King's College London

This paper is part of an examination of the College counting towards the award of a degree. Examinations are governed by the College Regulations under the authority of the Academic Board.

PG Cert/PG Dip/MSc Examination

7PADSPRI Research Skills: Synoptic Project I

Final word count: 3995
(word limit: 4000 words)

Please note that a 5% penalty will be applied when essays exceed the word limit. There is no additional 10% allowance in this submission. Markers will stop reading once they reach the word limit. This will have an impact on the grade awarded, so please respect the word limit.

Plagiarism Statement

This exam script was written by me using my own words. It does not use text from work I, or any other student/researcher, has already submitted for formative or summative assessment in another module, another course, publication or source anywhere in the world.

The university Academic Honesty and Integrity Policy can be found here:
http://www.kcl.ac.uk/governancezone/Assessment/Academic-Honesty-Integrity.aspx

Student ID Number 19071714 Date 21 February 2022

Student ID Number: 19071714

The role of the insula in processing fair and unfair distributions in the light of the reward-risk prediction model

Introduction

A sense of fairness has been found to emerge in early childhood (Fehr et al., 2008; Cowell et al., 2019) and observed across countries and cultures, although with some variations (Henrich et al., 2006; Blake et al., 2015; Huppert et al., 2019).

Behaviourally, fairness has been studied with experimental games such as the Ultimatum Game (UG) (Güth, et al., 1982). In this task, a Player 1 (the Proposer) is endowed with a certain amount of money that he must share with a Player 2 (the Responder). This latter has two options: accepting or rejecting the offer made. If the responder accepts the offer, the amount of money will be divided following the proposition of Player 1. In the case of a rejection, both receive nothing.

In this social dilemma, the *Homo Economicus* model (Mill, 1895) and utilitarian perspective would assume that people will be motivated by their self-interest to maximise their own profit (Von Neumann & Morgenstern, 1944). Therefore, Player 2 would never reject the offer and conversely, Player 1 would offer the smallest amount possible.

Nevertheless, behavioural studies showed that lower offers (often around 20% of the amount) have 50% chance to be rejected (Camerer, 2003; Sanfey, 2007) in line with the modern ideas of behavioural economics that human decisions are far less rational than predicted in the classical theory (Kahneman et al., 1986).

Rejection of an unfair offer has been said to be a pursuit of fairness and punishment over self-interest (Fehr & Gächter, 2002).

Over the last two decades, functional magnetic resonance imaging (fMRI) combined with economics games found that receiving an unfair offer was associated with the activation of the anterior insula (AI) along with the anterior cingulate cortex (ACC) and the ventromedial prefrontal cortex (vmPFC) (Sanfey et al., 2003). As the AI has been first found to be correlated with negative emotions (Craig, 2009), an early interpretation of these results has proposed that the AI might respond to the negative emotion elicited by unfair offers leading to their rejection (Sanfey, 2007).

Research in the non-social context of gambling games found that the insula is associated with the risk prediction (RiP) and risk prediction error (RiPE) in the reward-related decision-making processing (Preuschoff et al., 2008).
Later interpretations based on this model proposed that the same mechanism of error prediction could be extended to the social context of the UG to detect norm violation (Civai, 2013).
Nevertheless, there is debate as to whether the AI is actually responding to fairness *per se* (Fehr & Krajbich, 2014).

The aim of this critical review is to assess the literature on the role of the AI in encoding fairness and predicting error (Part 1). The research question is therefore to clarify the role played by the AI in fairness and reward/risk prediction in the context of experimental games with healthy controls.

Ethical considerations will be developed based on the studies mentioned (Part 2).

Research strategy – Searches on online platforms (PubMed, Ovid) were performed using these following terms: “fair*”, “Ultimatum Game”, “fMRI”, “risk prediction error”, “insula”, “Decision Making” [MeSH], “Reward” [MeSH], “risk” [MeSH] and by varying Boolean operators.

