Current Directions in Psychological Science
2024, Vol. 33(2) 93–99

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09637214231217678
www.psychologicalscience.org/CDPS

Hidden Reward: Affect and Its Prediction Errors as Windows Into Subjective Value

Marius C. Vollberg[1,2,3] **and David Sander**[2,3]
[1]Department of Psychology, University of Amsterdam; [2]Swiss Center for Affective Sciences, University of Geneva; and [3]Department of Psychology, FPSE, University of Geneva

Abstract
Scientists increasingly apply concepts from reinforcement learning to affect, but which concepts should apply? And what can their application reveal that we cannot know from directly observable states? An important reinforcement learning concept is the difference between reward expectations and outcomes. Such reward prediction errors have become foundational to research on adaptive behavior in humans, animals, and machines. Owing to historical focus on animal models and observable reward (e.g., food or money), however, relatively little attention has been paid to the fact that humans can additionally report correspondingly expected and experienced affect (e.g., feelings). Reflecting a broader "rise of affectivism," attention has started to shift, revealing explanatory power of expected and experienced feelings—including prediction errors—above and beyond observable reward. We propose that applying concepts from reinforcement learning to affect holds promise for elucidating subjective value. Simultaneously, we urge scientists to test—rather than inherit—concepts that may not apply directly.

Keywords
affect, prediction errors, reinforcement learning, subjective value

Imagine your "friends" dragged you to a casino, and now you find yourself in front of two slot machines. You have heard that some slot machines have higher reward probabilities than others, and, reluctantly, you spin each machine once. You may have started out with equal predictions for both slot machines (e.g., that you will halve your $1 investment). Imagine slot machine A turns your $1 into $0.25, and slot machine B leaves you with $0.75. In other words, you made a prediction error of $0.25 in both cases—to the negative for A and to the positive for B. Next, you decide to go for a third spin—which slot machine do you choose? Presumably your choice will be influenced by your experience on the first two rounds. Of course, your experience is still quite limited at this point, but the prediction errors you have made might slightly incline you toward the slot machine that turned out better than expected and away from the one that turned out worse than expected. This logic of reinforcement learning based on observable reward has proved tremendously useful in understanding and guiding adaptive behavior across animals, humans, and machines.

The value we assign to options is influenced not only by their observable, objective properties but also by the composition of states that is subjective to us. To illustrate this point, imagine the preceding example but from the perspective of someone who has $3 compared with that of someone who has $3 million. The same objective outcome would likely lead to different subjective valuations. In this case, the source of the influence on subjective value—one's prior endowment—is external to the perceiver. But subjective value can also be influenced by factors internal to the perceiver, including the motivational (e.g., wanting to prove that casinos always win) and the incidental (e.g., feeling regretful about your choice of friends). All those factors have in common that they contribute to the transformation of objective into subjective value (Juechems & Summerfield, 2019; Kahneman et al., 1997).

The space of factors contributing to subjective value is vast, which poses a challenge for measurement, and affect may serve as a useful output approximation. Objective value can be reduced to, say, expected and obtained dollar amounts, but how might we capture subjective value? Intriguingly, recent innovations reframe

Corresponding Author:
Marius C. Vollberg, Department of Psychology, University of Amsterdam
Email: mariusvollberg@gmail.com

a means of approximation that we will explore in detail here: verbally reported affect (i.e., feelings) and its prediction errors. (Note that we consider it a reframing rather than a discovery, because feelings have long been studied as indices for individual preferences, and we will merely highlight the potentially generative framing of applying a reinforcement learning lens to affect.)

We integrate across recent developments in psychological science that implicitly or explicitly speak to the relationship between affect and subjective value from a reinforcement learning perspective. We place particular emphasis on the similarities and differences between prediction errors about objective and subjective value. Last, we outline open questions for future research aimed at elucidating subjective value via affect and its prediction errors.

