Mentalizing in value-based social decision-making: shaping expectations and social norms

Claudia Civai[1] & Alan Sanfey[2]

[1]Division of Psychology, School of Applied Science, London South Bank University, London UK

[2]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

Abstract

In this chapter, we take a neuroeconomic perspective to explore how the ability to understand mental states of other and predict their behaviour, termed mentalizing, is crucial in value-based social decision-making. These types of choice involve attributing value to social stimuli and motivations in order to inform decisions. Here, (i) we define the concept of value and value-based choice; then, (ii) explain the ways in which mentalizing is integrated into the computation of these choices in social interaction. In (iii) and (iv), we outline the link between mentalizing and social expectations, and how our ability to learn from social interactions and predict behaviour shape our social norms and, therefore, our ability to make optimal decisions in social contexts. To conclude, (v) we analyse how mentalizing allows for flexibility in social expectations and for context-dependent decision-making processes, and (vi) how individual differences in mentalizing ability help explain variability in social decision-making. Overall, we argue that mentalizing is an essential component of social decision-making, and also should be taken into account in applied settings, such as clinical and forensic.

Value-based decision-making

Everyday life is defined by choice: deciding what to have for breakfast, what clothes to wear, what career to pursue, or who to befriend. All complex cognitive processes that involve evaluating alternatives and eventually selecting one option that is deemed preferable. Understanding how value is assigned to each alternative is central to explaining the psychological mechanisms of decision-making. One recently proposed, and effective, means to investigate value-based choices is a neuroeconomic approach, which integrates strengths of different disciplines, providing a clear operationalisation of value using utility functions (from economics) and fine-grained and multi-layered explanations of the cognitive and neural mechanisms involved in the evaluation process (from psychology and

neuroscience) (Rangel, Camerer and Montague, 2008). From this rich body of literature, the existence of a neural circuit underpinning value computation has emerged, encompassing subcortical and cortical areas; a recent meta-analysis (Clithero & Rangel, 2014) found that the posterior cingulate cortex (PCC), the ventral striatum (VStr), and the medial part of the orbitofrontal, also referred to ventromedial prefrontal cortex (vmPFC)[1] play crucial roles in integrating external and internal information, and thus eventually determining the subjective value of a choice option. This system underpins the evaluation of any stimulus characterised by relevant valence in the context of the decision process, with the system is commonly activated for primary and secondary incentives, such as food or money (McClure, Ericson, Laibson, Lowenstein, & Cohen 2007; Knutson, Westdorp, Kaiser, & Hommer, 2000), as well as social incentives, such as praise or good reputation (Rilling et al, 2002; Izuma, Saito, & Sadato, 2008).

Whether social decision-making, e.g., to decide to cooperate with another, as opposed to individual decision-making, e.g., to decide what to eat for dinner, should be considered as a separate aspect of the cognitive system, or whether the same basic processes apply to both individual and social decisions, is an issue of debate; these two accounts are, however, not necessarily mutually exclusive (see Ruff & Fehr, 2014). On the one hand, there are cognitive mechanisms whose existence are potentially uniquely associated with socio-cognitive processing, such as Theory of Mind (ToM) and empathy, in that their involvement in cognitive processing depends on the presence of other individuals. Supporting the idea that social cognition is a unique mechanism, there is evidence of brain areas, such as the temporo-parietal junction (TPJ), that are specifically associated with the multidimensional domain of ToM (see Schurz, Radua, Aichhorn, Richlan, & Perner, 2014 for a meta-analysis; Lee & Seo, 2016). On the other hand, when considering incentives and motivation, the neuroscientific findings are more in line with the idea of a common mechanism that

[1] Dixon, Thiruchselvam, Todd, and Christoff (2017), in their recent exhaustive review on the anatomical and functional parcellation of the prefrontal cortex, distinguish a lateral OFC, underpinning the evaluation of external stimuli in relation to internal goals, and a medial OFC, or ventromedial prefrontal cortex (vmPFC), integrating the evaluation of the external stimuli with internally-generated scenarios and contributing to value-based decision-making. In relation to mentalizing, while the OFC evaluates others on the basis of external features, more dorsal areas of the medial prefrontal cortex are involved in evaluating others' mental states.

evaluates both social and non-social rewards and guides both types of behaviour accordingly (Fehr & Camerer, 2007; Clithero & Rangel, 2008; Ruff & Fehr, 2014).

Economic theory, in particular behavioural economics, has formalised the social aspect of these processes by incorporating the other agent into the utility function, which describes mathematically the value attached to a decision outcome. Theories of other-regarding preferences are so-called because they consider the presence of other individuals and their status (e.g., economic payoff) as an integral part of the scenario that a person evaluates when making a social decision. These models incorporate the other's payoff as a parameter in summarizing the final outcome of the decision: for example, the parameter value representing the payoff to the other agent is crucial in determining whether, and how much, an individual is averse to inequality (Fehr & Schmidt, 1999); similarly, assessing the importance ascribed to the other's intentions is central to understanding decisions to reciprocate good, as well as bad, behaviour (Charness & Rabin, 2002; Dufwenberg & Kirchsteiger, 2004). The utility associated with these complex decision outputs is associated with neural signals in the vmPFC and striatum, key areas for encoding both social and non-social rewards. Tricomi and colleagues, for instance, found that activation in the ventral striatum was stronger when people could increase the payoff of a disadvantaged group, at a cost to themselves, and re-establish equality in the exchange (Tricomi, Rangel, Camerer, & O'Doherty, 2010); similarly, the striatum showed increased activated when people were given the chance to engage in costly punishment of injustice and unfairness (de Quervain, et al, 2004; Strobel et al, 2011; Stallen et al, 2018).

