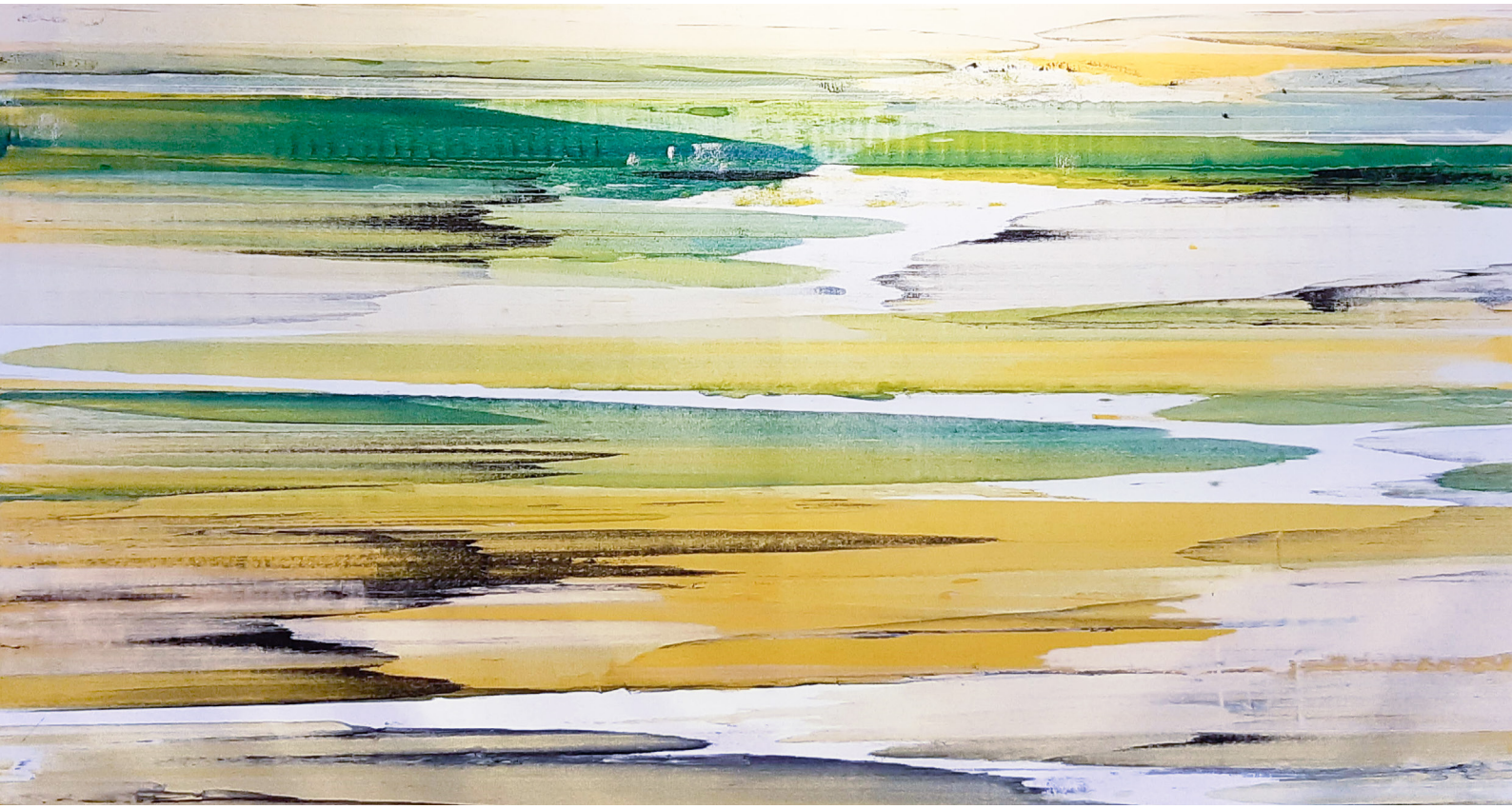
ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Центр междисциплинарных исследований человеческого потенциала

Как стратегия выбора зависит от контекста?

Научный дайджест № 8 (13) • 2022



Научный центр мирового уровня «Центр междисциплинарных исследований человеческого потенциала» создан в ноябре 2020 г. в рамках Национального проекта «Наука» как консорциум четырех организаций – лидеров в данной области научного знания: НИУ ВШЭ, РАНХиГС, МГИМО МИД России и Института этнологии и антропологии им. Н.Н. Миклухо-Маклая РАН.

