中国科学技术大学

硕士学位论文

基于人类决策有效性的人机混合决策方法研究

作者姓名： 游诗艺

学科专业： 控制科学与工程

导师姓名： 康宇教授 赵云波教授

完成时间： 二〇二二年五月二十六日

University of Science and Technology of China

A dissertation for master's degree

Research on Human-Machine Hybrid Decision-Making Method Based on the Effectiveness of Human Decision

Author: You Shiyi

Speciality: Control Science and Engineering

Supervisors: Prof. Yu Kang, Prof. Yun-Bo Zhao

Finished time: May 26, 2022

中国科学技术大学学位论文原创性声明

本人声明所呈交的学位论文，是本人在导师指导下进行研究工作所取得的成果。除已特别加以标注和致谢的地方外，论文中不包含任何他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的贡献均已在论文中作了明确的说明。

作者签名：_____ 签字日期：_____

中国科学技术大学学位论文授权使用声明

作为申请学位的条件之一，学位论文著作权拥有者授权中国科学技术大学拥有学位论文的部分使用权，即：学校有权按有关规定向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅，可以将学位论文编入《中国学位论文全文数据库》等有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。本人提交的电子文档的内容和纸质论文的内容相一致。

控阅的学位论文在解密后也遵守此规定。

☑ 公开 ☐ 控阅（____ 年）

作者签名：_____ 导师签名：_____

签字日期：_____ 签字日期：_____

摘　要

随着人工智能技术的发展，机器的自主能力不断地提高，智能机器在各行各业的应用和发展日益深入。在此进程中，不可避免地会遇到智能机器无法应对实际任务的复杂性和不可预测性的情况，许多系统在未来仍将需要人类在监督、目标设定、应急响应等方面与机器进行持续、密切的交互，研究此种场景下如何混合人类决策和机器决策以达到更好的决策效果也因此尤为重要和有意义。

在人机混合决策中，人类决策是否有效，即人的决策是否促进任务的完成并有效地反映人类的真实意图，从两方面影响着最终的决策性能。一方面在于一方决策失效将导致混合性能的下降；另一方面在于智能机器常常无法直接得知人的意图，而需先根据人类决策推测意图，再做出决策辅助人完成该意图，人类决策的失效可能导致意图推理的失效，进而导致人机混合决策方法的失效和任务失败。因此本文以人机混合决策方法为研究对象，基于人类决策的有效性，从人类决策全时有效和人类决策非全时有效两个方面展开研究，提出基于强化学习的人机混合决策方法来改善决策性能。本文的研究工作主要包括以下两个方面：

（1）针对人类决策全时有效的情况，提出一种基于最小干预原则的人机混合决策方法，在优化整体系统性能的基础上，进一步考虑人对于人机系统满意度的相关指标。通过将最小干预原则引入人机混合决策，设置人机决策融合的自适应阈值，该方法能够以最小程度的干预为人类提供最大程度的帮助，并能在实时变化的环境中保持最优，同时提升和改善系统性能和人类满意度两类指标，为后续的优化设计方案提供基础性方法。

（2）针对人类决策非全时有效的情况，提出一种基于人类决策有效性评估机制的人机混合决策方法，以避免人的无效决策损害系统性能。通过利用强化学习算法判断人类决策的有效性，识别人的意图是否改变，该方法能够在人类决策无效时由机器单独完成任务，使得系统在人类决策非全时有效的情况下，仍能完成正确的任务目标，有效提升了人机混合决策质量和系统性能。

关键词：人机系统；混合决策；决策有效性；仲裁；强化学习

ABSTRACT

With the development of artificial intelligence technology, the autonomous ability of machines has been continuously improved, enabling intelligent machines to be applied and developed in all walks of life. In this process, it is inevitable to encounter situations where intelligent machines cannot cope with the complexity and unpredictability of real tasks, and many systems will still require humans to continuously work with machines in supervision, goal setting, emergency response, etc. in the future. Therefore, it is particularly important and meaningful to study how to mix human decision-making and machine decision-making to achieve better decision-making effects in this scenario.

In human-machine hybrid decision-making, whether human decision is effective, that is, whether human decision promotes the completion of tasks and effectively reflects human's true intentions, affects the final decision-making performance from two aspects. On the one hand, the failure of one decision-making will lead to the decline of the hybrid performance. On the other hand, intelligent machines usually cannot directly know the intention of human, but need to infer the intention based on human decisions, and then make decisions to assist human to complete the intention. The failure of human decision-making may lead to the failure of intention inference, which in turn leads to the overall failure of machine decision-making and human-machine hybrid decision-making methods. Therefore, taking the reinforcement learning algorithms as the tool and the human-machine hybrid decision-making methods as the research object, this dissertation conducts research from two aspects: the full-time effective human decision-making and the non-full-time effective human decision-making, and proposes human-machine hybrid decision-making methods based on reinforcement learning to improve the decision performance. The research work of this dissertation mainly includes the following two aspects:

(1) Aiming at the situation that human decisions are always effective, a human-machine hybrid decision-making method based on reinforcement learning and following the principle of minimal intervention is proposed. It is noted that most of the existing methods only optimize the system performance, while ignoring the indicators of human satisfaction with the human-machine system. This method introduces the principle of minimal intervention into the human-machine hybrid decision-making, sets the adaptive threshold of human-machine decisions fusion, and provides the greatest help for humans with minimal intervention, so that the method can remain optimal in the

changing environment, and improves the indicators of system performance and human satisfaction at the same time, providing a basic method for the subsequent optimization design scheme .

(2) Aiming at the situation that human decisions may be invalid, a human-machine hybrid decision-making method based on reinforcement learning is proposed. It is noted that most of the existing methods assume that human decisions are always effective. This method considers the situation that human decisions may be ineffective. By judging the effectiveness of human decisions and identifying whether the human's intention has changed, the method can make the machine complete the task alone when the human decision is ineffective and avoid ineffective decisions from damaging the system performance, so that the system can still complete the correct task goal in the case of long-term ineffective human decisions, which effectively improves the task success rate.

Key Words: Human-Machine Systems; Hybrid Decision-Making ; Effectiveness of Decision; Arbitration; Reinforcement Learning

目　　录

插 图 清 单

表 格 清 单

第 1 章　绪　　论

本章首先介绍了人机混合决策的产生背景和重要研究意义，然后阐述了人机混合决策在自动化领域的研究现状和不足，最后给出了全文的主要工作和结构安排。

1.1　研究背景和意义

得益于机器学习算法的突破，互联网和物联网支持下大数据的可得，和以 GPU 计算为代表的计算能力的提升，以深度学习为基础的人工智能技术在近些年蓬勃发展开来。人工智能技术的发展使得由其赋能的机器具有强大的智能自主能力，它们能够构建自己的行为策略，包括目标预测、战略规划和行动执行，而非单纯地执行预先定义的行为。这些智能机器已被应用于越来越多的领域中，比如：

（1）监测人类驾驶员的注意力从而提升驾驶的安全性。据中国疾控中心机动车安全处报道，五分之一的交通事故是由于驾驶员注意力不集中而发生的，这也导致超过 425000 人受伤，大约 3000 人丧生。未来的自动驾驶很可能依然需要人的深度参与，在这种人机共同参与的驾驶中，有可能通过机器的参与提升驾驶的安全性，比如，可以利用装置在汽车前挡风玻璃处的摄像头采集人脸信息，使用基于深度学习的图像识别技术判断人的注意力是否集中，在人精神状态饱满的时候将车辆交由人控制，而在人的驾驶出现危险的情况下允许辅助驾驶系统的智能介入[1]，如图1.1a所示①。

（2）遥操作微创外科手术。达芬奇外科手术系统是一套由人操控机器人进行微创手术的外科手术系统，可用于心脏瓣膜修复等手术过程，如图1.1b所示②。截止 2012 年该手术机器人已进行了超过 20 万次手术。在手术过程中，病人躺在特制的手术车中，外科医生则在独立的操作台上通过观看患处放大后的 3D 高清视频，操控手术车上的机械臂进行手术。这一系统充分利用了外科医生的医疗知识（知道在何处进行何种操作）和 3D 高清视频技术对患处的放大及机械手在极小尺度下的精确度，达到了单纯人或机器都无法达到的效果：人很难在微创的极小尺度下精确操作，而机器人则难以判断如何进行手术[2]。

然而，将深度学习驱动的人工智能技术应用于自动控制领域存在着一些缺陷，控制的要求和深度学习的特性之间存在若干难以调和的矛盾。具体包括：

（1）动态实时要求和计算复杂度的矛盾[3-4]。自动化控制应用具有动态和实

① 见：http://m.cheyuansu.com.cn/news_show_250.html。访问日期：2022 年 3 月 14 日。
② 见：https://www.huxiu.com/article/340275.html。访问日期：2022 年 3 月 14 日。

(a) 驾驶行为识别和疲劳检测系统　　**(b) 达芬奇外科手术系统**

图 1.1　智能机器在当前生活中的应用场景

时性的特点，所需的信息如果错过了当前时刻，往往就会变得难以利用甚至毫无价值。现有的 AI 算法在面向动态实时的要求时存在着本质的限制。首先，硬件设施的计算能力和能耗各方面在面向复杂的 AI 算法和海量数据时具有限制；其次，自动化控制系统的动态性质使得算法依赖的系统模型可能随时间而演化，因此可能有必要针对系统底层模型的变化重新进行 AI 算法的建模和训练，这对硬件的计算能力和系统的实时性要求都提出很大的挑战。

