Title

What zombies can't do: A social cognitive neuroscience approach to the irreducibility of reflective consciousness

ISBN

9780199230167

Author

Lieberman, Matthew D

Publication Date

2009-01-29

DOI

10.1093/acprof:oso/9780199230167.003.0013

Chapter 13

What zombies can't do: A social cognitive neuroscience approach to the irreducibility of reflective consciousness

Matthew D. Lieberman

For the past 30 years, philosophers have debated whether zombies could exist (Kirk, 1974a). Far from being a detour into the world of horror films, this debate asks a serious question: Could an individual act and speak just like other individuals without having any internal conscious experience? Belief in the possibility of these so-called philosophical zombies serves as a litmus test for whether someone believes in some form of mind-body dualism or materialism. Here, I would like to focus on a related hypothesis that is emerging within psychology which I will refer to as the *psychological zombie hypothesis* (zombie will be used to distinguish the psychological variant for the remainder of the chapter). This hypothesis suggests that our behaviors and judgments are produced by an 'inner-zombie' whose mental work does not depend on conscious awareness and that those mental operations that are typically accompanied by conscious awareness do not rely on awareness to generate the operations and their outputs. In fact, this hypothesis suggests, mental operations that are typically accompanied by conscious awareness can be produced in the absence of conscious awareness, thus demonstrating the superfluousness of awareness.

At the outset, it is useful to distinguish between two additional terms, reflective and non-reflective, that can be applied to describe consciousness, awareness, and mental processes more generally (Lieberman et al., 2002). These terms will be addressed at length below (see Table 13.1). Generally, reflective awareness involves focusing attention on, considering, or manipulating that previously was part of the moment-to-moment stream of consciousness. If one is looking at a picture of a slightly sad face, one could explicitly think about whether the individual is actually sad and, if so, why. This explicit thought would be considered a reflective process. If on the other hand, one thought about which supermarket to stop at on the drive home from work, then the emotional aspects of the picture that are encoded would be the result of non-reflective processes. It is important to note that non-reflective processes include both those that involve phenomenal awareness such as this example (i.e. the individual saw the face, it just wasn't thought about) and others that do not involve phenomenal awareness (i.e. if the face was presented subliminally, outside of consciousness awareness).

Table 13.1 Characteristics of the X-system and C-system

	X-system	C-system
Phenomenological characteristics	Non-reflective consciousness Feels spontaneous or intuitive Outputs experienced as reality	Reflective consciousness Feels intentional and deliberative Outputs experienced as self-generated
Processing characteristics	Parallel processing Fast operating Slow learning Implicit learning of associations Pattern matching and pattern completion	Serial processing Slow operating Fast learning Explicit learning of rules Symbolic logic and propositional
Representational characteristics	Typically sensory Representation of symmetric relations Representation of common cases Representations are not tagged for time, place, ownership, identity	Typically linguistic Representation of asymmetric and conditional relations Representation of special cases (e.g. exceptions) Representation of abstract features that distinguish (e.g. negation, time, ownership, identity)
Evolutionary characteristics	Phylogenetically older Similar across species	Phylogenetically newer Different in primates or humans
Moderator effects	Sensitive to subliminal presentations Relation to behavior unaffected by cognitive load Facilitated by high arousal	Insensitive to subliminal presentations Relation to behavior altered by cognitive load Impaired by high arousal
Brain regions	Amygdala, ventral striatum, ventromedial PFC, dorsal ACC, lateral temporal cortex	Lateral PFC, medial PFC, latera PPC, medial PPC, rostral ACC, medial temporal lobe

Support for the psychological zombie hypothesis

To read the New York Times, one might be forgiven for believing the psychological zombie hypothesis should be re-termed the 'law of psychological zombies'. A recent story titled 'Who's minding the mind' (Carey, 2007) drew the conclusion that 'the subconscious brain is far more active, purposeful, and independent than previously known … The brain appears to use the very same neural circuits to execute an unconscious act as it does a conscious on.' This conclusion was probably read by more

than a million people and has significant implications for how we understand human behavior.

Only occasionally has the psychological zombie hypothesis been so explicitly posited and defended within the scientific community (Velmans, 1991). Nevertheless, in the past few decades, this hypothesis has been gaining steam as neuroscience and social cognition have both interjected themselves into this debate by shedding empirical light on the hypothesis. While neither field has claimed to have created full-blown stand alone zombies, both fields have produced what might be termed partial zombies, individuals with impairments in the ability or tendency to consciously reflect on some aspect of one's own experience. Neuroscience has examined individuals with particular forms of brain damage that render an individual unable to report in a particular domain what would ordinarily be experienced by others, whereas social cognition has used priming paradigms (e.g. subliminal exposure of words related to a concept) to activate mental representations without the individual's awareness, preventing the individual from engaging in any conscious reflective work on those specific representations. In either case, the individuals appear to be zombie with respect to some particular domain of cognition or some particular set of mental representations, at least temporarily. In both cases, the general conclusion has been that partial zombies can be made to perform just as if they were not zombies, affirming the zombie hypothesis and casting doubt on the relevance of reflective consciousness.

Blindsight is one of the neurological conditions described most often to support the zombie hypothesis (Velmans, 1991). Blindsight individuals have damage to visual cortical areas associated with conscious perception of the world, but the damage is limited to a region that corresponds to a particular spatial extent in their perception (Weiskrantz et al., 1974). In other words, there is a particular part of the visual field that does not give rise to conscious experience and if an object is placed in that part of the visual field, the patient will not report seeing the object. What many studies have shown, initially to establish the phenomenon and later to rule out artifactual explanations, is that blindsight patients can guess quite accurately which of two objects is in their blind spot despite a lack of conscious experience of the object or the ability to reflect on the identity of the object. Thus, in this small part of visual space, blindsight patients appear to be partial zombies in that they have no reflective awareness of what is in this part of space and yet they show a preserved ability to function like those who have the awareness by accurately identifying what is in the blindsight area of space.

Within social cognition, a different approach to the zombie hypothesis has been taken. As mentioned, social cognition has used priming techniques to activate mental representations of different types (goals, stereotypes, affect) to determine whether non-reflective activation of these representations produces similar consequences as when these same representations are activated explicitly in a way that allows for reflective conscious processing.

A study of memory encoding provides one of the cleanest instantiations of this approach. Chartrand and Bargh (1996) examined the effects of memory versus social encoding mindsets on memory for a written passage. This study was a replication of

an earlier study by Hamilton et al. (1980) which found that subjects instructed to form an impression of the individual in the passage ('social encoding mindset') without any memory instructions subsequently demonstrated better memory for the passage than subjects who were instructed to memorize the passage and were informed that there would be a memory test ('memory mindset'). Chartrand and Bargh's zombie-relevant twist was that half of their subjects were given the encoding mindset outside of conscious awareness through priming procedures. Despite the fact that these 'zombie' subjects had no idea that they were induced into either memory encoding or social encoding mindsets, they produced the same memory performance as subjects who were explicitly induced into one of these mindsets. Thus, even though these individuals were zombies with respect to the encoding goal and thus were prevented from reflecting on this goal and intentionally reading the passage in a way that facilitated the goal, they behaved just like non-zombies.

The blindsight and social cognition examples given here are just two of many studies that have produced similar results. These studies commonly prevent subjects from engaging in reflective processing of certain inputs to the system and demonstrate that the lack of reflective awareness does not lead to a change in behavioral performance. In essence, zombie (don't) see, zombie do.

Implications of zombie studies

Findings from the zombie studies are extremely exciting because they counter our naïve expectations about what the brain can do when the conscious mind is not overtly running the show. Indeed, much of the work in social and cognitive psychology over the past two decades clearly establishes an impressive array of computations that can be performed without conscious direction in the forms of implicit memory, implicit learning, and automatic behavior. This work changes our fundamental understanding of the unmonitored mind. Yet the implied subtext of zombie studies goes much further than this. Although many of the researchers conducting 'zombie success' studies are careful to provide circumscribed discussions of the implications of these findings, these studies together, without any study necessarily explicitly stating it, imply that the zombie hypothesis is correct and that perhaps reflective conscious processes have little or no pragmatic value.

