REGULARISED NEURAL NETWORKS MIMIC HUMAN INSIGHT

Anika T. Löwe[1,2] **Léo Touzo**[3] **Paul S. Muhle-Karbe**[4,5,6] **Andrew M. Saxe**[7,8,9]

Christopher Summerfield*[4] **Nicolas W. Schuck***[1,2,10]

February 23, 2023

1 Max Planck Research Group NeuroCode, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany
2 Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Lentzeallee 94, 14195 Berlin, Germany
3 Department of Physics, Ecole Normale Supérieure, Paris, France 75005
4 Department of Experimental Psychology, University of Oxford, Oxford OX2 6GG, UK
5 School of Psychology, University of Birmingham, Birmingham B15 2SA, UK
6 Centre for Human Brain Health, University of Birmingham, Birmingham B15 2TT, UK
7 Gatsby Computational Neuroscience Unit, University College London, London, UK W1T 4JG
8 Sainsbury Wellcome Centre, University College London, London, UK W1T 4JG
9 CIFAR Azrieli Global Scholar, CIFAR, Toronto, Canada
10 Institute of Psychology, Universität Hamburg, Von-Melle-Park 5, 20254 Hamburg, Germany

*equal contribution

E-mail: loewe@mpib-berlin.mpg.de

ABSTRACT

Humans sometimes show sudden improvements in task performance that have been linked to moments of insight. Such insight-related performance improvements appear special because they are preceded by an extended period of impasse, are unusually abrupt, and occur only in some, but not all, learners. Here, we ask whether insight-like behaviour also occurs in artificial neural networks trained with gradient descent algorithms. We compared learning dynamics in humans and regularised neural networks in a perceptual decision task that provided a hidden opportunity which allowed to solve the task more efficiently. We show that humans tend to discover this regularity through insight, rather than gradually. Notably, neural networks with regularised gate modulation closely mimicked behavioural characteristics of human insights, exhibiting delay of insight, suddenness and selective occurrence. Analyses of network learning dynamics revealed that insight-like behaviour crucially depended on noise added to gradient updates, and was preceded by "silent knowledge" that is initially suppressed by regularised (attentional) gating. This suggests that insights can arise naturally from gradual learning, where they reflect the combined influences of noise, attentional gating and regularisation.

Keywords Insights · Neural Networks · Learning

1 Introduction

he ability to learn from experience is common to all animals and some artificial agents. Neural networks trained with stochastic gradient descent (SGD) are a current theory of human learning that can account for a wide range of learning phenomena, but while standard networks seem to imply that all learning is gradual, humans may sometimes learn in an abrupt manner.

Such non-linear improvements in task performance or problem solving have been described as insights or aha-moments Köhler [1925], Durstewitz et al. [2010], and are often thought to reflect a qualitatively different, discrete learning mechanism Stuyck et al. [2021], Weisberg [2015]. One prominent idea, dating back to Gestalt psychology Köhler [1925], is that an insight occurs when an agent has found a novel problem solution by restructuring an existing task representation Kounios and Beeman [2014]. It has also been noted that humans often lack the ability to trace back the cognitive process leading up to an insight Jung-Beeman et al. [2004], suggesting that insights involve unconscious processes becoming conscious. Moreover, so called "aha-moments" can sometimes even be accompanied by a feeling of relief or pleasure in humans Kounios and Beeman [2014], Danek et al. [2014], Kounios and Beeman [2015]. Such putative uniqueness of the insight phenomenon would also be in line with work that has related insights to brain regions distinct from those associated with gradual learning Shen et al. [2018], Jung-Beeman et al. [2004]. These include, for instance, the anterior temporal gyrus Jung-Beeman et al. [2004], Tik et al. [2018], as well as subcortical areas such as the left amygdala or right hippocampal gyrus Shen et al. [2018]. Altogether, these findings have led psychologists and neuroscientists to propose that insights are governed by a distinct learning process Jung-Beeman et al. [2004], that cannot be accounted for by current common theories of learning.

Here, we show that insight-like phenomena can occur without dedicated mechanisms for re-representation or a division of labour between conscious and unconscious processes. Our argument does not concern the subjective experiences related to insights, but focuses on showing how insight-like behaviour can emerge from gradual learning algorithms. Specifically, we aim to explain the following three main observations Schuck et al. [2015, 2022], Gaschler et al. [2019, 2013, 2015]: First, insights trigger abrupt behavioural changes, accompanied by meta-cognitive suddenness (a "sudden and unexpected flash") Bowden et al. [2005], Gaschler et al. [2013], Metcalfe and Wiebe [1987], Weisberg [2015]. These abrupt behavioural changes are often accompanied by fast neural transitions, which have been observed in humans as well as animals Durstewitz et al. [2010], Karlsson et al. [2012], Miller and Katz [2010], Schuck et al. [2015], Allegra et al. [2020]. Second, insights occur selectively in some subjects, while for others improvement in task performance arises only gradually, or never Schuck et al. [2015]. Finally, insights occur "spontaneously", i.e. without the help of external cues Friston et al. [2017], and are therefore observed after a seemingly random duration of impasse Ohlsson [1992] or delay after a change in environmental contingencies for different participants. In other words, participants seem to be "blind" to the new solution for an extended period of time, before it suddenly occurs to them. Insights are thus characterised by suddenness, selectivity, and delay.

The idea that insight-like behaviour can arise naturally from gradual learning is supported by previous work on neural networks trained with gradient descent Power et al. [2022]. Saxe and collaegues Saxe et al. [2014], for instance, have shown that non-linear learning dynamics, i.e. suddenness in the form of saddle points and stage-like transitions, can result from gradient descent even in linear neural networks, which could explain sudden behavioural improvements. Other work has shown a delayed or stage-like mode of learning in neural networks that is reminiscent of the period of impasse observed in humans, and reflected for instance in the structure of the input data Saxe et al. [2019a], Schapiro and McClelland [2009], McClelland and Rogers [2003], or information compression of features that at some point seemed task-irrelevant Flesch et al. [2022], Saxe et al. [2019b]. Finally, previous work has also found substantial individual differences between neural network instances that are induced by random differences in weight initialisation, noise, or the order of training examples Bengio et al. [2009], Flesch et al. [2018], which can become larger with training Mehrer et al. [2020].

Two factors that influence discontinuities in learning in neural networks are regularisation and gating. Regularisation plays a key role in the suppression of input features. While this avoids overfitting and can help a network to escape a local minimum Liu et al. [2020], it might also cause above mentioned "blindness" to a solution that involves inputs which were once erroneously deemed irrelevant. Gating, on the other hand, is known to cause exponential transitions in learning that are widely seen in multiplicative dynamical systems like the logistic growth model. Both techniques are widely used in artificial neural networks Bishop [2006], Krishnamurthy et al. [2022], Jozefowicz et al. [2015], and are inspired by biological brains Groschner et al. [2022], Poggio et al. [1985], Costa et al. [2017]. Regularisation and gating could therefore be important aspects of network structure and training that are related to the temporary impasse followed by a sudden performance change, akin to insight-like behaviour.

Based on these findings, we hypothesised that insight-like behaviour – as characterised by suddenness, selectivity, and delay – can occur in simple neural networks trained with gradient descent. As indicated above, a simple neural network architecture with multiplicative gates and regularisation served as our candidate model. We predicted that due to the multiplicative nature of gating, regularising gates during training could lead to blindness of some relevant features that are key to a solution. We focused specifically on L1-regularisation because it forces gates of irrelevant inputs most strongly towards 0, compared to the less aggressive L2-regularisation. We reason that applying L1-regularisation, besides creating non-linear learning dynamics due to the multiplicative nature of the weights and gates, will lead to a sustained suppression period before the fast transition, similar to the delay observed in humans.

Results

To study insight-like learning dynamics, 99 participants and 99 neural networks, matched in their behavioural performance to their human counterparts (see below for details), performed a decision task that required a binary choice about circular arrays of moving dots Rajananda et al. [2018] for humans and a symbolic version in which inputs were two scalars for networks. Dots were characterised by two features with different degrees of noise, (1) a motion direction (four possible orthogonal directions: NW, NE, SW, SE) and (2) a colour (orange or purple) (Fig.1A). Participants and networks had to learn the correct choice in response to each stimulus from trial-wise binary feedback, and were not instructed which features of the stimulus to pay attention to.

Importantly, the task provided a hidden opportunity to improve one's decision strategy that could be discovered through insight, similar to the spontaneous strategy switch task developed earlier Schuck et al. [2015]. Participants first underwent an initial training phase (4 blocks, 100 trials each in humans, 8 blocks/800 trials in networks), during which only the motion direction predicted the correct choice, while stimulus colour was random (*motion phase*, see Fig.1D). Without any announcement, stimulus colour became predictive of the correct response in a later phase, such that from then on both features could be used to determine choice (*motion and colour phase*, 5 blocks for humans and networks, Fig.1D). Such unannounced changes in feature relevance elicit insights, i.e. behaviour exhibits changes that are sudden, delayed and selective, and post-experimental verbal questionnaires indicate that these changes go hand in hand with gaining consciousness about the new regularity Gaschler et al. [2019].

To test whether and when participants employed the hidden colour insight, we assessed whether choices were sensitive to the motion direction (using the colour insight meant that stimulus motion could be ignored). Specifically, following an initial pre-training period (see Methods) the amount of motion noise varied randomly in five levels of motion coherence (5%, 10%, 20%, 30% or 45%, noise variability started in the last two blocks before the onset of the *motion and colour phase*). Behaviour in trials with the highest amount of noise in dot motion (5% coherence, 30 trials per block) was then used to test whether participants had an insight about the usefulness of the colour, as high performance in these trials could only be achieved by using the colour information Schuck et al. [2015]. Colour difficulty was constant and consistently allowed participants and networks to identify colour easily. A second measure that we used to investigate insight was a post-experimental questionnaire, in which participants were asked whether (1) they had noticed a rule in the experiment, (2) how long it took them to notice the rule, (3) whether they had paid attention to colour during choice. The questionnaire was administered after the *motion and colour phase*, and was followed by a instruction block that served as a sanity check (see Methods).

Human Behaviour

Data from the *training phase*, during which motion directions were highly coherent and colours changed randomly (Block 1-2, dark grey tiles in Fig 1D), showed that participants learned the response mapping for the four motion directions well (78% correct, t-test against chance: $t(98) = 30.8$, $p < .001$). In the following task phase, noise was added to the motion, while the colour remained uncorrelated (*motion phase*, blocks 3-4, grey tiles in Fig. 1D). This resulted in an accuracy gradient that depended on noise level (linear mixed effects model of accuracy: $\chi^2(1) = 726.36$, $p < .001$; RTs: $\chi^2(1) = 365.07$, $p < .001$; N = 99, Fig.2A). Crucially, performance during this phase was heavily diminished in the conditions with the largest amounts of motion noise, i.e. the two lowest coherence conditions: the percentage of correct choices was at only 60% and 63% in the two lowest coherence conditions, and did not change over time (paired t-test block 3 vs 4: $t(195.9) = -1.13$, $p = 0.3$, $d = 0.16$). Hence, performance (improvements) largely beyond these low baseline levels can only be attributed to colour use, rather than heightened motion sensitivity.

The noise level continued to influence performance in the *motion and colour phase*, as evidenced by a difference between performance in high vs. low coherence trials (20, 30 & 45% vs 5 & 10 % coherent motion, respectively; $M = 93 \pm 6\%$ vs $M = 77 \pm 12\%$; $t(140.9) = 12.5$, $p < .001$, $d = 1.78$, see Fig.2A-B). Notably, however, the onset of the colour correlation triggered performance improvements across all coherence levels ($t(187.2) = -12.4$, $p < .001$, $d = 1.8$; end of *motion phase*: $M = 78 \pm 7\%$ vs. end of *motion and colour phase*: $M = 91 \pm 8\%$), contrasting the stable performance found during the motion phase and suggesting that at least some participants leveraged colour information once available.

We asked whether these improvements are related to gaining conscious insight by analysing the post-experimental questionnaire. Results show that conscious knowledge about the colour regularity arose in some, but not all, participants: 57.6% (57/99) reported in the questionnaire to have used colour, while 42.4% indicated to not have noticed or used the colour. We then checked whether these conscious insights were related to the key behavioural characteristics of suddenness, selectivity, and variable delay. To test for suddenness, we fitted each participant's time course of accuracy on low coherence trials by either (1) a linear ramp or (2) a sigmoid function. While a linear model (free parameters: intercept y_0 and slope m) can capture gradual improvements in behaviour that might occur without insight, a better fit

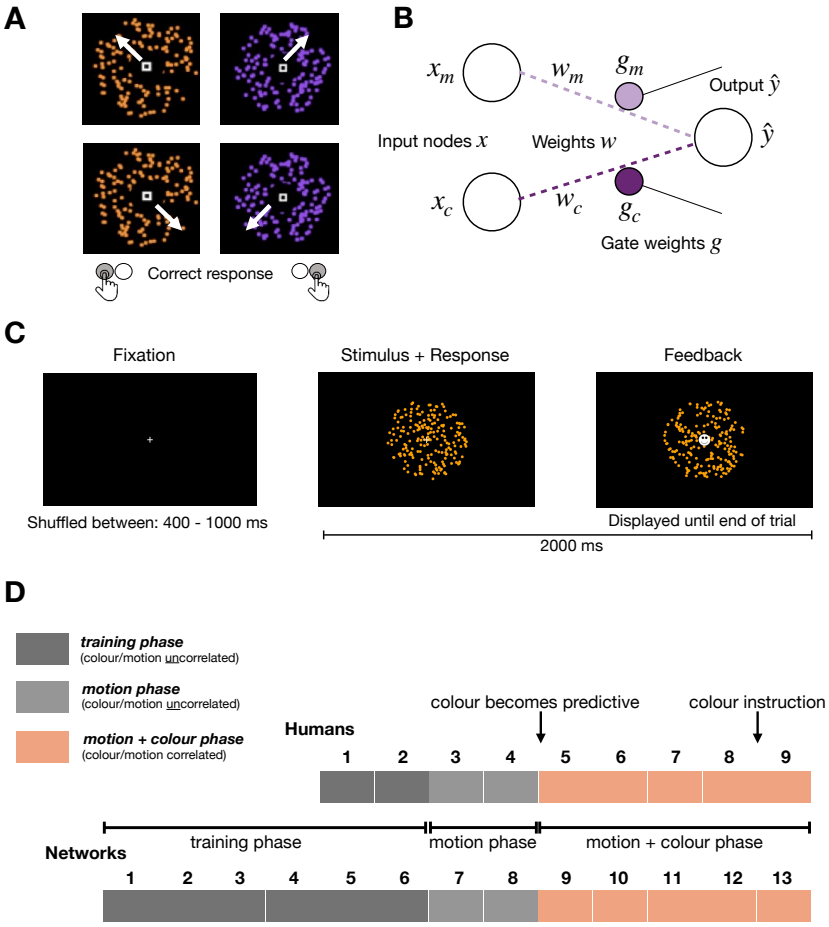


Figure 1: Stimuli and task design **(A)** Stimuli and stimulus-response mapping: dot clouds were either coloured in orange or purple and moved to one of the four directions NW, NE, SE, SW with varying coherence. A left response key, "X", corresponded to the NW/SE motion directions, while a right response key "M" corresponded to NE/SW directions. **(B)** Schematic of simple neural network with regularised gate modulation with colour codes corresponding to respective colour and motion *weights* and *gates*. Number of nodes shown is the exact number of nodes used in the neural network simulations. **(C)** Trial structure: a fixation cue is shown for a duration that is shuffled between 400, 600, 800 and 1000 ms. The random dot cloud stimulus is displayed for 2000 ms. A response can be made during these entire 2000 ms and a feedback cue will replace the fixation cue at the centre of the stimulus until the end of the stimulus display duration. **(D)** Task structure of the two-alternative forced choice task for humans and neural networks: each block consisted of 100 trials. Colour was predictive of correct choices and correlated with motion directions as well as correct response buttons in the last five blocks (*motion and colour phase*). In the last block, humans and networks were instructed to use colour inputs to respond. In the *motion phase*, colour changed randomly and was not predictive. A first training block only for humans contained only 100% motion coherence trials to familiarise subjects with the S-R mapping. The remaining training blocks contained only high coherence (0.2, 0.3, 0.45) trials.

of a non-linear sigmoid function indicates sudden behavioural transitions (free parameters: slope m, inflection point t_s and function maximum y_{max}). Performance across participants on low coherence trials was best fit by a non-linear sigmoid function, indicating at least a subsection of putative insight participants (BIC sigmoid function: $M = -6.7$, $SD = 0.7$, protected exceedance probability: 1, BIC linear function: $M = -6.4$, $SD = 0.5$, protected exceedance probability: 0). The sigmoid function also outperformed a step function with free parameters inflection point t_s and function maximum y_{max} (BIC step function: $M = -6.5$, $SD = 0.6$, protected exceedance probability: 0) (Fig.2D-E, Fig.S2).

