological syndromes, patients may show tacit awareness of stimuli that cannot be consciously recollected or identified."[25] In an experiment, they presented P.S., a patient who manifested left visuo-spatial neglect, simultaneously with two line drawings of a house that were almost the same, except that the left side of one of them was on fire. Across eleven trials, P.S. was asked if she could identify any difference between drawings, and which house she would prefer to live in. She reported that both drawings were identical and that she thought the second question was silly ("because they're the same"); yet quite remarkably, when she was forced to choose, she preferred the house that was not burning on 9 out of 11 trials. Similar results have been obtained in patients with other neurological syndromes like 'blindsight',[26]

25. Marshall, J.C. & Halligan, P.W.. "Blindsight and insight into visuo-spatial neglect." In: *Nature*, 366 (1998), 766-767.
26. See Marshall paper for reference

prosopagnosia,[27] and brain tumors, where subjects correctly answered questions about the similarities between two stimuli, despite reporting only being able to see one of them (and like P.S., the patients found the tasks "silly" since they were being asked to compare two stimuli when they could only detect one!).[28]

This gives important support to the notion that pre-conscious cognitive mechanisms might be the source of much of human agency; our ability to make elaborate, decisions in the abstract need not be attributed exclusively to consciousness. We have good reason to believe that our cognitive structure might be similar in some sense to that of the bees I discussed earlier, whose behavior was dictated by two different, independent processes. I said that the bee's mind was composed of at least two independent cognitive processes, each of which supported a different kind of agency. The one caused the bee to exert a very rigid agency; it caused the bee to invariably travel along a vector that it memorized or learned from a conspecific. The latter allowed the bee much more sophisticated and flexible agency; it allowed the bee to locate itself on a Euclidian map-like representation, and calculate a new trajectory to its desired destination. Of course, human agency is probably never this invariant or dichotomized; our mind is much more self-regulating, which allows us to engage in nuanced behavior that shows we are sensitive to subtle variations in our environment.

As McGeer and Pettit point out, language, which requires consciousness, allows for this kind of self-regulation. Just as consciousness allows for Korsgaard's normative self-regulation, language allows for more general self-regulation with regard to non-normative constraints, including those imposed by our non-moral intentional states. But recent evidence obtained from studies on rhesus monkeys (*Maccaca mulatta*) suggests that pre-conscious psychological mechanisms are also highly self-regulating in this sense. The monkeys exhibit behavior which cannot be accounted for unless one grants that their minds meet the three conditions for self-regulation which McGeer and Pettit laid down: being capable of (1) attending to the content of one's intentional states, (2) identifying constraints

27. Ibid
28. Ibid

on the formation of coherent and adequate intentional states, and (3) implementation of those constraints in the process of forming intentional states. For example, rhesus monkeys appear to be able to link knowing and behaving, at least with respect to themselves. They engage in behavior directed at changing the state of their own knowledge—thus meeting conditions (2) & (3)—which implies that they somehow realize that the knowledge that they have about their environment might be inadequate to perform a task—thus meeting conditions (1) & (2). This was demonstrated through an experiment where monkeys had to correctly select the tube that contained a reward out of a total of four possible tubes. The design only permitted the subject one choice per trial, which gave the monkeys an incentive to choose a tube only if they knew it contained the reward. In the experimental conditions, a monkey had either seen or had not seen the tube being bated with a reward (seen vs. unseen), but in both conditions the monkey could learn where the reward was by looking down the tubes. All of the monkeys looked down the tubes on some of both the seen and the unseen trials, but they looked significantly more often in the unseen trials.[29]

Rhesus monkeys are also capable of making judgments about their own intentional states, which, according to McGeer and Pettit, is a hallmark of a self-regulating mind.[30] Son and Kornell demonstrated this by training two rhesus monkeys to perform two cognitive tasks: the first involved selecting the image of the longest line out of nine possibilities, all of which were displayed simultaneously in a touch-sensitive monitor (each trial could be made easier or harder by manipulating how similar in length the alternatives were to the correct answer); the second was identical except that they had to choose the image which contained the largest/smallest numerical quantity of dots. After the monkeys had made their choice, they were allowed to place a bet on the likelihood that they had chosen correctly (they could choose either a high-risk or low-risk bet). The results showed that they were significantly more likely to place high-risk bets on trials where they had effectively given the correct answer, and more likely to place a low-risk bet on trials in which their choice had been

