On (Ir)rationality

Kamron Soldozy, Jovana Kondic

1. Introduction

Despite the ongoing research of sociologists, economists, psychologists, philosophers, and biologists, why and how humans are irrational remains poorly understood. Indeed, scholars have historically disagreed even on the meaning of the term: although the psychologist Albert Ellis (1975) defined irrationality as “any thought, emotion, or behavior that leads to self-defeating or self-destructive consequences”, researchers have failed to agree on what constitutes irrationality, as opposed to, for example, a “cognitive illusion” (see Cohen, 1981). Additionally, relatively modern research reveals that apparently irrational behaviors may be grounded in evolutionarily optimal neurobiological processes (see, for example, Tsetsos et al., 2016).

In the first section of this report, we first briefly describe a sample of behaviors widely labeled as irrational by psychologists, neuroscientists, and economists. Drawing especially on the Reinforcement Learning (RL) literature, we also include computational models explaining these behaviors and speak to their neural plausibility. In the second half, we consider the practicalities of being “irrational” humans: are our emotions fundamentally at-odd with our aspirations for rationality, or is irrationality not nearly as “self-destructive” as Ellis pointed it out to be? What are emotions and how do they arise? Should we wish to discard our own moods and emotions, or can we leverage these affective states to do more good than harm?

2. Drivers Of (Ir)rational Decision-Making

2.1. The Behavior

Are you acting rationally? It depends on who you ask. Epistemic rationality, as defined by logicians, is reflected by skepticism toward unfounded belief. Behavioral game theorists, on the other hand, label decision-making as rational so long as it maximizes one's expected utility. Within economists, definitions span from instrumental rationality - taking the means necessary for achieving one's ends - to the axiomatic approach - being logically consistent within one's preferences and beliefs. Given such diverse definitions, an attempt to standardize human fallibility included the development of various simple experiments to evaluate one's reasoning.

As revealed by the popular false-positive paradox, **base rate fallacy** appears to fog our analytical reason in the general low prevalence - high true positivity rate scenarios, yielding unexpectedly frequent false positives that often defy our intuition. Furthermore, as studied by Kahneman and Tversky (1974), we often evaluate the occurrence of a single event to be less likely than its joint occurrence with another event, in numerous instances of **conjunction fallacy**. Led by **availability bias**, we violate laws of probability and bet on our favorite player losing the first set but winning the match, although just losing the first set is generally more likely.

We introduced the class to the **framing effect** bias by polling half of the participants in a reward-oriented manner, and polling the other half using a loss-oriented approach. It is interesting to note that ⅔ of the class-wide response yielded both risk-avoidance for positive framing and risk-proneness for negative framing, in accordance with the expected results obtained from the general population.

In summary, while standardized tests based on above-mentioned fallacy examples yield compelling results, they are often argued to be an insufficient methodology for determining irrational behavior. Some psychologists suggest that humans are rational in principle but err in practice, and others are proponents of breaking away from the use of just standard rules of logic, probability theory, or rational choice theory as norms of good reasoning (see Gigerenzer, 2001).

2.2. Computational Models and Neurological Implementation

A variety of models and heuristics have been developed to explain irrational behaviors, including those mentioned in section 2.1. Consider, for example, the principle of loss aversion: the observation that many people may prefer to avoid loss more strongly than they would acquire an equivalent gain. In a task where a participant may opt to receive $20 every trial (**option A**), or 0% half of the time and $40 the other half (**option B**), a majority - but not all - participants opt for the former, safer option. What algorithms explain this deviation from the anticipated 50/50 split a simple valuation would predict?

An initial explanation - **subjective utility** - draws on practical experience: perhaps the perceived value of $40 is less than twice as high as $20. For example, if a college student were to **need** $20 to purchase a new book, they would benefit decreasingly from excess quantities of money.

Two similar algorithmic explanations draw on temporal difference (TD) learning. The first argues that TD learning, as is, can account for this phenomenon. Imagine repetitively choosing between options A and B for multiple trials. You are not told the values of options A and B before the experiment and must learn them for yourself. You explore both options, and after a few trials, have learned that option A has a value of $20, and have received $40, 0$, and $40 from option B (a valuation of $26.67). You choose option B twice more, getting $0 both times, and subsequently never choose option B again: why do so when you perceive it to be valued $16, a smaller value than that of option A?

