Prevalence-Induced Concept Change: Universal or Context-Dependent? Implications for Social Psychology and AI Cognition

Rebecca Albrecht (rebecca.albrecht@cognition.uni-freiburg.de)
Department of Psychology, Center for Cognitive Science
University of Freiburg, Germany

Mikhail S. Spektor (mikhail@spektor.ch)
College of Arts and Sciences
VinUniversity, Vietnam

Abstract

Prevalence-induced concept change (PICC) occurs when reduced category prevalence increases the likelihood that ambiguous stimuli are classified as belonging to the now-minority category. PICC has been observed across perceptual and social domains and persists despite instructions or incentives to suppress it. However, other findings suggest its expression is instead shaped by social context. If AI models are to be treated as theories of cognition, they should exhibit PICC as well. We show that a standard AI model for sequential learning, trained on dynamic category distributions, does not display PICC, instead favoring the now-majority category. This opposite-PICC effect suggests that simple sequential learning may be insufficient to produce PICC. Additional mechanisms, such as structured priors, contextual sensitivity, or internal feedback, seem necessary for its emergence. Our findings contribute to the understanding of PICC and its implications for categorization theories, AI-driven decision-making, and the role of media exposure in shaping social perceptions.

Keywords: Category Learning; Social Categorization; Artificial Intelligence and Cognition; Unified Theories of Cognition; AI and Psychological Theory.

Introduction

When making categorization judgments, do people rely on fixed decision rules or do they adjust their category boundaries in response to environmental contingencies and domain-specific information? The former assumes that categorization follows stable, context-independent principles, ensuring consistency in line with normative decision-making frameworks (Luce, 1959). In contrast, the latter suggests that category boundaries can shift in response to statistical changes in the environment, consistent with efficient coding theories (Bhui & Gershman, 2018). If categorization is flexible, this raises an important question: Are these adjustments driven by fundamental *cognitive* principles or do they depend on additional high-level considerations, such as domain-specific learning and social influences?

The recently popularized idea of prevalence-induced concept change (PICC) provides a prominent example of a phenomenon that results from flexible categorization. Levari and colleagues (2018) demonstrated that when the prevalence of a category decreases, people are more likely to classify ambiguous stimuli as belonging to the now-minority category (see Figure 1). Notably, this effect appears highly robust, persisting even when participants are explicitly instructed to resist it or when they are given incentives to maintain stable category boundaries. Such findings have been interpreted as evidence that PICC reflects a domain-independent cognitive mechanism, driven by fundamental principles of adaptive categorization.

In addition to having being demonstrated in low-level perceptual tasks (e.g., color categorization), PICC has also been observed in social judgments, such as morality, facial trustworthiness, and body-image perception (Devine et al., 2022, 2024; Levari et al., 2018). Unlike perceptual categorization, social judgments are often shaped by high-level considerations such as social norms, prior beliefs, and cultural expectations (Mendoza-Denton et al., 2001; Henrich et al., 2010; Rhodes et al., 2018). For example, PICC seems to occur in judgments of female but not male body size, aligning with media-driven stereotypes (Devine et al., 2022, 2024). If PICC is moderated by socially acquired priors, it would cast into question that PICC reflects a general principle of adaptive categorization.

Whether or not PICC is a fundamental cognitive process has profound implications for psychology and AI as a theory of cognition. If PICC is a domain-independent response to the statistical properties of an environment, its emergence in artificial systems designed to approximate human category learning would support its status as a fundamental cognitive mechanism (Lake et al., 2017; Tenenbaum et al., 2011). Foreshadowing our results, they suggest that tractable AI systems such as recurrent neural networks or long short-term memory models (LSTM) do not exhibit PICC under relevant statistical-learning conditions and model specifications. This absence indicates that simple sequential learning mechanisms are insufficient to reproduce human-like adaptive boundary shifts. For psychology, this highlights the need for systematic investigation into whether PICC operates as a general mechanism or whether its effects are shaped by domain-specific factors. For AI as a theory of cognition (Binz et al., 2024), it calls for further exploration of which additional mechanisms—such as internal feedback loops or contextual sensitivity—are necessary to capture the flexible categorization behaviors observed in humans.

