**Model-Based Learning Influences Model-Free Value**

**Notes/comments**

* Max: have adjusted incon/con to common rare. I have left the figures as they are for now since I thought we might want to change their presentation anyway

**Abstract**

Reinforcement learning (RL) is widely regarded as divisible into two distinct strategies which have been theorised computationally. Model-free learning is a simple RL process in which outcome is associated with actions, whereas model-based learning relies on the formation of internal models of the environment to maximise reward. An unusual model-based reward prediction error (RPE) has recently been discovered that is not explained by current model-based computational architectures and raises the possibility of interaction between the two learning systems. Here, using an adapted two-stage decision task, I present evidence that features of model-based learning can alter model-free value in healthy individuals. Model-based predictions biased passive accumulation of model-free value in an attention-independent manner, an effect which correlated with model-based performance. These findings question the canonical view that these learning systems operate independently and provides evidence for a more integrated approach.

**Introduction**

Model-free learning involves simply reinforcing actions that lead to reward[[1]](#endnote-2), without any further internal model of cause and effect[[2]](#endnote-3). Although it is computationally cheap, model-free learning is slow at reacting to dynamic environments where contingencies constantly shift[[3]](#endnote-4),[[4]](#endnote-5). As early as Tolman[[5]](#endnote-6), it was proposed that animals are not merely restricted to simple RL but are capable of using complex internal models of the environment to guide behaviour. Model-based learning formalises Tolman’s theory, describing the acquisition of a cognitive map detailing how environmental states are linked to each other. Modern theoretical frameworks outline how action values for different paths through this map may be calculated to choose an optimal strategy for reward maximisation[[6]](#endnote-7).

Key variables in model-free learning are value and RPE. Value is the predicted reward associated with an action. If the predicted reward does not match the actual reward, an RPE acts as a learning signal to improve the accuracy of the value estimate3. Various studies have implicated the mesostriatal system in model-free learning[[7]](#endnote-8),[[8]](#endnote-9),[[9]](#endnote-10), which is in keeping with the dopaminergic-RPE hypothesis in RL[[10]](#endnote-11). Evidence is consistent with the proposed signature of RPE within the ventral striatum (VS)[[11]](#endnote-12),[[12]](#endnote-13) and action value within the dorsolateral striatum (DLS)1,12,[[13]](#endnote-14),[[14]](#endnote-15),[[15]](#endnote-16).

Model-based learning, on the other hand, is thought to involve predominantly the prefrontal cortex (PFC)[[16]](#endnote-17),[[17]](#endnote-18),[[18]](#endnote-19),[[19]](#endnote-20) and the dorsomedial striatum (DMS)[[20]](#endnote-21),[[21]](#endnote-22). The utility of the model-based system arises from the formation of an accurate map of environmental states which can be used prospectively to assess future action sequences, in contrast to model-free learning, which relies on historical, cached values. Model-based value variables are state-specific rather than action-specific as in model-free learning. A key variable in driving these environmental maps in model-based learning is thought to be a state prediction error (SPE), observed in the intraparietal sulcus20 and the lateral PFC20, which is produced when there is a conflict between the external structure of the environment and the internal model.

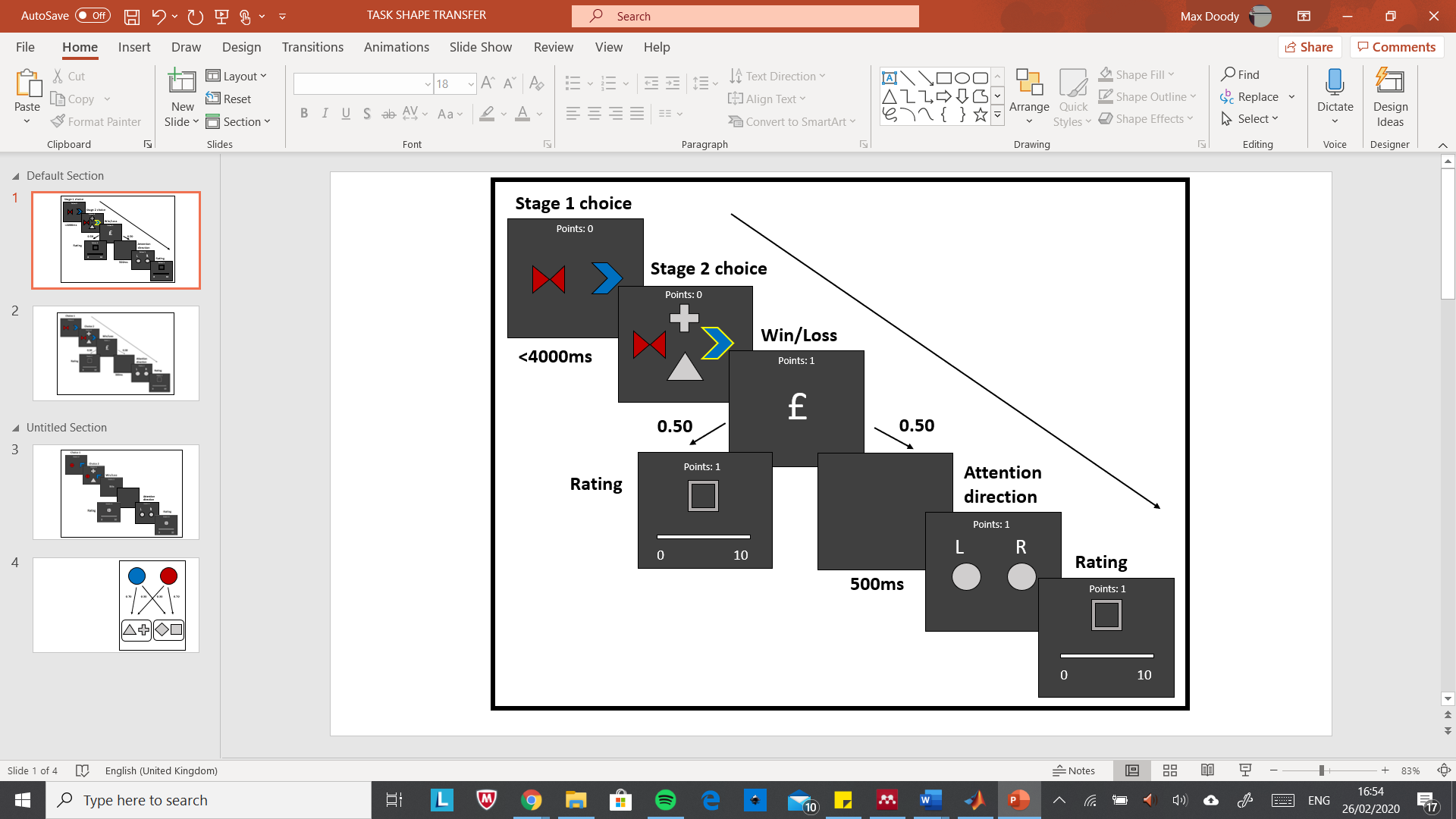
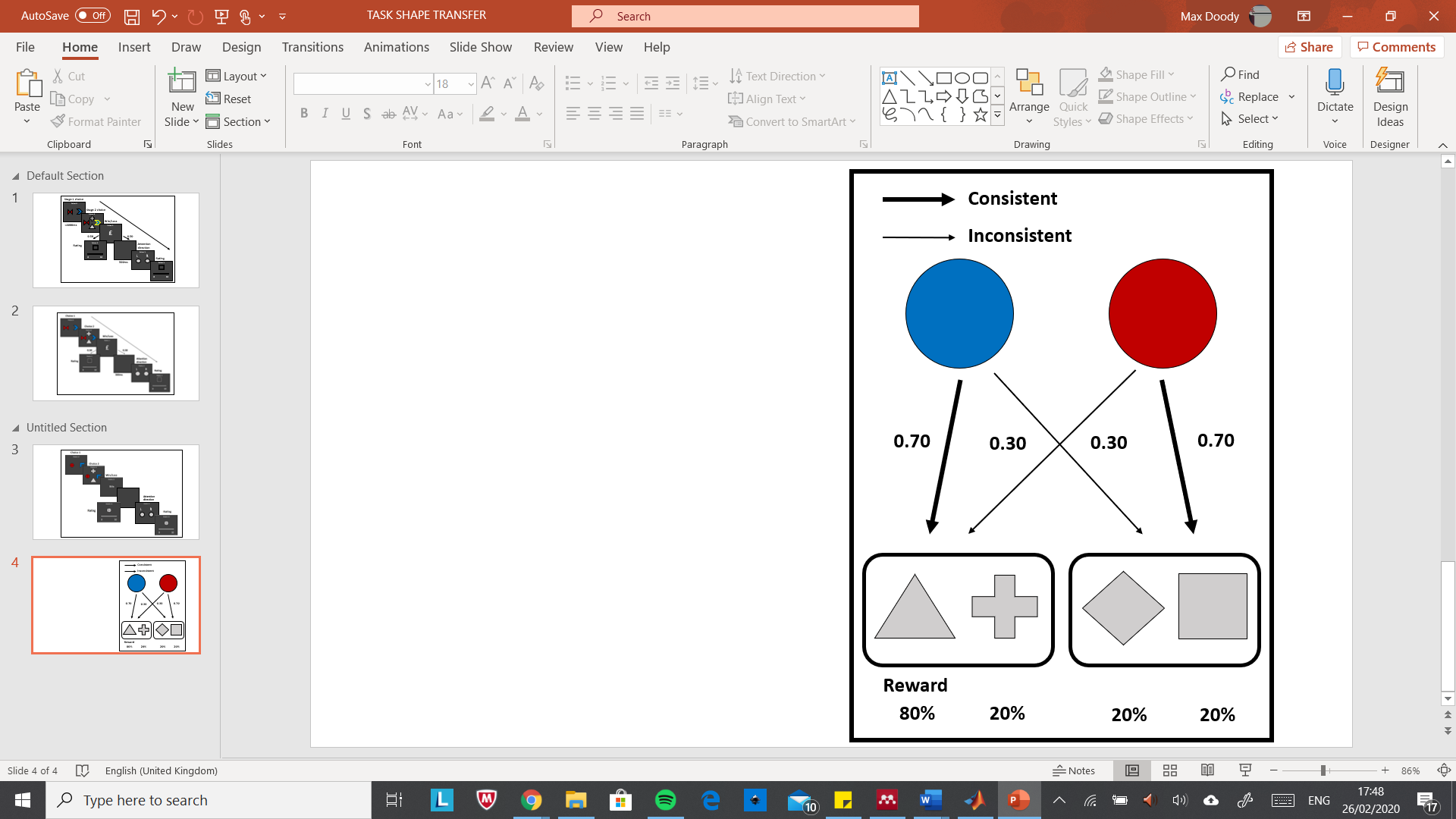
To investigate the neural basis of model-free and model-based learning, Daw et al.11 devised a two-stage decision task in conjunction with fMRI. In this Markov decision task, participants make two sequential decisions, both between two shapes, to obtain a reward. However, the transitions between the first and second stage are probabilistic, with one pathway more common than the other. If participants are driven by simple RL, they will repeat decisions in the first-stage that led to a reward, whereas a model-based agent will repeat an action that *would* have led to a win, via a common, but not an uncommon transition. Consequently, the mixture of model-free and model-based approaches used by an agent can produce opposing behaviours that are quantifiable. Daw found evidence of an RPE signature within the VS and the medial PFC during this task. One component of this RPE was best explained by a model-based regressor, suggesting that a part of this RPE was model-based. In other words, in line with an internal model this area signals the values at the time of the transition relative to what would be expected given the first-stage choice. Similar signals have been observed over frontal electrodes in EEG[[22]](#endnote-23). In contrast to SPEs, the generation of a model-based RPE is unexpected, since it is not canonically needed for model-based learning. Given the unlikelihood of such a signal being epiphenomenal, either our account of model-based learning needs to change, or else the model-based RPE is utilised by another system. I set out to investigate if this model-based RPE influences model-free learning, given that the VS is associated with this process11,12,[[23]](#endnote-24),[[24]](#endnote-25). Such influence would allow internal models to alter value thought previously to be assigned solely through model-free association. One way to accomplish this is to probe how cached value is attributed to stimuli during the two-stage task. When participants are asked to report an estimate of stimulus value, this estimate might be influenced by the model-based state-outcome RPE – to which the model-free system is by definition blind – only if there is a transfer of information from the model-based system.

