OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

10

Morphing Morals

NEUROCHEMICAL MODULATION OF MORAL JUDGMENT AND BEHAVIOR

Molly J. Crockett

HOW DOES BRAIN chemistry affect morality? Simple observations and introspection reveal that people sometimes feel motivated to help others, but not always. Moral sentiments are diverse and dynamic: they differ between individuals, and within individuals over time. This observation suggests that moral sentiments could be shaped by *neuromodulators*, brain chemicals that modify neuronal dynamics, excitability, and synaptic function. Here I will review recent studies examining how the neuromodulator serotonin influences moral judgment and behavior.

In our work, we focused on serotonin for several reasons (Crockett et al. 2008). If one were to engineer a neurotransmitter system for the purpose of governing something as complex as moral behavior, one would want a system that is evolutionarily old yet easily modifiable, widely distributed in the brain, with many receptor subtypes to allow for maximum flexibility in control of behavior. The serotonin system fulfills all of these requirements (Insel and Winslow 1998), and has been implicated in regulating a wide range of social behaviors across species. In both primates and humans, serotonin function tends to covary positively with "prosocial" behaviors, such as grooming, cooperation, and affiliation, and tends to covary negatively with "antisocial" behaviors such as aggression and social isolation (Higley and Linnoila 1997; Higley et al. 1996; Higley et al. 1996; Knutson et al. 1998; Krakowski 2003; Linnoila et al. 1983; Moskowitz et al. 2001; Raleigh et al. 1991). Such prosocial and antisocial behaviors are likely precursors to human morality (Brosnan and Waal 2012; Proctor et al. 2013; Stevens et al. 2005).

acprof-9780199357666.indd 237 5/24/2016 5:03:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

In one study, we examined how serotonin influences moral judgments in the domain of harm and care. Humans have a basic aversion to harming others that infuses moral judgment and moral behavior. Such harm aversion is thought to underlie deontological moral judgments, for example, judging it to be morally unacceptable to kill one person in order to save five others. One particularly elegant demonstration of how harm aversion is related to deontological moral judgment comes from a study by Cushman and colleagues. In the lab, subjects were asked to perform and observe various "fake" harmful acts, such as shooting a fake gun and smashing a doll's head against a table. Subjects' physiological reactivity to both performing and observing the harmful acts was correlated with subsequent deontological moral judgments (Cushman et al. 2012).

How does serotonin affect deontological moral judgments? We conducted a study in which participants judged the moral permissibility of harmful actions in hypothetical scenarios on three separate occasions (Crockett et al. 2010). On one occasion, they received the drug citalopram, a selective serotonin reuptake inhibitor. Citalopram enhances serotonin function by blocking its reuptake after it has been released into the synapse, thus prolonging serotonin's actions on postsynaptic receptors. On another occasion, they received a different drug, atomoxetine, a noradrenaline reuptake inhibitor. Finally, on a third occasion, they received a placebo pill.

We compared the effects of citalopram, atomoxetine, and placebo on judgments in three types of scenarios: neutral scenarios that contained no moral content, "personal" moral scenarios in which harmful actions were emotionally salient, and "impersonal" moral scenarios in which harmful actions were not emotionally salient. There were no differences across treatments on judgments in the neutral scenarios or the impersonal scenarios. However, in the personal scenarios, we found that citalopram seemed to increase harm aversion—participants were more deontological, that is, less likely to endorse harming one to save many others.

We also examined how the drug effects interacted with individual differences in empathy. We might expect that individuals who have a stronger baseline level of harm aversion could be more susceptible to the effects of serotonin manipulations. Our data supported this prediction. We split participants into low- and high-empathy groups, defined by scores in the Interpersonal Reactivity Index (Davis 1983). We found no effect of citalopram on judgments in the low-empathy group, but a strong effect in the high-empathy group. In our sample, the most harm-averse individuals were those who scored high on empathy and also received citalopram.

In another set of studies, we examined how serotonin shapes morality in the domain of fairness and reciprocity. Humans care deeply about fairness, to the extent that they are willing to incur personal costs in order to punish unfair behavior and enforce fair outcomes. Such "costly punishment" plays an important role in encouraging and sustaining cooperation (Fehr and Fischbacher 2003).

In our studies, we measured costly punishment behavior using the Ultimatum game (Güth et al. 1982). This game has two players, a proposer and a responder. They

acprof-9780199357666.indd 238 5/24/2016 5:03:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

have to agree on how to split a sum of money, or neither of them gets any money. The proposer makes an offer to the responder. The responder can accept the offer, in which case both players are paid accordingly. Or he can reject the offer, in which case neither player is paid. A receiver motivated solely by self-interest will accept any offer, because something is better than nothing. But many studies have shown that most receivers would rather have nothing than let the proposer get away with taking the lion's share.

We were interested in how costly punishment in the ultimatum game would be sensitive to manipulations of the serotonin system. In our first study (Crockett et al. 2008), we used a technique called acute tryptophan depletion, which temporarily lowers the amount of serotonin that is available to the brain, allowing us to observe behavior when the brain is in a low-serotonin state. We compare the effects of acute tryptophan depletion to a placebo treatment in a double-blind study.

Participants played the role of responder in a series of one-shot ultimatum games, interacting with a different proposer on each round. We deliberately used one-shot games to rule out strategic motivations for rejecting unfair offers. In a repeated game, rejecting an unfair offer can induce the proposer to offer higher amounts in subsequent rounds. We were interested in rejection behavior that is motivated solely by the desire to punish unfair behavior, which can be measured cleanly in one-shot games.

We found that impairing serotonin function with acute tryptophan depletion increased the rejection of unfair offers, without affecting the (low) rejection of fair offers. In a second study, we examined the effects of enhancing serotonin function with citalopram on costly punishment in the ultimatum game (Crockett et al. 2010). We compared the effects of citalopram with those of atomoxetine and placebo. As with the moral judgment study, we found that citalopram influenced decision-making, but atomoxetine had no effect. Specifically, citalopram reduced the rejection of unfair offers, producing the opposite effect to acute tryptophan depletion. Again, the effects of citalopram were strongest in participants high in empathy.

Note that the effects of serotonin manipulations on moral judgment and behavior cannot be explained by changes in mood. In all studies, we were careful to measure mood at baseline and postmanipulation. We did not observe any reliable effects of serotonin manipulations on subjects' mood, and subjects were blind to the treatment they received. Despite this, we found significant effects of serotonin manipulations on moral judgment and behavior.

In light of these behavioral findings, we became interested in better understanding the motivational processes that give rise to these effects. Although costly punishment is usually framed in terms of fairness and reciprocity, it is worth considering that punishing a norm violation requires harming the norm violator—whether economically (as in ultimatum games), emotionally (as in scolding or gossiping), or even physically (as in corporal punishment). Decisions *not* to punish (which were enhanced by citalopram) could therefore be driven by some form of harm aversion.

acprof-9780199357666.indd 239 5/24/2016 5:03:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

There is indeed converging evidence that similar sorts of motivational processes could influence both moral judgments in hypothetical scenarios, and costly punishment decisions. In our studies, citalopram made people less likely to reject unfair offers and less likely to endorse harmful actions in personal moral scenarios. Meanwhile, patients with damage to the ventromedial prefrontal cortex show the opposite pattern—they reject *more* unfair offers (Koenigs and Tranel 2007) and are *more* likely to endorse harmful actions in personal moral scenarios (Koenigs et al. 2007). Finally, psychopaths and healthy people with psychopathic traits show a behavioral pattern similar to that of the ventromedial patients—they also reject more unfair offers (Koenigs et al. 2010) and are more likely to endorse personal harms (Bartels and Pizarro 2011; Koenigs et al. 2011). Together these studies support the notion that costly punishment and moral judgment are similarly governed by a medial fronto-striatal circuit concerned with "moral sentiments" like harm aversion.

In our most recent study, we investigated how depleting serotonin shapes the neural circuitry of costly punishment behavior (Crockett et al. 2013). One advantage of examining brain activations is that they can provide clues about the motivational processes that drive behavior. Specifically, we tested whether serotonin influences punishment behavior through affecting a more "altruistic" motive to enforce fairness norms, versus a more "antisocial" motive for revenge. These competing explanations generate different predictions about the effects of serotonin manipulations on brain activity.

Previous work has identified brain regions associated with motives for revenge. In one classic neuroimaging study, participants first played a series of trust games with two confederates. One confederate played fairly, and the other played unfairly. Next, participants watched as the fair and unfair confederates received electric shocks.

When participants passively observed the unfair players receive shocks, they showed activation in the ventral striatum, and this activation was correlated with self-reported desire for revenge (Singer et al. 2006). Other studies have reported ventral striatum activation when sports fans observed fans of their rival team suffer (Cikara et al. 2011; Hein et al. 2010). These findings suggest that watching a rival suffer has motivational value. There is also evidence that actively delivering punishment has motivational value. When people punish unfair behavior, they show increased activity in the dorsal striatum (de Quervain et al. 2004; Strobel et al. 2011).

Together these findings suggest that if depleting serotonin increases costly punishment behavior by enhancing the motivational value of punishment, we should see that serotonin depletion increases activity in the striatum when people punish. This is precisely what we observed. Serotonin depletion increased activity in the dorsal striatum during punishment, and individual differences in the neural effects of depletion were correlated with individual differences in the behavioral effects of depletion.

An alternative (though not mutually exclusive) possibility is that serotonin depletion increases punishment by enhancing altruistic motives to enforce fairness norms. Previous studies have shown that fair outcomes activate value-processing regions, including the ventral striatum and medial prefrontal cortex (Tabibnia et al. 2008; Tricomi

acprof-9780199357666.indd 240 5/24/2016 5:03:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

et al. 2010). Thus, if serotonin depletion amplifies preferences for fairness, we should expect that serotonin depletion would increase fairness-related responses in these regions. Instead, we observed the opposite effect: serotonin depletion blunted ventral striatal responses to fair outcomes, suggesting that depletion made fairness goals less salient.

We observed a similar, albeit weaker, effect in the medial prefrontal cortex. The cluster showing the interaction was located in anterior MPFC (Brodmann area 10), a region that has been implicated in the representation of abstract and social rewards (Haber and Knutson 2009; Rademacher et al. 2010), mentalizing (Amodio and Frith 2006), and moral cognition (Moll et al. 2011). It may therefore be involved in assessing the long-term benefits arising from mutual cooperation (Rilling and Sanfey 2011). Supporting this view, patients with damage to the MPFC show impaired prosocial emotions (Moll et al. 2011) and are less likely to reciprocate trust (Krajbich et al. 2009). Thus, serotonin may normally increase cooperation (and inhibit retaliation, which can damage social relationships) by enhancing the subjective value of the distant rewards associated with repeated cooperation. This would be consistent with our previous finding that serotonin depletion increased impulsive choice in tandem with its effects on costly punishment (Crockett et al. 2010) and could also explain the observation that serotonin depletion reduces cooperation in the repeated prisoner's dilemma (Wood et al. 2006), while enhancing serotonin function has the opposite effect (Tse and Bond 2002). An obvious next step would be to test whether enhancing serotonin function promotes positive reciprocity by boosting ventral striatal and MPFC responses to mutual cooperation.

The pattern of results we observed suggests that serotonin depletion decreased the subjective value of social exchange while simultaneously enhancing the subjective value of retaliation via punishment. More broadly, our observations are compatible with the hypothesis that serotonin regulates social preferences, where enhancing (versus impairing) serotonin function leads individuals to value the outcomes of others more positively (versus negatively; Siegel and Crockett 2013). This hypothesis unifies a range of empirical data describing positive associations between serotonin function and prosocial behavior on the one hand, and negative associations between serotonin function and antisocial behavior on the other hand. To test this hypothesis, future work will combine more precise computational models of social preferences with pharmacological manipulations and neuroimaging.

Studies demonstrating effects of neurochemical manipulations on morality may have potential normative implications. We have shown that moral judgments and decisions are sensitive to fluctuations in brain chemistry; others have discovered similar effects of stress (Starcke et al. 2012; Youssef et al. 2012) and even time of day (Danziger et al. 2011). Together these studies have shown that moral judgments respond to factors that are clearly nonnormative. However, certain kinds of moral judgments appear to be more sensitive to nonnormative factors than others. In particular, moral judgments about physical harms (e.g., Greene's "personal" scenarios in which harms are emotionally salient) seem to be more susceptible to the influence of

acprof-9780199357666.indd 241 5/24/2016 5:03:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

neuromodulators (Crockett et al. 2010; Terbeck et al. 2013) and acute stress (Starcke et al. 2012; Youssef et al. 2012). This raises the question of whether we should generally be more skeptical of judgments in such "personal" cases, compared with judgments in "impersonal" cases involving more indirect harms, which appear to be less susceptible to nonnormative influences.

In addition, some individuals may be more vulnerable to influence by nonnormative factors than others; for instance, we have shown that the effects of manipulating serotonin on moral judgments interact with individual differences in empathy (Crockett et al. 2010). This is hardly surprising, given that genotypic variability affects how an individual's nervous system reacts to perturbations in neurotransmitter levels (Rogers 2011). But this individual variability raises important ethical issues, particularly in the realm of applied ethics: if certain individuals are less influenced by nonnormative factors than others, by way of their genetic makeup or some other stable underlying trait, should the judgments of these individuals be privileged above those who are more influenced by such factors? Take, for example, a recent study showing that judges are more likely to grant parole if the parole hearing takes place immediately after a snack break (Danziger et al. 2011). Suppose we discover a biomarker that reliably detects an individual's susceptibility to the influence of snack breaks on judgments. Should such evidence be used to bar certain individuals from making judicial decisions? Obviously we are a long way off from such a test, but the fact that different people can be more or less sensitive to nonnormative factors begs the question of whether we should be more skeptical of the former's judgments compared to those of the latter.

The evidence reviewed here shows that moral judgments are not fixed, but malleable and contingent on neuromodulator levels, stress, and so on. Many of us would rather believe that such judgments are not so dependent on the vagaries of neurochemistry. This raises the question whether there exists any "neutral" physiological state from which one can generate reliable ethical principles (Crockett and Rini 2015). Although the idea of finding such a state is certainly attractive, in practice it is far from straightforward. Neuromodulator levels are constantly in flux and extremely difficult to measure in humans; for this reason there is not even a scientific consensus on what constitutes a "healthy" level of serotonin, dopamine, and so on, if such a state even exists. Moreover, determining which state is the "neutral" one, for the purpose of establishing a neutral ethical baseline, is itself a sort of value judgment that would be subject to influence by the same neurochemical factors—an inescapable problem. In the coming decade it will be important to systematically investigate these questions and debate their significance for morality.