（2）可信要求和不确定性的矛盾[2,5]。深度学习算法自身缺乏可解释性，本质上固有不确定性，这意味着深度学习算法的引入可能在决策层面引入了额外的不可信因素。尽管反馈控制本身可以有效处理环境和模型的不确定性，但针对深度学习算法所引入的不确定性，目前仍然缺乏统一的方法框架进行处理。从深度学习自身出发，有基于贝叶斯框架的不确定性刻画和分析方法，对其可解释性的讨论和探究，以及 Trusted AI 的相关研究，但现有研究仅在起步阶段，离问题的真正解决路程尚远，甚至对问题是否能够得到彻底解决仍存有争议。

（3）鲁棒要求和攻击脆弱性的矛盾[6-7]。因为所处环境的不确定性，自动化控制应用需要满足很强的鲁棒性，大多数自动控制系统在存在一定程度的有界噪声时都可良好运行。但深度学习算法有着特殊的脆弱性，比如对于人眼无法分辨区别的原图片和加以噪声的干扰图片，深度学习算法可能会给出完全不同的分类结果：如图1.2所示，Goodfellow et al.[6]提出在原本正确分类的熊猫图片的基础上加 0.07 倍的噪声，人类肉眼无法察觉这种细微的调整，但算法以 99.3% 的高置信度将其分类为长臂猿。人工智能算法和人的大脑的工作方式不同，倘若让人描述如何识别图片中的物体是熊猫，人可能会描述目标特征，如黑白皮毛、体型大小、目标背景、目标姿势等，而深度学习算法则根据每个像素值和神经网络的权重，通过公式进行计算给出答案。换句话说，只要精确调整图片的像素值，便可精确引导算法的计算结果。攻击者便可利用这种算法的脆弱性，对人脸识别、目标检测等重要任务发起攻击。这种脆弱性自身，以及在开放应用中受到针对性攻击时，如何保障使用深度学习算法的自动化控制应用的鲁棒性，是一个

在现有的鲁棒性框架下难以回答的问题。

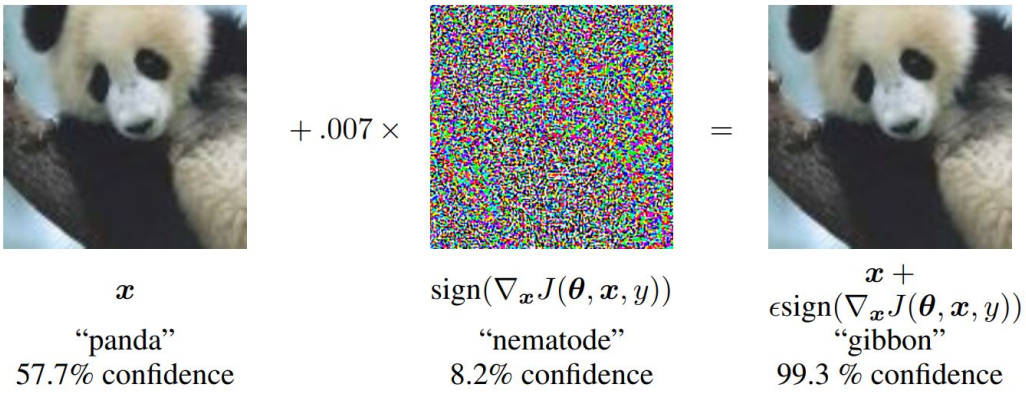


图 1.2　深度学习算法脆弱性的示意图

另一方面，人在自动控制系统中有着其不可替代性，使得自动化无法、也不能完全取代人类智能，包括：

（1）若不显式考虑人的因素，或没有人的深度介入，则系统就不能达到所设的目标，或者根本不可能（没有人的因素的显式考虑，遥操作的微创外科技术就是不可能的），或者效果不好（霍金的轮椅设计如果不全面考虑霍金本人需求，则使用效果必然会打折扣）。

（2）人在人机系统中有其特殊作用和地位[8-9]。一方面，人是控制系统目标的终极来源，任何系统的设计总是为人服务的，人是控制系统存在的原因，为控制系统设定目标、赋予价值。另一方面，人的一些特殊能力仍然是现有机器无法达到的，比如直觉，系统需要利用人的能力达到更好的性能。

出于以上因素，由人和 AI 驱动的机器进行混合决策成为解决实际任务的复杂性和不可预测性的可行方案[10]。人的智能和机器智能在自动化控制领域的共融共存导致了所谓的“人机混合智能系统”的出现[11-14]，其典型框架如图1.3所示。这一新型的系统形式和智能形式在两方面具有本质的重要性：一方面，从自动化控制角度来说，人机混合决策所代表的系统结构形式是传统自动化控制系统应对 AI 赋能的机器智能变革的必然发展形式；另一方面，从智能科学的角度来说，人机混合决策所代表的智能形式也成为人工智能未来发展的重要甚至是唯一的终极形式。这两方面本质上的重要性使得建立相关领域的理论和方法框架变得极为迫切和重要。人机混合决策关注于实现人的决策和机器决策的有效融合，以达到更好的决策效果[15-17]，现已被广泛应用于许多领域，如机器遥操作[11,18]、半自主驾驶[19-20]、康复外骨骼[7,21]等。由于人工智能的快速发展，这一领域正受到越来越多的关注。

在人机混合决策中，人类决策是否有效，即人的决策是否促进任务的完成并有效地反映人类的真实意图，从两方面影响着最终的决策性能。一方面在于一方决策失效将导致混合性能的下降；另一方面在于智能机器常常无法直接得知人

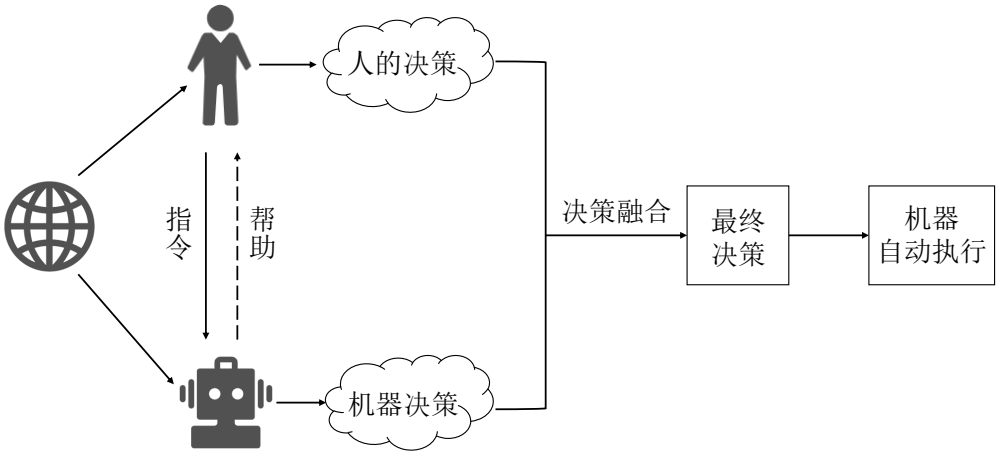


图 1.3　人机混合智能系统框架

的意图，而需先根据人类决策推测意图，再做出决策辅助人完成该意图，人类决策的失效可能导致意图推理的失效，进而导致机器决策和人机混合决策方法的整体失效。这促使本文以人机混合决策方法为研究对象，基于人类决策的有效性，从人类决策全时有效和人类决策非全时有效两个方面展开研究，提出新的人机混合决策方法来改善决策性能。

1.2　国内外研究现状

近年来随着人工智能的发展，智能机器已然能够建立自己的行为策略，包括目标预测、战略规划和行动执行，而非仅仅执行由人预先定义的行为。然而，由机器单独完成任务的完全自主仍未实现，我们距离用自动化取代人力征途尚远。主要原因是环境甚至系统本身都在动态变化，很难一劳永逸地设计系统，许多自动控制系统仍然要求人类在监督、目标设定、应急响应等方面与智能机器进行持续、密切的交互。人类和智能机器的混合决策是解决实际任务的复杂性和不可预测性的可行方案，也因此受到越来越多的关注。

人机混合决策在自动化领域的现有研究大体上可以分为介入控制和共享控制两类，这两类控制模式对人类决策的有效性有着不同的要求。介入控制即人和机器在感知-行动的环路中通过一方干预另一方的方式完成某一动态任务，这一动态任务是人或机器在理想环境都可以独立完成的[22]。介入控制是基于一定评价机制或任务目标的强决策者对弱决策者的干预，人机为非平等的主从关系，具体又可分为机器介入人和人介入机器两个子类，人类决策的有效性在其中有明确的限定。共享控制即人和机器在感知-行动的环路中以平等协作的方式完成某

一动态任务，更多强调二者针对当前情形给出的决策的融合，以取得优于双方单独控制的效果，因此其基本假设为人类决策始终有效。

共享控制和介入控制的本质区别在于人与机器的交互方式是“平等协作”还是“一方干预另一方”。这一交互方式的不同造成了采用这两种控制方式的人机混合智能系统在控制目标、系统结构、人机主次地位、人机相互影响、系统设计要求等各方面的不同，具体的特点对比如表1.1所示[8]。

表 1.1　介入控制和共享控制的特点对比

特点	介入控制	共享控制
控制目标	主要为了防止人机系统发生不可接受的后果，以系统的稳定和安全为主	主要为了系统性能的提升，以系统的优化为主
系统结构	一般只需要一方观测另一方的状态并决定介入时机，因而不需要人机之外的额外机制	可能需要高于人机的总的“仲裁机构”执行人机共享策略
主次地位	人机具有不同的地位, 介入方比被介入方具有更高的地位，拥有更大的决策权	人机地位平等，不存在某一方的地位始终高于另一方
相互影响	单向或双向影响。比如辅助驾驶系统既可以由人接管机器驾驶，也可机器接管人的驾驶；防碰撞系统则是机器介入人的单方面行为。	共享控制是人机合作协同，相互影响
设计要求	执行机制简单，对介入的时机和强度的要求较为严格，介入失误可能导致系统失控	执行机制复杂，对决策的设计较为宽容，失误的后果多是优化不力或性能次优而非系统损毁

