Cognitive Critique

THE DANGERS OF DUALISM: IMPLICATIONS OF THE MULTIPLE DECISION-MAKING SYSTEM THEORY FOR FREE WILL AND RESPONSIBILITY

A. DAVID REDISH

Department of Neuroscience
University of Minnesota, Minneapolis, Minnesota

EMAIL: redish@umn.edu

Accepted May 28, 2013

KEYWORDS
neuroscience, will, philosophy of mind, intention, responsibility

ABSTRACT

Recent work on decision-making suggests that we are a conglomeration of multiple decision-making information-processing systems. In this paper, I address consequences of this work on decision-making processes for three philosophical conceptualizations that can be categorized as forms of dualism.

1. *A rejection of Cartesian dualism*. Although most scientists reject the existence of a separate non-physical being, the importance of this still has not been fully appreciated in many other fields, such as legal or philosophical scholarship.

2. *A rejection of the software analogy*. Many researchers still argue that we are software running on neural hardware. I will argue that this is a modern form of dualism and that this hypothesis is not compatible with the data.

3. *A rejection of Augustinian dualism*. Many researchers identify human cognition with only one of the multiple decision-making systems, which leads to concepts such as *emotion made me do it*. I will argue that this is a poor description of the human individual.

All three of these errors are still used in making decisions in current law and legal scholarship. As scientific results begin to undercut these dualist interpretations, it becomes dangerously easy to reject intention, free will, and responsibility. I argue that taking a more realistic perspective on human decision-making will provide a more reasonable basis for legal decisions.

INTRODUCTION

The fundamental problem of free will is that it is (on the surface) incompatible with the concept that the world is predictable — that there is causal structure within the world.[1] As causal mechanisms become recognized, there is less and less room for extra-causal mechanisms to explain events. Standard definitions of free will depend on an ability to create changes that are uncaused. The idea that free will is incompatible with causality traces scientifically back to Galileo, Newton, Descartes, and Laplace (Laplace 1814; Earman 1986). However, the complex relationship between mind and behavior has allowed the notion of free will as a driving force in human behavior to persist.[2]

The issue of free will has returned to become an important question for legal and philosophical scholarship because neuroscience threatens to bring it to the fore once again. Although we are influenced by our genetics, the physical environment we experience, and the social structures in which we have lived, the connection between these influences and behavior is multi-dimensional and complex. In contrast, neuroscience provides a *transparent bottleneck* (Greene and Cohen 2004): under the assumption that there is no external supernatural force driving our actions beyond our brains, our behavior is driven directly (if in a complex manner) through neural function. This means that, eventually, it is likely that we will be able

to decipher intention and predict behavior through the observations of neural signals.

This is the problem through which causality is incompatible with free will. Because the neural system causes behavior and we can observe that neural system, we will eventually reach the point where we can predict behavior before the person takes the action. Wherein then lies the *freedom to will*?

Traditionally, there have been three attempts to reconcile this incompatibility.

First, *Cartesian dualism*[3] says that there is a separate being, somehow separable from the physical nature of our brains. This is a descriptive[4] hypothesis, and the evidence against the existence of a non-physical component to cognition or decision-making is overwhelming (see Redish 2013, for review). Dualism says that free will and full causality are incompatible, and that full causality implies determinism, and a loss of free will.

Second, *Determinism* says that free will is an illusion, that our actions are all predetermined by the laws of physics and the state of the world. Of course, current physics suggests that subsequent states of the world are not fully determined, but include random components (quantum randomness).[5] Nevertheless, saying that you are a consequence of predetermined states plus random actions is no more satisfying than true determinism (Greene and Cohen 2004).

Third, *compatibilism* says that we are beings with free will who live in a physical world with causality, and that, somehow, these two concepts are not incompatible. The problem with compatibilism is that it is unclear how a mechanistic machine can be free. I believe that a large part of the problem with incompatibility is that we do not yet have machines complex enough to appear to have free will. As we begin to understand the mechanisms and complexity of decision-making systems, we may find that assuming a socially-cognizant, separable being is a useful conceptualization.

Several authors have suggested that mental states are a separably-understandable software running on a physical hardware, but that the software is explicitly different from the hardware (Sperry 1969; Hofstadter 1985; Newell 1990), and that these *mental states* are somehow fundamentally different from the underlying *physical states* (Morse 2010). I will argue that this is a modern form of dualism and is a hypothesis that does not fit the available data.

THE PHYSICAL NATURE OF COGNITION AND THE DECISION-MAKING SYSTEM(S)

Over the last several decades, a new view of cognition and neural processing has been developed based on the concepts of algorithm, representation, computation, and information processing (Marr 1982; Hofstadter 1985, 2008; Churchland and Sejnowski 1994; Damasio 2010; Redish 2013). Within this theoretical framework, psychological constructs are computational processes occurring across physical neural systems.

A good example of this computational paradigm is the concept of *intention*. Under the computational framework, intention is a representation of a planned action, occurring before that action. As such, there is nothing that says that a physical machine cannot have an intention. To identify that intention one would have to be able to measure and decode the neural representation. This is currently possible using modern neuroscience technology, such as neural ensemble recordings (Georgopoulos et al. 1988; Hatsopoulos et al. 1998; Averbeck et al. 2003; Yang and Shadlen 2007; Johnson et al. 2009). An important consequence of the computational framework, however, is that the physical nature of the intention does not diminish the psychological computational construct served by that intention process.

Fundamentally, free will is about the ability to make decisions, and, as such, the key question first becomes *What is a decision?* In order to study decisions, we need to operationalize them to a form that can be measured. In Redish (2013), I operationalize decision-making as action-selection. This definition is not meant to restrict decisions to simple motor actions such as lifting a finger or ducking a blow, but rather to allow decisions to include any action that affects the world. For example, the decision of which college to attend eventually consists of sending an email or signing a letter to one college saying *Yes, I will attend,* and, hopefully if one is polite, to the other colleges saying *I'm sorry, but I will be going to college elsewhere*. The decision to get married ends in a speech act of standing in front of a society with one's partner and saying *I do*. This definition captures both fast decisions (e.g., whether to swing at a pitch or not) and long, deliberative decisions (e.g., where to go to college). What this definition does is change the question to *What action should I select to do?*

Following the computational framework, one can describe the decision-making system in the mammalian brain as an interaction between four action-selection systems (*Reflexes*, *Pavlovian*, *Deliberative*, and *Procedural*) and four support systems (*Perceptual*, *Motoric*, *Situation-Recognition*, and *Motivational*). Each action-selection system selects actions through a different information-processing algorithm, which will be advantageous in some situations and disadvantageous in others. Mechanisms must also exist with which to decide which action-selection system should drive behavior when they are in conflict. Although there is not space here to provide an in-depth discussion of these systems, nor for the justification or citation list to identify them, I will provide a brief definition of each of the action-selection systems. For further details, see Redish (2013).