A close relationship between working memory (WM) and model-based performance has been reported. Model-based learning requires attentional resources[[25]](#endnote-26),[[26]](#endnote-27), and is vulnerable to transcranial magnetic stimulation in left dorsolateral PFC[[27]](#endnote-28). I therefore hypothesised that transfer value from the model-based to the model-free system may be an attention-dependent phenomenon. If so, we might expect transfer to also correlate with WM capacity and for attentional manipulations to affect transfer.

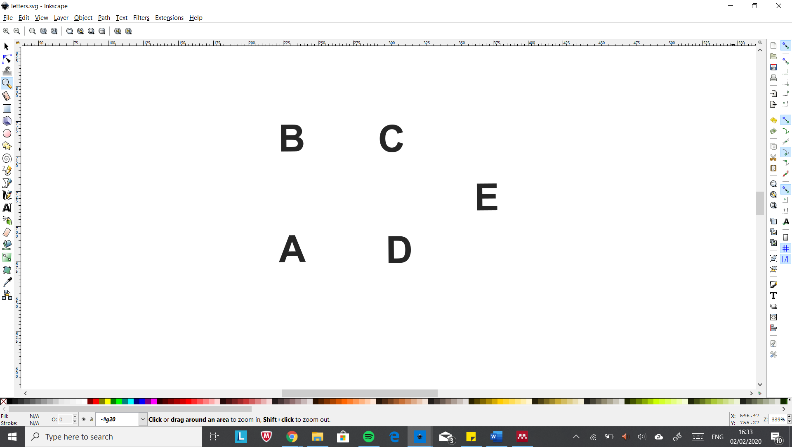
**Experiment 1: Healthy participants show model-based transfer of value**

**Method 1**

A cohort of healthy subjects (n=30) aged 18-40 were recruited to assess the presence of model-based transfer. The task performed was an adapted version of the two-stage decision task11, which has repeatedly been used to differentiate model-free and model-based components in decision-making20,22,24,29,[[28]](#endnote-29),[[29]](#endnote-30). All experiments were performed using a touchscreen PC running MATLAB R2018b. Subjects made two sequential choices (**fig.1A**). In the first stage a choice was made between two coloured shapes, leading to one of two second-stage states requiring a choice to be made on one of the two pairs of grey shapes. During any period, only one of the four possible grey shapes shown was likely to be rewarded. The probability that the colour of the first shape leads to a particular pair of grey shapes was fixed at the proportion 70/30 (**fig.1B**), such that each colour produced a common and rare transition. Participants were explicitly told that shape colour determines the stage-two shapes presented. The shape itself was not relevant to the outcome. A time limit of 4 seconds was applied to the first choice. Participants were then told whether they had won, with points banked visibly at the top of the screen. The rewarding stage-two shape was changed randomly every 32 trials. To encourage model-based learning, in line with theory16,[[30]](#endnote-31), the probability of the currently rewarding grey shape to produce a win was 0.8, whereas other shapes had a 0.2 chance of winning. The rewarding location was randomised between both pairs of shapes every 32 trials to avoid a model-based signature from the model-free system[[31]](#endnote-32). Likewise, to counterbalance, the left-right positions of stage-one and two shapes were randomised.



**Fig. 1** | **A:** Structure of the task. **B:** The probabilistic relationship between stage-one colour and stage-two shapes.

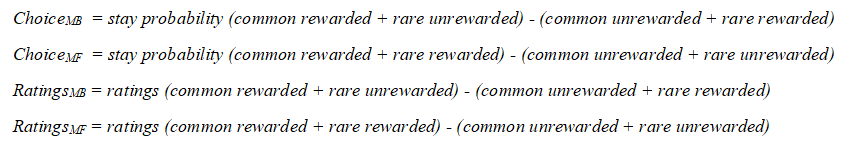


Two new features were added to the standard task. First, I assessed transfer of value to the stage-one shapes. By probing value trial-by-trial, a nuanced assessment of value is possible[[32]](#endnote-33). Participants were shown one of the five stage-one shapes, presented in grey, and asked to “assign a value between 1 and 10 depending on how likely the shape shown was to be rewarding at that time” using a number line. However, the shapes had no effect on reward since colour alone determined the transition to stage-two shapes. Shapes shown on the previous or current trial were never assessed to deter subjects from holding the shape in WM.

The second addition was the inclusion of an attention manipulation. On half of the trials, subjects were asked to select the side of the screen (left or right) on which their chosen first-stage shape was positioned. This aimed to direct the attention of the participant back to their stage-one decision without referencing features that directly related to reward or were probed for transfer. In the other half of trials, subjects did not see this screen and proceeded to the rating screen. In total subjects completed 4 blocks of 64 trials to give a total of 256 trials.

Four equations were used to assess the degree to which choices and shape ratings followed predicted model-based and model-free patterns in line with previous work31(see equations 1-4). If the participant makes use of the task structure (is model-based) and won on the previous trial via an rare transition they should not re-select the same colour on the next trial in order to maximise reward. Conversely, if the participant is unaware of the transitions (is model-free) we would predict that they return to the colour that led to a win on the next trial (**fig.2A/B**). By assigning stay probabilities split by conditions on the previous trial (rewarded-unrewarded/common -rare) it is possible to establish an estimate of the relative contributions of these strategies to choice.