Affect and Subjective Value

Subjective valuation of objective value may differ across perceivers and time, and measures of affect may help us to understand these differences by approximating subjective value. Whether a millionaire devalues slot machine returns, two people derive vastly different levels of pleasure from the same hobby, or a sated animal displays reduced effort to obtain food—the same objective value can be of varied subjective value to different perceivers, who appraise objective value through their idiosyncratic composition of states (Brosch & Sander, 2016; Chew et al., 2021; Juechems & Summerfield, 2019; Sander & Nummenmaa, 2021; Wuensch et al., 2019). Unlike objective value, subjective value cannot be observed directly. However, recent investigations have begun to formalize verbal measures of affect (i.e., feelings) as a proxy for subjective value. These investigations fall under at least two broad categories: verbally reported affect as (a) an outcome and as (b) a predictor of behavior. We will outline both in the following to situate our focus on (b), affect as a predictor of behavior.

One way of probing how affect relates to subjective value is to treat affect as a downstream result of expected and obtained objective reward. Prominent examples of this approach have found, for example, that momentary subjective well-being (i.e., happiness) is proportional to obtained objective reward relative to expectations (i.e., reward prediction errors; Rutledge et al., 2014). That is, all else being equal, the lower people's expectations, the happier they are about an outcome. Moreover, changes in happiness might be better explained by the learning that such prediction errors afford rather than by differences in reward themselves (Blain & Rutledge, 2020). In other words, getting more than expected tends to make us feel better but perhaps more because of what we learn than because of what we gain. Although this research formalizes verbally reported affect as a proxy for subjective valuation of objective reward, it does not focus on the link between affect and future behavior.

Another approach is to treat expected and experienced affect as predictors of future behavior. A subtype of this approach asks whether our choices are better accounted for by the feelings we experience around the time of choice or by the feelings we anticipate experiencing afterward (e.g., Dewall et al., 2016; Jäger et al., 2020; Lerner et al., 2015). Answers to such questions would thus address, for instance, whether your decision to gamble is better explained by your current feelings or by the regret you anticipate feeling after halving your investment. By effectively testing the competition between two different target time points, this approach goes beyond the present scope: the role of affect in the subjective valuation regarding the same time point. Another subtype that also deserves mention as a point of contradistinction is research on "liking," which has historically emphasized "objective affective reactions" (e.g., the facial expression an infant makes when fed sucrose; Berridge & Robinson, 2003, p. 509). Liking has been increasingly subjected to a reinforcement learning lens (Dayan, 2022), too, and is intertwined with subjective affective reactions (E. Pool et al., 2016), which underlie our present focus: verbally reported affect. An example of using verbally reported affective expectations and outcomes to investigate the role of subjective valuation regarding the same target comes from Charpentier and colleagues (2016). Across two tasks, participants provided responses to dichotomous objective reward outcomes (e.g., earn £4 versus £0). In two distinct "feelings" blocks, participants reported either expected or experienced feelings in response to gains or losses, using a continuous scale from *extremely unhappy* to *extremely happy*. In the second task (order randomized), participants could choose on each trial between a dichotomous gamble and a safe option, all involving the amounts shown in the feelings blocks. By separating expected affect, experienced affect, and gambling choices, this design allowed identifying choice variance explained by expected and experienced feelings above and beyond objective monetary reward. Crucially, gambling choices were better predicted by models that included feeling ratings about objective value outcomes than by models that included only objective value outcomes. Moreover, in line with previously documented nonlinearities in value-based decision making (Kahneman & Tversky, 1979), feelings related to losses exerted a stronger influence on choices than feelings related to gains.

Affect can predict behavior beyond what can be predicted by objective value, plausibly because affect partially reflects the subjective valuation of a given objective value. In the context of objective value, behavior has often been effectively predicted using differences between expectations and outcomes, or prediction errors. Knowing about the predictive role of expected and experienced affect, then, is there anything to be gained by predicting behavior using affective prediction errors—the difference between expected and experienced affect?