- *Reflexes* form a stimulus-response system that reacts quickly to immediately sensed stimuli. Reflexes are information processing algorithms that made animals with them more likely to survive and procreate. As such, reflexes can be thought of as having been learned over evolutionary time-scales, and generally do not entail learning over the course of an individual's lifespan. The classic example is pulling your hand from a burning flame.

- *Pavlovian action-selection* is now known to be based on causal hypotheses of the processes of the world (Rescorla 1988; Bouton 2007).[6] These hypotheses lead to an expectation that the world is in a specific state, which leads to an already established response to that expectation. Unlike the deliberative system, below, these expectations are not about the outcome of the action, but rather about the relationship between the cues and the state of the world. A good example is running from a lion — you can learn that when a lion is stalking you there will be a rustle in the grass, which leads you to fear the rustle in the grass, producing the appropriate behavior — RUN! Pavlovian systems are closely related to emotions (LeDoux 1996, 2012) and turn out to be critical to normal social interactions (Damasio 1994; Sanfey 2007). While the Pavlovian system can learn that cues imply a specific situation (e.g., a rustle in the grass implies a lion may be stalking you), the set of actions that makes up the response is generally not learned

within a given lifespan (preventing running requires other systems, Cavanagh et al. 2013).

- *Deliberative action-selection* is now known to entail an actual imagination and evaluation process (Buckner and Carroll 2007; Johnson and Redish 2007; Hassabis and Maguire 2011; Schacter and Addis 2011; van der Meer et al. 2012). It is particularly useful for behaviors that one cannot (or should not) try multiple times (such as deciding on which college to attend), and for and for novel situations that have never been encountered before. It is slow, conscious, and entails episodic representations of future possibilities.

- *Procedural action-selection* entails releasing action-chains in response to categorized situations. In a sense, procedural learning caches the work into a situation-recognition system, allowing a fast, reliable response (Daw et al. 2005; Dezfouli and Balleine 2012). Although this learning is often discussed in terms of *habit*, it also includes sports behaviors, such as a quarterback deciding whether or not to throw and a batter deciding whether or not to swing. Motor control systems can then take over, ensuring that the behavior is released correctly for the parameters of the situation (Cisek and Kalaska 2010).

CARTESIAN DUALISM: THE DESPERATE DILEMMA

> "Je pense, donc je suis, étoit si ferme et si assurée... je jugeai que je pouvois la recevoir sans scrupule pour le premier principe."
>
> "I think, therefore I am, was so certain... that I decided I could take it as the first principle."
> (Descartes, 1647)

An experienced anatomist, Descartes knew that human neuroanatomy showed a strong similarity to that of other animals (Descartes 1637b). An experienced mathematician and scientist, Descartes knew that the hypothesis of external movers of physical things, like angels moving planets (Aquinas 1274), was no longer tenable given the discoveries of the time (Descartes 1637a). Desperate to deny determinism, Descartes simply stated that a ma-

chine cannot think, that God would not provide him the illusion of free will without it, and that humans have free will (even if animals do not).

But this leaves us with the problem of defining *cognition* and *thought* rather than free will. If we had a machine that could cogitate or think, it would be free by Descartes' definition. Certainly, evidence is mounting that neural signals within the brains of other animals (particularly other mammals, such as rats and monkeys) include representations that we identify with cognition, such as planning signals (Georgopoulos et al. 1989; Johnson and Redish 2007; van der Meer et al. 2012) and regret-related counterfactuals (Abe and Lee 2011; Steiner and Redish 2012). These signals occur under similar conditions, with similar time courses, and in the homologous brain structures that humans need for such cognitions (Shepard and Metzler 1971; Coricelli et al. 2005; Hassabis and Maguire 2011; Schacter and Addis 2011).

Mathematical definitions of cognition as representations decoupled from immediate sensory or motor signals suggest a mechanistic account of cognition (Johnson et al. 2009). There is no reason to think that a mechanistic account of cognition is incompatible with our own conscious experiences (Gray 2004; Damasio 2010; Gazzaniga 2011).

The evidence that cognitive events are physical neural signals is overwhelming; neural signals track cognitive events in both humans and animals, and manipulation of those physical neural signals change cognitive events. For example, monkeys can report switching representations in binocular rivalry, and cellular firing in visual cortex tracks those supposedly conscious events (Logothetis 1998). Rats pausing and making planning decisions represent potential paths ahead of them serially, as if deliberating (Johnson and Redish 2007), and represent potential future outcomes, as if evaluating that deliberation (van der Meer and Redish 2009). These signals appear in structures known to be primarily involved in the deliberative decision-making systems but not in those known to be primarily involved in the procedural decision-making systems (van der Meer et al. 2010). In humans, imagination entails representation in visual cortices (Kosslyn 1994). It is possible to predict what images human subjects will report having dreamed from the activity in their extra-striate visual cortices (Horikawa et al. 2013). Similarly, manipulation of those physical neural signals changes actions consistent with implications of changes in those cognitive

events (Salzman et al. 1992). Lesions of the neural tissue thought to be critical for the computation lead to inability to perform that computation. For example, parietal-lesioned patients show hemineglect of the visual field, even in imagined and remembered spaces (Bisiach and Luzatti 1978; Bisiach and Vallar 1988).