This explanation, however, is flawed. It does not explain why the majority of subjects exhibit risk averse behavior. Further, it is contingent on a policy that does not adequately explore to eventually perceive the value of option B. An alternative account is **risk-sensitive TD-Learning**, in which there exist two different learning rates for positive and negative prediction errors. This explanation intrinsically implements subjective utility and is neurologically plausible, given that Niv et al. (2012) have demonstrated it to better explain fMRI data in a similar task than an untouched TD learning model.

In favor of alternatives to valuation, there also exist various heuristics (as well as non-heuristic processes) that can be used to accurately reflect and explain our decision-making. One example is the **priority heuristic**. Relatively formulaic, the heuristic argues that people iteratively compare outcomes across multiple “dimensions”. For example, one might compare the minimum gain of $20 in option A to $0 in option B, and determine if the difference between the two ($20) is greater than 10% of the maximum possible gain ($4). Since the answer is yes, people would choose the option with the best minimum gain (option A). If this weren’t the case, a similar logic would be applied comparing the probabilities for minimum gain, and then comparing the maximum gains. An obvious criticism is that the comparison threshold of 10% is arbitrary. Additionally, neurological evidence for this algorithm is nonexistent. Nonetheless, it aptly explains human behavior in complex decision-making situations.

Traditionally, our actions, motivations, and preferences are all ultimately driven by perceived associated value. In this account, assigning a universal scalar quantity to options at hand is the first necessary step towards meaningful comparison. While most of the discussion on rationality centers around the algorithmic side of learning and value maximization, the very **existence of value** is often unquestioned and assumed for granted. As Hayden and Niv (2020), suggest, just because we associate values with options in our everyday decision-making, it does not simply follow that our brain actually does the same.

fMRI data collected from nucleus accumbens have been shown to highly correlate with value predictions of our computational models. However, the alignment between predicted values and neural recordings does not dispute the hypothesis that these recorded signals represent attention, or, perhaps, plans, or preferences.

Contrary to popular assumption, values don't just sit in our brain but rather involve an active process that takes place as decisions are being made. While values can change, so can our conditional preferences, which makes them practically impossible to decouple from one another.

Ultimately, however, in debating the existence of values, one could imagine that some combination of value-driven decision-making and heuristic driven decision-making exists: it is undeniable that certain paradigms are dramatically more conducive to value-based decisions than heuristical ones, and vice versa. Until the mechanisms for implementing heuristic-driven methods are better understood, however, little progress can be made in reconciling these two options.

3. Do You Want To Be Right Or Happy?

3.1. The Behavior

"Get a grip of yourself!"
This idiom speaks to the perception of emotions and mood as a general inconvenience, perhaps even a source of irrational behaviors. In direct contrast to this sentiment, evolutionary psychologists have argued that they might instead be adaptive tools for optimizing decision-making. For example, whereas humans have a tendency to depreciate the value of distant rewards (also known as **delay discounting**, da Matta et al., 2012), emotions like determination and motivation may incentivize behavior devaluing immediate rewards (Forgas, 2012). Similarly, love and guilt may assist in remaining loyal to a loved one. In this framework, love, guilt, determination, and other emotions are not intrinsically irrational, but rather correspond to different states under which specific behaviors and decision-making processes are made more favorable. These states may be thought of as irrational when they are undesirably present or more or less intensive than is desired by their human host.

Comparably, while optimal decision-making may necessitate optimal learning in order to maximize the expected utility, certain emotions such as happiness, in fact, necessitate the opposite. It has been shown that happiness does not depend on the cumulative earnings themselves, but rather on whether the cumulative utility is greater than expected. In particular, contrary to the goal of successful predictive learning, it was found that self-reported happiness directly depends on the presence of positive reward prediction errors.