Создание Центра стало самой масштабной в России инициативой в области социальных и гуманитарных наук за последние десятилетия. Среди его основных задач – не только проведение исследований мирового уровня в области развития человеческого потенциала, но и налаживание кооперации с зарубежными организациями-лидерами, запуск образовательных программ, создание передовой научной инфраструктуры, обеспечение трансфера полученных результатов в практику государственного управления и образование.

Центр реализует 78 научных проектов. Программа исследований охватывает ключевые аспекты человеческого потенциала, актуальные сегодня в глобальной повестке:

- социальное и гуманитарное измерение человеческого потенциала
- нейрокогнитивные механизмы социального поведения
- демографические и социальные факторы активного долголетия
- природно-климатические детерминанты устойчивого развития
- занятость, социальная активность и формирование ключевых навыков и компетенций
- человеческий потенциал и безопасность в глобальном мире
- человек в эпоху технологических трансформаций

Научный дайджест подготовлен в рамках проекта «Контекстно-зависимое кодирование в обучении с подкреплением».

Руководитель проекта «Научный дайджест НЦМУ ЦМИЧП»: О.В. Ворон
Авторы: С. Палминтери, М. Лебретон
Редакторы: А. Андрианова, А. Гамбарян

Введение

Со времен появления в психологии Гештальт теории [1] и теории полезности [2] ученые исследовали зависимость перцептивного восприятия и оценки субъективной ценности от контекста. Благодаря десятилетиям исследований в психологии, экономике и нейронауке, в процессе изучения мозговых механизмов восприятия и принятия решений, не осталось никаких сомнений в том, что перцептивные и экономические решения находятся в зависимости от контекста [3]. Оказалась, что в основе данной зависимости лежат два различных вычислительных механизма: центрирование вокруг точки отсчета и диапазонная адаптация [4–6].

В реальной жизни в процессе принятия решения результаты прошлого выбора подвержены ретроспективному пересмотрению [7]. Тем не менее, нейромоделирование принятия решений, основанных на опыте (обучение с подкреплением), практически игнорирует проблему контекстной зависимости [8, 9]. Данный обзор содержит анализ последних экспериментальных исследований, показывающих, что в обучении с подкреплением у человека результат выбора кодируется и запоминается в зависимости от контекста обучения.

Основываясь на ранних работах по перцептивному принятию решений – ситуации неопределенности (самых простых решений, например, о том, куда движутся точки, налево или направо, если речь идет о зрительном восприятии), мы представляем контекстную зависимость результата выбора как выражение адаптивного кодирования. Последнее означает, что нейрорепрезентация определённой переменной (кодирование переменной мозгом) ограничена ее статистическим распределением (контекст [4, 5]). Аналогичным образом, в процессе обучении с подкреплением кодирование результата выбора зависит от распределения результатов выбора, пережитых в том или ином контексте.

Зависимость от точки отсчета в обучении с подкреплением

Согласно теории уровней адаптации Гарри Хелсона (1898–1977) [10], перцептивные характеристики стимула (освещённость, громкость, вес и др.) оцениваются относительно нормы (или уровню адаптации) следующим образом:

