Interactive Explanations: Diagnosis and Repair of Reinforcement Learning Based Agent Behaviors

Christian Arzate Cruz and Takeo Igarashi

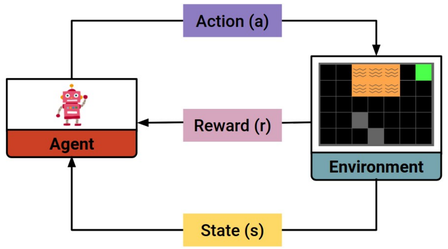
東京大学 THE UNIVERSITY OF TOKYO

[2021] IEEE Conference on Games

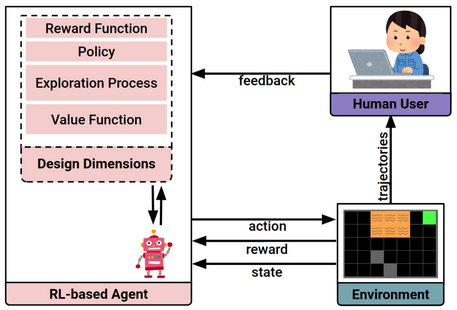
Motivation Making artificial intelligence (AI) models more accessible and easy to fix or personalize through human-computer interaction (HCI) techniques.

What are we doing? A novel interaction method that uses interactive explanations.

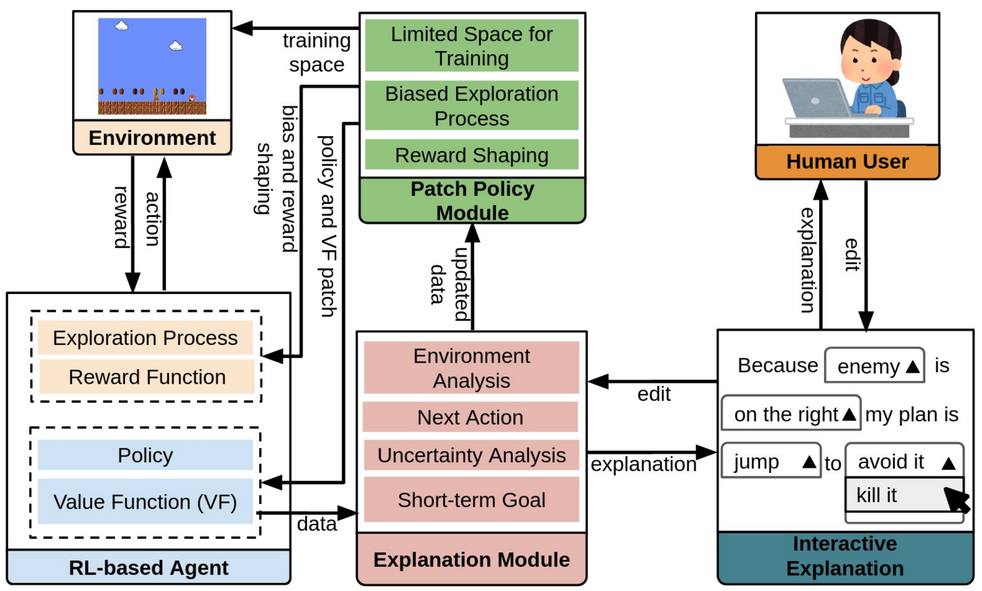
How are we doing? We use an interactive reinforcement learning (RL) method that exposes the thinking procedure of the learning agent. In this setting, users can align the bot's behavior to fit their intentions by creating policy patches.



