Journal of Neuroscience, Psychology, and Economics
2011, Vol. 4, No. 3, 174–180

1937-321X/11/$12.00 DOI: 10.1037/a0024654

The Dark Side of Social Encounters: Prospects for a Neuroscience of Human Evil

Martin Reimann
University of Southern California

Philip G. Zimbardo
Stanford University

This article discusses how findings from social, cognitive, and affective neuroscience might contribute to our understanding of human evil. Integrating theories of personality and social psychology as well as the notions of deindividuation and dehumanization with recent neuroscientific insight, the authors elaborate on the nature of human evil and its potential roots in brain systems associated with affective processing and cognitive control.

Keywords: human evil, aggression, deindividuation, dehumanization, social, cognitive, affective neuroscience

While much neuroscientific research focuses on the neural basis of positive social responses such as empathy, fairness, and trust, the neural correlates of negative social encounters are largely unknown. Specifically, what are the neurophysiological substrates of human evil? Or, in other words, which neural mechanisms underlie this basic social phenomenon? For several decades, personality and social psychology has informed us of two processes that have been named central to human evil: deindividuation and dehumanization (Bandura, 1982; Pines & Solomon, 1977; Zimbardo, 1969; Zimbardo, 2007). Whereas deindividuation refers to the process that facilitates the perception of others as anonymous (Zimbardo, 1969), dehumanization is at the core of much evil, occurring when one excludes another from the moral order as a human being, or less than human, as animallike (Zimbardo, 2007). Dehumanization can be conceptualized as a "cortical cataract," that blurs our perception of others as having any similarity to our kind or us. It is also a central process in genocide, mass murder, rape as a terror tactic, and in prejudice.

In this research, we propose how findings from social, cognitive, and affective neuroscience might contribute to the study of deindividuation and dehumanization and, therefore, at least two aspects of human evil. We outline several ideas advanced by leading neuroscientists that are provocative and in need of further experimental testing. We also present a framework derived from aggression research, which might serve for testing of hypotheses on the neuroscience of human evil. Although the neuroscience of human evil is in its infancy, programmatic research can yield interesting insights on one basic form of human social interaction.

While researchers have put forth calls for a better neural understanding of evil, few investigators have conducted empirical research. One reason for this inattention may be the absence of work that bridges concepts from personality and social psychology with insights from neuroscience. Such a conceptual integration is the main objective of the current research. We start with a general framework rooted in aggression research that is relevant to the understanding of human evil. This framework provides a context for forming empirical questions and testing hypotheses on a neuroscience of human evil. On the basis of this framework, we develop a preliminary model on the neural correlates of deindividuation and dehumanization, essential for subsequent experimental analyses.

This article was published Online First July 18, 2011.

Martin Reimann, Department of Psychology, University of Southern California; Philip G. Zimbardo, Psychology Department, Stanford University.

Correspondence concerning this article should be addressed to Martin Reimann, University of Southern California, Department of Psychology, Seeley G. Mudd Building, 3620 McClintock Avenue, Los Angeles, CA 90089-1061. E-mail: mreimann@usc.edu

A Framework From Aggression Research

We build on extant research on aggression to argue that a unique neural framework underlies the mechanisms of human evil. Besides conducting psychiatric research on the neurobiological roots of aggression such as genes and neurotransmitters (e.g., Volavka, 1999), investigators have also focused on the functional neuroanatomy of aggression. Generally, human aggression research offers generic concepts on antisocial behavior (Bandura, 1973; Buss, 1961; Buss & Durkee, 1957; Caprara, Barbaranelli, & Zimbardo, 1996), with a number of definitions and several views of the construct existing in the literature (for a review on the social, cognitive, and affective dimensions of human aggression, see Caprara et al., 1996).

One frequently cited definition characterizes human aggression as "a response that delivers noxious stimuli to another organism" (Buss, 1961, p. 1), and a widely used concept of aggression consists of seven dimensions: assault, indirect aggression, irritability, negativism, resentment, suspicion, and verbal aggression (Buss & Durkee, 1957). However, measurement issues with this first aggression inventory resulted in the introduction of a condensed four-factor model of aggression that includes hostility, anger, verbal aggression, and physical aggression (Buss & Perry, 1992). Hostility consists of ill will and a sense of injustice and embodies the cognitive component of behavior. It can be defined as the stubborn refusal to accept proof that the perceptions of one's surrounding are false. Anger entails physiological arousal (e.g., increased heart rate, blood pressure, and levels of adrenaline and noradrenaline) and preparation for verbal and physical aggression. Anger thus represents the emotional or affective component of behavior. It is also expressed externally and is apparent in facial expressions and body language (for a guideline on how to detect subtle anger in facial expressions, see Ekman & Friesen, 2003). Verbal aggression and physical aggression, the third and fourth components of aggression, both involve attempts to hurt or harm others and, as such, represent the instrumental or motor component of behavior (Buss & Perry, 1992).

