Temporal discounting correlates with directed exploration but not with random exploration

Hashem Sadeghiyeh[1,3], Siyu Wang[1], Maxwell R. Alberhasky[1], Hannah M. Kyllo[1], Amitai Shenhav[4], and Robert C. Wilson[1,2]

[1]Department of Psychology, University of Arizona
[2]Cognitive Science Program, University of Arizona
[3]Department of Psychological Science, Missouri University of Science and Technology
[3]Department of Cognitive, Linguistic, & Psychological Sciences, Brown University

September 26, 2019

Abstract

The explore-exploit dilemma describes the trade off that occurs any time we must choose between exploring unknown options and exploiting options we know well. Implicit in this trade off is how we value future rewards — exploiting is usually better in the short term, but in the longer term the benefits of exploration can be huge. Thus, in theory there should be a tight connection between how much people value future rewards, i.e. how much they discount future rewards relative to immediate rewards, and how likely they are to explore, with less 'temporal discounting' associated with more exploration. By measuring individual differences in temporal discounting and correlating them with explore-exploit behavior, we tested whether this theoretical prediction holds in practice. We used the 27-item Delay-Discounting Questionnaire (Kirby et al., 1999) to estimate temporal discounting and the Horizon Task (Wilson et al. 2014) to quantify two strategies of explore-exploit behavior: directed exploration, where information drives exploration by choice, and random exploration, where behavioral variability drives exploration by chance. We find a clear correlation between temporal discounting and directed exploration, with more temporal discounting leading to less directed exploration. Conversely, we find no relationship between temporal discounting and random exploration. Unexpectedly, we find that the relationship with directed exploration appears to be driven by a correlation between temporal discounting and uncertainty seeking at short time horizons, rather than information seeking at long horizons. Taken together our results suggest a nuanced relationship between temporal discounting and explore-exploit behavior that may be mediated by multiple factors.

Keywords— exploration; exploitation; explore-exploit dilemma; temporal discounting; delay; time

Introduction

The explore-exploit dilemma refers to a ubiquitous problem in reinforcement learning in which an agent has to decide between exploiting options it knows to be good and exploring options whose rewards are unknown (Sutton & Barto, 2018). For example, when ordering sushi at a favorite restaurant, should we exploit our usual favorite (the Rainbow Roll), which is guaranteed to be good, or explore the Burrito Roll, which could be delicious, disgusting or somewhere in between. As anyone who has agonized over a dining decision will know, making explore-exploit choices can be hard, and there is considerable interest in how these decisions are made by humans and other animals (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006).

Recently, a number of studies have shown that people make explore-exploit decisions using a mixture of two strategies: directed exploration and random exploration (Wilson, Geana, White, Ludvig, & Cohen, 2014; Gershman, 2018; Frank, Doll, Oas-Terpstra, & Moreno, 2009; Schulz & Gershman, 2019; Wyart & Koechlin, 2016; Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018). In directed exploration, choices are biased towards more informative options by an 'information bonus,' that increases the relative value of unknown options (Gittins, 1979). In random exploration, behavioral variability, perhaps driven by random noise processes in the brain, causes exploratory options to be chosen by chance (Watkins, 1989; Sutton & Barto, 2018). Further work suggests these two types of exploration have different computational properties (Gershman, 2018), age dependence (Somerville et al., 2017), and may be controlled by different systems in the brain (Zajkowski, Kossut, & Wilson, 2017; Blanchard & Gershman, 2018; Gershman & Tzovaras, 2018; Warren et al., 2017).

Regardless of the type of exploration, the benefits of exploring over exploiting lie in the possibility of earning larger rewards in the future. For example, in our restaurant example, if the Rainbow Roll is an above-average item on the menu, then, in the short term, exploiting it will usually be best. In the longer term, however, if the Burrito Roll turns out to be sublime, then we could order this roll again and again for years to come. Thus, how much we care about future rewards, that is how we discount them relative to immediate rewards, should play a critical role in how we make our explore-exploit choice.

