

A small cognitive core explains human behavior at scale

Marcel Binz^{1,2*}, Eric Schulz¹

¹Institute for Human-Centered AI, Helmholtz Munich, Neuherberg, Germany

²Max Planck Institute for Biological Cybernetics, Tübingen, Germany

*To whom correspondence should be addressed; E-mail: marcel.binz@tuebingen.mpg.de.

Psychology has proposed many principles of cognition. Each, however, was developed in isolation, leaving their combined explanatory power unknown. We assembled a small set of established principles into one model, which we call COntextual Retrieval and Encoding (CORE). CORE consists of a hierarchy of episodic memories into which prediction errors are written and whose encoding, retrieval, and forgetting depend on context. Fitted jointly to 7,036,442 choices from 137 experiments, the model predicts human choices better than black-box models and reproduces 75% of reported behavioral effects when simulated. Moreover, its latent variables show classic signatures of Bayesian learning and event segmentation. Taken together, these findings suggest that a small cognitive core accounts for much of human behavior. They also open a path toward making psychology more than the sum of its parts, by building computational models on the entire experimental record.

Introduction

A chief aim of cognitive science is to discover the computational principles that give rise to human behavior. More than a century of research has produced many candidates, and there is much we already know about individual mechanisms that seem important. For example, we know that associations strengthen when events co-occur (1), that learning is driven by prediction errors (2), and that representations are organized hierarchically (3). What is missing, however, is an estimate of their combined explanatory power: how much of human behavior can these principles, taken together, explain?

The most direct way to answer this question is to build a single computational model based on a small set of such principles and test it against human behavior across a broad range of experiments (4, 5). The ingredients for this are now finally in place: large-scale data sets of human behavior, expressive and meaningful computational models, and the compute to actually train them. A model of this kind must meet two requirements. It must be built from cognitively meaningful principles, and it must capture behavior across paradigms. To date, no model meets this standard. Existing approaches fall into two categories: domain-specific cognitive models that are interpretable but confined to a single paradigm (2, 6, 7) and large-scale predictive models that capture behavior across paradigms but remain black-box engines (8).

We introduce the **CO**ntextual **R**etrieval and **E**ncoding (CORE) model, combining the interpretability of cognitive models with the breadth of predictive models. CORE consists of a hierarchy of episodic memories into which prediction errors are written and whose encoding, retrieval, and forgetting are all context dependent. Each individual component has a long history in cognitive science (1–3, 9–12), but their joint explanatory power has not previously been tested across a broad range of human behavior. CORE operates directly on natural language transcripts of experimental sessions, allowing the same model to be fitted across paradigms.

We show that this small set of principles actually brings us quite far in explaining human behavior. Fitted jointly to 137 experiments involving 7,036,442 choices in 80,452 experimental sessions, CORE predicts human choices well compared to standard machine learning architectures, such as Transformers (13), and black-box foundation models of cognition, such as Centaur (8). Furthermore, we find that, when simulated, CORE reproduces 75% (50 out of 67) of behavioral effects identified in the corresponding studies.

Taken together, our work suggests that a surprisingly small cognitive core can account for much of human behavior, and establishes a framework for testing theories of cognition not one experiment at a time, but against large portions of the behavioral record at once. We believe that this approach can move psychology from a loose collection of theories toward an actually unified account of the human mind.

Results

Contextual Retrieval and Encoding (CORE). Let us suppose we wanted to build a cognitively meaningful model that captures human behavior across the full range of psychological experiments. Which components should such a model contain?

First, it needs to handle experiments that differ widely in their stimuli, instructions, and response formats. Natural language provides an input format that allows us to express many different experimental paradigms, making it a suitable choice for the model’s representation. Next, the model needs a memory that can store experiences and retrieve them when they become relevant. Importantly, what is being stored, retrieved, and forgotten should depend on the current context, because what is relevant may change over time. The model should also rely on what it already knows instead of repeatedly storing the same information. Finally, it should represent information in an efficient, hierarchical manner, from simple low-level features to more abstract concepts.

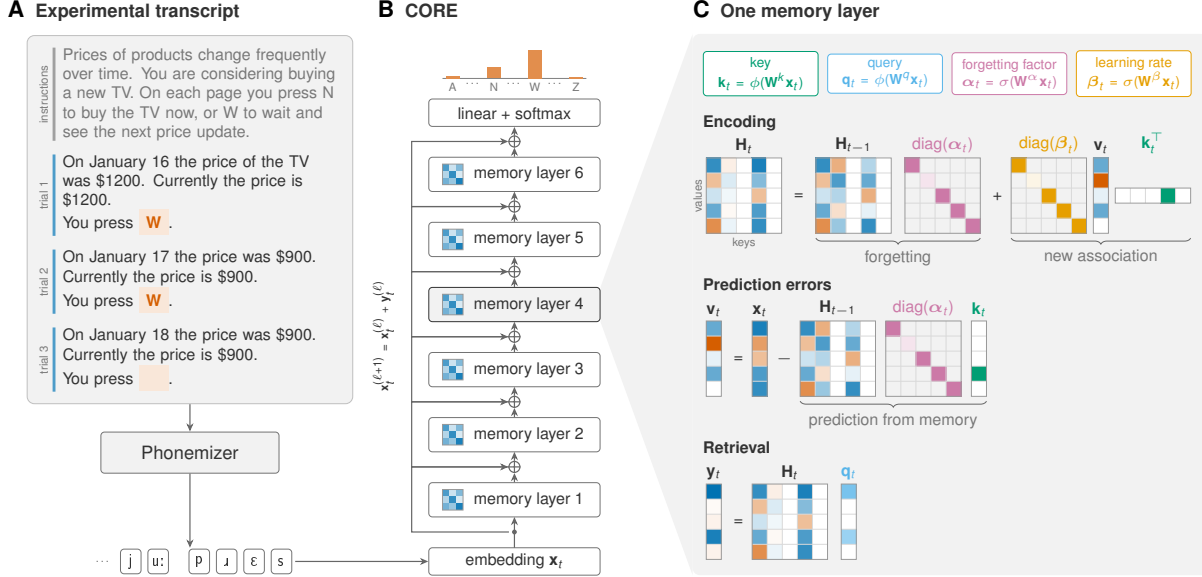


Fig. 1. The CORE model. (A) CORE reads the transcript of an experimental session (instructions, stimuli, feedback, and the participant’s own responses) converted into a sequence of phonemes. (B) The phonemes are embedded and passed through a stack of memory layers, each of which adds its output to its input; a linear read-out followed by a softmax predicts the next button press. (C) Each memory layer computes a key, a query, a forgetting factor, and a learning rate from its current input, which makes every operation contextual. Encoding: the memory is decayed feature-wise and a new association between a value and a key is added, scaled by the learning rate (Eq. 1). Prediction errors: the value is the difference between the input and what the memory predicts for the current key (Eq. 4). Retrieval: the query reads out the written values whose keys match it (Eq. 3).