References

Amodio, D. M., and C. D. Frith. 2006. "Meeting of Minds: The Medial Frontal Cortex and Social Cognition." *Nature Reviews Neuroscience* 7, no. 4: 268–277. doi:10.1038/nrn1884.

acprof-9780199357666.indd 242 5/24/2016 5:03:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

Bartels, D. M., and D. A. Pizarro. 2011. "The Mismeasure of Morals: Antisocial Personality Traits Predict Utilitarian Responses to Moral Dilemmas." *Cognition* 121, no. 1: 154–161. doi:10.1016/j.cognition.2011.05.010.

Brosnan, S. F., and F. B. M. de Waal. 2012. "Fairness in Animals: Where to from Here?" *Social Justice Research* 25, no. 3: 336–351. doi:10.1007/s11211-012-0165-8.

Cikara, M., M. M. Botvinick, and S. T. Fiske. 2011. "Us versus Them: Social Identity Shapes Neural Responses to Intergroup Competition and Harm." *Psychological Science* 22, no. 3: 306–313. doi:10.1177/0956797610397667.

Crockett, M. J., A. Apergis-Schoute, B. Herrmann, M. D. Lieberman, U. Muller, T. W. Robbins, and L. Clark. 2013. "Serotonin Modulates Striatal Responses to Fairness and Retaliation in Humans." *Journal of Neuroscience* 33, no. 8: 3505–3513. doi:10.1523/JNEUROSCI.2761-12.2013.

Crockett, M. J., L. Clark, M. D. Hauser, and T. W. Robbins. 2010. "From the Cover: Serotonin Selectively Influences Moral Judgment and Behavior through Effects on Harm Aversion." *Proceedings of the National Academy of Sciences* 107, no. 40: 17433–17438. doi:10.1073/pnas.1009396107.

Crockett, M. J., L. Clark, M. D. Lieberman, G. Tabibnia, and T. W. Robbins. 2010. "Impulsive Choice and Altruistic Punishment Are Correlated and Increase in Tandem with Serotonin Depletion." *Emotion* 10, no. 6: 855–862. doi:10.1037/a0019861.

Crockett, M. J., L. Clark, G. Tabibnia, M. D. Lieberman, and T. W. Robbins. 2008. "Serotonin Modulates Behavioral Reactions to Unfairness." *Science* 320, no. 5884: 1739–1739. doi:10.1126/science.1155577.

Crockett, M. J., and R. A. Rini. 2015. "Neuromodulators and the (In)stability of Moral Cognition." In *The Moral Brain*, edited by J. Decety and T. Wheatley, 221–235. Cambridge, MA: MIT Press,

Cushman, F., K. Gray, A. Gaffey, and W. B. Mendes. 2012. "Simulating Murder: The Aversion to Harmful Action." *Emotion* 12, no. 1: 2.

Danziger, S., J. Levav, and L. Avnaim-Pesso. 2011. "Extraneous Factors in Judicial Decisions." *Proceedings of the National Academy of Sciences* 108, no. 17: 6889–6892.

Davis, M. H. 1983. "Measuring Individual Differences in Empathy: Evidence for a Multidimensional Approach." *Journal of Personality and Social Psychology* 44, no. 1: 113–126. doi:10.1037/0022-3514.44.1.113.

De Quervain, D. J.-F., U. Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder, A. Buck, and E. Fehr. 2004. "The Neural Basis of Altruistic Punishment." *Science* 305, no. 5688: 1254–1258. doi:10.1126/science.1100735.

Fehr, E., and U. Fischbacher. 2003. "The Nature of Human Altruism." *Nature* 425, no. 6960: 785–791. doi:10.1038/nature02043.

Güth, W., R. Schmittberger, and B. Schwarze. 1982. "An Experimental Analysis of Ultimatum Bargaining." *Journal of Economic Behavior and Organization* 3, no. 4: 367–388. doi:10.1016/0167-2681(82)90011-7.

Haber, S. N., and B. Knutson. 2009. "The Reward Circuit: Linking Primate Anatomy and Human Imaging." *Neuropsychopharmacology* 35, no. 1: 4–26. doi:10.1038/npp.2009.129.

Hein, G., G. Silani, K. Preuschoff, C. D. Batson, and T. Singer. 2010. "Neural Responses to Ingroup and Outgroup Members' Suffering Predict Individual Differences in Costly Helping." *Neuron* 68, no. 1: 149–160. doi:10.1016/j.neuron.2010.09.003.

acprof-9780199357666.indd 243 5/24/2016 5:03:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

Higley, J. D., S. T. King Jr., M. F. Hasert, M. Champoux, S. J. Suomi, and M. Linnoila. 1996. "Stability of Interindividual Differences in Serotonin Function and Its Relationship to Severe Aggression and Competent Social Behavior in Rhesus Macaque Females." *Neuropsychopharmacology* 14, no. 1: 67–76. doi:10.1016/S0893-133X(96)80060-1.

Higley, J. D., and M. Linnoila. 1997. "Low Central Nervous System Serotonergic Activity Is Traitlike and Correlates with Impulsive Behavior." *Annals of the New York Academy of Sciences* 836, no. 1: 39–56. doi:10.1111/j.1749-6632.1997.tb52354.x

Higley, J. D., P. T. Mehlman, R. E. Poland, D. M. Taub, J. Vickers, S. J. Suomi, and M. Linnoila. 1996. "CSF Testosterone and 5-HIAA Correlate with Different Types of Aggressive Behaviors." *Biological Psychiatry* 40, no. 11: 1067–1082. doi:10.1016/S0006-3223(95)00675-3.

Insel, T. R., and J. T. Winslow. 1998. "Serotonin and Neuropeptides in Affiliative Behaviors." *Biological Psychiatry* 44, no. 3: 207–219. doi:10.1016/S0006-3223(98)00094-8.

Knutson, B., O. M. Wolkowitz, S. W. Cole, T. Chan, E. A. Moore, R. C. Johnson, . . . V. I. Reus. 1998. "Selective Alteration of Personality and Social Behavior by Serotonergic Intervention." *American Journal of Psychiatry* 155, no. 3: 373–379.

Koenigs, M., M. Kruepke, and J. P. Newman. 2010. "Economic Decision-Making in Psychopathy: A Comparison with Ventromedial Prefrontal Lesion Patients." *Neuropsychologia* 48, no. 7: 2198–2204. doi:10.1016/j.neuropsychologia.2010.04.012.

Koenigs, M., M. Kruepke, J. Zeier, and J. P. Newman. 2011. "Utilitarian Moral Judgment in Psychopathy." *Social Cognitive and Affective Neuroscience* 7, no. 6: 708–714. doi:10.1093/scan/nsr048.

Koenigs, M., and D. Tranel. 2007. "Irrational Economic Decision-Making after Ventromedial Prefrontal Damage: Evidence from the Ultimatum Game." *Journal of Neuroscience* 27, no. 4: 951–956. doi:10.1523/JNEUROSCI.4606-06.2007.

Koenigs, M., L. Young, R. Adolphs, D. Tranel, F. Cushman, M. Hauser, and A. Damasio. 2007. "Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements." *Nature* 446, no. 7138: 908–911. doi:10.1038/nature05631.

Krajbich, I., R. Adolphs, D. Tranel, N. L. Denburg, and C. F. Camerer. 2009. "Economic Games Quantify Diminished Sense of Guilt in Patients with Damage to the Prefrontal Cortex." *Journal of Neuroscience* 29, no. 7: 2188–2192. doi:10.1523/JNEUROSCI.5086-08.2009.

Krakowski, M. 2003. "Violence and Serotonin: Influence of Impulse Control, Affect Regulation, and Social Functioning." *Journal of Neuropsychiatry and Clinical Neurosciences* 15, no. 3: 294–305. doi:10.1176/appi.neuropsych.15.3.294.

Linnoila, M., M. Virkkunen, M. Scheinin, A. Nuutila, R. Rimon, and F. K. Goodwin. 1983. "Low Cerebrospinal Fluid 5-hydroxyindoleacetic Acid Concentration Differentiates Impulsive from Nonimpulsive Violent Behavior." *Life Sciences* 33, no. 26: 2609–2614. doi:10.1016/0024-3205(83)90344-2.

Moll, J., R. Zahn, R. de Oliveira-Souza, I. E. Bramati, F. Krueger, B. Tura, . . . J. Grafman. 2011. "Impairment of Prosocial Sentiments Is Associated with Frontopolar and Septal Damage in Frontotemporal Dementia." *NeuroImage* 54, no. 2: 1735–1742. doi:10.1016/j.neuroimage.2010.08.026.

Moskowitz, D. S., G. Pinard, D. C. Zuroff, L. Annable, and S. N. Young. 2001. "The Effect of Tryptophan on Social Interaction in Everyday Life: A Placebo-Controlled Study." *Neuropsychopharmacology* 25, no. 2: 277–289.

acprof-9780199357666.indd 244 5/24/2016 5:03:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

Proctor, D., R. A. Williamson, F. B. M. de Waal, and S. F. Brosnan. 2013. "Chimpanzees Play the Ultimatum Game." *Proceedings of the National Academy of Sciences* 110(6: 2070–2075. doi:10.1073/pnas.1220806110.

Rademacher, L., S. Krach, G. Kohls, A. Irmak, G. Gründer, and K. N. Spreckelmeyer. 2010. "Dissociation of Neural Networks for Anticipation and Consumption of Monetary and Social Rewards." *NeuroImage* 49, no. 4: 3276–3285. doi:10.1016/j.neuroimage.2009.10.089.

Raleigh, M. J., McM. T. Guire, G. L. Brammer, D. B. Pollack, and A. Yuwiler. 1991. "Serotonergic Mechanisms Promote Dominance Acquisition in Adult Male Vervet Monkeys." *Brain Research* 559, no. 2: 181–190. doi:10.1016/0006-8993(91)90001-C.

Rilling, J. K., and A. G. Sanfey. 2011. "The Neuroscience of Social Decision-Making." *Annual Review of Psychology* 62, no. 1: 23–48. doi:10.1146/annurev.psych.121208.131647.

Rogers, R. D. 2011. "The Roles of Dopamine and Serotonin in Decision Making: Evidence from Pharmacological Experiments in Humans." *Neuropsychopharmacology* 36, no. 1: 114–132. doi:10.1038/npp.2010.165.

Siegel, J. Z., and M. J. Crockett. 2013. "How Serotonin Shapes Moral Judgment and Behavior." *Annals of the New York Academy of Sciences* 1299, no. 1: 42–51.

Singer, T., B. Seymour, O'J. P. Doherty, K. E. Stephan, R. J. Dolan, and C. D. Frith. 2006. "Empathic Neural Responses Are Modulated by the Perceived Fairness of Others." *Nature* 439, no. 7075: 466–469. doi:10.1038/nature04271.

Starcke, K., A.-C. Ludwig, and M. Brand. 2012. "Anticipatory Stress Interferes with Utilitarian Moral Judgment." *Judgment and Decision Making* 7, no. 1: 61–68.

Stevens, J. R., F. A. Cushman, and M. D. Hauser. 2005. "Evolving the Psychological Mechanisms for Cooperation." *Annual Review of Ecology, Evolution, and Systematics* 36: 499–518.

Strobel, A., J. Zimmermann, A. Schmitz, M. Reuter, S. Lis, S. Windmann, and P. Kirsch. 2011. "Beyond Revenge: Neural and Genetic Bases of Altruistic Punishment." *NeuroImage* 54, no. 1: 671–680. doi:10.1016/j.neuroimage.2010.07.051.

Tabibnia, G., A. B. Satpute, and M. D. Lieberman. 2008. "The Sunny Side of Fairness: Preference for Fairness Activates Reward Circuitry (and Disregarding Unfairness Activates Self-Control Circuitry)." *Psychological Science* 19, no. 4: 339–347. doi:10.1111/j.1467-9280.2008.02091.x.

Terbeck, S., G. Kahane, McS. Tavish, J. Savulescu, N. Levy, M. Hewstone, and P. J. Cowen. 2013. "Beta Adrenergic Blockade Reduces Utilitarian Judgement." *Biological Psychology* 92, no. 2: 323–328. doi:10.1016/j.biopsycho.2012.09.005.

Tricomi, E., A. Rangel, C. F. Camerer, and J. P. O'Doherty. 2010. "Neural Evidence for Inequality-Averse Social Preferences." *Nature* 463, no. 7284: 1089–1091. doi:10.1038/nature08785.

Tse, W., and A. Bond. 2002. "Serotonergic Intervention Affects Both Social Dominance and Affiliative Behaviour." *Psychopharmacology* 161, no. 3: 324–330. doi:10.1007/s00213-002-1049-7.

Wood, R. M., J. K. Rilling, A. G. Sanfey, Z. Bhagwagar, and R. D. Rogers. 2006. "Effects of Tryptophan Depletion on the Performance of an Iterated Prisoner's Dilemma Game in Healthy Adults." *Neuropsychopharmacology* 31, no. 5: 1075–1084.

Youssef, F. F., K. Dookeeram, V. Basdeo, E. Francis, M. Doman, D. Mamed, . . . G. Legall. 2012. "Stress Alters Personal Moral Decision Making." *Psychoneuroendocrinology* 37, no. 4: 491–498. doi:10.1016/j.psyneuen.2011.07.017.

acprof-9780199357666.indd 245 5/24/2016 5:03:07 PM

> Hendrix and Pearson pulled Romona into a dilapidated house. They took her down to a basement where they tortured her, beat her up, and raped her for four days. They took off her clothes and put her in chains. They sodomized her. They tried to saw off her hands and feet. They beat her face with a hammer when she screamed. They cut the webbing between her fingers. They burned her face with a lit cigarette. Then they kept her under a tarp. When a friend stopped by, they brought him to basement where Ramona was being kept. As they were smoking pot on the couch, Hendrix said out loud: "Say hi, bitch." The friend asked: "Who y'all talking to?" Hendrix and Pearson pulled up a tarp on the floor to show their proud work. They made Ramona recount everything they had done to her to their friend. Then they dropped the tarp back to the floor. Romona was eventually bludgeoned to death with a hammer and a barbell. Hendrix and Pearson were later reported to say their interaction with Ramona was "fun."
>
> —A compilation from Gardiner 2008; Katz 2006; Ginsberg 2006.

> Every murder raises terrible questions that no trial, no law, no punishment can answer. What forces make it possible for one human being to take the life of another? . . . Scholars ranging from theologians and psychologists to evolutionary biologists have offered theories about murder—theories of evil, theories of disease, theories of disposition—but the analytical burden placed on any general discussion of murder, freighted, as it is, with atrocity, is nearly unbearable. Nothing suffices, or can.
>
> —Lepore 2009

11

Of Mice and Men

THE INFLUENCE OF RODENT MODELS OF EMPATHY ON HUMAN MODELS OF HARM PREVENTION

Jana Schaich Borg

THE ACTS HENDRIX and Pearson committed in the story above were morally very wrong. For many, it seems impossible to even imagine doing such horrible things to another being. This raises fundamental questions. One question discussed in this volume

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

is why we *judge* or *believe* it to be morally wrong to cause harm to others. An equally essential question, however, is why some people (hopefully most people) *act* morally while other people, like Hendrix and Pearson, *act* so immorally. In other words, what leads to moral or immoral action, as opposed to its related cousins, moral judgment and moral belief?

