25 Moral Reasoning: A Network Neuroscience Perspective

Evan D. Anderson and Aron K. Barbey

Introduction

Humans are moral creatures – uniquely capable of drawing on beliefs about morality in order to distinguish between actions that are right or wrong. Neuroscience research has investigated the neural substrates that underlie this reasoning process, identifying multiple regions of the brain that are recruited to facilitate performance on moral reasoning tasks. Current evidence indicates that reasoning about morality recruits a distributed collection of brain areas functioning variously in conflict or agreement (Pascual, Rodrigues, and Gallardo-Pujol, 2013).

Recently, dual-process frameworks have been employed to operationalize moral reasoning in terms of two distinct cognitive systems (Crockett, 2013; Cushman, 2013; Greene, 2017; Greene et al., 2001). This dual-process framework posits that, when faced with a moral dilemma, individuals will predominantly rely on one of two reasoning systems (heuristic, or model-based) to produce deontological or consequentialist moral judgments. Fundamental questions remain, however, as to whether these two reasoning processes depend on distinct neurobiological substrates, and to what extent moral cognitions may result from dynamic interactions between these two cognitive systems. To expand our understanding of how judgments about morality might depend on the brain's systems and organization, this chapter considers moral reasoning in terms of the topology and dynamics of brain networks.

The emerging field of network neuroscience uses tools and methods from the network sciences to characterize the topology and dynamics of brain networks (see Bassett and Sporns, 2017; Bressler and Menon, 2010). Regions of neural tissue are defined as topological elements (nodes) in a network, and physiological signatures are measured and used to infer their relationships (edges). Evidence from this literature suggests that the brain is a complex, self-organized structure constructed to be at once highly efficient, flexible, and specialized.

Network neuroscience's set of conceptual and empirical tools affords the opportunity to investigate neural circuits for moral reasoning at a new level of resolution – that of communication and interactions occurring within and between large-scale functional brain networks (see Chiong et al., 2013; Young and Saxe, 2008). Current neuroscience accounts recognize that moral reasoning depends on relationships between a number of regions and circuits in the brain, but have yet to address the

relevance of these functional interrelationships to the integration of reasoning systems within the dual-process framework. A network neuroscience perspective affords the opportunity to theorize about the underlying topology and dynamics of network interactions that mediate either heuristic or model-focused processing. In turn, better characterizing the neurobiological mechanisms of these systems provides guidance for future research into how individuals reason when they need to deploy elements of both processing systems.

In this chapter, we review evidence for dual-process accounts of moral reasoning from a network neuroscience perspective. Neuroscientific evidence suggests that moral reasoning (consequentialist, and deontological ethics) depends on activity in distinct, functionally interacting brain networks (frontoparietal and limbic; Chiong et al, 2013, Jeurissen et al., 2014; Moll et al., 2008). Evidence also suggests that the ability to combine processing from these two systems is specifically dependent on activity in a particular anatomical region (ventromedial prefrontal cortex; Ciaramelli et al., 2007; Shenhav and Greene, 2014). This process of integration involves simulating (imagining) a hypothetical model of the world, and is mediated by activity in a third brain network, the default-mode (Chiong et al., 2013; Greene et al., 2001). All three of these functional networks share a single network hub (ventromedial prefrontal cortex), and research suggests that interactions between that hub and those networks serve to influence the ultimate form of reasoning (consequentialist or deontological) we tend to rely on when forming a moral judgment (e.g. Cima, Tonnaer, and Hauser, 2010; Friesdorf, Conway, and Gawronski, 2015; Koenigs et al., 2007). We begin by discussing moral reasoning itself, and then review the literature on dual-process theories, and on the neuroscience of moral reasoning. We then review network neuroscientific evidence that bears on a dual-process theory of moral reasoning and discuss implications for future research into its cognitive and neural foundations.

The Trolley Problem

Imagine a train leaving the station at thirty-two miles an hour, headed along a straight path. You meet it in the middle of its journey and find it barreling down a track toward five people, who are tied to the rails and in imminent peril. Situated nearby is a switch box with a lever on top and pulling the lever will divert the train onto a second track, where only one person is tied. In the absence of any intervention, the train will continue toward the five, flipping the switch will effectively save four lives in net. Would you intervene, and pull the lever, or not? *What course of action would you take?*

When presented with this problem (see also Chapter 24), people often decide to pull the lever, choosing to save the greatest number of lives (Greene et al., 2001; Thomson, 1985). Concrete problems of this sort assess moral reasoning – they ask participants to reason logically about what practical decision they would pursue, based on what they think should be done, or ought to be done, according to their definition of morally acceptable behavior. The moral decision to sacrifice one life

and save five, for example, reflects reasoning through a consequentialist model of ethics. In this model, a utilitarian value is assigned to each outcome (i.e. saving one vs. five lives) and the morally preferable choice is selected on the basis of their relative utility (i.e. choosing the action that saves the greatest number of lives). Consequentialism values the morality of our actions in terms of the outcomes those actions will produce, and thus, electing to save four lives in net is a utilitarian response. Alternate versions of the Trolley problem, however, reveal that our moral reasoning is influenced by the way in which the problem is framed.

In a formally equivalent variant of the Trolley problem, a "fat man" stands on a bridge as a train barrels toward five people below (Greene et al., 2001; Thomson, 1985). The man can be pushed onto the rails from above, presenting a significant obstacle for the train that is capable of halting it in its tracks. Thus, instead of pulling a lever to divert a train onto a second track and therefore causing one person to die, the participant is confronted with the decision to actively push a fat man off a bridge and onto the tracks to block the train from killing five people. Critically, the net utilitarian outcome of this decision is identical to the original problem: One person is sacrificed to save five. Participants will now judge this choice to be an immoral action. In contrast to a utilitarian approach, this decision reflects a deontological model of ethics, which values choices in terms of moral beliefs about the actions themselves. Pushing a person off a bridge is an inherently harmful act – as opposed to pulling a lever – and is therefore judged to be an immoral action.

The two versions of this problem illustrate an interesting feature of moral reasoning: When presented with equivalent choices under different framings, people will employ different normative ethical standards and arrive at different decisions. Moral reasoning, then, is the set of cognitive processes that lead to the retrieval of a relevant model in the face of a particular moral problem and the use of that model to weight and select between actions on the basis of their moral value. These processes depend critically then on the ability to represent and manipulate valid *representations* of the external world – that is, to recollect and imagine the moral value of individual actions, or of those action's causal consequences.

