This is a preprint of a chapter published in Johan De Smedt & Helen De Cruz & (eds.), *Empirically Engaged Evolutionary Ethics*, Springer – Synthese Library. The final published version is available at: https://link.springer.com/chapter/10.1007/978-3-030-68802-8_5

Morality as an Evolutionary Exaptation

Marcus Arvan
University of Tampa

What did moral cognition evolve for—that is, what is its evolutionary function? The dominant answer to this question across anthropology (Curry, 2016; Cosmides and Tooby, 1992; Henrich and Henrich, 2007), evolutionary biology (Alexander, 1987; de Waal, 2006), philosophy (Carruthers and James, 2008; Joyce, 2006, 2007; Kitcher, 1998, 2005, 2011; Prinz 2007, p. 185; Sinclair, 2012, p. 14; Sterelny and Fraser 2016; Wisdom 2017), and psychology (Casebeer, 2003; Greene 2015; Tomasello and Vaish, 2013) is that moral cognition is a biological adaptation to foster social cooperation. This chapter argues, to the contrary, that moral cognition is likely an evolutionary *exaptation* (Gould, 1991): a form of cognition where neurobiological capacities selected for in our evolutionary history for a variety of different reasons—many unrelated to social cooperation—were put to a new, prosocial use *after* the fact through individual rationality, learning, and the development and transmission of social norms.

My argument has three steps. First, I provide a brief overview of the emerging behavioral neuroscience of moral cognition. I then outline a theory of moral cognition that I have argued explains these findings better than alternatives (Arvan, 2020). Finally, I demonstrate how the evidence for this theory of moral cognition and human evolutionary history together suggest that moral cognition is likely not a biological adaptation. Instead, like reading sheet music or riding a bicycle, moral cognition is something that individuals *learn* to do—in this case, in response to *sociocultural norms* created in our ancestral history and passed down through the ages to enable cooperative living. This chapter thus aims to set evolutionary ethics on a new path, identifying the evolutionary function of moral cognition with a complex interplay between neurobiological and cultural evolution.

1. How to Do Evolutionary Ethics: Four Sequential Questions

Evolutionary ethicists standardly use the method of *telling plausible evolutionary stories* of how capacities seemingly involved in moral cognition—such as altruism (Kitcher, 1998), caring for others (Churchland, 2011), 'moral emotions' such as empathy, spite, shame, and guilt (Frank, 1988), 'universal human values' (Curry, 2016; Haidt and Joseph, 2004; Haidt, 2012), or particular judgment-types such as moral beliefs (May, 2018, chapter 3)—were likely selected for in our ancestral history. Evolutionary ethicists thus generally assume that they have a clear enough idea of what moral cognition involves (e.g. altruism, particular emotions or values, etc.) in order to theorize properly about its evolutionary origins and function. However, this is a mistake. First, as Smyth (2017, p. 1127) argues:

> The collection of practices, beliefs, and dispositions we call 'morality' is far more functionally complex than the standard story would have us believe. Morality may indeed reduce social tensions in certain contexts, but it may also inflame them in others, and it probably plays a variety of other distinct roles in human societies.

To take just one example, moral language and beliefs are often used in ways, such as moral grandstanding (Tosi and Warmke, 2016), that are conducive to group polarization—a phenomenon linked to *less* cooperative (and immoral) behaviors ranging from war to genocide and other forms of mass violence (Arvan, 2019). Second, there is a deeper problem. Meta-ethicists disagree substantially over what *constitutes* morality and by extension moral cognition. For example, Kantians hold that morality properly understood involves conformity to the Categorical Imperative—a principle that Kant argues is normatively binding due to transcendental freedom, not 'moral emotions', 'universal human values', or any empirical effects of morality, such as social cooperation (Kant, 1785, 4:387-4:392, 4:394; Wood, 2008; Cf. Korsgaard, 2008, 2009; Luco, 2016). Moral cognition for Kantians thus involves a *very* specific kind of

reasoning: namely, cognizing (at least implicitly) the Categorical Imperative, and acting upon universalizable maxims that respect the 'humanity' of oneself and others (Kant, 1785, 4:421, 4:429). However, other metaethicists defend very different pictures of morality and moral cognition. For example, moral realists often argue that morality involves having and conforming to *moral intuitions*: immediate, non-inferential, and potentially affectively-laden (Haidt, 2001) judgments that X (an action, action-type, etc.) is right, wrong, good, or bad (Audi, 2016; Prichard, 1912; Ross, 1930). However, some apparently universal types of intuitions (involving norms of purity and respect for authority) may foster social cooperation in some contexts yet profoundly undermine it in others (Greene, 2013). For example, Hitler and the Nazis were obsessed with racial purity, regarding it as a moral imperative (Hitler, 1925, pp. 215, 282). Yet this belief, along with the belief that Germans should respect Hitler's absolute authority as Führer (Trueman, 2020), served to further genocide and World War II—immoral actions antithetical to social cooperation. Still other meta-ethicists argue that morality is reducible to *prudence*, that is, to what makes an individual's life tend to go better over the course of life as a whole (Aristotle, 1984, Book II sections 6–9, Book IV sections 5, 11–13, and Book X; Arvan, 2016, 2020). Yet, as we will see, if this is correct, then moral cognition fundamentally involves long-term planning capacities that *may* be used to foster social cooperation, but also to undermine it—including anti-social behaviors antithetical to social cooperation that would have plausibly increased the evolutionary fitness of our ancestors.

Consequently, the theory that morality is a biological adaptation for social cooperation appears to be based upon highly uncertain foundations. There are not only many different metaethical theories of the nature of morality and moral cognition (see Arvan, 2016, pp. 30-5 and Arvan, 2020, pp. 106-18 for overviews of influential accounts). On at least some such theories, 'the function' of moral cognition may not be social cooperation, but rather something

else entirely: long-term prudential planning, transcendental freedom, conformity to categorical normative reasons, and so on. Accordingly, in order to determine what the evolutionary function of moral cognition really is, we must be more careful. First, we must determine which metaethical criterion of morality has the best evidence in its favor. That criterion—if it exists—will enable us to determine with greater certainty what counts as moral (as opposed to non-moral) cognition. Second, once we determine what moral cognition is, we must establish which brain regions and associated cognitive capacities are involved in it.[1] Third, we need evidence of *how* the brain regions and capacities involved in moral cognition function within it: specifically, whether particular brain regions engage in moral cognition innately, or whether moral cognition is instead something we learn to do in response to features of our surrounding environment. Finally, we need evidence of how the brain regions and capacities involved in moral cognition were likely selected for in evolutionary history. Were particular brain regions involved in moral cognition selected as biological adaptations to foster social cooperation, or were they selected for in evolutionary history for very different reasons and only harnessed later (via learning and constructed sociocultural norms) for a prosocial use, *qua* exaptation?

In sum, to determine the true evolutionary function of moral cognition, we must carefully address the following four issues in order:

1. What morality *is*, and by extension what counts as *moral cognition*.
2. Precisely *which* brain regions and associated capacities are implicated in moral cognition.
3. *How* they function in moral cognition.
4. *How and why* they were selected for in evolutionary history.

[1] I do not mean to endorse neuroessentialism here, the view that specific capacities are located in or identical to the functions of particular brain regions. I merely affirm scientific findings that particular brain regions are associated with particular cognitive functions.

As we have seen, there is widespread metaethical disagreement about the very first issue: what morality is. One possible response to this problem is to try to provide such a broad definition of morality (as altruism, etc.) that the definition will seem uncontroversial (see Frank, 1988; Joyce, 2006; Kitcher, 2011). However, we have seen that any such definition will offend the metaethical sensibilities of those who defend a narrower definition (e.g. as conformity to Kant's Categorical Imperative, etc.). The lesson here, I believe, is that in doing evolutionary ethics, there is no way around taking controversial metaethical stances on the nature of morality and moral cognition. Accordingly, this will be my approach. I will outline an account of morality and moral cognition that I have defended and refined across two books (Arvan, 2016, 2020), and which I have argued to be the best explanation of a variety of empirical and normative data (Arvan, 2016, chapter 8; Arvan, 2020, chapter 4). I will then argue that on this theory, moral cognition is likely not a biological adaptation but rather a form of learned cognition that individuals engage in due to *sociocultural norms* originally created in our ancestral past on the basis of rational deliberation, which have been subsequently transmitted and enforced in stable cultures to this day.

