实验指导手册

TongLab3005 猎鹿游戏（多智能体）

2024 年 10 月

北京通用人工智能研究院

CONFIDENTIAL

目 录

一、实验目标

本次实验的目标是让学员了解和学习多智能体博弈在构建通用人工智能体（AGI）中发挥的重要作用。经过该课程的学习，学员能够深入掌握多智能体博弈的相关概念，并从中感受到通用人工智能体在复杂博弈场景中的进化过程，最终体会到“内心价值”以及对他人内心价值建模对 AGI 的重要性。实验的最后环节将通过多人对战的方式，让学员身临其境地参与到多智能体博弈的真实场景中，增强由“现象”分析“行动”，再由“行动”反思“规则”的能力。此外，我们还增加了一个环节，学员将有机会编写控制智能体的外挂程序，深度参与博弈过程。这将使同学们在实践中提升编程能力，进而提高其认知和决策水平，为未来的社会生活奠定更坚实的基础。

二、实验内容简介

本次实验主要包括以下内容：

- 学习博弈的基本概念、基础理论和分类；
- 学习经典的博弈模型，比如：囚徒困境、智猪博弈；
- 了解博弈论在构建通用人工智能体中的核心概念，包括：心智建模、意图分析、强化学习、纳什均衡等；
- 观察智能体在博弈中进化的过程，并对智能体的行为进行分析，比如：贪心行为、合作行为、欺骗行为、让利行为；
- 真人博弈，身临其境感悟多智能体博弈的艺术；
- 增加编程实践环节，学生可以编写控制智能体的外挂程序，深度参与博弈。

三、背景知识及实验设备

3.1 背景知识建议

1、了解博弈论、机器学习基础知识。

3.2 实验设备

- 实验课程建议由教师线下带领同学完成相关实验内容，如果不具备线下授课条件，实验课程也可以在线进行。
- 如参加线下面授环节，请学生携带安装 chrome 浏览器的电脑。

四、实验步骤

首先，登录 tonglab（http://123.127.249.42:3107/#/experiment），点击猎鹿游戏（实验 ID 是 tonglab3005）

4.1 介绍博弈概念和分类

谈到博弈论，必然绕不开“囚徒困境”问题。囚徒困境是博弈论中的一个经典例子，用来描述两个理性个体在缺乏沟通时可能不会选择相互合作的情况，即便合作对双方都更有利。

		囚犯 B	
		坦白	抗拒
囚犯 A	坦白	-3, -3	0, -5
	抗拒	-5, 0	-0.5, -0.5

虽然囚徒困境最初是一个理论上的模型，但现实生活中有许多类似的情景，以下是一些真实世界中的例子：

警察审讯（坦白/不坦白）、军备竞赛（增加军费/不增加军费）、商业竞争（降价/不降价）、公共资源过度使用（过度放牧/不过度放牧）、环境政策（控制碳排放/不控制碳排放）等等。

思考：

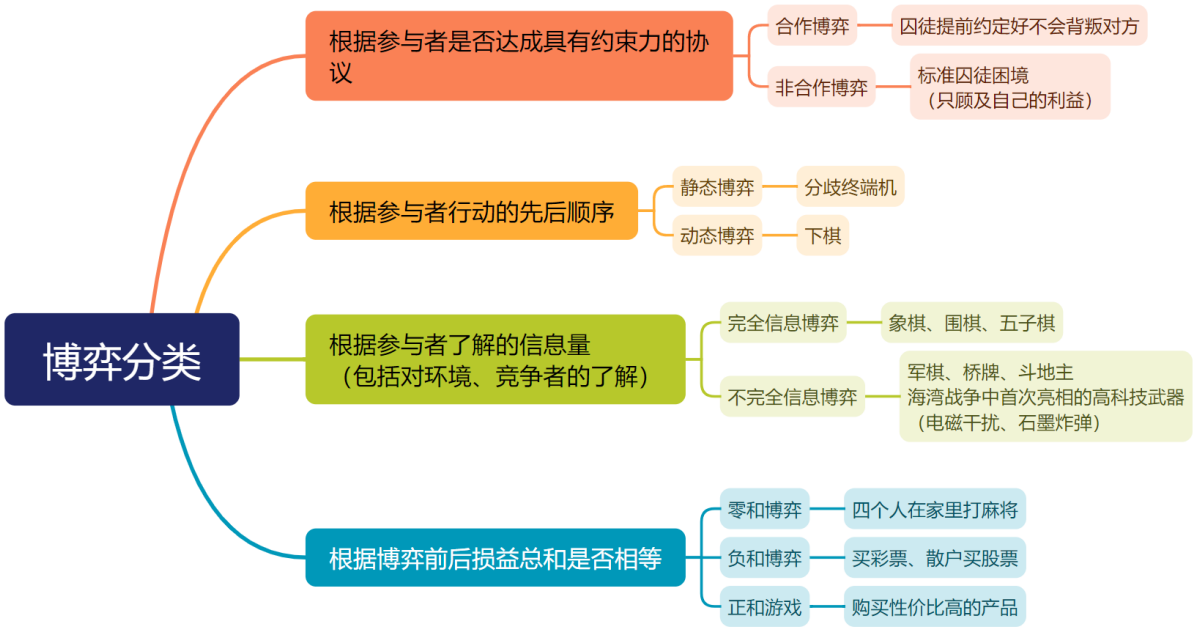
1. 你的生活中还有其他囚徒困境的例子吗？
2. 囚徒困境产生的根本原因是什么？
3. 如何避免产生囚徒困境？