$$J_i = S_i - \bar{S}$$

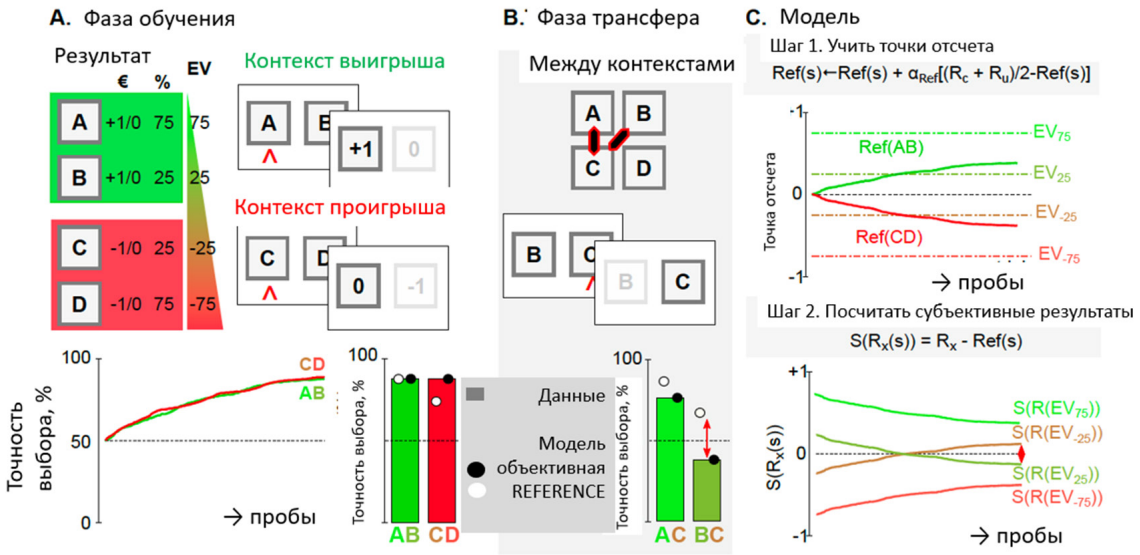
где J_i является оценкой стимула i по определённому атрибуту (свойству), S_i является объективной оценкой того же стимула по тому же атрибуту, и $\bar{S}$ является нормой, арифметическим средним оценок подобных стимулов относительно контекста. Норма является точкой отсчета, обычно определяемой как скользящее среднее сходных стимулов, попавших в недавнюю или текущую выборку (так называемое центрирование). Как следует из приведенной выше формулы, норма используется нашим мозгом для оценки текущего стимула). Используя основную идею теории уровней адаптации в области экономических решений, Канеман и Тверски предположили, что полезность ожидаемого результата отражает не его объективную ценность, а ощущение выгоды или потери относительно какой-то точки отсчета. Зависимость от точки отсчета является внутренним свойством предложенной Канеманом и Тверски «Теории перспектив» ([11, 12]).

Мы протестировали, влияет ли зависимость от точки отсчета на то, как кодируется (и запоминается) результат выбора в ходе обучения с подкреплением у человека [13]. Схема нашего поведенческого эксперимента объединяет фазу обучения с фазой перехода (на рис: трансфера) между контекстами (выигрыш или проигрыш) [14, 15]. Сначала в фазе обучения участники эксперимента выбирают между двумя опциями, представленными как фиксированные пары стимулов, ассоциированных с вероятностными денежными выигрышами (или проигрышами). Тип результата определял контекст обучения: “выигрыш” (максимизация вознаграждений) или “проигрыш” (минимизация наказания) (Рис. 1A). В фазе перехода участникам эксперимента нужно было выразить свои предпочтения в каждой паре возможных комбинаций всех опций, использованных в фазе обучения (Рис. 1B). При анализе поведения испытуемых мы получили два ключевых результата: 1) В фазе обучения точность выбора была значительно выше уровня случайных выбора и была одинаковой и в контексте выигрыша, и в контексте проигрыша. 2) В фазе перехода испытуемые предпочитали опции небольших потерь относительно опций небольших выигрышей. Результаты продемонстрировали нарушение строгого монотонного ряда опций по ожидаемой ценности (Рис. 1A-B). При этом данные два результата противоречат предсказанию (выбора) стандартного Q-обучения. В частности, в фазе обучения стандартная модель Q-обучения предсказывает более низкий уровень выполнения задачи в контексте проигрышей по сравнению с контекстом выигрышей. Последнее обусловлено внутренней асимметрией нормы обучения с подкреплением в контексте и выигрышей, и проигрышей (так называемый парадокс обучения наказанию [16–18]. А в фазе перехода стандартная модель предсказывает строгое монотонное распределение предпочтения опций как функций от их объективной ценности.

Следуя логике теории уровней адаптации и теории перспектив, мы предложили модель, которая обучается ценности точки отсчета и использует ее для динамического центрирования результатов выборов до того, как посчитать ошибку предсказания для конкретной опции (Рис. 1C). Мы называли данный алгоритм моделью REFERENCE, и она превзошла стандартное Q-обучение в широком ряде условий. Данный результат был воспроизведен не только в нашей лаборатории, но и в других исследованиях: в экспериментах с разными дизайнами, включая социальное обучение [19], а также в экспериментах с разными взаимозависимостями опций, последовательностей и манипуляций [20–22].