We next tested insight selectivity, i.e. whether all participants, or only a subset, showed abrupt behavioural transitions, as indicated by participants' self-reports. Chance level of suddenness was determined by an out-of-sample null distribution of sigmoid steepness derived from a control experiment (N = 20), in which participants performed an identical task, except that colour never started to correlate with motion, and hence no insight was possible. Fitting the same sigmoid

function to this data, we derived a baseline distribution of the steepness (see Methods for details). Comparing the steepness values (at the inflection point) obtained in our experimental sample to the baseline distribution derived from the control group with no colour correlation, showed that about half of participants (48/99, 48.5%) had values larger than the 100% percentile of the control distribution. This thus suggests that truly abrupt insight occurred selectively in these "insight participants" (Fig.2F). 79.2% of the participants classified as insight subjects also self-reported to have used colour to make correct choices (Fig. S6A-B). Hence, our behavioural marker of unexpectedly sudden performance changes can serve as a valid indicator for insight.

We validated our behavioural metric of selectivity through additional analyses. Splitting behaviour into two separate insight (participants with steepness values larger than the 100% percentile of the control distribution) and no-insight groups showed that, as expected based on the dependency of accuracy and our behavioural metric, insight subjects started to perform significantly better in the lowest coherence trials once the *motion and colour phase* (Fig.2C) started, (mean proportion correct in *motion and colour phase*: $M = 83 \pm 10\%$), compared to participants without insight ($M = 66 \pm 8\%$) ($t(92) = 9.5$, $p < .001$, $d = 1.9$). Unsurprisingly, a difference in behavioural accuracy between insight participants and no-insight participants also held when all coherence levels were included ($M = 91 \pm 5\%$ vs. $M = 83 \pm 7\%$, respectively, t-test: $t(95.4) = 6.9$, $p < .001$, $d = 1.4$). Interestingly, accuracy in the *motion phase*, which was not used in steepness fitting, did not differ between groups (low coherence trials: $M = 59\%$, vs. $M = 62\%$; $t(94.4) = -1.9$, $p = 0.07$, $d = 0.38$; all noise levels: $M = 76\%$ vs $M = 76\%$,$t(96) = 0.45$, $p = 0.7$, $d = 0.09$). Reaction times, which are independent from the choices used in model fitting and thus served as a sanity check for our behavioural metric split, reflected the same improvements upon switching to the colour strategy. Subjects that showed insight about the colour rule ($M = 748.47 \pm 171.1$ ms) were significantly faster ($t(96.9) = -4.9$, $p < .001$, $d = 0.97$) than subjects that did not ($M = 924.2 \pm 188.9$ ms) on low coherence trials, as well as over all noise levels ($t(97) = -3.8$, $p < .001$, $d = 0.87$) ($M = 675.7 \pm 133$ ms and $M = 798.7 \pm 150.3$ ms, respectively).

Finally, we asked whether insights occurred with random delays, as reported earlier. To quantify this key characteristic, insight moments were defined as the time points of inflection of the fitted sigmoid function, i.e. when performance exhibited abrupt increases (see Methods). We verified the precision of our switch point identification by time-locking the data to the individually fitted switch points. This showed that accuracy steeply increased between the halved task block (50 trials) immediately before vs. after the switch, as expected ($M = 62\%$ vs $M = 83\%$ $t(89) = -11.2$, $p < .001$, $d = 2.34$, Fig.2C, Fig. S5A). Additionally, reaction times dropped steeply from pre- to post-switch ($M = 971.63$ ms vs. $M = 818.77$ ms, $t(87) = 3.34$, $p < .001$, $d = 0.7$). The average delay of insight onset was 1.3 task blocks (130 trials) (± 95 trials / 0.95 blocks, Fig.2G). The distribution of delays among insight participants ranged from 0 to 3 blocks after the start of the *motion and colour phase*, and statistically did not differ from a normal distribution taking into account the hazard rate (Exact two-sided Kolmogorov-Smirnov test: $D(48) = 0.15$, $p = 0.69$).

Hence, the behaviour of human subjects showed all characteristics of insight: sudden improvements in performance that occurred only in a subgroup and with variable delays.

Neural Network Behaviour

To probe whether insight-like behaviour can arise in simple neural networks trained with gradient descent, we simulated 99 network models performing the same decision making task. The networks had two input nodes (x_c, x_m, for colour and motion, respectively), two input-specific gates (g_m, g_c) and weights (w_m, w_c), and one output node ($\hat{y}$, Fig.1B). Network weights and gates were initialised at 0.01. The stimulus features motion and colour were reduced to one input node each, which encoded colour/motion direction of each trial by taking on either a positive or a negative value. More precisely, given the correct decision $y = \pm$, the activities of the input nodes were sampled from i.i.d. normal distributions with means yM_m and yM_c and standard deviations $\sigma_m = 0.01$ and $\sigma_c = 0.01$ for colour and motion respectively. Hence M_m and M_c determine the signal to noise ratio in each input. We fixed the colour mean shift $M_c = 0.22$, while the mean shifts of the motion node differed by noise level and were fitted individually such that each human participant had one matched network with comparable pre-insight task accuracy in each motion noise condition (see below).

The network multiplied each input node by two parameters, a corresponding weight, and a gate, and returned a decision based on the output node's sign $\hat{y}$:

$$\hat{y} = \text{sign}(g_m w_m x_m + g_c w_c x_c + \eta) \tag{1}$$

where $\eta \sim \mathcal{N}(0, \sigma)$ is Gaussian noise, and weights and gates are the parameters learned online through gradient descent.

To train L1-networks we used a simple squared loss function with L1-regularisation of gate weights:

$$\mathcal{L} = \frac{1}{2}(g_m w_m x_m + g_c w_c x_c + \eta - y)^2 + \lambda(|g_m| + |g_c|) \tag{2}$$

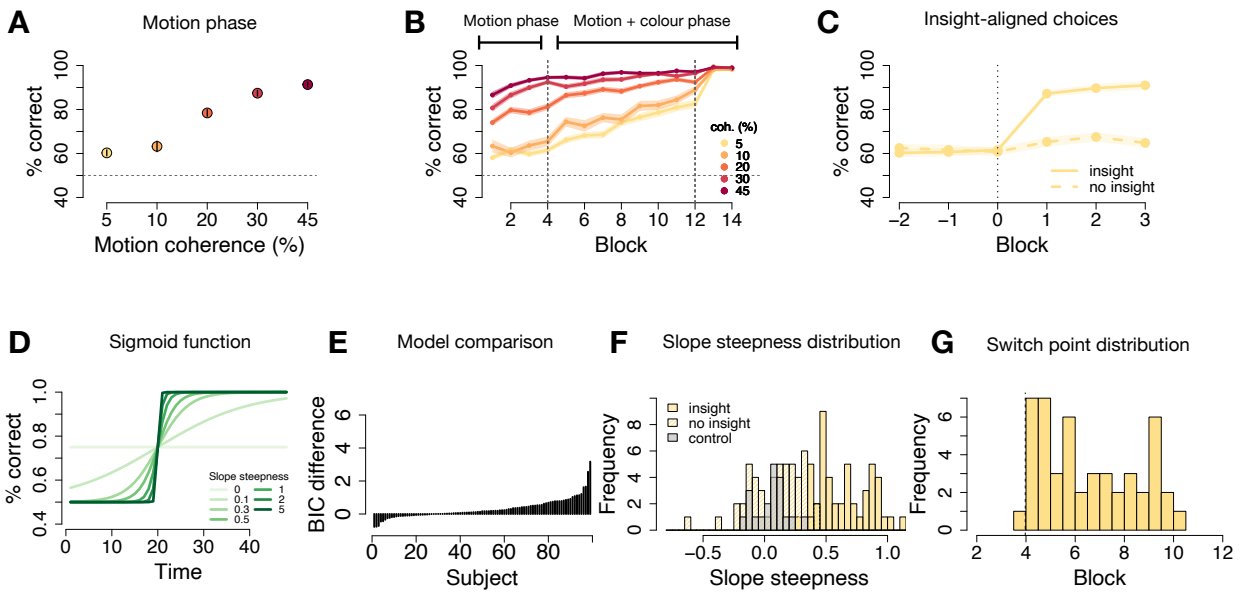


Figure 2: Humans: task performance and insight-like strategy switches **(A)** Accuracy (% correct) during the *motion phase* increases with increasing motion coherence. N = 99, error bars signify standard error of the mean (SEM). **(B)** Accuracy (% correct) over the course of the experiment for all motion coherence levels. First dashed vertical line marks the onset of the colour predictiveness (*motion and colour phase*), second dashed vertical line the "instruction" about colour predictiveness. Blocks shown are halved task blocks (50 trials each). N = 99, error shadows signify SEM. **(C)** Switch point-aligned accuracy on lowest motion coherence level for insight (48/99) and no-insight (51/99) subjects. Blocks shown are halved task blocks (50 trials each). Error shadow signifies SEM. **(D)** Illustration of the sigmoid function for different slope steepness parameters. **(E)** Difference between BICs of the linear and sigmoid function for each human subject. N = 99. **(F)** Distributions of fitted slope steepness at inflection point parameter for control experiment and classified insight and no-insight groups. **(G)** Distribution of switch points. Dashed vertical line marks onset of colour predictiveness. Blocks shown are halved task blocks (50 trials each).

with a fixed level of regularisation $\lambda = 0.07$.

During training, Gaussian noise was added to each gradient update to mimic learning noise and induce variability between individual networks (same gradient noise level for all networks). $\xi \sim \mathcal{N}(\mu_\xi = 0, \sigma_\xi = 0.05)$ was added to each gradient update, yielding the following update equations for noisy SGD of the network's weights

$$\Delta w_m = -\alpha x_m g_m (x_m g_m w_m + x_c g_c w_c + \eta - y) + \xi_{w_m}, \tag{3}$$

and gates,

$$\begin{aligned} \Delta g_m = &-\alpha x_m w_m (x_m g_m w_m + x_c g_c w_c + \eta - y) \\ &-\alpha \lambda \mathrm{sign}(g_m) + \xi_{g_m} \end{aligned} \tag{4}$$

where we have not notated the dependence of all quantities on trial index t for clarity; and analogous equations hold for colour weights and gates with all noise factors ξ_{g_m} ξ_{w_m} etc, following the same distribution.

Using this setup, we studied whether L1-regularisation would lead the network to show key characteristics of insight-like behaviour. Specifically, we reasoned that L1-regularisation of the gate weights would introduce competitive dynamics between the input channels that can lead to non-linear learning dynamics. We focused on L1-regularisation because it forces gates of irrelevant inputs most strongly towards 0, compared to L2-regularisation, which is less aggressive in particular once gates are already very small. While the multiplicative nature of the weights and gates results in non-linear quadratic and cubic gradient dynamics, applying L1-regularisation will lead to a sustained suppression period before the fast transition (see Methods).

Networks received an extended pre-task training phase of 6 blocks, but then underwent a training curriculum precisely matched to the human task (2 blocks of 100 trials in the *motion phase* and 5 blocks in the *motion and colour phase*, see Fig. 1D). We adjusted direction specificity of motion inputs (i.e. difference in distribution means from which x_m was drawn for left vs right trials) separately for each participant and coherence condition, such that performance in the motion phase was equated between each pair of human and network (Fig.3A, see Methods). Moreover, the colour and

motion input sequences used for network training were sampled from the same ten input sequences that humans were exposed to. A learning rate of $\alpha = 0.6$ (same for all participants) was selected to match average learning speed.

L1-regularised Neural Networks

Networks learned the motion direction-response mapping well in the training phase, during which colour inputs changed randomly and output should therefore depend only on motion inputs (*motion phase*, 75% correct, t-test against chance: $t(98) = 33.1$, $p < .001$, the accuracy of humans in this phase was $M = 76 \pm 6\%$). As in humans, adding noise to the motion inputs (*motion phase*) resulted in an accuracy gradient that depended on noise level (linear mixed effects model of accuracy: $\chi^2(1) = 165.61$, $p < .001$; N = 99, Fig.3A), as expected given that input distributions were set such that network performance would equate to human accuracy (Fig.3A-B). Networks also exhibited low and relatively stable performance levels in the two lowest coherence conditions (58% and 60%, paired t-test to assess stability in the *motion phase*: $t(98) = -0.7$, $p = 0.49$, $d = 0.02$), and had a large performance difference between high vs low coherence trials ($M = 88\% \pm 6\%$ vs. $M = 74 \pm 13\%$, $t(137.3) = 9.6$, $p < .001$, $d = 1.36$ for high, i.e. $\geq 20\%$ coherence, vs. low trials). Finally, humans and networks also performed comparably well at the end of learning (last block of the *colour and motion phase*: $M(nets) = 79\% \pm 17\%$ vs. $M(humans) = 82 \pm 17\%$, $t(195.8) = 1.1$, $p = 0.27$, $d = 0.16$, Fig. S8C), suggesting that at least some networks did start to use colour inputs. Hence, networks' baseline performance and learning were successfully matched to humans.

To look for characteristics of insight in network performance, we employed the same approach used for modelling human behaviour, and investigated suddenness, selectivity, and delay. To identify sudden performance improvements, we fitted each network's time course of accuracy on low coherence trials by (1) a linear model and (2) a non-linear sigmoid function, which would indicate gradual performance increases or insight-like behaviour, respectively. As in humans, network performance on low coherence trials was best fit by a non-linear sigmoid function, indicating at least a subsection of putative "insight networks" (BIC sigmoid function: $M = -10$, $SD = 1.9$, protected exceedance probability: 1, BIC linear function: $M = -9$, $SD = 2.4$, protected exceedance probability: 0)(Fig.3D).

We then tested whether insight-like behaviour occurred only in a subset of networks (selectivity) by assessing in how many networks the steepness of the performance increase exceeded a chance level defined by a baseline distribution of the steepness. As in humans, we ran simulations of 99 control networks with the same architecture, which were trained on the same task except that during the *motion and colour phase*, the two inputs remained uncorrelated. About half of networks (48/99, 48.5%) had steepness values larger than the 100% percentile of the control distribution, matching exactly the value we observed in the human sample. The L1-networks that showed sudden performance improvements were not matched to insight humans more often than chance ($\chi^2(47) = 27.9$, $p = 0.99$), suggesting that network variability did not originate from baseline performance levels or trial orders. Hence, a random subset of networks showed sudden performance improvements comparable to those observed during insight moments in humans (Fig.3E).

For simplicity reasons in comparing network behaviour to humans, we will refer to the two groups as "insight and no-insight networks". Analysing behaviour separately for the insight and no-insight networks showed that switches to the colour strategy improved the networks' performance on the lowest coherence trials once the *motion and colour phase* started, as compared to networks that did not show a strategy shift ($M = 83 \pm 11\%$, vs. $M = 64 \pm 9\%$, respectively, $t(89.8) = 9.2$, $p < .001$, $d = 1.9$, see Fig.3C). The same performance difference between insight and no-insight networks applied when all coherence levels of the *motion and colour phase* were included ($M = 88 \pm 7\%$ vs. $M = 77 \pm 6\%$, $t(93.4) = 7.8$, $p < .001$, $d = 1.57$). Unexpectedly, insight networks performed slightly worse on low coherence trials in the motion phase, i.e. before the change in predictiveness of the features, ($t(97) = -3.1$, $p = 0.003$, $d = 0.62$) (insight networks: $M = 58 \pm 8\%$; no-insight networks: $M = 64 \pm 9\%$), and in contrast to the lack of pre-insight differences we found in humans.

Finally we asked whether insight-like behaviour occurred with random delays in neural networks, again scrutinising the time points of inflection of the fitted sigmoid function, i.e. when performance exhibited abrupt increases (see Methods). Time-locking the data to these individually fitted switch points verified that, as in humans, the insight-like performance increase was particularly evident around the switch points: accuracy was significantly increased between the halved task blocks preceding and following the insight-like behavioural switch, for colour switching networks ($M = 66 \pm 8\%$ vs. $M = 86 \pm 7\%$, $t(91.6) = -12.7$, $p < .001$, $d = 2.6$, see Fig.3C, Fig. S5B).

