The Human Mirror Neuron System (MNS): Toward a Motivated Autonomous Agent

David B. Newlin, Ph.D.
Research Triangle Institute (RTI International)
6801 Eastern Avenue, Suite 203, Baltimore, MD 21224
dnewlin@rti.org

Abstract

The purpose for this discussion is to provide information on the human mirror neuron system (MNS) to aid in the design of artificial autonomous agents. Specifically, we emphasize the motivational and affective aspects of the MNS to allow the design of agents that themselves exhibit motivated behavior. Current evidence indicates that the MNS assigns motives to observed behavior. A tentative working model of the MNS is presented to organize this information and to highlight important issues that remain unresolved, such as the architecture of hemispheric asymmetry of function in the left and right MNSs and mesolimbic motivation systems. We conclude with recommendations—particularly in terms of dual, functionally asymmetric control systems—for designing motivated autonomous agents using information learned about the human MNS.

What Are Mirror Neurons?

Imitation(?)

Mirror neurons fire both when a person performs a specific action and when that person observes or imagines someone else performing the same behavior (Oberman & Ramachandran, 2007; Rizzolatti & Craighero, 2004). That is, the mirror neuron system (MNS) responds when one's behavior is "mirrored" in others. For example, if one is drinking from a cup, a specific array of neurons is activated. This is ordinary. What is extraordinary is that this same MNS also responds when one sees someone else drinking from a cup. This characteristic has led many brain scientists to view the MNS as the neurobiological foundation for imitative learning (Oberman & Ramachandran, 2007; Rizzolatti & Craighero, 2004). For our purposes, we suggest that autonomous agents may need the equivalent of mirror neurons to acquire the capacity for learning through mimicry and for social learning, both desirable traits. These abilities appear highly adaptive for humans. Some have even suggested that the proliferation of mirror neurons in the hominid brain presented an advantage, particularly in terms of tool use, language, and social organization, that led directly to the evolutionary transition to modern humans (*Homo sapiens)*.

The MNS is not unique to humans. In fact, the original discovery of these brain cells was in macaque monkeys (Di Pellegrino et al., 1992; Gallese et al., 1996). The obvious interpretation of this surprising discovery was that mirror neurons support imitation. However, there is a fatal problem with this notion: monkeys simply do not engage in mimicry or imitative learning (although humans certainly do). This indicates the need for new hypotheses about the psychological functions that the MNS serves in monkeys, though it does not imply that the MNS cannot be the biological substrate for mimicry in people. Lyons et al. (2006) argued that in Rhesus monkeys (macaques), mirror neurons represent a social processing system that "extracts the goal structure of observed action" (Lyons et al., 2006). In other words, mirror neurons assess and predict the consequences and behavioral motivations of others. This is far more complex processing than simple motor mimicry.

The primary regions of mirror neurons studied in the Rhesus monkey are the ventral premotor cortex (area F5), the superior temporal sulcus (STS), and the rostral part of the inferior parietal lobe (area PF). While motor neurons have been measured typically with single-cell electrophysiology in monkeys, the MNS has been mapped in human brains through neuroimaging techniques, particularly functional magnetic resonance imaging (fMRI).

The Human MNS

The MNS is more broadly distributed in the human brain than was originally thought. For example, Dinstein et al. (2007) mapped the MNS in humans using fMRI in overlapping brain regions during both executed movements and observed movements (their proposed definition of the MNS). These included the ventral premotor cortex and several areas in the parietal lobe (anterior intraparietal sulcus, anterior intraparietal sulcus, superior intraparietal sulcus, and posterior intraparietal sulcus), as well as an area within the lateral occipital cortex.

The MNS in humans also has been assessed using scalp electrophysiology (EEG). Suppression or desynchronization of the rolandic *mu* rhythm (8-13 Hz. recorded over the sensorimotor cortex; roughly electrode sites C_3, C_z and C_4) has been interpreted as a sensitive measure of the MNS. The *mu* rhythm is suppressed during both execution and observation of motor movements. Tognoli et al. (2007) found *mu* suppression over central rolandic sites (and occipital alpha rhythm suppression) during social interaction between two simultaneously-recorded people. *Mu* suppression indicating MNS activation did not depend on whether their interaction was coordinated with each

other (i.e., dependent or independent). In contrast, central parietal *phi* rhythms (9.2 to 11.5 Hz.) in the right hemisphere were strongly related to the degree of coordination or interdependence of their social interaction. If replicated, these findings represent an important advance in understanding the relationship between the frontal and parietal components of the MNS. This study would be difficult to replicate with fMRI or other whole-brain neuroimaging techniques because it requires imaging two different people at the same time.