CORE brings these components together in a single model, building on ideas with a long history in cognitive science (Fig. 1). It reads a transcript of an experimental session that contains all the information the participant had access to (instructions, stimuli, previous responses, and feedback) and responds with button presses (A–Z). When processing an experiment, CORE first converts its transcript into a sequence of phonemes, in line with models of speech perception that operate on phoneme-based representations (12). Each phoneme s_t is then mapped onto a learned d -dimensional embedding $\mathbf{x}_t$ (14, 15).

Following distributed accounts of human memory, information is written into an episodic

memory that takes the form of an associative matrix $\mathbf{H}_t \in \mathbb{R}^{d \times d}$ (9–11). In each time-step, the existing memory content $\mathbf{H}_{t-1}$ is multiplied by a feature-wise forgetting factor α_t and a new association between a value $\mathbf{v}_t$ and a key $\mathbf{k}_t$ is encoded using a Hebbian outer-product rule (1, 9, 10, 16):

$$\mathbf{H}_t = \mathbf{H}_{t-1} \text{diag}(\alpha_t) + \text{diag}(\beta_t) \mathbf{v}_t \mathbf{k}_t^\top \quad (1)$$

What is written, and how strongly, depends on context (17, 18). Keys, forgetting factors, and learning rates β_t are all computed from the current input:

$$\begin{aligned} \mathbf{k}_t &= \phi(\mathbf{W}^k \mathbf{x}_t) \\ \alpha_t &= \sigma(\mathbf{W}^\alpha \mathbf{x}_t) \\ \beta_t &= \sigma(\mathbf{W}^\beta \mathbf{x}_t) \end{aligned} \quad (2)$$

where ϕ denotes a nonlinearity followed by normalization to unit length, and σ a nonlinearity that maps onto values between zero and one.

To read from memory, CORE computes a context-dependent query $\mathbf{q}_t = \phi(\mathbf{W}^q \mathbf{x}_t)$. Echoing classic models of memory retrieval (19), it then returns the stored values, weighted by the similarity between their keys and the query:

$$\begin{aligned} \mathbf{y}_t = \mathbf{H}_t \mathbf{q}_t &= \left(\sum_{i=1}^t \text{diag}(\beta_i) \mathbf{v}_i \bar{\mathbf{k}}_i^\top \right) \mathbf{q}_t = \sum_{i=1}^t \text{diag}(\beta_i) \mathbf{v}_i \left(\bar{\mathbf{k}}_i^\top \mathbf{q}_t \right) \\ \bar{\mathbf{k}}_i &= \text{diag}(\alpha_t) \cdots \text{diag}(\alpha_{i+1}) \mathbf{k}_i \end{aligned} \quad (3)$$

where $\bar{\mathbf{k}}_i$ is the key written at step i , with each feature decayed by the forgetting that has happened since.

What is written into memory are not the inputs themselves but — as proposed in associative learning theory (2) — prediction errors between the input and what the memory retrieves for the current key (20, 21):

$$\mathbf{v}_t = \mathbf{x}_t - \underbrace{\mathbf{H}_{t-1} \text{diag}(\alpha_t) \mathbf{k}_t}_{\text{prediction from memory}} \quad (4)$$

Lastly, CORE stacks multiple of these associative memories on top of each other, reflecting the idea that cognition operates over multiple levels of representation (3). The output of the final layer is transformed with a linear mapping followed by a softmax into a probability for each button press.

We fit CORE using maximum a posteriori estimation on `Psych-201`, a large-scale collection of psychological experiments from domains such as reinforcement learning, decision-making, category learning, and social interaction (22). For our analyses, we focus on experiments in which participants respond via button presses (137 experiments from 62 studies; the training set comprises 7,036,442 choices in 80,452 experimental sessions).

To evaluate the resulting models, we rely on two metrics. The negative log marginal likelihood (NLML) measures how improbable the data are when averaging over all possible weights of a model, weighted by their prior probability (23, 24). The predictive negative log-likelihood (PNLL) measures how well the fitted model predicts human choices in held-out sessions. We compare CORE against standard machine learning architectures — such as Transformers (13) — and black-box foundation models of cognition — such as Centaur (8) — as well as against ablated versions of itself.

CORE captures human choices across paradigms. We found that CORE predicts human choices well compared to several baselines (Fig. 2A–B). It achieves both a lower negative log marginal likelihood (3,498,133 vs. 4,592,031) and a lower predictive negative log-likelihood (262,572 vs. 329,043) than a Transformer-based sequence model of matched configuration. CORE also predicts held-out choices better than Centaur (262,572 vs. 389,151) — a large language model fine-tuned on human behavior that has been shown to outperform domain-specific cognitive models. These results hold broadly across studies: CORE predicts held-out choices better than the Transformer in all 62 studies, and better than Centaur in 47 of them (Fig. 2C–D).