Among insight networks, the delay distribution ranged from 1 to 4 blocks after the start of the *motion and colour phase*, and did not differ from a normal distribution taking into account the hazard rate (Exact two-sided Kolmogorov-Smirnov test: $D(48) = 0.13$, $p = 0.85$). The average delay of insight-like switches was 1.75 task blocks (± 1.05), corresponding to 175 trials (Fig.3F). The insight networks' delay was thus slightly longer than for humans ($M = 130 \pm 95$ trials vs. $M = 175 \pm 105$ trials, $t(92.7) = -2.1$, $p = 0.04$, $d = 0.42$). The variance of insight induced strategy switch onsets as well as the relative variance in the abruptness of the switch onsets thus qualitatively matched our behavioural results observed in human participants. The behaviour of L1-regularised neural networks therefore showed all characteristics

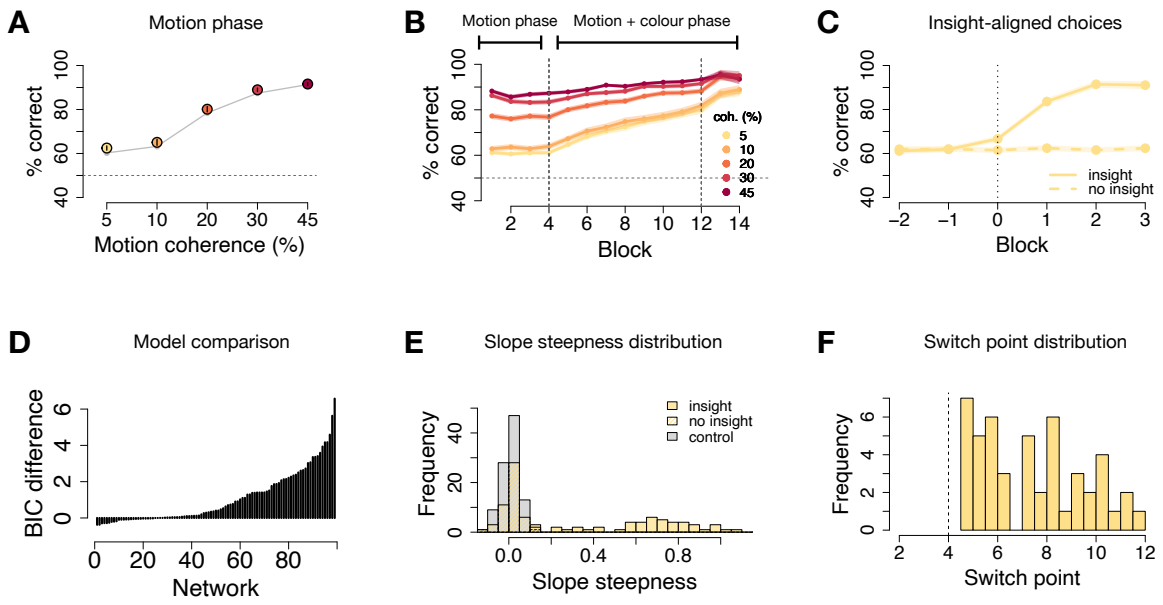


Figure 3: L1-regularised neural networks: task performance and insight-like strategy switches **(A)** Accuracy (% correct) during the *motion phase* increases with increasing motion coherence. N = 99, error bars signify SEM. Grey line is human data for comparison. **(B)** Accuracy (% correct) over the course of the experiment for all motion coherence levels. First dashed vertical line marks the onset of the colour predictiveness (*motion and colour phase*), second dashed vertical line the "instruction" about colour predictiveness. Blocks shown are halved task blocks (50 trials each). N = 99, error shadows signify SEM. **(C)** Switch point-aligned accuracy on lowest motion coherence level for insight (48/99) and no-insight (51/99) networks. Blocks shown are halved task blocks (50 trials each). Error shadow signifies SEM. **(D)** Difference between BICs of the linear model and sigmoid function for each network. **(E)** Distributions of fitted slope steepness at inflection point parameter for control networks and classified insight and no-insight groups. **(F)** Distribution of switch points. Dashed vertical line marks onset of colour predictiveness. Blocks shown are halved task blocks (50 trials each).

of human insight: sudden improvements in performance that occurred selectively only in a subgroup with variable random delays.

L2-regularised Neural Networks

Following our observation that L1-regularised networks exhibited human-like insight behaviour, we investigated whether this was specific to the form of regularisation. We therefore trained otherwise identical networks with a L2-regularisation term on the gate weights. We hypothesised that L2-regularisation would also lead to competitiveness between input nodes, but to a lower extent than L1-regularisation. We reasoned that in particular the fact that during the *motion phase* the networks motion weights would not shrink as close to 0 would lead to more frequent and earlier insight-like behavioural switches.

While L2-regularised gate weights led to switches that were similar to those previously observed in their abruptness (Fig. S7C), such insight-like behaviours were indeed much more frequent and clustered: 96% of networks switched to a colour strategy, with a switch point distribution that was much more centred around the onset of the colour predictiveness (Fig. S7F, average delay of 1 task block ($SD = 1.1$) corresponding to 100 trials after onset of the colour correlation (*motion and colour phase*). This was significantly shorter than for L1-regularised networks ($M = 1.05 \pm 1.1$ vs. $M = 1.75 \pm 1.05$, $t(59.6) = 4$, $p < 0.001$, $d = 0.9$) and also differed from a normal distribution taking into account the hazard rate (Exact two-sided Kolmogorov-Smirnov test: $D(95) = 0.26$, $p = 0.005$). Additionally, performance on the lowest coherence level in the last block of the *colour and motion phase* before colour instruction was centred just below ceiling and thus did not indicate a range of colour use like humans and L1-regularised networks ($M(L2 - networks) = 97\% \pm 2\%$ vs. $M(humans) = 82 \pm 17\%$, $t(101.6) = -8.8$, $p < .001$, $d = 1.25$, Fig. S8C).

While L2-regularised networks thus showed abrupt behavioural transitions, they failed to show the other two key characteristics of insight: selectivity and delay.

Non-regularised Neural Networks

In non-regularised networks, the effects observed in L2-regularised networks are enhanced. 99% of the networks started using colour inputs (Fig. S8A), but colour use occurred in a more linear, less abrupt way than for L1- or L2-regularised networks. Additionally, there was very little delay of only 0.7 task blocks (70 trials, (± 0.25)) between onset of the *motion and colour phase* and the start of the networks making use of the colour input predictiveness (Fig. S8B). As for L2-networks, this delay was significantly shorter than for L1-regularised networks ($M = 0.7 \pm 0.55$ vs. $M = 1.75 \pm 1.05$, $t(49.3) = 6.6$, $p < 0.001$, $d = 1.6$) and also differed from a normal distribution taking into account the hazard rate (Exact two-sided Kolmogorov-Smirnov test: $D(98) = 0.35$, $p < .001$). Similarly, performance on the lowest coherence level in the last block indicated that all networks used colour inputs ($M = 100\% \pm 0.3\%$ vs. $M = 82 \pm 17\%$, $t(98) = -10.4$, $p < .001$, $d = 1.5$, Fig. S8C). Thus non-regularised networks also did not show the insight key behavioural characteristics of selectivity and delay.

Origins of Insight-like Behaviour in Neural Networks

Having established the behavioural similarity between L1-networks and humans in an insight task, we asked what gave rise to insight-like switches in some networks, but not others. We therefore investigated the dynamics of gate weights and the effects of noise in insight vs. no-insight networks, and the role of regularisation strength parameter λ.

Colour Gradients Increase after Colour Becomes Predictive

Our first question was how learning about stimulus colour differed between insight and no-insight L1 networks, as expressed by the dynamics of network gradients. We time-locked the time courses of gradients to each network's individual switch point. Right when the switch occurred (at t of the estimated switch), colour gate weight gradients were significantly larger in insight compared to no-insight L1-networks ($M = 0.06 \pm 0.06$ vs. $M = 0.02 \pm 0.03$, $t(73.2) = 5.1$, $p < .001$, $d = 1.05$), while this was not true for motion gate weight gradients ($M = 0.18 \pm 0.16$ vs. $M = 0.16 \pm 0.16$, $t(97) = 0.7$, $p = 0.5$, $d = 0.13$).

Notably, insight networks had larger colour gate weight gradients even before any behavioural changes were apparent, right at the beginning of the *motion and colour phase* (first 5 trials of *motion and colour phase*: $M = 0.05 \pm 0.07$ vs. $M = 0.01 \pm 0.01$; $t(320) = 8.7$, $p < .001$), whereas motion gradients did not differ ($t(576.5) = -0.1$, $p = 0.95$). This increase in colour gate weight gradients for insight networks happened within a few trials after correlation onset (colour gradient last trial of *motion phase*: $M = 0 \pm 0$ vs. 5th trial of *motion and colour phase*: $M = 0.06 \pm 0.08$; $t(47) = -5.6$, $p < .001$, $d = 1.13$), and suggests that insight networks start early to silently learn more about colour inputs compared to their no-insight counterparts. A change point analysis considering the mean and variance of the gradients confirmed the onset of the *motion and colour phase* to be the change point of the colour gradient mean, with a difference of 0.04 between the consecutive pre-change and change time points for insight networks vs 0.005 for no-insight networks (with a change point detected two trials later), indicating considerable learning about colour for insight networks.

"Silent" Colour Knowledge Precedes Insight-like Behaviour

A core feature of our network architecture is that inputs were multiplied by two factors, a gate g, and a weight w, but only gates were regularised. This meant that some networks might have developed larger colour weights, but still showed no signs of colour use, because the gates were very small. This could explain the early differences in gradients reported above. To test this idea, we investigated the absolute size of colour gates and weights of insight vs no-insight L1-networks before and after insight-like switches had occurred.

Comparing gates at the start of learning (first trial of the *motion and colour phase*), there were no differences between insight and no-insight networks for either motion or colour gates (colour gates: $M = 0 \pm 0.01$ vs. $M = 0 \pm 0.01$; $t(95.3) = 0.8$, $p = 0.44$, motion gates: $M = 0.5 \pm 0.3$ vs. $M = 0.6 \pm 0.3$; $t(93.1) = -1.7$, $p = 0.09$, see Fig.4A, Fig.4H,J). Around the individually fitted switch points, however, the gates of insight and no-insight networks differed only for colour gates (colour gates: 0.2 ± 0.2 vs 0.01 ± 0.02 for insight vs no-insight networks, $t(48) = 6.7$, $p < 0.001$, $d = 1.4$, motion gates: 0.5 ± 0.3 vs 0.5 ± 0.3 for insight vs no-insight networks, $t(95.6) = 0.2$, $p = 0.9$, $d = 0.04$). Insight networks' increased use of colour inputs was particularly evident at the end of learning (last trial of the *motion and colour phase*) and reflected in larger colour gates (0.7 ± 0.3 vs 0.07 ± 0.2 for insight vs no-insight networks, $t(73.7) = 13.4$, $p < 0.001$, $d = 2.7$) while the reverse was true for motion gates ($M = 0.2 \pm 0.2$ vs $M = 0.5 \pm 0.3$, respectively, $t(81) = -7.5$, $p < 0.001$, $d = 1.5$, see Fig.4B, Fig.4H,J). Hence, differences in gating between network subgroups were only present after, but not before learning, and did not explain the above reported gradient differences or which network would show insight-like behaviour.

A different pattern emerged when investigating the weights of the networks. Among insight networks colour weights were significantly larger already at the start of learning (first trial of the *motion and colour phase*), as compared to no-insight networks (insight: $M = 1.2 \pm 0.6$; no-insight: $M = 0.4 \pm 0.3$, $t(66.2) = 8.1$, $p < .001$, $d = 1.7$, see Fig.4C, Fig.4G,I). This was not true for motion weights (insight: $M = 3.4 \pm 0.7$; no-insight: $M = 3.5 \pm 0.5$, $t(89.5) = -1.1$, $p = 0.3$, $d = 0.2$, see Fig.4C, Fig.4G,I). Thus, colour information appeared to be encoded in the weights of insight networks already before any insight-like switches occurred. Because the colour gates were suppressed through the L1-regularisation mechanism before learning, the networks did not differ in any observable colour sensitivity. An increase of colour gates reported above could then unlock the "silent knowlegde" of colour relevance.

To experimentally test the effect of pre-learning colour weights, we ran a new sample of L1-networks ($N = 99$), and adjusted the colour and motion weight of each respective network to the mean absolute colour and motion weight size we observed in insight networks at start of learning (first trial of *motion and colour phase*). Gates were left untouched. This increased the number of insight networks from 48.5% to 70.7%, confirming that encoding of colour information at an early stage was an important factor for later switches, but also not sufficient to cause insight-like behaviour in all networks. Note that before weights adjustments were made, the performance of the new networks did not differ from the original L1-networks ($M = 0.8 \pm 0.07$ vs $M = 0.8 \pm 0.07$, $t(195) = 0.2$, $p = 0.9$, $d = 0.03$). In our new sample, networks that would later show insight-like behaviour or not also did not differ from each other (insight: $M = 0.7 \pm 0.07$ vs $M = 0.7 \pm 0.07$, $t(100.9) = 1.4$, $p = 0.2$, $d = 0.3$, no-insight: $M = 0.8 \pm 0.05$ vs $M = 0.8 \pm 0.07$, $t(71) = 0.9$, $p = 0.4$, $d = 0.2$). Weight and gate differences between L1- and L2-networks are reported in the Supplementary Material (see also Fig.4E-F).

Noise is Needed For Insight-like Behaviour

One possible factor that could explain the early differences between the weights of network subgroups is noise. The networks were exposed to noise at two levels: on each trial noise was added at the output stage ($\eta \sim \mathcal{N}(0,\, \sigma_\eta^2)$), and to the gate and weight gradients during updating ($\xi \sim \mathcal{N}(0,\, \sigma_\xi^2)$).

We probed whether varying the level of noise added during gradient updating, i.e. σ_ξ, would affect the proportion of networks exhibiting insight-like behaviour. Parametrically varying the variance of noise added to colour and motion gates and weights led to increases in insight-like behaviour, from no single insight network when no noise was added to 100% insight networks when σ_{ξ_g} reached values of larger than approx. 0.05 (Fig.5A). Since gate and weight updates were coupled (see Eq. 4-7), noise during one gradient update could in principle affect other updates as well. We therefore separately manipulated the noise added to updates of colour gates and weights, motion gates and weights, all weights and all gates. This showed that adding noise to only weights during the updates was sufficient to induce insight-like behaviour (Fig.5B). In principle, adding noise to only gates was sufficient for insight-like switches as well, although noise applied to the gates had to be relatively larger to achieve the same effect as applying noise to weight gradients (Fig.5B), presumably due the effect of regularisation. Adding noise only to the gradients of motion gates or weights, but not to the colour gradients, was not sufficient to induce insight-like switches (Fig.5B). On the other hand, noise added only to the colour parameter updates quickly led to substantial amounts of insight-like behavioural switches (Fig.5B).

An analysis of *cumulative* noise showed that the effects reported above are mostly about momentary noise fluctuations: cumulative noise added to the output did not differ between insight and no-insight networks at either the start (first trial of the *motion and colour phase*) or end of learning (last trial of the *motion and colour phase*) (start: $M = -0.3 \pm 4.7$ vs. $M = -0.6 \pm 3.9$; $t(91.2) = 0.4$, $p = 0.7$, end: $M = 0.6 \pm 7.1$ vs. $M = 0.5 \pm 7.1$; $t(96.7) = 0.07$, $p = 1$), and the same was true for cumulative noise added during the gradient updates to weights and gates (see Supplementary Material for details).

We therefore conclude that Gaussian noise added to updates of particularly colour gate weights, in combination with "silent knowledge" about colour information stored in suppressed weights, is a crucial factor for insight-like behavioural changes.

Regularisation Parameter λ Affects Insight Delay and Frequency

In our previous results, the regularisation parameter λ was arbitrarily set to 0.07. We next tested the effect of of λ on insight-like behaviour. The number of L1-regularised insight networks linearly decreased with increasing λ (Fig.5C). Lambda further had an effect on the delay of the insight-like switches, with smaller λ values leading to decreased average delays of switching to a colour strategy after predictiveness of the inputs had changed (Fig.5D). The regularisation parameter λ thus affects two of the key characteristics of human insight – selectivity and delay.

