Suphx:用深度强化学习掌握麻将*

、1 山田宗越、2 叶祁伟、1、3、4 杨、5 李昭、1
陶秦，1 个刘铁燕，1 个萧-乌恩

1 微软亚洲研究院
2 京都大学
3 中国科学技术大学
清华大学
5 南开大学

摘要

人工智能(AI)已经在许多领域取得了巨大的成功，游戏 AI 被广泛认为是自 AI 出现以来的滩头阵地。近年来，对游戏人工智能的研究逐渐从相对简单的环境(例如，完全信息游戏如围棋、象棋、shogi 或两人不完全信息游戏如单挑德州扑克)发展到更复杂的环境(例如，多人不完全信息游戏如多人德州扑克和 StartCraft II)。麻将是世界范围内流行的多人不完全信息游戏，但由于其复杂的游戏规则和丰富的隐藏信息，对人工智能的研究极具挑战性。我们设计了一个麻将人工智能，命名为 Suphx，基于深度强化学习和一些新引入的技术，包括全局奖励预测，oracle 指导和运行时策略适应。在稳定排名方面，Suphx 表现出了比大多数顶级人类玩家更强的表现，并在 Tenhou 平台所有官方排名的人类玩家中获得了 99.99%以上的评级。这是计算机程序第一次在麻将中胜过大多数顶级人类玩家。

1 介绍

为游戏构建超人程序是人工智能(AI)的一个长期目标。游戏人工智能在过去的二十年里取得了巨大的进步(2，3，

*这项工作在微软亚洲研究院进行。第二、第四、第五和第六作者当时是微软亚洲研究院的实习生。

11, 13, 15, 16, 18).最近的研究已经逐渐从相对简单的完全信息或两人游戏(例如，shogi、chess、Go和单挑德州扑克)发展到更复杂的不完全信息多人游戏(例如，契约桥(12)、Dota (1)、星际争霸II (21)和多人德州扑克)
德州扑克(4))。

麻将，一个多回合的瓷砖为基础的游戏，具有不完善的信息和
多玩家，在全球数亿玩家中非常受欢迎。在麻将游戏的每一轮中，四名玩家相互竞争，以第一次完成获胜的一手牌。建立一个强大的麻将程序对当前的游戏人工智能研究提出了巨大的挑战。

第一，麻将计分规则复杂。每局麻将包含多轮，一局的最终排名(以及奖励)由这些轮的累积分数决定。输掉一轮并不总是意味着玩家在该轮中表现不佳(例如，如果他/她在前几轮中具有很大优势，则玩家可能在战术上输掉最后一轮以确保游戏的排名1)，因此我们不能直接使用该轮分数作为学习的反馈信号。此外，麻将有大量可能获胜的手牌。这些获胜的手牌可以彼此非常不同，并且不同的手牌导致该轮的不同获胜回合分数。这种计分规则比以前研究的游戏复杂得多，包括国际象棋、围棋等。一个职业玩家需要仔细选择形成什么样的赢手牌，才能权衡该轮的赢概率和赢分。

第二，在麻将中，每个玩家手里有多达13张其他玩家看不到的私人牌，并且在死墙中有14个标题，它们在整个游戏中对所有玩家都是不可见的，而在活墙中有70张牌，它们一旦被玩家抽取和丢弃就会变得可见。结果，平均来说，对于每个信息集(一个玩家的决策点)，有超过1048个隐藏状态是他/她无法区分的。如此大量的隐藏信息使得麻将比以前研究过的德州扑克更难成为不完全信息游戏。麻将玩家很难仅根据他/她自己的私人牌来确定哪个动作是好的，因为动作的好坏高度依赖于其他玩家的私人牌和每个人都看不见的墙牌。因此，AI也很难将奖励信号与观察到的信息联系起来。

第三，麻将的玩法是复杂的:(1)有不同类型的动作，包括理一、周、乓、孔、弃牌，以及(2)在打一手牌(乓或孔)、打麻将(宣布一手赢牌)或抢孔时，可以打断规则的玩法顺序。因为每个玩家可以有多达13个私人牌，很难预测这些中断，因此我们甚至不能建立一个常规的游戏树；即使我们建立了一个游戏树，这样的树在一个玩家的连续动作之间会有大量的路径。这阻碍了以前成功技术的直接应用
对于游戏，如蒙特卡罗树搜索(14，15)和反事实后悔最小化(3，4)。

由于上述挑战，尽管有几次尝试(7–9，20)，最好的麻将 AI 仍然远远落后于顶级人类玩家。

在这项工作中，我们建立了 Suphx(超级凤凰的缩写)，一个用于 4 人日本麻将(日式麻将)的人工智能系统，它拥有最大的
世界麻将社区。Suphx 采用深度卷积神经网络作为其模型。该网络首先通过从人类职业选手的日志中的监督学习来训练，然后通过自玩强化学习(RL)来增强，其中网络作为策略。我们使用
用于自播放 RL 的流行策略梯度算法(17)引入了几种技术来解决上述挑战。

1. 全局奖励预测训练预测器根据当前和先前回合的信息来预测游戏的最终奖励(在未来几个回合之后)。该预测器提供有效的学习信号，从而可以执行策略网络的训练。此外，我们设计了前瞻功能，以编码不同获胜手牌的丰富可能性及其本轮获胜分数，作为对 RL 代理决策的支持。

2. Oracle guiding 引入了一个 Oracle 代理，它可以看到完美的信息，包括其他玩家的私人瓷砖和墙砖。由于(不公平的)完美的信息访问，这个 oracle 代理是一个超级强大的麻将 AI。在我们的 RL 训练过程中，我们逐渐丢弃 oracle agent 中的完美信息，最终将其转换为只接受可观测信息作为输入的普通 agent。在 oracle 代理的帮助下，我们的普通代理比仅利用可观察信息的标准 RL 训练提高得更快。

3. 由于复杂的麻将游戏规则导致不规则的博弈树，并阻碍了蒙特卡罗树搜索技术的应用，我们引入了参数蒙特卡罗策略自适应(pMCPA)来提高我们的代理的运行时性能。pMCPA 在一轮进行中，当有更多可观察的信息(例如四个玩家丢弃的公共牌)时，逐渐修改和适应离线训练的策略以适应在线播放阶段中的特定一轮。

我们在最受欢迎和最具竞争力的麻将平台 Tenhou (19)上对 Suphx 进行了评估，该平台拥有超过 35 万的活跃用户。Suphx 在 Tenhou 达到了 10 dan，其描述一个玩家长期平均表现的稳定排名超过了大多数顶级人类玩家。

2 Suphx 概述

在本节中，我们首先描述 Suphx 的决策流程，然后描述 Suphx 中使用的网络结构和功能。

模型	功能
丢弃模型	决定在正常情况下丢弃哪张牌
日一模型周	决定是否宣布决定是否制作一个 Chow
模型庞模型	决定是否制作一个 Pong 决定是否制作
孔模型	一个 Kong

表 Suphx 中的五种型号

2.1 决策流程

由于麻将的复杂玩法，Suphx 学习了五种模式来处理不同的情况:弃牌模式、理一模式、周模式、乓模式和孔模式，如表中所示 1.