Although animal studies have provided evidence in the understanding of the mechanisms, string to exclude animal studies was included due to the time and length constraints of this project.

Part 1 - Literature review

1.1. The role of the insula in the UG and the early emotional model

The neural basis of the UG was first examined by Sanfey and colleagues (2003). This important study laid the foundations of the understanding of the neural correlates underlying the substantial rejection rates of unfair offers.

When undergoing a fMRI, the participants played a UG where offers were predetermined by an algorithm (with half of the trials were the result from an equal split on $10, and half of them, from an unequal split (two offers of $9:$1, two offers of $8:$2 and one of $7:$3)).

Consistent with behavioural studies, the investigators found a higher proportion of rejections when the offers deviated from the equal split.

Neuroimaging data showed a greater activation in the AI, DLPFC and ACC for unfair offers compared to fair offers. Greater activation of the AI was also found to be correlated with the degree of unfairness (greater activation for a $9:$1 offer compared to a $8:$2 offer). Participants with stronger activation of the AI have been recorded to reject a higher proportion of unfair offers.

The activation of the AI has been therefore suggested to not only be associated with fairness but also lead to the rejection of unfair offers.

As the insula has been said to be associated with negative emotional states, such as physical pain, disgust, and other subjective feelings (Craig, 2009), the early interpretation of these results associated the rejection of the unfair offers with the negative emotion arousal (Sanfey, 2007), supporting the theory that affective processing is an important element in decision making.

This is in line with the wounded pride/spite model found in previous behavioural studies on the UG which suggest that anger and frustrations lead to rejection of unfair offers (Pillutla & Murnighan, 1996) and the inequity aversion theory which assumes that individuals tend to dislike unfair distribution and are willing to take action to seek fair outcomes (Fehr & Schmidt, 1999).

A subsequent study where fairness was distinguished from the final financial outcome found similarly an increased activity in the insula to be correlated with the

tendency to reject unfair offers and that the reward pathway was involved in processing fair offers (Tabibnia et al., 2008).

However, an interesting study comparing a group of experienced mindfulness meditators to controls during an UG task, found meditators were willing to accept asymmetrical offers at a higher proportion than controls (Kirk et al., 2011). Moreover, a major difference has been noted when comparing the two groups regarding to the AI: no significant activation for fair and unfair offers was observed in meditators and there was no correlation found between AI activity and rejections. The then interpretation of these results suggested that experienced meditators were able to recruit a different neural network compared to controls, enabling them to avoid generating negative emotions linked to unfair offers and punishment response.

Taken together, these results show that if the insula was consistently found to be activated in response to unfairness in healthy controls, these results are not confirmed when working with other groups with different cognitive abilities. Moreover, a later meta-analysis noted no systematic correlation between the strength of the anterior insula activation and the rejection rate (Gabay et al., 2014). This could suggest another role of the insula beyond representing negative emotions in the UG (Civai, 2013).

1.2. The alternative prediction error paradigm

As the activation in the insula has also been interpreted to represent RiP and RiPE in decision-making, a complementary hypothesis is whether it could be encoding prediction error of the player's own payoff in the social context of fairness.

1. The RiP and RiPE mechanism in non-social contexts

Prediction errors have been studied in the context of decision-making in gamble games (Preuschoff et al., 2008). He proposed that the reward-based reinforcement learning framework (Sutton & Barto, 1998) should not only encode reward prediction error, but also a RiP and a RiPE (Preuschoff & Bossaerts, 2007). This model is based on the analogous mechanism of reward-processing which requires a reward prediction signal and a reward prediction error signal. Similarly, there would be a RiP (which is measured in terms of variance in decision theory) and a RiPE (which arises when the RiP is misjudged).