One prominent incident in medical history marks a starting point for the neuroscientific study of aggression: the case of a nineteenth-century railroad worker, Phineas Gage, who is best remembered for surviving an accident in which an iron rod was driven through his head, damaging large sections of his frontal lobe. The photograph in Figure 1 is believed to be one of Phineas Gage, holding the iron rod (Wilgus & Wilgus, 2009).

After this accident, Phineas Gage's physician noted dramatic changes in his personality, social conduct, judgment, and decision-making (Harlow, 1848, 1868). Gage refused to show respect for social conventions and offended other people with profanities, and he also became aggressive in his behavior (Harlow, 1868). While previous researchers investigating Gage's case had to rely on Harlow's notes to draw conclusions, developments in technology have allowed more recent researchers to obtain evidence confirming the notion that Gage's aggression was linked to specific frontal lobe damages. Using neuroimaging techniques to examine Gage's preserved skull, investigators found that the lesion Gage experienced affected the prefrontal cortex (Damasio, Grabowski, Frank, Galaburda, & Damasio, 1994). Additionally, research in lesion patients as well as healthy in-

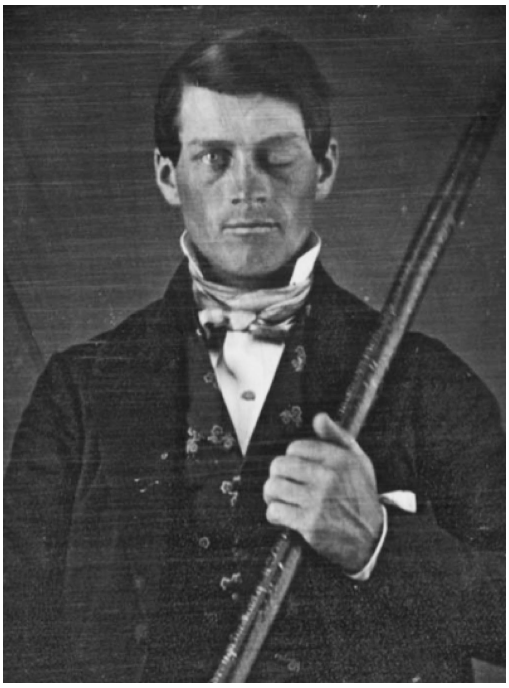

Figure 1. A case of frontal lobe damage and aggression: Phineas Gage. Printed with permission of B. Wilgus.

dividuals undergoing neuroimaging has confirmed the relationship between frontal lobe damage and aggression. In particular, studies have shown that patients with prefrontal cortex lesions can show disinhibited and socially inappropriate behaviors such as aggression (e.g., Rolls, Hornak, Wade, & McGrath, 1994). Furthermore, extant neuroimaging research has identified functional deactivation in the ventromedial prefrontal cortex of individuals responding to pictures that show aggressive behavior (e.g., Pietrini, Guazzelli, Basso, Jaffe, & Grafman, 2000).

Through a content analysis of several neuroimaging studies, researchers studied the relationship between frontal lobe pathology and aggressive behavior (Bufkin & Luttrell, 2005). Across a number of different neuroimaging methodologies (including single-photon emission computed tomography, positron emission tomography, and functional magnetic resonance imaging), the investigators report a strong link between deficits in frontal functioning and aggression. In particular, analyses of specific regions in the medial prefrontal cortex revealed that individuals who were aggressive had significantly lower prefrontal activity in the orbitofrontal cortex, anterior medial cortex, medial frontal cortex, and/or superior frontal cortex. The content analysis also identified the temporal lobe and several subcortical structures as being involved in aggression, although these regions were not the focus of most studies in the review (Bufkin & Luttrell, 2005). Overall, extant research on aggression suggests that decreased activation of frontal lobe structures, particularly the prefrontal cortex, or lesioning of this brain area can be a central cause for aggression, although another review identified the amygdala as a central subcortical structure associated with aggression (Siever, 2008).