29. Hampton, R.R., Zivin, A., & Murray, E.A. "Rhesus monkeys (Macaca mulatta) discriminate between knowing and not knowing and collect information as needed before acting." In: *Animal Cognition* 7 (2004), 239-246.
30. McGeer & Pettit, "The Self Regulating Mind."

mistaken.[31] Thus, in line with McGeer and Pettit's arguments, the mind that makes these confidence judgments must be self-regulating in all the ways they claim language allows the human mind to be self-regulating.

So once again, either we grant that rhesus macaques are conscious in the same way we are, or we grant that this kind of consciousness is not necessary for very sophisticated, self-regulating agency. The problem with embracing the latter view is that it blurs the distinction between a routinized and a self-regulating mind—it seems to allow for highly complex aspects of the mind which allow the expression of a very sophisticated and self-regulating agency that functions without conscious mediation. It seems like it would be appropriate to speak of pre-conscious, self-regulating parts of the mind which are nevertheless irreflexive—or at least not sufficiently reflexive that they generate the self-consciousness which is allegedly so characteristic of human beings.

Therefore, conscious deliberation is not the only source of sophisticated agency in our minds. The next natural question, which takes us back to our discussion regarding the traditional view of the soul, asks about the nature of the relationship between these pre-conscious sources of agency and that which results from conscious deliberation. In particular, we need to address the puzzling fact that it certainly feel as if much, if not most, of our behavior were governed by the reflexive, self-conscious parts of our minds. I think Daniel Wegner's analysis of this phenomenon is correct: he claims that we so often experience our actions as the result of conscious deliberation that we assume that this is always the case.[32] And I agree with his insistence that this impression isn't therefore necessarily accurate—there is overwhelming evidence (such Marshall's work with P.S.) that it is in fact grossly inaccurate much of the time. Michael Gazzaniga's work on patients with a 'split brain' is especially convincing.

In one experiment, he flashed an image of a chicken claw to the left hemisphere of a

31. Son, L.K. & N. Kornell. 2005. "Metaconfidence judgments in rhesus macaques: Explicit versus implicit mechanisms." In: *The Missing Link in Cognition: Origins of the Self-Reflective Consciousness* (H.S. Terrace & J. Metcalf, Eds.), pp. 296-320. Oxford University Press, New York.

32. Wegner, D. M. (2003). The mind's best trick: How we experience conscious will. In: *Trends in Cognitive Sciences.* Vol 7(2), 65-69.

patient's brain (right eye), and one of a snow scene to the right hemisphere (left eye). He then asked the patient to choose from an array of pictures arranged such that some were lateralized to the right, and some were lateralized to the left. The patient responded by choosing a picture of a chicken with his right hand, and a shovel with his left. When asked why he chose those images, he answered "Oh, that's simple, the chicken claw goes with the chicken, and you need a shovel to clean out the chicken shed."[33] Now, we know the left hemisphere is endowed with what Gazzaniga calls an 'interpreter system' that operates on the activities of other cognitive processes, and is devoted to giving explanations. In this case, the obvious associations were the chicken for the chicken claw, and the shovel for the snow scene. However, since the left hemisphere did not have access to the workings of the right hemisphere (which is a consequence of having a split brain—hence the name), it confabulated a reasonable—but false—causal story. This experiment shows that the behavior of the left hand is clearly governed by a cognitive process which can respond to verbal instructions and dictate behavior on the basis of abstract representations, but none of this is accessible to conscious reflection. Gazzaniga argues that the cognitive process in the left hemisphere, which takes the operations of other cognitive processes as the set of data on which it performs its own operations, generated an explanation for the left hand's behavior that cohered with the information it had available regarding the right hand's behavior. We, however, know that this information was incomplete, and thus only allowed for an inaccurate (if coherent) account of the actual process by which the right hemisphere chose the shovel.

