scientific reports

OPEN

Neural superstatistics for Bayesian estimation of dynamic cognitive models

Lukas Schumacher[1]✉, Paul-Christian Bürkner[2], Andreas Voss[1], Ullrich Köthe[3] & Stefan T. Radev[4]

Mathematical models of cognition are often memoryless and ignore potential fluctuations of their parameters. However, human cognition is inherently dynamic. Thus, we propose to augment mechanistic cognitive models with a temporal dimension and estimate the resulting dynamics from a superstatistics perspective. Such a model entails a hierarchy between a low-level observation model and a high-level transition model. The observation model describes the local behavior of a system, and the transition model specifies how the parameters of the observation model evolve over time. To overcome the estimation challenges resulting from the complexity of superstatistical models, we develop and validate a simulation-based deep learning method for Bayesian inference, which can recover both time-varying and time-invariant parameters. We first benchmark our method against two existing frameworks capable of estimating time-varying parameters. We then apply our method to fit a dynamic version of the diffusion decision model to long time series of human response times data. Our results show that the deep learning approach is very efficient in capturing the temporal dynamics of the model. Furthermore, we show that the erroneous assumption of static or homogeneous parameters will hide important temporal information.

Mathematical models are important tools for conceptualizing human cognition and predicting observable behavior. Such models aim to provide a mathematical formalization of cognitive processes by mapping latent cognitive constructs to model parameters and specifying how these generate manifest data[1]. The surge of cognitive model applications has made it possible to test precise mechanistic hypotheses and to predict performance in various domains, such as decision-making[2,3], learning[4,5], or memory[6,7].

The majority of cognitive models treat human data as independent and identically distributed (IID) observations. The IID assumption implies that these models largely ignore the temporal changes of latent cognitive constructs. However, such constructs are inherently dynamic, regardless of a particular time scale[8–11]. For instance, there is little dispute that constructs, such as working memory capacity[12] or mental speed[13], change over the human life span. These constructs also vary on much shorter time scales, for example, within experimental sessions[14,15].

In psychological experiments, cognitive affordances are influenced not only by external task demands but also by internal mental processes and brain states that change over time. There are many possible explanations for the resulting systematic and unsystematic fluctuations, for instance, fatigue[16,17], practice[18,19], mind-wandering[20,21], or motivational factors[22,23]. In this article, we argue that cognitive mechanisms should be treated as complex dynamic systems and that cognitive models should account for the dynamics of their components to fully understand and capture the rich structure of empirical human data.

Ignoring temporal fluctuations and changes in cognitive parameters can have drastic consequences for the descriptive, explanatory, and predictive merits of cognitive models. Consider a simple inverted U-shape hypothetical trajectory of a single parameter, as depicted in Fig. 1. Typical cognitive models assuming IID observations[2,6] would estimate a flat trajectory (depicted in blue) whose uncertainty would match the width of the marginal parameter distribution (depicted in gray). Differently, dynamic models would account for temporal change and achieve a much greater information gain (depicted in red). Indeed, this is not just a hypothetical scenario, and we subsequently demonstrate its consequences in a real data application (cf. Fig. 8).

[1]Institute of Psychology, Heidelberg University, Heidelberg, Germany. [2]Department of Statistics, TU Dortmund University, Dortmund, Germany. [3]Computer Vision and Learning Lab, Heidelberg University, Heidelberg, Germany. [4]Cluster of Excellence STRUCTURES, Heidelberg University, Heidelberg, Germany. ✉email: lukas.schumacher@psychologie.uni-heidelberg.de

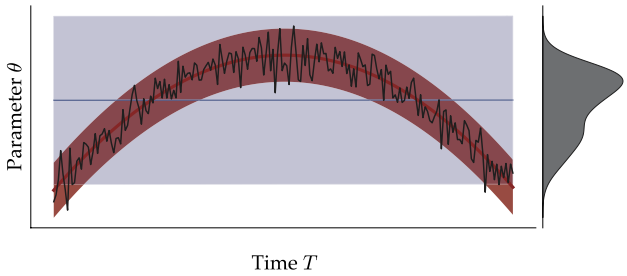

Figure 1. Conceptual illustration of a hypothetical parameter θ varying over time (solid black line). The solid blue line and shaded blue region depict the posterior mean and the 95% CI of a static model, respectively. The solid red line and shaded red region depict the posterior mean and 95% CI of a dynamic model, respectively. Treating the parameter as static (i.e., stationary) by marginalizing out the effects of time leads to inflated uncertainty estimates (matching the width of the marginal distribution, depicted in gray) and obscures the underlying change.

One way to mathematically formalize dynamic systems is by treating them as stochastic generative processes that produce data with temporal dependencies (i.e., time series data). As most complex systems are inherently non-linear, these time series often do not exhibit simple fluctuations around a stable mean with a fixed variance, but resemble a heterogeneous random walk[24]. Beck and Cohen[25] coined the term *superstatistics*, which refers to a superposition of multiple stochastic processes on different temporal scales that can describe heterogeneous temporal dynamics[26]. Thus, instead of assuming static model parameters, a superstatistics modeling approach introduces a hierarchy of at least two models: A low-level (i.e., observation or microscopic) model that formalizes the local behavior of a system and a high-level (i.e., transition or macroscopic) model that describes the parameter dynamics of the low-level model. Note that there is no absolute time scale for low- and high-level processes. The meaning of these terms is relative and always depends on the scale relevant to the research question.

A viable approach for modeling parameter transitions is offered by hidden Markov models (HMMs). For instance[27], accounted for different response states during a decision-making task by combining a HMM with an evidence accumulation model of decision-making. This model combination allows for discontinuous changes on longer time scales and continuous changes on shorter time scales. Similarly[28], extended a hierarchical version of the same decision-making model with a HMM and applied it to three existing long time series of response time and choice data. Both studies demonstrate that the HMM approach can reveal plausible fluctuations of decision model parameters in cognitive tasks.

However, the superstatistics framework is far more general and flexible in representing macroscopic fluctuations. First, it does not require modelers to pre-define a small set of possible modes (i.e., distinct system behaviors). Further, models within the superstatistics framework can be agnostic about the concrete dynamics of the model parameters—the most plausible dynamic can be directly estimated in a data-driven fashion. For example, using a superstatistics framework[29], demonstrated that the transition between different sleep stages is less abrupt than previously suggested.

The superstatistics framework has been utilized in physics[30–32], the life-sciences[33] and economics[34,35], but it has not yet been disseminated in the cognitive sciences. Under the assumption that cognitive processes are dynamic and complex, it seems natural to equip existing cognitive models with superstatistical aspects. However, to our knowledge, no previous study besides[29] has employed superstatistical methods for studying the dynamic aspects of cognitive parameters. Existing dynamic models of cognition fit stationary time series models (e.g., autoregressive models) to the observed behavior[9] but do not incorporate a low-level mechanistic model that formalizes the underlying cognitive process(es). Thus, these time series models describe how behavior changes over time but do not explain how behavior occurs at a specific point in time. On the other hand, popular mechanistic models tailored to describe local behavior, such as diffusion decision models (DDM[2,36,37]), either ignore the dynamic aspects of their parameters entirely or represent parameters as deterministic functions of time[38–42].

In this work, we argue that the superstatistics framework can reveal a more nuanced picture of cognitive dynamics and behavioral fluctuations. This is possible because we formalize the dynamic aspect of the low-level parameters as a higher-order stochastic process. Consequently, we estimate the low-level parameters at each time step directly from the data. Thus, their temporal evolution is only constrained by the modeler's choice of prior distributions and by the high-level transition model. Nevertheless, superstatistical models can be rigorously validated in the same way as their static counterparts, using standard model criticism methods, such as simulation-based calibration (SBC) to assess computational faithfulness, parameter recovery for inferential calibration, posterior re-simulation checks for assessing model adequacy, as well as cross-validation for assessing predictive performance[43,44]. Superstatistical models allow us to address questions about how cognitive systems undergo distinct transitions in various settings[27]. Further, one can examine which model parameters explain behavioral fluctuations without predefined equations that fix the hypothesized temporal evolution of specific parameters.

Superstatistical models can be quite challenging to estimate and compare for a number of reasons, especially in a Bayesian framework for principled uncertainty quantification[24]. First, both the high-level and low-level models are stochastic, so there is considerable uncertainty about the values of all model parameters (i.e., static and dynamic) given a finite number of observations. Second, the low-level models might be complex and non-linear so that there is not always a closed-form analytic expression relating model parameters to data (i.e., the

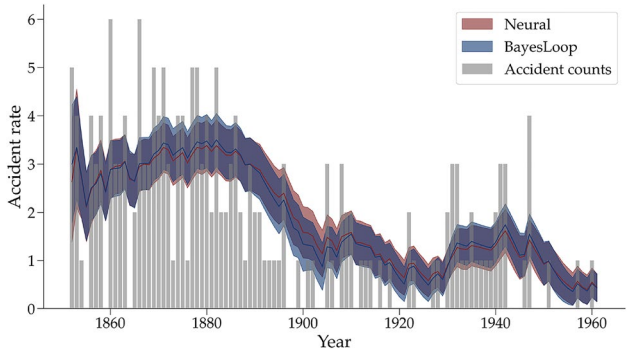

Figure 2. Coal mining disasters in the United Kingdoms between 1852 and 1962. The annual reported accident counts are depicted using gray bars. The mean posterior of the rate parameter λ of a Poisson process with Gaussian fluctuation is shown with solid lines for both estimation methods separately. The shaded area represents ± 1 posterior standard deviation.

likelihood function is *intractable*), or the likelihood might be computationally very expensive to evaluate. Finally, even for stationary low-level models, the computational cost might become insurmountable when these models are applied to multiple data sets, since standard Bayesian methods are not amortized and thus need to be re-run sequentially (unless massively parallelized) and from scratch for each data set[45,46].

Indeed, estimation challenges may be the main reason for the underrepresentation of superstatistical models in psychology and the cognitive sciences. However, we argue that recent advances in (amortized) simulation-based inference (SBI[45,47,48]) render estimation challenges secondary and allow researchers to create and test high-fidelity models of cognition without worrying about analytic tractability. SBI encompasses methods that use synthetic data to approximate intractable posterior distributions of unknown parameters. Moreover, amortized SBI with neural networks represents a particularly efficient way to perform posterior estimation on multiple data sets by investing the primary computational effort in a relatively expensive training phase[47,48]. Once simulation-based training has converged, the trained networks can be applied to any number of observations or set of observations consistent with the model's structure.

