arXiv:2301.08144v1 [cs.IR] 17 Jan 2023

面向协作式人类-AI任务的以用户为中心的策略推荐系统设计

Lakshita Dodeja*, **Pradyumna Tambwekar***, **ErinHedlund-Botti**, **Matthew Gombolay**佐治亚理工学院交互计算学院，亚特兰大，乔治亚州 {ldodeja3, pradyumnat, erin.botti}@gatech.edu, matthew.gombolay@cc.gatech.edu

摘要

人类正借助人工智能协作解决搜救、制造业等领域的复杂任务。理解用户偏好，并向人类推荐解决特定任务的不同策略，有助于实现高效的团队合作。以往研究主要关注在电子商务或社交网络背景下，为相对成熟的任务个性化推荐系统。本文旨在探究设计以用户为中心的决策策略推荐系统时需要考虑的重要因素。我们开展了一项人类受试者实验（n=60），测量具有不同人格类型的用户对不同策略推荐系统的偏好。实验涵盖了以往研究确立的四种策略推荐模式：（1）单一策略推荐；（2）多个相似推荐；（3）多个多样化推荐；（4）所有可能策略的推荐。尽管以往研究分别探讨过这些策略推荐方式，本研究的创新之处在于同时采用所有这些方式，并将其用于策略推荐，从而深入了解人们对不同策略推荐系统的感知。我们发现，某些人格特质（例如尽责性）会显著影响用户对特定系统类型的偏好（$p < 0.01$）。最后，我们报告了可用性、对齐和感知智能之间一种有趣的关系：推荐与用户自身偏好的感知对齐程度越高，感知智能（$p < 0.01$）和可用性（$p < 0.01$）也越高。

关键词： 智能 用户界面 · 交互式决策支持系统 · 创新设计与评估 交互式系统

1 引言

人工智能系统能力的不断提升，使其在许多领域得到广泛应用。然而，在搜救（Murphy，2004年；Heintzman 等人，2021年）、航空（Li 等人，2021年）和医疗保健（Magrabi 等人，2019年）等安全关键领域，相较于完全依赖人工智能系统，人类与人工智能合作更受青睐。为了实现有效协作，人类需要能够说明自己的意图，即人工智能系统应如何执行协作任务，从而使人类与人工智能伙伴朝着相同的目标努力（Gombolay 等人，2017年；Tabrez 等人，2020年；Nikolaidis 和 Shah，2012年；Johnson 和 Bradshaw，2021年）。另一方面，由于缺乏将意图整理成适当格式所需的技术专长，人类未必总能将自己的意图表达为可执行指令。在这种情况下，人工智能系统可以询问用户的内在偏好，并提出一项可执行策略供人类接受、拒绝或修改，这或许会有所帮助。这一过程将使非专家也能协调

*作者贡献相同

他们试图协作完成任务时所采用的策略。受先前研究启发，我们将策略表述为完成任务所需的目标和约束（Tambwekar 等人，2022年）。在本研究中，我们使用“策略推荐系统”一词来定义这类系统。

要设计此类策略推荐系统，我们需要考虑并纳入最终用户的功能性需求和倾向性需求。关于推荐系统和人机协作的既有研究阐明了此类系统采用以用户为中心的设计以促成有效协作的必要性。研究表明，用户更不愿意使用通用人工智能系统，因为这类系统无法满足用户独特的特征和需求（Longoni 等人，2019）。个性化推荐是使人工智能系统能够提供针对特定用户的建议的一种方法（Huang 等人，2022）。事实证明，个性化已能有效提升消费者满意度（Xiao 等人，2019）和企业收入（Behera 等人，2020）。此前，基于用户偏好的推荐系统内容个性化是通过协同过滤、基于内容的过滤和混合方法实现的（Kumar 和 Thakur，2018；He 等人，2017；Schafer 等人，2007；Lops 等人，2011）。

开发以用户为中心的推荐系统，不仅涉及对推荐内容进行个性化，还包括通过个性化用户界面或交互过程向用户呈现内容的方式。尽管内容个性化方面的算法取得了这些进展，人类仍有细致多样的需求，这些需求会影响他们与推荐系统的交互。有时，人们会觉得，获得与自身需求和偏好相关的项目推荐很有帮助，例如在电子学习（Tarus 等人，2018年）和电子商务（Linden 等人，2003年）中。另一些时候，人们似乎更喜欢收到一组多样的选项供其选择（Kunaver 和 Požrl，2017年）。研究还表明，人格特质会影响人们对推荐中项目特定排列结构的偏好（Nguyen 等人，2018年）。例如，Nguyen 等人发现，内向程度较高的用户更偏好多样化推荐，而非相似推荐。此外，开放性较低的用户更偏好与自己以往消费的数据更一致的推荐。然而，这些发现未必适用于人机协作推荐系统。因此，研究推荐中项目的结构如何影响人们对策略推荐系统的感知十分重要。

本文提出了一项新颖的人类受试者实验，以了解在安全关键领域中，为人类-AI协作设计有效策略推荐系统时需要考虑的因素。我们的研究旨在探讨如何为人类个性化地收集策略推荐。以往研究在推荐系统中探讨个性化，关注的是推荐内容本身的相关性；与之不同，我们旨在了解如何最好地选择策略推荐的组合，以适应用户的性格与性情。我们希望确定策略推荐的性质和结构是否会影响用户对系统的感知。为此，我们开发了一项经验证的度量指标，用于将用户的游戏偏好转化为可执行策略，并运用该指标根据用户的偏好向其提供推荐。在实验中，我们借鉴以往研究（Tan 等人，2008年；Szpektor 等人，2013年；Bollen 等人，2010年）提出的四种基线方案，但采用被试间设计，并置于策略推荐系统的情境中：（1）推荐单一策略，该策略与用户偏好的对齐程度最高；（2）推荐多个与用户偏好相似的策略；（3）推荐多个策略，其中既包括与用户偏好相似对齐的策略，也包括与其相反对齐的策略；（4）推荐全部可能的策略。被试间设计的优势在于，它使我们能够通过相同的指标比较所有策略推荐系统。我们研究人格类型以及偏好特定类型推荐的倾向如何影响用户对任一策略推荐类型的对齐程度、偏好和感知智能。我们的主要贡献如下：

1. 我们设计了一项新颖的用户研究，以了解用户对高难度任务策略推荐的偏好。2. 我们开发并验证了一项指标，只需三个问题即可将用户偏好的游戏风格转化为可执行的策略（$p < 0.05$）。3. 我们从可用性和系统感知智能、任务工作负荷，以及推荐与用户策略的对齐程度等方面评估用户偏好（$p < 0.01$）。4. 我们研究推荐列表的感知对齐程度如何影响人们对某类推荐的整体感知。我们发现，推荐选项与用户偏好的感知对齐程度会显著影响系统的可用性（$p < 0.01$）和感知智能（$p < 0.01$）。

我们开展了首个此类人类受试者实验，深入了解人们在与用于决策的策略推荐系统交互时，哪些因素会影响其偏好。这些见解为设计能够适应用户偏好的策略推荐系统奠定了基础。

2 相关工作

本节将首先介绍推荐系统中现有的个性化方法，随后探讨不同研究如何考察推荐系统中项目的呈现方式对用户以及用户对推荐系统的感知所产生的影响。

2.1 推荐系统与个性化 n

推荐系统可以根据任务需求进行个性化或非个性化处理。推荐中的个性化主要通过协同过滤（Sarwar et al., 2001; Kluver et al., 2018）和基于内容的过滤（Van Meteren and Van Someren, 2000; Vanetti et al., 2010）实现。协同过滤会寻找另一位偏好相似的用户，并根据相邻用户的偏好进行推荐；基于内容的过滤则根据用户过去的偏好进行推荐（Koren et al., 2022）。协同过滤可通过邻域方法或潜因子模型实现。邻域方法通过寻找相似用户或相似项目发挥作用，而潜因子模型则尝试将用户和项目分解到同一嵌入空间中，以发现相似特征。基于内容的过滤涉及发现新的数据源（如关联开放数据、用户生成内容），或采用新的算法方法（如基于元路径的方法、元数据编码和深度学习）（Lops et al., 2019）。显式项目评分、自适应对话和表单，以及基于比较的技术，均已用于获取用户偏好。可视化、解释和用户反馈等交互技术还有助于进一步根据用户偏好调整推荐界面（Jugovac and Jannach, 2017）。推荐系统的个性化可以通过用户界面、内容或交互过程实现（Zanker et al., 2019）。与其研究呈现给用户的内容在算法准确性方面的表现，我们的研究旨在探究如何为策略推荐系统实现用户界面的个性化。

2.2 推荐中条目的呈现

推荐系统通常会呈现符合用户偏好的选项（Tan 等人，2008；Chen 和 Yin，2006；Linden 等人，2003），但在某些情况下，呈现与用户偏好不同的选项也可能有所裨益。研究证明，推荐中的“新颖性与多样性”有助于提升个性化问题推荐系统的性能（Szpektor 等人，2013）。一项关于音乐推荐系统的最新研究也指出，需要采用能够考虑多样性的算法，换言之，算法应能向用户推荐相关且多样的音乐选项（Anderson 等人，2020）。

