Appendix

This Appendix explain the search algorithms in more detail. These equations use the same Kalman Filtering equations used in Gershman (2018, see also: Bishop, 2006; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Pearson, Hayden, Raghavachari, & Platt, 2009). Table A1 shows the data from an individual participant, and how V, RU, TU, and V/TU were calculated.

The choice can be either the optimal (O) or suboptimal (S). *r* is the outcome (pain) on a given trial. Q is the posterior means of the pain for the optimal and suboptimal options. σ^2 is the posterior variances for each of the options. α is the learning rate parameter, also known as gain, for the two options. As already explained, $V = Q_s - Q_o$, $RU = \sigma_s - \sigma_o$, and $TU = \text{sqrt}(\sigma_s^2 + \sigma_o^2)$.

To calculate, V, RU, and TU for each trial, the following three equations are run iteratively. The updating equations are run separately for the optimal and suboptimal choices, and are only run when a given option was chosen; if it was not chosen then the numbers remain the same as the prior trial.

$$\alpha_{t+1} = \frac{\sigma_t^2}{\sigma_t^2 + \tau^2} \quad \text{EQ 1}$$

$$Q_{t+1} = Q_t + \alpha_t[p_t - Q_t] \quad \text{EQ 2}$$

$$\sigma_{t+1}^2 = \sigma_t^2 - \alpha_t \sigma_t^2 \quad \text{EQ 3}$$

Prior to the first choice, σ^2 starts at τ_0^2, which is the theoretical variance of the mean of the distributions from which the optimal and suboptimal choices were drawn from. τ^2 represents the prior variance of the pain experiences around the mean pain for each choice. For both τ_0^2 and τ^2, participants were not given any prior knowledge, so we set them to 20^2 to reflect a fairly high degree of uncertainty. This is why α starts at .5 on Trial 1. (the model was repeated with considerably higher (30^2) and lower (1^2) parameters, with all iterations converging on the same best fitting model). The *Q* for both choices were set to 34 prior to the first trial, which represents the prior mean of the pain; 34 is close to the posterior pain rating for the optimal and suboptimal options together.

After V, RU, V/TU, and PC are calculated, the search algorithms can make a choice for the subsequent trial. In Table A1, the choices were determined by a participant. EQ4 shows how the participants' choices can be modelled by the search algorithms. Specifically, EQ4 provides an equation showing the hybrid model that has all four components, V, V/TU, RU, and PC. Each one has its own regression coefficient. Simpler models with fewer components simply involve making these coefficients zero. EQ 4 provides the probability that the choice (action = *a*) at time *t* is O instead of S. Φ represents the cumulative distribution function of the Normal distribution, so choice is modelled as a probit regression.

$$P(a_t = O) = \Phi\left(B_0 + B_1 V + B_2 \frac{V}{TU} + B_3 RU + B_4 PC\right) \quad \text{EQ 4}$$

Table A1. Calculating V, RU, V/TU, and PC for an individual participant.

		Optimal Choice				Suboptimal Choice				Model Components				
Trial	Choice	*r*	α	Q	σ^2	*r*	α	Q	σ^2	V	RU	TU	V/TU	PC
0				34.00	400.00			34.00	400.00	0.00	0.00	28.28	0.00	
1	O	36	0.50	35.00	200.00		0.50	34.00	400.00	-1.00	-5.86	24.49	-0.04	1
2	S		0.33	35.00	200.00	39	0.50	36.50	200.00	1.50	0.00	20.00	0.08	0
3	O	34	0.33	34.67	133.33		0.33	36.50	200.00	1.83	-2.60	18.26	0.10	1
4	S		0.25	34.67	133.33	38	0.33	37.00	133.33	2.33	0.00	16.33	0.14	0
5	S		0.25	34.67	133.33	50	0.25	40.25	100.00	5.58	1.55	15.28	0.37	0
6	O	25	0.25	32.25	100.00		0.20	40.25	100.00	8.00	0.00	14.14	0.57	1
7	S		0.20	32.25	100.00	49	0.20	42.00	80.00	9.75	1.06	13.42	0.73	0
8	O	13	0.20	28.40	80.00		0.17	42.00	80.00	13.60	0.00	12.65	1.08	1
9	O	13	0.17	25.83	66.67		0.17	42.00	80.00	16.17	-0.78	12.11	1.33	1
10	O	9	0.14	23.43	57.14		0.17	42.00	80.00	18.57	-1.38	11.71	1.59	1
11	O	6	0.13	21.25	50.00		0.17	42.00	80.00	20.75	-1.87	11.40	1.82	1
12	S		0.11	21.25	50.00	35	0.17	40.83	66.67	19.58	-1.09	10.80	1.81	0
13	O	43	0.11	23.67	44.44		0.14	40.83	66.67	17.17	-1.50	10.54	1.63	1
14	O	42	0.10	25.50	40.00		0.14	40.83	66.67	15.33	-1.84	10.33	1.48	1
15	S		0.09	25.50	40.00	65	0.14	44.29	57.14	18.79	-1.23	9.86	1.91	0
16	O	35	0.09	26.36	36.36		0.13	44.29	57.14	17.92	-1.53	9.67	1.85	1
17	S		0.08	26.36	36.36	55	0.13	45.63	50.00	19.26	-1.04	9.29	2.07	0
18	O	44	0.08	27.83	33.33		0.11	45.63	50.00	17.79	-1.30	9.13	1.95	1
19	O	17	0.08	27.00	30.77		0.11	45.63	50.00	18.63	-1.52	8.99	2.07	1
20	O	29	0.07	27.14	28.57		0.11	45.63	50.00	18.48	-1.73	8.86	2.09	1
21	O	45	0.07	28.33	26.67		0.11	45.63	50.00	17.29	-1.91	8.76	1.97	1
22	S		0.06	28.33	26.67	73	0.11	48.67	44.44	20.33	-1.50	8.43	2.41	0
23	O	39	0.06	29.00	25.00		0.10	48.67	44.44	19.67	-1.67	8.33	2.36	1
24	O	20	0.06	28.46	23.53		0.10	48.67	44.44	20.21	-1.81	8.24	2.45	1

