frontiers in Human Neuroscience

PERSPECTIVE
published: 24 March 2022
doi: 10.3389/fnhum.2022.824085

The Role of Predictions, Their Confirmation, and Reward in Maintaining the Self-Concept

Aviv Mokady[1]* and *Niv Reggev*[1,2]*

[1]Department of Psychology, Ben-Gurion University of the Negev, Be'er Sheva, Israel, [2]Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Be'er Sheva, Israel

The predictive processing framework posits that people continuously use predictive principles when interacting with, learning from, and interpreting their surroundings. Here, we suggest that the same framework may help explain how people process self-relevant knowledge and maintain a stable and positive self-concept. Specifically, we recast two prominent self-relevant motivations, self-verification and self-enhancement, in predictive processing (PP) terms. We suggest that these self-relevant motivations interact with the self-concept (i.e., priors) to create strong predictions. These predictions, in turn, influence how people interpret information about themselves. In particular, we argue that these strong self-relevant predictions dictate how prediction error, the deviation from the original prediction, is processed. In contrast to many implementations of the PP framework, we suggest that predictions and priors emanating from stable constructs (such as the self-concept) cultivate belief-maintaining, rather than belief-updating, dynamics. Based on recent findings, we also postulate that evidence supporting a predicted model of the self (or interpreted as such) triggers subjective reward responses, potentially reinforcing existing beliefs. Characterizing the role of rewards in self-belief maintenance and reframing self-relevant motivations and rewards in predictive processing terms offers novel insights into how the self is maintained in neurotypical adults, as well as in pathological populations, potentially pointing to therapeutic implications.

Keywords: predictive processing, belief maintenance, self-concept, motivations, reward, self-verification, self-enhancement

OPEN ACCESS

Edited by:
Anila Maria D'Mello,
Massachusetts Institute of Technology, United States

Reviewed by:
Jacob Elder,
University of California, Riverside, United States

****Correspondence:***
Aviv Mokady
avivmok@post.bgu.ac.il
Niv Reggev
reggevn@bgu.ac.il

Specialty section:
This article was submitted to Cognitive Neuroscience, a section of the journal Frontiers in Human Neuroscience

Received: 28 November 2021
Accepted: 01 March 2022
Published: 24 March 2022

Citation:
Mokady A and Reggev N (2022) The Role of Predictions, Their Confirmation, and Reward in Maintaining the Self-Concept. Front. Hum. Neurosci. 16:824085. doi: 10.3389/fnhum.2022.824085

INTRODUCTION

Predictive processing is a theoretical framework for understanding the principles guiding human behavior, as illustrated in this special issue (Clark, 2013; Hohwy, 2013; Ueda et al., 2021). The predictive processing (PP) framework posits that people constantly create predictions about the sensory and interoceptive inputs they expect to receive to facilitate their perception of their surroundings. These predictions are then set against the actual input received from the world to create a prediction error (PE), defined as the difference between the predicted and received information. Common PP interpretations assert that perceivers strive to minimize PE to facilitate fluent interaction with their surroundings (Gilead et al., 2020; Hohwy, 2020). To minimize PE, people usually employ one of two methods. The first and more common application of PP principles involves updating the prior beliefs driving the prediction, thus improving the correspondence between future predictions and reality (e.g., Friston et al., 2009; Nassar et al., 2010; Sharot and Garrett, 2016; Vlasceanu et al., 2021; Elder et al., 2021). The second method involves changing the way people perceive reality ("active inference" in PP terms; Friston, 2010; Hohwy, 2020; Yon et al., 2021), for example by reinterpreting incoming inputs to better align with their predictions (e.g., motivated reasoning; Kunda, 1990; Epley and Gilovich, 2016). To date, most PP theories have focused on the belief-updating process, its effects, and PE minimization (e.g., Rao and Ballard, 1999; Friston et al., 2009; Griffiths et al., 2018; Elder et al., 2021; Kube et al., 2021). However, only a handful of contributions have examined how PP can lead to belief-maintaining processes of stable constructs (see, for instance, Gershman and Cikara, 2021). This perspective manuscript suggests that predictive processing principles guide the maintenance of a stable self-concept, whether positive in neurotypical adults or negative in specific pathological populations. To do so, we examine how the PP framework extends prior notions of self-relevant motivations with a novel emphasis on the role of subjective rewards in self-belief maintenance when no PE ensues.

Predictive Processing and Self-Related Cognition

Principles of PP can explain several stable aspects of personality and social functions that mediate the maintenance of beliefs central to oneself (Yon et al., 2019). For instance, hard-to-falsify religious or supernatural beliefs are often held with very high precision (i.e., individuals attribute substantial weight to these prior beliefs; e.g., Harris and Corriveau, 2021). Such high precision, in turn, leads individuals to interpret inputs in ways that maintain these prior beliefs (i.e., to engage in PE-minimization *via* active inference; van Elk and Aleman, 2017; Gershman, 2019). Similar effects were demonstrated for stereotypical beliefs, whereby seeing an individual who conforms to the stereotype (triggering a small PE) strengthens prior convictions. In contrast, an individual diverging from the stereotype (generating a large PE) can be categorized into a "subtype," i.e., a member of a subcategory with distinctive features. This subcategory then prevents changing the primary category's parameters, thus avoiding changing prior beliefs (Kunda and Oleson, 1995; Gershman, 2019; Westra, 2019; Gershman and Cikara, 2021). In the domain of the self, Hohwy and Michael (2017) suggested that people perceive and maintain their self-hood by an internal model of hierarchical endogenous (hidden) causes. The interaction between high-level causes such as desires or long-term goals and low-level causes such as actions generates top-down predictions and minimizes bottom-up PEs. Finally, Moutoussis et al. (2014) go even further to suggest that the predictive brain and Bayesian inference together shape how people understand their own self-concept by taking actions that will most probably fulfill goals of desired self-representations. These studies suggest that, in cases of high-precision beliefs, individuals minimize PE by shaping reality or reinterpreting new inputs to match and maintain these beliefs.