既有研究还发现，算法缺乏透明度且难以预测，有时会让用户在使用自动化系统时感到无助（Jhaver 等人，2018）；而研究证明，系统的专业水平有助于提升人们对系统的认可（Hou 和 Jung，2021）。因此，展示各种策略或许能帮助用户更好地理解推荐系统，进而减轻使用系统时的焦虑。反过来，这也可能导致“选择过载”（Bollen 等人，2010）。当推荐系统提供大量优质选项供用户选择时，就会出现选择过载。研究表明，随着选项数量增加，用户满意度会略有提升，随后便开始下降（Reutskaja 和 Hogarth，2009）。

在我们的研究中，我们在设计向用户呈现的四种推荐类型时，考虑了这些现象：即单一对齐推荐、相似的对齐推荐、由相似和相反对齐策略共同组成的三种多样化推荐，以及全部可能的推荐。然而，与以往的研究不同，我们同时比较了所有类型的推荐系统，并将研究置于策略推荐的背景下。这使我们能够比较和对照不同的策略推荐系统，以及不同用户对它们的感知。我们还分析了用户的人格特质与其对特定推荐系统的偏好之间的相关性。我们希望研究，关于多样性、焦虑和选择过载的既有发现，是否会体现在用户与策略推荐的交互中。

2.3 研究人们对推荐系统的感知

推荐系统不仅应能够准确估计两个项目之间的相似与差异，还应能够生成以用户为中心的建议。人类-推荐系统交互这一概念主张从终端用户的角度重新设计推荐系统，以更好地满足他们的需求（McNee 等，2006）。此后，人们开展了许多用户研究，以了解用户如何看待不同的推荐系统，以及推荐系统会如何影响用户的决策。相关研究探讨了推荐列表应如何呈现、列表是否应个性化以及适用于哪些情境（Tam 和 Ho，2003），以及列表是否应具有多样性（Willemsen 等，2016；Ziegler 等，2005）。以往的研究还开展了

实验，分析实体机器人和屏幕上的智能体所提供的推荐对人类决策的影响（Shinozawa 等人，2005）。研究表明，在传达推荐时，三维形象并非总是优于二维形象，这也取决于所使用的交互环境。

研究人员还制定了各种指导方针，以帮助开展这些用户研究，确定要研究的系统方面、所用因变量等（Knijnenburg 和 Willemsen，2015）。人类受试者实验也是评估推荐系统有效性的重要方法（Shani 和 Gunawardana，2011）。我们可以向人类推荐系统生成的项目，并检查人们是否认为这些项目比随机推荐的项目或其他算法推荐的项目更好，从而测试新系统。我们希望借鉴这些先前实验的见解，设计我们的人类受试者实验，并分析人们对策略推荐系统的感知和可用性。在我们的研究中，我们开展了首个此类人类受试者实验，以评估并了解一种以用户为中心的策略推荐系统。

3 实验设计

我们设计了一项新颖的人类受试者研究，以了解用户对游戏玩法策略推荐的偏好。通过这项研究，我们希望了解哪些因素会影响人们喜欢以何种方式获得策略推荐，例如他们对特定推荐类型的先天倾向，或其个性。我们是首个研究用户对定义明确的任务（如电子商务或社交媒体推荐）以外的推荐系统感知的实验。我们希望为面向人机协作的推荐系统提供设计见解，使人工智能系统能够提出策略推荐，从而减轻人类协作者的认知负荷。

我们的策略推荐以目标和约束来定义。目标是指希望实现的理想状态集合，而约束则是在实现这些状态时施加的条件（策略示例见图2）。我们采用目标和约束这一框架，是因为它们是对人工智能智能体行为进行编程的有效机制。它们可以轻松应用于学习和基于规划的方法，以指定智能体的任务。对于一项策略，六个目标中的每一个都可取“不利”“中性”或“有利”三个值，用以表示该目标对整体策略的重要性。约束则以独立陈述的形式表示，例如“我需要4支部队才能有效防守一个国家”或“我必须守护阿斯加德的边界”。

3.1 环境

我们的实验采用了桌游《风险》。由于在《风险》游戏中可以采用多种截然不同的策略来赢得游戏，因此《风险》是本研究的理想环境。该环境也足够复杂，因而有必要使用策略推荐系统：如果没有任何推荐，玩家就需要玩上好几轮，才能总结出自己的策略。玩《风险》涉及完成各种资源分配、调度和规划任务，而这些都是许多现实世界任务的关键组成部分。此外，《风险》是一个随机环境，与金融交易、灾害应对、搜救、机器人操作等现实世界情境相契合。与《星际争霸》或《帝国时代》等其他调度或资源分配类游戏不同，人们能够更直观地为《风险》制定并理解策略。《风险》的规则集明显更简单，因此人们无需具备深厚的领域专长，也能制定和理解策略。这些特性使《风险》成为适合的环境，可用于了解为现实世界任务推荐人工智能策略的潜在趋势。

我们采用的 Risk 版本是一款由三名玩家参与的回合制游戏，包含四个主要阶段：

1. **选地**：在游戏地图上选择你的初始领土，并部署最初分配给你的部队。 2. **增援**：向你现有的领土部署额外部队。 3. **进攻**：与敌方领土交战，进攻你控制的领土。 4. **调动**：在你控制的两个领土之间调动部队。

每位玩家只在游戏开始时进行一次选地。在所有玩家完成初始征兵阶段后，每位玩家的回合由依次完成增援、进攻和调动 阶段组成。玩家征服地图上的全部领土后即获胜。在本实验中，我们改编了先前研究 Tambwekar 等人（2022年）中的 Risk 游戏模拟器。我们在该模拟器中创建了游戏代理，使其能够遵循本研究制定的策略，以便向参与者实际展示部署任一策略的效果。我们的 Risk模拟器如图1所示。

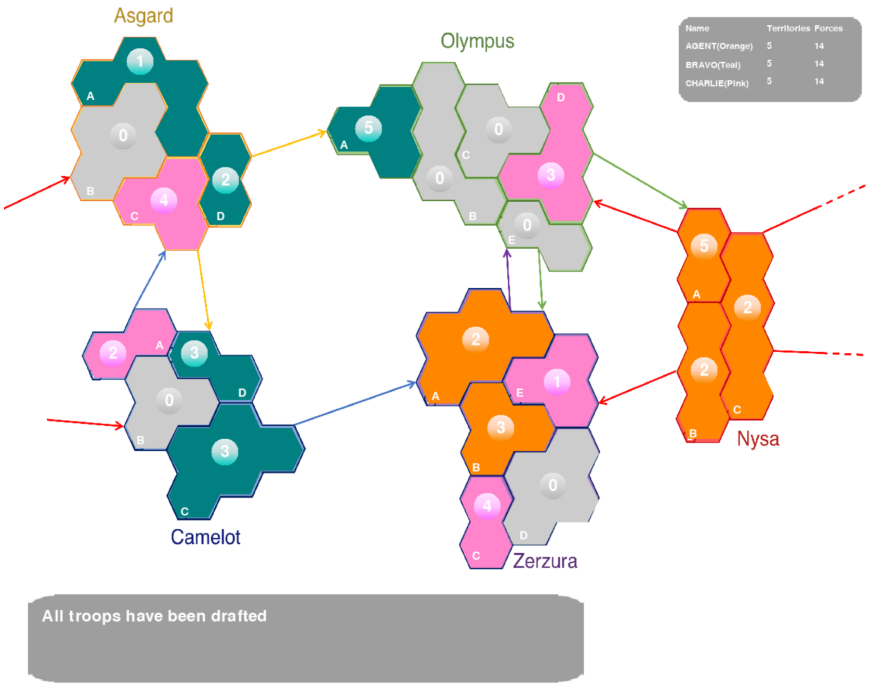


图1：该图展示了本研究使用的 Risk 模拟器。推荐策略的模拟由橙色玩家（Agent）执行，该玩家与青绿色玩家（Bravo）和粉色玩家（Charlie）对战。我们还加入了图例，方便参与者查看所有玩家的兵力和领土。每个行动都在屏幕底部的文本框中附有文字说明。

3.2 研究问题

通过本实验，我们旨在回答以下研究问题：

研究问题1——我们如何自动校准出符合参与者游戏偏好的可执行策略？研究问题2——我们的研究条件，即向用户推荐的策略类型，是否会影响推荐系统的因变量？研究问题3——人格因素和人口统计信息如何影响本研究中的因变量？研究问题4——对推荐列表的感知对齐程度如何影响人们对某种推荐类型的整体感知？

我们的实验分为两个阶段。在第一阶段，我们验证使策略符合参与者偏好的机制（研究问题*1*）。在第二阶段，我们研究不同策略推荐机制的感知，并确定影响这些感知的因素（研究问题*2*至*4*）。

4 学习偏好策略 - 阶段1

我们实验的第一步是确定一种机制，将参与者与生俱来的游戏偏好编码为与之匹配、可付诸行动的策略（研究问题*1*）。为实现这一目标，我们需要一套有意义的问题，以确定参与者的游戏偏好。首先，我们制定了八种不同的策略，涵盖《Risk》中多样的游戏策略。接下来，我们设计了一份问卷，参与者需要回答三个问题，说明他们在《Risk》的虚构情境中会采取什么行动。这份问卷采用深度为三的决策树结构，因此参与者收到的问题取决于他们对前一个问题的回答（见图3）。根据参与者对这三个问题的回答，我们可以确定八种策略（即$2^3 = 8$选项）中，哪一种最符合他们的游戏方式。在将问卷用于研究之前，我们首先开展了一项研究，以验证我们的方法能否准确校准参与者偏好的《Risk》