Complementary Actions

Another important discovery (Newman-Norlund et al., 2007) is that only a subset (perhaps one third) of the MNS is activated in strictly mirror acts, while a larger brain area (two-thirds) of the MNS is activated in complementary actions. An example of complementary behavior is when one hands someone else a cup and then sees the other person receiving it. Complementary actions differ from the conventional motor acts associated with MNS activation in which the **same** motor behavior is executed and also observed in another individual. A small region in the right inferior frontal gyrus and large areas of the bilateral inferior parietal lobes were activated by complementary actions, but not by strictly mirror actions (Newman-Norlund et al., 2007). These results underscore the importance of self-other coordination in the functions of the MNS, rather then mere imitation.

The Human Self

Uddin et al. (2007) and Wheatley et al. (2007) argued that two different brain systems are primarily related to a person's representation of self: (1) the fronto-parietal MNS and (2) a midline cortical network for social cognition, consisting of superior temporal and medial prefrontal cortex (PFC), fusiform gyrus, posterior cingulate, insula, and amygdala. Note first that these systems taken together include much of the cortex, excluding most primary and secondary sensory areas and the frontal pole. These authors proposed that the two brain systems, rather than being separate and independent, cooperate closely in constructing self-representation and guiding social interaction (self-other relations). In this scheme, the MNS helps achieve understanding of others by internally simulating their behavior, emotions, and motives. For example, Wheatley et al. (2007) found that observing and imagining moving shapes engaged the MNS, but the midline social network of the brain was activated only when those same moving shapes were reinterpreted as animate (living, moving humans or animals). In terms of recognizing one's own face(Uddin et al., 2005) versus that of others (an important aspect of self-representation), self-face recognition activates right hemisphere MNS structures, while other-face recognition activates only the well-described "default/resting state" brain network, consisting in this case of medial PFC and precuneus.

Table 1. Proposed hierarchy of neuro-cognitive capacities.

Theory of Mind
social motives
relational self
self-recognition
self-other distinction

Functions of the MNS

Table 1 illustrates neurocognitive functions, listed with more basic capacities at the bottom, that are required to support Theory of Mind (Gallese, 2007; Gallese & Goldman, 1998; Schulte-Ruther et al., 2007), which is listed at the top of the table. Theory of Mind is defined as the awareness that other individuals also have minds, with consciousness, emotions, and intentions not unlike one's own. Theory of Mind is a capacity that is recruited to estimate the intentions of others and to interpret their behavior. Not only must the organism make a clear distinction between self and others, it must recognize itself (e.g., in a glass mirror) and view itself in relation to others. Associated with these capacities is the ability to make assessments and predictions concerning the motives of others' behavior. This "mind-reading" ability (Gallese, 2007; Gallese & Goldman, 1998; Schulte-Ruther et al., 2007) is a critical component of Theory of Mind with clear evolutionary advantages. Whether the 'other' is friend or foe, predator or prey, biological relative or stranger, the need to predict motivations reasonably accurately based on observable information and social context appears critical to survival and reproduction in a social world. It has been argued that these capacities are based on the human MNS (Agnew et al, 2007).

From a broader biological perspective, we note that most elements in life have multiple functions, whether they are genes, cells, organs, or behaviors. Even if these elements originally evolved in relation to a single function, the optimizing sieve of natural selection would tend to recruit them to "solve" a variety of problems whose functions may overlap, but are sometimes wholly different. This applies to the MNS as well in that it may be misleading to search for the **singular** function of this brain system, particularly in humans. A more inclusive strategy is to seek to discover the **multiple** functions of the MNS, and to study the interrelations, if they exist, among these.

Table 2. Functions associated with the human MNS.

Action	**Emotion**	**Motivation**
motor mimicry	facial synchronization	inferred intent
imitative learning	empathy & affective decoding	inferred social motivation
complementary action	predictive social relations	interactional social goals
predictive action outcome	social agency	**Theory of Mind**

Table 2 lists several proposed functions of the MNS in humans, organized in terms of motor action, affect, and intentions. In this table, just as the level of abstraction

(toward higher cognitive functions) increases from left to right and from top to bottom in relation to functioning, so too does the uncertainty of the interpretation of empirical results increase in these directions. Moreover, we propose that developmentally, change with maturation is in the same directions, from motor mimicry at very early ages to the acquisition of Theory of Mind at later ages. We add that so-called "higher" animals are more likely to have capacities toward the right and bottom of **Table 2** than are "lower" organisms.

The MNS and Motivation

Is Motivation an Emergent Property?