These studies give the impression that anything that can be done reflectively can be done non-reflectively and promote the conclusion that reflective conscious processes are identical to non-reflective and non-conscious processes except for the addition of awareness. This statement warrants unpacking before proceeding further because two separate and important implications follow from this single statement. First, this statement implies a neutered reflective consciousness that is valuable only in that it provides our ticket into the theater where we watch the movie of our life. If the added awareness associated with reflective processes has no causal teeth, and it must not if the same input-output relations can be preserved in the absence of awareness, then reflective consciousness is merely epiphenomenal. It may appear to us that our conscious thoughts about our plans, goals, and behaviors have consequences, but a zombie could do the same without having the experience of these thoughts.

AQ - Nisbett and Wilson, 1977; please provide reference

Indeed, there is extensive evidence to suggest that the relation of our reflective conscious thoughts about our behavior is often poorly correlated with our behavior and the causes of our behavior (Nisbett and Wilson, 1977; Wegner, 2003).

The second implication focuses on non-reflective processes. The evidence supporting equivalent input-output relations in reflective and non-reflective processes has been used to suggest that non-reflective processes are actually reflective after all, capable of propositional and intentional processes operating on a host of symbolic representations of the self, others, ones attitudes, beliefs, goals, and motivations—only without the awareness component (Dijksterhuis and Nordgren, 2006; Velmans, 1991). Note that this is a different claim about non-reflective processes than the one made above about the impressive array of computations that non-reflective processes can support. Here, the claim is much more Freudian in the sense that there seems to be 'an intelligence' or a central voice in the non-conscious capable of explicit thought, only it is a voice that the reflective mind cannot hear.

Together these claims suggest that the reflective mind is less human than we naively believe and that the non-reflective mind is more human. It is interesting to note that the skills now imputed to the non-reflective mind in terms of symbolic processing and propositional logic are simultaneously derived from the putative skills of the reflective mind and yet also used to deny that the reflective mind is really doing those things.

Limitations of zombie studies

While the excitement grows about the power of the non-reflective mind to do anything and everything that the reflective mind does, because it is indeed the same mind minus some mechanism that supports epiphenomenal awareness, there has been very little questioning of whether the zombie studies really provide evidence in support of the zombie hypothesis. Although these studies are fascinating and important, they ultimately fall short of supporting the assumptions that are seeping into our collective understanding of the mind.

Consider the example of blindsight. When asked to discriminate between two potential stimuli in the blindsighted area, these individuals can perform the task successfully. However, there is no report of a blindsighted individual spontaneously identifying or using the information present in that field. These individuals haven't commented to a researcher, 'I know we're not in the middle of a testing session, but did you by chance just hide a $100 bill in my blind spot? I'm not having an experience of it, but I'm feeling compelled to grab for my blind spot as if there was a large denomination bill there.'

In fact, when food is placed in the blind spot of a food-deprived monkey, they make no attempt to reach for it or approach it in any way (Cowey and Weiskrantz, 1963). This finding is rarely noted and yet gives a more balanced picture of blindsight as it relates to the zombie hypothesis. These individuals can clearly do some forms of information processing based on information coming through the retina that does not reach visual cortex, but they just as clearly cannot process or use the information in other ways. The information cannot be spontaneously considered and used in light

of the individual's current goals and concerns. Thus, blindsight is as much of an indicator of what zombies can't do as what they can do.

More generally, these findings raise the question of whether zombie studies in general are selectively choosing independent and dependent variables such that zombies perform like non-zombies. To demonstrate that a third grader and an adult can both add 3 + 3 does not demonstrate that they have the same math abilities, or even that they use the same mechanisms when performing this particular calculation. To examine whether these individuals have different mathematical abilities, one would want to devise tests where they are likely to differ (e.g. geometry, algebra). Similarly, strong tests of the zombie hypothesis need to consider input-output relations that opponents of the hypothesis would predict zombies to be unable to perform.

I have previously written about two neurocognitive systems, the X-system and the C-system, that are hypothesized to be largely responsible for non-reflective and reflective social cognition, respectively (Lieberman, 2007a; Lieberman et al., 2002; Satpute and Lieberman, 2006). These systems will be described in detail (see Table 13.1), however, just noting a few of the distinguishing features of the two systems is instructive as far as devising clearer tests of the zombie hypothesis. For instance, one claim about the X-system is that it slowly extracts associative relationships present during perception and logical thought, whereas the C-system is capable of extracting these relationships in a single experience. Thus, the fact that elderly primes lead to slower walking behavior, as if the subject is enacting elderly behavior (Bargh et al., 1996), may be making use of associations that have been repeatedly presented over a lifetime. What if, instead, subjects were exposed to a novel group and information about the group's characteristics? For instance, if subjects learned that a newly discovered tribe, the Nochmani, in Indonesia tend to have a bouncier gate than Americans. Would priming 'Nochmani' lead to a bouncier step in Americans? It's hard to imagine that it would but the zombie hypothesis would need to predict this. To be clear, Bargh et al. have not suggested this would be the case and are not defending the zombie hypothesis. Nevertheless, the concern is how these results are interpreted more broadly within and beyond the scientific community.

Another claim is that the X-system handles affirmative associations (X and Y are associated) but does not explicitly represent non-associations (X and Y are unrelated). Thus, in the C-system 'not tense' and 'relaxed' might be interchangeable descriptors of an individual but in the X-system, 'not tense' cannot be parsed because 'not' is purely symbolic and has no 'associative/sensory referents'.

Deutsch et al. (2006) have provided strong support for the inability of 'not' and negations more generally to be processed non-reflectively. In one study, subjects had to either indicate the valence of a term (e.g. 'a party'; positive valence) or the valence of the negated term (e.g. 'no party'; negative valence). With extensive practice, both tasks became faster suggesting that non-reflective processes which are slow learning but fast acting were at work. Nevertheless, the negation task maintained a constant reaction time disadvantage relative to the basic task. In other words, there appears to have been automatization of several features common to both tasks, but the negation component showed no evidence of becoming automatized. Similarly, in a speeded evaluative priming task, primes had the same valenced priming effects whether shown

in normal ('a party') or negated ('no party') form suggesting that non-reflective cognition was unable to integrate negations into its computations.

Another claim is that the C-system, but not the X-system, is capable of propositional reasoning and the representation of asymmetrical relations (i.e. 'If A then B' does not imply 'If B than A'). Although priming of a goal or trait has been shown to affect performance in prime consistent ways (Dijksterhuis and Bargh, 2001), DeWall and colleagues (in press) have also shown that prime words related to logic and reasoning do not improve logic and reasoning at all, despite leading to semantic activation of these categories. For a broader discussion, see Evans (2008).

AQ - DeWall and colleagues (in press). Please check/correct; is this different from DeWall et al., detailed in references as 2007

Looking to the brain

Another approach to the zombie hypothesis is to examine the neural mechanisms that are at work when zombies and non-zombies perform a task. There are at least three hypotheses of what this data should look like if the zombie hypothesis is correct. The first is a bit of a straw man and suggests that the same brain mechanisms are activated in identical ways whether a task is performed reflectively or non-reflectively. Although a believer in Cartesian dualism might expect this, I doubt many others would expect this to be how the zombie hypothesis would correspond to brain function. Thus, the first formal hypothesis (*zombie neural hypothesis #1)* is that the same brain mechanisms are activated for both reflective and non-reflective task performance, however, reflective processing is associated with greater activity in these regions (see Figure 13.1a). The implicit assumption behind this claim is that awareness is a function of activation levels. If a neural process or mental representation is more activated then we are more likely to be aware of it because this is the source of awareness. Note that this hypothesis also implies that while greater activity might lead to performance changes because more neural work is being done, none of this additional work is due to the awareness associated with heightened neural activity. The second formal hypothesis (*zombie neural hypothesis* #2) states that additional brain regions may be recruited during reflective processing, compared with non-reflective processing, however, (a) only the brain activity associated with non-reflective processing will relate to performance outcomes regardless of whether the task is performed reflectively or non-reflectively and (b) brain regions associated with non-reflective processing will be at least as active during reflective processing as during non-reflective processing (see Figure 13.1b). This account suggests that awareness may be associated with different brain regions than those involved in non-reflective processing, but that only those involved in non-reflective processing are doing mental work with causal consequences and thus these processes and the activity in the brain regions supporting them should be present regardless of whether reflective awareness is also present.