表 1：此表展示了我们在一致性问卷中使用的李克特陈述。每条陈述均采用七级响应格式，选项从「非常不同意」到「非常同意」。

	李克特量表
1	给我的建议**与**我的策略一致
2	这些建议**未能**反映我的策略
3	这些建议**关注了**我的策略
4	这些建议**忽视了** 我偏好的策略

策略。本节分析了试点研究，这些研究为验证研究的最终结构以及用于验证问卷的最终流程提供了依据。

4.1 试点研究

在试点阶段，我们多次迭代校准阶段的研究设计，以找出并弥补研究中的不足。第一次迭代中，我们只向参与者展示决策树生成的策略（实际策略（AS）），并请他们填写一致性问卷（见表 1），但很快意识到这可能会导致研究中出现确认偏差（n=12，均值=5.4375，SD=1.28）。第二次迭代中，我们在研究中同时纳入实际策略及其反向策略（RS）。首先向参与者推荐其中一种策略，并请他们填写该策略的一致性问卷；随后对另一种策略重复相同流程。由于参与者无法比较两种策略，他们最终都给两种策略打出了较高的一致性评分（n=3，AS：均值=6.33，SD=0.62；RS：均值=5.42，SD=1.19）。因此，在下一阶段，我们将两种策略并排展示（图2）。此时，参与者能够比较和对照这两种策略。偏好反向策略的人无法连贯地解释自己的选择（n=5，AS：均值=4.45，SD=1.94；RS：均值=5.5，SD=1.20）。因此，在研究的最后阶段，我们先请参与者在偏好问卷中解释其策略选择依据，然后再向他们展示推荐策略。此外，我们招募了有玩过 RISK、CATAN 等策略游戏经验的参与者，以便他们更好地验证我们的偏好问题。如果参与者仍然偏好反向策略而非实际策略，我们便向他们提出一些半结构化访谈问题（附录C）。

4.2 校准问卷的验证

我们的研究首先通过一份 Risk 规则教程，让参与者熟悉环境规则以及我们对游戏所做的简化。接下来，参与者可以使用 Risk 模拟器，熟悉游戏玩法。研究管理员会引导参与者逐一体验模拟器中的各个游戏阶段，随后参与者可以继续游戏，直到确信自己理解了游戏。与每项试点研究一样，我们的问卷采用二叉树结构，由若干基于 Risk 虚构场景的多项选择题组成。每道题有两个选项，参与者的回答会决定他们接下来看到的问题。每道题都旨在了解参与者在游戏偏好方面的特定情况，例如：“他们是进攻型玩家吗？”“他们喜欢让部队保持紧密集结吗？”“他们更倾向于尽可能多地参与战斗，还是控制更多国家？”等等。附录中列出了树中的七道问题以及树本身的结构。参与者回答三个问题后，我们的方法便能得出一个独特的策略档案，反映该参与者会如何玩 Risk。我们总共设计了八种游戏玩法档案，并为每种档案配备了一个基于启发式的游戏智能体，以模拟其在实际游戏中的表现。

接下来，我们向每位参与者展示两个选项：(1)“实际”策略，即与其问卷回答最为一致的策略；(2)“反向”策略，即与其问卷回答相反的策略，也就是位于决策树另一侧的策略。我们将这两种策略并排展示给参与者。参与者交替先收到实际策略或反向策略作为建议策略 1（见图2）。随后，我们向参与者展示一段人工智能风险玩家采用给定策略进行模拟的视频，并请他们回答一份由四个问题组成的一致性问卷（见表 1），以评估该策略与其偏好的契合程度。这项调查采用 7 点、4 项李克特量表进行评估，选项从“强烈不同意”到“强烈同意”。最后，我们请参与者回答一些主观问题，谈谈他们对建议策略的喜好之处和不喜欢之处。我们依次对两种策略执行这一流程，即先对策略 1 进行模拟、一致性评估和主观评价，再对策略 2 进行模拟、一致性评估和主观评价。我们舍弃了一个数据点

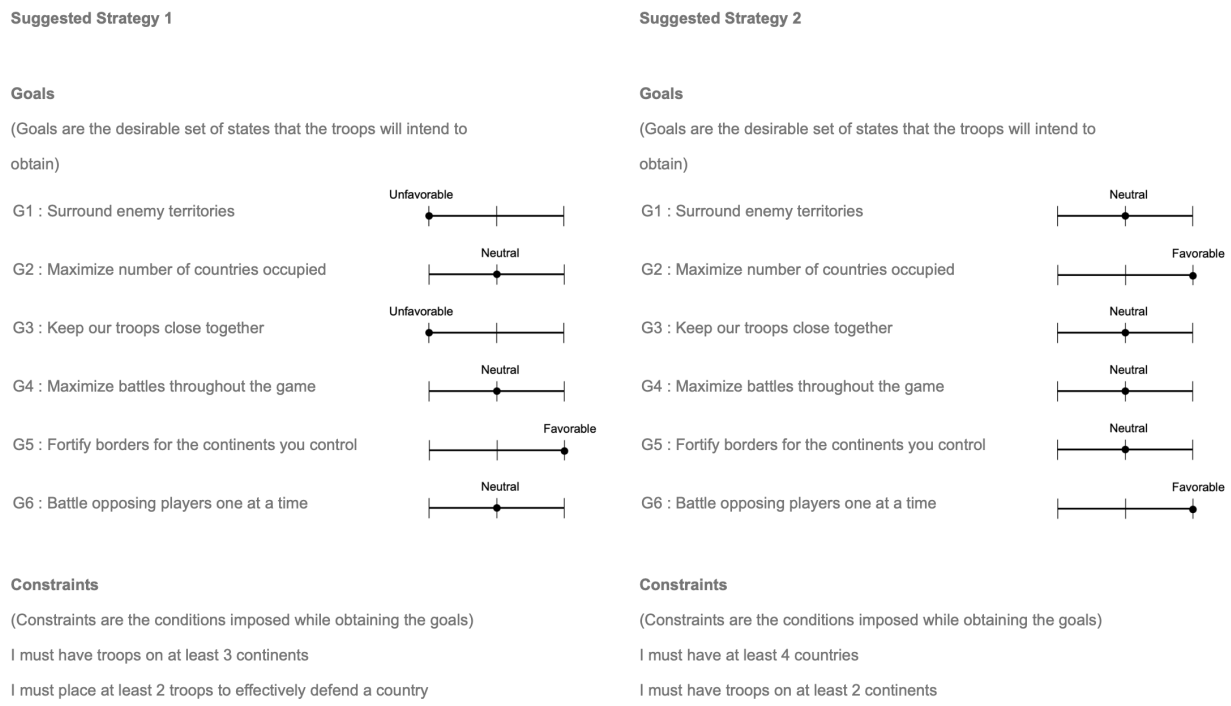


图2：该图展示了在校准研究中向参与者呈现实际策略和反向策略的方式。前半数参与者看到的实际策略标为“策略1”，后半数参与者看到的反向策略标为“策略1”。

在该数据点中，参与者承认，实际策略更能体现自己在这些情境中的选择，但给反向策略的评分却更高。

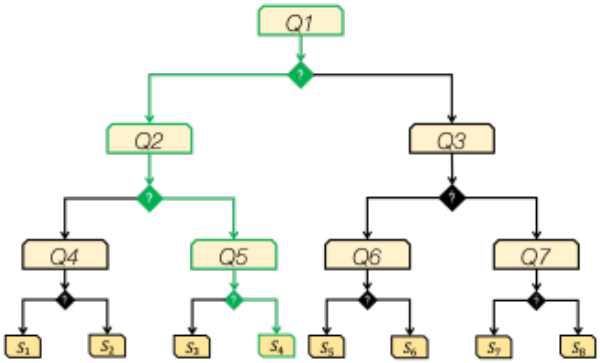
4.3 结果

本节将介绍校准研究的统计结果，以验证研究问题*1*。我们对16名参与者进行了校准研究，但不得不舍弃其中一名参与者的数据。在最终纳入的15名参与者中，8人看到的实际策略标为策略1（图2左侧），另有7人看到的反向策略标为策略1。参与者对实际策略的对齐程度评分（均值 = 5.18，标准差 = 1.28）平均高于对反向策略的评分（均值 = 3.93，标准差 = 1.78）。数据未通过正态性夏皮罗-威尔克检验，因此我们采用了威尔科克森符号秩检验。检验结果证实，这一差异具有统计学显著性，$Z = -3.22$，$p < 0.05$。一致性问卷的克朗巴赫α系数为 $\alpha = 0.94$。威尔科克森符号秩检验的统计显著性证实，问卷推荐的策略与参与者的偏好准确对齐。该结果验证了我们的假设：该问卷能够将偏好编码为可执行的《风险》策略，用于实验的第二阶段。

5 第二阶段 - 策略推荐偏好研究

验证偏好问卷后，我们进入第二个实验，以验证研究问题*2*至*4*。我们以现场实验的形式开展了第二项研究，采用 1×4 被试间设计，其中研究条件是向用户推荐策略的方法。我们首先通过第 4 节中的偏好问卷，校准参与者的游戏偏好。完成校准后，我们以以下四种形式之一向用户推荐策略：