Mentalizing in strategic and prosocial value-based interactions

How is the other-regarding element integrated into the valuation process? Computing social signals and integrating them into subjective valuation involves mentalizing: choosing to set aside one's self interest to be generous and charitable, or simply fair, requires the ability to take the other's perspective into consideration, and to understand the feelings and beliefs of the other. As previously mentioned, these cognitive mechanisms, which are intrinsically linked to the social context, are underpinned by specific brain areas, encompassing both

posterior, i.e., temporo-parietal junction (TPJ) and PCC, and anterior, i.e., medial prefrontal cortex (mPFC), regions (Schurtz et al, 2014).

These mentalizing areas have been implicated in different aspects of social decision-making tasks, and the evaluation of other-regarding preferences in particular. For example, the activity of mPFC, and especially its dorsal region, has been positively linked to understanding and correctly predicting others' preferences (Kang, Lee, Sil & Kim, 2013), as well as other-regarding values in a reward-based task, i.e., the utility that others would derive from a specific choice made by the decision maker themselves (Sul et al, 2015). Interestingly, Sul and colleagues found a gradient in the mPFC, whereby dorsomedial areas (dmPFC) represented other-regarding values and ventromedial areas (vmPFC) correlated with self-regarding values. Selfish individuals showed a clear regional differentiation, with vmPFC active for self-regarding values and dmPFC active for other-regarding values; conversely, prosocial individuals, while showing a higher vmPFC for personal rewards, lacked this gradient for other-regarding values, instead demonstrating equal strength of activation in both regions. This may be in line with the hypothesis that vmPFC computes an overall value-signal after integrating different pieces of information (e.g., Roy, Shohamy & Wager, 2012), which, in the case of other-regarding values, is higher for prosocial as compared to selfish individuals. This interpretation would concur with other findings, such as those by Suzuki et al (2012), who showed that vmPFC encodes the shared representation of self and other reward-prediction error, defined as the difference between what one gets and what one expected to get; and by Hutcherson and colleagues, who identified vmPFC[2] as the region that encoded both self and other rewards in a simple Dictator Game (see Box 1), where participants can decide how to split a sum of money between themselves and another powerless player (Hutcherson, Bushong & Rangel, 2015). Conversely, dmPFC may be an area that is specifically recruited to compute other-regarding values.

There is ample evidence that supports the involvement of TPJ, specifically in the right hemisphere, in considering other-regarding preferences. For example, Hutcherson et al's neurocomputational model found that the activation of this area positively correlated with

[2] Examining the coordinates reported in this study suggests that the area labelled by the authors as vmPFC, in fact, encompasses both dorsal and ventral mPFC, as defined by Sul et al (2015). Hutcherson et al's findings may therefore support the idea that this area represents a shared value-computation, although the specificity of the spatial gradient is not analysed.

the amount of money allocated to the other person in the Dictator Game, suggesting that an other-regarding value signal is already encoded here. Morishima and colleagues performed a voxel-based morphometry analysis and found that grey matter volume of the right TPJ was positively associated with people's altruistic preferences in advantageous inequality situations i.e., when participants were better off than their task partner (Morishima, Schunk, Bruhin, Ruff & Fehr, 2012). Other findings support the link between value encoding and mentalizing, suggesting that the subjective evaluation of social stimuli depends on the strength of the functional connectivity between subjective-value areas (vmPFC) and social cognition areas (TPJ) (Smith, Clithero, Boltuck & Huettel 2014). For example, Strombach and colleagues found that the connectivity between these two areas was greater when people chose a generous rather than a selfish option in a social decision-making task, suggesting the integration of social signals into the final subjective evaluation in order to guide decisions (Strombach et al, 2015).

---------- INSERT BOX 1 ----------

Mentalizing shapes expectations

Why does the other's perspective need to be integrated into the subjective valuation that eventually determines decisions in social contexts? The goal of the valuation process is to determine the optimal outcome for the decision-maker; hence, being able to predict the various outcomes of all potential choice options is fundamental in order to select the best one. By making the other's state of mind available to the decision-maker, mentalizing allows for the prediction of the other's behavioural reactions in different scenarios, adding a crucial element to the choice process. Let us consider, for example, the case of the Ultimatum Game (See Box 1; Guth, Schmittberger & Schwarze, 1982): in this task, the first player (proposer) is asked to divide a sum of money, for example $10, with the second player (responder), who can decide whether to accept the offer of the proposer or reject it. If the offer is rejected, both players end up with zero, with no possibility to reopen negotiation. The Nash equilibrium for this game predicts that the proposer will offer the smallest amount of money that the responder is willing to accept; if the proposer thinks that the responder is an economically rational player and will accept anything higher than zero, then they will offer the minimum, e.g., $1, or even less. On the other hand, if the proposer thinks that the

responder is strictly egalitarian, they will offer half of the sum, because they fear that anything less will be rejected. Whichever solution is chosen by the proposer, it is clear that their allocation decision is based on the proposer's evaluation of the responder's perspective, and the prediction of how this will subsequently drive their behaviour to either accept or reject. In more general terms, it is possible to say that decisions are driven by the expectations that we hold about the outcome of a certain social interaction (e.g., keeping money because the responder has accepted our Ultimatum Game's offer), and these expectations are in turn shaped by our ability to take the other's perspective and predict their behaviour.