除了这五个学习模型，Suphx 还采用了另一个基于规则的获胜模型来决定是否宣布一手获胜牌并赢得这一轮。它主要检查是否可以从其他玩家丢弃的牌或从墙上抽出的牌中形成一手获胜的牌，然后根据以下简单规则做出决定:

- 如果这不是游戏的最后一轮，则声明并赢得该轮；
- 如果这是游戏的最后一轮，
  - 如果在宣布一手获胜牌后，整局游戏的累积回合分数是四名玩家中最低的，则不要宣布；
  - 否则，宣布并赢得该回合。

麻将玩家需要采取行动的情况有两种，我们的 AI Suphx 也是如此(见图 1):

抽牌场景:Suphx 从墙上抽一张牌。如果它的私人牌可以与抽取的牌形成一手获胜牌，获胜的模型决定是否宣布获胜。如果是，它就声明，这一轮结束。否则，

1. 孔步骤:如果私有图块可以与绘制的图块形成封闭孔或添加孔，则孔模型决定是形成封闭孔还是添加孔。如果没有，转到 Riichi 步骤；否则，有两种子情况:
   (a) 如果这是一个封闭的孔，使封闭的孔，并回到平局的情况。
   (b) 如果它是一个 AddKong，其他玩家可以使用此 AddKong 牌赢得此回合。如果其他玩家赢了，这一轮就结束了；否则，进行 AddKong 并回到听牌的情况。

2. Riichi 步骤:如果私有瓷砖可以用绘制的瓷砖制作 Riichi，则 Riichi 模型决定是否声明 Riichi。如果没有，转到丢弃步骤；否则，声明Riichi，然后转到丢弃步骤。
3. 丢弃步骤:丢弃模型选择要丢弃的牌。之后，轮到其他玩家采取行动，否则这一轮结束时不会留下任何墙砖。[1]

其他人弃牌的情况:其他玩家弃牌。如果 Suphx 能用这张牌组成一手赢牌，则由赢牌模式决定是否宣布赢牌。如果是，它就声明，这一轮结束。否则，它检查是否可以用丢弃的牌来制作 Chow、Pong 或 Kong。如果没有，则轮到其他玩家采取行动；否则，Chow、Pong 或 Kong 模型决定采取什么动作:

1. 如果三个模型都没有建议行动，则轮到其他玩家采取行动，否则这一轮结束时不会留下任何墙砖。
2. 如果建议了一个或多个动作，Suphx 会建议具有最高置信度得分(由这些模型输出)的动作。如果提议的行动没有被其他玩家的更高优先级的行动打断，Suphx 采取行动，然后进入第一种情况下的弃牌步骤。否则，提议的动作被中断，并且轮到其他玩家采取动作。

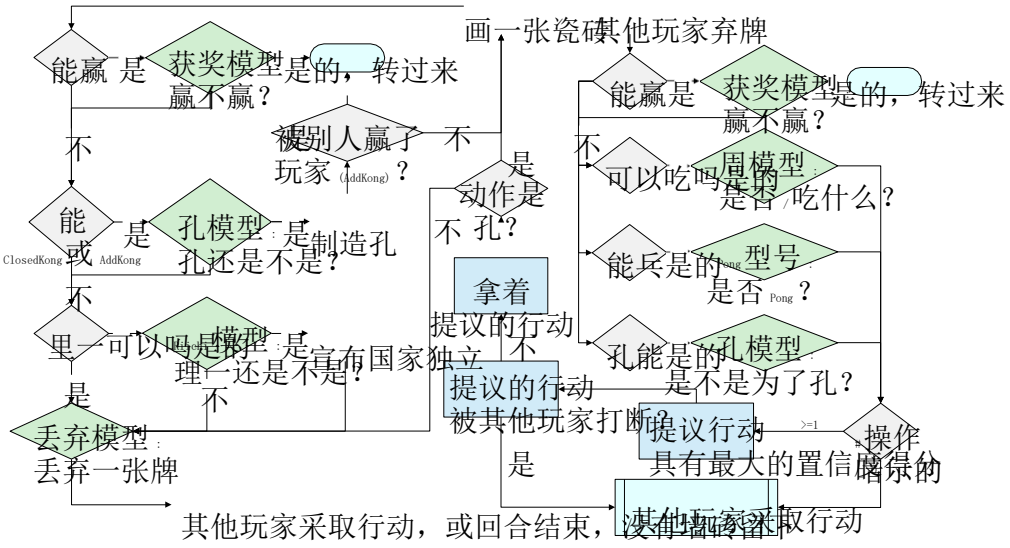

图 Suphx 的决策流程。

1 如果一轮结束时没有剩下墙砖，这也称为穷尽式抽牌。If

2.2 特征和模型结构

由于深度卷积神经网络(CNN)已经展示了强大的表示能力，并且在诸如国际象棋、shogi 和 Go 等游戏中得到验证，因此 Suphx 也采用深度 CNN 作为其策略的模型架构。

与围棋和国际象棋等棋盘游戏不同，信息(如图所示 2) 提供给麻将玩家的不是自然格式的图像。我们精心设计了一组特征，将观察到的信息编码成 CNN 能够消化的通道。

图 2:状态的例子。麻将的状态包含几种类型的信息:(1)牌组，包括私人牌、开放牌和 doras，(2)丢弃牌的顺序，(3)整数特征，包括四个玩家的累积回合分数和活动墙中剩余的牌的数量，以及(4)分类特征，包括回合 id、发牌者、重复发牌者的计数器和 Riichi 下注。

由于在日本 Mahong 中有 34 个唯一的图块，所以我们使用多个 34 1 通道来表示一个州。如图 2 所示 3, 我们使用四个通道来编码玩家的私人牌。开手牌、doras 和被丢弃的牌的序列被类似地编码到其他通道中。分类特征被编码到多个通道中，每个通道要么全为 0，要么全为 1。整数特征被划分到桶中，每个桶使用全为 0 或全为 1 的通道进行编码

除了可直接观察到的信息，我们还设计了一些外观-

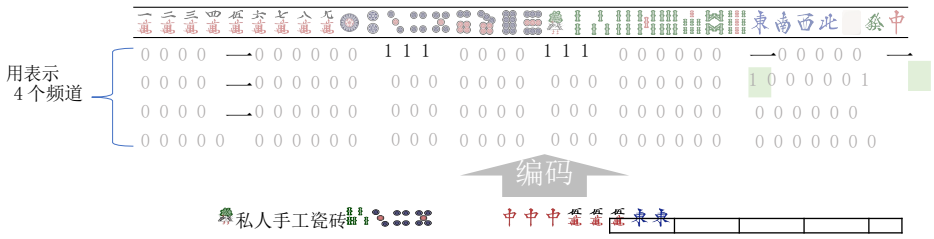

图 3:私有图块的编码。我们将玩家的私人牌编码成四个通道。有 4 行 34 列，每行对应
到一个通道，并且每列指示一种类型的图块。第 n 个通道中的第 m 列表示手中是否有 n 张第 m 种类型的牌。

提前特征，其指示如果我们从当前手牌中丢弃特定的牌，然后从墙上抽取牌来替换一些其他手牌，则赢得一手牌的概率和回合分数。在日本麻将中，一手赢了 14 张牌的牌包括四张牌和一对牌。有 89 种融合牌和 34 种对子，这导致了大量不同的可能获胜牌。此外，根据复杂的计分规则，不同的手牌会产生不同的获胜分数。[2]无法列举不同弃牌/抽牌行为和赢牌的所有组合。因此，为了降低计算复杂度，我们在提取前瞻特征时进行了若干简化:(1)我们执行深度优先搜索以找到可能的获胜手牌。(2)我们忽略对手的行为，只考虑自己代理人的抽牌和弃牌行为。通过这些简化，我们获得了 100 多个前瞻特征，每个特征对应于一个 34 维向量。例如，一个特征表示丢弃一个特定的牌是否可以导致一手 12，000 回合得分的获胜牌，用从墙上抽取的牌或其他玩家丢弃的牌替换 3 手牌。

在 Suphx 中，所有模型(即 discard/Riichi/Chow/Pong/Kong 模型)
使用类似的网络结构(图 4 和 5)，除了输入和输出层的尺寸(表 2).丢弃模型具有 34 个输出神经元，对应于 34 个独特的瓦片，Richii/Chow/Pong/Kong 模型仅具有两个输出神经元，对应于是否采取某个动作。除了状态信息和前瞻特征之外，Chow/Pong/Kong 模型的输入还包含关于用什么牌来制作 Chow/Pong/Kong 的信息。请注意，在我们的模型中没有池层，因为通道的每一列都有其语义含义，池化将导致信息丢失。