1. **单一** - 向参与者推荐一种最符合其 Risk 游戏偏好的策略。 2. **相似** - 向参与者推荐三种彼此相似且符合其游戏偏好的策略。



Condition	Strategies recommended
Single	$[S_4]$
Similar	$[S_2, S_3, S_4]$
Diverse	$[S_2, S_4, S_8]$
All	$[S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8]$

图3：根据分配给参与者的研究条件推荐策略的示意图。在此示例中，根据参与者对每个问题的回答，其理想策略为 S_4，如绿色标示的路径所示。在“单一”条件下，我们只推荐 S_4。在“相似”条件下，向参与者推荐3种策略，即兄弟策略 S_3，以及它的一种“表亲”策略 S_2。在“多样”条件下，同样向参与者推荐3种策略，但不推荐兄弟策略，而是推荐树另一侧的一种策略，即 S_8。最后，在“全部”条件下，参与者会看到全部八种策略。完整的校准问卷见附录B.1。

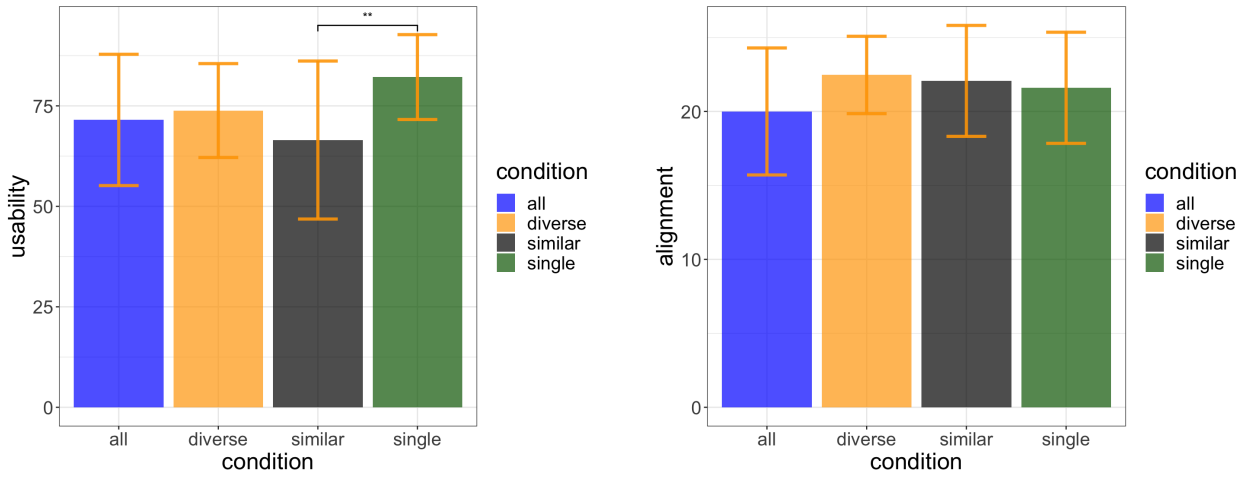
3. **多样** - 向参与者推荐三种策略，其中既包括与其游戏偏好相似的策略，也包括与其游戏偏好相反的策略。 4. **全部** - 向参与者推荐所有可用于玩 Risk 的策略。在我们的研究中，共有八种策略。

5.1 流程

我们以与校准研究相似的方式开展了这项研究：为参与者提供 RISK 和 RISK模拟器的教程，让他们探索并体验游戏。在推荐任何策略之前，我们请参与者填写 Mini-IPIP人格问卷，以测量其人格类型（Donnellan等，2006）。

完成教程和预调查环节后，参与者便可以开始研究。研究开始时，参与者填写上述校准后的偏好问卷，以确定他们理想的策略。在提供策略推荐之前，我们请参与者口头说明自己在每种情境中做出选择的理由。这一环节旨在让参与者根据自己的选择重新校准整体策略，同时也有助于筛除犯了错误或不再认同原先选择的参与者。接下来，我们根据参与者所分配到的研究条件，向其展示一条、三条或八条策略推荐。图3展示了如何根据参与者在策略问卷中对各个问题的回答来推荐策略。策略按照相关性从高到低排列，即从最符合其游戏偏好的策略到最不符合其偏好的策略，但我们没有告知参与者这一排序。除了介绍每种策略的目标和约束外，我们还提供了一张图片，展示采用该策略的玩家在选地阶段会如何部署部队（附录图2）。最后，对于每条推荐策略，参与者都可以选择观看一段模拟，了解采用该策略的 Agent 在游戏前两个回合中的表现。如果参与者难以理解策略的目标和约束，这段模拟有助于他们直观地了解相应策略。我们将模拟限制在两个回合，是为了避免参与者因游戏 Agent 是否获胜而产生偏见。我们的目的是鼓励参与者根据推荐策略与自己游戏方式的对齐程度来进行评估，而非根据策略的成功与否进行评估。如果参与者看到某个策略失败了——这可能是由于 Risk 的随机性所致——他们就可能不太愿意承认该策略符合自己的游戏方式。

在参与者确认自己已理解每项推荐策略后，我们会请他们回答四份实验后问卷，以了解他们使用我们推荐界面的体验。首先，他们填写一份一致性问卷，该问卷与校准研究中使用的问卷相同（见表 1）。接下来，参与者填写先前研究中的系统可用性问卷（Brooke 等人，1996年），以评估他们认为推荐界面的可用性。完成可用性评估后，参与者填写 NASA 任务负荷指数（TLX）问卷（Hart 和 Staveland，1988年），用于测量任务的工作负荷。随后，参与者填写 Godspeed 感知智能问卷（Bartneck 等人，2009年），以评估推荐过程的智能程度。最后一份问卷是一份新设计的偏好问卷，



(a) (b)

图4：两张条形图，分别展示了各研究条件在（a）可用性和（b）对齐方面的表现。

旨在了解参与者在接收单一、相似、多样或全部 推荐方面的一般偏好（附录B.2）。每个类别包含四个相关问题，共计16个李克特量表项目。该问卷采用7点李克特量表，从“强烈反对”到“强烈赞同”。我们还加入了两个主观问题，即：（1）你喜欢推荐界面的哪些方面？（2）你不喜欢推荐界面的哪些方面？通过这些问题，我们希望深入了解参与者对每种推荐类型的看法及体验。这些问题安排在新设计的偏好问卷之前，以消除任何偏差。

5.2 结果

本节报告为回答本文提出的研究问题，在校准研究和主研究中进行的统计检验。我们共招募了60名参与者开展实验。研究结束后，我们通过一项调查收集了参与者的其他人口统计信息。在提交反馈的58名参与者中，21人为女性，36人为男性，1人为非二元性别。参与者的平均年龄为23.5岁（标准差为2.95），他们在1至5分量表上的策略游戏平均专业水平为3.08（标准差为0.98）。本研究预计耗时45分钟，参与者获得了15美元的报酬。我们将介绍针对各项研究问题进行的统计检验，并结合用户对策略推荐的偏好分析研究结果。

为回答有关用户对策略推荐的偏好的研究问题（研究问题*2 - 4*），我们针对每个因变量进行了多元回归分析。对于所有线性回归模型，我们都检验了残差正态性和方差齐性的假设。我们使用莱文检验衡量方差齐性，结果发现所有模型均满足方差齐性。所有模型及假设检验的摘要见附录。我们使用夏皮罗-威尔克检验检验各模型残差的正态性，但发现部分模型的残差不服从正态分布。先前研究表明，F检验对非正态性具有稳健性（Cochran，1947年；Blanca Mena 等，2017年；Hack，1958年；Glass 等，1972年）。因此，我们选择继续进行线性回归分析。为每项测量确定合适的模型时，我们采用 AICc 作为奥卡姆剃刀。由于我们的样本量相对较小（n=60），我们使用 AICc；它在标准 AIC 的基础上增加了校正项，以避免过拟合。我们进行了单因素方差分析，以衡量各项测量对因变量的显著性影响。此外，我们还进行了 Tukey HSD 事后检验，以确定作为因素构建的自变量各取值之间的两两比较显著性。

5.2.1 研究问题2

我们首先旨在衡量研究条件（即向参与者推荐策略的方法）对可用性、感知智能和工作负荷的影响。对可用性的线性回归模型进行方差分析后发现，不同推荐类型之间的可用性存在显著差异（$F(3, 56) = 4.1609$，$p < 0.05$）。图基事后检验表明，单一策略推荐的可用性显著高于相似策略推荐（$p < 0.01$）。单一策略的评分也高于另外两种条件（多样、全部），