Affective Prediction Errors

The observation that people make errors when predicting how they will feel in the future is far from new, but those errors have rarely been used to predict future behavior. For decisions varied in size and time horizon, people tend to forecast incorrectly how a given outcome will make them feel (Wilson & Gilbert, 2005). In diverse contexts, such as dormitory assignments, romantic relationships, and health outcomes, individuals anticipate that desirable outcomes will bring them more happiness, and undesirable ones more unhappiness, than they actually end up experiencing. When people forecast their affect incorrectly, they may not update their predictions on the basis of those errors, because people remember their predictions incorrectly as well (Meyvis et al., 2010). Notably, according to this research program, how people end up feeling depends more on what happened than on what they expected to happen (Golub et al., 2009). Although extant research establishes that affective predictions can be inaccurate, it is less clear what affective prediction errors related to current behavior can tell us about how people adapt their behavior going forward.

Prediction errors are deemed central to adaptive behavior in the context of reinforcement learning from objective value, raising questions regarding their role in subjective counterparts. Corroborated by tremendously successful applications to artificial intelligence, the ability to adjust one's model of the world based on reward prediction errors is thought to be essential to natural intelligence. Some even propose that reward is "enough"; that is, it may be at the core of all adaptive behavior across domains (e.g., Silver et al., 2021). Presumably owing to a historic emphasis on model organisms and machines, much of the corresponding research focuses more on objective value than on its subjective counterpart. Unlike model organisms and machines, however, humans can report their feelings, which might serve as a proxy for subjective value. If we entertain this idea, we may ask if affective prediction errors relate to behavior similarly and perhaps in complement to the relationship already observed between objective value prediction errors and behavior.

Reflecting a broader "rise of affectivism" (Dukes et al., 2021), a recent article by Heffner and colleagues (2021) showed that affective prediction errors can explain behavior in complement to objective prediction errors. Across four studies and different economic decision-making games involving monetary offers, participants reported both the offer they expected to receive and their expected affect about the offer. In these studies, affective reports were collected on a two-dimensional, continuous grid reflecting valence (x-axis) and arousal (y-axis). Participants then received unfair offers and reported their actual affect on the same grid before making a choice. Choices in one task were limited to accepting or rejecting the offer, whereas another task additionally allowed participants to reverse the offer (cost-free punishment) or increase their own payout to match that of the proposer (nonpunitive compensation). Independent of task details, this design allows modeling variance in choice behavior using differences between expected and actual offers (objective-value prediction errors) as well as differences between expected and actual affect (proxy for subjective-value prediction errors). Importantly, across these different types of punitive and nonpunitive choices, variance in behavior was predicted by affective-valence prediction errors (i.e., feeling worse or better than expected) above and beyond variance explained by objective-reward prediction errors. Concretely, in models jointly including objective and affective prediction errors, feeling worse than expected about an offer significantly and positively predicted rejection or punishment and offer-reversal rates. This suggests that verbally reported affect can be usefully framed in terms of prediction errors, putatively approximating a component of subjective value that is not captured by objective value alone.

Similar Concepts, Different Rules

There are inherent differences between affective prediction errors and objective-reward prediction errors, which has implications for their relationship to behavior. Prediction errors have been studied in the context of objective reward with such relative depth that it might seem intuitive to transpose those insights onto hypothesized subjective counterparts. However, to apply any approach analogously, the underlying context needs to be sufficiently analogous, too. Is it? To answer this question, we must examine the similarities and differences between the components constituting each type of prediction error as well as their relation