[2]不同得分模式的 2A 通用列表可在以下网址找到 Ahttp://arcturus.su/wiki/List_of_yaku

	抛弃	理一	食物	恶臭	孔
投入	34	838	34	958	
输出	34	2			

表 2:不同型号的输入/输出尺寸

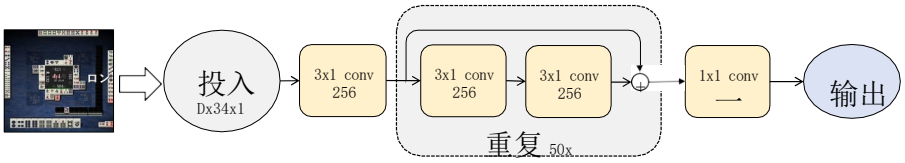

图 4:丢弃模型的结构

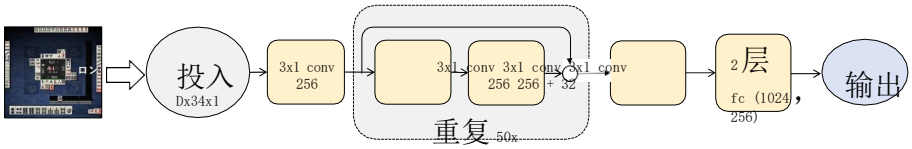

图 Riichi、Chow、Pong 和 Kong 模型的结构

3 学习算法

Suphx 的学习包含三个主要步骤。首先，我们通过监督学习训练 Suphx 的五个模型，使用从 Tenhou 平台收集的顶级人类玩家的(状态，动作)对。其次，以模型为策略，通过自玩强化学习对监督模型进行改进。我们采用流行的策略梯度算法(第 3.1) 并引入全局奖励预测(第 3.2)和 oracle guiding(部分 3.3) 处理麻将的独特挑战。第三，在在线游戏期间，我们采用运行时策略适应(第 3.4) 利用对当前回合的新观察，以便表现得更好。

3.1 基于熵正则化的分布式强化学习

Suphx 的训练基于分布式强化学习。特别地，我们采用策略梯度方法并利用重要性采样来处理由于异步分布式训练引起的轨迹陈旧性:

$$(\theta) = s, aE' \quad \frac{\pi\ \theta\ (a\ \ s)\ A\pi\ \theta\ (s, a),\ (1)}{\pi\ \theta'\ (a\ \ s)}$$

其中 θ’是生成训练轨迹的旧策略(的参数),
θ 是要更新的最新策略，Aπ θ (s，A)是动作 A 在 at 的优势

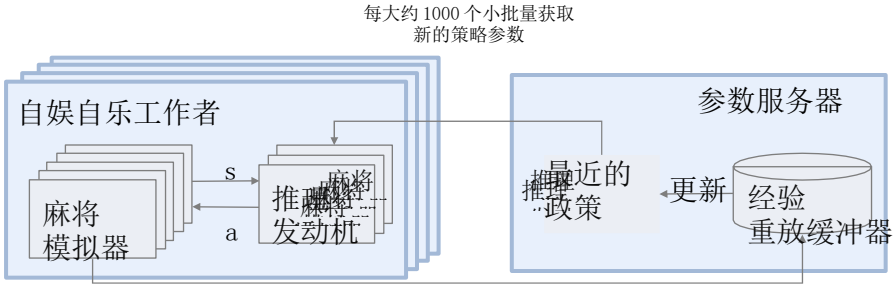

图 Suphx 中的分布式 RL 系统

关于政策 π θ 的状态 s。

我们发现 RL 训练对策略的熵很敏感。如果熵太小，RL 训练收敛快，自玩对策略没有显著改善；如果熵太大，RL 训练变得不稳定，并且学习的策略具有大的方差。因此，我们在 RL 训练期间如下正则化策略的熵:

$$\nabla J(\pi) = E_{\pi\ \theta(\text{对数}\ \pi}(a\ \ s)\ A\pi\ \theta\ (s,\ a) + \alpha \nabla H(\pi) \quad (2)$$

$$s,\ a)$$

$$\nabla$$

其中 H(ω θ)是策略 ω θ 的熵， α ＞ 0 是权衡系数。为了确保稳定的探索，如果我们的策略的熵在最近一段时间内小于/大于目标 Htarget，则我们动态地调整 α 以增加/减少熵项:

$$\alpha \leftarrow \alpha + \beta H\ target\ H(\omega\ \theta),$$

(3) 其中 H(ω θ)是最近一段时间轨迹的经验熵，以及
$\beta > 0$ 是一个小步长。

Suphx 使用的分布式 RL 系统如图所示 6. 该系统由多个自玩工作者组成，每个工作者包含一组基于 CPU 的麻将模拟器和一组基于 GPU 的推理机来生成轨迹。策略 ω θ 的更新与轨迹的生成是分离的:参数服务器用于使用基于重放缓冲器的多个 GPU 来更新策略。训练时，每台麻将模拟器随机初始化一个游戏，由我们的 RL 代理作为玩家和其他三个对手。当四个玩家中的任何一个需要采取行动时，模拟器将当前状态(由特征向量表示)发送到 GPU 推理引擎，然后该引擎将行动返回给模拟器。GPU 推理引擎定期从参数服务器获取最新的策略 ω θ，以确保自播放策略足够接近最新的策略 ω θ。

3.2 全球奖励预测

在麻将中，每局包含多轮，例如在 Tenhou 中有 8-12 轮。[3] 一轮从发给每个玩家 13 张私人牌开始，依次玩家抽牌和弃牌，这一轮结束，直到其中一个玩家完成一手好牌或者墙上没有牌，然后每个玩家得到一轮分数。例如，形成获胜手牌的玩家获得正回合分数，其他玩家获得零或负回合分数。当所有回合结束时，每个玩家根据累积回合分数的排名获得游戏奖励。玩家每轮结束时获得回合分数，8-12 轮后获得游戏奖励。然而，无论是回合分数还是游戏奖励都不是 RL 训练的好信号:

> 由于同一游戏中的多个回合共享相同的游戏奖励，使用游戏奖励作为反馈信号不能区分玩得好的回合和玩得差的回合。因此，应该更好地分别衡量每一轮的表现。
>
> 虽然回合分数是为每一回合计算的，但它可能无法反映动作的好坏，尤其是对顶级职业选手而言。例如，在游戏的最后一轮或两轮中，在累积回合分数方面领先很多的排名第一的玩家通常会变得更加保守，并且可能故意让排名第三或排名第四的玩家赢得这一轮，以便他/她可以安全地保持排名第一。也就是说，负的回合分数不一定意味着糟糕的政策:它有时可能反映了某些策略，因此对应于相当好的政策。

因此，为了给 RL 训练提供有效的信号，我们需要将最终的游戏奖励(一个全局奖励)恰当地归属于游戏的每一轮。为此，我们引入了一个全局奖励预测器 ϕ，它在给定当前回合和该游戏所有先前回合的信息的情况下，预测最终的游戏奖励。在 Suphx 中，报酬预测器 ϕ 是一个递归神经网络，更具体地说，是一个两层门控递归单元(GRU),后面跟着两个全连接层，如图所示 7.

这个奖励预测器 ϕ 的训练数据来自于天后顶级人类玩家的日志， ϕ 是通过最小化以下均方误差来训练的*新基*

$$\text{最小}\,1\sigma 1\sigma\,(\phi\,(x1,\ XJ)R)2, \tag{4}$$

其中 N 表示训练数据中的游戏数，Ri 表示第 I 个游戏的最终游戏奖励，Ki 表示第 I 个游戏的回合数，xk 表示第 I 个游戏的第 k 个回合的特征向量，包括