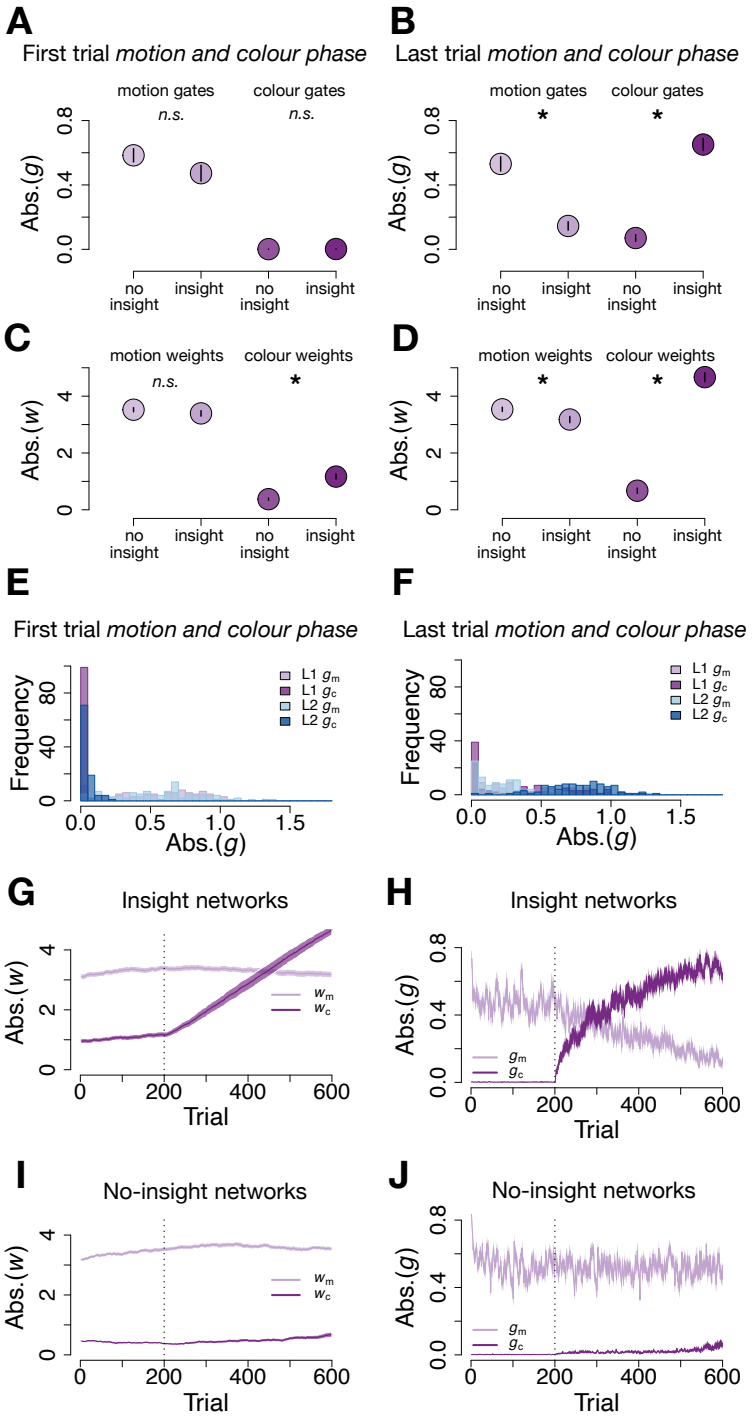


Figure 4: Gate and weight size differences at the start and end of learning and dynamics. Colour and motion gates at **(A)** the first trial and **(B)** the last trial of the *motion and colour phase*. **(C)** Colour and motion weights at the first trial and **(D)** the last trial of the *motion and colour phase*. Error bars signify SEM. **(E)** Gate weight sizes for colour and motion gate weights at the first trial and **(F)** the last trial of the *motion and colour phase* for L1- and L2-regularised networks. **(G)** Weights of insight L1-networks. The dashed vertical line marks the onset of the *motion and colour phase*. Error shadows signify SEM. **(H)** Gates of insight L1-networks. The dashed vertical line marks the onset of the *motion and colour phase*. Error shadows signify SEM. **(I)** Weights of no-insight L1-networks. The dashed vertical line marks the onset of the *motion and colour phase*. Error shadows signify SEM. **(J)** Gates of no-insight L1-networks. The dashed vertical line marks the onset of the *motion and colour phase*. Error shadows signify SEM.

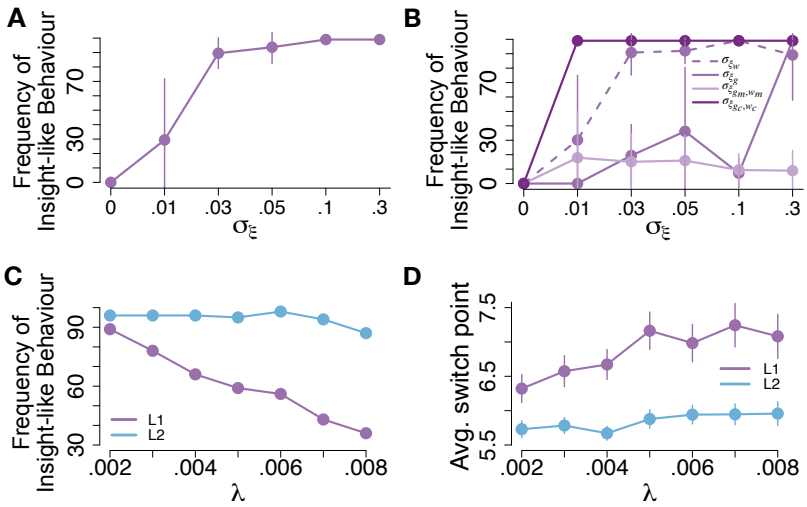


Figure 5: Influence of Gaussian noise distribution variance σ_ξ and regularisation parameter λ on insight-like switches in L1-regularised networks **(A)** Influence of noise standard deviation (σ_ξ) applied to all gradient updates on the frequency of switches to a colour strategy (number of networks defined as having "insight"). The frequency of insight-like switches increases gradually with σ_ξ until it plateaus. Error bars are SD. We ran 10 x 99 simulations. **(B)** Effects of noise added only to either all weights (σ_{ξ_w}), all gates (σ_{ξ_g}), all motion parameters (i.e. motion weight and motion gates, $\sigma_{\xi_{g_m,w_m}}$) and all colour parameters ($\sigma_{\xi_{g_c,w_c}}$) on the frequency of insight-like switches when it is only applied to the network *weights* and/or *gates*. The frequency of insight-like switches increases gradually with σ_{ξ_w} until it plateaus (dashed purple line), while it jumps abruptly after relatively high levels of σ_{ξ_g} (solid purple line). $\sigma_{\xi_{g_m,w_m}}$ on motion alone is not sufficient for insight-like switches (lightest purple shade), but small $\sigma_{\xi_{g_c,w_c}}$ is sufficient for the frequency of insight networks to plateau (darkest purple shade). Error bars are SD. We ran 10 x 99 simulations. Colour scheme as in Fig. 1B **(C)** Influence of λ on the frequency of switches to a colour strategy (number of networks defined as having "insight"). The frequency of insight-like switches declines with increasing λ for L1-regularised networks, but is largely unaffected for L2-regularised networks. **(D)** Influence of λ on the averaged switch points. The averaged switch point occurs later in the task with increasing λ for both L1 and L2-regularised networks. Error bars signify SEM.

Discussion

We investigated insight-like learning behaviour in humans and neural networks. In a binary decision making-task with a hidden regularity that entailed an alternative way to solve the task more efficiently, a subset of regularised neural networks with multiplicative gates of their input channels (as an attention mechanism) displayed spontaneous, jump-like learning that signified the sudden discovery of the hidden regularity – mysterious insight moments boiled down to the simplest expression.

Networks exhibited all key characteristics of human insight-like behaviour in the same task (suddenness, selectivity, delay). Crucially, neural networks were trained with standard stochastic gradient descent that is often associated with gradual learning. Our results therefore suggest that the behavioural characteristics of aha-moments can arise from gradual learning mechanisms, and hence suffice to mimic human insight.

Network analyses identified the factors which caused insight-like behaviour in L1-networks: noise added during the gradient computations accumulated to non-zero weights in some networks. As long as colour information was not useful yet, i.e. prior to the onset of the hidden regularity, close-to-0 colour gates rendered these weights "silent", such that no effects on behaviour can be observed. Once the hidden colour regularity became available, the non-zero colour weights helped to trigger non-linear learning dynamics that arise during gradient updating, and depend on the starting point. Hence, our results hint at important roles of "attentional" gating, noise, and regularisation as the computational origins of sudden, insight-like behavioural changes. We report several findings that are in line with this interpretation: addition of gradient noise ξ in particular to the colour weights and gates, pre-learning adjustment of colour weights and a reduction of the regularisation parameter λ all increased insight-like behaviour. We note that our networks did not have a hidden layer, witnessing the fact that no hidden layer is needed to produce non-linear learning dynamics.

Our findings have implications for the conception of insight phenomena in humans. While present-day machines clearly do not have the capacity to have aha-moments due to their lack of meta-cognitive awareness, our results show that the remarkable behavioural signatures of insights by themselves do not necessitate a dedicated process. This raises

the possibility that sudden behavioural changes which occur even during gradual learning could in turn lead to the subjective effects that accompany insights Frensch et al. [2003], Esser et al. [2022].

Our results also highlight noise and regularisation as aspects of brain function that are involved in the generation of insights. Cellular and synaptic noise is omnipresent in brain activity Faisal et al. [2008], Waschke et al. [2021], and has a number of known benefits, such as stochastic resonance and robustness that comes with probabilistic firing of neurons based on statistical fluctuations due to Poissonian neural spike timing Rolls et al. [2008]. It has also been noted that noise plays an important role in jumps between brain states, when noise provokes transitioning between attractor states Rolls and Deco [2012]. Previous studies have therefore noted that stochastic brain dynamics can be advantageous, allowing e.g. for creative problem solving (as in our case), exploratory behaviour, and accurate decision making Rolls and Deco [2012], Faisal et al. [2008], Garrett et al. [2013], Waschke et al. [2021]. Our work adds a computationally precise explanation of how noise can lead to insights to this literature. Questions about whether inter-individual differences in neural variability predict insights Garrett et al. [2013], or about whether noise that occurs during synaptic updating is crucial remain an interesting topic for future research.

Previous work has also suggested the occurrence and possible usefulness of regularisation in the brain. Regularisation has for instance been implied in synaptic scaling, which helps to adjust synaptic weights in order to maintain a global firing homeostasis Lee et al. [2019], thereby aiding energy requirements and reducing memory interference Tononi and Cirelli [2014], De Vivo et al. [2017]. It has also been proposed that regularisation modulates the threshold for induction of long-term potentation Lee et al. [2019]. These mechanisms therefore present possible synaptic factors that contribute to insight-like behaviour in humans and animals. We note that synaptic scaling has often been linked to sleep Tononi and Cirelli [2014], and regularisation during sleep has also been suggested to help avoid overfitting to experiences made during the day, and therefore generalisation Hoel [2021]. Since our experiments were conduced in an uninterrupted fashion during daylight, our findings could not reflect any sleep effects. The findings above nevertheless suggests a possible link between sleep, synaptic scaling and insight Wagner et al. [2004], Lacaux et al. [2021].

On a more cognitive level, regularisation has been implied in the context of heuristics. In this notion, regularisation has been proposed to function as an infinitely strong prior in a Bayesian inference framework Parpart et al. [2018]. This infinitely strong prior would work as a sort of attention mechanism and regularise input and information in a way that is congruent with the specific prior, whereas a finite prior would under this assumption enable learning from experience Parpart et al. [2018]. Another account regards cognitive control as regularised optimisation Ritz et al. [2022]. According to this theory, better transfer learning is supported by effort costs regularising towards more task-general policies. It therefore seems possible that the factors that impact regularisation during learning can also lead to a neural switch between states that might be more or less likely to govern insights.

The occurrence of insight-like behaviour with the same characteristics as found in humans was specific to L1-regularised networks, while no comparable similarity occurred in L2- or non-regularised networks. Although L2-regularised neural networks learned to suppress initially irrelevant colour feature inputs and showed abrupt performance increases reminiscent of insights, only L1 networks exhibited a wide distribution of time points when the insight-like switches occur (delay) as well as a selectivity of the phenomenon to a subgroup of networks, as found in humans. We note that L2- and non-regularised networks technically performed better on the task, because they collectively improve their behavioural efficiency sooner. One important question therefore remains under which circumstances L1 would be the most beneficial form of regularisation. One possibility could be that the task is too simple for L1-regularisation to be beneficial. It is conceivable that L1-regularisation only starts being advantageous in more complex task settings when generalisation across task sets is required and a segregation of task dimensions to learn about at a given time would prove useful.

Taken together, gradual training of neural networks with gate modulation leads to insight-like behaviour as observed in humans, and points to roles of regularisation, noise and "silent knowledge" in this process. These results make an important contribution to the general understanding of learning dynamics and representation formation in environments with non-stationary feature relevance in both biological and artificial agents.

Methods

Task

Stimuli

We employed a perceptual decision task that required a binary choice about circular arrays of moving dots Rajananda et al. [2018], similar to the spontaneous strategy switch task developed earlier Schuck et al. [2015]. Dots were characterised by two features, (1) a motion direction (four possible orthogonal directions: NW, NE, SW, SE) and (2) a

colour (orange or purple, Fig.1A). The noise level of the motion feature was varied in 5 steps (5%, 10%, 20%, 30% or 45% coherent motion), making motion judgement relatively harder or easier. Colour difficulty was constant, thus consistently allowing easy identification of the stimulus colour. The condition with most noise (5% coherence) occurred slightly more frequently than the other conditions (30 trial per 100, vs 10, 20, 20, 20 for the other conditions).

The task was coded in JavaScript and made use of the jsPsych 6.1.0 plugins. Participants were restricted to use desktops (no tablets or mobile phones) of at least 13 inch width diagonally. Subjects were further restricted to use either a Firefox or Google Chrome browser to run the experiment.

On every trial, participants were presented a cloud of 200 moving dots with a radius of 7 pixels each. In order to avoid tracking of individual dots, dots had a lifetime of 10 frames before they were replaced. Within the circle shape of 400 pixel width, a single dot moved 6 pixel lengths in a given frame. Each dot was either designated to be coherent or incoherent and remained so throughout all frames in the display, whereby each incoherent dot followed a randomly designated alternative direction of motion.

The trial duration was 2000 ms and a response could be made at any point during that time window. After a response had been made via one of the two button presses, the white fixation cross at the centre of the stimulus would turn into a binary feedback symbol (happy or sad smiley) that would be displayed until the end of the trial (Fig.1C). An inter trial interval (ITI) of either 400, 600, 800 or 1000 ms was randomly selected. If no response was made, a "TOO SLOW" feedback was displayed for 300 ms before being replaced by the fixation cross for the remaining time of the ITI.

Task Design

For the first 400 trials, the *motion phase*, the correct binary choice was only related to stimulus motion (two directions each on a diagonal were mapped onto one choice), while the colour changed randomly from trial to trial (Fig.1D). For the binary choice, participants were given two response keys, "X" and "M". The NW and SE motion directions corresponded to a left key press ("X"), while NE and SW corresponded to a right key press ("M") (Fig.1A). Participants received trial-wise binary feedback (correct or incorrect), and therefore could learn which choice they had to make in response to which motion direction (Fig.1C).

We did not specifically instruct participants to pay attention to the motion direction. Instead, we instructed them to learn how to classify the moving dot clouds using the two response keys, so that they would maximise their number of correct choices. To ensure that participants would pick up on the motion relevance and the correct stimulus-response mapping, motion coherence was set to be at 100% in the first block (100 trials), meaning that all dots moved towards one coherent direction. Participants learned this mapping well and performed close to ceiling (87% correct, t-test against chance: $t(98) = 37.4$, $p < .001$). In the second task block, we introduced the lowest, and therefore easiest, three levels of motion noise (20%, 30% and 45% coherent motion), before starting to use all five noise levels in block 3. Since choices during this phase should become solely dependent on motion, they should be affected by the level of motion noise. We assessed how well participants had learned to discriminate the motion direction after the fourth block. Participants that did not reach an accuracy level of at least 85% in the three lowest motion noise levels during this last task block of the pre-training were excluded from the *motion and colour phase*. All subjects were notified before starting the experiment, that they could only advance to the second task phase (*motion and colour phase*, although this was not communicated to participants) if they performed well enough in the first phase and that they would be paid accordingly for either one or two completed task phases.