To test this paradigm, the participants were asked to play a card game where two cards were drawn consecutively. While undergoing a fMRI, the participants guessed whether the second card would be lower or higher than the first card; these predictions were recorded twice, one before the first card and one before the second card. This design allows risk to be distinct from expected reward to avoid confounding factors. Thus, after seeing the first card, the mathematical function of

the expected reward, which is the probability of reward, should describe a linear ascending line while the function associated with risk should create an inverted U-shape pattern (Preuschoff et al., 2006) showing the increase of the risk probability from 0 until 50% of chance of winning (high risk) and then the decrease of the risk until 0 at the end of the trial.

Results from the fMRI found two types of activations, with different timings (a rapid and a delayed activations), in the insula corresponding to the non-linear U-shaped pattern. This suggests that the insula is not encoding expected reward but is associated with the risk prediction model, one signal reflecting RiP and one reflecting RiPE. More in-depth analysis will be further described in the second part of the project.

Following this finding, some of the interpretations regarding the neuroimaging results of the UG can be reread in light of the concept of error prediction.

2. **Predicting error in social contexts**

Error detection in the UG – A study based on Montague and Lohrenz's model of norm violation (2007) shed light on the relationship between fairness as a social norm of interest and prediction error (Xiang, Lohrenz & Montague, 2013).
This model postulates that in the UG, the Responder will compare the offer made with a fairness norm and error signals would be generated correlated with deviations

from this norm. In that situation, two signals have been suggested to be at play: a norm prediction error, the deviation of the offer from the mean, that could be related to dopamine targeted areas (O'Doherty et al., 2003) and the variance prediction error equivalent to the deviation of the square of the prediction error from the estimated variance which could be associated with RiPE involving the AI (Preuschoff et al., 2008).

Using a norm-training mathematical model, the norm was shifted to vary the participants expectations in the study. Given that unfair offers were usually associated with lower offers, the researchers preadapted the conditions to either high or low offer so that it would be possible to study the difference in rejection rates for identical offers.

Results show that activity in the ventral striatum and vmPFC was correlated with norm prediction error and activity in the AI was associated with the norm variance prediction error. Moreover, the particular U-shape response of the AI to the norm prediction error has also been observed, similarly to the pattern found by Preuschoff (2008).

It has been concluded that the regions involved in detecting norm violations in social context are the same as the ones identified in reward and risk prediction non-social decision making.

Towards an integrative model – Together with the role of the insula in representing emotional arousal, these findings could support an integrative model of a common role of the insula in feelings, empathy and uncertainty proposed by Singer and

colleagues (2009). In this model, the insula integrates bodily, affective and sensory information with uncertainty to improve learning and guide decision making. Thus, the involvement of the AI in tracking risk is not necessarily in contradiction with its role in processing feelings (Bossaerts, 2010) in line with the *somatic marker hypothesis* (Bechara & Damasio, 2005) which proposes that rational decisions are influenced by emotional responses. This is consistent with the structure and the position of the insula located between the frontal and temporal lobes which projects to different structures implicated in decision making. In this context, it has been suggested that the insula may respond to the emotional arousal created by deviation from the social norm (Xiang et al. 2013).

However, this view is not as straightforward as it seems to be. An important interrogation is as to whether these results from the UG could be encoding fairness itself and not the prediction of the responder's own payoff.

3. Results from studies decorrelating fairness from participant's self-interest

In all the studies mentioned above, the person who can accept or reject the offer made by the proposer is the same as the one to whom the fair or unfair treatment is targeted. This is an important confounding factor inherent to the classic version of the UG. To test for the hypothesis as to whether the activation of the insula is not simply a negative prediction error for the participant's own payoff, there is a need to look at games where fairness and self-interest are uncorrelated (Ruff & Fehr, 2014).

Using a modified version of the UG where the participants would play, in one condition for themselves with the same rules as the classical UG and in another condition on the behalf of other players (Civai et al., 2010), Corradi-Dell'Acqua and colleagues (2013) attempted to decorrelate self versus third party fairness and found that the AI was involved in both myself and third-party condition.