[3] 一场比赛的回合数是不固定的，取决于每一回合的输赢。The 更多详细信息，请访问 https://tenhou.net/man/。

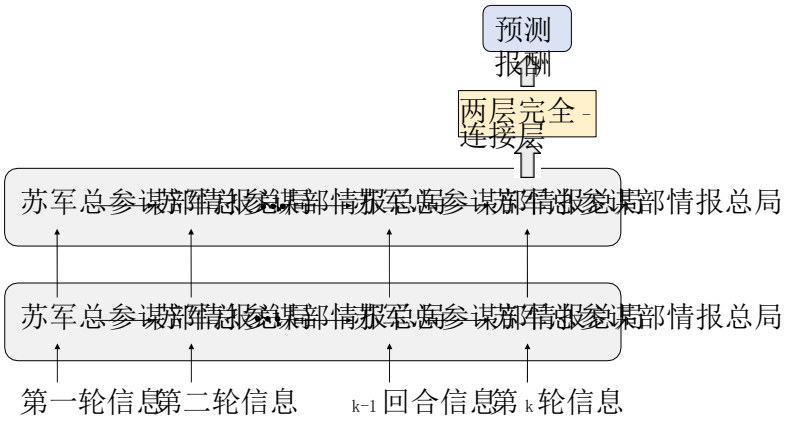

图 7:奖励预测:GRU 网络

本轮得分、当前累积回合得分、庄家位置、重复庄家和日市下注的计数器。

在 Φ 训练好的情况下，对于一个 K 回合的自玩游戏，我们用 Φ(xk) Φ(xk 1)作为 RL 训练第 K 回合的奖励。

3.3 Oracle Guiding

麻将中隐藏着丰富的信息(如其他玩家的私牌和墙牌)。没有这些隐藏的信息，就很难采取好的行动。这是麻将难打的一个根本原因。在这种情况下，虽然代理可以通过强化学习来学习策略，但是学习可能会非常慢。为了加速 RL 训练，我们引入了一个 oracle 代理，它可以看到关于一个状态的所有完美信息:(1)玩家的私人牌，(2)所有玩家的开放(以前丢弃的)牌，(3)其他公共信息，如累积的回合分数和 Riichi 赌注，(4)其他三个玩家的私人牌，以及(5)墙中的牌。只有(1)(2)和(3)可用于普通代理，[4]而(4)和(5)是附加的“完美”信息，只对 oracle 可用。有了(不正当的)获取完美信息的途径，甲骨文代理人经过 RL 训练后会很容易成为麻将高手。这里的挑战是如何利用 oracle 代理来指导和加速我们普通代理的培训。根据我们的研究，简单的知识提炼并不能很好地工作:对于一个只有有限信息访问的普通代理来说，很难模仿一个训练有素的 oracle 代理的行为，Oracle 代理超级强大，远远超出了普通代理的能力。因此，我们需要更聪明的方法来用神谕指引我们的普通代理人。

4 这里的“正常”是指代理人无法获得完美的信息。“Normal”

为此，可能有不同的方法。在 Suphx 中，我们所做的是首先通过强化学习训练 oracle 代理，使用包括完美特性在内的所有特性。然后，我们逐渐放弃完美的功能，以便 oracle 代理最终过渡到普通代理:

$$E_{s,a\sim\pi'}\left[\nabla_\theta \log \pi_\theta(a \mid x_n(s), x_o(s))\, A_{\pi_\theta}(x_n(s), x_o(s), a)\right], \quad (5)$$

其中 xn(s)表示状态 s 的正常特征，xo(s)表示状态 s 的附加完美特征，δt 是第 t 次迭代时的丢失矩阵，其元素是伯努利变量，P(δt(i，j) = 1) = γt，我们逐渐将 γt 从 1 衰减到 0。当 γt = 0 时，所有的完美特征都被丢弃，模型从 oracle 代理过渡到普通代理。

在 γt 变为零之后，我们继续对正常代理进行一定次数的迭代训练。在持续训练中，我们采用了两个技巧。首先，我们将学习率衰减到十分之一。第二，如果重要性权重大于预定义的阈值，我们拒绝一些状态-动作对。根据我们的实验，没有这些技巧，持续的训练是不稳定的，也不会导致进一步的提高。

3.4 参数蒙特卡罗策略适应

当一个顶级人类玩家的初始手牌(私人牌)发生变化时，他/她的策略会非常不同。例如，如果初始手牌很好，他/她会激进地玩以赢得更多，如果初始手牌不好，他/她会保守地玩以输掉更少。这与之前包括围棋、星际争霸在内的游戏有很大不同。因此，我们相信，如果我们能够在运行时适应离线训练的策略，我们可以构建一个更强大的麻将代理。

蒙特卡罗树搜索(MCTS)是一种在围棋(14)等游戏中用于提高运行时性能的成熟技术。不幸的是，如上所述，麻将的游戏顺序是不固定的，很难建立一个规则的游戏树。因此，MCTS 不能直接应用于麻将。在这项工作中，我们设计了一种新的方法，称为参数蒙特卡罗策略适应(pMCPA)。

当一轮开始，第一手私人牌发给我们的代理人时，我们将离线训练的策略调整到给定的第一手牌，如下所示:

1. 模拟:从瓷砖池中随机抽取三个对手的私人瓷砖和墙壁瓷砖(不包括我们自己的私人瓷砖),然后使用离线训练的策略推出并完成整个轨迹。这样总共生成了 K 条轨迹。
2. 适应:使用卷展轨迹执行梯度更新，以微调离线策略。
3. 推论:使用微调策略在本回合中与其他玩家对战。

设 h 表示我们的代理在一轮中的私人手牌， θ o 表示离线训练的策略的参数， θ a 表示适应于这一轮的新策略的参数。那我们有了

$$\theta a = \arg\max_{\theta} \sum_{\tau(h)} \sum_{\tau} \left(\frac{p(\tau;\ \theta)}{p(\tau;\ \theta o)} R(\tau), \right.$$

其中(h)是前缀为 h 的轨迹集， p(τ； θ)是政策 θ 生成轨迹 τ 的概率。

根据我们的研究，模拟/轨迹的数量 K 不需要非常大， pMCPA 不需要为这一轮的所有后续状态收集统计数据。由于 pMCPA 是一种参数方法，更新的策略(使用 K 个模拟)也可以导致在模拟中未访问的那些状态的更新的估计。也就是说，这种运行时适应有助于将我们从有限模拟中获得的知识推广到未知状态。

请注意，针对每一轮独立地执行策略调整。也就是说，在我们在当前回合中调整我们的代理的策略之后，对于下一回合，我们将再次从离线训练的策略重新开始。

4 离线评估

在本节中，我们通过线下实验报告 Suphx 各技术组件的有效性。

4.1 监督学习

在 Suphx 中，这五个模型首先分别通过监督学习进行训练。每个训练样本都是从人类职业选手那里收集的状态-动作对，状态作为输入，动作作为监督学习的标签。例如，对于丢弃模型的训练，样本的输入是状态的所有可观察信息(和头部特征)，标签是人类玩家采取的动作，即在该状态丢弃的牌。

表中报告了训练数据大小和测试精度 3.对于所有模型，验证数据和测试数据的大小分别为 10K 和 50K。由于丢弃模型解决了 34 类分类问题，我们为它收集了更多的训练样本。从表中可以看出，我们对 discard 模型实现了 76.7%的准确度，对 Riichi 模型实现了 85.7%的准确度，对 Chow 模型实现了 95.0%的准确度，对 Pong 模型实现了 91.9%的准确度，对 Kong 实现了 94.0%的准确度模特。我们还列出了以前的作品(6)所达到的精度作为参考。[5]

5 我们想指出的是，由于不同的训练/测试数据和模型结构，我们的数字不能与之前的数字直接比较。We would like to point out that our numbers are not directly comparable to previous

模型	训练数据大小	测量精度	以前的精确度(6)
丢弃模型	15M	76.7%	68.8 %
里一模型	5M	85.7 %	-
周模型	10M	95.0%	90.4%
Pong 模型	10M	91.9 %	88.2%
孔模型	4M	94.0 %	-

表 3:监督学习的结果

4.2 强化学习

为了展示 Suphx 中每个 RL 组件的价值，我们培训了几个麻将代理:

SL:监督学习代理。如前一小节所述，这个代理(具有所有五个模型)是以监督的方式训练的。

SL-weak:SL 代理的一个训练不足的版本，在评估其他代理时充当对手模型。

RL-basic:强化学习代理的基本版本。在 RL- basic 中，丢弃模型由 SL 丢弃模型初始化，然后通过以回合分数为奖励的策略梯度法和熵正则化进行提升。Riichi、Chow、Pong 和 Kong 模型与 SL 代理的模型保持相同。[6]

RL-1:用全局回报预测增强 RL-basic 的 RL 代理。奖励预测器通过使用来自 Tenhou 的人类游戏日志的监督学习来训练。

RL-2:通过 oracle guiding 进一步增强 RL-1 的 RL 代理。请注意，在 RL-1 和 RL-2 中，我们也仅使用 RL 来训练丢弃模型，而将其他四个模型保留为与 SL 代理的模型相同。

最初的私人牌具有很大的随机性，会极大地影响游戏的输赢。为了减少由初始私人牌引起的差异，在离线评估期间，我们随机生成了一百万个游戏。在这些游戏中，每个代理与 3 个 SL 弱代理进行游戏。在这样的设定下，一个代理的评测用了 20 个特斯拉 K80 GPUs 两天。对于评估指标，我们按照 Tenhou 规则计算了代理的稳定等级(参见附录 C)。为了减少稳定等级的方差，对于每个代理，我们从一百万个游戏中随机抽样 80 万个游戏，进行 1000 次。