2. Morality as Prudential Risk-Aversion

Across two books, I have argued that moral philosophy should be based on (A) empirical psychology and (B) a simple 'means-ends', instrumental theory of normativity according to which people rationally ought to adopt the best means for achieving their ends (Arvan 2016, chapters 1-3; Arvan, 2020, chapters 2-3). The basic rationale for this approach is as follows. First, whereas traditional philosophical methods have been argued to face serious epistemic problems (Brennan 2010; Arvan, 2016, chapter 1), empirical psychology promises demonstrable knowledge of human cognition (Arvan, 2020, chapters 1 and 4), recent replication issues aside (Maxwell et al., 2015). Second, whereas other forms of normativity—such as categorical normativity (Kant, 1785), metaphysically primitive moral reasons (Parfit, 2011; Scanlon, 1998,

2014), and so on—are deeply controversial, instrumental normativity enjoys virtually universal acceptance across academic theorizing (Anand, 1995; Hansson, 2005; Peterson, 2017) and everyday life (Arvan, 2016, pp. 24-7, Arvan, 2020, pp. 37-45, 66, 104, 132-3). The typical person recognizes that if X is their goal (or end), and Y is the best means to achieve X, then there is a clear sense (a 'means-end' sense) in which they *ought* to do Y. For example, students can recognize that if they want to perform well on an exam and studying hard is the best means to do well, then they *ought* to study hard. This is not only true of the typical person. Importantly, it is true even of individuals who may be skeptical of or otherwise insensitive to moral norms. For example, young children who misbehave, wanton criminals, and psychopaths all routinely recognize normative requirements of instrumental rationality. A thief or murderer can recognize that *if* they have committed a crime, they want to avoid detection, and the best means to avoid detection is to take careful steps to hide evidence, then there is a clear sense in which they *should* take those steps. Similarly, even very young children can understand that if they want to stay out of trouble with their parents or other authority figures (such as schoolteachers), there are things they *should* and *shouldn't* do (such as not get into schoolyard fights). Finally, instrumentalism and empirical psychology together promise a uniquely strong, unified, and parsimonious explanation of a wide variety of phenomena, normatively reducing morality to prudence and descriptively reducing moral cognition to prudential cognition (Arvan, 2020, chapters 2-4). Allow me to explain.

My theory of prudence and morality begins with these assumptions, as well as the further assumptions—also widely accepted in the philosophical literature (Bricker, 1980; Bruckner 2003, pp. 34–5; Haybron, 2011, Section 1; Price, 2002; Pettigrew, 2020) and in ordinary life (Aristotle 1984; Arvan, 2020, pp. 27-8)—that because human beings normally want to live happy lives, prudence (for humans) is a matter of making instrumentally optimal choices that maximize one's

own expected lifetime utility (Arvan 2020, chapter 2, section 1). I then argue that prudence involves *mental time-travel*, the capacity to mentally simulate different possible pasts and futures—as this is vital to learning from past prudential errors and deliberating about the future (Arvan, 2020, pp. 32-50). Third, following Donald Bruckner (2003), I argue that because life is profoundly uncertain over the long run, prudent individuals learn to act in ways that treat life this way: as consisting of decisions made in radical ignorance of lifetime probabilities (Arvan 2020, pp. 30-32). Importantly, I contend that we learn this primarily from *socialization*: from seeing risky violations of social norms punished by others around us, including authority figures such as parents, school officials, and law-enforcement (Arvan, 2020, pp. 37-45). Fourth, I argue that once we learn from socialization to treat life as highly uncertain, the internalized attitudes this generates ('moral risk-aversion') make it instrumentally rational to engage in *other-perspective taking* (OPT). We learn it is prudent to imaginatively simulate how our actions might affect others—including how others might reward or punish us, and how we might feel guilt or remorse—as a long-term strategy for minimizing severe regret: an end that prudent individuals have grounds to want to avoid given radical lifetime uncertainty (Bruckner 2003; Arvan, 2020, pp. 63-5; Arvan, 2016, pp. 118-28). Fifth, I argue that this form of prudential other-perspective-taking makes it rational to obey Four Principles of Fairness: a deontological principle of coercion-minimization, a consequentialist principle of mutual assistance, a contractualist principle of fair negotiation, and virtue-theoretic principle of internalizing the first three principles as standing cognitive and behavioral dispositions (Arvan, 2020, pp. 68-72; Arvan, 2016, chapters 5 and 6). While I cannot summarize these principles or their derivation here in detail, I have argued that they plausibly unify the moral domain, reconciling the competing insights of traditional moral frameworks, while also supporting Rawlsian frameworks for domestic, international, and global justice, both in 'ideal theory' and 'nonideal theory' (Arvan, 2020, pp. 83-7). Sixth, I argue that once

a person fully internalizes moral risk-aversion and the above principles of fairness (through socialization), the person comes to treat moral norms *as though* they are 'categorical' normative requirements, with categorical moral attitudes coming to comprise our 'conscience' (Arvan, 2020, pp. 42-50, 71; Arvan, 2016, pp. 110-11, 122, 177-80).

Notice that my account is broadly Hobbesian. In *Leviathan*, Hobbes argues that moral cognition is not naturally instilled in us biologically (Hobbes, 1651, chapter XIII). Although Hobbes allows that people in nature may have various 'pre-moral' capacities—such as concern for kin, empathy, and so on (Hobbes, 1651, chapters XIII and X)—for Hobbes our 'natural condition' revolves around purely instrumental planning, or seeking to effectively pursue our desires (Hobbes, 1651, chapter VI). Hobbes then argues that moral cognition (viz. Laws of Nature) is an achievement of instrumental reasoning and sociopolitical enforcement, as he holds that moral norms are only rational to obey when enforced by a sovereign authority (Hobbes, 1651, chapters XIV-XV). Importantly, Hobbes argues that even when enforced, moral laws are ultimately *prudential* laws—that they are merely "conclusions or theorems concerning what conduceth to [a person's] conservation and defense of themselves" (Hobbes, 1658, p. 47) My account is similar. It holds that sociocultural norms—originally learned in our ancestral past to enable social cooperation, and transmitted and incentivized in stable societies to this day—make it rational for children, adolescents, and adults to *learn* to use a variety of 'pre-moral' capacities that were not biologically selected for social cooperation in a novel, prosocial way.

We can begin to see this more clearly by first considering some evidence sometimes taken to favor the hypothesis that moral cognition is an innate biological adaptation. First, human infants, adults, and a variety of nonhuman animals demonstrate a rudimentary sense of fairness (Brosnan, 2006; Brosnan and de Waal, 2014; Geraci and Surian, 2011); Schmidt and Sommerville, 2011). Second, human infants and children display preferences for altruism (Barragan et al., 2020;

Schmidt and Sommerville, 2011) and retribution for antisocial behavior (Hamlin, 2013). Third, five 'moral foundations' (values of care, fairness, loyalty, respect for authority, and purity) have been argued to be universal across human societies (Doğruyol et al., 2019)—though serious questions have been raised about these claims (Graham et al., 2013; Suhler and Churchland, 2011). All of these findings might appear to suggest that moral cognition is innate and social cooperation its evolutionary function. However, this is a spurious inference. Although dogs, mice, and human infants all display a rudimentary sense of fairness, infants have other prosocial dispositions, and dogs can cooperate in small packs, we do not treat any of these creatures as *morally responsible* agents, blaming them for unfair or selfish behavior. Why? The answer is twofold. First, they lack the *mental time-travel* capacities necessary for appreciating the long-term consequences of their actions (Kennett and Matthews, 2009; Levy, 2007; Suddendorf and Corballis, 2007). Second, genuine moral responsibility also requires *recursion*: the capacity to apply moral rules to a potentially infinite variety of new cases, including cases where they individual is *not* inclined altruistically or fairly—as people are when tempted to behave immorally (Arvan, 2016, pp. 5-7, 96, 109). Crucially, only human adults appear to have either of these capacities—mental time-travel and recursion—in any robust degree (Corballis, 2007; Suddendorf and Corballis, 2007).