Some researchers have suggested that Cartesian dualism can still be consistent with these observations if we hypothesize that the extra-causal, non-material object interacts with the world in some way (e.g., Eccles et al. 1965; Eccles 1994). These hypotheses have recently tended to concentrate on the potential for this non-material object to affect quantum states (Penrose 1980; Eccles 1994; Hameroff 1994). But this is just another example of Descartes' desperate hypothesis. These theories need to explain why this non-material object connects to humans and not to other animals. Similarly, they need to explain why the non-material object connects to these components of the brain and not to others. Why can we not affect quantum things at a distance (Randi 1982; Kaiser 2011)? At the point where the non-material object is receiving input from the world and providing output to the world, it becomes just another computational object, and there is no reason to assume that that computational object is non-material (Turing 1950).

Finally, these theories cannot explain why decision-making takes time. The frontispiece of the book by Eccles (1994) is of van Leeuwenhoek lost in thought, designing an experiment. But there is no need to hypothesize an external, non-material, extra-causal object taking time to come up with the right decision. Unlike the Cartesian dualist theory, the computational theory explains why planning takes time — due to the computational complexity of the calculations involved. We now have examples of non-human animals that pause to consider options (Tolman 1932; Johnson and Redish 2007; van der Meer et al. 2010) and of machines that take time to find the correct answer to our queries (Nilsson 1980; Russell and Norvig 2002).

THE SOFTWARE/HARDWARE ANALOGY AS A MODERN DUALISM

The computational explanation of deliberation suggests that the key to free will is a computational one. This has led some scholars to argue that there is a separate being hidden within the machine — that we are software running on neural hardware (Sperry 1969;

Hofstadter 1979; Marr 1982; Newell 1990; McHugh and Slavney 1998; Kurzweil 2006; Gazzaniga 2011).

I will argue here that this is a modern version of a dualist hypothesis. The software suggestion is that there are mental states that are dissociable from the physical states underlying them (Morse 2010; Searle 2006; Gazzaniga 2011).[7] Although these authors generally argue that these mental states occur on a physical substrate, they also argue that the separate software can emerge as a truly different entity from the underlying hardware. Unlike modern computers where software can be run on any machine (e.g., reading this text, written on one computer, but read, without change to the text, on another), the translation of information from one neural system to another entails fundamental changes in representation and kind. For example, information consolidated from short-term to long-term memory entails a transformation from episodic representations to semantic representations (Nadel and Moscovitch 1997; Redish and Touretzky 1998; Nadel and Bohbot 2001; Corkin 2002). When the hippocampal-lesioned patient H.M. learned new tasks, he used a different memory system which made decisions based on different information processing mechanisms (Milner et al. 1968; Cohen and Squire 1980). This implies that, although neural systems are information processing machines, the information is tightly coupled to the neural hardware.

Searle (2006) argues that there could be an emergent effect of networks that produces an agency neither visible nor derivable from the lower-level components. He gives the example of a wheel that has properties (it rolls) different from the properties of the individual molecules. However, this observation does not preclude the possibility (or the fact) that one can derive the emergent property from the interactions of the component molecules. Gazzaniga (2011) makes a similar point that interactions between objects can have effects that are not obvious from their component parts. However, one should still be able to reliably predict those effects from expected interactions between known parts. Gazzaniga's example of traffic patterns as not being explainable from nuts and bolts of cars does not preclude the fact that traffic patterns are predictable from knowledge of the physical properties of cars, roads, traffic lights, and the psychology of human behavior.

In any case, these emergent effects of neurons, bodies, and interactions may produce a different being than simply having a set of neurons sitting in a dish, but this does not get at the fundamental

problem of how that network can achieve *freedom*, nor does it get at the fundamental problem of what we are going to do when we can read the information represented within that network.

AUGUSTINIAN DUALISM: DELIBERATION AND EMOTION

> "Make me chaste, O Lord... but not today."
> (Augustine of Hippo, Saint Augustine, 398)

Although Cartesian dualism no longer plays a strong role in neuroscientific discussions, there remains another form of dualism that pervades much of neurophilosophical scholarship (Wechsler 1962; Sapolsky 2004; Haidt 2006; Eagleman 2011; Gazzaniga 2011, see Jones et al., forthcoming, for additional review). We now know that there are multiple systems which can drive the actions we take. Some of these systems include conscious planning, whereas others do not (Cohen and Squire 1980; Mishkin and Appenzeller 1987; Redish et al. 2008; Kurzban 2010; Kahneman 2011; Bargh et al. 2012). These systems are separable in that they depend on different brain structures and select their actions based on different algorithms that process information in different ways (O'Keefe and Nadel 1978; Daw et al. 2005; Cohen and Squire 1980; Redish et al. 2008; Montague et al. 2012; van der Meer et al. 2012; Redish 2013).

Many neuroscientists define decision-making as deliberative decision-making only. Often this is stated in terms of *conscious* versus *unconscious* actions (Greene and Cohen 2004; Haidt 2006; Eagleman 2011), and it leads to a major conceptual problem where some intentions are identifiable before or without conscious thought (Libet et al. 1979; Kurzban 2010; Gazzaniga 2011). The fact that some actions are performed deliberatively and others non-deliberatively has been known since ancient times (Plato 4th century BCE; Augustine of Hippo 398). In modern terminology, this division is often called the dual-system hypothesis (Bechara and van der Linden 2005; McClure et al. 2004) and often described in terms of a rider trying to control a horse or other powerful creature (e.g., an elephant, Haidt 2006).

I call this *Augustinian* dualism from Augustine's psychological dichotomy between the animal (emotional) and reason (rational, *cognitive*). In particular, Augustine argues that the key to being a

successful human is the ability of reason to overwhelm and control the animal emotionality (Augustine of Hippo 398, 427). As pointed out by Lehrer (2009, see also Eagleman 2011), this duality reflects Plato's concept of a charioteer driving two horses, a wild, emotional, Dionysian horse and an intellectual, Apollonian horse (Plato 4th century BCE), and Freud's three components of the *triune* brain (Freud 1923; MacLean 1990). In the modern version, there are only two beings — an Apollonian (deliberative) rider trying to control a Dionysian (emotional) horse.

The problem with this description is that purely rational beings make fundamentally different decisions than typical humans and, in fact, often appear pathological to other humans (Damasio 1994; Singer 2008; Zak 2008; Smith 2009). In games such as the ultimatum game, emotional (Pavlovian) responses are critical to normal behavior (Sanfey 2007), and cognitive load (which interferes with deliberative systems, see Kurth-Nelson and Redish 2012) drives players to be more, not less, fair (Schulz et al., in press).