Optimal models of explore-exploit decision making formalize this relationship between temporal discounting exploration, at least for directed exploration (Gittins, 1979). In these models, the explore-exploit choice is made by choosing the option that maximizes the expected discounted future reward. Because this maximizing behavior is deterministic (apart from rare cases in which options are tied), optimal models do not exhibit random exploration. Thus, while they predict a negative relationship between temporal discounting and directed exploration, they say nothing about the relationship with random exploration. Sub-optimal models of explore-exploit decision making do include random exploration, but most of them predict no relationship with temporal discounting (Thompson, 1933; Watkins, 1989; Sutton & Barto, 2018).

Thus, in theory, one might predict a negative relationship between temporal discounting and directed exploration, and no relationship between temporal discounting and random exploration. In practice, however, previous experimental work suggests a more nuanced picture because of how temporal discounting covaries with our attitudes toward risk. In particular, high temporal discounting is associated with greater impulsivity (Wittmann & Paulus, 2008), and higher impulsivity is associated with greater risk taking (Zuckerman & Kuhlman, 2000). This suggests that more temporal discounting is associated with more risk seeking (Madden & Bickel, 2010; Hill, Jenkins, & Farmer, 2008). In most explore-exploit paradigms, such increased risk taking would look a lot like increased directed exploration, because the more informative option is usually more uncertain, i.e. risky, too. Thus, while theory might predict a negative correlation between temporal discounting and directed exploration, this effect could be countered by a positive correlation between temporal discounting and risk taking.

In the current study, we investigated the correlation between temporal discounting and the two kinds of exploration using an individual differences approach. That is, we asked whether people with higher temporal discounting show less directed and/or random exploration. We used the 27-item Delay Discounting Questionnaire (Kirby & Maraković, 1996) to measure temporal discounting. In this question-

naire, participants choose between between small but immediate amounts of money and a larger but delayed amounts of money (e.g. $11 now or $30 in two weeks). Based on participants' pattern of choosing between immediate and delayed options, a parameter k (Kaplan et al., 2016) is calculated for each participant which estimates their average discounting rate for delayed rewards.

We used the Horizon Task (Wilson et al., 2014) to measure directed and random exploration. In this task participants make a series of choices between two slot machines (one-armed bandits). When played, each machine pays out a reward from a Gaussian distribution. The average payout is different for each machine such that one option is always better on average. Thus, to maximize their rewards, participants need to exploit the option with the highest average payout, but can only find out which option is best by exploring both options first. By manipulating key parameters in this task (distribution of rewards, time horizon, and the amount of uncertainty for each bandit), the Horizon Task allows us to quantify directed and random exploration, and, crucially, to dissociate them from baseline risk seeking and behavioral variability.

Thus, by comparing individual differences in behavior on the Horizon Task with individual differences in temporal discounting, we aimed to quantify the relationship between the two types of exploration and temporal discounting.

Methods

Participants

We collected data from a total of 82 participants (ages 18-25, average = 19.10; Females = 47, Males = 35). Participants were recruited through the Psychology subject pool at the University of Arizona and received course credit for their participation. All participants gave informed consent and the study was approved by the Institutional Review Board at the University of Arizona.

Temporal Discounting measure

To measure temporal discounting we used the Delay Discounting Questionnaire developed by (Kirby, Petry, & Bickel, 1999). In this instrument there are 27 questions asking participants' preferences between two hypothetical monetary rewards: one of which pays immediately but is smaller, and the other pays more but is delayed. For example, one item asks: Do you prefer $11 now or $30 in 7 days? The amount of sooner-immediate reward ("today" option), later-delayed reward ("later" option) and the delay (in terms of days) vary in those 27 questions ("today" reward between $11 - $80; "later" reward between $25 - $85; Delay between 7 - 186 days). The exact values are reported in Kirby et al. (1999)-Table 3.

To quantify temporal discounting we used a number of different measures. The simplest was just the number of today options chosen, with greater temporal discounting associated with larger number of "today" choices.