Notably, most implementations of the PP framework to self-relevant judgments have focused either on characterizing the mechanisms supporting belief-updating or on the actions people take to pre-emptively minimize PE (e.g., Moutoussis et al., 2014; Sharot and Garrett, 2016; Kube and Rozenkrantz, 2021). From an evolutionary standpoint, to ensure optimal adaptation to local surroundings, people should indeed be motivated to minimize PE by updating their beliefs in general and their beliefs about themselves in particular (Okasha, 2013). In the current article, we propose that belief-updating is but one of several approaches people employ when encountering self-relevant information. Specifically, we suggest that self-relevant belief maintenance plays an equally important role in shaping one's cognitions, feelings, and motivations. In the following sections, we briefly describe previous conceptualizations of self-related constructs and explore how a PP framework can apply to processes involving these constructs. We then characterize the role reward plays in maintaining self-relevant beliefs using PP principles. We conclude by discussing the implication of our framework.

Conceptualization of the Self-Concept and Supporting Motivations

To apply PP principles to the self-concept, we first need to understand its nature and predictive features. Epstein (1973) suggested that individuals continuously gather self-relevant information to gradually construct a "self-theory," or an inner model from which people make their predictions about themselves. Individuals gather such self-relevant information from various sources, including social (Cooley, 1902; Mead, 1934) and personal (Bem, 1972) cues. Thus, before establishing a stable self-concept, people update their beliefs about themselves according to inputs from their surroundings. Once the theory of the self is consolidated, individuals gradually assign increasing weights to their self-concept to facilitate its maintenance at the expense of continuous updating. Inner models (priors) such as the self-concept and the predictions they make affect, in turn, how people interact with the world and how such interactions affect their previous priors (for a review see Briñol and Petty, 2021). The priors and their precision come into play, for instance, when credible or non-credible sources validate (or invalidate) previous beliefs (Tormala et al., 2006) or when predicting other's actions (for a review see Bach and Schenke, 2017). Priors also affect the judged fit of a message with one's goal and situation (Cesario et al., 2004) or the potential fit of a decision with one's (cued) identity (Oyserman, 2009; Oyserman et al., 2012). In this sense, the nature of self can be seen as a generative model consisting of beliefs (priors) and predictions that, in turn, interact with the world (Hohwy and Michael, 2017; Van de Cruys and Van Dessel, 2021).

Past research has characterized several motivations that govern self-relevant beliefs. According to self-verification, people strive to experience their surroundings and interactions as confirming their self-concepts, thus maintaining their beliefs about themselves (Swann, 2011), in line with a general need for cognitive consistency (Kruglanski et al., 2018). Complementarily, self-enhancement motivates people to pursue positive self-evaluations under the umbrella of people's general endeavor to feel good about themselves (Taylor and Brown, 1988). Self-enhancement theory suggests that people over estimate their abilities, positive attributes, general self-worth (the above-average effect; Sedikides and Gregg, 2008), as well as the likelihood of positive future events (Sharot, 2011). To satisfy both motivations, people employ various behavioral and interpretive strategies before, during, and after interactions (Swann and Read, 1981). People opt to interact with partners that will provide feedback satisfying their self-motivation (Swann and Read, 1981; Swann et al., 1989, 1994; Burke and Stets, 1999), and, when possible, choose to receive positive information (Sedikides, 1993; Charpentier et al., 2018). Additionally, self-enhancement leads individuals to attribute positive outcomes to the self and negative consequences to external factors, such as other people or situational circumstances (the fundamental attribution error; Ross et al., 1977; Ross, 2018). Accordingly, people allocate more attention to and better recall information according to their motivations (self-verifying; Swann and Read, 1981; Maheshwari et al., 2021; self-enhancing; Sedikides and Green, 2000). In sum, people use coalescing mechanisms to align future inputs (and the interpretations of these inputs) with self-relevant predictions, both for predictions aligned with the self-concept and for more positive predictions.

The Impact of Self-Relevant Motivations on Predictive Processing

As highlighted above, individuals constantly rely on their self-concept (i.e., their prior) to predict the inputs and responses of their environment. We suggest that the self-verifying and self-enhancing motivations heavily impact these predictions and the weight given to information pertinent to these predictions (i.e., the input from the world—the evidence; see **Figure 1**). For example, when the motivation to self-verify dominates, predictions will be consistent with the self-concept; this will be the case even if the relevant self-related features are negative. In contrast, when self-enhancement motivation prevails, predictions will entail evaluations that are more positive and optimistic than the self-concept. When individuals engage in self-relevant interactions, these predictions are compared with incoming inputs. A matching input—self-consistent information for a self-verifying prediction or an input more positive than the self-concept for a self-enhancing prediction—triggers a minimal PE; the information complements prior predictions and thus increases the precision of future predictions. Furthermore, as we elaborate below, we suggest that information congruent with self-relevant predictions triggers a reward response, thus strengthening the strive to minimize PE in the future (see **Figure 1**).