The main purpose of this article is two-fold. First, we demonstrate and validate the use of superstatistics in cognitive modeling via an out-of-the-box extension of a popular mechanistic cognitive model, namely, the DDM. Second, we develop and validate a novel Bayesian estimation method grounded in the `BayesFlow` framework for amortized neural SBI[45]. To this end, we first perform benchmark comparisons with existing frameworks on simulated data. We then specify a non-stationary DDM and fit it to long time series of response times obtained from human participants. Moreover, with this application, we empirically demonstrate how stationary models assuming IID observations can hide a number of interesting dynamic patterns and fluctuations present in behavioral data.

Results

Benchmark studies. To ensure the trustworthiness of our method, we first benchmark its performance against two existing Bayesian frameworks which use different estimation algorithms: `bayesloop`[24] and `Stan`[49]. The former employs grid approximation for low-dimensional problems, whereas the latter relies on Hamiltonian Monte Carlo (HMC[50]) sampling. Both frameworks operate in a non-amortized way and can only estimate superstatistical models with closed-form likelihoods.

Coal mining accidents. Currently, `bayesloop` cannot fit low-level models as complex as the DDM, nor high-level models such as the Gaussian process. Therefore, we compare the estimation performance of our method on a simpler example based on the coal mining accident data (freely available from[24]). These data comprise counts of coal mining accidents in the United Kingdom between 1852 and 1962. The low-level model is a simple Poisson distribution with a parameter λ that corresponds to the accident rate. One can assume that the accident rate in coal mines was not constant during this more than a century-long period. Therefore, the accident rate λ is allowed to fluctuate over time according to the Gaussian random walk transition model (cf. Eq. 3). Both estimation methods use the same informative prior distribution for the low-level parameter $\lambda_0 \sim \text{Exp}(0.5)$ and high-level parameter $\sigma \sim \text{Beta}(1, 25)$.

Using the `bayesloop` software, we approximated a grid with 4000 equally spaced points ranging from 0 to 15 for λ and from 0 to 1 for σ, respectively. This calculation lasted approximately 38 minutes on a standard desktop computer. Training the neural network for 20 epochs took approx. 18 minutes, and obtaining 4000 posterior samples took less than a second. Thus, in this case, the training effort amortizes even after *a single* data set.

Figure 2 shows the annual count of coal mining accidents overlaid with the estimated dynamic accident rate λ (posterior mean and ± 1 standard deviation). Both methods estimate an almost identical latent trajectory for the low-level model parameter λ. Between the years 1880 and 1900, we observe a decrease in coal mining accidents followed by two temporary increases around the years 1905 and 1930. The estimated parameter dynamic

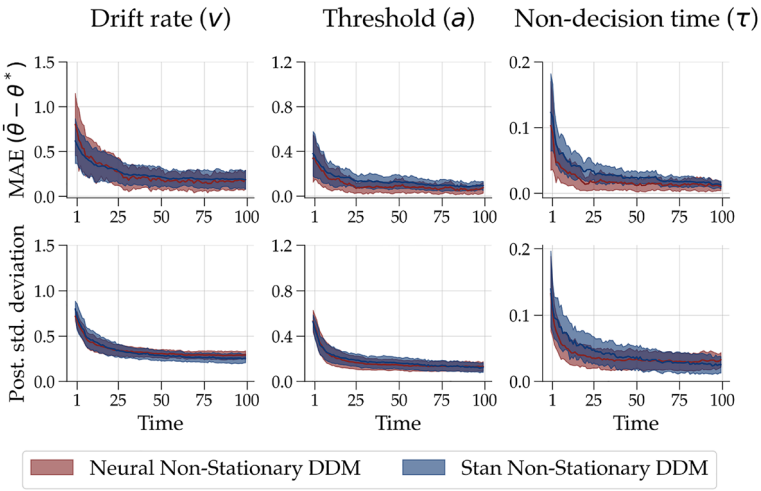

Figure 3. Comparison between the neural and Stan estimation method. First row: Median absolute error (MAE) between the ground truth data-generating parameters and the estimated posterior means across the 100 simulations over time. Second row: Posterior standard deviation aggregated across the 100 simulated data sets over time (solid lines). The shaded area depicts the median absolute deviation (MAD).

closely follows these data patterns. Thus, we conclude that our neural method can estimate a plausible parameter dynamic for a simple low-level model and performs equally well compared to `bayesloop`.

Static diffusion decision model. As a second benchmark, we compare our neural method to `Stan` in terms of Bayesian updating, assuming a "true" DDM with time-invariant parameters. This benchmark serves two goals. Firstly, it aims to compare the estimation performance of our method with that of `Stan`, which is regarded as the gold standard for sampling-based Bayesian inference. Secondly, it aims to assure that our method can correctly identify stationary parameters when fitting a dynamic model on data generated from a stationary process (i.e., it does not estimate "pseudo-dynamics").

To this end, we simulated 100 data sets with 100 observations, each using a static DDM with 3 free parameters (see "A non-stationary diffusion decision model" section) without parameter fluctuation over time. Then, we fit a non-stationary DDM with a Gaussian random walk transition model (cf. Eq. 3) to all 100 data sets using both estimation methods. Again, we use the same prior distributions (see Appendix) to ensure comparability. We compared the two methods based on the following two performance metrics: (i) the median absolute error (MAE) between the estimated posterior means and the data generating stationary parameters averaged across all 100 simulations, and (ii) the average posterior standard deviation over time. These two metrics are common indicators for inferential model calibration, which aims to analyze the global behavior of the posterior distribution given possible observations from the prior predictive distribution[51]. The former metric informs us how well the posterior recovers the true model configurations (analogous to posterior z-scores). The latter metric indicates how much the posterior is informed by the data beyond the prior knowledge that was encoded in the prior distribution (analogous to posterior contraction)[51].

The upper panel of Fig. 3 depicts the absolute difference between the true data generating parameters and the dynamically estimated posterior means over time, averaged over all 100 simulations for both methods separately. On average, the posterior means show a relatively large deviation from the true data generating parameters on early trials of the data. This difference then quickly decreases and flattens after approximately 25 trials. The performance of both methods concerning this metric is almost indistinguishable.

The lower panel of Fig. 3 displays the median posterior contraction measured as posterior standard deviation over time for all 3 parameters separately. We observe considerable posterior contraction within the first 25 time points. Again, the performance of both methods is nearly identical. However, there is a large difference in estimation time between the two methods. As we are interested in the filtering posterior distributions, the `Stan` model has to be refit with every additional observation of a time series. Hence, we fit the `Stan` model to each simulated $x_{1:t}, t = 1, \ldots, T$, which amounted to $T = 100$ re-fits per simulated data set. Fitting the model to all 100 synthetic data sets resulted in 100×100 model fits. This procedure took over 1 week of non-stop computing on a standard desktop computer—whereas training the neural network lasted approximately 8 h with almost instantaneous fit to the 100 data sets thereafter. This is a non-negligible difference that will grow with longer time series, more data sets, or increased complexity until reaching a point where models can no longer be estimated with `Stan` due to limited processing resources or time constraints (see next section).

In summary, our method closely approximates the results obtained from `bayesloop` and `Stan` on the considered benchmark examples. Note, however, that our method is primarily designed for models where the above frameworks cannot be applied—higher dimensional models, possibly lacking a closed-form likelihood (i.e., available only as stochastic simulators), or many data sets consisting of long time series. The next application we present could be tackled with our neural approximators, but not with the above two frameworks.

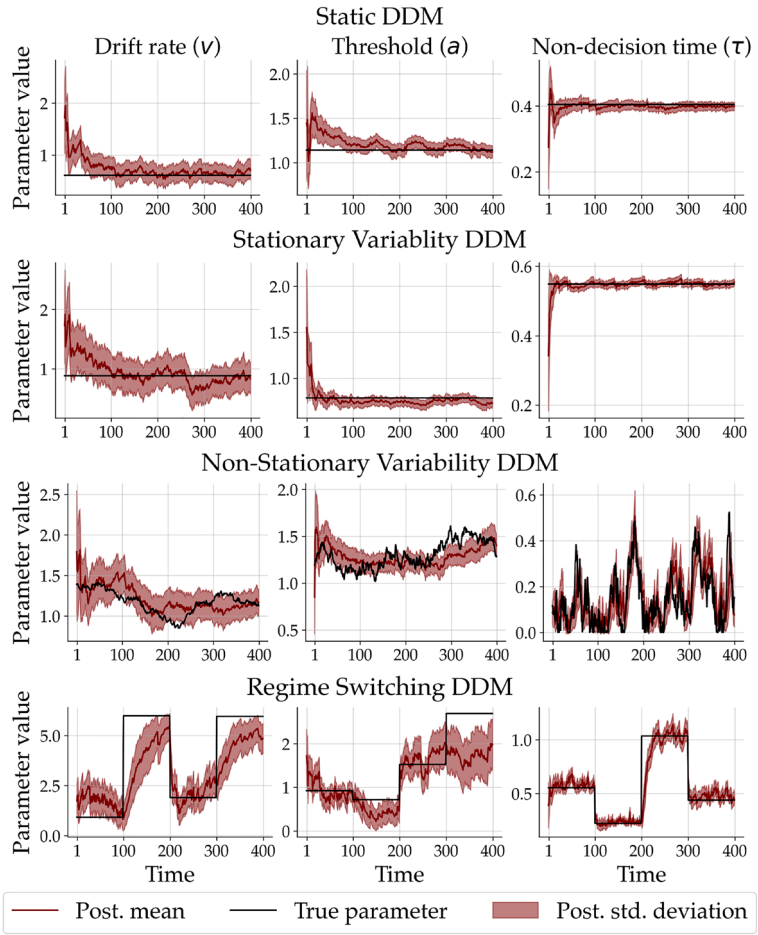

Figure 4. Example time-varying parameters estimated by our neural method in each scenario of the simulation study. Each row depicts the posterior estimates obtained from a single simulated person. The third row corresponds to the dynamic model used for training the network (i.e., well-specified case). The first, second, and fourth rows correspond to model variants not seen during training (i.e., misspecified cases).