1.2.1　机器介入人

机器介入人即系统在正常运行时由人的智能自主和机器的自动化执行共同起作用，而在特定条件下允许机器的智能强制剥夺人的自主性，以避免人在此时的自主决策所可能导致的严重后果[23-24]。在某些人机混合决策场景中，人一旦超出了所被允许的自主性边界，将会引发严重的后果，在这种情况下，允许机器智能的强制性介入、甚至临时剥夺人的自主性，成为可行的人机协作策略。比如

各种以安全性为目标的驾驶辅助系统，在人分心驾驶或遇到人无法处理的紧急情况时，允许辅助驾驶系统的强制接管避免潜在的危险。这类策略明确考虑了人类决策无效的情况，并将人类决策有效性的判定转化为人的自主性边界的计算，这也是机器介入策略设计的本质难点所在。

Broad et al.[25]通过预测系统的安全性来计算人的自主性边界，并提出在定义机器的安全约束以维护系统安全的同时，不应限制人实现不确定行为的能力。方法使用 Koopman 算子估计系统模型，用模型预测控制的方法实现混合决策。Lam et al.[26]提出传统的人工辅助功能只靠监控机器的状态来提高安全性，但下一代系统应同时监视人的状态，并在必要时候加以接管。方法将辅助驾驶建模为部分客观马尔可夫决策过程，以状态转移概率和观测概率的已知为前提条件。Bruemmer et al.[27]则认为在移动机器人搜索应用程序中，如果机器人认为人类的命令具有潜在危险，它应保留否决权。

1.2.2　人介入机器

人介入机器即系统在正常运行时由机器的智能自主和机器的自动化执行共同起作用，而在特定条件下允许人的智能强制剥夺机器的自主性，以避免机器在此时的自主决策可能导致的严重后果。很多人机混合决策场景赋予人以最终决策权，比如在移动搜索和救援任务中[28-29]，最终权威往往在于人类；军事领域中人对自主武器保持最终决断；远程遥操作机器人一般也以人的决策为行动标准。这类策略则假设人类决策始终有效，一旦人类做出决策行为，系统将无条件地执行，策略设计的核心问题是如何准确判定机器是否超出了所被允许的自主性边界。

考虑一类人主要为辅助机器完成目标而存在的人机系统[30-31]，典型的例子如汽车驾驶中的人机共驾系统（半自动驾驶）、人对自主武器系统的最终控制等。这类人机系统大多数时候由机器进行控制，但在机器行为明显违背系统安全性等关键指标时，人需要及时介入以保证人机系统的安全性和其他可能的性能指标。机器出现错误的原因，可能是由于所采用的机器智能的本质缺陷（如深度学习算法的弱鲁棒性），也可能是由于人机系统遇到了机器智能从未遇见过的突发情况，但在任何场景下，人类智能的独特优势都提供了改进人机系统整体性能的可能性[31-32]。根据 Phillips-Grafflin et al.[23]的观点, 人和机器均可单独控制被控系统的动作，在人启动任务或输入决策动作后，机器通过跟随输入来自主执行任务，而人负责监视机器的操作效果。Marcano et al.[33]则提出了一个人通过手势实施介入控制的远程操作系统，系统利用感知、评估等模块，根据物体姿势和形状来规划机器人的运动轨迹。

1.2.3　人机共享控制

由于环境的部分可观测性和系统参数的不精确性，许多控制任务对机器来说存在困难，而由于人的有限理性[34]和物理限制（如缺乏多维控制能力）等原因，人类也很难单独完成。共享控制通过结合人机互补的能力来解决此类问题[25,35]，也因此未对人类决策的有效性做明确要求。关于共享控制的定义，目前学术界尚未形成统一架构，这里给出一些较为相关的定义说明。Broad et al.[25]认为共享控制是一种范式，它将一个自主伙伴整合到机器人系统的控制回路中，以帮助人类伙伴完成他们自己无法完成的任务；Oh et al.[36]认为共享控制是定义一个如何混和人的控制和自治控制的策略；Li et al.[37]认为共享控制将机器的角色从被动的跟随者或执行者提升为控制权的协作伙伴，共享控制既可以利用人类在动态、不确定环境中做出决策的适应能力，又可以利用机器的自动化能力，从而更快更轻松地完成任务，并减少对人类的身心需求。Flemisch et al.[38]对比了共享控制和人机协同的概念，指出前者关注控制权的共享，后者关注任务和情境的共享。Marcano et al.[3]则详细介绍了共享控制在自动驾驶领域的理论发展和技术现状，并将其定义为“如何将驾驶员和自动辅助系统的动作良好结合以实现最优的控制效果”。

有部分文章将共享控制定义为“人和机器同时在同一任务上工作”。这种相互作用可以通过两种方式实现：（1）机器扩展人的能力，比如微创外科手术中机器能够在微小的创口中完成精细的操作；（2）机器减轻操作员的总任务工作量，比如具有触觉反馈的方向盘，通过维持方向盘的角度来缓解驾驶员在生理和心理上的负担。“同时”一词区分了共享控制和介入控制，介入控制通常为某一方出于某种考量介入进来，强制剥夺另一方的控制权，比如防碰撞系统的自动刹车，驾驶员和智能机器只有一方参与控制。“同一任务”一词则区分了共享控制和合作控制，即某个任务被划分为多个子任务交由人和机器分别执行，比如人控制方向盘，机器控制油门和刹车。Abbink et al.[22]在这一定义的基础上对共享控制做了更详细的描绘：

- 人和机器同时完成同一任务，排除介入控制和合作控制；
- 人和机器同时参与感知-行动环路，排除各种警报系统，因其只进行感知，没有对被控系统采取控制动作；
- 该任务是人或机器在理想情况下可以独立完成的。

本文认同这一定义，并在此基础上构建了后续研究。当前的人机共享控制方法大多有两个基本组件[8-9]：

（1）目标推理。在一类传统人机系统中，系统的设计目标是由机器辅助实现人计划实现的目标，其中的一个经典问题是实现人的意图推测，因为对机器来

讲，人机系统中人的意图往往是未知且时变的，而如果知道了人类意图，则剩下的问题就不依赖人机系统的框架了。比如人通过无人机上装置的摄像头进行观察，无人机的自动系统可保证自身稳定飞行，但它并不知道人类的观察目标，因而对下一步如何运动缺乏预期，也进而影响了无人机自身飞行动态的优化。实现意图推测大致有两类方法。第一类方法是已知一个可能的目标集（如机械臂抓取任务中所有可以抓取的物体），其中存在人的真正目标，机器根据人的历史控制行为计算目标集上的概率分布，概率最大者即为人的目标[11-12]；第二类方法是已知一组可能的行为（如向左移动、向右移动等），通过行为推断所有可能的任务（如跟踪轨迹、拾取物体等），计算任务的概率分布，概率最大者即为人的意图[13]。

（2）仲裁。考虑人机协作控制时，仲裁[8,14]是一个核心概念，它决定何时由机器决策、何时由人类决策、以及如何融合机器行为和人类行为。适当的仲裁策略可在正确的时间为各方提供适当的控制权，以最大程度地发挥其优势，并最大限度地减少其劣势。在很多人机协作方法中，机器通过观察人类决策行为推断人类意图达到的目标[39]，机器代理结合预测目标和自身策略，对被控对象的实时环境状态给出行为判断。然后，机器决策行为和人类决策行为同时进入仲裁阶段，由仲裁函数给出最终的决策行为。目前常用的仲裁形式为对人和机器的控制行为的线性组合，其因设置简单、性能优良被广泛应用于共享控制系统中[9]。

1.3　本文工作和结构安排

人机混合决策任务具有序贯决策的特性，这使得其常被建模为马尔可夫决策过程，而环境、人、机器及其相互之间的交互广泛存在不确定性，人与机器的状态和行为也往往难以全面准确观测，这使得部分可观马尔可夫决策过程等模型成为人机混合决策任务中解决很多问题的常用工具。但基于这此类模型的方法通常需要已知状态转移函数、观测概率函数和奖赏函数等先验知识，这限制了方法的适应性和通用性：状态转移概率在很多任务中无法获得或因人而异，而对系统目标的固定表示（如离散的可抓取对象）降低了系统执行任务的灵活性，再者，较大的算力需求也影响了方法在复杂场景中的实时控制。强化学习方法无需其他数据、在与环境交互的过程中便可学习策略，在消除对先验知识的依赖上具有独特的优势，并且强化学习算法能够估计每个决策动作的价值大小，为评估人类决策的有效性提供了可能。因此本文研究如何利用强化学习算法解决人机混合决策在不同的人类决策有效性下存在的问题，试图提供实现人机混合智能系统的另一条途径。

全文共五章，具体结构安排描述如下：

第一章阐述本文的研究背景和研究意义，介绍人机混合决策相关研究的分类及当前进展，给出文章的组织结构和各章内容。

第二章介绍本文研究涉及的相关概念和基础知识，后续研究基于这些概念展开，部分方法基于现有算法提出，故做简单介绍为后续打好基础，具体包括马尔可夫决策过程、强化学习和深度强化学习。