However, neither literary descriptions of human behavior, nor our own introspective experiences, separate the emotional and deliberative systems into two separate beings. People describe themselves as being *afraid* in conditions that engender fear (such as standing in the open field and hearing the rustle in the grass). Similarly, sports stars talk about themselves as being *in the zone* not *out of body* when selecting actions successfully (when their procedural systems are working well). Introspectively, most people feel a sense of accomplishment when they swing the bat and hit the home run.

Neurophysiologically, one of the most important components of the deliberative system is the prefrontal cortex. This means that with prefrontal damage the other action-selection systems tend to come to the fore more often. A surprising number of neuroscientists have argued that a being with prefrontal damage no longer has free will (Greene and Cohen 2004; Sapolsky 2004, see Jones et al., forthcoming, for additional examples).

In my opinion, identifying the *self* with the deliberative system diminishes who we are and reduces our actual decision-making processes. We are both the horse and the rider. We are our reflexes, our Pavlovian, emotional responses, our deliberative decisions, and our learned procedural action-chains, all working together as a single being.

IMPLICATIONS FOR AGENCY

The idea that we are physical beings who consist of multiple decision-making systems, each of which processes information differently, has important implications for agency, particularly for the legal concept of *mens rea*, which is based on the idea that a lack of requisite awareness and intention reduces culpability (Wechsler 1962, see Jones et al., forthcoming, for additional review). In order to address these issues, we need to examine four potential departures from agency: (1) that of being forced through threat or manipulation, (2) that of a true, uncontrolled accident, (3) that of a dysfunction in the underlying physical nature of the neural system, and (4) that of actions being driven by other (non-deliberative) decision-making systems.

PHYSICAL ISSUES OF AGENCY

BEING FORCED TO ACT THROUGH THREAT OR MANIPULATION.

The classic examples of this are situations in which one has a gun held to one's head, which changes the set of available options. In such a situation even a rational actor may decide to commit the crime. One might say that this rational actor has a guilty mind, but makes an understandable choice, given the available options.

Similarly, one can include in this category situations in which the agent has incorrect or even disinformation. The question of culpability certainly has to depend on the information available to the agent. Of course, this raises questions of whether the agent is lying by saying that it did not know the previous information.[8] In addition, the question of incorrect information raises questions of whether the agent should have (or could have) determined whether the information was correct.

Another example of this first lack of agency is when a person is pushed into another person. Imagine a situation in which a car is stopped at a traffic light and another car hits the first car from behind, driving the first car into a crossing pedestrian. We would not want to blame the first driver for the death of the pedestrian. Of course, if the first driver has the car in neutral and does not have the brakes on, we might be less inclined to be so forgiving.

A TRUE, UNCONTROLLED ACCIDENT.

If a driver hits the brakes appropriately, but skids uncontrollably

through an icy intersection and kills a pedestrian, we do not declare the incident equivalent to an intentional murder. Questions of culpability in these situations tend to relate to issues of recklessness. In such situations, questions of culpability depend on whether the agent was doing something incorrect that led to the error. For example, if the driver was going too fast for the road conditions, or had refused to replace tires known to have poor traction.

But sometimes the error is in the learning or the lack of learning. Should we blame the driver who skidded on the ice if that driver has not learned how to correctly respond to a skid? *One implication of including the procedural system as part of the agent is that sometimes the crime is in the learning.* This is the point of licensing. Certainly, we blame a kid who takes a car out for a spin, without a driver's license or any of the (minimal) training that requires. Pilots, doctors, and other professionals who perform dangerous jobs are required to be licensed to ensure that they have the necessary training.

I have spent time on these classic, physical issues of culpability because they set the stage for neural issues of culpability.

NEURAL ISSUES OF AGENCY

Neural systems are physical systems. As previously discussed, there is no evidence of a non-physical being controlling our neural system, and, in fact, there is extensive evidence against it (see discussion of Cartesian dualism, above). However, nothing in our discussion suggests a revision of the concept of individuality: I am a different individual from you, and our physical nature does not imply a lack of separation of our being. Thus, there is a difference between an outside agent physically acting upon the person and neural processing errors occurring within the person.

A DYSFUNCTION IN THE PHYSICAL NATURE OF THE NEURAL SYSTEM.

The classic example of a physical dysfunction is that of an epileptic who hits someone during a seizure, or someone who passes out while driving due to a syncope event. Of course, these examples raise the same issues of recklessness and learning as do physical accidents. If someone knew that they were prone to syncope, then perhaps they should not be handling dangerous equipment. Researchers are now working on epileptic seizure warning systems (Mormann et al. 2007; Iasemidis 2011), which would allow a driver to pull over before the seizure occurs.

Much of neural-related legal scholarship is concerned with the medicalization of neural dysfunction (Eagleman 2011; Jones et al., forthcoming). The classic case is that of a tumor in the brain; a patient may act differently with the tumor present than before the tumor appeared or after the tumor is treated.[9] Some defendants have argued that they should not be blamed for actions that occurred when they had their tumor (See Jones et al., forthcoming, for cases and discussion).

The problem with this logic is that there is no non-physical person apart from the physical brain. One option is to declare that there are two people separated temporally — the one *before the tumor* and the one *with the tumor*. This is dangerous because it provides opportunities for a legal fiction that does not recognize the danger of the tumor (or other effect) returning.

More usefully, one can try to address the issues of the purpose of retribution. Extensive new research examining societies and multi-agent interactions suggests that retribution serves an important purpose of reducing interpersonal defections (Sober and Wilson 1998; Wilson 2002; Zak 2008; Smith 2009). In situations where treatment could prevent future conditions better than retribution, it might make sense to deal with treatment rather than retribution, or to find some reasonable combination of the two. One strategy is to insist that one be responsible for amelioration of one's own disabilities, as in an addict having to provide continued negative urine samples as a condition for future employment, or a tumor victim reporting for regular checkups (Bonnie 2002).

Both of these examples, however, are dysfunctions in the physical nature of the body, not dysfunctions in the information processing of the neural system itself. This leads us to the final example, in which actions are driven by other decision-making systems.

