Confirmation Bias Is Generalizable Across Pain, Negative Emotion, and Cognitive Effort

Aryan Yazdanpanah (aryan.yazdanpanah.gr@dartmouth.edu)
Heejung Jung (heejung.jung.gr@dartmouth.edu)
Alireza Soltani (alireza.soltani@dartmouth.edu)
Tor D. Wager (tor.d.wager@dartmouth.edu)
Psychological and Brain Sciences, Dartmouth College, 3 Maynard Street
Hanover, NH 03755 USA

Abstract

Expectations have strong influences on perception, cognition, and behavior. With subsequent learning that relies on prediction errors, one can flexibly update the association between expectation and experience, leading to a fine-tuned representation of the experience. However, this update could resist change, due to "confirmation bias", i.e., when learning is strengthened by evidence that supports expectations and attenuated by evidence that contradicts them. Despite prior research on confirmation bias, their shared underlying mechanisms are unclear due to studies focusing on a single domain. To overcome this, we performed a large study on the effects of expectation on somatic pain, vicarious pain, and cognitive effort within the same participants. Using a combination of model-free and model-based approaches, we found evidence for domain-general expectancy effects. Moreover, the confirmation bias within individuals was correlated between somatic pain and cognitive effort and between vicarious pain and cognitive effort. Overall, our results provide evidence for some consistency of confirmation bias and shared mechanisms across cognitive and affective domains.

Keywords: Confirmation bias, Domain-generality, Reinforcement Learning, Expectation

Introduction

Expectations effects are potent enough to change one's subjective experience and is a robust phenomenon across clinical (Bingel et al., 2011; Benedetti et al., 2003) and cognitive domains (Parong, Seitz, Jaeggi, & Green, 2022; Oken et al., 2007). In turn, the experience generates feedback that dynamically changes the expectations (Roy et al., 2014), ideally resulting in fine-tuned representations of the expectations. However, this is not always the case. Learning from evidence that is congruent with expectations is stronger compared to learning from evidence that is incongruent with expectations, leading to resistance to change in the beliefs and the existence of confirmation bias (Doll, Hutchison, & Frank, 2011; Jepma, Koban, van Doorn, Jones, & Wager, 2018; Palminteri, Lefebvre, Kilford, & Blakemore, 2017). Despite recent evidence for confirmation bias in single domains, it is still unknown whether these phenomena rely on similar mechanisms. Using a large-scale experiment with three tasks, conducted within the same participants, here, we explored the question of domain-generality and domain-specificity of confirmation bias across somatic pain, vicarious pain, and cognitive effort experiments.

Method

Participants and experimental design. 101 participants partook in the study spanning three domains: somatic pain ("Pain"), vicarious pain ("Vicarious"), and cognitive effort ("Cognitive"). Each task consisted of 2 cues (high/low) × 3 stimulus intensity levels (low/medium/high) factorial designs, to examine the influence of expectation on sensory perception. Each trial consists of four epochs: cue, expectation rating, stimulus, outcome rating (**Fig. 1**). Initially, participants were shown cues categorized as high or low, with ratings from prior participants depicted as scattered data points ("cue"; **Cue**). Then, participants reported their expectations for the upcoming experience ("expectation ratings"; **E**), followed by the stimulus ("Stimulus"; **S**), and lastly, reported their subjective experience ("Outcome"; **O**). We included participants with at least 24 trials; the intersection across three domains resulted in N=88 for the experiment.

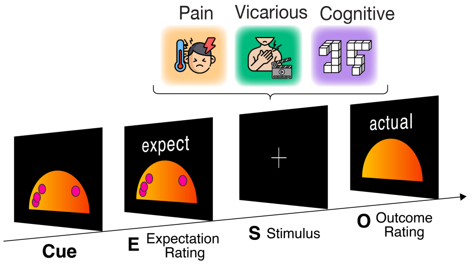

Figure 1: **Experimental design.** Schematic of one trial, identical across three domains of somatic pain, vicarious pain, and a cognitive effort task.

Computational model. The computational model includes three key components: (1) outcome rating, (2) expectation rating, and (3) confirmation bias.