After the *motion phase*, in the *motion and colour phase*, the colour feature became predictive of the correct choice in addition to the motion feature (Fig.1D). This meant that each response key, and thus motion direction diagonal, was consistently paired with one colour, and that colour was fully predictive of the required choice. Orange henceforth corresponded to a correct "X" key press and a NW/SE motion direction, while purple was predictive of a correct "M" key press and NE/SW motion direction (Fig.1A). This change in feature relevance was not announced to participants, and the task continued for another 400 trials as before - the only change being the predictiveness of colour.

Before the last task block we asked participants whether they 1) noticed a rule in the experiment, 2) how long it took until they noticed it, 3) whether they used the colour feature to make their choices and 4) to replicate the mapping between stimulus colour and motion directions. We then instructed them about the correct colour mapping and asked them to rely on colour for the last task block. This served as a proof that subjects were in principle able to do the task based on the colour feature and to show that, based on this easier task strategy, accuracy should be near ceiling for all participants in the last instructed block.

Human Participants

Participants between eighteen and 30 years of age were recruited online through Prolific.

Participation in the study was contingent on showing learning of the stimulus classification. Hence, to assess whether participants had learned to correctly identify motion directions of the moving dots, we probed their accuracy on the three easiest, least noisiest coherence levels in the last block of the uncorrelated task phase. If subjects reached an accuracy level of at least 85%, they were selected for participation in the experiment.

Ninety-six participants were excluded due to insufficient accuracy levels after the *motion phase* as described above. 99 participants learned to classify the dots' motion direction, passed the accuracy criterion and completed both task phases. These subjects make up the final sample included in all analyses. 34 participants were excluded due to various technical problems or premature quitting of the experiment. All participants gave informed consent prior to beginning the experiment. The study protocol was approved by the local ethics committee of the Max Planck Institute for Human Development. Participants received 3£ for completing only the first task phase and 7£ for completing both task phases.

Neural Networks

L1-regularised Neural Networks

We utilise a simple neural network model to reproduce the observations of the human behavioural data in a simplified supervised learning regression setting. We trained a simple neural network with two input nodes, two input gates and one output node on the same decision making task (Fig.1B).

The network received two inputs, x_m and x_c, corresponding to the stimulus motion and colour, respectively, and had one output, $\hat{y}$. Importantly, each input had one associated multiplicative gate (g_m, g_c) such that output activation was defined as $\hat{y} = \text{sign}(g_m w_m x_m + g_c w_c x_c + \eta)$ where $\eta \sim \mathcal{N}(0, \sigma)$ is Gaussian noise (Fig.1B).

To introduce competitive dynamics between the input channels, we added L1-regularisation on the gate weights g, resulting in the following loss function:

$$\mathcal{L} = \frac{1}{2}(g_m w_m x_m + g_c w_c x_c + \eta - y)^2 + \lambda(|g_m| + |g_c|) \tag{5}$$

The network was trained in a gradual fashion through online gradient descent with Gaussian white noise ξ added to the gradient update and a fixed learning rate α. Given the loss function, this yields the following update equations for noisy stochastic gradient descent (SGD):

$$\Delta w_m = -\alpha x_m g_m (x_m g_m w_m + x_c g_c w_c + \eta - y) + \xi_{w_m} \tag{6}$$

$$\Delta g_m = -\alpha x_m w_m (x_m g_m w_m + x_c g_c w_c + \eta - y) - \alpha\lambda \text{sign}(\text{g}_\text{m}) + \xi_{\text{g}_\text{m}} \tag{7}$$

$$\Delta w_c = -\alpha x_c g_c (x_c g_c w_c + x_m g_m w_m + \eta - y) + \xi_{w_c} \tag{8}$$

$$\Delta g_c = -\alpha x_c w_c (x_c g_c w_c + x_m g_m w_m + \eta - y) - \alpha\lambda \text{sign}(\text{g}_\text{c}) + \xi_{\text{g}_\text{c}} \tag{9}$$

with λ = 0.07, α = 0.6 and ξ = 0.05.

This implies that the evolution of the colour weights and gates will exhibit non-linear quadratic and cubic dynamics, driven by the interaction of w_c and g_c. Multiplying the weights w with the regularised gate weights g leads to smaller weights and therefore initially slower increases of the colour weights w_c and respective gate weights g_c after colour has become predictive of correct choices.

To understand this effect of non-linearity analytically, we used a simplified setup of the same model without gate weights:

$$\mathcal{L} = [w_m x_m + w_c x_c + \eta - y]^2 \tag{10}$$

Using this model, we observe exponential increases of the colour weights w_c after the onset of the *motion and colour phase*. This confirms that the interaction of w_c and g_c, as well as the regularisation applied to g_c are necessary for the insight-like non-linear dynamics including a distribution of insight onsets as well as variety in slope steepness of insight-like switches.

Note that because the regularisation term is non-differentiable at 0, we cannot take the limit $\alpha \to 0$, but averaged over the data instead. To avoid oscillations of the coefficients around 0 due to the non-differentiability, we added the following rules after each update of the gates: (1) if the gate g^t was zero before the update, a regularisation term $-\min(\alpha\lambda, |\text{g}^{\text{t+1}}|)\text{sign}(\text{g}^{\text{t+1}})$ was added and (2) if the gate changed sign during the update, the value was set to 0.

The accuracy is given by:

$$\begin{aligned}&\mathbb{P}[\hat{y} = y|w_m, g_m, w_c, g_c] \\ &= \frac{1}{2}[1 + \mathrm{erf}(\frac{\mathrm{g_m w_m x_m} + \mathrm{g_c w_c x_c}}{\sqrt{2((\mathrm{g_m w_m}\sigma_\mathrm{m})^2 + (\mathrm{g_c w_c}\sigma_\mathrm{c})^2 + \sigma^2)}})]\end{aligned} \tag{11}$$

We trained the network on a curriculum precisely matched to the human task, and adjusted hyperparameters (noise levels), such that baseline network performance and learning speed were carefully equated between humans and networks.

Specifically, we simulated the same number of networks than humans were included in the final analysis sample ($N = 99$). We matched the motion noise based performance variance of a given simulation to a respective human subject using a non-linear COBYLA optimiser. While the mean of the colour input distribution (0.22) as well as the standard deviations of both input distributions were fixed (0.01 for colour and 0.1 for motion), the respective motion input distribution mean values were individually fitted for each single simulation as described above.

The input sequences the networks received were sampled from the same ten input sequences that humans were exposed to in task phase two. This means that for the task part where colour was predictive of the correct binary choice, *motion and colour phase* (500 trials in total), networks and humans received the same input sequences.

The networks were given a slightly longer *training phase* of six blocks (600 trials) in comparison to the two blocks *training phase* that human subjects were exposed to (Fig.1D). Furthermore, human participants first completed a block with 100% motion coherence before doing one block with low motion noise. The networks received six *training phase* blocks containing the three highest motion coherence levels. Both human subjects and networks completed two blocks including all noise levels in the *motion phase* before colour became predictive in the *motion and colour phase*.

L2-regularised Neural Networks

To probe the effect of the aggressiveness of the regulariser on insight-like switch behaviour in networks, we compared our L1-regularised networks with models of the same architecture, but added L2-regularisation on the gate weights g. This yielded the following loss function:

$$\mathcal{L} = \frac{1}{2}(g_m w_m x_m + g_c w_c x_c + \eta - y)^2 + \frac{\lambda}{2}(|g_m| + |g_c|)^2 \tag{12}$$

From the loss function we can again derive the following update equations for noisy stochastic gradient descent (SGD):

$$\Delta w_m = -\alpha x_m g_m (x_m g_m w_m + x_c g_c w_c + \eta - y) + \xi_{w_m} \tag{13}$$

$$\begin{aligned}\Delta g_m = &-\alpha x_m w_m (x_m g_m w_m + x_c g_c w_c + \eta - y) \\ &- \alpha\lambda \mathrm{sign}(\mathrm{g_m})\mathrm{abs}(\mathrm{g_m}) + \xi_{\mathrm{g_m}}\end{aligned} \tag{14}$$

$$\Delta w_c = -\alpha x_c g_c (x_c g_c w_c + x_m g_m w_m + \eta - y) + \xi_{w_c} \tag{15}$$

$$\begin{aligned}\Delta g_c = &-\alpha x_c w_c (x_c g_c w_c + x_m g_m w_m + \eta - y) \\ &- \alpha\lambda \mathrm{sign}(\mathrm{g_c})(\mathrm{g_m})\mathrm{abs}(\mathrm{g_c}) + \xi_{\mathrm{g_c}}\end{aligned} \tag{16}$$

with λ = 0.07, α = 0.6 and ξ = 0.05.

The training is otherwise the same as for the L1-regularised networks.

Modelling of insight-like switches

Models of colour use

In order to probe whether strategy switches in low coherence trials occurred abruptly, we compared three different models with different assumptions about the form of the data. First, we fitted a linear model with two free parameters:

$$y = mt + y_0$$

where m is the slope, y_0 the y-intercept and t is time (here, task blocks)(Fig. S2). This model should fit no-insight participants' data well where colour use either increases linearly over the course of the experiment or stays at a constant level.

Contrasting the assumptions of the linear model, we next tested whether colour-based responses increased abruptly by fitting a step model with three free parameters, a switch point t_s, the step size s and a maximum value y_{max} (Fig. S2), so that

$$y = \begin{cases} y_{max} - s & \text{if } t < t_s \\ y_{max} & \text{if } t \geq t_s \end{cases}$$

We also included a sigmoid function with three free parameters as a smoother approximation of the step model:

$$y = y_{max} - y_{min}\frac{1}{+}e^{-m(t-t_s)} + y_{min}$$

where y_{max} is the fitted maximum value of the function, m is the slope and t_s is the inflection point (Fig. S2). y_{min} was given by each individual's averaged accuracy on 5% motion coherence trials in block 3-4.

Comparing the model fits across all subjects using the Bayesian Information Criterion (BIC) and protected exceedance probabilities yielded a preference for the sigmoid function over both a step and linear model, for both humans (Fig.2E) and L1-regularised neural networks (Fig.3D). On the one hand, this supports our hypothesis that insight-like strategy switches do not occur in an incremental linear fashion, but abruptly, with variance in the steepness of the switch. Secondly, this implies that at least a subset of subjects shows evidence for an insight-like strategy switch.

Human participants

To investigate these insight-like strategy adaptations, we modelled human participants' data using the individually fitted sigmoid functions (Fig. S3). The criterion we defined in order to assess whether a subject had switched to the colour strategy, was the slope at the inflection point, expressing how steep the performance jump was after having an insight about colour. We obtained this value by taking the sigmoid function's partial derivative of time

$$\frac{\partial y}{\partial t} = (y_{max} - y_{min})\frac{me^{-m(t-t_s)}}{(1+e^{-m(t-t_s)})^2}$$

and then evaluating the above equation for the fitted switch point, $t = t_s$, which yields:

$$y'(t_s) = \frac{1}{4}m(y_{max} - y_{min})$$

Switch misclassifications can happen that are caused by irregularities and small jumps in the data - irrespective of a colour strategy switch. We therefore corrected for a general fit of the data to the model by subtracting the individually assessed general model fit from the slope steepness at the inflection point. Insight subjects were then classified as those participants whose corrected slope steepness at inflection point parameters were outside of the 100% percentile of a control group's (no change in predictiveness of colour) distribution of that same parameter. By definition, insights about a colour rule cannot occur in this control condition, hence our derived out-of-sample distribution evidences abrupt strategy improvements hinting at insight (Fig.3F).

Before the last task block we asked participants whether they used the colour feature to make their choices. 57.6% of participants indicated that they used colour to press correctly. The 48.5% insight participants we identified using our classification method overlapped to 79.2% with participants' self reports.

Neural Networks

We used the same classification procedure for neural networks. All individual sigmoid function fits for L1-regularised networks can be found in the Supplementary Material (Fig. S4).

Acknowledgements

ATL is supported by the International Max Planck Research School on Computational Methods in Psychiatry and Ageing Research (IMPRS COMPPPSYCH, www.mps.ucl-centre.mpg.de). PMK was funded by the Wellcome Trust (award: 210849/Z/18/Z). AMS was supported by a Sir Henry Dale Fellowship from the Wellcome Trust and Royal Society (216386/Z/19/Z), and the Sainsbury Wellcome Centre Core Grant from Wellcome (219627/Z/19/Z) and the Gatsby Charitable Foundation (GAT3755). AMS is a CIFAR Azrieli Global Scholar in the Learning in Machines & Brains program. CS was funded by the European Research Council (ERC Consolidator awards 725937) and Special Grant Agreement No. 945539 (Human Brain Project SGA). NWS was funded by the Federal Government of Germany and the State of Hamburg as part of the Excellence Initiative, a Starting Grant from the European Union

(ERC-StG-REPLAY-852669), and an Independent Max Planck Research Group grant awarded by the Max Planck Society (M.TN.A.BILD0004). The funding parties had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

We thank Robert Gaschler for helpful comments on this manuscript.

References

Wolfgang Köhler. *The Mentality of Apes*. Kegan Paul, Trench, Trubner & Co. ; Harcourt, Brace & Co., 1925.

Daniel Durstewitz, Nicole M. Vittoz, Stan B. Floresco, and Jeremy K. Seamans. Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. *Neuron*, 66(3):438–448, 2010. ISSN 08966273. doi:10.1016/j.neuron.2010.03.029. URL http://dx.doi.org/10.1016/j.neuron.2010.03.029.

Hans Stuyck, Bart Aben, Axel Cleeremans, and Eva Van den Bussche. The Aha! moment: Is insight a different form of problem solving? *Consciousness and Cognition*, 90(April 2020):103055, 2021. ISSN 10902376. doi:10.1016/j.concog.2020.103055. URL https://doi.org/10.1016/j.concog.2020.103055.

Robert W. Weisberg. Toward an integrated theory of insight in problem solving. *Thinking and Reasoning*, 21(1):5–39, 2015. ISSN 14640708. doi:10.1080/13546783.2014.886625. URL http://dx.doi.org/10.1080/13546783.2014.886625.

John Kounios and Mark Beeman. The cognitive neuroscience of insight. *Annual Review of Psychology*, 65:71–93, 2014. ISSN 15452085. doi:10.1146/annurev-psych-010213-115154.

Mark Jung-Beeman, Edward M. Bowden, Jason Haberman, Jennifer L. Frymiare, Stella Arambel-Liu, Richard Greenblatt, Paul J. Reber, and John Kounios. Neural activity when people solve verbal problems with insight. *PLoS Biology*, 2(4):500–510, 2004. ISSN 15449173. doi:10.1371/journal.pbio.0020097.

Amory H. Danek, Thomas Fraps, Albrecht von Müller, Benedikt Grothe, and Michael Öllinger. It's a kind of magic-what self-reports can reveal about the phenomenology of insight problem solving. *Frontiers in Psychology*, 5(DEC):1–11, 2014. ISSN 16641078. doi:10.3389/fpsyg.2014.01408.

John Kounios and Mark Beeman. *The eureka factor: Aha moments, creative insight, and the brain.* Random House, New York, 2015. ISBN 9781400068548.

Wangbing Shen, Yu Tong, Feng Li, Yuan Yuan, Bernhard Hommel, Chang Liu, and Jing Luo. Tracking the neurodynamics of insight: A meta-analysis of neuroimaging studies. *Biological Psychology*, 138(January):189–198, 2018. ISSN 18736246. doi:10.1016/j.biopsycho.2018.08.018. URL https://doi.org/10.1016/j.biopsycho.2018.08.018.

Martin Tik, Ronald Sladky, Caroline Di Bernardi Luft, David Willinger, André Hoffmann, Michael J. Banissy, Joydeep Bhattacharya, and Christian Windischberger. Ultra-high-field fMRI insights on insight: Neural correlates of the Aha!-moment. *Human Brain Mapping*, 39(8):3241–3252, 2018. ISSN 10970193. doi:10.1002/hbm.24073.