ACTIONS DRIVEN BY OTHER (NON-CONSCIOUS) DECISION-MAKING SYSTEMS.

It has been known for millennia that humans sometimes take actions that were not initiated by a rational, conscious mind. In ancient times, it was assumed that another being was controlling behavior when one behaved irrationally, whether it be an animal past (Augustine of Hippo 398), or some supernatural being (Boyer 2002). Contemporary interpretations admit the unified, physical nature of the being, but continue to suggest a rider attempting to control a wild animal (Freud 1923; Haidt 2006; Eagleman 2011).

The multiple decision-making systems theory, however, implies that the person we identify as a given individual consists of all of these systems working together (Kurzban 2010; Redish 2013). Each of these action-selection systems processes information differently and has advantages and disadvantages. For example, *deliberative* systems are capable of planning novel paths (O'Keefe and Nadel 1978; Gupta et al. 2010), of attending to different aspects of the available options (Hill 2008), and, thus, of taking immediate motivation factors into account (Niv et al. 2006). On the other hand, deliberative systems are slow and prone to variability (van der Meer et al. 2012). *Procedural* learning allows the caching of action-sequences (Dezfouli and Balleine 2012), but leads to habits that can be hard to break (Niv et al. 2006; Redish et al. 2008). *Pavlovian* systems allow complex species-specific behaviors that do not have to be learned. For example, appropriate conspecific human social interaction depends on normal Pavlovian behaviors (Damasio 1994; Sanfey 2007), but can also drive inappropriate attention to reward-predictive cues (e.g., sign-tracking vs. goal-tracking, Flagel et al. 2011).

The discussion of neural implications for legal scholarship tend to be based on the Augustinian dualism error that suggests that non-conscious components of decision-making are irrational and not part of our free will (Sapolsky 2004; Greene and Cohen 2004; Bechara and van der Linden 2005; Eagleman 2011). Similarly, legal scholarship suggests that *mens rea* depends on conscious, rational deliberation (Wechsler 1962), which (as noted above) causes problems when a crime is committed under clearly non-deliberative control.

Nevertheless, common law has consistently derived different retributive responses to *emotional* and *rational* crimes (e.g., passion-driven vs. premeditated). I would argue that the legal concept of different retributive responses to these different crimes is both reasonable and consistent, when reinterpreted in the context of multiple decision-making systems. It is not that the person is *less culpable* or that the person is unable to *control him or herself*, but rather that we punish Pavlovian errors (crimes) differently from deliberative errors (crimes). If we include recklessness, or lack of training, in the mix, then we would say that we punish procedural errors (crimes) differently, as well.

This reinterpretation is particularly important in situations where there is *diminished capacity*, for example, after prefron-

tal cortical damage (which diminishes the ability for self-control (Baumeister et al. 1994)). Current discussions are based on the idea that a person with prefrontal damage is less culpable (Sapolsky 2004; Greene and Cohen 2004). However, the multiple decision-making system theory implies that such a person is more likely to respond with a Pavlovian or procedural system. Perhaps we should deal with them accordingly.

SUMMARY

In this paper, I have addressed the implications of three dualist errors — the Cartesian error (*If we are machines, then wherein lies the decision-making?*), the emergent software error (*The self emerges from the hardware but is separable from it.*), and the Augustinian error (*Non-conscious decisions are not under "your" control.*). These errors have implications for the concepts of free will and responsibility.

Many neuroscientists have come to an answer to the Cartesian dilemma (*How can one be both a machine and free?*) that lies in the fact that machines can make decisions (Kishida 2012; Redish 2013) and in the computational conceptual framework in which psychological phenomena emerge from interactions of underlying computational processes (MacCorquodale and Meehl 1954; Marr 1982; Churchland and Sejnowski 1994; Gazzaniga 2011).

However, many neuroscientists have also suggested that the unconscious, automatic processing that drives much of decision-making (Kahneman 2011; Bargh et al. 2012) is not reflective of the being in question (e.g., *zombie processes*, Eagleman 2011), and that the conscious, intentional being (the one with *mens rea*) is only a superficial surface of an underlying maelstrom (Kurzban 2010; Eagleman 2011; Gazzaniga 2011).[10]

In my opinion, this Augustinian error is also a dualistic fallacy, in that these other decision-making processes (Pavlovian systems, procedural systems, perceptual systems, etc.) are also part of the decision-making system that is the being in question. These systems can learn, process input, and make decisions. And therein, I contend, lies the key to responsibility.

As pointed out by Gazzaniga (2011), if something breaks in your car, it is useless to punish the car. Instead, one should simply fix it. One cannot *blame* the car. If people are also mechanistic machines, then how can one blame the perpetrator of a crime?

However, Gazzaniga notes that if your transportation vehicle were a horse rather than a car, then one might be more tempted to punish it. I argue that, fundamentally, this is because a horse is a learning system, and retributive punishment can change its behavior. As pointed out by Kishida (2012), it is now possible to build algorithms that learn from reward and punishment. Whether we are physical machines or not is irrelevant to the question. We may be machines, but we are machines with decision-making abilities, including both intention and free will.

Whether one should punish crimes, medicalize them (treat the underlying cause), or some combination thereof depends on specific circumstances. I contend that taking a more complete account of the decision-making processes that underlie human behavior will be an important part of understanding those specific circumstances. As pointed out by Bonnie (2002), the correct response to impaired decision-making processes requires differentiating responsibility for the impaired process, responsibility for the action at the time, and responsibility for ameliorating the impaired process. In particular, both retribution (as deterrent and as punishment to drive learning) and treatment (to improve decision-making processes) are important components. Importantly, however, one should not lose sight of the significance of compassion, forgiveness, and mercy, all of which play important roles in the social interactive contract in which we live (Sober and Wilson 1998; Boyer 2002; Wilson 2002).

ENDNOTES

1. The difference between predictability (that an event A is likely to follow a preceding event B) and causality (that event A occurs as a consequence of event B) is a large and complex topic. In part, it depends on potential explanations for mechanism (Ben-Ari 2005). Recent work has suggested that Bayesian analysis and latent variables can be used to infer causality via Causal Model Networks (Pearl 1988, 2009). However, the identification of appropriate latent variables as explanations for mechanism is complex and subtle. This subtlety is not critical to the issues being discussed here.