It is important to underscore here just how much evidence there is for the centrality of mental time-travel to moral cognition and responsibility. First, human adults—who we ordinarily consider to be morally-responsible agents—typically have robust mental time-travel capacities (Suddendorf and Corballis, 2007). Second, sub-classes of humans exhibiting diminished moral capacities—children, adolescents, and psychopaths—have underdeveloped mental time-travel capacities and neural-circuitry (Blair, 2003; Casey et al., 2008; Giedd et al., 1999; Kennett and Matthews 2009; Levy, 2007; Stuss et al., 1992; Weber et al., 2008; Yang and Raine, 2009), making them less able to appreciate the consequences of their actions (Baskin-Sommers et al., 2016; Hare,

1999; Hart and Dempster, 1997; Litton, 2008; Moffit, 1993; Moffit et al., 2011; Shoemaker, 2011). Third, mental time-travel is directly linked to moral performance: (1) lack of imaginative vividness of the future predicts psychopathy (Hosking et al., 2017) and criminal delinquency (Van Gelder et al., 2013), (2) the ability to project oneself into the future is negatively related to unethical behavior (Hershfield et al., 2012), (3) experimental interventions priming imagination of the future decrease willingness to violate moral norms (Van Gelder et al., 2013), and (4) experimental inhibitions of mental time-travel (via transcranial magnetic stimulation) result not only in greater impulsivity but also greater egocentricity, selfishness, deficits in other-perspective-taking, and less-prosocial behavior (Soutschek et al., 2016). Finally, nonhuman animals in general—who we do not treat as morally responsible agents—appear to lack any robust mental time-travel capacities (Suddendorf and Corballis, 2007). Although some evidence suggests that other hominids (great apes) and crows may possess some mental time-travel capacities, these capacities appear to be far more limited than ours (Kabadayi and Osvath, 2017).

The point then is this: although human infants, dogs, mice, and other animals have certain prosocial inclinations (viz. fairness, altruism, etc.), they are simply not moral agents. They lack cognitive capacities (mental time-travel, recursion, etc.) necessary for *genuine moral agency and moral cognition*. First, they lack capacities necessary for understanding why they should avoid immoral behavior in cases where they lack dispositions to behave morally (which is what mental time-travel and OPT enable in humans via long-term instrumental planning). Second, animals lack the ability to represent and extend moral principles to *new cases* (viz. recursion). To put it more simply, children and nonhuman animals are not moral agents—they do not engage in genuine moral cognition—because they lack capacities to *regulate* their behavior according to moral norms in cases where they lack prosocial inclinations (viz. temptations to behave selfishly rather than fairly or altruistically).

Similar considerations show why cross-cultural 'moral foundations' (or 'universal values') are insufficient for full-fledged moral cognition. Even if humans have evolved to naturally value care, fairness, loyalty, respect for authority, and purity, none of these are sufficient for moral responsibility or moral cognition. Adult human beings are morally responsible for our actions because, in addition to valuing particular things, we possess robust capacities for *regulating our behavior* according to moral norms via mental time-travel, OPT, and recursion (see May, 2018). It is thus simply a mistake to infer from the universality of values or prosocial dispositions in infants or animals that *moral cognition* is an innate biological capacity.

2. The Elements of Moral Cognition

If genuine moral cognition involves more than innate beliefs or values, then what exactly does it involve? The emerging behavioral neuroscience coheres extremely well with my theory of prudence and morality outlined above. On my account, moral cognition involves (i) mental time-travel, (ii) other-perspective-taking, and (iii) risk-aversion. We learn to care about other people's perspectives and interests in a distinctly *moral* way by learning (across childhood, adolescence, and adulthood) that other people typically *reward* us in the future for treating them well, and *punish* us for treating them poorly. These patterns of social reward and punishment—embodied in culturally-evolved norms (including laws)—lead us to *worry instrumentally* about violating social norms, viz. risk-aversion (Arvan, 2020, chapter 2). This form of risk-aversion then leads us to mentally simulate how others are likely to react to our actions (viz. mental time-travel), leading to represent and care about how our actions affect others (viz. OPT)—all of which makes it rational to obey *moral principles* (Arvan, 2020, chapter 3; Arvan, 2016, chapters 3-6).

Bearing this model of moral cognition in mind and the fact that evolution by natural selection is an incremental process wherein new biological capacities emerge and are selected for at different times in evolutionary history for different reasons, consider the following

empirical findings. First, moral cognition has indeed been found to centrally involve mental time-travel (Kennett and Matthews, 2009; Levy, 2007), other-perspective-taking (Viganò, 2017, pp. 219-21; Cf. Benoit et al., 2011; Peters and Büchel, 2010; Daniel et al., 2013; Singer and Tusche, 2014; Singer and Lamm, 2009), and risk-aversion (Ito et al., 1998; Baumeister et al., 2001; Kahneman and Tversky, 1979). Second, prudential and moral cognition have been found to be neurofunctionally intertwined in the ways my theory hypothesizes. Stimulating forward-looking mental time-travel results in greater prudential saving behaviors and greater fairness toward others via other-perspective-taking (Ersner-Hershfield, Wimmer, and Knutson, 2009; Ersner-Hershfield et al., 2009; Hershfield et al., 2011; Van Gelder et al., 2013), whereas inhibiting mental-time-travel degrades prudential behavior and fairness to others (Soutschek et al., 2016). Third, all of the following regions of the brain's Default Mode Network (DMN) have been implicated in moral judgment (i.e. moral belief) across a wide variety of tasks[2]:

a. *Ventromedial prefrontal cortex (vmPFC)*: processes risk and uncertainty, and is involved in learning from mistakes and applying moral judgments to one's own behavior (Fellows and Farah, 2007), as well as emotional regulation (Koenigs et al., 2007). Deficits lead to lack of empathy, irresponsibility, and poor decisionmaking (Motzkin et al., 2011), causing patients to choose immediate rewards ignoring future consequences (Bechara et al., 2000). Also implicated in 'extinction', wherein previously reinforced behaviors gradually cease when reinforcement no longer occurs (Milad et al., 2005).

b. *Dorsomedial prefrontal cortex (dmPFC)***:** involved sense of the self (Gusnard et al., 2001). and theory of mind, i.e. understanding others' mental states (Isoda and Nirotake, 2013).

[2] The following overview of DMN regions is from Arvan (2020), pp. 12-13. As I argue in Arvan (2020), chapter 4, although the DMN is involved in many cognitive tasks other than moral cognition, my account provides a powerful normative and descriptive explanation of why and how some of the main cognitive functions associated with these DMN regions *should and do* interact to generate moral cognition. Cf. Pascual et al. (2013).

c. *Temporoparietal junction (TPJ):* involved in sympathy and empathy through representing different possible perspectives on a single situation (Decety and Lamm, (2007). Also implicated in 'out of body experiences', where one's first-personal perspective occupies what is ordinarily a third-personal standpoint (Blanke et al., 2005). Also involved in mental time-travel and empathy with one's own future selves, viz. representing one's *own* perspective and emotional-affective reactions in possible future situations (Soutschek et al., (2016). Also involved in processing the order of events in time (Davis et al., 2009).
    a. Includes *Wernicke's area*, associated with 'inner monologue' (Shergill et al., 2001).
    b. Includes the *angular gyrus,* which processes attention to salient visual features of situations and mediates episodic memory retrieval to infer the intentions of other people (Seghier, 2013), and is involved in representing the mental states of individuals in cartoons and stories (Gallagher et al., 2000).
d. *Middle temporal gyrus (MTG)***:** involved in contemplating distance from oneself, facial-recognition, and word-meaning while reading (Acheson and Hagoort, 2013).
e. *Superior temporal sulcus (STC)***:** involved in social perception, including where others are gazing (viz. joint attention) and direction of others' emotions (Campbell et al., 1990).
f. *Middle occipital gyrus (MOG):* contains topographical maps of external world and engages in spatial processing (Renier et al., 2010).
g. *Temporal pole (TP):* involved in conceptual knowledge (Lambon Ralph et al., 2008), semantic memory of objects, people, words, and facts (Bonner and Price, 2013), and facial recognition, theory of mind, and visceral emotional responses (Olson et al., 2007).

