IBRO/IBE-UNESCO Science of Learning Briefings

Ontogeny and brain representations of moral decisions

Jean-Claude Dreher

Title	Ontogeny and brain representations of moral decisions
Series	IBRO/IBE-UNESCO Science of Learning Briefings
IBE Director	Dr. Mmantsetsa Marope
Author of brief	Dr Jean-Claude Dreher
Author affiliation	CNRS, Institute of Cognitive Science, Lyon, France
Date	November 16th 2021

Dr Jean-Claude Dreher is CNRS research director of the lab 'Neuroeconomics, Reward and Decision Making' (https://dreherteam.wixsite.com/neuroeconomics) at the Institut des Sciences Cognitives Marc Jeannerod (UMR 5229, Lyon, France). He investigates the neural mechanisms underlying decision making, motivation and reward processing in humans, using concepts from cognitive neuroscience, psychology and behavioral economics. He uses tools such as model-based functional Magnetic Resonance Imaging to understand the neurocomputational processes involved when making individual and social decisions.

Executive Summary

- ➔ Moral behavior is essential for efficient living in groups.
- ➔ Social preferences include altruism, reciprocity, intrinsic pleasure in helping others, aversion to inequity, empathic concern and moral decisions weighing monetary self-interest against moral benefits to others.
- ➔ The field of decision neuroscience investigates how social preferences are represented in the brain of children and adults.
- ➔ Developmental psychology helps to understand the origins of early altruistic behavior.
- ➔ Early capacities for moral evaluations observed in toddlers and children, including helping behavior, fairness, aversion to inequity, third-party-punishment and reputation concerns has been determined.
- ➔ A recent approach in adults combines fMRI with computational modeling. This allows us to identify the neural computations engaged in moral decisions.
- ➔ Moral dilemmas can be modeled as decisions that weigh self-interest against moral costs/harm to others.
- ➔ Decision value computation presiding moral choices engages the brain valuation system as well as additional brain regions.
- ➔ Reputation concerns (audience effects), influence moral and amoral behavior in children and adults and their neural bases has been determined.
- ➔ Understanding how moral choices are made and how moral values are acquired by the brain have important implications for education.

Introduction

Morality can be viewed as the set of customs and values that are embraced by a cultural group to guide social conduct (Decety & Wheatley, 2015; Moll et al., 2005). Moral cognition concerns rules and behaviors that define what is good or bad and what is encouraged or punished within a society. Cognitive neuroscientists, behavioral economists and developmental psychologists investigate social preferences which are related to the individuals' concern toward others' well-being, and the tendency of individuals to adhere to moral principles such as honesty or not harming others. These preferences include motives such as altruism, reciprocity, the intrinsic pleasure in helping others (pure altruism), aversion to inequity, empathic concern and ethical commitments that induce people to help others beyond simply maximizing personal wealth or material payoffs (Decety et al., 2021). These motives lead to different types of observable behaviors such as helping, cooperating, sharing, comforting and rescuing others.

Developing an understanding of right and wrong is an important aspect of early childhood, but there are controversies between researchers regarding the emergence (Smetana et al., 2018) and meaning of this ability. Some researchers support the view that infants and toddlers have an innate "moral sense" (Hamlin et al., 2013; Van de Vondervoort & Hamlin, 2017). For instance, infants as young as 3 months of age prefer ''helpers" to ''hinderers" as assessed using looking time measures in visual scenarios. In contrast, researchers supporting the social domain theory perspective (Smetana et al., 2014; Turiel, 1983) have proposed that moral judgments are constructed through social interactions during early childhood. Other researchers propose that moral evaluation and judgments are the product of an integration of general processes such as attention allocation and approach and avoidance (Cowell & Decety, 2015). According to the last two views, infants' desire to participate in adult activities is an important developmental precursor to morality, but this desire does not constitute a moral concern. These views consider that the understanding of morality in toddlers and children might not be evident before 3 or 4 years of age, although children as young as 18 months have been reported to help others to achieve their goals in different situations (Warneken & Tomasello, 2006). Regardless of these controversies, there are practical implications for early development of social preferences. Indeed, infants' performance in moral evaluation (mean age = 12 months) has been reported to predict social and behavioral adjustment later during preschool (at 4 years of age), (Tan et al., 2018). Thus, there is a developmental continuity in the sociomoral domain and infants' early behavioral tendencies may be building blocks for subsequent socio-moral development.

Below, we present recent findings from developmental psychology and moral neuroscience on the ontogeny and neural correlates of morality. We examine early capacities for moral evaluations that are observed in toddlers and children, including helping behavior, fairness and aversion to inequity, third-party-punishment and reputation concerns. These distinct motivations rely on domain-specific

mechanisms for social preference, as well as domain general social cognitive systems responsible for attention, theory of mind and executive function. Developmental neuroscience research is critical to understand the foundations of moral decision making and helps to identify the mechanisms that guide prosocial behavior. In adults, we present a recent neurocomputational framework, known as model-based fMRI, which combines computational models and fMRI to offer a mechanistic account of moral behavior (Frost & McNaughton, 2017; Lopez-Persem et al., 2017; Rangel et al., 2008; Sescousse et al., 2013).

Neurodevelopmental changes in helping behavior and antisocial behavior

The degree of generosity clearly increases with age in children, and this effect is also sensitive to different culture (Cowell et al., 2017) (**Figure 1**). Four phases have been distinguished in the development of human altruism: (a) interest in social interactions, (b) preference for others' goal completion, (c) concern with others' well-being, and (d) a normative stance toward altruistic actions (Dahl et al., 2017). One can distinguish helping behavior as first-party (i.e. helping another person) or as a third-party, making moral judgment as an observer, but not involved in the social interaction between other individuals. When acting as first-party, infants' helping behavior may be based on a desire to participate in social interactions which might not necessarily be accompanied by the moral judgment that helping is good (Dahl & Paulus, 2019; Kahn, 1992; Killen & Turiel, 1998; Miller et al., 1990; Turiel, 2015). When viewing third-party social interactions with puppets, 9-months-old infants have the ability to distinguish between helping or comforting and antisocial actions such as hitting, they express a preference for individuals who act prosocially compared to those who act antisocially (Decety, 2020; Ting et al., 2019). Orientations toward helping undergo transformations between infancy and late childhood (Dahl et al., 2017). Around 3–4 years of age, children make categorical moral judgments based on concerns with welfare and rights (Dahl & Kim, 2014; Josephs & Rakoczy, 2016; Killen & Smetana, 2015; Nucci & Weber, 1995; Schmidt et al., 2012; Smetana et al., 1999).

