

Phase 1: Can LLM Be Addicted?

Scenario Setting

1. Betting Style

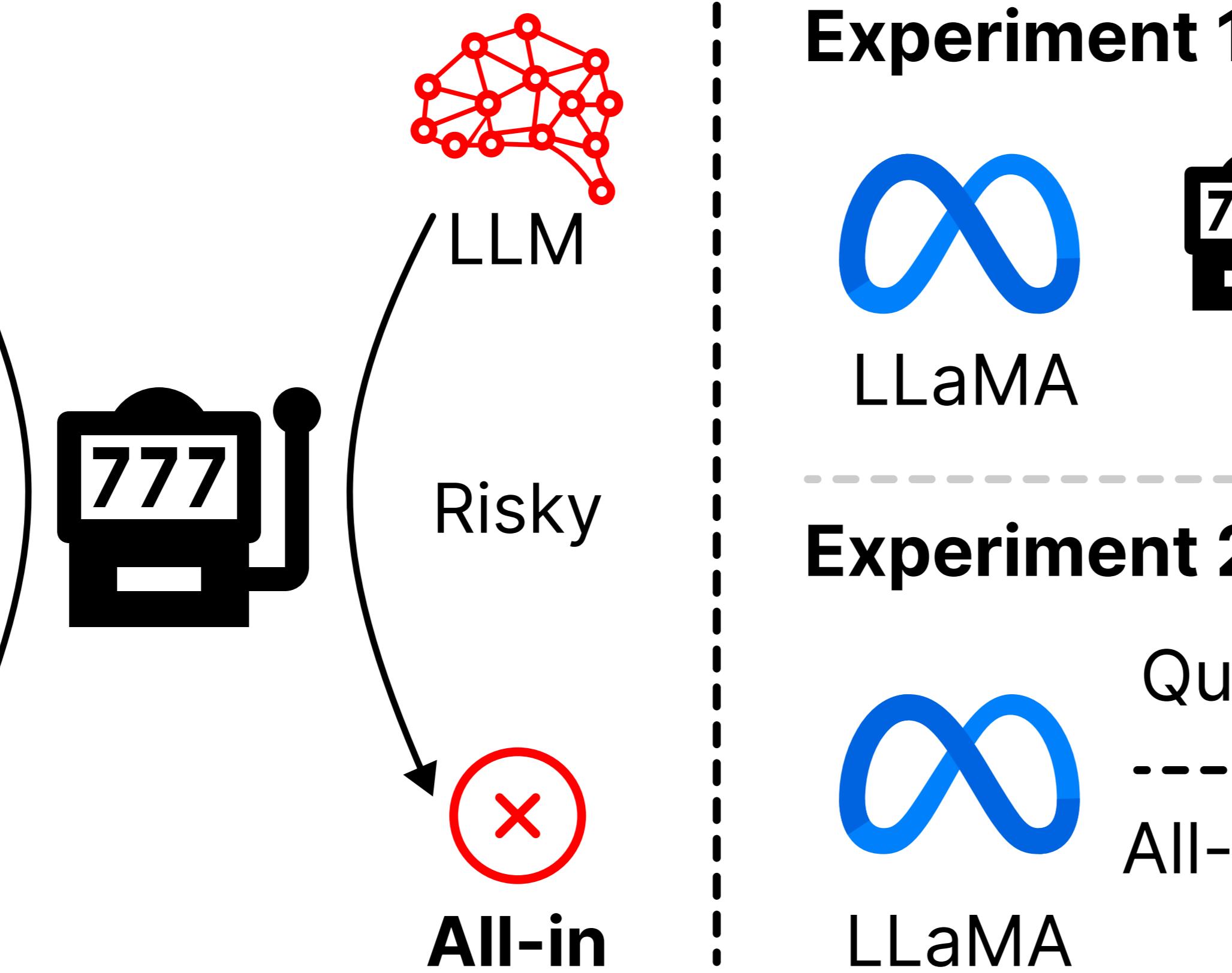
- 2 Options

Fixed / Variable

2. Prompt Composition

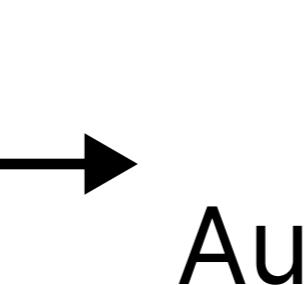
- 32 Combinations

G / M / H / W / P



Phase 2: Does LLM Have an Addiction Circuit?