Dual-Process Theories of Moral Reasoning

Dual-process theories of human inference (see Evans, 2008) propose that there is a distinction between automatic and deliberate processing streams during reasoning tasks, and have been argued to be one framework that supports the pattern of responses observed across variants of the Trolley problem. From a dual-process perspective, human inference depends on dissociable cognitive systems for intuition (e.g. based on an emotional response) and deliberation (e.g. based on critical thought and evaluation). Evidence suggests this dual-process system is engaged during tasks of moral reasoning. Cohen (2005) reviews evidence establishing the interactive role of deliberative and emotional systems in moral reasoning, demonstrating that controlled processing during reasoning tasks can be easily overridden by more salient emotional cognition. That is – intuitive, emotional, intuition-based processing will

override deliberative, considered processing during moral reasoning tasks, consistent with many dual-process theories.

Research in the psychological sciences has specifically considered moral reasoning (and the Trolley problem) through the lens of a dual-process theory of human inference (e.g. Cushman, 2013). The dual-process theory proposes that moral reasoning engages two distinct sets of information-processing systems: one heuristic system that operates through mechanisms that are fast and automatic, and another system that is rule-based and reflects slow, deliberative mechanisms. According to this framework, utilitarian and deontological ethics reflect the respective engagement of two distinct deliberative and heuristic systems. More recently, these two processing systems have been suggested to reflect specific computational approaches to decision-making: model-based (deliberative) and model-free (heuristic) systems (see Daw et al., 2011; Dayan and Berridge, 2014; Gläscher et al., 2010).

Model-Based and Model-Free Learners

Model-based and model-free systems acquire representations of the world through learning; however, they differ in the way they represent and process that information. Model-based learning involves encoding schemas and causal relationships from prior experience, which allow the respondents to construct causal models of possible actions and to reason about moral value of their consequences. This is precisely how consequentialist ethics operates: The ultimate consequences of actions are determined and valued, and the choice that maximizes utility is selected. In contrast, model-free systems do not employ a causal model of action and consequence. Instead, model-free systems employ a sparse and heuristic approach that retrieves the values of possible actions and chooses between them on the basis of that information alone. This is the operation that defines deontological ethics, as well: The moral value of actions is considered, and the consequences of those actions are not.

Together, these two systems provide a framework for understanding the cognitive differences between consequentialist (model-based) vs. deontological (model-free) ethics in moral reasoning – one reflects slow, effortful, goal-directed deliberation, the other reflects fast, automatic, heuristic processing. Participants value the actions of flipping a switch and pushing a man to his death differently; the dual-process framework suggests the aversion to pushing people in front of trolleys and affecting their demise to be an automatic, heuristic tendency. Participants are driven to recollect or imagine the morality of an action in that particular situation, engaging a cognitive system that overrides the judgments of causal models about the consequences of actions. Critically, this process suggests that moral reasoning requires interactions between model-based and model-free systems during moral reasoning: In order to evaluate model-free representations of actions that have (presumably) never been directly performed (e.g. pushing a man to his death, committing incest with a sibling), model-based representations would first be required to construct or imagine such a simulation of that scenario (Crockett, 2013) to then serve as input to a model-free system. In particular, better understanding how properties of brain

networks support this interaction between model-based and model-free systems will further our understanding of how neural systems converge on a particular decision-making strategy.

The Neuroscience of Moral Reasoning

Research in neuroscience has investigated the neurobiological foundations of reasoning – specifically, the brain regions and networks responsible for implementing model-based and model-free systems (Dayan, 2012; Wunderlich, Dayan, and Dolan, 2012). Across several species, this research variously identifies model-based reasoning within prefrontal, striatal, parietal, and default-mode structures, and identifies model-free reasoning within striatal and other subcortical structures. In general, neuroscience research suggests that model-based reasoning recruits brain areas involved in reasoning and cognitive control, and that model-free reasoning recruits brain areas involved in emotional processing.

Further neuroscience research has specifically investigated the neural correlates of moral reasoning in humans. In their review of the cognitive neuroscience literature, Pascual, Rodrigues, and Gallardo-Pujol (2013) conclude that moral reasoning depends on several overlapping networks and processes, suggesting that moral reasoning does not depend on the activity of a single neural substrate or system. Many of the neurobiological structures identified in this meta-review (Figure 25.1) overlap with regions associated with model-based or model-free systems as components of the limbic, frontoparietal, and default-mode networks.

Pascual et al. (2013) suggest that moral reasoning is implemented within a broadly distributed network of regions that are functionally integrated across the brain. The number and complexity of brain regions known to be engaged during moral reasoning suggests that neural mechanisms governing the coordination and integration of large-scale brain networks play a role in cognition during moral reasoning tasks. A central question remains: What neural and cognitive mechanisms are engaged in cases in which individuals rely on a mixture of model-based and model-free systems?

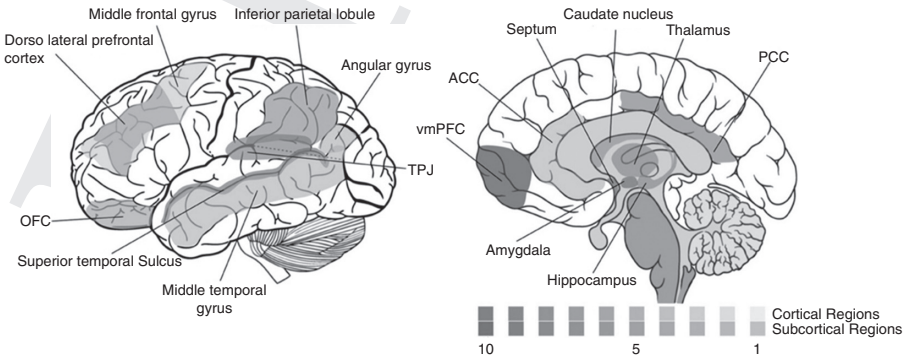

Figure 25.1 *Brain regions associated with tasks of moral reasoning*

The diversity of neurobiological regions involved in moral reasoning, and the putative network interactions required to coordinate them, resist explanations that appeal to single regions or mechanistic circuits (for additional reviews, see Garrigan, Adlam, and Langdon, 2016 and Young and Dungan, 2012). A network neuroscience framework provides an integrative path forward by allowing us to conceptualize moral reasoning in terms of emergent interactions at the level of brain networks.

Functional Brain Networks in Moral Reasoning

Network neuroscience examines information processing in the brain in terms of neurobiological principles that govern the structure and communication of brain networks. Evidence from the field suggests that the brain is organized according to an interactive *hierarchy* of components (Meunier, Lambiotte, and Bullmore, 2010), with network attributes at one scale constraining attributes at another. For example, the organization of structural brain networks (properties of how the brain is physically wired) will support dynamic patterns of connectivity and organization that are present in functional brain networks (properties of how the brain is connected) (Sporns, Tononi, and Edelman, 2000).

