Lecture 6
Mechanism's Bugaboos: Freewill, Values, and Human Dignity

For many people, scientists, philosophers, and lay persons, the proposal that our mind/brain is a causal mechanism, even if an unimaginably complex one, is personally unacceptable. It seems to deny too many things most people hold dear about human life—that we are free agents, that we have and direct our lives according to our values, and that we have human dignity. Theorists otherwise committed to treating the human mind/brain as a object of scientific inquiry, such as Kant and James, contended that we must adopt a different framework when it comes to living our lives as human beings. To live our lives, they contended, we must view ourselves as free, autonomous agents and deny that we are mechanisms.

My goal in this lecture is to try to show that worries stemming from conceiving of ourselves as mechanisms are misguided—they are bugaboos, not something really to be feared. Our mind/brains could be complex mechanisms and yet we could still be autonomous, responsible agents capable of identifying and pursuing values and living lives of dignity. And, in a sense that is important to those enterprises, we could still be free. And those things are true of us not in spite of our being mechanisms but in virtue of the kind of mechanism that we are.

The previous lecture has prepared an important component of the analysis of mechanism that enables us to note at the outset that even if our mind/brain, and body in general, is *only* a complex mechanism, we are not just what our internal components do. We are the whole mechanism and our activities are those that of the whole mechanism. Mechanisms perform activities that engage them with other entities, including other mechanisms. Moreover, they are often embedded in other systems, other mechanisms, and are in part shaped and altered by what is happening in those systems. As mechanisms, we engage other mechanisms and are embedded in social structures.[1] As a result of being so embedded, we are capable of being participants in a culture, both affecting that culture through out own activities and being affected, including having our values shaped, by the culture. Moreover, we routinely employ features of our environment in our activities, allowing us to accomplish more than we might on our own. To take just one example, in preparing this text I interact with external representations I have constructed, both writing words and sentences, but also re-reading and revising them.

Perhaps the most basic concern about construing ourselves as mechanisms is that mechanisms are deterministic—the activities they perform are determined by their components and what they do, their organization, and the factors impinging on the mechanism. There is no possibility of a mechanism doing something other than what it is determined to do. For us to be autonomous, responsible agents, many think it must be possible for us to do something other than what we are determined to do.

[1] Just how appropriate and fruitful it is to view social systems as mechanisms is a topic beyond these lectures. But clearly some social systems have characteristics of mechanisms—they are coordinated systems comprised of individuals that carry out activities and, of special importance, feedback back upon their constituents and constrain their behavior.

Although it is possible to envisage a mechanism whose components operate in a probabilistic rather than a deterministic manner, there is little to be gained in that direction in overcoming the bugaboos of mechanism. Probabilistic behavior, in itself, will do little to secure what is needed to provide for autonomous and responsible agency. To do that, I must confront head-on the main objection and show how deterministic mechanisms are compatible with autonomy and responsibility.

Before beginning the defense of the mechanistic construal of us against these bugaboos, let me offer one clarification. Not all mechanisms are capable of autonomous, responsible action. It takes a very special kind of mechanism—not just one of great complexity, but one with the right kind of complexity. I am not going to offer an analysis here of what kind of complexity is needed. I don’t think we are yet in a position to do that. At this point we only know one mechanism that seems to be capable of meeting the requirements—human beings. Although members of some other species may come close, we do not, as a matter of fact, construe any of them as agents responsible for their behavior. And we have not yet built any artifacts that we hold responsible. The strategy I recommend for understanding responsible agency is not to analyze those notions in an a priori fashion, but to pursue the sort of naturalism that I have advocated throughout these lectures. This will involve both examining under what conditions we hold humans to be responsible agents and determining what it is about the mechanisms within them that enables them to meet the demands of moral agency. We are just at the beginning stages of this endeavor, although we have already made some significant progress. As I will discuss below, research has already shown the mistaken of isolating reason from emotion—responsible agency appears to require the coordinated engagement of both reason and emotion. But without offering an account of what is required for autonomous, responsible agency, let me try to show why such agency is not incompatible with us being a certain kind of mechanism.