Moral judgment and moral action must be differentiated; what people think or report is morally wrong or obligatory does not always correspond with what people do (Blasi 1980, 1983), or even what they say they will do (Schaich Borg 2006). The past decade of neuroscience research on morality has taught us a lot about the neural systems involved in making moral judgments. The next decade of neuroscience research needs to focus on delineating the brain systems involved in morally relevant action. In particular, it should focus on a particularly devastating type of morally relevant action: unjustified physical violence.

A major reason we do not yet know how to prevent unjustified violence is that biomedical studies of human antisocial behavior are rife with ethical and practical challenges (Ward and Willis 2010). To start, one of the most efficient ways to identify study participants who are likely to commit violence in the future is to identify people who have committed violence in the past; thus, prisoners are an informative population to study. However, federal rules severely limit the amount of biomedical research that can be carried out in prison populations, especially if such research does not benefit the research participants themselves (to protect prisoners from coercion) (Osganian 2008). Research that ultimately benefits incarcerated research participants is often restricted as well, because some (including granting agencies) feel that individuals who have already committed immoral actions should be punished without an opportunity for treatment (Taylor 2011; Beck 2010; Eastman 1999). Such ethical concerns motivate some researchers to study antisocial behavior outside of the prison, but antisocial research outside the prison is plagued by its own difficulties. People who consistently commit violent or immoral action do not usually seek treatment or research participation voluntarily, and when they do, clinical service providers may be obligated to turn them away to avoid the inherent physical risks of working with such a dangerous population (Howells and Day 2007). Finally, there is concern that studying the biological basis of violent behavior will undermine notions of responsibility, culpability, and free will, and that identifying or treating potential offenders can lead to discrimination or stigmatization (Beck 2010; Pickersgill 2011; Pustilnik 2009). Perhaps these challenges explain why so few attempts have been made to develop or test biomedical treatments for violent or immoral behavior (Gibbon et al. 2010; Khalifa 2010), and why no biomedical treatments currently exist (Salekin et al. 2010).

Ethical concerns surrounding the study of human immoral and antisocial action are very important, but they must be balanced with our obligations to protect the rights of our fellow members of society, including the right to live a life without unjustified physical or emotional harm. We now know that the most persistent 5 percent

acprof-9780199357666.indd 247 5/24/2016 5:03:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

to 8 percent of offenders are responsible for between 50 percent and 70 percent of documented violent crimes (Farrington et al. 1986; Moffitt 1993; Vaughn et al. 2011). We also now know that a considerable amount of variance in immoral and antisocial behavior can be accounted for by one's genetics (Ferguson 2010; Gunter et al. 2010). This raises the strong possibility that there is something genetically and biologically different about the most persistent violent offenders that, when fully understood, may shed some light on how to change their violent, immoral behavior. Given the disproportionate amount of damage this small population of destructive individuals causes, a change in these individuals' behavior could have a dramatic impact on how much violence is committed worldwide. If it is feasible to develop treatments for the biological causes of violent behavior in an ethically appropriate way, we should try to do so.

One critical advance from the past decade that has made the development of treatments for unjustified violent behavior a realistic possibility is the development of rodent models of empathy (which will be renamed "negative intersubjectivity" in later parts of this chapter). As discussed in sections 11.1 and 11.2 of this chapter, empathy is a phenomenon that contributes to violence aversion, and we now know some of the brain regions that likely contribute to human empathy. However, these advances, by themselves, have not been sufficient to make significant progress toward developing methods for treating biological causes of violent behavior. Fortunately, rodent models of empathy have become available at the same time as field-changing technologies that allow us to dissect rodent neural circuits with unprecedented accuracy and efficiency. Section 11.3 will describe these rodent models and explain how they provide powerful and realistic opportunities for neuroscience to help prevent unjustified violence in more mechanistic ways than were previously thought possible. Sections 11.1–11.3 will collectively provide evidence that the most efficient way to understand moral action and prevent immoral action is to pursue rodent research and human research in parallel.

11.1. "Negative Intersubjectivity"?

"Negative intersubjectivity" will be defined as *feeling negative affect when another feels negative affect*. In the story recounted earlier, Hendrix and Pearson certainly did not have negative intersubjectivity because they did not feel negative when they saw Ramona's wounds or heard her screaming in pain. Instead, they were entertained by, and eventually bored by, her distress. Ramona would be alive and unharmed if Hendrix and Pearson had felt aversion to Ramona's distress cues. This illustrates a fundamental principle: if we could learn how to induce negative intersubjectivity, or aversion to actions that cause pain and distress in others, we could prevent the type of violent behavior performed by Hendrix and Pearson.

acprof-9780199357666.indd 248 5/24/2016 5:03:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

11.1.1. NEGATIVE INTERSUBJECTIVITY'S RELATIONSHIP TO "EMPATHY"

"Empathy" is one of the first words that comes to mind to describe the phenomenon of feeling negative in response to another person's distress, and any discussion of negative intersubjectivity would be remiss to ignore the term. However, despite this natural association, the term "empathy" is not very useful because (despite much effort) there remains little or no consensus about what it means in either a scientific or colloquial context. As some researchers in the field wrote, "Psychologists are known for using terms loosely, but in our use of empathy we have outdone ourselves" (Batson et al. 1987, 19). Providing support for the spirit of that exclamation, all of the following subprocesses have been incorporated into some published definitions of empathy:

(1) **Motor mimicry:** adopting the posture, position, or facial expression of an observed other. *Example*: grabbing your own elbow when you see a friend fall and grab his elbow.

(2) **Personal distress:** a self-oriented reactive emotion in response to the perception or recognition of another's negative emotion or situation. Personal distress does not need to be congruent with the emotion or situation of the observed other, but it does need to describe a negative personal state in the observer. *Example*: feeling anxious about your own job upon hearing that your friend was laid off.

(3) **Emotional contagion:** feeling the same emotion as the emotion perceived or recognized in another. Most theorists specify that emotional contagion does not require observers to be able to distinguish between themselves and the person they are observing. *Example*: a newborn infant's reactive cry to the distress cry of another infant.

(4) **Affective intersubjectivity:** feeling emotions that are "more congruent with another's situation than with one's own situation" (Hoffman 2000, 30). This affective reaction does not necessarily require an accompanying state of understanding. Unlike emotional contagion, affective intersubjectivity does not require that the subject and target feel the exact same emotions, although it does tend to require that the subject and target feel emotions of the same valence. *Example*: feeling sad when you hear that your friend's parent died.

(5) **Theory of mind:** the ability to understand another's perspective (Wellman 2010). This can be broken down into (*a*) "perceptual perspective-taking," the ability to understand and report what another is likely to perceive, (*b*) "emotional perspective-taking," the ability to understand and report what another is likely to feel (sometimes this is also referred to as "empathic accuracy"), and (*c*) "cognitive perspective-taking," the ability to understand and report what another is likely to think. *Examples*: Knowing that someone sitting with his back to the wall is not likely to see a sign posted on the wall above his head

acprof-9780199357666.indd 249 5/24/2016 5:03:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

(perceptual perspective-taking); knowing that somebody will be angry if he is not reimbursed the money he lent (emotional perspective-taking); knowing that somebody thinks the meeting is scheduled for Monday because he was not told that it had been rescheduled for Tuesday (cognitive perspective-taking).

(6) **Sympathy:** caring for another's well-being as it relates to his emotion or situation (Darwall 1998). This term usually implies that an observer is motivated to alleviate another's suffering, or has judged that another's suffering should be alleviated (Wispé 1986; Eisenberg and Strayer 1987). *Example*: Wanting to give money to a homeless person on the street because you believe he's hungry.

(7) **Moral judgment or motivation:** principles of right or wrong that motivate and structure behaviors within social environments. It is under debate whether these principles exist because humans have innate emotions toward their perceptions of how other people feel or because humans have unemotional and impartial cognitive processes that allow them to deduce relevant social and behavioral rules (Haid 2004, 2001; Greene et al. 2001). *Example*: Most people believe that it is wrong to physically beat children and consequently may be motivated to prevent other people from physically beating children.

(8) **Altruism:** the intention of increasing another's welfare, even at the expense of harm to oneself. Colloquial conceptions of empathy often incorporate altruistic intentions, but empathy and altruism are usually differentiated in the academic literature (Batson et al. 2011, 417). *Example*: Giving your food to a homeless person, even though you are very hungry yourself, because you want to end the homeless person's hunger.

Using different combinations of these processes, empathy has been defined with various descriptions including:

> When we have (a) an affective state, (b) which is isomorphic to another person's affective state, (c) which was elicited by observing or imagining another person's affective state, and (d) when we know that the other person's affective state is the source of our own affective state (de Vignemont and Singer 2006, 435). [This definition includes some aspects of emotional contagion, may include either personal distress or theory of mind, and has an extra knowledge component that requires the observer to know the other person's state is the source of one's own state.]

> The capacity to a) be affected by and share the emotional state of another, b) assess the reasons for the other's state, and c) identify with the other, adopting his or her perspective. This definition extends beyond what exists in many animals, but I employ the term "empathy" even if only the first criterion is met as I believe all of these elements are evolutionarily connected (de Waal 2008, 281). [This definition can include any of the subprocesses listed above.]

acprof-9780199357666.indd 250 5/24/2016 5:03:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

> Empathy is defined as an affective state that stems from the apprehension of another's emotional state or condition, and that is congruent with it. Thus, empathy can include emotional matching and the vicarious experiencing of a range of emotions consistent with those of others (Eisenberg and Miller 1987, 91). [This definition specifies that empathy cannot only be theory of mind because it requires affect, and suggests that empathy is usually differentiated from sympathy.]
>
> "The intellectual or imaginative apprehension of another's condition or state of mind [which] is central for understanding a broad range of social phenomena including, in particular, moral development. Within this latter context, an empathic disposition can be regarded as the capacity to adopt a broad moral perspective, that is, to take "the moral point of view" (Hogan 1969, 307). [This definition says empathy requires theory of mind and moral judgment.]
>
> Empathy is the ability to experience and understand what others feel without confusion between oneself and others (Decety and Lamm 2006, 1146). [This definition specifies that empathy is not emotional contagion and likely not personal distress, since personal distress is often elicited via confusion between oneself and others.]

11.1.2. "Negative Intersubjectivity" is Related to, but Not Identical to, Definitions of "Empathy"

The phenomenon critical for this chapter is *feeling negative affect when another feels negative affect*. In order to discuss this phenomenon without bringing conflicting definitions of "empathy" to bear, the term "negative intersubjectivity" will be used instead of "empathy" to refer to this phenomenon of interest. The word "intersubjectivity" in this phrase refers to subjective experiences initiated by interactions with others, and the word "negative" specifies that those subjective experiences must have a negative valence. It is therefore similar to "affective intersubjectivity" (described earlier), but does not include subjective experiences with a positive valence. It departs from many colloquial definitions of empathy because it does not require the recognition of moral content or altruistic intentions. Negative intersubjectivity, said simply, is disliking (for any reason, selfish or not) when another individual feels bad. This term represents the phenomenon relevant to antisocial and prosocial action better than other terms currently available.

To further clarify how this phenomenon relates to other academic definitions of empathy, negative intersubjectivity as it is being operationalized here *does* require negative affect in both an observer and a receiver. Since negative intersubjectivity requires affect, neither (1) motor mimicry nor (5) theory of mind is sufficient on its own to meet the criteria of negative intersubjectivity, and (7) moral judgment would only be sufficient if it involved negative affect. Beyond that, however, negative intersubjectivity may or may not involve mimicry systems, personal distress, identical emotions between the observer and receiver, or cognitive apprehension of what is happening in either a receiver or

acprof-9780199357666.indd 251 5/24/2016 5:03:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

oneself. Most importantly, again, negative intersubjectivity *does not require* motivation to help the individual in distress or feelings or judgments about what is right or wrong.

Although the term "empathy" will be avoided whenever possible, many studies relevant to the goals of this chapter utilize published measures of "empathy" or "callousness." These measures ask participants to rate items such as "I often have tender, concerned feelings for people less fortunate than me" (taken from the Interpersonal Reactivity Index) (Davis 1980), and are clearly related to negative intersubjectivity. As a consequence, the only times the word "empathy" will be used in this chapter will be when describing results of these studies that explicitly report empathy measurements. The following points must be kept in mind when considering these studies. First, as already discussed, empathy self-report tools incorporate processes (1)–(7) described above to different degrees, so it is difficult to quantify how much self-reported "empathy" scores uniquely represent negative intersubjectivity as opposed to other phenomena like theory of mind (Lovett and Sheffield 2007). Second, it is unclear how well such self-report questions map onto negative intersubjectivity, given that empathy self-reports can be poorly correlated with nonverbal measures of responses to others' distress, such as autonomic outputs or facial expressions (Eisenberg and Fabes 1990). Third, it is uncertain how honest people are when responding to these questionnaires. Nonetheless, studies using empathy self-report scales provide important evidence corroborating that negative intersubjectivity consistently correlates with human moral behavior, and therefore deserve attention in section 11.2, which summarizes what is known about this critical link.

11.2. Human Negative Intersubjectivity and Moral Behavior

11.2.1. Negative Intersubjectivity Predicts Antisocial and Prosocial Behavior

One of the most consistent lines of evidence linking negative intersubjectivity to behavior comes from studies referencing the "affective" components of trait empathy self-reports. Affective components attempt to capture one's "emotional" responses to others' experiences, and therefore come close to indexing negative intersubjectivity as I have defined it. Indeed, scores on the affective components of self-reported empathy correlate with antisocial behavior, or behavior that imposes harm on others (Jolliffe and Murray 2012; Björkqvist et al. 2000; Jolliffe and Farrington 2004; LeSure-Lester 2000). Similar evidence comes from research with "callous" personality traits in children and adolescents. Callousness is when an individual does not feel guilt, appears more concerned about the effects of his actions on himself than on others (even when his actions result in substantial harm to others), and generally disregards others' feelings (American Psychiatric Association's Proposed Draft Revisions to DSM Disorders and Criteria). Callous traits often cluster with "unemotional" traits (found in individuals who have low levels of affect overall), so the two traits are usually combined and called "callous-unemotional" (CU) traits. Consistent with the correlations found with self-reports of trait empathy, CU traits are good predictors of the persistence and intensity

acprof-9780199357666.indd 252 5/24/2016 5:03:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

of violence within juvenile and adolescent delinquents, and some evidence suggests that CU traits also predict offenders' resistance to treatment (Frick and White 2008; Hawes and Dadds 2005). In addition, persistent violent offenders are disproportionately likely to have high CU traits when they are adolescents compared to other offenders (Vaughn and DeLisi 2008). CU traits in adults can be indexed by items in the "affective factor" of the Psychopathy Checklist (Revised, PCL-R), a semistructured interview used to diagnose psychopathy. When present simultaneously with high impulsivity (another trait of psychopathy), scores on the affective factor of the PCL-R correlate with increased violence in adult psychopaths (Hare 1991, 2003). Conversely, the personality traits that lead to low scores on the affective items of the PCL-R may be protective against the influence of impulsive traits on violent behavior (Walsh and Kosson 2008). Finally, scores on the newly developed Inventory of Callous Unemotional Traits, a self-report assessment designed for both adolescents and young adults (Essau et al. 2006; Kimonis et al. 2008), correlate with number of arrests and charges for violent crime (Kahn et al. 2012). In fact, a study in young adults reported that callousness, on its own and separated from unemotional traits, correlated with repeated and violent offending even after controlling for prior criminal behavior and other well-established risk factors (Kahn et al. 2012). Thus, there is a well-documented correlation between negative intersubjectivity deficits and persistent antisocial behavior, suggesting that if we could determine how to appropriately modulate negative intersubjectivity, we might be able to decrease antisocial behavior.