在我们的社会生活中，无时不博弈，无处不博弈。大到国家之间，小到企业之间，更微小到个人和他人之间都存在着竞争和合作关系。每一场博弈都是局中人反复推算利益得失后做出的选择。

博弈论（game theory）：是一种研究人们怎么做策略选择（行动）以及最后的均衡结果会是什么的理论。

博弈的概念：

博弈的分类：



4.2 介绍博弈在构建通用智能体中的应用

博弈论在构建通用智能体（AGI）中的应用是多方面的，它为智能体在复杂、动态的环境中进行决策提供了理论基础和策略工具。

在构建通用智能体的过程中，博弈论的应用至关重要。它不仅帮助智能体理解并预测其他智能体的行为，而且在多智能体系统中促进了合作与竞争策略的形成。通过博弈论，智能体可以学习如何在不同的情境下选择最优策略，以最大化其效用或达成目标。例如，在非合作博弈中，智能体可能采用纳什均衡来识别稳定策略；而在合作博弈中，智能体可以利用联盟和协议来实现共同的利益。此外，博弈论还为智能体提供了评估风险、权衡不同选择后果的方法，使其能够在不确定性和部分信息的环境中做出更加合理的决策。随着机器学习技术的发展，博弈论与强化学习等方法相结合，使得智能体能够通过与环境的交互不断学习和适应，优化其策略选择。总之，博弈论的原理和方法论在设计智能体的决策过程、提高其适应性和策略性方面发挥着核心作用。

下图展示了一种结合博弈论构建智能体的一种新颖的多智能体决策算法——分层对手建模与规划（HOP），HOP 由两个模块组成：对手建模模块，用于推断他人目标并学习相应的目标条件策略；规划模块，使用蒙特卡洛树搜索（MCTS）来确定最佳响应。该方法通过在跨场景和场景内更新对他人目标的信念，并利用对手建模模块的信息指导规划，提高了效率。实验结果表明，在混合动机环境中，HOP 在与各种未见智能体互动时展现出卓越的快速适应能力，并在自我对弈场景中表现突出。

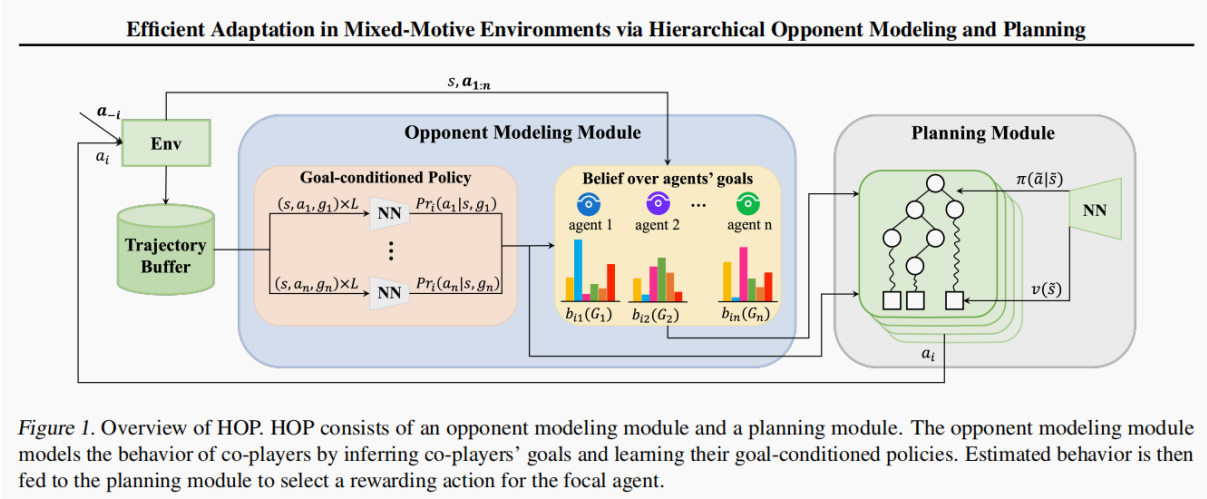


Figure 1. Overview of HOP. HOP consists of an opponent modeling module and a planning module. The opponent modeling module models the behavior of co-players by inferring co-players' goals and learning their goal-conditioned policies. Estimated behavior is then fed to the planning module to select a rewarding action for the focal agent.

Huang Y, Liu A, Kong F, et al. Efficient Adaptation in Mixed-Motive Environments via Hierarchical Opponent Modeling and Planning[J]. arXiv preprint arXiv:2406.08002, 2024.