Predicting outcomes is therefore vitally important in order to select the right strategy, where, the "right strategy" is one that delivers the preferred outcome in that context (e.g., self-interest, altruistic/prosocial, egalitarian, etc). A useful example to clarify this concept comes from developmental science. Mentalizing ability develops with age, and younger children are generally less generous than older children (Benenson, Pascoe & Radmore, 2007). However, if taking someone else's perspective automatically resulted in greater generosity, then we should expect generosity to increase with perspective taking and mentalizing abilities. However, the results of a study by Cowell and colleagues challenge this position: in fact, when 3 to 5 year-olds were asked to play as proposers, or dictators, in a Dictator Game, which is similar to the Ultimatum Game with the crucial exception that responders passively receive offers without the opportunity to change the outcome, their sharing behaviour negatively correlated with ToM abilities. This suggests that these children were able to understand the other's perspective, but purposely choose not to be altruistic in a situation that did not involve reciprocity (Cowell, Samek, List & Decety, 2015). Conversely, when playing the Ultimatum Game, a situation that involves reciprocity, children with higher ToM abilities made fairer offers, suggesting that they were able to understand that unfair offers were more likely to be rejected (Takagishi, Kameshima, Schug, Koizumi & Yamagishi, 2010).

Neuroscientific evidence is also inconclusive with respect to the directionality of the relationship between brain activation associated with mentalizing abilities and altruistic behaviour. For example, Chang and colleagues found that TPJ was more active when people decided to reciprocate trust of the investor in a Trust Game (see Box 1; Chang, Smith,

Dufwenberg & Sanfey, 2011); on the other hand, van Baar and colleagues reported that another ToM key area, the posterior superior temporal sulcus (pSTS), was more active when people chose to not reciprocate trust (van Baar, Chang & Sanfey, 2019). Similarly, some studies have found TPJ to be involved with the decision to react to unfair behaviour (David, Hu, Krueger & Weber, 2017), whereas others found the area to be associated with the decision to refrain from punishing said behaviour (Stallen et al, 2018). Moreover, Buckholtz et al (2008) investigated punishing decisions and responsibility assessment in a legal context and found TPJ to be involved in assessing all levels of criminal responsibility. In conclusion, mentalizing is clearly associated with the evaluation of the other's perspective, but not necessarily in a way that predicts the directionality of behavioural outcomes.

We will see in the next section how expectations shape social norms, and how mentalizing allows us to adapt these norms to different situations.

From mentalizing to social norms

The Stanford Encyclopaedia of Philosophy defines social norms as "the informal rules that govern behaviour in groups and society, [… and] the unplanned result of individuals' interaction" (Bicchieri, Muldoon & Sountuoso, 2018). Social norms, such as fairness, cooperation, or trust, can therefore be interpreted as rules based on acquired expectations of the outcomes of social exchange, which have been learnt through repeated interactions. For this reason, mentalizing is crucial to the acquisition of social norms, as it is via this process that we are able to predict others' mental states and associated behaviours and, as a consequence, apply the correct social rule.

As mentioned above, social norms are rules that have been acquired through repeated exposure to social interactions. How does this learning happen? As proposed by Lee and Seo (2016), reinforcement learning theory can explain how we learn to respond to social tasks (i.e., apply social norms) that require decision-making. Model-free algorithms, where the likely outcome of each option of a decision task is compared to the value of the pre-decision state, can work in simple situations, such as when we have to choose between an apple and an orange. However, these algorithms are not suitable for complex and ever-changing

environments such as social ones, where the many variables involved are constantly changing through time; this is because model-free algorithms require many iterations of events to update the response to any small change in the context, and therefore complex situations would require too many repeated interactions for the learning to take place. For this reason, model-free algorithms might not correctly capture the way in which social norms are acquired. Model-based algorithms, on the other hand, compute the value of each option based on the decision-maker's knowledge of the situation, such as the other's beliefs, thoughts and emotional state. Computationally, these model-based algorithms are more complex, but, thanks to their flexibility, are also more suitable to successfully describe and predict how we make decisions in the social environment, and therefore better explain how social norms are acquired. Mentalizing makes model-based algorithms of social decision-making psychologically feasible and, importantly, this is also the mechanism that distinguishes social and non-social learning and decision-making. Neurophysiological evidence supports this distinction, in that neural areas specific to the evaluation of the other's outcomes (TPJ, pSTS) are involved in updating social prediction-error (Suzuki et al, 2012); in other words, mentalizing processes contribute to shaping expectations on others' social behaviours.