Whereas antisocial behavior harms others, prosocial behavior is defined as voluntary behavior that benefits others, either intentionally or unintentionally (Eisenberg 1986). Consistent with the studies above suggesting that increased negative intersubjectivity correlates with reduced antisocial behavior, it is well documented that increased negative intersubjectivity (or reduced callousness) also correlates with increased helping and prosocial behavior (Batson et al. 2011, 417; Eisenberg and Miller 1987; Hoffman 2008; Eisenberg 2008; Trommsdorff et al. 2007; Batson and Shaw 1991; Batson and Oleson 1991; Barraza and Zak 2009). For example, self-reports of empathy when viewing videos of emotional scenes correlate positively with helping (preparing gifts to send people in the emotional movies) in children (Miller et al. 1996), and positively with more generous monetary offers in an ultimatum game in adults (Barraza and Zak 2009). Further, educational training programs aimed to enhance children's empathy and empathy-related skills have been reported to increase cooperation, helping, and generosity (Feshbach 1979; Feshbach and Feshbach 1982). Of note, many reports suggest there may be a nonlinear relationship such that although moderate levels of negative intersubjectivity correlate with increased helping as summarized above, extremely high levels of negative intersubjectivity may actually inhibit prosocial behavior to redirect efforts toward minimizing one's own distress (rather than others' distress) (Batson et al. 2011, 417; Eisenberg 2008; Feshbach and Feshbach 1982). Taking this into account, if we succeed in developing methods to regulate negative intersubjectivity appropriately, the literature suggests we may be able to increase prosocial behavior.

acprof-9780199357666.indd 253 5/24/2016 5:03:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

11.2.2. NEURAL BASIS OF HUMAN NEGATIVE INTERSUBJECTIVITY

Understanding the mechanisms that cause negative intersubjectivity helps provide insight into how to manipulate negative intersubjectivity. Toward this end, much effort has been dedicated to elucidating the neural basis of human negative intersubjectivity. This now fairly large body of work has been discussed in many excellent reviews (Shirtcliff et al. 2009; Walter 2012; Bernhardt and Singer 2012; Decety 2011), so rather than duplicate these reviews, the three primary themes of human negative intersubjectivity neuroscience research will be summarized below.

Overlapping, but not identical, brain regions are involved in receiving pain oneself and observing pain in another. Two recent meta-analyses of "empathy" functional magnetic resonance imaging (fMRI) studies converged on identifying the anterior cingulate (ACC) and the anterior insula (AI) as brain regions consistently activated both when receiving pain and when observing others in pain (Fan et al. 2011; Lamm et al. 2007). Further, Lamm et al. calculated that approximately 60 percent of studies measuring self-reported acute empathy in response to task stimuli, and many studies measuring general trait empathy using self-report questionnaires, show that self-reported empathy positively correlates with AI and ACC activity evoked when participants observe others in pain. Of note, similar correlations have been reported when observing people in negative affective states other than physical pain, such as during social exclusion (Meyer et al. 2013) or while smelling a disgusting odor (Jabbi et al. 2008; Wicker et al 2003). These observations have led to the "observing is feeling" hypothesis: observing pain in others activates the same neural circuitry as when you are in pain yourself, and activation of this neural circuitry mediates empathy.

The established relationships between AI and ACC activity and negative intersubjectivity are qualified by two considerations. First, despite the consistent overlap in the AI and ACC, there are also many differences in the neural representation of observing and receiving pain. Even within the insula, there is a topographical gradient localizing the sensory components of physical pain experience to the more posterior regions, and the affective components of physical pain experience to the more anterior regions (Fan et al. 2011; Lamm et al. 2007; Corradi-Dell'Acqua 2011). Vicarious pain only elicits activity in the more anterior regions. Second, the role of the AI in negative intersubjectivity might be more consistent than the role of the ACC. Whereas the AI seems to be preferentially active in response to all vicarious pain conditions, the ACC is only preferentially active when responding to others in pain under certain high-attention conditions, suggesting it might be more involved in the conscious voluntary control of behaviors relevant to negative intersubjectivity than negative intersubjectivity itself (Gu et al. 2010; Azevedo et al. 2013; Gu et al. 2012). Perhaps through this putative attentional role, the ACC is also hypothesized to help coordinate the recruitment of other task-relevant brain networks as well, like those involved in understanding another person's actions or mental states (Bruneau et al. 2012). In sum, there is compelling evidence that the AI and ACC are

acprof-9780199357666.indd 254 5/24/2016 5:03:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

involved in negative intersubjectivity in some way, but without methods for manipulating activity in brain regions located deep within the human brain, it is not known whether AI or ACC activity is either necessary or sufficient for executing moral actions.

Mirror neurons might have a role in responses to others' pain. Drawn from a completely different set of observations and principles, another popular hypothesis posits that "mirroring mechanisms" mediate negative intersubjectivity (Gallese et al. 2011; Spunt and Lieberman 2012). Mirror neurons are neurons in the parietal and premotor cortices of macaque monkeys that fire both when a monkey performs a particular action and when a monkey observes a human or another monkey intentionally perform the same action (Rizzolatti and Craighero 2004). Encouraged by the fMRI studies reviewed above that show affective brain areas are active both when humans observe and when they receive pain, proponents of mirror-neuron explanations of negative intersubjectivity have postulated that there may be "affective mirror neurons" in brain areas identified in those fMRI studies—including the ACC or AI—that mediate physical pain to oneself and aversion to pain in others (Iacoboni 2009). According to this view, witnessing somebody else in pain activates the same neurons in the ACC or AI that fire when you yourself are in physical pain, reinforcing the notion introduced earlier that at least to some extent, "empathetic" pain is a type of physical pain.

Two primary potential roles for mirror neurons in negative intersubjectivity have been discussed. First, we may automatically *physically* imitate the actions or facial expressions of others in distress using motor mirror neurons. Then, second, doing so may make it easier for us to automatically *affectively* imitate the feelings of others in distress using affective mirror neurons. In partial support, some studies report that the magnitude of motor "mirroring," defined by observable motor imitation or activity in one's own motor system when observing another person perform an action, correlates with self-reports of state empathy (Sonnby-Borgström 2002; Sonnby-Borgstrom et al. 2008; Gazzola et al. 2006; Kaplan and Iacoboni 2006; Dimberg et al. 2011; Chartrand and Bargh 1999). A recent fMRI study provided some additional compelling evidence using multivariate pattern analysis (MVPA) showing that the anterior insula, and perhaps the middle insula and cingulate cortex, had identical patterns of hemodynamic activity both when receiving pain and when observing pain (Corradi-Dell'Acqua et al. 2011). Of course, fMRI studies cannot provide single-cell resolution and are only correlational. Thus, though the "mirror-neuron hypothesis of negative intersubjectivity" remains a very popular (and very controversial) hypothesis in academic and popular literature (Gallese et al. 2011; Baird et al. 2011; Hickok 2009), it has never been tested directly.

Oxytocin's potential role in negative intersubjectivity is unclear, and likely indirect. Oxytocin is popularly called the "love" or "empathy" hormone (see lovehormone.org). More accurately, it is a neuropeptide secreted from the paraventricular nucleus, accessory magnocellular nuclei, and the supraoptic nucleus of the hypothalamus during many social behaviors (such as birthing or sex). In rodents, oxytocin receptor expression patterns correlate strongly with sociality and pair-bonding.[92] In humans, one study

acprof-9780199357666.indd 255 5/24/2016 5:03:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

demonstrated that oxytocin levels in the blood correlated positively with self-reported positive feelings of "empathy" and negative feelings of distress when people watched videos of a father describing his son's terminal brain cancer (Barraza and Zak 2009). Oxytocin nasal inhalations have also been shown to increase generosity in a one-shot game asking participants to split a sum of money with a stranger (Zak et al. 2007) and change brain activity in response to infant cries (Riem et al. 2011). In addition, some oxytocin gene variants correlate with self-reports of trait empathy (Rodrigues et al. 2009; Wu et al. 2012). Reports like these have led to speculations that oxytocin could potentially be used to enhance prosocial behavior (Walter 2012; Yamasue et al. 2012; McKaughan 2012).

Enthusiasm for using oxytocin as a "love hormone" to increase negative intersubjectivity and prosocial behavior has been tempered by conflicting evidence about whether such applications would work, and more evidence that even if oxytocin did affect negative intersubjectivity or prosocial behavior, it would likely do so in an indirect fashion. First, some studies fail to detect any effects of oxytocin on human negative intersubjectivity or responses to others' pain (Singer et al. 2008; Bartz et al. 2011), and others have observed oxytocin effects that are in the opposite direction of ideal social behavior (Zhong et al. 2012; Radke and 2012; Sheng et al. 2013; De Dreu 2012). Second, other studies report qualified interactions such that oxytocin only affected subgroups of people with either a specific gender, set of development conditions, or baseline social abilities (Wu et al. 2012; Bartz et al. 2011; Riem et al. 2013; Hurlemann et al. 2010; Bartz et al. 2010; Huffmeijer et al. 2012).

Part of the explanation for these complex results is likely based in oxytocin anatomy, which has been characterized mostly through rodent research rather than human research. Regardless of the task, most studies find that the observed oxytocin effects are mediated by decreased activity in the amygdala, a limbic brain region involved in many aspects of emotion (Singer et al. 2008; Hurlemann et al. 2010; Striepens et al. 2012; Hirosawa et al. 2012; Tost et al. 2010). Indeed, oxytocin receptors are densely expressed in the amygdala (Veinante and Freund-Mercier 1998; Loup et al. 1991). Importantly, at least in rats, they are much less densely expressed in the insula, and perhaps not expressed at all in the ACC (Tribollet et al. 1992). This hints that oxytocin's role in social behavior may not be directly through negative intersubjectivity circuits, but rather through amygdala-dependent social circuits that are modulated or preferentially enlisted in specific contexts, including (but not limited to) circuits related to general anxiety and/or fear (Churchland and Winkielman 2012). Thus, despite great interest in the role of oxytocin in negative intersubjectivity and prosocial behavior, its potential therapeutic applications remain controversial.

To conclude, efforts to map out the neural bases of intersubjectivity have led to many interesting and compelling biological models of how negative intersubjectivity might manifest. Unfortunately, two of the reigning hypotheses about the neural basis of negative intersubjectivity have yet to be proven or refuted: (1) the putative overlap between

acprof-9780199357666.indd 256 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

neural circuits involved in observing and receiving distress (are some of the same cells involved? If so, which ones, and why?), and (2) the related "mirror-neuron hypothesis of negative intersubjectivity" (do cells exist that respond to both being in distress and affectively imitating another's distress, and do these cells have any relation to the motor system?). Slightly more progress has been made testing the oxytocin hypothesis of negative intersubjectivity, but efforts have yielded conflicting results. Up to this point, our ability to more definitively test these hypotheses of the neural basis of negative intersubjectivity has been obstructed by the technological constraints of human research. The last section of this chapter will discuss how rodent research can rectify this problem.

11.3. Evidence of Negative Intersubjectivity in Rodents

11.3.1. Humans versus Rodents as Experimental Subjects

The fact that morality research has almost exclusively used humans as experimental subjects is understandable given the popular view that humans are the only species to have morality (although some argue that nonhuman primates, dolphins, and elephants have rudimentary moral systems) (de Waal 2009, 2008). Furthermore, a lot of morality research aims to study emotions or beliefs, and it is challenging to develop methods that allow us to infer the internal state or beliefs of animals with whom we cannot communicate. The downside of restricting morality research to humans, of course, is that it will only be able to advance as far as techniques available in humans and human research ethics permit. Neuroscience tools available in humans index a wide variety of biological functions and can cover many brain or body regions at once, but they also all have poor temporal and spatial resolution for the types of processes we believe are responsible for cognition and behavior (Homberg 2013; de Waal 2012).

To illustrate, consider the limitations of using fMRI to address the observing-is-feeling or the mirror-neuron hypothesis of negative intersubjectivity. The following details of brain anatomy might at first seem detached from complex phenomena like morality, but they are extremely important for understanding why we have not yet been able to test popular theories of the neural basis of negative intersubjectivity. Brain tissue is populated by heterogeneous populations of cells. Brain cells can be excitatory or inhibitory (or be different types of neurons within these classifications), can have diverse anatomical projections that allow them to participate in different anatomical circuits, can use different neurotransmitters or peptides, and most of the time aren't even neurons! (Allen and Barres 2009). Critical for the methods of human research, the BOLD (blood oxygen level dependent) signal measured by fMRI is due to hemodynamics that respond to activity of any of these cells, regardless of whether they are stimulating or inhibiting functioning in a brain area, and regardless of whether or not they are neurons (Schulz et al. 2012). Therefore, the BOLD signal is too imprecise a measurement to provide us information about many types of neural circuits. Furthermore, the BOLD signal can sometime be too insensitive a measurement. Very small populations of cells within

acprof-9780199357666.indd 257 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

a given brain area may have incredibly important roles in an animal's behavior (Witten et al. 2010), but not be detectable with fMRI because the hemodynamic response of the brain area is dominated by the metabolic needs of surrounding cells. In sum, if the neural processes underlying moral behavior are not mediated by relatively large, homogeneous populations of cells firing in a relatively homogeneous way, we will not be able to detect or study those mechanisms using human fMRI studies, and such studies may actually mislead us. Neither the observing-is-feeling nor the mirror-neuron hypothesis of negative intersubjectivity makes any claims about whether relevant cells are in large homogeneous populations.

