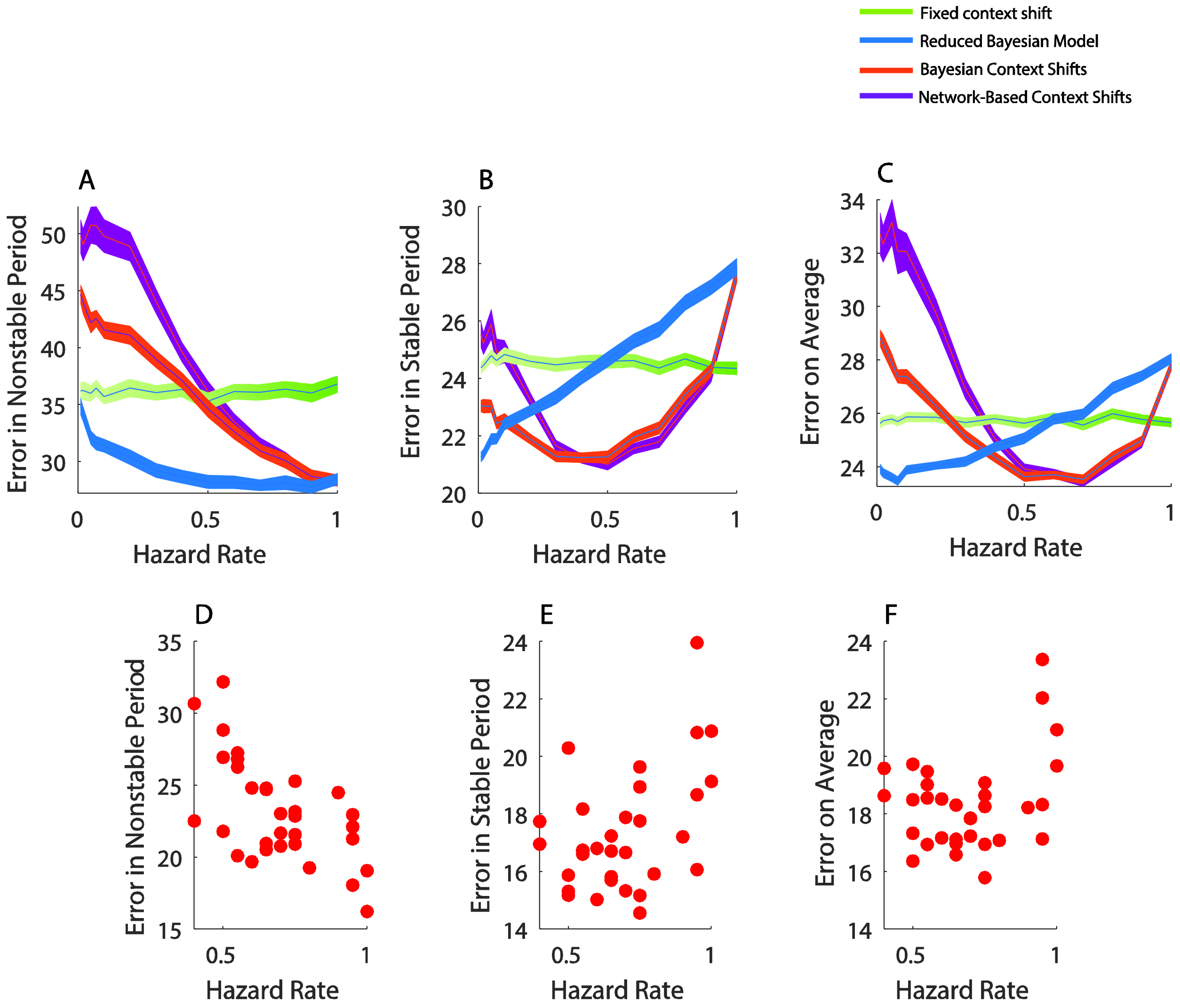
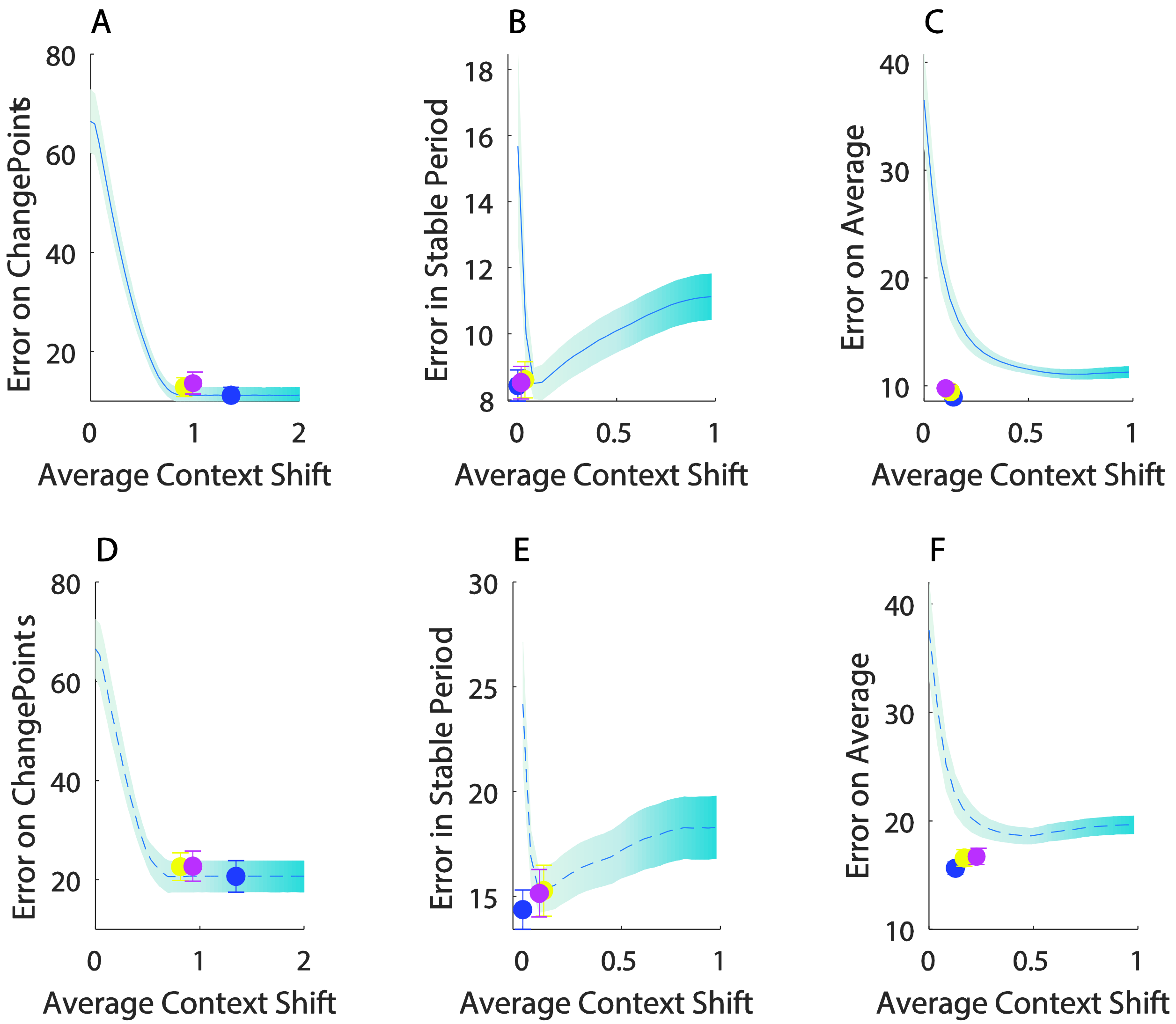