Beyond its theoretical implications, understanding PICC is particularly relevant in today's social and informational landscape. If category boundaries shift based on prevalence, this could influence how people perceive social groups, respond to media-driven stereotypes, or adapt their moral judgments in changing environments. This trend may be further amplified

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

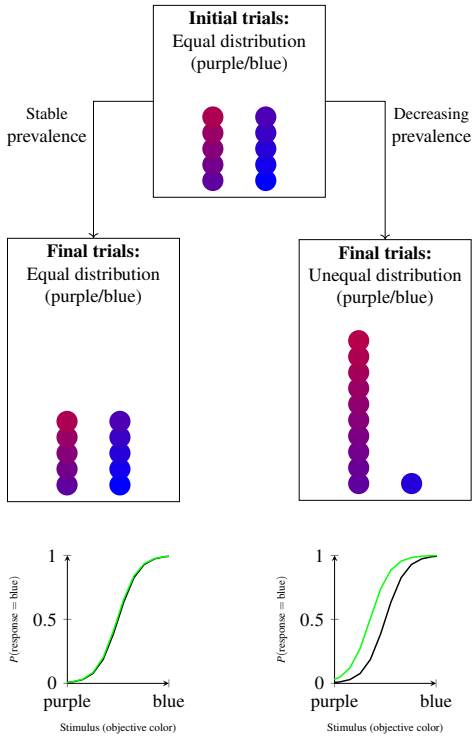

Figure 1: Idealized PICC results: In the stable prevalence condition, the probability distributions remain unchanged between initial (black line) and final trials (green line). In the decreasing prevalence condition, the probability distribution shifts toward the now-minority category (blue dots) in most studies.

by the increasing ability of individuals to curate their digital environments through social media. As they select their exposure through content engagement and as algorithms further fine-tune the content people are exposed to, the statistical occurrence of opposing viewpoints can dramatically diminish, creating a feedback loop and a conceptual change of categories in turn. Similarly, in AI-as-a-unified-theory-of-cognition development, understanding how models categorize information under changing distributions and social contexts is essential for designing systems that interact effectively with human users. Addressing whether PICC is domain-independent or shaped by social priors could help refine both psychological theories of categorization and AI approaches to adaptive learning.

Prevalence-induced concept change: A domain-independent phenomenon?

The original demonstrations of PICC relied on a two-alternative forced-choice task in which participants categorized stimuli that varied along a single continuous dimension. To illustrate, Levari et al. (2018) presented participants with a sequence of colored dots, each shown one at a time and ranging from blue to purple, and asked them to classify them as either "blue" or "purple".

In the *stable-prevalence* condition, blue and purple stimuli (as defined relative to the midpoint of the range of hues) appeared with equal frequency throughout the experiment. In the *decreasing-prevalence* condition, blue stimuli became progressively rarer, reaching a prevalence as low as 6%. Participants in the decreasing-prevalence condition systematically expanded their "blue" category boundary over time such that ambiguous stimuli were increasingly classified as being "blue" (Levari et al., 2018).

This effect was remarkably persistent: sudden shifts in stimulus prevalence, explicit warnings, explanations, and even monetary incentives failed to eliminate it. Furthermore, this pattern replicated across multiple domains, including facial trustworthiness, moral evaluations, and other socially relevant judgments (Levari et al., 2018). These findings have been interpreted as evidence that individuals unconsciously adjust their conceptual boundaries in response to statistical prevalence shifts, supporting the notion that categorization is adaptively flexible rather than strictly rule-based. This aligns with frameworks such as decision-by-sampling and efficient coding that propose that cognitive systems rely on relative comparisons rather than fixed decision rules (Bhui & Gershman, 2018; Woodford, 2020).

Cognitive modeling of prevalence-induced concept change

From a computational perspective, PICC challenges standard models of categorization which typically assume stable category boundaries. Many well-established frameworks—including probabilistic statistical models such as logistic regressions, similarity-based categorization models, and reinforcement-learning models—are unable to predict the boundary shifts observed due to PICC (Ashby et al., 1992; Nosofsky, 1986; Griffiths & Tenenbaum, 2007; Levari, 2022). In fact, models that rely on static decision thresholds or optimize responses based on error minimization tend to develop a bias in the opposite direction, such that the now-majority category is favored rather than the now-minority category (Green & Swets, 1966; Griffiths & Tenenbaum, 2007).