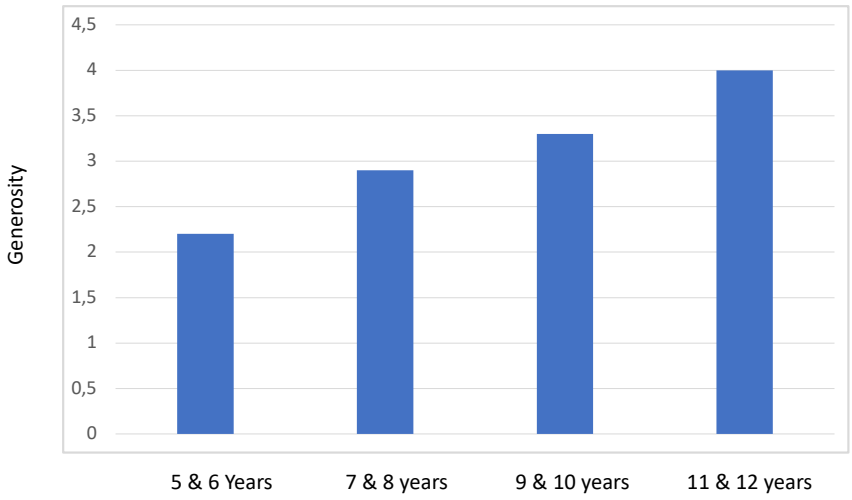

Figure 1. Age-related changes in generosity in a sample of n=999 children over 5 countries (age range= 5-12 years old). Adapted from (Cowell et al., 2017) Cowell et al., *Dev. Science*. 2017

Most of the work published so far in the field of developmental neuroscience on the neural bases of morality in children has used electroencephalography (EEG) because it is simpler to implement than fMRI. This approach allows researchers to investigate evoked-related potentials (ERPs) occurring at the scalp for specific cognitive events. For example, a recent EEG study investigated the neural dynamics of third-party moral scenarios in infants and toddlers (between ages 12 to 24 months) (Cowell & Decety, 2015). Children preferred looking at prosocial cartoon characters rather than antisocial characters, and a late ERP difference was observed at the central scalp location in the time window 300–500 ms, when observing characters helping others compared to those hindering others. Moreover, children with greater ERP negativity for the perception of prosocial characters compared with antisocial characters in this time window also tended to exhibit behavioral preference for the prosocial character, choosing to reach for it. Interestingly, ERPs were also influenced by parental values regarding justice and fairness (Cowell & Decety, 2015). Overall, this developmental neuroscience study shows that precursors to prosocial behavior and moral evaluation appear very early in development but do not necessarily reflects that infants and toddlers possess a "moral sense". Instead, it suggests that early moral cognition is

embedded in general cognitive processes such as attention and approach–withdrawal behavior and is also susceptible to parental influence.

Neural bases of third-party punishment and inequity aversion

Empirical evidence from developmental psychology and cognitive neuroscience supports that third-party punishment (TPP) is an important aspect for constructing morality. This work shows that sensitivity to interpersonal harm emerges early during development, as reflected by both the capacity for implicit social evaluation and an aversion to antisocial agents that harm an ingroup victim (Decety et al., 2021). For example, infants (1 year old) and toddlers (2.5 years old) expect individuals to refrain from helping an ingroup victim's aggressor. In such studies, an indirect form of TPP is used: that is, children withheld help after they saw a wrongdoer steal a toy from a victim while a bystander watched. Immediately after the wrongdoer needed assistance with a task, and the bystander either helped or hindered them. The group memberships of the wrongdoer and the victim were varied relative to that of the bystander. When the victim belonged to the same group as the bystander, children expected TPP: they detected a violation when the bystander chose to help the wrongdoer. Children thus expect third-party punishment selectively to perpetrators of harm to ingroup members. This aversion to antisocial actions may constitute a rudimentary element of morality (Decety & Cowell, 2018). Importantly, such predispositions are in place before children integrate the social conventions and norms of their culture. Later, an understanding that harmful actions cause suffering emerges, followed by the integration of rules that can depend on social contexts and cultures.