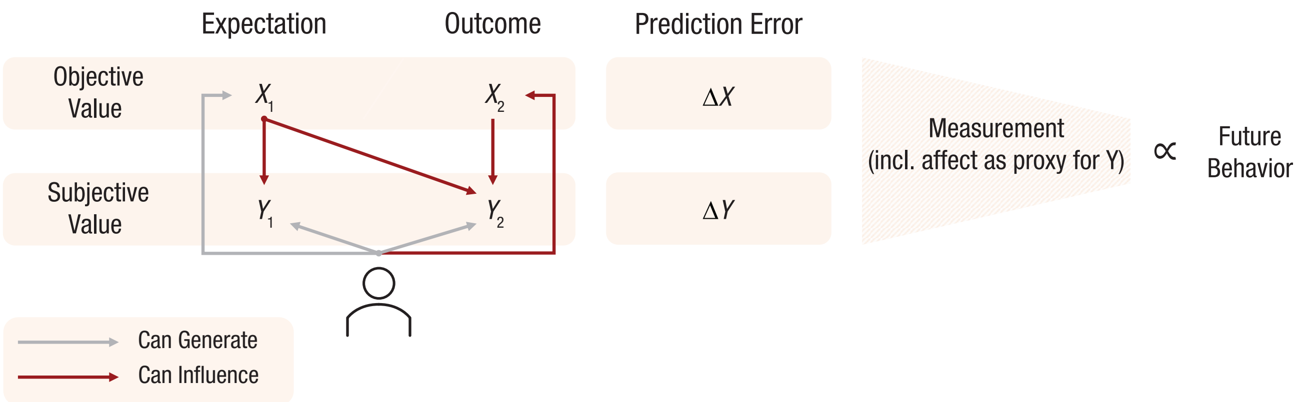

Fig. 1. Schematic illustration of principled differences between objective and subjective prediction errors, both of which are subject to measurement and predictive of behavior. *Objective value* refers to observable states, be it monetary reward, food, or a ballet. We can entirely generate expectations (X_1) about, but usually at best influence (e.g., poker) or trigger (e.g., slot machines), objective outcome states (X_2). *Subjective value* is the result of the appraisal of objective value via idiosyncratic compositions of external and internal states. On the subjective side, we can generate both expectations about how a given objective outcome will be valued subjectively (Y_1) and our actual subjective valuation (Y_2) following the actual outcome. Across expectations and outcomes, subjective value is influenced—but not strictly determined—by objective value. For both objective and subjective value, prediction errors (ΔX and ΔY) are defined as the difference between outcomes and expectations. The constituent parts of prediction errors, except for objective outcome value, are not directly observable and necessitate measurement. In this context, "measurement" primarily pertains to self-report. Whereas the measurement of objective-value expectations is relatively constrained by context (e.g., reported dollar amounts), the measurement of subjective value—approximated via affect—may rely on a wider range of measures, such as feeling ratings. Based on such measurement, both objective and subjective prediction errors about current outcomes have been shown to predict future behavior (e.g., repetition of behaviors that previously led to positive prediction errors).

to behavior (see Fig. 1). Starting with value expectations, both the objective and the subjective kind are generated by the perceiver. For value outcomes, however, the perceiver generates the subjective kind but typically not the objective kind. Although there are contexts allowing for more control than slot machines (which we can merely trigger), we can usually, at best, influence—but not generate—objective-value outcomes. So, what does this mean for the difference between objective and subjective prediction errors? Objective prediction errors consist of one perceiver-generated component (objective-value expectations) and a component that is, at best, perceiver influenced (objective-value outcomes). In contrast, subjective prediction errors consist of two perceiver-generated components (both subjective-value expectations and outcomes). Consequently, the components of subjective prediction errors are causally connected by a shared latent cause (i.e., the perceiver). On the objective side, although reasonable objective-value expectations may correlate with objective-value outcomes, they are not causally connected in this way. On the subjective side, however, this separation between expectations and outcomes is not a logical necessity, because of the shared latent cause. Note that, in addition to the indirect connection via a shared latent cause, there may also be direct influence of subjective-value expectations on subjective-value outcomes. Yet, this influence is difficult to isolate, rendering the shared latent cause a more parsimonious basis for distinction. Put simply, how we subjectively value a slot machine outcome can be connected to how we expected to value it, but the objective payout cannot be connected to our expectations. This difference is important, because prediction errors combine expectations and outcomes, each thought to contain unique information. However, as with any variable, the less separate the components are, the less likely they are to contain unique information relative to one another. Given the separation between objective-value expectations and outcomes, this may not be much of a concern for objective-value prediction errors. In contrast, subjective-value expectations and outcomes are more closely intertwined. This leads to a crucial implication: Subjective-value prediction errors, due to their interlinked components, may not always enhance our understanding beyond either constituent part alone, including when predicting future behavior.