An emergent property is one that "emerges" from the interaction of two—or more often—many factors in ways that cannot be predicted from those same factors considered singly. For example, motivation may emerge in a complex, artificial neural network from many inputs, none of which would be considered "motivational" when entered into the network. The alternative would be a highly evolved and specific biological system whose primary or only function is motivation. The dopaminergic corticomesolimbic "reward system" (see below) in the mammalian brain is thought to be a reinforcement system for driving adaptive behavior. The "design" in this case is from evolutionary forces such as natural selection, although it appears to be a specific control system rather than an emergent property of diffusely distributed and undifferentiated neurons. These considerations become crucial when designing an artificial motivational system such as one based on the MNS.

Motivation and Goal-Directed Behavior

The science of human motivation is exceedingly complex, much of it beyond the scope of this discussion. In an effort to reduce the topic to its most fundamental aspects, we assert that the most basic—and universal—motivations are the Darwinian ones of survival and reproductive fitness (Newlin, 2002, 2007). Surprisingly, behavioral and cognitive psychology have overlooked these seemingly obvious motives for human behavior. Behavioral psychologists speak of reward, punishment, and reinforcement (whether positive or negative), but not survival or reproductive fitness. Cognitive psychology, which seems poorly suited to motivational theorizing, uses incentives (often monetary) and social motives (without reference to survival and reproductive fitness), or simple reinforcement, to provide engines for behavior. The result on a theoretical level is an organism that is like "a train without a locomotive."

Universality among individuals and between peoples of the world implies the existence of biological primitives that provide a genetic foundation for variable behavioral and cultural expressions. We emphasize in this discussion that the brain's slate is **not** blank, nor should the intellectual and motivational origins of behavior in artificial autonomous agents be empty at conception (or construction). Two obvious indications of this, although far from the only ones, are the universality of the MNS in neurologically-intact individuals, and the gross division of the cerebral hemispheres, specialized along the lateral dimension—the left versus right hemispheres—for linguistic and prosodic aspects of speech, respectively, and for different aspects of motivation and emotion (see below). Nonspecific notions of behavioral reinforcement simply do not explain specialization of function for individual brain regions, networks, or hemispheres.

Motivation and Hemispheric Asymmetry

Shallice (2004) proposed laterally asymmetric executive functions for the left and right PFC. Rather than the PFC controlling working memory, long an assumption of cognitive neuroscience, he relegated working memory functions to the parietotemporal cortex. He proposed instead that the left PFC exerts supervisory (top-down) control over lower-level systems of the brain, such as working memory and verbal communication. In contrast, the right PFC maintains control over errant mentation and behavior that does not accord with task goals.

Shallice's (2004) analysis, which is specific to the PFC, implies that MNS functions may differ between the right and left hemispheres. Specifically, the formal language functions of the left MNS, which is neuro-anatomically distinct from the right MNS, would be a higher-level system that is controlled by anterior regions of the PFC in the furtherance of current and future goals, i.e., goal-directed motivation. For example, the capacity to form internal representations of the verbal speech of one's self and that of others, particularly as they intersect (mirror or complement each other), would allow more anterior (PFC) supervisory regions of the brain to guide social communication toward goals consistent with enhanced survival and reproductive fitness. At the same time, these internal representations provide "grist for the mill" of the right frontal MNS to detect and correct deviations in communication away from the same goals. A further

Table 3. Summary of laterally asymmetric affective and motivational functions of the prefrontal cortex (PFC).

	LEFT Prefrontal (PFC)	**RIGHT Prefrontal (PFC)**
Cognition	expressive speech	visuospatial
Emotion	positive affect [plus anger]	negative affect [except anger]
Executive	top-down strategic modulation	checking on reaching task goals
Motivation	**approach**	**avoidance**
Temperament	undercontrolled	negative affectivity
Personality	sensation seeking novelty seeking	anxiety, depression social inhibition

example is that face recognition, whether it is one's own face or someone else's, is primarily a right frontal function that is part of the MNS (Uddin et al., 2005). It determines social context and whether communication is appropriate to that social situation. The evolutionary advantages of asymmetric supervisory control systems, as proposed by Shallice (2004) are unclear at the present time, but may be important in the design of autonomous agents.