Determinism and Agency

There is a common conception of what is required for autonomy and responsibility is that agents be free in the sense that their actions are not caused. The challenge for the advocate of this position is to make clear what it would be for an activity to be uncaused and to show that such an activity has the right features to account for responsible action. Let's try to conceive of something happening that has no causal ancestry. Perhaps a red balloon will just pop into existence in front of us in a moment, or this lectern will just fall over, without any determining cause. The first is problematic since it does not just involve a change happening with no cause, but something coming into existence *de novo*. That would not seem to give us any grip on free action. The lectern falling over seems a better candidate. When an inanimate object starts doing things that we cannot explain, we begin to think it is really animate, or enchanted, or something similar. But why do we do so? Perhaps because we thought it had a mind and was up to something. It might, for example, be expressing its profound disagreement with this talk. Notice, though, that if we do this, we are providing a reason for the lectern falling over. On many standard interpretations, reasons are causes—conditions which increase the likelihood of certain consequences. If we pursue this strategy, however, we are not thinking of the lectern's behavior as uncaused. Rather, we are thinking of it as having a certain kind of cause.

What this seems to indicate is that if someone does something for no reason, but just does it, we are not presented with an exemplar of autonomous, responsible agency. As Hume noted, "Where [actions] proceed not from some cause in the characters and dispositions of the person, who perform'd them, they infix not themselves upon him, and can neither redound to his honor if good, nor infamy, if evil" (Hume, 1739, p. 411). To understand autonomous agency, this suggests that we should not reject causation, but focus on the type of causation that is involved.

But what has led so many people, philosophers and others, to see causal determination to be incompatible with autonomous, responsible agency? A common way of characterizing what it is for an agent to be morally responsible for her actions is to maintain that whatever action the agent actually performed, she *could have done otherwise*. In many respects, this is a very reasonable demand. If the agent was forced to behave as he or she did, then it seems unfair to hold them responsible. But it is not straight-forward to explicate the notion of *being forced* and *could have done otherwise*. When one rejects an action as a reflection of autonomous, responsible agency on the grounds of being forced, one typically has in mind something like externally applied force. If someone else makes you give your money to charity, either by forcing your hand to take hold of your wallet and to give it to someone or by threatening you, you do not get moral credit for giving to charity. Likewise, if someone directly controls your mind-brain, either through mind-control techniques or through direct chemical or electrical operation on your brain, the resulting behavior is not an expression of autonomous, responsible agency.

But what about a case where one could not have done otherwise than to give to charity because that is just the kind of person one is? Let's not envisage an obsessive giver but one who gives in moderation but does so as a result of her deepest values. This person, we might imagine, will justify her actions if asked in terms of the importance of insuring the welfare of the less fortunate and if further pressed will develop a detailed and coherent moral and political philosophy. She is able to understand and even articulate reasons for not giving to charity, but rejects them with cogent counter-reasons of her own. Such a person might not have been able to do otherwise than to give to charity, at least without fundamentally changing the person she is. But even though she could not have done otherwise in this sense, this person seems to be an exemplar of autonomous, responsible agency.

The requirement of being able to do otherwise, therefore, must be appropriately constrained. Perhaps the needed constraint is something like, one would have done otherwise if, after appropriate deliberation, one had chosen to do otherwise. Not all acts of autonomous, responsible agents are the products of deliberation. We have neither the time nor resources to deliberate in all cases of action and, and James noted, must rely on developed habits. But we are capable of deliberating, and when our actions are responsive so such deliberation, then perhaps it makes sense to construe our actions as an expression of our autonomous, responsible agency. Surely this requirement must be explicated far more carefully than I have done here. But this is sufficient for the main conclusion I need to establish—that autonomous, responsible agency does not depend upon decisions being uncaused. What matters is the kind of causation.

Perhaps a major reason why causation is so often taken to be inimical to agency is that it seems to render the decisions of the agent predictable. If others can completely predict how we will behave, then it seems wrong to hold us responsible. The basic insight here seems to be correct. If behaviorist learning theory were generally correct, then autonomous, responsible agency would be undercut. After one has conditioned an animal to behave in a certain way, it hardly seems appropriate to hold it responsible for behaving in that way. We, as humans, are not immune from conditioning, either classical or operant. The Garcia effect is perhaps the best known exemplar of classical conditioning and if one has become ill a certain period after eating an unusual food, one will develop an aversion to that food. The mechanism underlying this effect is robust and it hardly seems right to hold someone responsible for this aversion. Likewise, appropriate schedules of reinforcement can be powerful determinants of human behavior and if someone behaves in a certain way after being subjected to such a schedule of reinforcement, his or her behavior does not seem to be an expression of autonomous, responsible agency.