**Equations 1-4**



**Results 1**

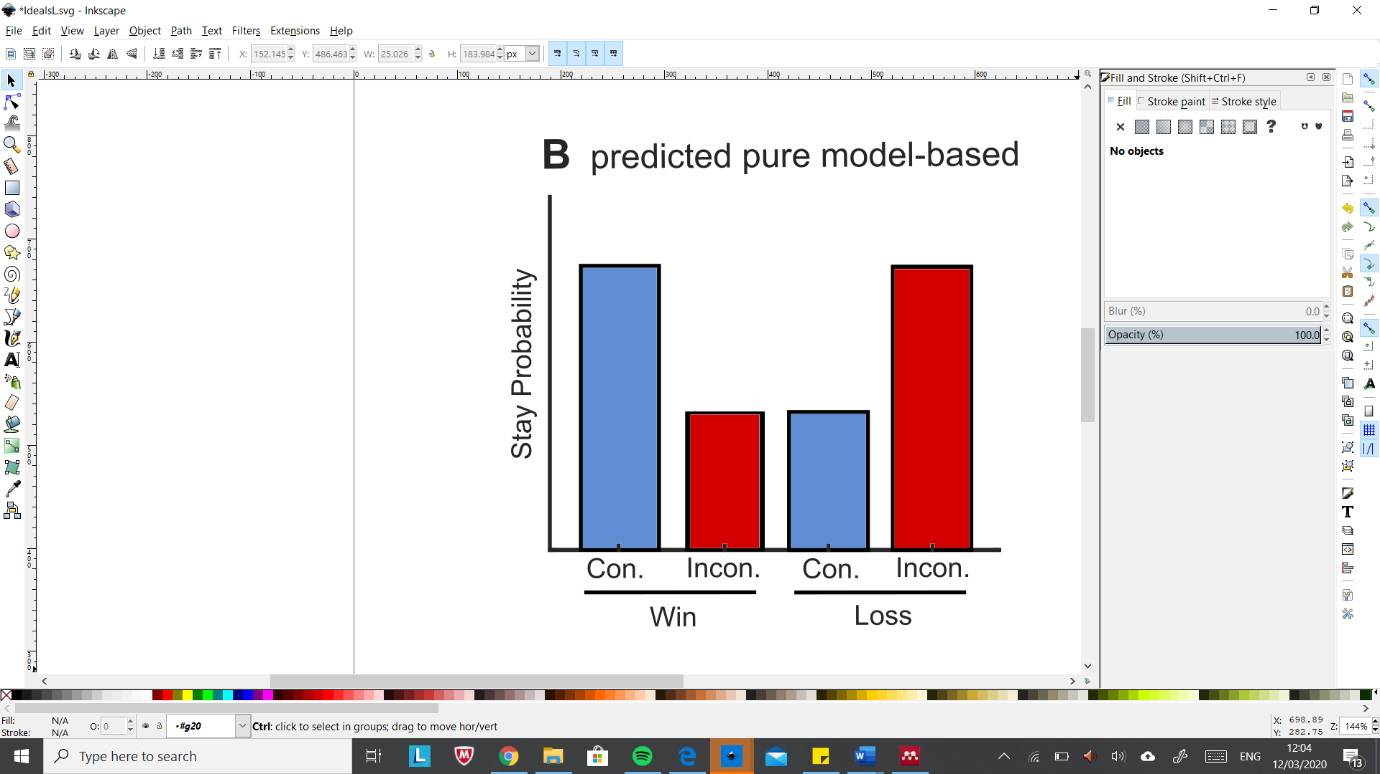
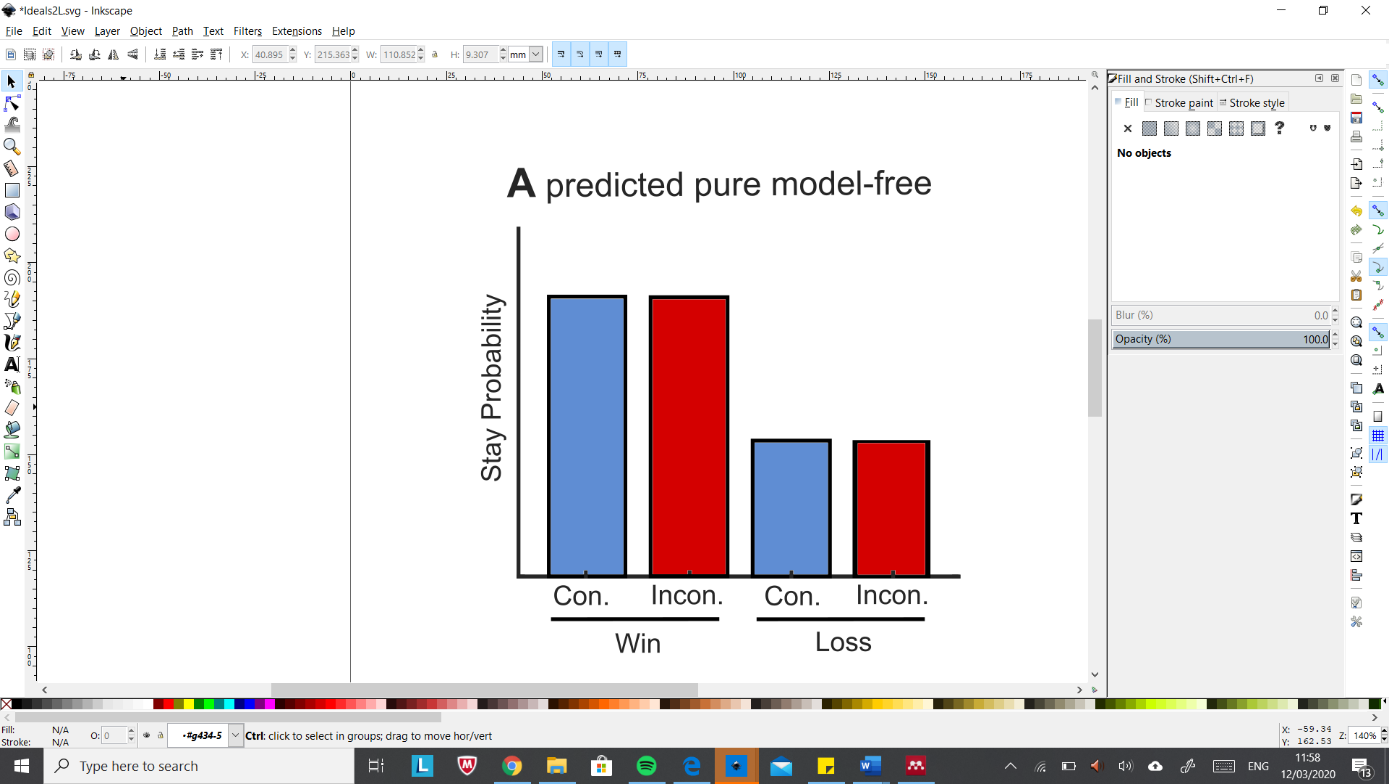
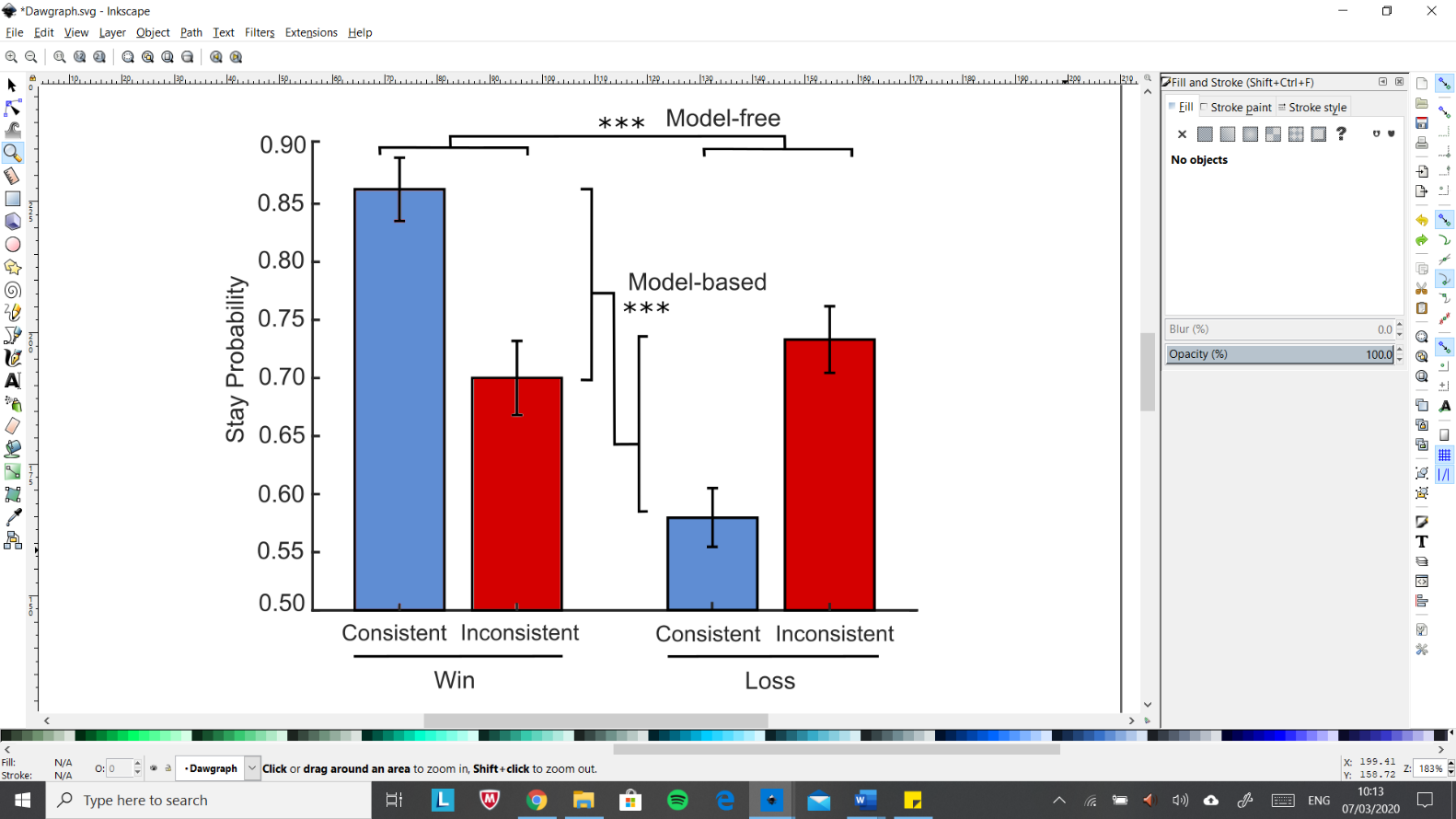
Analysis of the results was conducted using MATLAB 2018b and SPSS. Ratings were organised by rank to ensure even contributions by all participants to overall results irrespective of their ratings range and reduce the influence of outliers. The definition of ‘being model-based’ was taken as *ChoiceMB*>0; participants who failed to behave in a model-based fashion were excluded. This resulted in the removal of 3 participants from the core analysis. Furthermore, one participant was excluded due to their ratings having a standard deviation of less than 20 pixels (2.4 SD below the population mean). Due to the error in width of the human finger it is difficult to ascertain if any variation in ratings is deliberate. A further participant was excluded due to a computer error, meaning in total 25 data sets were analysed.

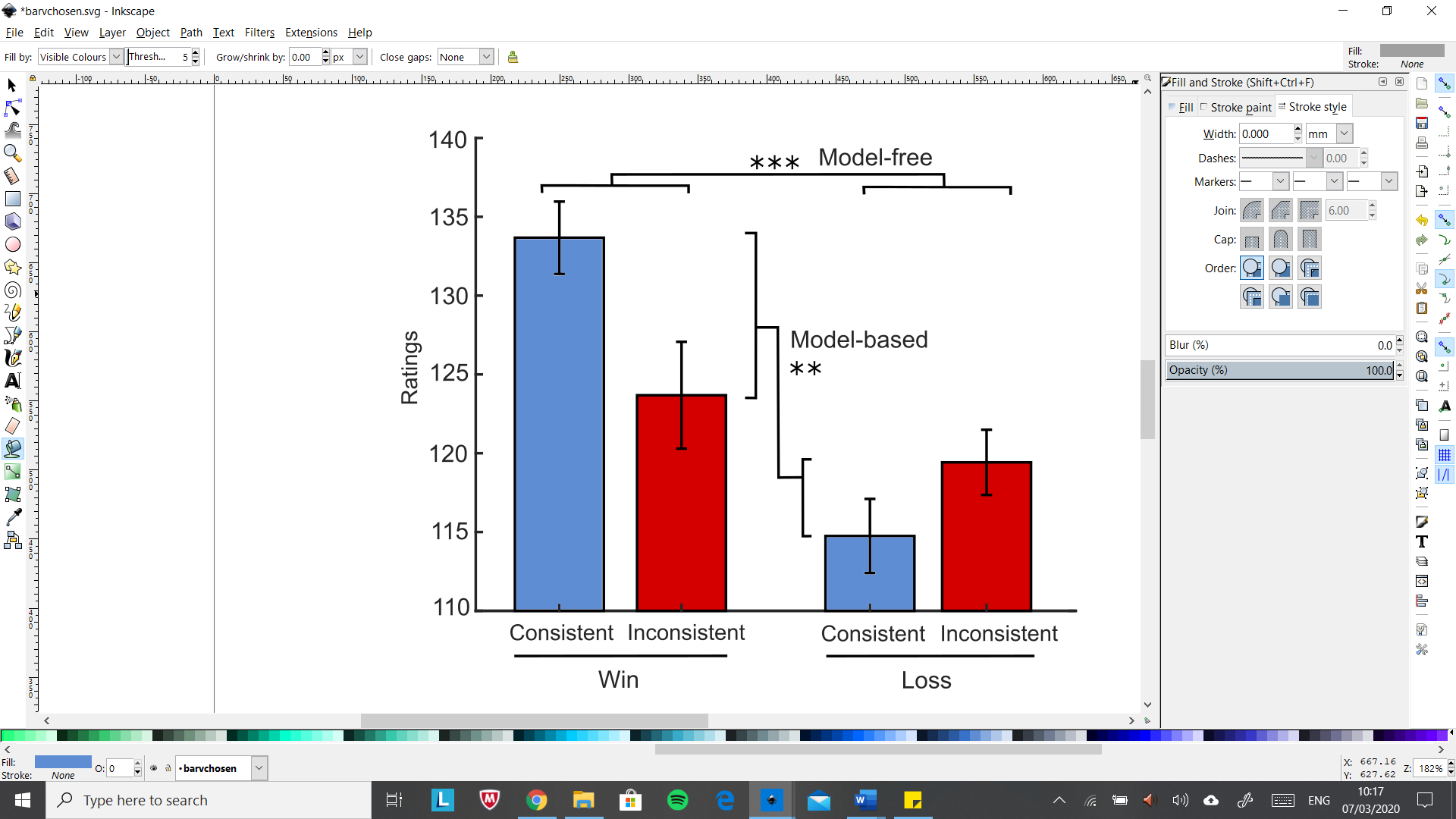
In line with previous studies, subjects showed a mixture of model-based and model-free strategies (**fig.2C**). Two-way ANOVA revealed a significant effect of reward (F(1, 27) = 29.4, p<.0001) and a significant interaction between consistency and reward (F(1, 43) = 75.5, p<.0001). Participants were more likely to stick to a given colour in a trial if they had won rather than lost on the previous trial, indicating model-free influence on their choices. Subjects were also more likely to stick to a colour if the win on a previous trial involved a common transition and showed the converse for loss. These indicate that the model-based system contributed to the decision process.