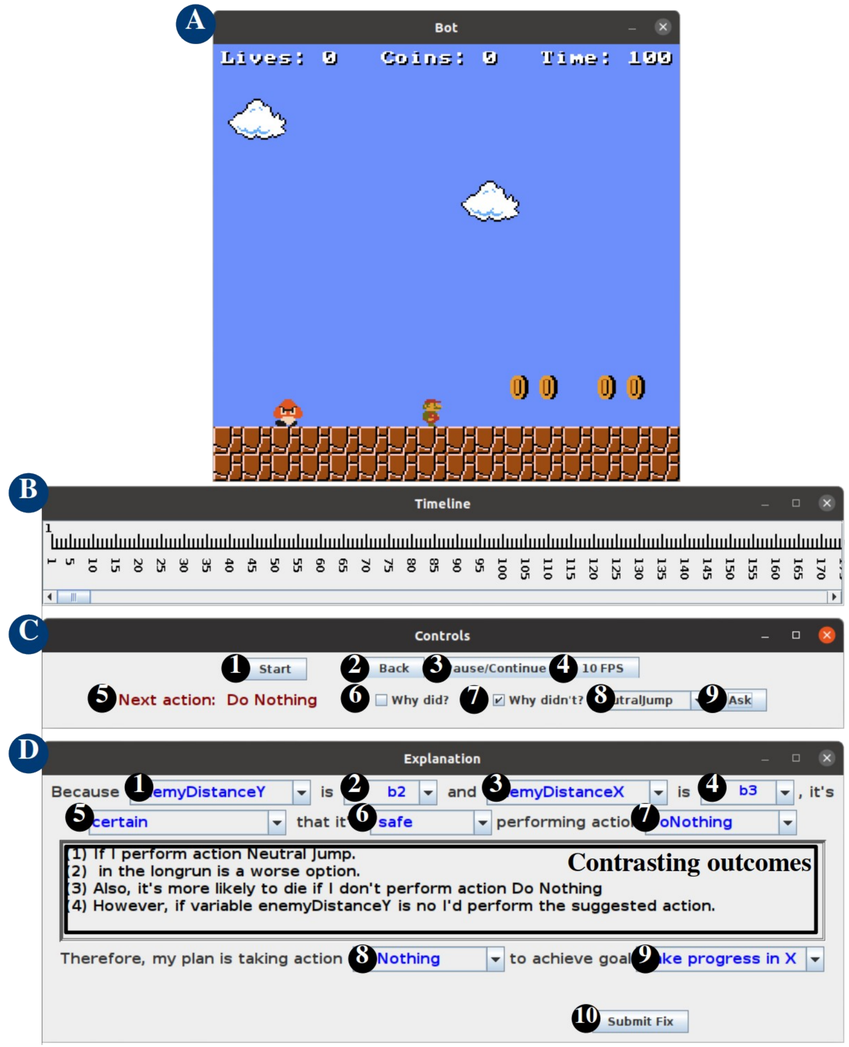
RL is a paradigm that is based on the idea of an agent that learns by interacting with its environment.



Interactive RL is a framework that adds a human-in-the-loop that can adapt the underlying learning agent.



Interactive Explanations Our interactive explanations let users ask "Why" and "Why not" questions to the bot which provides answers using a natural language template.



Interface The image above shows the interface of our interactive explanation implementation for Super Mario Bros.

Results The results of our user test are promising; users felt comfortable using our interactive explanations system.

Limitations One limitation of our method is that we require a base policy for which we can create patches that solve bugs or make small adjustments to the base behavior but this can make it difficult to make a global change.

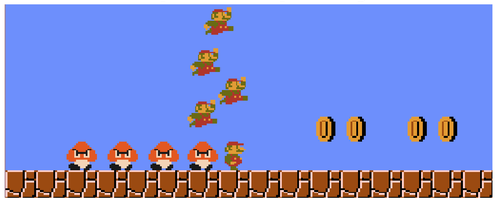
Next Steps Some users were not able to make all the changes to the policy that they wanted. To solve this, we would need to give users the ability to create arbitrary goals using as a reference any element on the screen.

Original policy

Interactive explanation
Because EnemyDistanceX is b3 and Box5Type is air, it is certain that it's safe performing action RunRight. Therefore, my plan is taking action RunRight to achieve goal Make Progress in X.

Contrasting Outcome Why didn't NeutralJump?
If I perform action NeutralJump in the long-run is a worse option. Also, it's more likely to die if I don't perform action RunRight

Updated policy



Fix
Because EnemyDistanceY is b2 and EnemyDistanceX is b3, it is certain that it's safe performing action NeutralJump. Therefore, my plan is taking action NeutralJump to achieve goal Kill an Enemy.

Contrasting Outcome Why didn't RunRight?
If I perform action RunRight in the long-run is a worse option. Also, it's more likely to die if I don't perform action NeutralJump. However, if variable EnemyDistanceY is no I'd perform the suggested action.