More sophisticated measures of temporal discounting were obtained by fitting a hyperbolic discount factors to the data. In particular, we assume that future reward, A, arriving after a delay D, is discounted according to a hyperbolic discount function (Mazur, 1987)

$$V = A/(1 + kD) \tag{1}$$

where k is the subject-specific discount factor. Fitting k was done using the spreadsheet provided by (Kaplan, Lemley, Reed, & Jarmolowicz, 2014) based on the method described in (Kaplan et al., 2016). In addition to computing an overall k using all 27 items, this approach also computes separate discount factors for small, medium and large reward items, based on the idea that delay discounting may be different for different range of rewards, and the geometric mean of the small, medium and large ks. Thus we have six measures of temporal discounting for each subject: the fraction of "today" choices, overall k, small k, medium k, large k, and the geometric mean of small, medium and large ks. Importantly, as shown below, the correlations with explore-exploit behavior are almost identical for all of these measures.

Horizon Task

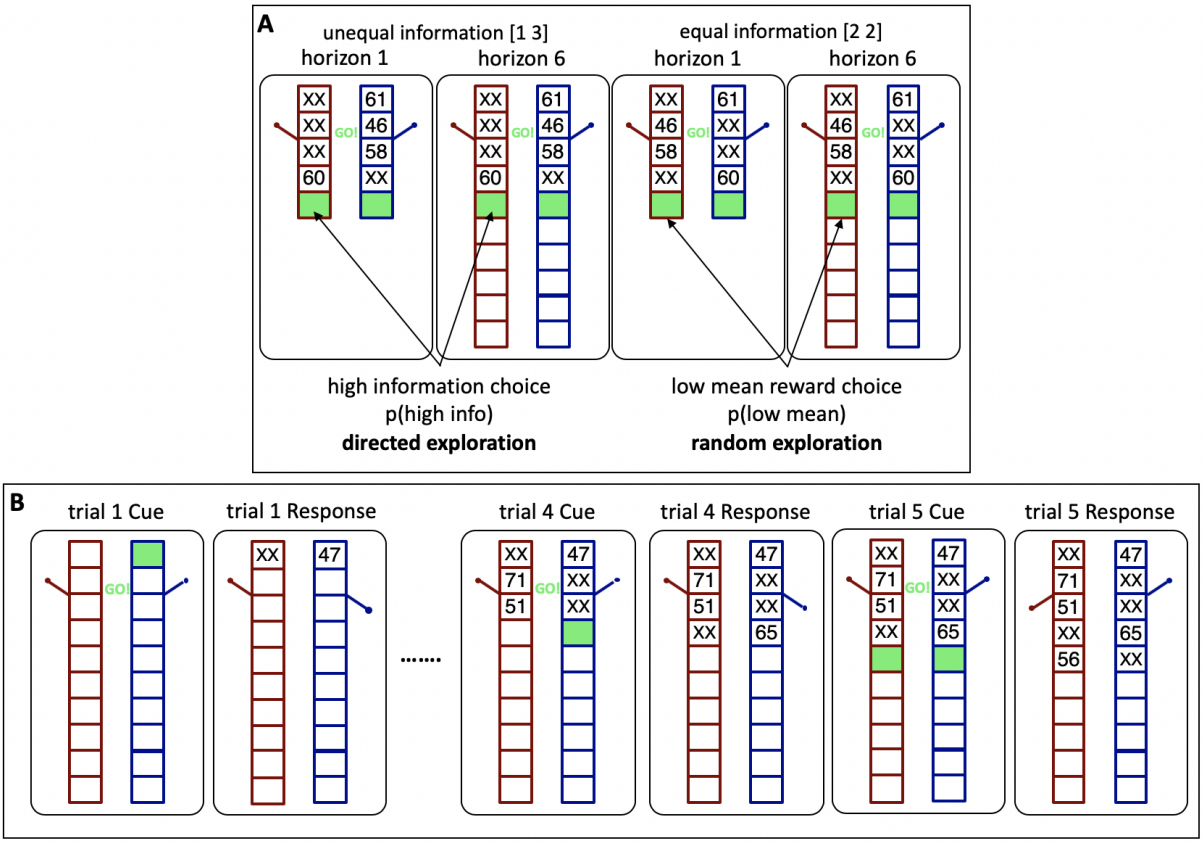

Figure 1: A)Horizon task: the four forced trials set up one of two information conditions (unequal [1 3] and equal [2 2] information) and two horizon conditions (1 vs 6) before participants make their first free choice. B) Active version of the horizon task