The subjective value ratings for the irrelevant stage-one shapes were split according to the conditions (rewarded-unrewarded/common -rare) during the last trial in which that shape had been chosen in stage-one. If people transferred model-based value to the shape, we would expect that when a shape was chosen, and a reward received on a common transition, subsequent ratings of that shape should be higher than if a reward was received on an rare transition; the opposite would be true for loss. Remarkably, ratings showed a similar pattern of model-based and model-free influences (**fig.2D**). Two-way ANOVA revealed significant effects of reward (F(1, 62) = 18.2, p<.0001) and a significant interaction between consistency and reward (F(1, 62) = 7.29, p<.01). The model-based (*RatingsMB*) and model-free (*RatingsMF*) influence on ratings were (15.95±6.50%, mean±SEM) and (23.8±8.10%, mean±SEM) respectively. Overall, these results suggest that when value is assigned on receiving or not receiving a reward, both model-free and model-based properties are encoded. Model-based value thus influenced shape value ratings, despite the fact that stage-one shapes were outcome-irrelevant, and they were seen at least one trial before.

Comparisons between trials in which attention was or was not manipulated revealed no significant effect of attention on ratings (supplementary figure 1). Surprisingly, attention did not have a significant effect on model-based behaviour (supplementary figure 2), nor on model-free choice. The transfer of value to the unchosen stage-one shape was also assessed. Two-way ANOVA revealed no significant effect of reward (F(1, 36) = 1.05, p=0.312), consistency (F(1, 39) = 3.01, p=0.087) or interaction between them (F(1, 19) = 2.27, p=0.148). Four-way ANOVA showed a weak effect of attention on ratings of the unchosen stage-one shape (F(1, 274) = 4.95, p<.05), with attention on average increasing the ratings of unchosen shapes. No evidence was found of a correlation between digit span tests and model-based choice or transfer.





****

**Fig. 2** | **Healthy subjects showed mixed model-based/model-free choice and transfer.** Idealised results of a pure **A** model-free and **B** model-based approach to the task in rewarded-unrewarded/common (con.)-rare (incon.) trials for illustrative purposes. **C:** Contributions of model-free and model-based learning to stay probability. **D:** Contributions of model-free and model-based learning to value ratings of stage-one shapes. Error bars represent within-subject standard error of the mean.

**Discussion 1**

Since the probe in this task is assessing perceived value of an object’s shape irrespective of context this is taken to be a model-free assessment. Consequently, the results can be interpreted as transfer of information from the model-based to the model-free system, in line with a hypothesis first suggested by Daw et al.11 to explain a model-based RPE. Whilst we cannot rule out that ratings simply reflect model-based valuations, there are several reasons why this is unlikely. Firstly, by definition the model-based system does not learn shape value, instead it learns the environmental structure6. Secondly, stage-one shapes were not outcome-relevant. If the model-based system were accurately valuing stage-one shapes, we would expect all shapes to be rated equally as they did not affect reward probability. Participants were explicitly told that colour alone determines transition probability. The most parsimonious explanation for the results seen is that a value learning system that was blind to the task structure was nevertheless influenced by it. Thirdly, the task design ensured that the shape probed was not shown during choice one of that trial or of the trial before. Given that the mean distance between choosing a shape and it being probed was 5.31 trials, an average of 10.62 stage-one shapes were shown between choice and ratings, making it implausible that details of that trial were held in the model-based structure.

One possibility is that the transfer probe may not reflect model-free value but instead the ‘post-arbitration’ value. Arbitration is a putative process that controls the relative contributions of model-free and model-based learning to decision-making. The degree of model-based and model-free choice in the two-stage task varies with projections from the ventromedial PFC to the medial striatum30, suggesting that the medial striatum may play a role in arbitration. Given that previous studies lacked the spatial sensitivity to rule out the presence of the model-based RPE in the medial striatum, it is plausible the RPE may feed into arbitration11,22. However, there was no effect of consistency or reward on the probability of sticking to a stage-one shape on the next trial. This is presumably because the decision is dominated by colour, which was explicitly relevant, demonstrating that whilst stage-one shapes were rated in a model-based manner this did not affect decision-making during the task and thus cannot reflect post-arbitration value.

There were no significant effects of attention manipulation on chosen shape transfer. This could be because the need to recall location was insufficient to direct attention. However, an alternative interpretation is that model-based transfer does not require attention. In support of this, digit span tests did not correlate with transfer. The major cost of model-based transfer is likely the search over possible actions within the environmental model, rather than the value and RPE generated against it, which is equally trivial for model-free and model-based RPEs. Since application of the model did not show a correlation with WM and attention, it is perhaps unsurprising that the generation of the model-based RPE, presumed to be less costly, is also not correlated with WM and attention. Given this, it could be asked to what extent transfer requires awareness of model-based values. Therefore, a second experiment was designed to test whether model-based transfer arises without explicit model instructions.

**Experiment 2: No model-based learning with implicit rules**

**Method 2**

A second cohort (n=15) aged 18-40 was tested without explicit knowledge of the relationship between stage-one shape colour and transitions to stage-two shapes. They were instructed that the stage-one coloured shape they chose would affect the stage-two shapes shown, but not which feature lead to the transitions. Stimuli, design and analysis were as per experiment 1.

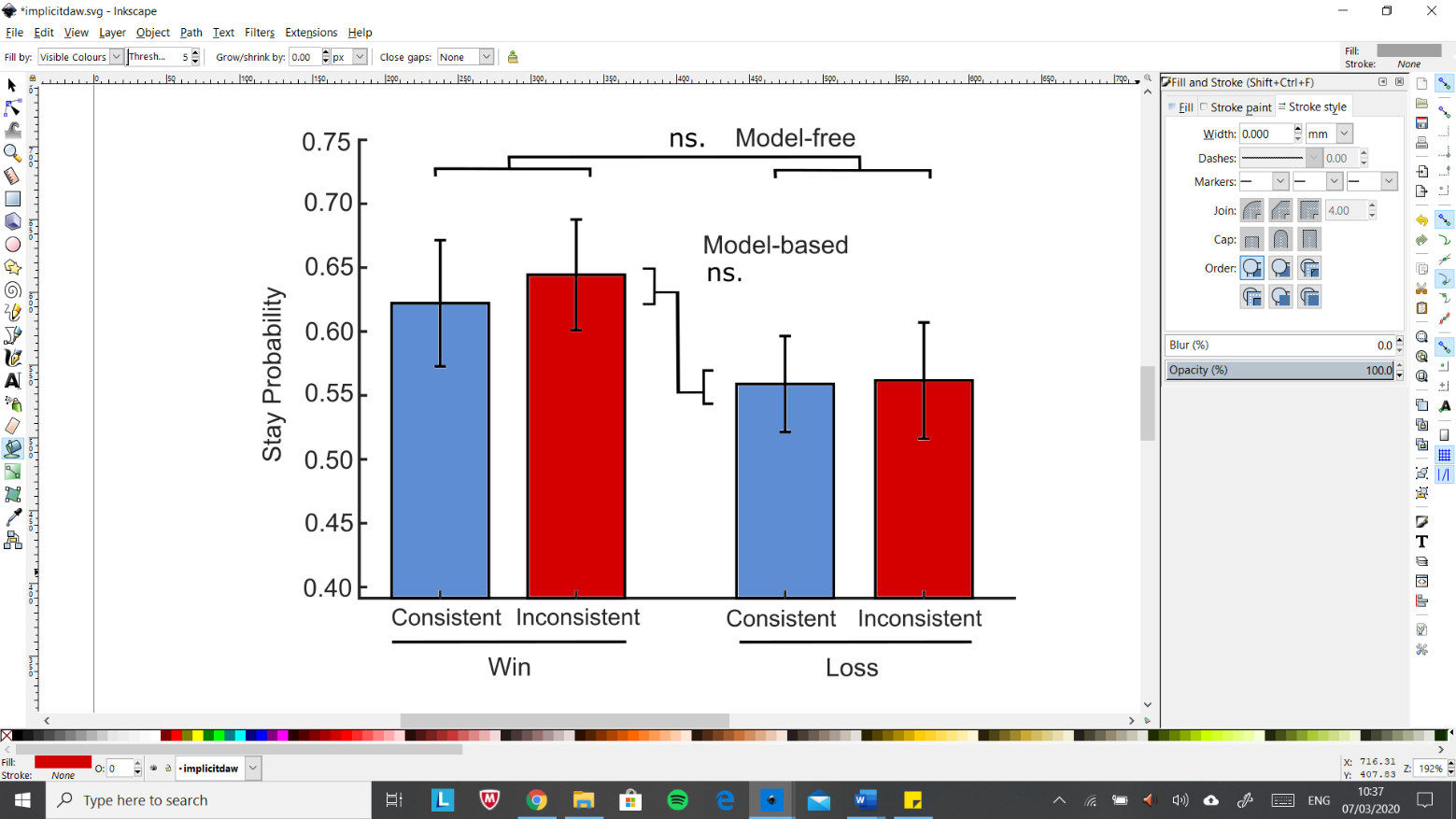
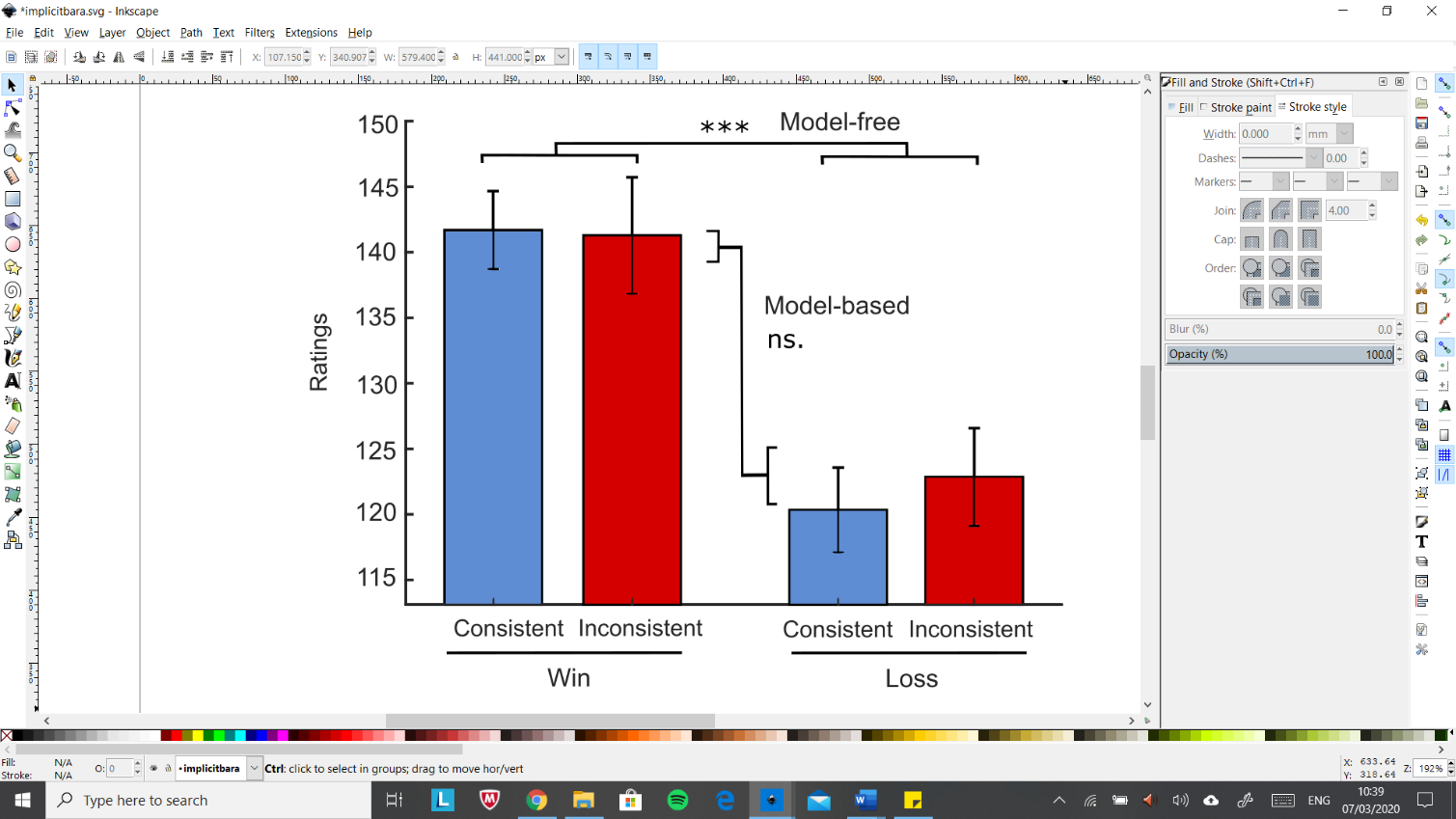
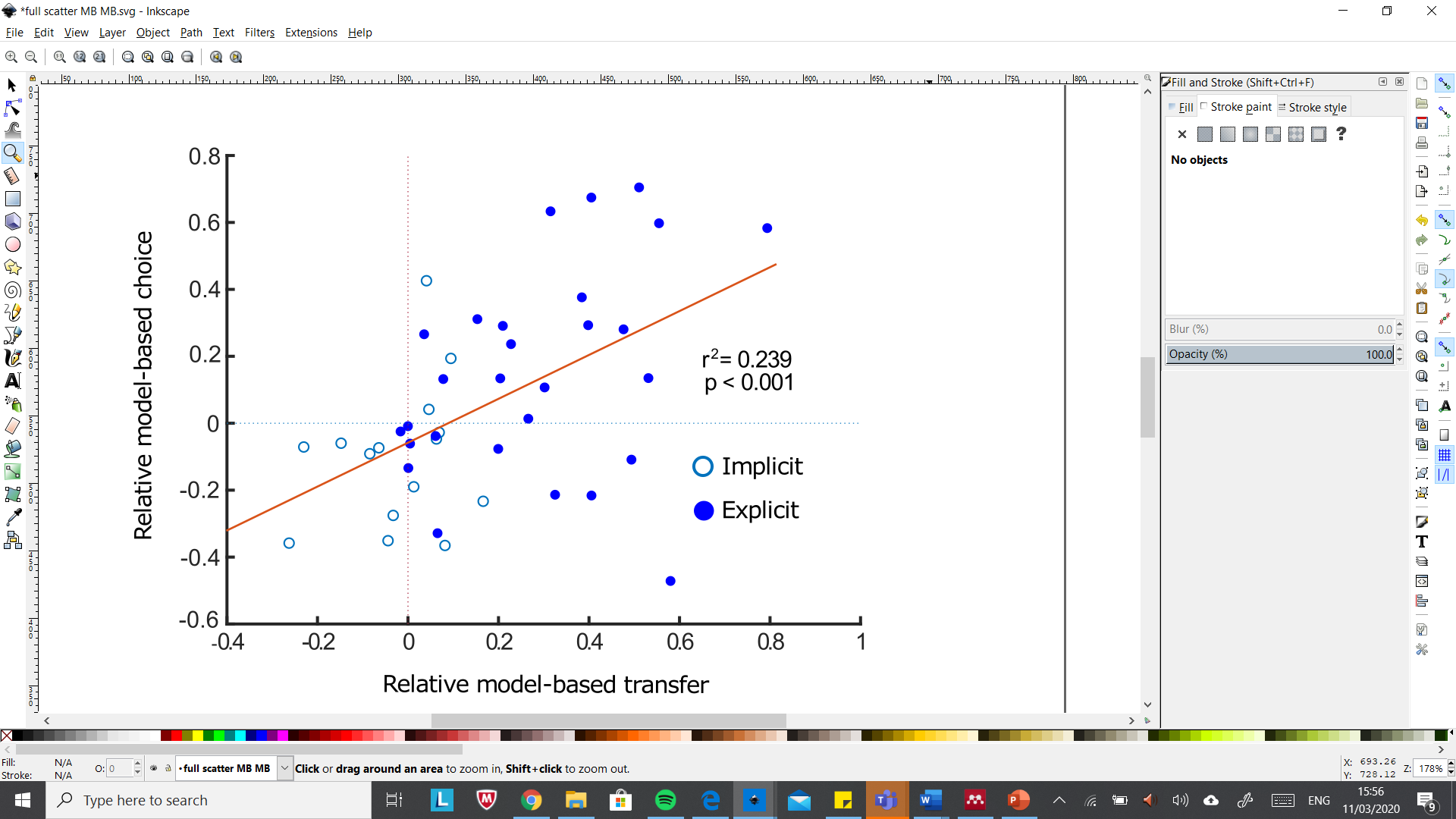
**Results 2**