Рисунок 1

Зависимость от точки отсчета в обучения с подкреплением: задание, результаты и модельные переменные



(A) Контексты фазы обучения (на рисунке вверху) и типичное поведение (на рисунке внизу). Участникам предъявлялся ряд проб с двумя контекстами обучения: AB (контекст максимизации выигрыша) и CD (контекст минимизации проигрыша). Обратная связь (вознаграждение) имела носила вероятностный характер. Точность выбора прогрессивно нарастала с уровня случайного выбора пока не достигала одинакового уровня в двух контекстах.

(B) Контексты фазы перехода или трансфера (на рисунке вверху) и типичное поведение (на рисунке внизу). После фазы обучения символы переорганизованы в новых комбинациях. В иллюстративных целях мы фокусируемся на самых информативных комбинациях (AC и BC). Признаком зависимости результата выбора от точки отсчета является предпочтение C относительно B в пробах сравнений BC (зеленый столбик). Предсказание выбора моделью REFERENCE (черные точки). Предпочтения в пробах сравнений AC также информативны и указывают, что эффект только частичный.

(C) Эволюция контекстных переменных (на рисунке вверху) и субъективных результатов (на рисунке внизу). Верхняя часть рисунка иллюстрирует динамику эволюции точки отсчета в контекстах выигрыша и проигрыша. На полпути обучения точка отсчета пересекает ожидаемую ценность маленьких выигрышных/проигрышных опций. Нижняя часть рисунка иллюстрирует динамику средних субъективных результатов для каждой опции. Симметрично к верхней панели, примерно на полпути обучения, субъективная ценность опций EV_{25} и EV_{-25} начинает субъективно восприниматься как негативная и положительная, соответственно.

Диапазонная адаптация результатов выбора в обучении с подкреплением

В конце 20-го века Аллен Пардучи выявил присутствие зависимости от контекста в аффективной оценке счастья, удовольствия и боли и сформулировал теорию диапазонной частоты [23]. В нашем обзоре нам было особенно важно коснуться принципа диапазонов, который описывает субъективную оценку стимула J_i как:

$$J_i = \frac{S_i - S_{min}}{S_{max} - S_{min}}$$

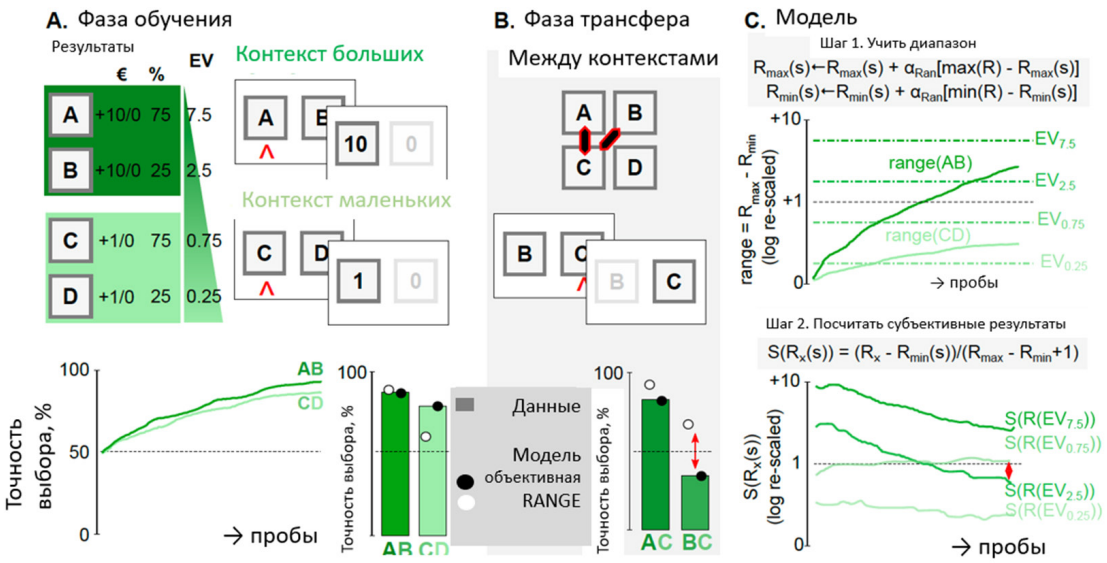
где S_i является объективной ценностью стимула i по поводу его рассматриваемой характеристики, а S_{max} и S_{min} являются самыми высокими и низкими ценностями, представленными в данном контексте, которые замыкают интервал возможных результатов выбора. Принцип диапазонов утверждает, что субъективное оценивание адаптируется к распределению стимула через нормализацию. Недавно Контек и Левандовски транслировали эту идею в область исследования принятия решений и предложили модель полезности, зависящей от диапазона, как альтернативу теории перспектив [24]. Эта модель предполагает, что оценка лотерей зависит от диапазона. Такой подход объясняет несколько известных поведенческих парадоксов, связанные с когнитивными искажениями и смещениями оценок при принятии решений [25].