What underlies these cases, however, is not the operation of a mechanism per se, but of one of a certain kind—one that is very predictable. If in general our brains were such easily predicable mechanisms, then construing us as autonomous, responsible agents would seem to be misguided. But must a mechanism's behavior be predictable?

Causal mechanisms and predestination

On first blush, it may seem just obvious that if our brains are mechanisms then, if someone knew the inputs to them, one could predict their behavior. Perhaps the computations might be difficult and one might need to rely on very powerful computers to carry them out, but prediction, at least in principle, seems guaranteed. For those theologically inclined, it means that God could have known before setting the universe into motion exactly what we would do. For many, this prospect seems to diminish any role we might play in producing our own behavior. My goal in this section is to show that this particular worry is unfounded.

To begin to see why, let's consider the magnitude of the computation that is required. At least in part, the brain is an electrical switching system. The production of action potentials in individual neurons is at least one fundamental activity of the brain. Whether an action potential develops in a particular neuron is determined by action potentials in neurons that synapse on that neuron. The simplest assumption is that an action potential will develop in a particular neuron if a there are action potentials in a particular time frame in sufficient numbers of these synapsing neurons. If this is the case, we can write the equation describing the response of any given neuron. Now all we need to do is calculate on a moment by moment basis the propagation of electrical activity through the system. Mathematical modeling of artificial neural networks has very much this character—the activity of each unit in the network is calculated from the activity of the units providing inputs to the unit (generally by applying a non-linear activation function) and this is determined successively for each unit in the system.

For simple artificial networks of a few dozen, hundred, or thousands of units, these equations are routinely solved on current digital computers. But as the number of units and the number of connections grows, these calculations become ever more complex and time consuming. The real brain is vastly more complex than any artificial network yet simulated—it is estimated to have approximately 10^{12} neurons, each with on average 10^{3} connections to other neurons. Even if these neurons operate in accord with the simple assumptions of artificial neural networks, calculating the behavior of the system for a single time interval on the most powerful digital computers yet built will take enormous processing time. But these assumptions are almost certainly false and the computational simulation of real neurons is far more complex. For one thing, the processes determining whether an action potential will be generated in any given neuron is more complex than just a summation of activity in neurons that synapse onto it. For another, the behavior of synapses is constantly changing as a result of chemical alterations accompanying synaptic discharges, requiring continual updating of the equations governing neural responses. Further, there is no common clock determining the update process for each unit. When this is combined with the fact that many of the actual processes in the brain are recurrent, the computation becomes ever more complex.

What do human engineers do when the equations describing a system they are contemplating become too complex? Typically they begin with approximations to get a sense of the type of behavior a particular system will likely exhibit. But when greater precision is required, they turn to building the system itself. Even if we assume that the equations we employ are accurate and that the measurements of the inputs to the system are precise, it is simply far more efficient to build a complex system and observe its behavior than to calculate how it will behave.

Given the complexity of the brain, especially in light of its non-linear dynamics, we may not have to worry much about our behavior being predictable, even if we are deterministic machines. Although we cannot conceive an omniscient intelligence reasons, it may turn out that even God would operate in the same way as human engineers if he sought to know how a human brain of a certain kind would behave—he would build the device and observe its operation. This would not be due to limits on his cognitive capacities but the selection of the more efficient reasoning strategy.

Casual determination in a mechanism, therefore, does not entail predictability. Even if one knew the mechanism operating in us and the inputs we are receiving, by far the simplest way of determining how we would behave would be to let us process the inputs and behave. What this does, of course, is put us back in the casual pathway. It is our brains that carry out the processing so that *we* can behave.

Causal mechanisms and values

The worries about causal determination and predictability, however, are just preliminaries to the real concern about mechanism. Far more central is the concern about values and their role in the determination of behavior. Is there a place for values in a totally mechanical system? To see that there is, let's consider what it would take to get a mechanical system to make decisions.