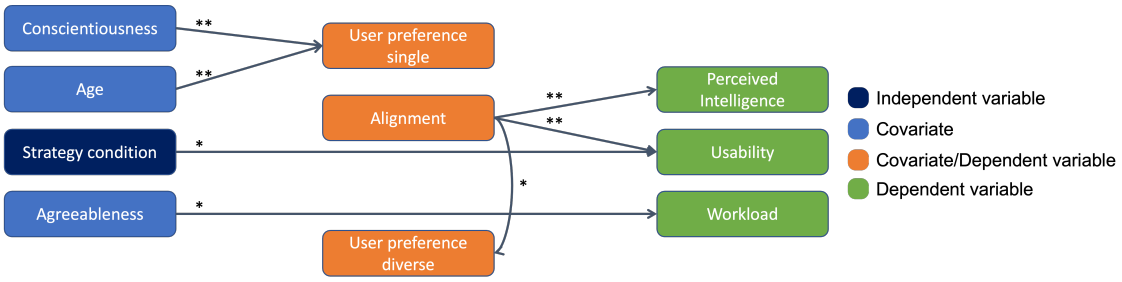
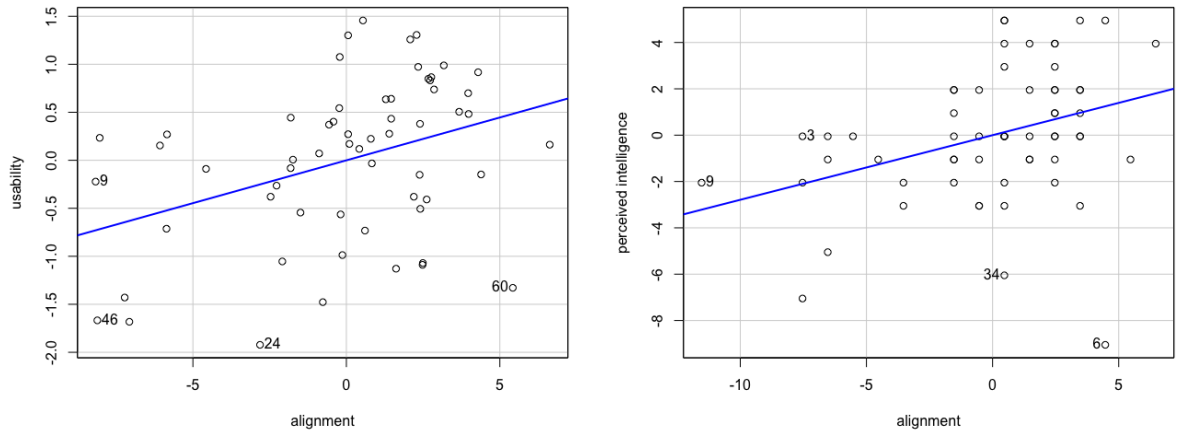


图5：本研究中所有显著结果的总结。



(a) (b)

图6： 这些图展示了感知对齐对可用性和感知智能的影响。

但两两比较的差异未达到显著水平。推荐策略的方法对感知智能或工作负荷的影响均未达到显著水平。

5.2.2 研究问题3

接下来，我们旨在了解内在性格类型如何影响对特定策略的偏好。首先，尽责性显著影响了策略推荐系统的整体可用性。尽责性程度较高的参与者（$F(1,58) = 9.6539$，$p < 0.01$）往往会显著地认为推荐系统更具可用性。喜欢做好准备并关注细节的参与者，可能会觉得通过推荐策略来解决复杂任务更有用，因为他们愿意花时间全面评估各种选项。宜人性特质较高的参与者，其工作负荷显著更高（$F(1, 58) = 4.3898$，$p < 0.05$）。这一发现合乎情理，因为宜人性较高的参与者更可能逐一模拟每种策略，以便充分理解每项推荐，从而承担更高的工作负荷。此外，我们发现年龄与对单一策略推荐系统的偏好呈正相关（$F(1, 58) = 14.6446$，$p < 0.01$）。这可能意味着年长者对自己的策略更有信心，不愿浪费时间分析其他选项。

5.2.3 研究问题4

接下来，我们报告一些关于推荐与参与者固有偏好之间感知对齐的发现。研究发现，对齐程度显著提升了可用性（$F(1, 58) = 8.6770$，$p < 0.01$）。根据第 5.2.1 节，我们知道，单一策略条件被认为是最易用的。这一发现与个性化领域的既有研究一致（Tan 等人，2008；Chen 和 Yin，2006；Linden 等人，2003），这些研究指出，人们通常希望推荐选项能体现自己的偏好。对齐程度也显著影响了感知智能（$F(1,58) = 9.3313$，$p < 0.01$）。我们发现，认为所提供的推荐

与自身固有偏好更为对齐的参与者，认为该系统更加智能。这合乎逻辑：认为系统能够准确理解并将其游戏偏好编码到推荐中的参与者，会认为该系统更智能。

此外，在通过偏好问卷对多样化策略的一般偏好进行建模时，对齐是一个显著的协变量（$F(1, 58) = 5.0068$，$p < 0.05$）。更高的感知对齐度提升了参与者对多样化策略的一般偏好，这表明，当参与者收到的推荐被认为更准确地反映其偏好时，他们会更愿意接受其他具有对比性的策略。

5.3 讨论

我们的实验凸显了策略推荐系统设计中的若干关键趋势。与此前关于一般推荐系统的研究（Nguyen 等人，2018）一致，我们的结果强调，在设计人工智能策略推荐时考虑人格因素的重要性。我们发现，尽责性显著降低了参与者对单一策略呈现模式的偏好，而宜人性可能增加其所感受到的工作负荷。这一发现暗示，最终用户的内在特质可以用于预先设计更适合的策略推荐系统。例如，既有研究表明，高尽责性对于外科医生或直升机飞行员等高压力、高技能岗位非常重要（Dickens，2013；Grice 和 Katz，2006；Mullola 等，2018）。我们的结果可能表明，从事此类职业的人类可能不偏好单一策略条件下的推荐。此外，在私营部门工作的医生，如果其专业领域是全科医学或职业健康，也可能具有较高水平的宜人性（Mullola 等，2018）。这些领域的专业人员在使用策略推荐系统时可能会经历更高的工作负荷。此外，相较于相似策略推荐系统，人们似乎总体上更偏好单策略推荐系统。因此，如果推荐系统能够推断用户偏好，与其让用户从一系列类似于最相关策略的策略中进行选择，用户可能会认为仅呈现最相关的策略更易于使用。

以往研究表明，人们对人工智能助手的要求高于对人类助手的要求（Chen 等，2021）。为了设计出人们愿意采纳、适用于人类-AI协作任务的推荐系统，我们需要更深入地了解人们对这类系统的感知。我们的研究结果揭示了策略推荐系统在可用性和感知智能方面的重要见解。认为自身偏好与系统推荐相一致的参与者，会觉得系统更智能、更易用。本研究验证了使推荐选项与用户偏好相一致所带来的益处；未来研究应探索如何大规模自动推断用户的内在偏好，并据此调整策略推荐。此外，我们的定性问卷表明，在策略推荐系统中纳入可解释人工智能机制（Zhu 和 Williams，2020年；Ehsan 等，2019；Silva 等，2022）或许还能进一步增强用户对系统的信心。例如，一位参与者评论道：“我希望界面能更详细地解释我的选择是如何转化为策略的，这样我就会更有信心。”我们还揭示了研究条件对上述各项因变量的影响，从而为这些因变量的建模提供更多参考。下一步的一项重要工作是从用户采纳的角度研究这些指标，以确定这三个属性（对齐、可用性和感知智能）分别如何影响用户采纳特定策略推荐模式的意愿。

6 局限性

首先，尽管我们为参与者提供了教程以及 Risk 游戏玩法模拟，但无法确定参与者是否充分理解了 Risk。我们会口头确认参与者对游戏感到熟悉，但不会专门对此进行测试。如果参与者没有充分理解 Risk 的玩法，他们可能无法建立自己会采用何种策略的心理模型，因此可能难以评估策略推荐。尽管这一点值得指出，但鉴于参与实验的参与者人数以及教程和模拟内容的全面性，这并不会削弱我们的研究发现和分析。

其次，本研究尚未对我们设计的偏好问卷进行因子分析。我们的分析发现了与参与者偏好相关的显著趋势，为问卷的有效性提供了支持。然而，在未来的工作中，我们希望进行因子分析，以确保各因素（单一、相似、多样、全部）相关问题的可靠性。最后，尽管本研究对广义策略推荐偏好提供了新的见解，但我们的实验并非在沉浸式现实世界任务中进行。未来，我们希望开展一项在现实世界中与机器人协作的实验，由机器人向一名

参与者，而人类则利用这些推荐来指定机器人应如何完成协作任务。未来，我们还将在更多样化的人群中开展这项人类受试者实验。

7 结论

随着人类与人工智能代理协作以解决更多高难度任务，人类可能无法将自身偏好转化为人工智能代理能够执行的可执行策略。为解决这一问题，我们需要开发能够接收用户偏好并推荐定义明确的策略的策略推荐系统。以往关于以用户为中心的推荐系统的研究，探讨了用户的人格类型对其推荐系统感知的影响，以及推荐列表的一般结构。然而，如果当前任务更加复杂精细，这些偏好可能会发生变化。本文开展了一项新颖的人类受试者实验，以了解不同类型的推荐对可用性、感知智能、工作负荷，以及对任一推荐方式的偏好会产生怎样的影响。我们创建并验证了一种新方法，通过三个简单的游戏问题，将参与者的偏好编码为适用于Risk游戏的可执行策略。我们还开发了一份新问卷，用于衡量用户对接收单一、相似、多样或全部 类型推荐的偏好。分析结果显示，某些人格特质（例如：尽责性）会显著影响用户对特定类型推荐系统的偏好。最后，我们报告了偏好、对齐和可用性之间的相关性，指出有必要进一步开发机制，以自动推断用户偏好，并使策略推荐与之对齐。我们的工作揭示了用户对策略推荐系统的偏好，可用于未来设计面向人类-AI组队的个性化系统。个性化进而可以提高用户满意度和策略推荐系统的采用率。