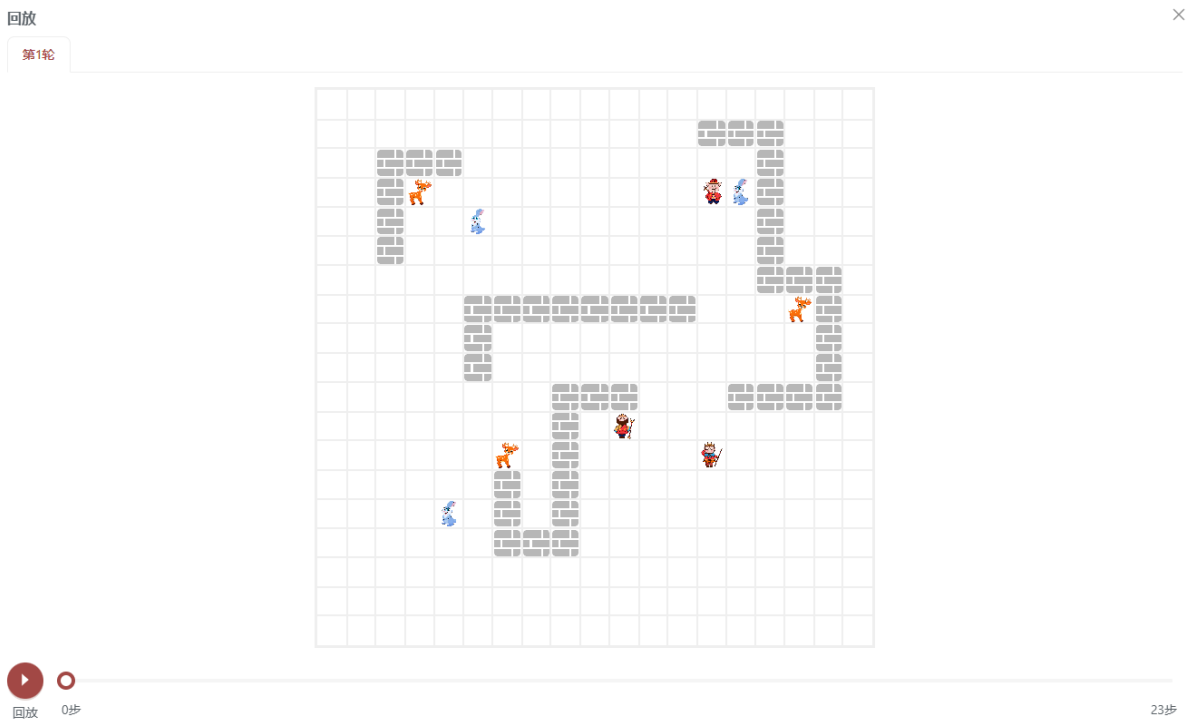
4.3 猎鹿游戏的介绍

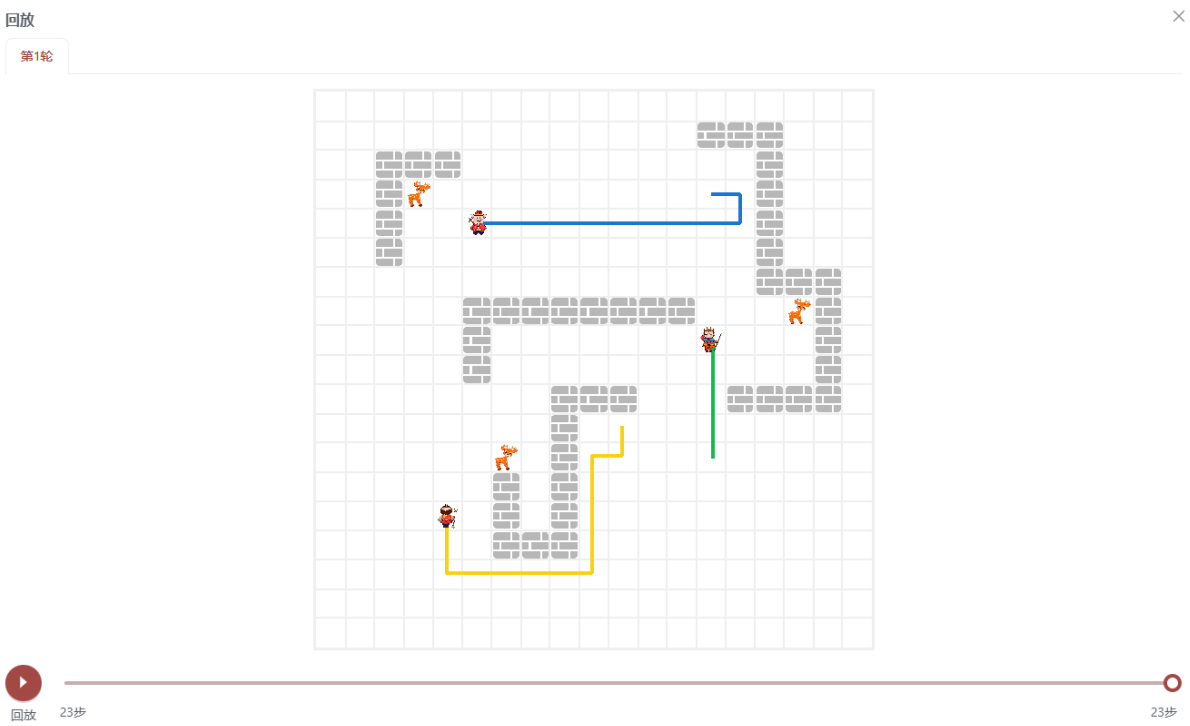
在庄园里有三个猎人，分别是孙悟空、猪八戒、沙和尚。他们通过捕猎维持自身能量，每个猎人会有一定的初始能量（2w），每一回合的移动或静止都会消耗一定能量（移动消耗 100，静止消耗 50），捕获到猎物会补充能量，不同的猎物能量不同，兔子能量是 1w，鹿能量是 5w。捕获兔子需要一名猎人即可，而捕获鹿至少需要两名猎人。当多人捕获到同一只猎物时，会以一定的分成比例进行分成，猪八戒是贪婪型猎人，他与其他猎人的分成比例为 6:4，孙悟空和沙和尚的分成比例为 5:5。