Recent research suggests that although affect may approximate subjective value and predict behavior beyond objective value, outcome affect alone may sometimes be just as predictive as affective prediction errors (Vollberg & Cikara, 2022). Participants could choose between aggressive and nonaggressive behaviors that resulted in personal gain (i.e., earn or steal) or not (i.e., create or destroy), depending on the task used in a given experiment. The actual amount earned, stolen, created, or destroyed always corresponded exactly to participants' choices (i.e., no objective-value prediction errors). What could vary between expectations and outcomes, however, was participants' affect, measured on a continuous, one-dimensional scale from *not at all good* to *extremely good*. Crucially, this variability in affective prediction errors was predictive of choice behavior, such that participants chose increasing amounts of (i.e., escalated) behaviors that had recently felt better than expected. However, especially when looking at coarser, dichotomous measures of aggression, models including only outcome affect provided a better fit than models including affective prediction errors. Although the explanatory power of prediction errors in the context of objective value may lead us to expect otherwise, affective prediction errors may not always add information beyond outcome affect. In sum, there are substantial differences between the composition of prediction errors about subjective and objective value, which should inform how we conceptualize and test their relationship to behavior.

Prediction errors are just one example of the analogies between objective and subjective value whose extent remains in need of investigation. Beyond the constituent parts of their prediction errors, objective and subjective value also differ in the degree to which they are accessible via measurement. Expectations (objective and subjective) and subjective outcome value cannot be observed directly and thus require some form of psychological measurement. Leaving nonverbal measures aside, measurement of objective-value expectations is relatively more constrained (e.g., for a monetary slot machine outcome, we can measure monetary expectations). On the subjective side, however, measurement options are numerous, including, for example, arousal, control, relevance, valence, or specific emotions. Valence may play a unique role in capturing subjective value, as it has been shown to more robustly predict behavior above and beyond objective value compared with arousal (Heffner et al., 2021). Relatedly, nascent evidence from machine vision suggests human valence ratings cannot be approximated as precisely via neural networks as ratings of arousal or beauty (Conwell et al., 2022), likely reflecting processing that transcends linear combinations of objective stimulus features. However, we do not know exactly how subjective value is reflected in affective responses, and appraisal patterns or full-blown emotions may provide additional granularity for understanding patterns in behavior (Heffner & FeldmanHall, 2022; Lerner et al., 2015).

Measures of subjective value may share another potential distinction from those of objective value: introspection dependence. If the slot machine yielded $100, that would probably beat your objective-value expectations, and you might learn from the prediction error in a way that could transfer to other tasks. That is, if the slot machine kept generating positive returns, you would likely be less surprised if other games at that casino do so, too. Notably, this holds whether you are explicitly asked about your expectations or not. Are approximations of subjective value equally introspection independent? Given our focus on verbally reported affect, this question might seem nonsensical: Self-report measures are per definition introspection dependent! However, at least on short time scales, people appear to get better at predicting how they will feel (Vollberg & Cikara, 2022), allowing us to ask whether such reductions in affective prediction errors can still occur after people sampled options *without* reporting their affect. Degrees of introspection dependence of affective measures (e.g., affective prediction errors getting smaller only when people report them) relative to introspection independence of behavior (behavior remaining unchanged whether self-reported affect is measured or not) could help us to better understand convergence and divergence between affective measures and subjective value.