The Horizon Task (Wilson et al., 2014) is a recently developed task that allows for the measurement of directed and random exploration. The key manipulation in the Horizon Task is the time horizon, the number of trials participants will make in the future. The idea being, that in a long time horizon, people should explore, while in a short time horizon, people should exploit. Thus the change in behavior between short and long horizons can be used to quantify directed and random exploration.

More specifically, in the Horizon Task participants choose between two one-armed bandits. When chosen, the bandits pay out rewards sampled from a Gaussian distribution whose standard deviation is always fixed at 8 points, but whose mean is different for each bandit and can change from game to game. Each game lasts for 5 or 10 trials and participants' job is to make multiple choices between the two bandits to try to maximize their reward. Because they know nothing about the mean of each bandit at the start of each game, they can only find out which option is best by exploring.

To control the amount of information, the first four trials of each game are predetermined (Figure 1-B). Participants are instructed to pick either the left or right bandit during these four "forced trials". By changing the number of forced choices for each bandit, we manipulate the amount of "uncertainty" or information participants have about the payoffs from each bandit. In the unequal uncertainty (or [1 3] condition) participants are forced to choose one option once and the other three times; whereas in the equal uncertainty (or [2 2] condition) participants play both options twice. After the forced-choice trials, the rest of trials are "free trials" in which participants make their own choice. The number of free trials

varies between horizon conditions with 1 free choice in the horizon 1 condition and 6 free choices in the horizon 6 condition.

These two information conditions allow us to quantify directed and random exploration by looking at the first free choice in each game, immediately after the four forced choices (Figure 1-A). Because directed exploration involves information seeking, it can be quantified as the probability of choosing the more informative option in the [1 3] condition, p(high info). Conversely, because random exploration involves decision noise, it correlates with choosing the low mean option in the [2 2] condition, p(low mean). Computing these measures separately for each horizon condition allows us to quantify four key properties of explore-exploit behavior:

- uncertainty preference as p(high info) in horizon 1
- baseline behavioral variability as p(low mean) in horizon 1
- directed exploration as Δp(high info), the change in information seeking with horizon
- random exploration as Δp(low mean), the change in variability with horizon

Results

Behavior on the Horizon Task

Behavior on the Horizon Task was consistent with that previously reported in (Wilson et al., 2014). Specifically we see a significant increase in p(low mean) with horizon (p(low mean)h1_average = 0.2883; p(low mean)h6_average = 0.3554; t(81) = 3.87; p < 0.001) and we see a clear trend (but not significant) in p(high info) with horizon (p(high info)h1_average = 0.5146; p(high info)h6_average = 0.5486; t(81) = 1.75; p = 0.084), consistent with participants using both types of exploration in this paradigm.