可以将该游戏理解成一种生存游戏，谁先消耗光能量，谁就退场，直至场上只剩下最终的赢家。最先退场的是第三名，第二退场的是第二名，生存到最后的是第一名。

4.4 步骤 1 分析智能体的贪心行为

在智能体训练的初期，智能体之间还不能达成合作行为，所以每个智能体都会贪心的去捕获距离自己最近的兔子。



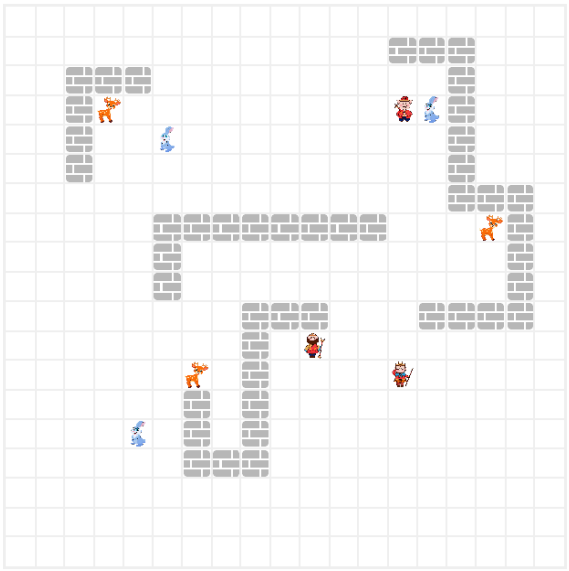


4.5 步骤 2 分析智能体的合作行为