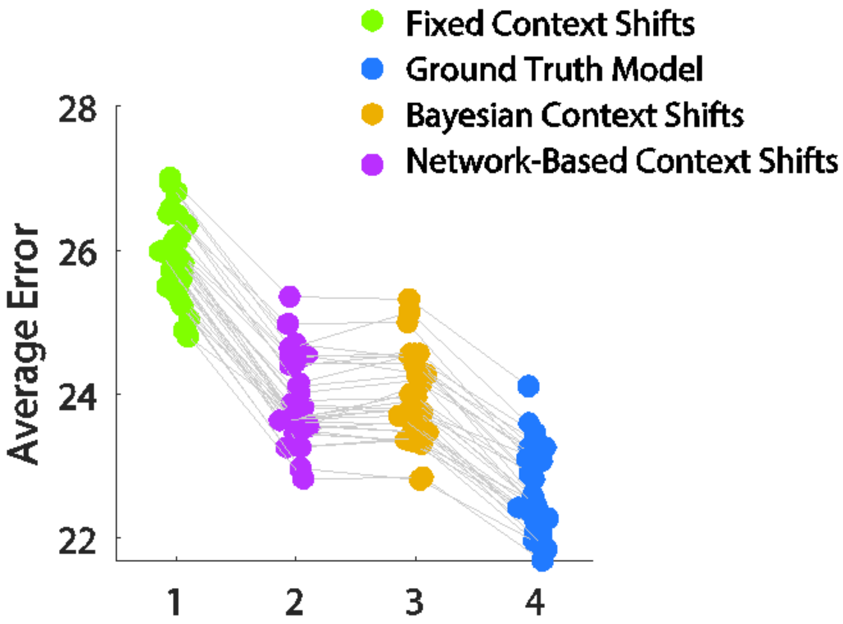