However, many inputs from one's environment involve information that mismatches the self-verifying or self-enhancing predictions, thus triggering a PE. Subsequently, perceivers attempt to minimize the PE *via* one of three possible types of reactions. One type of PE minimization prominent in self-verification (Swann, 2011) and self-enhancement theories (Taylor and Brown, 1988) involves maintaining self-beliefs by altering the perception of reality to conform to self-relevant predictions (i.e., active inference). As the self-theory builds on a lifetime of accumulated evidence, the precision of the predictions it creates is typically much higher than that of a new input incongruent with these predictions. Therefore, to explain such incongruent inputs, perceivers employ interpretive strategies to *post hoc* explain how such inputs cohere with the predictions. These strategies include, for instance, developing auxiliary hypotheses (Gershman, 2019) or attributing the input to external situational circumstances that do not call for a model update (e.g., the fundamental attribution error; Campbell and Sedikides, 1999; Ross, 2018). Other PE minimization options involve updating self-beliefs to match the information [in line with Festinger's cognitive dissonance theory (Festinger, 1957)] or dismissing and ignoring the mismatching event altogether. However, both options are significantly less likely than active inference. To navigate their lives, people need a stable self-concept; frequent updates will result in inefficient use of resources and a high potential for erroneous updates. Similarly, people are unlikely to completely ignore information unless the input is very unlikely (e.g., telling a tall person she is short; Kube et al., 2021). Thus, although all these strategies lead to the end goal of minimizing PE, perceivers are more likely to engage in active inference to solve the PE and maintain prior knowledge in the case of stable constructs such as the self.

The Reward Value of Confirming Self-Relevant Predictions

Strategies that employ PE minimization can explain many human behaviors, including self-relevant belief maintenance (see, for example, Moutoussis et al., 2014). However, frameworks that emphasize how individuals reckon with PE often overlook what happens when inputs match predictions—when no PE ensues. Typical PP frameworks implicitly assume that when PE equals zero, no action takes place. In contrast, both recent and classic theoretical accounts suggest that individuals strive to maintain a self-consistent worldview (Thorndike, 1911; Festinger, 1957; Theriault et al., 2021)—i.e., to actively keep their PE minimized. For example, people prefer to feel emotions that will maintain a desired (therefore predicted) state, serving a long-term goal (e.g., standing your ground), even if that means feeling an unpleasant emotion (e.g., anger; Millgram et al., 2015; Tamir et al., 2017). Furthermore, people dislike individuals succeeding in stereotype-incongruent roles, both for gender stereotypes (Heilman et al., 2004; Rudman et al., 2012; Moss-Racusin and Johnson, 2016) and ethnic stereotypes (Mendes et al., 2007), leading to negative social interactions and evaluations.

Motivation to maintain a self-consistent perception of the world should go hand-in-hand with the motivation to achieve the goal of a minimal PE. Drawing on classic literature, obtaining a goals hould generate a subjective reward response (Reiss, 2004; Bromberg-Martin and Sharot, 2020). Numerous studies have shown that an event that satisfies an organism's goal triggers subjective feelings of reward as well as activity in neural structures associated with a reward response such as the ventral striatum and the medial prefrontal cortex (O'Doherty, 2004; Delgado, 2007). Goal achievement triggers a reward response for survival-related goals, such as nutrition and reproduction,

