L1: Initial Belief Distribution **Beliefs** Deceptive Region (v<0, a>0) Alignment Signal (a) True Value (v)