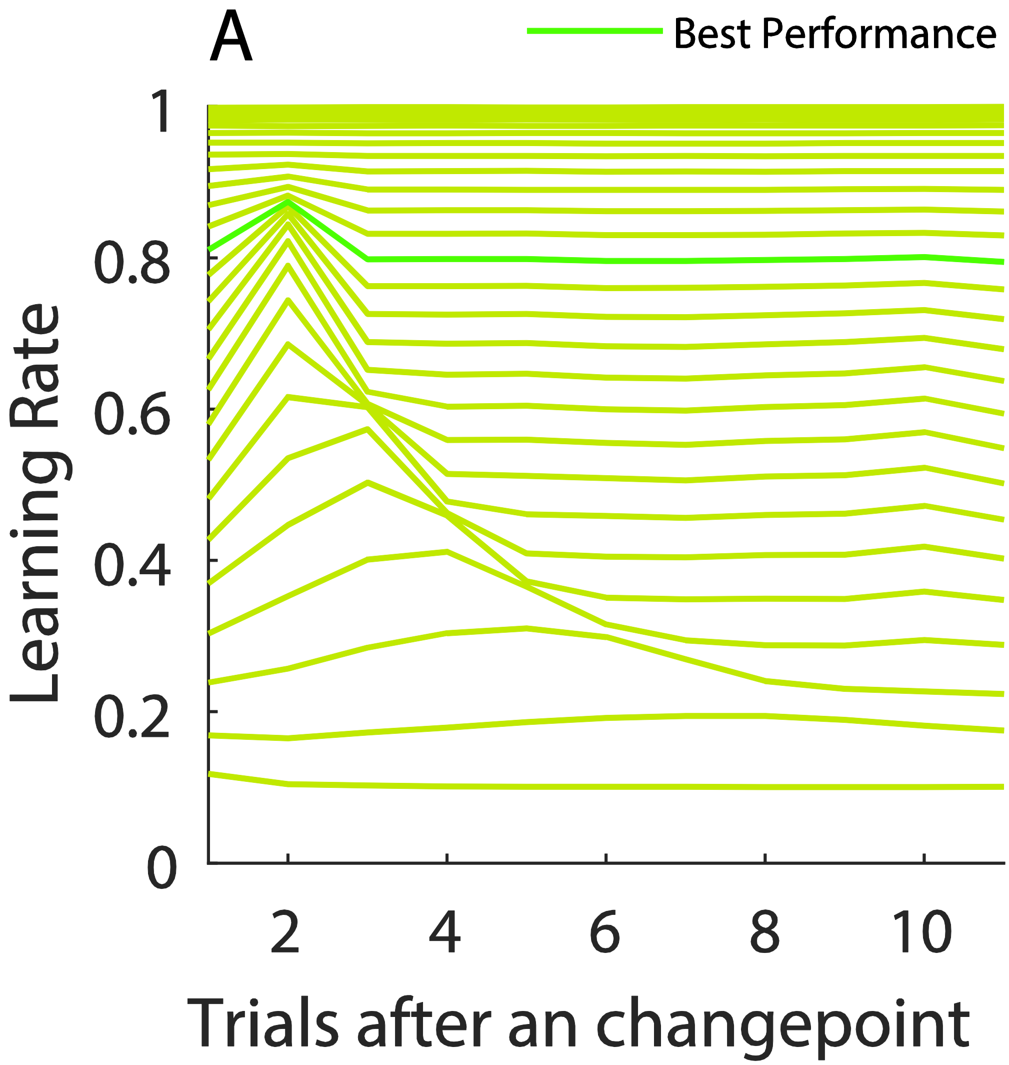
Supplementary Figure 1: **Top row: Performance of an example fixed context shift model, the Bayesian, network-based dynamic context shifts and the reduced Bayesian models. Bottom row: The relationship between fitted Hazard rate and performance in human subjects** Performance is shown separately for A) changepoint trials B) Non-changepoint trials and C) all trials on average. A fixed hazard rate of 0.6 hazard rate minimized errors for both the Bayesian and Network-based context shift models, and thus was used for all simulations of these models in the main paper. D) Humans with higher fitted hazard rate show lower error during non-stable period $(\mathrm{R}^2 = 0.61, \mathrm{p} - \mathrm{value} = 0.0002)$ (E)Subjects with higher fitted hazard rate show more error during stable period. $(\mathrm{R}^2 = 0.45, \mathrm{p} - \mathrm{value} = 0.0094)$ F) Error on average does not show a statistically significant pattern $(\mathrm{R}^2 = 0.31, \mathrm{p} - \mathrm{value} = 0.07)$.