8 致谢

本研究获得美国海军研究办公室资助（项目编号 N00014-19-2076），并得到柯尼卡美能达向佐治亚理工学院研究基金会捐赠的支持。

参考文献

Ashton Anderson、Lucas Maystre、Ian Anderson、Rishabh Mehrotra 和 Mounia Lalmas。2020年。算法对 Spotify 消费多样性的影响。载于*2020*年万维网大会论文集。2155–2165。Christoph Bartneck、Dana Kuli´c、Elizabeth Croft 和 Susana Zoghbi。2009年。机器人拟人化、生命感、喜爱度、感知智能和感知安全性的测量工具。国际社会机器人学期刊1，1（2009年），71–81。Rajat Kumar Behera、Angappa Gunasekaran、Shivam Gupta、Shampy Kamboj 和 Pradip Kumar Bala。2020年。个性化数字营销推荐引擎。《零售与消费者服务期刊》53（2020年），101799。María José Blanca Mena、Rafael Alarcón Postigo、Jaume Arnau Gras、Roser Bono Cabré、Rebecca Bendayan 等。2017年。非正态数据：方差分析仍然是有效选项吗？《心理主题》（2017年）。Dirk Bollen、Bart P Knijnenburg、Martijn C Willemsen 和 Mark Graus。2010年。理解推荐系统中的选择过载。载于第四届 *ACM* 推荐系统大会论文集。63–70。John Brooke 等人。1996年。SUS——一种快速简便的可用性量表。工业可用性评估189，194（1996年），4–7。Jian Chen 和 Jian Yin。2006年。基于影响集的推荐。载于网络挖掘与网络使用分析研讨会论文集。CiteSeer。Nuoya Chen、Smaraki Mohanty、Jinfeng Jiao 和 Xiucheng Fan。2021年。人非圣贤：服务失误时应宽容人类而非机器。《零售与消费者服务期刊》59（2021年），102363。William G Cochran。1947年。当方差分析的假设不成立时的一些后果。《生物统计学》3，1（1947年），22–38。Paul Dickens。2013年。寻找 Chuck 和 PT：评估直升机飞行员的循证方法。载于第*17*届国际航空心理学研讨会。651。M Brent Donnellan、Frederick L Oswald、Brendan M Baird 和 Richard E Lucas。2006年。迷你国际人格项目量表：简短而有效的大五人格因素测量工具。心理评估18，2（2006年），192。Upol Ehsan、Pradyumna Tambwekar、Larry Chan、Brent Harrison 和 Mark O Riedl。2019年。自动化理由生成：一种可解释人工智能技术及其对人类感知的影响。载于第*24*届国际智能用户界面会议论文集。263–274。

Gene V Glass、Percy D Peckham 和 James R Sanders。1972年。未满足固定效应方差与协方差分析基本假设的后果。《教育研究评论》42, 3 (1972年), 237–288。Matthew Gombolay、Anna Bair、Cindy Huang 和 Julie Shah。2017年。兼顾人为因素（态势感知、工作负荷和工作流程偏好）的混合主动式人机协作计算设计。《国际机器人研究杂志》36, 5-7 (2017年), 597–617。Robert Grice 和 Lawrence C Katz。2006年。不同任务平台上经验丰富的美国陆军飞行员的人格特征画像。技术报告。美国陆军行为与社会科学研究所，拉克堡，阿拉巴马州……HRB Hack。1958年。对来自两个非正态总体的样本中 F 比率分布的实证研究。《生物统计学》45, 1/2 (1958年), 260–265。Sandra G Hart 和 Lowell E Staveland。1988年。NASA-TLX（任务负荷指数）的开发：实证与理论研究结果。载于《心理学进展》。第52卷。爱思唯尔，139–183。Xiangnan He、Lizi Liao、Hanwang Zhang、Liqiang Nie、Xia Hu 和 Tat-Seng Chua。2017年。神经协同过滤。载于第*26*届国际万维网大会论文集。173–182。Larkin Heintzman、Amanda Hashimoto、Nicole Abaid 和 Ryan K Williams。2021年。人机协同搜救中的前瞻性规划与动态失踪人员模型。载于*2021*年 *IEEE* 国际机器人与自动化会议（*ICRA*）。IEEE，8252–8258。Yoyo Tsung-Yu Hou 和 Malte F Jung。2021年。谁是专家？调和人工智能支持决策中的算法厌恶与算法欣赏。《*ACM*人机交互汇刊》5, CSCW2 (2021年), 1–25。Anna YQ Huang、Owen HT Lu 和 Stephen JH Yang。2022年。人工智能赋能的个性化推荐对翻转课堂中学习者学习投入、动机和学习成果的影响。《计算机与教育》(2022年), 104684。Shagun Jhaver 、 Yoni Karpfen 和 Judd Antin。2018 年。Airbnb 房东的算法焦虑与应对策略。载于*2018*年计算系统中的人为因素 *CHI* 会议论文集。1–12。Matthew Johnson 和 Jeffrey M Bradshaw。2021年。相互依存如何解释团队合作的世界。载于《人工智能系统工程》。Springer，122–146。Michael Jugovac 和 Dietmar Jannach。2017年。与推荐系统交互——概述与研究方向。《*ACM*交互式智能系统汇刊》（*TiiS*）7, 3 (2017年), 1–46。Daniel Kluver、Michael D Ekstrand 和 Joseph A Konstan。2018年。基于评分的协同过滤：算法与评估。《社会信息获取》(2018年), 344–390。Bart P Knijnenburg 和 Martijn C Willemsen。2015年。通过用户实验评估推荐系统。载于《推荐系统手册》。Springer，309–352。Yehuda Koren、Steffen Rendle 和 Robert Bell。2022年。协同过滤进展。《推荐系统手册》(2022年), 91–142。Pushpendra Kumar 和 Ramjeevan Singh Thakur。2018年。推荐系统技术及相关问题：综述。《国际信息技术杂志》10, 4 (2018年), 495–501。Matevž Kunaver 和 Tomaž Požrl。2017年。推荐系统多样性——综述。《知识型系统》123 (2017年), 154–162。Qinbiao Li、Kam KH Ng、Zhijun Fan、Xin Yuan、Heshan Liu 和 Lingguo Bu。2021年。一种基于功能性近红外光谱、用于空中交通管制环境中自适应决策的以人为中心的方法：案例研究。《高级工程信息学》49 (2021年), 101325。Greg Linden、Brent Smith 和 Jeremy York。2003年。亚马逊网站推荐：基于物品的协同过滤。《*IEEE*互联网计算》7, 1 (2003年), 76–80。Chiara Longoni、Andrea Bonezzi 和 Carey K Morewedge。2019年。对医疗人工智能的抵触。《消费者研究杂志》46, 4 (2019年), 629–650。Pasquale Lops、Marco de Gemmis 和 Giovanni Semeraro。2011年。基于内容的推荐系统：最新进展与趋势。《推荐系统手册》(2011年), 73–105。Pasquale Lops、Dietmar Jannach、Cataldo Musto、Toine Bogers 和 Marijn Koolen。2019年。基于内容推荐的发展趋势。《用户建模与用户适应性交互》29, 2 (2019年), 239–249。