Directions Ahead

By providing a window into subjective value, adequate conceptualization of affect and its prediction errors may prove valuable not just in complement to observable reward but also in its stead. In order to assess whether affect explains variance beyond observable reward, initial investigations necessarily measure both. Future research may expand beyond this point of departure, using affect and its prediction errors to complement our understanding of valuation in contexts where value is less tangible than monetary reward (Lindström et al., 2021; E. R. Pool et al., 2022; Tamir & Mitchell, 2012), encompassing policy decisions, social interactions, and more. The potential breadth of this approach is made evident in a recent study that ventured beyond tangible reward by leveraging affective prediction errors to study wishful thinking (Melnikoff & Strohminger, 2023).

Conclusion

Applying a reinforcement learning lens to affect and probing affective prediction errors hold promise for making subjective value tractable and thus for better understanding behavior. At the same time, charting this territory with parsimony will continue to require inspiration from relevant frameworks (e.g., prediction errors) without presupposing them to map one-to-one.

Recommended Reading

Charpentier, C. J., De Neve, J.-E., Li, X., Roiser, J. P., & Sharot, T. (2016). (See References). Empirical study on expected and experienced affect as separate predictors of choice.

Heffner, J., Son, J.-Y., & FeldmanHall, O. (2021). (See References). Empirical study on the difference between expected and experienced affect as a composite predictor of choice.

Juechems, K., & Summerfield, C. (2019). (See References). Interdisciplinary piece on how objective value gets transformed into subjective value.

Transparency

Action Editor: Robert L. Goldstone
Editor: Robert L. Goldstone
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.
Funding
This work has been supported by a fellowship from the Swiss National Science Foundation awarded to M. C. Vollberg.

ORCID iD

Marius C. Vollberg https://orcid.org/0000-0003-2301-3552

Acknowledgments

We would like to thank Daniel Gilbert and Todd Hare for helpful conversations.

References

Berridge, K. T., & Robinson, T. E. (2003). Parsing reward. *Trends in Neurosciences, 26*(43), 507–513. https://doi.org/10.1016/S0166-2236(03)00233-9

Blain, B., & Rutledge, R. B. (2020). Momentary subjective well-being depends on learning and not reward. *eLife*, *9*, Article e57977. https://doi.org/10.7554/eLife.57977

Brosch, T., & Sander, D. (Eds.). (2016). *Handbook of value: The affective sciences of values and valuation*. Oxford University Press.

Charpentier, C. J., De Neve, J.-E., Li, X., Roiser, J. P., & Sharot, T. (2016). Models of affective decision making: How do feelings predict choice? *Psychological Science*, *27*(6), 763–775. https://doi.org/10.1177/0956797616634654

Chew, B., Blain, B., Dolan, R. J., & Rutledge, R. B. (2021). A neurocomputational model for intrinsic reward. *Journal of Neuroscience*, *41*(43), 8963–8971. https://doi.org/10.1523/JNEUROSCI.0858-20.2021

Conwell, C., Konkle, T., Graham, D., & Vessel, E. A. (2022). *The perceptual primacy of feeling*. https://doi.org/10.31234/osf.io/5wg4s

Dayan, P. (2022). "Liking" as an early and editable draft of long-run affective value. *PLOS Biology*, *20*, Article e3001476. https://doi.org/10.1371/journal.pbio.3001476

Dewall, C. N., Baumeister, R. F., Chester, D. S., & Bushman, B. J. (2016). How often does currently felt emotion predict social behavior and judgment ? A meta-analytic test of two theories. *Emotion Review*, *8*(2), 136–143. https://doi.org/10.1177/175407391557269

Dukes, D., Abrams, K., Adolphs, R., Ahmed, M. E., Beatty, A., Berridge, K. C., Broomhall, S., Brosch, T., Campos, J. J., Clay, Z., Clément, F., Cunningham, W. A., Damasio, A., Damasio, H., D'Arms, J., Davidson, J. W., de Gelder, B., Deonna, J., de Sousa, R., . . . Sander, D. (2021). The rise of affectivism. *Nature Human Behaviour*, *5*(7), 816–820. https://doi.org/10.1038/s41562-021-01130-8