The results from Heijne and Sanfey (2015) clearly illustrate this distinction between social and non-social learning. The authors investigated the mechanisms of decision-making in a stay/leave social situation, in which participants had to choose whether or not to leave either a social or a non-social partner in order to succeed in a cooperative task; two studies were run, one in which participants had no information about their partner, and one in which they were given prior knowledge to shape their beliefs. The findings showed that, as expected, the (non)cooperative behaviour of the partner influenced the decision to stay or leave the relationship. Prioir beliefs also had an effect, biasing the decisions, though sometimes in a maladaptive way such as situations when beliefs about behaviour did not match actual behaviour: for example, when a partner was presented as cooperative, but their actual behaviour was non-cooperative, choosing to stay in the relationship would be considered maladaptive. Importantly, these results also showed that prior beliefs had a relatively weaker effect on the social choice compared to the non-social one: people used both their prior knowledge about their partner and their partner's actual behaviour in order

to make a decision about whether or not to stick with that partner; on the other hand, in the non-social context, participants were much more driven by their prior expectations. This supports the idea that social value-based decision-making cannot be fully explained using simple non-social reinforcement learning and reward prediction theory; other processes must be also accounted for, such as mentalizing, which allows us to understand that people might change their preferences.

Although being repeatedly exposed to other people's behaviours and mental states are a useful way to learn and follow social norms, observing another person's behaviour and predicting these behaviours are two different processes, with mentalizing playing a major role in the latter. Haroush and Williams (2015) found that non-human primates (rhesus monkeys) playing the Prisoner's Dilemma (see Box 1), a game in which mutual cooperation is required in order to achieve the best outcome for both players, would reciprocate both cooperative and non-cooperative choices; but, somewhat surprisingly, they found that a group of neurons in the dorsal anterior cingulate cortex (dACC) of the decision-maker would specifically encode and predict the other monkey's decision to cooperate before the decision was shown (players were required to decide simultaneously). Therefore, when the monkey saw their opponent's decision before choosing an action themselves, they successfully reciprocated; however, when the other's decision was unknown, monkeys chose to cooperate significantly more, suggesting that cooperation was the default norm. Importantly, these neurons exclusively encoded the predicted cooperative choice of the other, not of oneself; moreover, they were sensitive to social context, firing more when the monkeys were in the same room rather than in separate rooms. These interesting results suggest that (i) engaging in mentalizing, rather than simply observing behaviour, may lead to higher levels of cooperation, possibly because compliance with social norms is highly expected; (ii) social context is needed in order to trigger the social element of the decision process (Sanfey, Civai & Vavra, 2015). The latter is also supported by findings showing that the willingness to engage in fair and altruistic, but costly, behaviours diminishes when the all parties are guaranteed anonymity, hence eliminating reputational concerns (Kurtzban, DeScioli & O'Brien, 2007).

Expectations in flexible social environments

As previously mentioned, model-based algorithms of decision-making are better suited to deal with the unique complexity and flexibility of the social environment. Indeed, when it comes to predicting social norm compliance, expected behavioural outcomes vary dramatically depending on many different variables. For example, in order to determine whether an outcome is fair or unfair, intentionality plays a very important role. Findings showed that when responders in an Ultimatum Game know that the proposer was required to make an unfair offer (i.e. a 'no-alternative' condition), rejections of unfairness decrease significantly (Sutter, 2007). Interestingly, the anterior insula (AI), a key area in detecting social norm violations (Chang & Sanfey, 2011; Corradi-Dell'Acqua, Civai, Rumiati & Fink, 2012), was more active when participants rejected unfair offers in the no-alternative condition, and accepted unfair offers in the fair-alternative condition (Guroglu, van den Bos, Rombouts & Crone, 2011), suggesting that the act of rejecting an unfair offer when there is no alternative ,and accepting an unfair offer when the fair alternative is available, are both perceived as violations of a social norm. This indicates that fairness norms are context-dependent, and that, in the no-alternative condition, accepting unfair offers represents the social norm; therefore, AI here signals a behavioural deviation from the norm when rejecting unfairness. Noticeably, TPJ and mPFC are also more active when rejecting unfairness in the no-alternative condition, stressing the importance of mentalizing in adapting the norm to the context.

Other variables also influence our perception of fair outcomes: wealth and need, for example, are considered when deciding how to share resources, and people tend to prefer unequal outcomes that favour poorer groups (Tricomi et al, 2010); merit and effort are also integrated in evaluating context-dependent fairness: for instance, these variables determine the amount that people are willing to sacrifice in a Dictator Game where the amount of money to share is determined by the work of players (Frohlich, Oppenheimer & Kurki, 2004).

Expectations regarding the type of environment in which a decision is made also play a fundamental role. Sanfey (2009) and Chang and Sanfey (2011) show that manipulating expectations of responders in the Ultimatum Game change the likelihood of rejecting unfair

offers. Here, before playing the game, participants were led to believe that proposers would be either fair or unfair; as predicted, those who expected fair offers were much more likely to reject unfairness compared to those that expected unfair offers. As an extension to this, Vavra and colleagues show that not only the average, but also the variance of the expected distribution can influence participants' choices: specifically, the mean offer amount determined the threshold for accepting offers, whereas the variance of the offers determined how strictly participants adhered to this threshold (Vavra, Chang & Sanfey, 2018).