One of the main reasons for designing artificial intelligence (AI) systems is to improve on human decision making by arriving at decisions more rapidly or more accurately than human decision makers. (In part this is due to known limitations on human decision makers, who are often subject to a variety of biases, such as over-emphasizing certain evidence and failing to consider base rates.) To get even simple AI systems to make decisions, the system must represent the goals to be accomplished. To play a credible game of chess, a computer system must represent the goal of checkmate and evaluate possible strategies in terms of their likelihood of achieving that goal. To take an even simpler problem solving task, in order to solve a problem such as the Tower of Hanoi problem (the problem of moving rings of different sizes from one peg to another without ever violating the rule of never putting a lager ring on top of a smaller ring), the computer must represent the goal. The representation of a goal becomes a determinate in the system's behavior. Observing such a system in operation, one might be inclined to attribute to the system the value of achieving the goal. (When restrictions are placed on the ways in which such a system will achieve its goals, the resulting behavior seems to respect other values specified in those prescriptions.)

The manner in which goals are built into AI systems strike many as profoundly different from our activities of valuing. Part of what seems different is that the operation of these normative principles does not seem to engage the system in the way normative principles engage us. Values for us do not seem to be just another set of rules governing our operation. Value issues seem to engage us as agents in a far more immediate manner than such rules engage the computer. One way to appreciate this fact is to recognize that values are not always realized in behavior. Not infrequently values are violated. Sometimes this is due to value conflicts where, to promote one valued result one violates another value. Even though one values winning a game, one may choose to loose in order to retain a friendship, something one values even more. Even though one generally respects private property, one may destroy another's property as a form of protest. Other times one may fail to act in accord with one's values for less noble reasons, even reasons that one cannot explain. The phenomenon of weakness of will is all too familiar.

It may be possible to build AI devices whose behavior reflects these features of human valuing, they still do not seem to be really engaged in valuing. What lies behind this powerful intuition? In some sense, such systems do not seem to *care* about their values. They have no passion. Part of engagement with values for humans seems to be affective. We *care* about our values. It is this emotional engagement that seems to be lacking in AI systems. (They are not just abstract principles. This may explain why some people find a disconnection between value theory and values. What utilitarian theory, for example, implies we should do may seem disconnected from our actual value commitments.)

Many theorists, including many philosophers have drawn a sharp contrast between reason and emotion and have construed reason as itself able to direct behavior and indeed the proper guide to behavior. For Plato, for example, emotions were disruptive and needed to be controlled. But Plato did not deny them a role. The appetitive part of the soul, properly directed by reason, was necessary to achieve ends. In the domain of morality Kant went further, denying any role for the emotions in determining moral reasoning. It is just this denial of affective, however, that has made Kantian ethics seem so unnatural to many.

Recent neuroscience thinking has begun to identify a critical role for the emotions in human decision making, including moral decision making. As is often the case, critical insights came for the study of patients with specific deficits which had surprising consequences. Perhaps the most famous is the 19th century railroad worker Phineas Gage who, in an accidental explosion, and a tamping pole thrust through the orbital regions of his frontal cortex. At first it appeared that Gage survived his injury unscathed. Not only did he live, but his reasoning ability seemed normal. But in other respects he was anything but normal. Previously a very responsible individual, Gage became irresponsible, unable to hold his job, maintain his marriage, etc. Something about the damage to his frontal cortex seemed to have produced dramatic changes in his character.

The contribution of the orbitofrontal areas damaged in Gage's brain remained a mystery for over a century after his accident. Since no clear cognitive function was attributable to this region, it was an area surgeons would remove if tumors occurred there. EVR was such a patient who had a tumor removed from the ventromedial part of his frontal lobes that resulted in bilateral lesions. Prior to surgery, EVR had an IQ of about 140, which was not diminished as a result of the surgery. Like Gage, in terms of his thinking, he seemed to emerge from surgery unharmed. But outside the laboratory his life was severely impacted. His performance at work suffered as he showed up late, failed to complete tasks, etc. For over a decade EVR has been studied by Alberto and Hannah Damsio. As the Damasios' investigated him, they discovered that although he often knew what would be the rational choice, he would act otherwise.

