Fading Faces: When Agents Forget Who You Are

David Freire-Obregón [1]

[1]SIANI, Universidad de Las Palmas de Gran Canaria, Spain
david.freire@ulpgc.es

Abstract

As artificial agents form their cognitive depiction of others, their visual recognition of what they once knew becomes fragile. In this work, we discuss perceptual forgetting via a minimalist simulation of embodied agents. Agents have distinct visual identities (faces) and personal convolutional neural networks (CNN) learned online to identify neighbors. Initially, faces appear in high detail, but gradually decay to minimal figures, modeling cognitive aging or perceptual decline. We see a precipitous drop in recognition accuracy while symbolic identity remains intact, illustrating a dissociation between knowing who a person is and recognizing persons. Our results place forgetting not as erasure but as perception fade, a symbolic-perceptual dissociation resonant with actual cognitive mortality, prompting concerns about embodiment, memory, and decay in artificial life.

Introduction

Recognition is a basic type of social memory (Wang and Zhan, 2022). For humans and machines alike, knowing who a person or object is depends upon a tenuous chain of perceptual, symbolic, and embodied procedures. However, whereas machine recognition tends to presuppose ideal, noise-free situations, real perception deteriorates, transmutes, or disappears altogether. We model perceptual forgetting with a minimal simulation where agents on a 2D grid each have a distinct face. Using a fine-tuned ResNet18, they learn to recognize neighbors during encounters. Over time, faces degrade from detailed images to sketches, eroding perception: identities remain unchanged, yet recognition fails. Our setup draws its impulse from cognitive theories of mortality. Forgetting is not dichotomic in biological life: one can recall a name but not a face, or recall an emotion but not the individual. Similarly, in our agents, symbolic identity (an unforgettable, unique ID) itself is retained, but perceptual access to such identity decays. The effect is a deteriorating form of mortality, more worried about erosion of familiarity than about vanishing. This process falls within configurations within autopoiesis (Froese et al., 2023), homeodynamic systems (Yates, 1994), and representational drift (Driscoll, 2022), where internal models dissociate from external reality. Dissociation, here, does not originate from an agent's body, but from perception by a body of others. We thereby show how minimal agents, equipped with neural perception and short-term memory, can be exposed to symbolic decay.

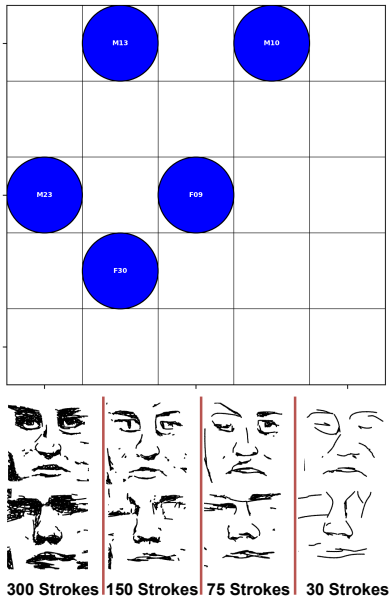

Figure 1: Snapshot of the 5×5 agent environment, with five agents randomly distributed in the grid. At the bottom, two sample identities are shown across four abstraction levels (300, 150, 75, and 30 strokes), illustrating the visual degradation process. These increasingly abstract depictions simulate the memory decay agents face when recognizing others under harsher perceptual conditions.

Experimental Setup

We implement an agent-based simulation using the Mesa framework. Each agent is assigned an initial facial image, randomly selected from the KDEF dataset, restricted to neutral expressions and one image per identity (Lundqvist et al., 1998). Faces are rendered with varying degrees of abstrac-

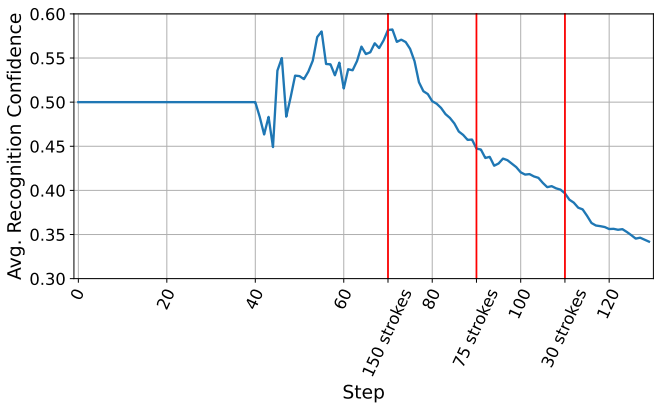

Figure 2: Average recognition confidence of agents over time. Red lines indicate degradation steps.