h. *Fusiform gyrus (FG):* involved in facial and visual-word recognition (George et al., 1999; McCandliss et al., 2003).

i. *Inferior temporal gyrus (ITG):* involved in object recognition (Spiridon et al., 2006); and facial recognition (Meadows, 1974; Purves et al., 2001, p. 622).

j. *Precunus (PC):* a neural correlate of consciousness (Vogt and Laureys, 2005) involved in self-awareness (Kjaer et al., 2002), episodic memory (Lundstrom et al., 2003) including past-events affecting oneself (Lou et al., 2004), and visual imagery and attention, particularly representing other people's points-of-view (Vogeley et al., 2004), which has been implicated in empathy and forgiveness (Farrow et al., 2001).

Many of the same DMN regions are also implicated in *moral sensitivity***,** the capacity to monitor and recognize morally salient details of a given situation (Han, 2017, p. 98) However, the following additional DMN regions are also involved in moral sensitivity:

k. *Cingulate gyrus (CG)*: involved in emotion processing, memory, and learning, particularly the linking outcomes to motivational behavior (Hayden and Platt, 2010).

l. *Orbitofrontal cortex (OFC)*: processes cross-temporal (i.e. diachronic) contingencies and representation of the relative subjective value of outcomes (Fettes et al., 2017). Is also involved in processing reward and punishment, and learning from counterfactual prediction errors (Kringelbach and Rolls, 2004), as well as reversing behavior (Walton et al., 2010). Also involved in autonomic nervous system regulation including heartbeat and sexual arousal (Barbas 2007), and behavioral inhibition related to moral behavior (Fuster, 2001). Damage is known to produce extreme changes in personality, most famously associated with Phineas Gage, who dramatically transformed from a prudent and moral individual into a reckless person unable to resist morally base impulses (Harlow, 1848; Damasio et al., 1994).

m. *Lingual gyrus (LG)*: involved in visual processing in memories and dreams (Bogousslavsky et al., 1987), including memories of parts of faces (McCarthy et al., 1999).

n. *Cuneus:* involved in visual processing and behavioral inhibition (Haldane et al., 2008)., but also pathological gambling in those with high activity in the dorsal visual processing system (Crockford et al., 2005).

o. *Amygdala*: involved in long-term emotional-memory consolidation, specifically fear conditioning (Maren, 1999) but also positive, reward-based conditioning (Paton et al., 2006). Also implicated in decisionmaking involving fear, anger, sadness, and anxiety (Amunts et al., 2005), as well as in using emotional valence (positive or negative) to motivate behavior more generally (Nieh et al., 2013).

The behavioral neuroscience thus indicates that moral cognition involves a truly wide variety of capacities—capacities that, in the broadest sense, are capacities useful for *long-term planning*, in conformity with my theory of prudence and morality. I will now argue that *none* of the above capacities are distinctly 'moral' or inherently conducive to social cooperation, and that they were each plausibly selected in evolutionary history for *amoral* reasons: as capacities that enable fitness advantages irrespective of whether they are used to general moral actions conducive to social cooperation or not. Consequently, I will conclude that moral cognition is almost certainly not a biological adaptation for social cooperation.

4. The Diverse Evolutionary Advantages of Our Moral Capacities

As we have seen, at least seventeen brain regions and capacities are involved in moral judgment and sensitivity across a wide variety of tasks. I will now argue that (1) there is *good historical evidence* that different capacities involved in moral cognition emerged at different times in our evolutionary history, some long before the emergence of robust social cooperation; and (2) each brain region and capacity involved in moral cognition would have conferred *particular kinds of*

fitness advantages on our ancestors. These two types of facts together should enable us to pin down each brain region's likely *etiological function*, or reason why each region and associated capacities were selected and retained in evolutionary history (Millikan, 1989). This finally, should enable us to determine whether moral cognition is a biological adaptation for social cooperation.

Let us begin with *mental time-travel*, the capacity (associated with several regions of the Default Mode Network) to imaginatively simulate different possible pasts and futures. Mental time-travel is neither sufficient for moral cognition, nor plausibly 'for' social cooperation. Considered by itself, it is an amoral capacity: one that confers obvious fitness advantages on organisms irrespective of the moral status of their actions. This is because mental time-travel serves as a *long-term planning capacity*. It would have enabled our ancestors to imaginatively recall the effects of their past actions—such as which types of plants are poisonous, and how to catch prey—and to imagine different possible future outcomes for their actions (such as what will happen if one eats a particular plant in the future). None of these obvious fitness advantages (avoiding poisonous things, solving problems, etc.) concern social cooperation *per se*: they merely would have enabled out distant ancestors to plan more effectively in general. Second, although people can learn how to use mental time-travel in distinctly moral ways conducive to social cooperation (see Arvan, 2020, chapters 2 and 3), mental time-travel equally enables individuals—and would have enabled our distant ancestors—to harmfully exploit other people, contrary to morality and cooperation. This is true even today. Consider a capitalist exploiting sweatshop labor, a tyrannical dictator maintaining their power through mass murder, or a spouse engaging in infidelity. All of these immoral actions are enabled by mental time-travel, as the ability to imagine different possible futures enables people to plan how to harm others for one's own personal advantage. Third, the kinds of immoral behavior mental time-travel can give rise to plausibly generated *fitness advantages* for our ancestors, as those who gain or maintain

power through immoral means (e.g. despots, warlords, etc.) can plausibly sire more offspring than those they dominate or murder. Fourth, mental time-travel appears to have emerged in evolutionary history *far before* evidence of robust social cooperation. Mental time-travel appears to have emerged at least 400,000 years ago (Suddendorf and Corballis 2007, p. 312), as it appears necessary for inventing complex tools and using fire, both of which archaeological evidence suggests first emerged during the middle Pleistocene period (Hallos, 2005; Boaz et al., 2004). Robust forms of social cooperation (e.g. stable groups and societies), on the other hand, appear to have emerged only in the last 200,000 years (Apicella and Silk, 2019). Fifth, the use of complex tools, fire and social cooperation all appear to presuppose the development of *norms*, which requires normative capacities to represent how things should or shouldn't be done (Braddock and Rosenberg, 2017, pp. 65-71. Cf. Arvan, 2020, Chapter 2 and Hobbes 1651). Yet, the capacity to flexibly extend normative judgments to new cases requires recursion, which appears to have emerged in our ancestral history between 150-200 *million* years ago (Barceló-Coblijn, 2012, especially p. 178)—far before the emergence of mental time-travel *or* robust social cooperation. Given that mental time-travel (A) is demonstrably critical to moral cognition, (B) afforded our ancestors plausible fitness advantages *independent of and prior to* social cooperation, and (C) using it in pro-social, cooperative ways appears to be predicated upon the development and transmission of *sociocultural norms* (Arvan 2020, Chapter 2) that emerged only the past 200,000 years or perhaps even in the last 50,000 years (Kitcher, 2011, p. 97, fn. 37), it follows that a central feature of moral cognition—mental time-travel—was likely not selected for in evolutionary history as a biological adaptation for social cooperation.