Neuroscientific evidence suggests that model-free and model-based systems depend on activity in three distinct functional brain networks, the frontoparietal, limbic, and default-mode. Research indicates that the brain is organized according to several such intrinsic connectivity networks (e.g. Power et al., 2011; Yeo et al., 2011), in which intrinsic connectivity reflects the default, resting state organization of functional brain regions across the cortex. At rest, activity in these networks will fluctuate in an anticorrelated manner. During a task, however, brain networks modify their activity to produce patterns of functional connectivity that dynamically respond to task demands. This emergent connectivity is constrained by the brain's underlying neural architecture, and in this way, the structural (and functional) connectivity of brain networks at rest informs the set of possible task-based configurations networks can achieve. That is, the ease of any required network reconfigurations during task-based activity depends on a network's specific topological properties (see Cole et al. 2014, 2016). The extent to which model-based and model-free systems can integrate their processing together, or operate in isolation, therefore depends in part on the underlying properties of brain networks.

Two prominent intrinsic connectivity networks have been widely implicated in moral reasoning – the frontoparietal control network **(**Spreng et al., 2010; Vincent et al., 2008) and the limbic network **(**Carmichael and Price, 1995; Morgane, Galler, and Mokler, 2005**)**. Network neuroscience also identifies a third network, the default-mode (Chiong et al., 2013; Spreng et al., 2010), that is specifically recruited during situations in which model-based and model-free systems must integrate their activity to imagine and simulate hypothetical events.

The frontoparietal control network is responsible for orchestrating cognitive control and supporting goal-directed behavior (Barbey et al., 2014; Cole et al.,

2012). Collectively, frontoparietal control systems are engaged during general tasks of value-based decision-making – assessing the utilitarian value of possible actions and making a choice between them (Domenech et al., 2018; Gläscher et al., 2012; Polanía et al., 2015). During moral reasoning, multiple components of the frontoparietal network, including portions of the dorsolateral prefrontal (dlPFC) and ventromedial prefrontal (vmPFC) cortices, are engaged to facilitate cognitive control and valuation during decision-making (Harenski et al., 2010; Jeurissen et al., 2014; Kédia et al. 2008; see Figure 25.1). Cognitively, these operations entail the same processes ascribed to consequentialist, model-based reasoning: consciously deliberating about causal models and evaluating actions based on their consequences. Moral reasoning has been specifically associated with activity in the frontoparietal control network during model-based processing of consequentialist decisions (i.e. flipping a lever to save four lives in net; Chiong et al., 2013).

A second intrinsic connectivity network recruited during moral reasoning is the limbic network, a system comprised of orbitofrontal cortex, insular cortex, ventral striatum, and deep-brain nuclei. This network is primarily engaged during the representation and processing of emotion (Power et al., 2011; Rajmohan and Mohandas, 2007; Schneider et al., 2013) and is responsible for processing associations between stimuli and their immediate rewards (Everitt et al., 1991) – all processes engaged during model-free moral reasoning. During moral reasoning, orbitofrontal and subcortical components of the limbic network demonstrate functional integration, facilitating representations of moral value for individual actions and enabling assessments of their respective utility (Greene et al., 2001; Moll et al., 2008; Pascual, Rodrigues, and Gallardo-Pujol, 2013).

One difficulty for dual-process theories arises from considering when and how the two (nominally distinct) processing systems they posit would need to coordinate activity, for example to produce complex deontological reasoning. The model-free system, for example, would need to engage with the model-based system in order to evaluate hypothetical events not encoded by previous experience (see Crockett, 2013). In this situation, model-based representations would first be required to construct or imagine a simulation of that scenario, to then serve as input to model-free systems (also see Cushman, 2013). Neuroscientific evidence provides support for such a process, mediated by a specific anatomical region: the ventromedial prefrontal cortex, whose activity is associated with deontological judgments of personal harms (electing to push the fat man off the bridge; Greene et al., 2001). To represent this hypothetical model of the world, the brain must maintain a description of agents and their interactions for manipulation outside of the normal frontoparietal mechanisms that mediate learned, model-based systems. Network neuroscience identifies this process with activity in a third brain network, the default-mode (Chiong et al., 2013). One cognitive process associated with this form of task-based, default-mode activity is imagination – accessing or envisioning imagery and knowledge through internal simulation of models (e.g. Østby et al., 2012). Thus, in addition to model-free and model-based system employed during moral reasoning, activity in the default-mode network also mediates a third system for *model-*

simulation, possibly distinguishing moral reasoning from other (more domain-general) systems for assessing value.

Functional Connectivity of the Ventromedial Prefrontal Cortex

How do these three networks coordinate their activity during moral reasoning? Neuroscientific evidence suggests that interactions between model-based and model-free systems during moral reasoning may be mediated by a central network hub, the ventromedial prefrontal cortex (Figure 25.2). The orbitofrontal component of the limbic network overlaps anatomically with the vmPFC, a region with structural projections to the dlPFC and other prefrontal regions. The vmPFC is also considered a part of the default-mode network. The vmPFC has been previously identified as an important functional hub for the integration of representations between limbic and frontocortical structures (e.g. Benoit et al., 2014; Roy et al., 2012), involved in propagating information between limbic system and regions of the frontoparietal control network. Several studies indicate the vmPFC/OFC serves a critical role in several decision-making processes that involve determining the values of potential choices (Hare et al., 2009; Rangel and Hare, 2012; Levy and Glimcher 2012; Sescousse et al., 2013). The vmPFC has also been more specifically associated with moral reasoning abilities (Raine and Yang, 2006), and with encoding value during both model-based (Ballenie and O'Doherty, 2010) and model-free (Daw et al, 2011)

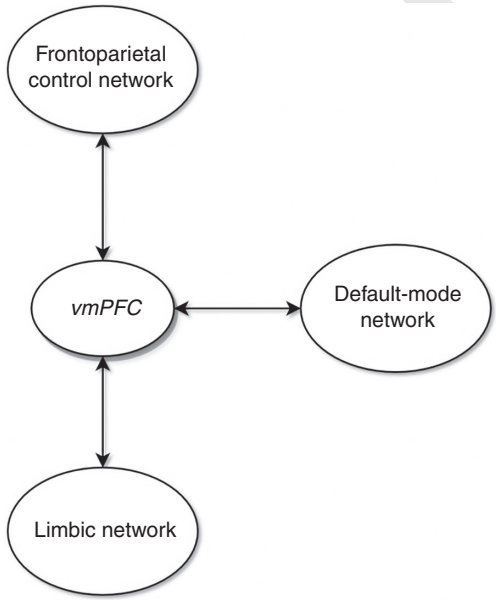

Figure 25.2 *Interactions between vmPFC and functional ICNs figure prominently during moral reasoning*

learning and reasoning. That is, evidence suggests the vmPFC may be an important hub region between the networks involved in moral reasoning, capable of transferring information between a processing intensive, model-based system mediated by prefrontal areas, and a more heuristic, model-free system dependent on limbic areas. Granger causality analyses indicate that the vmPFC also indirectly mediates the recruitment of default-mode processing during moral reasoning, facilitating the imagination of events through an interaction between model-based and model-free systems (Chiong et al., 2013).