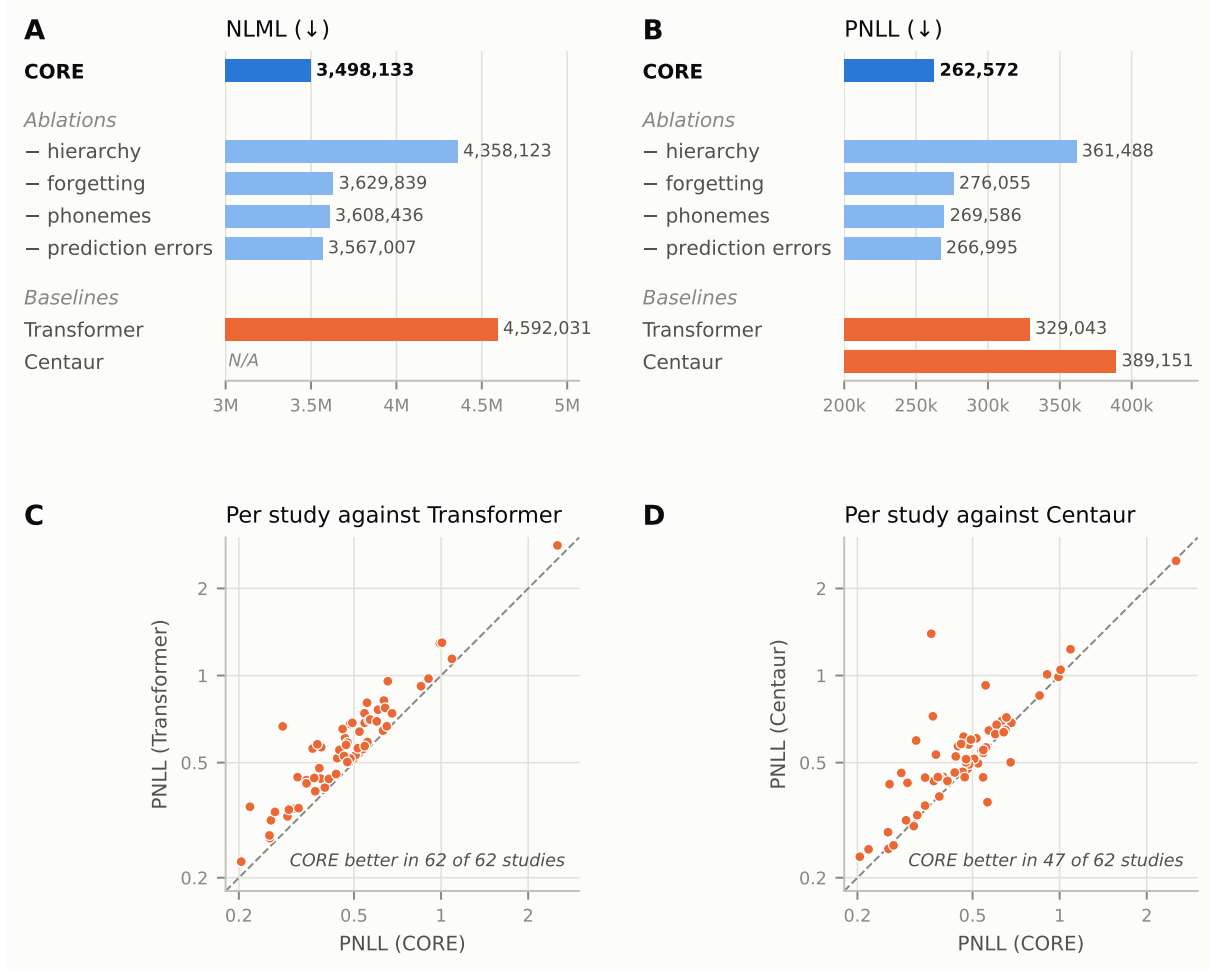


Fig. 2. Model comparison. (A) Negative log marginal likelihood (NLML) and (B) predictive negative log-likelihood (PNLL) of CORE, its ablations, and the baselines; lower is better. (C) PNLL per choice of CORE against the Transformer and (D) against Centaur in each of the 62 studies; points above the diagonal are studies in which CORE predicts held-out choices better.

CORE relies on all of its components. To test whether each of CORE’s components is necessary, we removed them one at a time and compared CORE against each ablated version (Fig. 2A–B). The hierarchy proved to be the most important component (Δ NLML = 859,990, Δ PNLL = 98,916). That is because, with a single layer, the model lacks the expressivity to capture human behavior. Forgetting came in second (Δ NLML = 131,706, Δ PNLL = 13,483). In its absence, old associations never fade and can only be overwritten by new ones. Phoneme-based encoding outperformed byte-pair encoding (25), the standard tokenization scheme in modern language models (Δ NLML = 110,303, Δ PNLL = 7,014). Even though the byte-pair model fit the training data better (negative log-likelihood of 2,904,051 vs. 3,066,566), it generalized worse and used considerably more embedding parameters (1,750,016 vs. 11,136). Finally, writing prediction errors instead of raw inputs improved performance as well (Δ NLML = 68,874, Δ PNLL = 4,423). Whereas writing raw inputs adds familiar content again, prediction errors contain only what the memory has not yet learned, thereby freeing up capacity for other content.

CORE reproduces behavioral effects. Thus far, we have shown that CORE fits human choices well, which is a necessary but not sufficient requirement for a model of human cognition. Equally important is that CORE, when simulated, reproduces the behavioral effects observed in humans (26). We therefore simulated CORE on the experiments of `Psych-201` and tested whether its simulated data show the same behavioral effects reported in the original studies. To do this at scale, we used an agentic pipeline that implemented a simulator for each experiment and an analysis script that tests the paper’s main effects on the resulting data. In total, we obtained 67 effects from 47 experiments of 32 studies after filtering. We simulated as many participants as in the original study and counted an effect as reproduced if it was statistically significant in the direction reported in the paper (or, for predicted null effects, if it was not significant).

Domain	Reproduced	Experiment	Not reproduced
Exploration	11/11 (100%)	—	—
Reinforcement learning	10/11 (91%)	chambon.2020_information (27)	Counterfactual feedback improves performance in a two-armed bandit.
Risky and intertemporal choice	9/13 (69%)	olschewski.2024_frequent (28)	Preference for the more frequent winner in a sampling task with equal-value options.
		olschewski.2024_frequent (28)	Preference for the more frequent winner in a sampling task with identical outcome distributions.
		olschewski.2024_frequent (28)	Preference for the right-skewed option in a sampling task without a frequent winner.
		spektor.2024_absolute (29)	Fewer risky choices in a loss-averse than in a gain-seeking context.
Planning	4/6 (67%)	kool.2016_when (30)	Model-based planning signature in a two-step task.
		feherdasilva.2023_rethinking (31)	Intuitive task instructions increase model-based planning in a two-step task.
Function and category learning	9/14 (64%)	flesch.2018_comparing (32)	Blocked training reduces interference between rules in context-dependent categorization (cardinal rules).
		flesch.2018_comparing (32)	Blocked training reduces interference between rules in context-dependent categorization (diagonal rules).
		flesch.2018_comparing (32)	Blocked training beats interleaved training in context-dependent categorization.
		binz.2022_heuristics (33)	No increase in accuracies over trials in multi-cue paired comparison with four unknown cues.
		binz.2022_heuristics (33)	Increased accuracies over trials in multi-cue paired comparison.
Language and social judgment	7/12 (58%)	gnther.2022_patterns (34)	Existing compounds interpreted correctly in a relational judgment task.
		gnther.2022_patterns (34)	Novel compounds interpreted systematically in a relational judgment task.
		dentella.2023_systematic (35)	Ungrammatical sentences detected above chance in grammaticality judgment.
		cheung.2024_large (36)	Framing effect in sacrificial moral dilemmas.
		evangelidis.2022_upscaling (37)	Inferior third option shifts preference toward quality over price in consumer choice.
Total	50/67 (75%)		

Table 1. Behavioral effects reproduced by simulating CORE, grouped by domain, with every effect that was not reproduced listed. Missed effects are colored by failure mode: **requires knowledge from outside the experiment**, **learning dynamics**, or **right direction but not significant**.