Somatic marker theory (Damasio, 1994, 1996; Damasio, Tranel, & Damasio, 1991; Reimann & Bechara, 2010) provides the theoretical basis of the relationship between the ventromedial prefrontal cortex and other brain areas, which include both brain stem structures (i.e., hypothalamus and periaqueductal gray) and the amygdala. According to this theory, emotion-related signals (i.e., somatic markers, also sometimes termed bodily markers), which are indexed changes in the visceral state (e.g., changes in heart rate, blood pressure, gut motility, and glandular secretion), interact with cognitive processes. These changes in the visceral state can be considered a form of anticipation of the bodily impact of events in the world that allows an organism to maximize the survival value of particular situations. These situations could promote homeostasis, such as an opportunity to engage in social interaction, as well as events that disrupt homeostasis, such as a signal of deindividuation (e.g., being treated anonymously rather than as an individual) or dehumanization (e.g., a serious threat of being excluded as a human being). Overall, these visceral responses are one component of a broader emotional response system that also includes changes in the endocrine and skeletomotor systems, and within the brain (Damasio, 1994).

Somatic marker theory further proposes several neural structures that studies have shown to be key components of the neural circuitry underlying somatic state activation. The amygdala and the ventromedial prefrontal cortex region (i.e., medial frontal and orbitofrontal cortex) are critical structures for triggering somatic states. Whereas the amygdala seems to be more important for triggering somatic states from emotional events that occur in the external environment, the ventromedial prefrontal cortex region seems to be more vital for triggering somatic states from the internal cortical environment of memories and knowledge (Bechara & Damasio, 2005).

Toward a Neuroscience of Human Evil

In personality and social psychology, two processes have been named central to human evil: deindividuation and dehumanization (Bandura, 1982; Pines & Solomon, 1977; Zimbardo, 1969; Zimbardo, 2007). The deindividuation and dehumanization phenomena offer a framework for explaining the antisocial behavior of aggressive crowds, such as hooligans and the lynch mob (Postmes, Spears, & Lea, 1998). Deindividuation theory has also been applied to help explain antisocial behavior in computer-mediated communication (Kiesler, Siegel, & McGuire, 1984; Kiesler & Sproull, 1992) and group decision support systems (Jessup, Connolly, & Tansik, 1990; for a history and meta-analysis of deindividuation theory, see Postmes et al., 1998).

Deindividuation

The idea of deindividuation goes back to the work of Le Bon (1897), was reintroduced by Festinger, Pepitone, and Newcomb (1952), and then extended and developed by Zimbardo (1969). The term refers to the process that facilitates the perception of others as lacking in personal identity (Pines & Solomon, 1977), and people engaging in deindividuation treat others as anonymous, not as individuals (Zimbardo, 2007). Deindividuation also includes processes that encourage a person to mask his or her identity in various ways, thereby reducing a sense of personal accountability. Decades of research have demonstrated that deindividuation strongly predicts the odds of antisocial behavior such as aggression. Social circumstances that promote anonymity increase the probability of eliciting antisocial behaviors when individuals perceive they have permission to be hostile or aggressive or to break social norms. Deindividuation can result from a variety of factors in addition to anonymity and loss of personal accountability, including sensory overload, unstructured situations, and substance abuse.

Social psychologists have long recognized the crucial role of anonymity in antisocial behavior. Beginning with the Stanford Prison Experiment in 1971 (Haney, Banks, & Zimbardo, 1973), numerous studies have demonstrated that a feeling of anonymity can often cause antisocial behavior. In the Stanford Prison Experiment, physically fit and mentally stable young men were randomly assigned to play the roles either of "guard" or "prisoner" for two weeks in the basement of the psychology department of Stanford University. With the cooperation of the local police department, prisoners were unexpectedly arrested, brought to the police station for finger printing and registration, and then placed in a detention cell. After being blindfolded, each prisoner was guided to the mock prison at Stanford University. To promote anonymity of the subjects, each group was issued identical uniforms and identification numbers replaced their names. Guards' uniforms consisted of plain khaki shirts and trousers, whistle, wooden baton, and reflecting sunglasses, the last making eye contact impossible. Prisoners' uniforms were loose-fitting muslin smocks with identification numbers on the front and back and rubber sandals. No underwear was worn under the smocks and a chain and lock were placed around one ankle. After only 6 days, the experiment had to be stopped because the behavior of the guards toward the prisoners had become increasingly aggressive and the prisoners were experiencing severe emotional distress. Since this experiment, research on deindividuation has been interpreted as suggesting that any situation that makes individuals feel anonymous and reduces their sense of personal accountability can potentially lead to evil behavior.

