OVEREXPECTATION

Authored by

mohammad looti

October 27, 2025

OVEREXPECTATION

Primary Disciplinary Field(s): Classical Conditioning, Learning Theory, Experimental Psychology

1. Core Definition

The concept of **overexpectation** describes a specific phenomenon observed in Pavlovian conditioning experiments, characterized by a measurable reduction in the learned associative strength of individual conditioned stimuli (CSs) after they have been presented together as a compound stimulus (CS-AB) and paired with an unconditioned stimulus (US) that is equal in magnitude to the US previously presented with each individual component. Essentially, the organism develops an associative strength for the compound stimulus that is greater than the actual magnitude of the subsequent US it receives. This discrepancy, often termed a negative **prediction error**, results in the organism effectively "over-expecting" the outcome, leading to a subsequent devaluation or lessening of the reaction to the component conditioned stimuli when tested individually later.

The mechanism of overexpectation is most highly correlated with methods utilized in associative learning theory, particularly those modeled by influential mathematical frameworks like the Rescorla-Wagner model. This model posits that the amount of learning that occurs on any given trial is directly proportional to the difference between the expected outcome and the actual outcome. When two maximally conditioned stimuli (A and B) are compounded and paired with an expected outcome (US) that only satisfies the association strength of one component, the sum of the expected strengths ($V_A + V_B$) is double the actual received strength (V_{US}), creating a large negative error term that drives the strength of both V_A and V_B downward.

2. Theoretical Context: Pavlovian Conditioning

Overexpectation serves as a critical demonstration of the nature of association formation within the framework of **classical conditioning**. It provides evidence that associative learning is not simply a passive pairing process but is fundamentally driven by predictive relationships and the computation of error. When an organism learns that Stimulus A predicts Outcome X, and Stimulus B also predicts Outcome X, combining A and B leads the organism to predict an outcome greater than X, perhaps 2X. If the resulting outcome is only X, the failure to meet the aggregated expectation forces a recalibration of the predictive value of the component stimuli.

This phenomenon contrasts sharply with earlier, simpler views of conditioning that emphasized mere contiguity. Overexpectation highlights the role of stimulus competition in determining the final associative strength of various elements within a learning environment. Because the associative strength is finite (limited by the magnitude of the US), the components of the compound stimulus

must compete for that limited associative strength when the outcome is unexpectedly small relative to the aggregate prediction. This dynamic process underscores the active, adaptive nature of associative learning systems.

3. Mechanism and Experimental Protocol

The induction of the **overexpectation effect** requires a precise two-stage experimental design to ensure that the individual components possess robust and equal associative strength prior to compounding. The typical experimental protocol proceeds as follows, often utilizing procedures developed in animal models (such as pigeons or rats) using appetitive stimuli (e.g., food or water reward):

Phase 1: Individual Conditioning (Acquisition): Two distinct conditioned stimuli, CS-A and CS-B, are separately and repeatedly paired with a moderate or strong unconditioned stimulus (US). This phase establishes maximal associative strength (V_A and V_B) for both components, meaning that the presentation of A or B reliably elicits a conditioned response (CR).
Phase 2: Compound Conditioning (Overexpectation Trials): The two conditioned stimuli are then presented simultaneously as a compound (CS-AB), but crucially, this compound is paired with a US of the same magnitude used in Phase 1 (US_X). The organism expects an outcome proportional to $V_A + V_B$, but only receives US_X. Since $V_A + V_B > V_{US_X}$, a large negative prediction error is generated.
Phase 3: Testing: The individual components (CS-A and CS-B) are presented separately again. The key finding is a significant reduction in the magnitude or frequency of the conditioned response (CR) to both A and B compared to the response level established in Phase 1 or compared to control groups that did not undergo the compounded Phase 2.

This observed reduction in conditioned responding confirms the **devaluation** of the component stimuli. The organism has learned, via the negative prediction error, that the individual stimuli are less predictive of the outcome than initially believed, given the context of the compound pairing. This outcome is compelling evidence that predictive value is distributed across stimuli based on the total outcome predicted versus the total outcome received.

4. Relationship to Blocking and Other Learning Phenomena

Overexpectation is often discussed alongside other fundamental phenomena in associative learning, most notably the blocking effect. While both blocking and overexpectation involve competition for associative strength and depend upon the principle of prediction error, they arise from distinct experimental histories and involve different types of prediction errors:

Blocking: Occurs when prior conditioning of CS-A with the US prevents or "blocks" subsequent conditioning of CS-B when CS-A and CS-B are paired together with the US (CS-AB $\rightarrow$ US). In

blocking, the prediction error for the new stimulus B is zero because A already fully predicts the US.

Overexpectation: Occurs after separate maximal conditioning of CS-A and CS-B. When the compound CS-AB is paired with the original US, the organism expects a super-additive outcome (V_A + V_B) but receives only the simple additive outcome (V_US). This generates a **negative prediction error**, causing the associative strengths of A and B to decrease.

In essence, blocking demonstrates that learning only occurs when the outcome is surprising (i.e., when expectation is not met), while overexpectation demonstrates that when the outcome is less than the sum of the parts predicted, the predictive value of those parts must be actively reduced. Both effects highlight the sophisticated, cognitive nature of classical conditioning, where learning is about updating predictions rather than merely forming habits.

5. Broader Applications and Implications

Although historically studied within the confines of animal learning experiments, the underlying principle of overexpectation--that predictions based on the sum of component parts can be disproportionately high compared to the actual outcome--has applications across various arenas of psychology and cognitive science. The general psychological principle is that when multiple cues point toward a very strong outcome, but the eventual outcome is only moderately strong, the predictive value of all initial cues is degraded.

In human cognition and decision-making, analogous effects can be seen in situations involving **over-optimism** or inflated expectations. For instance, if an individual believes two highly capable employees (A and B) will generate a synergistic, super-additive result (2X), but the actual project outcome is merely successful (X), the perceived competence or value of employees A and B may temporarily decrease due to the disappointment of the unmet prediction. The overexpectation model thus provides a fundamental framework for understanding how highly predictive cues can lose their associative value when they are compounded but fail to deliver a proportional combined outcome.

Further Reading

Rescorla-Wagner Model of Conditioning
Classical Conditioning
Associative Learning