Now turn to *other-perspective taking* (OPT). When combined with well-developed capacities for empathy—the kind we are socialized across childhood, adolescence, and adulthood to engage in (via. mental time-travel)—OPT plays a central role in moral cognition

(see Arvan 2020, chapters 2-4). However, is that what the capacity evolved *for*? On its own, OPT is most plausibly construed, just like mental time-travel, as a *planning* capacity. It enables us (and would have enabled our ancestors) to understand how other people experience a single situation, including our role in that situation and how they might react to our actions. To this extent, OPT is clearly *not* a 'moral capacity.' Being able to understand other people's perspectives is something that we can equally use to *exploit* them—as a con man does when he exploits other people's trust for personal gain. Further, it is easy to imagine numerous ways in which other-perspective-taking would confer fitness-advantages upon individuals (and populations they are part of) regardless of whether it is used in moral or immoral ways. Again, consider infidelity, an individual-level behavior that can increase the fitness of the individual who engages in it by enabling them to sire a larger number of offspring. Other-perspective-taking can be used to enable infidelity by enabling the person to understand and exploit the other person's trust. Consequently, OPT is also unlikely to have been biologically selected in our evolutionary history 'for' social cooperation: it plausibly increased the fitness of our ancestors when used in moral *and* immoral ways.

Now turn to *risk-aversion* and the underlying neurobiology that leads people to overweight negative outcomes relative to positive outcomes. As noted in Section 2, on my account of morality risk-aversion plays a central role in moral cognition—at least when we are socialized to avoid risking violating moral norms. Yet, as others have pointed out, risk-aversion *per se* is not a moral capacity: rather, it is something that helps a person *preserve themselves*, enabling them to survive, bear offspring, and so on (Viganò, 2017, pp. 218, 222).

Now turn to the *ventromedial prefrontal cortex* (*vmPFC)*, which is associated with processing risk and uncertainty. On my model, moral cognition emerges out of *prudential calculations* of risk and reward. Prudent individuals learn through socialization to engage in

forms of mental time-travel and other-perspective-taking that make conformity to moral principles rational. The vmPFC thus plays a clear role in moral cognition on this picture. However, the vmPFC is *clearly* not a 'moral capacity' in and by itself. To see how, consider the behavior of warlords and gangs in failed states or the behavior of religious extremists (such as the members of ISIS/Islamic State). These individuals can have fully functioning *vmPFC*'s, weighing risks and rewards. Yet, many of them appear to lack a moral conscience, and are instead willing to murder and oppress others with abandon. Why? Because, given their environment, they have learned they can personally benefit from it, obtaining more resources for themselves, siring more survivable offspring than those they murder or oppress. The difference between people of 'moral conscience'—people who engage in *moral cognition*—and warlords or political dictators (who don't) thus appears to be environmental and a matter of *learning and reasoning*. You and I have learned through socialization to care about how our actions might negatively affect others. Dictators and religious extremists learn the opposite: that they can *benefit* from using the *vmPFC* in immoral ways. Consequently, the vmPFC was not plausibly selected in evolutionary history for social cooperation either. It offers and would have offered our ancestors plausible fitness advantages both to those who cooperate in prosocial ways, but also to individuals who use its associated capacities in immoral, anti-cooperative ways for their own reproductive benefit.

Now consider the *dorsomedial prefrontal cortex*, which is involved the sense of the self and theory of mind (i.e. understanding the mental states of others), and the *temporoparietal junction (TPJ)*, which is involved in sympathy and empathy, specifically the ability to represent different possible perspectives on a single situation, as well as 'out of body experiences' and processing the order of events in time. Are these 'moral capacities'? Although they are involved in sympathy and empathy, they do so by way of a *general mechanism*: the capacity to represent

a single situation from multiple perspectives. So we need to ask: is that mechanism, by itself, a 'moral' capacity? The answer is no. Being able to understand other people's mental states and appreciate many different perspectives on a single situation is a *predictive planning* capacity—one that enables us to predict how others around us will respond to our actions. Well-socialized individuals learn to use this capacity to *empathize* with others. However, dictators, wanton criminals, or warlords learn differently. They learn to use the capacity to represent the same situation from many different perspectives *to exploit or murder people*. For example, Hitler's capacity to represent multiple perspectives plausibly enabled him to take advantage of Neville Chamberlain. Chamberlain bet that the Munich Pact would appease Hitler—yet Hitler understood and exploited these expectations to do the opposite, enabling the Nazis' murderous march across Europe.

Now consider other brain regions implicated in moral cognition:

k. *Middle temporal gyrus (MTG)***:** involved in contemplating distance from oneself, facial-recognition, and word-meaning while reading.

l. *Superior temporal sulcus (STC)***:** involved in social perception, including where others are gazing (viz. joint attention) and direction of others' emotions.

m. *Middle occipital gyrus (MOG):* contains topographical maps of external world and engages in spatial processing.

n. *Temporal pole (TP):* involved in conceptual knowledge, semantic memory of objects, people, words, and facts, facial recognition, theory of mind, and visceral emotional responses.

o. *Fusiform gyrus (FG):* involved in facial and visual-word recognition.

None of the capacities associated with these brain regions are distinctly 'moral' capacities plausibly selected in evolutionary history for social cooperation. Rather, they are perceptual and

conceptual capacities—the kinds of capacities that enable well-socialized individuals like you and I use to construct moral principles in language and thought, apply them to concrete situations in, and so on: capacities which almost certainly provided a wide variety of non-moral fitness-advantages to our ancestors.

In sum, moral cognition involves a wide variety of different brain regions and associated capacities ranging from long-term planning capacities to perceptual capacities. Second, because evolution is a gradual process, those brain regions were likely selected at different times in evolutionary history for different reasons. Third, as we have seen, some capacities central to moral cognition—including mental time-travel, other-perspective-taking, and recursion—would have offered our ancestors fitness advantages *irrespective* of social cooperation. Fourth, some of the above capacities (mental time-travel, recursion, etc.) appear to have emerged in our evolutionary history long before social cooperation. Fifth, the theory of moral cognition that I have argued best explains the behavioral neuroscience holds that moral cognition is something we *learn* to do through socialization.

5. Conclusion

These facts indicate that moral cognition is unlikely to be a biological adaptation for social cooperation. First, moral cognition involves a variety of long-term planning capacities that would have conferred fitness-advantages on our ancestors irrespective of whether our ancestors used those capacities for cooperation. Second, moral cognition (viz. social cooperation) is something people *learn* to do via individual-level rationality and socialization. Stable groups and large-scale societies were the result of individuals learning to cooperate in our ancestral past (Braddock and Rosenberg, 2012). These groups then developed, transmitted, and enforced norms that *socialize* individuals to grasp the rationality of obeying moral principles (Arvan 2020, chapters 2 and 3). If this is correct, then moral cognition is not a biological adaptation for social cooperation but

instead an *exaptation*. Indeed, moral cognition is not a biological capacity at all. Rather, it is something we *learn* to engage in as a result of individual-level reasoning and socialization—processes that put capacities selected in evolutionary history for many reasons to a novel, prosocial use.

References

Acheson, D. J., & Hagoort, P. (2013). Stimulating the brain's language network: syntactic ambiguity resolution after TMS to the inferior frontal gyrus and middle temporal gyrus. *Journal of Cognitive Neuroscience*, *25*(10), 1664-77.

Alexander, R. D. (1987). *The biology of moral systems*. Transaction.

Amunts, K., Kedo, O., Kindler, M., Pieperhoff, P., Mohlberg, H., Shah, N. J., ... & Zilles, K. (2005). Cytoarchitectonic mapping of the human amygdala, hippocampal region and entorhinal cortex: intersubject variability and probability maps. *Anatomy and embryology*, *210*(5-6), 343-52.

Anand, P. (1995). *Foundations of rational choice under risk*. Oxford University Press.

Apicella, C .L., & Silk, J. B. (2019). The evolution of human cooperation. *Current Biology*, *29*(11), R447-R450.

Aristotle [1984]. *Nicomachean Ethics*. In J. Barnes (Ed.), *The complete works of Aristotle: The revised Oxford translation*. Princeton University Press.

Arvan, M. (2020). *Neurofunctional prudence and morality: A philosophical theory*, Routledge.

Arvan, M. (2019). The dark side of morality: Group polarization and moral epistemology. *The Philosophical Forum*, *50*(1), 87-115.

Arvan, M. (2016). *Rightness as fairness: A moral and political theory*, Palgrave MacMillan.