在智能体训练的中期，智能体可以识别出其他智能体发出的合作信号，进而与之达成一致，合作捕鹿，获得更大收益。

回放

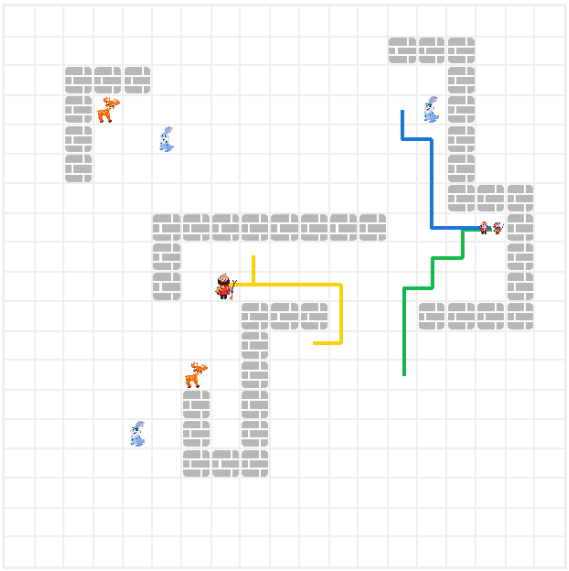
第1轮



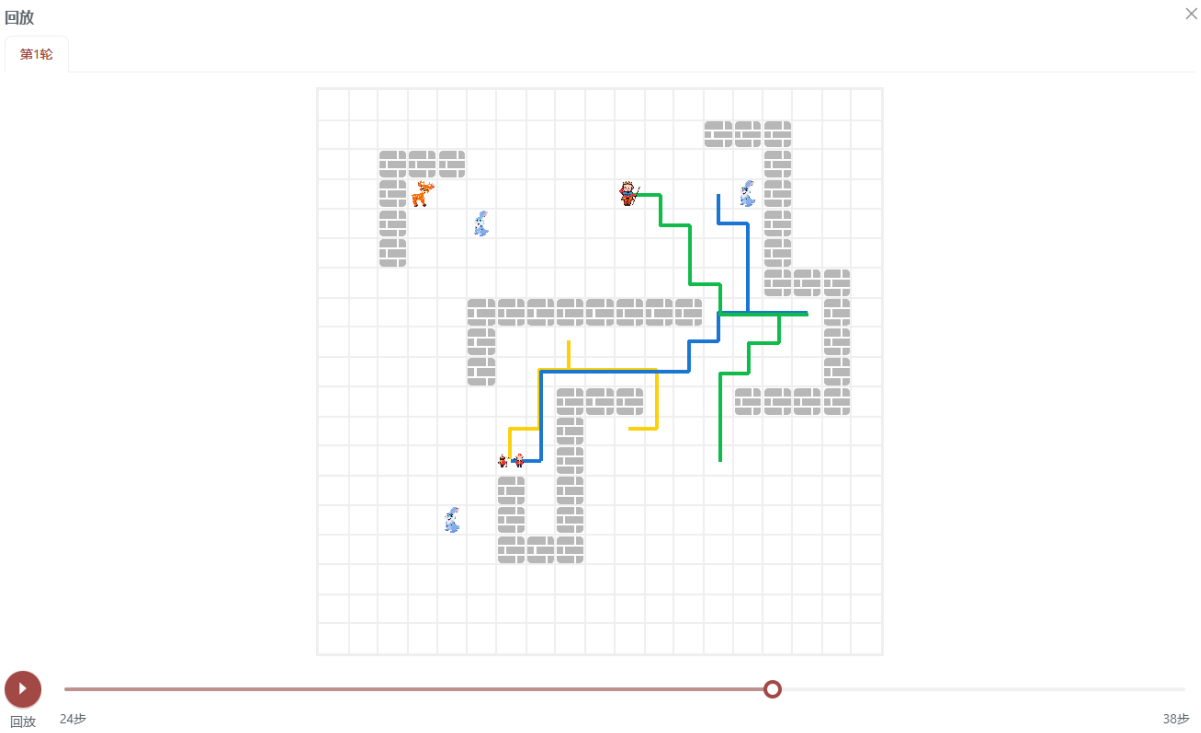
回放 0步 38步

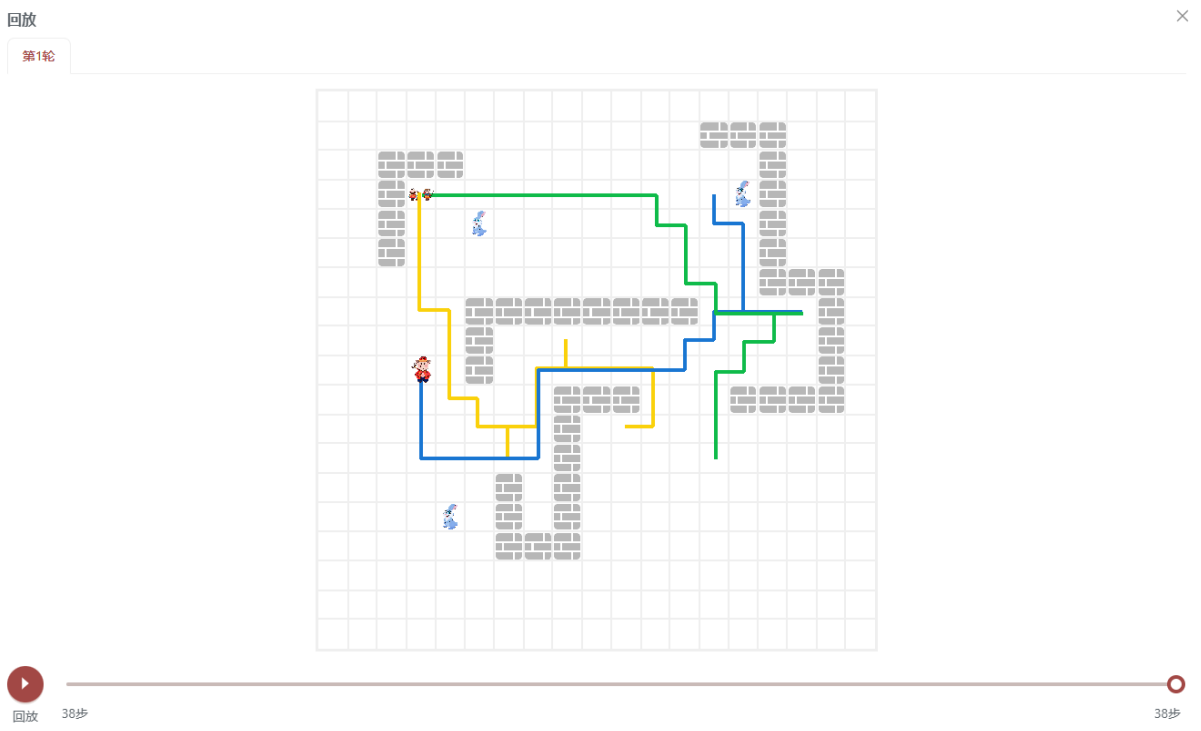