The neural bases of TPP remain to be determined in children. A neurodevelomental study investigated intentional harm in a large cohort of participants aged between 4 and 37 years by presenting scenarios that depicted intentional *versus* accidental actions that caused harm/damage. Intentional harm was evaluated as equally wrong across all participants, however, ratings of deserved punishments and malevolent intent were more differentiated with age. An age-related increase in activity was observed in the ventromedial prefrontal cortex in response to intentional harm to people (Decety et al., 2012). In adults, a number of neuroimaging studies have investigated the neural basis of TPP using behavioral economics experiments. In this field, TPP is often viewed as a potential explanation for social cooperation: some individuals not involved in social interactions between two other individuals A and B, are ready to enforce social norms by applying punishment (Boyd & Richerson, 1988; Fehr & Fischbacher, 2003; Riedl et al., 2012). This phenomenon has been widely explored using a modified version of a game known as the Ultimatum Game. In this game, participants, as observers, may pay to punish unfair allocations of endowments by one of two others. Such TPP is interesting as it is costly to the third party who herself receives no material benefit (Fehr & Fischbacher, 2004; Riedl et al., 2012). Inequity aversion has been proposed to be a key motive driving TPP (Blake et al., 2015; Fehr & Fischbacher, 2004; Raihani & McAuliffe, 2012), with payoff difference between proposer and receiver defining the extent of inequity (Zhong et al., 2016). Early neuroimaging studies reported that participants in the role of arbiter derive satisfaction from punishing norm violations, an effect engaging the ventral striatum, which is known to be engaged in reward processing (De Quervain et al., 2004). More recent neuroimaging studies have identified brain regions representing TPP which are similar to those representing inequity aversion (Sanfey et al., 2003), including the anterior cingulate cortex (ACC) and the anterior Insula (Zhong et al., 2016) (**Figure 2**). These brain regions are positively associated with detection of distributional inequity, while the anterior dorsolateral prefrontal cortex (dlPFC), was associated with assessment of intentionality to the norm violator. A two-system network has also been proposed on the basis of meta-analyses, including the anterior insula and ventromedial prefrontal cortex (vmPFC) for rapid evaluation of norm violations, and the dorsal ACC engaging cognitive control mechanisms to resolve the conflict between the norm violations and self-interest (Bellucci et al., 2020; Feng et al., 2015). Several other studies have implicated the mentalizing network in TPP, including medial prefrontal cortex (mPFC) and the TPJ, which reflects empathy for the victim of norm violation and evaluation of legal responsibility (Batson et al., 2007; Baumgartner et al., 2012; Bellucci et al., 2017). The dlPFC may convert the blame signal into a specific punishment signal (Krueger & Hoffman, 2016). Together, these developmental psychology and neuroimaging studies indicate that precursors of morality may act in concert with the emergence of social equality abilities, such as the fair and equal treatment of others and the willingness to punish norm violators, even when one is not directly involved as a victim of the norm violation.

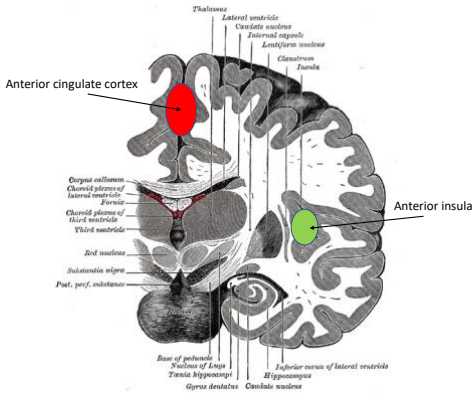

Figure 2. Brain regions engaged in third-party punishment partially overlap with those engaged by inequity aversion such as the anterior cingulate cortex (red) and the bilateral insula (green).

Mechanistic understanding of moral choices using model-based fMRI

Model-based fMRI consists of combining fMRI with computational models of the behavior observed inside the scanner while participants are performing a cognitive task. This approach has flourished in the past decade in decision neuroscience because it allows researchers to address 'how' a particular cognitive operation is implemented in a given brain area in terms of the underlying computational processes (Dunne & O'Doherty, 2013). To do this, researchers identify which brain regions covary with the output variables of the best computational model accounting for behavior among those tested. This is a step forward from performing simple comparisons between two conditions A and B (i.e. simple mapping of brain regions more engaged by A>B). Although the neural circuitry involved in moral cognition has been studied in adults using the simple brain mapping approach (i.e. comparison between 2 conditions) with hypothetical moral dilemmas (Greene et al., 2001, 2004; Moll et al., 2002), more recent model-based fMRI studies allow us to describe the neurocomputational mechanisms engaged in moral choices. This provides insights to understand how underspecified cognitive processes can be mapped to neural computations.

Most neuroimaging research on moral choices using the model-based fMRI approach has concentrated on cost/benefit tradeoffs between moral values and monetary payoff. These studies borrow a key concept from decision neuroscience and behavioral economics, which is the concept of utility, also called decision value. According to this concept, a moral choice is made after a valuation stage considering the subjective value of each option under consideration. A choice is made after the values of options are compared: the option having the highest value is then selected. This principle of value computation has proven successful in reliably identifying a brain valuation system that includes the ventromedial prefrontal cortex (vmPFC) and ventral striatum. This system is known to be engaged in value-based decision making which depends only upon individual preferences (eg. deciding between an apple and an orange). It has also been reported to be engaged in variouss social choices situations (Konovalov et al., 2018; Park et al., 2017; Suzuki & O'Doherty, 2020) and with processing social rewards such as good reputation or being cooperative (Izuma et al., 2008; Rilling et al., 2002; Zaki & Mitchell, 2011) (**Figure 3**).

One key question is to know whether moral value computations engage only this classical brain valuation system or whether moral considerations also engage additional brain systems. There are still controversies regarding how moral considerations can or cannot be incorporated in the valuation system. According to a first hypothesis, computing moral values relies on the same neurocomputational mechanisms as those involved in non-moral value computation. Thus, the brain valuation network would also be engaged for moral decisions during choices coupling financial rewards with moral consequences (**Figure 3**). Supporting this view, several fMRI studies report that the brain has developed the capacity to incorporate moral considerations into its standard valuation circuitry (Crockett et al., 2017a; Hare et al., 2010a; Hutcherson et al., 2015a; Qu et al., 2019a, 2020a).