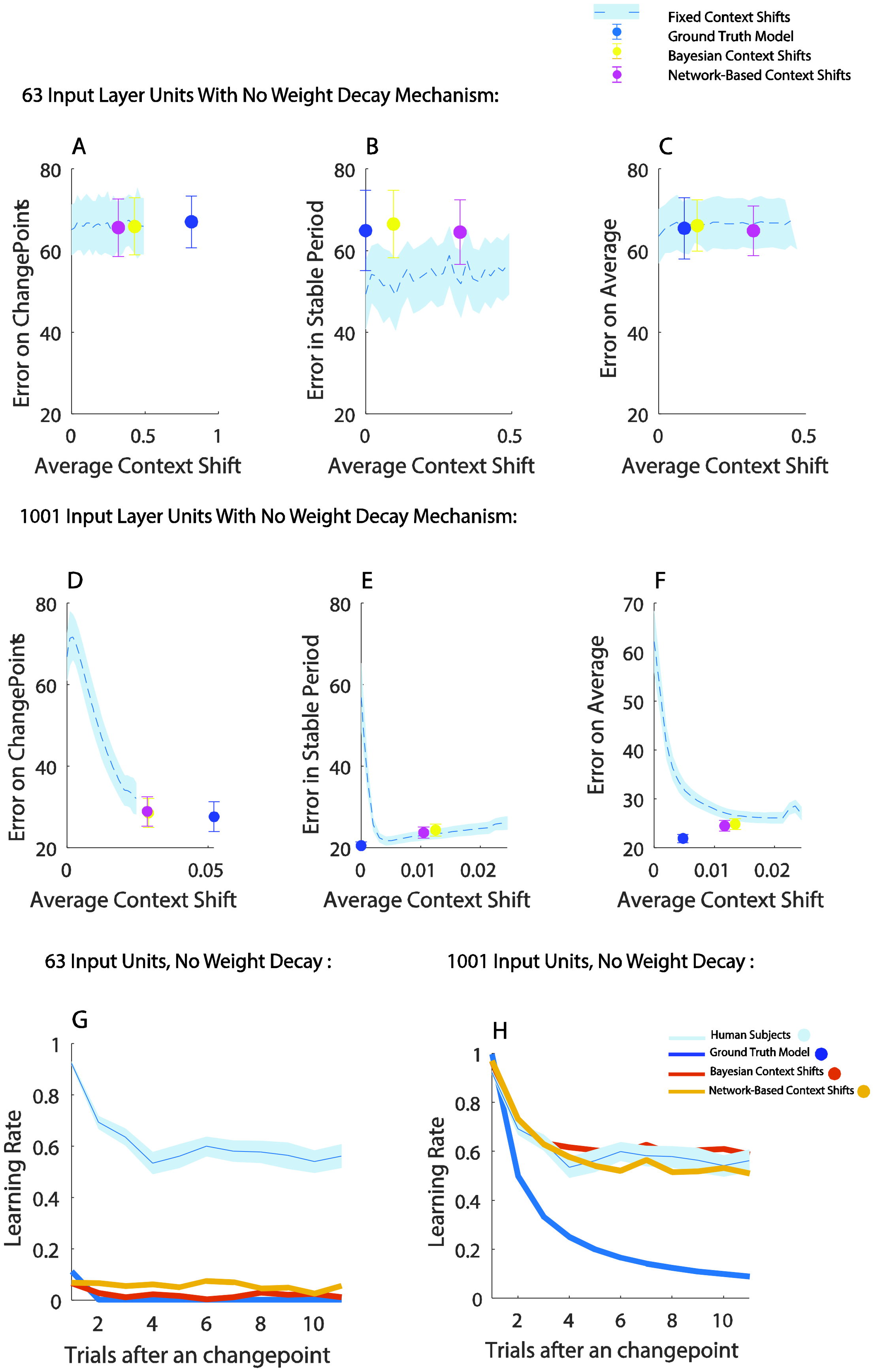
Supplementary Figure 2: Dynamic context shifts improve performance across noise conditions. This figure provides the same analysis as figure 3 c-e but with different noise levels. For each plot error (ordinate) is plotted against context shift size (abscissa) for fixed (line/shading) and dynamic (points) context shift models. A-D: performance of models in a predictive inference task with noise = 10 for A) changepoint trials, B) stable trials, and C) on aggregate. Averaged across all trials, the performance of the dynamic context shift models was better than the best fixed context shift model ($Bayesian\ Context\ Shift : t = 21.9. df = 31. p < 10^{-16}$ D-F. Network-Based Context Shift: $t = 20.48, , df = 31. p < 10^{-16}$). D-F: performance of models in a task that includes both noise conditions from McGuire 2014 (noise = 10 or 25) in alternation. Performance of the dynamic models was again better than that of the best fixed context shift model on aggregate ($Bayesian\ Context\ Shift: t = 20.62. df = 31. p < 10^{-16}. Network - Based\ Context\ Shift: = t = 15.91. df = 31. p < 10^{-16}$) suggesting that our results hold across the range of experimental noise conditions.