数字 8 显示四分位数范围[7]超过 1000 人的稳定队伍-那些特工的电话。注意，为了公平比较，每个 RL 代理都经过培训

6 这四个模型也可以通过强化学习来改进，尽管不如丢弃模型重要。因此，为了减少训练时间，我们继承了 SL 模型。7 根据维基百科，“在描述统计学中，四分位距(IQR)是统计离差的一种度量，等于第 75 和第 25 之间的差 These

使用 150 万个游戏。每个代理的训练需要花费 44 个 GPU(参数服务器 4 个 Titan XP，自玩工作者 40 个 Tesla K80)和两天时间。可以看出，RL-basic 相对于 SL 有很好的改进，RL-1 优于 RL-basic，RL-2 相对于 RL-1 有额外的增益。这些实验结果清楚地证明了强化学习的价值，以及全局奖励预测和 oracle 指导的附加价值。

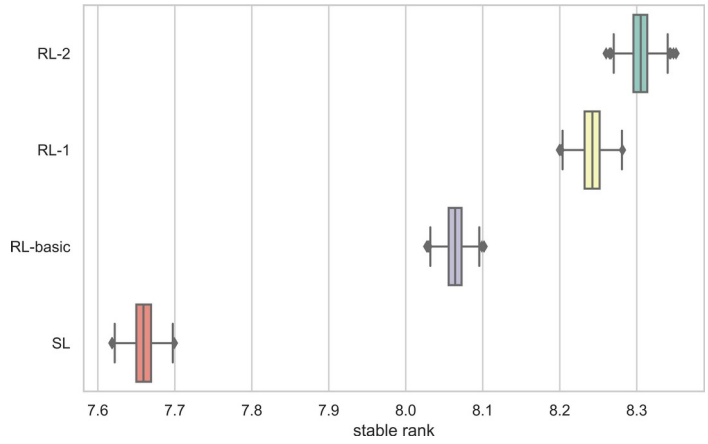

图 8:稳定排名过百万游戏统计。该图显示了分布的三个四分位值以及极值。“触须”延伸到位于上下四分位数 1.5 IQRs 内的点，然后独立显示超出该范围的观察值。

通过全局奖励预测器将游戏奖励分配给每一轮，训练有素的代理可以更好地最大化最终游戏奖励，而不是轮得分。例如，在图中 9，我们的经纪人(南方球员)在最后一轮比赛中凭借一手好牌大幅领先。根据目前四位玩家的累积回合分数，赢得这一回合只能获得边际奖励，而输掉这一回合将会受到很大的惩罚。因此，我们的代理人不是积极地玩来赢得这一轮，而是保守地玩，选择最安全的牌来丢弃，并最终获得这场游戏的第一名/排名。相比之下，RL-basic 弃另一张牌来赢得回合，这带来了失去整个游戏第 1 排名的很大风险。

百分位数，或上下四分位数之间，IQR = Q3 - Q1。换句话说，IQR 是第三个四分位数减去第一个四分位数；这些四分位数可以在数据的箱线图上清楚地看到。它是一个微调估计量，定义为 25%微调范围，是一个常用的稳健尺度。″

图 9:使用全局奖励预测，我们的代理人(南方玩家)在一局游戏的最后一轮中，当其累积回合分数大幅领先时，即使其手牌不错，且有一定概率赢得这一轮时，也会保守出牌。RL-basic 代理放弃红框牌以赢得这一轮，但放弃此牌是有风险的，因为在这一轮中没有任何玩家放弃相同的牌。相比之下，RL-1 和 RL-2 代理以防御模式进行游戏，并丢弃蓝框牌，这是一个安全的牌，因为相同的牌刚刚被西方玩家丢弃。

4.3 运行时策略适应的评估

除了测试对离线 RL 训练的增强，我们还测试了运行时策略适应。实验设置描述如下。

当一轮开始，私人牌发给我们的代理人时，

1. 数据生成:我们修复我们的代理的手瓷砖，并模拟 100K 轨迹。在每个轨迹中，其他三个玩家的手牌和墙牌是随机生成的，我们使用四份我们代理的副本来滚动和完成轨迹。
2. 策略调整:我们对离线培训的策略进行微调和更新

这些 100K 轨迹通过使用基本的政策梯度方法。

3. 调整后的策略的测试:我们的代理在另一个 10K 测试集上使用更新后的策略与其他三个玩家进行游戏，其中我们的代理的私有瓷砖仍然是固定的。由于我们的代理的初始私有瓦片是固定的，所以在这个测试集上的适配代理的性能可以告诉我们这样的运行时策略适配是否真的使我们的代理适配并更好地为当前私有瓦片工作。

请注意，由于部署和在线学习，运行时策略调整非常耗时。因此，在目前阶段，我们只在数百个初始回合上测试了这种技术。RL-2 的适配版本相对于其未适配版本的获胜率是 66%，这展示了运行时策略适配的优势。

策略调整使我们的代理更好地为当前的私人手牌工作，特别是在游戏的最后 1 或 2 轮。数字 10 显示了最后一轮游戏的示例。通过模拟，代理人了解到，虽然很容易以一个漂亮的回合分数赢得这一回合，但不幸的是，这不足以避免以第四名结束游戏。因此，在适应之后，代理人玩得更积极，承担更多风险，并最终以大得多的回合得分赢得回合，并成功避免以第 4 名结束游戏。

5 在线评估

来评价我们麻将 AI Suphx 的真实表现[8]，我们让它在 Tenhou.net 上玩，这是日本最受欢迎的在线麻将平台。天后宫有两个主要房间，专家室和凤凰室。专家室对 AI 和 4 丹及以上的人类玩家开放，凤凰室只对 7+丹的人类玩家开放。按照这个政策，Suphx 只能在专家室玩。

Suphx 在专家室玩了 5000+游戏，在记录排名方面达到了 10 dan[9]稳定排名 8.74 丹。[10]这是天后第一个也是唯一一个在记录等级上达到 10 级的 AI。

我们将 Suphx 与表中的几个 AI/人类玩家进行了比较 4:

Bakuuchi (10):这是东京大学基于蒙特卡洛模拟和对手建模设计的麻将 AI。它不使用强化学习。

8Suphx 相当于用 250 万左右的游戏训练出来的 RL-2。鉴于 Tenhou.net 对每个动作都有时间限制，运行时策略适应在 Tenhou.net 上测试时没有集成到 Suphx 中，因为它很耗时。我们相信运行时策略适应的集成将进一步改进 Suphx。Suphx

9 记录等级是玩家在 Tenhou 中获得的最高等级。如所示附录 B，玩家的排名是动态的，通常会随着时间而变化。比如他/她最近发挥不好，排名就会下降。

10 稳定等级的定义见附录 C。

图 10:在这个例子中，要走出游戏的第 4 名，代理人需要在这一轮中赢得 12，000 以上的回合分数。通过模拟，代理学习知道丢弃红框瓷砖容易赢得这一轮；但是对应的获胜回合分数会少于 12000。改编后，代理人丢弃蓝框瓷砖，这导致获胜概率较低，但一旦获胜，将获得 12，000 多个获胜回合分数。通过这样做，它承担了风险，并成功地从第四名中脱颖而出。

全国截肢残疾人高尔夫球协会[11]:这是 Dwango Media Village 基于深度卷积神经网络设计的麻将 AI。它也不使用强化学习。

我们也比较了 Suphx 和在记录排名方面达到 10 dan 的顶级人类玩家。为了公平起见，我们只比较了他们在达到 10 dan 后在专家室玩的游戏。由于这些顶尖的人类玩家大部分时间都呆在凤凰屋(部分是因为它更友好的计分规则),只有在他们达到 10 丹后才偶尔在专家屋玩，我们很难计算出一个可靠的稳定

[11]https://dmv.nico/ja/articles/mahjong_ai_naga/

给他们每个人排名。[12]因此，我们将它们视为一个宏观参与者，以进行统计上合理的比较。

我们可以看到，在稳定排位方面，Suphx 比 Suphx 之前最好的两个麻将 ai，Bakuuchi 和 NAGA，都要好 2 dan 左右。虽然这些顶尖的人类选手都取得了和 Suphx 一样的战绩排名(10 丹)，但是从稳定排名来说，都不如 Suphx 强。数字 11 绘制记录等级的分布[13]这表明 Suphx 在天厚 99.99%的人类玩家之上。