Human Emotion and Hemispheric Asymmetry

Davidson and his colleagues (Davidson, 1999, 2003) proposed that the left and right human PFCs are laterally asymmetric in terms of "approach" (left PFC) versus "avoidance" motivation (right PFC), as well as in terms of affect. Davidson (1999, 2003) reviewed evidence that left PFC brain regions in humans are associated with "approach" orientation toward external stimuli, while contralateral right anterior structures are related to "avoidance" motivation away from noxious or threatening stimuli. The original conclusion that the left PFC was associated with positive affect and the right PFC with negative emotion gave way to the approach versus avoidance dimension based in large part on evidence that anger, a negative emotion but having an approach orientation, also activates the left PFC (Harmon-Jones, 2003). These associations are listed in **Table 3**.

Are the MNSs Laterally Asymmetric in Function?

Of course, the first issue is the role played by the MNS in receptive and expressive speech. The linguistic aspects of speech (as opposed to prosody) are strongly lateralized to the left hemisphere in most right-handed people. In fact, Rhesus monkey area F5, which was the region of the original discovery of mirror neurons, is homologous to Broca's expressive speech area in humans. There are functional parallels as well, particularly concerning language-related gestures and the motoric aspects of speech (Fogassi & Ferrari, 2007). Fogassi and Ferrari raised the question of whether in humans the MNS has been recruited (in an evolutionary sense) specifically for language processing. This issue is far from resolved (Aziz-Zadeh et al., 2006) but will engender intense scientific scrutiny because of the importance of speech to human functioning.

Motivation and Emotion

It is difficult to conceptualize emotion except in the context of motivation. The antecedents of affect are usually responses to changes in the status of goal-directed behavior. For example, we generally understand joy in terms of meeting a personal goal and anger as a frustrated response to thwarted goals. Affective responses can also lead to changes in motivation. Sadness, which tends to reduce personal motivation, usually stems from failure to meet social expectations, and fear leads to strong motivation, unless it is extreme.

Opponent Process Theory of Acquired Motivation

We have discussed "innate" motivation, particularly the Darwinian motivations to enhance one's survival fitness and to be reproductively fit (Newlin, 2002, 2007). Opponent process theory (Solomon, 1980) has had a profound influence on motivation theory, in part because it provides a compelling theoretical account of how new motivations are acquired in the life of the individual. In this model, intense affective stimuli produce an unconditioned response, the "a" process, which is relatively invariant with repeated elicitations. Opponent process theory also proposes there is a reflexive "b" process that is opposite in emotional valence to "a", that is recruited more slowly than the "a" process, grows stronger with repeated elicitations of "a", and has a slower offset. Solomon (1980) assumed that the "a" and "b" processes are additive and opposite in direction, so they tend to partially cancel each other when elicited in overlapping time. The result is that intense emotional responses are progressively and partially canceled by the "b" process because "a" and "b" are opposite in direction. Moreover, the more slowly decaying "b" process has full expression after the "a" process has abated because it is then unopposed.

For example, drinking a large quantity of alcohol can produce a euphoric response—intoxication—but also reflexively elicits an opponent "b" process, in this case dysphoria. The "b" process progressively cancels the "a" process and is fully manifest as a dysphoric reaction later in the same bout with the drug. Partial cancellation corresponds to tolerance to the drug, a hallmark of alcohol dependence. The dysphoric response is typical of the time period when alcohol levels in the blood and brain are descending and the drug effect is "wearing off." Progressive recruitment of the "b" process can lead in some individuals to alcohol dependence as dysphoric "b" processes may engender further drinking to cancel them by "a" processes.

What is Adaptive Motivation?

Opponent process theory represents a form of neurobiological adaptation. It is unfortunate in this case because it is thought to lead to drug addiction. In most cases, though, opponent processes produce neuroadaptation that is highly "adaptive" in an evolutionary sense because it attenuates strong responsivity to repetitive affective stimuli. Without partial cancellation of these intense responses, the result would be behavior that is disproportionate and destructive. In other words, excessive responding may impair biological fitness. For example, a fear reaction to a brief threatening stimulus might be initially adaptive because it motivates behavior to deal effectively with the threat, but would be highly maladaptive if it were unremitting or always elicited in the same magnitude to similar fearful stimuli presented many times.

Opponent Process Theory and the MNS

The implications of opponent process theory for the functions of the MNS have not been investigated directly. Since opponent process theory concerns intense affective responses and their compensatory counter-responses, we consider how emotional systems of the brain might be organized in ways that would conform to this theory. We noted above that motivational systems, at least of the PFC, are laterally asymmetric with left PFC associated with approach orientation and right PFC with avoidance of external stimuli. The driving system in terms of motivation is thought to be mesolimbic (chiefly ventral striatum, consisting of the ventral tegmental area, nucleus accumbens, and other regions of the extended amygdala). Dopamine is the primary neurotransmitter in this system, although it is modulated by a host of other neural signaling systems. Unfortunately, lateral asymmetry of function in this mesolimbic region has been grossly understudied. This is despite decades of intensive research on the functional roles of mesolimbic dopamine and other neurotransmitters in reward processes.