Working with the Damasios, Antione Bechara developed an experimental procedure that helped reveal the nature of EVR's deficit. A subject is presented with four decks of cards and is free to choose to turn over cards on each round from one of the four decks. As the subject turns over the cards, he or she receives or loses the amount of money specified on the card. The penalty cards are dispersed through the decks so that the subject cannot anticipate when they will show up. Two of the decks have cards with relatively low payouts and low penalties, designed such that over the long-term a subject will make money by choosing cards from those decks. The other two decks have cards with higher payouts, but even higher penalties, so that over the long-term a subject will loose money by choosing cards from those decks. Normal subjects learn after 15-20 trials to choose cards primarily from the low payout/low penalty decks, while EVR and other patients with lesions in the ventromedial frontal areas continue to choose cards from the higher-payout/even-higher-penalty decks. It is not, however, that EVR cannot figure out which decks are to his advantage. He can report which is the more rational strategy; but he acts otherwise.

Other research has revealed that the difference between patients such as EVR and other subjects has to do with the connection between reason and emotion. When skin conductance was measured with normal subjects, they started to exhibit a skin conductance response when they reached for the bad cards on trials even before they started to reliably reject the bad cards. As noted, around trial 15-20 they started to avoid the bad cards, and reported a feeling that something was "funny" about the bad decks. Only after about 50 trials could they articulate what was the winning strategy. EVR and other ventromedial patients, however, never showed the skin conductance response and, although they did figure out what was the winning strategy, they did not adhere to it.

The ventromedial areas of frontal cortex are known to be areas with projections both to limbic areas, thought to be involved in emotional responses, and cortical areas thought to be critical for higher level reasoning. Lacking these connections, EVR does not exhibit the normal emotional responses and, on the Damasios' interpretation, it is this failure that accounts for his abnormal responses. Without this, he is not able to put his knowledge into action.

It is important to note that for normal subjects, the skin conductance response preceded the reasoned analysis of the situation. This suggests that emotion is not just a handmaiden of reason, but itself potentially an important guide to action. Many of the areas of the limbic system that are involved in neural processing of emotions are evolutionarily very early. They began to develop in nervous systems that were primarily directed at monitoring and regulating internal organs, including those in the alimentary canal. Even in more complex organisms, in which much brain activity, including emotional responses, is directed outwards towards the environment, there is still a close connection between limbic processes and our internal organs. There is a traditional theory of the emotions, due to James and Lange, according to which emotions first involve changes in the body and only subsequently registration in the mind/brain. Without taking a stance on this hypothesis, it is nonetheless noteworthy that emotional responses are highly integrated with other physiological processing in our bodies, including digestive activity in particular. The colloquial expression "gut response" seems to have a foundation in our neuroanatomy. Minus such somatic responses, our cognitive systems seem unable to execute what our reason dictates.

The disorders exhibited by patients such as Gage and EVR provide a first clue as to what is required for our brain mechanisms to undergird valuing. The system must be affective as well as cognitive. But, as already noted, early brains were in the business of regulating internal organs associated with the alimentary canal, and the parts of the brain that are most directly involved in such regulation are components of the limbic system. The more cognitive components of the brain, the neocortex generally and the frontal areas more specifically, are later phylogenetic developments in brains already developed to utilize affect in guiding behavior. While philosophy and AI have been tempted down the path of segregating reason and treating it as isolated from affect, that is not how brains seem to work.

There is a great deal we do not yet know about how valuing is realized in the brain. What we do have, though, is a suggestion as to why our values are so central to our identity (and why talk of values with computers seems ill-founded). Valuing involves not such reasoning, but our affective processes. As a result, they are not detached from us in the way reason, including moral theorizing, often seems to be. Moreover, by recognizing that our brains originated and remain systems for regulating organic life, including our most basic physiological processes, we can see how it is that a mechanism can be engaged in something so central to our being.

Decision Making By Complex Machine Brains

A requirement on any proposed mechanistic explanation is to account for what we know about the behavior of the system, including any information we have about the manner in which the system performs its activity. If, then, one puts forward a mechanistic model of human decision making, it must account not only for the decisions humans actually make, but what we know about the process of making decisions, including the phenomenological features of that process.