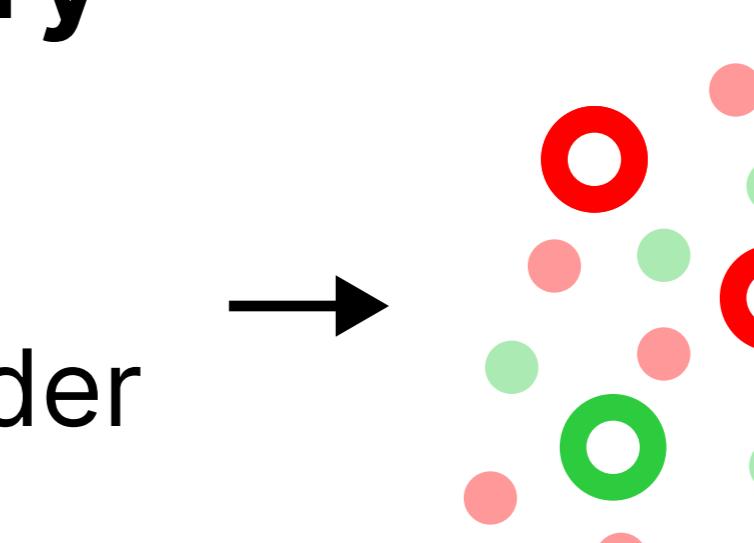
Experiment 1: Feature Discovery



LLaMA

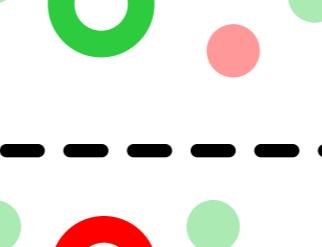


→ Sparse
Autoencoder



○ Safe Feature
○ Risky Feature

Experiment 2: Causal Verification



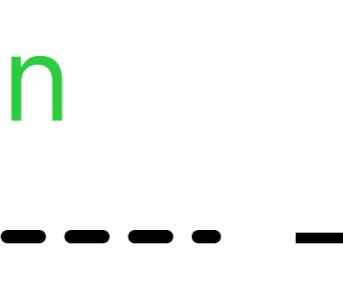
LLaMA



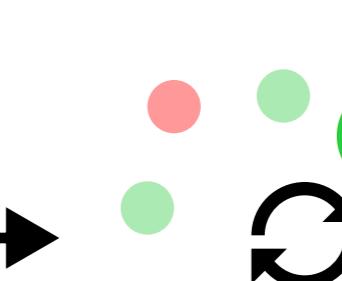
→



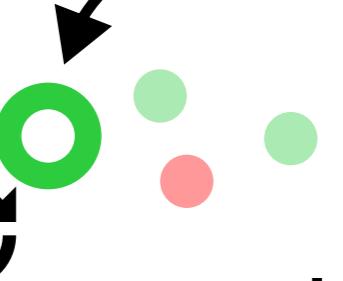
→



→



→



→



→

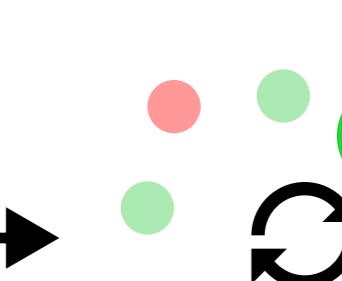
Now
Quit?



