Appendix I
Strategy-Critical States

Previous work has looked at generating explanations for a human of when an autonomous agent thinks it's very important to take a particular action to not fail at the task as a series of states, called *critical states*. Given a stochastic policy π, the original paper searches the set of critical states $\mathcal{C}_\pi$ where the agent prefers a small set of actions over all others:

$$\mathcal{C}_\pi = \{s \mid \mathcal{H}(\pi(\cdot \mid s)) < t\} \tag{8}$$

where $\mathcal{H}(\pi(\cdot \mid s))$ is the entropy of the policy's action distribution at state s and $t \in \mathbb{R}$ is a threshold for criticality.

We modify their approach to try and generate explanations for when the agent thinks it's very important to take particular actions to end up taking particular strategies. In particular, we consider finding the set of *strtategy-critical states* $\mathcal{SC}_\pi$ where the agent prefers taking a small set of *different actions* for *each strategy*:

$$\mathcal{SC}_\pi = \{s \mid \frac{1}{k^2} \sum_{z_i^*, z_j^*} D_{KL}(\pi(\cdot|s, z_i)) \parallel \pi(\cdot|s, z_j) > t\}. \tag{9}$$

where D_{KL} is the KL-divergence. This method will search for states where the average KL-divergence between the action distributions for the policy under each potential strategy cluster are sufficiently large.

Using this approach, we find that strategy-critical states are branching points between strategies, i.e. the last possible common state between different strategies. Most strategy-critical states show the two agents just before committing to a particular strategy. We use these strategy-critical states to determine the initial strategy landmark when generating the full explanation.

Appendix II
Textual Explanation Generation

Game Description Prompt:

A two-player game in a 10 by 10 grid. Row 0, column 0 is the top left corner. This is the starting state of the game.

- H represents Player 1 at row 3, column 4
- R represents Player 2 at row 7, column 6
- J1 represents the jewel at row 3, column 9
- J2 represents the jewel at row 7, column 2
- B1 represents the button at row 2, column 5
- B2 represents the button at row 9, column 7
- E1 represents the exit at row 2, column 2
- E2 represents the exit at row 9, column 9
- D1 represents the door at row 4, column 9
- D2 represents the door at row 6, column 2

Here are the rules of the game. There is a jewel at row 3, column 9, but it is blocked by a door at row 4, column 9. There is a second jewel at row 7, column 2, but it's blocked by a door at row 6, column 2. In order to open the door at row 4, column 9, one of the two players must stand at the location of a button, which is located at row 2, column 5. When one of the players stands on the button, the door will open, and the other player can collect the jewel. In order to open the door at row 6, column 2, one of the two players must stand at the location of a second button, which is located at row 9, column 7. Additionally, once a player has collected a jewel, that player cannot collect the other jewel. The player who has already collected a jewel must help the other player to collect the other jewel. Once both jewels have been collected, one player needs to move to the location of an exit, located at row 2, column 2. And the other player needs to move to the location of a second exit, located at row 9, column 9. When both players are located at each of the exits simultaneously, the game ends.

Landmark Description Prompt: State B is the following. In State B, Player R is located at row 8, column 6. Player H is located at row 3, column 4. Next we will describe a different state, State C. In State C, Player H is at the location of the button at row 2, column 5. Player R is at row 4, column 9, and is holding a jewel. Describe succinctly (¡15 words) and intuitively in present tense what happened to get from State B to State C.

ChatGPT Response: H moves up to Upper button; R moves up and right to collect Upper jewel.