If not humans, in what species should we study negative intersubjectivity? At first, nonhuman primates might seem like the ideal model organism. Monkeys and apes have a close evolutionarily relationship to humans and under certain conditions can be used to examine many detailed mechanisms of cognition. Primate work has its own set of methodological challenges, however. First, the financial cost and regulation governing the use of research monkeys dictates that only two or three animals can be used per experiment. If monkeys have the same diversity in social behavior as humans, it would be unwise to infer general mechanisms of social decision-making from such a small number of animals. Second, although it is controversial to cause distress in any experimental species, it is particularly controversial to cause distress or pain in monkeys (Suran and Wolinsky 2009), making it difficult to study how monkeys respond to other monkeys in distress or pain. Third, even if both of these challenges are surmounted, monkeys are a difficult species to use to study the molecular or anatomical mechanisms underlying behavior because it is costly and ethically contentious to genetically alter primates or collect their brains. To be most effective, moral neuroscience needs to be able to study negative intersubjectivity and social behavior in a model species that can be used in large numbers, that allows us to record from and manipulate the activity of single neurons, and that is compatible with techniques for robust and controlled circuit dissection and manipulation. The most efficient way to meet these criteria, despite nonhuman primates' evolutionary proximity to humans, is to study negative intersubjectivity in the dominant mammalian species used in basic neuroscience research: rodents.

In stark contrast to human research, rodent research permits exquisite control over environmental and genetic factors affecting its participants and affords data with extremely high spatial and temporal resolution. In rodents we can (with enough time) not only determine precisely what neurons fire in response to a certain stimulus or action, but also the genetic identity of those neurons, what those neurons look like, where those neurons project to or receive projections from, and how those neurons share information with other neurons. This is the type of information needed to address the observing-is-feeling and mirror-neuron hypotheses of negative intersubjectivity. This is also the type of information that permitted (preliminary) explanations of the confusing data collected while testing the oxytocin hypothesis of negative intersubjectivity.

acprof-9780199357666.indd 258 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

Another strong advantage of using rodents as model organisms is that they offer a relatively unique opportunity to study the causal relationship between neural properties and associated phenotypes. Popular science writing has exposed most people to the idea that we can lesion or temporarily inactivate specific brain regions in rodents, and either knock out or knock in specific genes. What might be less known is that we can actually knock out or knock in specific genes in specific brain regions at specific times during a behavior or development (Luo et al. 2008). Furthermore, thanks to a new technology called optogenetics, we can now activate or inhibit genetically targeted populations of cells with precise temporal precision (Fenno et al. 2011). We can also use advances in brain-computer interfaces to train algorithms that can determine how and when to make specific neurons fire to create precise behaviors, like grasping an object or even kicking a soccer ball (Thorsten and Christian 2011; Nicolelis 2012). It would be terribly exciting and perhaps revolutionary to apply all of these techniques to the study of moral behavior and moral decision-making. Of course, the primary challenge in using rodents to study morality is this: do rodents do or think anything that is morally relevant?

11.3.2. Requirements of a Rodent Model of Negative Intersubjectivity

Evidence supporting a strong relationship between negative intersubjectivity and moral behavior has been presented, and the limitations of studying moral behavior exclusively in humans has been discussed. The groundwork has been laid to argue that a primary goal of moral neuroscience over the next decade should be to develop, establish, and use a standardized rodent model of negative intersubjectivity to (a) identify the biological systems and neural processes that lead to humans *disliking* pain and distress in other humans, and (b) understand how these systems and processes lead to *avoidance* of causing pain and distress in other humans, or stated in reverse, how dysfunction in these systems and processes lead to persistent callous or violent behavior. The next step towards making this argument is to delineate what such a model would look like.

To address (a), a rodent behavioral model of negative intersubjectivity needs to convincingly demonstrate that rodents *dislike* witnessing other rodents in distress. To interpret the literature discussed in the following section, it is important to know that rodents (and some would argue humans) have three natural defense mechanisms against something they don't like: flight (avoidance), freezing, or fight (aggressive attack). In theory, any one of these defense mechanisms could be used to index negative affect. However, to simultaneously address (b), a rodent behavioral model of negative intersubjectivity must demonstrate that rodents are motivated to actively *avoid* distress in other rodents. Given the purposes of the modeling negative intersubjectivity, clearly a model that induces attack in an animal observing another animal's distress would not be appropriate. When considering the other two defense mechanisms, it is useful to consider their ethological functions. Avoidance is the defense mechanism typically used when

acprof-9780199357666.indd 259 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

threats are defined, present, and avoidable. Freezing, on the other hand, is the defense mechanism typically used when threats are undefined, not yet present, and unavoidable (Eilam 2005). If we want to understand how negative intersubjectivity contributes to behavior when a person has an option to prevent violence from happening, avoidance would be the most appropriate natural defense mechanism. This becomes even more clear when one considers that avoidance and freezing are behaviorally mutually exclusive (because you can't be fleeing if you are freezing) and are known to be mediated by different neural circuits (Eilam 2005; Fanselow 1994). For these reasons, a rodent model of intersubjective avoidance would be a better model for human negative intersubjectivity than a rodent model of intersubjective freezing (the importance of this distinction will be become clear later). Putting these concepts together, an ideal rodent model of intersubjective avoidance would do the following:

a. Convincingly show that rodents do not like witnessing distress in other rodents
b. Incorporate an active avoidance or choice behavior, rather than a freezing behavior
c. Have an analogue test that could be run in humans for validation and comparison

It is important to recognize that such a model of negative intersubjectivity would not meet many definitions of empathy, because it does not make any assumptions about intentions, especially altruistic intentions. The next section reviews where the field is in developing such a model.

11.3.3. Progress in Developing Rodent Models of Negative Intersubjectivity

Although interest in rodent negative intersubjectivity has increased over the past decade, evidence that rodents might be useful for studying intersubjectivity has been around for a long time. It is commonly known in animal husbandry that exposing rodents to stress cues from other rodents can increase baseline stress, even if no harm is incurred to the observing animals themselves. This vicarious stress is so well accepted that the American Veterinary Medical Association recommends rodent euthanasia be performed away from locations where other animals are housed to prevent potential stress in the noneuthanized animals (AVMA Guidelines on Euthanasia, 2007) (Špinka 2012). Furthermore, in support of these practical guidelines, experimental observations have documented that observing, smelling, and hearing other mice get shocked for an extended period of time can increase morphine self-administration (Kuzmin et al. 1996), cocaine self-administration (Ramsey and Van Ree 1993), saccharine preference (Pijlman et al. 2003), locomotor activity (Pijlman et al. 2003; Gutiérrez-García et al. 2006), and immobility in a forced-swim test (Gutiérrez-García and Contreras 2009; Gutiérrez-García et al. 2007),

acprof-9780199357666.indd 260 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

and disrupt sleep (Cui et al. 2008). These studies make it clear that rodents perceive at least some cues from other animals in distress that can affect their overall behavior. That said, these documented observations only prove that rodents perceive distress cues in other rodents; they don't yet prove that rodents have negative emotions that motivate them to avoid other rodents in distress.

Fortunately, more explicit behavioral reports of rodents perceiving other rodents' distress have been documented as well. Male mice will modulate writhing pain behaviors when they see other mice exhibiting the same writhing pain behaviors, a response modulated by the physical proximity between the mice being tested (Langford et al. 2006, 2011). In addition, rats show increased locomotor activity and startle amplitude when they are returned to a home cage with a rat who was recently shocked in a separate room (Knapska 2006), and their ability to learn to associate a shock with a conditioned stimulus will be improved by observing a conspecific receive shock (Chen et al. 2009; Guzman et al. 2009; Kiyokawa et al. 2009; Bruchey et al. 2010; but see Bredy and Barad 2009) or get worse if they are paired with a conspecific who did *not* receive shock in association with that same stimulus (Guzman et al. 2009; Kiyokawa et al. 2009). Outside of an explicit learning context, rats and mice who have experienced foot shock themselves will freeze upon witnessing a cagemate experience foot shock (Atsak et al. 2011) or upon witnessing a cagemate vocalize in response to a stimulus previously paired with foot shock (Kim et al. 2010). Perhaps most relevant, even if an observing mouse has never experienced foot shock itself, it may freeze upon witnessing a noncagemate experience foot shock, as long as the noncagemate is shocked with very intense strength and frequency (Jeon et al. 2010). These data demonstrate that distress cues from other rodents will affect freezing responses in observing rodents, which in turn suggests that the experience of perceiving other rodents' distress cues can have a negative valence. This is encouraging for developing a rodent model of negative intersubjectivity, but for the reasons stated earlier, freezing is not the ideal behavior for studying the role of negative intersubjectivity in violence. We must look elsewhere in the literature to find examples of behaviors relevant to avoiding distress in others.

Interestingly, most of the studies suggesting that rodents will perform active behaviors to avoid another rodents' distress were published half a century ago and were implemented in rats rather than mice. The most famous of these was published by Russell Church in 1959. The explicit goal of this study was to determine whether the "sympathetic" response that "animals and people" often have to the emotional states of others could be explained by others' emotional states acting as conditioned stimuli, or neutral stimuli that are learned through experience to be associated with other valenced (in this case negatively valenced) stimuli to oneself. To address this, Church taught a group of hungry observer rats to press a lever to obtain food. Then Church conditioned the rats. In the first conditioning sessions, a light would come on for one minute followed by another minute of a rat in an adjacent box getting shocked. This taught the observers to associate the light with the other rat getting shocked. In the next set of conditioning

acprof-9780199357666.indd 261 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

sessions, the rat in the adjacent box would get shocked for thirty seconds followed by a one-second shock to the observer rat. This taught the observers to associate the other rat getting shocked with shock to themselves. Then Church tested how much the observer rats would press a lever to obtain food in the presence of the rat in the adjacent box getting shocked, even if the observers were no longer shocked themselves. He found that, indeed, shock to another rat initially served as an effective conditioned stimulus: observers that were conditioned to associate shock to another rat in the adjacent chamber with shock to themselves reduced their lever-pressing for food during testing days. However, this decrease in lever pressing habituated each day over the course of ten days. Church also found that a control group who never received shocks to themselves during training did not decrease their lever pressing at all during testing (and, if anything, increased their lever pressing). Based on this data the authors concluded, "The difference between the experimental and unshocked control group was considered support for a conditioned-response interpretation of some cases of 'sympathy' " (Church 1959, 134). In other words, Church hypothesized that humans may dislike observing distress in others because we learn through experience that distress in others is usually associated with some type of distress in ourselves. Of note, even if this hypothesis is true, humans (or rats) would still have negative intersubjectivity as it is defined in this chapter.

The Church (1959) study is famous because it is one of the first studies designed to assess whether rats change their behavior in response to other rats' distress. However, the results were not particularly encouraging as a model of negative intersubjectivity because observers' response to other rats' distress habituated quite rapidly and was therefore not very robust. In addition, the primary observation was related to a lack of lever pressing rather than an increase in lever pressing. This makes it difficult to interpret whether the observing rats were just freezing when the other rats were getting shocked or whether they were intentionally withholding lever presses because doing so would prevent further distress.

The first study to address whether rodents are motivated to perform a directed action to avoid distress in another rat was published by Rice and Gainer in 1962. In a rather strange experimental paradigm, they began by training observer rats to press a lever to avoid (or escape) a shock to themselves. A five-second signal light preceded the escapable shock. At the same time as the observer rats were performing this avoidance task, they could see a styrofoam block suspended from a hoist being raised up and down adjacent to the training chamber. Whenever an observer rat pressed the lever to avoid or escape a shock, the block was lowered to the floor, where it remained for fifteen seconds before being raised again. The signal light that preceded the shock to the observer was presented when the block was at its zenith. After learning this avoidance task, all observer rats underwent extinction training until bar pressing in response to the signal light disappeared. This suggested that the rats no longer associated the signal light with a future experience of shock to themselves. After extinction was verified, testing commenced. During testing, the procedure changed in two ways. First, unlike training, the observer

acprof-9780199357666.indd 262 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

rats never got shocked themselves. Second, the styrofoam block was replaced with a live rat who "typically squealed and wriggled satisfactorily while suspended, and if it did not, it was prodded with a sharp pencil until it exhibited signs of discomfort" (Rice and Gainer 1962, 123). Although the observer rats would not be shocked during testing, as in the conditioning phase, whenever an observer rat pressed the lever in their chamber, the hoist would be lowered until the hoisted rat had all four feet on the ground for fifteen seconds. Under these conditions, Rice and Gainer compared the bar pressing of observer rats who did and did not receive previous avoidance training. They also examined the bar pressing of observer rats who did and did not receive previous avoidance training when the suspended rat was replaced with a styrofoam block. They made the following observations: *both* groups of observers who witnessed a rat being suspended from the hoist, regardless of previous avoidance training, pressed the lever significantly more times than their matched controls. Over the course of ten minutes, the experimental groups with a hoisted rat pressed the lever an average of fifteen and eighteen times (for the trained and untrained rats, respectively), whereas the control groups with a hoisted styrofoam block pressed an average of one and five times. This clearly demonstrated that rats will perform active actions when they observe another rat in distress. Since the study suggested that rats might be willing to press a lever to bring a suspended rat to safety, the authors controversially concluded, "It is suggested that this behavior might operationally be termed altruistic" (Rice and Gainer 1962, 124).

Half a century later, a relevant paradigm was reported by Inbal Ben-Ami Bartal, Jean Decety, and Peggy Mason (2011). The purpose of this paradigm was to test whether rats had enough prosocial motivation to release a trapped cagemate from captivity. To test this, they placed two rats in an open arena. The observer was free to run around, while its cagemate was trapped in a plastic restrainer. If the observer pressed the door to the restrainer with enough force, it could open the restrainer and release its cagemate. During training, the observer and receiver were placed in the arena for ninety minutes and the latency to door opening was recorded. If the observer did not open the door of the restraining chamber by a certain time, it was opened by the experimenter and the rats were allowed to interact before being returned to their shared home cage. Control experiments were run in which (a) an observer was placed in the arena with an empty restrainer, but no other rat, present, (b) an observer was placed in the arena with an empty restrainer present, and a cagemate in a separate adjacent compartment, (c) the observer could open the restrainer, but the cagemate was released into a separate compartment, (d) the observer was placed in the arena with one restrainer holding its cagemate and another restrainer holding chocolate chips, and (e) the observer was placed in the arena with one restrainer holding chocolate chips and another empty restrainer. The following observations were made: observers took an average of seven days to learn how to open the door to the restrainer to "free" the restrained rat before the time limit, and by the end of this learning curve the observers opened the door between five and ten minutes after the beginning of the session. The average opening rate was similar for the experimental

acprof-9780199357666.indd 263 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

group as for the group when the observer could open the restrainer but the cagemate was released into a separate compartment (suggesting the observer wasn't just opening the restrainer to have social interaction), but much higher than either group with empty restrainers. Finally, when given a choice between opening a restrainer with a rat or a restrainer with chocolate chips, observers opened both restrainers equally quickly, but when given a choice between opening a restrainer with chocolate chips or an empty restrainer, it opened the restrainer with the chocolate chips faster (although it also opened the empty restrainer). The authors concluded, "The most parsimonious interpretation of the observed helping behavior is that rats free their cagemate in order to end distress, either their own or that of the trapped rat, that is associated with the circumstances of the trapped cagemate" (Bartal et al. 2011, 1430).