Simulation study. Next, we probe the parameter recoverability of a non-stationary DDM under different induced misspecifications (i.e., models that differ from the one used for training the network). To this end, we performed an extensive study for which we simulated data sets consisting of $T = 400$ time points in four different scenarios: (i) A static DDM with constant parameters; (ii) a DDM with stationary variability (commonly referred to as "inter-trial variability") where the 3 DDM parameter fluctuate randomly around a constant value; (iii) a non-stationary DDM with a Gaussian random walk transition model; (iv) and a DDM with constant parameters that jump abruptly and uniformly at three predefined time points (i.e., a regime switching model). Crucially, we trained the neural approximator only with simulations from the non-stationary model. However, during amortized inference, we applied the network to 200 data sets from each of the four scenarios. Thus, we could investigate the network's response in the *open world* setting where the true data generator may differ from the reference model used during the training phase.

Figure 4 shows an exemplar fit of the non-stationary DDM with a random walk transition model to data sets from each of the four simulation scenarios. In the top row, we see that the estimated parameter trajectories converge to the constant ground-truth parameters. A similar pattern emerges when the ground-truth parameters randomly fluctuate around a constant value (second row), yet we observe less uncertainty reduction. The third row depicts the posterior estimates based on a data set simulated from the reference non-stationary DDM (i.e., the well-specified case). Besides some local deviations from the ground-truth parameter trajectory, the model is able recover the overall trend of the dynamics. In the fourth row, we can inspect the posterior estimates from a data set simulated from the regime switching DDM which allows the parameters to "jump" uniformly at three time points to any value within the parameter bounds. Despite the severe misspecification, the random walk DDM is able to recover the discontinuous trajectories surprisingly well; still, the gradual change implied by the random walk transition does not allow for a rapid adaptation and exhibits a notable lag after each switch.

Figure 5 depicts the true data generating and the estimated posterior means at time point $T = 99$ (right before the first jump of the regime switching transition model). We observe excellent recovery performance for all 3 parameters in all 4 simulation scenarios at the selected time point. The recovery performance at other time points as well as further details and analyses (i.e., MAE over time) can be found in the Appendix.

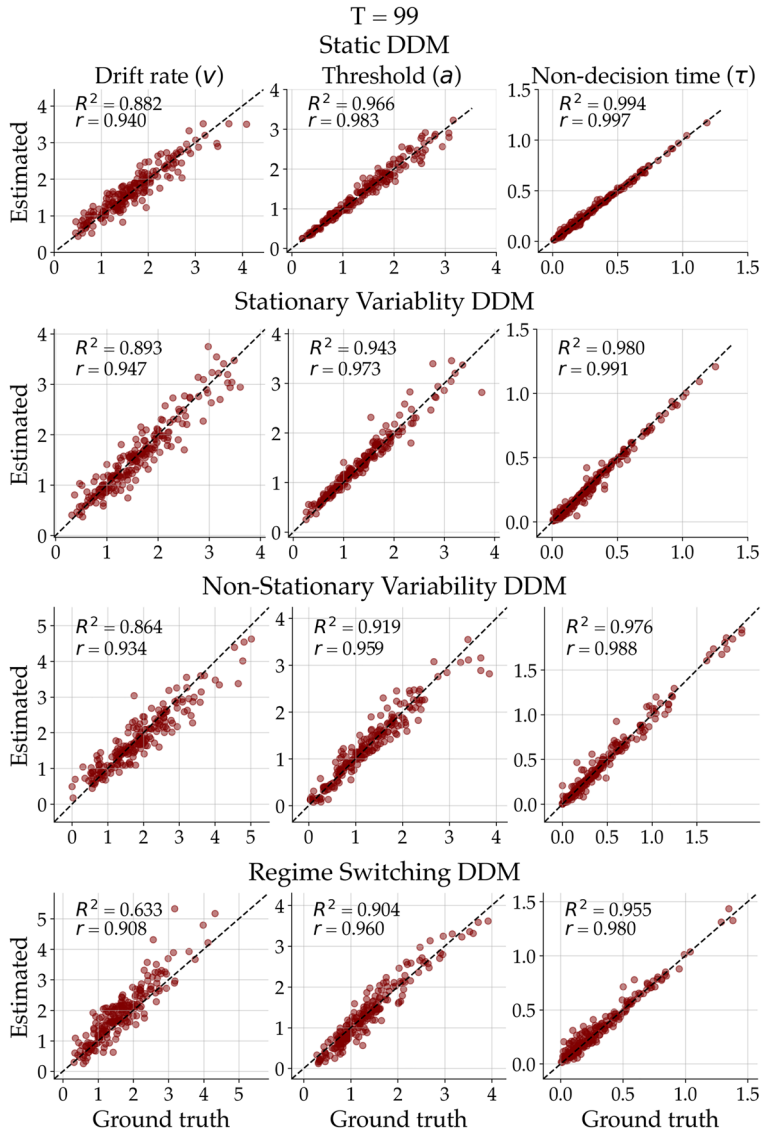

Figure 5. Ground truth-data generating parameters plotted against posterior means for all 3 parameters and simulation scenarios separately at time point $T = 99$ (just before the change of regime of the regime switching DDM).

Human data applications. Following our benchmarking and simulation studies, we applied non-stationary versions of the DDM to two separate data sets collected from response time (RT) experiments: (i) A standard random-dot motion task (a maximum of $T = 1320$ trials per participant), and (ii) very long time series (a maximum of $T = 3200$ trials per participant) from a lexical decision task. The first application serves as a starting point with data stemming from a popular task in experimental psychology. The second application showcases the utility of our method to estimate a complex non-stationary DDM with a Gaussian process (GP) transition model and multiple drift rate parameters for different difficulty conditions. Before fitting a model to empirical data, it is imperative to assess the faithfulness of the approximation method[43,52]. To this end, we perform simulation-based calibration (SBC[53,54]). These analyses suggest that our neural Bayesian method exhibits reasonable calibration, with slightly miscalibrated posteriors for the non-decision time parameter (see Appendix for more details on calibration).

Random-dot motion task. First, we fit a non-stationary DDM with a Gaussian random walk transition model to a data set retrieved from the experimental study of[55]. We chose this data set because the purpose of the original study was to investigate the decline of the threshold parameter over time. The experiment had a 3 (*Low*, *Medium*, and *High* feedback) by 2 (*Time* and *Trial* condition) factorial between-subject design. Differently from our approach,[55] subdivided the time series into trial bins and fitted a stationary hierarchical Bayesian DDM to each bin separately. Therefore, we can compare the parameter trajectories recovered by our neural superstatistics method with the estimates obtained by the original authors using Markov chain Monte Carlo (MCMC).

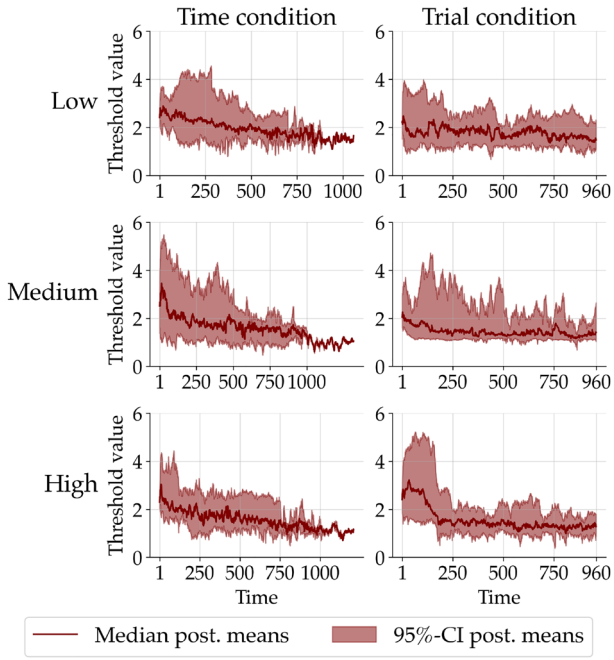

Figure 6. Estimated trajectories of the DDM threshold parameter aggregated across all individuals for each between-subject experimental condition. The first column corresponds to the *Time* and the second to the *Trial* condition. The rows correspond to the three feedback conditions, *Low*, *Medium*, and *High*, respectively. The red solid lines depict the median of the individual posterior means and the red shaded area the 95% credibility interval of these posterior means.

Figure 6 depicts the trajectory of the threshold parameter aggregated across all individuals in a separate panel for each experimental condition. Note, that in the *Time* condition participants had a fixed amount of time they could spend on the task resulting in different time intervals. When we compare our estimates to those obtained by[55], it becomes evident that both approaches yield similar qualitative and quantitative patterns. This result complements our promising results "in silico" and points to the convergent validity of our superstatistics approach in applications with real data.

Lexical decision task. We fit the non-stationary DDM with a GP transition model (cf. Eq. 5) to human behavioral data originating from a lexical decision-making task. The data consist of long RT and choice time series from four experimental conditions. For this application, we use four separate drift rates—one for each experimental condition. The length of these time series made it impossible to estimate the model with `Stan` (due to memory limitations and infeasible compute time). Thus, to increase the trustworthiness of the results obtained with our neural method, we resort to the established `fast-dm` software[36] as a benchmark, which is capable of estimating homogeneous (block) trial-by-trial fluctuations (i.e., inter-trial variabilities). We then compare the goodness of absolute fit in terms of re-simulation accuracy between both estimation methods and investigate the multi-horizon predictive performance of our method. Further, we analyze the main advantage of the non-stationary DDM, that is, the inferred trial-by-trial parameter dynamics, and compare those to the static `fast-dm` parameter estimates. Note that `fast-dm` is not a Bayesian method and is thus not included in our previous benchmark studies.

The left panel of Fig. 7 depicts the empirical RT time series data of an individual participant in black (Figures for the remaining participants are available in the Appendix). To evaluate whether the non-stationary DDM is capable of capturing the empirical data, we perform posterior re-simulations on the first 3 blocks of the experiment (trials 1–2500). To this end, we draw 100 samples from the posterior distributions over $\theta_{0:2499}$ to simulate 100 posterior re-simulated data sets. The resulting RT time series are then summarized with the median and the 95% credibility interval (CI) across simulations and depicted in red color. We smooth the trial-by-trial empirical data and model outputs via a simple moving average (SMA) with a period of 5 to ease visual inspection of potential trends. Note, that the re-simulation from the `fast-dm` model is only shown in the marginal RT distribution on the right panel to avoid visual clutter.