As a matter of fact, in the classroom, we set about documenting the effect of certain emotions on decision-making. Students in the class were tasked with completing the Ultimatum Game. Here is a brief description:

*You are the **proposer**. You have been given $100. You are tasked with splitting your money with a stranger, the **responder**. If the responder accepts the split that you propose, you both keep the money after the game ends. If the responder does not accept no one keeps the money.*

The question: how much money do you decide to offer the responder?

Splitting them into two groups, half of the participants had 5 seconds to respond to the question, whereas the other half had 30 seconds. After completing the task, participants were asked about their current moods. In our analysis, we compared the quantity of money offered to the responder between the two groups, finding that the amounts were statistically insignificantly different (Figure 1a). The same was true when comparing the amount of money offered by proposers reporting positive versus negative moods in the 30 second group (Figure 1b). Importantly, however, participants reporting positive (and neutral) emotions offered less money to the responder than participants reporting negative emotions in the 5 second group (Figure 1c). Albeit with small sample sizes, our findings suggest that emotions can situationally affect decision-making, like when respondents have little time to make a decision. It remains unclear, however, why respondents in the positive-mood group may have chosen to offer less money in the 5 second condition: perhaps they were more optimistic about their prospects.

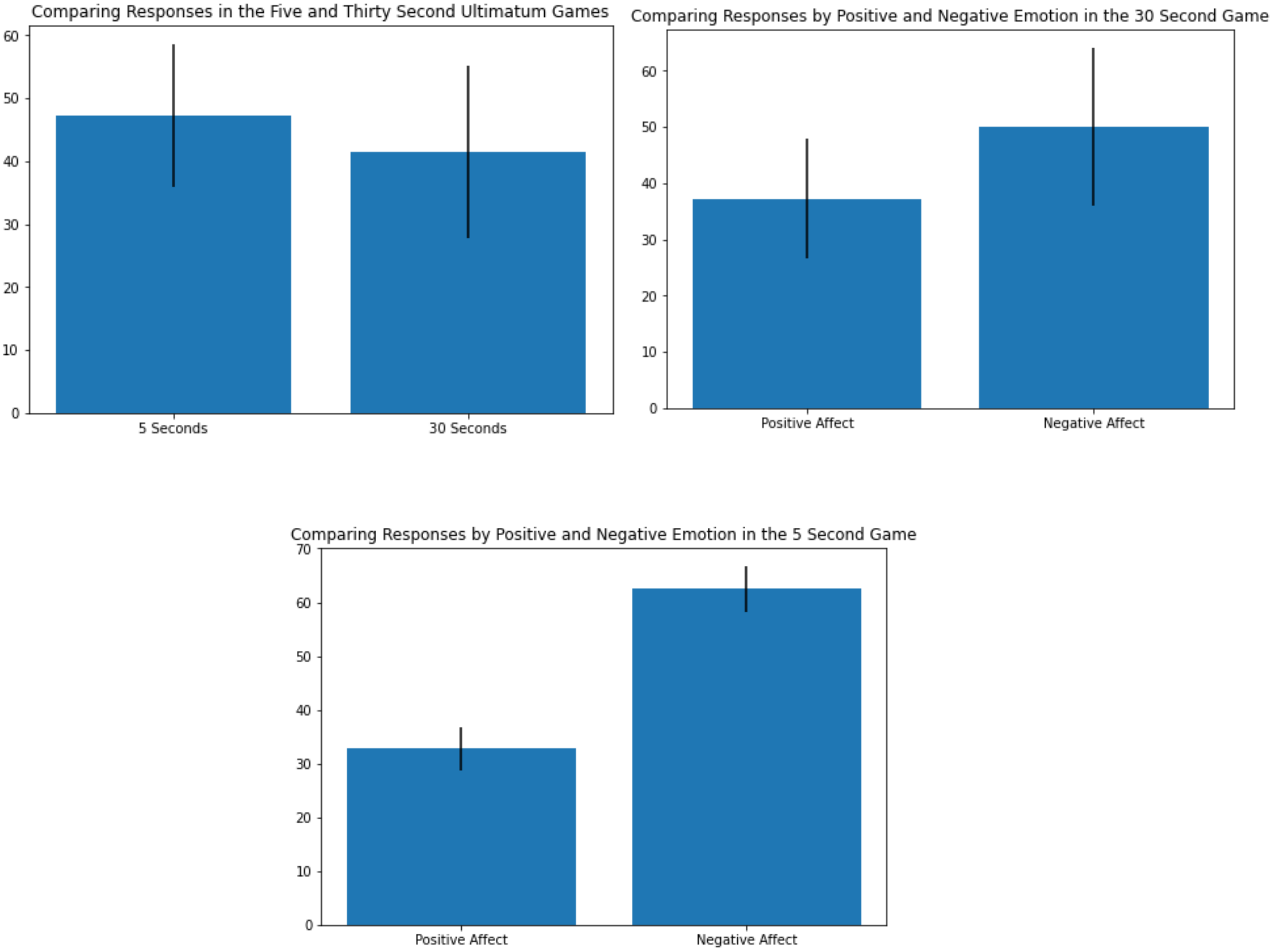

Figure 1. Comparing Amounts of Money Offered to a Theoretical Respondent Across various Categories. **(A-C)**. Panels A-C are organized from left to right and top to bottom. The amount of money offered is compared between the 5 and 30 second games, the positive and negative emotional categories in the 30 second group, and the positive and negative emotional categories in the 5 second group, respectively. Black bars indicate double standard error bars, and the y-axis is the amount of money proposed by the participants (USD).

3.2. Computational Models and Neurological Implementation

Thus, in contrast to the lay opinions that our moods and emotions are irrational remnants of our evolutionary ancestors, neuroscientists are aware of the importance of emotions and aim to better characterize the influence of mood on decision-making. Thus far, it is clear that positive mood induces risk-taking, depressed mood increases attention to negative information, and current perceived mode of thought is biased by the mood category. To explain the evident correlations, computational models of mood dynamics suggest that mood can be used to approximate average reward value, as well as its **momentum** - the accelerator of learning (see Eldar et al., 2015).