Functional Architecture of the Mesolimbic Motivation System

Given this paucity of empirical results concerning opponent process theory and specific neural systems, we begin with several plausible assumptions:

1) first, the corticomesolimbic motivational system is **not** a "reward" center, pathway, or system as the field has proposed. Newlin (2002, 2007) summarized evidence that it is instead a survival and reproductive fitness motivation system that, among other things, is also activated by novelty and aversive, noxious stimuli in addition to "rewarding" stimuli. The sensitivity to stressful stimulation is sharply inconsistent with viewing it as a "reward pathway." Darwinian motivations handily trump hedonics in the hierarchy of motives. If faced with the choice between boosting fitness or experiencing "pleasure," a human or animal will choose to enhance or preserve its chances for survival or reproductive fitness;
2) the ventral striatum exhibits laterally asymmetric specialization of function, consistent with its major neural connections and functional congruence with asymmetric PFC systems that control approach and avoidance motivation;
3) the left ventral striatum is associated with euphoria and approach motivation (including anger), and the right ventral striatum with aversive responding and avoidance motivation;
4) the left ventral striatum exploits opportunities to boost Darwinian fitness, such as a "rewarding" stimulus or a novel one that presents an ambiguous opportunity for "reward" or a threat to survival;
5) the right ventral striatum seeks to minimize threats to fitness; it mobilizes motivation to make the best of a bad situation. If the threatening stimulus is intense and uncontrollable or very prolonged, the system shifts into a more passive mode that is actually associated with depression of right ventral striatum activity;
6) top-down control from the left and right PFC on this survival and reproductive fitness motivation system is primarily inhibitory, and the right PFC has more inhibitory control than does the left PFC. Consistent with Shallice's (2004) proposed dual PFC control architecture (see above), inhibition of affect and behavior, and the regulation of emotion now appear to be primarily right PFC functions (Johnstone et al., 2007).

Ventral Striatum

Given these (arguable) assumptions, we can then ask how the left and right ventral striatum are related to each other. With a paucity of research concerning this question, several possibilities for dynamic relationships are evident:

a) independence—the two sides of the corticomesolimbic dopamine system operate without reference to each other;
b) dominance—one side of the ventral striatum typically inhibits the contralateral side;
c) reciprocal responsivity—the relationship between left and right is reciprocal. That is, the dominance of one side diminishes as that of the other side increases during changing circumstances;
d) mutual inhibition—similar in some ways to reciprocal responsivity, this model of the relationship between left and right mesolimbic activity assumes active tonic inhibition of right by left activity, and left by right.
e) mutual excitation—this is the converse of mutual inhibition. Activation of the left mesolimbic area excites the right and vice versa. This architecture will produce strong swings toward excitation or inhibition that are commensurate between left and right.

These alternatives are not necessarily mutually exclusive. As an arbitrary example, it is possible that the left side is dominant over the right for extended periods of time, but that the system switches to mutual inhibition at others. Flexibility such as this would underscore the importance of supervisory control systems, probably PFC or MNS, that determine the nature of these relationships in different contexts.

The same organizing principles (*a* through *e*, above) may apply to the left and right MNSs as well as the ventral striatum. This issue is fundamental to our understanding of the MNS, but has not been addressed in a way that is mathematically sophisticated. fMRI provides the means to test these alternative models of asymmetric functioning.

As noted above, the neurobiological underpinnings of opponent process theory have received little research attention. The possible exception to this is the rodent literature on drug abuse and addiction. However, despite

over two decades of intensive research on precisely this corticomesolimbic system—including powerful techniques such as microdialysis, voltammetry, and small-animal neuroimaging—the issue of lateral asymmetry of function in this deep midline system has seldom been studied. Although a number of studies have used more than one microdialysis probe or voltammetric sensor in the same primate or rodent brain, almost none have sampled dopamine or other neurotransmitter dynamics in parallel locations on the right and left sides of the ventral striatum. It simply has not been considered an interesting question! [Note the possible exception Nielson and his colleagues' (1999) work.] Yet we consider this probable functional asymmetry in the ventral striatum critical to understanding opponent processes in motivational and affective function. There is pervasive evidence of lateral specialization of function for virtually all other brain structures; it therefore seems unlikely that the two sides of the mesolimbic area would be functionally homogenous.