CORE reproduced 50 of the 67 effects (75%; Table 1), indicating that it also mirrors qualitative patterns of human behavior. We observed the best match in exploration and reinforcement learning, likely because these paradigms are among the most prevalent in the training data. CORE performed worst on language and social judgment. We believe that this may be due to the fact that many of these tasks rely on knowledge that people acquire outside the laboratory, about language, numbers, and the world. CORE does not have access to such background knowledge because it has only ever seen transcripts of psychological experiments. CORE also missed several effects that concern learning dynamics: the model-based planning signature in one of three two-step experiments, the benefits of blocked over interleaved training in context-dependent categorization, and learning over trials in a condition where people did not learn. Finally, six of the missed effects were in the right direction but did not reach significance. We found that some of them are fragile in the human data as well: resampling the human participants at the original sample size reproduced them in only 65–85% of cases.

Looking at the ablated versions, we found that a variant without hierarchy reproduced only 22 of the 67 effects: it retained simple choice preferences, but lost most effects that require learning within an experiment. This is expected, since a single layer can only recognize familiar tokens but not learn sequential contingencies between choices and outcomes. The variant without forgetting reproduced 42 of the 67 effects. It mainly lost effects that depend on the information of the current trial or context, such as the dependence of risk taking on the current context, learning in contextual bandits, and the use of the current gamble’s features in risky choice.

CORE’s latent variables are psychologically meaningful. CORE’s latent variables have a direct psychological interpretation. To test whether they are indeed meaningful, we recorded and analyzed them while the model performed the simulated experiments.

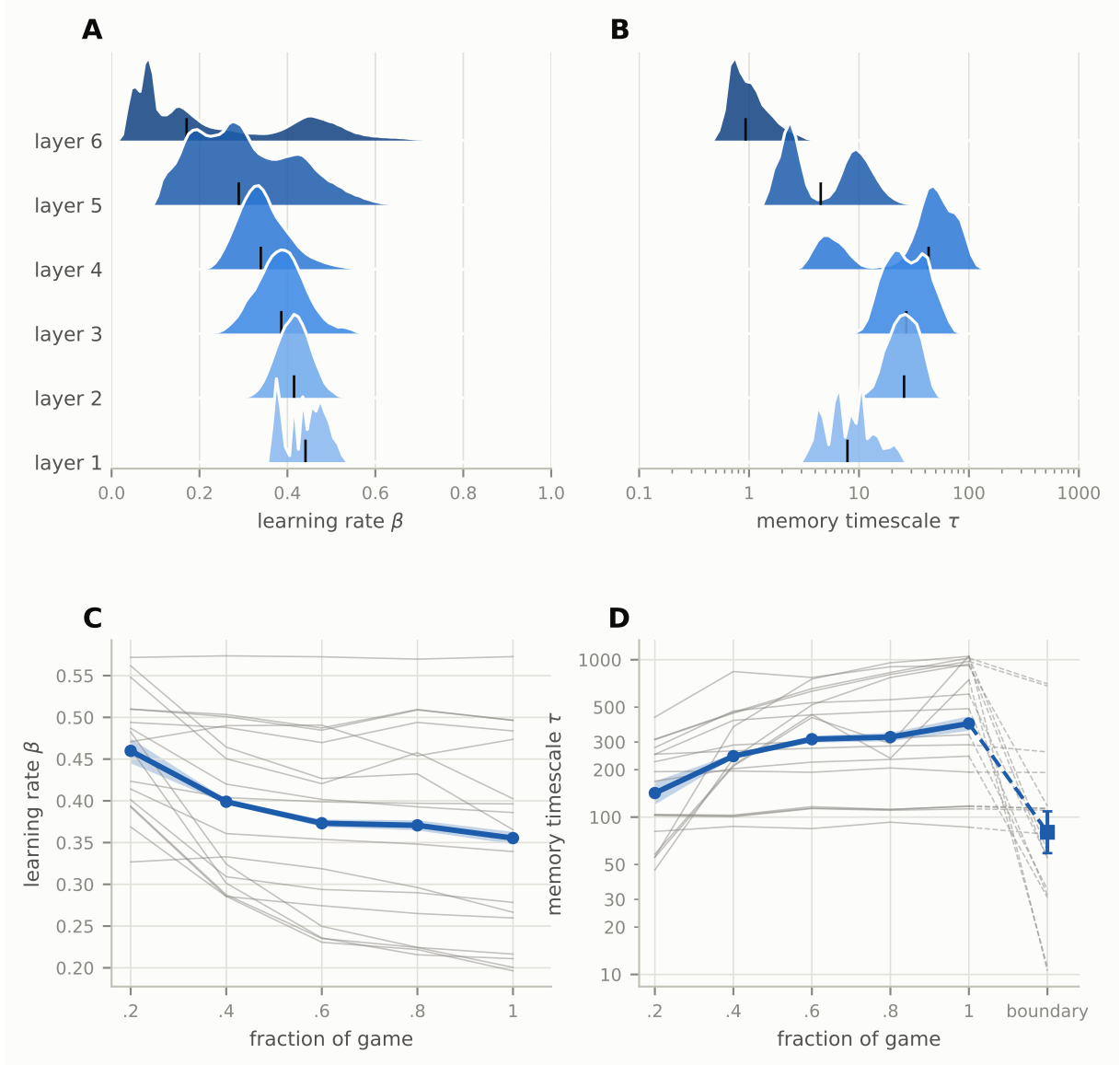


Fig. 3. Latent variable analysis. (A) Learning rate β_t and (B) memory timescale $\tau = -1/\log \alpha_t$ in tokens (log scale) per layer, over all tokens of the simulated experiments; black ticks mark the medians. (C) Learning rate and (D) memory timescale (log scale) of a single dimension of the fifth layer over the course of a game, each the dimension with the largest consistent change (see Methods). Values are averaged over the feedback that follows each choice. Each game is divided into five bins of equal relative length. The boundary is the text from the end of a game's last feedback up to the first choice of the next game. Thick lines show the mean over experiments and thin lines show single experiments. Bands and error bars show $\pm$ SEM over experiments of the change within participants.