Data that is inconsistent with these formal hypotheses would provide challenges for them. In addition, we can consider a formal hypothesis that opposes the zombie claim (*anti-zombie neural hypothesis*; see Figure 13.1c). This hypothesis differs from zombie neural hypothesis #2 in two ways. First, this hypothesis claims that activity in brain regions supporting reflective and non-reflective processes are each independently

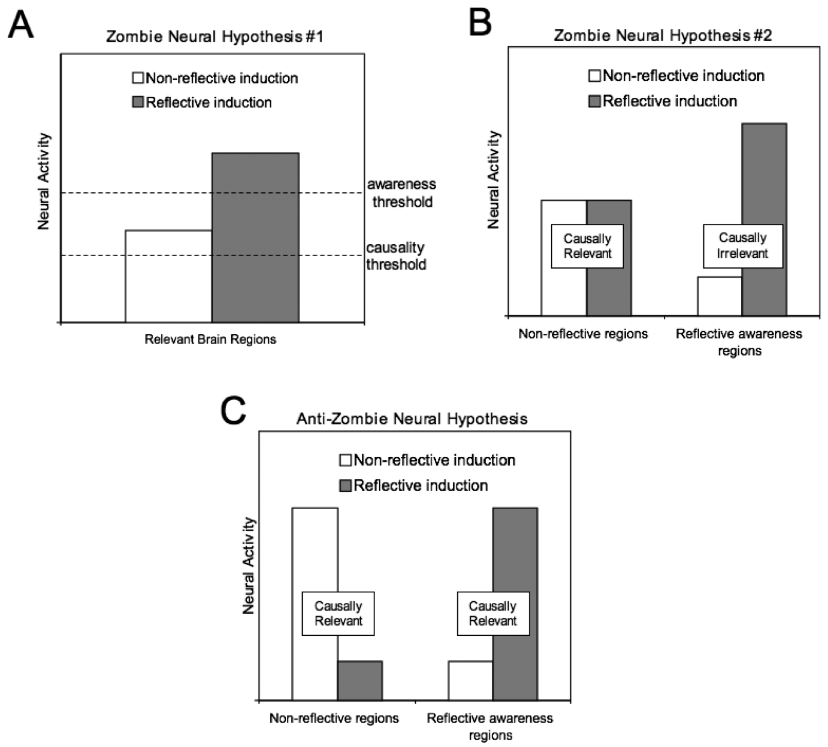

Fig. 13.1 Hypothetical brain activations associated with the zombie hypothesis. Panels A and B depict two possible zombie neural hypotheses. Panel C depicts a neural pattern that would conflict with the zombie hypothesis.

related to performance outcomes under appropriate conditions. Second, this hypothesis claims that under conditions that promote reflective processing, activity in brain regions that support non-reflective processing may be diminished relative to conditions that promote non-reflective processing.

The remainder of this chapter is focused on data, largely from my lab, that bears on the (anti-) zombie neural hypotheses. This work examines how reflective and non-reflective social cognition is instantiated in the brain. In order to transition to this work a bit of exposition is necessary. First, I will describe in more detail the putative differences between reflective and non-reflective social cognition. Second, I will describe the neural model my colleagues and I have developed to examine reflective and non-reflective social cognition in the brain. I will then discuss findings on implicit learning, self-knowledge, emotional processing, social well-being, and attitude change.

Reflective versus non-reflective social cognition

The first step in determining how reflective and non-reflective social cognition are implemented in the brain requires us to declare the characteristics of each system in

terms of operating characteristics, mental representation, phenomenology, and the moderators that facilitate or interfere with each system. Our account (Lieberman et al., 2002; Lieberman, 2007a, b; Satpute and Lieberman, 2006) is greatly influenced by dual-process models (for review see, Chaiken and Trope, 1999) and dual-system models (McClelland et al., 1995; Metcalfe and Mischel, 1999; Sloman, 1996; Smith and DeCoster, 2000). In each of these accounts, one system or process is thought to act quickly, potentially without intention, effort, or awareness, based on associations formed incrementally over numerous experiences. Thus, this system can be thought of as a slow learning, fast acting system. In contrast, the second system is thought to act more slowly, typically with intention, effort, and awareness, based on the application of symbolically represented rules that can be learned in a single moment of experience. This system would then be thought of as a fast learning, slow acting system.

AQ –Smith and DeCoster, 2000. Please check/correct year; detailed as 1998 in references

Our characterization of the two systems (see Table 13.1) takes these distinguishing characteristics as a starting point and considers a number of additional distinctions between the systems. First, there is a greater focus on the phenomenological aspects of processes emanating from the two systems. Non-reflective processes feel like reality. When we observe one person shoving another, this is experienced as an aggressive act, with the aggressiveness experienced as objectively out there in the world, even though the aggressiveness is constructed psychologically based on a number of characteristics separate from the act itself (Kunda and Sherman-Williams, 1993). That is, the aggressiveness is 'seen' as out there in the world, and is not felt as constructed in any part by our minds. In contrast, we typically feel a sense of ownership, construction, and responsibility for our reflective processes. If we see an aggressive behavior and then enumerate the reasons why this behavior may have been morally justified, we believe that this enumeration is our own specific contribution to our understanding of the behavior and the person. We know that this contribution has come from inside ourselves and we are open to the idea that this may not represent reality but our own internal processing.

Although this characterization of the processing characteristics (i.e. how representations are processed) largely conforms to other dual processing accounts, there are also a number of representational features (i.e. what is represented) that are hypothesized to differ in the systems. The non-reflective system tends to trade in more sensory cues (e.g. images, sounds) and associations that are closely wed to these concrete sensations. In contrast the reflective system is most efficient in dealing with linguistic representations. Sensory and linguistic representations are not exclusively tied to one system or the other, but rather each system appears to be optimized to deal better with one than the other and in any combination. That is, the reflective system is optimized to deal with the overall meaning of multi-word phrases and statements but is less capable of dealing with multiple sensory cues that vary along subtle or complex gradations. In contrast, the non-reflective system is optimized in the opposite way (Deutsch et al., 2006; Greenwald and Liu, 1985).

The reflective system is also able to tag various symbolic aspects of a represented entity that distinguish it from other entities in the same class or category, whereas the non-reflective system largely represents commonalities across members of a class or category. The reflective system can represent conditional aspects of an entity

(e.g. A has feature B in situation 1 but not situation 2), distinguishing non-present features of an entity (e.g. 'Although he is a professor, unlike most professors he did not graduate from high school'); temporary information (e.g. 'Today my car is parked in a different spot than it usually is'); asymmetrical relationships (i.e. A causes B, but B does not cause A); ownership relations (e.g. 'Jim owns that car'); and explicit representation of identity relations ('That car is a Volvo').

Finally, there are a number of moderators that are thought to affect which system is likely to handle the lion's share of processing at a particular moment. On the one hand, physiological arousal, cognitive load, and subliminal cue presentations will enhance the dominance of non-reflective processing (if non-reflective processes can represent the relevant information), whereas the use of purely symbolic cues and propositional information will enhance the dominance of reflective processing.