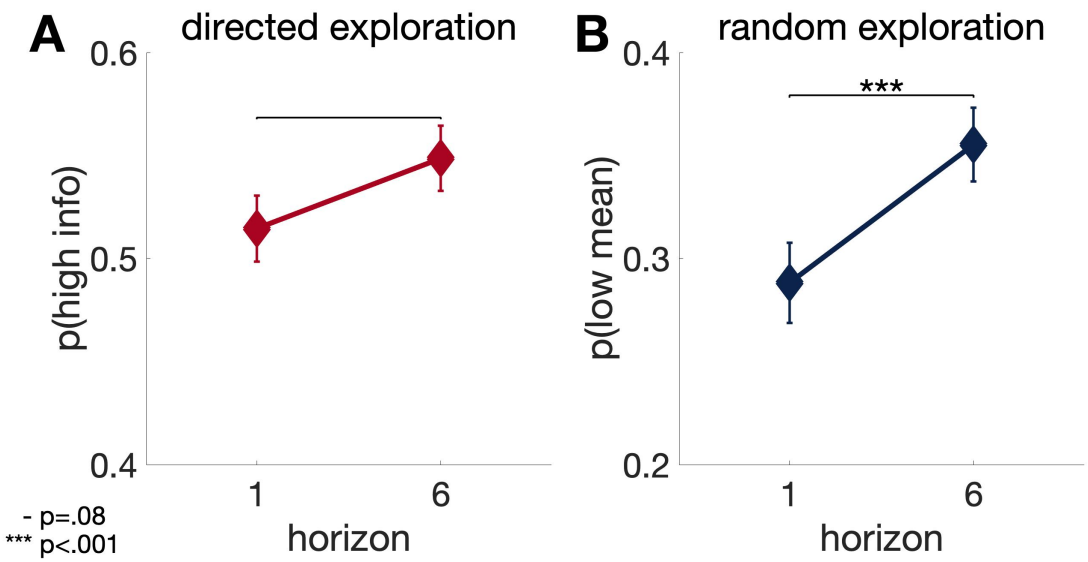

Figure 2: The average of p(high info) (A) and p(low mean) (B) for 82 participants on each horizon condition. The increase in p(high info) and p(low mean) from horizon 1 to horizon 6 follows the typical pattern observed in our previous studies and shows the use of both directed and random exploration.

Behavior on the Temporal Discounting Task

For the temporal discounting measure we obtained 5 different k values for each participant as a measure of how much they discount future reward. We also can simply estimate that measure just by counting the number of times participants chose the immediate versus delayed reward. In our research, it turned out that all of these indices are highly correlated with each other and all have very similar relationship with directed and random exploration. Table 1 shows the range and average of temporal discounting indices (k's and # today items) in 82 participants of our study which is similar to previous studies using the same measure (Kirby et al., 1999; Kirby & Maraković, 1996).

Table 1: Minimum, maximum and the linear average of different k values and # today items for 82 participants

	min	max	average
Overall k	0.0004	0.2494	0.0303
Small k	0.0016	0.2468	0.0476
Medium k	0.0002	0.25	0.0297
Large k	0.0002	0.2488	0.0201
Geomean k	0.0004	0.2485	0.0276
# today items	4	27	15.85

Correlation between temporal discounting and explore-exploit behavior

Table 2 shows the correlation between measures of temporal discounting and the horizon task parameters: directed and random exploration, p(high info) & p(low mean) at horizons 1 & 6, reaction times and accuracy (the percentage of times the "accurate" option (the higher mean option) was chosen for each horizon (1 & 6) conditions. We found a significant negative correlation between between temporal discounting and directed exploration, with more temporal discounting associated with less directed exploration. Closer inspection revealed that this negative correlation was driven by a positive correlation between temporal discounting and p(high info) at horizon 1 and a zero correlation between temporal discounting and p(high info) at horizon 6.

Table 2: Correlation between task parameters and log (k overall)

	r	p
directed exploration	**-0.30**	0.007
random exploration	0.04	0.720
p(high info) h1	**0.35**	0.001
p(high info) h6	-0.01	0.924
p(low mean) h1	0.22	0.052
p(low mean) h6	**0.27**	0.014
accuracy h1	**-0.33**	0.003
accuracy h6	-0.17	0.130
reaction time h1	-0.05	0.669
reaction time h6	-0.12	0.304

In contrast to directed exploration, temporal discounting did not correlate with random exploration. There was, however, a positive correlation between temporal discounting and overall behavioral variability, p(low mean) in both horizon conditions. This suggests that people with higher temporal discounting perform worse on the task overall.

Finally, to demonstrate that the significant correlations were not driven by outliers, we plot the correlations between measures of directed and random exploration and the number of today items chosen in Figure 3.