Overall, these results stress some important aspects . What is considered as 'fair' changes depending on different contextual variables (e.g., intentionality, merit, effort), and our value-based decisions change accordingly (e.g., we prefer an unequal outcome if we know that our opponent needs the resources more than we do); however, even when our perception of a fair outcome remains the same (e.g., in Sanfey (2009) people would still consider the equal outcome, 50:50, to be a fair share), our prior expectations regarding the chances of obtaining the preferred outcome also can shift our decision threshold: this means that we may decide to accept an unfair offer, even though we would still prefer a fair one, if unfairness is what we expect in a specific context.

Individual differences in mentalizing and social norm compliance

It is difficult to clearly understand the involvement of mentalizing in social value-based decision-making without considering individual differences. Quantifying the average behaviour of the population in specific social interactions can be very useful, for example to devise large-scale interventions such as social policies. Nevertheless, in order to understand the psychological roots underlying the multifaceted and multidimensional decision-making mechanisms, it is important to investigate the complex interactions between individual and contextual variables. For instance, in an Ultimatum Game people on average prefer fairness and offer an equal split most of the time; however, as previously mentioned, the amount of money that each individual proposer chooses to offer will depend on their own beliefs about their game partner and the situation more generally. Typically, how people react to social norm violations involves the interaction of many different variables. Some of these

variables are context-dependent, such as relative inequality of outcomes, reputation effects, need, merit, anonymity etc. Other variables are more directly tied to the decision-maker, such as age (Murnighan & Saxon, 1998; Bailey, Ruffman & Rendell, 2012), gender (Solnick, 2001), or political beliefs (Zettler & Hilbig, 2010). As far as mentalizing and perspective talking are concerned, studies show that when required to choose between punishing a perpetrator or assisting a victim of an injustice, people with higher empathic traits show an increased disposition towards helping behaviour (Leliveld, van Dijk & van Beest, 2012), with this attitude correlating with activation in TPJ (Hu et al, 2015). In a recent study, Civai and colleagues show that people classified as compensators, based on their preference to compensate the victim of an injustice rather than punish the perpetrator in a punishment/compensation task (see Box 1), have a stronger activation in TPJ as compared to people classified as punishers, i.e., people who prefer to punish a perpetrator rather than compensate a victim (Civai, Huijsmans & Sanfey, 2019). Somewhat in contrast to this, van den Bos and colleagues found that TPJ activation in a Trust Game was modulated by participants' subjective value orientation: prosocial participants, who care about their own as well as the other's gain, showed a higher TPJ activation when defecting, whereas proself individuals, who focus only on their own gain ignoring the other's, showed this association when reciprocating, suggesting that more prosocial individuals attended more to the need of the others when defecting their trust (van den Bos, van Dijk, Westenberg, Rombouts & Crone, 2009).

As previously mentioned, neither psychological nor neuroscientific evidence point to a clear relationship between mentalizing abilities and social preferences. However, moving from a localization view towards considering functional connectivity between regions can be a productive approach to explore this relationship (Vavra, van Baar & Sanfey, 2018). For example, van Baar et al (2019), in their version of the Trust Game, where participants must decide whether or not to reciprocate the trust of the investor, identified two types of reciprocators: those who reciprocate because they behave according to their own internal fairness norm (inequity-averse players), and those who take into account the investor's perspective, reciprocating in order to match the investor's expectations (guilt-averse players). The guilt-averse players show a stronger functional connectivity between TPJ and the vmPFC (as in Strombach et al, 2015) as compared to the other group, suggesting that

players with this specific approach to social interactions (i.e., avoiding guilt) integrate the other's perspective into the final value calculation, whereas players that use other strategies do not.

Conclusions

Mentalizing is an essential mechanism which allows us to evaluate available choice options in a social context and then make adecision: this process allows individuals to integrate the perspective of others in an attempt to better predict each of the potential outcomes and, eventually, to select the optimal solution. Neuroscientific evidence supports the idea that a specialised neural circuit encodes this information, and that the derived signal is then integrated with other aspects into an overall value signal that informs the decision-maker about the preferred option. Importantly, whilst the ability to mentalise and take the other's perspective is crucial in order to build a predictive model of the other's behaviour, it is not straightforward to correlate this to specific behavioural outcomes: for example, mentalizing and generosity are not always positively related. In order to understand the psychological roots of these mechanisms, individual differences must be taken into account to explain the multifaceted motivational drives that lead to the broad spectrum of behavioural outcomes.

To conclude, the ability to predict others' beliefs, emotions, and states of mind is fundamental to successful social decision-making; these observations have important implications when considering the effects that abnormal functioning of these mechanisms, either via brain damage or certain personality spectra, may have on people's ability to make optimal, or at least predictable, decisions. In the clinical setting, for example, suboptimal behaviour may get in the the way of rehabilitation, preventing a good recovery unless different strategies are adopted (Bechara, 2005); in the forensic setting, the issue of criminal responsibility is tightly linked to the concept of mental ability, and therefore establishing the extent of this trait, and any contributing factors, is extremely important (Gazzaniga, 2008). Therefore, it is important to consider these implications in settings where abnormal behaviour needs to be explained and taken into account for successfully addressing the issues at hand.

