Properties and predictiveness of affective prediction errors

Marius Vollberg
University of Amsterdam, Amsterdam, Netherlands

Abstract

Do verbally reported feelings follow reinforcement learning principles? Prediction errors—differences between expectations and outcomes—are key in models of learning across humans, animals, and machines. Historically, the emphasis has been on outcomes in the environment (e.g., money or food), focusing relatively less on the fact that humans can also report correspondingly expected and experienced affect (i.e., feelings). Recent research suggests that expected and experienced affect, including prediction errors, can explain behavior beyond outcomes in the environment alone. However, the properties of affective prediction errors underlying this explanatory power are unknown. We address this gap across two studies. We show that affective prediction errors can decrease over time, but that the decrease depends on introspection (Study 1). We then replicate this finding while additionally documenting transfer effects across tasks (Study 2). Crucially, decreases in affective prediction errors generally occurred independent of changes in behavior.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*