第三章研究了人类决策全时有效下的基于最小干预原则的人机混合决策方法。讨论了在人类决策均有效的约束下，如何将人的决策和机器决策相结合以获得更高质量的决策动作，从而实现更好的决策效果。具体涉及最小干预原则的引入、共享控制模式的适用、仲裁机制的设计等，实现了算法的仿真验证，为后续的优化设计方案提供基础性方法。

第四章研究了人类决策非全时有效下的基于人类决策有效性评估的人机混合决策方法。讨论了在人类行为可能无效的约束下，如何将人的决策和机器决策进行混合以获得更高质量的决策动作，包含人介入机器、机器介入人和人机共享控制三种模式。具体涉及人类决策有效性的评估、对无效人类决策的处理、三种控制模式的切换等，并实现了算法的仿真验证。

第五章总结全文的内容和贡献，简述本文研究存在的不足和对未来工作的展望。

第 2 章　相关基础知识

本章介绍本文研究所涉及的相关概念和基础知识，后续研究基于这些概念展开，部分方法基于现有算法提出，故在此做基础性介绍，具体包括马尔可夫决策过程、强化学习和深度强化学习。

2.1　马尔可夫决策过程

人机混合决策任务大多为序贯决策过程。序贯决策是一类具有时序和多阶段特点的决策问题，其中智能体（Agent）在每个离散的时间步与动态系统进行迭代式交互，其决策过程可形式化表示为图2.1。在每个时间步 t 系统处于状态 s_t，智能体根据当前系统状态，按照当前策略 π_t 选择并执行动作 a_t，系统将到达下一个状态 s_{t+1}，智能体获得环境给予的反馈收益 r_t 并更新策略 π_{t+1}。如此循环进行动作选择和状态演化，智能体的目的是使自己获得的收益最大化。序贯决策与其他决策问题最大的不同在于，其他决策问题大多为了优化当前时刻的指标，序贯决策则需优化未来一段时间内的累积收益，而这也是序贯决策的难点所在：需要基于过去和现在的记录预测未来收益，在即时收益和累积收益之前进行权衡，且求解空间随着时间长度的增加呈指数增加。

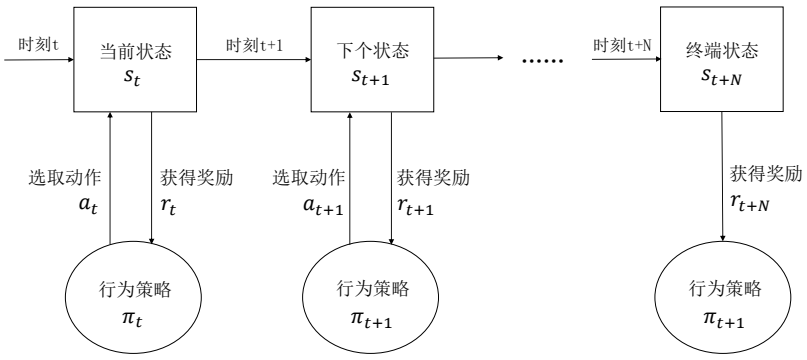


图 2.1　序贯决策的状态转移示意图

序贯决策问题可以分为两类：随机序贯决策和确定性序贯决策[40]。其中确定性序贯决策即决策后系统状态是确定的，s_t 以概率 1 转移至新状态 s_{t+1}，其状态转移方程和损失函数也是确定的。故在此类问题中，可根据当前状态和决策得到确定的状态和收益链，从最后一个状态向前逆推，直至推得第一个状态为止，便可确定整个决策问题的最优解。一个典型的例子是最短路径问题。

随机性序贯决策即系统以一定的概率分布转移至新状态，做出决策后形成随机过程而非得到确定的状态和动作链[41-42]。此类问题一般通过状态转移概率和损失函数利用动态规划方法进行求解，倘若状态转移概率未知，则需要智能体执行一系列的决策，获取关于系统环境的信息，对假设的先验概率进行修正，并求解得到后验概率进而做出最优决策。倘若假设系统的状态只与上一时刻的状态和动作有关，而和更早的信息无关，即决策过程满足马尔可夫性，则形成马尔可夫决策过程（Markov Decision Process, MDP）。马尔可夫决策过程因其灵活的表达能力和建模能力而被广泛应用于序贯决策问题中。

马尔可夫决策过程可表示为四元组 (S, A, T, R)[43-45]，元组中每个元素的具体含义说明如下：

- $S = \{s_0, s_1, \ldots, s_N\}$ 为状态集合，包含决策过程中系统可能出现的所有状态；
- $A = \{a_0, a_1, \ldots, a_N\}$ 为动作集合，包含决策过程中智能体可能执行的所有动作；
- T 为状态转移函数，其中包含概率 $P(s_{t+1}|s_t, a_t) \in [0, 1]$ 表示智能体在系统状态 s_t 下执行动作 a_t，系统转移到状态 s_{t+1} 的概率。状态转移过程满足马尔可夫性即满足公式2.1：

$$P(s_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}, ..., s_0, a_0) = P(s_{t+1}|s_t, a_t) \tag{2.1}$$

状态转移函数应满足公式2.2：

$$\sum_{s_t, s_{t+1} \in S} P(s_{t+1}|s_t, a_t) = 1 \tag{2.2}$$

- R 为收益函数，每个时刻的即时收益 $r_t(s_t, a_t)$ 表示智能体在系统状态 s_t 下执行动作 a_t 获得的反馈收益。

Shachter 等人将马尔可夫决策过程形象化地描述为图2.2所示[46]。图中的结点为状态或动作，边表示影响关系，每一时刻的状态和收益都由上一时刻的状态和动作共同决定。

在 MDP 模型中系统状态可被完全观测到，即其对于智能体是完全可知的，但在实际任务中，系统状态往往不可知或部分可知，智能体需根据观察到的状态对系统状态进行推断再做出决策，这种模型被称为部分可观马尔可夫决策过程（Partially Observable MDP,POMDP）[47]。POMDP 是对 MDP 的泛化，其系统状态的演变同样具有马尔可夫性，只是智能体只能得到状态的部分观测值，比如在实际任务中传感器只能采集到有限的环境信息，故 POMDP 模型比 MDP 更加接近于实际工程。

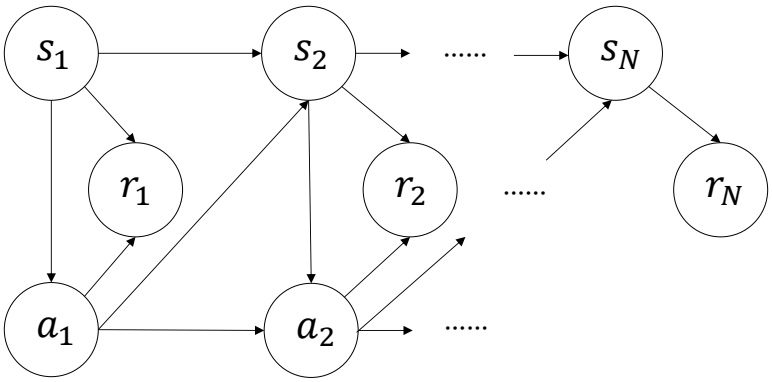


图 2.2　马尔可夫决策过程示意图

POMDP 常被表示为六元组 (S, A, T, R, Z, O)[48-49]，元组中每个元素的具体含义说明如下：

- $S = \{s_0, s_1, \ldots, s_N\}$ 为隐藏的系统状态集合，包含决策过程中系统可能出现的所有隐藏状态；
- $A = \{a_0, a_1, \ldots, a_N\}$ 为动作集合，具体含义与 MDP 相同；
- T 为隐藏状态的转移函数，具体含义与 MDP 相同；
- R 为收益函数，具体含义与 MDP 相同；
- $Z = \{z_0, z_1, \ldots, z_N\}$ 为智能体观测到的状态的集合;
- O 为观测概率函数，包含概率 $P(z_{t+1}|s_{t+1}, a_t)$，表示智能体在执行动作 a_t，系统状态转移到 s_{t+1} 后观测到状态 z_{t+1} 的概率。观测概率函数应满足公式2.3：

$$\sum_{z \in Z} P(z|s, a) = 1, \forall s \in S \tag{2.3}$$

POMDP 模型使用信念状态（belief state）b 描述系统隐藏的状态空间，即 $b(s)$ 表示当前系统隐藏状态为 $s \in S$ 的概率。智能体的策略 π 为从信念状态到决策动作的映射，即 $\pi(b) \in A$。在每一个时间步，系统隐藏状态为 $s \in S$，智能体根据策略和信念状态选择动作 $a = \pi(b)$，收到反馈收益 $r(s, a)$，系统状态演变至新状态 s'，智能体观测到状态 z'，则信念状态的更新为：

$$\begin{aligned} b'(s') &= P(s'|z', a, b) \\ &= \frac{P(z'|s', a, b)P(s'|a, b)}{P(z'|a, b)} \\ &= \frac{P(z'|s', a)\sum_{s \in S} P(s'|s, a)b(s)}{P(z'|a, b)} \end{aligned} \tag{2.4}$$

其中 $P(z'|a, b)$ 可以看作归一化项。

2.2　强化学习

人机混合智能任务具有序贯决策的特性，这使得其常被建模为 MDP 问题，而环境、人、机器及其相互之间的交互广泛存在不确定性，人与机器的状态和行为也往往难以全面准确观测，这使得 POMDP 模型成为人机协同任务中解决很多问题的常用工具。比如 Javdani et al.[9]和 Javdani et al.[50]将共享控制建模为人的目标具有不确定性的 POMDP；Lam et al.[26]使用 POMDP 为人在回路控制系统构建了一个统一的框架，以便系统监控人和机器的状态，并在必要时提供反馈。但基于这两种模型的方法通常需要已知状态转移函数、观测概率函数和奖赏函数等先验知识，这限制了方法的适应性和通用性：状态转移概率在很多任务中无法获得或因人而异，而对系统目标的固定表示（如离散的可抓取对象）降低了系统执行任务的灵活性，再者，较大的算力需求也影响了方法在复杂场景中的实时控制。应对这些困难，无需其他数据、在与环境交互的过程中便可学习策略的强化学习方法具有独特的优势。