	#游戏	记录等级	稳定等级
巴库希	30,516	9 丹	6.59
全国截肢残疾人高尔夫球协会	9,649	8 丹	6.64
顶级人类	8,031	10 丹	7.46
Suphx	5,760	10 丹	**8.74**

表 4:与其他人工智能和顶级人类玩家的比较。

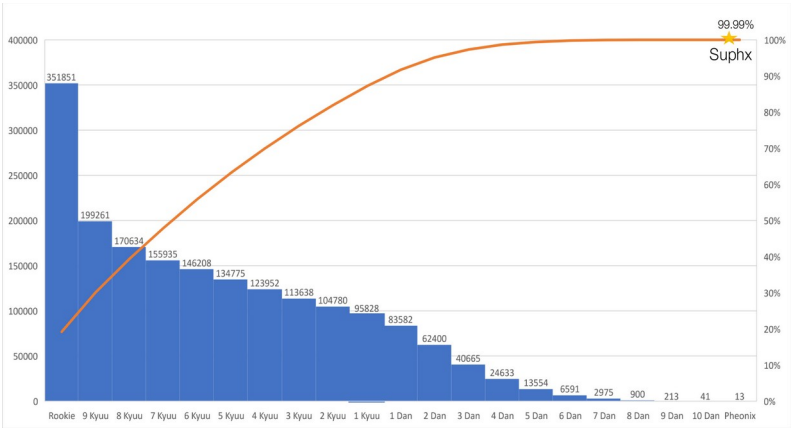

图 11:天后人类玩家记录等级分布。每一个条形表示天后中某一等级以上的人类玩家数量。

正如附录 B 中所讨论的，战绩排名有时候并不能反映一个玩家的真实水平:比如天厚历史上有 100+个战绩排名 10 丹的玩家，但是他们的真实水平可以相差很大。这

12 我们选择根据顶级人类玩家在专家室而不是凤凰室的游戏表现来比较他们的表现，因为这两个房间有不同的评分规则，稳定的等级不能直接比较。We chose to compare with the performance of top human players according to their game.

13Phoenix 在 Tenhou 是一个荣誉称号，当一个 10 段选手获得 4000 排名点。天厚历史上只有 13 位玩家(和 14 个账号)获得过这个 4 人麻将的荣誉称号。

稳定等级比记录等级更稳定(根据其定义)且粒度更细；然而，它也可能有很大的差异，尤其是当玩家在 Tenhou 没有玩够游戏次数时。因此，为了进行更丰富和可靠的比较，我们进行如下。对于每个人工智能/人类玩家，我们从他/她在专家室的日志中随机抽取 K 个游戏，并使用这些 K 个游戏计算稳定排名。我们进行 N 次这样的采样，并在图中显示每个玩家对应的 N 个稳定等级的统计数据 12.可以看出，Suphx 大大超过了其他两个人工智能和顶级人类职业球员的平均表现。

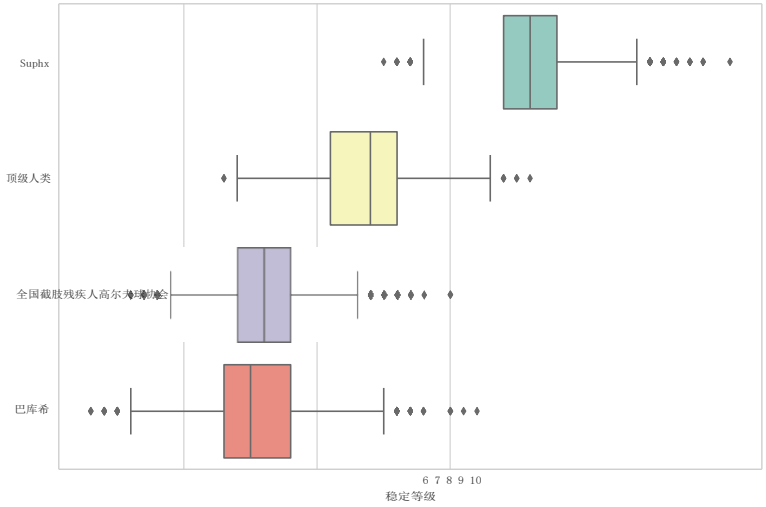

图 12:K = 2000，N = 5000 的稳定秩统计。

我们在表中进一步展示了这些人工智能/人类玩家的更多统计数据 5.我们从表格中得到了一些有趣的观察结果:

Suphx 防守很强，成交率很低。Suphx 上人类顶级玩家的评论也证实了这一点。[14]

Suphx 的 4 级率非常低，根据其评分规则，这是在天厚获得高稳定排名的关键。

Suphx 发展了自己的打法，得到了人类顶级玩家的认可。例如，Suphx 非常擅长保持安全的瓷砖，更喜欢

14 引自凤凰人类玩家的话可以在 Ahttps://twitter.com/Futokunaio_Sota/status/1142399895577325568

	第一 军阶	第二 军阶	第三 军阶	第四 军阶	胜利 速度	交易 速度
巴库奇	28.0%	26.2%	23.2%	22.4%	23.07%	12.16%
那加	25.6%	27.2%	25.9%	21.1%	22.69%	11.42%
顶级人类	28.0%	26.8%	24.7%	20.5%	-	-
上海	29.3%	27.5%	24.4%	**18.7%**	22.83%	**10.06%**

表 5:更多统计数据:等级分布和成功/成交率

用半同花赢了一手牌[15]等。数字 13 就是 Suphx 保持安全牌平衡未来攻防的一个例子[16]。

6 结论和讨论

Suphx 是迄今为止最强的麻将 AI 系统，也是日本著名麻将在线平台 Tenhou.net 第一个超越大多数顶级人类玩家的麻将 AI。由于麻将的复杂性和独特的挑战性，我们认为即使 Suphx 已经表现得非常好，但仍有很大的进一步改进空间。

我们在 Suphx 中引入了全局奖励预测。在当前系统中，奖励预测器将有限的信息作为其输入。显然，更多的信息会导致更好的奖励信号。比如一轮由于我们初始手牌运气好非常容易赢，赢了这一轮并不能体现我们政策的优越性，不应该奖励太多；相比之下，赢得一轮艰难的比赛应该得到更多的奖励。也就是说，在设计奖励信号时，应该考虑游戏难度。我们正在研究如何利用完美的信息(例如，通过比较不同玩家的私人初始手牌)来衡量一轮/一局游戏的难度，然后提高奖励预测值。

我们引入了 oracle guiding 的概念，并通过从 oracle 代理到普通代理的逐步过渡(通过完美的特性删除)实例化了这一概念。除此之外，可能还有其他方法来利用完美信息。例如，我们可以同时训练一个 oracle 代理和一个普通代理，让 oracle 代理向普通代理提取其知识，同时限制这两个代理之间的距离。根据我们的初步实验，这种方法也非常有效。再举一个例子，我们可以考虑设计一个 oracle critic，它提供更有效的