The ultimate reliance on either a consequentialist or deontological mode of judgment during moral reasoning may depend on the relative network connectivity of the vmPFC to frontoparietal vs. limbic structures. This hypothesis would align with previous research on the role of vmPFC connectivity in impaired moral reasoning (e.g. Koenigs et al., 2007), and explain contradictory patterns of results previously reported in the field (e.g. Chiong et al., 2013). Through its network-level connectivity, the vmPFC is positioned to act as an influential node in coordinating or routing activity between prefrontal and limbic regions of the brain. One topological function of the vmPFC may be to direct processing during moral reasoning toward brain networks that implement model-based and model-free systems, determining what sort of cognitive processing (effortful, or heuristic) is engaged during moral reasoning tasks (also see Pessoa, 2008). How readily individuals can deploy model-based or model-free reasoning may therefore depend on how readily the underlying topology and connectivity of vmPFC supports communication with these brain networks. Relative differences in the connectivity of vmPFC to other brain networks (thought to index the ability of vmPFC/OFC to send and receive information) may produce different patterns of dynamic connectivity during moral reasoning tasks – for example, by altering the energy cost of routing communications through different network topologies, serving to index the relative susceptibility to model-free vs model-based reasoning.

Activation in the vmPFC is known to conflict with prefrontal activity – interfering with ongoing utilitarian (model-based) reasoning and instead facilitating the emotional processing thought to mediate model-free reasoning (Shenhav and Greene., 2014). Activation of vmPFC appears to be associated with transferring information between brain networks, allowing information from previously *segregated* model-based and model-free systems to propagate through to other areas. This phenomenon has been studied in the context of moral reasoning in patients with focal brain lesions to the vmPFC, who sit at one (clinical) extreme of network disconnectivity. In patients with a focal vmPFC lesion, impaired moral reasoning is often observed, such that the absence of a connecting network hub leads subjects to resist employing model-free reasoning during moral dilemmas (Koenigs et al., 2007). That is, vmPFC patients perform consequentialist, model-based reasoning in the same manner as healthy controls, and also apply model-based reasoning in situations in which healthy controls would employ model-free reasoning (Ciaramelli et al., 2007), utilizing a utilitarian judgment system and always electing to push the fat man off the bridge.

This selective deficit in model-free reasoning observed with vmPFC lesions (and limbic/frontoparietal disconnectivity) can be further contrasted with other network connectivity profiles that alter the moral reasoning process. An

interesting pattern of findings has emerged from comparisons between vmPFC lesion patients and psychopaths – individuals who share the same behavioral deficits in everyday life (blunted emotions, antisocial behavior), but who possess an intact, if dysfunctional, vmPFC (see Koenigs, 2012; Motzkin et al., 2011). Individuals with psychopathic personalities have been observed to display reduced emotional responses during both model-based and model-free moral reasoning. These individuals simultaneously experience a similar amount of disconnectivity between the vmPFC and areas of the frontoparietal, default-mode, and limbic networks (Motzkin et al., 2011). Critically however, an intact vmPFC still allows for communication between these networks for model-based and model-free systems, such that psychopaths present identical patterns of model-based and model-free reasoning to healthy individuals when considering moral dilemmas in a laboratory setting (Cima, Tonnaer, and Hauser, 2010; but see Koenigs et al., 2012). Though dysfunctional in emotional processing and more weakly connected to all other relevant structures, an intact vmPFC is still capable of serving as a hub for communication, such that psychopaths present with patterns of utilitarian and deontological judgment that match those of normal controls – despite their dysregulated application of these same judgments to everyday life. Intact network hubs in healthy controls, and connectivity-reduced, dysfunctional hubs in psychopathic individuals both facilitate equivalent amounts of connectivity between limbic and prefrontal structures; both produce identical patterns of consequentialist and deontological reasoning during moral reasoning tasks. A lesioned vmPFC prevents the integration of model-based and model-free systems; consequently, vmPFC patients are more reliant on model model-free, deontological processing.

Research into neurodegenerative disorders further elucidates the relationship between vmPFC connectivity and utilitarian judgments, suggesting that employing model-free systems is specifically sensitive to relative differences in the connectivity of vmPFC with limbic structures. Behavioral Variant Frontotemporal Dementia (bvFTD) and Alzheimer's disease (AD) selectively display tropism for distinct brain networks, with AD producing damage to the posterior default-mode network, and bvFTD producing damage to vmPFC and limbic structures (Mendez and Shapira, 2009). In both cases, frontoparietal structures are largely spared, and individuals present with normal rates of utilitarian judgment on moral reasoning tasks. Individuals with AD will experience default-mode degeneration, impairing their ability to simulate models during personal moral dilemmas (i.e. when pushing a fat man from a bridge). Strikingly, AD patients still respond to these sorts of moral dilemmas in the same way as healthy controls. bvFTD patients, experiencing degeneration of the vmPFC, striatum, and paralimbic structures, do present with impaired moral reasoning, behaving identically to patients with vmPFC lesions (Chiong et al., 2013). This evidence suggests that neurodegeneration in the default-mode network still preserves healthy patterns of moral reasoning – instead, losing the ability to integrate model-based and model-free systems results from damage to medial prefrontal structures that connect all three networks.