Supplementary Figure 3. The Ground truth model performs better than all other models as number trials goes to infinity. Each dot represents average error for each model over all trials in one simulation of the experiment (here 1000 trials). The Y axis is average error and the x axis is the four models used in the paper. (Fixed context shift, Network-based Context Shift Bayesian Context Shift and ground truth models)



Supplementary Figure 4: Effective learning rate dynamics for different fixed context shift models. A) In extreme cases where the context shift is close to zero or one, effective learning rate is independent of trial after a changepoint. However, with moderate levels of context shift, the model shows increased learning one (or a few) trial(s) after a changepoint which is consistent with the fact that it takes a few time steps for the model to dissociate from previous context and transition to a new context.



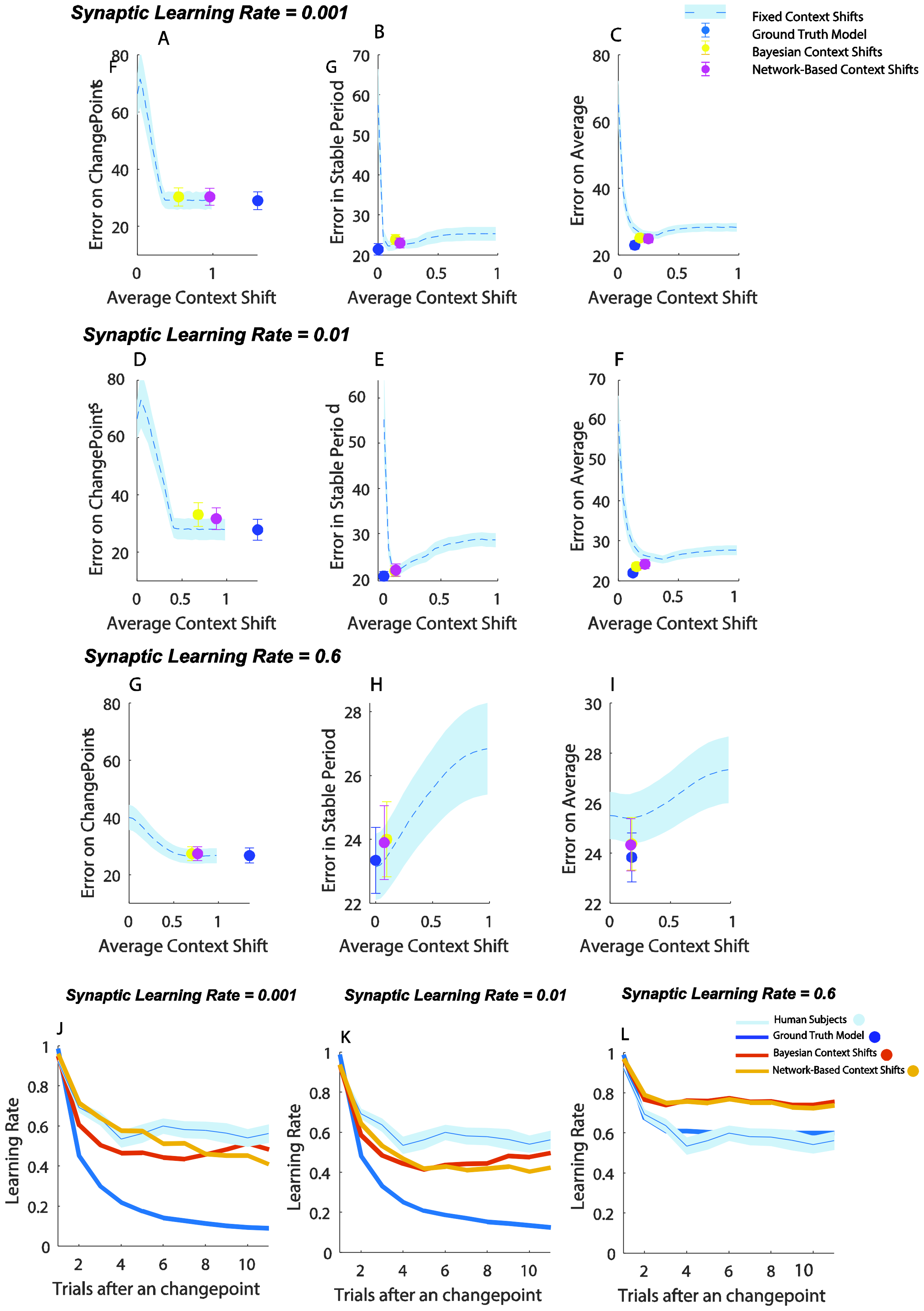
63 Input Layer Units With No Weight Decay Mechanism:

1001 Input Layer Units With No Weight Decay Mechanism:

63 Input Units, No Weight Decay :

1001 Input Units, No Weight Decay :

Supplementary Figure 4. Model suffers interference when all mechanisms for weight decay are removed, but this can be rescued by extending the size of the input layer. A-C: Performance of a model containing 63 input layer units but without weight decay on A) trials following a changepoint B) non-changepoint trials and C) on average. D-F: Performance of the same model , but with 1001 input layer units on D) trials following a changepoint E) non-changepoint trials and F) on average, Within this simulation, the performance of the Bayesian context shift