Nicolas W. Schuck, Robert Gaschler, Dorit Wenke, Jakob Heinzle, Peter A. Frensch, John Dylan Haynes, and Carlo Reverberi. Medial prefrontal cortex predicts internally driven strategy shifts. *Neuron*, 86(1):331–340, 2015. ISSN 10974199. doi:10.1016/j.neuron.2015.03.015. URL http://dx.doi.org/10.1016/j.neuron.2015.03.015.

Nicolas W. Schuck, Amy X. Li, Dorit Wenke, Destina S. Ay-Bryson, Anika T. Loewe, Robert Gaschler, and Yee Lee Shing. Spontaneous discovery of novel task solutions in children. *PloS ONE*, 17(5):e0266253, 2022. doi:10.1371/journal.pone.0266253. URL http://dx.doi.org/10.1371/journal.pone.0266253.

Robert Gaschler, Nicolas W. Schuck, Carlo Reverberi, Peter A. Frensch, and Dorit Wenke. Incidental covariation learning leading to strategy change. *PLoS ONE*, 14(1):1–32, 2019. ISSN 19326203. doi:10.1371/journal.pone.0210597.

Robert Gaschler, Bianca Vaterrodt, Peter A. Frensch, Alexandra Eichler, and Hilde Haider. Spontaneous Usage of Different Shortcuts Based on the Commutativity Principle. *PLoS ONE*, 8(9):1–13, 2013. ISSN 19326203. doi:10.1371/journal.pone.0074972.

Robert Gaschler, Julian N. Marewski, and Peter A. Frensch. Once and for all—How people change strategy to ignore irrelevant information in visual tasks. *Quarterly Journal of Experimental Psychology*, 68(3):543–567, 2015. ISSN 17470226. doi:10.1080/17470218.2014.961933. URL http://dx.doi.org/10.1080/17470218.2014.961933.

Edward M. Bowden, Mark Jung-Beeman, Jessica Fleck, and John Kounios. New approaches to demystifying insight. *Trends in Cognitive Sciences*, 9(7):322–328, 2005. ISSN 13646613. doi:10.1016/j.tics.2005.05.012.

Janet Metcalfe and David Wiebe. Intuition in insight and noninsight problem solving. *Memory & Cognition*, 15(3):238–246, 1987. ISSN 0090502X. doi:10.3758/BF03197722.

Mattias P. Karlsson, Dougal G.R. Tervo, and Alla Y. Karpova. Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. *Science*, 338(6103):135–139, 2012. ISSN 10959203. doi:10.1126/science.1226518.

Paul Miller and Donald B. Katz. Stochastic transitions between neural states in taste processing and decision-making. *Journal of Neuroscience*, 30(7):2559–2570, 2010. ISSN 02706474. doi:10.1523/JNEUROSCI.3047-09.2010.

Michele Allegra, Shima Seyed-Allaei, Nicolas W. Schuck, Daniele Amati, Alessandro Laio, and Carlo Reverberi. Brain network dynamics during spontaneous strategy shifts and incremental task optimization. *NeuroImage*, 217(January): 116854, 2020. ISSN 10959572. doi:10.1016/j.neuroimage.2020.116854. URL `https://doi.org/10.1016/j.neuroimage.2020.116854`.

Karl J Friston, Marco Lin, Christopher D Frith, Giovanni Pezzulo, J. Allan Hobson, and Sasha Ondobaka. Active Inference , Curiosity and Insight. *Neural Computation*, 29:2633–2683, 2017. doi:10.1162/neco.

Stellan Ohlsson. Information-processing explanations of insight and related phenomena. In *Advances in the Psychology of Thinking*. Harvester Wheatseaf, 1992.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets. *arXiv*, pages 1–10, 2022. URL `http://arxiv.org/abs/2201.02177`.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *Proceedings of the International Conference on Learning Representations 2014.*, pages 1–22, 2014.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 166(23): 11537–11546, 2019a. ISSN 10916490. doi:10.1073/pnas.1820226116.

Anna C. Schapiro and James L. McClelland. A connectionist model of a continuous developmental transition in the balance scale task. *Cognition*, 110(3):395–411, 2009. ISSN 00100277. doi:10.1016/j.cognition.2008.11.017.

James L. McClelland and Timothy T. Rogers. The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, 4(4):310–322, 2003. ISSN 14710048. doi:10.1038/nrn1076.

Timo Flesch, Keno Juechems, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, 110(7): 1258–1270, 2022. ISSN 10974199. doi:10.1016/j.neuron.2022.01.005. URL `https://doi.org/10.1016/j.neuron.2022.01.005`.

Andrew M. Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D. Tracey, and David D. Cox. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12), 2019b. ISSN 17425468. doi:10.1088/1742-5468/ab3985.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum Learning. In *In: Proceedings of International Conference on Machine Learning*, pages 41–48, 2009. URL `http://arxiv.org/abs/1611.06204`.

Timo Flesch, Jan Balaguer, Ronald Dekker, Hamed Nili, and Christopher Summerfield. Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences of the United States of America*, 115(44):E10313–E10322, 2018. ISSN 10916490. doi:10.1073/pnas.1800755115.

Johannes Mehrer, Courtney J Spoerer, Nikolaus Kriegeskorte, and Tim C Kietzmann. Individual differences among deep neural network models. *Nature Communications*, 11(5725):1–12, 2020. ISSN 2041-1723. doi:10.1038/s41467-020-19632-w. URL `http://dx.doi.org/10.1038/s41467-020-19632-w`.

Shengchao Liu, Dimitris Papailiopoulos, and Dimitris Achlioptas. Bad global minima exist and SGD can reach them. *Advances in Neural Information Processing Systems*, 2020-Decem(NeurIPS), 2020. ISSN 10495258.

Chrisopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer US, 2006. ISBN 9780387310732. doi:10.1007/978-3-030-57077-4_11.

Kamesh Krishnamurthy, Tankut Can, and David J. Schwab. Theory of Gating in Recurrent Neural Networks. *Physical Review X*, 12(1):11011, 2022. ISSN 21603308. doi:10.1103/PhysRevX.12.011011. URL `https://doi.org/10.1103/PhysRevX.12.011011`.

Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. An empirical exploration of Recurrent Network architectures. *32nd International Conference on Machine Learning, ICML 2015*, 3:2332–2340, 2015.

Lukas N. Groschner, Jonatan G. Malis, Birte Zuidinga, and Alexander Borst. A biophysical account of multiplication by a single neuron. *Nature*, 603(7899):119–123, 2022. ISSN 14764687. doi:10.1038/s41586-022-04428-3.

Tomaso Poggio, Vincent Torre, and Christof Koch. Computational vision and regularization theory. *Nature*, 317(6035): 314–319, 1985. ISSN 00280836. doi:10.1038/317314a0.

Rui Ponte Costa, Yannis M. Assael, Brendan Shillingford, Nando De Freitas, and Tim P. Vogels. Cortical microcircuits as gated-recurrent neural networks. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips 2017): 272–283, 2017. ISSN 10495258.

Sivananda Rajananda, Hakwan Lau, and Brian Odegaard. A random-dot kinematogram for web-based vision research. *Journal of Open Research Software*, 6(1), 2018. ISSN 20499647. doi:10.5334/jors.194.

P. A. Frensch, H. Haider, D. Rünger, U. Neugebauer, S. Voigt, and J. Werg. The route from implicit learning to verbal expression of what has been learned: Verbal report of incidentally experienced environmental regularity. In L. Jimenez, editor, *Attention and implicit learning*, pages 335–366. John Benjamins Publishing Company, 2003.

Sarah Esser, Clarissa Lustig, and Hilde Haider. What triggers explicit awareness in implicit sequence learning? Implications from theories of consciousness. *Psychological Research*, 86(5):1442–1457, 2022. ISSN 14302772. doi:10.1007/s00426-021-01594-3. URL https://doi.org/10.1007/s00426-021-01594-3.

A. Aldo Faisal, Luc P.J. Selen, and Daniel M. Wolpert. Noise in the nervous system. *Nature Reviews Neuroscience*, 9 (4):292–303, 2008. ISSN 1471003X. doi:10.1038/nrn2258.

Leonhard Waschke, Niels A. Kloosterman, Jonas Obleser, and Douglas D. Garrett. Behavior needs neural variability. *Neuron*, 109(5):751–766, 2021. ISSN 10974199. doi:10.1016/j.neuron.2021.01.023. URL https://doi.org/10.1016/j.neuron.2021.01.023.

Edmund T. Rolls, James M. Tromans, and Simon M. Stringer. Spatial scene representations formed by self-organizing learning in a hippocampal extension of the ventral visual system. *European Journal of Neuroscience*, 28(10): 2116–2127, 2008. ISSN 0953816X. doi:10.1111/j.1460-9568.2008.06486.x.

Edmund T. Rolls and Gustavo Deco. *The Noisy Brain: Stochastic dynamics as a principle of brain function*. Oxford University Press, 2012. ISBN 9780191702471. doi:10.1093/acprof:oso/9780199587865.001.0001.

Douglas D Garrett, Gregory R Samanez-larkin, Stuart W S Macdonald, Ulman Lindenberger, Anthony R Mcintosh, and Cheryl L Grady. Neuroscience and Biobehavioral Reviews Moment-to-moment brain signal variability : A next frontier in human brain mapping ? *Neuroscience and Biobehavioral Reviews*, 37(4):610–624, 2013. ISSN 0149-7634. doi:10.1016/j.neubiorev.2013.02.015. URL http://dx.doi.org/10.1016/j.neubiorev.2013.02.015.

Hey-kyoung Lee, Alfredo Kirkwood, and Hey-kyoung Lee. Mechanisms of Homeostatic Synaptic Plasticity in vivo. *PNAS*, 13(December):1–7, 2019. doi:10.3389/fncel.2019.00520.

Giulio Tononi and Chiara Cirelli. Sleep and the Price of Plasticity: From Synaptic and Cellular Homeostasis to Memory Consolidation and Integration. *Neuron*, 81(1):12–34, 2014. ISSN 08966273. doi:10.1016/j.neuron.2013.12.025. URL http://dx.doi.org/10.1016/j.neuron.2013.12.025.

Luisa De Vivo, Michele Bellesi, William Marshall, Eric A. Bushong, Mark H. Ellisman, Giulio Tononi, and Chiara Cirelli. Ultrastructural evidence for synaptic scaling across the wake/sleep cycle. *Science*, 355(6324):507–510, 2017. ISSN 10959203. doi:10.1126/science.aah5982.

Erik Hoel. The overfitted brain : Dreams evolved to assist generalization. *Patterns*, 2(5):100244, 2021. ISSN 2666-3899. doi:10.1016/j.patter.2021.100244. URL https://doi.org/10.1016/j.patter.2021.100244.

Ullrich Wagner, Steffen Gais, Hilde Haider, Rolf Verleger, and Jan Born. Sleep inspires insight. *Nature*, 427(6972): 352–355, 2004. ISSN 00280836. doi:10.1038/nature02223.

Célia Lacaux, Thomas Andrillon, Céleste Bastoul, Yannis Idir, Alexandrine Fonteix-galet, Isabelle Arnulf, and Delphine Oudiette. Sleep onset is a creative sweet spot. *Science Advances*, 5866(December):1–10, 2021.

Paula Parpart, Matt Jones, and Bradley C Love. Heuristics as Bayesian inference under extreme priors. *Cognitive Psychology*, 102(March):127–144, 2018. ISSN 0010-0285. doi:10.1016/j.cogpsych.2017.11.006. URL https://doi.org/10.1016/j.cogpsych.2017.11.006.

Harrison Ritz, Xiamin Leng, and Amitai Shenhav. Cognitive control as a multivariate optimization problem. *Journal of Cognitive Neuroscience*, 34(4):569–591, 2022.

Hidden layer model

In order to verify that our results were not merely an artefact of the oversimplified models we used, we tested the task on a more complex neural network model that had one additional hidden layer of fully connected linear units.

The linear neural network received two inputs, x_m and x_c, corresponding to the stimulus motion and colour, respectively, and had two output nodes, $\hat{y}$, as well as one hidden layer of 48 units. Importantly, each weight connecting the inputs with a hidden unit had one associated multiplicative gate g. To introduce competitive dynamics between the input channels, we again applied L1-regularisation on the gate weights g.

The network was trained on the Cross Entropy loss using stochastic gradient descent with $\lambda = 0.002$ and $\alpha = 0.1$.

As for the one-layer network, we trained this network on a curriculum precisely matched to the human task, and adjusted hyperparameters (noise levels), such that baseline network performance and learning speed were carefully equated between humans and networks (see Methods).

We employed the same analysis approach to detect insight-like behaviour (see Methods for details) by running simulations of a "control" network of the same architecture, but without correlated features and therefore without colour predictiveness in the *motion and colour phase*. We found that when we applied L1-regularisation with a regularisation parameter of $\lambda = 0.002$ on the gate weights, 18.2% of the networks exhibited *abrupt* and *delayed* learning dynamics, resembling insight-like behaviour in humans (Fig.1A) and thereby replicating the key insight characteristics suddenness and selectivity. Insight-like switches to the colour strategy thereby again improved the networks' performance significantly. Using the same parameters, experimental setup and analyses, but applying L2-regularisation on the gate weights g, yielded an insight-like switch rate of 51.5% (Fig.1B).

We again also observed a wider distribution of delays, the time point when the switches in the *motion and colour phase* occurred in insight networks, for L1-regularised networks with a hidden layer (Fig.1C-D).

Taken together, these results mirror our observations from network simulations with a simplified setup. We can thereby confirm that our results of L1-regularised neural networks' behaviour exhibiting all key characteristics of human insight behaviour (suddenness, selectivity and delay) are not an artefact of the one-layer linearity.

Weight and Gate Differences between L1- and L2-regularised Networks

At correlation onset (first trial of *motion and colour phase*), neither motion nor colour weights differed (motion: $M = 3.5 \pm 0.6$ vs $M = 3.4 \pm 0.5$, $t(192.7) = 1.2$, $p = 0.2$, $d = 0.2$, colour: $M = 0.8 \pm 0.6$ vs $M = 0.8 \pm 0.5$, $t(189.2) = 0.4$, $p = 0.7$, $d = 0.1$). After learning, however, i.e. at the last trial of the *motion and colour phase*, the average absolute size of the colour weights was higher in L2- compared to L1-networks ($M = 2.6 \pm 2.2$ vs $M = 4.7 \pm 0.7$, $t(115.1) = -9$, $p < .001$, $d = 1.3$), while the reverse was true for motion weights ($M = 3.4 \pm 0.7$ vs $M = 2.8 \pm 0.6$, $t(194.9) = 5.6$, $p < .001$, $d = 0.8$). For gate weights, differences between L1- and L2-networks are already apparent at correlation onset (first trial of *motion and colour phase*), where the mean of the motion gate was 0.53 for L1-networks and 0.58 for L2-networks, and hence lower in L1 networks,

albeit not significantly ($t(195.1) = -1$, $p = 0.3$, $d = 0.1$, see Fig. 4E). In addition, the average absolute size of the colour gate weights was higher in L2- compared to L1-networks ($M = 0.04 \pm 0.05$ vs $M = 0.002 \pm 0.006$, respectively, $t(100.6) = -7.2$, $p < 0.001$, $d = 1$). The respective distributions also reflected these effects. L1-networks had a much more narrow distribution for colour gates and just slightly narrower distribution for motion gates (L1: colour gates: 0 to 0.04, motion gates: 0 to 1.3, L2: colour gates: 0 to 0.2, motion gates: 0 to 1.4) After learning, i.e. at the last trial of the *motion and colour phase*, the mean colour gate size still was lower in L1- compared to L2-regularised networks ($M = 0.4 \pm 0.4$ vs $M = 0.8 \pm 0.2$, $t(169.1) = -9.3$, $p < 0.001$, $d = 1.3$), while the reverse was true for motion gates ($M = 0.3 \pm 0.3$ vs $M = 0.2 \pm 0.2$, $t(152.4) = 3.9$, $p < 0.001$, $d = 0.6$, see Fig. 4F). This was again also reflected in the respective distributions with L1-networks having much wider distributions for motion and slightly shorter width for colour gates (L1: colour gates: 0 to 1.2, motion gates: 0 to 1.3, L2: colour gates: 0 to 1.3, motion gates: 0 to 0.7).