But we needn't restrict ourselves to evidence obtain from patients with abnormal brain functioning due to injury, or surgery, etc. Nisbett and Wilson (1977) reviewed a large amount of psychological studies in which normal subjects were asked to report on their behavior, which had been influenced by experimental manipulations without their knowledge. Examples included things like misattribution of emotional states to placebos and vice-versa, erroneous beliefs regarding the effect of reassurance on willingness to take

33. Gazaniga, M. *The Mind's Past*. Berkley: University of California Press, 1998.

electrical shocks,[34] and global evaluations of an individual affecting the evaluations of particular attributes like mannerisms or accent.[35] In one particularly telling experiment, subjects were presented with four identical stockings simultaneously, and asked to say which particular article of clothing was of the best quality. There was a pronounced left-to-right position effect, such that the right-most stocking was heavily over-chosen (by a factor of almost four to one). However, when subjects were asked why they had chosen the article they had, no subject mentioned the position of the article. And, when asked directly about the possibility of this effect, virtually all subjects denied it, "usually with a worried glance at the interviewer suggesting that they felt either that they had misunderstood the question or were dealing with a mad-man."[36] They therefore concluded that:

> Ssubjects are sometimes (a) unaware of the existence of a stimulus that importantly influenced a response, (b) unwarare of the existence of the response, and (c) unaware that the stimulus has affected the response…[thus], when people attempt to report on the effects of a stimulus on a response, they do not do so on the basis of any true introspection. Instead, their reports are based on a priori, implicit causal theories, or judgments about the extent to which a particular stimulus is a plausible cause of a given response. [37]

They argue that "people have little awareness of the nature or even the existence of the cognitive processes that mediate judgments, inferences, and the production of complex social behavior."[38]

These conclusions are given further support by Uri Simonsohn and George Lowenstein's work regarding people's housing preferences. Their results strongly suggest that upon moving to a new city, people's price preference and commuting tolerance are strongly affected by what rent they were used to paying in the city they left, and how long

34. Nisbett, R., & Wilson, T. "Telling more than we can know: Verbal reports on mental processes." In: *Psychological Review,* 84 (1977), 231-259.
35. Nisbett, R. E., & Wilson, T. D. "The halo effect: Evidence for unconscious alteration of judgments." In: *Journal of Personality and Social Psychology*, in press.
36. Nisbett & Wilson, "Telling more than we can know: Verbal reports on mental processes."
37. Nisbett & Wilson, "Telling more than we can know: Verbal reports on mental processes."
38. Nisbett & Wilson, "The halo effect: Evidence for unconscious alteration of judgments."

their commute used to be, instead of being determined by a rational deliberation regarding their 'true' preferences.[39] For example, upon moving to a new city, households coming from more expensive cities reliably chose more expensive apartments in the new city than did households coming from less expensive cities—even after controlling for potential income effects. Similarly, individuals who were used to longer commutes in their old city chose longer commutes in the new city when compared to individuals who were used to shorter commutes in their old city. In addition, they observed that people who moved again within the new city revised both their commute length and their housing expenditures, thus countering the initial impact of previous prices and commutes. These additional observations give further support to notion that their initial choices did not reflect their 'true' preferences, but that subjects were simply drawing on salient cues which in some cases might not have been "normatively defensible."[40]

Notice that these subjects had presumably not suffered brain injury of any kind, and were not in a contrived, ecologically questionable laboratory setting. The observations were taken from normal individuals making important, real-life decisions. It is even reasonable to assume that these individuals all engaged in some kind of conscious deliberation during the process of choosing a living location, and yet the most reliable predictor of both price of rent and length commute location was how much they had paid and how long they had commuted previously. This strongly suggests that we do not always make our decisions, even important decisions, through conscious deliberation. Our conscious cognitive process is not always the source of our agency—many times it is completely ignorant of the forces that influence our decisions. It simply confabulates on these occasions.

So are we never consciously in charge? Are instances of seemingly conscious decision-making simply *post hoc* rationalizations or tale-spinning? I do not think any of the evidence

39 . Simonsohn, U. "New Yorkers Commute More Everywhere: contrast effects in the field." In: *The Review of Economics and Statistics* 58 (2006), 1.