2. Obviously, humans are partially predictable. We have personalities. Our social structure is fundamentally dependent on being able to predict what other humans will do. However, these predictions occur within a probabilistic range, and free will can still be assumed to shift decisions within that personality.

3. Smart (1961) and Greene and Cohen (2004) refer to this as libertarianism, but I will avoid the term since it could be confused with political libertarianism, which is an entirely different beast.

4. That is, one that attempts to describe the world as it is, as compared to a prescriptive one that attempts to describe the world as it should be. By definition, descriptive hypotheses must be compatible with our observations of the world (Sagan 1997; Dawkins 2004; Ben-Ari 2005).

5. It is extremely important to differentiate quantum randomness (in which one state of the world can proceed to multiple other states and that one cannot predict with 100% accuracy which of those subsequent states it will be) from the Hameroff (1994), Penrose (1980), and Eccles (1994) hypotheses that an external being is manipulating those quantum effects.

6. Calling this second action-selection system Pavlovian is controversial. However, the mathematical description of the information processing of this system is very well defined; responses predefined within the agent (presumably through evolution) are released because of an expected causal structure predicted from the world (Montague et al. 2012; van der Meer et al. 2012; Cavanagh et al. 2013). See Redish (2013) for a detailed description of the specifics of how this system processes information to select actions and for additional discussion of the appropriateness of the terminology.

7. These authors often argue for determinism and explicitly reject the concept of an extra-causal, non-material soul that is separate from the physical world (Searle 2006; Morse 2010; Gazzaniga 2011).

8. The issue of lie-detection is a continuing, contentious question that depends on the mechanisms of memory and semantic information processing (Loftus and Palmer 1974; Wells and Loftus 1984; Schacter 2001). Since our discussion in this paper is about decision-making and not memory, I will avoid getting side-tracked by the discussion of how one determines whether the agent really did or did not know the previous information.

9. Note the importance of physicality here — the tumor affects the person, there is no Cartesian dualism!

10. Eagleman (2011) suggests an analogy to a young CEO inheriting an already smoothly-running company. Kurzban (2010) suggests an analogy to a press secretary trying to explain the company's behavior. Gazzaniga (2011) suggests an analogy to an interpreter trying to interpret not only stimuli from the outside world but also one's inscrutable behavior.

ACKNOWLEDGMENTS

I would like to thank Steve Kelley and the Consortium on Law and Values in the Health, Environment, and Life Sciences whose discussion session on Laurence Steinberg's work first pricked my interest in this topic, and to thank Francis Shen, for providing me an early draft of the Jones/Schall/Shen book, which helped provide for me a starting point in the legal scholarship. I would also like to thank Laurence Steinberg, Susanna Blumenthal, Nate Powell, and Francis Shen for discussions, and Francis Shen, Adam Johnson, Seiichiro Amemiya, Nate Powell, and Adam Steiner, as well as two anonymous reviewers, for comments on an early draft of this paper. Any errors, of course, remain my own.

REFERENCES

Abe H, Lee D (2011) Distributed coding of actual and hypothetical outcomes in the orbital and dorsolateral prefrontal cortex. Neuron 70:731–741

Aquinas T (1274/1947) *Summa Theologica*. Fathers of the English Dominican Province (trans). Retrieved from http://www.sacred-texts.com/chr/aquinas/summa/index.htm

Augustine of Hippo (398/1961) *Confessions*. Pine-Coffin RS (trans). Penguin, New York, NY

Augustine of Hippo (427/1972) *The City of God*. Bettenson H (trans). Penguin, New York, NY

Averbeck BB, Crowe DA, Chafee MV, Georgopoulos AP (2003) Neural activity in prefrontal cortex during copying geometrical shapes II. Decoding shape segments from neural ensembles. Exp Brain Res 150:142–153

Bargh JA, Schwader KL, Hailey SE, Dyer RL, Boothby EJ (2012) Automaticity in social cognitive processes. Trends Cogn Sci 16:593–605

Baumeister RF, Heatherton TF, Tice DM (1994) *Losing Control: How and why people fail at self-regulation.* Academic Press, San Diego, CA

Bechara A, van der Linden M (2005) Decision-making and impulse control after frontal lobe injuries. Curr Opin Neurol 18:734–739

Ben-Ari M (2005) *Just a Theory: Exploring the nature of science*. Prometheus Books, Amherst, NY

Bisiach E, Luzatti C (1978) Unilateral neglect of representational space. Cortex 14:129–133

Bisiach E, Vallar G (1988) Hemineglect in humans. In: Boller F, Grafman J (eds) *Handbook of Neuropsychology*. Elsevier, New York, NY, pp 195–222

Bonnie RJ (2002) Responsibility for addiction. J Amer Acad Psychiatry Law 30:405-413

Bouton ME (2007) *Learning and Behavior: A contemporary synthesis*. Sinauer Associates, Sunderland, MA

Boyer P (2002) *Religion Explained*. Basic Books, New York, NY

Buckner RL, Carroll DC (2007) Self-projection and the brain. Trends Cogn Sci 11:49–57

Cavanagh JF, Eisenberg I, Guitart-Masip M, Huys Q, Frank MJ (2013) Frontal theta overrides Pavlovian learning biases. J Neurosci 33:8541–8548

Churchland P, Sejnowski TJ (1994) *The Computational Brain*. MIT Press, Cambridge, MA

Cisek P, Kalaska JF (2010) Neural mechanisms for interacting with a world full of action choices. Ann Rev Neurosci 33:269–298

Cohen NJ, Squire LR (1980) Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. Science 210:207–210

Coricelli G, Critchley HD, Joffily M, O'Doherty JP, Sirigu A, Dolan RJ (2005) Regret and its avoidance: A neuroimaging study of choice behavior. Nature Neurosci 8:1255–1262

Corkin S (2002) What's new with the amnesic patient H.M.? Nature Rev Neurosci 3:153–160

Damasio A (1994) *Descartes' Error: Emotion, reason, and the human brain*. Grosset/Putnam, New York, NY