Gaussian Noise Differences at Weights and Gates between Insight and No-Insight Networks

Comparing Gaussian noise $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$ at the weights and gates around the individually fitted switch points revealed no differences between insight and no-insight networks for either motion or colour weights (colour weights: $M = -0.08 \pm 1$ vs. $M = 0.04 \pm 0.8$; $t(89.5) = -0.6$, $p = 0.5$, motion weights: $M = 0.5 \pm 0.3$ vs. $M = 0.6 \pm 0.3$; $t(93.1) = -1.7$, $p = 0.09$) or gates (colour gates: $M = -0.1 \pm 0.9$ vs. $M = 0.1 \pm 0.9$; $t(95.3) = 0.8$, $p = 0.44$, motion gates: $M = 0.2 \pm 0.6$ vs. $M = -0.3 \pm 0.8$; $t(94.4) = 2$, $p = 0.05$). There also were no σ_ξ differences at either the start of learning (first trial of the *motion and colour phase*) (colour weights: $M = -0.06 \pm 0.8$ vs. $M = -0.03 \pm 0.5$; $t(78.1) = -0.2$, $p = 0.8$, motion weights: $M = 0.08 \pm 0.7$ vs. $M = 0.07 \pm 0.7$; $t(96.7) = 1$, $p = 0.3$, colour gates: $M = 0 \pm 0.6$ vs. $M = -0.2 \pm 0.7$; $t(97) = 1.6$, $p = 0.1$, motion gates: $M = -0.04 \pm 0.6$ vs. $M = -0.07 \pm 0.7$; $t(97) = 0.2$, $p = 0.8$) or end of learning (last trial of the *motion and colour phase*)(colour weights: $M = 0.05 \pm 1.3$ vs. $M = 0.08 \pm 1.1$; $t(92.7) = -0.1$, $p = 0.9$, motion weights: $M = 0 \pm 1.2$ vs. $M = -0.02 \pm 1.1$; $t(95.6) = 0.04$, $p = 1$, colour gates: $M = 0.2 \pm 1.1$ vs. $M = -0.2 \pm 1.2$; $t(97) = 1.7$, $p = 0.09$, motion gates: $M = -0.1 \pm 1.3$ vs. $M = 0.05 \pm 1.3$; $t(96) = -0.7$, $p = 0.5$).

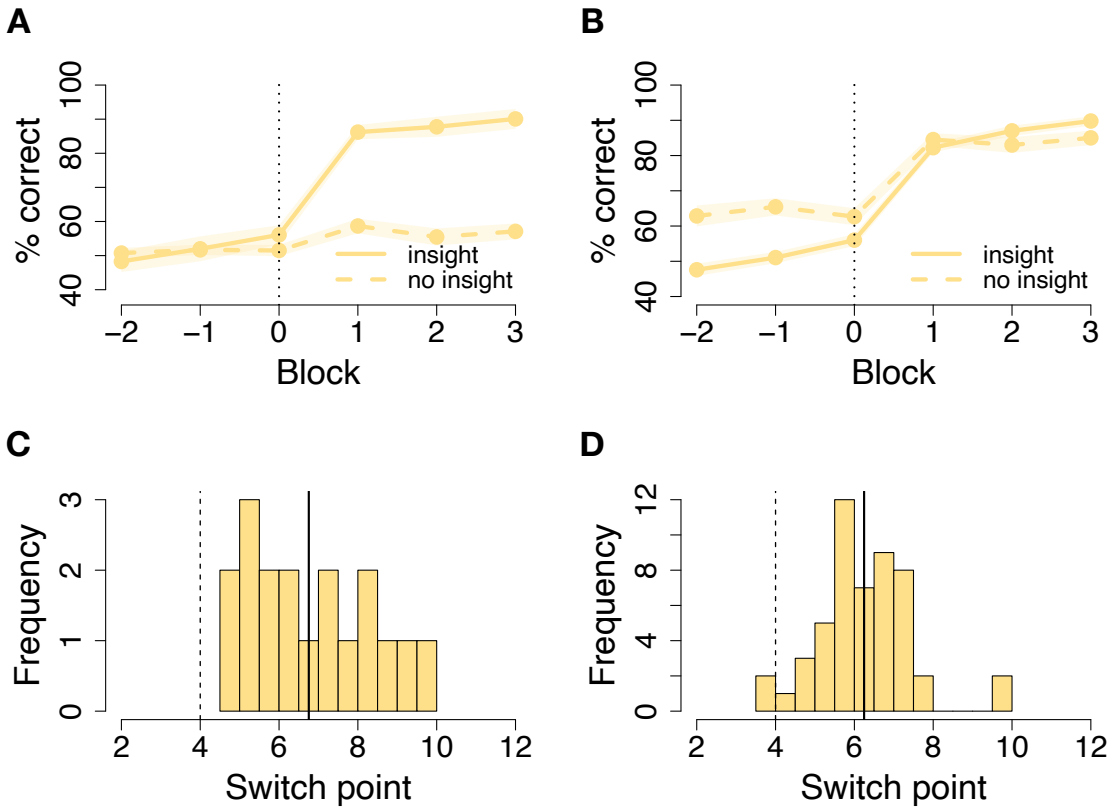


Fig. 1: Switch-aligned performance and switch point distributions for L1- and L2-regularised neural networks with a 48 unit hidden layer each. Blocks shown are halved task blocks (50 trials each). Error shadows signify SEM. **(A)** Switch-aligned performance for insight (18/99) and no-insight groups (81/99) respectively for L1-regularised networks with a hidden layer. **(B)** Switch-aligned performance for insight (51/99) and no-insight (48/99) L2-regularised neural networks with a hidden layer. **(C)** Switch point distributions for L1-regularised insight networks with a hidden layer. Dashed vertical line marks onset of colour predictiveness. **(D)** Switch point distributions for L2-regularised insight neural networks. Dashed vertical line marks onset of colour predictiveness.

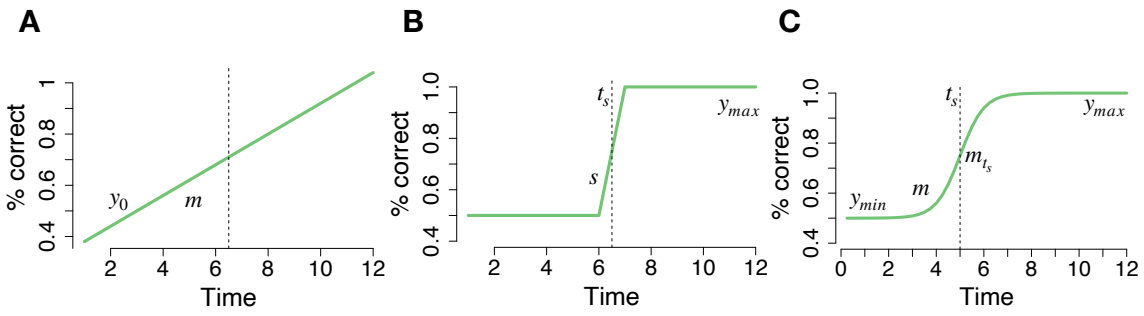
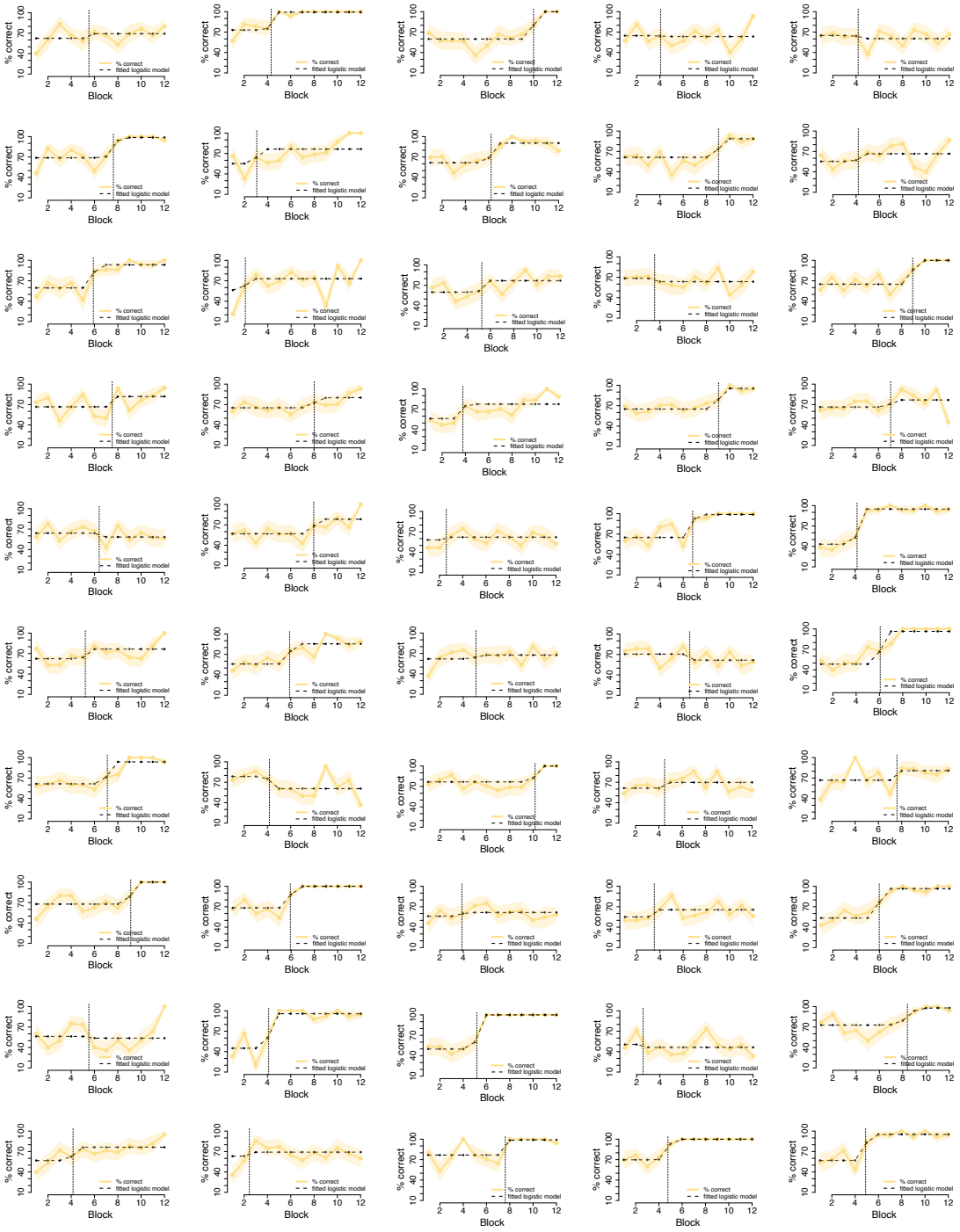


Fig. 2: Illustrations of models and respective parameters. **(A)** Linear function with free parameters intercept y_0 and slope m. **(B)** Step function with free parameters inflection point t_s and function maximum y_{max}. **(B)** Generalised logistic regression function with free parameters slope m, inflection point t_s and function maximum y_{max}.


% correct
Block
% correct
fitted logistic model

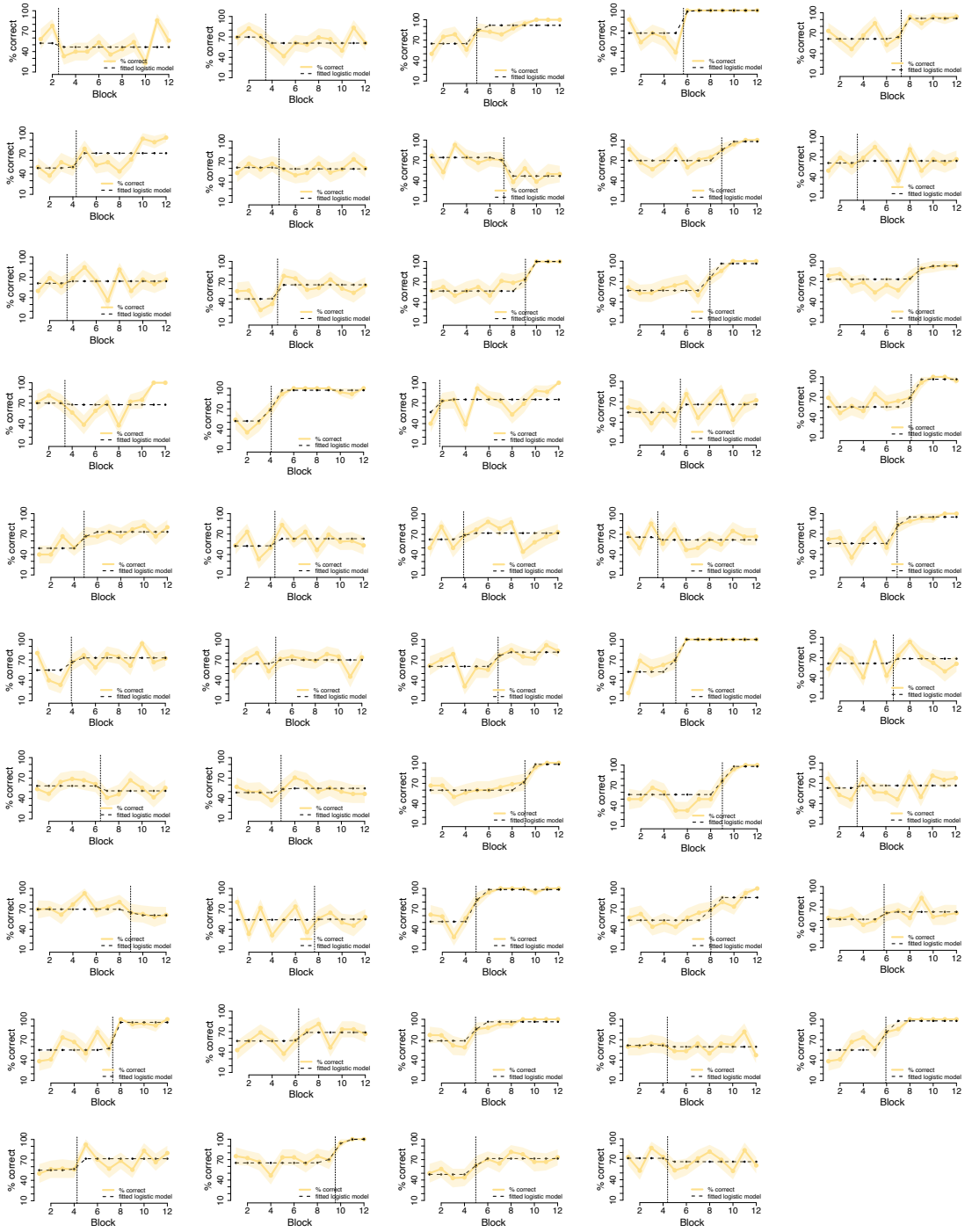
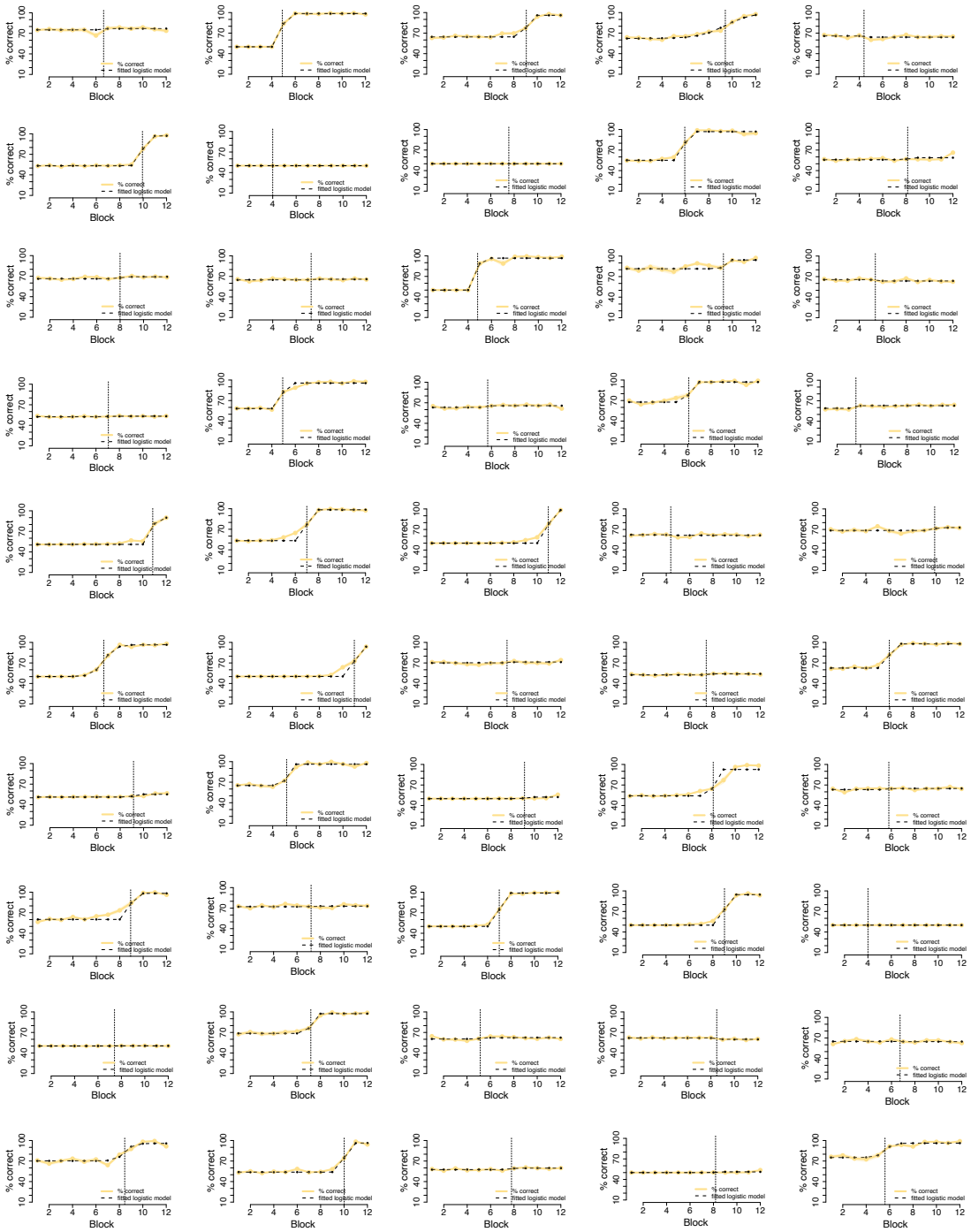


Fig. 3: Performance on highest motion noise trials (yellow) and model predictions (black) for every human participant. Blocks shown are halved task blocks (50 trials each). Error shadows signify SEM.