While neuroscientific research on deindividuation is nonexistent in the literature, extant neuroimaging research investigating how social gestures affect brain activation provides some limited clues to potential neural markers associated with deindividuation. In a recent study, participants underwent functional MRI while viewing various social gestures, including fascist saluting or simple waving (Knutson, McClellan, & Grafman, 2008). The investigators found that highly provocative social gestures (i.e., fascist saluting) compared to less provocative but still socially meaningful gestures (i.e., waving) are associated with activation in the medial prefrontal cortex, among other areas. The researchers also compared brain activation of participants viewing multiple gesturing actors with activation while viewing a single gesturing actor and found an involvement of the frontal lobe for viewing multiple gesturing actors. However, results were ambiguous as to whether this effect was due to participants being deindividuated because they felt they were an anonymous member of a group with a resulting loss in personal accountability (as argued by these investigators) or whether this effect was due to the perception of many different motor movements.

Dehumanization

Dehumanization occurs when one person considers others to be excluded from the moral order of being a human being (Zimbardo, 2007). It involves the denial of uniquely human attributes and is often accompanied by contempt and disgust and by a tendency to explain others' behavior in terms of one's own desires rather than cognitive states (Haslam, 2006). Dehumanization is at the core of human evil, because it depicts a person as less than human. It occurs in a variety of stressful human encounters, such

as when a large flow of people has to be managed or processed efficiently (Bandura, 1977; Pines & Solomon, 1977; Zimbardo, 1969), and can be considered an everyday social phenomenon (Haslam, 2006).

Some pioneering neuroimaging research has examined dehumanization. In particular, a recent neuroimaging study reports decreased brain activation in the medial prefrontal cortex and increased activation in the amygdala and the insula when participants viewed pictures of stereotypically hostile and stereotypically incompetent groups (Harris & Fiske, 2006). The investigators used functional MRI in a between-subjects design, showing photographs of social groups to one group and pictures of objects to another group. Results demonstrated that perceiving extreme out-groups (e.g., homeless people and drug addicts) results in decreased medial frontal cortex activation and increased amygdala and insula activation. The authors argue that these extreme out-groups may be perceived as less than human, or dehumanized (Harris & Fiske, 2006). This finding suggests that when humans dehumanize others, less medial prefrontal cortex activation but more amygdala and insula activation is involved. In sum, these findings point in the same direction as previous neuroscientific research on aggression, which predicts less activation in the ventromedial prefrontal cortex in association with aggressive behavior. The results are also in line with the case of Phineas Gage, which revealed that ventromedial prefrontal cortex lesioning can result in a loss of "cognitive control" and the rise of aggressive behavior.

Although drawing conclusions regarding the neurophysiological markers of deindividuation and dehumanization would be premature, extant research on aggression points toward the prefrontal cortex as a crucial mechanism in human evil. Besides the prefrontal cortex, prior studies have named other subcortical brain regions, such as the amygdala, as being important. We borrow from somatic marker theory (Damasio, 1994, 1996; Damasio et al., 1991; Reimann & Bechara, 2010) to derive a preliminary model of deindividuation and dehumanization, their potential neural correlates, and downstream aggression. As Figure 2 shows, being deindividuated and/or dehumanized could potentially involve a network of brain areas, including the ventromedial prefrontal cortex, the amygdala, and brainstem structures (i.e., hypothalamus and periaqueductal gray). As reviewed above, somatic marker theory predicts that the amygdala is more important for triggering somatic states from emotional events that occur in the environment and that the ventromedial prefrontal cortex region is more vital for triggering somatic states from memories and