Bayesian models adjust categorization based on prior probabilities but do not inherently predict an over-classification of ambiguous stimuli into the now-minority category (Griffiths & Tenenbaum, 2007). Similarly, reinforcement-learning models that update decision policies based on action–reward contingencies typically converge towards the now-majority category unless explicitly trained to shift category boundaries. Unlike human participants in PICC tasks, standard reinforcement-learning models do not spontaneously adjust category boundaries in response to prevalence changes, as their decision policies are driven by error minimization rather than adaptive responses to distributional shifts (Botvinick et al., 2019).

This observation is highly general: models that optimize for classification accuracy tend to favor now-majority category responses unless explicitly adjusted through mechanisms such as category weighting or cost-sensitive learning.

Without such adjustments, these models prioritize overall accuracy, leading to a bias toward the majority category. The difficulty of accounting for PICC within these standard computational frameworks suggests the need for alternative modeling approaches that better capture the adaptive nature of human categorization.

Range-frequency theory predicts shifts in categorization thresholds One influential explanation for PICC is range–frequency theory (Parducci, 1995) which assumes that judgments are based on a combination of range normalization (rescaling stimulus values within the observed minimum and maximum range) and frequency weighting (ranking stimuli relative to others in the local context). A related approach can be found in decision by sampling (Stewart et al., 2006), which posits that judgments emerge through relative comparisons rather than absolute decision boundaries.

The key mechanism in these models is that as the prevalence of a category decreases, the distribution of experienced values in the local context shifts accordingly. This shift alters the relative position of ambiguous stimuli, making them appear more extreme within a context where one category is rarely observed. As a result, ambiguous stimuli become more likely to be classified into the now-minority category, as they fall closer to the remaining instances of that category within the shifted distribution. In support of this account, Levari (2022) report that simultaneously expanding the range of the now-minority category (so that the relative position of ambiguous stimuli remains constant) reduces or eliminates the PICC effect.

Despite range–frequency based accounts providing a compelling descriptive account of how category boundaries shift in response to changing stimulus distributions, the implications for long-term learning remain unclear (Levari et al., 2018; Levari, 2022). While decision by sampling includes an updating mechanism of the stored category representations (Stewart et al., 2006), it is not clear whether it leads to a lasting restructuring of categories or if it is a transient effect dependent on local stimulus context. The answer to this question has implications for whether PICC should generalize across domains or remain sensitive to domain-specific influences such as learned priors and social expectations (Bhui & Gershman, 2018).

A key strength of the range–frequency account is that it explains why PICC effects persist despite explicit instructions to maintain stable category boundaries. Unlike models that rely on explicit rule-based classification, range–frequency theory assumes that judgments are anchored to the local stimulus context (Levari et al., 2018; Stewart et al., 2006). Since this adjustment happens automatically, it remains unaffected by top-down control or explicit incentives. However, this feature conflicts with empirical findings from social psychology (Devine et al., 2022, 2024). Here, PICC appears to occur in body-size judgments for women but not for men, aligning with media-driven stereotypes rather than purely statistical adaptation. If PICC was domain-independent, such asymmetries would not be observed.

One possibility is that PICC is a fundamental mechanism, but its expression is moderated by pre-existing category boundaries shaped by cultural exposure and learned priors. Alternatively, this discrepancy might indicate that range–frequency theory is unable to account for how social priors influence the adaptation of category boundaries in complex cognitive domains.

Testing prevalence-induced concept change in a long term-short term memory network

To shed light on the question whether PICC is a fundamental cognitive process or requires high-level knowledge, we turn to AI models designed to approximate human category learning. If PICC emerges as a general principle of categorization, we might expect it to appear in domain-general learning models such as LSTMs. However, if PICC relies on structured priors or real-world contexts, then it would be absent in such models. In the following section, we present a computational demonstration testing whether a simple LSTM trained on dynamic category distributions exhibits PICC.

Unlike feed-forward neural networks, which rely solely on static input–output mappings, LSTMs explicitly process the order of information and maintain a form of short-term memory (Elman, 1990; Hochreiter & Schmidhuber, 1997), allowing them to integrate prior observations into their categorization. This makes LSTMs conceptually quite close to range–frequency theory.

Thus, the central question we ask is whether an LSTM trained on dynamic prevalence distributions exhibits PICC. If so, this would suggest that history-dependent learning alone can give rise to the effect, providing a computational stepping stone between simple decision models and more flexible adaptive systems. However, if the LSTM fails to replicate PICC, this would indicate that additional mechanisms, such as structured priors or internal feedback mechanisms, are necessary to explain why human participants show category boundary shifts under changing prevalence conditions.

