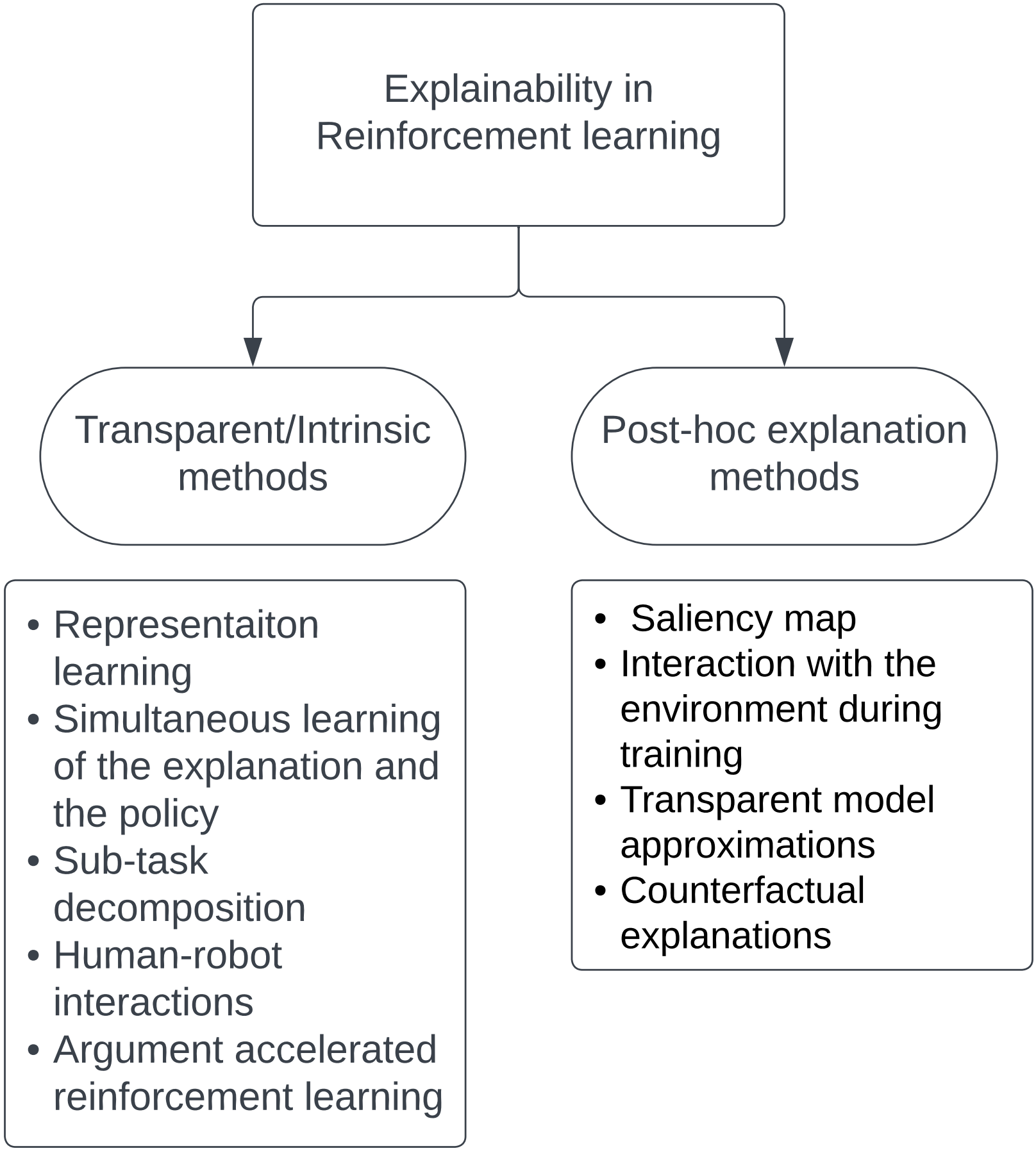
Explainability in Reinforcement learning
Transparent/Intrinsic methods
Post-hoc explanation methods
• Representaiton learning
• Simultaneous learning of the explanation and the policy
• Sub-task decomposition
• Human-robot interactions
• Argument accelerated reinforcement learning
• Saliency map
• Interaction with the environment during training
• Transparent model approximations
• Counterfactual explanations