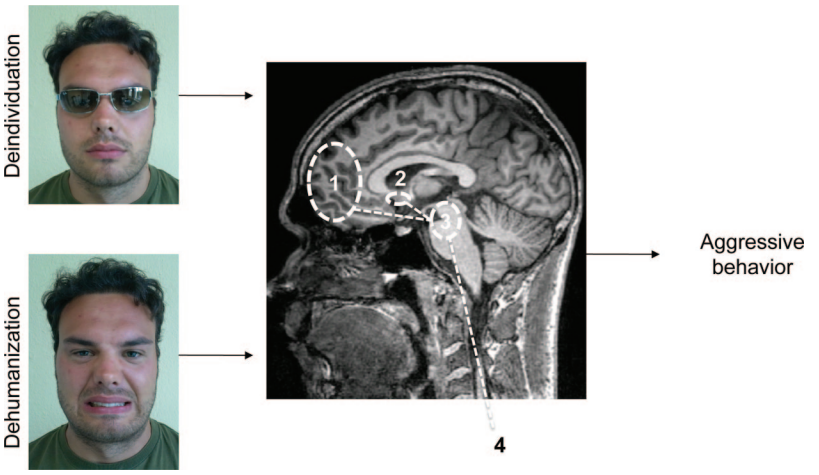

Figure 2. A preliminary model of neurophysiological and behavioral correlates of deindividuation and dehumanization. Note. (1) Ventromedial prefrontal cortex, (2) amygdala, (3) brainstem structures, and (4) visceral responses such as changes in heart rate, blood pressure, gut motility, and glandular secretion.

knowledge (Bechara & Damasio, 2005). In addition, deindividuation and dehumanization could involve visceral responses, such as changes in heart rate, blood pressure, gut motility, and glandular secretion, which could serve as important somatic markers for evaluating social encounters and for controlling aggressive behavior.

In the case of deindividuation, a decrease in activation of the ventromedial prefrontal cortex could possibly result from a "feeling" of anonymity and a loss of personal accountability, because somatic states are not triggered by memories and knowledge of social norms and, thus, lead to disinhibited, antisocial behaviors. In the case of dehumanization, decreased ventromedial prefrontal cortex activation could also be accompanied by an increase in amygdala activation (Harris & Fiske, 2006). Being excluded from the moral order of human beings (that is, being dehumanized) could pose an immediate threat that leads to increased processing in the amygdala, which in turn triggers somatic states in the viscera via the brainstem.

Conclusion

With the goal of better understanding core processes of human evil, this analysis brings together several disparate themes from personality and social psychology as well as social, cognitive, and affective neuroscience. The prospects for a neuroscience of human evil are promising. Applying ideas from investigations of aggression as well as preliminary findings from neuroimaging studies of deindividuation and dehumanization may guide the development of future hypotheses that test the cognitive and affective mechanisms underlying human evil.

Of the many issues within this domain of research, the most pressing is the need for empirical testing. Here, the differentiation between deindividuation and dehumanization in terms of functional neuroanatomy will pose certain methodological challenges. Are the same or different neural processes involved for deindividuation and dehumanization, and to what extent? Another interesting issue worth pursuing is the distinction between the person who deindividuates and/or dehumanizes and the one being deindividuated and/or dehumanized. Reports from the Stanford Prison Experiment state that both guards and prisoners showed negative, hostile, and affrontive attitudes and behaviors when encountering each other (Haney et al., 1973). Thus, does it matter, in terms of the underlying neural mechanisms, which perspective one takes—the perspective of the guard or the perspective of the prisoner? Could even the guards, who were obviously in power, be subject to the same neural mechanisms because of their uniform appearance, which resulted in anonymity and loss of personal accountability? In addition to addressing these questions, other worthwhile investigations include the causality between deindividuation/dehumanization, brain mechanisms, and aggressive behaviors. Researchers could also examine the possible mediating role of these brain processes in the relationship between deindividuation and dehumanization and aggression.

Beyond using functional neuroimaging in healthy individuals, researchers should also conduct experimental work in patients with brain damage. While functional neuroimaging studies will yield a broader picture of the neural processes underlying deindividuation and dehumanization, lesion studies could focus on prefrontal cortex patients and, therefore, help to further pinpoint a brain area that seems to be implicated in human evil. Moreover, investigating the neural correlates of human evil may advance existing areas of inquiry, such as aggression research and somatic marker theory, and also reveal new findings and further research questions.

References

Bandura, A. (1973). *Aggression: A social learning analysis.* Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84,* 191–215.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37,* 122–147.

Bechara, A., & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior, 52,* 336–372.

Bufkin, J. L., & Luttrell, V. R. (2005). Neuroimaging studies of aggressive and violent behavior. *Trauma, Violence, & Abuse, 6,* 176–191.

Buss, A. H. (1961). *The psychology of aggression.* New York, NY: Wiley.

Buss, A. H., & Durkee, A. (1957). An inventory for assessing different kinds of hostility. *Journal of Consulting Psychology, 21,* 343–349.