These results suggest that the AI reacts not only to self-directed unfairness but also to unfairness affecting an unknown other person. This has been interpreted to be in line with the above cited work suggesting that AI may encode deviations from expectations of an equal distribution (Civai, 2013).

To further examine the neural representation of fairness for the self compared to fairness for others, a recent study implemented another variation of the UG where participants' self-interest could directly compete with fairness for others (Yoder & Decety, 2020).

The experimental task involved three parties, the Proposer, a neutral observer and the Responder. Participants to the study played the role of the Responder who could accept or reject the distribution made by the Proposer of $12 between the three parties (as in the classic UG). Four types of offers were used: SelfFair-OtherFair (4:4:4), SelfFair-OtherUnfair (6:1:5), SelfUnfair-OtherFair (6:5:1), and SelfUnfair-OtherUnfair(10:1:1), allowing self-interest to conflict with other's fairness in some conditions.

Neuroimaging results show that activity in the AI was detected only in Self-Fairness conditions in response to unfair offers and not in Other-Fairness conditions.

Moreover, no overlap was found in the neural basis of processing fairness when comparing the self to other fairness situations.

These results suggest therefore that fairness for the self and fairness for others are computed differently and that the insula is seen to be only involved in self-serving fairness to signal unfair offers. This absence of the activation of the AI in the OtherFairness conditions has been interpreted by the authors in consideration of the role of the insula in the salience network (Seeley et al., 2007); when self-interest competes with third-party fairness, one's own payoff would become more salient, diminishing the AI response of the third-party fairness.

An additional interpretation can be suggested based on the involvement of the insula in the risk prediction model in the reward-based decision-making mechanism as these results could provide evidence that the insula might be activated not in response to violation to fairness *per se* but to signal error when the player's own payoff is involved. This can be viewed in light of recent work on risk prediction model which proposes that RiPE could indicate surprise and violations of expectations (Lauffs et al., 2020; Loued-Khenissi et al., 2020). Together these results could suggest that the insula might play a role in signaling a mismatch associated with the expectations of the player on its own payoff.

Nevertheless, it has not been explicitly examined and new experimental designs could integrate different gambling tasks, adding explicit measures (of RiP and RiPE especially) as well as comparisons between groups to manipulate the involvement of self-interest to clarify the role of the insula in the relationship between fairness and the reward-risk processing mechanism.

Further analysis will be carried out in the second part of the project to compare and contrast the measures found in these studies.

Part 2 - Ethical considerations

2.1. General guidelines for behavioural research with healthy participants

Neuroeconomics research described above involves healthy controls. General ethical guidelines require that participants have been given their valid consent (2.1.1.), a principle that might be threatened by the use of deceptive practices (2.1.2.).

2.1.1. Valid consent

Obtaining a valid consent is at the center of ethical considerations when conducting research involving human participants. The British Psychological Society (BPS) Code of Human Research Ethics (2021) states that participants from whom data are gathered have been given their free consent to take part to the research after receiving sufficient information. Valid consent is therefore based on prior adequate information which entails giving the necessary information (i.e. aim(s) of the project, the data collection process, the risks associated, confidentiality) in a clear and accessible manner.

An essential point about valid consent is the ability for the participants to give their assent freely, without any coercion and to withdraw their consent at any moment. This rule requires therefore to take cautions to not undermine the freedom of the participants when using payments for example, as incentives play an interesting role in experimental games considering its salience (Shram & Ule, 2019).

2.1.2. Deception

Traditionally, deception has been excluded in experimental economics (Shram & Ule, 2019). Nevertheless, ethics committees allow situations where participants give their informed consent while researchers make use of deception. The potential risks and harm caused by deception are then to be weighted with the research contributions and the absolute necessity of using deception.

Deception can be valuable in cases where disclosing the truth could have an impact on the participants' behaviours (e.g. hiding the real purpose of the study). Nevertheless, deception should not be used if participants would have refused to take part in the study should they have been informed of its true aim.
Other situations extend deception to delivering purposely false information (e.g. studies that used computerised offers where participants were told that they will be playing with a human proposer whereas they responded to predetermined offers (Corradi-Del-Acqua, 2013; Xiang, 2013)).