В нескольких экспериментах мы проверили, применяется ли принцип диапазонов к кодированию результатов выбора и ретроспективному запоминанию в обучении с подкреплением [26, 27]. Основываясь на прошлых поведенческих экспериментах по систематическому манипулированию размера результатов выбора, мы сгенерировали контексты обучения с разными диапазонами результатов выбора. Как и в прошлом исследовании, за фазой обучения последовала фаза перехода, которая включала новые комбинации альтернатив (Рис. 2A-B). Здесь также были выявлены два эффекта: 1) точность выполнения задания была одинаковой в контекстах маленьких и больших размеров результатов выбора; 2) в фазе трансфера предпочтения участников эксперимента не соответствовали объективным ценностям результатов выбора. Важно отметить, что альтернативы, которые были локально лучше в контексте маленьких выигрышей/проигрышей, систематически выбирались чаще тех, которые были локально хуже в контексте больших размеров результатов выбора несмотря на то, что объективная ожидаемая ценность последних была выше. Стандартное Q-обучение (с объективными результатами выборов и soft-max правилом принятия решений [28]) не смогло предсказать такое поведение игрока. В соответствии с теорией частоты диапазона, мы предложили модель, которая обучается диапазону возможных результатов и использует его для динамического масштабирования результатов выбора до того, как посчитать ошибку предсказания конкретной альтернативы (Рис. 2C). Эта модель, которую мы назвали RANGE, удовлетворительно предсказала поведенческие эффекты. В нашем последнем исследовании [27] мы изменили сложность фазы обучения двумя способами: 1) манипулируя информацией о результате выбора (частичная или полная обратная связь) и 2) манипулируя структурой задачи (блок-дизайн или смешанный дизайн). Мы выявили, что адаптация по диапазону

результатов выбора была особенно выраженной в легких условиях задачи (блок-дизайн с полной обратной связью). Последнее соответствует идее, что простые условия позволили участникам легче определить контекстно-зависимые переменные. Как предсказывает модель RANGE, такой результат сопутствует снижению способности экстраполировать ценности альтернатив выбора в процессе смены контекста. Данное открытие сильно противоречит общепринятому пониманию, что снижение сложности задания ведет к повышению точности выполнения и рациональности поведения [29, 30]. Другое исследование, использовавшее обучение с подкреплением с одной определенной (без риска) и одной вероятностной (с рисками) опциями показало, что диапазон результатов выбора влияет на субъективную ценность результатов выбора [31]. Предпочтения более рискованных лотерей сильно зависели от повышения заметности крайних опций (например, возможный максимум и минимум), когда они представлены в одном контексте, чем когда прикреплены к конкретному объективному результату.

Рисунок 2

Диапазонная адаптация в обучении с подкреплением: задание, результаты и модельные переменные



(A) Контексты фазы обучения (на рисунке вверху) и типичное поведение (на рисунке внизу). Участникам представлены несколько проб с двумя контекстами обучения: AB (контекст большого размера результатов) и CD (контекст маленького размера результатов). Обратная связь носила вероятностный характер. Точность выбора испытуемыми обычно начиналась с уровня случайного выбора и прогрессивно увеличивалась, достигая одинакового уровня в двух контекстах.

(B) Контексты фазы перехода (трансфера) (на рисунке вверху) и типичное поведение (на рисунке внизу). После фазы обучения символы переорганизовались в новых комбинациях. В иллюстративных целях мы фокусируемся на самых информативных комбинациях (AC и BC). Признаком диапазонной адаптации является предпочтение C относительно B в пробах сравнений BC (зеленый столбик). Предсказания модели, предполагающей объективное кодирование (белые точки). Предсказание выборов моделью RANGE (черные точки). Предпочтения в пробах сравнений AC также информативны и указывают, что эффект только частичный.