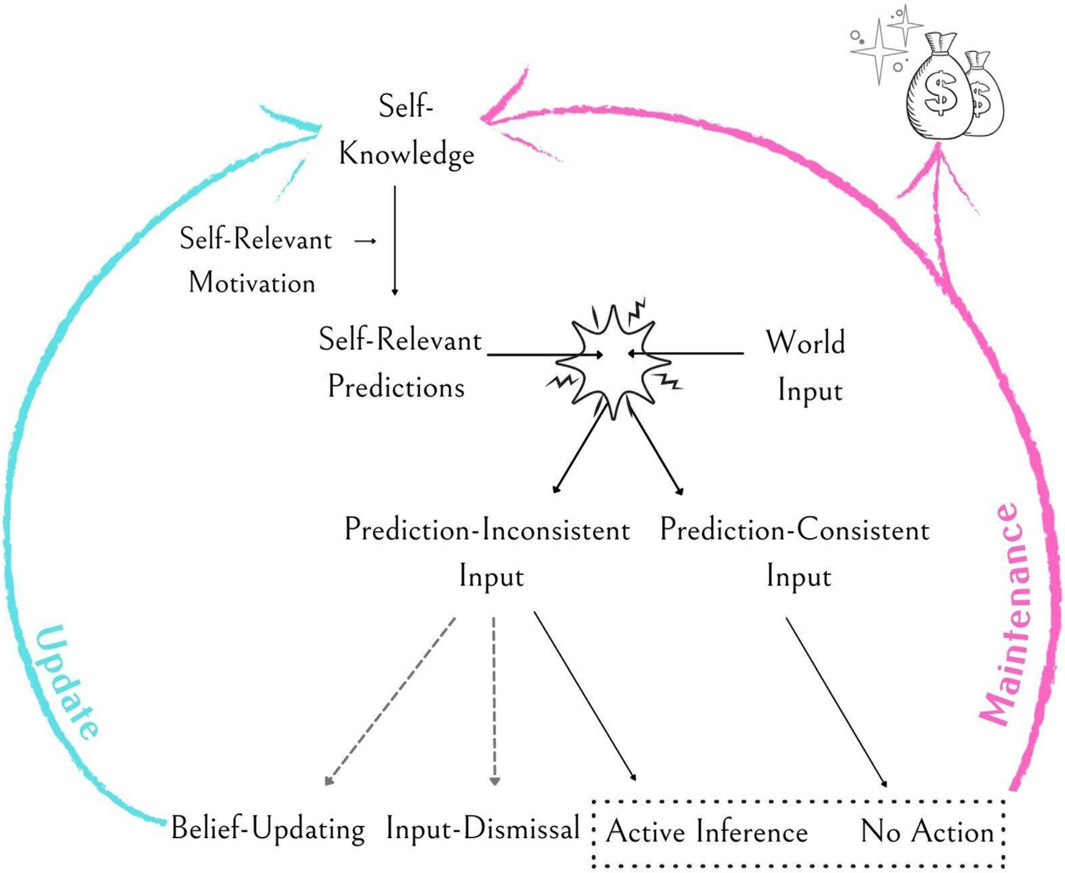

FIGURE 1 | Illustration of the proposed model. Self-Knowledge serves as an inner model of the self (i.e., prior) from which predictions are derived. The (context-dependent) dominant self-relevant motivation, be it self-verification, self-enhancement, or another form of motivation, mediates these predictions. These predictions are compared with inputs from the world, leading to a match or a mismatch. The outcomes of these comparisons lead to one of several possible outcomes. The less likely outcomes (indicated by broken lines in the figure) include model updating (in the unlikely case of belief-updating following prediction-inconsistent information) and input-dismissal (when receiving information that seems extremely irrelevant to the self). More likely outcomes (indicated by solid lines) include self-knowledge maintenance *via* active inference or prediction-consistent input. The more likely outcomes (grouped by the broken lined box) also generate a reward response (money bag), thus further increasing their likelihood to be used in the future.

and higher-level goals, such as securing resources for future generations and subjective well-being. If maintaining beliefs supporting a predicted world is one's goal, then the individual holding the beliefs should experience reward whenever they obtain that goal, even when no PE minimization is required. The rewarding experience, in turn, should create a cascade of downstream effects that subsequently reinforces the goal of belief consistency.

Several recent neural and behavioral studies provide initial support for the rewarding aspect of prediction confirmation across domains. Sensory input (in the form of musical sounds) that matches the predicted information triggers an intrinsic reward response (Salimpoor et al., 2015). Similarly, observing targets that match people's stereotypical predictions is also rewarding (Reggev et al., 2021). From a broader perspective, perceiving information congruent with people's beliefs, which makes their beliefs more certain, is also rewarding (Bromberg-Martin and Sharot, 2020). These findings suggest that minimal PE (i.e., prediction congruent input) trigger a reward response, highlighting the goal of consistency with predictions. In turn, such reward responses could explain why individuals strive to engage with inputs that have minimal PE, over and above traditional explanations such as fluency (i.e., the ease of processing information; see, for example, Kahl and Kopp, 2017) and free energy (i.e., minimizing "surprise"; Friston, 2010).

If perceiving a minimal-PE input is rewarding in the sensory, social, and cognitive domains, we hypothesize that similar effects should be evident for self-relevant beliefs. Specifically, we suggest that perceiving inputs that produce minimal-PE with self-relevant predictions trigger a subjectively rewarding response. Importantly, we suggest that this reward response can occur regardless of the nature of the specific prediction; a positive input matching a self-enhancing prediction will be as rewarding as a self-consistent input matching a self-verifying prediction. Importantly, we suggest that the reward response for a self-consistent input will occur even if the prediction and its corresponding input are negative. Initial neural findings suggest that self-enhancing processes indeed correlate with reward-related brain regions such as the ventral striatum (Parrish et al., 2022). A similar confirming-related reward can ensue when individuals successfully reinterpret information that initially does not match the prediction, as reinterpreted inputs often conform to the respective predictions. Together, the current framework portrays a vital role for reward and consistency-motivation in people's tendency to engage with prediction-consistent information and to minimize PE *via* belief-maintaining rather than belief-updating. This role complements the classic PP interpretations that posit that a minimal PE is the end goal of active inference, and thus once it is reached, no more interaction with the stimuli is needed. Additionally, a related reward response may explain why people continue to interact with such information even after the PE is minimized and the prior beliefs and predictions stabilize. Supporting this notion, the value of a reward generated after reaching a desired state has been recently shown to gradually habituate and drive people to actively maintain that desired state to "regenerate" the reward (Dubey et al., 2021). In the context of the current manuscript, after reaching the desired state—a stable self-concept and a predicted world—and experiencing reward caused by prediction confirmation, habituation can kick in and drive people to regenerate reward by reconfirming predictions and thus triggering a reinforcing cycle.

Implications

Applying reward-oriented processing and PP principles to the motivation to confirm self-relevant predictions offers several exciting implications. First, such an account explains the impact of self-esteem on the balance between self-verification and self-enhancing (Swann and Read, 1981; Taylor and Brown, 1988; Sedikides, 1993; Swann, 2011). People with higher self-esteem tend to self-enhance more frequently and generally expect more positive feedback than people with low self-esteem (Hepper et al., 2011). In contrast, people with low self-esteem experience conflict between wanting to receive positive evaluations (i.e., self-enhancement motivation) and striving for accurate assessments (i.e., self-verification motivation). Typically, such people tend to form accurate-negative predictions and evaluations of essential elements of the self (Ronde and Swann, 1993). Interestingly, as low self-esteem is strongly related to depression (Sowislo and Orth, 2013), people suffering from depression may demonstrate an exacerbated tendency to form negative self-verifying predictions. Recasting previous analyses of depressive automatic thoughts and expectations (Beck, 1963, 1976) in terms of our framework, we go further to suggest that individuals in a state of depression seek inputs consistent with their negative predictions because such inputs trigger rewarding experiences. The reward experience, in turn, may initiate a continuous reinforcement cycle that perpetuates depression and the related phenomenon of depressive realism (Alloy and Abramson, 1979; Moore and Fresco, 2012). If correct, this analysis implies that depression treatment may benefit from examining the reinforcing connection between subjective rewards and evaluations that verify negative predictions or the diminished rewards for positive self-evaluations (see also Van de Cruys and Van Dessel, 2021).