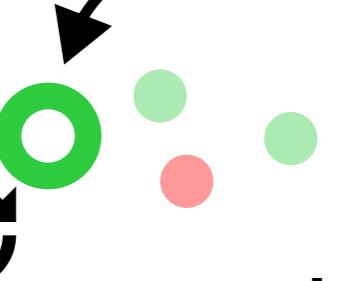
→



→



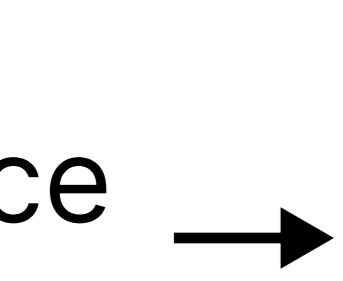
→



→



→



→



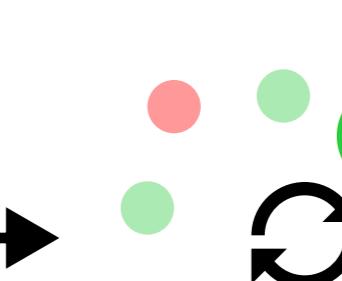
→



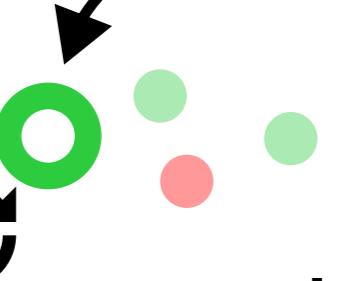
→



→



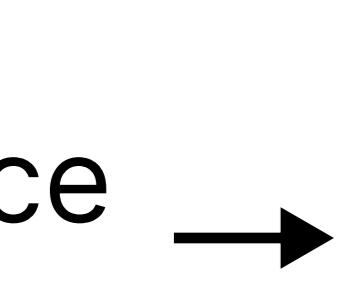
→



→



→



→



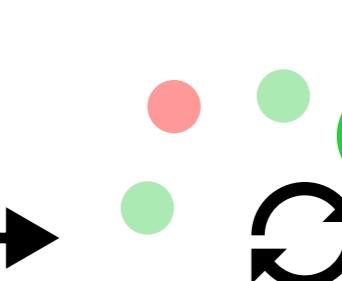
→



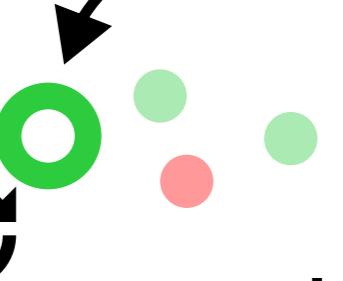
→



→



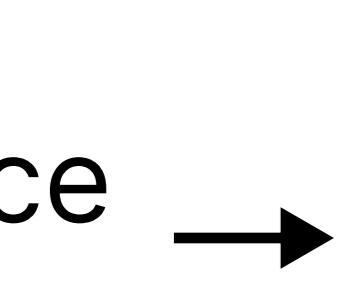
→



→



→



→



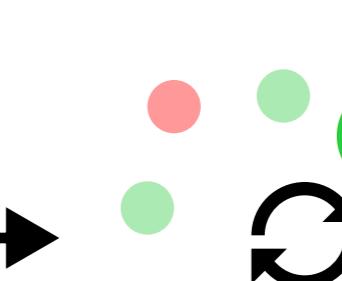
→



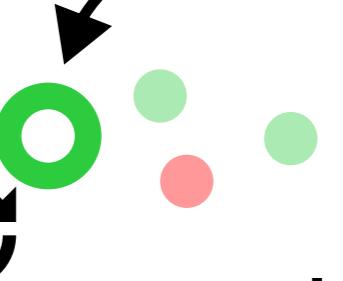
→



→



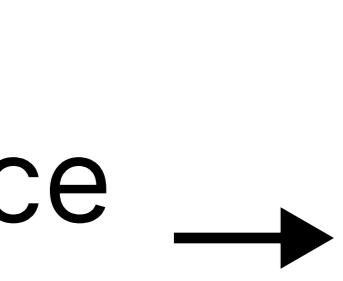
→



→



→



→



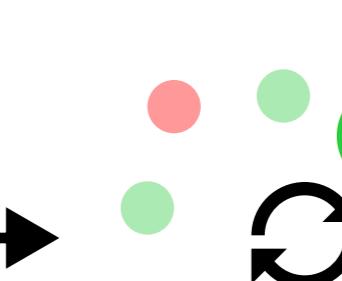
→