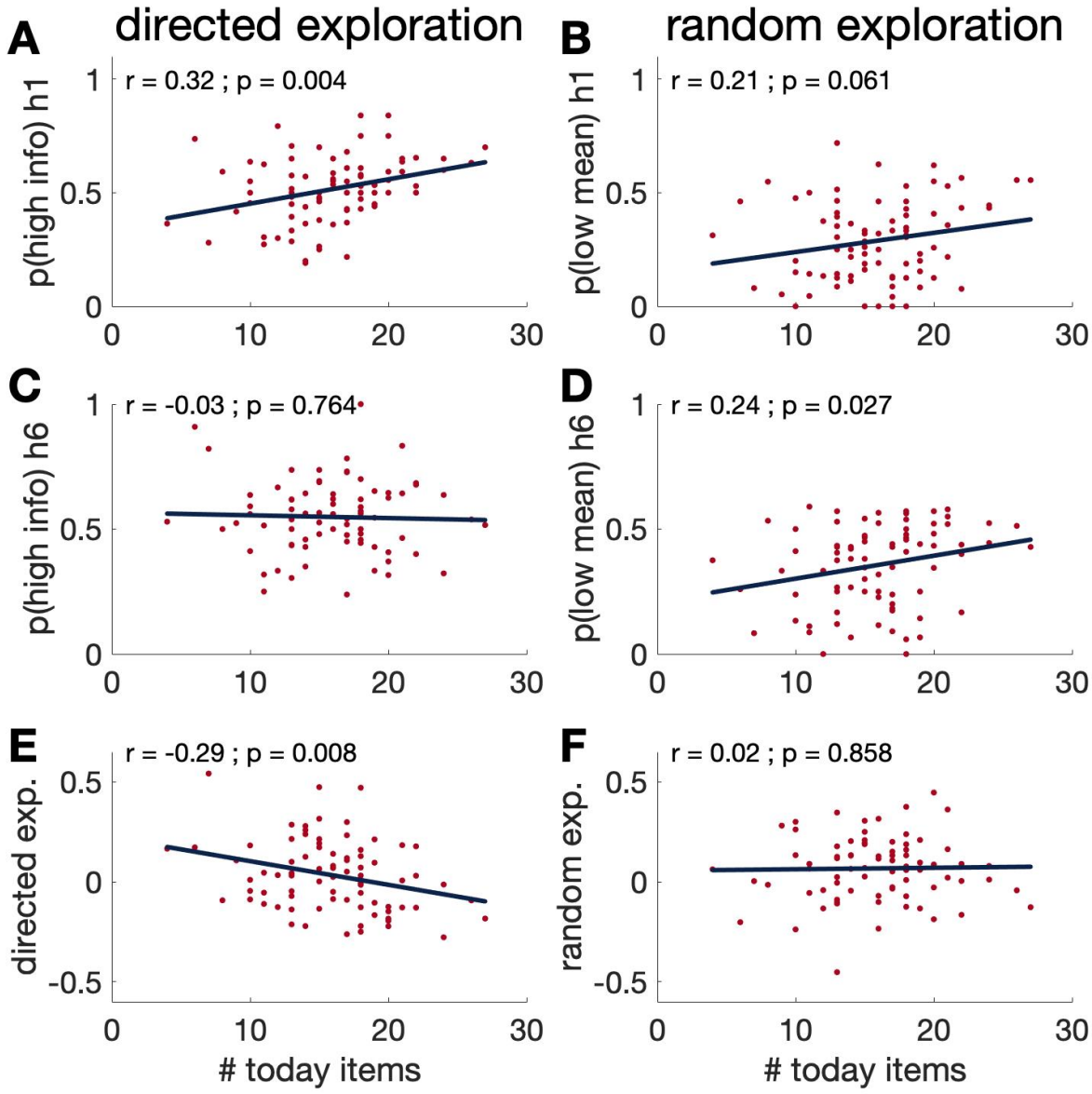

Figure 3: Scatter plots for (A) p(high info) h1, (B) p(low mean) h1, (C) p(high info) h6, (D) p(low mean) h6, (E) directed exploration and (F) random exploration over a temporal discounting measure (# today items). It clearly shows that the negative correlation between temporal discounting and directed exploration is driven by a positive correlation

Discussion

In this study we investigated the correlation between temporal discounting measured by a monetary choice questionnaire (Kirby et al., 1999) and two types of exploration (directed and random) measured

by the Horizon Task (Wilson et al., 2014). We found a negative correlation between temporal discounting and directed exploration that was driven by a positive correlation between temporal discounting and uncertainty seeking in horizon 1. Conversely, we found no correlation between temporal discounting and random exploration, although we did see a positive correlation between temporal discounting and overall behavioral variability.