We found that both learning rates and memory timescales $\tau = -1/\log \alpha_t$ (the number of tokens after which a stored association decays to $1/e$) varied within as well as across layers (Fig. 3A–B). Lower layers showed narrow, unimodal distributions, suggesting domain-general processes that are invariant across contexts. Higher layers showed broad and often multimodal distributions, suggesting more specialized, context-specific processes. Learning rates decreased with depth (Fig. 3A; median 0.44 in the first vs. 0.17 in the last layer), indicating that lower layers track the most recent input, whereas higher layers form running averages that capture the structure of an experiment. Memory timescales increased from the first to the fourth layer (Fig. 3B; median 8 vs. 43 tokens). Like for the learning rates, this points to a hierarchy from fast, local processing to slow, integrative processing.

Lastly, we asked whether CORE’s latent variables show signatures of classical learning and memory phenomena. For this, we focused on experiments that consist of a sequence of independent games, such as bandit problems with new reward probabilities or learning problems with a new underlying rule. For every choice, we averaged the learning rate and memory timescale of the penultimate layer over the feedback that followed it.

Bayesian models of learning, such as the Kalman filter, predict that learning rates should decrease as evidence accumulates (38, 39). CORE’s learning rates also exhibited this decline. Fig. 3C shows the dimension in which this effect was strongest. For this dimension, learning rates decreased in 15 of the 17 experiments. The decline was not limited to this dimension alone: learning rates decreased on average in 89 of the 128 dimensions of the layer.

Event segmentation theory holds that people maintain a model of the current event, which they build up while the event unfolds and reset when a new event begins (40). Likewise, CORE’s memory timescales grew within a game and collapsed at the boundaries between games. Fig. 3D shows the dimension for which this effect was strongest: its timescale dropped by 80% at the boundary to the next game. This effect was robust across experiments (17 of 17) and generalized

across the layer: on average, 94 of the 128 dimensions had a shorter timescale at the boundary than at the end of the game.

Discussion

Psychology has no shortage of principles of human cognition, and new ones are added every year. Many of them, however, are narrow in explanatory power, i.e., they fail to generalize beyond the specific paradigm they were developed for (41). Instead of adding more principles, now is the time to ask a fundamentally different question: how far do the principles we already have take us in understanding the human mind?

To address this question, we took some of psychology’s most established principles, assembled them into a single model, and measured how much of human behavior they can actually explain. We found that CORE predicts human choices better than standard machine learning architectures and a foundation model of human cognition with five orders of magnitude more parameters. Even more strikingly, CORE reproduced 75% of the behavioral effects reported in the original studies, indicating that a small set of carefully selected principles can bring us quite far.

CORE’s main components — phonemic input, associative memory, context dependence, prediction errors, and hierarchical representations — are all well established in cognitive modeling and neuroscience. The concept of storing knowledge in associative memories has a long history, including the linear associator (9, 10), Hebbian learning (1), Hopfield networks (16, 42), the context maintenance and retrieval model (18), and others (17). Prediction errors are central to many theories of human learning, from general frameworks such as predictive coding (43) down to specific implementations like the Rescorla–Wagner model (2). Hierarchical representations feature in many lines of research, including the classic work on receptive fields and modern connectionist models (3). CORE’s components also map onto proposals for how the

brain implements memory (44, 45). For instance, Kozachkov et al. have shown how associative memories — like the one CORE uses — could be implemented by networks of neurons and astrocytes (46, 47). More broadly, CORE repeats a single type of memory layer across its hierarchy, much like the neocortex is thought to repeat a canonical local circuit across areas (48, 49).

CORE furthermore borrows ideas from machine learning, in particular linear attention (50), fast weight programmers (20, 51), and Gated Delta Networks (21, 52). What distinguishes CORE from these models is what it leaves out: it contains no computation besides writing to and retrieving from memory, which makes it possible to attribute its behavior to a handful of principles, and nothing else. Parameter estimation at this scale relied on the flash-linear-attention library (53), extended with a custom kernel that cut the run time from about 10 days to about 24 hours on a single GPU.

We have shown that a few established principles can take us quite far, but there is still some way to go. What is needed to capture the remaining 25%? CORE struggles with model-based planning, as shown by our model simulations (54). This makes sense, given that it contains no explicit planning mechanism. Future model iterations could incorporate such a mechanism in several ways, for example using ideas from looped Transformers (55). There are also mechanisms that our method cannot detect at present, simply because the data do not support it. Long-term memory is one example. Experiments in our data set typically span less than an hour, so how participants learn and store information over weeks and months is never observed. Longitudinal data would allow us to add a long-term memory to CORE, which would make it possible to study how such a memory differs from CORE’s episodic memory and how consolidation transfers information from one to the other (56).

Longitudinal data of this kind are easiest to collect outside the laboratory, in naturalistic environments that people revisit over weeks and months. While traditional cognitive models resist such environments, CORE is stimulus-computable (57): it operates directly on natural

language and could be readily extended to other sensory modalities.

CORE's principles are interpretable, but its computations are not. Instead of a full mechanistic account, CORE is, in its current form, best seen as an intermediate step between classical cognitive models and black-box systems (58). There are still components of CORE that lack a theoretical grounding, e.g., how it computes keys, values, learning rates, and forgetting. Future models should retain the broad behavioral coverage of CORE while replacing learned transformations with independently motivated computational principles, thereby moving from an interpretable architecture toward a complete mechanistic account.

Conclusion

Imagine a future in which every behavioral experiment ever conducted is available in a unified format. How should we develop theories of the mind in such a world? For most of its history, psychology has tested its theories one experiment at a time (59). This leaves most of the evidence unused: each theory is constrained by a few experiments, not by the entire behavioral record of psychology. What we have pursued amounts to a radically different paradigm: principles are no longer tested in isolation, but assembled into a single model that is fitted jointly to many experiments.

CORE is an initial step in this direction. It shows that a small set of established cognitive mechanisms can account for a surprisingly broad range of human behavior when fitted jointly across experiments. The opportunity, however, is larger than any single model. If every experiment ever conducted constrains a shared model of the mind, psychology may finally become more than the sum of its experiments.

Materials and Methods

Data. We used a version of `Psych-201` re-transcribed by an agentic pipeline. Each experiment was converted into a natural-language transcript using `DeepSeek-V4-Flash` (`deepseek-v4-flash-0731`) (60) and verified using `Claude Fable 5.1` (`claude-fable-5-1`) (61). We release this version on the Hugging Face Hub: <https://huggingface.co/datasets/marcelbinz/Psych-201-discrete-agentic> (training set) and <https://huggingface.co/datasets/marcelbinz/Psych-201-discrete-agentic-test> (test set). Transcripts contain all the information participants had access to and were never truncated. Human choices are enclosed in response markers (`[HUMAN_RESPONSE]...[/HUMAN_RESPONSE]`). We kept these markers only around button presses and removed them around all other responses, which remain in the transcript as text but are not predicted; sessions without any button press were dropped. The resulting data set comprises 137 experiments from 62 studies (27–37, 62–112), with 7,634,315 choices in 83,352 sessions. Each button press was represented by a single token, and the loss was computed on these tokens only. To this end, we added one separate token per button (A–Z) to the vocabulary of the tokenizer, so that a button press is never split into phonemes or subwords. For each study, we held out $\min(100, \lfloor 0.1 n \rfloor)$ of its n sessions as a test set (597,873 choices in 2,900 sessions).