强化学习问题在数学上的理想形式为 MDP 模型，MDP 是通过交互式学习实现目标的理论框架，在这个框架下可以进行精确的理论说明[43]。强化学习问题关心作为强化学习本体的智能体（Agent）在所处环境中的行动，问题的一般描述可借由四元组 (s, a, p, r) 表示，其中状态 s 表示智能体的位置、速度等状态信息；动作 a 表示智能体所可能采取的左移、前进等动作；状态转移概率 p 为智能体在某一状态采取某一动作后转移到下一状态的概率；奖赏 r 是智能体执行动作后得到的环境给予的反馈信号。智能体以离散时间步与环境进行交互，它在每个时间步上依据当前状态 s 选择动作 a 执行，在动作 a 下状态以概率 p 移至 s'，并获得奖励 r，其结构示意图如图2.3所示。然后继续选择动作，循环往复，直至任务结束。

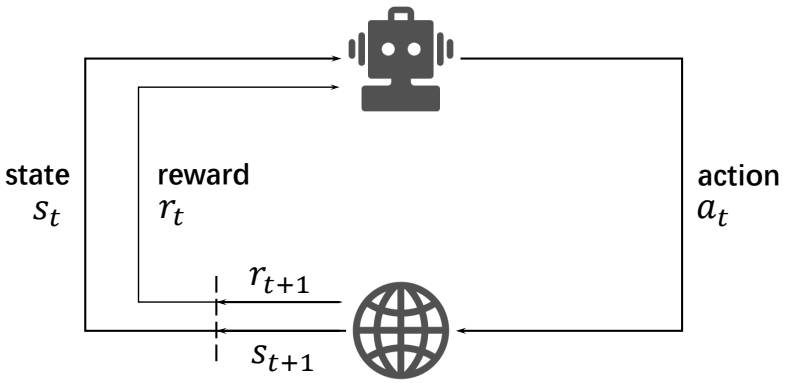


图 2.3　强化学习结构示意图

为了达到强化学习问题的最终目标，智能体一般需要执行多步动作，状态也

经过多步转移。强化学习问题的一般设计目标，是选取合适的行动序列（或称策略 π），使得智能体在达成最终目标过程中的多步行动和状态转变所导致的奖赏的和（或称累积奖赏），达到其最大值。

为了求取累积奖赏最大的最优策略 π^*，一般会首先定义状态价值函数 $V_\pi(s)$ 和动作价值函数 $Q_\pi(s,a)$。前者是智能体从状态 s 开始按照策略 π 进行决策能够获得的累积奖赏的期望值，后者是智能体在状态 s 时执行动作 a，后续按照策略 π 进行决策能够获得的累积奖赏的期望值。然后利用贝尔曼方程对这两个函数进行形式化表示, 如式2.5所示，并进而优化求解。具体求解方法可分为三类，即动态规划法、蒙特卡洛法和时间差分法。

$$\begin{aligned} V_\pi(s) &= \sum_{a\in A}\pi(s,a)\sum_{s'\in S}P(s'|s,a)(r+\gamma V_\pi(s')) \\ Q_\pi(s,a) &= \sum_{s'\in S}P(s'|s,a)(r+\gamma V_\pi(s')) \end{aligned} \tag{2.5}$$

- 动态规划法在其求解过程中，一般先将问题分解为子问题，由子问题的最优解构成原问题的最优解，并通过记住求解过的子问题来节省时间，这要求两个性质：（1）整个问题的最优解可以通过求解子问题得到；（2）子问题的求解结果可以存储下来并再次使用。强化学习任务满足上述性质：贝尔曼方程给出了递归分解方法，最优贝尔曼方程的唯一解即是最优值函数，如式2.6所示；值函数可以作为子问题的求解结果。将动态规划法应用于强化学习中，首先通过策略评估计算给定策略 π 的优劣程度，然后计算策略 π 的最优状态价值函数 $V_\pi^*(s)$，根据最优状态价值函数 $V_\pi^*(s)$ 进而确定最优策略 π^*。尽管动态规划法具有方法简单、优化结果更好的优点，但其求解以已知完整环境模型为前提，这大大限制了该方法在实际中的应用范围。

$$\begin{aligned} V^*(s) &= \max_{a\in A}\sum_{s'\in S}P(s'|s,a)(r+\gamma V^*(s')) \\ Q^*(s,a) &= \sum_{s'\in S}P(s'|s,a)(r+\gamma \max_{a'\in A}Q^*(s',a')) \end{aligned} \tag{2.6}$$

- 蒙特卡洛法和时间差分法是无模型算法，可用于状态转移概率或奖赏值未知等信息不完全情况下最优策略的求解。在蒙特卡洛法中，根据样本求解最优策略。比如在初始状态 s 遵循策略 π 最终获得奖赏值 R 为一个样本，根据许多个样本便可估计在状态 s 下遵循策略 π 的期望回报，蒙特卡洛法即依靠样本的平均回报解决学习问题。但该方法存在一些不足，比如数据方差大、收敛速度慢等，导致其在实际任务中的运行效果并不理想。
- 时间差分法则结合了蒙特卡洛和动态规划的优点，能够更准确高效地求解强化学习任务。时间差分法和蒙特卡洛一样从样本中学习，和动态规划一

样基于已经学习过的状态估计新状态，因此时间差分可以学习不完整的样本，即任务尚未完成，未获得总回报 R 时，时间差分法可基于已有状态推测任务结果，同时持续更新这个推测，而蒙特卡洛法只能在任务结束后进行学习。时间差分法主要有同轨策略（on-policy）的 Sarsa 算法和离轨策略（off-policy）的 Q-learning 算法两种，算法流程分别如算法2.1和算法2.2所示。这两种算法的区别和联系如下：

- 由算法2.1可以看出，在 Sarsa 算法中，当智能体处于状态 s 时，根据当前 $Q(s,a)$ 及一定的策略选取动作 a，得到一下步的状态 s' 和奖赏值 r，并再次根据当前 $Q(s,a)$ 及相同策略选择动作 a'。即 Sarsa 算法中动作价值函数的每一次更新都需已知五元组 (s,a,r,s',a')，选择动作时遵循的策略和更新函数时遵循的策略是相同的。
- 由算法2.2可以看出，在 Q-learning 算法中，当前步的 $Q(s,a)$ 更新完毕再根据新状态 s' 选取动作 a'，即函数更新时采用的策略不同于选择动作时采用的策略，动作价值函数的每一次更新只需已知四元组 (s,a,r,s') 即可。
- 由上可知，Sarsa 为同轨策略，每一次参数更新都需要同环境交互，采集新的经验样本进行学习；而 Q-learning 为离轨策略，可以学习过往的经验和数据，比 Sarsa 算法有更高的样本效率。整体效果上 Q-learning 的学习效果更好，但 Sarsa 收敛更快。本节对这些方法不做进一步介绍，感兴趣的读者可阅读 Sutton et al.[43]。

算法 2.1　Sarsa 算法步骤

初始化参数步长 α 和折扣因子 γ；
初始化动作价值函数 $Q(s,a)$；
while *episode=1,2,...,M* **do**
    初始化状态 s；
    基于动作价值 $Q(s,a)$ 选取当前状态 s 下的动作 a；
    while *t=1,2,...,T* **do**
        执行动作 a，获得奖赏值 r 和下一状态 s'；
        基于 $Q(s,a)$ 选取当前状态 s' 下的动作 a'；
        $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$；
        $s = s'$，$a = a'$
    end
end

算法 2.2　Q learning 算法步骤

初始化参数步长 α 和折扣因子 γ；
初始化动作价值函数 $Q(s,a)$；
while *episode=1,2,...,M* **do**
    初始化状态 s;
    while *t=1,2,...,T* **do**
        基于动作价值 $Q(s,a)$ 选取当前状态 s 下的动作 a；
        执行动作 a，获得奖赏值 r 和下一状态 s'；
        $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$；
        $s = s'$
    end
end

2.3　深度强化学习

早期的强化学习算法主要基于表格的方式求解状态集和动作集离散且有限的任务，比如 Q-learning 和 Sarsa 算法。表的横纵坐标分别为状态和动作，每一格代表在当前状态 s 下执行动作 a 能够获得的奖赏值 $r(s,a)$，完善此表即可找到累积奖赏最大的决策链，即智能体完成了学习。但在实际环境中，大部分任务的状态集和动作集都有较高维度，以至于无法使用表格进行记录和索引，因此难以采用传统的强化学习算法进行求解。

深度强化学习结合了强化学习和深度学习，使用强化学习定义问题和优化目标，使用深度学习求解策略函数或价值函数，并使用反向传播算法优化目标函数。深度强化学习基于已获得的智能体和环境的交互数据训练神经网络拟合价值函数，网络输入是从系统观测到的状态，输出是动作空间中的每个动作在当前状态能够获得的累积奖赏值，奖赏值的分布构成当前时刻的行为策略，进而完成动作的选取和执行。深度强化学习无需在表格中记录具体 Q 值，而是使用神经网络估计和预测 Q 值，通过最小化损失函数来更新网络，从而学习最优策略，其流程结构如图2.4所示。