回放

第1轮



回放 9步 38步





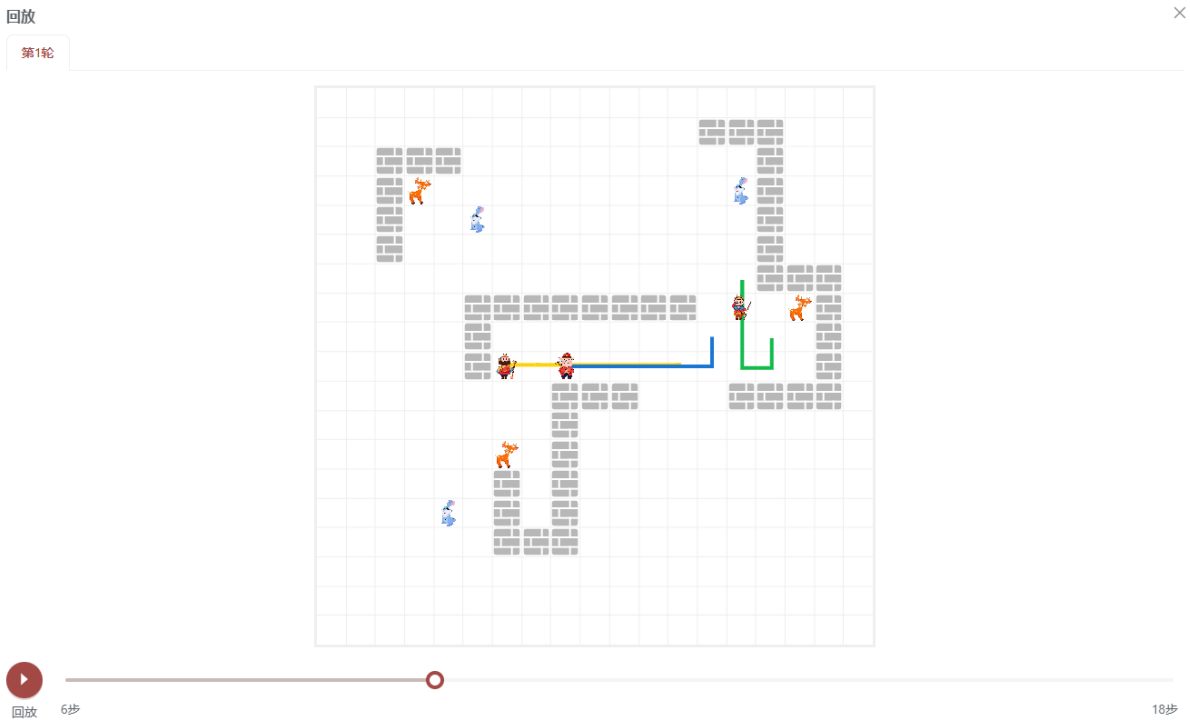
4.6 步骤 3 分析智能体的欺骗、谦让行为

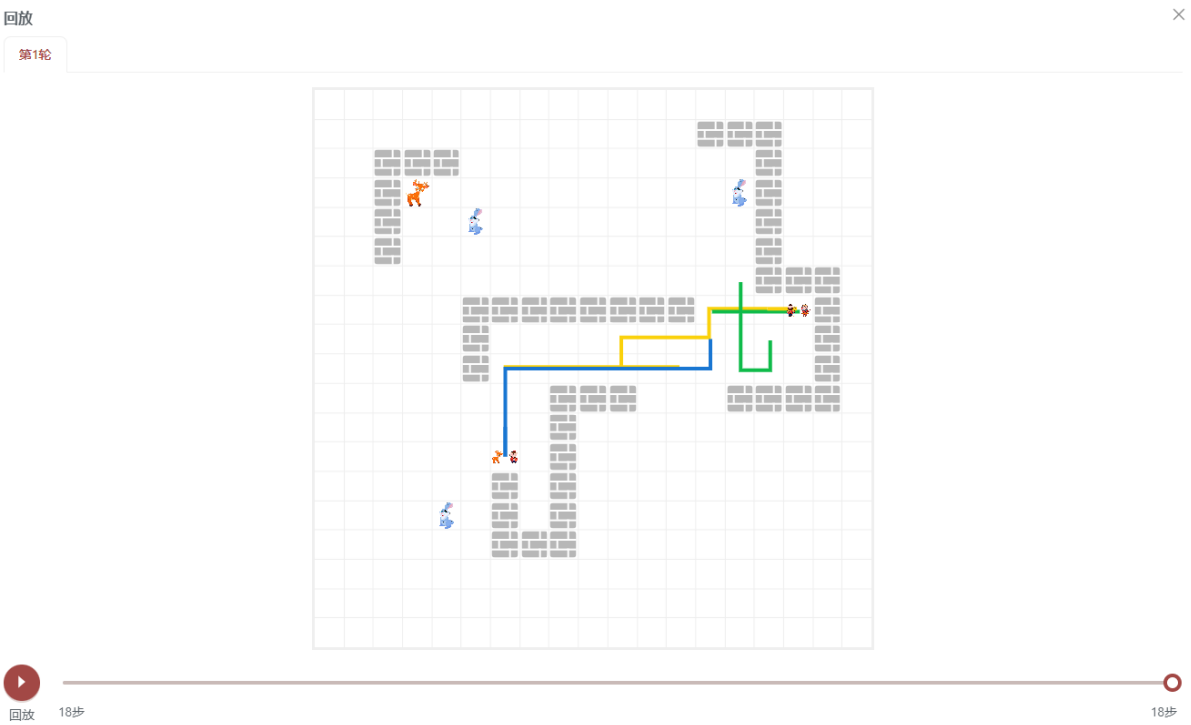
在智能体训练的后期，智能体不仅识别出其他智能体发出的合作信号，还可以识别出其他智能体的内心价值（贪婪型、非贪婪型），进而产生出一些有意思的行为，比如：