The overall time series show that the individual's RTs decrease over time. Furthermore, the variability of the RTs, which is most pronounced in the first session, decreases considerably over time. The non-stationary DDM not only captures both of these overall trends, but also represents the shorter time oscillations within the empirical RT time series. The data also exhibits various sudden "jumps" in RTs, probably due to fluctuations in non-decisional processes, such as inattention. Unsurprisingly, these jumps are not fully accounted for by our non-stationary DDM since the high-level model (GP with squared exponential kernel) does not allow for sudden large changes in the low-level parameters.

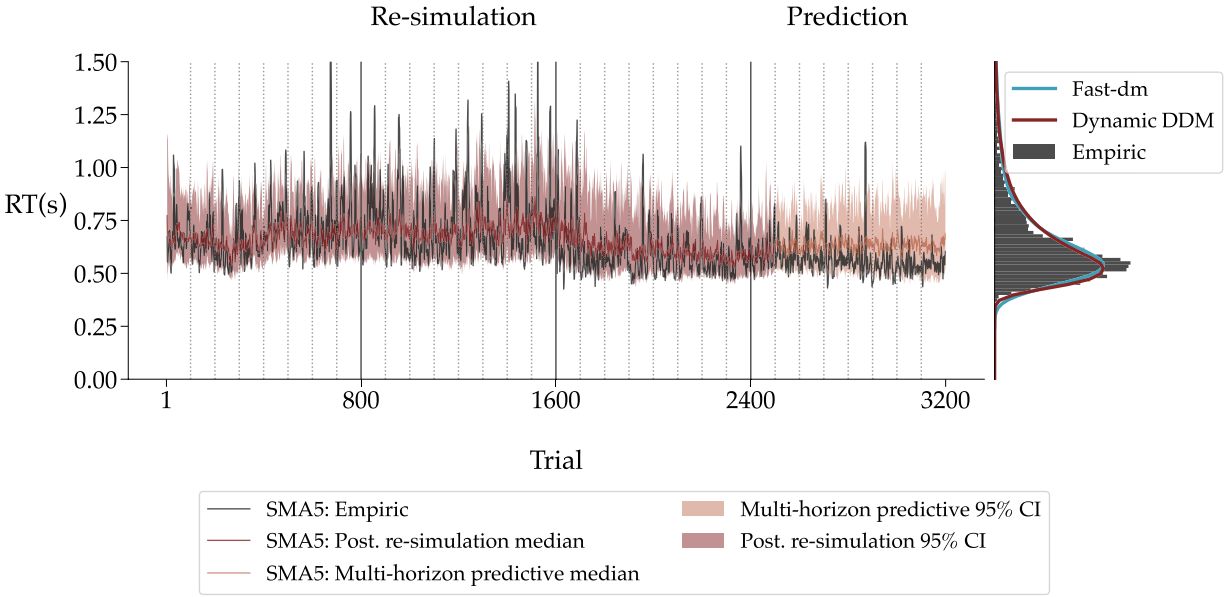

Figure 7. Model fit to human data. **Left panel** The empirical RT time series of a single individual in black. From trial 1 to 2500, the median posterior re-simulation (aka *retrodictive check*) using the non-stationary DDM is shown in red. The models' multi-horizon prediction is depicted for the remaining trials in orange. The shaded areas for the posterior re-simulation and multi-horizon prediction correspond to 95% credibility intervals. All the time series were smoothed via a simple moving average (SMA) with a period of 5. The dotted vertical lines indicate the end of an experimental block, and the solid vertical lines the end of an experimental session. **Right panel** The raw RT distribution is plotted as a histogram in black. The re-simulated RT distributions from the non-stationary DDM and reference re-simulations from the static DDM using `fast-dm` are shown as kernel density estimates (KDEs) in red and blue, respectively.

We purposefully leave out the remaining 700 trials from the posterior re-simulation analysis to also test the predictive capabilities of the non-stationary DDM against held-out empirical data[56,57]. To this end, we generate 100 new parameter dynamics according to Eq. (5) with randomly drawn posterior samples of θ_{2499} as initial parameter values and posterior samples of the high-level Gaussian process parameters η. Then, we simulate 100 novel RT time series for the remaining 700 trials using the simulated parameter trajectories. The resulting RT time series are summarized in the same manner as before (median, 95% CI) and again smoothed with an SMA. The corresponding multi-horizon posterior predictions are depicted in Fig. 7 with an orange color. The dynamic model yields accurate predictions on the held-out data and thus does not overfit the training data. Moreover, the held-out time series remain in the 95% CI of the multi-horizon prediction, which is the case for all individual data sets (see Appendix).

The right panel of Fig. 7 depicts the empirical RT distributions (black) along with the data generated by the non-stationary DDM (red) and the static DDM (blue). Note that the three empirical RT distributions show a substantial overlap. Since the `fast-dm` re-simulations serve as a benchmark for the non-stationary DDM, it is essential to quantify if there are pronounced deviations between the re-simulated and the empirical RT distributions. To this end, we estimate the pairwise maximum mean discrepancy (MMD) between the three distributions for each individual separately and then average the resulting values across participants. MMD is a kernel-based statistical metric of equality between distributions[58].

Accordingly, our analysis reveals no pronounced differences between the three distributions. The average MMD between the empirical RT distributions and the ones predicted by the non-stationary DDM ($\overline{MMD} = 0.026, SD = 0.008$) is lower than between the empirical and the ones predicted by the `fast-dm` model ($\overline{MMD} = 0.042, SD = 0.027$). The SDs of the average MMD values indicate that data generated with the non-stationary DDM are not only slightly more accurate on average but also more consistent compared to data generated from the standard DDM. For the sake of completeness, we also compare both re-simulated RT distributions ($\overline{MMD} = 0.035, SD = 0.019$). This comparison reveals that the re-simulated RT distributions of the static DDM are more similar to the one obtained by the non-stationary DDM than to the empirical RT distribution. Altogether, both models can reproduce the empirical RT distributions with high fidelity, but the non-stationary DDM fits the data slightly better than the static DDM estimated with `fast-dm`.

In summary, our non-stationary DDM can closely re-simulate and predict the temporal trajectory of empirical RT time series as well as corresponding raw RT distributions from all individuals (see Appendix). Even though the standard DDM also accounts for the marginal RT distribution, it cannot generate the observed heterogeneous RT time series data (cf. Fig. 7).

However, the most decisive advantage of our non-stationary DDM over its stationary counterpart is that it can recover parameter dynamics directly from the empirical data. As the static parameters of `fast-dm` can only vary homogeneously around their mean, we cannot detect any systematic changes in the parameters over time. However, the dynamic parameters estimated with our neural method strongly suggest such systematic

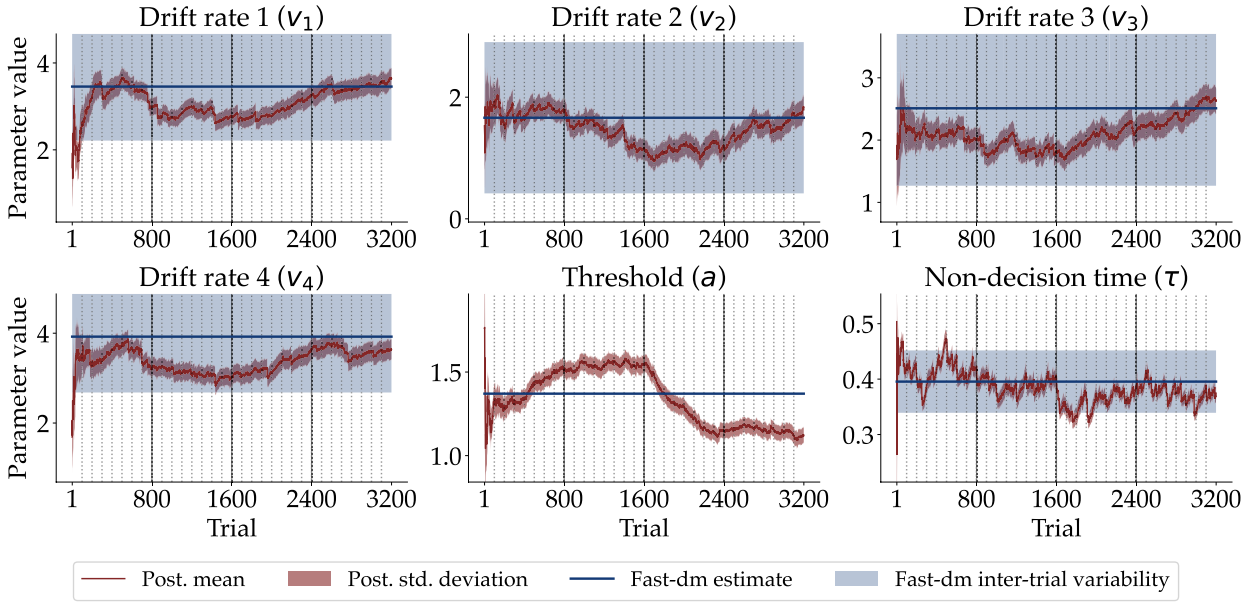

Figure 8. Estimated parameter dynamics. The trial-wise posterior mean and ±1 standard deviation for all six parameters, namely the four drift rates v_1–v_4 (one for each experimental condition), the threshold a, and the non-decision time τ of an individual participant. The point estimates of the stationary DDM parameters and the corresponding inter-trial variabilities (except for the threshold a) are shown in solid blue lines and blue shaded areas, respectively.

changes. Figure 8 depicts the dynamics of the estimated trial-by-trial posterior means and ±1 standard deviation for all DDM parameters of the same participant as above in red (see Appendix for the parameter dynamics of the remaining participants as well as the average parameter dynamic). The corresponding point estimates (solid line) and inter-trial variabilities (shaded area) obtained with `fast-dm` are shown in blue.