X-system and C-system

The aforementioned descriptions of reflective and non-reflective processing are hypothetical models of two systems. These descriptions are not, in themselves, evidence for the existence of the two systems. Instead, these criteria were used to initially identify candidate brain regions for the two systems. The majority of studies that were used to select the candidate brain regions were minimally social or non-social in nature. Although the brain regions are mostly assigned in the same fashion that they were originally (Lieberman et al., 2002), there have been changes along the way in light of numerous studies that have been reported in the field since the initial assignments (see Lieberman, 2007a,b; Satpute and Lieberman, 2006).

AQ - Lieberman et al., 2007. Please provide reference

The X-system is hypothesized to handle non-reflective social cognition. The brain regions that are associated with the X-system are the amygdala, basal ganglia (including ventral striatum), ventromedial prefrontal cortex (ventromedial PFC), dorsal anterior cingulate cortex (ACC), and lateral temporal cortex (including superior temporal sulcus, inferotemporal cortex, and the temporal pole) (see Figure 13.2). The C-system is hypothesized to process reflective social cognition and the brain regions associated with the C-system are lateral PFC (both dorsolateral and ventrolateral PFC), medial PFC (including dorsomedial PFC), lateral posterior parietal cortex (lateral PPC), medial PPC, rostral ACC, and medial temporal lobe (including the hippocampus and surrounding structures). The particular contributions of these brain regions have been discussed extensively elsewhere (Lieberman, 2007a; Satpute and Lieberman, 2006). Here, it is only critical that these regions are separated into two sets.

Given this hypothetical division, we can then examine whether studies manipulate variables that alter the tendency for one system or the other to dominate processing produce neural responses that conform to one of the zombie neural hypotheses or instead fit the anti-zombie pattern. Zombie neural hypothesis #1 (Figure 13.1a) predicts that the X- and C-systems will not be differentially involved in task performance, but rather a single set of relevant brain regions will be activated under both reflective and non-reflective task induction. In both cases, the activity in these brain regions would surpass 'causal threshold' allowing the neural activity to appropriately transform inputs into outputs, whereas activity in the same brain regions would typically surpass a higher 'awareness threshold' only during reflective task inductions.

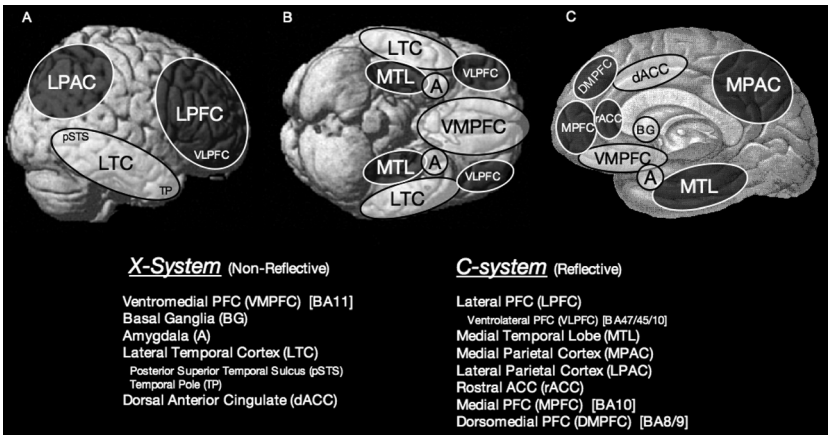

Fig. 13.2 The X-system and C-system responsible for non-reflective and reflective social cognition, respectively. X-system regions are shown in white ovals with black borders. C-system regions are shown in grey ovals with white borders. Panels A–C present lateral, ventral, and mid-sagittal views, respectively.

Zombie neural hypothesis #2 (Figure 13.1b) allows for separate neural networks linked to the X- and C-system. Here regions in the X-system would be expected to be similarly active under reflective and non-reflective task inductions and to be the driving causal force in both cases. C-system activity would be expected to be much greater under reflective than non-reflective task inductions and associated with reflective awareness but not causally relevant to transforming inputs into outputs. The *anti-zombie neural hypothesis* (see Figure 13.1c) differs from hypothesis #2 in two critical ways. First, both X-system and C-system activity can be related to task performance. Second, in at least some cases, X-system activity is reduced during reflective task inductions relative to non-reflective task inductions. This latter point would imply that the brain regions supporting performances under non-reflective task inductions are not still supporting performances under reflective task inductions, even if the same input-output relations are found. Stated differently, this hypothesis, if supported, would indicate that reflective processing would generate performances through different neural mechanisms than those that support non-reflective processes.

Evidence against the zombie hypothesis

Implicit versus explicit learning

The first domain under consideration is implicit and explicit learning. This comparison involves non-reflective and reflective cognition, respectively, but not social cognition per se. However, the results are relevant enough to the zombie hypothesis to deserve inclusion and the distinction has clear implications for social cognition (Lieberman, 2000). Numerous behavioral studies have shown that individuals can

learn cue sequences that contain task relevant information without having awareness that they are doing so (Reber, 1967). Indeed, amnesic patients who are generally unable to remember previous episodes of experience show normal learning on these tasks (Knowlton et al., 1996). In a commonly used implicit learning task, the artificial grammar task, individuals are presented with 'words' from an invented language. The construction of the words follow a set of rules defined in a Markovian grammar chain, but the rules are usually impossible to learn explicitly. After exposure to several examples from the language, individuals are able to successfully guess whether new letter strings constitute legal words in this new language or not. Knowlton and Squire (1996) designed a variant of this task that assesses implicit learning of the grammar as well as explicit learning of simple letter dyads and triads within the words which could be used to make judgments, albeit incorrectly. For instance, a subject might explicitly realize they saw the letter pair 'TV' in some of the training trials and then assume any new string with TV in it is a legal string. This cue may not actually predict legal strings, but it is explicit learning that the subject can reflectively perform and use.

We conducted an fMRI study (Lieberman et al., 2004) in which we examined the neural responses that were more activated on test trials (e.g. 'JTVP') for which explicitly learned cues could be applied (e.g. the test string contains 'TV') but no implicit sequence cues were present relative to other trials in which the implicit sequence cues were present but no explicitly learned cues were present (e.g. 'JVPT'). In other words, the first kind of trial could only be responded to on the basis of reflective knowledge acquired during training and the second kind of trial could only be responded to on the basis of non-reflective knowledge.

In the implicit learning condition, we observed greater basal ganglia activity to valid 'words' than to invalid 'words', when the explicit cues were absent. This X-system activation is consistent with previous motor implicit learning studies (Grafton et al., 1995) which have also implicated the basal ganglia in implicit learning. No C-system activity was observed in this analysis. In contrast, when explicit cues were present, but implicit cues were absent, hippocampal and medial temporal lobe activity was present, both C-system regions, however, no X-system activity was present. Additionally, like a pattern observed in a number of other studies (Packard et al., 1989; Poldrack et al., 2001), we found that activity in the basal ganglia and medial temporal lobes were inversely correlated with one another such that greater activity in one was associated with diminished activity in the other. Thus, while the zombie hypothesis argues that basal ganglia activity, associated here with non-reflective processing, should be equally or more active during reflective processing conditions was not observed at all during reflective processing and was less active to the extent C-system activity was present.

Foerde et al. (2006) provide even more compelling evidence. Subjects trained on an explicit learning task either freely, allowing for reflective processing, or under cognitive load, which impairs reflective cognition but allows non-reflective cognition to proceed. Task performance at test was correlated with brain activity during the training phase. Task performance at test was not significantly different, however, the brain activity during training that predicted test performance differed by training condition.

When training occurred without cognitive load, test performance correlated with medial temporal load activity during training. When training occurred with cognitive load, test performance instead correlated with basal ganglia activity during training. Importantly, basal ganglia activity in the no load training did not correlate with test performance. This finding directly conflicts with the zombie hypothesis in any of its forms as the region associated with non-reflective performances was not predictive of performances under reflective task conditions.