Model. CORE converts each transcript into phonemes using the eSpeak-based phoneme tokenizer of `wav2vec 2.0` (15), and maps each phoneme onto a learned 128-dimensional embedding. The embeddings pass through six memory layers, each of which normalizes its input using `RMSNorm` (113) and adds the output of its memory to it. Each layer consists of four heads, i.e., independent memory modules that operate in parallel. The memory matrix of each head starts empty ($\mathbf{H}_0 = \mathbf{0}$) at the beginning of every session. Each head computes its own keys, queries, learning rates, and forgetting factors (32 dimensions each) from the full 128-

dimensional input; the outputs of the four heads are concatenated to form the output of the layer. Keys and queries are computed with a SiLU nonlinearity (114) followed by normalization to unit length, the learning rate β_t with a logistic function, and the forgetting factor α_t as $\exp(-a \cdot \text{softplus}(z + b))$ with learned $a > 0$ and b per feature. A linear read-out followed by a softmax maps the output of the last layer onto the response buttons. The model contains no other components and has 452,608 parameters in total. The memory update was implemented as a chunked Triton kernel extending the flash-linear-attention library (53).

Training. All parameters were fitted jointly on the training set by maximum a posteriori estimation, using schedule-free AdamW (115) with a learning rate of 10^{-3} and a decoupled weight decay of 0.5. We trained for 100 epochs with batches of 304,500 tokens and a single random seed. Training took about 24 hours on a single NVIDIA H100 GPU and used the `transformers` library (116).

Model comparison. We approximated the log marginal likelihood with a Laplace approximation around the fitted weights, using a diagonal empirical Fisher information matrix over all parameters and a prior precision per layer that was optimized after training (117). The predictive log-likelihood was the summed log-probability of all held-out choices. The Transformer baseline used the same configuration as CORE (128 dimensions, six layers, four heads) but a byte-pair tokenizer (25). The ablations removed one component each: a single memory layer instead of six, no forgetting ($\alpha_t = 1$), Hebbian instead of error-driven encoding ($\mathbf{v}_t = \mathbf{x}_t$), and a byte-pair tokenizer instead of phonemes. We did not compute the marginal likelihood for Centaur, because it is unclear how its pretraining should be accounted for. Per-study likelihoods, behavioral effects, and example transcripts for all models can be explored at <https://core-cognition.github.io/core-model/results/>.

Experiment simulations. For each experiment, an agentic pipeline produced a simulator of the task, which generates transcripts in the same format as the data, and an analysis script that tests the main effects reported in the original paper. We retained experiments for simulation and subsequent analysis based on the following criteria. First, we excluded experiments whose simulator or original transcripts involve other responses, such as numbers, words, or free text. We then excluded effects whose analysis depended on information the simulators cannot produce, such as reaction times, age, gender, ratings, or clinical group membership. Finally, we kept only effects that the original human data reproduce under the same analysis. In total, this yielded 67 effects from 47 experiments of 32 studies. CORE was simulated by sampling its predicted distribution with a temperature of 1.0. We simulated as many participants as in the original study, up to 1,000. Failure modes from Table 1 were assigned by hand after scoring.

Latent-variable analysis. We analyzed the latent variables of CORE while it performed the simulated experiments. For each simulated session, we recorded the learning rate β_t and the forgetting factor α_t of every dimension and layer at every token. Forgetting factors were converted into memory timescales $\tau = -1/\log \alpha_t$ (i.e., the number of tokens after which a stored association has decayed to $1/e$ of its strength) and averages of memory timescales were computed on a logarithmic scale.

For Fig. 3A–B, we averaged the learning rates and log forgetting factors over the dimensions of each head. For each experiment, we computed histograms over all tokens and heads of a layer, and normalized them to sum to one. We then averaged the histograms over the simulated experiments, so that each experiment has the same weight regardless of its number of participants or tokens. Medians were computed from the averaged histograms.

To follow the latent variables over the course of a game, we used the simulated experiments that consist of a sequence of games (at least four games per participant with at least five choices

in a typical game). This resulted in 17 experiments from 9 studies. For Fig. 3C–D, we kept the individual dimensions instead of averaging over the dimensions of a head. For each choice, we averaged the latent variables over the feedback that follows it, from the choice to the end of the line that reports the outcome. For the boundary between two games, we averaged over the text between the feedback of a game’s last choice and the first choice of the next game, which introduces the new game and presents its first stimulus. We assigned each choice to one of five bins according to its position in the game, relative to the length of the game, so that the first and last choices of every game fall into the first and last bin; games with fewer than five choices were excluded. We averaged within participants, then over the participants of an experiment, and then over experiments, so that each experiment has the same weight. The shaded bands and error bars in Fig. 3C–D show the standard error over experiments of the change within participants, that is, after we subtracted each participant’s mean over the five bins.

References and Notes

1. D. O. Hebb, *The organization of behavior: A neuropsychological theory* (Wiley, New York, 1949).
2. R. A. Rescorla, *Classical conditioning, Current research and theory* **2**, 64 (1972).
3. D. E. Rumelhart, J. L. McClelland, P. R. Group, *et al.*, *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations* (The MIT press, 1986).
4. A. Newell, *Unified theories of cognition* (Harvard University Press, 1994).
5. J. R. Anderson, *et al.*, *Psychological review* **111**, 1036 (2004).
6. R. M. Nosofsky, *Formal approaches in categorization* pp. 18–39 (2011).