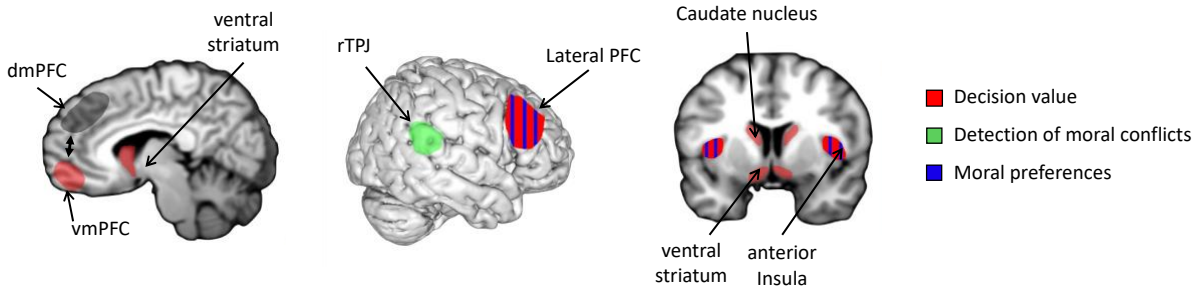

Figure 3. Brain networks involved in moral decision-making. Brain network computing decision value in moral context including the vmPFC, striatum, lPFC and anterior insula (Crockett et al., 2017b; Hare et al., 2010b; Hutcherson et al., 2015b; Qu et al., 2019b, 2020b). The rTPJ signals a moral conflict reflecting the discrepancy between one's self-interest and moral rules (green) (Obeso et al., 2018), and the lPFC encodes moral preferences, reflecting the individual's degree of adherence to moral rules (blue) (Crockett et al., 2017b; Gao et al., 2018; Qu et al., 2020b; Zhu et al., 2014). Translating moral norms into moral behavior involves changes in functional connectivity between brain regions, as reflected by the vmPFC, that computes the decision value of an immoral offer, and enhances its functional coupling with components of the mentalizing network (dmPFC), depending upon the beneficiary of an immoral action (Qu et al., 2020b). Abbreviations: rTPJ, right temporo-parietal junction;; vmPFC ventromedial prefrontal cortex; lPFC, lateral prefrontal cortex, dmPFC dorsomedial prefrontal cortex.

According to a second account, in addition to the classical valuation system, there may also be distinct neural substrates engaged by moral value computation, which preside choices that weigh moral against monetary cost/benefit. Thus, even if the computational principles underlying valuation of moral and value-based decisions are similar (weighing self-monetary profits against moral costs/harm), moral decisions also engage brain regions not observed in non-moral value-based decision making (**Figure 1**). In one recent fMRI study, we investigated how the brain weighs the benefits and costs of moral and monetary payoffs when moral values and monetary payoffs are at odds in two situations: when deciding whether to earn money by contributing to a 'bad cause' (amoral behavior) and when deciding whether to sacrifice money to contribute to a 'good cause' (prosocial behavior) (Qu et al., 2019b). Using a neurocomputational model of decision value and fMRI, we showed that similar principles of decision value computations were used to solve these dilemmas, but that they engaged two distinct valuation systems. When weighing monetary benefits and moral costs, people were willing to trade their moral values in exchange for money, an effect accompanied by decision value computation engaging the anterior insula and the lateral prefrontal cortex. In contrast, weighing monetary costs against compliance with one's moral values engaged the ventral putamen. This is consistent with the proposal that there are distinct valuation systems for two types of considerations: one treating violations of moral norms as aversive outcomes and another treating compliance with moral rules as a rewarding outcome (Rangel et al., 2008). Another recent fMRI study also indicated that moral considerations do not simply engage the standard valuation brain system, since the rTPJ was observed to be specifically engaged in encoding moral values (Ugazio et al., 2019). These findings indicate that similar computational rules are applied by brain systems outside of the classical brain valuation system.

In a follow up paper (Obeso et al., 2018), we disentangled three possible functions of the right temporo-parietal junction (rTPJ) for human altruism, namely: implementing the motivation to help, signaling conflicts between moral and material values, or representing social reputation concerns (**Figure 4**). Again, we used a donation-decision task consisting of decisions requiring trade-offs between either positive moral values and monetary cost when donating to a good cause, or negative moral values and monetary benefits when sending money to a bad cause. We used a technique known as transcranial magnetic stimulation (TMS) over the the right Temporo-parietal Junction (rTPJ), to test the causal role of this brain region in specific moral processes. Disrupting the rTPJ using TMS did not change the general motivation to give or to react to social reputation cues, but specifically reduced the behavioral impact of moral-material conflicts (**Figure 4**). This finding reveals that signaling moral-material conflict is a core rTPJ mechanism that may contribute to a variety of human moral behaviors.

Additional evidence supports that moral decision computations require nodes outside the classical brain valuation system, including the dlPFC, insula and the rTPJ. For example, the lPFC responds more strongly when harming others for a small relative to a larger profit (Crockett et al., 2017a), agreeing with previous work showing that lPFC responds to moral norm violations (Chang & Koban, 2013; Ruff et al., 2013). Altruistic people, who show higher positive moral preference scores, have to

overcome a stronger subjective moral cost to accept offers that profit themselves at the expense of their moral values, and this behavioral effect is associated with stronger dlPFC signals (Qu et al., 2020a). Together, these approaches indicate that neural computations engaged in moral tradeoffs do not simply engage the brain valuation system, but that other areas are recruited when performing moral decision computations.

Audience effect in moral and immoral behavior

Humans value not only extrinsic monetary rewards, but also their own morality and their image in the eyes of others. However, violation of moral norms is frequent, especially when people know that they are not under scrutiny. In fact, across many social animals, behavior is strongly influenced by whether or not actions are visible to others. Humans tend to behave in a more egoistic manner under guaranteed anonymity (Ariely et al., 2009; Bohnet & Frey, 1999), and more pro-socially when observed by others (Ariely et al., 2009; Izuma et al., 2010). Recent economic theories of prosocial behavior combine heterogeneity in individual greed and altruism with social image concerns, i.e. the extent to which we value how others think of us (Bénabou & Tirole, 2006). In these models, motivation is threefold: extrinsic (the material rewards associated with the action), intrinsic (the moral benefits associated with the action), and attached to image (the concerns for what others think of us). Thus, according to these models, humans exhibit preferences for anti- or pro-social behavior not because they are intrinsically bad or good, but because they weigh a mixture of these different sources of motivation. We have investigated the influence of audience (being observed by others), both with moral and amoral choices (Qu et al., 2019b) (**Figure 4**). There were two types of choices: to decide whether to earn money by contributing to a 'bad cause' or to decide whether to sacrifice money to contribute to a 'good cause'. Behaviorally, participants were more likely to choose the prosocial option when they were observed (i.e. making a donation in public, but earning money at a moral cost in private). Regardless of the type of dilemma, a brain network including anterior cingulate cortex, anterior insula and the rTPJ was more engaged in public than in private settings. These findings identify how the brain processes three sources of motivation when weighing extrinsic rewards, moral values and concerns for image.