DQN（Deep Q Network）[51-52]是将强化学习和深度学习成功结合的开端，它将卷积神经网络和 Q-learning 相结合。网络的输入是环境状态向量，输出是所有动作在该状态下的 Q 值，进而得到将要执行的动作，实现了从环境状态到动作的端到端映射。

将深度学习应用到强化学习中有诸多挑战，其中之一便是深度学习通常假设数据样本独立同分布，而强化学习中作为训练样本的状态通常是高度相关的序列。DQN 的关键技术之一就是采用了经验回放，将每次和环境交互得到的奖

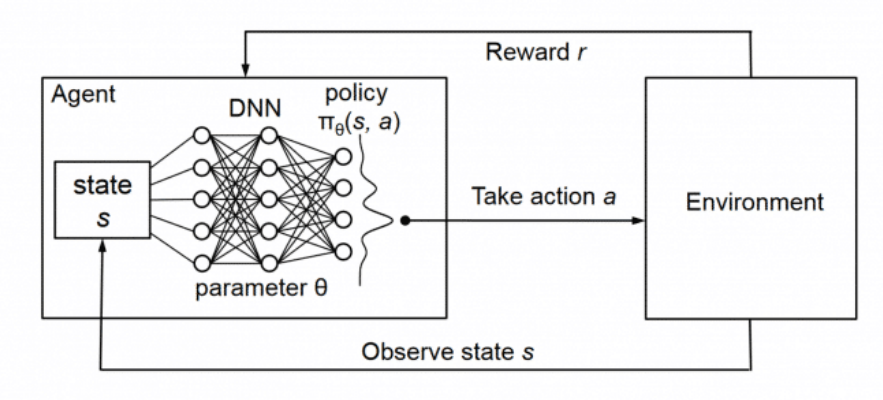


图 2.4　深度强化学习结构示意图

励和下一步状态以四元组 (s_t, a_t, r_t, s_{t+1}) 的形式存储在大小有限的经验池中，数据记录满后从中随机均匀采样作为训练样本进行网络更新，从而打破了数据之间的关联性。经验池的更新为覆盖更新，即下一个四元组会覆盖第一个四元组。

神经网络的训练是一个最优化问题，即最小化损失函数。DQN 中损失函数为目标 Q 值和当前真实 Q 值的差平方，即：

$$\begin{aligned}\text{Target}Q &= r + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta) \\ L(\theta) &= E[(\text{Target}Q - Q(s_t, a_t; \theta))^2] \\ &= E[(r + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta) - Q(s_t, a_t; \theta))^2]\end{aligned}$$

其中 s_{t+1} 为状态 s_t 下执行动作 a_t 得到的下一步状态。第一版 DQN 在 2013 年由 Deepmind 提出[51]，需要用待训练的网络参数计算目标 Q 值，然后再用目标 Q 值进行参数的更新。两者循环依赖，相关性较强，不利于算法的收敛。2015 年 Deepmind 在 Nature 上发表论文[52]，提出用两个结构相同参数不同的神经网络：评估网络和目标网络，来解决这一问题。评估网络用于计算当前 Q 值，使用反向传播算法进行参数实时更新，目标网络用于计算目标 Q 值，每隔一段时间从评估网络复制权重，即延迟更新，以此减少二者之间的相关性。其算法流程如算法2.3所示。

上述两种 DQN 都无法克服 Q-learning 的固有缺陷——会在特定状态下高估某些动作的价值，进而导致过于乐观的值函数估计。Van Hasselt et al.[53]证明了在实际任务中这种高估是常见现象并且会损害算法性能。该文献同时提出 Double DQN 的方法，将动作的选择和动作值函数估计用两个 Q 网络分别进行学习。具体算法上将 2015 年的 Nature DQN 算法中目标值的计算步骤拆分为两步，其余不变。

（1）通过评估网络获得值函数最大的动作 a:

$$a^{\max}(\phi_{j+1}; \theta) = \max_a Q(\phi_{j+1}, a; \theta)$$

算法 2.3　深度强化学习求解的双网络 DQN 算法

初始化容量大小为 N 的经验池 D；
初始化评估网络 Q，随机生成权重 θ；
初始化目标网络 $\hat{Q}$，权重 $\theta^{-}=\theta$;
while *episode=1,2,...,M* **do**
    初始化状态 s_t，其状态向量为 $\phi_t=\phi(s_t)$；
    while *t=1,2,...,T* **do**
        以概率 ϵ 选取随机动作 a_t；
        否则选取 Q 值最大的动作 $a_t=\max_a Q^*(\phi(s_t),a;\theta)$；
        执行动作 a_t，获得奖赏值 r_t 和新状态 s_{t+1}，新状态向量 $\phi_{t+1}=\phi(s_{t+1})$；
        将四元组 $(\phi_t,a_t,r_t,\phi_{t+1})$ 存入经验池 D；
        从经验池 D 中采集 m 个样本 $(\phi_j,a_j,r_j,\phi_{j+1})$，j=1,2,...,m；
        计算当前样本的目标 Q 值:

$$y_j=\begin{cases} r_j & \phi_{j+1}\text{为终止状态} \\ r_j+\gamma\max_{a'}\hat{Q}(\phi_{j+1},a';\theta^{-}) & \phi_{j+1}\text{非终止状态} \end{cases}$$

        对损失函数 $(y_j-Q(\phi_j,a_j;\theta))^2$ 做梯度反向传播以更新评估网络参数 θ；
        每 C 步更新目标网络 $\hat{Q}$ 参数 $\theta^{-}=\theta$；
    end
end

（2）通过目标网络获得（1）中动作 a 的目标值:

$$y_j=r_j+\gamma\hat{Q}(\phi_{j+1},a^{\max}(\phi_{j+1};\theta);\theta^{-})$$

与上述算法不同，Dueling DQN[54]从网络结构上进行优化，在全连接层中将值函数分为两部分：状态值函数 $V(s)$ 和动作优势函数 $A(s,a)$。即：

$$Q(s,a,\theta,\alpha,\beta)=V^{\pi}(s,\theta,\alpha)+A^{\pi}(s,a,\theta,\beta)$$

其中 θ 是公共的网络参数，α 是状态值函数独有部分的网络参数，β 是动作优势函数独有部分的网络参数。由于 Dueling DQN 只涉及网络结构的改进，故其原则上可以和上面任意一个 DQN 算法结合，只需要用 Dueling DQN 的模型结构去替换上面任意一个 DQN 网络的模型结构。

以上介绍的都是在原始 DQN 的基础上进行单方面改进，而 Rainbow[55]则结合了对 DQN 算法的六个改进，包括 Double DQN、优先回放、Dueling DQN、多

步学习、分布式 RL 和噪声网络。优先回放即认为经验池中的数据应有不同的采样权重，而非均匀采样，TD Error 越大的样本，算法能从中获得越大的进步，因此该样本应以更大的概率被采样用以更新网络，从而提高了数据效率。多步学习即获得多步的即时奖励来计算 Q 值，因此训练前期的 Q 值可以得到更准确的估计，从而加快了训练速度；分布式 RL 使网络的输出为 Q 值的价值分布，而非单个的期望值，从而得到更多的信息，获得更好、更稳定的结果；噪声网络为对网络参数添加噪声来增强模型的探索能力。Rainbow 将六种改进集成在一个智能体上，因此相对于传统 DQN 和只进行单方面改进的 DQN，Rainbow 的算法性能有巨大提升，在不同场景的适用性和鲁棒性也有巨大进步，是目前性能最好的算法之一。

第 3 章　人类决策全时有效下的人机混合决策方法

本章基于强化学习方法提出了一种人类决策全时有效下的人机混合决策方法。讨论了在人类决策均有效的约束下，如何将人的决策和机器决策相结合以获得更高质量的决策动作，从而实现更好的决策效果。具体涉及最小干预原则的引入、共享控制模式的应用、仲裁机制的设计等，实现了算法的仿真验证，为后续的优化设计方案提供基础性方法。

3.1　引言

在人机混合决策中，人类和智能机器以互补的能力共同完成实时控制任务，以获得比他们各自单独决策更好的性能[14,56]。以无人机着陆为例：人类能够灵活应对实时变化的因素，但难以实现多个维度上的精确控制；机器在处理重复性任务方面具有高精度和长耐力的优势，但难以应对不同的复杂情况。混合决策则结合了人的策略和机器的策略来解决问题。

本章假设人的决策始终有效，即人的决策行为促进任务的完成并有效地反映人类的真实意图[11,57-58]，机器智能用于弥补人在物理能力上的不足（比如提高操作的精确度、弥补人的反应时间等），以及用于减轻人的工作量（比如辅助驾驶系统）。在这一约束下，共享控制模式拥有最好的混合决策性能，人机混合决策方法也因此分为三步：对人类意图的推断（由于意图无法显式表达或来不及传达等原因，机器通常无法直接获得人的意图）、机器基于推断的意图做出决策、以及对机器的决策和人类决策之间的仲裁[8-9,13]。仲裁决定了如何在人和机器之间分配控制权以最大限度地整合人类智能和机器智能，如图3.1所示，比如人类和机器共同控制机械臂末端执行器的速度，仲裁决定了各自动作对最终执行动作的影响程度，对该策略的定义一直是人机混合决策的基本问题之一[7]。仲裁可被分为四种类型[7]：

- 分治：人类和智能机器分别完成不同的子任务，例如人类控制机械臂的方向，机器控制末端执行器的速度。
- 主从：人类和机器有各自的自主权，但当它们发生冲突时，人类保留最终的决策权。
- 师生：用智能机器来训练人类，主要包括康复机器人，系统不断尝试以减少机器协助的次数。
- 协作：人类和机器是平等的合作者，这也是本章的主要研究领域。