7. R. Ratcliff, *Psychological review* **85**, 59 (1978).
8. M. Binz, *et al.*, *Nature* **644**, 1002 (2025).
9. J. A. Anderson, *Mathematical biosciences* **14**, 197 (1972).
10. T. Kohonen, *IEEE Transactions on Computers* **C-21**, 353 (1972).
11. B. B. Murdock, *Psychological Review* **89**, 609 (1982).
12. J. L. McClelland, J. L. Elman, *Cognitive psychology* **18**, 1 (1986).
13. A. Vaswani, *et al.*, *Advances in neural information processing systems* **30** (2017).
14. M. Bernard, H. Titeux, *Journal of Open Source Software* **6**, 3958 (2021).
15. Q. Xu, A. Baevski, M. Auli, *arXiv preprint arXiv:2109.11680* (2021).
16. J. J. Hopfield, *Proceedings of the national academy of sciences* **79**, 2554 (1982).
17. M. W. Howard, M. J. Kahana, *Journal of mathematical psychology* **46**, 269 (2002).
18. S. M. Polyn, K. A. Norman, M. J. Kahana, *Psychological review* **116**, 129 (2009).
19. D. L. Hintzman, *Psychological review* **93**, 411 (1986).
20. I. Schlag, K. Irie, J. Schmidhuber, *International conference on machine learning* (PMLR, 2021), pp. 9355–9366.
21. S. Yang, J. Kautz, A. Hatamizadeh, *International Conference on Learning Representations* (2025), vol. 2025, pp. 29687–29707.
22. M. Binz, *et al.*, *arXiv preprint arXiv:2605.07632* (2026).
23. D. J. MacKay, *Neural computation* **4**, 448 (1992).

24. A. Immer, M. Bauer, V. Fortuin, G. Rätsch, K. M. Emtiyaz, *International Conference on Machine Learning* (PMLR, 2021), pp. 4563–4573.
25. R. Sennrich, B. Haddow, A. Birch, *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)* (2016), pp. 1715–1725.
26. S. Palminteri, V. Wyart, E. Koechlin, *Trends in cognitive sciences* **21**, 425 (2017).
27. V. Chambon, *et al.*, *Nature Human Behaviour* **4**, 1067 (2020).
28. S. Olschewski, M. S. Spektor, G. Le Mens, *Proceedings of the National Academy of Sciences* **121**, e2317751121 (2024).
29. M. S. Spektor, D. Kellen, J. Rieskamp, K. C. Klauer, *Journal of Experimental Psychology: General* **153**, 454 (2024).
30. W. Kool, F. A. Cushman, S. J. Gershman, *PLoS computational biology* **12**, e1005090 (2016).
31. C. Feher da Silva, G. Lombardi, M. Edelson, T. A. Hare, *Nature Human Behaviour* **7**, 956 (2023).
32. T. Flesch, J. Balaguer, R. Dekker, H. Nili, C. Summerfield, *Proceedings of the National Academy of Sciences* **115**, E10313 (2018).
33. M. Binz, S. J. Gershman, E. Schulz, D. Endres, *Psychological review* **129**, 1042 (2022).
34. F. Günther, M. Marelli, *Cognitive Psychology* **134**, 101471 (2022).
35. V. Dentella, F. Günther, E. Leivada, *Proceedings of the National Academy of Sciences* **120**, e2309583120 (2023).

36. V. Cheung, M. Maier, F. Lieder, *Proceedings of the National Academy of Sciences* **122**, e2412015122 (2025).
37. I. Evangelidis, J. Levav, I. Simonson, *Journal of Consumer Research* **50**, 492 (2023).
38. R. E. Kalman (1960).
39. P. Dayan, S. Kakade, *Advances in neural information processing systems* **13** (2000).
40. J. M. Zacks, N. K. Speer, K. M. Swallow, T. S. Braver, J. R. Reynolds, *Psychological bulletin* **133**, 273 (2007).
41. T. Yarkoni, *The Behavioral and brain sciences* **45**, e1 (2020).
42. D. Krotov, J. J. Hopfield, *Advances in neural information processing systems* **29** (2016).
43. R. P. Rao, D. H. Ballard, *Nature neuroscience* **2**, 79 (1999).
44. S. J. Gershman, I. Fiete, K. Irie, *Neuron* **113**, 1694 (2025).
45. K. Irie, S. J. Gershman, *arXiv preprint arXiv:2508.08435* (2025).
46. L. Kozachkov, K. V. Kastanenko, D. Krotov, *Proceedings of the National Academy of Sciences* **120**, e2219150120 (2023).
47. L. Kozachkov, J.-J. Slotine, D. Krotov, *Proceedings of the National Academy of Sciences* **122**, e2417788122 (2025).
48. R. J. Douglas, K. A. Martin, *Annu. Rev. Neurosci.* **27**, 419 (2004).
49. A. M. Bastos, *et al.*, *Neuron* **76**, 695 (2012).
50. A. Katharopoulos, A. Vyas, N. Pappas, F. Fleuret, *International conference on machine learning* (PMLR, 2020), pp. 5156–5165.

51. J. Schmidhuber, *Neural Computation* **4**, 131 (1992).
52. A. Hatamizadeh, Y. Choi, J. Kautz, *arXiv preprint arXiv:2605.22791* (2026).
53. S. Yang, Y. Zhang, FLA: A Triton-based library for hardware-efficient implementations of linear attention mechanism, <https://github.com/fla-org/flash-linear-attention> (2024).
54. N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, R. J. Dolan, *Neuron* **69**, 1204 (2011).
55. M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, Ł. Kaiser, *arXiv preprint arXiv:1807.03819* (2018).
56. J. L. McClelland, B. L. McNaughton, R. C. O'Reilly, *Psychological review* **102**, 419 (1995).
57. D. L. Yamins, J. J. DiCarlo, *Nature neuroscience* **19**, 356 (2016).
58. J. C. Peterson, D. D. Bourgin, M. Agrawal, D. Reichman, T. L. Griffiths, *Science* **372**, 1209 (2021).
59. A. Newell, *Machine intelligence* (Routledge, 2012), pp. 121–146.
60. A. Xu, *et al.*, *arXiv preprint arXiv:2606.19348* (2026).
61. Anthropic, System card: Claude Fable 5.1 & Claude Mythos 5.1, <https://www.anthropic.com/claude-fable-and-mythos-5-1> (2026).
62. E. Akata, *et al.*, *Nature Human Behaviour* **9**, 1380 (2025).
63. F. Anvari, S. Billinger, P. P. Analytis, V. R. Franco, D. Marchiori, *Nature Communications* **15**, 7721 (2024).