In any case, researchers should provide justification for the absence of alternatives to the use of deception and verify that participants would not exhibit any discomfort when the use of deception will be explained in the debriefing to mitigate potential harm (which is required for any research but all the more important in presence of deceptive practices).

2.2. Guidelines applied to neuroimaging studies

Most of the studies mentioned used neuroimaging techniques, and especially fMRI for data collection. While fMRI is considered to be a safe procedure, it requires additional ethical considerations relative to the contraindications inherent to this technology (2.2.1.), the important issue of confidentiality (2.2.2.) as well as the question of incidental finding and giving advice (2.2.3.).

2.2.1. Contraindications, information and screenings

As MRI involved electromagnetic field, undergoing the scan with certain conditions can be harmful for the participants. A proper health prescreening shall be done to determine whether the history of participant's health could be incompatible with the use of fMRI such as, without being exhaustive, the presence of a pacemaker, brain operations with insertion of metal, or any metal implant as well as history of epilepsy, or risk of pregnancy, as clearly mentioned in one of the studies cited (Yoder & Decety, 2020).

As with any research, information and informed consent is essential with the emphasis of potential risks of taking part in fMRI study. Mention of the confined and noisy environment of the scanner should also be included in the information sheet.

2.2.2. Confidentiality

While the principle of confidentiality is a general guideline applicable for all research projects, it takes on a certain importance when considering research using neuroimaging techniques. In addition to the legislation on data protection, participants must be assured that the information they provided along with the data collected will be treated confidentially. They will also need to be assured that the data will not be used outside of the study purpose.

2.2.3. Incidental finding and giving advice

The question as to whether findings of potential significance for participants' health should be disclosed has raised an important discussion, becoming even more relevant as the number of neuroimaging research increases.

On one hand, it can be said that disclosure of anomalies discovered during imaging can be beneficial and life-saving in certain cases for participants (Ross, 2005). On the

other hand, such disclosure might lead to more challenges without necessarily benefit for the participant (e.g. anxiety, long-term follow-ups, harmful medical procedures, financial costs etc…) (Booth, 2010).

No standardised rules about incidental findings exist; guidelines vary depending on the institutions and no general obligation to disclose incidental findings to the participants exist (The Royal College of Radiologists, 2013).
In case of disclosure, it has been suggested to be best performed by a medical practioner (Booth, 2010). At the same time, it has been recognised that this practice can have serious practical implications for the researchers (The Royal College of Radiologists, 2013).

In any case, it is of utmost importance for such questions to be handled beforehand. The course of action should be anticipated, and the investigators shall have planned and established a process to handle and report such findings before the start of the study.

Conclusion

Understanding how fairness is processed in the brain has made significant advances thanks to the emergence of the interdisciplinary field of neuroeconomics. Drawing models from economics, neuroscience and psychology, the studies mentioned use a unique combination of economics models, experimental games, and neuroimaging

techniques to try to decipher the neural representation of fairness. This intertwining relationship between these various fields has implications in terms of ethics as these research must not only follow the general rules laid down for human research (e.g. valid consent, confidentiality) but also the conventions of certain disciplines as well as the specificities of particular techniques (i.e. neuroimaging) to minimise the risk of harm and respect the general rules of conduct when doing research.

The literature search on fairness led to studies that consistently show an activation of the insula among the general population in response to unfair offers in the context of the classical design of the UG (Sanfey et al. 2003). Given the role of the insula in processing emotions, early interpretation posits that it was responding to the negative emotions created by these unfair offers. A complementary interpretation based on the involvement of the insula in RiP and RiPE found in the non-social decision-making process, suggests that the activation of the insula could be corresponding to detecting norm violation which used the same error detection mechanism in the social context of the UG (Xiang et al. 2013). An integrative model proposed therefore that the insula may respond to the emotional arousal created by deviation from the social norm.