Bibliography

1. Bailey, P. E., Ruffman, T., & Rendell, P. G. (2012). Age-related differences in social economic decision making: the ultimatum game. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *68*(3), 356-363.

2. Bechara, A. (2005). Decision making, impulse control and loss of willpower to resist drugs: a neurocognitive perspective. *Nature neuroscience*, *8*(11), 1458.

3. Benenson, J. F., Pascoe, J., & Radmore, N. (2007). Children's altruistic behavior in the dictator game. *Evolution and Human Behavior*, *28*(3), 168–175. https://doi.org/10.1016/j.evolhumbehav.2006.10.003

4. Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, *10*(1), 122-142.

5. Bicchieri, C., Muldoon, R., & Sontuoso, A. (2018). Social Norms. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2018/entries/social-norms/>.

6. Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The Neural Correlates of Third-Party Punishment. *Neuron*, *60*(5), 930–940. https://doi.org/10.1016/j.neuron.2008.10.016

7. Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

8. Chang, L. J., & Sanfey, A. G. (2011). Great expectations: neural computations underlying the use of social norms in decision-making. *Social cognitive and affective neuroscience*, *8*(3), 277-284.

9. Chang, L. J., Smith, A., Dufwenberg, M., & Sanfey, A. G. (2011). Triangulating the Neural, Psychological, and Economic Bases of Guilt Aversion. *Neuron*, *70*(3), 560–572. https://doi.org/10.1016/j.neuron.2011.02.056

10. Charness, G., & Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, *117*(3), 817–869. https://doi.org/10.1162/003355302760193904

11. Civai, C., & Hawes, D. R. (2016). Game theory in neuroeconomics. In *Neuroeconomics* (pp. 13-37). Springer, Berlin, Heidelberg.

12. Civai, C., Huijsmans, I., & Sanfey, A. G. (2019). Neurocognitive mechanisms of reactions to second-and third-party justice violations. *Scientific reports*, *9*(1), 9271.

13. Corradi-Dell'Acqua, C., Civai, C., Rumiati, R. I., & Fink, G. R. (2012). Disentangling self- and fairness-related neural mechanisms involved in the ultimatum game: an fMRI study. *Social cognitive and affective neuroscience*, *8*(4), 424-431.

14. Cowell, J. M., Samek, A., List, J., & Decety, J. (2015). The Curious Relation between Theory of Mind and Sharing in Preschool Age Children. *PLOS ONE*, *10*(2), e0117947. https://doi.org/10.1371/journal.pone.0117947

15. David, B., Hu, Y., Krüger, F., & Weber, B. (2017). Other-regarding attention focus modulates third-party altruistic choice: An fMRI study. *Scientific Reports*, *7*, 43024. https://doi.org/10.1038/srep43024

16. 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 (New York, N.Y.)*, *305*(5688), 1254–1258. https://doi.org/10.1126/science.1100735

17. Dixon, M. L., Thiruchselvam, R., Todd, R., & Christoff, K. (2017). Emotion and the prefrontal cortex: An integrative review. *Psychological bulletin*, *143*(10), 1033.

18. Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*(2), 268–298. https://doi.org/10.1016/j.geb.2003.06.003

19. Edele, A., Dziobek, I., & Keller, M. (2013). Explaining altruistic sharing in the dictator game: The role of affective empathy, cognitive empathy, and justice sensitivity. *Learning and individual differences*, *24*, 96-102.

20. Falk, A., Fehr, E., & Fischbacher, U. (2003). On the nature of fair behavior. *Economic inquiry*, *41*(1), 20-26.

21. Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, *114*(3), 817-868.

22. Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and human behavior*, *25*(2), 63-87.

23. Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends in cognitive sciences*, *11*(10), 419-427.

24. Frohlich, N., Oppenheimer, J., & Kurki, A. (2004). Modeling Other-Regarding Preferences and an Experimental Test. *Public Choice*, *119*(1), 91–117. https://doi.org/10.1023/B:PUCH.0000024169.08329.eb

25. Gazzaniga, M. S. (2008). The law and neuroscience. *Neuron*, *60*(3), 412-415.

26. Güroğlu, B., van den Bos, W., Rombouts, S. A. R. B., & Crone, E. A. (2010). Unfair? It depends: Neural correlates of fairness in social context. *Social Cognitive and Affective Neuroscience*, *5*(4), 414–423. https://doi.org/10.1093/scan/nsq013

27. Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367–388. https://doi.org/10.1016/0167-2681(82)90011-7

28. Hu, Y., Strang, S., & Weber, B. (2015). Helping or punishing strangers: neural correlates of altruistic decisions as third-party and of its relation to empathic concern. *Frontiers in Behavioral Neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00024

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

30. 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

31. Kang, P., Lee, J., Sul, S., & Kim, H. (2013). Dorsomedial prefrontal cortex activity predicts the accuracy in estimating others' preferences. *Frontiers in human neuroscience*, *7*, 686.