Specifically, we look to the dynamics (in the sense of change over time) of these deep midline structures, the ventral striatum and extended amygdala, to draw inferences about how opponent process theory might be actualized in the human and animal brain. Opponent processes likely determine how the MNS recruits these (asymmetric) deep structures to form internal representations of mirror and complementary agency, particularly its affective components.

We have suggested (above) several possible functional relationships. The two types of architecture that are most clearly consistent with opponent process theory are (a) reciprocal responsivity and (b) mutual inhibition. A straightforward experiment to examine these relationships is an fMRI neuroimaging study in which healthy humans are exposed to emotionally-laden stimuli. These stimuli would differ **not** in terms of positive or negative valence (the usual case in fMRI studies), but in relation to approach versus avoidance motivation (Davidson, 1999, 2003). We then identify regions of interest on the left and right sides of the mesolimbic region based on neuroanatomical co-registry. It is then possible to model mathematically the precise nature of these dynamic relations between activations in comparable left and right structures. The mathematical tools are readily available to test reciprocity and mutual inhibition. It is not enough to simply ask, "What lights up during approach stimuli versus avoidance stimuli?" This latter question would fail to address the **dynamic relationships** between left and right structures, although it would provide compelling evidence for differences in functional specialization between left and right mesolimbic areas. A comparable fMRI study of functional asymmetry of the MNSs would include approach and avoidance activity by both oneself and another agent. In fact, careful examination of the extant neuroimaging literature would likely yield several old datasets that could address these critical questions, and could be reanalyzed retrospectively to address issues of reciprocity and mutual inhibition.

A Working Model

A working model is one that is tentative, but can guide research. The value of a working model is determined by the extent to which the research questions or hypotheses derived from the model lead to new research that contributes meaningfully to the field of inquiry. Working models are modified as new data are obtained. Here we attempt to integrate some of what is known about several disparate lines of research: the MNS, motivation and emotion theory, theory of self, asymmetry of function between the cerebral hemispheres, and opponent processes.

This working model of the MNS concerns motivation as an essential component. We begin with the assumption that the MNS demonstrates significant asymmetry of hemispheric function in humans. This is based on strong evidence that the frontoparietal structures that have been identified as MNS (or complementary) are themselves functionally asymmetric. For both the left and right MNS, the anterior components have greater supervisory control, while the posterior parts of the MNS perform much of the cognitive work. We propose that the left MNS promotes social interaction on the basis of more formal, declarative aspects of linguistic speech and social rules—that is, it is largely prosocial. This is consistent with Davidson's (2003) conclusion that the left PFC supports approach motivation. It must be noted that anger is an emotion that also activates the left PFC. However, anger is clearly not an inhibitory affect, while other negative emotions (fear, sorrow, guilt, sadness, disgust) have generally inhibitory functions.

Conversely, our working model assigns to the right MNS supervisory control over social affect that is largely inhibitory, and prosodic aspects of speech, including nonverbal expressive and receptive communication. Specifically, the right PFC exerts a "braking" effect on disinhibitory behavior and mentation or affect that deviates from the individual's goals or from social norms. To perform these avoidance motivational functions, the right MNS must have the capacity to recognize and predict the motives and (generally adverse) consequences of deviant behavior or behavior that is inconsistent with personal goals.

Internal Representations of Agency

These MNS functions are constructive or reconstructive—the processing consists primarily of creating (or re-creating) internal **representations** of the actions, affect, and motives of others by referring to the corresponding internal representations of one's own behavior, emotion, and motivation. If the latter representations do not exist, then they are constructed. If they do exist, they are utilized to draw inferences about the other person, animal, or animate entity (e.g., a cartoon character).

These internal representations are **not** templates. Instead, they are **active creations** of the MNS, although re-creation is typically faster and easier than initial creation. For example, mammals develop (create) over time an internal model of the effects of gravity (Indovina et al., 2005) that is flexible enough to deal with a variety of different postures, body motions, etc.; this adaptability would be uncharacteristic of a template for gravity. If we observed another person violating the rules of gravity (such as in a movie or on the moon), then our internal re-creation of this representation would be noticeably disturbed.

Similarly, if we observe an individual obviously behaving in a self-defeating or socially inappropriate manner, our internal representation of their motives by our MNS will be violated. We react emotionally in these cases as if **we** were behaving in this manner, and we experience the fear of condemnation or reprisal that the other person may also (or perhaps should) experience. This emotional embodiment (Niedenthal, 2007) also includes adopting the facial expressions, tone of language, and posture of the individual who is observed. This empathic responding is squarely in the context of inferring motivational processes.