Evidence for sex differences in healthy individuals also suggests that variation in the underlying connectivity of vmPFC to the frontoparietal and limbic networks index the relative reliance on consequentialist and deontological reasoning. Sex differences exist in the prefrontal cortex, such that males exhibit greater connectivity within prefrontal cortex regions (Chuang and Sun, 2014), increasing the density of connections between vmPFC and other portions of the frontoparietal control network. Sex differences in emotional processing have also been observed, such that females possess a larger vmPFC (Welborn et al., 2009), overall greater density of gray matter connectivity (Tomasi and Volkow, 2012), and higher evoked, task-based connectivity between orbitofrontal cortex and limbic structures (Koch et al., 2007). This suggests that males may experience greater connectivity between vmPFC and prefrontal structures, and females may experience greater connectivity between vmPFC and limbic structures. Selective differences in healthy vmPFC connectivity may be associated with a relative reliance on model-based or model-free systems – in contrast with psychopathic individuals, who experience universally decreased vmPFC connectivity and no differences in utilization of model-free reasoning. Gender differences in vmPFC connectivity may therefore explain the robust finding that females are less likely to employ model-based, utilitarian moral reasoning than males, and are instead more likely to employ deontological, model-free reasoning (Friesdorf, Conway, and Gawronski, 2015).

Dynamic Network Connectivity during Moral Reasoning

Research in network neuroscience suggests that neural activity is an adaptive and *dynamic* phenomenon, reflecting the capacity of brain regions to create new information-processing networks by altering their connectivity in a task-dependent manner (Braun et al., 2015; Shine et al., 2016). Frontoparietal regions of the brain, for example, are either specialized or flexible in their task-based engagement. Specialized regions generally serve a fixed set of specific information-processing functions. More flexible regions are instead capable of functionally modifying their network membership through task-based coupling with other regions, altering network connectivity in response to task demands (Yeo et al., 2015).

These same organizational properties can be used to describe many "real-world" networks – subway systems, commercial airline traffic, social groups – in which multiple entities are communicating with some cost. The biological cost of building and maintaining a neural architecture imposes pressure to optimize the brain across all scales, resulting in efficient organization that respects the physiological and metabolic costs inherent in assembling, maintaining, and operating neural tissue (Bullmore and Sporns, 2012). Models of cortical function constructed to respect these biological constraints (across multiple brain imaging modalities) have arrived at modular, hierarchical organizations of both structural and functional dynamics as a highly efficient way to optimize the brain's architecture (e.g. Gray and Robinson 2013). The network properties facilitated by this hierarchical structure provide an optimal balance between communication at a local and global scale, enabling more

efficient transitions between functional brain networks (see Gallos, Makse, and Sigman, 2012).

One mechanism through which the relative network connectivity of vmPFC to frontal, limbic, and default-mode structures could produce individual differences in rates of consequentialist and deontological reasoning is through relative differences in the underlying topology of structural connections that affect the difficulty of network transitions required to coordinate activity between model-based and model-free systems. The dynamic brain-network reconfigurations evoked during moral reasoning (imagining pushing a man from a bridge through modal-based systems and communicating that information to model-free systems) are necessarily constrained by the underlying structural anatomy and topology of relevant brain regions. Approaches that model this underlying structure, such as network control theory (Gu et al., 2015), analyze structural connections between brain regions and characterize how well-connected nodes (hubs) to exert influence on the functional trajectory of brain states (Figure 25.3). As one such hub node, the vmPFC asymmetrically affects network controllability to exert influence over the set of functional brain state transitions possible from moment to moment (Kerr et al., 2012; also see Betzel et al., 2016). In this way, differences in the network connectivity of vmPFC to limbic

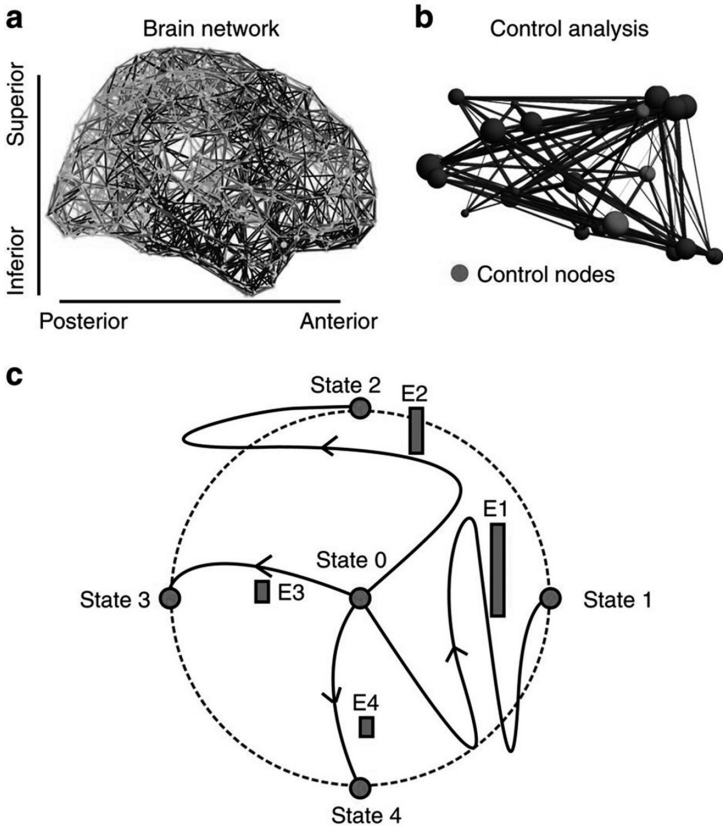

Figure 25.3 *Structural network controllability*

and prefrontal areas may influence the ability of the vmPFC to control the trajectories of task-based networks, affecting the relative ease of deploying consequentialist or deontological reasoning when faced with a moral dilemma.

Network control theory offers insight into the types of network operations required to assemble model-based and model-free representations. Model-based reasoning requires slow, effortful processing to assemble and manipulate mental models. This process is cognitively demanding and sensitive to additional sources of cognitive load. When placed under cognitive load (for example during stress, or while required to perform additional cognitively demanding tasks), subjects will forego model-based reasoning and rely instead on a less demanding, heuristic system for model-free reasoning (Economides et al., 2015). The effortful construction of causal models for moral reasoning may therefore require a more demanding series of network operations and transitions. That is, model-based systems may depend of a series of dynamic shifts in connectivity that transition the brain away from a less demanding, more baseline processing system (model-free) to a more demanding, difficult-to-reach network state (model-based). Critically, determining whether to pursue model-based processing or to engage a model-free system would depend on the relative ability to inhibit activity in the vmPFC, facilitating persistence in model-based reasoning by maintaining an effortful, difficult-to-reach network state.