40. Simonsohn, U. & Loewenstein, G. "Mistake # 37: The effects of previously encountered prices on current housing demand." In: *The Economic Journal* 116 (2006), 175-199.

gives support to these assumptions either; they simply disprove the other extreme. They simply show that consciousness is not necessary, and is not always sufficient, for self-regulation or self- determination. So the question regarding what or who actually dictates our behavior—regarding what determines the kind of agency we exercise—therefore remains obscure; the obvious answer, 'cognition', is as unhelpful and vague as the question itself. But why is the question so vague? Why are we unable to ask something more concrete—something that can be answered without resorting to waves of our hand and vague concepts (like "the subconscious") or to trivial tautologies (like "cognition")? I believe that what is missing is clarity regarding the whole concept of agency, its sources, and its exercise. I will therefore proceed to answer a question that is generally not considered by philosophers writing on human agency; viz., what is the structure of the source our agency—what is the structure of our soul?

The Structure of our Cognitive Architecture

And if we must say that this element possesses reason, then the element with reason will also have two parts, one, in the strict sense, possessing it in itself the other ready to listen to reason as one is ready to listen to the reason of one's father.

-Aristotle[41]

Let's review where we are. Frankfurt and Korsgaard believe that our reflexive faculties face us with a problem regarding our own agency that no other animal has to face: they make it so that we cannot help but turning our attention to our own intentional states. As such, we become aware that we have intentional states which we would prefer not to have, we find that we do not always identify with many of our motives, desires, appetites, inclinations, etc., and this puts us in a position where we can decide which of these we would have be effective in moving us to action, and which ones we would not. They argue that through this process of reflexive self-evaluation, we construct a self that is identified by

41. Aristotle. *Nicomachean Ethics*. Cambridge: Cambridge University Press, 2000.

those desires which we treat as reason-giving, and this autonomous process of identifying with some of our desires gives rise to an agential identity that commits us to exercising our agency according to those reasons. As such, we create a self with a practical identity through this process of reflexive endorsement, which should be understood as the source of our freedom, and of the normative constraints which we autonomously choose to impose on ourselves. If they are correct, it has to be appropriate to say that human beings are the kinds of beings whose agency is limited by autonomously self-imposed constraints. However, there is much empirical evidence that suggests this is not the case. It seems like we many times respond to constraints without being consciously aware that they exist (Marshall's work with P.S.; Nisbett and Wilson's experiment with the stockings, etc.); that upon reflection, we give incorrect reports regarding which constraints determined how we behaved (Gazzaniga's work with 'split brain' patients); and that we even make important decisions that probably involved a good deal of conscious deliberation on the basis of simple biases which we probably would not endorse as being the relevant constraints for that kind of a decisions (Simonsohn and Loewenstein's studies on rent and commuting preferences).

What this means is that many times, the internal principles of action which dictate our behavior are not accessible for reflective evaluation. It is not simply the case that we might not have conscious control over which of our motives and desires are effective—many of the forces that govern our behavior might not even be available for reflective evaluation. The problem with how we relate to our own agency turns out to be more complicated than the already puzzling reality that we often encounter desires that seem completely alien and contrary to our self-conception—we must add to this picture the fact that we many times do *not* encounter desires, motives, appetites, passions, inclinations, dispositions, preferences, etc., that we would presumably then have to decide whether to endorse or reject. However, much of the difficulty is not actually inherent in the problem itself. Instead, it results from how philosophers traditionally approach the problem. I believe that many of the difficulties result from the prevalence of traditional views regarding the sources of human agency.

According to these views, which can be traced back to Aristotle's dichotomized view of the soul (as presented in the quote at the beginning of this sub-section), reason and appetite are conceptually independent faculties whose only relationship to each other lies in the fact that reason should govern the appetites.

I do not think it is too controversial to maintain that the human soul, understood as the fundamental source of our capacity for intentional, goal directed activity, is embodied in our cognitive faculties. In any case, it is my view that cognition is *the* source of agency. Agency—the capacity to act in an intentional, goal-directed manner—at the very least requires the capacity to represent the goals in pursuit of which agency is exercised, and the environment in which those goals are pursued. The process by which these representations are generated is called cognition. If we understand the structure of our cognitive architecture, we will be in a position to assess what can be true of human agency and the degree to which any given characterization of our agency is plausible in light of the properties of those faculties which support our agency altogether.