A Return to the Self

We now return to self-theory, in part because the human self provides an overarching organization for the processes under discussion. It may be useful to view the self as an **emergent property** of higher-level systems such as the MNS, the social network (Uddin et al., 2007; Wheatley et al., 2007), the cortico-mesolimbic motivation system, and the default/resting network. In this way, the self emerges from the interrelationships among these systems. This view may be replaced in the event that future research identifies "self neurons," or those comparable to mirror neurons but with uniquely self attributes.

Summary and Conclusions

We conclude this discussion with a set of "recommendations" for designing autonomous agents. These can be viewed also as possible experiments in design or starting points for constructing autonomous agents that exhibit motivated behavior. The goal may not be to replicate human motivation in silicon or some other engineering substrate, such as biological materials. It may be instead to create new "life" forms that have interesting and useful characteristics, but are unlike humans. In any case, the human brain (or mammalian brain) is a remarkable biochemical machine that provides a rich mine for modeling elements. We do argue, though, that motivation is a key aspect of artificial intelligence that may require consideration of MNS design characteristics to achieve animacy and agency:

1) There is now strong evidence that the MNS has functions that go far beyond motor mimicry. This underscores the importance and potential of the MNS as a design element for autonomous agents.
2) An appropriate but lofty goal of designing effective autonomous agents is for them to achieve Theory of Mind. There are gradations of neurocognitive skills or capacities (see **Tables 1 and 2**) that agents may need to achieve before Theory of Mind is possible. These subcapacities are important design elements in their own rights.
3) Of critical importance is that the MNS is **representational** rather than using neural templates, and it is **active** and **constructive** rather than passive and matching to sample.
4) A designer should strongly consider engineering parallel functionality. Gross neuroanatomy, the division of the two cerebral hemispheres, determines parallel systems in the human brain. Although the precise evolutionary advantages of dual architecture remain hazy, there can be little doubt that these advantages are real. In fact, research with artificial autonomous agents may help reveal how dual systems benefit design; this is a particularly difficult question for research with human beings.
5) In designing an artificial MNS (or two parallel MNSs), it is important that these system(s) further motivated, goal-directed behavior. The internal construction of representations, though potentially reflexive, appears critical to a person's understanding of the motives of others. It is equally important that the dual MNSs in humans may differ in terms of approach (left) versus avoidance (right) motivation. The natural inclination for us to dichotomize affect in terms of positive and negative does not match neurphysiological reality. Specifically, the angry autonomous agent might engage the same artificial MNS as when it is joyful or euphoric, but not that of the fearful or depressed agent. Therefore, motivational orientation organizes affective responses.
6) Research on both human (using fMRI) and artificial autonomous agents is needed to determine the nature or architecture of relationships between these parallel and functionally asymmetric systems. The answers to this question are likely to shed much light on the evolutionary advantages of having two functionally asymmetric MNSs, and motivational-affective systems.
7) For the time being, the designer may benefit from allowing the "self" to emerge from the interactions of higher-level subsystems, including the MNS(s), rather than attempting to design it.

References

Agnew, Z.K., Bhakoo, K.K. and Puri, B.K. 2007. The Human Mirror System: A Motor Resonance Theory of Mind-Reading. *Brain Research Reviews* 54(2): 286-293.

Aziz-Zadeh, L., Koski, L., Zaidel, E., Mazziotta, J., and Iacoboni, M. 2006. Lateralization of the Human Mirror

Neuron System. *The Journal of Neuroscience* 26(11): 2964-2970.

Davidson, R. J. 2003. Affective Neuroscience and Psychophysiology: Toward a Synthesis. *Psychophysiology* 40(5): 655-665.

Davidson, R. J. 1999. The Functional Neuroanatomy of Emotion and Affective Style. *Trends in Cognitive Sciences* 3: 11-21.

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., and Rizzolatti, G. 1992. Understanding Motor Events: A Neurophysiological Study. *Experimental Brain Research* 91: 176-180.

Dinstein, I., Hasson, U., Rubin, N., and Heeger, D.J. 2007. Brain Areas Selective for Both Observed and Executed Movements. *Journal of Neurophysiology* (in press).

Fogassi, L., and Ferrari, P. F. 2007. Mirror Neurons and the Evolution of Embodied Language. *Current Directions in Psychological Science* 16(3): 136-141.

Gallese, V. 2007. Before and Below 'Theory of Mind': Embodied Simulation and the Neural Correlates of Social Cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences* 362(1480): 659-669.

Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. 1996. Action Recognition in the Premotor Cortex. *Brain* 119: 593-609.

Gallese, V., and Goldman, A. 1998. Mirror Neurons and the Simulation Theory of Mind-Reading. *Trends in Cognitive Sciences* 2(12): 493-501.