Note. Sample data from Participant 7. O = optimal choice, and S = suboptimal choice. For this participant, O always resulted in a shock of 50% and S resulted in a shock of 60%.

As can be seen in Table A1, the learning rate α decreases each time that an option is chosen. The posteriors of the two options diverge over time; Q for the optimal choice ends up being lower than Q for the suboptimal choice since the pain is higher for the suboptimal choice. σ^2 decreases each time an option is chosen to reflect increasing certainty about that option.

With regards to the model components (V, RU, and V/TU, and PC), they are coded in a way so that a higher (more positive) number predicts a stronger choice in favor of the optimal choice. V (difference in posterior values of the two choices) initially starts at 0, reflecting the lack of knowledge of which choice produces lower pain. V fairly quickly becomes positive, which reflects that the participants' pain ratings associated with the suboptimal option tend to be about 20 points higher than the pain for the optimal option. Higher numbers for V represent a stronger preference for O in Thompson sampling. This is how making a decision partially based on V leads to exploiting the better option.

RU (relative uncertainty) initially starts at 0 reflecting that both options have been chosen the same number of times. After the optimal choice is selected on Trial 1, RU goes down to -5.86, which reflects that something is now known about O but not about S; a negative value suggests choosing S. After Trial 2 in which S was selected, both options have been selected once so RU goes back to 0. There are three important things to note for RU. First, it is positive if S has

been selected more times than O, negative if the reverse, and 0 if they have been selected equally. Second, at the beginning there are some trials for which the absolute value of RU is fairly large; the difference in relative uncertainty having selected O one time and S zero times (after Trial 1; RU=-5.86) is bigger than the difference between selecting O 5 times and S 4 times (after Trial 9; RU=-.79). This is why an agent guided only by RU would be somewhat similar to the alternating strategy – such an agent would alternate if one option was chosen 1 more time than the other, but would pick randomly if the two options were chosen the same number of times. Third, over time RU tends to become negative. The reason for this is that exploiting O results in more being known about O than S. This means that whereas RU drives the choice towards S for continued exploring, V drives the choice to O for exploiting. However, when V is large it is typically able to override RU.

TU decreases with each additional trial, regardless of which option is chosen. It decreases faster at the beginning when nothing is known about either option. (Note, unlike the other model components, TU is always positive because it reflects total uncertainty about both options, not relative uncertainty comparing the two.) V/TU over time gradually increases to reflect the preference for O in two senses: O is believed to be better than S, and over time more experiences have been accumulated to make this comparison with higher confidence.

PC (prior choice) is 1 if the prior choice was O, and 0 if the prior choice was S. Alternation predicts a negative influence of prior choice, so if the prior choice was O the next choice would tend to favor S, and vice versa.

References

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441*(7095), 876-879. doi:https://doi.org/10.1038/nature04766

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition, 173*, 34-42. doi:https://doi.org/10.1016/j.cognition.2017.12.014

Pearson, J. M., Hayden, B. Y., Raghavachari, S., & Platt, M. L. (2009). Neurons in Posterior Cingulate Cortex Signal Exploratory Decisions in a Dynamic Multioption Choice Task. *Current Biology, 19*(18), 1532-1537. doi:https://doi.org/10.1016/j.cub.2009.07.048