In the implicit cohort only 6 participants were model-based. Because this revealed no significant effects analysis was expanded to include those with *ChoiceMB*<0. In these 15 participants no significant model-based or model-free pattern of choice was observed in the stay probability graphs (**fig.3A**). Assessment of ratings revealed no significant evidence of model-based transfer, however, a highly significant pattern of model-free transfer was observed (**fig.3B**). Two-way ANOVA showed subjects rated shapes more highly if a reward had been received the last time the shape was chosen (F(1, 20) = 18.0, p<.001). The model-based influence on ratings was weakly negative, *RatingsMB*(-9.88±5.56%, mean±SEM), but strong for model-free influence on ratings, *RatingsMF* (39.47±13.9%, mean±SEM). To see if participants used an alternative strategy, we examined the probability of participants returning to the side of the screen chosen in the first choice of the previous trial after a win as compared to the same probability after a loss. A significant difference in was observed (ANOVA, F(1, 32) = 7.516, p=.01) (supplementary figure 3) indicating a model-free location-learning strategy. This was not observed in the explicit group.

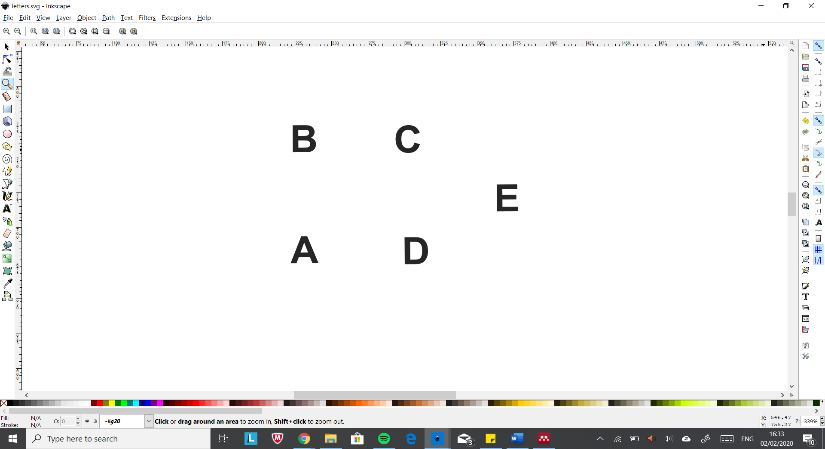
Across both groups, model-based choice correlated with the degree of model-based transfer to ratings (r=0.3985, p<0.01) (**fig.3C**). This finding was not replicated for model-free choice and transfer (supplementary figure 4), nor within the groups individually.

**Discussion 2**

The implicit task failed to produce accurate model-based learning and instead, participants formed spurious models of the task structure. During de-briefing, it became clear that many participants had invented complex rules such as odd-sided stage-one shapes being more rewarding. Since this manipulation abolished accurate model-based learning, no model-based transfer would be expected. It may be possible to produce model-based behaviour in implicit conditions, however the design of this task strongly encourages participants to look for stage-one *shape*-related effects on reward, unless they are explicitly told that colour determines the transitions.



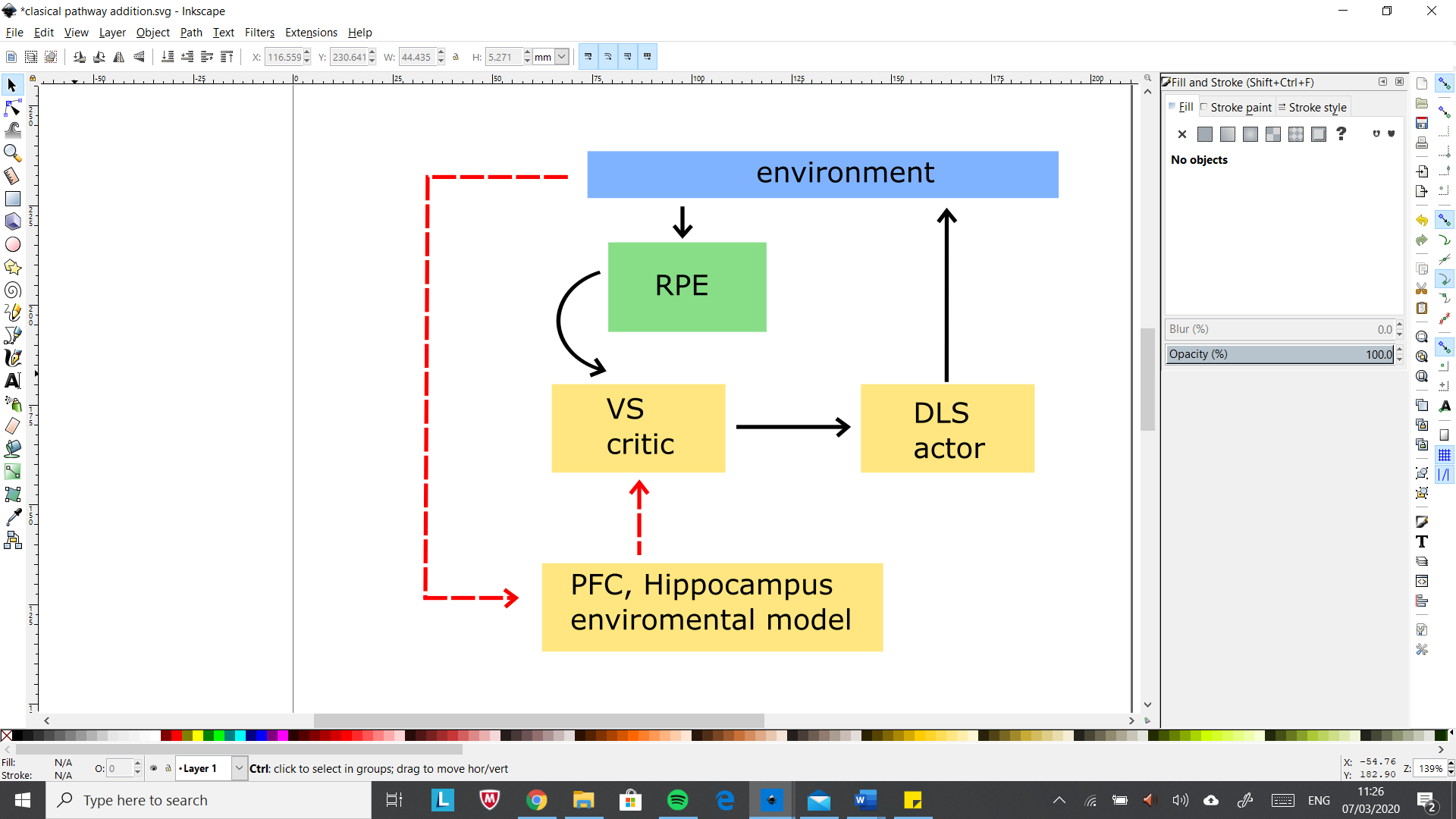
**Fig. 3** | **Subjects with explicit instructions show poor learning but strong model-free transfer.** **A:** No significant model-based or model-free choice. **B:** Shape ratings revealed significant model-free transfer but no model-based transfer. **C:** Correlation between model-based choice and transfer in implicit and explicit healthy groups including those with *ChoiceMB*<0. Error bars represent within-subject standard error of the mean.