Method

We tested whether LSTM neural networks exhibit PICC when trained on dynamically changing category distributions. Unlike static classifiers such as logistic regressions or feed-forward neural networks, LSTMs incorporate sequential dependencies and maintain a form of short-term memory, making them conceptually closer to models relying on local relative comparisons. If PICC arises from memory-based decision adjustments independent of domain-specific features, these models should, in principle, predict boundary shifts similar to those observed in humans.

Model architecture and training. We implemented an LSTM model in PyTorch (version 2.6.0) with two variations:

- **Single-output model**: A network with one output node trained to classify the current input and assign a discrete label.

- **Dual-output model**: An extended version with an additional output node predicting the next (future) label. This setup encourages the network to integrate sequential information through its recurrent connections rather than treating each input independently.

The model included a single recurrent layer (hidden size = 1) connected to one or two output nodes, depending on the model version. Despite the differences in the number of output nodes, both models produced qualitatively similar results. In the following analyses, we focus on the results of the dual-output model to highlight the network's use of sequential information.

To evaluate how the model responds to prevalence shifts, we tested four different learning environments. In all cases we generated synthetic sequences of length $T = 5$ (length of considered memory sequence). The final input from each such sequence was used to compute a binary label via a sigmoid/logistic function:

$$p(\text{label} = 1) = \frac{1}{1 + e^{-k \cdot (x - x_0)}} \tag{1}$$

where x is the input, k controls the slope, and x_0 is the midpoint. In the analysis presented here, k was set to 20 and x_0 was set to .5. For the next-label output, labels were shifted one step forward in the sequence. Supervised training was done using cross-entropy loss summed across the two outputs (current and next). The Adam optimizer (learning rate 0.01) trained the model for $1,000$ epochs.

We trained the model in four distinct learning environments, each comprising of $1,000$ observations:

1. **Balanced Environment:** The model was trained on a stable, balanced distribution in which all inputs were drawn from a standard uniform distribution $\mathcal{U}(0,1)$. The category labels were approximately balanced throughout.

2. **Imbalanced Environment:** In the imbalanced environment, 90% of the inputs were drawn from $\mathcal{U}(0,.5)$, whereas the remaining 10% were drawn from $\mathcal{U}(.5,1)$, without any changes in the baseline prevalence throughout the $1,000$ observations.

3. **PICC Environment:** This is the environment that mimics typical PICC experiments with human participants. The inputs for the first 500 observations were drawn from the balanced environment and the final 500 observations were drawn from the imbalanced environment.

4. **Shifting Environment:** The shifting environment comprised of two smaller PICC environments, where the sequence was balanced/imbalanced/balanced/imbalanced à 250 observations, for a total of three change points.

Model Evaluation. After training, the model was evaluated in two distinct ways. First, we assessed whether it had successfully learned to classify stimuli in the same environment it was trained on. Second, we tested it on all other environments to examine whether any kind of concept change emerged in these different prevalence conditions. In all environments, evaluation was conducted without feedback, enabling us to determine whether the model adapted its categorization behavior based solely on exposure to shifting statistical distributions. Since the learning environment did not lead to qualitatively different outcomes, we focus here on results obtained with the PICC training environment.

We tested various hyperparameter configurations (learning rate, sequence length T, hidden layer size, and number of training epochs) across all models to ensure that our findings were not artifacts of specific parameter choices. All configurations produced qualitatively similar results. Most notably, the results did not depend on hidden size or sequence length T. The reported parameters represent a configuration in which the target function was learned effectively.

To evaluate the model's success, we relied on standard metrics from machine learning to assess its classification performance. The confusion matrix is a contingency table of true category labels (based on the objective values of the inputs, in this case category '1' if input $\geq .5$) and model-predicted category labels. The confusion matrix can be interpreted similarly to signal detection theory contingency tables and is helpful in identifying the cases that a model mis-classifies. From the confusion matrix, we derive precision, recall, and the F1-score. Precision measures the proportion of items predicted as a certain category (e.g., minority) that actually belong to that category, whereas recall indicates the proportion of true instances of that category that the model correctly identifies. The F1-score is the harmonic mean of precision and recall, providing a single measure that balances both. A high F1-score indicates strong performance in correctly identifying instances while limiting false positives. A PICC effect is characterized by high recall (a high rate of true positives) and low precision (a high rate of false positives) for the minority category, indicating that more stimuli are classified as belonging to the minority category than is objectively warranted.