Harmon-Jones, E. 2003. Early Career Award. Clarifying the Emotive Functions of Asymmetrical Frontal Cortical Activity. *Psychophysiology* 40(6): 838-848.

Indovina, I., Maffei, V., Gianfranco, B., Zago, M., Macaluso, E., and Lacquaniti, F. 2005. Representation of Visual Gravitational Motion in the Human Vestibular Cortex. *Science* 308: 416-419.

Johnstone, T., van Reekum, C.M., Urry, H. L., Kalin, N. H., and Davidson, R.J. 2007. Failure to Regulate: Counterproductive Recruitment of Top-Down Prefrontal-Subcortical Circuitry in Major Depression. *The Journal of Neuroscience* 27(33): 8877-8884.

Lyons, D.E., Santos, L.R., and Keil, F.C. 2006. Reflections of Other Minds: How Primate Social Cognition Can Inform the Function of Mirror Neurons. *Current Opinion in Neurobiology* 16: 230-234.

Newlin, D. B. 2002. The Self-Perceived Survival Ability and Reproductive Fitness (SPFit) Theory of Substance Use Disorders. *Addiction* 97: 427-446.

Newlin, D. B. 2006. Self-Perceived Survival and Reproductive Fitness (SPFit) Theory: Substance Use Disorders, Evolutionary Game Theory, and the Brain. In Platek, S., Keenan, J. P., and Shackelford, T. (Eds.) *Evolutionary Cognitive Neuroscience*, 285-326. Cambridge, Mass.: MIT Press.

Newman-Norland, R.D., van Schie, H.T., van Zuijlen, A. M.J., and Bekkering, H. 2007. The Mirror Neuron System Is More Active During Complementary Compared With Imitative Action. *Nature Neuroscience* 10: 817-818.

Niedenthal, P.M. 2007. Embodying Emotion. *Science* 316: 1002-1005.

Nielsen, D.M., Crosley, K.J., Keller, R.W., Glick, S.D., Carlson, J.N. 1999. Rotation, Locomotor Activity and Individual Differences in Voluntary Ethanol Consumption. *Brain Research* 27: 80-87.

Oberman, L.M., and Ramachandran, V.S. 2007. The Simulating Social Mind: The Role of the Mirror Neuron System and Simulation in the Social and Communicative Deficits of Autism Spectrum Disorders. *Psychological Bulletin* 133(2): 310-327.

Rizzolatti, G., and Craighero, L. 2004. The Mirror-Neuron System. *Annual Review of Neuroscience* 27: 169-192.

Schulte-Ruther, M., Markowitsch, H. J., Fink, G. R., and Piefke, M. 2007. Mirror Neuron and Theory of Mind Mechanisms Involved in Face-to-Face Interactions: A Functional Magnetic Resonance Imaging Approach to Empathy. *Journal of Cognitive Neuroscience* 19(8): 1354-1372.

Shallice, T. 2004. The Fractionation of Supervisory Control. In Gazzaniga, M. S. (Ed.) *The Cognitive Neurosciences III*, 943-956. Cambridge, Mass.: MIT Press.

Solomon, R. L. 1980. The Opponent-Process Theory of Acquired Motivation: The Costs of Pleasure and the Benefits of Pain. *American Psychologist* 35(8):691-712.

Tognoli, E., Lagarde, J., DeGuzman G. C., and Kelso, J. A. S. 2007. The Phi Complex as a Neuromarker of Human Social Coordination. *Proceedings of the National Academy of Sciences of the United States of America* 104(19): 8190-8195.

Uddin, L. Q., Iacoboni, M., Lange, C., and Keenan, J. P. 2007. The Self and Social Cognition: The Role of Cortical Midline Structures and Mirror Neurons. *TRENDS in Cognitive Sciences* 11(4): 153-157.

Uddin, L. Q., Kaplan, J. T., Molnar-Szakacs, I., Zaidel, E., and Iacoboni, M. 2005. Self-Face Recognition Activates a Frontoparietal "Mirror" Network in the Right Hemisphere: An Event-Related fMRI Study. *NeuroImage* 25: 926-935.

Wheatley, T., Milleville, S. C., and Martin, A. 2007. Understanding Animate Agents: Distinct Roles for the Social Network and Mirror System. *Psychological Science* 18(6): 469-474.

Acknowledgements: Kevin Strubler and Diana Fishbein provided helpful comments on a draft of this paper, and Alex Aversano gave valuable technical assistance. This work was supported by funding from RTI International and NIDA grant #5R21DA020592.