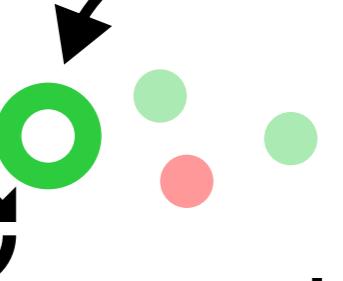
→



→



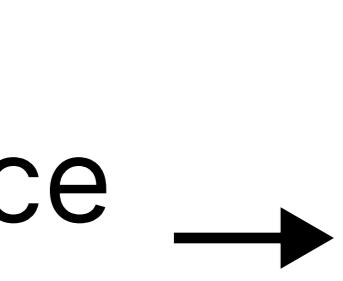
→



→



→



→



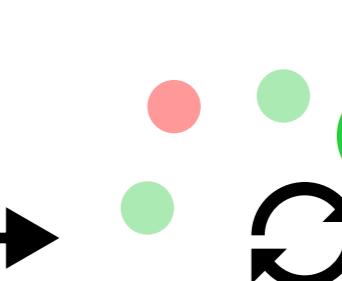
→



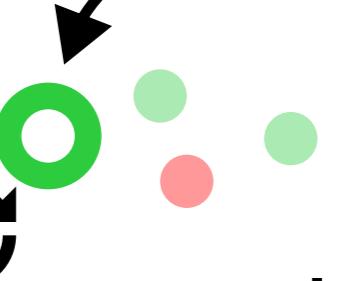
→



→



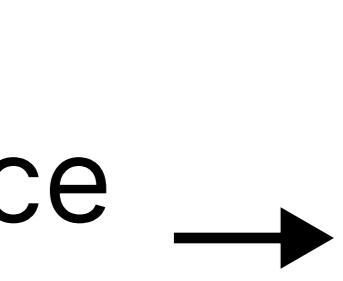
→



→



→



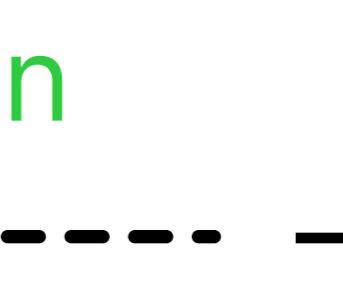
→



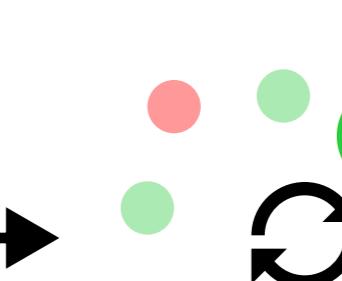
→



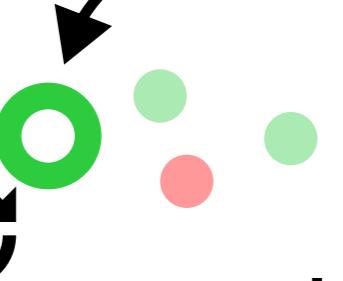
→



→



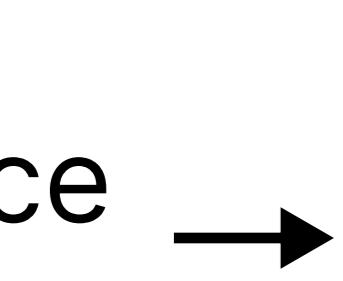
→



→



→



→



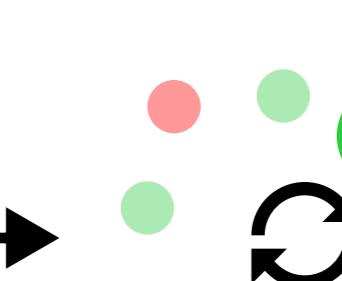
→



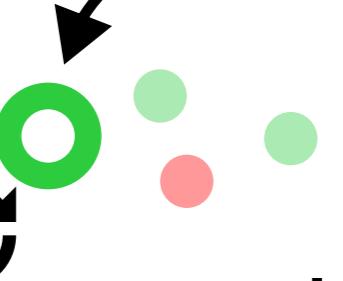
→



→



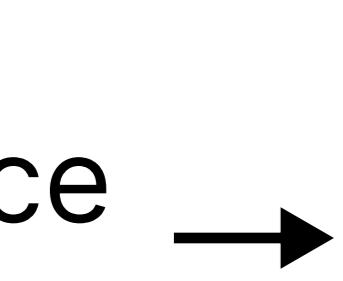
→



→



→



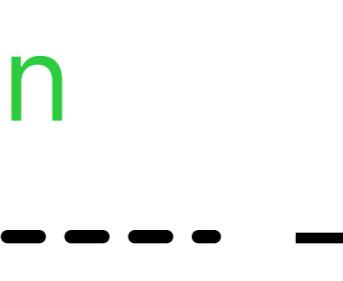