and network-based context shift are significantly better than the best fixed context shift $Bayesian\ Context\ Shift\ : t = -7.2\ . df = 31. p = 4.2\ \times 10^{-8}\ . Network - Based\ Context\ Shift: t\ =\ -10.49. df = 31. p = 9.95\ \times 10^{-12}$. (The x axis (i.e. context shift size) in all figures A-F is in radians). G-H: Effective learning rate (ordinate) as a function of trials after a changepoint (abscissa for non-decay (G) and non-decay with extended input layer (H) simulations. Note that key results reported in the main text hold for a model without any form of weight decay so long as that model contains a sufficiently large population of input units.


Synaptic Learning Rate = 0.001
A
B
C
Fixed Context Shifts
Ground Truth Model
Bayesian Context Shifts
Network-Based Context Shifts
80
60
40
20
F
Error on ChangePoints
0
1
Average Context Shift
70
60
50
40
30
20
G
Error in Stable Period
0
0.5
1
Average Context Shift
80
60
40
20
Error on Average
0
0.5
1
Average Context Shift
Synaptic Learning Rate = 0.01
D
E
F
80
60
40
20
Error on ChangePoints
0
0.5
1
Average Context Shift
60
50
40
30
20
Error in Stable Perio d
0
0.5
1
Average Context Shift
70
60
50
40
30
20
Error on Average
0
0.5
1
Average Context Shift
Synaptic Learning Rate = 0.6
G
H
I
80
60
40
20
Error on ChangePoints
0
0.5
1
Average Context Shift
28
26
24
22
Error in Stable Period
0
0.5
1
Average Context Shift
30
28
26
24
22
Error on Average
0
0.5
1
Average Context Shift
Synaptic Learning Rate = 0.001
Synaptic Learning Rate = 0.01
Synaptic Learning Rate = 0.6
J
K
L
Human Subjects
Ground Truth Model
Bayesian Context Shifts
Network-Based Context Shifts
1
0.8
0.6
0.4
0.2
0
Learning Rate
2
4
6
8
10
Trials after an changepoint