What are some of the critical features of human decision making? One feature follows closely upon the discussion above of valuing. Human decision making does not seem to be just a process of abstract reason as rational decision theory often characterizes it. It is not infrequent that we find ourselves in situations where reason tells us to do one thing, but we deeply want to do something else. In such cases we sometimes follow reason, but experience deep regret at the benefits of an alternative choice that we have foregone. Other times we *follow our guts*, hoping that we do not later regret ignoring the dictates of reason. One way to account for this opposition is to recognize that brains involve multiple processes operating simultaneously. Without embracing modularity in any strict form, we can nonetheless recognize relative separation of processes occurring in different parts of the brain.

Another frequent feature of human decision making is a process of vacillation before arriving at a decision (or even afterwards). Although vacillation would not be expected in simple machines, we now have plenty of models of complex mechanisms in which something resembling vacillation is the norm. Non-linear interactions between components of a system frequently give rise to dynamical systems that exhibit complex trajectories, including ones that oscillate between different semi-stable states.

To put some flesh on this conception, we can again consider artificial neural networks. Interactive networks are often construed as constraint satisfaction devices in which the constraints are soft since not all can be satisfied simultaneously. Such networks *settle* into states in which some constraints are satisfied while others are violated. If the internal node in such a system have internal dynamics (e.g., they are harmonic oscillators), then the system may only partially settle. The internal activities of the components may lead the system to spontaneously break out of a relatively stable configuration, settle into another configuration, only to break out once again. Such systems exhibit *metastability*. These systems are mechanisms, albeit ones with complex internal dynamics. Although work on such systems is at present highly theoretical, exploring the properties of such systems provides a means to appreciate that the phenomenological features of decision making may be realized in a mechanism.

Causal Mechanisms and Our Self

Having tried to address some of the reasons one may think that mechanism is incompatible with autonomous, responsible agency, let me turn now to one feature that seems to be necessary to such agency—a self that is the agent. This is key to the conception of autonomy as self-governing and of responsibility as being accountable for one's actions. A person must be an agent—a self. Where does a *self* emerge in a mechanism?

An approach that is likely to be fatal is to seek a self as a component of the mechanism. A number of theories of human cognition introduce the idea of a *central executive* which carries out the highest level processing and exercises executive control over other parts of the brain. If it makes sense to treat certain parts of the brain as such an executive, then one might be tempted to construe it as the self. But this has all the negative features traditionally associated with homuncular theories, especially the challenge of explaining how the homunculus can itself perform all the operations attributed to it. The point of decomposition is to divide the activity of a system into component operations which are *simpler* than the activity of the whole system.

A far more promising approach is not to localize the self in a part of the neural mechanism, but identify the self, the agent, with the brain as a whole, or perhaps with the person as a whole. In the end, that is the position I maintain we should adopt. It is whole persons that we hold responsible for actions. The vocabulary for describing agents, so-called folk psychology, is pitched at the level of persons. It is persons who have beliefs, desires, values, etc. The mechanistic explanation is directed toward explaining how agents are able to carry out the activity of agents—it is not to identify agency as a part of the system.

That said, though, there are features of our conception of our selves that may draw on the processing capacities of more localized parts of the brain. As I will suggest in the final section, part of the challenge of being a successful agent is to integrate these different capacities, and this integration may represent more of a project rather than something already realized at any point in one's life. But before turning to that, are there aspects of the self which are separately realized in the brain?

To pursue this analysis, I am going to follow Ulric Neisser (1988) in differentiating different aspects of the self, although I will depart from his analysis in identifying an aspect potentially more fundamental than those he distinguishes. This first component consists of the self-regulative, autonomic processes highlighted in the previous section. Although our earliest evolutionary ancestors presumably had no awareness of themselves, they were individual entities whose nervous system served to maintain them in homeostatic states. In the fashion first described by Claude Bernard, the self regulative activities of the autonomic nervous system enabled the organism to resist external forces that opposed it. It thereby enabled the organism to maintain itself as a separate system. One virtue of starting with the processes of the autonomic nervous system is that we begin with processes that tie a feature of one's self intimately to one's body. This fits with the experience of many that their self conception is in part tied to their body and that fundamental changes in their body alter their sense of self.