All parameters of the non-stationary DDM seem to exhibit considerable fluctuations and notable oscillations throughout the experiment. Due to the assumption of homogeneous variation, the inter-trial variabilities inferred with `fast-dm` vastly overestimate the uncertainty in parameter estimates (cf. Fig. 8). The dynamic drift rates fluctuate roughly within the uncertainty corridors spanned by the homogeneous inter-trial variabilities, but exhibit much tighter error bars. As a consequence, local drift rates are much less uncertain than the homogeneous variability parameters indicate. On the other hand, the dynamic non-decision time τ fluctuates more than the corresponding flat inter-trial variability. Note that `fast-dm` does not support estimating inter-trial variability of the threshold a, so we only report the estimates of our neural method, suggesting a substantial decrease of the threshold parameter throughout the experiment. Notably, we observe a considerable mismatch between heterogeneous and homogeneous dynamics in almost all individuals (see Appendix).

Discussion

In this work, we explored the merits of superstatistics for representing non-stationary dynamics in cognitive processes, along with the utility of a neural Bayesian method for estimating superstatistical models. We verified the computational faithfulness and adequacy of our method using simulations and two benchmark studies. We then applied our method to a dynamic, non-stationary diffusion decision model and estimated the temporal trajectories of its key parameters, namely, drift rates, decision threshold, and non-decision time from the data of two experiments. We showed that such a non-stationary model (i) can indeed be fit to long time series of human data with high fidelity and (ii) that the inferred heterogeneous dynamics reveal patterns that would have remained hidden by traditional stationary models[2,6]. To our knowledge, this is the first attempt to augment a stationary cognitive model by employing a superstatistics framework.

Previous research has suggested that response times often exhibit heterogeneous dynamics[9,10]. It has also been shown that even the history of past choices can influence specific parameters of the DDM[40]. Hence, it seems plausible that the cognitive processes represented by the DDM parameters vary over time even within an experimental session due to internal psychological factors. This is exactly what was implied by the individual parameter dynamics inferred from the lexical decision task data set. However, as the data originates from an experiment that was not designed explicitly to test dynamic modeling, we need to be wary of any *ad hoc* interpretations concerning the estimated parameter dynamics.

Nevertheless, some of the recovered patterns may suggest interpretable underlying changes. For instance, the threshold parameter seemed to decrease within an experimental session for many individuals. This indicates that participants generally responded less cautiously toward the end of an experimental session. A plausible explanation for this change in response caution might be that participants became increasingly bored during a session and started to decrease their ambitions. Note that current DDM modeling approaches rarely account for

such variation in the threshold parameter. Further, the drift rates generally tended to increase over time, suggesting that participants' increased their information processing speed over time. A change in the average rate of information uptake typically results in shorter RTs, which is precisely what we observed in most individual data sets (cf. Appendix). These increases in drift rates over time could imply the occurrence of learning effects. An important next step will be to tailor experiments with systematic manipulations from which we expect specific changes in some cognitive process and test whether the estimated parameter dynamics exhibit these changes.

Notwithstanding, our neural method has certain limitations. As can be seen in Fig. 3, the values for most parameters change strongly at the beginning of the time series. One could be tempted to (falsely) claim that the psychological constructs mapped to these parameters drastically change at the beginning of the first session of the experiment. However, these early parameter trajectories should be interpreted with great caution as they can be quite dependent on the initial prior. As a result, we cannot easily differentiate between initially large Bayesian updates to move away from the prior or actual changes in the underlying process. As is the case for any dynamic process, our modeling approach may also not be sensible for data sets with few observations. In the context of psychological experiments, a possible remedy could be to use burn-in trials at the beginning of an experiment that only serves the purpose of having some data points to inform the plausible parameter values. At the same time, these could serve as practice trials during which participants get accustomed to the task.

Furthermore, the simulation study has demonstrated that the non-stationary DDM exhibits a good performance in recovering parameters across various scenarios. However, it is essential to acknowledge that there still exists an error between the true and estimated parameters. Especially for the drift rate parameter errors around 0.25 have been observed frequently. Consequently, interpreting small local changes in parameter values requires caution. Despite this limitation, we firmly believe that the proposed method excels particularly in scenarios where moderately large changes in parameters are expected to occur over the course of a couple of time steps.

Another limitation concerns the implementation of the low-level mechanistic model, that is, the DDM itself. We assumed four different drift rates—one for each stimulus type—which is the standard procedure used in the application of stationary DDMs[2]. This parameter is usually regarded as a proxy for average information uptake speed. However, in theory, there should only be one drift rate per participant[3] that changes over time, for instance, due to experimental manipulation. Thus, a non-stationary DDM could also incorporate only one drift rate parameter. In our experiment, the manipulation (i.e., four conditions) was randomized throughout the experiment. This implies that besides fluctuation stemming from other sources, the drift rate would "jump" from trial to trial based on this change in task difficulty. To account for these jumps, we would need a different high-level transition model whose changes can be bigger than what a smooth Gaussian process or Gaussian random walk allows. In order to keep the content of this article manageable, we decided against proposing a novel transition model.

Finally, there are numerous degrees of freedom when implementing a computational model – not only with respect to the low-level observation model, but also regarding the high-level transition model. Exploring different model specifications and then deciding which is the most sensible for the type of task and data at hand requires Bayesian model comparison. Concerning dynamic cognitive models, it would be of particular interest to test which high-level transition model specification is most plausible for a given setting[24]. Since Bayesian model comparison is a topic in its own right, future studies should investigate the utility of simulation-based methods[59,60] for comparing competing superstatistical models.

We acknowledge that our study may not provide a definitive argument for when and why a non-stationary DDM is superior to a static DDM. The primary objective of this article is to showcase the implementation of non-stationary parameters within a superstatistics framework. However, we believe that the superstatistics framework, coupled with powerful neural approximators, gives rise to many new modeling opportunities and makes it possible to augment virtually any computational model with time-varying parameters. We think that there are many interesting research questions out there that could be investigated with the approach we propose in this work. Future studies can use our method to estimate even more challenging cognitive models than the DDM explored in this work and further extend its scope beyond cognitive science and psychology.

Methods

Experimental tasks. *Random-dot motion task.* The data set used in this study was adopted from[55]. It includes data from 58 individuals, after excluding participants with a response accuracy below 70%. Each individual was randomly assigned to one of six groups, which were formed by two factors: the *time* versus *trial* condition and three levels of feedback details. During the experiment, participants solved a total of 24 blocks of the task. In the *trial* condition, each block comprised 40 trials, whereas in the *time* condition, each block lasted for 1 minute. In each trial, participants were presented with a random dot kinematogram and were required to determine if some of the dots coherently moved to the top-left or top-right direction. For more in-depth information about the experimental setup and methodology, refer to the comprehensive details provided in[55].

Lexical decision task. A total of 11 students from Heidelberg University participated in the experiment. Their average age was 23.81 ($SD = 3.30$) and 10 of the participants were female. All individuals gave written informed consent to the study, which was approved by the local ethics committee. The study was conducted according to the ethical declarations of Helsinki.

The participants performed a lexical decision-making task. On each trial, they had to assess if a presented letter string was a German word. As stimuli, we used high and low-frequency words, pseudo words that were generated by replacing vowels of existing words, and random letter strings. These four experimental conditions were pseudo-randomly presented throughout 3200 trials. All participants solved their task on 4 separate days (sessions) consisting of 800 trials each. The sessions were further split into 8 blocks of 100 trials with short breaks

between them. On each trial, participants' choice (German word; non-German word) and response time was recorded.

Model family. Following[24], we consider dynamic models that entail a low-level model with time-dependent parameters θ_t, which vary according to a high-level model with static parameters η. The low-level model is defined by a likelihood function $\mathcal{L}$, and the high-level model consists of a transition function $\mathcal{T}$.

In this work, we aim to tackle general superstatistical models for which the low-level model likelihood $\mathcal{L}$ may not be available in closed-form. Such models are implemented as randomized stateful simulators that generate observable trajectories $\{x_t\}_{t=1}^{T}$ via the following (very general) recurrent system:

$$\theta_t = \mathcal{T}(\theta_{0:t-1}, \eta, \xi_t) \quad \text{with} \quad \xi_t \sim p(\xi|\eta) \tag{1}$$

$$x_t = \mathcal{G}(x_{1:t-1}, \theta_t, z_t) \quad \text{with} \quad z_t \sim p(z|\theta_t). \tag{2}$$

In the above equation, $\mathcal{T}$ is an arbitrary high-level transition function parameterized by η, $\mathcal{G}$ stands for an arbitrary (non-linear) transformation which encodes the functional assumptions of the low-level model. $\xi_t \sim p(\xi)$ and $z_t \sim p(z)$ are sources of random noise. The initial parameter configuration θ_0 follows a prior distribution $\theta_0 \sim p(\theta)$ which encodes available information about plausible parameter values.

One example of a transition model $\mathcal{T}$ is a convolution with a Gaussian distribution, which implies a gradual change in the low-level model's parameters resembling a random walk:

$$\mathcal{T}(\theta_{t-1}, \eta, \xi_t) = \theta_{t-1} + \eta\, \xi_t \quad \text{with} \quad \xi_t \sim \mathcal{N}(0, 1). \tag{3}$$

Another similar example would be a convolution with a fat-tailed distribution, allowing for abrupt changes in the parameter space. Furthermore, since our simulation-based setting is not limited to transition models with a Markov property, we can also test more complex transitions, such as a vector autoregression (VAR[61]):

$$\mathcal{T}(\theta_{t-p:t-1}, \eta, \xi_t) = c + A_1\theta_{t-1} + \cdots + A_p\theta_{t-p} + \xi_t, \tag{4}$$

where p is the order of the VAR model (i.e., its look-back period), $\xi_t \sim \mathcal{N}(0, \sigma)$, and $\eta = \{c, A_1, \ldots, A_p, \sigma\}$ are the high-level parameters of the model.