64. S. P. Badham, A. N. Sanborn, E. A. Maylor, *Psychology and Aging* **32**, 473 (2017).
65. B. Bahrami, J. Navajas, Y. Wang, S. Wang, 4 arm bandit task dataset, <https://osf.io/f3t2a/> (2020).
66. J. Barnby, N. Raihani, P. Dayan, *Cognition* **225**, 105098 (2022).
67. S. Bavard, M. Lebreton, M. Khamassi, G. Coricelli, S. Palminteri, *Nature communications* **9**, 4503 (2018).
68. S. Bavard, A. Rustichini, S. Palminteri, *Science Advances* **7**, eabe0340 (2021).
69. A. D. Breslav, *et al.*, *Psychological Science* **33**, 550 (2022).
70. X. Chen, *et al.*, *PLoS Computational Biology* **14**, e1006304 (2018).
71. B. Christian, E. M. Russek, T. L. Griffiths, *Proceedings of the National Academy of Sciences* **123**, e2509612123 (2026).
72. S. Ciranka, W. van den Bos, *Communications Psychology* **3**, 137 (2025).
73. A. O. Cohen, K. Nussenbaum, H. M. Dorfman, S. J. Gershman, C. A. Hartley, *npj Science of Learning* **5**, 16 (2020).
74. G. E. Cox, P. Hemmer, W. R. Aue, A. H. Criss, *Journal of Experimental Psychology: General* **147**, 545 (2018).
75. J. H. Decker, A. R. Otto, N. D. Daw, C. A. Hartley, *Psychological science* **27**, 848 (2016).
76. A. Dezfouli, K. Griffiths, F. Ramos, P. Dayan, B. W. Balleine, *PLoS computational biology* **15**, e1006903 (2019).
77. M. Dubois, T. U. Hauser, *Nature Communications* **13**, 4542 (2022).

78. H. Fan, S. J. Gershman, E. A. Phelps, *Nature Human Behaviour* **7**, 102 (2023).
79. S. F. Feng, S. Wang, S. Zarnescu, R. C. Wilson, *Scientific reports* **11**, 3077 (2021).
80. R. Frey, A. Pedroni, R. Mata, J. Rieskamp, R. Hertwig, *Science advances* **3**, e1701381 (2017).
81. B. Garcia, M. Lebreton, S. Bourgeois-Gironde, S. Palminteri, *Nature human behaviour* **7**, 611 (2023).
82. S. J. Gershman, *Cognition* **173**, 34 (2018).
83. F. Günther, M. Marelli, *Quarterly Journal of Experimental Psychology* **73**, 1082 (2020).
84. M. P. Gunadi, I. Evangelidis, *Journal of Marketing Research* **59**, 623 (2022).
85. N. Haines, *et al.*, *Clinical Psychological Science* **8**, 1017 (2020).
86. J. Heffner, O. FeldmanHall, *Nature communications* **13**, 1718 (2022).
87. A. Jagadish, M. Binz, T. Saanum, J. Wang, E. Schulz (2023).
88. R. A. Jansen, A. N. Rafferty, T. L. Griffiths, *Nature Human Behaviour* **5**, 756 (2021).
89. W. Kool, S. J. Gershman, F. A. Cushman, *Psychological science* **28**, 1321 (2017).
90. P. M. Krueger, F. Callaway, S. Gul, T. L. Griffiths, F. Lieder, *Psychological Review* **131**, 905 (2024).
91. K. R. Levering, N. Conaway, K. J. Kurtz, *Memory & cognition* **48**, 335 (2020).
92. M. Moutoussis, *et al.*, *PLoS computational biology* **14**, e1006679 (2018).
93. K. Nussenbaum, *et al.*, *Elife* **12**, e84260 (2023).

94. K. Nussenbaum, P. L. Katzman, H. Lu, S. Zorowitz, C. A. Hartley, *Psychological Science* **35**, 933 (2024).
95. S. Olschewski, T. L. Mullett, N. Stewart, *Cognitive Psychology* **157**, 101716 (2025).
96. A. Pirrone, Subjective utility modulates the effect of overall stimulus intensity on decision-making, <https://osf.io/xvnjd/> (2024). Unpublished manuscript.
97. O. Plonsky, *et al.*, *Nature Human Behaviour* **9**, 2271 (2025).
98. E. M. Russek, R. Moran, Y. Liu, R. J. Dolan, Q. J. Huys, *Nature Communications* **15**, 4269 (2024).
99. H. Sadeghiyeh, *et al.*, *Scientific reports* **10**, 4020 (2020).
100. E. Schulz, N. T. Franklin, S. J. Gershman, *Cognitive psychology* **119**, 101261 (2020).
101. N. Shahar, *et al.*, *PLoS computational biology* **15**, e1006803 (2019).
102. K. Singh, P. Aggarwal, P. Rajivan, C. Gonzalez, *Proceedings of the human factors and ergonomics society annual meeting* (SAGE Publications Sage CA: Los Angeles, CA, 2019), vol. 63, pp. 453–457.
103. H. Steingroever, *et al.*, *Journal of Open Psychology Data* **3** (2015).
104. P. Suthaharan, *et al.*, *Nature human behaviour* **5**, 1190 (2021).
105. A. I. Thoma, B. R. Newell, C. Schulze, *Journal of Experimental Psychology: General* **154**, 1523 (2025).
106. H. Vandendriessche, *et al.*, *Psychological Medicine* **53**, 4696 (2023).

107. B. Van Tiel, M. Franke, U. Sauerland, *Proceedings of the National Academy of Sciences* **118**, e2005453118 (2021).
108. K. Witte, M. Thalmann, E. Schulz (2024).
109. S. Wu, N. Éltető, I. Dasgupta, E. Schulz, *Scientific reports* **13**, 7680 (2023).
110. H. A. Xu, A. Modirshanechi, M. P. Lehmann, W. Gerstner, M. H. Herzog, *PLOS Computational Biology* **17**, e1009070 (2021).
111. S. Xu, X. Zhang, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), pp. 1–16.
112. S. Zorowitz, Y. Niv, Data from two-step task pilots, <https://osf.io/2tgjd/> (2023). Unpublished data.
113. B. Zhang, R. Sennrich, *Advances in neural information processing systems* **32** (2019).
114. S. Elfving, E. Uchibe, K. Doya, *Neural networks* **107**, 3 (2018).
115. A. Defazio, *et al.*, *Advances in Neural Information Processing Systems* **37**, 9974 (2024).
116. T. Wolf, *et al.*, *arXiv preprint arXiv:1910.03771* (2019).
117. E. Daxberger, *et al.*, *Advances in neural information processing systems* **34**, 20089 (2021).

Acknowledgments

Funding: This work was supported by the Helmholtz Association’s Initiative and Networking Fund on the HAICORE@FZJ partition. E.S. is supported by the European Research Council (ERC) through the TACOS grant (grant agreement No. 101116962). **Author contributions:** M.B. conceived the study, developed the model, performed the analyses, and wrote

the manuscript. E.S. edited the final draft and acquired funding. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The Psych-201 training and test sets are available on the Hugging Face Hub at <https://huggingface.co/datasets/marcelbinz/Psych-201-discrete-agentic> and <https://huggingface.co/datasets/marcelbinz/Psych-201-discrete-agentic-test>. Code is available at <https://github.com/CORE-COGNITION/core-model>.