The foundational level in Neisser's concept of self is what he terms the *ecological self*. In this he recognizes, following Gibson, that a fundamental feature of self for most animals is the prospect of action and that crucial for action are perceptual processes that specify where one is in one's environment and what are the possibilities for action possible in that environment. It is crucial for understanding this conception of self that perception for Gibson and Neisser is not directed at providing an objective, organism independent portrayal of the environment, but a decidedly egocentric perspective, specifying where things are related to oneself and what the possibilities for the particular organism, with its particular motor capacities, to act on that environment. As Neisser has further developed this aspect of the self, it is closely tied in us to the dorsal visual pathway (see lecture 2) which enables us to analyze the environment in ways directly tied to action. As I developed in lecture 3, it is unlikely that this processing figures directly in phenomenal awareness.

Animals, including humans, live lives interrelated to other organisms, especially conspecifics, and in part the identity of an individual is fixed by its relations to other organisms. Especially important in this respect are familial relations through which one relates to parents, siblings, and offspring. An organism is dependent on some organisms, has particular responsibilities to others, and this nexus of relations determines a set of roles for the individual. Neisser characterizes these as constituting the *interpersonal* self. Like the ecological self, the interpersonal self is inherently relational. Recognizing this is important, for at these foundational stages we position the self in relation to things outside the organism, especially larger social systems.

When we turn to humans in particular, one of the most important features in terms of which an individual identifies himself or herself is in terms of memory of one's life. As Neisser's label *extended* self makes clear, this aspect of oneself extends one into the past, to those events in which one has participated. Tulving dubbed this sort of memory *episodic memory*, emphasizing that it enables one to *revisit* episodes in one's life history. He contrasts it with semantic memory, memory of general information that is not linked to one's own experience of particular events. Episodic memory enables one to define oneself in terms of what one has done in the past. As numerous tragic cases of amnesics who lose their memories of or cannot acquire memories of their past make clear, such memories are critical to our self construal. As Jerome Bruner makes clear, self narratives as well as group narratives that place us in a broader social context are central in giving substances to our self. Neisser's characterization makes clear, though, that just as one's identity extends to the past, it also extends to the future and to projects we can envisage pursuing in the future.

Neisser's fourth self, the *private* self, captures an important feature of conscious mental life to which James drew attention—that our conscious state is uniquely ours and is private. Part of what is private is our phenomenal experience, which I have identified with stages of processing in our perception of the world. This experience represents our egocentric perspective on the world as we experience it through our senses. It is partly that this experience arises as part of our processing of sensory experience but it not the outcome of that processing that makes it hard to fully describe our phenomenal experience to others and has rendered such experience philosophically mysterious. Another part of our private self, though, arises once we have *internalized* the use of language, acquiring the ability to continually talk to ourselves and use linguistic representations as tools in our thinking. We can relate this internal monologue to others, but it occurs inside our skulls. At this point in time we have only the crudest understanding of how language is processed by our brains, but presumably this aspect of our private self involves running offline the processing we utilize in interpersonal language use.

Finally, Neisser identifies the *conceptual* self—one's representation of one's self. Our self concept is linked in many ways to our episodic memory, of how we remember our self. But it goes beyond that to include how we think of ourselves as agents—as having values, planning actions, etc. Like any representation, it may not accurately characterize our self. An important part of our self representation is that when we act, we first decide to act, and that decision is causally responsible for our actions. We deliberate, we choose, and then we act. That, after all, is a major part of what we think it is to act responsibly.