Mood can be particularly useful for learning about an environment, as opposed to an individual state. In a common scenario in which current changes in reward predict its later changes, a positive mood, for example, as a result of inference of a positive momentum, biases the perception of subsequent rewards upwards, thus updating expectations accordingly to catch up the agents perceptions with the rising rewards.

While neuroeconomics has led to fundamental changes in the understanding of how humans make decisions, many important behavior-influencing motives are not included in the strategic analysis. Given the recent formalizations of the mood-action dependency, it is critical that the reasoning about optimal (equilibrium) solutions, accordingly, considers parameters beyond the standard beliefs and utilities.

In a similar manner, Tamarit et al. (2016) adapted the standard Ultimatum game to reflect the findings of Kahneman on the effects of cognitive (System 2) and emotional (System 1) impulses on decision-making. In contrast with standard models, this utility function includes emotions as characterized by a psychological model. They suggest that, if no emotion is triggered during the game, the judgement is entirely determined by System 2. In the case when emotions are triggered, a purely rational decision can be overcome as a result of the extent of its influence on the utility function (that is characteristic of each individual).

Of course, how the brain even generates emotions has been a long standing question in neuroscience and psychology. Two directly competing theories posit that distinct brain regions correspond to distinct emotions (**locationist theory**) or that emotional categories are encoded by functional networks commonly employed across various emotions (**psychological constructionist approach**) (Lindquist et al., 2012). In favor of the psychological constructionist theory, Raz et al. (2016) found that the intensity of various emotions experienced as subjects viewed movie scenes was positively associated with the functional connectivity strength between two existing networks: the ventrolateral amygdala network (within the default mode network, or DMN) and the dorsal salience network. The former region is known to be recruited during emotional experiences, and the latter is thought to be responsible for detecting and filtering stimuli as well as recruiting other functional networks.

Although the authors don't speak to the practical relevance of emotions given their findings, one might imagine that the intensity of emotions could modulate key cognitive functions underlying decision-making. Indeed, Pessoa (2017) makes a similar claim, arguing that "Emotions … mobilize brain responses", which likely occurs through the activity of various functional networks. Pessoa explains that, for example, the close connection between the amygdala (known as the "danger detector" or "information gathering system" and the hypothalamus enables the control of neuroendocrine signaling, and that the connections between the amygdala and ventral striatum enable the emotional modulation of reward-related behavior.