(C) Эволюция контекстных переменных (на рисунке вверху) и субъективных результатов (на рисунке внизу). Верхняя часть рисунка иллюстрирует динамику эволюции диапазона в контекстах больших и маленьких результатов выбора. Изменения R_{max} и R_{min} зависят от тог, что R> R_{max} и R <R_{min}, соответственно. Нижняя часть рисунка иллюстрирует динамику средних субъективных результатов для каждой опции. Примерно на полпути обучения, субъективная ценность опций EV_{25} и EV_{-25} пересекаются.

Функциональная роль контекстной зависимости кодирования результатов выбора в процессе обучении с подкреплением

Зависимость кодирования результатов выбора от контекста систематически приводит к суб-оптимальности выбора, когда опции экстраполируются за пределы изначального контекста обучения в фазу трансфера (Рисунок 1-2). Мы показали, что зависимость от контекста может улучшить обучение в отдельных условиях (избежание потерь, маленькие размеры результатов). Тем не менее, большинство таких эффектов могли быть достигнуты путем нормализации сигналов ценности на этапе выбора, а не на этапе обучения и запоминания, таким образом, избегая издержек иррациональных предпочтений в фазе трансфера. Мы предполагаем две функциональные роли для искажений в обучении. Во-первых, зависимость результатов выбора от контекста могла вытекать из адаптивного и эффективного нейрокодирования, что оптимизирует обработку информации во время обучения [4, 5]. Во-вторых, несмотря на то, что зависимость от контекста приводит к суб-оптимальности выборов, они могут быть эволюционно рациональны, т.е. в среднем генерируют оптимальное выполнение задач именно в тех средах, где они возникают. Например, в ситуации, где ресурсы крайне изменчивы [32, 33].

Вопросы для дальнейшего исследования

Мы показали, что зависимость кодирования результатов выбора от контекста основывается на сочетании фазы обучения и фазы перехода или трансфера между контекстами проигрышей и выигрышей. Другое недавнее исследование показало, что система обучения по Павлову демонстрирует похожие ограничения кодирования результатов выбора [34]. Будущие исследования должны ответить на вопрос, распространяется ли описанный нами механизм на другие системы обучения (обучение по Павлову, инструментальное, целенаправленное) и репрезентации (декларативная, эпизодическая) [35, 36]. Несмотря на то, что наш обзор сфокусирован на ситуации, где контекстно-зависимое обучение с подкреплением соединяет проблемы обучения и распространения результатов обучения на другие контексты, для полноты исследования нужно обращаться к ситуациям, где акцент будет сделан на распространении результатов обучения на другие контексты, т. е. на фазу трансфера.

Выявленные механизмы и свойства зависимости от точки отсчета и диапазонной адаптации являются ключевыми для оценки нейробиологических основ обучения и субъективной оценки опций в процессе выбора [13 , 37, 38].

Литература

1. Fechner GT: Elemente der psychophysik. Breitkopf und Härtel; 1860.
2. Bernoulli D: Specimen theoriae novae de mensura sortis. Comment Acad Sci Imp Petropolitanae 1738, 5:175–192.
3. Kahneman D: Maps of Bounded Rationality: Psychology for Behavioral Economics. Am Econ Rev 2003, 93:1449–1475.
4. Carandini M, Heeger DJ: Normalization as a canonical neural computation. Nat Rev Neurosci 2012, 13:51–62.
5. Louie K, Glimcher PW: Efficient coding and the neural representation of value. Ann N Y Acad Sci 2012, 1251:13–32.
6. Rangel A, Clithero JA: Value normalization in decision making: theory and evidence. Curr Opin Neurobiol 2012, 22:970–981.
7. Rangel A, Camerer C, Montague PR: A framework for studying the neurobiology of value-based decision making. Nat Rev Neurosci 2008, 9:545–556.
8. Garcia B, Cerrotti F, Palminteri S: The description–experience gap: a challenge for the neuroeconomics of decision-making under uncertainty. Philos Trans R Soc B Biol Sci 2021, 376:20190665.
9. Hertwig R, Erev I: The description–experience gap in risky choice. Trends Cogn Sci 2009, 13:517–523.
10. Helson H: Adaptation-level theory: an experimental and systematic approach to behavior. New York; 1964.
11. Kahneman D, Tversky A: Prospect Theory: An Analysis of Decision under Risk. Econometrica 1979, 47:263.
12. Ruggeri K, Alí S, Berge ML, Bertoldo G, Bjørndal LD, Cortijos-Bernabeu A, Davison C, Demić E, Esteban-Serna C, Friedemann M, et al.: Replicating patterns of prospect theory for decision under risk. Nat Hum Behav 2020, 4:622–633.
13. Palminteri S, Khamassi M, Joffily M, Coricelli G: Contextual modulation of value signals in reward and punishment learning. Nat Commun 2015, 6:8096.
14. Frank MJ, Seeberger LC, O'Reilly RC: By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. Science 2004, 306:1940–1943.
15. Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD: Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature 2006, 442:1042–1045.
16. Moutoussis M, Bentall RP, Williams J, Dayan P: A temporal difference account of avoidance learning. Netw Comput Neural Syst 2008, 19:137–160.
17. Maia TV: Two-factor theory, the actor-critic model, and conditioned avoidance. Learn Behav 2010, 38:50–67.
18. Mowrer OH: Learning Theory and Behavior. John Wiley & Sons; 1960.
19. Burke CJ, Baddeley M, Tobler PN, Schultz W: Partial Adaptation of Obtained and Observed Value Signals Preserves Information about Gains and Losses. J Neurosci 2016, 36:10016–10025.