While the negative correlation between temporal discounting and directed exploration (i.e. $\Delta p(\text{high info})$ is consistent with the theory, the correlation with $p(\text{high info})$ in each horizon condition is not. In particular, normative models predict a negative correlation between temporal discounting and $p(\text{high info})$ in horizon 6 and no correlation in horizon 1. Conversely, we found no correlation with horizon 6 behavior and a positive correlation with horizon 1 behavior.

One reason for this discrepancy could be the positive association between temporal discounting and risk taking (Madden & Bickel, 2010; Hill et al., 2008). In both horizon conditions in the Horizon Task, the more informative option is also the more uncertain, riskier option. Thus, by this account, people who discount more would show greater $p(\text{high info})$ in both horizon conditions, but this would be counteracted by a negative relationship between temporal discounting and directed exploration in horizon 6. That is, in horizon 1, directed exploration is not present, and so the positive association with temporal discounting is revealed. In horizon 6, directed exploration is present, and this negative relationship with temporal discounting counteracts the positive relationship with risk taking leaving no correlation overall.

The fact that random exploration does not correlate with temporal discounting is also consistent with theories of random exploration (Watkins, 1989; Sutton & Barto, 2018). Moreover, this apparent dissociation between directed and random exploration is consistent with other findings showing that directed and random exploration have different computational properties ((Gershman, 2018), different age dependence (Somerville et al., 2017), and may rely on dissociable neural systems (Zajkowski et al., 2017; Gershman & Tzovaras, 2018; Warren et al., 2017). In this regard it is notable that directed exploration appears to rely on the same frontal systems thought to underlie temporal discounting (Frank et al., 2009, Gershman & Tzovaras, 2018, Zajkowski et al., 2017; Doya, 2002; McClure, Laibson, Loewenstein, & Cohen, 2004; McClure, Ericson, Laibson, Loewenstein, & Cohen, 2007), while random exploration does not. Thus, an intriguing prediction is that the relationship between directed exploration and temporal discounting may be mediated by the integrity of frontal circuits, something that future neuroimaging studies could address.

Acknowledgements

The authors thank Shlishaa Savita and Kathryn Lui Kellohen for their help in collecting and organizing data.

Author Contributions Statements

HS, SW, AS and RCW designed the experiment. HS, MRA and HMK ran the experiment. HS and SW analyzed the data with supervision from RCW. HS and RCW wrote the manuscript with input from all other authors.

Competing Interests Statement

The authors declare no competing interests.

References

Blanchard, T. C., & Gershman, S. J. (2018, 2). Pure correlates of exploration and exploitation in the human brain. *Cognitive, Affective and Behavioral Neuroscience*, *18*(1), 117–126. doi: 10.3758/s13415-017-0556-2

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006, 6). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. doi: 10.1038/nature04766

Doya, K. (2002, 6). Metalearning and neuromodulation. *Neural Networks*, *15*(4-6), 495–506. doi: 10.1016/S0893-6080(02)00044-8

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009, 8). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, *12*(8), 1062–1068. doi: 10.1038/nn.2342

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42. doi: 10.1016/j.cognition.2017.12.014

Gershman, S. J., & Tzovaras, B. G. (2018). Dopaminergic genes are associated with both directed and random exploration. *Neuropsychologia*, *120*, 97–104. doi: 10.1016/j.neuropsychologia.2018.10.009

Gittins, J. C. (1979). Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*(2), 148–177.

Hill, E. M., Jenkins, J., & Farmer, L. (2008). Family unpredictability, future discounting, and risk taking. *The Journal of Socio-Economics*, *37*(4), 1381–1396.

