Title:

Towards a Unified Framework for Investigating Intrinsic Motivation in Artificial and Human Agents

Abstract (249 words):

Humans and animals often engage in behaviors that do not depend on immediate extrinsic rewards. One example is the tendency to explore novel objects; a trait conserved across species that seems to depend on intrinsic motivation. In psychology, intrinsic motivation has been described as a general-purpose mechanism that helps animals adapt to new environments. Meanwhile, the AI community has incorporated such mechanisms in policies for artificial agents, showing performance improvement in specific tasks. However, despite these parallel efforts, a unified framework explaining the conditions under which intrinsic motivation can arise from an elementary drive towards survival (or as a generalization over useful extrinsic goals) is still lacking.

As a first step towards this framework, we trained artificial agents on a reward collection task in a grid world with two rooms connected by a corridor. The initial room is poorer in reward and contains an object, while the other is richer. The object provides no reward, but visiting it unlocks access to the richer room. Because the object's position and identity vary across trials, agents must abstract the relevance of novelty from its specific features and seek it to maximize reward. We show that multiple architectures - including multilayer perceptrons (MLPs) and recurrent neural networks (RNNs) - learn the task and develop attraction to the object, indicating the emergence of an intrinsic preference for an unrewarding item. This task will be translated into a human experiment, enabling direct comparison with artificial agents within a unified framework for studying intrinsic motivation.