But we must be cautious not to over intellectualize our understanding of our self. An important aspect of our self is our habits of response built up consciously or unconsciously over time. As James emphasized, the routines of our lives play an important structuring role, limiting the need for high-level conceptual deliberation. There are even tantalizing claims in the cognitive neuroscience literature that the whole conception of us as consciously deliberating and, on the basis of deliberation, initiating action, is a misrepresentation. Neural firings that initiate the sequence of actions may, at least in some cases, *precede* the resolution of our conscious deliberation. On first blush findings such as these seem to confirm the concerns that I am trying to counter in this lecture—the concern that understanding us as a mechanism may undercut our status as autonomous, responsible agents. But that is to conflate one aspect of our self, our self concept, with our self, something I have consistently warned against. Even if these claims are true, our actions resulted from our brains. It would mean that our self concept does not figure in the generation of action in the manner we might think. Does that render our self concept epiphenomenal? It need not. Even if our reasoning about our self and our actions does not figure in the generation of current actions, it may still play a critical role in shaping who we are and how we engage the world in the long run. That is, such conscious deliberation may have its efficacy in shaping our habits and emotions, which may then figure more directly in our generation of action.

Different components of our mind-brain play central roles in these different aspects of our self. Techniques such as neuroimaging are beginning to provide evidence as to what parts of the brain are most active when we do such things as recollect, deliberate, or decide what to do. The goal of such decomposition of self, however, is not to single out one or another of these brain regions as *the self* but to understand how different operations in the mechanism of the brain makes possible our operation as a self. As we understand how the brain realizes these different aspects of a self, we understand that there is no conflict between us being a complex mechanism and being a self in the sense needed for being an autonomous and responsible agent. The components interact so as to enable us to be agents.

Unity of Self and Knowing Oneself

By differentiating different aspects of self in the previous section it becomes easier to understand how a brain might realize a self, but it also raises the prospect that the different aspects of one's self may be at odds with one another. This, however, is not just an idle worry, but a feature of life. Far short of pathology, we recognize different tendencies within us. We desire a long-term objective (a certain bodily appearance, a certain style of living, being able to perform an activity). And we may know what it takes to achieve such goals (resisting eating particular foods, saving our income, practicing regularly). But other desires at the moment lead us not to do what we know we need to do to obtain these actions. We act against our own self interests.

Although that is our predicament, that does not mean we are helpless. Part of what we are able to do is reflect on our self and our behavior. We can evaluate what who we are, what we value, and what we do. In performing these activities, of course, we cannot literally stand outside ourselves and utilize resources other than those that comprise us. Rather, we do so using the same resources we do in other cognitive activities. Even the desire to reflect critically on ourselves must come from within. But if we have such a desire, we can undertake critical reflection. For many, part of that critical reflection is the desire to be autonomous, responsible agents. My contention has been that seeing ourselves as mechanisms does not undercut that project. But neither does it assure its success. Becoming an autonomous, responsible agent involves treating our lives as ongoing projects. Part of becoming such agents is to unify our selves so as to achieve what we most want to achieve. One of the things this involves is constraining parts of ourselves, including some of our desires. This possibility may seem beyond the range of a mechanism, but recall the claim of the previous lecture. Components of a mechanism can be affected by the conditions of the mechanism in which they reside. Our various activities alter conditions within us, allowing for just this sort of top-down control.

The old Greek adage, *know thyself*, remains important advice. If we are to become the self we want to become, we need to know what we can and cannot obtain, and what aspects of ourselves need to be changed to become the person we want to be. Does knowing our brain help us know ourselves? Not in any direct way. There is at present little prospect for changing ourselves by operating directly upon the mechanism that constitutes us by, for example, inserting new neural circuitry into our brains. But there are contexts in which what we know about the mechanisms within us can contribute. One of the most obvious contexts is with psychotropic drugs. Although the changes remain relatively crude, if we are suffering from depression, antidepressant drugs can *be part* of our project of making our self into what we want it to be.

More generally, knowing about our brains, and especially about the psychological processes occurring in brains, can provide guidance for choices we make in life. For example, we may learn that certain situations trigger behavior in us that we cannot, at that point, control. Good practical advice then is to avoid such situations. But this is something we could learn from Homer. We don't need cognitive neuroscience. Does that mean cognitive neuroscience won't help us know ourselves? In the sense in which knowing ourselves is critical for action, cognitive neuroscience is not where we should be focusing. The project of engaging the world is a project for our selves as whole mechanisms. This is an activity we perform. Understanding the mechanism within does not supplant knowledge of the activity the mechanism performs. But neither does it threaten our project of operating in the world as autonomous, responsible selves. And that is all I have been concerned to establish in this lecture.