(1) Outcome rating (Pain, Vicarious, Cognitive). In each trial t, the perceived outcome (O) is calculated as the current stimulus combined with the cue-dependent expectation in the current trial $E_{\text{Cue}_i}(t)$ (**Eq. 1**)

$$O(t) = (1-w)\times S(t) + w\times E_{Cue_i}(t) \tag{1}$$

In the model, w is the relative weight of cue-dependent expectation (E) to stimulus intensity (S), and is an adjustable free parameter. Stimulus intensity for each trial is calculated as the average of subjective outcome ratings across different levels of stimulus intensity.

(2) Expectation updates. The update for each cue is performed using the delta rule of the standard RL model (**Eq. 2**).

$$E_{Cue_i}(t+1) = E_{Cue_i}(t) + \alpha \times PE(t) \tag{2}$$

The teaching signal is the outcome rating (**Eq. 3**).

$$PE(t) = O(t) - E_{Cue_i}(t) \tag{3}$$

(3) Confirmation bias in expectation updating. In the presence of confirmation bias updating expectation, learning congruent information should be stronger than learning about incongruent ones, i.e. $\alpha_c > \alpha_i$ (Jepma, Koban, van Doorn, et al., 2018; Palminteri et al., 2017). This is implemented as Equation 4 (**Fig. 2**):

$$\alpha = \begin{cases} \alpha_c & \text{if “PE} > 0 \text{ \& Cue = high” OR “PE} < 0 \text{ \& Cue = low”} \\ \alpha_i & \text{if “PE} < 0 \text{ \& Cue = high” OR “PE} > 0 \text{ \& Cue = low”} \end{cases} \tag{4}$$

For comparison, we fitted models with and without confirmation bias, and a model excluding the learning process. The model with confirmation bias showed a better fit compared to the others based on both AIC and Bayesian Model Selection (BMS) methods.

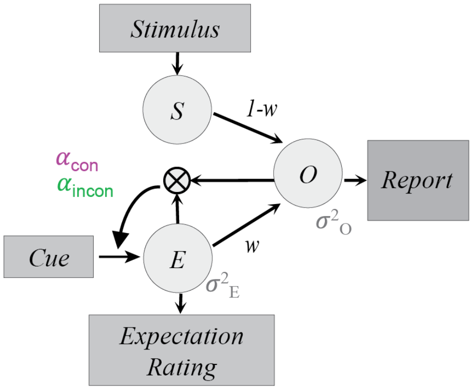

Figure 2: **Confirmation bias model.** The expectation update mechanism with confirmation bias in learning.

Results and Discussion

The effect of stimulus and cue on outcome ratings

Cue and stimulus intensity effects were statistically significant, across three domains. First, outcome ratings were, on average higher, with increasing levels of stimulus intensity (Wilcoxon two-sided signed rank test, Pain: Z = 8.45, p <.001; Vicarious: Z = 8.63, p <.001; Cognitive: Z = 7.85, p <.001.) Next, more importantly, the outcome ratings were significantly higher for high versus low cues across all three domains (Pain: Z = 7.31, p <.001; Vicarious: Z = 8.19, p <.001; Cognitive: Z = 8.07, p <.001; **Fig. 3a-c**).

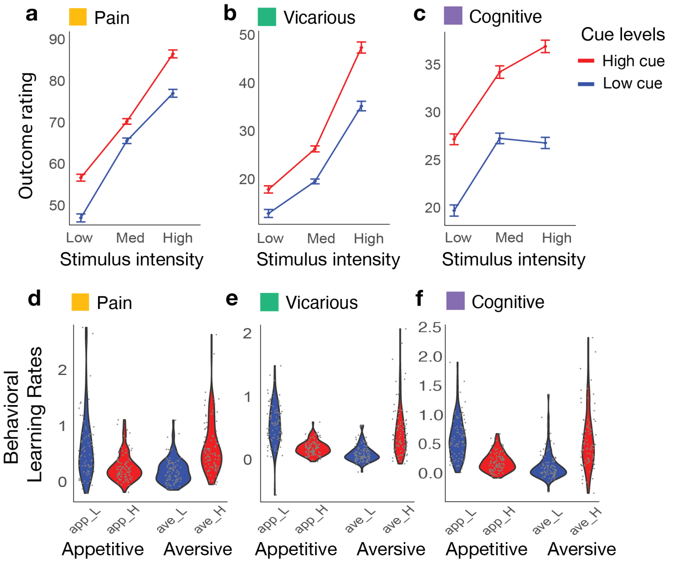

Figure 3: **Cue, stimulus, and confirmation bias effect across three domains. a-c.** Effects of cue and stimulus intensity on outcome rating. **d-f.** Behavioral learning rates in high/low appetitive/aversive trials.