4. Conclusion

Surprisingly, then, a cursory overview of studies exploring seemingly *ir*rational human behaviors reveal that many of the most fundamental human idiosyncrasies - like emotions - remain poorly understood. Although the relatively recent dawn of the neurosciences came with promises of groundbreaking insights in fundamental human behaviors and emotions, progress remains minimal and the error-learning theories stemming from Rescorla-Wagner remain among some of the most capable methods for modeling human behavior. Although it is difficult to judge, it seems that a critical impeding assumption to the development of this field is finally losing its prominence: indeed, the belief that nonoptimal behaviors are simply a result of computational inefficiencies of the human brain, rather than corresponding to dedicated, unique, and meaningful (both in the personal and evolutionary sense) brain mechanisms, is beginning to fade.

5. References

Bullmore, E., & Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3), 186–198. https://doi.org/10.1038/nrn2575

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *THE BEHAVIORAL AND BRAIN SCIENCES*, 54.

da Matta, A., Gonçalves, F. L., & Bizarro, L. (2012). Delay discounting: Concepts and measures. *Psychology & Neuroscience*, 5(2), 135. https://doi.org/10.3922/j.psns.2012.2.03

Eldar, E., Rutledge, R. B., Dolan, R. J., & Niv, Y. (2016). Mood as Representation of Momentum. *Trends in Cognitive Sciences*, 20(1), 15–24. https://doi.org/10.1016/j.tics.2015.07.010

Forgas, J. P. (2012). *Affect in Social Thinking and Behavior*. Psychology Press.

Gigerenzer, G. (2001). Decision Making: Nonrational Theories. *International Encyclopedia of the Social and Behavioral Sciences*, 5. https://doi.org/10.1016/B0-08-043076-7/01612-0

Hayden, B., & Niv, Y. (2020). *The case against economic values in the brain*. PsyArXiv. https://doi.org/10.31234/osf.io/7hgup

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. *The Behavioral and Brain Sciences*, 35(3), 121–143. https://doi.org/10.1017/S0140525X11000446

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. *Journal of Neuroscience*, 32(2), 551–562. https://doi.org/10.1523/JNEUROSCI.5498-10.2012

Pessoa, L. (2017). A Network Model of the Emotional Brain. *Trends in Cognitive Sciences*, *21*(5), 357–371. https://doi.org/10.1016/j.tics.2017.03.002

Raz, G., Touroutoglou, A., Wilson-Mendenhall, C., Gilam, G., Lin, T., Gonen, T., Jacob, Y., Atzil, S., Admon, R., Bleich-Cohen, M., Maron-Katz, A., Hendler, T., & Barrett, L. F. (2016). Functional connectivity dynamics during film viewing reveal common networks for different emotional experiences. *Cognitive, Affective, & Behavioral Neuroscience*, *16*(4), 709–723. https://doi.org/10.3758/s13415-016-0425-4

Ståhl, T., & van Prooijen, J.-W. (2018). Epistemic rationality: Skepticism toward unfounded beliefs requires sufficient cognitive ability and motivation to be rational. *Personality and Individual Differences*, *122*, 155–163. https://doi.org/10.1016/j.paid.2017.10.026

Tamarit, I., & Sánchez, A. (2016). Emotions and Strategic Behaviour: The Case of the Ultimatum Game. *PLoS ONE*, *11*(7). https://doi.org/10.1371/journal.pone.0158733

Tsetsos, K., Moran, R., Moreland, J. C., Chater, N., Usher, M., & Summerfield, C. (2016). Reply to Davis-Stober et al.: Violations of rationality in a psychophysical task are not aggregation artifacts. *Proceedings of the National Academy of Sciences*, *113*(33), E4764–E4766. https://doi.org/10.1073/pnas.1608989113

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Wallace, R. J. (2001). Instrumental Rationality—An overview | ScienceDirect Topics. *International Encyclopedia of the Social and Behavioral Sciences*. https://www-sciencedirect-com.ezproxy.princeton.edu/topics/social-sciences/instrumental-rationality