We formalized the learning problem that participants face in our experiments as a form of Bayesian concept learning (Tenenbaum, 1999; Goodman, 2006), represented graphically in Fig. X. The goal is to learn a concept *theta*, which is a set of probabilities for independent binary features $\theta_{1,2,..,n}$, where n is the number of features. Over the course of a block, the learner receives information about θ by observing exemplars y: instantiations of $\bar{\theta}$, where each feature $y_{1,2,..,n}$ is either on or off. Each feature θ_i and its corresponding exemplar y_i form a Beta-Bernoulli process:

$$p(\theta_i) \sim Beta(\alpha_i, \beta_i) \tag{1}$$

$$p(y_i|\theta_i) \sim Bernoulli(\theta_i) \tag{2}$$

Since the features are independent, this relationship holds for the entire concept θ. However, to model the time course of attention, we do not want to assume that information is encoded perfectly and instantaneously. Instead, we suggest that participants gather repeated noisy samples $\bar{z}$ from the exemplars. For any sample z from an exemplar y there is a small probability ϵ to misperceive the feature as off when it was actually on, and vice versa. Therefore, by making noisy observations $\bar{z}$, the learner obtains information about the true identity of the exemplar y, and by extension, about the concept $\bar{theta}$. By Bayes' rule:

$$P(\theta|\bar{z}) \quad = p(\bar{z}|y)p(y|\theta)p(\theta)/p(\bar{z}) \tag{3}$$

where $p(\bar{z}|y_i)$ is fully described by ϵ, and $p(y|\theta)$ by Bernoulli processes as in Eq. 2.

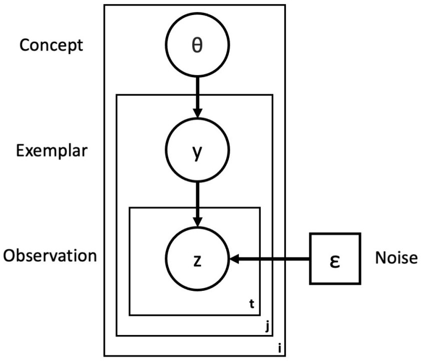

Figure 1: Graphical representation of our model. Circles indicate random variables. The squares indicate fixed model parameters.

Like in our experiment, the learner's task is to decide when to stop sampling. If they do so rationally, then they should anchor their sampling behavior to the expected information gain (EIG) of the next sample. We compute EIG by weighing the information gain from each possible next observation by the probability of that observation. We defined information gain as the KL-divergence between the hypothetical posterior after observing a sample z_{t+1} and the current posterior:

$$EIG(z_{t+1}) = \sum_{z_{t+1}\in[0,1]} p(z_{t+1}|\theta_t) * KL(\theta_{t+1}, p(\theta_t)) \tag{4}$$

Finally, to get actual sampling behavior from the model, it has to convert EIG into a binary decision about whether continue looking at the current sample, or to advance to the next trial. The model does so using a luce choice between the EIG from the next sample and a constant EIG from looking away.

$$p(look) = \frac{EIG(z_{t+1})}{EIG(z_{t+1}) + EIG(world)} \tag{5}$$

We also studied the behavior of the model when replacing EIG with other linking hypotheses, such as surprisal (the probability of a given z under the $P(\theta_t)$) and KL-divergence between the posterior $p(\theta_t)$ and the prior $p(\theta_{t-1})$.