While it is now readily admitted that reputational concerns promote prosociality in adults, their ontogenetic origins remain poorly understood (Ahmed et al., 2020; Engelmann & Rapp, 2018; Leimgruber et al., 2012). Recent studies proposed that at about 5 years of age, children become concerned for their reputations and that they become more prosocial in public compared to private settings. In middle childhood, at approximately 8 years of age, children acquire further abilities to control the image they project and start to reason explicitly with concern for their reputation. In adolescents, researchers have investigated peer influence on prosocial behavior using a paradigm where participants could revise their initial donation decisions after learning about the donations of others (age range: 11-35 years) (Ahmed et al., 2020; Chierchia et al., 2020). These studies indicate that social influence on prosocial behavior was stronger in young adolescents than adults.

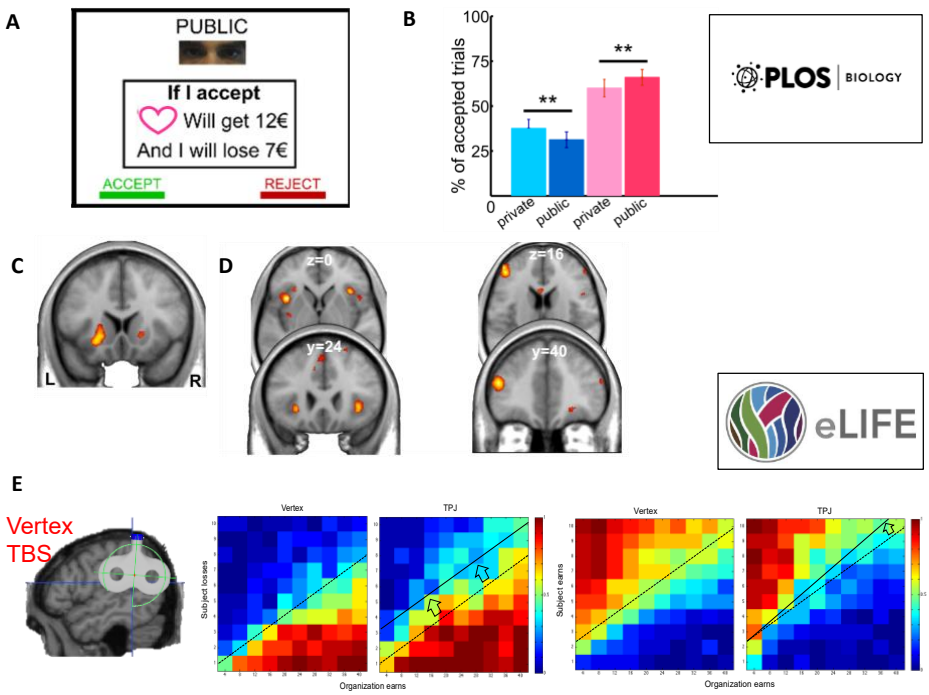

Figure 4. A-D. Two distinct valuation systems are engaged when weighing monetary benefits and moral costs (decision value computation engaging the anterior insula and the lateral prefrontal cortex) while weighing monetary costs against compliance with one's moral values engaged the ventral putamen. At the behavioural level, people were willing to trade their moral values in exchange for money, and there was an audience effect (people donating more to charity in public and less to a negatively valued organization) (Qu et al., 2019b). E. Stimulation of the right Temporo Parietal Junction (rTPJ) using TMS reduced the behavioral impact of moral-material conflicts. These findings reveal that signalling moral-material conflict is a core rTPJ mechanism (Obeso et al., 2018). The color matrix represent the probability of accepting the monetary transfer (more redish color) at control site (vertex) versus rTPJ cTBS stimulation.

Conclusions and Implications for education, policy, teaching and learning

Understanding the development and neural mechanisms underlying moral behavior is important from a fundamental knowledge perspective. However, one can ask what are the practical implications of this neuroscience research for moral education and for helping children learn moral rules to become caring adults latter. One primary aspect is to transmit moral cognition neuroscience knowledge to educators and teachers because they are important role models for students **(Han, 2019)**. However, because the direct implementation of the neuroscience of morality in education may be difficult in practice, another aspect is to derive useful heuristics from our neuroscience knowledge to enhance teachers' moral educational activities. For example, a moral education program that used close-other exemplars (e.g., friends and family members) was directly inspired by neuroimaging findings **(Han et al., 2017)**. This example presents educators with how to obtain heuristics from neuroscience while developing educational activities. A third aspect is that, as presented above, the neuroscience of moral decision making has recently benefited from a neurocomputational approach developed in adults. Understanding the neurocomputational bases underlying the development of moral orientations is only beginning. We believe that this approach can now be applied to children and adolescents to understand the neurocomputational mechanisms underlying moral choices. Finally, although this brief focuses on moral decision making, understanding how moral values are learned by the brain could greatly benefit from a neurocomputational approach based on reinforcement learning (see other brief on reinforcement learning for the classroom). This approach suggests that associative learning principles can help to understand moral learning behavior **(FeldmanHall & Dunsmoor, 2019; Qu et al., in press)**. We believe that developmental social neuroscience research, combined with a computational neuroscience approach (eg. model-based fMRI) has the potential to provide new directions for the study of moral development, and to develop more effective moral educational interventions based on neuroscience findings and developmental psychology.