The receiver operating characteristic (ROC) curve illustrates how the model's classification performance varies across different decision thresholds and can be interpreted the same way as ROC curves in signal detection theory. The area under the ROC curve (AUC) summarizes this performance: An AUC of 1 indicates perfect classification, while an AUC of .5 means the model performs no better than chance.

An additional (red) dot on the ROC curve represents the model's performance at the standard (unbiased) threshold of .5, where inputs are assigned to categories based on whether their predicted probability exceeds this value. If the model adapts its classification boundary dynamically in response to prevalence shifts, we would expect changes in the ROC curve and decision thresholds across different conditions.

Results. Both models successfully learned the target function in all learning environments, achieving high classification accuracy on both training and balanced test sets as indi-

cated by the ROC AUC (all ROC AUCs $\geq$ .96). To assess whether the model exhibited a bias toward the now-minority category, which would reflect a PICC, we analyzed the confusion matrix and ROC curve. For the balanced test (conducted on the observations from the balanced learning environment), the F1-score for both categories was .94, indicating high accuracy and no discernible bias. However, when evaluated on an imbalanced test set, a clear bias towards the *majority* category emerged, reflecting the opposite of PICC (Table 1). While the precision and recall for the now-majority category were high, recall was notably lower for the now-minority category. This indicates that, compared to the now-minority category, relatively few now-majority category items were misclassified (see also Figure 2).

Category	Precision	Recall	F1-score	Support
Majority	.93	1.00	.96	838
Minority	.99	.61	.76	161
Accuracy			.94	999
Macro Avg	.96	.81	.86	999
Weighted Avg	.94	.94	.93	999

Table 1: Classification report for the current label classification for the model trained on on the PICC environment and tested on an imbalanced set.

Discussion

When the prevalence of stimuli changes, do people adapt the classification of said stimuli into categories accordingly? The phenomenon of PICC suggests that they do, in a way that they tend to over-classify ambiguous stimuli into the now-minority category. The present work explored to what extent PICC is a fundamental property of sequential adaptation to distribution changes, and might therefore arise in AI architectures aiming to mimic human cognition, or if it requires additional high-level properties. To answer this question, we used the LSTM neural-network architecture to test whether PICC emerges in such artificial agents. Across a variety of model instances and learning environments, all models produced an inverse PICC, raising questions about whether simple bottom-up learning is sufficient to produce PICC.

The lack of PICC in our LSTM simulations yields two key insights. First, although LSTM architectures incorporate short-term dependencies, they consistently produce boundary shifts in the opposite direction of PICC, toward the majority category consistent with an error-minimization objective. In this sense, the model adapts to changing input statistics, but only in a way that reduces mis-classifications, not in a way that reflects human-like category expansion. Second, this opposite shift highlights a critical divergence from human behavior: the model does not broaden the now-minority category under prevalence change, suggesting that mechanisms beyond sequential learning—such as internal feedback, representational goals, or epistemic uncertainty reduction—may be required to reproduce human-like adaptive categorization.

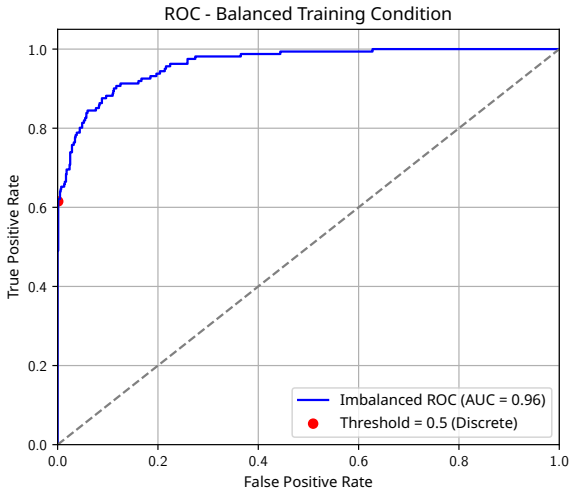

Figure 2: Receiver-operating characteristic curve for the long short-term memory model classification output. The red dot represents discrete predictions at a decision threshold of .5, showing a clear bias towards the now-majority category.

Is categorization domain-independent or socially shaped?