We can even test transition models which depend on the entire history of the process, such as a Gaussian process (GP[62])

$$\theta_{1:T} \sim \mathcal{GP}(\mu_\theta, K_\theta) \tag{5}$$

with mean function μ_θ and covariance function K_θ defined through the vector of time indices. The high-level parameters η in this case would be the free kernel parameters, such as the amplitude σ or the length-scale l of a Gaussian kernel

$$k(\theta_t, \theta_{t'}) = \sigma^2 \exp\left(\frac{||\theta_t - \theta_{t'}||^2}{2l^2}\right). \tag{6}$$

A typical task in Bayesian analysis of dynamic systems is to recover both the entire trajectory of dynamic parameters $\{\theta_t\}_{t=1}^{T}$ as well as the vector of static parameters η. Since for many discrete dynamic systems, the current data point x_t depends on the current parameter configuration θ_t as well as on the observable history of the system $x_{1:t-1}$, we can write the (implicit) point-wise likelihood as

$$\mathcal{L}_t = p(x_t|x_{1:t-1}, \theta_t). \tag{7}$$

The point-wise likelihood describes the probability of each data point, given the parameter values of the same time step and all past data points[24]. Notably, we do not require this likelihood to be available in closed-form; we only need the ability to generate random draws through the forward-time generative process specified by Eq. (1).

Assuming the above factorization of the likelihood is possible, we aim to estimate the joint *filtering* posterior distribution of θ_t and η up to each discrete time-step t

$$p(\theta_t, \eta|x_{1:t}) \propto \mathcal{L}_t\, p(\theta_t|x_{1:t-1}, \eta)\, p(\eta|x_{1:t-1}). \tag{8}$$

This posterior encodes the reduction in uncertainty regarding the dynamic states evolving over time and the static parameter values being increasingly constrained by the data. From this joint distribution, we can derive the corresponding marginal posteriors as follows:

$$p(\theta_t|x_{1:t}) = \int p(\theta_t, \eta|x_{1:t})\, d\eta, \tag{9}$$

$$p(\eta|x_{1:t}) = \int p(\theta_t, \eta|x_{1:t})\, d\theta_t. \tag{10}$$

These distributions describe the average parameter dynamics over all possible high-level parameters and the best estimate for the high-level parameters up to discrete time-step t, respectively. Thus, learning both distributions amounts to standard Bayesian updating with an additional uncertainty factor due to the high-level transition model $\mathcal{T}$. Thus, posterior contraction over time will strongly depend on the form of the transition model and

may even increase in some cases, such as models allowing for sudden "jumps" in their parameters (i.e., regime switching behavior).

Neural Bayesian estimation. Various methods for estimating dynamic models have been proposed in the literature. Markov chain Monte Carlo (MCMC) methods offer a viable but computationally demanding approach based on random draws from the posterior[63]. Variational inference (VI) methods approximate the true target posterior with simple, tractable densities and thus are a faster alternative to MCMC at the cost of a potential loss of posterior accuracy[63]. A recent promising approach for low-dimensional problems is the grid-based method of[24], which represents parameter distribution on discrete lattices and enables efficient approximation of model evidence.

However, the above methods all depend on the ability to evaluate the likelihood function $\mathcal{L}_t$ at each time point explicitly. This restriction makes it impossible to efficiently test the growing number of simulator-based or non-analytic models of cognition to observed data[45,64]. Furthermore, MCMC and standard variational methods cannot leverage experience and require the same repeated computational effort for every new data set. For instance, when multiple participants complete a cognitive task, the same estimation procedures need to be repeated for each participant from scratch. Differently, hierarchical Bayesian models can be employed to jointly estimate group- and participant-level parameters, but they come with high computational costs and also rely on a closed-form likelihood function.

In contrast, *amortized inference* refers to methods with a "pre-paid" computational cost - after an expensive optimization or training phase, the same procedure can be instantly applied to any data set whose structure is compatible with the model[45,46]. As a useful "side effect", amortization allows us to easily perform extensive checks of computational faithfulness and parameter recoverability "in silico", since we can obtain posterior samples from hundreds or even thousands of simulated data sets by applying the same pre-trained network. Amortized Bayesian inference is typically realized by specialized neural networks, which are trained to become estimation experts from repeated model simulations[45,65]. The architecture of these networks can easily encode the probabilistic symmetry of the data, for instance, recurrent networks for temporal data[66] or permutation-invariant networks for IID data[67].

Crucially, dynamic models with time-varying parameters present a challenge to existing neural architectures since they induce a new joint posterior at each time-step $p(\theta_t, \eta|x_{1:t})$. However, most previous architectures can only estimate a single set of parameters with no temporal information[45,47,65]. Thus, we propose to use a recurrent probabilistic neural architecture that estimates the joint posterior over all static and dynamic parameters for all discrete time points in a single forward pass.

Recurrent estimation method. Our proposed architecture consists of several neural components. First, a recurrent neural network (RNN) with learnable parameters $\psi^{(r)}$ embodying long short-term memory (LSTM) consumes the observed data sequentially:

$$h_t = \text{LSTM}\left(x_t, h_{t-1}; \psi^{(r)}\right), \tag{11}$$

where the hidden state h_t at each time point t represents the internal memory of the network over arbitrary temporal intervals. Thus, we can treat h_t as a compact representation of the observable history up to time point t. We employ a standard LSTM network, which consists of three gates: an input gate, an output gate, and a forget gate. These gates are responsible for weighing and integrating old and new information. Importantly, LSTM networks can naturally deal with sequences of varying length, which enables them to process streams of "online" data[66].

In order to recover the time-varying parameters θ_t of the low-level model as well as the static high-level parameters η, we use the hidden state h_t as a conditioning vector for a generative neural network with trainable weights $\psi^{(g)}$. This network can be implemented as a conditional variant of any popular generative architecture for inference, such as coupling networks[68], autoregressive flows[69], or standard neural networks with probabilistic outputs[70]. The generative network is responsible for approximating the current joint posterior up to time step t given the outputs of the recurrent summary network: $q(\theta_t, \eta|x_{1:t}, \psi) \equiv q(\theta_t, \eta|h_t, \psi)$. To reduce notational clutter, we set $\psi = (\psi^{(r)}, \psi^{(g)})$ and assume that h_t is expressive enough to encode all information contained in the data for correctly updating the prior (i.e., h_t is a maximally informative *summary statistic* of $x_{1:t}$).

Alternatively, we can also directly target one of the two equivalent factorizations of the joint posterior, namely:

$$p(\theta_t, \eta|x_{1:t}) = p(\theta_t|x_{1:t}, \eta)\, p(\eta|x_{1:t}) \tag{12}$$

$$= p(\eta|x_{1:t}, \theta_t)\, p(\theta_t|x_{1:t}). \tag{13}$$

While being mathematically equal, these factorizations imply different neural architectures and corresponding ancestral sampling schemes. The former factorization (Eq. 12) requires a generative network for first sampling the high-level parameters from $p(\eta|x_{1:t})$ and then sampling the low-level parameters from $p(\theta_t|x_{1:t}, \eta)$, conditional on the sampled high-level parameters. On the other hand, the latter factorization (Eq. 13) requires a generative network for first sampling the low-level parameters from $p(\theta_t|x_{1:t})$ and then sampling the high-level parameters from $p(\eta|x_{1:t}, \theta_t)$, conditional on the sampled low-level parameters.

In the current work, we consistently target the factorization in Eq. (12), but we were able to obtain comparable filtering results with either ancestral sampling strategy. In practice, we can either assume a multivariate Gaussian posterior for $q(\theta_t|x_{1:t}, \eta, \psi)$ and $q(\eta|x_{1:t}, \psi)$ as a dynamic extension of the basic method in[71] or estimate

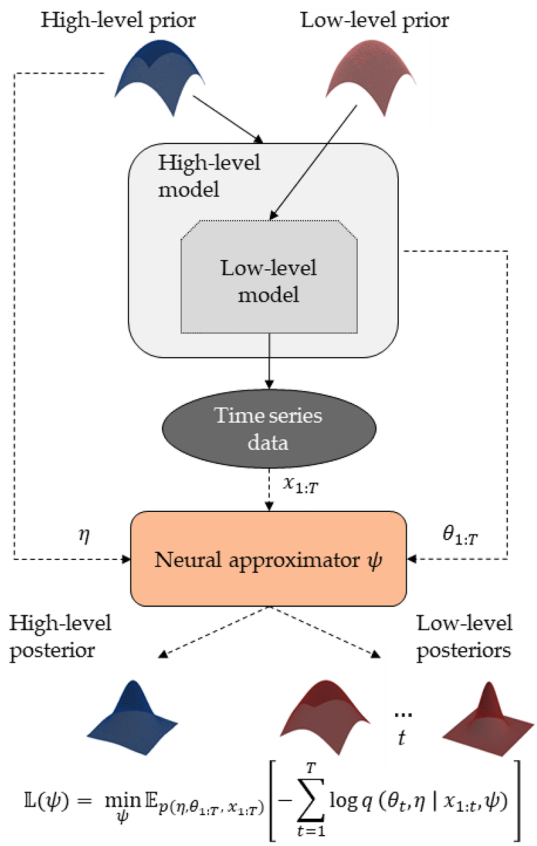

Figure 9. A graphical illustration of our neural inference method. A recurrent neural approximator updates the posterior of the low-level model parameters θ_t each time step t and yields the posterior over the high-level model parameters η considering all available data. The low-level prior constrains the initial dynamic parameter values θ_0, which then get passed to the high-level transition model. Together, the two priors and the two models comprise a stochastic simulator that trains the neural approximator to perform amortized Bayesian updating.

free-form posteriors as a dynamic extension of the BayesFlow method[45]. We use the former approach for the toy Coal Mining benchmark and the latter approach for all other experiments in this work.

Simulation-based training. Figure 9 graphically illustrates the rationale of our simulation-based inference approach. To train the networks, we treat the forward-time generative model as a simulator and employ Eq. (1) to generate multiple sets of simulated parameters and trajectories $(\eta, \theta_{1:T}, x_{1:T})$. We then minimize the Monte Carlo estimate of the following criterion

$$\mathbb{L}(\psi) = \min_{\psi} \mathbb{E}_{p(\eta,\theta_{1:T},x_{1:T})}\left[-\sum_{t=1}^{T}\log q(\theta_t,\eta|x_{1:t},\psi)\right], \tag{14}$$

where $\mathbb{E}[\cdot]$ denotes an expectation over the dynamic generative model and $\psi = (\psi^{(r)}, \psi^{(g)})$ denotes the collection of all trainable neural network parameters. This criterion ensures that the approximate posteriors match the analytic posteriors induced by the dynamic model and can be minimized either via online (i.e., generating dynamic simulations on the fly) or via offline training (i.e., using a set of pre-computed dynamic simulations).