32. Knutson, B., Westdorp, A., Kaiser, E., & Hommer, D. (2000). FMRI Visualization of

Brain Activity during a Monetary Incentive Delay Task. *NeuroImage*, *12*(1), 20–27. https://doi.org/10.1006/nimg.2000.0593

33. Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*(2), 75–84. https://doi.org/10.1016/j.evolhumbehav.2006.06.001

34. Lee, D., & Seo, H. (2016). Neural Basis of Strategic Decision Making. *Trends in Neurosciences*, *39*(1), 40–48. https://doi.org/10.1016/j.tins.2015.11.002

35. Leliveld, M. C., Dijk, E. van, & Beest, I. van. (2012). Punishing and compensating others at your own expense: The role of empathic concern on reactions to distributive injustice. *European Journal of Social Psychology*, *42*(2), 135–140. https://doi.org/10.1002/ejsp.872

36. McClure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2007). Time Discounting for Primary Rewards. *Journal of Neuroscience*, *27*(21), 5796–5804. https://doi.org/10.1523/JNEUROSCI.4246-06.2007

37. Morishima, Y., Schunk, D., Bruhin, A., Ruff, C. C., & Fehr, E. (2012). Linking Brain Structure and Activation in Temporoparietal Junction to Explain the Neurobiology of Human Altruism. *Neuron*, *75*(1), 73–79. https://doi.org/10.1016/j.neuron.2012.05.021

38. Murnighan, J. K., & Saxon, M. S. (1998). Ultimatum bargaining by children and adults. *Journal of Economic Psychology*, *19*(4), 415-445.

39. 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.

40. Rangel, A., & Clithero, J. A. (2014). Chapter 8 - The Computation of Stimulus Values in Simple Choice. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics (Second Edition)* (pp. 125–148). San Diego: Academic Press. https://doi.org/10.1016/B978-0-12-416008-8.00008-5

41. 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

42. Roy, M., Shohamy, D., & Wager, T. D. (2012). Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends in cognitive sciences*, *16*(3), 147-156.

43. Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*(8), 549.

44. Sanfey, A. G. (2009). Expectations and social decision-making: biasing effects of prior knowledge on Ultimatum responses. *Mind & Society*, *8*(1), 93-107.

45. Sanfey, A. G., Civai, C., & Vavra, P. (2015). Predicting the other in cooperative interactions. *Trends in cognitive sciences*, *19*(7), 364-365.

46. Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *42*, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009

47. Smith, D. V., Clithero, J. A., Boltuck, S. E., & Huettel, S. A. (2014). Functional connectivity with ventromedial prefrontal cortex reflects subjective value for social rewards. *Social Cognitive and Affective Neuroscience*, *9*(12), 2017–2025. https://doi.org/10.1093/scan/nsu005

48. Solnick, S. J. (2001). Gender differences in the ultimatum game. *Economic Inquiry*, *39*(2), 189-200.

49. Stallen, M., Rossi, F., Heijne, A., Smidts, A., Dreu, C. K. W. D., & Sanfey, A. G. (2018). Neurobiological Mechanisms of Responding to Injustice. *Journal of Neuroscience*, *38*(12), 2944–2954. https://doi.org/10.1523/JNEUROSCI.1242-17.2018

50. Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., & Kirsch, P. (2011). Beyond revenge: Neural and genetic bases of altruistic punishment. *NeuroImage*, *54*(1), 671–680. https://doi.org/10.1016/j.neuroimage.2010.07.051

51. Strombach, T., Weber, B., Hangebrauk, Z., Kenning, P., Karipidis, I. I., Tobler, P. N., &

Kalenscher, T. (2015). Social discounting involves modulation of neural value signals by temporoparietal junction. *Proceedings of the National Academy of Sciences*, *112*(5), 1619–1624. https://doi.org/10.1073/pnas.1414715112

52. Sul, S., Tobler, P. N., Hein, G., Leiberg, S., Jung, D., Fehr, E., & Kim, H. (2015). Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. *Proceedings of the National Academy of Sciences*, 201423895. https://doi.org/10.1073/pnas.1423895112

53. Sutter, M. (2007). Outcomes versus intentions: On the nature of fair behavior and its development with age. *Journal of Economic Psychology*, *28*(1), 69-78.

54. Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., … Nakahara, H. (2012). Learning to Simulate Others' Decisions. *Neuron*, *74*(6), 1125–1137. https://doi.org/10.1016/j.neuron.2012.04.030

55. Takagishi, Haruto, Shinya Kameshima, Joanna Schug, Michiko Koizumi, and Toshio Yamagishi. "Theory of mind enhances preference for fairness." *Journal of experimental child psychology* 105, no. 1-2 (2010): 130-137.

56. Tricomi, E., Rangel, A., Camerer, C. F., & O'Doherty, J. P. (2010). Neural evidence for inequality-averse social preferences. *Nature*, *463*(7284), 1089–1091. https://doi.org/10.1038/nature08785

57. Vavra, P., Chang, L. J., & Sanfey, A. G. (2018). Expectations in the Ultimatum Game: Distinct Effects of Mean and Variance of Expected Offers. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00992

58. Vavra, P., van Baar, J., & Sanfey, A. (2017). The Neural Basis of Fairness. In M. Li & D. P. Tracer (Eds.), *Interdisciplinary Perspectives on Fairness, Equity, and Justice* (pp. 9–31). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-58993-0_2

59. van Baar, J. M., Chang, L. J., & Sanfey, A. G. (In Press). The computational and neural substrates of moral strategies in social decision-making. Nature Communications

60. van den Bos, W., van Dijk, E., Westenberg, M., Rombouts, S. A., & Crone, E. A. (2009). What motivates repayment? Neural correlates of reciprocity in the Trust Game. *Social cognitive and affective neuroscience*, *4*(3), 294-304.