Another field of study that may benefit from these ideas is the study of social anxiety disorder (SAD), a disorder characterized as an ongoing fear of scrutiny by others that will lead to a negative evaluation (Heimberg et al., 2010; Kashdan et al., 2013). At the center of the Cognitive Behavioral Model of Social Anxiety Disorder (Heimberg et al., 2014) lies a "mental representation of the self as seen by others," i.e., priors based on the self-concept generating predictions of evaluations by others. Indeed, people with SAD see themselves more negatively in social situations and expect interaction partners to see them as such (Kashdan and Savostyanova, 2011). In social interaction, people with SAD allocate attention toward evidence of being evaluated and their flaws (Heimberg et al., 2014) and try to conceal their anxiety by suppressing emotional expression (Butler et al., 2003; Kashdan et al., 2013). For people with SAD, these and additional mechanisms (for a review see: Hofmann, 2007) manifest in most social situations, hindering interactions and contributing to the confirmation of predictions of being perceived negatively. As individuals with SAD keep employing such strategies despite recurrent failures, we propose that the motivation to confirm their predictions (and its associated reward) plays an essential role in maintaining the social anxiety cycle and its underlying negative social self-concept (Van de Cruys and Van Dessel, 2021).

Self-concept maintaining within the PP framework may also explain other social phenomena in which a person has an actively maintained negative or pathological self-concept. For instance, system justification theory (Jost, 2019) posits that socio-economically underprivileged people justify systematic obstacles that maintain the existing social hierarchy and thus their underprivileged position. Our account suggests that the person internalizes her group affiliation as part of her self-concept (similar to Stereotype Threat Theory; Steele, 1992, 1997; Steele and Aronson, 1995). To date, SJT was explained by several joint mechanisms, including the ego, intergroup conflict, and status quo rationalization (Jost et al., 2004). We suggest a more parsimonious explanation building on the predictive self-concept. A reward response to self-concept verification that reifies such underprivileged situations can explain why people act to maintain their current social status, even when it is a disadvantageous one. More broadly, the process of maintaining one's self-concept *via* the mechanism we suggest here can be applied to many social contexts in which individuals co-opt societal circumstances into their self-concept.

CONCLUSION

This perspective exemplifies how the PP framework can be applied to understand the self-concept, emphasizing self-concept maintenance rather than updating. Building on past studies of reward and consistency motives, we suggest that the motivation to maintain rather than update the self-concept manifests as (and is reinforced by) a reward in response to prediction-congruent evidence. Understanding the motives and mechanisms underlying how people perceive themselves may shed new light on behavioral research regarding the self-concept, its development and maintenance, and how it shapes people's interaction with their surroundings. In addition, it could be key for understanding and changing behaviors characterizing individuals with negative self-views or psychopathological conditions.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

AM and NR have conceived the conceptual framework. AM drafted the first version of the manuscript. AM and NR have prepared and edited the manuscript and have approved its final version. All authors contributed to the article and approved the submitted version.

FUNDING

This work was partially supported by the Israel Science Foundation (grant No. 540/20, to NR).

ACKNOWLEDGMENTS

We like to thank Dr. Michael Gilead, Prof. Gideon Anholt, Dr. Elena Stephan, Mr. Inon Raz, Ms. Hadas Wardi, and to the reviewer for their help in discussing and refining earlier versions of this manuscript.

REFERENCES

Alloy, L. B., and Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: sadder but wiser? *J. Exp. Psychol. Gen.* 108, 441–485. doi: 10.1037//0096-3445.108.4.441

Bach, P., and Schenke, K. C. (2017). Predictive social perception: towards a unifying framework from action observation to person knowledge. *Soc. Personal. Psychol. Compass* 11:e12312. doi: 10.1111/spc3.12312

Beck, A. T. (1963). Thinking and depression: I. idiosyncratic content and cognitive distortions. *Arch. Gen. Psychiatry* 9, 324–333. doi: 10.1001/archpsyc.1963.01720160014002

Beck, A. T. (1976). *Cognitive Therapy and the Emotional Disorders.* New York, NY: International Universities Press.

Bem, D. J. (1972). Self-perception theory. *Adv. Exp. Soc. Psychol.* 6, 1–62. doi: 10.1016/S0065-2601(08)60024-6

Briñol, P., and Petty, R. E. (2021). Self-validation theory: an integrative framework for understanding when thoughts become consequential. *Psychol. Rev.* doi: 10.1037/rev0000340. [Online ahead of print].

Bromberg-Martin, E. S., and Sharot, T. (2020). The value of beliefs. *Neuron* 106, 561–565. doi: 10.1016/j.neuron.2020.05.001

Burke, P. J., and Stets, J. E. (1999). Trust and commitment through self-verification. *Soc. Psychol. Q.* 62, 347–366. doi: 10.2307/2695833

Butler, E. A., Egloff, B., Wilhelm, F. H., Smith, N. C., Erickson, E. A., and Gross, J. J. (2003). The social consequences of expressive suppression. *Emotion* 3, 48–67. doi: 10.1037/1528-3542.3.1.48

Campbell, W. K., and Sedikides, C. (1999). Self-threat magnifies the self-serving bias: a meta-analytic integration. *Rev. Gen. Psychol.* 3, 23–43. doi: 10.1037/1089-2680.3.1.23

Cesario, J., Grant, H., and Higgins, E. T. (2004). Regulatory fit and persuasion: transfer from "Feeling Right". *J. Pers. Soc. Psychol.* 86, 388–404. doi: 10.1037/0022-3514.86.3.388

Charpentier, C. J., Bromberg-Martin, E. S., and Sharot, T. (2018). Valuation of knowledge and ignorance in mesolimbic reward circuitry. *Proc. Natl. Acad. Sci. U S A* 115, E7255–E7264. doi: 10.1073/pnas.1800547115

Clark, A. (2013). Whatever next? Predictive brains, situated agents and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204. doi: 10.1017/S0140525X12000477

Cooley, C. H. (1902). *Human Nature and the Social Order.* New York, NY: Charles Scribner's Sons.