[15] https://en.wikipedia.org/wiki/Japanese_Mahjong_yaku

16 游戏回放可在以下网址找到 https://tenhou.net/3/?log=2019070722gm-0029-0000-3bee4a7e&tw=3。

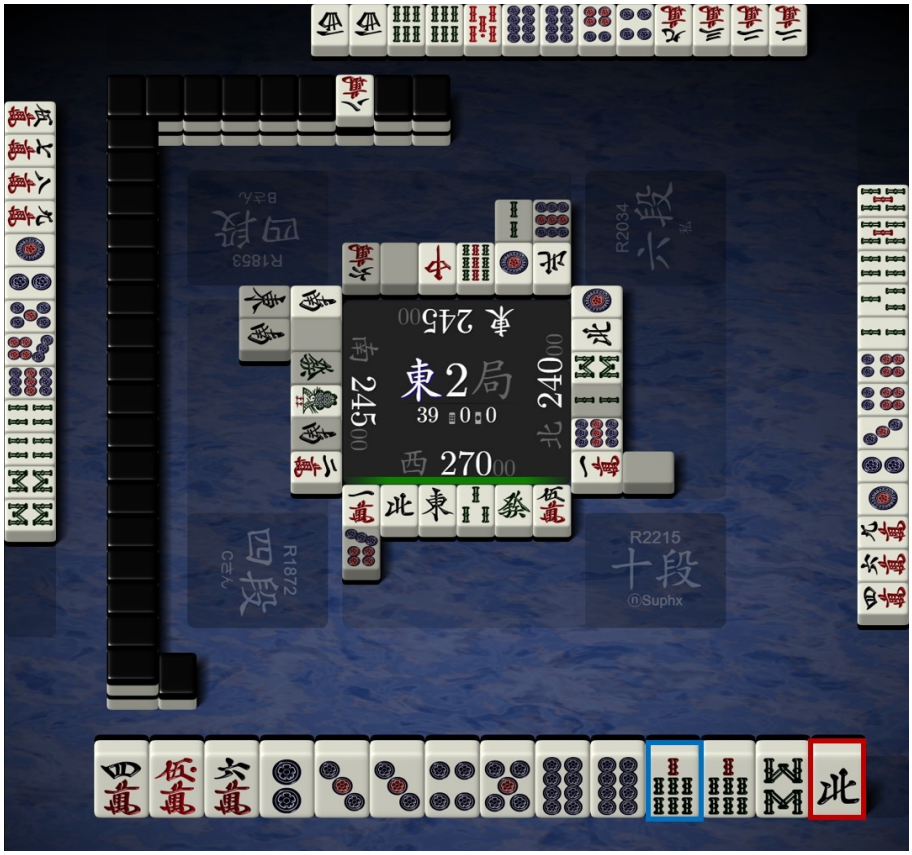

图 13: Suphx 在 st 状态下保持一个安全牌，以平衡未来的攻防。虽然在 st 状态下丢弃红框牌是安全的(实际上这确实是大多数人类玩家会丢弃的牌)，但 Suphx 将此牌保留在手中；相反，它会丢弃蓝框牌，这可能会减慢形成一手好牌的过程。这样做可以使未来的状态更加灵活，并且可以更好地平衡未来的攻击和防御。考虑一个未来状态 st+k，其中另一个玩家声明 Riichi，这是我们的代理人意想不到的。在这种情况下，Suphx 可以丢弃保存在状态 st 的安全牌，并且不会破坏它试图形成的获胜牌。相比之下，如果 Suphx 在状态 st 时弃掉该安全牌，则在 st+k 时它没有其他安全牌可弃，因此可能不得不打破其手牌中靠近获胜手牌的一张或一对牌，从而导致较小的获胜概率。

状态级即时反馈(而不是循环级反馈)加速了基于完美信息的策略函数的训练。

- 对于运行时策略适应，在当前的 Suphx 系统中，我们做到了

在每轮开始时，当私人牌发给我们的代理人时进行模拟。实际上，我们也可以在每个牌被任何玩家丢弃后进行模拟。也就是说，我们可以随着游戏的进行和越来越多的信息变得可以观察到，而不是只根据最初的手牌来调整策略。这样做应该能够进一步改善我们政策的执行情况。此外，由于我们逐渐调整我们的政策，我们不需要在每一步进行太多的抽样和推广。换句话说，我们可以将政策调整的计算复杂性分摊到整个回合中。由此，甚至有可能在具有负担得起的计算资源的在线游戏中使用策略适应。

Suphx 是一个不断学习和提高的代理。今天，Suphx 在 Tenhou.net 取得的成就仅仅是一个开始。展望未来，我们将为 Suphx 引入更多新颖的技术，并继续推动麻将人工智能和不完全信息游戏的前沿。

大多数现实世界的问题，如金融市场预测和物流优化，与麻将而不是围棋/象棋有着相同的特征
-操作/奖励规则复杂，信息不完善等。我们相信，我们在 Suphx 中为麻将设计的技术，包括全局奖励预测、oracle 引导和参数化蒙特卡罗策略适应，具有极大的潜力，有利于广泛的现实世界应用。

确认

我们真诚地感谢恒田顺吾和 Tenhou.net 为我们的实验提供了专业的游戏日志和在线平台。我们要感谢 Tenhou.net 的玩家和 Suphx 一起玩游戏。我们要感谢墨源帮助我们收集人类职业选手的统计数据。我们还要感谢我们的实习生和季晓红，以及我们的同事李亚涛、、、和，他们为开发 Suphx 的学习算法和训练系统做出了贡献。

参考

1. Christopher Berner、Greg Brockman、Brooke Chan、Vicki Cheung、Prze-mysaw d biak、Christy Dennison、David Farhi、Quirin Fischer、Shariq Hashme、Chris Hesse 等。Dota 2 与大规模深度强化学习。arXiv 预印本 arXiv:1912.06680，2019。

2. 迈克尔·鲍林、尼尔·伯奇、迈克尔·约翰逊和奥斯卡里·塔梅林。单挑限注德州扑克解决了。Commun。ACM，60(11):81–88，2017 年 10 月。

3. 诺姆·布朗和图马斯·桑德霍尔姆。用于单挑无限注扑克的超人 AI:Libra tus 击败顶级专业人士。科学，359(6374):418–424，2018 年 1 月。

4. 诺姆·布朗和图马斯·桑德霍尔姆。多人扑克的超人人工智能。
理科，365(6456):885–890，2019。

5. 欧洲麻将协会。日本麻将的规则。http: //mahjong-europe.org/portal/images/docs/Riichi-rules-2016-EN.pdf。

6. 高士奇，奥谷文典，川原义弘，鹤冈义正。麻将游戏中不完全信息数据的监督学习
通过深度卷积神经网络。日本信息处理学会，2018。

7. 高士奇，奥谷文典，川原义弘，鹤冈义正。通过深度卷积神经网络构建计算机麻将播放器。2019 年 6 月。

8. 栗田墨玉和国仁霍基。多人麻将游戏中抽象马尔可夫决策过程的人工智能玩家构造方法。2019 年 4 月。

9. n 水见和 Y 鹤冈。建立一个基于蒙特卡罗模拟和对手模型的电脑麻将游戏。2015 年 IEEE 计算智能和游戏会议(CIG)，第 275–283 页，2015 年 8 月。

10. 水上直树和鹤冈义正。建立一个基于蒙特卡罗模拟和对手模型的电脑麻将游戏。2015 年 IEEE 计算智能和游戏会议(CIG)，第 275–283 页。IEEE，2015。

11. Matej Morav、Martin Schmid、Neil Burch、Viliam Lisý、Dustin Morrill、Nolan Bard、Trevor Davis、Kevin Waugh、Michael Johanson 和 Michael Bowling。DeepStack:单挑无限注扑克中的专家级人工智能。科学，356(6337):508–513，2017 年 5 月。

12. 姜戎、陶秦和薄安。基于深度神经网络的竞争性桥牌投标。《第 18 届自主智能体和多智能体系统国际会议论文集》，第 16-24 页。自主智能体和多智能体系统国际基金会，2019。

13. David Silver、Aja Huang、Chris J、Arthur Guez、Laurent Sifre、George van den Driessche、Julian Schrittwieser、Ioannis Antonoglou、Veda Panneershelvam、Marc Lanctot、Sander Dieleman、张秀坤·格雷韦、John Nham、Nal Kalchbrenner、Ilya Sutskever、Timothy Lillicrap、Madeleine Leach、Koray Kavukcuoglu、托雷·格雷佩尔和戴密斯·哈萨比斯。掌握
具有深度神经网络和树搜索的围棋游戏。自然，529(7587):484–489，2016 年 1 月。

14. David Silver，Aja Huang，Chris J，Arthur Guez，Laurent Sifre，George Van Den Driessche，Julian Schrittwieser，Ioannis Antonoglou，Veda Panneershelvam，Marc Lanctot，等人用深度神经网络和树搜索掌握围棋。自然，529(7587):484，2016。

15. David Silver、Thomas Hubert、Julian Schrittwieser、Ioannis Antonoglou、Matthew Lai、Arthur Guez、Marc Lanctot、Laurent Sifre、Dharshan Kumaran、托雷·格雷佩尔、Timothy Lillicrap、卡伦·西蒙扬和戴密斯·哈萨比斯。A
通用强化学习算法，掌握国际象棋，日本象棋，并通过自我发挥。科学，362(6419):1140–1144，2018年12月。

16. David Silver、Julian Schrittwieser、、Ioannis Antonoglou、Aja Huang、Arthur Guez、Thomas Hubert、Lucas Baker、Matthew Lai、Adrian Bolton、陈玉田、Timothy Lillicrap、范辉、Laurent Sifre、George
范登德里斯切，托雷格雷佩尔和戴密斯·哈萨比斯。在没有人类知识的情况下掌握围棋。自然，550(7676):354–359，2017年10月。