Buss, A. H., & Perry, M. (1992). The aggression questionnaire. *Journal of Personality and Social Psychology, 63,* 452–459.

Caprara, G. V., Barbaranelli, C., & Zimbardo, P. (1996). Understanding the complexity of human aggression: Affective, cognitive, and social dimensions of individual differences in propensity toward aggression. *European Journal of Personality, 10,* 133–155.

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

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences, 351,* 1413–1420.

Damasio, A. R., Tranel, D., & Damasio, H. (1991). Somatic markers and the guidance of behavior: Theory and preliminary testing. In H. S. Levin, H. M. Eisenberg, & A. L. Benton (Eds.), *Frontal lobe function and dysfunction* (pp. 217–229). New York, NY: Oxford University Press.

Damasio, H., Grabowski, T., Frank, R., Galaburda, A. M., & Damasio, A. R. (1994). The return of Phineas Gage: Clues about the brain from the skull of a famous patient. *Science, 264,* 1102–1105.

Ekman, P., & Friesen, W. V. (2003). *Unmasking the face: A guide to recognizing emotions from facial clues.* Cambridge, MA: Malor Books.

Festinger, L., Pepitone, A., & Newcomb, T. (1952). Some consequences of de-individuation in a group. *Journal of Abnormal and Social Psychology, 47,* 382–389.

Haney, C., Banks, C., & Zimbardo, P. G. (1973). Interpersonal dynamics in a simulated prison. *International Journal of Criminology & Penology, 1,* 69–97.

Harlow, J. M. (1848). Passage of an iron rod through the head. *Boston Medical and Surgical Journal, 39,* 389–393.

Harlow, J. M. (1868). Recovery from the passage of an iron bar through the head. *Publications of the Massachusetts Medical Society 2,* 327–347.

Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the lowest of the low: Neuroimaging responses to extreme out-groups. *Psychological Science, 17,* 847–853.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10,* 252–264.

Jessup, L. M., Connolly, T., & Tansik, D. A. (1990). Toward a theory of automated group work: The deindividuating effects of anonymity. *Small Group Research, 21,* 333–348.

Kiesler, S., Siegel, J., & McGuire, T. W. (1984). Social psychological aspects of computer-mediated interaction. *American Psychologist, 39,* 1123–1134.

Kiesler, S., & Sproull, L. (1992). Group decision making and communication technology. *Organizational Behavior and Human Decision Processes, 52,* 96–123.

Knutson, K. M., McClellan, E. M., & Grafman, J. (2008). Observing social gestures: An fMRI study. *Experimental Brain Research, 188,* 187–198.

Le Bon, G. (1897). *The crows: A study of popular mind.* New York, NY: The Macmillan Company.

Pietrini, P., Guazzelli, M., Basso, G., Jaffe, K., & Grafman, J. (2000). Neural correlates of imaginal aggressive behavior assessed by positron emission tomography in healthy subjects. *American Journal of Psychiatry, 157,* 1772.

Pines, A., & Solomon, T. (1977). Perception of self as a mediator in the dehumanization process. *Personality and Social Psychology Bulletin, 3,* 219.

Postmes, T., Spears, R., & Lea, M. (1998). Breaching or building social boundaries?: Side-effects of computer-mediated communication. *Communication Research, 25,* 689–715.

Reimann, M., & Bechara, A. (2010). The somatic marker framework as a neurological theory of decision-making: Review, conceptual comparisons, and future neuroeconomic research. *Journal of Economic Psychology, 31,* 767–776.

Rolls, E., Hornak, J., Wade, D., & McGrath, J. (1994). Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. *Journal of Neurology, Neurosurgery & Psychiatry, 57,* 1518.

Siever, L. J. (2008). Neurobiology of aggression and violence. *American Journal of Psychiatry, 165,* 429–442.

Volavka, J. (1999). The neurobiology of violence: An update. *Journal of Neuropsychiatry and Clinical Neurosciences, 11,* 307.

Wilgus, J., & Wilgus, B. (2009). Face to face with Phineas Gage. *Journal of the History of the Neurosciences, 18,* 340–345.

Zimbardo, P. (1969). The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In W. J. Arnold & D. Levine (Eds.), *Nebraska symposium on motivation* (pp. 237–307): University of Nebraska Press.

Zimbardo, P. G. (2007). *The Lucifer effect: Understanding why good people turn evil.* New York, NY: Random House.