Reflective versus non-reflective self-knowledge

A number of fMRI studies have examined the neural correlates of self-knowledge (for review, see Lieberman, 2007a). Typically medial PFC, in the C-system, is more active when individuals are reflecting on their own personal characteristics than when they are judging the characteristics of another person (Kelley et al., 2002). On its face, this appears to be a reflective process and thus it is not surprising that a C-system structure would be involved. Nevertheless, not all self-knowledge judgments require reflective processing. When Tiger Woods is asked if he is athletic, his response is likely to be automatic and non-reflective. The key question with respect to the zombie hypothesis is whether the same brain regions are active when making self-judgments that do or do not require reflection.

Social psychologists have examined this distinction in the context of self-schemas. In self-schematic domains, like golf or athletics would be for Tiger Woods, a person has a great deal of prior experience and can make judgments about themselves in that domain very quickly (Markus, 1977). This characterization is consistent with the non-reflective system as being slow learning, but fast acting. We conducted an fMRI study (Lieberman et al., 2004) to examine the neural correlates of self-knowledge in schematic domains and non-schematic domains to probe the non-reflective and reflective self-knowledge, respectively. Individuals were selected to participate in the study either because they were lifelong competitive athletes (i.e. played on the UCLA soccer team) or were long-time actors (i.e. working in Los Angeles), but not both. Additionally, individuals were only classified as schematic if they demonstrated a substantial reaction time advantage for self-judgments in the high experience domain rather than the low experience domain. We then examined which brain regions were differentially activated in the self-schematic and non-schematic domains.

Recall, that in both zombie neural hypotheses, the brain regions invoked during non-reflective processes should be active as much or more during reflective processes, whereas the anti-zombie neural hypothesis predicts that in at least some cases, brain regions that are activated during non-reflective processing will become less activated during reflective processing (because other mechanisms are really at work during reflective processing).

Consistent with all hypotheses, a wide-array of X-system regions were active during self-judgments in the schematic domain, including ventromedial PFC, ventral striatum in the basal ganglia, amygdala and lateral temporal cortex. However, all of these regions were less active during self-judgments in the non-schematic domain and instead, activations in the hippocampus and dorsomedial PFC were present,

both C-system regions. This result is inconsistent with either of the zombie neural hypotheses and this study has been largely replicated (Rameson and Lieberman, 2007).

We attacked the same question from a different angle by comparing general self-knowledge judgments in children and adults (Pfeifer et al., 2007). Our assumption is that in general, adult self-judgments are less likely to require reflection than the self-judgments of children. In this fMRI study, children (average age = 10.2) and adults (average age = 26.1) reported whether short phrases (e.g. 'I like reading') described themselves or Harry Potter. Harry Potter was chosen as the target of social cognition because both children and adults are familiar with the character and his personality. Both children and adults produced greater activity in medial PFC when making self-judgments than social judgments, however, children produced significantly greater activity in this region than adults consistent with our hypothesis that this task requires more reflective processing for children than adults. Additionally, lateral temporal cortex was significantly active in adults, but not in children. Thus, this X-system region which was more activated for adults in schematic than non-schematic domains (Lieberman et al., 2004) was less active in children for whom self-knowledge judgments are thought to be more reflective. In other words, a region that supports self-knowledge processes in adults to the extent that the process is non-reflective, is not present in children. This, combined with the greater medial PFC activity in children than adults suggests that there may be different mechanisms responsible for self-knowledge judgments to the extent that the judgments invoke reflective or non-reflective processes. As with the self-schema study, these results are inconsistent with either zombie neural hypothesis.

Reflective and non-reflective emotional processing

One of the clearest pieces of evidence we have regarding the zombie hypothesis comes from research on affect labeling (Hariri et al., 2000; Lieberman et al., 2007; Lieberman et al., 2005). In these studies, the emotional aspects of emotionally evocative pictures are either processes reflectively or non-reflectively. The zombie hypothesis would argue that as long as the emotional information gets into the brain, the same brain regions active during non-reflective emotional processing should be at least as active during reflective emotional processing. However, these studies show that merely switching from non-reflective to reflective modes of emotional processing leads to reductions in the neural, physiological, and subjective responses associated with non-reflective emotional processing (for review, see Lieberman, 2007b).

In one of these studies (Lieberman et al., 2007), participants were shown negative emotional faces. In the reflective emotional processing condition, participants had to choose from two emotion words presented on the screen with the face which word described the emotion in the face. In the non-reflective emotional processing condition, participants had to choose which of two names was gender-appropriate for the face. In the non-reflective condition, the emotional stimulus is still present but the emotional aspect is incidental to the task and thus not likely to be reflected upon.

During non-reflective emotional processing, there was a significant response in the amygdala, an X-system region, similar to that seen in previous studies when

emotional images are presented subliminally (Whalen et al., 1998) and thus could not have been processed reflectively. However, during reflective emotional processing, the amygdala response was significantly diminished relative to the non-reflective condition. Indeed, all X-system regions (amygdala, VMPFC, ventral striatum in the basal ganglia, dACC, and LTC) were less active during reflective emotional processing than non-reflective emotional processing. In contrast, only a single region of the brain was more active during reflective emotional processing, right ventrolateral PFC, a C-system region. Moreover there was an inverse relationship between the activity in right ventrolateral PFC and the amygdala such that subjects who activated this prefrontal region more during reflective emotional processing also tended to show less activity in the amygdala during reflective processing.

These results are thus inconsistent with the zombie neural hypotheses on two separate accounts. First, all of the X-system regions activated during non-reflective emotional processing showed reduced activation during reflective processing. Second, as in the implicit learning studies above, regions associated with reflective processing appear to be competing with and dampening down activity in regions associated with non-reflective processing. In other words, reflective processing activations appear to occur at the expense of non-reflective processes.

Reflective and non-reflective aspects of social well-being

Kahneman and colleagues (Fredrickson and Kahneman, 1993; Kahneman et al., 1993; Redelmeier and Kahneman, 1996) have demonstrated in a number of contexts that well-being or distress that an individual reports during the individual moments of an experience ('momentary well-being') often do not correspond in expected ways with an individual's memory of aggregate well-being throughout the entire episode ('retrospective well-being'). One possible reason for this is that the two kinds of self-reports may rely on different processes. Momentary well-being may be the result of a fast, inuitive, non-reflective judgment whereas retrospective well-being may depend on more reflective processes both at retrieval and encoding. That is, retrospection on one's prior well-being integrated over several moments may be a reflective process, but the ability to retrieve those 'to-be-aggregated' moments may depend on those moments having been the subject of reflective processing when they occurred, as this would promote deeper encoding and thus better retrieval (Craik and Tulving, 1975).

We examined this dual-process account of momentary and retrospective well-being in a study (Eisenberger et al., 2007) that combined an fMRI assessment of neural responses to social rejection and an experience sampling study. In the fMRI session, subjects played a virtual game of catch with what they believed were two other players also in MRI scanners. In reality, the two other players were computer players controlled experimentally and programmed to stop throwing the ball to the subject at a set time in order to make the subject feel socially rejected. We had previously observed (Eisenberger et al., 2003) that to the extent that subjects felt 'social pain' during the moment of rejection that they had greater activity in the dorsal ACC which has been associated with physical pain distress in numerous studies (Peyron et al., 2000; Rainville et al., 1997).

AQ - Peyron et al., 2000. Please check/correct; detailed as 1999 in references

At a separate point in time, the same subjects participated in a 10-day experience sampling study. During this time, subjects carried Palm Pilots. Subjects were beeped several times a day and asked to rate their social well-being during their most recent social interaction. At the end of the day, subjects also made retrospective reports of social well-being aggregated over the entire day. Our logic for the study was that individual differences in the neural responses during social rejection in the fMRI session would serve as a proxy for individual differences in neural responses during the individual episodes of experience that subjects were asked to comment on during the experience sampling study. For instance, we expected that an individual who produced high levels of dorsal ACC activity to rejection in the scanner would also tend to report less social well-being during social interactions in their everyday life.