Consistency of the cue and stimulus effects across different domains

Overall, the cue effects were highly consistent between different experiments, (Spearman rank correlation, Pain & Vicarious: r =.54, p <.001; Pain & Cognitive r =.29, p =.0050; Vicarious & Cognitive: r =.43, p <.001).

Confirmation bias in expectation learning in different domains

In both computational models and behavioral effects, confirmation bias was observed across all tasks; learning rates were greater when signed PE aligned with the cue direction, i.e. appetitive PE (PE <0) & low cue trials or aversive PE (PE > 0) & high cue trials (repeated measures ANOVA, Pain: F = 84.69, p <.001; Vicarious: F = 201.80, p <.001; Cognitive: F = 142.53, p <.001; **Fig. 3d-f**).

Consistency of the confirmation bias in expectation learning in different domains

Finally, we investigated whether the confirmation bias originates from the same behavioral mechanisms or not. The confirmation bias in learning was consistent between the Pain & Cognitive (Spearman rank correlation, r = .28, p <.001), and Vicarious & Cognitive (r = .29, p <.001), but not between Pain & Vicarious (r =.15, p =.15).

Conclusion

Overall, we found evidence for the consistent effects of expectation on learning across cognitive and affective domains, pointing to the existence of shared underlying mechanisms. Furthermore, we observed confirmation bias across various domains, displaying some consistent effects, though not uniformly across all domains. These results lay the foundation for further exploration of the neural correlates and substrates of confirmation bias toenrich our understanding of self-reinforcing expectancy effects.

Acknowledgments

This study was funded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB; RO1EB026549) and National Institute of Mental Health (NIMH; R01MH076136), and conducted using the facilities at the Dartmouth Brain Imaging Center (DBIC).

References

Benedetti, F., Pollo, A., Lopiano, L., Lanotte, M., Vighetti, S., & Rainero, I. (2003). Conscious expectation and unconscious conditioning in analgesic, motor, and hormonal placebo/nocebo responses. *The Journal of Neuroscience*, *23*(10), 4315–4323. doi: https://doi.org/10.1523/jneurosci.23-10-04315.2003

Bingel, U., Wanigasekera, V., Wiech, K., Ni Mhuircheartaigh, R., Lee, M. C., Ploner, M., & Tracey, I. (2011). The effect of treatment expectation on drug efficacy: Imaging the analgesic benefit of the opioid remifentanil. *Science Translational Medicine*, *3*(70). doi: https://doi.org/10.1126/scitranslmed.3001244

Doll, B. B., Hutchison, K. E., & Frank, M. J. (2011). Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *The Journal of Neuroscience*, *31*(16), 6188–6198. doi: 10.1523/jneurosci.6486-10.2011

Jepma, M., Koban, L., van Doorn, J., Jones, M., & Wager, T. D. (2018). Behavioural and neural evidence for self-reinforcing expectancy effects on pain. *Nature Human Behaviour*, *2*(11), 838–855. doi: 10.1038/s41562-018-0455-8

Jepma, M., Koban, v. D. J., L., & Wager, T. D. (2018). Behavioural and neural evidence for self-reinforcing expectancy effects on pain. *Nature Human Behavior*, *2*(11), 838–855.

Oken, B. S., Flegal, K., Zajdel, D., Kishiyama, S., Haas, M., & Peters, D. (2007). Expectancy effect: Impact of pill administration on cognitive performance in healthy seniors. *Journal of Clinical and Experimental Neuropsychology*, *30*(1), 7–17. doi: https://doi.org/10.1080/13803390701775428

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLOS Computational Biology*, *13*(8), e1005684. doi: 10.1371/journal.pcbi.1005684

Parong, J., Seitz, A. R., Jaeggi, S. M., & Green, C. S. (2022). Expectation effects in working memory training. *Proceedings of the National Academy of Sciences*, *119*(37). doi: 10.1073/pnas.2209308119

Roy, M., Shohamy, D., Daw, N., Jepma, M., Wimmer, G. E., & Wager, T. D. (2014). Representation of aversive prediction errors in the human periaqueductal gray. *Nature Neuroscience*, *17*(11), 1607–1612. doi: 10.1038/nn.3832