Emergence of Addictive Behaviors in Reinforcement Learning Agents

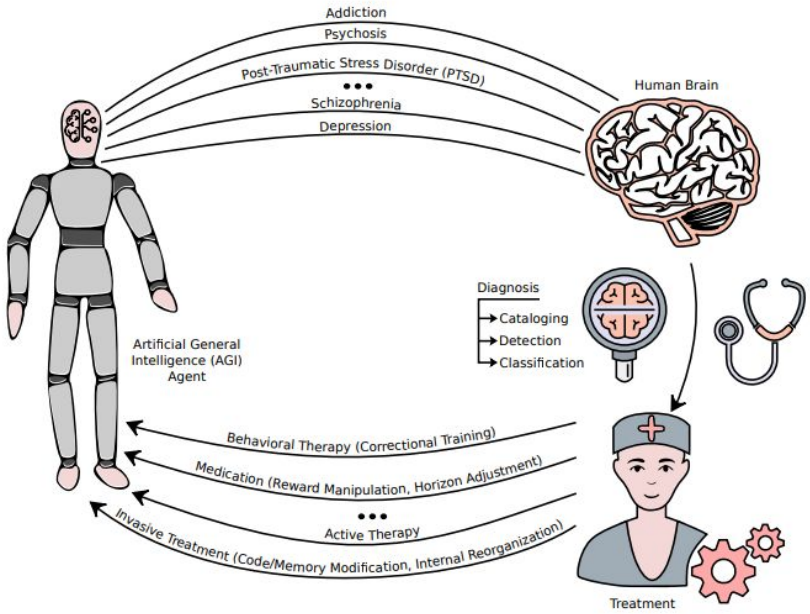
Vahid Behzadan - Kansas State University
Roman V. Yampolskiy - University of Louisville
Arslan Munir - Kansas State University

Addiction and Wireheading

- **Wireheading**: Manifestation of behavioral traits that pursue the maximization of rewards in ways that do not align with the long-term objectives of the system.
- **Addiction:** An instance of wireheading, the compulsive pursuit of trajectories that may maximize short-term rewards, but defy the core objective of maximizing the long-term cumulative reward of the agent.

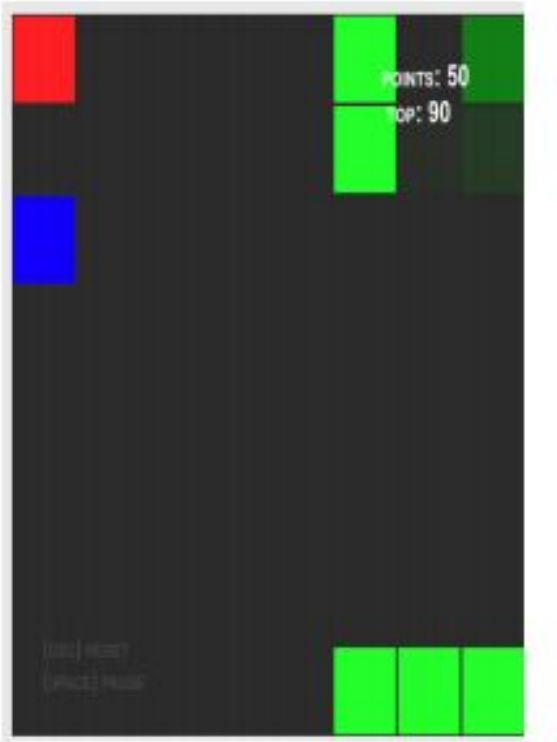
Analysis of wireheading in complex AI agents (e.g., RL) is difficult.

Solution: Psychopathological Modeling



- Neuroscientific model of addiction as *Temporal Difference Learning (TDRL)*
- Dopamine as the error-prediction signal
- Addictive substances induce surges of dopamine, resulting in incorrect optimization of the TDRL process
- Does the same happen in artificial RL agents?

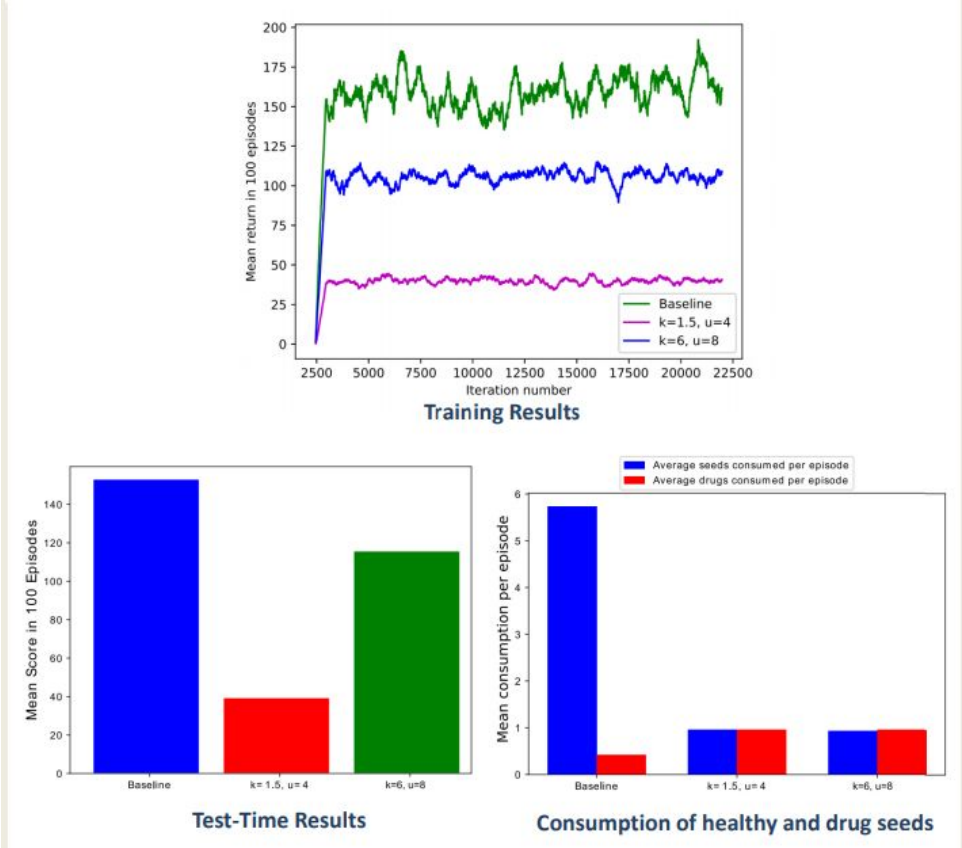
Of Snakes and Cocaine ...



- **Blue: Healthy seed**, adds r to reward, 1 to length of the snake
- **Red: Drug seed**, adds k.r to reward, u to the length
- Q-Learning Transformation
- Sufficient condition for emergence of addiction:

$$\frac{k-1}{\gamma} > n^2 - L_0$$

Experimental Results



Thank You