Although both studies by Rice and Gainer (1962) and Bartal et al. (2011) show clearly that rats' active behaviors are affected by other rats in distress, they also share a characteristic that confounds their interpretation: general arousal (Lucke and Baton 1980; Lavery and Foley 1963). In order for these behaviors to be useful models of negative intersubjectivity, the behaviors need to convincingly show that the observers feel negative affect when other rats are in distress, and no plausible explanations for the observed behavior should exist that do not require the subjective state of the observer to have a negative valence. To the contrary, both behaviors reported by Rice and Gainer (1962) and Bartal et al. (2011) could potentially be explained by increased general arousal without corresponding negative valence. Generalized arousal is the large-scale readiness of many potential behavioral responses to respond to environmental conditions (Pfaff et al. 2008). (Arousal, itself, is unvalenced, but it can interact with either positive or negative valence circuits to affect subjective states or behavior; Kensinge 2004). Arousal manifests in rodents as increased locomotor behavior when no other stimulus is present to initiate motivated action (Pfaff et al. 2008). Thus, if a suspended rat or a restrained rat increased arousal without "dislike" in an observing rat, an observing rat should increase its locomotion and interact with its environment in a nonspecific way. Bartal et al. measured locomotor activity in their study and reported that observers were more active before learning to open the restrainer, consistent with this prediction.

The best way to determine whether the reported responses were specifically linked to distress in another animal in either the Rice and Gainer study or the Bartal et al. study would have been to give observers the opportunity to perform a task-irrelevant action in the testing environment. For example, an inactive lever could have been provided in the Rice and Gainer study, or a separate empty restrainer could have been provided in the Bartal et al. study; researchers could then have tested whether responding toward the lever associated with lowering the suspended rat or the restrainer with the rat, as opposed to the unpaired level or the empty restrainer, was selectively increased. Neither study reported such a control. Even without these controls, there is reason to be skeptical of a negative intersubjectivity interpretation of the Bartal et al. study because both when (d) the observer was placed in the arena with one restrainer holding its cagemate

acprof-9780199357666.indd 264 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

and another restrainer holding chocolate chips and (e) the observer was placed in the arena with one restrainer holding chocolate chips and another empty restrainer, the observer opened both restrainers. This suggests that, indeed, the observer was more likely to respond to many cues, not just the restrainer with a rat in it. Further, inspection of the videos published with the Bartal et al. study suggests that neither the observer nor the restrained rat avoided the restrainer once the restrained rat was released. In fact, as soon as the restrained rat was released, both rats competed to enter the restrainer again, perhaps using the restrainer as a toy. This suggests that being in the restrainer was not likely very aversive (or aversive at all), which in turn suggests that the observer may never have had an opportunity to perceive distress cues from the other rat. In sum, the Rice and Gainer and Bartal et al. studies provide convincing evidence that witnessing another rat be suspended or be restrained is a salient stimulus for rats, which may be relevant to negative intersubjectivity in some way. However, they are still not ideal behaviors for modeling how negative intersubjectivity inhibits violence. If a violent human commits violence against a victim, we don't want the violent human to get generally aroused and start causing more damage to everything in his environment, including the victim. Rather, we want the violent human to perceive the victim's distress cues and feel so negative about those distress cues that he stops the violence he is currently performing and goes out of his way to avoid causing victims' distress cues in the future. Neither the Rice and Gainer nor the Bartal et al. study provides convincing evidence that rats can have these avoidance responses to other rats' distress.

Five other published studies do provide convincing evidence that rodents will avoid other rodents in distress, but unlike Rice and Gainer and Bartal et al. studies, none of them are well known or cited. The first of these studies put rats in a T-maze and trained them to go to either end of the T to receive a sucrose award. When receiving sucrose at one end of the T was paired with another rat getting shocked, the observers selectively avoided that end of the T and sought sucrose at the other end (Evans and Braud 1969). A similar behavior was reported in mice who avoided odors of a distressed mouse (distressed by hypertonic saline injection or by foot shock) in a shuttle box or tube where the odors were only pumped into one side of the box or tube (Rottman and Snowdon 1972; Zalaquett and Thiessen 1991). Perhaps the most convincing demonstration of rodent negative intersubjectivity, though, was reported by Simonov and colleagues using rats (Preobrazhenskaya and Simonov 1970; Wetzel and Simonov 1978). They designed a three-chamber apparatus such that an observer could be put in two chambers, one having a roof and the other having no roof. The observer could move freely between both of these chambers, and they were arranged in an "L" around a third receiver chamber. The observers could see, but not interact with, a receiver placed in the inner chamber from either the covered or the uncovered area. Observers and receivers were placed in this apparatus for five minutes daily, and the amount of time the observers spent in the covered versus uncovered chambers was recorded. During baseline, most observers preferred the covered chamber (although the researchers did not test baseline preference in

acprof-9780199357666.indd 265 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

all animals). During testing, whenever an observer went into the covered chamber, the rat in the inside chamber would get shocked until the observer entered the uncovered chamber. In twelve rats, Preobrazhenskaya and Simonov showed that some observers dramatically avoided the covered chamber under these conditions. Further, when the observers avoided the covered area during testing, they did so consistently over many days (sometimes over twenty days in a row). The authors also made the following interesting observations: (*a*) four of the twelve animals started to prefer the uncovered area after only two or three days of testing, (*b*) five of the twelve animals started to prefer the uncovered area only after they had been used as the shocked rat for three or five days, (*c*) three of the twelve rats never preferred the uncovered chamber, even after having been used as the shocked rat, and (*d*) the four rats that avoided the covered chamber without having experienced shock themselves tended to be less anxious as measured by an open field test, while the three rats that never avoided the covered chamber were very anxious. Further, at the end of the conditioning experiments, Preobrazhenskaya and Simonov tested observers in an "aggression" test. In this test, pairs of rats were placed in a compartment with a wire floor and observed as electrical charge was gradually introduced into the floor. The voltage at which the rats responded by attacking their neighbor was recorded. The four rats that avoided the covered chamber without having experienced shock themselves in the three-chamber experiment tended to respond to the shocked floor by cooperating with the other rat to short-circuit the nociceptive stimulation with their hind limbs. The three rats that never avoided the covered chamber showed aggressive responses at low voltages.

These studies provide five independent reports that mice and rats will change their physical location to avoid distress cues from other mice or rats. These behaviors aren't easily explained by general arousal, because the animals' change in location is specific to where the distress cues originate. Furthermore, the study by Preobrazhenskaya and Simonov suggests that rats will change their location in a dramatic and persistent way that may not habiutate even after weeks of training. Even more impressively, how much rats change their physical location to avoid distress cues from other mice or rats may negatively correlate with how likely they are to engage in aggression as a defensive reaction, suggesting that even in rats, avoidance of other's pain may correlate with decreased general aggression.

In sum, these studies support the hypothesis that it is possible to meet the criteria for an ideal rodent model of intersubjective avoidance. Amazingly, rodents avoid others' pain, a behavior human violent offenders do not perform and psychopaths may not have the affective neural machinery to perform (Aniskiewicz 1979; House and Milligan 1976; Blair et al. 1997). To clarify, these studies do *not* indicate *why* observer rodents feel negative in response to other rodents' distress (are they scared? do they feel sympathy?), nor whether observer rodents have conscious intentions toward other rodents' distress (do they want to help, or do they simply want to get away?). However, as discussed in sections 11.1 and 11.2 of this chapter, neither a specific origin nor a specific intention is necessary

acprof-9780199357666.indd 266 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

for negative intersubjectivity to mediate violence aversion. What is most important is simply that observers avoid others' distress cues, as opposed to ignoring them or freezing upon perceiving them. Indeed, the studies described above convincingly show that rodents can and will avoid other rodents' distress. The studies published thus far have not designed their behavioral paradigms to be able to quantify how much each rodent does not like witnessing distress in other rodents, but the paradigms could easily be adapted to do so by simply varying the intensity of negative stimulus rats are required to endure to avoid another rat in distress (intensity of light in the Preobrazhenskaya and Simonov studies) or varying the amount of positive stimulus rats are required to give up in order to avoid another rat in distress (amount of sucrose reward in the Evans and Braud study). Once the exact paradigm is established, it would be relatively easy to implement parallel studies in humans for validation and comparison. A few labs are actively working on developing these kinds of behavioral models for rodents, so the question is no longer *if* a robust rodent model of negative intersubjectivity can be established, but rather *when* such models will be standardized.

11.3.4. The Future of Rodent Models of Negative Intersubjectivity

When rodent intersubjective avoidance models become serviceable, multiple kinds of information will become instantly attainable that may facilitate treatments for antisocial behavior. As a few examples, the observing-is-feeling and mirror-neuron hypotheses of negative intersubjectivity can finally be proved or disproved by recording from single cells or populations of cells while rodents are avoiding other animals' distress. The hypothesis that oxytocin is involved in intersubjective avoidance can finally be tested rigorously using causal manipulations of specific populations or types of oxytocin cells. The pharmacological basis of negative intersubjectivity can be characterized to help identify potential pharmaceutical treatments for antisocial behavior in humans. High- or low-intersubjectivity rat lines can be bred to provide insight into what types of traits or biological markers tend to correlate with high or low intersubjective avoidance, and to help identify candidate genes responsible for prosocial or antisocial behavior in humans. More nuanced questions can also be addressed, such as whether the anatomical connections to putative mirror neurons help clarify why pain to oneself still feels subjectively different than observing pain in another, even if both pain experiences activate some of the same neurons. And, of course, new models of negative intersubjectivity can be developed and refined. The past decade of neuroscience has yielded more information than we imagined possible, and, similarly, the next decade of neuroscience has the potential to use rodent models to understand violence aversion in creative and insightful ways that far surpass what is described here. Very little—if any—of this progress will be achieved if antisocial behavior is only studied in humans.

This chapter has focused on how rodent research can inform our understanding of violence aversion. Will rodent research tell us anything about other types of moral

acprof-9780199357666.indd 267 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

action? As I final thought, I'd like to propose that *we should find out* (and I suspect the answer will be yes). There are two main mechanisms by which rodent negative intersubjectivity research is likely to influence human morality research: (1) by providing an increasingly sophisticated knowledge base to operationalize biologically based definitions and questions about moral judgment and moral action in general, and (2) by motivating very specific, testable hypotheses about what causes and influences judgments and behaviors that we believe have moral relevance. In particular, rodent research will be helpful both for delineating the different computations that contribute to moral judgments and for understanding subconscious forces that influence our moral behaviors. Such empirical observations may then, in turn, influence theoretical accounts of morality by providing opportunities to better define terms like "emotion" and "reason," and by providing more clarity to what "causes" us to commit to a moral judgment or perform a moral action. Of course, it is likely that some aspects of human morality cannot be studied in other species. However, it will be just as useful to determine what aspects of human behavior are unique to humans as it will be to determine what aspects of human behavior are shared with other species. In fact, such comparisons may be instrumental for understanding what makes humans feel something is "moral" in the first place.

To be clear, not all aspects of morality can be understood with empirical research, and rodent research cannot answer all empirical questions about morality. We should not perform all future morally relevant research in rodents. Rather, just like most other fields of biomedical research, rodent research and human research should be pursued in parallel with a commitment to compare and combine relevant findings. Rodent models should be used to generate hypotheses about moral action or judgment, and human studies should be used to test these hypotheses in verifiably moral situations (or vice versa). The results of these tests will then be used to refine and revise hypotheses in rodents (or humans), and the cycle will begin again. This approach has helped us make great advances in our understanding and treatment of phenomena as complex as addiction (Gardner 2010), and it may do the same for morality research. Most importantly, if the moral decision-making field commits to taking a multispecies, mechanism-grounded strategy to understanding moral action, we may dramatically improve human lives by reducing violence and increasing prosocial behavior. This, I believe, is the next frontier of morality research.

Our evolved reverence for the sanctity of life makes random violence and cruelty one of the most devastating of our strange behaviors, and one that seems like it will never be explainable. Caleb Mallery, a victim of violence in 1780, is reported to have cried out to his murderer in between blows, "Tell me what you do it for!" (Lepore 2009). While the past decade has done much to advance our understanding of moral judgment, the next frontier is to find an answer to Caleb's question. We must learn how to explain immoral action.

acprof-9780199357666.indd 268 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

References

Allen, N. J., and B. A. Barres. 2009. "Neuroscience: Glia—More Than Just Brain Glue." *Nature* 457, no. 7230: 675–677.

Aniskiewicz, A. S. 1979. "Autonomic Components of Vicarious Conditioning and Psychopathy." *Journal of Clinical Psychology* 35, no. 1: 60–67.

Atsak, P., et al. 2011. "Experience Modulates Vicarious Freezing in Rats: A Model for Empathy." *PLoS One* 6, no. 7: e21855.

Azevedo, R. T., et al. 2013. "Their Pain Is Not Our Pain: Brain and Autonomic Correlates of Empathic Resonance with the Pain of Same and Different Race Individuals." *Human Brain Mapping* 34, no. 12: 3168–3181.

Baird, A. D., I. E. Scheffer, and S. J. Wilson. 2011. "Mirror Neuron System Involvement in Empathy: A Critical Look at the Evidence." *Social Neuroscience* 6, no. 4: 327–335.

Barraza, J. A., and P. J. Zak. 2009. "Empathy toward Strangers Triggers Oxytocin Release and Subsequent Generosity." *Annals of the New York Academy of Sciences* 1167, no. 1: 182–189.

Bartal, I. B. A., J. Decety, and P. Mason. 2011. "Empathy and Pro-social Behavior in Rats." *Science* 334, no. 6061: 1427–1430.

Bartz, J. A., et al. 2010. "Oxytocin Selectively Improves Empathic Accuracy." *Psychological Science* 21, no. 10: 1426–1428.

Bartz, J. A., et al. 2011. "Social Effects of Oxytocin in Humans: Context and Person Matter." *Trends in Cognitive Sciences* 15, no. 7: 301–309.

Batson, C. D., N. Ahmad, and D. A. Lishner. 2011. "Empathy and Altruism." In *The Oxford Handbook of Positive Psychology*, edited by Shane J. Lopez and C. R. Snyder, 417–426. New York: Oxford University Press.

Batson, C. D., J. Fultz, and P. A. Schoenrade. 1987. "Distress and Empathy: Two Qualitatively Distinct Vicarious Emotions with Different Motivational Consequences." *Journal of Personality* 55, no. 1: 19–39.

Batson, C. D., and K. C. Oleson. 1991. "Current Status of the Empathy-Altruism Hypothesis." In *Prosocial behavior: Review of Personality and Social Psychology*, edited by M. S. Clark, 62–85. Newbury Park, CA: Sage.