I will use Cheney and Seyfarth's definition of cognition: it is the ability to relate different unconnected pieces of information in new ways and to apply the resulting knowledge in an adaptive manner (*How Monkeys See the World*, 1990). Let me build on what it means to "relate" different unconnected pieces of information. A cognitive process could be said to identify the particular way in which particular types of stimuli relate to each other. Formally, cognition could be said to generate an abstract data type (ADT)—cognition could be said to be a process of identifying relationships or links (edges) between elements of the set of relevant stimuli (the set of potential vertices). In other words, cognition could be defined as a set of faculties which construct graphs to represent the relationships between the stimuli which are relevant to the organism. A graph is an abstract representation of a set of objects (vertices) where some pairs of the objects are connected by links (edges). A detailed description of a particular cognitive process should include the exact properties of the ADT—the properties of the links between the vertices, and the types of stimuli that

can be taken as vertices.[42]

The resulting mind can best be understood as the aggregate of multiple, highly interconnected—but fundamentally independent—cognitive processes, each of which is exclusively dedicated to constructing the graph which relates the elements of very specific sets of stimuli to each other. The result is a network of multiple, specialized cognitive processes, each of which is responsible for generating and responding to a particular part of the representation which determines how we exercise our agency. Some units, for example, might be devoted to processing different parts of the information collected by our sensorial apparatus, some others might be devoted to processing information about our own body, and some might be devoted to processing the information presented by certain other cognitive processes. These latter cognitive processes could be viewed as being responsible for evaluating and integrating the representations of those units on which they perform their operations, thereby providing a richer, more nuanced intentional representation. But there is no reason to assume that all the units will be fully integrated with each other at any level. As such, it is possible that some units in one part of the network are more integrated with each other than they are with units in another part of the network. If each of these less integrated parts is sufficiently complex to generate a sophisticated, self-regulating representation, they could each independently support the exercise of very sophisticated human agency. If these sub-components of the network are not sufficiently interconnected with each other such that their individual representations are not integrated with those of other sub-components, the mind which is embodied by this network would contain multiple sources of agency whose dictates would not necessarily coincide. Thus, any one sub-component could, as it

42. The relationships represented in a 'directed' graph, for example, will specify a particular order between the vertices. Namely, the relationship (x,y) is different from (y,x). The significance is very intuitive if taken out of the formal sphere: (John loves Mary) is different from (Mary loves John). Moreover, the edges in a directed graph can have 'weights' assigned to them. The weight of a relationship could be understood as the intensity of the relationship or the probability with which the relationship will be established (John loves Mary very much/little—or—John is very likely/unlikely to love Mary).

I will be using 'stimuli' in a very general sense of the word. By stimuli I mean any information to which an organism could be sensitive. Non-observable mental states will be understood to fall into this category. Similarly, 'types of stimuli' simply refers to an arbitrary subset of the set of stimuli to which an organism is sensitive: visual stimuli, conspecifics, behavior, etc., could all be different types of stimuli.

were, temporarily commandeer control of the entire agent's behavior. This leaves us with a fragmented agent—one whose agency is the product of several, disjoined sources who do not necessarily find the same stimuli to be relevant, and who perhaps would direct behavior towards conflicting (perhaps even mutually exclusive) goals.

It must be admitted that this account of human cognition is not entirely uncontroversial, but the traditional views are no more robust. Though the connectionist assumptions on which it relies are not universally accepted,[43] the debate seems to be leaning in favor of a model that is consistent with those assumptions given that they align very nicely with the neurological architecture of the brain, and that they allow for a theory of cognition that is more plausible than the classical alternatives.[44] The resulting model further provides us with a view of the mind that can be reconciled with the body of odd behavioral evidence that I have presented, which can be made sense of if we allow that the source of our agency is less than unified.

Though it might still seem like one could superimpose this picture of our cognitive architecture onto the Aristotelian map of the soul such that the appetites are generated by one set of cognitive processes, and reason is embodied by another set cognitive process, I believe this would be a stubborn attempt to adhere to a fundamentally misconstrued understanding of the relationship between reason and all the other potential principles of action. As I suggested earlier, reflexivity can, and must, be built into this cognitive structure to allow for the self-regulation that is required by an intentional system that is capable of deliberately acting on its own intentional states. Now, as Frankfurt points out, it is not entirely clear what order of reflexivity must be attained for consciousness to arise,[45] but

43. Fodor, J. A., & Pylyshyn, Z. "Connectionism and cognitive architecture: a critical analysis." In: *Cognition*, 28 (1988), 3-71.