Network transitions that drive the brain into difficult-to-reach states are in general associated with activity within the frontoparietal control network (Dosenbach et al., 2008; Gu et al., 2015), and the ability of frontoparietal regions to control network state transitions is specifically indexed by regional measures of global connectivity (Cole et al., 2012; van den Heuvel et al., 2009). The increased cognitive demands associated with model-based reasoning suggest that the energy costs, topological requirements, and processing time required to engage in model-based reasoning are higher than model-free reasoning, and that increased frontoparietal connectivity should better facilitate the maintenance of these difficult-to-reach brain states. Conversely, enhanced connectivity between vmPFC and limbic structures should lower the difficulty of network-state transitions that facilitate model-free reasoning, which should depend on easier-to-reach network states. Dynamic integration between these two brain networks (the operations required to imagine pushing a man in front of a train, and then heuristically judge that action to be a moral wrong) should depend critically on the joint connectivity of vmPFC to limbic and prefrontal regions.

Future Directions and Conclusions

The current neuroscience literature supports eight conclusions about the nature of moral reasoning. (1) Dual-process reasoning is implemented within multiple functional brain networks. (2) Model-based reasoning is implemented by the frontoparietal control network. (3) Model-free reasoning is implemented by the limbic network. (4) Model-simulating interactions are implemented by the default-mode network. (5) The vmPFC is a hub node involved in mediating responses across

these three networks. (6) The relative connectivity strength of the vmPFC to frontoparietal and limbic networks indexes susceptibility to model-based or model-free reasoning. (7) Neurobiological changes that alter the connectivity of vmPFC to limbic network will alter rates of utilitarian moral reasoning. (8) Neurobiological changes that do not affect the relative connectivity of vmPFC to limbic regions will not alter rates of utilitarian reasoning.

The neurobiological account provided above offers an explanation of how model-based and model-free systems interact. We suggest that systems which support moral reasoning are mediated by activity within specific brain networks, and that the rate of consequentialist and deontological judgments during moral reasoning is influenced by the underlying topological relationships between those networks.

This model motivates two novel predictions for future research. Neuroscientific evidence indicates that structural and resting-state functional connectivity constrain the network transitions and states the brain can manifest during any task. Neuroscience research has yet to directly investigate how the topology of prefrontal and limbic networks might influence rates of utilitarian response to moral dilemmas. Reliance on a particular form of moral reasoning should reflect the availability of existing cognitive models and the ability of brain networks to support their manipulation. For example, individuals who engage in more utilitarian, model-based valuations during real-world decision-making may be more resistant to model-free reasoning when reasoning about moral dilemmas and should display greater connectivity within the prefrontal cortex. A second prediction addresses the inferential strength of moral reasoning – for example, how confidently a moral judgment can be drawn. A network neuroscience framework predicts that the speed with which moral judgments can be drawn will depend on the ease of network transitions and functional integration between requisite prefrontal, default-mode, and limbic structures. Processes that underlie moral reasoning depend in part on modifying networks in response to the environment to derive a task-based network – so, the strength and speed of moral inferences should depend on the degree to which topological properties of the vmPFC and related networks lead to easy or difficult transitions into relevant brain states.

Resolving moral dilemmas involves imagining novel or hypothetical scenarios, a cognitive process that depends on constructing, representing, and manipulating causal models as part of the reasoning process. Greene et al. (2001) identified unique contributions to this process made by rational and emotional systems and suggested that their interactions play an important role in producing moral judgments. Today, network neuroscience is interested in studying a similar phenomenon – how networks in the brain interact to produce emergent behavior. To the extent that models for moral reasoning posit multiple areas, networks, or systems for generating judgments and inferences, network neuroscience has an important role to play in studying how the dynamic and topological properties of distributed brain networks enable them to interact and collectively facilitate cognition. By representing neural activity as a distributed and dynamic system, network neuroscience can advance our understanding of the mechanisms that govern moral judgments, helping elucidating the relationship between moral reasoning and the neurobiological mechanisms that facilitate it.

References

Barbey, A. K., Colom, R., Paul, E. J., and Grafman, J. (2014). Architecture of Fluid Intelligence and Working Memory Revealed by Lesion Mapping. *Brain Structure & Function*, *219*(2), 485–494.

Bassett, D. S., and Sporns, O. (2017). Network Neuroscience. *Nature Neuroscience*, *20*(3), 353–364.

Benoit, R. G., Szpunar, K. K., and Schacter, D. L. (2014). Ventromedial Prefrontal Cortex Supports Affective Future Simulation by Integrating Distributed Knowledge. *Proceedings of the National Academy of Sciences*, *111*(46), 16550–16555.

Betzel, R. F., Gu, S., Medaglia, J. D., Pasqualetti, F., and Bassett, D. S. (2016). Optimally Controlling the Human Connectome: The Role of Network Topology. *Scientific Reports*, *6*, 30770.

Braun, U., Schäfer, A., Walter, H., et al. (2015). Dynamic Reconfiguration of Frontal Brain Networks during Executive Cognition in Humans. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(37), 11678–11683.

Bressler, S. L., and Menon, V. (2010). Large-Scale Brain Networks in Cognition: Emerging Methods and Principles. *Trends in Cognitive Sciences*, *14*(6), 277–290.

Bullmore, E., and Sporns, O. (2009). Complex Brain Networks: Graph Theoretical Analysis of Structural and Functional Systems. *Nature Reviews Neuroscience*, 10(3), 186–198.

Carmichael, S. T., and Price, J. L. (1995). Limbic Connections of the Orbital and Medial Prefrontal Cortex in Macaque Monkeys. *The Journal of Comparative Neurology*, *363*(4), 615–641.

Chiong, W., Wilson, S. M., D'Esposito, M., et al. (2013). The Salience Network Causally Influences Default Mode Network Activity during Moral Reasoning. *Brain*, *136*(6), 1929–1941.

Chuang, C.-C., and Sun, C.-W. (2014). Gender-Related Effects of Prefrontal Cortex Connectivity: A Resting-State Functional Optical Tomography Study. *Biomedical Optics Express*, *5*(8), 2503–2516.

Ciaramelli, E., Muccioli, M., Làdavas, E., and di Pellegrino, G. (2007). Selective Deficit in Personal Moral Judgment Following Damage to Ventromedial Prefrontal Cortex. *Social Cognitive and Affective Neuroscience*, *2*(2), 84–92.

Cima, M., Tonnaer, F., and Hauser, M. D. (2010). Psychopaths Know Right from Wrong but Don't Care. *Social Cognitive and Affective Neuroscience*, *5*(1), 59–67.

Cohen, Jonathan, D. (2005). The Vulcanization of the Human Brain: A Neural Perspective on Interactions between Cognition and Emotion. *Journal of Economic Perspectives*, *19* (*4*): 3–24.