X-system regions dorsal ACC, amygdala, and basal ganglia had activity during rejection in the fMRI session that each predicted the degree of social distress reported when rating their momentary experiences during the day in the experience sampling study. Thus, in a relatively non-reflective task, a host of X-system regions seen in other studies of non-reflective social cognition were again involved. These regions were not correlated with the degree of social distress reported when subjected rated their aggregate experience over the course of the day. Instead, social distress in the aggregate self-reports was associated with activity in the hippocampus and medial PFC, both C-system regions involved in autobiographical memory. Its important to keep in mind that our fMRI data in study came during an episode of rejection, not during attempts to recall episodes of rejection so the aggregate analyses are probably more indicative of reflective processes present during the rejection episode itself that produce deeper encoding and thus render the episodic events more retrievable at the time of aggregate judgments.

Once again, we see the brain activations present during non-reflective social cognition, absent during similar judgments involving reflective processing. This is not a process pure manipulation and thus other accounts could be given of the data, but certainly, these data are not consistent with the zombie neural hypotheses. It is also notable that these activations were not associated trying to make judgments of one kind or another, but were associated with the observable outcomes of those judgments. An individual might report, without much reflection, at four separate times during the day that his last social interaction caused him social distress and then at the end of the day, reflect on those different episodes and report that he experienced a great deal of social distress during the day. From the observed self-report alone, one could not be blamed for thinking the same process is at work in both cases, however, the fMRI data suggests that the reflective and non-reflective conclusions of low social well-being were reached through quite different mechanisms.

Reflective and non-reflective attitude change

For half a century, social psychologists have studied post-decisional attitude change in terms of cognitive dissonance reduction (Festinger, 1957). The basic finding is that when people make decisions that conflict with existing attitudes, the attitudes tend to change to fit with the decision giving the appearance of rationalization to outside observers. In one common paradigm (Brehm, 1956), participants rank their

preferences for each member in a set of items (e.g. a set of appliances or a set of art prints). The experimenter then selects two closely ranked items and asks the subject to chose which of these to take home as part of their payment for the experiment. People often, but not always, choose the item that was originally ranked slightly higher. Rationally, people should be thinking to themselves, 'I liked X slightly more than Y, so that's why I chose X over Y. It was a tough choice because the items were similarly matched, but I like X slightly more'. However, when people are asked to re-rank all of the items after the decision has been made, the chosen item goes up in the rankings, and the rejected item goes down. This sudden change in preferences makes what might have been a tough choice in the moment seem like an obvious choice in hindsight.

While this attitude change clearly looks like rationalization from the outside ('Just yesterday, John couldn't decide which job to take because they were so evenly matched, but today he's acting as if he always though the job he took was the better job'), its unclear how it is experienced from the inside. When this attitude change is occurring, are people reflectively aware that they are shifting their attitudes or is some non-reflective process at work? Of course, the zombie hypothesis suggests that it does not matter and we will return to that shortly.

Most of the classic theories of cognitive dissonance suggest that this form of attitude change is the result of reflective attitude change processes (see Gawronski and Strack, 2004). These theories suggest that attitude change results from becoming reflectively aware of the conflict between one's prior attitudes and the decision that is at odds with them and then doing reflective cognitive work to change the attitudes to fit with the decision (for review of these theories with respect to reflective processing, see Lieberman et al., 2001). However, some studies suggest that reflective processing may not be necessary for post-decisional attitude change. Bem and McConnell (1970) found that after subjects' attitudes changed, they had no memory for ever having held different attitudes at the beginning of the study. This suggests that the attitude change took place without reflective processing of the change process. Similarly, Lieberman et al. (2001) found that both amnesic patients who cannot form new memories and individuals under cognitive load showed as much attitude change in a dissonance paradigm as normal control subjects. Thus, amnesic patients who do not remember ranking the items before making a decision and do not remember making a decision that conflicts with their prior attitudes still produced normal levels of post-decisional attitude change suggesting that at least in some cases, reflective processing is not involved in the attitude change process.

We recently examined reflective and non-reflective aspects of post-decisional attitude change in an fMRI study (Jarcho et al., 2007). In the study, we scanned participants while they made several decisions about different pairs of art prints. Each pair of prints had been previously rated by the subject to be roughly evenly matched, thus rendering the choices somewhat in conflict with existing attitudes. After getting out of the scanner, subjects rated all of the prints again so that we could assess attitude change that occurred as a function of their choices. Subjects then saw each art print a final time and were asked to remember what initial attitude towards the art print had been at the beginning of the study.

We looked at the neural response to just those trials for which attitude change occurred and further subdivided these trials into those for which the subjects remembered that they had previously held a different attitude and those for which the subjects reported that they had always had the same attitude towards the art print that they now hold. Our thinking was that trials for which subjects were aware of the attitude change that took place would be associated with more reflective processing than trials for which they were unaware. The unaware attitude change trials were associated with increased activity in the dorsal ACC and the amygdala, two X-system regions. Repeating the pattern seen through the studies in this chapter, these regions were not active on attitude change trials for which subjects were aware that attitude change had taken place. On the aware trials, which presumably were associated with more reflective cognition during the attitude change process, there was increased activity in several C-system regions including rostral ACC, lateral PFC, and the medial temporal lobe.

One particularly nice feature of this study with respect to the zombie hypothesis is that both the aware and unaware trials produce similar levels of attitude change on average. Thus, whether subjects were reflectively aware that they were changing their attitudes or were zombies with respect to the attitude change, the behavioral effects were similar, but the neural systems supporting the two forms of attitude change are largely distinct from one another.

Other evidence

The preceding sections have provided neurocognitive evidence that I have amassed that relates to the zombie hypothesis and to dual systems of social cognition more generally. Of course, the lion's share of data that bears on this hypothesis comes from other labs. Many of the other studies were not designed with these objectives in mind but because of the methods applied, they either promoted reflective or non-reflective social cognition of some type. This data has been recently reviewed (Lieberman, 2007a) and across 21 domains of social cognition, the activations largely conform to the hypothesized distinction between the X-system and the C-system. Figure 13.3 displays a summary of this review. The large gray ovals with which borders represent C-system regions and the large white ovals with black borders represent X-system regions. The small circles represent domains of social cognition with the circles places in regions only if these activations appeared in multiple studies from that domain. The small black circles represent tasks that induce reflective processing and the small white circles represent tasks that induce non-reflective processing. Of 53 regions of activity associated with reflective and non-reflective processes in 21 domains of social cognition, 49 of the 53 separated into the X- and C-systems as hypothesized.

Evaluating the zombie hypothesis

Conceptually, the zombie hypothesis suggests that reflective cognition is only different from non-reflective cognition in epiphenomenal ways: it might feel different, but the engine under the hood is doing the same things in either case, at least insofar as observable outcomes are concerned. In the terms of this chapter, it suggests that

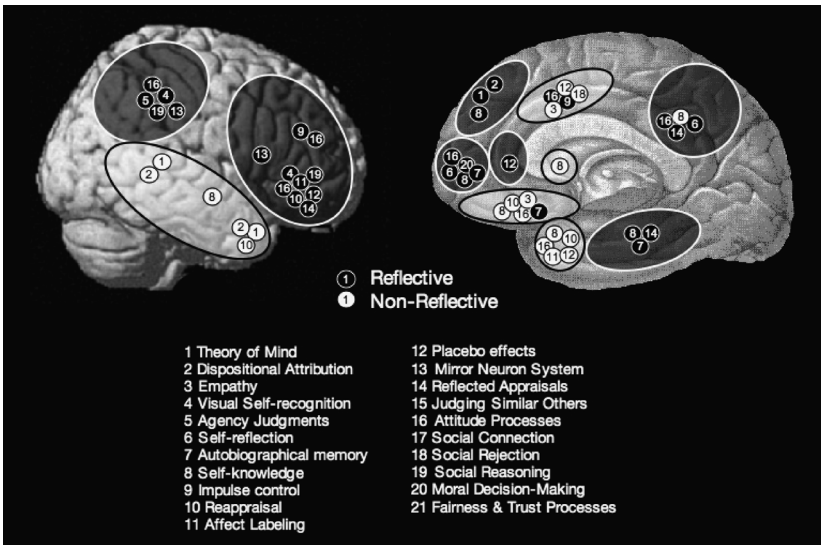

Fig. 13.3 Neural correlates of reflective and non-reflective processes from 21 domains of social cognition overlaid onto the X-system and C-system regions displayed in Figure 13.2.

reflective cognition does not make any contributions that cannot be duplicated by non-reflective cognition. To be honest, I'm not sure who if anyone is truly promoting the zombie hypothesis among empirical psychologists. I am confident, however, that the subtext being read into many studies of non-reflective processing is that the zombie hypothesis is true. A recent story in the New York Times (Carey, 2007) described these studies as revealing 'a subconscious brain that is far more active, purposeful and independent than previously known …. The brain appears to use the very same neural circuits to execute an unconscious act as it does a conscious one.' Thus, even if those doing zombie studies are conservative in the conclusions they draw, the lesson learned by the research community and the world at large is more expansive.