Damasio A (2010) *Self Comes to Mind: Constructing the conscious brain*. Pantheon, New York, NY

Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature Neurosci 8:1704–1711

Dawkins R (2004) *The Ancestor's Tale: A pilgrimage to the dawn of evolution*. Houghton Mifflin, New York, NY

Descartes R (1637a/1886/2008) *La Géométrie (The Geometry)*. Hermann A (ed). Project Gutenberg, ebook

Descartes R (1637b/2008) *Discourse on the Method of Rightly Conducting One's Reason and of Seeking Truth*. Veitch J (trans). Project Gutenberg, ebook

Descartes R (1647/1989) *The Passions of the Soul*. Voss SH (trans). Hackett, Indianapolis, IN

Dezfouli A, Balleine B (2012) Habits, action sequences and reinforcement learning. Euro J Neurosci 35:1036–1051

Eagleman D (2011) *Incognito: The secret lives of the brain*. Vintage, New York, NY

Earman J (1986) *A Primer on Determinism*. Springer, Berlin, Germany

Eccles JC (1994) *How the SELF Controls Its BRAIN*. Springer-Verlag, Berlin, Germany

Eccles JC, et al. (1965) Final discussion. In: Eccles JC (ed) *Brain and Conscious Experience*. Springer-Verlag, Berlin, Germany, pp 548–574

Flagel SB, Clark JJ, Robinson TE, Mayo L, Czuj A, Willuhn I, Akers CA, Clinton SM, Phillips PEM, Akil H (2011) A selective role for dopamine in stimulus-reward learning. Nature 469:53–57

Freud S (1923/1961) *The Ego and the Id*. In: Strachey J (ed and trans) *The Standard Edition of the Complete Psychological Works of Sigmund Freud (vol.19)*. Hogarth Press, London, UK

Gazzaniga M (2011) *Who's in Charge?* HarperCollins, New York, NY

Georgopoulos AP, Kettner RE, Schwartz AB (1988) Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. J Neurosci 8:2928–2937

Georgopoulos AP, Lurito JT, Petrides M, Schwartz AB, Massey JT (1989) Mental rotation of the neuronal population vector. Science 243:234–236

Gray J (2004) *Consciousness: Creeping up on the hard problem*. Oxford University Press, New York, NY

Greene J, Cohen J (2004) For the law, neuroscience changes nothing and everything. Phil Trans R Soc Lond B 359:1775–1785

Gupta AS, van der Meer MAA, Touretzky DS, Redish AD (2010) Hippocampal replay is not a simple function of experience. Neuron 65:695–705

Haidt J (2006) *The Happiness Hypothesis*. Basic Books, New York, NY

Hameroff SR (1994) Quantum coherence in microtubules: A neural basis for emergent consciousness? J Consciousness Stud 1:91–118

Hassabis D, Maguire EA (2011) The construction system in the brain. In: Bar M (ed) *Predictions in the Brain: using our past to generate a future*. Oxford University Press, New York, NY, pp 70-82

Hatsopoulos NG, Ojakangas CL, Paninski L, Donoghue JP (1998) Information about movement direction obtained from synchronous activity of motor cortical neurons. Proc Nat Acad Sci 95:15706–15711

Hill C (2008) The rationality of preference construction (and the irrationality of rational choice). Minnesota J Law Sci Technol 9:689–742

Hofstadter DR (1979) *Gödel, Escher, Bach: An eternal golden braid*. Basic Books, New York, NY

Hofstadter DR (1985) *Metamagical Themas: Questing for the essence of mind and pattern*. Basic Books, New York, NY

Hofstadter DR (2008) *I am a Strange Loop*. Basic Books, New York, NY

Horikawa T, Tamaki M, Miyawaki Y, Kamitani Y (2013) Neural decoding of visual imagery during sleep. Science 340:639–642

Iasemidis LD (2011) Seizure prediction and its applications. Neurosurgery Clinics N Amer 22:489–506

Johnson A, Fenton AA, Kentros C, Redish AD (2009) Looking for cognition in the structure in the noise. Trends Cogn Sci 13:55–64

Johnson A, Redish AD (2007) Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. J Neurosci 27:12176–12189

Jones OD, Schall JD, Shen FX, eds (forthcoming) *Law and Neuroscience*. Aspen Publishers, New York, NY

Kahneman D (2011) *Thinking, Fast, and Slow*. Farrar, Straus and Giroux, New York, NY

Kaiser D (2011) *How the Hippies Saved Physics: Science, counterculture, and the quantum revival*. W. W. Norton, New York, NY

Kishida KT (2012) A computational approach to "free will" constrained by the games we play. Frontiers Integrative Neurosci 6:1–6

Kosslyn SM (1994) *Image and Brain*. MIT Press, Cambridge, MA

Kurth-Nelson Z, Redish AD (2012) A theoretical account of cognitive effects in delay discounting. Euro J Neurosci 35:1052–1064

Kurzban R (2010) *Why Everyone (else) is a Hypocrite*. Princeton University Press, Princeton, NJ

Kurzweil R (2006) *The Singularity is Near: When humans transcend biology*. Penguin, New York, NY

Laplace PS (1814/1902) A Philosophical Essay on Probabilities. Truscott FW, Emory FL (trans). Chapman & Hall, London, UK; Wiley, New York, NY

LeDoux J (2012) Rethinking the emotional brain. Neuron 73:653–676

LeDoux JE (1996) *The Emotional Brain*. Simon & Schuster, New York, NY

Lehrer J (2009) *How We Decide*. Houghton Mifflin Harcourt, Boston, MA

Libet B, Wright EW, Feinstein B, Pearl DK (1979) Subjective referral of the timing for a conscious sensory experience. Brain 102:193–224

Loftus E, Palmer J (1974) Reconstruction of automobile destruction. J Verbal Learn Verbal Behav 13:585–589

Logothetis NK (1998) Single units and conscious vision. Phil Trans R Soc Lond B Biol Sci 353:1801–1888

MacCorquodale K, Meehl PE (1954) Edward C. Tolman. In: Estes W (ed) *Modern Learning Theory*. Appleton-Century-Crofts, New York, NY, pp 177–266