Delgado, M. R. (2007). Reward-related responses in the human striatum. *Ann. N. Y. Acad. Sci.* 1104, 70–88. doi: 10.1196/annals.1390.002

Dubey, R., Griffiths, T., and Dayan, P. (2021). *Why it is Hard to be Happy With What We Have: A Reinforcement Learning Perspective. PsyArXiv* [Preprint]. doi: 10.31234/OSF.IO/8JD2X

Elder, J., Davis, T., and Hughes, B. (2021). Learning about the self: motives for coherence and positivity constrain learning from self-relevant feedback. *PsyArXiv* [Preprint]. doi: 10.31234/OSF.IO/ETZ56

Epley, N., and Gilovich, T. (2016). The mechanics of motivated reasoning. *J. Econ. Perspect.* 30, 133–140. doi: 10.1257/jep.30.3.133

Epstein, S. (1973). The self-concept revisited. Or a theory of a theory. *Am. Psychol.* 28, 404–416. doi: 10.1037/h0034679

Festinger, L. (1957). *A Theory of Cognitive Dissonance.* Stanford, CA: Stanford University Press.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. J., Daunizeau, J., and Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLoS One* 4:e6421. doi: 10.1371/journal.pone.0006421

Gershman, S. J. (2019). How to never be wrong. *Psychon. Bull. Rev.* 26, 13–28. doi: 10.3758/s13423-018-1488-8

Gershman, S. J., and Cikara, M. (2021). Structure learning principles of stereotype change. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/52f9c

Gilead, M., Trope, Y., and Liberman, N. (2020). Above and beyond the concrete: the diverse representational substrates of the predictive brain. *Behav. Brain Sci.* 43:e121. doi: 10.1017/S0140525X19002000

Griffiths, O., Erlinger, M., Beesley, T., and Le Pelley, M. E. (2018). Outcome predictability biases cued search. *J. Exp. Psychol. Learn. Mem. Cogn.* 44, 1215–1223. doi: 10.1037/xlm0000529

Harris, P. L., and Corriveau, K. H. (2021). Beliefs of children and adults in religious and scientific phenomena. *Curr. Opin. Psychol.* 40, 20–23. doi: 10.1016/j.copsyc.2020.08.003

Heilman, M. E., Wallen, A. S., Fuchs, D., and Tamkins, M. M. (2004). Penalties for success: reactions to women who succeed at male gender-typed tasks. *J. Appl. Psychol.* 89, 416–427. doi: 10.1037/0021-9010.89.3.416

Heimberg, R. G., Brozovich, F. A., and Rapee, R. M. (2010). "A cognitive behavioral model of social anxiety disorder: update and extension," in *Social Anxiety: Clinical, Developmental, and Social Perspectives*, eds S. G. Hofmann and P. M. DiBartolo (Waltham, MA: Elsevier), 395–422. doi: 10.1016/B978-0-12-375096-9.00015-8

Heimberg, R. G., Brozovich, F. A., and Rapee, R. M. (2014). "A cognitive-behavioral model of social anxiety disorder," in *Social Anxiety: Clinical, Developmental, and Social Perspectives*, eds S. G. Hofmann and P. M. DiBartolo (Waltham, MA: Elsevier), 705–728. doi: 10.1016/B978-0-12-394427-6.00024-8

Hepper, E. G., Hart, C. M., Gregg, A. P., and Sedikides, C. (2011). Motivated expectations of positive feedback in social interactions. *J. Soc. Psychol.* 151, 455–477. doi: 10.1080/00224545.2010.503722

Hofmann, S. G. (2007). Cognitive factors that maintain social anxiety disorder: a comprehensive model and its treatment implications. *Cogn. Behav. Ther.* 36, 193–209. doi: 10.1080/16506070701421313

Hohwy, J. (2013). *The Predictive Mind.* Oxford, UK: Oxford University Press. doi: 10.1093/acprof:oso/9780199682737.001.0001

Hohwy, J. (2020). New directions in predictive processing. *Mind Lang.* 35, 209–223. doi: 10.1111/mila.12281

Hohwy, J., and Michael, J. (2017). "Why should any body have a self?," in *The Subject's Matter: Self-Consciousness and the Body*, eds F. De Vignemont and A. J. T. Alsmith (Cambridge, MA: The MIT Press), 364–387. doi: 10.7551/mitpress/10462.003.0020

Jost, J. T. (2019). A quarter century of system justification theory: questions, answers, criticisms and societal applications. *Br. J. Soc. Psychol.* 58, 263–314. doi: 10.1111/bjso.12297

Jost, J. T., Banaji, M. R., and Nosek, B. A. (2004). A decade of system justification theory: accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychol.* 25, 881–919. doi: 10.1111/j.1467-9221.2004.00402.x

Kahl, S., and Kopp, S. (2017). "Self-other distinction in the motor system during social interaction: a computational model based on predictive processing," in *Proceedings of the 39th Annual Conference of the Cognitive Science Society* 2350–2355.