Audi, R. (2015). Intuition and its place in ethics. *Journal of the American Philosophical Association 1*(1), 57-77.

Barbas, H. (2007). Flow of information for emotions through temporal and orbitofrontal pathways. *Journal of Anatomy*, *211*(2), 237-49.

Barceló-Coblijn, L. (2012). Evolutionary scenarios for the emergence of recursion. *Theoria et Historia Scientarium*, *Vol IX*, 171-99.

Barragan, R. C., Brooks, R. & Meltzoff, A .N. (2020). Altruistic food sharing behavior by human infants after a hunger manipulation. *Scientific Reports, 10*(1785) https://doi.org/10.1038/s41598-020-58645-9.

Baskin-Sommers, A., Stuppy-Sullivan, A. M., & Buckholtz, J. W. (2016). Psychopathic individuals exhibit but do not avoid regret during counterfactual decision making. *Proceedings of the National Academy of Sciences*, *113*(50), 14438-43.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323-70.

Bechara, A., Tranel, D., & Damasio, H. (2000). Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain*, *123*(11), 2189-202.

Benoit, R. G., Gilbert, S. J., Burgess, P. W. (2011). A neural mechanism mediating the impact of episodic prospection on farsighted decisions. *The Journal of Neuroscience, 31*(18), 6771–79.

Blair, R. J. R. (2003). Neurobiological basis of psychopathy. *The British Journal of Psychiatry, 182*(1), 5-7.

Blanke, O., Mohr, C., Michel, C. M., Pascual-Leone, A., Brugger, P., Seeck, M., Landis, T. & Thut, G. (2005). Linking out-of-body experience and self processing to mental own-body imagery at the temporoparietal junction. *Journal of Neuroscience*, *25*(3), 550-57.

Boaz, N. T., Ciochon, R. L., Xu, Q. & Liu, J. (2004) Mapping and taphonomic analysis of the Homo erectus loci at Locality 1 Zhoukoudian, China. *Journal of Human Evolution, 46*(5), 519–49.

Bonner, M. F., & Price, A. R. (2013). Where is the anterior temporal lobe and what does it do?. *Journal of Neuroscience*, *33*(10), 4213-15.

Bogousslavsky, J., Miklossy, J., Deruaz, J. P., Assal, G., & Regli, F. (1987). Lingual and fusiform gyri in visual processing: a clinico-pathologic study of superior altitudinal hemianopia. *Journal of Neurology, Neurosurgery & Psychiatry*, *50*(5), 607-14.

Braddock, M., & Rosenberg, A. (2012). Reconstruction in moral philosophy?. *Analyse & Kritik*, *34*(1), 63-80.

Brennan, J. (2010). Scepticism about philosophy. *Ratio, 23*(1), 1-16.

Bricker, P. (1980). Prudence. *The Journal of Philosophy*, *77*(7), 381–401.

Brosnan, S. F. (2006). Nonhuman species' reactions to inequity and their implications for fairness. *Social Justice Research*, *19*(2), 53-185.

Brosnan, S. F., & de Waal, F. B. (2014). Evolution of responses to (un)fairness. *Science, 346*(6207), 1251776.

Bruckner, D. (2003). A contractarian account of (part of) prudence. *American Philosophical Quarterly, 40*(1), 33-46.

Campbell, R., Heywood, C. A., Cowey, A., Regard, M., & Landis, T. (1990). Sensitivity to eye gaze in prosopagnosic patients and monkeys with superior temporal sulcus ablation. *Neuropsychologia*, *28*(11), 1123-1142.

Carruthers, P., & James, S. M. (2008). Evolution and the possibility of moral realism. *Philosophy and Phenomenological Research*, *77*(1), 237-44.

Casebeer, W. D. (2003). *Natural ethical facts: Evolution, connectionism, and moral cognition*. MIT Press.

Casey, B. J., Jones, R. M., Hare, T. A. (2008). The adolescent brain. *Annals of the New York Academy of Sciences, 1124*, 111-26.

Churchland, P. (2011). *Braintrust: What neuroscience tells us about morality*. Princeton University Press.

Corballis, M. C. (2007). The uniqueness of human recursive thinking: the ability to think about thinking may be the critical attribute that distinguishes us from all other species. *American Scientist, 95*(3), 240-8.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (pp. 163–228). Oxford University Press.

Crockford, D. N., Goodyear, B., Edwards, J., Quickfall, J., & el-Guebaly, N. (2005). Cue-induced brain activity in pathological gamblers. *Biological psychiatry*, *58*(10), 787-95.

Curry, O. S. (2016). Morality as cooperation: a problem-centred approach. In T.K. Shackelford & D. Hansen (Eds.), *The Evolution of Morality* (pp. 27-51). New York: Springer, Cham.

Damasio, H., Grabowski, T., Frank, R., Galaburda, A. M., & Damasio, A. R. (1994). The return of Phineas Gage: clues about the brain from the skull of a famous patient. *Science*, *264*(5162), 1102-5.

Daniel, T. O., Stanton, C. M., & Epstein, L. H. (2013). The future is now: comparing the effect of episodic future thinking on impulsivity in lean and obese individuals. *Appetite*, *71*(1), 120-5.

Davis, B., Christie, J., & Rorden, C. (2009). Temporal order judgments activate temporal parietal junction. *Journal of Neuroscience*, *29*(10), 3182-8.

Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. *The Neuroscientist*, *13*(6), 580-593.

de Waal, F. (2006). *Primates and philosophers: How morality evolved*. Princeton University Press.

Dietrich, E. (2011). There is no progress in philosophy. *Essays in Philosophy*, *12*(2), Article 9, 329-44.

Doğruyol, B., Alper, S., & Yilmaz, O. (2019). The five-factor model of the moral foundations theory is stable across WEIRD and non-WEIRD cultures. *Personality and Individual Differences*, *151*, 109547.

Ersner-Hershfield, H., Garton, M. T., Ballard, K., Samanez-Larkin, G. R., Knutson, B. (2009). Don't stop thinking about tomorrow: individual differences in future self-continuity account for saving. *Judgment and Decision Making, 4*, 280-6.

Ersner-Hershfield, H., Wimmer, G. E., Knutson, B. (2009). Saving for the future self: neural measures of future self-continuity predict temporal discounting. *Social Cognitive and Affective Neuroscience*, 4(1): 85-92.

Farrow, T. F., Ying Zheng, Y., Wilkinson, I. D., Spence, S. A., Deakin, J. F., Tarrier, N., … Woodruff, P. W. (2001). Investigating the functional anatomy of empathy and forgiveness. *Neuroreport, 12*(11), 2433-8.

Fellows, L. K., & Farah, M. J. (2007). The role of ventromedial prefrontal cortex in decision making: judgment under uncertainty or judgment per se?. *Cerebral Cortex*, *17*(11), 2669-74.

Fettes, P., Schulze, L., & Downar, J. (2017). Cortico-striatal-thalamic loop circuits of the orbitofrontal cortex: promising therapeutic targets in psychiatric illness. *Frontiers in systems neuroscience*, *11*(25), 1-23.

Frank, R. H. (1988). *Passions Within Reason: The Strategic Role of the Emotions*. WW Norton & Co.

Fuster, J. M. (2001). The prefrontal cortex—an update: time is of the essence. *Neuron*, *30*(2), 319-33.

Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia*, *38*(1), 11-21.

George, N., Dolan, R. J., Fink, G. R., Baylis, G. C., Russell, C., & Driver, J. (1999). Contrast polarity and face recognition in the human fusiform gyrus. *Nature neuroscience*, *2*(6), 574-80.

Geraci, A., & Surian, L. (2011). The developmental roots of fairness: infants' reactions to equal and unequal distributions of resources*. Developmental Science*, *14*, 1012–20.

Giedd, J. N., Blumenthal, J., Jeffries, N. O. (1999). Brain development during childhood and adolescence: a longitudinal MRI study. *Nature Neuroscience, 2*(10), 861-3.

