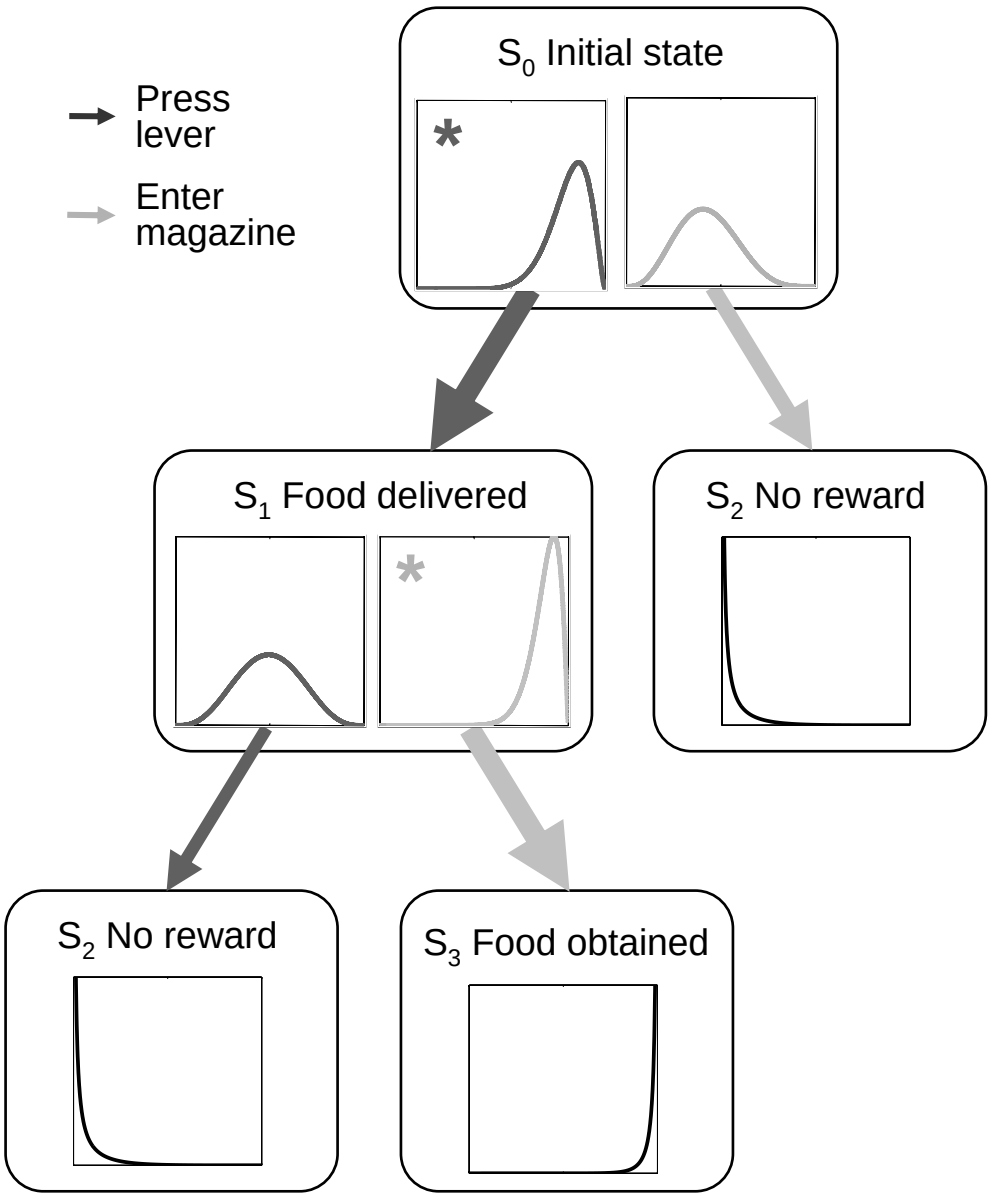

Supplementary Figure 1

Value propagation in tree search, after 50 steps of learning the task in **Figure 1a**. The inset plots show the distributions over state-action values, $\mathbf{Q}^{\text{tree}}_{s,a}$, computed by the tree system from the learned distributions over the values of the terminal states (shown in black) and over the transition structure of the task. The distributions are plotted as the probability assigned to each possible value q. Their moments are given by iteration on Equations 5 and 6 in **Supplementary Methods** — each value distribution is a function of the value distributions for the best action (marked with an asterisk) at each possible successor state. Arrows represent the most likely transition for each state and action, and their widths are proportional to the likelihoods (the full set of mean transition probabilities is illustrated in **Fig. 4**). The better actions were better explored and hence more certain (narrower value distributions); distributions at each state were similar to the distributions at the most likely successor state, and more so when transition to that state was more likely. As iterations progressed backwards, distributions got broader.