By reviewing a number of fMRI studies, I hoped to have at least cast doubt on the zombie hypothesis. We have repeatedly observed that the neural mechanisms invoked to support a non-reflective variant of a task are not present when the reflective variant of the same task is performed, even when the observable outputs of the two variants of similar. Instead, we find that the group of brain regions called the X-system tend to be active during non-reflective social cognition and a group of brain regions called the C-system tend to be active during reflective social cognition. Moreover in a number of studies, the degree to which C-system activity is invoked during reflective processes is inversely correlated with X-system activity during reflective processing. The zombie neural hypotheses suggest that the same processes doing the legwork during non-reflective social cognition should also be doing the legwork during reflective social cognition. Our data shows not only are different regions brought on line to do the legwork during reflective social cognition, but that the activation of these

reflective processes, far from relying on the neural mechanisms of non-reflective social cognition, seems to prevent the regions involved in non-reflective social cognition from doing any of the legwork.

For full disclosure, I must admit that there is indeed a fourth potential zombie neural hypothesis that was not rejected by the current review: however, on its face it requires tortured logic and suspension of disbelief. The fourth hypothesis is that separate neurocognitive systems (e.g. X- and C-systems) handle what are typically non-reflective and reflective processes, respectively. There are things that each of these systems can do computationally that the other cannot consistent with the description of processing and representational characteristics listed in Table 13.1. Although the neurocognitive systems that are invoked in reflective processing typically, if not always, occur along with some form of reflective awareness, it is possible that these same processes in the C-system could be invoked without reflective awareness under the right circumstances. This claim starts by admitting that the typical zombie studies do not really support the zombie hypothesis because there really are two systems. It then goes onto suggest that maybe, just maybe, the neurocognitive processes that are always associated with awareness—not those mimicked in typical zombie studies but those neurocognitive processes doing the work that is systematically correlated with reflective awareness—perhaps these processes could be produced without awareness. I must admit that there is no data that rules out this account, but it would be an enormous leap of faith because there is also no data to support this account either. Nevertheless, if someone wanted to provide evidence to support the zombie hypothesis, this is the sort of evidence that would be needed. Non-reflective cognition can mimic a number of the input-output relations produced by reflective cognition. Showing non-reflective production of a reflective process with C-system activation, rather than producing the same outputs, would be strong evidence.

This final zombie hypothesis notwithstanding, one remaining question is why would non-reflective processes so often mimic the outputs of reflective processes, thus giving rise to the empirical support for the zombie hypotheses? One way to think of non-reflective processes is as our own built in 'personal digital assistant' or PDA, like a Palm Pilot. There are various tasks that are relatively simple and straightforward that could each be done by a person but are a relief to hand off to the PDA. People remembered names, dates, addresses, and phone numbers before PDA and thus can do this, but people also forgot names, dates, addresses, and phone numbers. A PDA can only do certain things but what it can do it can do very reliably and it is designed to work to support our conscious goals and intentions. The X-system may be very much the same way, to serve the goals and intentions of the C-system by taking over repetitive tasks that would be effortful for the C-system to handle sequentially, freeing the C-system up to focus on things the X-system is not suited for. Thus, the X-system may be designed to mimic the C-system to the extent that it can give its own computational constraints and consequently in well-learned domains where associative processes can generate the outputs of interest, both the C-system and X-system are likely to produce results. However, when tasks tap into functions that fit the repertoire of one of the systems, but not the other, outputs should diverge.

Postscript: Interactions of the X- and C-systems

This chapter was intended to show that the neurocognitive systems supporting reflective and non-reflective social cognition could be distinguished from each other in ways that undermine the zombie hypothesis. The picture I've presented is one of independence or even of competition between the systems. Although the systems are separable and can be shown to compete, in many of life's daily experiences, these two systems work together quite closely to achieve the best outcomes. I've presented studies that were intended to differentiate the systems based on their distinguishing features. However, in everyday life, most tasks probably rely on both systems simultaneously.

For instance, there is evidence to suggest that medial PFC and lateral temporal cortex work together in theory of mind (ToM) tasks where the internal mental state of another must be taken into account (Frith and Frith, 2003). Lateral temporal cortex appears to support more automatic and non-reflective aspects of ToM (Lieberman, 2007a) by responding to facial cues such as facial expressions and eye gaze that relate to mental states in relatively straightforward ways, whereas medial PFC is invoked when an individual is explicitly thinking about the mental state of another. In this case, medial PFC is likely using inputs from lateral temporal cortex as part of the algebraic mental equation used in mental state inference. For instance, smiling is typically associated with positive emotional states without any reflective effort on the part of an observer. However, if the target being observed is a competitor in a poker tournament, then propositional logic may be used by medial PFC to combine these facial cues encoded in lateral temporal cortex with knowledge about situational context to reach the conclusion that this person might be smiling in order to bluff his competition. These high level inferences from medial PFC then may be fed back to lateral temporal cortex to filter incoming facial expressions for relevant information for additional hypothesis testing about the other person's intentions. So while it is possible to separate the contributions of medial PFC and lateral temporal cortex into reflective and non-reflective elements of ToM, respectively, they are also working together in harmony.

In conclusion, the zombie hypothesis may be tantalizing, but evidence from social cognitive neuroscience studies suggests this is not an accurate characterization of the mind. There are two systems, the X-system and the C-system, that are differentially responsible for non-reflective and reflective social cognition respectively. The systems appear to operate upon different principles and are invoked under different conditions. Although the two systems are capable of producing similar outputs in various situations, giving apparent support to the zombie hypothesis, neuroimaging shows that different mechanisms are indeed at work in these cases.

References

Bargh, J. A., Chen, M., and Burrows, L. (1996) Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, **71**, 230–44.

Bem, D. J., and McConnell, H. K. (1970) Testing the self-perception explanation of dissonance phenomena: On the salience of premanipulated attitudes. *Journal of Personality and Social Psychology*, **14**, 23–31.

Brehm, J. W. (1956) Postdecision changes in the desirability of alternatives. *The Journal of Abnormal and Social Psychology*, **52**, 384–89.

Carey, B. (2007) Who's minding the mind? *New York Times* (July 31, 2007)

Chaiken, S., and Trope, Y. (1999) *Dual-process theories in social psychology*. New York: Guilford Press.

Chartrand, T. L., and Bargh, J. A. (1996) Automatic activation of impression formation and memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions. *Journal of Personality & Social Psychology*, **71**, 464–78.

Cowey, A., and Weiskrantz, L. (1963) A perimetric study of visual field defects in monkeys. *Quarterly Journal of Experimental Psychology*, **15**, 91–115.

Craik, F. I. M., and Tulving, E. (1975) Depth of processing and retention of words in episodic memory. *Journal of Experimental Psychology: General*, **104**, 268–94.

Deutsch, R., Gawronski, B., and Strack, F. (2006) At the boundaries of automaticity: negation as reflective operation. *Journal of Personality and Social Psychology*, **91**, 385–405.