Farah Magrabi、Elske Ammenwerth、Jytte Brender McNair、Nicolet F De Keizer、Hannele Hyppönen、Pirkko Nykänen、Michael Rigby、Philip J Scott、Tuulikki Vehko、Zoie Shui-Yee Wong等。2019年。临床决策支持中的人工智能：评估人工智能的挑战及其实践意义。《医学信息学年鉴》28, 01 (2019年), 128–134。Sean M McNee、John Riedl和Joseph A Konstan。2006年。让推荐更好：人类-推荐系统交互的分析模型。载于*CHI'06*计算系统中的人为因素扩展摘要。1103–1108。Sari Mullola、Christian Hakulinen、Justin Presseau、David Gimeno Ruiz de Porras、Markus Jokela、Taina Hintsa和Marko Elovainio。2018年。芬兰医生的人格特质与职业选择：就业领域、临床患者接触、专科和专科变更。《*BMC*医学教育》18, 1 (2018年), 1–12。Robin R Murphy。2004年。救援机器人技术中的人机交互。《*IEEE*系统、人类与控制论汇刊*C*辑（应用与综述）》34, 2 (2004年), 138–153。Tien T Nguyen、F Maxwell Harper、Loren Terveen和Joseph A Konstan。2018年。用户个性与用户对推荐系统的满意度。《信息系统前沿》20, 6 (2018年), 1173–1189。Stefanos Nikolaidis和Julie Shah。2012年。运用共享心智模型开展人机团队协作。*ACM/IEEE*人机交互会议 (2012年)。Elena Reutskaja和Robin M Hogarth。2009年。选择满意度与备选项数量的关系：当"好东西让人满足"。《心理学与营销》26, 3 (2009年), 197–203。Badrul Sarwar、George Karypis、Joseph Konstan和John Riedl。2001年。基于项目的协同过滤推荐算法。载于第*10*届万维网国际会议论文集。285–295。J Ben Schafer、Dan Frankowski、Jon Herlocker和Shilad Sen。2007年。协同过滤推荐系统。载于《自适应网络》。Springer，291–324。Guy Shani和Asela Gunawardana。2011年。推荐系统评估。载于《推荐系统手册》。Springer，257–297。Kazuhiko Shinozawa、Futoshi Naya、Junji Yamato和Kiyoshi Kogure。2005年。机器人和屏幕智能体推荐对人类决策的影响差异。《国际人机交互研究杂志》62, 2 (2005年), 267–279。Andrew Silva、Mariah Schrum、Erin Hedlund-Botti、Nakul Gopalan和Matthew Gombolay。2022年。可解释人工智能：评估可解释人工智能对人类与智能体交互的客观和主观影响。《国际人机交互杂志》(2022年), 1–15。Idan Szpektor、Yoelle Maarek和Dan Pelleg。2013年。当相关性还不够：提升个性化问题推荐的多样性和新颖性。载于第*22*届万维网国际会议论文集。1249–1260。Aaquib Tabrez、Matthew B Luebbers和Bradley Hayes。2020年。人机团队协作中的心智建模技术综述。《机器人学最新报告》1, 4 (2020年), 259–267。Kar Yan Tam和Shuk Ying Ho。2003年。网络个性化：是否有效？《*IT*专业人员》5, 5 (2003年), 53–57。Pradyumna Tambwekar、Nathan Vaska、Lakshita Dodeja和Matthew Gombolay。2022年。指挥官意图：策略博弈中的人机任务规范数据集与建模方法。https://doi.org/10. 48550/ARXIV.2208.08374 Huiyi Tan、Junfei Guo和Yong Li。2008年。电子学习推荐系统。载于*2008*年国际计算机科学与软件工程会议，第5卷。IEEE，430–433。John K Tarus、Zhendong Niu和Ghulam Mustafa。2018年。基于知识的推荐：面向电子学习的本体推荐系统综述。《人工智能评论》50, 1 (2018年), 21–48。Robin Van Meteren和Maarten Van Someren。2000年。使用基于内容的过滤进行推荐。载于机器学习在新时代信息领域中的应用：*MLnet/ECML2000*研讨会论文集，第30卷。47–56。Marco Vanetti、Elisabetta Binaghi、Barbara Carminati、Moreno Carullo和Elena Ferrari。2010年。在线社交网络中的基于内容的过滤。载于数据挖掘与机器学习中的隐私和安全问题国际研讨会。Springer，127–140。Martijn C Willemsen、Mark P Graus和Bart P Knijnenburg。2016年。理解潜在特征多样化对选择难度和满意度的作用。《用户建模与用户适应性交互》26, 4 (2016年), 347–389。Wenyi Xiao、Huan Zhao、Haojie Pan、Yangqiu Song、Vincent W Zheng和Qiang Yang。2019年。超越个性化：促进创作者平等与消费者满意度的社交内容推荐。载于第*25*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集。235–245。

Markus Zanker、Laurens Rook 和 Dietmar Jannach。2019年。衡量在线个性化的影响：过去、现在与未来。《国际人机交互研究杂志》131（2019年），160–168。朱丽晓和 Thomas Williams。2020年。机器人主动解释对人机信任的影响。载于国际社交机器人会议。Springer，85–95。Cai-Nicolas Ziegler、Sean M McNee、Joseph A Konstan 和 Georg Lausen。2005年。通过主题多样化改进推荐列表。载于第*14*届万维网国际会议论文集。22–32。

A 附加结果

我们采用 AICc 作为奥卡姆剃刀，为每个因变量确定最终模型。对于每个因变量，我们首先建立一个模型，将每个单独的协变量和自变量纳入其中，并加入这些变量之间的所有两两交互效应。随后逐步移除变量，直到得到 AIC 分数最低的线性回归模型。本节报告一项额外的显著发现，该发现与我们的研究问题无关。

我们发现，对多样化策略的偏好与对相似策略的偏好呈负相关。在同时对相似策略偏好 ($F(1, 58) = 7.9319$, $p < 0.01$) 和多样化策略偏好 ($F(1, 58) = 7.6388$, $p < 0.01$) 进行建模时，这种相关性均达到显著水平（图7）。这一结果符合直觉：喜欢系统推荐符合自己偏好的策略的人，不会喜欢系统推荐多样化策略，反之亦然。

B 问卷

B.1 校准问卷

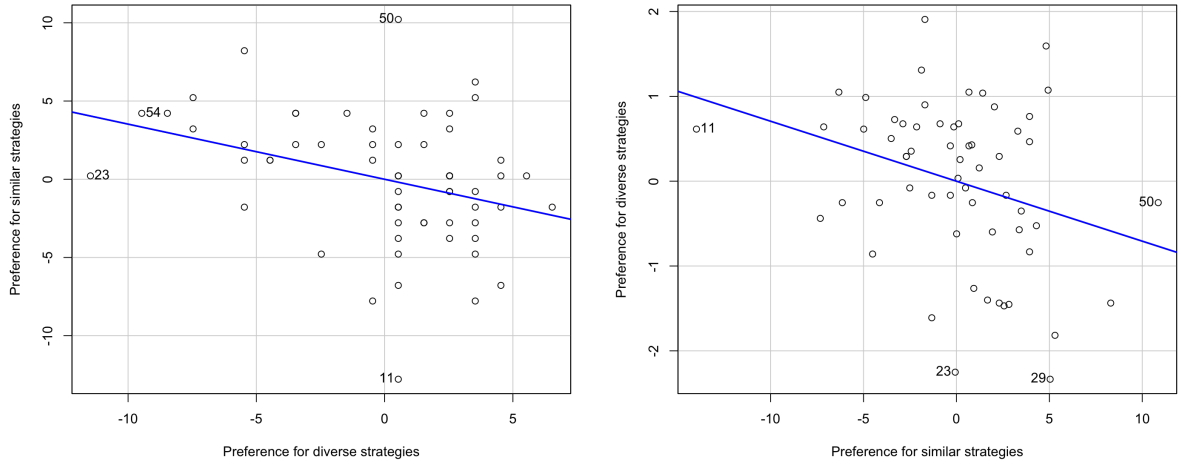
本节列出用于编码参与者偏好游戏风格的七个问题。校准后的游戏风格随后用于推荐策略（图3）。问卷包含七个问题，每个问题有两个选项可供选择（图8）。

B.2 偏好问卷

本节详细介绍我们新设计的偏好问卷中的问题及评分方式。我们设置了四个问题，用于衡量参与者对每种条件的偏好。针对某一条件的问题顺序经过随机化。

问卷 假设你正在完成一项具有挑战性的任务，需要制定一个明确且可行的策略。随后，你获得一个人工智能 Agent，它能够理解你的偏好，并准确地提供策略推荐（与本研究中你经历的过程类似）。在这种情况下，请根据你希望如何获得计划推荐回答以下问题，评分范围为 1 至 7 分，其中 1 分表示非常不同意，7 分表示非常同意。

> 1. 我希望只向我展示一个最能体现我的偏好的计划。2. 即使向我展示了一个最能体现我的偏好的计划，我仍希望看到其他选项，以供考虑。3. 如果向我展示的计划最能体现我的偏好，我就不需要多个计划。4. 评估多个计划所付出的额外工作不值得。5. 我希望看到多个计划：一个最能体现我的策略偏好，另外还有几个略有不同的备选计划。6. 我不希望看到多个符合我偏好的策略。其他选项应与之不同。7. 我不希望看到多样的计划，而是希望看到几个与我的策略偏好高度一致的计划。8. 从一组相似的计划中挑选，比从一组不同的计划中挑选更好。9. 我希望看到多个多样的计划：一个最能体现我的策略偏好，一个与我的偏好截然不同，还有一个与我的偏好既不相似也不相异。



(a) (b)

图7：这些图展示了对多样化策略的偏好与偏好相似策略之间的相关性。

10. 我希望向我展示的所有计划彼此相似，而非不同。11. 如果选项列表不太长，我希望能够考虑一组多样的选项。12. 我不希望考虑与我的偏好不同的计划。13. 我希望看到一份涵盖每一种可能计划的列表——每种可能的策略对应一个计划。14. 评估所有可能的计划不值得付出这么多努力。15. 了解规划可能性的广度，有助于找出最佳计划。16. 向我展示所有可能的计划，并不能帮助我实现目标。

评分 本节介绍我们如何计算问卷中各因素的得分。问卷中的部分题项在评分前需要反向计分。问卷中的每个因素至少包含一个反向题项，以确保参与者认真作答。题项 2、6、10、12、14 和 16 在评分前需要反向计分：

$1 = 7$
$2 = 6$
$3 = 5$
$4 = 4$
$5 = 3$
$6 = 2$
$7 = 1$

对单一策略推荐的偏好：题项 1、2r、3、4 的总和 对相似策略推荐的偏好：题项 5、6r、7、8 的总和 对多样化策略推荐的偏好：题项 9、10r、11、12r 的总和 对所有策略推荐的偏好：题项 13、14r、15、16r 的总和 此处的 r 表示反向计分题项