Cole, M. W., Bassett, D. S., Power, J. D., Braver, T. S., and Petersen, S. E. (2014). Intrinsic and Task-Evoked Network Architectures of the Human Brain. *Neuron*, *83*(1), 238–251.

Cole, M. W., Ito, T., Bassett, D. S., and Schultz, D. H. (2016). Activity Flow over Resting-State Networks Shapes Cognitive Task Activations. *Nature Neuroscience*, *19*(12), 1718–1726.

Cole, M. W., Yarkoni, T., Repovs, G., Anticevic, A., and Braver, T. S. (2012). Global Connectivity of Prefrontal Cortex Predicts Cognitive Control and Intelligence. *Journal of Neuroscience*, *32*(26), 8988–8999.

Crockett, M. J. (2013). Models of Morality. *Trends in Cognitive Sciences*, *17*(8), 363–366.

Cushman, F. (2013). Action, Outcome, and Value: A Dual-System Framework for Morality. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *17*(3), 273–292.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, *69*(6), 1204–1215.

Dayan, P. (2012). How to Set the Switches on this Thing. *Current Opinion in Neurobiology*, *22* (6), 1068–1074.

Dayan, P., and Berridge, K. C. (2014). Model-Based and Model-Free Pavlovian Reward Learning: Revaluation, Revision and Revelation. *Cognitive, Affective and Behavioral Neuroscience*, *14*(2), 473–492.

Domenech, P., Redouté, J., Koechlin, E., and Dreher, J.-C. (2018). The Neuro-Computational Architecture of Value-Based Selection in the Human Brain. *Cerebral Cortex*, *28*(2), 585–601.

Dosenbach, N. U. F., Fair, D. A., Cohen, A. L., Schlaggar, B. L., and Petersen, S. E. (2008). A Dual-Networks Architecture of Top-Down Control. *Trends in Cognitive Sciences*, *12*(3), 99–105.

Economides, M., Kurth-Nelson, Z., Lübbert, A., Guitart-Masip, M., and Dolan, R. J. (2015) Model-Based Reasoning in Humans Becomes Automatic with Training. *PLOS Computational Biology*, *11*(9), e1004463.

Evans, J.S., 2008. Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *The Annual Review of Psychology*, 59, 255–278.

Everitt, B. J., Morris, K. A., O'Brien, A., and Robbins, T. W. (1991). The Basolateral Amygdala-Ventral Striatal System and Conditioned Place Preference: Further Evidence of Limbic-Striatal Interactions Underlying Reward-Related Processes. *Neuroscience*, *42*(1), 1–18.

Friesdorf, R., Conway, P., and Gawronski, B. (2015). Gender Differences in Responses to Moral Dilemmas: A Process Dissociation Analysis. *Personality & Social Psychology Bulletin*, *41*(5), 696–713.

Gallos, L. K., Makse, H. A., and Sigman, M. (2012). A Small World of Weak Ties Provides Optimal Global Integration of Self-Similar Modules in Functional Brain Networks. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(8), 2825–2830.

Garrigan, B., Adlam, A. L. R., and Langdon, P. E. (2016). The Neural Correlates of Moral Decision-Making: A Systematic Review and Meta-Analysis of Moral Evaluations and Response Decision Judgements. *Brain and Cognition*, *108*, 88–97.

Gläscher, J., Adolphs, R., Damasio, H., et al. (2012). Lesion Mapping of Cognitive Control and Value-Based Decision Making in the Prefrontal Cortex. *Proceedings of the National Academy of Sciences*, *109*(36), 14681–14686.

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J. P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron*, *66*(4), 585–595.

Gray, R. T., and Robinson, P. A. (2013). Stability Constraints on Large-Scale Structural Brain Networks. *Frontiers in Computational Neuroscience*, *7*.

Greene, J. D. (2017). The Rat-a-Gorical Imperative: Moral Intuition and the Limits of Affective Learning. *Cognition*, *167*, 66–77.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, *293* (5537), 2105–2108.

Gu, S., Pasqualetti, F., Cieslak, M., et al. (2015). Controllability of Structural Brain Networks. *Nature Communications*, 6(1), 1–10.

Hare, T. A., Camerer, C. F., and Rangel, A. (2009). Self-Control in Decision-Making Involves Modulation of the vmPFC Valuation System. *Science (New York, N.Y.)*, *324*(5927), 646–648.

Harenski, C. L., Antonenko, O., Shane, M. S., and Kiehl, K. A. (2010). A Functional Imaging Investigation of Moral Deliberation and Moral Intuition. *NeuroImage*, *49*(3), 2707–2716.

Jeurissen, D., Sack, A. T., Roebroeck, A., Russ, B. E., and Pascual-Leone, A. (2014). TMS Affects Moral Judgment, Showing the Role of DLPFC and TPJ in Cognitive and Emotional Processing. *Frontiers in Neuroscience*, *8*.

Kédia, G., Berthoz, S., Wessa, M., Hilton, D., and Martinot, J.-L. (2008). An Agent Harms a Victim: A Functional Magnetic Resonance Imaging Study on Specific Moral Emotions. *Journal of Cognitive Neuroscience*, 20(10), 1788–1798.

Kerr, D. L., McLaren, D. G., Mathy, R. M., and Nitschke, J. B. (2012). Controllability Modulates the Anticipatory Response in the Human Ventromedial Prefrontal Cortex. *Frontiers in Psychology*, 3, 557.

Koch, K., Pauly, K., Kellermann, T., et al. (2007). Gender Differences in the Cognitive Control of Emotion: An fMRI Study. *Neuropsychologia*, *45*(12), 2744–2754.

Koenigs, M. (2012). The Role of Prefrontal Cortex in Psychopathy. *Reviews in the Neurosciences*, *23*(3), 253.

Koenigs, M., Kruepke, M., Zeier, J., and Newman, J. P. (2012). Utilitarian Moral Judgment in Psychopathy. *Social Cognitive and Affective Neuroscience*, *7*(6), 708–714.

Koenigs, M., Young, L., Adolphs, R., et al. (2007). Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements. *Nature*, *446*(7138), 908–911.

Kogler, L., Müller, V. I., Seidel, E.-M., et al. (2016). Sex Differences in the Functional Connectivity of the Amygdalae in Association with Cortisol. *NeuroImage*, *134*, 410–423.

Levy, D. J., and Glimcher, P. W. (2012). The Root of All Value: A Neural Common Currency for Choice. *Current Opinion in Neurobiology*, 22(6), 1027–1038.