**General Discussion**

There are several reasons to think value transfer may be advantageous. Sambrook et al.22 suggested that while the performance of a model-free system influenced by model-based information would still exhibit worse performance than a fully model-based approach, if the latter is interrupted due to increased cognitive load this extended model-free system could outperform simple RL. Recent evidence suggests model-based information can affect model-free learning rates[[33]](#endnote-34), likely by reactivation of model-free valuations during retrospective model-based inferences. Whilst this stops short of demonstrating value transfer, it supports the notion that model-based training of the model-free system may be of utility. In the real world many tasks that require knowledge of environmental structure may be split up into a number of subroutines involving simple model-free associations. Transfer could provide a mechanism by which a goal-based strategy composed of internal models exerts control over the model-free system[[34]](#endnote-35).

Actor-critic computational architectures[[35]](#endnote-36) for model-free learning are composed of three components: an actor which encodes the action policy, a critic which holds mappings between cues and their predicted value, and a learning signal. Structurally, the VS may hold value estimates of model-free predictions (the critic)12,40 that are trained by dopaminergic RPEs to increase accuracy via D1 and D2 receptor-mediated plasticity at medium-spiny neurons in the VS[[36]](#endnote-37). The value estimates made by the VS then inform the DLS (the actor)12 to guide actions.

These classic actor-critic architectures fail to incorporate various findings including those presented here. The VS has been shown to respond to various model-based properties due to input from areas including the PFC[[37]](#endnote-38) and hippocampus[[38]](#endnote-39). VS neurons have been shown to respond in a different way to stimuli that predict equally valued rewards that are sensorially distinct, which is unexpected given model-free system operates solely by reward magnitude41. Whilst it is clear the VS subserves multiple functions, the presence of model-based information in the VS supports the ‘multiplexing hypothesis’2, which argues that in addition to the model-free RPE, dopamine also encodes model-based reward error generated against internal models. These compound RPEs may then produce complex updates of ‘model-free’ stimulus value. The inhibitory projections from the VS to the ventral tegmental area, or direct cortical input, may support the production of a mixed model-based/model-free RPE which relays back to the VS to update value, incorporating the previously unaccounted for finding of Daw et al.11 (**fig.5**). Whilst we cannot rule out the possibility that the model-based component of the RPE serves another purpose, a signature essentially identical would likely be required for the results shown by this experiment.

**Fig. 5** |Schematic of traditional actor-critic framework for simple model-free learning with potential additions for model-based transfer marked by red arrows.

When knowledge of the task structure was poor (as in the implicit-cohort), reward influenced model-free value more strongly than when task structure was known (as in the explicit-cohort). Furthermore, the side on which the stage-one shapes were presented biased choice in the implicit but not the explicit group. This spatial bias has previously been observed in participants who display poor model-based choice[[39]](#endnote-40). An open question is whether this knowledge of task structure-dependent difference reflects competition between potentially-valued task dimensions during learning, during which all task features initially acquire value, before the outcome-relevant features (consistency) come to dominate outcome-irrelevant features as observed in the explicit group. Whilst in keeping with evidence that during habit formation, neurons in the DLS shift from encoding several task-relevant aspects to only the most relevant features over time[[40]](#endnote-41), further work is needed to uncover the factors which determine feature relevance. I intend to fit a reinforcement-learning model to estimate the contribution of trial-wise model-free and model-based prediction errors to the ratings. Given the variable interval between choosing and rating a stage-one shape this may uncover time-related transfer effects.

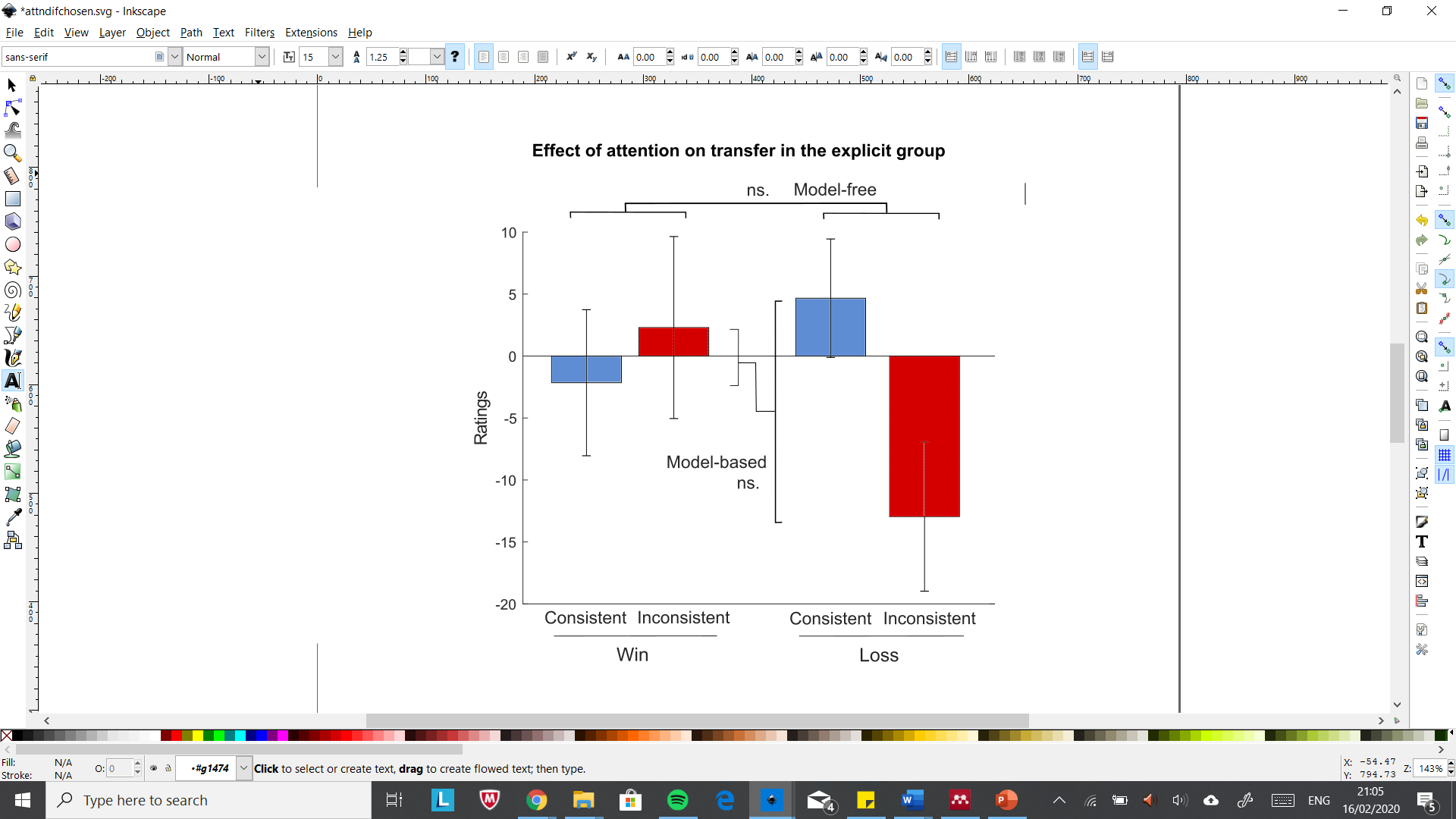
The finding in favour of transfer not only suggests a re-appraisal of the model-free/model-based dichotomy to include interactions, but also has widespread implications for accounts of emotions[[41]](#endnote-42) and morality[[42]](#endnote-43) that make use of this framework. Future work is necessary to support these findings, as well as assessing the timescale in which transfer persists and how far it can contribute to behaviour.

**Conclusion**

Results from healthy subjects provide evidence which supports the theory that model-based information influences value in model-free learning, probably utilising the model-based RPE observed previously11. This transfer to subjective ratings is likely to be independent of attention, however the attentional manipulation may not have been efficacious.