Batson, C. D., and L. L. Shaw. 1991. "Evidence for Altruism: Toward a Pluralism of Prosocial Motives." *Psychological Inquiry* 2, no. 2: 107–122.

Beck, J. C. 2010. "Dangerous Severe Personality Disorder: The Controversy Continues." *Behavioral Sciences and the Law* 28, no. 2: 277–288.

Bernhardt, B. C., and T. Singer. 2012. "The Neural Basis of Empathy." *Annual Review of Neuroscience* 35: 1–23.

Björkqvist, K., K. Österman, and A. Kaukiainen. 2000. "Social Intelligence – Empathy = Aggression?" *Aggression and Violent Behavior* 5, no. 2: 191–200.

Blair, J., et al. 1997. "The Psychopathic Individual: A Lack of Responsiveness to Distress Cues?" *Psychophysiology* 34, no. 2: 192–198.

Blasi, A. 1980. "Bridging Moral Cognition and Moral Action: A Critical Review of the Literature." *Psychological Bulletin* 88, no. 1: 1–45.

Blasi, A. 1983. "Moral Cognition and Moral Action: A Theoretical Perspective." *Developmental Review* 3, no. 2: 178–210.

acprof-9780199357666.indd 269 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

Bredy, T. W., and M. Barad. 2009. "Social Modulation of Associative Fear Learning by Pheromone Communication." *Learning and Memory* 16, no. 1: 12–18.

Bruchey, A. K., C. E. Jones, and M. H. Monfils. 2010. "Fear Conditioning by Proxy: Social Transmission of Fear during Memory Retrieval." *Behavioural Brain Research* 214, no. 1: 80–84.

Bruneau, E. G., A. Pluta, and R. Saxe. 2012. "Distinct Roles of the 'Shared Pain' and 'Theory of Mind' networks in Processing Others' Emotional Suffering." *Neuropsychologia* 50, no. 2: 219–231.

Chartrand, T. L., and J. A. Bargh. 1999. "The Chameleon Effect: The Perception-Behavior Link and Social Interaction." *Journal of Personality and Social Psychology* 76, no. 6: 893–910.

Chen, Q., J. B. Panksepp, and G. P. Lahvis. 2009. "Empathy Is Moderated by Genetic Background in Mice." *PLoS One* 4, no. 2: e4387.

Church, R. M. 1959. "Emotional Reactions of Rats to the Pain of Others." *Journal of Comparative and Physiological Psychology* 52, no. 2: 132–134.

Churchland, P. S., and P. Winkielman. 2012. "Modulating Social Behavior with Oxytocin: How Does It Work? What Does It Mean?" *Hormones and Behavior* 61, no. 3: 392–399.

Corradi-Dell'Acqua, C., C. Hofstetter, and P. Vuilleumier. 2011. "Felt and Seen Pain Evoke the Same Local Patterns of Cortical Activity in Insular and Cingulate Cortex." *Journal of Neuroscience* 31, no. 49: 17996–18006.

Cui, R., et al. 2008. "The Effects of Atropine on Changes in the Sleep Patterns Induced by Psychological Stress in Rats." *European Journal of Pharmacology* 579, no. 31–3: 153–159.

Darwall, S. 1998. "Empathy, Sympathy, Care." *Philosophical Studies* 89, no. 32: 261–282.

Davis, M. 1980. "A Multidimensional Approach to Individual Differences in Empathy." In *JSAS Catalog of Selected Documents in Psychology*. Washington, DC: American Psychological Association.

De Dreu, C. K. 2012."Oxytocin Modulates Cooperation within and Competition between Groups: An Integrative Review and Research Agenda." *Hormones and Behavior* 61, no. 3: 419–428.

de Vignemont, F., and T. Singer. 2006. "The Empathic Brain: How, When and Why?" *Trends in Cognitive Sciences* 10, no. 10: 435–441.

de Waal, F. B. 2008. "Putting the Altruism Back into Altruism: The Evolution of Empathy." *Annual Review of Psychology* 59, no. 1: 279–300.

de Waal, F. B. 2009. *The Age of Empathy*. New York: Harmony.

de Waal, F. B. 2012. "Research Chimpanzees May Get a Break." *PLoS Biology* 10, no. 3: e1001291.

Decety, J. 2011. "Dissecting the Neural Mechanisms Mediating Empathy." *Emotion Review* 3, no. 1: 92–108.

Decety, J., and C. Lamm. 2006. "Human Empathy through the Lens of Social Neuroscience." *Scientific World Journal* 6: 1146–1163.

Dimberg, U., P. Andréasson, and M. Thunberg. 2011. "Emotional Empathy and Facial Reactions to Facial Expressions." *Journal of Psychophysiology* 25, no. 1: 26–31.

Eastman, N. 1999. "Who Should Take Responsibility for Antisocial Personality Disorder? Fallon Suggests Emphasising Custody, but Psychiatrists' Future Role Remains Unclear." *BMJ: British Medical Journal* 318, no. 7178: 206–207.

Eilam, D. 2005. "Die Hard: A Blend of Freezing and Fleeing as a Dynamic Defense: Implications for the Control of Defensive Behavior." *Neuroscience and Biobehavioral Reviews* 29, no. 8: 1181–1191.

acprof-9780199357666.indd 270 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

Eisenberg, N. 1986. *Altruistic Emotion, Cognition, and Behavior*. Hillsdale, NJ: Erlbaum.

Eisenberg, N. 2008. "Empathy-Related Responding and Prosocial Behaviour." In *Empathy and Fairness*, edited by Greg Bock and Jamie Goode, 71–88. Chichester: John Wiley & Sons.

Eisenberg, N., N. D. Eggum, and L. Di Giunta. 2010. "Empathy-Related Responding: Associations with Prosocial Behavior, Aggression, and Intergroup Relations." *Social Issues and Policy Review* 4, no. 1: 143–180.

Eisenberg, N., and R. Fabes. 1990. "Empathy: Conceptualization, Measurement, and Relation to Prosocial Behavior." *Motivation and Emotion* 14, no. 2: 131–149.

Eisenberg, N., and P. A. Miller. 1987. "The Relation of Empathy to Prosocial and Related Behaviors." *Psychological Bulletin* 101, no. 1: 91–119.

Eisenberg, N., and J. Strayer. 1987. "Critical Issues in the Study of Empathy." In *Empathy and Its Development*, edited by N. Eisenberg and J. Strayer, 3–13. New York: Cambridge University Press.

Essau, C. A., S. Sasagawa, and P. J. Frick. 2006. "Callous-Unemotional Traits in a Community Sample of Adolescents." *Assessment* 13, no. 4: 454–469.

Evans, V. E., and W. G. Braud. 1969. "Avoidance of a Distressed Conspecific." *Psychonomic Science* 15, no. 3: 166.

Fan, Y., et al. 2011. "Is There a Core Neural Network in Empathy? An fMRI Based Quantitative Meta-analysis." *Neuroscience and Biobehavioral Reviews* 35, no. 3: 903–911.

Fanselow, M. 1994. "Neural Organization of the Defensive Behavior System Responsible for Fear." *Psychonomic Bulletin and Review* 1, no. 4: 429–438.

Farrington, D., Ohlin, L., and Wilson, J. Q. 1986. *Understanding and Controlling Crime*. New York: Springer-Verlag.

Fenno, L., O. Yizhar, and K. Deisseroth. 2011. "The Development and Application of Optogenetics." *Annual Review of Neuroscience* 34, no. 1: 389–412.

Ferguson, C. J. 2010. "Genetic Contributions to Antisocial Personality and Behavior: A Meta-analytic Review from an Evolutionary Perspective." *Journal of Social Psychology* 150, no. 2: 160–180.

Feshbach, N. D. 1979. "Empathy Training: A Field Study in Affective Education." In *Aggression and Behavior Change: Biological and Social Processes*, edited by Seymour Feshbach and Adam Fraczek, 234–249. New York: Praeger.

Feshbach, N. D., and S. Feshbach. 1982. "Empathy Training and the Regulation of Aggression: Potentialities and Limitations." *Academic Psychology Bulletin* 4: 399–413.

Frick, P. J., and S. F. White. 2008. "Research Review: The Importance of Callous-Unemotional Traits for Developmental Models of Aggressive and Antisocial Behavior." *Journal of Child Psychology and Psychiatry* 49, no. 4: 359–375.

Gallese, V., et al. 2011. "Mirror Neuron Forum." *Perspectives on Psychological Science* 6, no. 4: 369–407.

Gardiner, S. 2008. "NYPD Inaction over a Missing Black Woman Found Dead Sparks a Historic Racial-Bias Lawsuit." *Village Voice*, May 6.

Gardner, E. L. 2010. "What We Have Learned about Addiction from Animal Models of Drug Self-Administration." *American Journal on Addictions* 9, no. 4: 285–313.

Gazzola, V., L. Aziz-Zadeh, and C. Keysers. 2006. "Empathy and the Somatotopic Auditory Mirror System in Humans." *Current Biology* 16, no. 18: 1824–1829.

Gibbon, S., et al. 2010. "Psychological Interventions for Antisocial Personality Disorder." *Cochrane Database of Systematic Reviews*, doi:10.1002/14651858.CD007668.pub2.

acprof-9780199357666.indd 271 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

Ginsberg, A. 2006. "Killer's 'Fun' Day in Court." *New York Post*, November 21.

Greene, J. D., et al. 2001. "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science* 293, no. 5537: 2105–2108.

Gu, X., et al. 2010. "Functional Dissociation of the Frontoinsular and Anterior Cingulate Cortices in Empathy for Pain." *Journal of Neuroscience* 30, no. 10: 3739–3744.

Gu, X., et al. 2012. "Anterior Insular Cortex Is Necessary for Empathetic Pain Perception." *Brain* 135, no. 9: 2726–2735.

Gunter, T. D., M. G. Vaughn, and R. A. Philibert. 2010. "Behavioral Genetics in Antisocial Spectrum Disorders and Psychopathy: A Review of the Recent Literature." *Behavioral Sciences and the Law* 28, no. 2: 148–173.

Gutiérrez-García, A. G., and C. M. Contreras. 2009. "Stressors Can Affect Immobility Time and Response to Imipramine in the Rat Forced Swim Test." *Pharmacology Biochemistry and Behavior* 91, no. 4: 542–548.

Gutiérrez-García, A. G., et al. 2006. "A Single Session of Emotional Stress Produces Anxiety in Wistar Rats." *Behavioural Brain Research* 167, no. 1: 30–35.

Gutiérrez-García, A. G., et al. 2007. "Urine from Stressed Rats Increases Immobility in Receptor Rats Forced to Swim: Role of 2-Heptanone." *Physiology and Behavior* 91, no. 1: 166–172.

Guzman, Y. F., et al. 2009. "Social Modeling of Conditioned Fear in Mice by Non-fearful Conspecifics." *Behavioural Brain Research* 201, no. 1: 173–178.

Haidt, J. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108, no. 4: 814–834.

Haidt, J. 2004. "The Emotional Dog Gets Mistaken for a Possum." *Review of General Psychology* 8, no. 4: 283–290.

Hare, R. D. 1991. *The Hare Psychopathy Checklist—Revised (PCL-R)*. Toronto, ON: Multi-Health Systems.

Hare, R. D. 2003. *Manual for the Revised Hare Psychopathy Checklist*. 2nd ed. Toronto, ON: Multi-Health Systems.

Hawes, D. J., and M. R. Dadds. 2005. "The Treatment of Conduct Problems in Children with Callous-Unemotional Traits." *Journal of Consulting and Clinical Psychology* 73, no. 4: 737–741.

Hickok, G. 2009. "Eight Problems for the Mirror Neuron Theory of Action Understanding in Monkeys and Humans." *Journal of Cognitive Neuroscience* 21, no. 7: 1229–1243.

Hirosawa, T., et al. 2012. "Oxytocin Attenuates Feelings of Hostility Depending on Emotional Context and Individuals' Characteristics." *Scientific Reports* 2, doi: 10.1038/srep00384.

Hoffman, M. L. 2000. *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge: Cambridge University Press.

Hoffman, M. L. 2008. "Empathy and Prosocial Behavior." *Handbook of Emotions*, edited by Michael Lewis, Jeannette M. Haviland-Jones, and Lisa Feldman Barrett, 440–455. 3rd ed. New York: Guilford Press.

Hogan, R. 1969. "Development of an Empathy Scale." *Journal of Consulting and Clinical Psychology* 33, no. 3: 307–316.

Homberg, J. R. 2013. "Measuring Behaviour in Rodents: Towards Translational Neuropsychiatric Research." *Behavioural Brain Research* 236, no. 1: 295–306.

House, T. H., and W. L. Milligan. 1976. "Autonomic Responses to Modeled Distress in Prison Psychopaths." *Journal of Personality and Social Psychology* 34, no. 4: 556–560.

acprof-9780199357666.indd 272 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

Howells, K., and A. Day. 2007. "Readiness for Treatment in High Risk Offenders with Personality Disorders." *Psychology, Crime and Law* 13, no. 1: 47–56.

Huffmeijer, R., et al. 2012. "Asymmetric Frontal Brain Activity and Parental Rejection Predict Altruistic Behavior: Moderation of Oxytocin Effects." *Cognitive, Affective, and Behavioral Neuroscience* 12, no. 2: 382–392.

Hurlemann, R., et al. 2010. "Oxytocin Enhances Amygdala-Dependent, Socially Reinforced Learning and Emotional Empathy in Humans." *Journal of Neuroscience* 30, no. 14: 4999–5007.

Iacoboni, M. 2009. "Imitation, Empathy, and Mirror Neurons." *Annual Review of Psychology* 60: 653–670.

Insel, T. R. 2010. "The Challenge of Translation in Social Neuroscience: A Review of Oxytocin, Vasopressin, and Affiliative Behavior." *Neuron* 65, no. 6: 768–779.

Jabbi, M., J. Bastiaansen, and C. Keysers. 2008. "A Common Anterior Insula Representation of Disgust Observation, Experience and Imagination Shows Divergent Functional Connectivity Pathways." *PLoS One* 3, no. 8: e2939.

Jeon, D., et al. 2010. "Observational Fear Learning Involves Affective Pain System and Cav1.2 Ca2+ Channels in Acc." *Nature Neuroscience* 13, no. 4: 482–488.

Jolliffe, D., and D. P. Farrington. 2004. "Empathy and Offending: A Systematic Review and Meta-analysis." *Aggression and Violent Behavior* 9, no. 5: 441–476.

Jolliffe, D., and J. Murray. 2012. "Lack of Empathy and Offending." In *The Future of Criminology*, edited by R. Loeber and B. C. Welsh, 62–69. New York: Oxford University Press.