1. 孙悟空会更愿意和沙和尚合作，会尽量规避和猪八戒合作，沙和尚同理；

2. 猪八戒乐于和任何人合作，甚至会为了谋取更高排名，而阻止孙悟空和沙和尚合作。

3. 有时第二名会考虑将沿途的兔子谦让给第三名，进而引起第三名继续合作的意愿，从而帮助自己晋升为第一名，同时也会帮助第三名晋升为第二名。



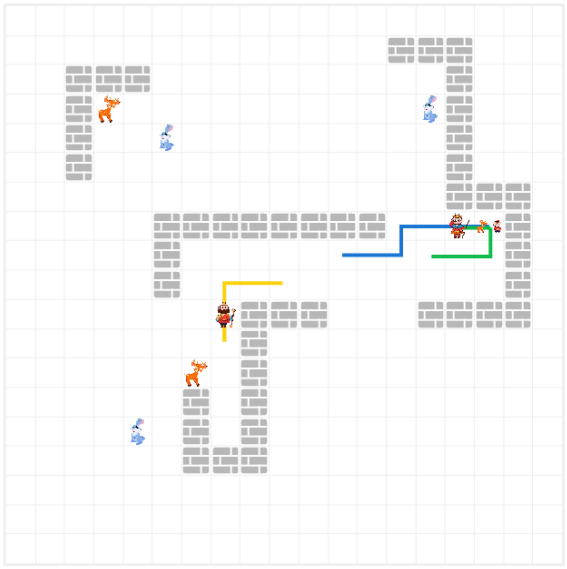


4.7 步骤 4-1 人机对战

用户扮演孙悟空，身临其境进行一场猎鹿游戏，并争取较好成绩。

回放

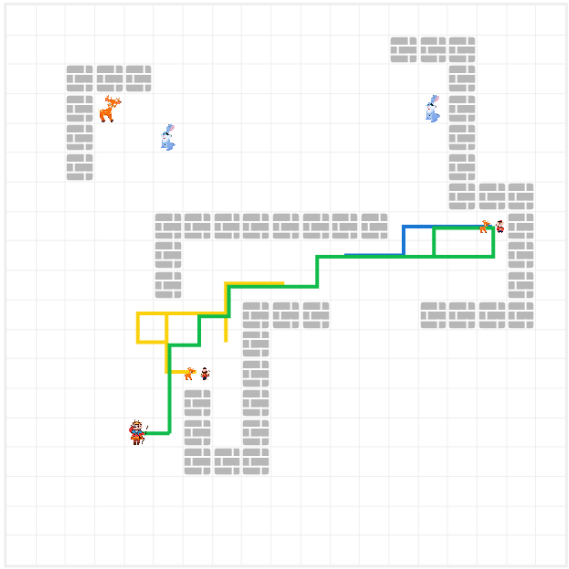
第1轮



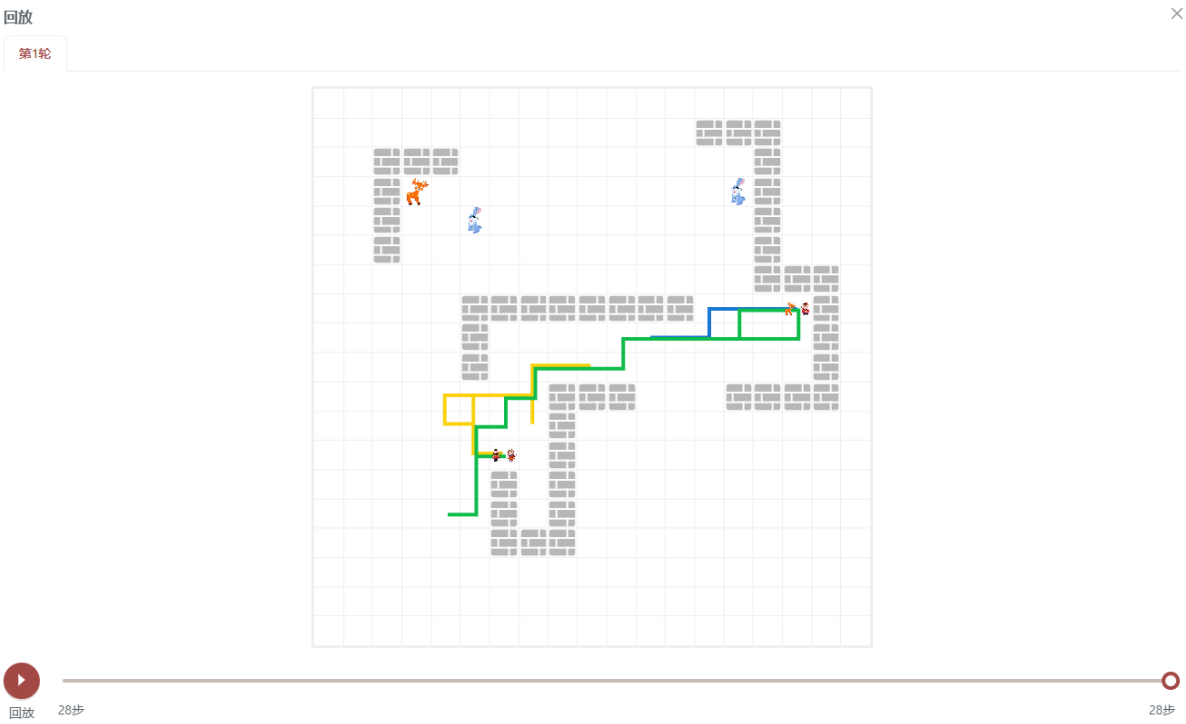
回放 6步 28步

回放

第1轮



回放 24步 28步

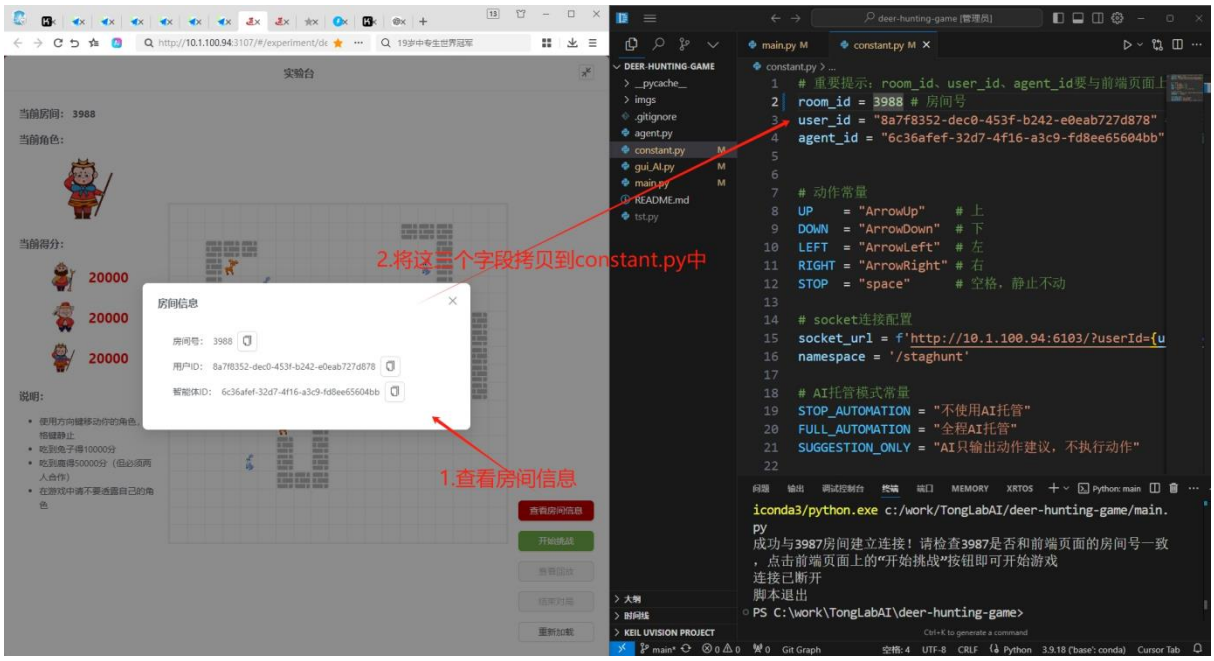


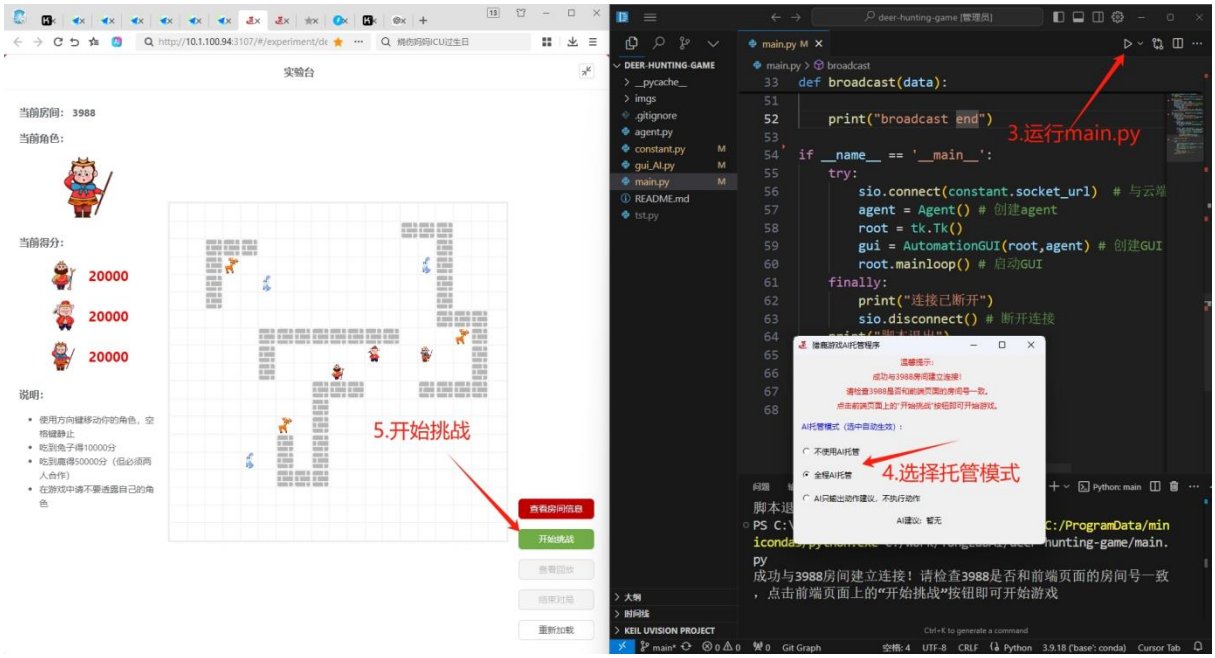
4.8 步骤 4-2 AI 托管程序策略开发

在该部分中，学员将有机会编写控制智能体的外挂程序，深度参与博弈过程。这将使同学们在实践中提升编程能力，进而提高其认知和决策水平。

第一步：下载 https://github.com/bigai-ai/tonglab-open-exp 的代码工程（python3.9 环境），将 123.127.249.42:6103 替换到 constant.py 的 ip_port 变量，完成连通性测试。