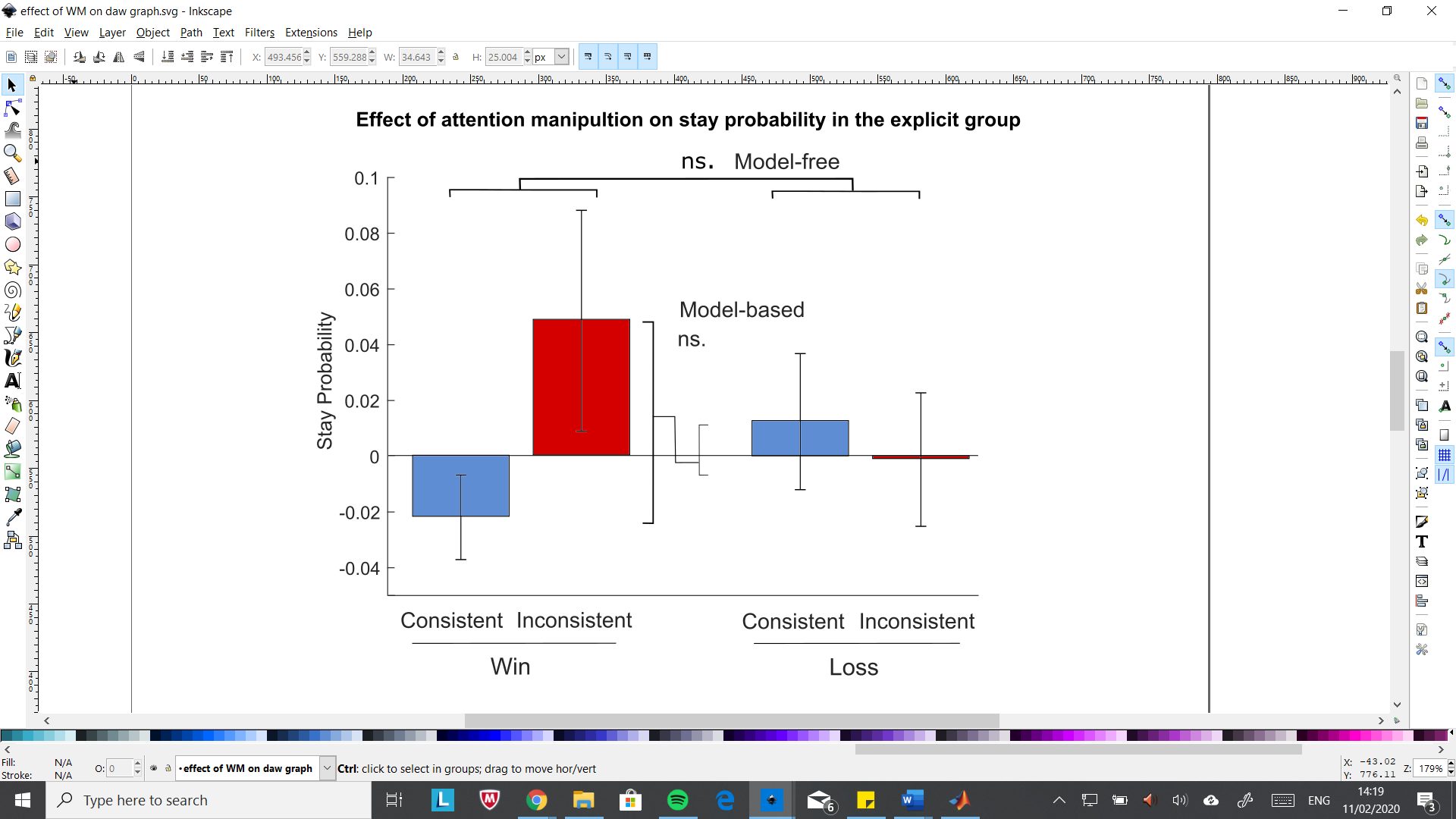
**Bibliography**

**Supplementary Figures**

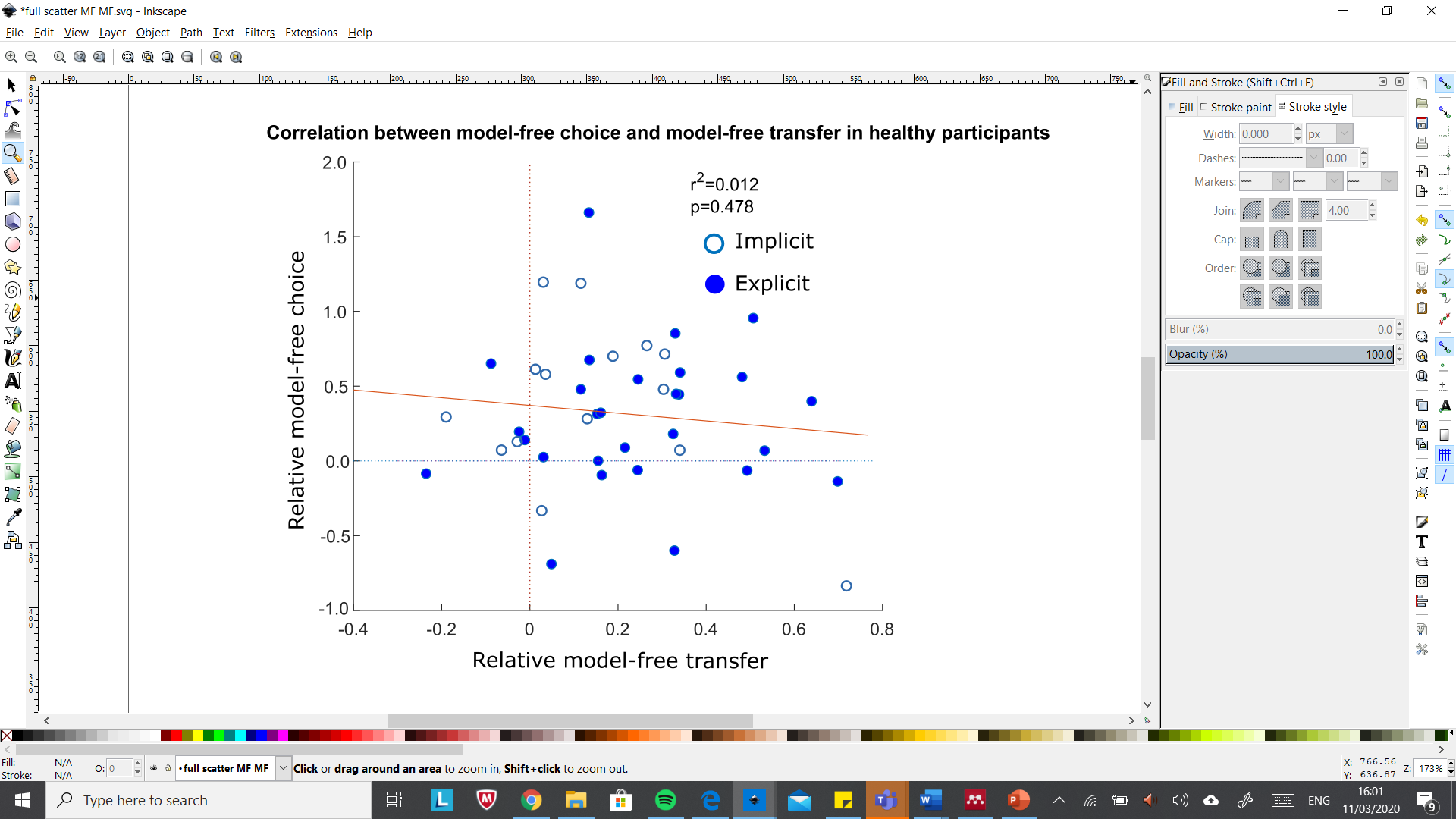


**Supp. Fig 1** | **Explicit group.** Attention had no significant effect on model-free or model-based transfer. Therefore, the inclusion of the attentional screen the last time a given shape was shown had no significant effect on the ratings when compared with trials in which the screen was not deployed. There is, however, a non-significant pattern of reduced model-based transfer.

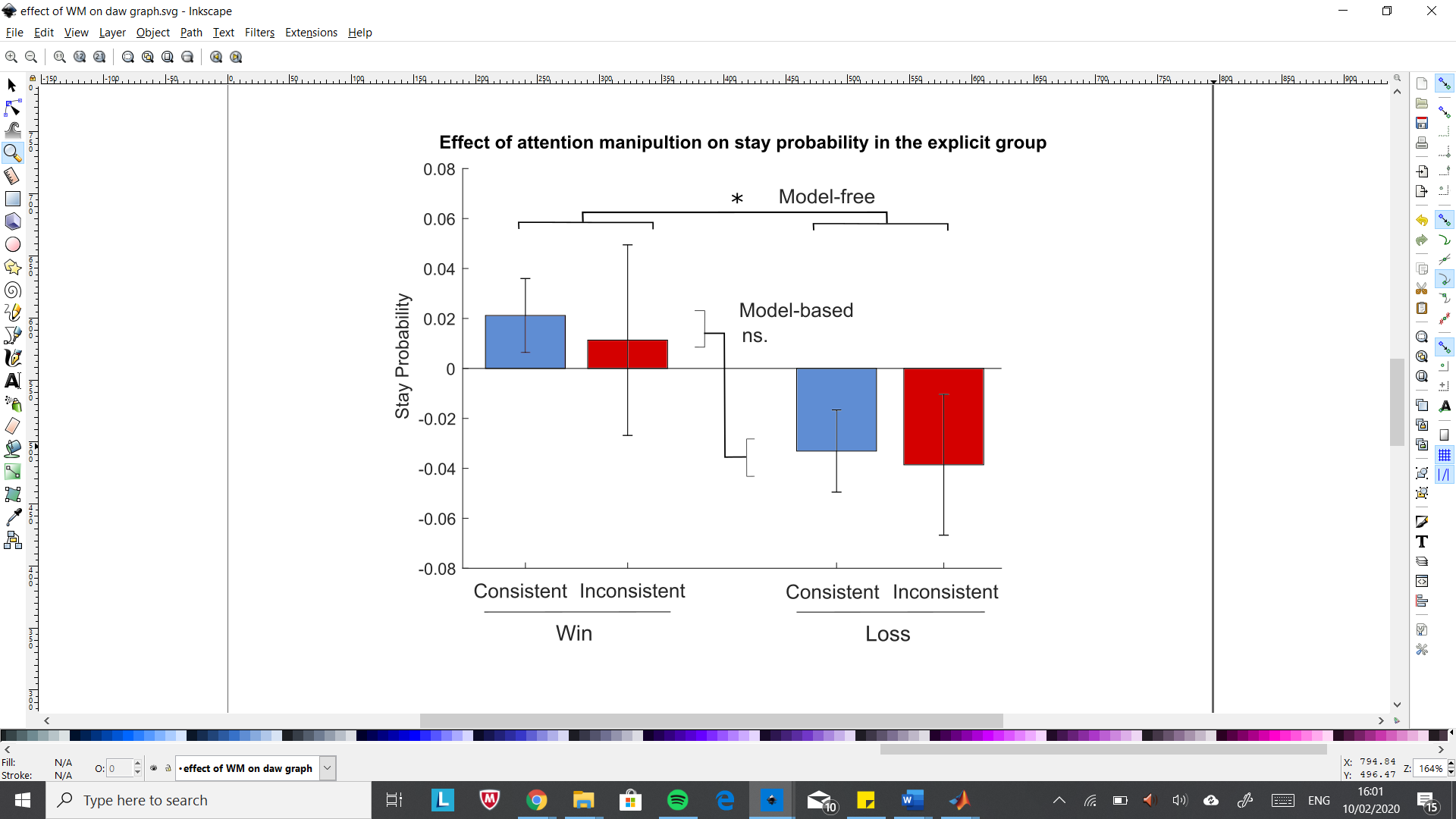
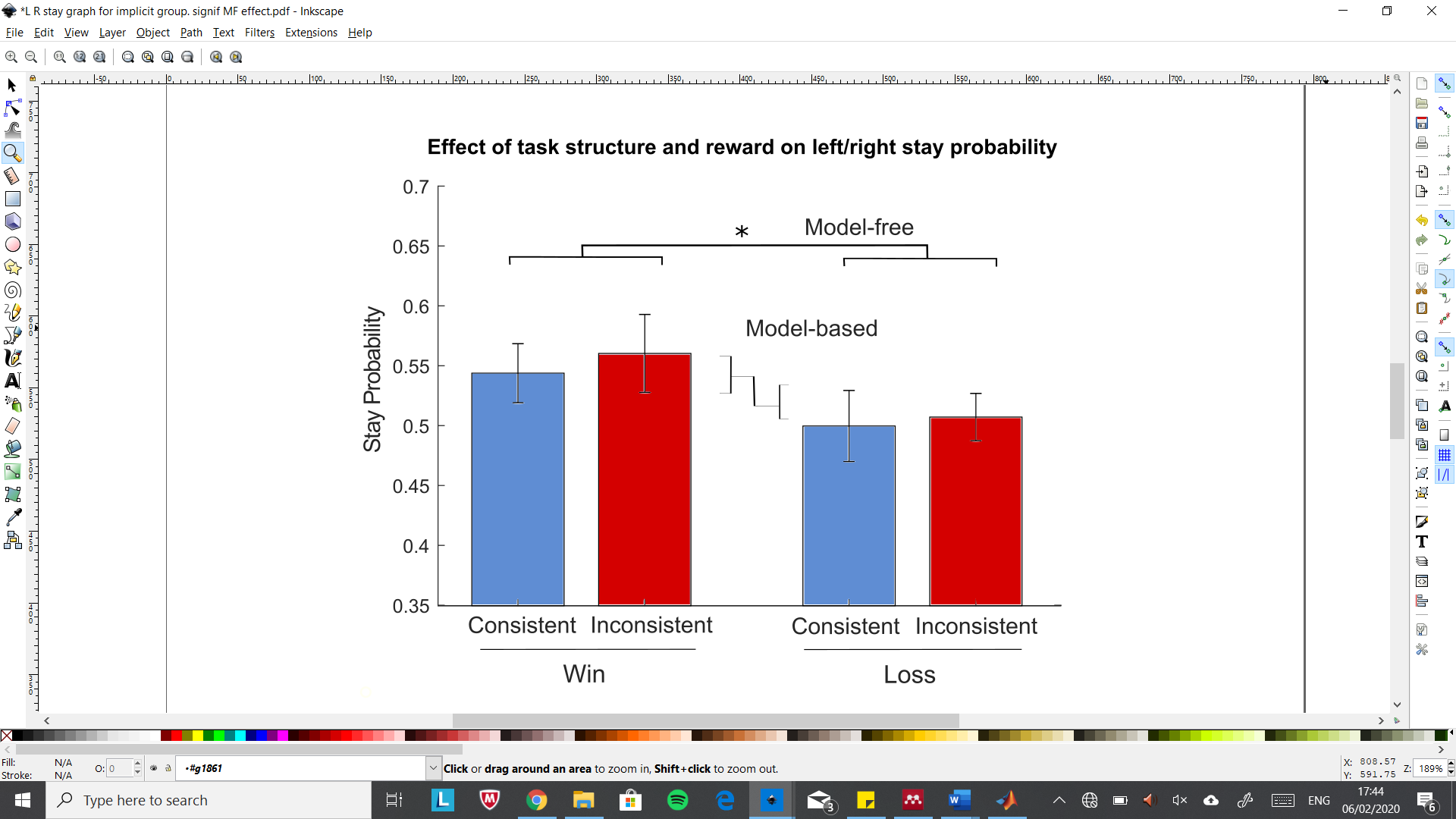
**Supp. Fig 2** | **Explicit group.** The attentional manipulation had no significant effect on stay probability. Therefore, for a given set of win/loss common /rare conditions on a trial, having the attention screen on that trial had no discernible effect on choice on the next trial in comparison to trials without the attention screen.