Golub, S. A., Gilbert, D. T., & Wilson, T. D. (2009). Anticipating one's troubles: The costs and benefits of negative expectations. *Emotion*, *9*(2), 277–281. https://doi.org/10.1037/a0014716

Heffner, J., & FeldmanHall, O. (2022). A probabilistic map of emotional experiences during competitive social interactions. *Nature Communications*, *13*, Article 1718. https://doi.org/10.1038/s41467-022-29372-8

Heffner, J., Son, J.-Y., & FeldmanHall, O. (2021). Emotion prediction errors guide socially adaptive behavior. *Nature Human Behavior*, *5*, 1391–1401. https://doi.org/10.1038/s41562-021-01213-6

Jäger, D. T., Boltzmann, M., Rollnik, J. D., & Rüsseler, J. (2020). Expected valence predicts choice in a recurrent decision task. *Frontiers in Psychology*, *14*, Article 580970. https://doi.org/10.3389/fnins.2020.580970

Juechems, K., & Summerfield, C. (2019). Where does value come from? *Trends in Cognitive Sciences*, *23*(10), 836–850. https://doi.org/10.1016/j.tics.2019.07.012

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–292.

Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *Quarterly Journal of Economics*, *112*(2), 375–405.

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology*, *66*, 799–823. https://doi.org/10.1146/annurev-psych-010213-115043

Lindström, B., Bellander, M., Schultner, D. T., Chang, A., Tobler, P. N., & Amodio, D. M. (2021). A computational reward learning account of social media engagement. *Nature Communications*, *12*, Article 1311. https://doi.org/10.1038/s41467-020-19607-x

Melnikoff, D., & Strohminger, N. (2023). *Bayesianism and wishful thinking are compatible*. https://doi.org/10.31234/osf.io/yhmvw

Meyvis, T., Ratner, R. K., & Levav, J. (2010). Why don't we learn to accurately forecast feelings? How misremembering our predictions blinds us to past forecasting errors. *Journal of Experimental Psychology: General*, *139*(4), 579–589. https://doi.org/10.1037/a0020285

Pool, E., Sennwald, V., Delplanque, S., Brosch, T., & Sander, D. (2016). Measuring wanting and liking from animals to humans: A systematic review. *Neuroscience and Biobehavioral Reviews*, *63*, 124–142. https://doi.org/10.1016/j.neubiorev.2016.01.006

Pool, E. R., Tord, D. M., Delplanque, S., Stussi, Y., Cereghetti, D., Vuilleumier, P., Sander, D. (2022). Differential contributions of ventral striatum subregions to the motivational and hedonic components of the affective processing of reward. *Journal of Neuroscience, 42*(13), 2716–2728. https://doi.org/10.1523/JNEUROSCI.1124-21.2022

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences, USA*, *111*(33), 12252–12257. https://doi.org/10.1073/pnas.1407535111

Sander, D., & Nummenmaa, L. (2021). Reward and emotion: An affective neuroscience approach. *Current Opinion in Behavioral Sciences*, *39*, 161–167. https://doi.org/10.1016/j.cobeha.2021.03.016

Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, *299*, 103535. https://doi.org/10.1016/j.artint.2021.103535

Tamir, D. I., & Mitchell, J. P. (2012). Disclosing information about the self is intrinsically rewarding. *Proceedings of the National Academy of Sciences, USA*, *109*(21), 8038–8043. https://doi.org/10.1073/pnas.1202129109

Vollberg, M. C., & Cikara, M. (2022). *Affective prediction errors in persistence and escalation of aggression*. https://doi.org/10.31234/osf.io/59k3c

Wilson, T. D., & Gilbert, D. T. (2005). Affective forecasting: Knowing what to want. *Current Directions in Psychological Science*, *14*(3), 131–134. https://doi.org/10.1111/j.0963-7214.2005.00355.x

Wuensch, L., Pool, E. R., & Sander, D. (2019). Individual differences in learning positive affective value. *Current Opinion in Behavioral Sciences*, *39*, 19–26. https://doi.org/10.1016/j.cobeha.2020.11.001