Our findings have important implications for categorization in the social sciences. While previous research has framed PICC as a domain-independent and robust cognitive mechanism (Levari et al., 2018), our findings raise questions about the assumption that PICC operates as a fully domain-general process. Moreover, our overview of research on PICC in social psychology suggests that the effect is not as universal as initially assumed, even in humans. PICC has been consistently observed in some domains such as perceptual categorization or moral judgments, yet it does not seem to emerge in some other contexts. Notably, PICC was reported to be absent in certain social judgments such as body-size categorization for men (Devine et al., 2022, 2024).

One interpretation is that PICC may not reflect a universal cognitive principle but rather an effect that emerges in specific domains due to structured priors and learned social expectations. Instead of a fundamental adaptation constrained by priors, PICC may depend entirely on pre-existing category structures, determining when and where it emerges. This view is consistent with research showing that categorization boundaries are highly sensitive to social context and stereotype-driven perception (Eberhardt et al., 2004). Notably, PICC has so far not been systematically tested under conditions where the prevalence of the socially *undesirable* category decreases or where in-group versus out-group distinctions are explicitly manipulated. Most studies assume that the prevalence of the desirable category decreases, leaving open the question of whether PICC relies in part on social desirability biases and group membership.

A challenge for AI cognition?

Our computational modeling results revealed that a standard LSTM did not exhibit PICC. A test with an imbalanced set after training revealed a bias towards the now-majority category, in the opposite direction of PICC. Whether or not this is a challenge for AI cognition depends on the perspective of what AI models are supposed to achieve within cognitive science. If AI models are to be unified theories of cognition (Binz et al., 2024), their failure to exhibit PICC suggests that current architectures may lack the mechanisms necessary to capture PICC.

An important consideration for interpreting our results is the role of feedback during learning and decision making. Recent findings suggest that PICC tends to emerge primarily in the absence of external feedback, where people are not told the 'correct' category labels after making a judgment. In contrast, the presence of feedback can produce opposite effects (Lyu et al., 2021). Our model was trained with and evaluated without feedback, analogous to experimental paradigms investigaing PICC effects in humans. Nevertheless, it did not exhibit PICC-like flexibility. One plausible explanation is that human adaptability under changing prevalence relies on internal feedback mechanisms, allowing individuals to adjust decision boundaries even without external correction. Supporting this view, studies on categorization of emotions have shown that both adults and children dramatically shift category boundaries when external feedback is removed and the prevalence of emotions shifts (Plate et al., 2019, 2023), despite prior successful learning of consistent boundaries. Future work might explore models incorporating additional processes to better capture the dynamics of human categorization.

More broadly, our findings contribute to ongoing concerns about the cognitive validity of behavioral mimicry in AI systems. Although artificial neural networks can often be tuned to reproduce specific behavioral patterns, such tuning alone does not imply that the underlying cognitive mechanisms are captured. Our exploration is consistent with concerns raised in both model evaluation research (Roberts & Pashler, 2000) and cognitive architecture theory (Sun, 2008) that simply achieving behavioral mimicry is not sufficient for establishing cognitive validity. This distinction becomes especially important in light of our findings, which illustrate that sequential learning alone does not reproduce key aspects of human adaptive categorization.

While AI models are in many ways motivated by human cognition, they still might be lacking key mechanisms of human categorization. Many current AI architectures, particularly those based on statistical learning, do not inherently incorporate the flexible, history-dependent adjustments observed in human categorization. This is consistent with theories like range–frequency theory and decision by sampling, according to which human categorization relies on relative comparisons within local contexts rather than solely on the absolute stimulus value. If AI models learn only static decision thresholds or frequency-based rankings, they will fail to exhibit PICC because the effect depends on a dynamic shift of the reference point rather than absolute prevalence shifts alone.

Limitations

This study presents a theoretical discussion of PICC and highlights open questions regarding its application to social categorization and AI cognition. However, some limitations should be acknowledged.

Firstly, although this paper raises theoretical concerns about the domain-independence of PICC, it does not provide direct empirical tests of alternative explanations. Experimental studies specifically designed to test whether PICC operates independently of learned social biases or whether it is shaped by structured priors and contextual influences are needed.

Secondly, the AI simulation presented here is somewhat limited in scope. While we relied on a standard and widely used sequential learning model, there are many different AI architectures and even within the same architecture, many variations and specifications are possible. It is impossible to rule out that more complex AI architectures could capture PICC under different training conditions. Nevertheless, the present work illustrates that PICC does not arise under reasonable specifications and simple bottom-up mechanisms within the recurrent neural network architecture and LSTM.