**Supp. Fig 3** | **Healthy cohort.** There was no significant correlation between model-free choice and model-free transfer.



**Supp. Fig 4** | **Implicit group.** The probability of returning to the left- or right-hand side of the screen on the next trial was impacted by reward, with participants more likely to return to a side if they had won. Consistency had no effect on the probability of returning to one side of the screen. The reward effect was not replicated in the explicit group.



1. Wunderlich, K., Dayan, P. & Dolan, R. J. Mapping value based planning and extensively trained choice in the human brain. *Nat. Neurosci*. **15**, 786–791 (2012). [↑](#endnote-ref-2)
2. Nakahara, H. Multiplexing signals in reinforcement learning with internal models and dopamine. *Current Opinion in Neurobiology* **25**, 123–129 (2014). [↑](#endnote-ref-3)
3. Dayan, P. & Niv, Y. Reinforcement learning: The good, the bad and the ugly*. Current Opinion in Neurobiology* **18**, 1–12 (2008). [↑](#endnote-ref-4)
4. Doll, B. B., Simon, D. A. & Daw, N. D. The ubiquity of model- based reinforcement learning. *Current Opinion in Neurobiology* **22**, 1075–1081 (2012). [↑](#endnote-ref-5)
5. Tolman, E. C. Cognitive maps in rats and men. *Psychological Review* **55**, 189–208 (1948). [↑](#endnote-ref-6)
6. Doya, K., Samejima, K., Katagiri, K. & Kawato, M. Multiple model- based reinforcement learning. *Neural Comput.* **14**, 1347–1369 (2002). [↑](#endnote-ref-7)
7. Knowlton, B. J., Mangels, J. A. & Squire, L. R. A neostriatal habit learning system in humans. *Science* **273**, 1399-1402 (1996). [↑](#endnote-ref-8)
8. Everitt, B. J. & Robbins, T. W. Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience* **8**, 1481-1489 (2005). [↑](#endnote-ref-9)
9. Cockburn, J., Collins, A. G. & Frank, M. J. A reinforcement learning mechanism responsible for the valuation of free choice. *Neuron* **83**, 551-557 (2014). [↑](#endnote-ref-10)
10. Glimcher, P. W. Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *PNAS***108**, 15647-15654 (2011). [↑](#endnote-ref-11)
11. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans’ choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011). [↑](#endnote-ref-12)
12. O’Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K. & Dolan, R. J. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454 (2004). [↑](#endnote-ref-13)
13. Balleine, B. W., Delgado, M. R. & Hikosaka, O. The role of the dorsal striatum in reward and decision-making. *J. Neurosci.* **27**, 8161–8165 (2007). [↑](#endnote-ref-14)
14. Jog, M. S., Kubota, Y., Connolly, C. I., Hillegaart, V. & Graybiel, A. M. Building neural representations of habits. *Science* **286**, 1745–1749 (1999). [↑](#endnote-ref-15)
15. Yin, H. H., Mulcare, S. P., Hilario, M. R., Clouse, E., Holloway, T., Davies, M. I., Hansson, A. C., Lovinger, D. M. & Costa, R. M. Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill*. Nat. Neurosci*. **12**, 333–341 (2009). [↑](#endnote-ref-16)
16. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci*. **8**, 1704–1711 (2005). [↑](#endnote-ref-17)
17. Hampton, A. N., Bossaerts, P. & O’Doherty, J. P. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.* **26**, 8360–8367 (2006). [↑](#endnote-ref-18)
18. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267–279 (2014). [↑](#endnote-ref-19)
19. Schuck, N. W., Cai, M. B., Wilson, R. C. & Niv, Y. Human orbitofrontal cortex represents a cognitive map of state space. *Neuron* **91**, 1402–1412 (2016). [↑](#endnote-ref-20)
20. Gläscher, J., Daw, N., Dayan, P. & O’Doherty, J. P. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010). [↑](#endnote-ref-21)
21. Balleine, B. W., O’Doherty & J. P. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology* **35**, 48-69 (2010). [↑](#endnote-ref-22)
22. Sambrook, T. D., Hardwick, B., Wills, A. J. & Goslin, J. Model-free and model-based reward prediction errors in EEG. *NeuroImage* **178**, 162–171 (2018). [↑](#endnote-ref-23)
23. Pagnoni, G., Zink, C. F., Montague, P. R. & Berns, G. S. Activity in human ventral striatum locked to errors of reward prediction. *Nat. Neurosci*. **5**, 97–98 (2002). [↑](#endnote-ref-24)
24. Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J. & Frith, C. D. Dopamine- dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* **442**, 1042–1045 (2006). [↑](#endnote-ref-25)
25. Otto, A. R., Gershman, S. J., Markman, A. B. & Daw, N. D. The Curse of Planning: Dissecting Multiple Reinforcement-Learning Systems by Taxing the Central Executive. *Psychological Science* ***24***, 751–761 (2013). [↑](#endnote-ref-26)
26. Schad, D. J., Jünger, E., Sebold, M., Garbusow, M., Bernhardt, N., Javadi, A. H., Zimmermann, U. S., Smolka, M. N., Heinz, A., Rapp, M. A. & Huys, Q. J. Processing speed enhances model-based over model-free reinforcement learning in the presence of high working memory functioning. *Front. Psychol.* **5**, 1450 (2014). [↑](#endnote-ref-27)
27. Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. & Dolan, R. J. Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron* **80**, 1–6 (2013). [↑](#endnote-ref-28)
28. Piray, P., Toni, I. & Cools, R. Human Choice Strategy Varies with Anatomical Projections from Ventromedial Prefrontal Cortex to Medial Striatum. *Journal of Neuroscience* **36**, 2857–2867 (2016). [↑](#endnote-ref-29)
29. Eppinger, B., Walter, M. & Li, S. C. Electrophysiological correlates reflect the integration of model-based and model-free decision information. *Cognitive, Affective and Behavioural Neuroscience* **17**, 406–421 (2017). [↑](#endnote-ref-30)
30. Dickinson, A. Actions and habits: the development of behavioural autonomy. *Philos. Trans. R. Soc. Lond. B Biol. Sci*. **308**, 67–78 (1985). [↑](#endnote-ref-31)
31. Da Silva, C. F. & Hare, T. A. A note on the analysis of two-stage task results: How changes in task structure affect what model-free and model-based strategies predict about the effects of reward and transition on the stay probability. *PLoS ONE*, **13**, 1–13 (2018). [↑](#endnote-ref-32)
32. Siegel, J. Z., Mathys, C., Rutledge, R. B. & Crockett, M. J. Beliefs about bad people are volatile. *Nature Human Behaviour*, **2**, 750-756 (2018). [↑](#endnote-ref-33)
33. Moran, R. & Dolan, R. J. Retrospective model-based inference guides model-free credit assignment. *Nature Communications* **10**, 750 (2019). [↑](#endnote-ref-34)
34. Rusu, S. I. & Pennartz, C. M. A. Learning , memory and consolidation mechanisms for behavioral control in hierarchically organized cortico-basal ganglia systems. *Hippocampus* **30**, 73-98(2020). [↑](#endnote-ref-35)
35. Barto, A. G., Sutton, R. S. &. Anderson, C. W. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics* **13**, 835-846 (1983). [↑](#endnote-ref-36)
36. Gerfen, C. R. & Surmeier, D. J. Modulation of striatal projection systems by dopamine. *Annu. Rev. Neurosci*. **34**, 4414–4466 (2011). [↑](#endnote-ref-37)
37. Takahashi, Y. K., Roesch, M. R., Wilson, R. C., Toreson, K., O'Donnell, P., Niv, Y. & Schoenbaum, G. Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. *Nat. Neurosci*. **14**, 1590-1597 (2011). [↑](#endnote-ref-38)
38. Ito, R., Robbins, T. W., Pennartz, C. M. & Everitt, B. J. Functional interaction between the hippocampus and nucleus accumbens shell is necessary for the acquisition of appetitive spatial context conditioning. *The* *Journal of Neuroscience* **28**, 6950–6959 (2008). [↑](#endnote-ref-39)
39. Shahar, N., Moran, R., Hauser, T. U., Kievit, R. A., McNamee, D., Moutoussis, M. & Dolan, R. J. Credit assignment to state-independent task representations and its relationship with model-based decision making. *Proc. Natl. Acad. Sci.* **116**, 15871–15876 (2019). [↑](#endnote-ref-40)
40. Sharpe, M. J., Stalnacker, T., Schuck, N. W., Killcross, S., Geoffrey, S. & Niv, Y. An Integrated Model of Action Selection : Distinct Modes of Cortical Control of Striatal Decision Making. *Annaul Review of Psychology* **70**, 53-76 (2019).                          [↑](#endnote-ref-41)
41. Huys, Q. J. M. & Renz, D. Review A Formal Valuation Framework for Emotions and Their Control*. Biological Psychiatry* **82**, 413–420 (2017). [↑](#endnote-ref-42)
42. Crockett, M. J. Models of morality. *Trends in Cognitive Sciences* **17**, 363-366. (2013). [↑](#endnote-ref-43)