61. Zettler, I., & Hilbig, B. E. (2010). Attitudes of the selfless: Explaining political orientation with altruism. *Personality and Individual Differences*, *48*(3), 338–342. https://doi.org/10.1016/j.paid.2009.11.002

Box 1: Experimental Tasks

Adapted from Vavra, van Baar and Sanfey (2017); for a review of the use of Game Theory and its paradigms in neuroeconomics, see Civai & Hawes (2016).

Behavioural economics and game theory offer a wide range of structured paradigms that can easily be adapted to a laboratory-based exploration of decision-making, usually involving a multi-player structure where participants are asked to make decisions. Some of the games that have been used in the context of investigation of social norms perception and compliance, and that are referred to in the current chapter, are described below.

The Ultimatum Game (UG; Güth et al, 1982) is a game where two players decide, sequentially, how to split a sum of money. The first player (proposer) is given a sum of money, e.g., $10, and has to decide how to divide this money with the second player (responder). The proposer makes offers to the responder, who has to either accept or reject these offers: if they accept, the money is divided as the proposer decided; if they reject, none of the players gets anything. Importantly, the game is often anonymous and played as a single round (one-shot game), therefore there is no room for negotiation. Some players may be considered strictly egalitarian (proposers always offering half of the share; responders always rejecting less than half of the share) or strictly rational (proposers always offering the smallest unit; responders always accepting any offer larger than 0). However, experimental evidence has been consistently showing that proposers tend to offer a fair share, and responders reject unfair offers (Camerer, 2003). This highlights the role played by mentalizing: the proposer will offer the smallest amount they believe will be accepted by the responder; on the other hand, the responder will accept any offer that is deemed fair, considering the circumstances and the proposer's intentions (Falk, Fehr, & Fischbacher, 2003).

The Dictator Game (DG) is very similar to the UG; the only difference is that the responder is powerless and does not have the chance to reject the offer. As a consequence, the first player (dictator) decides the allocation of the monetary sum and their decision does not have any consequence within the game. The strategic motivation for being fair is now removed, and therefore any monetary transfer in this game can be considered as genuine

generosity; in this case, mentalizing and particularly empathic concerns may explain the altruistic behaviour (Edele, Dziobek, & Keller, 2013).

Third party Games are often adapted versions of the UG or DG, where one additional player plays the role of the observer. In these games, the observer is required to decide whether or not to react to an injustice, by spending their own resources, when their payoff had not been affected by the injustice. The observer may react by punishing the perpetrator (e.g., an unfair dictator; Fehr & Fischbacher, 2004; Strobel et al, 2011) or by compensating the victim (Stallen et al, 2018). Mentalizing and other-regarding concerns are involved in these decisions, in particular when choosing to compensate the victim (Leliveld et al, 2012; Civai et al, 2019).

The Trust Game (TG; Berg, Dickhaut & McCabe, 1995) is a two-player game widely employed to investigate trust and reciprocity. One player (investor), is endowed with a sum of money, and can decide how much to transfer to the second player (trustee). The rules of the game establish that any amount transferred is multiplied by a fixed factor, e.g., four. For example, if the investor transfers $5, the trustee would receive $20; at this point, the trustee decides how much of this amount to transfer back to the investor, if any. Because trustee can decide to transfer nothing back, the decision of the investor to transfer something can be interpreted as trust in the second player. On the other hand, the trustee's decision to return any amount is interpreted as reciprocity. Similar to the UG, the investor will use mentalizing abilities to predict the trustee's willingness to reciprocate; in turn, the trustee may use the ability to predict the investor's mental state to decide how much to transfer back: this is true in particular for the guilt-free trustees who reciprocate to match the investor's expectations, as explained in the main text (van Baar et al, 2019).

The Prisoner's dilemma, first formalised in the 1950s by Flood and Dresher, is a game in which two players must decide whether to cooperate with each other or defect, knowing that cooperation would lead to the maximum outcome for both players. The game can be played simultaneously or sequentially, eliciting a tit-for-tat strategy; as shown by neuroimaging studies, mentalizing is one of the core mechanisms to guide the decision (Rilling et al, 2004).

As previously mentioned, all these games are often anonymous and one-shot. In neuroimaging studies, since it is necessary to have multiple observations, the so-called single-shot multi-round games are employed: each participant plays the game multiple times, on each round paired with a new partner. Repeated paradigms, i.e., having participants interacting with the same player more than once, are employed when the focus of the investigation is learning process. As in Heijne & Sanfey (2015) for example, interacting with the same partner more than once allows participants to learn whether to stay or leave the relationship. In conclusion, Game Theory offered a set of structured and flexible paradigms well suited for investigating social interaction in a laboratory context.