Kashdan, T. B., Farmer, A. S., Adams, L. M., Ferssizidis, P., McKnight, P. E., and Nezlek, J. B. (2013). Distinguishing healthy adults from people with social anxiety disorder: evidence for the value of experiential avoidance and positive emotions in everyday social interactions. *J. Abnorm. Psychol.* 122, 645–655. doi: 10.1037/a0032733

Kashdan, T. B., and Savostyanova, A. A. (2011). Capturing the biases of socially anxious people by addressing partner effects and situational parameters. *Behav. Ther.* 42, 211–223. doi: 10.1016/j.beth.2010.07.004

Kruglanski, A. W., Jasko, K., Milyavsky, M., Chernikova, M., Webber, D., Pierro, A., et al. (2018). Cognitive consistency theory in social psychology: a paradigm reconsidered. *Psychol. Inq.* 29, 45–59. doi: 10.1080/1047840X.2018.1480619

Kube, T., Kirchner, L., Lemmer, G., and Glombiewski, J. A. (2021). How the discrepancy between prior expectations and new information influences expectation updating in depression—the greater, the better? *Clin. Psychol. Sci.* doi: 10.1177/21677026211024644. [Online ahead of print].

Kube, T., and Rozenkrantz, L. (2021). When beliefs face reality: an integrative review of belief updating in mental health and illness. *Perspect. Psychol. Sci.* 16, 247–274. doi: 10.1177/1745691620931496

Kunda, Z. (1990). The case for motivated reasoning. *Psychol. Bull.* 108, 480–498. doi: 10.1037/0033-2909.108.3.480

Kunda, Z., and Oleson, K. C. (1995). Maintaining stereotypes in the face of disconfirmation: constructing grounds for subtyping deviants. *J. Pers. Soc. Psychol.* 68, 565–579. doi: 10.1037//0022-3514.68.4.565

Maheshwari, S., Kurmi, R., and Roy, S. (2021). Does memory bias help in maintaining self-esteem? Exploring the role of self-verification motive in memory bias. *J. Cogn. Psychol.* 33, 549–556. doi: 10.1080/20445911.2021.1926466

Mead, G. H. (1934). *Mind, Self and Society from the Standpoint of a Social Behaviorist.* Chicago, IL: University of Chicago Press.

Mendes, W. B., Blascovich, J., Hunter, S. B., Lickel, B., and Jost, J. T. (2007). Threatened by the unexpected: physiological responses during social interactions with expectancy-violating partners. *J. Pers. Soc. Psychol.* 92, 698–716. doi: 10.1037/0022-3514.92.4.698

Millgram, Y., Joormann, J., Huppert, J. D., and Tamir, M. (2015). Sad as a matter of choice? Emotion-regulation goals in depression. *Psychol. Sci.* 26, 1216–1228. doi: 10.1177/0956797615583295

Moore, M. T., and Fresco, D. M. (2012). Depressive realism: a meta-analytic review. *Clin. Psychol. Rev.* 32, 496–509. doi: 10.1016/j.cpr.2012.05.004

Moss-Racusin, C. A., and Johnson, E. R. (2016). Backlash against male elementary educators. *J. Appl. Soc. Psychol.* 46, 379–393. doi: 10.1111/jasp.12366

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J., and Friston, K. J. (2014). Bayesian inferences about the self (and others): a review. *Conscious. Cogn.* 25, 67–76. doi: 10.1016/j.concog.2014.01.009

Nassar, M. R., Wilson, R. C., Heasly, B., and Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *J. Neurosci.* 30, 12366–12378. doi: 10.1523/JNEUROSCI.0822-10.2010

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Curr. Opin. Neurobiol.* 14, 769–776. doi: 10.1016/j.conb.2004.10.016

Okasha, S. (2013). The evolution of Bayesian updating. *Philos. Sci.* 80, 745–757. doi: 10.1086/674058

Oyserman, D. (2009). Identity-based motivation: implications for action-readiness, procedural-readiness and consumer behavior. *J. Consum. Psychol.* 19, 250–260. doi: 10.1016/j.jcps.2009.05.008

Oyserman, D., Elmore, K., and Smith, G. (2012). "Self, self-concept and identity," in *Handbook of Self and Identity, 2nd ed*, eds M. R. Leary and J. P. Tangney (New York, NY: The Guilford Press), 69–104.

Parrish, M. H., Dutcher, J. M., Muscatell, K. A., Inagaki, T. K., Moieni, M., Irwin, M. R., et al. (2022). Frontostriatal functional connectivity underlies self-enhancement during social evaluation. *Soc. Cogn. Affect. Neurosci.* doi: 10.1093/scan/nsab139. [Online ahead of print].

Rao, R. P. N., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Reggev, N., Chowdhary, A., and Mitchell, J. P. (2021). Confirmation of interpersonal expectations is intrinsically rewarding. *Soc. Cogn. Affect. Neurosci.* 16, 1276–1287. doi: 10.1093/scan/nsab081

Reiss, S. (2004). Multifaceted nature of intrinsic motivation: the theory of 16 basic desires. *Rev. Gen. Psychol.* 8, 179–193. doi: 10.1037/1089-2680.8.3.179

Ronde, C., and Swann, W. B. (1993). "Caught in the crossfire: positivity and self-verification strivings among people with low self-esteem," in *Self-Esteem*, ed R. F. Baumeister (Boston, MA: Springer), 147–165. doi: 10.1007/978-1-4684-8956-9_8

Ross, L. (2018). From the fundamental attribution error to the truly fundamental attribution error and beyond: my research journey. *Perspect. Psychol. Sci.* 13, 750–769. doi: 10.1177/1745691618769855

Ross, L., Greene, D., and House, P. (1977). The "false consensus effect": an egocentric bias in social perception and attribution processes. *J. Exp. Soc. Psychol.* 13, 279–301. doi: 10.1016/0022-1031(77)90049-X

Rudman, L. A., Moss-Racusin, C. A., Phelan, J. E., and Nauts, S. (2012). Status incongruity and backlash effects: defending the gender hierarchy motivates prejudice against female leaders. *J. Exp. Soc. Psychol.* 48, 165–179. doi: 10.1016/j.jesp.2011.10.008