对人的决策和机器的决策做线性组合是一种常见的仲裁函数，其因简单的

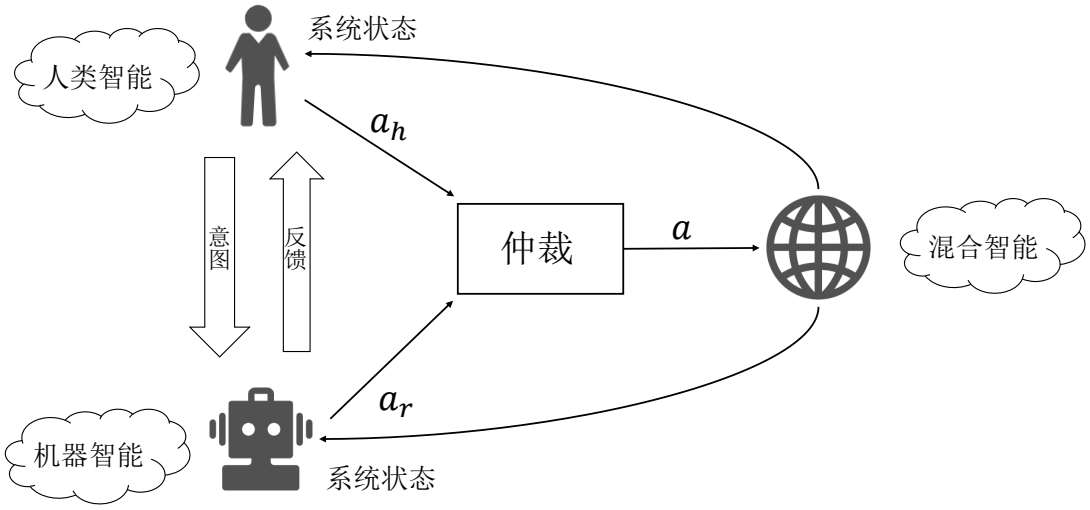


图 3.1　人机混合决策中仲裁示意图

形式和优良的效果已被广泛应用于许多人机混合决策系统中[9,14,50]。其中作为核心因素的仲裁权重大多由人预先调节和指定，而这一做法，即将加权的权重视为固定不变的超参数，无法使系统在不同的场景下始终保持最优[57]。为了应对这一挑战，一些其他方法基于意图推理的置信度来实时计算权重参数，当置信度较高时，机器被认为有较大的把握做出最优决策，人类因此会失去控制权[13]。这些方法最大限度地利用智能机器的性能，但对于需要人类做出最终决定的任务，尤其是在动态和不确定环境中的任务，这些方法可能会损伤性能。另一方面，机器的过度干预违背了人类对更多控制权的偏好，可能导致人类拒绝机器的协助而不是从中获得帮助，从而削弱系统性能和人类对于混合决策系统的满意度[15,50]。

综上所述，本章提出了一种人类决策全时有效下的人机混合决策方法，该方法基于强化学习，在保障系统性能的同时，最大限度地减少了最终决策与人的决策的偏差。该方法以最小程度的干预为人类提供最大程度的帮助，换句话说，当智能机器为了更好的性能进行干预时，他们应该尽可能少地修改人的决策，以增加人对帮助的接受度[25]。具体来说，本章使用长短时记忆网络来推断人类的意图并计算置信度，使用深度强化学习算法来估计所有决策动作（离散动作空间或对连续动作空间的采样）的控制效果。根据意图推理的置信度设置一个控制效果的自适应阈值，并在控制效果超过阈值的行为中选择最接近人类决策的行为作为最优行为执行，以平衡人类对控制权的需求和对性能的需求。

本章的研究内容主要可总结为三点：

- 研究能够在变化的环境中保持最优的行为选择方法；
- 研究遵循最小干预原则的混合决策仲裁方法，以提高系统性能；

- 研究基于深度强化学习的人机混合决策系统，使得无需已知被控系统的动态模型、人类的行为策略或关于人类能力的其他先验知识，便可获得更好的决策效果。这些先验知识在其他相关研究中可能是必要信息，但在实际任务中很难获得[11,16-17]。

本章的结构安排如下，第 3.2 节给出人机混合决策系统的问题描述，及如何用深度强化学习方法对其进行建模和求解，第 3.3 节介绍具体的混合决策仲裁方法的设计和实现，第 3.4 节给出仿真实验设计和结果分析，并讨论了控制权限的大小对人的满意度的影响，这些不仅是对本章人机混合决策算法的作证，也为后文的研究提供了部分基础。

3.2　问题建模

3.2.1　将人机混合决策建模为强化学习问题

本章研究的基础工作之一是用深度强化学习构建人机混合决策框架。之前的工作已将其建模为 POMDP[9]。假设人知道奖赏函数和任务目标 $g \in G$，机器知道奖赏函数 R、可能的目标集 G、状态转移函数 T 和人的行为策略 $\pi_h : S \times G \times A \to [0,1]$。机器在目标上的不确定性被建模为部分可观性，从而形成了 POMDP：状态空间和转移概率被任务目标所扩展 $\tilde{S} = S \times G, \tilde{T}((s_t, g)|s_t, g, a_t) = T(s_{t+1}|s_t, a_t)$，观测概率由人的行为策略给出 $O(s, a_h|a, g) = \pi_h(a_h|s, g)$，其中 a_h 为人的输入。这些方法假设目标空间 G、人的行为策略 π_h 和系统动力学模型 T 已知，并使用事后优化等近似方法求解 POMDP。

本章研究的方法使这些先验知识不再是必需品。状态空间和动作空间为决策任务的固有属性，故 S 和 A 为已知信息，本章使用深度强化学习计算的奖励值来衡量决策的优劣，并将奖励函数 R 分解为两部分：根据每个状态计算的即时奖励，和人在任务成功或失败时提供的最终奖励，其流程图如图3.2所示。

这种对奖励函数的分解使机器能够从密集的奖励信号中学习普遍有效的行为，比如保持系统正常运转，也能通过人的反馈适应不同的操作个体。在下式中，R_{general} 项计算机器已知的奖励，比如需要避障、需要尽快完成任务等；R_{feedback} 项计算机器未知，但可观测到的人的反馈。

$$R(s_t, a_t, s_{t+1}) = R_{\text{general}}(s_t, a_t, s_{t+1}) + R_{\text{feedback}}(s_t, a_t, s_{t+1}) \tag{3.1}$$

这一公式能够对已知的先验知识进行良好的融合和利用：

- 倘若已知人的行为策略 $\pi_h(a_h|s, g)$ 和目标空间 G，未知系统动力学模型 T，则可利用贝叶斯推断方法根据 π_h 推理出任务目标 g，将 R_{feedback} 参数化为 $R_{\text{feedback}}(s_t, a_t, s_{t+1}; g)$，当 $s_{t+1} = g$ 时 R_{feedback} 为高额奖励，否则为 0，这一

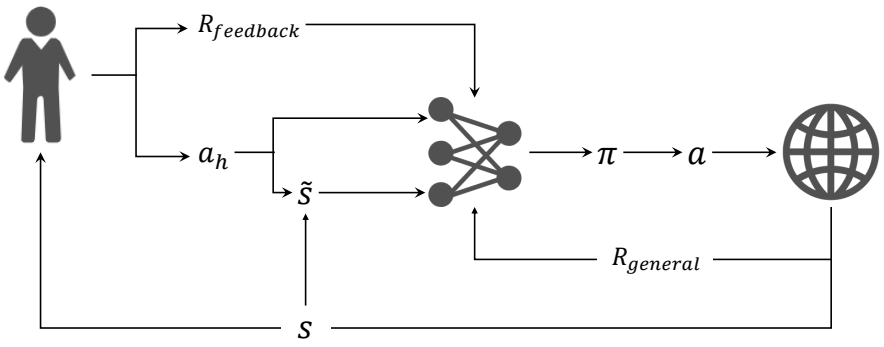


图 3.2　用深度强化学习构建人机混合决策的流程图

过程无需人的参与。

- 倘若已知目标空间 $\boldsymbol{G}$，未知人的行为策略 π_h 和系统动力学模型 $\boldsymbol{T}$，则可使用分类或回归方法从人的输入动作中预测目标，进而根据目标设置 $\boldsymbol{R}_{\text{feedback}}$。这种情况也是本章的研究假设。
- 倘若目标空间、人的行为策略和系统动力学模型均未知，则无法假定 $\boldsymbol{R}_{\text{feedback}}$ 的函数形式和奖赏规则，它通常是一个稀疏的终端奖励，由人给出，用于告诉机器任务是否成功。

基于深度强化学习的人机混合决策方法将系统状态、人的决策或任务目标（如果能够获得）作为输入，输出最优决策动作交由系统执行。因此机器观测到的状态应为系统观测状态 s_t 和来自人的信息 u_t，可公式化为：

$$\tilde{s}_t = \begin{bmatrix} s_t \\ u_t \end{bmatrix} \tag{3.2}$$

其中倘若可以推断出任务目标，则 u_t 为推断得到的目标 g，否则为人输入的决策动作 a_h。具体可分为三种情况：

- 倘若已知人的行为策略，则可用其计算目标 g 的最大后验概率。比如使用最大熵逆强化学习[59]进行贝叶斯目标推断。
- 倘若已知目标空间，则可使用监督预测方法来估计目标 g。比如使用一个单独的 LSTM 网络根据一系列的观测状态和人的决策来预测目标。
- 倘若人的行为策略和目标空间均未知，则无法使用显式的目标推断，而由机器收集人的决策输入，隐式解码人的意图并完成任务。对于机器而言，人是外部环境的一部分，人的决策是另一个观测来源，和其他来自传感器的信息无异。因为神经网络训练实现端到端映射，因此机器可以发现人的决策和环境状态之间的任务关系，而无需明确假设目标的存在。在这一情况