A non-stationary diffusion decision model. To illustrate the potential of our approach, we will reformulate in superstatistical terms a popular cognitive model for analyzing human response times (RTs) in binary decision tasks, namely the DDM. The standard DDM describes the microscopic dynamics of perceptual evidence accumulations via a simple stochastic ordinary differential equation (SDE). Accordingly, the accumulated evidence x_j in experimental task j follows a random walk with drift and Gaussian noise:

$$\mathrm{dx_j} = \mathrm{v}\mathrm{dt_s} + \mathrm{z}\sqrt{\mathrm{dt_s}} \quad \text{with} \quad \mathrm{z} \sim \mathcal{N}(0,1), \tag{15}$$

where t_s represents time on a continuous microscopic scale (i.e., during forced-choice decision making). A core assumption of the DDM is that task-relevant information (i.e., perceptual evidence) accumulates at a constant rate (v). This process runs in a corridor with two absorbing boundaries, which represent two decision alternatives. As

soon as the accumulated evidence x_j reaches either a pre-defined threshold (a) or 0, the model makes a categorical decision D_j for the alternative favored by the collected evidence:

$$D_j = \begin{cases} 1, & \text{if } x_j \geq a \\ 0, & \text{if } x_j \leq 0 \end{cases}. \tag{16}$$

Further, the model assumes a constant additive factor (τ) accounting for non-decision processes, such as encoding or motor responses. Thus, the standard (static) DDM has three key parameters $\theta = (v, a, \tau)$. The starting point of the decision process is either estimated as an additional parameter or fixed at $a/2$.

The typical assumption of the standard DDM is that the parameters θ remain stationary for the duration of a given cognitive task. In order to relax this restrictive assumption, the standard DDM has been extended to incorporate so-called inter-trial-variability for the drift rate and non-decision time parameters[72,73]. In this way, the extended DDM concedes that these cognitive parameters are not static but vary over time. However, the assumed variation is homogeneous and memoryless, and the generative model still yields IID data, that is, the transition model coincides with independent sampling and reduces to $\theta_t = \mathcal{T}(\eta, \xi_t)$.

In contrast, our superstatistical model assumes a stateful Gaussian process (GP) high-level model, which describes the trial-by-trial dynamics of the DDM parameters according to Eqs. (5) and (6) (see the Appendix for more details).

Thereby, we want to demonstrate that our estimation method can tackle very flexible transition models $\mathcal{T}$, as long as we can simulate data from the low-level model. However, we also fit a DDM with a simpler Gaussian random walk transition model to the data described in the "Human data application" section. This simpler model corroborates our findings by suggesting qualitatively similar parameter dynamics, but yields less sharp predictions on unseen data than its GP counterpart (see Appendix for more details).

Data and code availability

All models, data, and scripts for reproducing the results of this work are publicly available in the project's repository https://github.com/bayesflow-org/Neural-Superstatistics. The neural superstatistics method is implemented in the `BayesFlow` Python library for amortized Bayesian workflows[74].

Received: 15 March 2023; Accepted: 8 August 2023
Published online: 23 August 2023

References

1. Farrell, S. & Lewandowsky, S. *Computational Modeling of Cognition and Behavior*. (Cambridge University Press). https://doi.org/10.1017/CBO9781316272503 (2018).
2. Voss, A., Nagler, M. & Lerche, V. Diffusion models in experimental psychology. *Exp. Psychol.* **60**(6), 385–402. https://doi.org/10.1027/1618-3169/a000218 (2013).
3. Ratcliff, R., Smith, P. L., Brown, S. D. & McKoon, G. Diffusion decision model: Current issues and history. *Trends Cogn. Sci.* **20**(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007 (2016).
4. Eckstein, M. K. & Collins, A. G. E. Computational evidence for hierarchically structured reinforcement learning in humans. *Proc. Natl. Acad. Sci.* **117**(47), 29381–29389. https://doi.org/10.1073/pnas.1912330117 (2020).
5. Gershman, S. J. & Daw, N. D. Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Ann. Rev. Psychol.* **68**, 101–128. https://doi.org/10.1146/annurev-psych-122414-033625 (2017).
6. Oberauer, K. *et al.* Benchmarks for models of short-term and working memory. *Psychol. Bull.* **144**(9), 885–958. https://doi.org/10.1037/bul0000153 (2018).
7. Yoo, A. H. & Collins, A. G. E. How working memory and reinforcement learning are intertwined: A cognitive, neural, and computational perspective. *J. Cogn. Neurosci.* **34**(4), 551–568. https://doi.org/10.1162/jocn_a_01808 (2022).
8. Van Orden, G. C., Holden, J. G. & Turvey, M. T. Self-organization of cognitive performance. *J. Exp. Psychol. Gen.* **132**(3), 331–350. https://doi.org/10.1037/0096-3445.132.3.331 (2003).
9. Wagenmakers, E.-J., Farrell, S. & Ratcliff, R. Estimation and interpretation of 1/Fα noise in human cognition. *Psychon. Rev.* **11**(4), 579–615. https://doi.org/10.3758/BF03196615 (2004).
10. Gilden, D. L. Cognitive emissions of 1/f noise. *Psychol. Rev.* **108**(1), 33–56. https://doi.org/10.1037/0033-295x.108.1.33 (2001).
11. Collins, A. G. E. & Frank, M. J. Withinand across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proc. Natl. Acad. Sci.* **115**(10), 2502–2507. https://doi.org/10.1073/pnas.1720963115 (2018).
12. Brockmole, J. R. & Logie, R. H. Age-related change in visual working memory: A study of 55,753 participants aged 8–75. *Front. Psychol.* **4**, 12. https://doi.org/10.3389/fpsyg.2013.00012 (2013).
13. von Krause, M., Radev, S. T. & Voss, A. Mental speed is high until age 60 as revealed by analysis of over a million participants. *Nat. Hum. Behav.* **6**, 700–708. https://doi.org/10.1038/s41562-021-01282-7 (2022).
14. Riley, M. A. & Holden, J. G. Dynamics of cognition. *WIREs Cogn. Sci.* **3**(6), 593–606. https://doi.org/10.1002/wcs.1200 (2012).
15. Favela, L. H. Cognitive science as complexity science. *WIREs Cogn. Sci.* **11**(4), e1525. https://doi.org/10.1002/wcs.1525 (2020).
16. Ratcliff, R. & Van Dongen, H. P. Diffusion model for one-choice reaction-time tasks and the cognitive effects of sleep deprivation. *Proc. Natl. Acad. Sci.* **108**(27), 11285–11290. https://doi.org/10.1073/pnas.1100483108 (2011).
17. Walsh, M. M., Gunzelmann, G. & Van Dongen, H. Computational cognitive modeling of the temporal dynamics of fatigue from sleep loss. *Psychon. Bull. Rev.* **24**(6), 1785–1807. https://doi.org/10.3758/s13423-017-1243-6 (2017).
18. Kahana, M. J., Aggarwal, E. V. & Phan, T. D. The variability puzzle in human memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **44**(12), 1857. https://doi.org/10.1037/xlm0000553 (2018).
19. Evans, N. J., Brown, S. D., Mewhort, D. J. & Heathcote, A. Refining the law of practice. *Psychol. Rev.* **125**(4), 592. https://doi.org/10.1037/rev0000105 (2018).
20. Mittner, M., Hawkins, G. E., Boekel, W. & Forstmann, B. U. A neural model of mind wandering. *Trends Cogn. Sci.* **20**(8), 570–578. https://doi.org/10.1016/j.tics.2016.06.004 (2016).
21. Christoff, K., Irving, Z. C., Fox, K. C., Spreng, R. N. & Andrews-Hanna, J. R. Mind-wandering as spontaneous thought: A dynamic framework. *Nat. Rev. Neurosci.* **17**(11), 718–731. https://doi.org/10.1038/nrn.2016.113 (2016).
22. Brosowsky, N. P., DeGutis, J., Esterman, M., Smilek, D. & Seli, P. Mind wandering, motivation, & task performance over time: Evidence that motivation insulates people from the negative effects of MindWandering. *Psychol. Conscious. Theory Res. Pract.* https://doi.org/10.1037/cns0000263 *(2020).*