Kahn, R. E., A. L. Byrd, and D. A. Pardini. 2012. "Callous-Unemotional Traits Robustly Predict Future Criminal Offending in Young Men." *Law and Human Behavior*, advance online publication.

Kaplan, J. T., and M. Iacoboni. 2006. "Getting a Grip on Other Minds: Mirror Neurons, Intention Understanding, and Cognitive Empathy." *Social Neuroscience* 1, nos. 3–4: 175–183.

Katz, N. L. 2006. "Attorney Knifed as Suspects Bring Bloody Mayhem to Brooklyn Murder Trial." *New York Daily News*, January 20.

Kensinger, E. A. 2004. "Remembering Emotional Experiences: The Contribution of Valence and Arousal." *Reviews in the Neurosciences* 15, no. 4: 241–252.

Khalifa, N., et al. 2010. "Pharmacological Interventions for Antisocial Personality Disorder." *Cochrane Database of Systematic Reviews*, doi:10.1002/14651858.CD007667.pub2.

Kim, E. J., et al. 2010. "Social Transmission of Fear in Rats: The Role of 22-kHz Ultrasonic Distress Vocalization." *PLoS One* 5, no. 12: e15077.

Kimonis, E. R., et al. 2008. "Assessing Callous-Unemotional Traits in Adolescent Offenders: Validation of the Inventory of Callous-Unemotional Traits." *International Journal of Law and Psychiatry* 31, no. 3: 241–252.

Kiyokawa, Y., et al. 2009. "Main Olfactory System Mediates Social Buffering of Conditioned Fear Responses in Male Rats." *European Journal of Neuroscience* 29, no. 4: 777–785.

Knapska, E., et al. 2006. "Between-Subject Transfer of Emotional Information Evokes Specific Pattern of Amygdala Activation." *Proceedings of the National Academy of Sciences* 103, no. 10: 3858–3862.

Kuzmin, A., et al. 1996. "Enhancement of Morphine Self-Administration in Drug Naive, Inbred Strains of Mice by Acute Emotional Stress." *European Neuropsychopharmacology* 6, no. 1: 63–68.

Lamm, C., C. D. Batson, and J. Decety. 2007. "The Neural Substrate of Human Empathy: Effects of Perspective-Taking and Cognitive Appraisal." *Journal of Cognitive Neuroscience* 19, no. 1: 42–58.

acprof-9780199357666.indd 273 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

Langford, D. J., et al. 2006. "Social Modulation of Pain as Evidence for Empathy in Mice." *Science* 312, no. 5782: 1967–1970.

Langford, D. J., et al. 2011. "Varying Perceived Social Threat Modulates Pain Behavior in Male Mice." *Journal of Pain* 12, no. 1: 125–132.

Lavery, J. J., and P. J. Foley. 1963. "Altruism or Arousal in the Rat?" *Science* 140, no. 3563: 172–173.

Lepore, J. 2009. "Rap Sheet: Why Is American History So Murderous?" *New Yorker*, November 9.

LeSure-Lester, G. E. 2000. "Relation between Empathy and Aggression and Behavior Compliance among Abused Group Home Youth." *Child Psychiatry and Human Development* 31, no. 2: 153–161.

Loup, F., et al. 1991. "Localization of High-Affinity Binding Sites for Oxytocin and Vasopressin in the Human Brain: An Autoradiographic Study." *Brain Research* 555, no. 2: 220–232.

Lovett, B. J., and R. A. Sheffield. 2007. "Affective Empathy Deficits in Aggressive Children and Adolescents: A Critical Review." *Clinical Psychology Review* 27, no. 1: 1–13.

Lucke, J. F., and C. D. Baton. 1980. "Response Suppression to a Distressed Conspecific: Are Laboratory Rats Altruistic?" *Journal of Experimental Social Psychology* 16, no. 3: 214–227.

Luo, L., E. M. Callaway, and K. Svoboda. 2008. "Genetic Dissection of Neural Circuits." *Neuron* 57, no. 5: 634–660.

McKaughan, D. 2012. "Voles, Vasopressin, and Infidelity: A Molecular Basis for Monogamy, a Platform for Ethics, and More?" *Biology and Philosophy* 27, no. 4: 521–543.

Meyer, M. L., et al. 2013. "Empathy for the Social Suffering of Friends and Strangers Recruits Distinct Patterns of Brain Activation." *Social Cognitive and Affective Neuroscience* 8, no. 4: 446–454.

Miller, P. A., et al. 1996. "Relations of Moral Reasoning and Vicarious Emotion to Young Children's Prosocial Behavior toward Peers and Adults." *Developmental Psychology* 32, no. 2: 210.

Moffitt, T. E. 1993. "Adolescence-Limited and Life-Course-Persistent Antisocial Behavior: A Developmental Taxonomy." *Psychological Review* 100, no. 4: 674–701.

Nicolelis, M. A. 2012. "Mind in Motion." *Scientific American* 307, no. 3: 58–63.

Osganian, A. L. 2008. "Limitations on Biomedical and Behavioral Research Involving Prisoners: An Argument Supporting the Institute of Medicine's Recommendations to Revise Regulations." *New England Journal on Criminal and Civil Confinement* 34: 429.

Pfaff, D., et al. 2008. "Concepts and Mechanisms of Generalized Central Nervous System Arousal." *Annals of the New York Academy of Sciences* 1129, no. 1: 11–25.

Pickersgill, M. 2011. "'Promising' Therapies: Neuroscience, Clinical Practice, and the Treatment of Psychopathy." *Sociology of Health and Illness* 33, no. 3: 448–464.

Pijlman, F. T. A., G. Wolterink, and J. M. Van Ree. 2003. "Physical and Emotional Stress Have Differential Effects on Preference for Saccharine and Open Field Behaviour in Rats." *Behavioural Brain Research* 139, nos. 1–2: 131–138.

Preobrazhenskaya, L. A., and P. V. Simonov. 1970. "Conditioned Avoidance Responses to Nociceptive Stimulation of Another Individual." *Neuroscience and Behavioral Physiology* 4, no. 4: 15–20.

Pustilnik, A. C. 2009. "Violence on the Brain: A Critique of Neuroscience in Criminal Law." *Wake Forest Law Review* 44: 183–238.

acprof-9780199357666.indd 274 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

Radke, S., and E. R. De Bruijn. 2012. "The Other Side of the Coin: Oxytocin Decreases the Adherence to Fairness Norms." *Frontiers in Human Neuroscience* 6: Article 193.

Ramsey, N. F., and J. M. Van Ree. 1993. "Emotional but Not Physical Stress Enhances Intravenous Cocaine Self-Administration in Drug-Naive Rats." *Brain Research* 608, no. 2: 216–222.

Rice, G., and P. Gainer. 1962. "'Altruism' in the Albino Rat." *Journal of Comparative and Physiological Psychology* 55: 123–125.

Riem, M. M. E., et al. 2011. "Oxytocin Modulates Amygdala, Insula, and Inferior Frontal Gyrus Responses to Infant Crying: A Randomized Controlled Trial." *Biological Psychiatry* 70, no. 3: 291–297.

Riem, M. M. E., et al. 2013. "Does Intranasal Oxytocin Promote Prosocial Behavior to an Excluded Fellow Player? A Randomized-Controlled Trial with Cyberball." *Psychoneuroendocrinology* 38, no. 8: 1418–1425.

Rizzolatti, G., and L. Craighero. 2004. "The Mirror-Neuron System." *Annual Review of Neuroscience* 27, no. 1: 169–192.

Rodrigues, S. M., et al. 2009. "Oxytocin Receptor Genetic Variation Relates to Empathy and Stress Reactivity in Humans." *Proceedings of the National Academy of Sciences* 106, no. 50: 21437–21441.

Rottman, S. J., and C. T. Snowdon. 1972. "Demonstration and Analysis of an Alarm Pheromone in Mice." *Journal of Comparative and Physiological Psychology* 81, no. 3: 483–490.

Salekin, R. T., C. Worley, and R. D. Grimes. 2010. "Treatment of Psychopathy: A Review and Brief Introduction to the Mental Model Approach for Psychopathy." *Behavioral Sciences and the Law* 28, no. 2: 235–266.

Schaich Borg, J., et al. 2006. "Consequences, Action, and Intention as Factors in Moral Judgments: An fMRI Investigation." *Journal of Cognitive Neuroscience* 18, no. 25: 803–817.

Schulz, K., et al. 2012. "Simultaneous BOLD fMRI and Fiber-Optic Calcium Recording in Rat Neocortex." *Nature Methods* 9, no. 6: 597–602.

Sheng, F., et al. 2013. "Oxytocin Modulates the Racial Bias in Neural Responses to Others' Suffering." *Biological Psychology* 92, no. 2: 380–386.

Shirtcliff, E. A., et al. 2009. "Neurobiology of Empathy and Callousness: Implications for the Development of Antisocial Behavior." *Behavioral Sciences and the Law* 27, no. 2: 137–171.

Singer, T., et al. 2008. "Effects of Oxytocin and Prosocial Behavior On Brain Responses to Direct and Vicariously Experienced Pain." *Emotion* 8, no. 6: 781–91.

Sonnby-Borgström, M. 2002. "Automatic Mimicry Reactions as Related to Differences in Emotional Empathy." *Scandinavian Journal of Psychology* 43, no. 5: 433–443.

Sonnby-Borgstrom, M., P. Jönsson, and O. Svensson. 2008. "Gender Differences in Facial Imitation and Verbally Reported Emotional Contagion from Spontaneous to Emotionally Regulated Processing Levels." *Scandinavian Journal of Psychology* 49, no. 2: 111–122.

Špinka, M. 2012. "Social Dimension of Emotions and Its Implication for Animal Welfare." *Applied Animal Behaviour Science* 138, nos. 3–4: 170–181.

Spunt, R. P., and M. D. Lieberman. 2012. "An Integrative Model of the Neural Systems Supporting the Comprehension of Observed Emotional Behavior." *NeuroImage* 59, no. 3: 3050–3059.

Striepens, N., et al. 2012. "Oxytocin Facilitates Protective Responses to Aversive Social Stimuli in Males." *Proceedings of the National Academy of Sciences* 109, no. 44: 18144–18149.

acprof-9780199357666.indd 275 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

Suran, M., and H. Wolinsky. 2009. *The End of Monkey Research? New Legislation and Public Pressure Could Jeopardize Research with Primates in Both Europe and the USA*." EMBO reports, 10, no. 10: 1080.

Taylor, C. 2011. "Nothing Left to Lose? Freedom and Compulsion in the Treatment of Dangerous Offenders." *Psychodynamic Practice* 17, no. 3: 291–306.

Thorsten, O. Z., and K. Christian. 2011. "Towards Passive Brain-Computer Interfaces: Applying Brain-Computer Interface Technology to Human-Machine Systems in General." *Journal of Neural Engineering* 8, no. 2: 025005.

Tost, H., et al. 2010. "A Common Allele in the Oxytocin Receptor Gene (Oxtr) Impacts Prosocial Temperament and Human Hypothalamic-Limbic Structure and Function." *Proceedings of the National Academy of Sciences* 107, no. 31: 13936–13941.

Tribollet, E., et al. 1992. "Oxytocin Receptors in the Central Nervous System." *Annals of the New York Academy of Sciences* 652, no. 1: 29–38.

Trommsdorff, G., W. Friedlmeier, and B. Mayer. 2007. "Sympathy, Distress, and Prosocial Behavior of Preschool Children in Four Cultures." *International Journal of Behavioral Development* 31, no. 3: 284–293.

Vaughn, M. G., and M. DeLisi. 2008. "Were Wolfgang's Chronic Offenders Psychopaths? On the Convergent Validity between Psychopathy and Career Criminality." *Journal of Criminal Justice* 36, no. 1: 33–42.

Vaughn, M. G., et al. 2011. "The Severe 5%: A Latent Class Analysis of the Externalizing Behavior Spectrum in the United States." *Journal of Criminal Justice* 39, no. 1: 75–80.

Veinante, P., and M. J. Freund-Mercier. 1998. "Distribution of Oxytocin and Vasopressin Binding Sites in the Rat Extended Amygdala: A Histoautoradiographic Study." *Journal of Comparative Neurology* 383, no. 3: 305–325.

Walsh, Z., and D. S. Kosson. 2008. "Psychopathy and Violence: The Importance of Factor Level Interactions." *Psychological Assessment* 20, no. 2: 114–120.

Walter, H. 2012. "Social Cognitive Neuroscience of Empathy: Concepts, Circuits, and Genes." *Emotion Review* 4, no. 1: 9–17.

Ward, T., and G. Willis. 2010. "Ethical Issues in Forensic and Correctional Research." *Aggression and Violent Behavior* 15, no. 6: 399–409.

Wellman, H. M. 2010. "Developing a Theory of Mind." In *The Wiley-Blackwell Handbook of Childhood Cognitive Development*, edited by Usha Goswami, 258–284. 2nd ed. Malden, MA: Wiley-Blackwell.

Wetzel, W., and P. V. Simonov. 1978. "Avoidance Reaction to Painful Stimulation of Another Rat: Effect of Methylglucamine Orotate." *Pharmacology Biochemistry and Behavior* 9, no. 4: 401–404.

Wicker, B., et al. 2003. "Both of Us Disgusted in My Insula: The Common Neural Basis of Seeing and Feeling Disgust." *Neuron* 40, no. 3: 655–664.

Wispé, L. 1986. "The Distinction between Sympathy and Empathy: To Call Forth a Concept, a Word Is Needed." *Journal of Personality and Social Psychology* 50, no. 2: 314–321.

Witten, I. B., et al. 2010. "Cholinergic Interneurons Control Local Circuit Activity and Cocaine Conditioning." *Science Signalling* 330, no. 6011: 1677.

Wu, N., Z. Li, and Y. Su. 2012. "The Association between Oxytocin Receptor Gene Polymorphism (Oxtr) and Trait Empathy." *Journal of Affective Disorders* 138, no. 3: 468–472.

acprof-9780199357666.indd 276 5/24/2016 5:03:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue May 24 2016, NEWGEN

Yamasue, H., et al. 2012. "Integrative Approaches Utilizing Oxytocin to Enhance Prosocial Behavior: From Animal and Human Social Behavior to Autistic Social Dysfunction." *Journal of Neuroscience* 32, no. 41: 14109–14117.

Zak, P. J., A. A. Stanton, and S. Ahmadi. 2007. "Oxytocin Increases Generosity in Humans." *PLoS One* 2, no. 11: e1128.

Zalaquett, C., and D. Thiessen. 1991. "The Effects of Odors from Stressed Mice on Conspecific Behavior." *Physiology and Behavior* 50, no. 1: 221–227.

Zhong, S., et al. 2012. "U-Shaped Relation between Plasma Oxytocin Levels and Behavior in the Trust Game." *PLoS One* 7, no. 12: e51095.

acprof-9780199357666.indd 277 5/24/2016 5:03:08 PM