Conclusion

PICC has been widely interpreted as evidence for adaptive, domain-independent category learning, but findings from social psychology suggest that its expression may be shaped by learned priors and social constraints. Our results show that standard AI models of sequential learning fail to reproduce PICC under relevant statistical learning conditions, suggesting that current computational frameworks may not fully capture the necessary mechanisms. The theoretical status of PICC as a phenomenon remains an open question, with future studies needing to examine to what extent it is a fundamental cognitive bias or a context-dependent adaptation.

Code availability

The code that reproduces all analyses reported here is available on the Open Science Framework at `https://osf.io/w3c9n/`.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, U., & Waldron, E. M. (1992). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *99*(3), 411–433. doi: 10.1037/0033-295X.99.3.411

Bhui, R., & Gershman, S. J. (2018, November). Decision by sampling implements efficient coding of psychoeconomic functions. *Psychological Review*, *125*(6), 985–1001. doi: 10.1037/rev0000123

Binz, M., Saanum, T., Élteto, N., Dayan, P., & Schulz, E. (2024). Centaur: a foundation model of human cognition. *arXiv preprint arXiv:2410.20268*. Retrieved from https://arxiv.org/abs/2410.20268

Botvinick, M., Wang, J. X., & Dabney, W. (2019). Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, *23*(5), 408–422.

Devine, S., Germain, N., Ehrlich, S., & Eppinger, B. (2022). Changes in the prevalence of thin bodies bias young women's judgments about body size. *Psychological Science*, *33*(8), 1212–1225.

Devine, S., Germain, N., Kasprzyk, A., & Eppinger, B. (2024). Changes in the prevalence of muscular, but not thin, bodies bias young men's judgments about body size. *Psychology of Men & Masculinities*.

Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing black: race, crime, and visual processing. *Journal of personality and social psychology*, *87*(6), 876.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Griffiths, T. L., & Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. *Cognition*, *103*(2), 180–226. doi: 10.1016/j.cognition.2006.03.004

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. doi: 10.1017/S0140525X0999152X

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253.

Levari, D. E. (2022). Range-frequency effects can explain and eliminate prevalence-induced concept change. *Cognition*, *226*, 105196.

Levari, D. E., Gilbert, D. T., Wilson, T. D., Sievers, B., Amodio, D. M., & Wheatley, T. (2018). Prevalence-induced concept change in human judgment. *Science*, *360*(6396), 1465-1467. doi: 10.1126/science.aap8731

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, NY: Wiley.

Lyu, W., Levari, D. E., Nartker, M. S., Little, D. S., & Wolfe, J. M. (2021). Feedback moderates the effect of prevalence on perceptual decisions. *Psychonomic Bulletin & Review*, *28*(6), 1906–1914.

Mendoza-Denton, R., Ayduk, O. z., Mischel, W., Shoda, Y., & Testa, A. (2001). Cognitive-affective processing systems as a model for predicting race-based bias. *Journal of Personality and Social Psychology*, *81*(1), 85–100. doi: 10.1037/0022-3514.81.1.85

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. doi: 10.1037/0096-3445.115.1.39

Parducci, A. (1995). *Happiness, pleasure, and judgment: The contextual theory and its applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Plate, R. C., Wood, A., Woodard, K., & Pollak, S. D. (2019). Probabilistic learning of emotion categories. *Journal of Experimental Psychology: General*, *148*(10), 1814.

Plate, R. C., Woodard, K., & Pollak, S. D. (2023). Category flexibility in emotion learning. *Affective Science*, *4*(4), 722–730.

Rhodes, M., Leslie, S.-J., Bianchi, L., & Chalik, L. (2018). The role of generic language in the early development of social categorization. *Child Development*, *89*(1), 148–155. doi: 10.1111/cdev.12714

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? a comment on theory testing. *Psychological Review*, *107*(2), 358–367.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive psychology*, *53*(1), 1–26.

Sun, R. (2008). The importance of cognitive architectures: An analysis based on CLARION. *Philosophical Psychology*, *21*(2), 139–155.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Woodford, M. (2020). Modeling imprecision in perception, valuation, and choice. *Annual Review of Economics*, *12*, 579–601. doi: 10.1146/annurev-economics-102819-040518