Kaplan, B. A., Amlung, M., Reed, D. D., Jarmolowicz, D. P., McKerchar, T. L., & Lemley, S. M. (2016, 10). Automating Scoring of Delay Discounting for the 21- and 27-Item Monetary Choice Questionnaires. *Behavior Analyst*, *39*(2), 293–304. doi: 10.1007/s40614-016-0070-9

Kaplan, B. A., Lemley, S. M., Reed, D. D., & Jarmolowicz, D. P. (2014). *21- and 27- Item Monetary Choice Questionnaire Automated Scorer [spreadsheet application].* University of Kansas. Retrieved from `https://kuscholarworks.ku.edu/handle/1808/15424`

Kirby, K. N., & Maraković, N. N. (1996, 3). Delay-discounting probabilistic rewards: Rates decrease as amounts increase. *Psychonomic Bulletin and Review*, *3*(1), 100–104. doi: 10.3758/BF03210748

Kirby, K. N., Petry, N. M., & Bickel, W. K. (1999). Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. *Journal of Experimental Psychology: General*, *128*(1), 78–87. doi: 10.1037/0096-3445.128.1.78

Madden, G., & Bickel, W. (2010). *Impulsivity: The behavioral and neurological science of discounting.* Washington, DC: American Psychological Association.

Mazur, J. E. (1987). An adjusting procedure for studying delayed reinforcement. In M. L. Commons, J. E. Mazur, & J. A. Nevin (Eds.), *Quantitative analyses of behavior: vol. 5. the effect of delay and of intervening events on reinforcement value* (pp. 55–73). doi: 10.1016/j.pbb.2008.06.017

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. doi: 10.1523/JNEUROSCI.4246-06.2007

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate Neural Systems Value Immediate and DelayedMonetary Rewards: EBSCOhost. *Science*, *306*(October), 503–507. doi: 10.1126/science.1100907

Schulz, E., & Gershman, S. J. (2019). *The algorithmic architecture of exploration in the human brain* (Vol. 55). doi: 10.1016/j.conb.2018.11.003

Somerville, L. H., Sasse, S. F., Garrad, M. C., Drysdale, A. T., Abi Akar, N., Insel, C., & Wilson, R. C. (2017, 2). Charting the expansion of strategic exploratory behavior during adolescence. *Journal of experimental psychology. General*, *146*(2), 155–164. doi: 10.1037/xge0000250

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning : an introduction* (Second ed.). Cambridge, MA: MIT press.

Thompson, W. R. (1933, 12). On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, *25*(3/4), 285. doi: 10.2307/2332286

Warren, C. M., Wilson, R. C., Van Der Wee, N. J., Giltay, E. J., Van Noorden, M. S., Cohen, J. D., & Nieuwenhuis, S. (2017, 4). The effect of atomoxetine on random and directed exploration in humans. *PLOS ONE*, *12*(4), e0176034. doi: 10.1371/journal.pone.0176034

Watkins, C. (1989). *Learning from delayed rewards* (Unpublished doctoral dissertation). Cambridge University.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014, 12). Humans use directed and random exploration to solve the explore-exploit dilemma. *Journal of experimental psychology. General*, *143*(6), 2074–81. doi: 10.1037/a0038199

Wittmann, M., & Paulus, M. P. (2008). Decision making, impulsivity and time perception. *Trends in Cognitive Sciences*, *12*(1), 7–12. doi: 10.1016/j.tics.2007.10.004

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). *Generalization guides human exploration in vast decision spaces* (Vol. 2) (No. 12). doi: 10.1038/s41562-018-0467-4

Wyart, V., & Koechlin, E. (2016). Choice variability and suboptimality in uncertain environments. *Current Opinion in Behavioral Sciences*, *11*, 109–115. doi: 10.1016/j.cobeha.2016.07.003

Zajkowski, W. K., Kossut, M., & Wilson, R. C. (2017). A causal role for right frontopolar cortex in directed, but not random, exploration. *eLife*, *6*. doi: 10.7554/eLife.27430

Zuckerman, M., & Kuhlman, D. M. (2000). Personality and risk-taking: Common biosocial factors. *Journal of Personality*, *68*(6), 999–1029. doi: 10.1111/1467-6494.00124