20. Klein TA, Ullsperger M, Jocham G: Learning relative values in the striatum induces violations of normative decision making. Nat Commun 2017, 8:16033.

21. Lebreton M, Bacily K, Palminteri S, Engelmann JB: Contextual influence on confidence judgments in human reinforcement learning. PLOS Comput Biol 2019, 15:e1006973.

22. Ting C-C, Palminteri S, Lebreton M, Engelmann J: The elusive effects of incidental anxiety on reinforcement-learning. J Exp Psychol Gen in press, doi:10.31234/osf.io/7d4tc.

23. Parducci A: Happiness, pleasure, and judgment: The contextual theory and its applications. Lawrence Erlbaum Associates, Inc; 1995.

24. Kontek K, Lewandowski M: Range-Dependent Utility. Manag Sci 2017, 64:2812–2832.

25. Tversky A, Kahneman D: Advances in prospect theory: Cumulative representation of uncertainty. J Risk Uncertain 1992, 5:297–323.

26. Bavard S, Lebreton M, Khamassi M, Coricelli G, Palminteri S: Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences. Nat Commun 2018, 9:4503.

27. Bavard S, Rustichini A, Palminteri S: Two sides of the same coin: Beneficial and detrimental consequences of range adaptation in human reinforcement learning. Sci Adv 2021, 7:eabe0340.

28. Luce RD: Individual Choice Behavior: A Theoretical Analysis. Courier Corporation; 2012.

29. Day RH: Rational choice and economic behavior. Theory Decis 1971, 1:229–251.

30. McFadden DL: Rationality for Economists? J Risk Uncertain 1999, 19:73–105.

31. Ludvig EA, Madan CR, McMillan N, Xu Y, Spetch ML: Living near the edge: How extreme outcomes and their neighbors drive risky choice. J Exp Psychol Gen 2018, 147:1905–1918.

32. McNamara JM, Trimmer PC, Houston AI: The ecological rationality of state-dependent valuation. Psychol Rev 2012, 119:114.

33. McNamara JM, Fawcett TW, Houston AI: An Adaptive Response to Uncertainty Generates Positive and Negative Contrast Effects. Science 2013, 340:1084–1086.

34. Fontanesi L, Palminteri S, Lebreton M: Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: a meta-analytical approach using diffusion decision modeling. Cogn Affect Behav Neurosci 2019, 19:490–502.

35. Balleine BW, Daw ND, O'Doherty JP: Chapter 24 - Multiple Forms of Value Learning and the Function of Dopamine. In Neuroeconomics. Edited by Glimcher PW, Camerer CF, Fehr E, Poldrack RA. Academic Press; 2009:367–387.

36. Squire LR: Memory systems of the brain: A brief history and current perspective. Neurobiol Learn Mem 2004, 82:171–177.

37. Lebreton M, Bavard S, Daunizeau J, Palminteri S: Assessing inter-individual differences with task-related functional neuroimaging. Nat Hum Behav 2019, 3:897–905.

38. Cox KM, Kable JW: BOLD Subjective Value Signals Exhibit Robust Range Adaptation. J Neurosci 2014, 34:16533–16543.