44. Horgan, T. & Tienson, J. "Cognitive systems as dynamical systems." Topoi. 11 (1992), 27-43.; Smolensky, P. "On the proper treatment of connectionism." In: *Behavioral and Brain Sciences*, 11 (1988), 1-74.; van Gelder, T. J. "Compositionality: a connectionist variation on a classical theme." In: *Cognitive Science*, 14 (1990), 355-384.

45. However, it is not clear to me that this reflexivity which allows a system to be self-regulating with respect to its own intentional states by itself can account for the qualia of human consciousness. Whatever the case, I think it is safe to say that reflexivity allows for self-awareness. I am therefore inclined to agree with Frankfurt when he says that self-consciousness involves immanent reflexivity, at least if one restricts the meaning of the term self-consciousness to being coextensive with self-awareness. I will use these two, as well as consciousness and awareness more generally, interchangeably, and ignore the question of qualia.

consciousness necessarily involves "a secondary awareness of a primary response."[46] If the primary response is conceived simply as a lower-order cognitive process, consciousness involves a secondary awareness of the operations of a more primary cognitive process which performs its operations independently of the conscious evaluation.

The Construction of a Self Out of a Fragment of Our Agency

"I have done that," says my memory. "I could not have done that," says my pride, and remains inexorable. Eventually—memory yields.

-Nietzsche[47]

Frankfurt's theory is more in line with my account of the structure of those faculties which support our capacity for agency than most traditional views of the human soul, in so far as it recognizes that the characterizing essence of human agency lies more in the structure of our will than in our capacity to reason. But the manner in which he arranged the hierarchy between the different parts of this structure still falls prey to the traditional fallacy where conscious reasoning is assigned an erroneously independent and unwarrantedly authoritative position. But if it is in fact the case that consciousness arises as a result of the reflexive integration of the operations of more primary cognitive processes, then there is an important sense in which conscious reasoning is inseparable from, and dependent on, the internal principles of action which these primary processes embody. More simply, it would be the case that conscious deliberation only operates with reference to the representations and the determinations of the lower-order cognitive processes. As such, variations in the salience or vehemence of the determinations of one cognitive process would intrinsically affect the determinations at which a conscious deliberative process can arrive.

I think a short digression will make this claim somewhat clearer: Philosophers consistently exhibit a peculiar bias in favor of those conclusions at which we arrive during

46. Frankfurt, "Freedom of the Will and the Concept of a Person."
47. Nietzsche, F. "Beyond good and evil : prelude to a philosophy of the future." New York : Vintage Books, 1989.

'cool moments of deliberation,' over those at which we arrive in the 'heat of the moment.' The most common justification appeals to the lack of time that is characteristic of these 'hot' moments, which implies that the conclusions of the deliberative process would have resulted in different behavior had the agent been allowed more time to reason. Another line of defense makes an appeal to the fact that normal deliberation can be influenced by abnormal affective states. The argument here is that the deliberative process is influenced by excessive emotional responses, or that the abnormal affective reactions to the 'heat of the moment' alter the normal course of reasoning. But this also begs the question as to why the affective states characteristic of 'cool' moments should be the standard for normal deliberative process. Being 'cool' is as much an affective state as any other, and the philosophers have not provided a reason to assume that correct deliberation only occurs under the influence of those emotions and not any others. It certainly does not seem obvious to me that in the 'heat of the moment,' it is not better to have one's deliberation influenced by the affective states normal to *those* situations rather than those normal to 'cool moments.' Stress, arousal, anger, fear, infatuation, apathy, tranquility, satisfaction, etc., are all affective states that influence the way in which we deliberate—and we are probably rather content with that reality. It is surely not the case that we would prefer never to have our decision-making influenced by less than rational forces.