References

Ahmed, S., Foulkes, L., Leung, J. T., Griffin, C., Sakhardande, A., Bennett, M., Dunning, D. L., Griffiths, K., Parker, J., Kuyken, W., Williams, J. M. G., Dalgleish, T., & Blakemore, S. J. (2020). Susceptibility to prosocial and antisocial influence in adolescence. *Journal of Adolescence*, *84*, 56–68. https://doi.org/10.1016/j.adolescence.2020.07.012

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review*, *99*(1), 544–555.

Batson, C. D., Kennedy, C. L., Nord, L., Stocks, E., Fleming, D. A., Marzette, C. M., Lishner, D. A., Hayes, R. E., Kolchinsky, L. M., & Zerger, T. (2007). Anger at unfairness: Is it moral outrage? *European Journal of Social Psychology*, *37*(6), 1272–1285.

Baumgartner, T., Götte, L., Gügler, R., & Fehr, E. (2012). The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. *Human Brain Mapping*, *33*(6), 1452–1469.

Bellucci, G., Camilleri, J. A., Iyengar, V., Eickhoff, S. B., & Krueger, F. (2020). The emerging neuroscience of social punishment: Meta-analytic evidence. *Neuroscience & Biobehavioral Reviews*, *113*, 426–439. https://doi.org/10.1016/j.neubiorev.2020.04.011

Bellucci, G., Chernyak, S., Hoffman, M., Deshpande, G., Dal Monte, O., Knutson, K. M., Grafman, J., & Krueger, F. (2017). Effective connectivity of brain regions underlying third-party punishment: Functional MRI and Granger causality evidence. *Social Neuroscience*, *12*(2), 124–134.

Bénabou, R., & Tirole, J. (2006). Incentives and Prosocial Behavior. *American Economic Review*, *96*(5), 1652–1678. https://doi.org/10.1257/aer.96.5.1652

Blake, P., McAuliffe, K., Corbit, J., Callaghan, T., Barry, O., Bowie, A., Kleutsch, L., Kramer, K., Ross, E., & Vongsachang, H. (2015). The ontogeny of fairness in seven societies. *Nature*, *528*(7581), 258–261.

Bohnet, I., & Frey, B. S. (1999). The sound of silence in prisoner's dilemma and dictator games. *Journal of Economic Behavior & Organization*, *38*(1), 43–57.

Boyd, R., & Richerson, P. J. (1988). *Culture and the evolutionary process*. University of Chicago press.

Chang, L. J., & Koban, L. (2013). Modeling Emotion and Learning of Norms in Social Interactions. *Journal of Neuroscience*, *33*(18), 7615–7617. https://doi.org/10.1523/jneurosci.0973-13.2013

Chierchia, G., Piera Pi-Sunyer, B., & Blakemore, S.-J. (2020). Prosocial Influence and Opportunistic Conformity in Adolescents and Young Adults. *Psychological Science*, *31*(12), 1585–1601. https://doi.org/10.1177/0956797620957625

Cowell, J. M., & Decety, J. (2015). Precursors to morality in development as a complex interplay between neural, socioenvironmental, and behavioral facets. *Proceedings of the National Academy of Sciences*, *112*(41), 12657–12662. https://doi.org/10.1073/pnas.1508832112

Cowell, J. M., Lee, K., Malcolm-Smith, S., Selcuk, B., Zhou, X., & Decety, J. (2017). The development of generosity and moral cognition across five cultures. *Developmental Science*, *20*(4), e12403. https://doi.org/10.1111/desc.12403

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017a). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, *May*, 1–10. https://doi.org/10.1038/nn.4557

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017b). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, *20*(6), 879–885. https://doi.org/10.1038/nn.4557

Dahl, A., & Kim, L. (2014). Why is it bad to make a mess? Preschoolers' conceptions of pragmatic norms. *Cognitive Development*, *32*, 12–22.

Dahl, A., & Paulus, M. (2019). From Interest to Obligation: The Gradual Development of Human Altruism. *Child Development Perspectives*, *13*(1), 10–14. https://doi.org/10.1111/cdep.12298

Dahl, A., Waltzer, T., & Gross, R. L. (2017). Helping, Hitting and Developing: Toward a Constructivist–Interactionist Account of Early Morality. In *New perspectives on moral development* (pp. 33–54). Routledge.

De Quervain, D. J. F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The Neural Basis of Altruistic Punishment. *Science*, *305*(5688), 1254.

Decety, J. (2020). *The social brain: A developmental perspective*. MIT Press.

Decety, J., & Cowell, J. M. (2018). Interpersonal harm aversion as a necessary foundation for morality: A developmental neuroscience perspective. *Development and Psychopathology*, *30*(1), 153–164.

Decety, J., Michalska, K. J., & Kinzler, K. D. (2012). The contribution of emotion and cognition to moral sensitivity: A neurodevelopmental study. *Cerebral Cortex*, *22*(1), 209–220.

Decety, J., Steinbeis, N., & Cowell, J. M. (2021). The neurodevelopment of social preferences in early childhood. *Current Opinion in Neurobiology*, *68*, 23–28. https://doi.org/10.1016/j.conb.2020.12.009

Decety, J., & Wheatley, T. (2015). *The moral brain: A multidisciplinary perspective* (p. 327). The MIT Press.

Dunne, S., & O'Doherty, J. P. (2013). Insights from the application of computational neuroimaging to social neuroscience. *Current Opinion in Neurobiology*, *23*(3), 387–392. https://doi.org/10.1016/j.conb.2013.02.007

Engelmann, J. M., & Rapp, D. J. (2018). The influence of reputational concerns on children's prosociality. *Current Opinion in Psychology*, *20*, 92–95. https://doi.org/10.1016/j.copsyc.2017.08.024

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785–791.

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185–190.

