The Stabilization of the Past: Feedback Loops in Self-Memory Dynamics

In this paper, I argue that the relation between self-related information (the self-model) and episodic memories (EM) involves positive and negative feedback loops leading to the stabilization of mnemonic tendencies also known as self-related memory biases. The central claim is that self-related memory biases, such as the self-enhancing bias or the consistency bias (Schacter et al., 2023), are the result of control processes over memory recall subordinated to a hierarchy of values. In arguing for this claim, I draw from the self-memory system approach (Conway & Pleydell-Pearce, 2000; Conway et al., 2019) and perceptual control theory (Mansell & Marken, 2015).

The self-memory system (SMS) is a framework for the integration of episodic information and autobiographical/self-related knowledge resulting in memory retrieval. In line with generative approaches, the framework distinguishes between stored, long-term representations and contextually activated information, and treats both as crucial components (Conway et al., 2019) of memory retrieval. According to the SMS, single memories are constructed by integrating episodic information about specific events, autobiographical knowledge, and conceptual knowledge about the self, via recurrent and iterative patterns of activation of information until search criteria are met (Conway and Pleydell-Pearce, 2000).

Perceptual control theory (PCT) posits that behavior is a function of the attempt to maintain a desired state (Mansell & Marken, 2015). The currently experienced state is continuously compared to an internal goal specification (or reference signal); any discrepancy generates an error signal that drives behaviors to fix the perception. The desired state is influenced by a hierarchy of values, goals, and norms. For example, the value of social safety determines the evaluation of the currently experienced situation with reference to a desired level of perceived safety, and an eventual mismatch leads to taking actions to diminish the mismatch (Gucciardi et al., 2026).

In order to describe the reciprocal relations between self-model and episodic memory, it is necessary to distinguish between 1) the modulation exerted by the self-model on EM at retrieval, considered unidirectionally, 2) the iterative, bidirectional self-memory dynamics and, lastly, 3) the diachronic stabilization of self-memory dynamics resulting in self-related memory biases. With regard to 1), the modulation exerted by the self-model on EM is a process in which information from the self-model is integrated with other information and the memory trace, resulting in the retrieval of a specific memory. From the perspective of perceptual control

theory, the influence of the self-model on EM is the result of *negative feedback loops* that prevent and reduce the mismatch between the actual experience (the recall of a specific memory and related affect) and the desired state, in light of the values of maintaining a coherent/positive/etc. self-model. For example, retrieving a memory that contradicts core beliefs about the self contravenes the value of a coherent self and can cause deeply negative affect, resulting in a great mismatch between current and desired experience, consequently, control processes are involved in avoiding retrieving or modifying such a memory.

Conversely, retrieved memories can influence the self-model. For instance, frequent recollection of EMs representing professional successes might result in self-attribution of the trait of being successful encoded in the self-model. Taken together, the reciprocal influence between self-model and EM gives rise to self-memory dynamics (2). Self-memory dynamics involve iterative feedback loops, such that the self-model at a point in time, $S\text{-}M_t$, shapes EM_t, and EM_t feeds back into $S\text{-}M_{t+1}$.

In relation to 3), over time and without counterbalancing constraints, self-memory dynamics converge towards inflexible, recurrent, and self-validating patterns. Correspondence, i.e., the tendency to construct memories that are veridical or accurate with respect to factual events they represent (Dings et al., 2023), is one such constraint. Several biases have been indicated in the literature that can be modeled as the result of *positive feedback loops* in self-memory dynamics. The most investigated is the self-enhancement (or self-serving) bias (Demiray & Janssen, 2015; Schacter et al., 2023), which is the tendency to recall memories that reflect positively on the self. In such a case, a generally positive self-model ($S\text{-}M_t$) contributes to the construction of a positive memory (EM_t) which feeds back into the self-model ($S\text{-}M_{t+1}$) and further enhances it. Analogous processes can be assumed for the negative memory bias observed in depression (Marchetti et al., 2018), for the self-coherence or consistency bias (Conway, 2005), and others.

The stabilization of and interaction between self-related memory biases can be illuminated by PCT. In particular, the elusive relations between memory biases can be elucidated if understood as resulting from a hierarchy of values affecting control processes, as posited by PCT. For instance, whereas in the general population the bias for self-coherence is adaptive and functional in preserving a positive self-model, in depression, self-coherence supports the maintenance of a negative view of the self, despite the apparent disadvantages of doing so. Although this may seem counterintuitive, it can be readily explained by positing that a coherent identity is prioritized (i.e., it is higher in the hierarchy of values) over a positive self-model.

Indeed, evidence suggests that depressed individuals adopt behaviors that reinforce depressive symptoms (e.g., seeking negative feedback about the self from others or assuming self-defeating attitudes in social contexts) in order to validate their negative self-model (Hart et al., 2021). To summarize, I argue that PCT can provide fruitful insights for a systematic analysis of self-memory dynamics and the formation and stabilization of self-related memory biases.

Bibliography:

Conway, M. A. (2005). Memory and the self. Journal of Memory and Language 53: 594–628. https://doi.org/10.1016/j.jml.2005.08.005.

Conway, M. A., & Pleydell-Pearce C. W. (2000). The construction of autobiographical memories in the self-memory system. Psychological review 107.2: 261.

Demiray, B., & Janssen, S. M. J. (2015). The self-enhancement function of autobiographical memory. *Applied Cognitive Psychology, 29*(1), 49–60. https://doi.org/10.1002/acp.3074

Dings, R., & Newen, A. (2023). Constructing the past: The relevance of the narrative self in modulating episodic memory. Review of Philosophy and Psychology, 14, 87–112.

Gucciardi, D. F., Crane, M. F., Riddell, H., & Mansell, W. (2026). Toward an integrative perspective of personalised stress regulation: insights from perceptual control theory. *Health Psychology Review*, 1–25. https://doi.org/10.1080/17437199.2026.2619026

Hart, W., Breeden, C. J., Richardson, K., & Kinrade, C. (2021). Depression and the Adoption of Faux Depression Symptoms: Novel Evidence for a Self-Verification Perspective. Clinical Psychological Science, 9(4), 598–614. doi:10.1177/2167702621992337

Mansell, W., & Marken, R. S. (2015). The Origins and Future of Control Theory in Psychology. *Review of General Psychology*, *19*(4), 425-430. https://doi.org/10.1037/gpr0000057

Marchetti, I., Everaert, J., Dainer-Best, J., Loeys, T., Beevers, C. G., & Koster, E. H. W. (2018). Specificity and overlap of attention and memory biases in depression. *Journal of affective disorders*, *225*, 404–412. https://doi.org/10.1016/j.jad.2017.08.037

Schacter, D. L., Greene, C. M., & Murphy, G. (2023). Bias and constructive processes in a self-memory system. *Memory*, 1–10. https://doi.org/10.1080/09658211.2023.2232568