% correct
Block
% correct
fitted logistic model

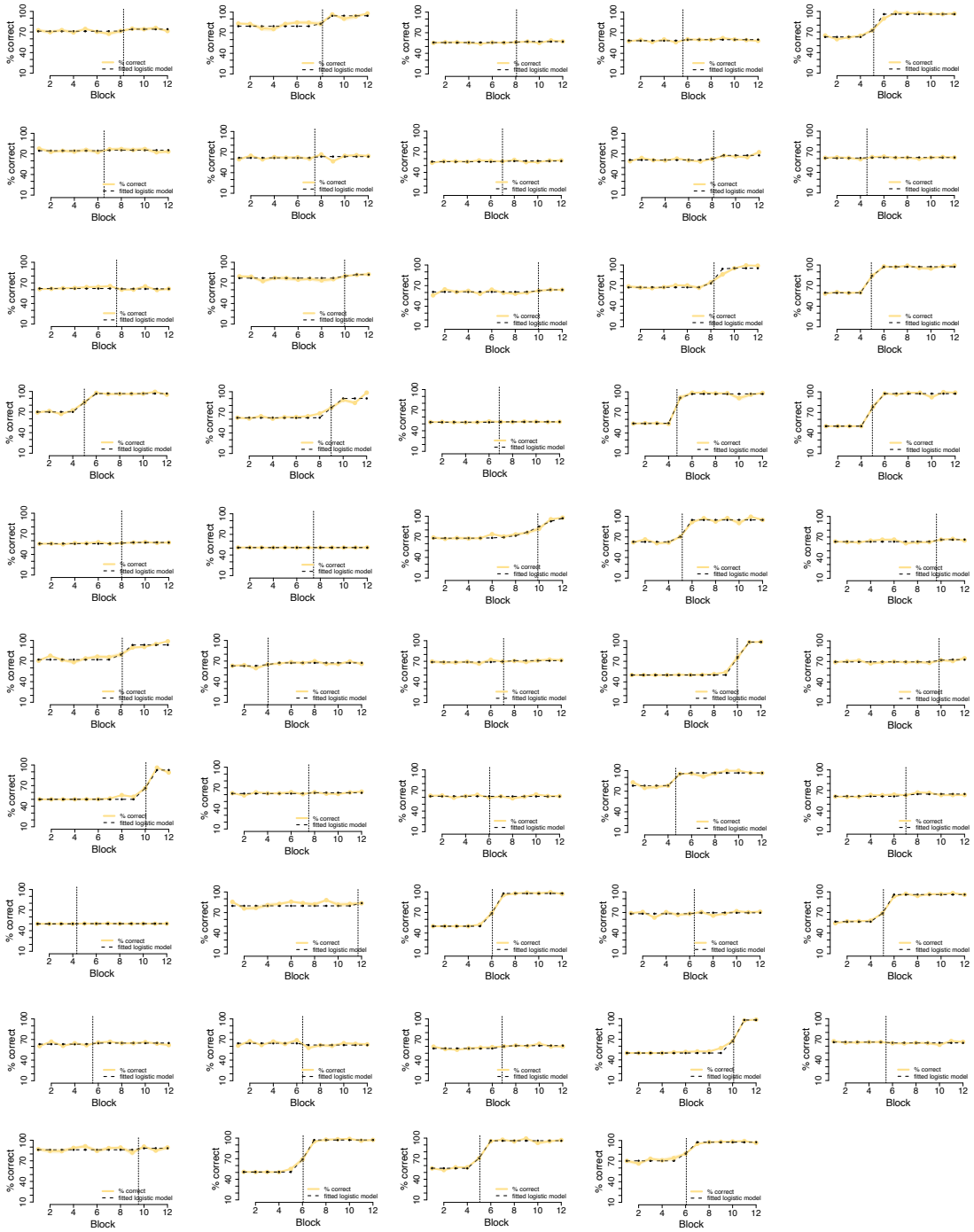


Fig. 4: Performance on Performance on highest motion noise trials (yellow) and model predictions (black) for every L1-regularised neural network. Blocks shown are halved task blocks (50 trials each). Error shadows signify SEM.

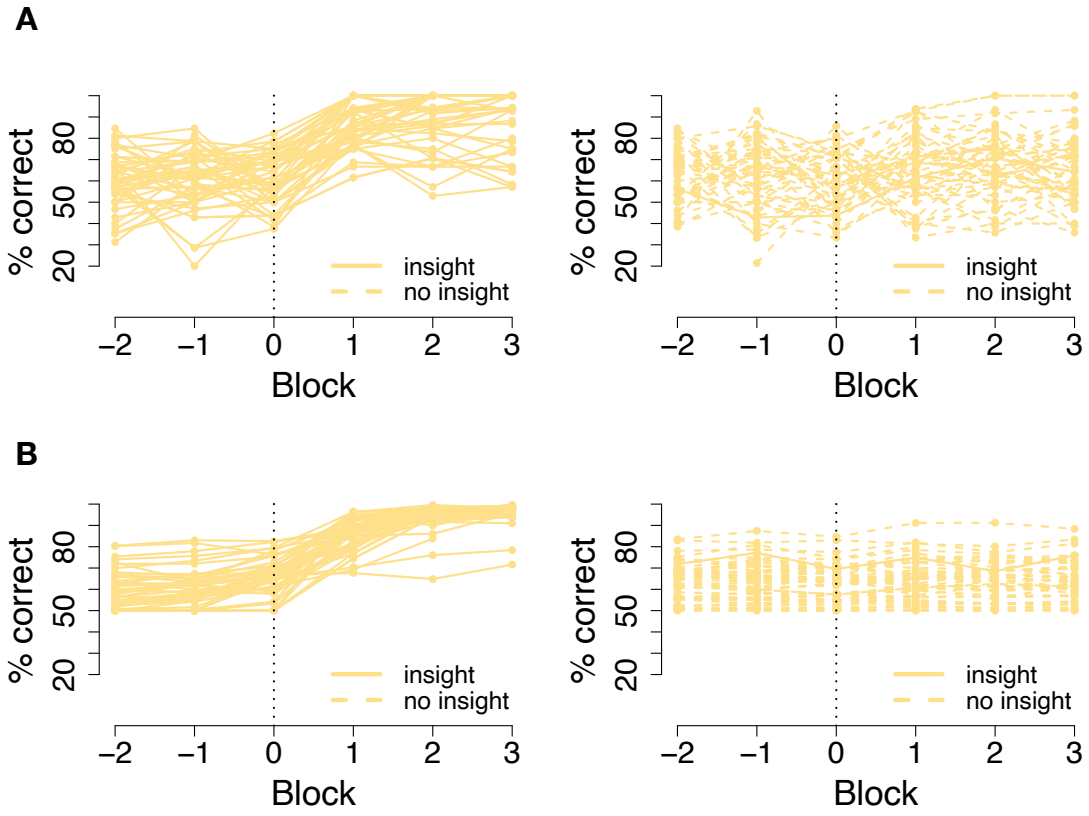


Fig. 5: Switch-aligned performance for insight group (left) and no-insight group (right). **(A)** Human insight group (48/99). **(B)** L1-regularised neural network insight group (48/99).

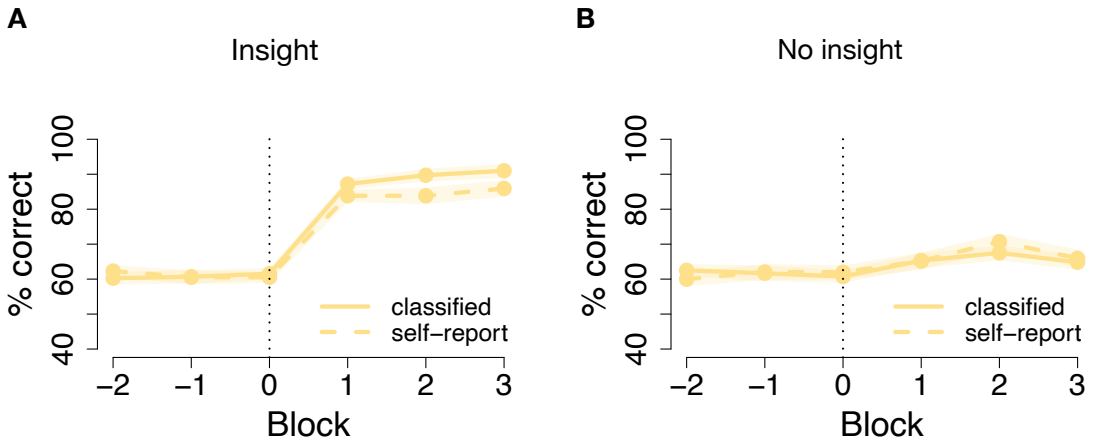


Fig. 6: Switch-aligned performance and overlap between classification and self-reported colour use. **(A)** Switch-aligned performance and overlap (38) between classified insight subjects (48/99) and self-reported colour use (57/99). **(B)** Switch-aligned performance and overlap (32) between classified no-insight subjects (51/99) and self-reported no colour use (42/99).

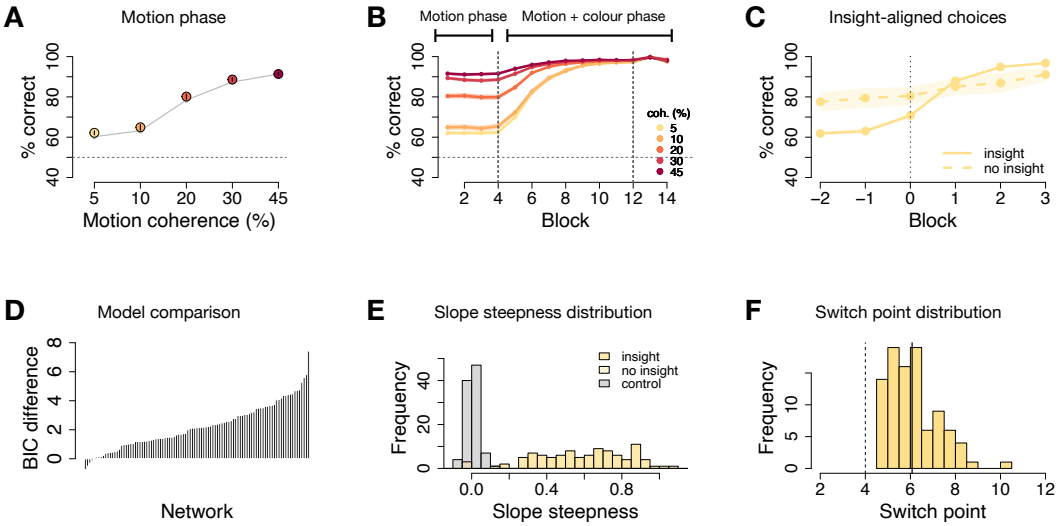


Fig. 7: L2 networks: Task performance and insight-like strategy switches **(A)** Accuracy (% correct) during the *motion phase* increases with increasing motion coherence. Blocks shown are halved task blocks (50 trials each). N = 99, error bars signify SEM. Grey line is human data for comparison. **(B)** Accuracy (% correct) over the course of the experiment for all motion coherence levels. First dashed vertical line marks the onset of the colour predictiveness (*motion and colour phase*), second dashed vertical line the "instruction" about colour predictiveness. N = 99, error shadows signify SEM. **(C)** Switch point-aligned accuracy on lowest motion coherence level for insight (95/99) and no-insight (4/99) networks. Blocks shown are halved task blocks (50 trials each). Error shadow signifies SEM. **(D)** Difference between BICs of the linear and sigmoid function for each network. **(E)** Distributions of fitted slope steepness at inflection point parameter for control networks and classified insight and no-insight groups. **(F)** Distribution of switch points for insight networks. Dashed vertical line marks onset of colour predictiveness. Blocks shown are halved task blocks (50 trials each).

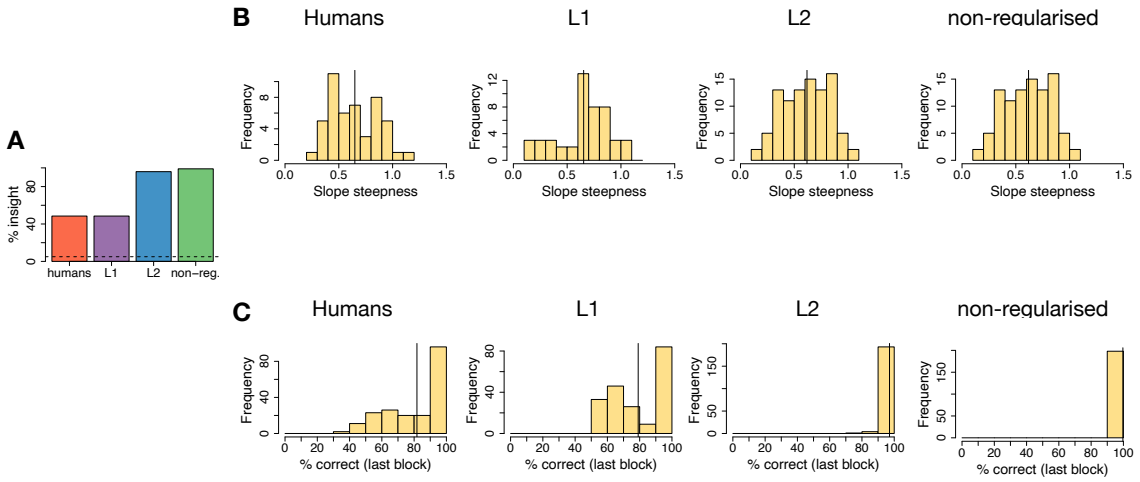


Fig. 8: Comparison of insight percentage and performance in the last task block across groups **(A)** % insight-like switches in humans, L1, L2 and non-regularised networks, respectively. Dashed line marks chance percentage of "insight". **(B)** Distributions of switch points for humans, L1, L2 and non-regularised networks. Blocks shown are halved task blocks (50 trials each). **(C)** Distributions of performance (% correct) for humans, L1, L2 and non-regularised networks for the last block of the *colour and motion phase* before the colour instruction.