Gould, S. J. (1991). Exaptation: A crucial tool for an evolutionary psychology. *Journal of Social Issues*, *47*(3), 43-65.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: the pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology* (Vol. 47, pp. 55-130). Academic Press.

Greene, J. D. (2015). The rise of moral cognition. *Cognition*, *135*, 39-42.

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin Books.

Gusnard, D. A., Akbudak, E., Shulman, G. L., & Raichle, M. E. (2001). Medial prefrontal cortex and self-referential mental activity: relation to a default mode of brain function. *Proceedings of the National Academy of Sciences*, *98*(7), 4259-4264.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Pantheon Books.

----- (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment, *Psychological Review*, *108*(4), 814-34.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4), 55-66.

Haldane, M., Cunningham, G., Androutsos, C., & Frangou, S. (2008). Structural brain correlates of response inhibition in Bipolar Disorder I. *Journal of Psychopharmacology*, *22*(2), 138-43.

Hallos, J. (2005). "15 minutes of fame:" Exploring the temporal dimension of Middle Pleistocene lithic technology. *Journal of Human Evolution, 49*, 155 –79.

Hamlin, J. K. (2013). Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core. *Current Directions in Psychological Science*, *22*(3), 186-93.

Han, H. (2017). Neural correlates of moral sensitivity and moral judgment associated with brain circuitries of selfhood: a meta-analysis. *Journal of Moral Education, 46*(2), 97-113.

Hansson, S. O. (2005). *Decision Theory: A Brief Introduction*, https://people.kth.se/~soh/decisiontheory.pdf, retrieved 7 December 2020.

Hare, R. D. (1999). *The Hare Psychopathy Checklist-Revised: PLC-R*. Multi-Health Systems.

Harlow, J. M. (1848). Passage of an iron rod through the head. *The Boston Medical and Surgical Journal (1828-1851)*, *39*(20), 0_1.

Hart, S. D. & Dempster, R. J. (1997). Impulsivity and Psychopathy. In C. D. Webster, M. A. Jackson (Eds.), *Impulsivity: Theory, Assessment, and Treatment* (pp. 212-32), The Guilford Press.

Haybron, D. (2011). Happiness. *The Stanford Encyclopedia of Philosophy*. E. N. Zalta (Ed.), https://plato.stanford.edu/archives/fall2011/entries/happiness/.

Hayden, B. Y., & Platt, M. L. (2010). Neurons in anterior cingulate cortex multiplex information about reward and action. *Journal of Neuroscience*, *30*(9), 3339-46.

Henrich, N., & Henrich, J. (2007). *Why humans cooperate: A cultural and evolutionary explanation*. Oxford University Press.

Hershfield, H. E., Cohen, T. R., & Thompson, L. (2012). Short horizons and tempting situations: lack of continuity to our future selves leads to unethical decision making and behavior. *Organizational Behavior and Human Decision Processes*, *117*, 298–310.

Hershfield, H. E., Goldstein, D. G., Sharpe, W. F., Fox, J., Yeykelis, L., Carstensen, L.L., & Bailenson, J. N. (2011). Increasing saving behavior through age-progressed renderings of the future self. *Journal of Marketing Research: November* 2011, *48*(SPL), S23-S37.

Hitler, A. (1925). *Mein Kampf*. Fairborne Publishing.

Hobbes, T. [1658]. *De Homine*. In B. Gert (ed.), *Man and Citizen*. Anchor Books, 1972.

----- [1651]. *Leviathan*, in Sir W. Molesworth (Ed.), *The English Works of Thomas Hobbes: Now First Collected and Edited* (Vol. 3, pp. ix-714), John Bohn, 1839-45.

Hosking, J. G., Kastman, E. K., Dorfman, H. M., Samanez-Larkin, G. R., Baskin-Sommers, A., Kiehl, K. A., ... & Buckholtz, J. W. (2017). Disrupted prefrontal regulation of striatal subjective value signals in psychopathy. *Neuron*, *95*(1), 221-31.

Isoda, M., & Noritake, A. (2013). What makes the dorsomedial frontal cortex active during reading the mental states of others?. *Frontiers in Neuroscience*, *7*, 232.

Ito, T. A., Larsen, J. T., Smith, N. K., Cacioppo, J. T. (1998). Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology, 75*(4), 887-900.

Joyce, R. (2007). *The myth of morality*. Cambridge University Press.

Joyce, R. (2006). Metaethics and the empirical sciences. *Philosophical Explorations*, *9*(1), 133-48.

Kabadayi, C., & Osvath, M. (2017). Ravens parallel great apes in flexible planning for tool-use and bartering. *Science*, *357*(6347), 202-4.

Kahneman, D. & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 4, 263–291.

Kant, I. [1785]. *Groundwork of the metaphysics of morals*, in M. J. Gregor (Ed.), *The Cambridge edition of the works of Immanuel Kant: Practical philosophy* (pp. 38-108). Cambridge University Press, 1996.

Kennett, J. & Matthews, S. (2009). Mental timetravel, agency and responsibility. In M. Broome and L. Bortolotti (Eds.), *Psychiatry as cognitive neuroscience: Philosophical perspectives* (pp. 327-50). Oxford: Oxford University Press.

Kitcher, P. (2011). *The Ethical Project*. Harvard University Press.

Kitcher, P. (2005). Biology and ethics. In D. Copp (Ed.). *The Oxford Handbook of Ethical Theory* (pp. 163-85). Oxford University Press.

Kitcher, P. (1998). Psychological altruism, evolutionary origins, and moral rules. *Philosophical Studies, 89*(2-3), 283-316.

Kjaer, T. W., Nowak, M., & Lou, H. C. (2002). Reflective self-awareness and conscious states: PET evidence for a common midline parietofrontal core. *Neuroimage*, *17*(2), 1080-6.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, *446*(7138), 908-11.

Korsgaard, C. M. (2009). *Self-constitution: Agency, identity, and integrity*. Oxford University Press.

Korsgaard, C. M. (2008). *The constitution of agency*. Oxford University Press.

Kringelbach, M. L., & Rolls, E. T. (2004). The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. *Progress in Neurobiology*, *72*(5), 341-72.

Lambon Ralph, M. A., Pobric, G., & Jefferies, E. (2008). Conceptual knowledge is underpinned by the temporal pole bilaterally: convergent evidence from rTMS. *Cerebral Cortex*, *19*(4), 832-8.

Levy, N. (2007). The responsibility of the psychopath revisited. *Philosophy, Psychiatry, and Psychology, 14*(2). 129-38.

Litton, P. (2008). Responsibility status of the psychopath: on moral reasoning and rational self-governance. *Rutgers Law Journal*, *39*(349), 350-92.

Lou, H. C., Luber, B., Crupain, M., Keenan, J. P., Nowak, M., Kjaer, T. W., Sackeim, H. A. & Lisanby, S. H. (2004). Parietal cortex and representation of the mental self. *Proceedings of the National Academy of Sciences*, *101*(17), 6827-32.

Luco, A. C. (2016). Non-negotiable: Why moral naturalism cannot do away with categorical reasons. *Philosophical Studies, 173*(9), 2511-28.

Lundstrom, B. N., Petersson, K. M., Andersson, J., Johansson, M., Fransson, P., & Ingvar, M. (2003). Isolating the retrieval of imagined pictures during episodic memory: activation of the left precuneus and left prefrontal cortex. *Neuroimage*, *20*(4), 1934-43.

Maren, S. (1999). Long-term potentiation in the amygdala: a mechanism for emotional learning and memory. *Trends in Neurosciences*, *22*(12), 561-7.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean?. *American Psychologist*, *70*(6), 487-98.

May, J. (2018). *Regard for reason in the moral mind*. Oxford University Press.

McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences*, *7*(7), 293-9.

McCarthy, G., Puce, A., Belger, A., & Allison, T. (1999). Electrophysiological studies of human face perception. II: Response properties of face-specific potentials generated in occipitotemporal cortex. *Cerebral cortex*, *9*(5), 431-44.

Meadows, J. C. (1974). The anatomical basis of prosopagnosia. *Journal of Neurology, Neurosurgery & Psychiatry*, *37*(5), 489-501.