But the point is that our preferences are actually irrelevant on this matter. If conscious deliberation intrinsically only functions with reference to the operations of the lower-order cognitive processes, then the determinations at which we can arrive through conscious deliberation will intrinsically be influenced by the relative salience of the particular determinations of those lower-order processes upon which it operates. As McGeer and Pettit point out, any capacity for the self-regulation of a system's intentional states is limited by virtue of the fact that one intentional state can only be evaluated with reference to another intentional state whose soundness is, at that instance of evaluation, taken for granted.[48] So the consistency of the outcome of our conscious deliberation will

48. McGeer & Pettit, "The Self Regulating Mind."

irrevocably be threatened on two fronts. One the one side, it is possible that variations in the salience of the determinations of different lower-order cognitive processes cause variations in the relative weights assigned to them during conscious deliberation. On the other, the dictates of our conscious deliberation will depend on what cognitive process provides the intentional state which we take as the standard against which we evaluate the other intentional states when we engage in reflexive self-evaluation.

That is why Frankfurt's example of the unwilling drug addict is not very helpful; it presents us with an impossibly impoverished depiction of the psychology of an agent who is reflecting on his desires in a single context. If we imagine that same individual in a different context, perhaps in one where he normally does drugs, it is not implausible to assume perhaps he would fully endorse his desire to take the drug—indeed these kinds of dramatic reversals are not at all uncommon amongst chronic heroin users, to name one example,[49] or amongst people who have become sexually aroused. Ariel and Loewenstein showed that peoples' judgments, preferences, and their reports on how they would behave, are all dramatically affected by the state of their sexual arousal—even though their knowledge was not influenced.[50] For example, in both across-subject and within-subject comparisons, subjects at higher levels of sexual arousal, were much more likely to report being willing to have unsafe sex with a hypothetical partner than were subjects assigned to at lower levels of arousal, even though subjects in both groups were equally likely to report they knew the inherent risk of contracting a sexually transmitted disease from that partner. So, Frankfurt's account would commit us to making the odd concession that we would be dealing with a different person at different times; stranger still, in the case of the drug addicts, we would have to recognize that it is possible that these agents are at times free and at times not free by virtue of acting on the *same* desire.

I think enough evidence of the significant problems with Frankfurt's theory has been provided. Though compelling, there is simply no evidence that suggests that we have

49. Bourgois, P. & Schonberg, J. *Righteous Dopefiend*. Los Angeles: California Series in Anthropology, 2009.
50. Ariel & Loewenstein. "The Heat of The Moment: The Effect of Sexual Arousal on Sexual Decision Making." In: *The Journal of Behavioral Decision Making*, 19 (2006), 87-98.

conscious access to many of the forces that determine how we behave such that we may carry out a comprehensive, reflexive self-evaluation that results in a view of the character of our agency that is accurate inter-temporally —or even at any given instant. Instead, I agree with David Velleman that this process of alleged self-definition is perhaps more a case of wishful thinking that results in self-deception.[51] I believe that what results from this process of reflective endorsement is not our true 'self', but rather a hopeful self-conception that is not representative the actual character of our agency.

The problem is that Frankfurt's view, all desires, impulses, motives, inclinations, dispositions, reasons, or any other denomination for our internal principles of action, are external to the individual and are only internalized if the agent reflectively endorses them. But one could view it the other way as well; one could grant the prerogative to all the unreflective processes such that all their dictates would be internal to the agent until they were externalized by her consciousness. In other words, one could just as easily view consciousness as a usurper of one's identity rather than its legitimate guardian (Nietzsche, for example, could certainly be read in this way).[52]

I think both positions are arbitrary and equally wrong. Philosophers sustain that consciousness is the sources of human agency, and I do not entirely disagree—I believe it is one of the sources of our agency. Yet I hope to have established that it is not the only source, nor the most powerful one. I claim that it is not the most legitimate source of our agency because it cannot take into account the dictates of all the other aspects of our agency at the same time, nor can we be sure that it gives their determinations sufficient weight. Ariely and Loewenstein's experiment is a case in point: they conclude that "sexual arousal seems to narrow the focus of motivation, creating a kind of tunnel-vision where goals other than sexual fulfillment become eclipsed."[53] And as I argued earlier, any bias in favor of those desires which we endorse in a 'cool moment of deliberation' over those which we endorse

51. Velleman, David. "Self to Self." *In Identification and Identity.* Cambridge: Cambridge University Press, 2005.
52. Nietzchse, "Beyond Good and Evil: the genealogy of morals."
53. Ariely, D. & Loewenstein, G. "The Heat of The Moment: The Effect of Sexual Arousal on Sexual Decision Making."

in 'the heat of the moment' is unwarranted. Thus, identification through reflexive self-evaluation will necessarily alienate us from aspects of our agency which not only *will* move to action at times, but will necessarily alienate us from aspects of our agency by which we would perhaps *want* to be moved to action in certain situations. But it is equally wrong to consider our self-consciousness an illegitimate source of agency.