tion using CLIPasso sketches (Vinker et al., 2022), to four levels: 300, 150, 75, and 30 strokes. The 300-stroke rendering is the most detailed image; low-stroke renderings progressively degrade visual clarity (Freire-Obregón et al., 2025). Agents randomly move inside a 2D toroidal grid. Upon entering neighboring cells, two agents attempt to recognize each other. They recognize each other based on a CNN (ResNet18), and each agent has its own network. During the first 70 steps, agents incrementally learn: whenever they see an unfamiliar face, they assign it a new class label and update their network (only the final, i.e., classification, layer is trained). Agents stop learning from step 71 and begin evaluating instead. At certain intervals, facial visual quality deteriorates (from 300 to 150, to 75, to 30 strokes). The identity of the agent never varies, but it becomes harder to classify its perceptual data. To model the recognition decay, we keep track of whether or not agent A correctly classifies agent B when they encounter each other. If the prediction corresponds to the actual label learned at training, we say recognition has succeeded; if not, we mark it as a failure. Each agent monitors its accuracy over time and reveals a gradual decline in recognition performance as visual data decays. Trust in agents, measured as correct predictions divided by the number of attempts, also varies with their visual size, signifying the level of confidence. Such degradation mirrors how humans negotiate identity with partial memory, showing forgetting as part of social interaction rather than a flaw.

Results and Analysis

The development of overall average trust for all agents is presented in Figure 2, beginning from step 40, once assessment commences. For the initial 40 steps, agents gather no visual knowledge other than through local interaction; no measure of recognition performance is yet conducted.

Once evaluation starts, we observe a rapid rise in average trust, reaching a peak near step 70. This indicates that agents successfully consolidate their recognition capabilities during the learning phase, particularly when the input images maintain high fidelity (300-stroke sketches). However, from step 70 onwards, when the sketches gradually deteriorate, trust levels start to come down. The decline accelerates when agents move from 150-stroke to 75-stroke and finally to 30-stroke representations. At step 130, average trust levels drop to less than 0.35, a severe degradation of recognition accuracy. This behavior favors our hypothesis: agents forget not by deleting internal representations, but by perceiving perceptual mismatch between stored representations and newly corrupted inputs. Although still maintaining their learned CNN parameters, agents cannot match more abstract sketches, leading to a symbolic form of forgetting.

Conclusions

This work provides a symbolic account of memory and forgetting in embodied agents through face-based social interaction. We show how agents, with localized CNNs and finite lifespans, acquire recognition abilities via situated interaction, but gradually deteriorate, as perceptual fidelity worsens. Importantly, forgetting in this scenario does not arise due to erasure from memory but due to a mismatch between internal representations and external inputs, instantiating a structural understanding of forgetting. These raise questions about how perceptual decline and symbolic loss shape not only artificial memory systems, but also the emergence of sociality and cognitive development, suggesting that mortality in agents mirrors the human condition of remembering and forgetting.

Acknowledgements. This work is partially funded funded by project PID2021-122402OB-C22/MICIU/AEI /10.13039/501100011033 FEDER, UE and by the ACIISI-Gobierno de Canarias and European FEDER funds under project ULPGC Facilities Net and Grant EIS 2021 04.

References

Driscoll, L. N. (2022). Representational drift: Emerging theories for continual learning and experimental future directions. *Current Opinion in Neurobiology*, 76:102609.

Freire-Obregón, D., Neves, J., Žiga Emeršič, Meden, B., Castrillón-Santana, M., and Proença, H. (2025). Synthesizing multilevel abstraction ear sketches for enhanced biometric recognition. *Image and Vision Computing*, 154:105424.

Froese, T., Weber, N., Shpurov, I., and Ikegami, T. (2023). From autopoiesis to self-optimization: Toward an enactive model of biological regulation. *bioRxiv*.

Lundqvist, D., Flykt, A., and Öhman, A. (1998). The karolinska directed emotional faces (kdef). CD-ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet. ISBN 91-630-7164-9.

Vinker, Y., Pajouheshgar, E., Bo, J. Y., Bachmann, R. C., Bermano, A. H., Cohen-Or, D., Zamir, A., and Shamir, A. (2022). CLIPasso: Semantically-Aware Object Sketching. *ACM Trans. Graph.*, 41(4).

Wang, X. and Zhan, Y. (2022). Regulation of social recognition memory in the hippocampal circuits. *Frontiers in Neural Circuits*, 16:839931.

Yates, F. (1994). Order and complexity in dynamical systems: Homeodynamics as a generalized mechanics for biology. *Mathematical and Computer Modelling*, 19(6):49–74.