FeldmanHall, O., & Dunsmoor, J. E. (2019). Viewing Adaptive Social Choice Through the Lens of Associative Learning. *Perspectives on Psychological Science*, *14*(2), 175–196. https://doi.org/10.1177/1745691618792261

Feng, C., Luo, Y.-J., & Krueger, F. (2015). Neural signatures of fairness-related normative decision making in the ultimatum game: A coordinate-based meta-analysis. *Human Brain Mapping*, *36*(2), 591–602. https://doi.org/10.1002/hbm.22649

Frost, R., & McNaughton, N. (2017). The neural basis of delay discounting: A review and preliminary model. *Neuroscience & Biobehavioral Reviews*, *79*, 48–65. https://doi.org/10.1016/j.neubiorev.2017.04.022

Gao, X., Yu, H., Sáez, I., Blue, P. R., Zhu, L., Hsu, M., & Zhou, X. (2018). Distinguishing neural correlates of context-dependent advantageous- and disadvantageous-inequity aversion. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(33), E7680–E7689. https://doi.org/10.1073/pnas.1802523115

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389–400. https://doi.org/10.1016/j.neuron.2004.09.027

Greene, J. D., Sommerville, R. B., & Nystrom, L. E. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science (New York, N.Y.)*, *293*, 2105–2108.

Hamlin, J. K., Mahajan, N., Liberman, Z., & Wynn, K. (2013). Not Like Me = Bad: Infants Prefer Those Who Harm Dissimilar Others. *Psychological Science*, *24*(4), 589–594. https://doi.org/10.1177/0956797612457785

Han, H. (2019). *Neuroscience of morality and teacher education*. SocArXiv. https://doi.org/10.31235/osf.io/97g3e

Han, H., Kim, J., Jeong, C., & Cohen, G. L. (2017). Attainable and Relevant Moral Exemplars Are More Effective than Extraordinary Exemplars in Promoting Voluntary Service Engagement. *Frontiers in Psychology*, *8*, 283. https://doi.org/10.3389/fpsyg.2017.00283

Hare, T. A., Camerer, C. F., Knoepfle, D. T., & Rangel, A. (2010a). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *Journal of Neuroscience*, *30*(2), 583–590. https://doi.org/10.1523/JNEUROSCI.4089-09.2010

Hare, T. A., Camerer, C. F., Knoepfle, D. T., & Rangel, A. (2010b). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *Journal of Neuroscience*, *30*(2), 583–590. https://doi.org/10.1523/JNEUROSCI.4089-09.2010

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015a). A Neurocomputational Model of Altruistic Choice and Its Implications. *Neuron*, *87*(2), 451–462. https://doi.org/10.1016/j.neuron.2015.06.031

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015b). A Neurocomputational Model of Altruistic Choice and Its Implications. *Neuron*, *87*(2), 451–462. https://doi.org/10.1016/j.neuron.2015.06.031

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron*, *58*(2), 284–294. https://doi.org/10.1016/j.neuron.2008.03.020

Izuma, K., Saito, D. N., & Sadato, N. (2010). Processing of the incentive for social approval in the ventral striatum during charitable donation. *Journal of Cognitive Neuroscience*, *22*(4), 621–631.

Josephs, M., & Rakoczy, H. (2016). Young children think you can opt out of social-conventional but not moral practices. *Cognitive Development*, *39*, 197–204.

Kahn Jr., P. H. (1992). Children's Obligatory and Discretionary Moral Judgments. *Child Development*, *63*(2), 416–430. https://doi.org/10.1111/j.1467-8624.1992.tb01637.x

Killen, M., & Smetana, J. G. (2015). *Origins and development of morality.*

Killen, M., & Turiel, E. (1998). Adolescents' and Young Adults' Evaluations of Helping and Sacrificing for Others. *Journal of Research on Adolescence*, *8*(3), 355–375. https://doi.org/10.1207/s15327795jra0803_4

Konovalov, A., Hu, J., & Ruff, C. C. (2018). Neurocomputational approaches to social behavior. *Current Opinion in Psychology*, *24*, 41–47. https://doi.org/10.1016/j.copsyc.2018.04.009

Krueger, F., & Hoffman, M. (2016). The emerging neuroscience of third-party punishment. *Trends in Neurosciences*, *39*(8), 499–501.

Leimgruber, K. L., Shaw, A., Santos, L. R., & Olson, K. R. (2012). Young children are more generous when others are aware of their actions. *PloS One*, *7*(10), e48292.

Lopez-Persem, A., Rigoux, L., Bourgeois-Gironde, S., Daunizeau, J., & Pessiglione, M. (2017). Choose, rate or squeeze: Comparison of economic value functions elicited by different behavioral tasks. *PLoS Computational Biology*, *13*(11), 1–18. https://doi.org/10.1371/journal.pcbi.1005848

Miller, J. G., Bersoff, D. M., & Harwood, R. L. (1990). Perceptions of social responsibilities in India and in the United States: Moral imperatives or personal decisions? *Journal of Personality and Social Psychology*, *58*(1), 33–47. https://doi.org/10.1037/0022-3514.58.1.33

Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourão-Miranda, J., Andreiuolo, P. A., & Pessoa, L. (2002). The Neural Correlates of Moral Sensitivity: A Functional Magnetic Resonance Imaging Investigation of Basic and Moral Emotions. *The Journal of Neuroscience*, *22*(7), 2730 LP – 2736. https://doi.org/10.1523/JNEUROSCI.22-07-02730.2002

Moll, J., Zahn, R., Oliveira-souza, R. D., & Krueger, F. (2005). The neural basis of human moral cognition. *Nature Reviews Neuroscience*, *6*, 799–809.

Nucci, L., & Weber, E. K. (1995). Social interactions in the home and the development of young children's conceptions of the personal. *Child Development*, *66*(5), 1438–1452.

Obeso, I., Moisa, M., Ruff, C. C., & Dreher, J.-C. (2018). A causal role for right temporo-parietal junction in signaling moral conflict. *ELife*, *7*, e40671. https://doi.org/10.7554/eLife.40671

Park, S. A., Goïame, S., O'Connor, D. A., & Dreher, J.-C. (2017). Integration of individual and social information for decision-making in groups of different sizes. *PLOS Biology*, *15*(6), e2001958. https://doi.org/10.1371/journal.pbio.2001958

Qu, C., Benistant, J., & Dreher, J.-C. (in press). Neurocomputational mechanisms engaged in moral choices and moral learning. *Neuroscience and Biobehavioral Reviews*.