Nevertheless, these studies do not necessarily confirm that the insula is responding to fairness. Thus, a study comparing healthy controls with meditators found no such activation of the insula in the latter group when playing the UG (Kirk et al., 2011). Moreover, a study where fairness and self-interest were decorrelated found a differentiated representation of fairness for self and other and that the insula was

only activated to signal unfair offers in the Self-Fairness conditions (Yoder and Decety, 2020). This could suggest that the greater activation of the insula found in rejections of unfair offers could not be responding to fairness *per se* but could be related to the RiP and RiPE processing of surprise and violation of expectations.

Further research is needed to clarify the role of the insula in experimental games involving fair/unfair distributions and will have implications not only for the general knowledge of our behaviours but also for clinical populations (as the insula has been found to be acting differently in borderline personality for example (King-Casas et al., 2008), as well as for the field of artificial intelligence which relies on neuroscience (Hassabis et al., 2017) and game theory (Conitzer et al., 2017) to model some of its algorithms.

The subsequent project will look in more detail at the analyses done in the studies to further compare and contrast them as well as indicate some limitations related to the methodology.

Reference list

Litterature review

Bechara, A., & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and economic behavior*, *52*(2), 336-372.

Blake, P. R., McAuliffe, K., Corbit, J., Callaghan, T. C., Barry, O., Bowie, A., ... & Warneken, F. (2015). The ontogeny of fairness in seven societies. *Nature*, *528*(7581), 258-261.

Bossaerts, P. (2010). Risk and risk prediction error signals in anterior insula. *Brain Structure and Function*, *214*(5), 645-653.

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction.* Russell Sage Foundation.

Camerer, C., Loewenstein, G., & Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of economic Literature*, *43*(1), 9-64.

Civai, C., Corradi-Dell'Acqua, C., Gamer, M., & Rumiati, R. I. (2010). Are irrational reactions to unfairness truly emotionally-driven? Dissociated behavioural and emotional responses in the Ultimatum Game task. *Cognition*, *114*(1), 89-95.

Civai, C. (2013). Rejecting unfairness: emotion-driven reaction or cognitive heuristic?. *Frontiers in human Neuroscience*, *7*, 126.

Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. (2017, February). Moral decision making frameworks for artificial intelligence. In *Thirty-first aaai conference on artificial intelligence*.

Corradi-Dell'Acqua, C., Civai, C., Rumiati, R. I., & Fink, G. R. (2013). Disentangling self-and fairness-related neural mechanisms involved in the ultimatum game: an fMRI study. *Social cognitive and affective neuroscience*, *8*(4), 424-431.

Cowell, J. M., Sommerville, J. A., & Decety, J. (2019). That's not fair: Children's neural computations of fairness and their impact on resource allocation behaviors and judgments. *Developmental psychology*, *55*(11), 2299.

Craig, A. D., & Craig, A. D. (2009). How do you feel--now? The anterior insula and human awareness. *Nature reviews neuroscience*, *10*(1).

d'Acremont, M., Lu, Z. L., Li, X., Van der Linden, M., & Bechara, A. (2009). Neural correlates of risk prediction error during reinforcement learning in humans. *Neuroimage*, *47*(4), 1929-1939.

Fehr, E., & Krajbich, I. (2014). Social preferences and the brain. In *Neuroeconomics* (pp. 193-218). Academic Press.

Fehr, E., Bernhard, H., & Rockenbach, B. (2008). Egalitarianism in young children. *Nature*, *454*(7208), 1079-1083.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137-140.

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

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.

Gabay, A. S., Radua, J., Kempton, M. J., & Mehta, M. A. (2014). The Ultimatum Game and the brain: A meta-analysis of neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, *47*, 549-558.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of economic behavior & organization*, *3*(4), 367-388.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245-258.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., ... & Ziker, J. (2006). Costly punishment across human societies. *Science*, *312*(5781), 1767-1770.