Salimpoor, V. N., Zald, D. H., Zatorre, R. J., Dagher, A., and McIntosh, A. R. (2015). Predictions and the brain: how musical sounds become rewarding. *Trends Cogn. Sci.* 19, 86–91. doi: 10.1016/j.tics.2014.12.001

Sedikides, C. (1993). Assessment, enhancement and verification determinants of the self-evaluation process. *J. Pers. Soc. Psychol.* 65, 317–338. doi: 10.1037/0022-3514.65.2.317

Sedikides, C., and Green, J. D. (2000). On the self-protective nature of inconsistency-negativity management: using the person memory paradigm to examine self-referent memory. *J. Pers. Soc. Psychol.* 79, 906–922. doi: 10.1037//0022-3514.79.6.906

Sedikides, C., and Gregg, A. P. (2008). Self-enhancement: food for thought. *Perspect. Psychol. Sci.* 3, 102–116. doi: 10.1111/j.1745-6916.2008.00068.x

Sharot, T. (2011). The optimism bias. *Curr. Biol.* 21, R941–R945. doi: 10.1016/j.cub.2011.10.030

Sharot, T., and Garrett, N. (2016). Forming beliefs: why valence matters. *Trends Cogn. Sci.* 20, 25–33. doi: 10.1016/j.tics.2015.11.002

Sowislo, J. F., and Orth, U. (2013). Does low self-esteem predict depression and anxiety? A meta-analysis of longitudinal studies. *Psychol. Bull.* 139, 213–240. doi: 10.1037/a0028931

Steele, C. M. (1992). Race and the schooling of black americans. *Atlantic* 269, 68–78.

Steele, C. M. (1997). A threat in the air: how stereotypes shape intellectual identity and performance. *Am. Psychol.* 52, 613–629. doi: 10.1037//0003-066x.52.6.613

Steele, C. M., and Aronson, J. (1995). Stereotype threat and the intellectual test performance of african americans. *J. Pers. Soc. Psychol.* 69, 797–811. doi: 10.1037//0022-3514.69.5.797

Swann, W. B. (2011). "Self-Verification Theory," in *Handbook of Theories of Social Psychology*, 2nd Edn., eds P. A. M. Van Lange and A. W. Kruglanski (Thousand Oaks, CA: SAGE Publications Ltd.), 23–42. doi: 10.4135/9781446249222.n27

Swann, W. B., De La Ronde, C., and Hixon, J. G. (1994). Authenticity and positivity strivings in marriage and courtship. *J. Pers. Soc. Psychol.* 66, 857–869. doi: 10.1037//0022-3514.66.5.857

Swann, W. B., Pelham, B. W., and Krull, D. S. (1989). Agreeable fancy or disagreeable truth? Reconciling self-enhancement and self-verification. *J. Pers. Soc. Psychol.* 57, 782–791. doi: 10.1037//0022-3514.57.5.782

Swann, W. B., and Read, S. J. (1981). Self-verification processes: how we sustain our self-conceptions. *J. Exp. Soc. Psychol.* 17, 351–372. doi: 10.1016/0022-1031(81)90043-3

Tamir, M., Schwartz, S. H., Oishi, S., and Kim, M. Y. (2017). The secret to happiness: feeling good or feeling right? *J. Exp. Psychol. Gen.* 146, 1448–1459. doi: 10.1037/xge0000303

Taylor, S. E., and Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychol. Bull.* 103, 193–210. doi: 10.1037/0033-2909.103.2.193

Theriault, J. E., Young, L., and Barrett, L. F. (2021). The sense of should: a biologically-based framework for modeling social pressure. *Phys. Life Rev.* 36, 100–136. doi: 10.1016/j.plrev.2020.01.004

Thorndike, E. L. (1911). *Animal Intelligence: Experimental Studies.* New York, NY: The Macmillan Company.

Tormala, Z. L., Briñol, P., and Petty, R. E. (2006). When credibility attacks: the reverse impact of source credibility on persuasion. *J. Exp. Soc. Psychol.* 42, 684–691. doi: 10.1016/j.jesp.2005.10.005

Ueda, S., Sato, T., and Kumada, T. (2021). Model-based prediction of operation consequences when driving a car to compensate for a partially restricted visual field by A-pillars. *Front. Hum. Neurosci.* 15:697295. doi: 10.3389/fnhum.2021.697295

Van de Cruys, S., and Van Dessel, P. (2021). Mental distress through the prism of predictive processing theory. *Curr. Opin. Psychol.* 41, 107–112. doi: 10.1016/j.copsyc.2021.07.006

van Elk, M., and Aleman, A. (2017). Brain mechanisms in religion and spirituality: an integrative predictive processing framework. *Neurosci. Biobehav. Rev.* 73, 359–378. doi: 10.1016/j.neubiorev.2016.12.031

Vlasceanu, M., Morais, M. J., and Coman, A. (2021). The effect of prediction error on belief update across the political spectrum. *Psychol. Sci.* 32, 916–933. doi: 10.1177/0956797621995208

Westra, E. (2019). Stereotypes, theory of mind and the action-prediction hierarchy. *Synthese* 196, 2821–2846. doi: 10.1007/s11229-017-1575-9

Yon, D., de Lange, F. P., and Press, C. (2019). The predictive brain as a stubborn scientist. *Trends Cogn. Sci.* 23, 6–8. doi: 10.1016/j.tics.2018.10.003

Yon, D., Zainzinger, V., de Lange, F. P., Eimer, M., and Press, C. (2021). Action biases perceptual decisions toward expected outcomes. *J. Exp. Psychol. Gen.* 150, 1225–1236. doi: 10.1037/xge0000826

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.