Qu, C., Hu, Y., Tang, Z., Derrington, E., & Dreher, J.-C. (2020a). Neurocomputational mechanisms underlying immoral decisions benefiting self or others. *Social Cognitive and Affective Neuroscience*, *January*, 135–149. https://doi.org/10.1093/scan/nsaa029

Qu, C., Hu, Y., Tang, Z., Derrington, E., & Dreher, J.-C. (2020b). Neurocomputational mechanisms underlying immoral decisions benefiting self or others. *Social Cognitive and Affective Neuroscience*, *15*(2), 135–149. https://doi.org/10.1093/scan/nsaa029

Qu, C., Météreau, E., Butera, L., Villeval, M. C., & Dreher, J.-C. (2019a). Neurocomputational mechanisms at play when weighing concerns for extrinsic rewards, moral values, and social image. *PLOS Biology*, *17*(6), e3000283. https://doi.org/10.1371/journal.pbio.3000283

Qu, C., Météreau, E., Butera, L., Villeval, M. C., & Dreher, J.-C. (2019b). Neurocomputational mechanisms at play when weighing concerns for extrinsic rewards, moral values, and social image. *PLOS Biology*, *17*(6), e3000283. https://doi.org/10.1371/journal.pbio.3000283

Raihani, N. J., & McAuliffe, K. (2012). Human punishment is motivated by inequity aversion, not a desire for reciprocity. *Biology Letters*, *8*(5), 802–804.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews. Neuroscience*, *9*(7), 545–556. https://doi.org/10.1038/nrn2357

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2012). No third-party punishment in chimpanzees. *Proceedings of the National Academy of Sciences*, *109*(37), 14824–14829.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A Neural Basis for Social Cooperation. *Neuron*, *35*(2), 395–405. https://doi.org/10.1016/S0896-6273(02)00755-9

Ruff, C. C., Ugazio, G., & Fehr, E. (2013). Changing social norm compliance with noninvasive brain stimulation. *Science (New York, N.Y.)*, *342*(6157), 482–484. https://doi.org/10.1126/science.1241399

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The Neural Basis of Economic Decision-Making in the Ultimatum Game. *Science*, *300*(5626), 1755–1758. https://doi.org/10.1126/science.1082976

Schmidt, M. F., Rakoczy, H., & Tomasello, M. (2012). Young children enforce social norms selectively depending on the violator's group affiliation. *Cognition*, *124*(3), 325–333.

Sescousse, G., Caldú, X., Segura, B., & 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), 681–696. https://doi.org/10.1016/j.neubiorev.2013.02.002

Smetana, J. G., Ball, C. L., Jambon, M., & Yoo, H. N. (2018). Are young children's preferences and evaluations of moral and conventional transgressors associated with domain distinctions in judgments? *Journal of Experimental Child Psychology*, *173*, 284–303. https://doi.org/10.1016/j.jecp.2018.04.008

Smetana, J. G., Jambon, M., & Ball, C. (2014). The social domain approach to children's moral and social judgments. *Handbook of Moral Development*, *2*, 23–45.

Smetana, J. G., Toth, S. L., Cicchetti, D., Bruce, J., Kane, P., & Daddis, C. (1999). Maltreated and nonmaltreated preschoolers' conceptions of hypothetical and actual moral transgressions. *Developmental Psychology*, *35*(1), 269.

Suzuki, S., & O'Doherty, J. P. (2020). Breaking human social decision making into multiple components and then putting them together again. *Cortex*, *127*, 221–230. https://doi.org/10.1016/j.cortex.2020.02.014

Tan, E., Mikami, A. Y., & Hamlin, J. K. (2018). Do infant sociomoral evaluation and action studies predict preschool social and behavioral adjustment? *Journal of Experimental Child Psychology*, *176*, 39–54. https://doi.org/10.1016/j.jecp.2018.07.003

Ting, F., Dawkins, M. B., Stavans, M., & Baillargeon, R. (2019). Principles and concepts in early moral cognition. *The Social Brain: A Developmental Perspective*.

Turiel, E. (1983). *The Development of Social Knowledge: Morality and Convention*. Cambridge University Press.

Turiel, E. (2015). Morality and prosocial judgments. *The Oxford Handbook of Prosocial Behavior*, 137–152.

Ugazio, G., Grueschow, M., Polania, R., Lamm, C., Tobler, P. N., & Ruff, C. C. (2019). Neuro-Computational Foundations of Moral Preferences. *BioRxiv*, 801936. https://doi.org/10.1101/801936

Van de Vondervoort, J. W., & Hamlin, J. K. (2017). Preschoolers' social and moral judgments of third-party helpers and hinderers align with infants' social evaluations. *Journal of Experimental Child Psychology*, *164*, 136–151. https://doi.org/10.1016/j.jecp.2017.07.004

Warneken, F., & Tomasello, M. (2006). Altruistic Helping in Human Infants and Young Chimpanzees. *Science*, *311*(5765), 1301–1303. https://doi.org/10.1126/science.1121448

Zaki, J., & Mitchell, J. P. (2011). Equitable decision making is associated with neural markers of intrinsic value. *Proceedings of the National Academy of Sciences of the United*

States of America, *108*(49), 19761–19766. PubMed. https://doi.org/10.1073/pnas.1112324108

Zhong, S., Chark, R., Hsu, M., & Chew, S. H. (2016). Computational substrates of social norm enforcement by unaffected third parties. *NeuroImage*, *129*, 95–104.

Zhu, L., Jenkins, A. C., Set, E., Scabini, D., Knight, R. T., Chiu, P. H., King-Casas, B., & Hsu, M. (2014). Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. *Nature Neuroscience*, *17*(10), 1319–1321. https://doi.org/10.1038/nn.3798