I believe that the correct view is to consider ourselves as fragmented or disjointed agents—as beings endowed with multiple, highly interconnected sources of agency that are nonetheless not fully integrated with, and significantly independent of, each other. And although some integration results from reflexive self-evaluation, it is circumspect to what degree it can truly allow us unify our selves; in the end, it seems like it will always fail to fully integrate the determinations of all the sources of our agency. For this reason, one can only treat the dictates of reason as the dictates of one of the sources of agency; perhaps a source of agency that has unified many others, but they are nevertheless the dictates of a fragment of our agency. As such, the agent *qua* agent cannot be identified with those dictates to the exclusion of the rest.

However, this means that if we accept Frankfurt's theory concerning freedom of the will, we will have to acknowledge that either we rarely enjoy that kind of freedom, or that we are a collection of many different free agents, each one of which sporadically and unpredictably comes to govern the behavior of our bodies. Similarly, if we accept Korsgaard's theory of normativity, we have to acknowledge that we are thoroughly inconsistent in abiding by our own normative commitments. It is simply the case that any conception of ourselves at which we arrive through a process of reflexive self-evaluation will fail to be an accurate description of the character of our agency. We do not limit the exercise of our agency according to those constraints which we reflexively endorse—in fact, we do not even consistently endorse the same constraints!

So *are* we the kinds of agents that we would consciously like to be, as Frankfurt would like to believe? Is Korsgaard right in her claim that obligation is an inescapable reality of

our lives? Would it be right to say that human beings are the kind of agents who live within the limits we believe we have reason to respect? My answer is that to a certain extent, we *can* be, but whether we actually are can only be decided by evaluating all the aspects of our agency. Whether we *are* the kind of agents that we conceive ourselves to be is an empirical question—it is a question cannot be settled through introspection. As things stand, we do not yet understand the structure and functioning of the mind sufficiently well to allow us to conclude which of its many faculties is the ultimate determinant of our behavior, if ever such a faculty is to be discovered. It seems that we are just as ignorant regarding the causes of our own behavior as we are regarding the causes of the universe in general. Here, David Hume's observations regarding the latter phenomenon seem perfectly applicable:

> We must be far removed from the smallest tendency to skepticism not to be apprehensive that we have here got quite beyond the reach of our faculties…we know not how far we ought to trust our vulgar methods of reasoning in such a subject… Were a man to abstract form everything which he knows or has seen, he would be altogether incapable, merely from his own ideas, to determine what kind of scene the universe [or in our case the mind] must be, or to give the preference to one state or situation of things above another. For as nothing which he clearly conceives could be esteemed impossible or imply a contradiction, every chimera of his fancy would be upon an equal footing; nor could he assign any just reason why he adheres to one idea or system, and rejects the others which are equally possible.[54]

Thus, I believe we can only settle this question about the character of our agency by observing our behavior, since only in practice are all the different aspects of our fragmented agencies allowed to express themselves. And if we sometimes act on desires which we wish were not a part of who we are, then so be it. It seems to me that Diderot was entirely right in his account of the difference between a human being and a clavichord: the sounds a clavichord makes depend exclusively on which keys are stricken by an external entity, whereas a philosopher can strike her own keys. But this leaves open the possibility that we

54. Hume, David. *Dialogues Concerning Natural Religion*. Cambridge: Hackett Publishing Company, 1980.

are many times not the composers of the music that we play, nor are we the interpreters. So perhaps my grandfather's skepticism on this matter was wiser than it seemed as I was growing up. Perhaps we would do well to follow his advice that the path one should take care to travel does not lead to self-discovery—but rather to self-acceptance.