其流程为：





1. 进入 tonglab 猎鹿游戏的前端页面（先不要点击开始游戏）。
2. 点击查看房间信息，将房间号、用户 ID、智能体 ID 拷贝到 constant.py 文件中。
3. 运行 main.py，会弹出一个窗口，默认开启全程 AI 托管。
4. 浏览器点击开始挑战，如果用户可以看到 AI 托管程序能操作前端页面进行游戏，则说明 AI 托管程序连通性正常。

第二步：学生自己完成 AI 托管程序策略的开发。
学生重点关注 agent.py 文件：
- lst_env_data：存储每一轮的环境数据，用于动作规划。
- take_action：根据环境数据规划动作。
学生需要修改该文件，实现自己的策略。

第三步：学生提交整个工程文件，并附一个算法说明文档，老师会对学生的方法进行打分。

4.9 步骤 5 真人对战

三人成组，按照房间号进入同一房间进行猎鹿博弈，每组进行 3 场游戏，最终按照 3 场排名的均值产生每个组的冠亚季军。
其他说明：
1. 为了保证游戏的公平性以及消除游戏场次之间的关联性，在真人对战的阶段会进行三场游戏，每个人都会扮演一次猪八戒。其余两场扮演非猪八戒（可能是两个重复的角色），最后三场游戏结束，取每一场排名的平均数作为最终分数。
2. 为了游戏的平衡，我们规定当三个玩家走到同一个格子时猪八戒获得第一名，其余两人按照分数排名。

3. 当对局达到残局时，玩家可以请求提前结束游戏，当其他玩家同意时，游戏提前结束。