→



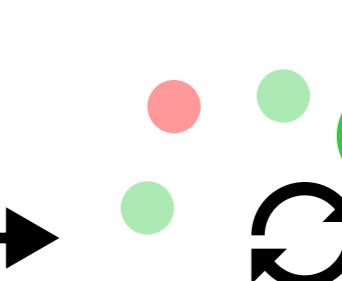
→



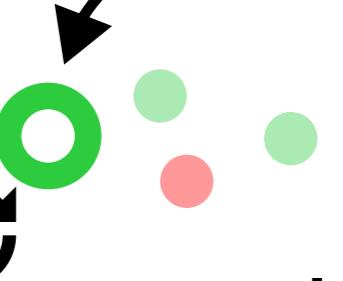
→



→



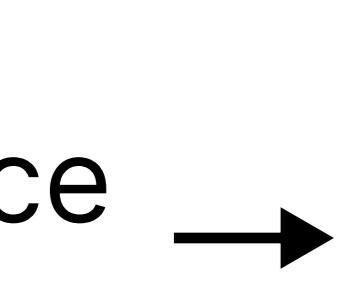
→



→



→



→



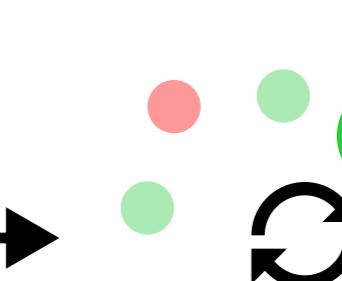
→



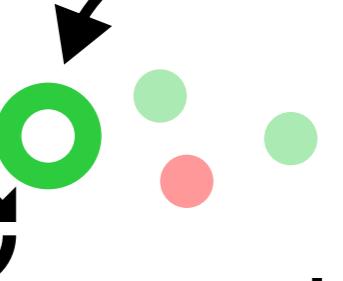
→



→



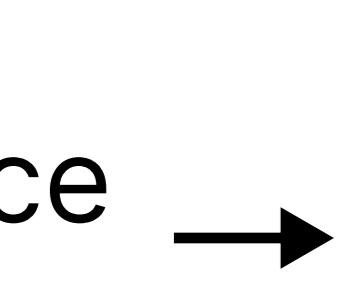
→



→



→



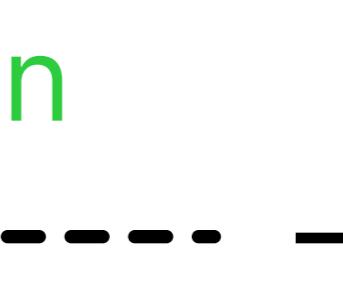
→



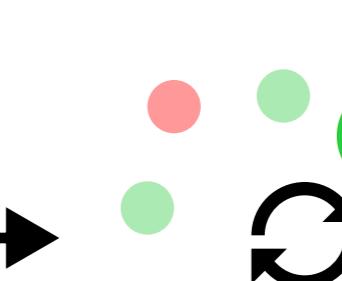
→



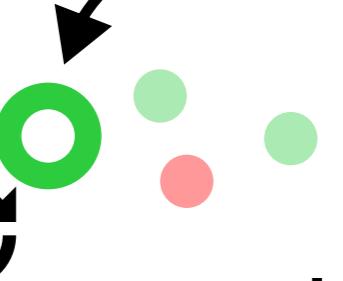
→



→



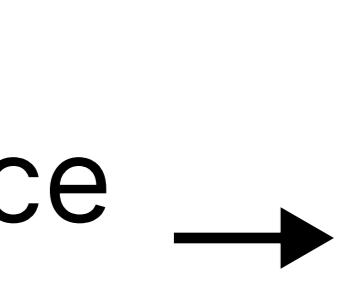
→



→



→



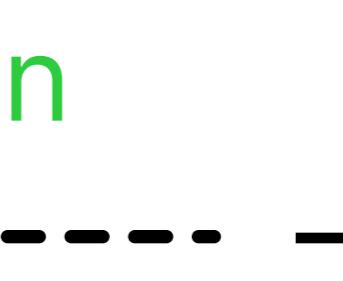
→



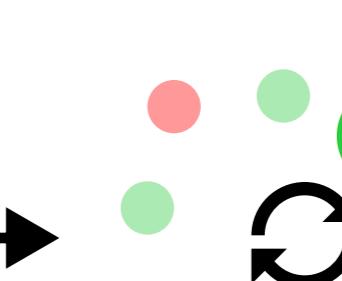
→



→



→



→