23. Kiuru, N. *et al.* The dynamics of motivation, emotion, and task performance in simulated achievement situations. *Learn. Individ. Differ.* **80**, 101873. https://doi.org/10.1016/j.lindif.2020.101873 (2020).
24. Mark, C. *et al.* Bayesian model selection for complex dynamic systems. *Nat. Commun.* **9**(1), 1803. https://doi.org/10.1038/s41467-018-04241-5 (2018).
25. Beck, C. & Cohen, E. G. D. Superstatistics. *Phys. A Stat. Mech. Appl.* **322**, 267–275. https://doi.org/10.1016/S0378-4371(03)00019-0 (2003).
26. Hanel, R., Thurner, S. & Gell-Mann, M. Generalized entropies and the transformation group of superstatistics. *Proc. Natl. Acad. Sci.* **108**(16), 6390–6394. https://doi.org/10.1073/pnas.1103539108 (2011).
27. Kucharský, Š, Tran, N.-H., Veldkamp, K., Raijmakers, M. & Visser, I. Hidden Markov models of evidence accumulation in speeded decision tasks. *Comput. Brain Behav.* **4**(4), 416–441. https://doi.org/10.1007/s42113-021-00115-0 (2021).
28. Gunawan, D., Hawkins, G. E., Kohn, R., Tran, M.-N. & Brown, S. D. Time-evolving psychological processes over repeated decisions. *Psychol. Rev.* **129**(3), 438. https://doi.org/10.1037/rev0000351 (2022).
29. Metzner, C., Schilling, A., Traxdorf, M., Schulze, H. & Krauss, P. Sleep as a random walk: A super-statistical analysis of EEG data across sleep stages. *Commun. Biol.* **4**(1), 1–11. https://doi.org/10.1038/s42003-021-02912-6 (2021).
30. Yalcin, G. C., Rabassa, P. & Beck, C. Extreme event statistics of daily rainfall: Dynamical systems approach. *J. Phys. A Math. Theor.* **49**(15), 154001. https://doi.org/10.1088/1751-8113/49/15/154001 (2016).
31. Rabassa, P. & Beck, C. Superstatistical analysis of sea-level fluctuations. *Phys. A Stat. Mech. Appl.* **417**, 18–28. https://doi.org/10.1016/j.physa.2014.08.068 (2015).
32. Williams, G., Schäfer, B. & Beck, C. Superstatistical approach to air pollution statistics. *Phys. Rev. Res.* **2**(1), 013019. https://doi.org/10.1103/PhysRevResearch.2.013019 (2020).
33. Bogachev, M. I., Markelov, O. A., Kayumov, A. R. & Bunde, A. Superstatistical model of bacterial DNA architecture. *Sci. Rep.* **7**(1), 43034. https://doi.org/10.1038/srep43034 (2017).
34. Van der Straeten, E. & Beck, C. Superstatistical fluctuations in time series: Applications to share- price dynamics and turbulence. *Phys. Rev. E* **80**(3), 036108. https://doi.org/10.1103/PhysRevE.80.036108 (2009).
35. Denys, M., Gubiec, T., Kutner, R., Jagielski, M. & Stanley, H. E. Universality of market superstatistics. *Phys. Rev. E* **94**(4), 042305. https://doi.org/10.1103/PhysRevE.94.042305 (2016).
36. Voss, A. & Voss, J. Fast-Dm: A free program for efficient diffusion model analysis. *Behav. Res. Methods* **39**(4), 767–775. https://doi.org/10.3758/BF03192967 (2007).
37. Ratcliff, R. A theory of memory retrieval. *Psychol. Rev.* **85**(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59 (1978).
38. von Krause, M. *et al.* Stability and change in diffusion model parameters over two years. *J. Intell.* **9**(2), 26. https://doi.org/10.3390/jintelligence9020026 (2021).
39. Diederich, A. & Busemeyer, J. R. Modeling the effects of payoff on response bias in a perceptual discrimination task: Bound-change, drift- rate-change, or two-stage-processing hypothesis. *Percept. Psychophys.* **68**(2), 194–207. https://doi.org/10.3758/BF03193669 (2006).
40. Urai, A. E., de Gee, J. W., Tsetsos, K. & Donner, T. H. Choice history biases subsequent evidence accumulation. *eLife* **8**, e46331. https://doi.org/10.7554/eLife.46331 (2019).
41. van Rooij, M. M. J. W., Favela, L. H., Malone, M. & Richardson, M. J. Modeling the dynamics of risky choice. *Ecol. Psychol.* **25**(3), 293–303. https://doi.org/10.1080/10407413.2013.810502 (2013).
42. Gasimova, F. *et al.* Dynamical systems analysis applied to working memory data. *Front. Psychol.* **5**, 687. https://doi.org/10.3389/fpsyg.2014.00687 (2014).
43. Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. Bayesian workflow. https://doi.org/10.48550/ARXIV.2011.01808 (2020).
44. van de Schoot, R. *et al.* Bayesian statistics and modelling. *Nat. Rev. Methods Prim.* **1**(1), 1–26. https://doi.org/10.1038/s43586-020-00001-2 (2021).
45. Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L. & Köthe, U. BayesFlow: Learning complex stochastic modelswith invertible neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **33**(4), 1452–1466. https://doi.org/10.1109/TNNLS.2020.3042395 (2020).
46. Mestdagh, M., Verdonck, S., Meers, K., Loossens, T. & Tuerlinckx, F. Prepaid parameter estimation without likelihoods. *PLoS Comput. Biol.* **15**(9), e1007181. https://doi.org/10.1371/journal.pcbi.1007181 (2019).
47. Cranmer, K., Brehmer, J. & Louppe, G. The frontier of simulation-based inference. *Proc. Natl. Acad. Sci.* **117**(48), 30055–30062. https://doi.org/10.1073/pnas.1912789117 (2020).
48. Bürkner, P.-C., Scholz, M., & Radev, S. Some models are useful, but how do we know which ones? Towards a unified Bayesian model taxonomy. https://doi.org/10.48550/arXiv.2209.02439 (2022).
49. Carpenter, B. *et al.* Stan: A probabilistic programming language. *J. Stat. Softw.* **76**(1), 1–32. https://doi.org/10.18637/jss.v076.i01 (2017).
50. Neal, R. M. *et al.* MCMC using Hamiltonian dynamics. *Handb. Markov Chain Monte Carlo* **2**(11), 2. https://doi.org/10.1201/b10905-7 (2011).
51. Betancourt, M. Calibrating model-based inferences and decisions. https://doi.org/10.48550/ARXIV.1803.08393 (2018).
52. Schad, D. J., Betancourt, M. & Vasishth, S. Toward a principled Bayesian workflow in cognitive science. *Psychol. Methods* **26**, 103–126. https://doi.org/10.1037/met0000275 (2021).
53. Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. Validating Bayesian inference algorithms with simulation-based calibration. https://doi.org/10.48550/ARXIV.1804.06788 (2018).
54. Säilynoja, T., Bürkner, P.-C., & Vehtari, A. Graphical test for discrete uniformity and its applications in goodness of fit evaluation and multiple sample comparison. https://doi.org/10.48550/ARXIV.2103.10522 (2021).
55. Evans, N. J. & Brown, S. D. People adopt optimal policies in simple decision-making, after practice and guidance. *Psychon. Bull. Rev.* **24**, 597–606. https://doi.org/10.3758/s13423-016-1135-1 (2017).
56. Yarkoni, T. & Westfall, J. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect. Psychol. Sci. J. Assoc. Psychol. Sci.* **12**(6), 1100–1122. https://doi.org/10.1177/1745691617693393 (2017).
57. Bürkner, P.-C., Gabry, J. & Vehtari, A. Approximate leave-future-out cross-validation for Bayesian time series models. *J. Stat. Comput. Simul.* **90**(14), 2499–2523. https://doi.org/10.1080/00949655.2020.1783262 (2020).
58. Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. A kernel two- sample test. *J. Mach. Learn. Res.* **13**(25), 723–773 (2012).
59. Radev, S. T., D'Alessandro, M., Mertens, U. K., Voss, A., Köthe U., & Bürkner, P. -C. Amortized Bayesian model comparison with evidential deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* **34**(8), 4903–4917. https://doi.org/10.1109/TNNLS.2021.3124052 (2021).
60. Schmitt, M., Radev, S. T., & Bürkner, P . -C. Meta-uncertainty in Bayesian model comparison. https://doi.org/10.48550/ARXIV.2210.07278 (2022).
61. Toda, H. Y. & Phillips, P. C. Vector autoregression and causality: A theoretical overview and simulation study. *Econ. Rev.* **13**(2), 259–285. https://doi.org/10.1080/07474939408800286 (1994).
62. Rasmussen, C. E. Gaussian processes in machine learning. In *Summer School on Machine Learning*. Vol. **3176** (eds Bousquet, O.) 63–71 (Springer, Berlin, Heidelberg). https://doi.org/10.1007/978-3-540-28650-9_4 (2003).

63. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: A review for statisticians. *J. A. Stat. Assoc.* **112**(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773 (2017).
64. van Rooij, I., Blokpoel, M., Kwisthout, J. & Wareham, T. Applications. *Intractability Guide Class. Parameterized Complex. Anal.* https://doi.org/10.1017/9781107358331 (2019).
65. Greenberg, D., Nonnenmacher, M., & Macke, J. Automatic posterior transformation for likelihood-free inference, in *International Conference on Machine Learning*, 2404–2414. https://doi.org/10.48550/arXiv.1905.07488 (2019)
66. Gers, F. A., Schmidhuber, J. & Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural Comput.* **12**(10), 2451–2471. https://doi.org/10.1049/cp:19991218 (2000).
67. Bloem-Reddy, B. & Teh, Y. W. Probabilistic symmetries and invariant neural networks. *J. Mach. Learn. Res.* **21**, 90–1 (2020).
68. Ardizzone, L., Lüth, C., Kruse, J., Rother, C., & Köthe, U. Guided image generation with conditional invertible neural networks. https://doi.org/10.48550/arXiv.1907.02392 (2019).
69. Papamakarios, G., Pavlakou, T., & Murray, I. Masked autoregressive flow for density estimation. *Adv. Neural Inf. Process. Syst.* **30** (2017).
70. Kendall, A. & Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Adv. Neural Inf. Process. Syst.* **30** (2017).
71. Radev, S. T., Mertens, U. K., Voss, A. & Köthe, U. Towards end-to-end likelihood-free inference with convolutional neural networks. *Br. J. Math. Stat. Psychol.* **73**(1), 23–43. https://doi.org/10.1111/bmsp.12159 (2020).
72. Ratcliff, R. & Rouder, J. N. Modeling response times for two-choice decisions. *Psychol. Sci.* **9**(5), 347–356. https://doi.org/10.1111/1467-9280.00067 (1998).
73. Ratcliff, R. & Tuerlinckx, F. Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychon. Bull. Rev.* **9**(3), 438–481. https://doi.org/10.3758/BF03196302 (2002).
74. Radev, S. T., Schmitt, M., Schumacher, L., Elsemüller, L., Pratz, V., Schälte, Y., Köthe, U., & Bürkner, P.-C. BayesFlow: Amortized Bayesian workflows with neural networks. https://doi.org/10.48550/arXiv.2306.16015 (2023).

Acknowledgements
We thank Daniel Durstewitz and Lasse Elsemüller for their helpful feedback on this project. We also thank Marie Wieschen for her efforts in data collection. L.S. and A.V. were supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; Grant Number GRK 2277 "Statistical Modeling in Psychology"). P.-C.B. was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC-2075 - 390740016 (the Stuttgart Cluster of Excellence SimTech). S.T.R was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC-2181 - 390900948 (the Heidelberg Cluster of Excellence STRUCTURES). For the publication fee we acknowledge financial support by Deutsche Forschungsgemeinschaft within the funding programme "Open Access Publikationskosten" as well as by Heidelberg University.

Author contributions
L.S., S.T.R., designed the research, created and applied the models, and wrote the initial draft of the manuscript. P.C.B. significantly contributed with his statistical and scientific expertise to the methods and the written content. A.V. and U.K. supervised the project and contributed to all stages.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-40278-3.

Correspondence and requests for materials should be addressed to L.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023