17. 理查德·萨顿、戴维·麦卡勒斯特、萨廷德·辛格和伊沙伊·曼苏尔。函数逼近强化学习的策略梯度方法。神经信息处理系统进展，1057-1063页，2000。

18. 杰拉尔德·泰索洛。时间差学习和TD-Gammon。社区。美国计算机学会，38(3):58-68，1995年。

19. 信吾常田。天厚。https://tenhou.net/。访问时间:2019-6-17。

20. Dwango媒体村。NAGA:深度学习麻将AI。https://dmv.nico/ja/articles/mahjong_ai_naga/。访问时间:2019-6-29。

21. Oriol Vinyals，Igor Babuschkin，Wojciech M Czarnecki，Michal Mathieu，Andrew Dudzik，Junyoung Chung，David H Choi，Richard Powell，Timo Ewalds，Petko乔尔杰夫等.星际争霸2中使用多代理强化学习的特级大师级别。自然，575(7782):350–354，2019。

附录 A:麻将规则

麻将是一种基于瓷砖的游戏，数百年前在中国发展起来，现在在世界范围内流行，拥有数亿玩家。麻将游戏本身在世界各地有许多变化，不同地区的麻将在规则和使用的牌上都不同。考虑到日本麻将在职业联赛中非常受欢迎，在本研究中，我们将重点放在 4 人日本麻将(日式麻将)上[17]日本顶尖选手的名单
并且它的规则(5)被清楚地定义并且被很好地接受。由于日本麻将的玩法/计分规则非常复杂，而这项工作的重点不是
为了对它们进行全面的介绍，这里我们做了一些选择和简化，并对那些规则做了简要的介绍。全面的介绍可以在麻将国际联赛规则(5)中找到。

日本麻将有 136 张牌，由 34 种不同的麻将牌组成
瓷砖，每种四块。这 34 张牌由三种花色组成，分别是竹子、人物和圆点，每种花色从 1 到 9，还有 7 张不同的荣誉牌。一个游戏包含多个回合，当一个玩家失去所有点数，或者触发一些获胜条件时，游戏结束。

在一场游戏中，每个玩家将从 25，000 点开始，四个玩家中的一个被指定为庄家。在每一轮开始时，所有的牌都被洗牌并排列成四面墙，每面墙有 34 张牌。52 张牌分发给 4 个玩家(每个玩家 13 张，作为他/她的私人手牌)，14 张牌形成死墙，除非玩家宣布空牌并抽取替换牌，否则永远不会使用，其余 70 张牌形成活墙。4 名玩家轮流抽牌和弃牌。[18]首先用至少一个 yaku 组成一手完整牌的玩家赢得这轮游戏，并获得由奖励规则计算的特定回合分数:

- 一手完整的牌是一套 4 张融合牌加上一对。一手牌可以是 Pong(三张相同的牌)、Kong(四张相同的牌)和 Chow(同一花色的三张连续的简单牌)。这对由任意两个组成
相同的瓷砖，但它不能与四个融合。一个玩家可以用(1)他/她自己从墙上抽出的一张牌做一个 Chow/Pong/Kong，在这种情况下，这张牌对其他人是隐藏的，或者(2)其他玩家丢弃的一张牌，在这种情况下，这张牌对其他人是暴露的。如果
一个洞是由从墙上画下来的瓷砖做成的，它被称为闭孔。玩家出了一个孔后，需要从死墙上额外抽一张牌来替换。一个名为 AddKong 的特例是，当玩家抽取与他/她所拥有的一个暴露的 Pong 相匹配的牌时，将一个暴露的 Pong 转换为 Kong。

- 一个雅库是玩家牌的某个模式或者某个特殊条件。雅库是决定回合得分的主要因素，雅库的数值因模式不同而不同。一手好牌可能包含几种不同的

[17] https://m-league.jp/
[18] 回合可以被空/空打断，并宣布赢一手牌。

	基本排名分数。	第一第二第三第四层以上				水平下降			
新手	0								0 20
9Kyu	0	+				20 +10 0 20 -			0
8Kyu	0	@正常				房间@普通房间			
		7Kyu	0			0	20	-	
		6Kyu	0			0	40	-	
4ku	0	5Kyu	0			0	60	-	
3Kyu	0	+40		+10					
2Kyu	0	@高级房间		@高级房间					
1ku	0								
1丹	200				+0				
2丹	400								
3丹	600	+50		+20					
		@专家室		@专家室					
		丹 800 -90 1600 是							
		丹 1000 -105 2000 是							
		6丹 1200 -120 2400 是							
7丹	1400								
8丹	1600	+60		+30		-135	2800	是	
		@凤凰房		@凤凰房		-150	3200	是	
		9丹	1800			-165	3600	是	
		10丹	2000			-180	4000	是	
		凤凰			荣誉称号			-	

表 6: Tenhou 排名系统:不同级别及其要求

雅库和最后一轮的分数将累积在所有的雅库手里。日本麻涌的不同变种有不同的雅库模式。一个普通的雅库列表包括 40 种不同的类型。此外，多拉是一种特殊的瓷砖，在抽瓷砖之前通过掷骰子来确定，它提供额外的点数作为奖励。

玩家与他/她的私人牌一起形成获胜牌的最后一张牌可以来自(1)他/她自己画的墙上的牌，或者(2)其他玩家丢弃的牌。对于第一种情况，所有其他玩家将向获胜者扣分。对于第二种情况，弃牌的玩家将会输给赢家。

一个特别的 yaku 是玩家可以在他/她的手牌离获胜的手牌只有一张牌的时候宣布 Riichi。一旦宣布一手牌，玩家只能从自己抽取的牌或其他玩家丢弃的牌中选择一手获胜牌，并且不能再改变他/她的手牌。[19]

玩家从游戏中获得的最终排名分数由他/她的级别和他/她在多轮游戏中累积的回合分数排名决定，如表所示 6.

附录 B: Tenhou 排名规则

天厚使用日本武术排名系统[20]，也就是从菜鸟开始，9 kyuu 降到 1 kyuu，然后 1 dan 到 10 dan。玩家赢/输

[19] http://mahjong.wikidot.com/riichi20https://en.wikipedia.org/wiki/Dan_(rank)

当他们赢/输排名游戏时的排名点数。赢得或输掉的金额取决于游戏结果(范围从 1 到 4)、玩家的当前级别以及玩家所在的房间。天厚有四种房型，普通房，高级房，专家房，凤凰房。输掉游戏的惩罚(即负排名点数)在这些房间中是相同的，但是赢得游戏的奖励(即正排名点数)是不同的。排名系统的设计是为了惩罚高水平的玩家，如果他/她输了一场比赛:一个高水平的玩家在第四场比赛中会比一个低水平的玩家失去更多的分数，例如，一个 10 段棋手 vs
-6 段选手 120。

当玩家玩游戏时，他/她从游戏中赢得/失去一些排名点数，并且他/她的总排名点数改变。如果总排名分数增加并达到下一级的要求，他/她的排名增加 1 级；如果总排名点减少到 0，他/她的排名减少 1 级。当他/她的等级改变时，他/她将在新的等级获得初始等级点数。表中列出了不同房间和级别的排名点详情 6.因此，玩家的等级不稳定并且经常
随着时间的推移而改变。我们用记录等级来表示玩家在 Tenhou 中达到的最高等级。

附录 C:稳定等级

Tenhou 用稳定排名来评价一个球员的长期平均表现。专家室中的稳定排名计算如下。[21]

设 n1 表示游戏者获得最高累积回合分数的游戏次数，n4 表示他/她获得最低累积回合分数的游戏次数，n2 和 n3 表示第二/第三高累积回合分数的游戏次数。那么玩家在 dan 方面的稳定排名是

$$\frac{5 \quad n1 + 2 \quad n2}{n4} - \qquad 2.(7)$$

由于游戏的累积回合分数不仅取决于玩家的技能，还取决于四个玩家的私人牌和墙牌，由于隐藏信息的随机性，稳定的排名可能具有很大的变化。再者，在天后玩的时候，对手是由天后系统随机分配的，这就带来了额外的随机性。于是，对于一个天厚的玩家来说，通常假设至少需要几千场才能得到一个相对可靠的稳定排名。

[21] https://tenhou.net/man/#RANKING