Supplementary Figure 5. Key results are robust to synaptic learning rate. A-I: Model simulations were performed with three synaptic learning rates (0.001, 0.01, 0.6) to examine the robustness of our primary results. Error (ordinate) is plotted as a function of average context shift (abscissa) separately for changepoint trials (A,D,G) and periods of stability (> 5 trials after a changepoint; B,E,H) as well as aggregated across all trials (C,F,I). J-L: Learning rate (ordinate) as a function of trials after changepoint (abscissa) for models with fixed synaptic learning rates of (J) 0.001, (K) 0.01 and (L) 0.6.

Appendix – 1

Changepoint condition:

The probability that a changepoint has occurred in the current trial is computed as the likelihood of current trial observation coming from a new helicopter position, which is randomly chosen from a uniform distribution between 0 and 300, over the sum of the likelihoods of the current observation coming from a non-changepoint distribution (Gaussian distribution centered around the most recent belief about helicopter positions) and changepoint distribution:

$$\text{CPP} = \Omega_t = \frac{U(X_t|0.300)\ H}{U(X_t|0.300)H + N(\delta_t; 0.\sigma_\mu^2 + \sigma_N^2)(1-H)}$$

Where Ω_t is changepoint probability, H is the hazard rate and is the same hazard rate used in the generative process for simulating outcomes, δ_t is the difference between the predicted and actual outcome, σ_N^2 is the variance of the distribution of bags around the helicopter, and σ_μ^2 is the variance on predictive distribution (estimation uncertainty) and is computed recursively after observing an outcome and using the changepoint probability of the current trial:

$$\sigma_\mu^2 = \Omega_t \sigma_N^2 + (1-\Omega_t)\,\tau_t \sigma_N^2 + \Omega_t(1-\Omega_t)(\tau_t + B_t(1-\tau_t) - X_t)$$

Where B_t is the current belief about helicopter position and X_t is the current trial outcome and τ_t is the relative uncertainty.

Relative uncertainty is computed as estimation uncertainty about bag locations as a fraction of total uncertainty that is sum the of estimated uncertainty and noise.

$$RU = \tau_{t+1} = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_N^2}$$

Oddball condition:

The probability of an observation being an oddball is computed in a similar manner using the prediction error of the model on each trial:

$$\text{OBP} = \Omega_t = \frac{U(X_t|0.300)\ H}{U(X_t|0.300)H + N(\delta_t; 0.\sigma_\mu^2 + \sigma_N^2)(1-H)}$$

Where H is the hazard rate of the oddball condition. Here, the numerator is the likelihood of current outcome being an oddball and the second term in the denominator is the likelihood of current outcome coming from a normal distribution with the mean of predicted outcome and variance of total uncertainty.

The estimation uncertainty is computed similar to change point condition, albeit with a few modifications:

$$\sigma_{\mu}^{2} = \Omega_t \frac{\sigma_N^2 \tau_t}{1 - \tau_t} + (1 - \Omega_t)\, \tau_t \sigma_N^2 + \Omega_t (1 - \Omega_t)(\delta_t \tau_t)^2 + \sigma_{drift}^2$$

Where the first term represents the contribution of an oddball observation to the uncertainty, the second term is the contribution of a nonoddball observation to the uncertainty, the third term is uncertainty proportional to the prediction error received from either an oddball or a nonodball trial and the last term reflects the uncertainty expected from the drift rate of the helicopter in between two trials.

Again similar to changepoint condition, relative uncertainty is computed as estimation uncertainty over total uncertainty:

$$RU = \tau_{t+1} = \frac{\sigma_{\mu}^{2}}{\sigma_{\mu}^{2} + \sigma_N^2}$$