Medaglia, J. D., Satterthwaite, T. D., Kelkar, A., et al. (2018). Brain State Expression and Transitions Are Related to Complex Executive Cognition in Normative Neurodevelopment. *NeuroImage*, 166, 293–306.

Mendez, M. F., and Shapira, J. S. (2009). Altered Emotional Morality in Frontotemporal Dementia. *Cognitive Neuropsychiatry*, *14*(3), 165–179.

Meunier, D., Lambiotte, R., and Bullmore, E. T. (2010). Modular and Hierarchically Modular Organization of Brain Networks. *Frontiers in Neuroscience*, 4, 200.

Moll, J., de Oliveira-Souza,R., Zahn, R., and Grafman, J. (2008). The Cognitive Neuroscience of Moral Emotions. In W. Sinnott-Armstrong (ed.), *Moral Psychology, Vol 3. The Neuroscience of Morality: Emotion, Brain Disorders, and Development*. Cambridge, MA: MIT Press, 1–17.

Morgane, P. J., Galler, J. R., and Mokler, D. J. (2005). A Review of Systems and Networks of the Limbic Forebrain/Limbic Midbrain. *Progress in Neurobiology*, 75(2), 143–160.

Motzkin, J. C., Newman, J. P., Kiehl, K. A., and Koenigs, M. (2011). Reduced Prefrontal Connectivity in Psychopathy. *Journal of Neuroscience*, *31*(48), 17348–17357.

Østby, Y., Walhovd, K., Tamnes, C., et al. (2012). Mental Time Travel and Default-Mode Network Functional Connectivity in the Developing Brain. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 16800–16804.

Pascual, L., Rodrigues, P., and Gallardo-Pujol, D. (2013). How Does Morality Work in the Brain? A Functional and Structural Perspective of Moral Behavior. *Frontiers in Integrative Neuroscience*, 7, 65.

Pessoa, L. (2008). On the Relationship between Emotion and Cognition. *Nature Reviews Neuroscience*, 9(2), 148–158.

Polanía, R., Moisa, M., Opitz, A., Grueschow, M., and Ruff, C. C. (2015). The Precision of Value-Based Choices Depends Causally on Fronto-Parietal Phase Coupling. *Nature Communications*, 6, 8090.

Power, J. D., Cohen, A. L., Nelson, S. M., et al. (2011). Functional Network Organization of the Human Brain. *Neuron*, 72(4), 665–678

Raine, A., and Yang, Y. (2006). Neural Foundations to Moral Reasoning and Antisocial Behavior. *Social Cognitive and Affective Neuroscience*, 1(3), 203–213.

Rajmohan, V., and Mohandas, E. (2007). The Limbic System. *Indian Journal of Psychiatry*, 49(2), 132–139.

Rangel, A., and Hare, T. (2010). Neural Computations Associated with Goal-Directed Choice. *Current Opinion in Neurobiology*, 20(2), 262–270.

Roy, M., Shohamy, D., and Wager, T. (2012). Ventromedial Prefrontal-Subcortical Systems and the Generation of Affective Meaning. *Trends in Cognitive Sciences*, 16(3): 147–56.

Schneider, K., Pauly, K. D., Gossen, A., et al. (2013). Neural Correlates of Moral Reasoning in Autism Spectrum Disorder. *Social Cognitive and Affective Neuroscience*, *8*(6), 702–710.

Sescousse, G., Caldú, X., Segura, B., and Dreher, J.-C. (2013). Processing of Primary and Secondary Rewards: A Quantitative Meta-Analysis and Review of Human Functional Neuroimaging Studies. *Neuroscience and Biobehavioral Reviews*, *37* (4).

Shenhav, A., and Greene, J. D. (2014). Integrative Moral Judgment: Dissociating the Roles of the Amygdala and Ventromedial Prefrontal Cortex. *Journal of Neuroscience*, *34* (13), 4741–4749.

Shine, J. M., Bissett, P. G., Bell, P.T., et al. (2016). The Dynamics of Functional Brain Networks: Integrated Network States during Cognitive Task Performance. *Neuron*, *92*(2), 544–554.

Sporns, O., Tononi, G., and Edelman, G. M. (2000). Theoretical Neuroanatomy: Relating Anatomical and Functional Connectivity in Graphs and Cortical Connection Matrices. *Cerebral Cortex*, *10*(2), 127–141.

Spreng, R. N., Stevens, W. D., Chamberlain, J. P., Gilmore, A. W., and Schacter, D. L. (2010). Default Network Activity, Coupled with the Frontoparietal Control Network, Supports Goal-Directed Ccognition. *NeuroImage*, *53*(1), 303–317.

Thomson, J. J. (1985). The Trolley Problem. *The Yale Law Journal*, *94*(6), 1395–1415.

Tomasi, D., and Volkow, N. D. (2012). Gender Differences in Brain Functional Connectivity Density. *Human Brain Mapping*, *33*(4), 849–860.

van den Heuvel, M. P., Stam, C. J., Kahn, R. S., and Hulshoff Pol, H. E. (2009). Efficiency of Functional Brain Networks and Intellectual Performance. *Journal of Neuroscience*, *29*(23), 7619–7624.

Vincent, J. L., Kahn, I., Snyder, A. Z., Raichle, M. E., and Buckner, R. L. (2008). Evidence for a Frontoparietal Control System Revealed by Intrinsic Functional Connectivity. *Journal of Neurophysiology*, *100*(6), 3328–3342.

Welborn, B. L., Papademetris, X., Reis, D. L., et al. (2009). Variation in Orbitofrontal Cortex Volume: Relation to Sex, Emotion Regulation and Affect. *Social Cognitive and Affective Neuroscience*, *4*(4), 328–339.

Wunderlich, K., Dayan, P., and Dolan, R. J. (2012). Mapping Value Based Planning and Extensively Trained Choice in the Human Brain. *Nature Neuroscience*, *15*(5), 786–791.

Yeo, B. T. T., Krienen, F. M., Eickhoff, S. B., et al. (2015). Functional Specialization and Flexibility in Human Association Cortex. *Cerebral Cortex (New York, NY)*, *25*(10), 3654–3672.

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., et al. (2011). The Organization of the Human Cerebral Cortex Estimated by Intrinsic Functional Connectivity. *Journal of Neurophysiology*, *106*(3), 1125–1165.

Young, L., and Dungan, J. (2012). Where in the Brain is Morality? Everywhere and Maybe Nowhere. *Social Neuroscience*, *7*(1), 1–10.

Young, L., and Saxe, R. (2008). The Neural Basis of Belief Encoding and Integration in Moral Judgment. *NeuroImage*, *40*(4), 1912–1920.