MacLean PD (1990) *The Triune Brain in Evolution*. Plenum Press, New York, NY

Marr D (1982) *Vision*. W. H. Freeman and Co, New York, NY

McClure SM, Laibson DI, Loewenstein G, Cohen JD (2004) Separate neural systems value immediate and delayed monetary rewards. Science 306:503–507

McHugh PR, Slavney PR (1998) *The Perspectives of Psychiatry*. Johns Hopkins University Press, Baltimore, MD

Milner B, Corkin S, Teuber H (1968) Further analysis of the hippocampal amnesia syndrome: 14-year follow-up study of H. M. Neuropsychologia 6:215–234

Mishkin M, Appenzeller T (1987) The anatomy of memory. Scientific American 256:80–89

Montague PR, Dolan RJ, Friston KJ, Dayan P (2012) Computational psychiatry. Trends Cogn Sci 16:72–80

Mormann F, Andrzejak RG, Elger CE, Lehnertz K (2007) Seizure prediction: The long and winding road. Brain 130:314–333

Morse SJ (2010) Lost in translation? An essay on law and neuroscience. In: Freeman M (ed) *Current Legal Issues 2010*. Oxford University Press, New York, NY, pp 530–562

Nadel L, Bohbot V (2001) Consolidation of memory. Hippocampus 11:56–60

Nadel L, Moscovitch M (1997) Memory consolidation, retrograde amnesia and the hippocampal complex. Curr Opin Neurobiol 7:217–227

Newell A (1990) *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA

Nilsson N (1980) *Principles of Artificial Intelligence*. Tioga Press, Palo Alto, CA

Niv Y, Joel D, Dayan P (2006) A normative perspective on motivation. Trends Cogn Sci 10:375–381

O'Keefe J, Nadel L (1978) *The Hippocampus as a Cognitive Map*. Clarendon Press, Oxford, UK

Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference*. Morgan Kaufmann, San Mateo, CA

Pearl J (2009) *Causality: Models, reasoning and inference*. Cambridge University Press, New York, NY

Penrose R (1980/1989) *The Emperor's New Mind*. Penguin Books, New York, NY

Plato (4th century BCE/2008) *Phaedrus*. Jowett B (trans). Project Gutenberg, ebook

Randi J (1982) *Flim-Flam!* Prometheus Books, Buffalo, NY

Redish AD (2013) *The Mind within the Brain: How we make decisions and how those decisions go wrong*. Oxford University Press, New York, NY

Redish AD, Jensen S, Johnson A (2008) A unified framework for addiction: Vulnerabilities in the decision process. Behav Brain Sci 31:415–487

Redish AD, Touretzky DS (1998) The role of the hippocampus in solving the Morris water maze. Neural Computation 10:73–111

Rescorla RA (1988) Pavlovian conditioning: it's not what you think it is. Amer Psychologist 43:151–160

Russell SJ, Norvig P (2002) *Artificial Intelligence: A modern approach*. Prentice Hall, Upper Saddle River, NJ

Sagan C (1997) *The Demon-Haunted World*. Random House, New York, NY

Salzman CD, Murasugi CM, Britten KH, Newsome WT (1992) Microstimulation in visual area MT: Effects on direction discrimination performance. J Neurosci 12:2331–2355

Sanfey AG (2007) Social decision-making: Insights from game theory and neuroscience. Science 318:598–602

Sapolsky RM (2004) The frontal cortex and the criminal justice system. Phil Trans R Soc Lond B 359:1787–1796

Schacter DL (2001) *The Seven Sins of Memory*. Houghton Mifflin, Boston, MA

Schacter DL, Addis DR (2011) On the nature of medial temporal lobe contributions to the constructive simulation of future events. In: Bar M (ed) *Predictions in the Brain: Using our past to generate a future.* Oxford University Press, New York, NY, pp 58–69

Schulz J, Fischbacher U, Thöni C, Utikal V (in press) Affect and fairness: Dictator games under cognitive load. J Econ Psychol

Searle J (2006) *Freedom and Neurobiology: Reflections on free will, language, and political power*. Columbia University Press, New York, NY

Shepard RN, Metzler J (1971) Mental rotation of three-dimensional objects. Science 171:701–703

Singer T (2008) Understanding others: brain mechanisms of theory of mind and empathy. In: Glimcher PW, Camerer C, Poldrack RA, Fehr E (eds) *Neuroeconomics: Decision making and the brain*. Academic Press, New York, NY, pp 251–268

Smart JJC (1961) Free will, praise and blame. Mind 70:291–306

Smith V (2009) *Rationality in Economics: Constructivist and ecological forms*. Cambridge University Press, Cambridge, UK

Sober E, Wilson DS (1998) *Unto Others: The evolution and psychology of unselfish behavior*. Harvard University Press, Cambridge, MA

Sperry RW (1969) A modified concept of consciousness. Psychol Rev 76:532–536

Steiner A, Redish AD (2012) Orbitofrontal cortical ensembles during deliberation and learning on a spatial decision-making task. Frontiers Decision Neurosci 6:131

Tolman EC (1932) *Purposive Behavior in Animals and Men*. Appleton-Century-Crofts, New York, NY

Turing AM (1950) Computing machinery and intelligence. Mind 59:433–460

van der Meer MAA, Johnson A, Schmitzer-Torbert NC, Redish AD (2010) Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. Neuron 67:25–32

van der Meer MAA, Kurth-Nelson Z, Redish AD (2012) Information processing in decision-making systems. The Neuroscientist 18:342–359

van der Meer MAA, Redish AD (2009) Covert expectation-of-reward in rat ventral striatum at decision points. Frontiers Integrative Neurosci 3:1–15

Wechsler H (1962) On culpability and crime: The treatment of mens rea in the Model Penal Code. Annals Amer Acad Political Social Sci 339:24–41

Wells GL, Loftus EF, eds (1984) *Eyewitness Testimony*. Cambridge University Press, New York, NY

Wilson DS (2002) *Darwin's Cathedral: Evolution, religion, and the nature of society*. University of Chicago Press, Chicago, IL

Yang T, Shadlen MN (2007) Probabilistic reasoning by neurons. Nature 447:1075–1080

Zak PJ, ed (2008) *Moral Markets: The critical role of values in the economy*. Princeton University Press, Princeton, NJ