DeWall, C. N., Baumeister, R. F., and Masicampo, E. J. (2007) Evidence that logical reasoning depends on conscious processing. *Unpublished manuscript.*

Dijksterhuis, A., and Bargh, J. A. (2001) The perception-behavior expressway. *Advances in Experimental Social Psychology*, **33**, 1–40.

Dijksterhuis, A., and Nordgren, L. F. (2006) A theory of unconscious thought. *Perspectives on Psychological Science*, **1**, 95–109.

AQ – Eisenberger – please provide page range

Eisenberger, N. I., Gable, S. L., and Lieberman, M. D. (2007) fMRI responses relate to differences in real-world social experience. *Emotion, 4, xxx–xxx.*

Eisenberger, N. I., Lieberman, M. D., and Williams, K. D. (2003) Does rejection hurt? An fMRI study of social exclusion. *Science*, **302**, 290–2.

Evans, J. St. B. T. (2008) Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology*, **59**, 255–78.

Festinger, L. (1957) *A theory of cognitive dissonance*. Evanston, Ill.: Row, Peterson.

Foerde, Knowlton, and Poldrack (2006 Foerde, K. E., Knowlton, B. J., and Poldrack, R. A. (in preparation) Secondary task effects on classification learning examined using fMRI.

Fredrickson, B. L., and Kahneman, D. (1993) Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, **65**, 45–55.

Frith, U., and Frith, C. D. (2003) Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London*, **358**, 459–73.

Gawronski, B. and Strack, F. (2004) On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology*, **40**, 535–42.

Grafton, S. T., Hazeltine, E., and Ivry, R. (1995) Functional mapping of sequence learning in normal humans. *Journal of Cognitive Neuroscience, 7*, 497–510.

Greenwald, A. G., and Liu, T. J. (1985) *Limited unconscious processing of meaning*. Presented at annual meeting of the Psychonomic Society, Boston.

Hamilton, D. L., Katz, L. B., and Leirer, V. O. (1980) Cognitive representation of personality impressions: organizational processes in first impression formation. *Journal of Personality and Social Psychology*, **39**, 1050–63.

Hariri, A. R., Bookheimer, S. Y., and Mazziotta, J. C. (2000) Modulating emotional response: Effects of a neocortical network on the limbic system. *NeuroReport*, **11**, 43–8.

Jarcho, J., Berkman, E., and Lieberman, M.D. (2007, May) *Neural correlates of post-decisional attitude change*. Poster presented at the Neural Systems of Social Behavior conference, Austin, TX.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., and Redelmeier, D. A. (1993) When more pain is preferred to less: Adding a better ending. *Psychological Science*, **4**, 401–5.

Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., and Heatherton, T. F. (2002) Finding the self?: An event-related fMRI study. *Journal of Cognitive Neuroscience*, **14(5)**, 785–94.

Kirk, R., 1974a, 'Sentience and Behaviour', Mind **83**: 43–60

Knowlton, B. J., Mangels, J. A., and Squire, L. R. (1996) A neostriatal habit learning system in humans. *Science*, **273**, 1399–1402.

Kunda, Z., and Sherman-Williams, B. (1993) Stereotypes and the construal of individuating information. *Personality and Social Psychology Bulletin*, **19**, 90–9.

Lieberman, M. D. (2000) Intuition: A social cognitive neuroscience approach. *Psychological Bulletin*, **126**, 109–37.

Lieberman, M. D. (2007a) Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology*, **58**, 259–89.

Lieberman, M. D. (2007b) The X- and C-systems: The neural basis of automatic and controlled social cognition. To appear in E. Harmon-Jones and P. Winkelman (Eds.), *Fundamentals of Social Neuroscience* (290–315). New York: Guilford.

Lieberman, M. D., Chang, G. Y., Chiao, J., Bookheimer, S. Y., and Knowlton, B. J. (2004) An event-related fMRI study of artificial grammar learning in a balanced chunk strength design. *Journal of Cognitive Neuroscience*, **126**, 427–38.

Lieberman, M. D., Gaunt, R., Gilbert, D. T., and Trope, Y. (2002) Reflection and reflexion: A social cognitive neuroscience approach to attributional inference. *Advances in Experimental Social Psychology*, **34**, 199–249.

Lieberman, M. D., Hariri, A., Jarcho, J. M., Eisenberger, N. I., and Bookheimer, S. Y. (2005) An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. *Nature Neuroscience*, **8**, 720–2.

Lieberman, M. D., Jarcho, J. M., and Satpute, A. B. (2004) Evidence-based and intuition-based self-knowledge: An fMRI study. *Journal of Personality and Social Psychology*, **87**, 421–35.

Lieberman, M. D., Ochsner, K. N., Gilbert, D. T., and Schacter, D. L. (2001) Do amnesics exhibit cognitive dissonance reduction? The role of explicit memory and attention in attitude change. *Psychological Science*, **12**, 135–40.

Markus, H. R. (1977) Self-schemata and processing information about the self. *Journal of Personalty and Social Psychology*, **35**, 63–78.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995) Why there ae complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, **102**, 419–57.

Metcalfe, J., and Mischel, W. (1999) A hot/cool system analysis of delay of gratification: dynamics of willpower. *Psychological Review*, **106**, 3–19.

Nisbett, R. and Ross, L. (1980) *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.

AQ – Kirk; please clarify reference

AQ - Nisbett, R. and Ross, L. (1980), can't find reference to in main text

Packard, M. G., Hirsh, R., and White, N. M. (1989) Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: Evidence for multiple memory systems. *Journal of Neuroscience*, **9**, 1465–72.

AQ – Peyron – correct year ?

Peyron, R., Garcia-Larrea, L., Gregoire, M., Costes, N., Convers, P., Lavenne, F., Mauguiere, F., Michel, D., and Laurent, B. (1999) Haemodynamic brain responses to acute pain in humans: Sensory and attentional networks. *Brain*, **122**, 1765–79.

Pfeifer, J. H., Lieberman, M. D., and Dapretto, M. (2007) 'I know you are but what am I?!': An fMRI study of self-knowledge retrieval during childhood. *Journal of Cognitive Neuroscience*, **19**, 1323–37.

Poldrack, R. A., Clark, J., ParÈ-Blagoev, E. J., Shohamy, D., Creso Moyano, J., Myers, C., and Gluck, M. A. (2001) Interactive memory systems in the human brain. *Nature*, **414**, 546–50.

Rainville, P., Duncan, G. H., Price, D. D., Carrier, B., and Bushnell, M. C. (1997) Pain affect encoded in human anterior cingulate but not somatosensory cortex. *Science*, **277**, 968–71.

Rameson, L. and Lieberman, M. D. (2007.) Thinking about the self from a social cognitive neuroscience perspective. *Psychological Inquiry*, **18**, 117–22.

Reber, A. S. (1967) Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, **6**, 855–63.

Redelmeier, D. A., and Kahneman, D. (1996) Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain*, **66**, 3–8.

Satpute, A.B., and Lieberman, M. D. (2006) Integrating automatic and controlled processing into neurocognitive models of social cognition. *Brain Research*, **1079**, 86–97.

Sloman, S. A. (1996) The empirical case for two systems of reasoning. *Psychological Bulletin*, **119**, 3–22.

Smith, E. R., and DeCoster, J. (1998) Knowledge acquisition, accessibility, and use in person perception and stereotyping: simulation with a recurrent connectionist network. *Journal of Personality and Social Psychology*, **74**, 21–35.

Velmans, M. (1991) Is human information processing conscious? Behavioral and Brain *Sciences*, **14**, 651–726.

Wegner, D. W. (2003) *The illusion of conscious will*. Cambridge, MA: MIT Press.

Weiskrantz, L., Warrington, E.K., Sanders, M.D. and Marshall,J. (1974) Visual capacity in the hemianopic field, following a restricted occipital ablation. *Brain*, **97**, 709–28.

Whalen, P. J., Rauch, S. L., Etcoff, N. L., McInerney, S. C., Lee, M. B., and Jenike, M. A. (1998) Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *The Journal of Neuroscience*, **18**, 411–18.