C 校准研究访谈问题

在校准阶段，如果参与者更倾向于采用反向策略（即与我们的决策树生成的策略相反），我们会向他们提出以下半结构化访谈问题，以了解其中的差距。

1. 你通常在游戏中采用什么策略？2. 哪种策略与你的策略最契合？

表 2：本表详细列出了每个模型的自变量、因变量和协变量，也列出了方差分析检验的假设及所应用的变换。

DV	变换	自变量/协变量	夏皮罗-威尔克检验	莱文检验
可用性	Box-Cox 变换	条件，尽责性，开放性，单一偏好策略，对齐	$p = 0.07245$	$p = 0.3248$
对齐	Box-Cox变换	偏好多样化策略	$p = 0.1011$	$p = 0.3682$
单一偏好策略	N/A	尽责性、年龄	$p = 0.6839$	$p = 0.8932$
偏好相似策略	N/A	偏好多样化策略	$p = 0.3213$	$p = 0.3786$
偏好多样策略	Box-Cox 变换	偏好相似策略，对齐	$p = 0.4235$	$p = 0.4235$
工作负荷	N/A	宜人性、尽责性，偏好单一策略、年龄	$p = 0.5515$	$p = 0.5909$
感知智能	Box-Cox 变换	对齐	$p = 0.4376$	$p = 0.8405$

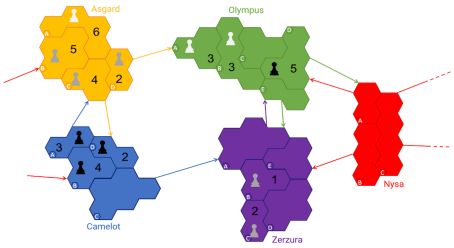
3. 你认为策略A与策略B有何不同？4. 请将以下策略按照与你的策略从最不对齐到最对齐进行排序。5. 你为什么选择选项X？6. 你认为这个问题想要表达什么？7. 设置中的哪些变化会让你改变答案？

D 数据筛选评分标准

本节包含用于筛选主要研究数据的评分标准。我们根据以下标准剔除了四个数据点——

1. 参与者没有阅读推荐策略的目标和约束。2. 参与者评估的是 Qualtrics 用户界面，而非推荐系统。3. 参与者没有评估策略，而只是评估模拟策略的表现如何。4. 参与者在进行用户研究时犯了一些错误。(a) 例如，参与者在初始校准问卷中选错了选项。(b) 例如，参与者忘记回答偏好问卷中的一个李克特量表项目

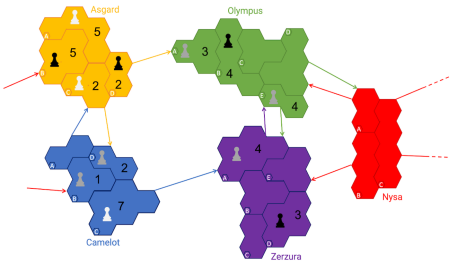
You are currently in the attack phase for the RISK board given below. What would be your next course of action ? (You are the white player playing against black and grey players)



- ◯ Battle Asgard D, Asgard C and Asgard B sequentially with troops in Asgard A
- ◯ Occupy empty territories on the Olympus continent

(a) 问题 1

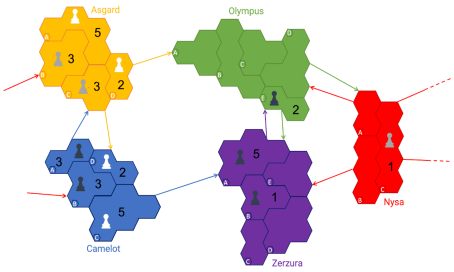
You are currently in the attack phase for the RISK board given below. What would be you next course of action ? (You are the white player playing against black and grey players)



- ◯ Target players individually i.e. Target black player in Asgard **or** grey player in Camelot
- ◯ Target whichever player is easier to defeat i.e. Target black player in Asgard D **and** grey player Camelot B

(b) 问题 2

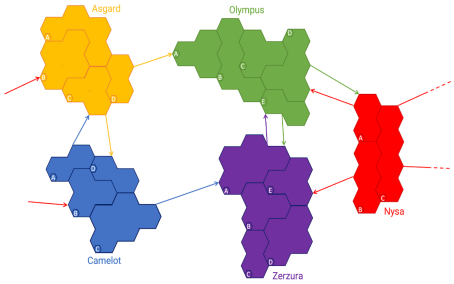
You are currently in the attack phase for the RISK board given below. What would be you next course of action ? (You are the white player playing against black and grey players)



- ◯ Attack the Grey player's troops on Asgard
- ◯ Attack the Black player's troops on Camelot

(c) 问题 4

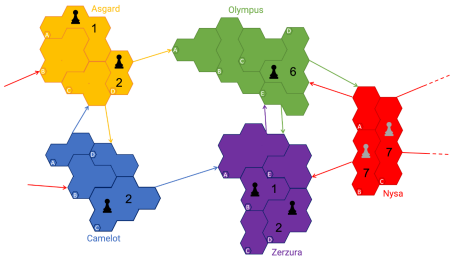
If you were playing the game of Risk, which continent do you think is more valuable to occupy ?



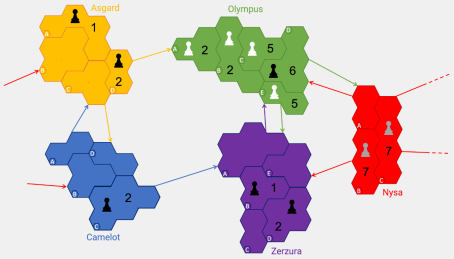
- ◯ Nysa Continent
- ◯ Zerzura Continent

(d) 问题 6

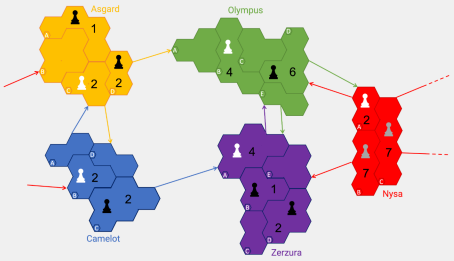
You are currently in the drafting phase and have 14 troops to draft for the RISK Board given below. Which one do you think is a better drafting strategy ? (You are the white player playing against black and grey players)



2 troops on Olympus A, 2 troops on Olympus B, 5 troops on Olympus C, 5 troops on Olympus E

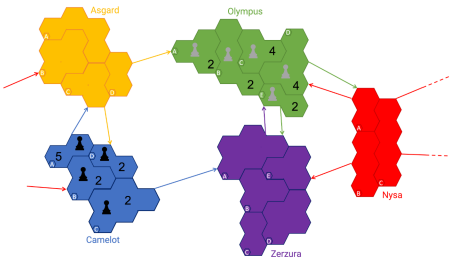


2 troops on Asgard C, 2 troops on Camelot B, 4 troops on Zerzura A, 4 troops on Olympus B, 2 troops on Nysa A

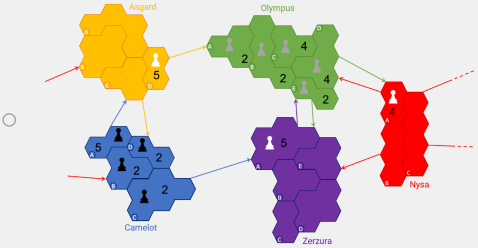


(e) 问题3

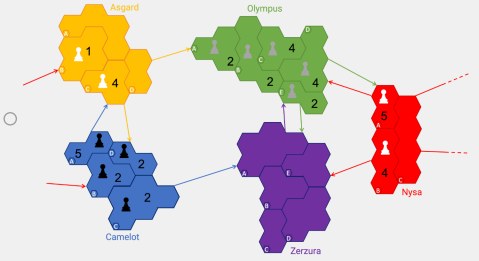
You are currently in the drafting phase and have 14 troops to draft for the RISK Board given below. Which one do you think is a better drafting strategy ? (You are the white player playing against black and grey players)



5 troops on Asgard D, 5 troops on Zerzura A, 4 troops on Nysa A

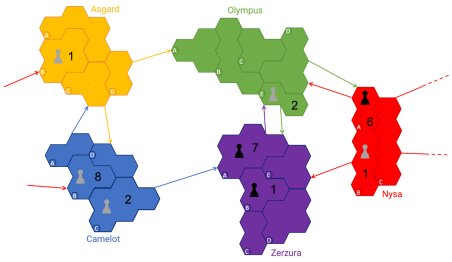


5 troops on Nysa A , 4 troops on Nysa B , 4 troops on Asgard C, 1 troop on Asgard B

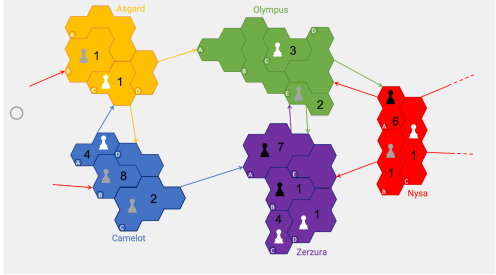


(f) 问题5

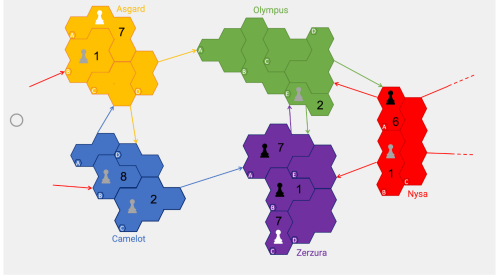
You are currently in the drafting phase and have 14 troops to draft for the RISK Board given below. Which one do you think is a better drafting strategy ? (You are the white player playing against black and grey players)



4 troops on Zerzura C, 1 troop on Zerzura D, 4 troops on Camelot A, 3 troops on Olympus C, 1 troop on Asgard C, 1 troop on Nysa C



7 troops on Zerzura C, 7 troops on Asgard A



（g）问题7

图8：校准问卷