Milad, M. R., Quinn, B. T., Pitman, R. K., Orr, S. P., Fischl, B., & Rauch, S. L. (2005). Thickness of ventromedial prefrontal cortex in humans is correlated with extinction memory. *Proceedings of the National Academy of Sciences*, *102*(30), 10706-11.

Millikan, R. G. (1989). In defense of proper functions. *Philosophy of Science 56*(2), 288-302.

Moffitt, T. E. (1993). Adolescence-limited and life-course persistent antisocial behavior: a developmental taxonomy. *Psychological Review*, *100*, 674–701.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., … & Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences*, *108*(7), 2693-8.

Motzkin, J. C., Newman, J. P., Kiehl, K. A., & Koenigs, M. (2011). Reduced prefrontal connectivity in psychopathy. *Journal of Neuroscience*, *31*(48), 17348-57.

Nieh, E. H., Kim, S. Y., Namburi, P., & Tye, K. M. (2013). Optogenetic dissection of neural circuits underlying emotional valence and motivated behaviors. *Brain research*, *1511*, 73-92.

Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007). The enigmatic temporal pole: a review of findings on social and emotional processing. *Brain*, *130*(7), 1718-31.

Parfit, D. (2011). *On what matters* (Vols. 1&2). Oxford University Press.

Pascual, L., Gallardo-Pujol, D., & Rodrigues, P. (2013). How does morality work in the brain? A functional and structural perspective of moral behavior. *Frontiers in Integrative Neuroscience*, *7*(65), 1-8.

Paton, J. J., Belova, M. A., Morrison, S. E., & Salzman, C. D. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature*, *439*(7078), 865-70.

Peters, J. & Büchel, C. (2010). Episodic future thinking reduces reward delay discounting through an enhancement of prefrontal-mediotemporal interactions. *Neuron*, *66*(1), 138–48.

Peterson, M. (2017). *An introduction to decision theory*, 2nd edition. Cambridge University Press.

Pettigrew, R. (2020). *Choosing for changing selves*. Oxford University Press.

Price, B. W. (2002). The worthwhileness theory of the prudentially rational life. *Journal of Philosophical Research*, *27*, 619–39.

Prichard, H. A. (1912). Does moral philosophy rest on a mistake?. *Mind*, *21*(81), 21-37.

Prinz, J. J. (2007). *The Emotional Construction of Morals*. Oxford University Press.

Purves D., Augustine G. J., Fitzpatrick D., Hall, W. C., LaMantia, A., McNamara, J. O., Williams, S. M. [Eds.] (2001). *Neuroscience*. 2nd edition. Sinauer Associates.

Renier, L. A., Anurova, I., De Volder, A. G., Carlson, S., VanMeter, J., & Rauschecker, J. P. (2010). Preserved functional specialization for spatial processing in the middle occipital gyrus of the early blind. *Neuron*, *68*(1), 138-48.

Ross, W. D. [1930]. *The right and the good*. Oxford University Press, 2002.

Russell, B. (1945). *A history of Western philosophy*. Simon and Schuster.

Scanlon, T. M. (2014). *Being realistic about reasons*. Oxford University Press.

Scanlon, T. M. (1998). *What we owe to each other*. Harvard University Press.

Schmidt, M. F., & Sommerville, J. A. (2011). Fairness expectations and altruistic sharing in 15-month-old human infants. PloS one, *6*(10), e23223.

Seghier, M. L. (2013). The angular gyrus: multiple functions and multiple subdivisions. *The Neuroscientist*, *19*(1), 43-61.

Shergill, S. S., Bullmore, E. T., Brammer, M. J., Williams, S. C. R., Murray, R. M., & McGuire, P. K. (2001). A functional study of auditory verbal imagery. *Psychological Medicine*, *31*(2), 241-53.

Shoemaker, D. W. (2011). Psychopathy, responsibility, and the moral/conventional distinction. *Southern Journal of Philosophy*, *49*(s1), 99-124.

Sinclair, N. (2012). Metaethics, teleosemantics and the function of moral judgments. *Biology and Philosophy*, *27*(5): 639-62.

Singer, T. & Lamm, C. (2009). The social neuroscience of empathy. *Annals of the New York Academy of Sciences* 1156, 81–96.

Singer, T. & Tusche. A. (2014). Understanding others: brain mechanisms of theory of mind and empathy. In P.W. Glimcher, Ernst Fehr (Eds.), *Neuroeconomics: Decision Making and the Brain*, 2nd edition (pp. 249-66). Academic Press.

Smyth, N. (2017). The function of morality. *Philosophical Studies* 174(5), 1127-44.

Sommer, M., Meinhardt, J., Rothmayr, C., Döhnel, K., Hajak, G., Rupprecht, R., & Sodian, B. (2014). Me or you? Neural correlates of moral reasoning in everyday conflict situations in adolescents and adults. *Social Neuroscience*, *9*(5), 452-70.

Soutschek, A., Ruff, C. C., Strombach, T., Kalenscher, T., & Tobler, P. N. (2016). Brain stimulation reveals crucial role of overcoming self-centeredness in self-control. *Science Advances*, *2*(10): e1600992.

Spiridon, M., Fischl, B., & Kanwisher, N. (2006). Location and spatial profile of category-specific regions in human extrastriate cortex. *Human brain mapping*, *27*(1), 77-89.

Sterelny, K., & Fraser, B. (2016). Evolution and Moral Realism. *The British Journal for the Philosophy of Science*. *68*(4), 981-1006.

Stuss D. T., Gow, C. A., Hetherington, C. R. (1992). 'No longer Gage': frontal lobe dysfunction and emotional changes. *Journal of Consulting and Clinical Psychology, 60*(3): 349-59.

Suddendorf, T. & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences, 30*(3), 299-313.

Suhler, C.L. & Churchland, P. (2011). Can innate, modular "foundations" explain morality? Challenges for Haidt's moral foundations theory. *Journal of Cognitive Neuroscience, 23*(9): 2103-16.

Tomasello, M., & Vaish, A. (2013). Origins of human cooperation and morality. *Annual Review of Psychology*, *64*, 231-55.

Tosi, J., & Warmke, B. (2016). Moral Grandstanding. *Philosophy & Public Affairs*, *44*(3), 197-217.

Trueman, C. N. (2020). The Fuhrer Principle. *History Learning Site*, https://www.historylearningsite.co.uk/nazi-germany/the-fuehrer-principle/, retrieved 7 December 2020.

Van Gelder, J. L., Hershfield, H. E., & Nordgren, L. F. (2013). Vividness of the future self predicts delinquency. *Psychological Science*, *24*(6), 974-80.

Viganò, E. (2017). Adam Smith's theory of prudence updated with neuroscientific and behavioral evidence. *Neuroethics, 10*(2), 215-33.

Vogeley, K., May, M., Ritzl, A., Falkai, P., Zilles, K., & Fink, G. R. (2004). Neural correlates of first-person perspective as one constituent of human self-consciousness. *Journal of cognitive neuroscience*, *16*(5), 817-27.

Vogt, B. A., & Laureys, S. (2005). Posterior cingulate, precuneal and retrosplenial cortices: cytology and components of the neural network correlates of consciousness. *Progress in brain research*, *150*, 205-17.

Walton, M. E., Behrens, T. E., Buckley, M. J., Rudebeck, P. H., & Rushworth, M. F. (2010). Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. *Neuron*, *65*(6), 927-39.

Weber, S., Habel, U., Amunts, K., Schnieder, F. (2008). Structural brain abnormalities in psychopaths—a review, *Behavioral Sciences & the Law, 26*(1), 7-28.

Wisdom, J. (2017). Proper-function moral realism. *European Journal of Philosophy*, *25*(4), 1660-74.

Wood, A. (2008). *Kantian ethics*. Cambridge University Press.

Yang, Y., Raine, A. (2009). Prefrontal structural and functional brain imaging findings in antisocial, violent, and psychopathic individuals: a meta-analysis. *Psychiatry Research, 174*(2), 81-8.