Huppert, E., Cowell, J. M., Cheng, Y., Contreras-Ibáñez, C., Gomez-Sicard, N., Gonzalez-Gadea, M. L., ... & Decety, J. (2019). The development of children's preferences for equality and equity across 13 individualistic and collectivist cultures. *Developmental science*, *22*(2), e12729.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the assumptions of economics. *Journal of business*, S285-S300.

Kirk, U., Downar, J., & Montague, P. R. (2011). Interoception drives increased rational decision-making in meditators playing the ultimatum game. *Frontiers in neuroscience*, *5*, 49.

Kishida, K. T., King-Casas, B., & Montague, P. R. (2010). Neuroeconomic approaches to mental disorders. *Neuron*, *67*(4), 543-554.

Lauffs, M. M., Geoghan, S. A., Favrod, O., Herzog, M. H., & Preuschoff, K. (2020). Risk prediction error signaling: A two-component response?. *NeuroImage*, *214*, 116766.

Loued-Khenissi, L., Pfeuffer, A., Einhäuser, W., & Preuschoff, K. (2020). Anterior insula reflects surprise in value-based decision-making and perception. *NeuroImage*, *210*, 116549.

Mill, J. S. (1895). Utilitarianism. Longmans, Green and Company

Montague, P. R., & Lohrenz, T. (2007). To detect and correct: norm violations and their enforcement. *Neuron*, *56*(1), 14-18.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences*, *16*(1), 72-80.

Pillutla, M. M., & Murnighan, J. K. (1996). Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational behavior and human decision processes*, *68*(3), 208-224.

Preuschoff, K., Bossaerts, P., & Quartz, S. R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron*, *51*(3), 381-390.

Preuschoff, K., & Bossaerts, P. (2007). Adding prediction risk to the theory of reward learning. *Annals of the New York Academy of Sciences*, *1104*(1), 135-146.

Preuschoff, K., Quartz, S. R., & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience*, *28*(11), 2745-2752.

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*(8), 549-562.

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.

Sanfey, A. G. (2007). Social decision-making: insights from game theory and neuroscience. *Science*, *318*(5850), 598-602.

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., ... & Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, *27*(9), 2349-2356.

Singer, T., Critchley, H. D., & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in cognitive sciences*, *13*(8), 334-340.

Tabibnia, G., Satpute, A. B., & Lieberman, M. D. (2008). The sunny side of fairness: preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychological science*, *19*(4), 339-347.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton university press.

Wang, Y., Zheng, D., Chen, J., Rao, L. L., Li, S., & Zhou, Y. (2019). Born for fairness: evidence of genetic contribution to a neural basis of fairness intuition. *Social cognitive and affective neuroscience*, *14*(5), 539-548.

Wu, Y., Zang, Y., Yuan, B., & Tian, X. (2015). Neural correlates of decision making after unfair treatment. *Frontiers in Human Neuroscience*, *9*, 123.

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *Journal of Neuroscience*, *33*(3), 1099-1108.

Yoder, K. J., & Decety, J. (2020). Me first: Neural representations of fairness during three-party interactions. *Neuropsychologia*, *147*, 107576.

Ethical considerations

Booth, T. C., Jackson, A., Wardlaw, J. M., Taylor, S. A., & Waldman, A. D. (2010). Incidental findings found in “healthy” volunteers during imaging performed for research: current legal and ethical implications. *The British journal of radiology*, *83*(990), 456-465.

British Psychological Society. (2021). *BPS Code of Ethics and Conduct*.

British Psychological Society. (2021). *BPS Code of Human Research Ethics*.

Ross K. (2005). When volunteers are not healthy. *EMBO reports*, *6*(12), 1116–1119.

Schram, A., & Ule, A. (Eds.). (2019). *Handbook of research methods and applications in experimental economics*. Edward Elgar Publishing Limited.

The Royal College of Radiologists. (2013) Management of incidental findings found during research imaging.