下，$u_t = a_h$。

3.2.2　将仲裁建模为线性函数

线性仲裁是目前最常用的仲裁形式，本节对线性仲裁进行建模和具体介绍，为后续做对比仿真实验打好基础。假设系统在时刻 t 的状态为 $s(t)$，人和智能机器输入到系统的决策动作分别为 $a_h(t)$ 和 $a_r(t)$，由其各自的策略 π_h 和 π_r 产生，即

$$\begin{aligned} a_h(t) &= \pi_h(s(t)) \\ a_r(t) &= \pi_r(s(t)) \end{aligned} \tag{3.3}$$

混合决策系统通过参数为 θ 的仲裁函数 $\beta(\cdot)$ 在人和机器的决策中进行仲裁，并产生决策命令 $a(t)$ 交由被控系统执行，如公式3.4所示，其中参数 θ 为多个时变参数的集合。

$$a(t) = \beta_\theta(a_h(t), a_r(t)) \tag{3.4}$$

基于该公式，寻找产生最佳人机交互和任务性能的参数集 θ 的问题可以被描述为一个最优控制问题。最优控制模型假设在任务执行过程中存在某种被优化的成本函数。一般来说，成本函数 J 可以写成公式3.5的形式。

$$J = h(s(t_f), t_f) + \int_{t_0}^{t_f} k(s(t), a(t), t)dt \tag{3.5}$$

其中 t_0 和 t_f 分别为初始时刻和终端时刻。公式第一项为终端成本，比如最终位置和目标位置之间的距离；第二项为内部成本，比如每个时刻花费的能源。本节并不试图确定成本函数的确切性质，影响成本函数的不可测量因素很多，确定成本函数的确切数学形式可能是一个棘手的问题，本节只做形式化表示，以对其结构进行说明。

在线性仲裁中，仲裁函数 $\beta(\cdot)$ 一般为公式3.6的形式，其中 α_θ 是参数为 θ 的函数。$\alpha_\theta = 0$ 为完全的人为控制，$\alpha_\theta = 1$ 为完全的自动化。

$$\beta_\theta(a_h(t), a_r(t)) = (1 - \alpha_\theta) \cdot a_h(t) + \alpha_\theta \cdot a_r(t) \tag{3.6}$$

大多数线性仲裁函数 α_θ 可以简化为图3.3所示的，被三个参数 $\{\theta_1, \theta_2, \theta_3\}$ 和变量 $c(t)$ 定义的函数形式。参数应满足 $\forall i, \theta_i \in [0, 1]$ 和 $\theta_1 \leqslant \theta_2$ 的约束。函数分为三段：

- $c(t) \leqslant \theta_1$ 时，仲裁权重 $\alpha_\theta = 0$，即系统由人单独控制；
- $\theta_1 \leqslant c(t) \leqslant \theta_2$ 时，仲裁权重 α_θ 和变量 $c(t)$ 呈线性关系，其斜率由参数 θ_2 和 θ_3 共同影响，斜率决定了机器参与控制的积极程度；

- $c(t) \geqslant \theta_2$ 时，仲裁权重保持不变，$\alpha_\theta = \theta_3$，即人和机器按固定比例共同控制被控系统。$\theta_3 = 0$ 时系统始终由人单独控制，而无论 $c(t)$ 的取值。

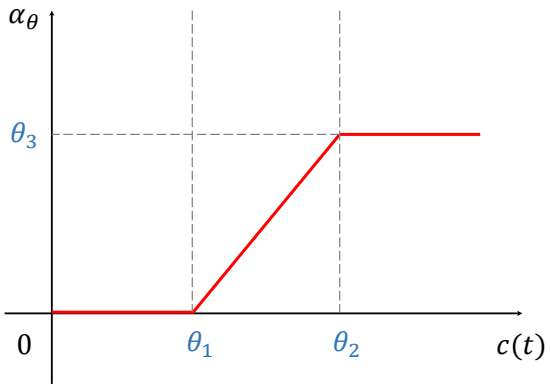


图 3.3　典型的线性仲裁函数形式

线性形式为仲裁函数的设计提供了良好的起点，但从长远来看，预定义的协助级别可能无法对变化的环境始终保持最优。现有研究大多出于优化性能的目的调整系统参数，最优性准则的直接选择是考虑与任务相关的性能度量，例如最小化所花费的时间、最小化成本和最大化任务的成功率等。这些方法无法捕获与人相关的指标，比如人的舒适度、独立性或对系统的满意度等，而这类指标在人机系统的设计中往往十分重要。这一缺陷促使本文提出新的方法，同时实现两类指标的改善和优化。

3.3　人类决策全时有效下的人机混合决策方法设计

基于第3.2节对人机混合决策问题的描述和建模，本节将具体介绍方法如何设计并实现。如图3.4所示，本章提出的方法的主要过程是人根据系统状态 s 给出决策动作 a_h，长短时记忆网络根据一系列的人的输入和系统运动轨迹推断目标 g 并计算置信度 c（第3.3.1节），深度 Q 网络根据推断目标 g 估计所有动作的控制效果 r（第3.3.2节），仲裁模块根据意图推理的置信度 c 和动作的控制效果 r 选择最优动作 a（第3.3.3节）。被控系统执行最优动作，并演变到下一状态。

3.3.1　意图推理

在人机混合决策系统中，人的输入是一系列按时间顺序排列的彼此高度相关的决策，系统根据人之前的输入呈现一个状态，人根据当前的当状态做出新的决策。递归神经网络（Recurrent Neural Network，RNN）是一种用于处理序列数据的神经网络。与其他要求数据独立同分布的神经网络相比，RNN 可以处理顺序变化的数据，例如，同一个词在不同的上下文中有不同的含义。长短时记忆网

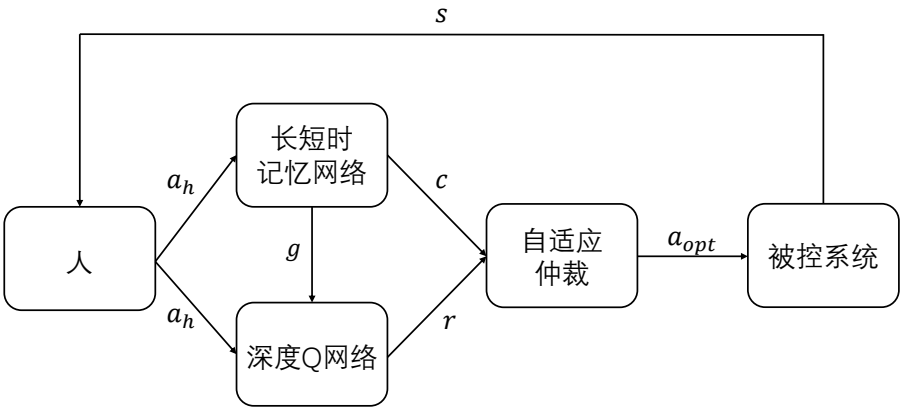


图 3.4　人类决策全时有效下的人机混合决策方法的流程示意图

络（Long Short-Term Memory，LSTM）是一种特殊的 RNN，可以解决长序列数据在训练过程中的梯度消失和梯度爆炸问题，在捕获长期时间依赖性方面既通用又有效，已被用于解决许多问题，如手写识别[60]、语言建模[61]和翻译[62]、音频分析[63]等，因此使用 LSTM 网络进行意图推理。

假设已知一组可能的目标集 G（例如无人机着陆任务中所有可能的着陆点），机器知道人类的目标存在于该目标集中，但不知道是具体哪一个。LSTM 网络将一系列人类决策和系统运动轨迹作为输入以预测目标 g_p。考虑到 LSTM 网络本身存在不确定性和计算误差，本章将已知目标集中最接近预测结果 g_p 的目标视为人的目标 g，并将目标集中所有目标与预测结果 g_p 之间的距离进行归一化，其结果即为目标集上的概率分布。

机器根据推理出的目标做出决策，推理越准确，做出的决策就越有效。同样，推理越不确定，做出错误选择的可能性就越大。因此，本章计算了意图推断的置信度，当置信度没有达到设定的阈值时，被控系统会直接执行人的决策动作，因为机器决策的错误概率太大而不应纳入考虑；置信度达到阈值时，仲裁模块在人的决策和机器的决策之间进行仲裁。意图推理的置信度是概率分布中的最大概率减去概率分布中的最小概率，这也是目前常被使用的一种计算方法[9]：

$$c = \max_{g'} p(g'|a_h) - \min_{g'} p(g'|a_h), c \in [0,1] \tag{3.7}$$

这里有两种极端情况：

- 目标集中有一个目标的概率为 1，其他目标的概率均为 0。这种情况即为 LSTM 网络准确推断出目标集中的某个目标，机器完全确定人的意图是什

么，因此意图推理的置信度为 1。

- 目标集中所有目标概率相等。即机器完全不确定人的意图是什么，因此意图推理的置信度为 0。

3.3.2　基于强化学习估计决策效果

强化学习算法的优化方向是使机器获得最优策略，能够在实现最终目标过程中通过多步动作产生的奖励总和达到最大。机器在每个时间步根据当前环境状态 s_t 和行为策略 π_t 做出决策，选择动作 a_t，即 $a_t = \pi_t(s_t)$，环境状态演化到 s_{t+1} 并给机器一个反馈奖励 r_t。这种四元组形式的转换会不断重复，直到系统达到终端状态或转换达到设定的最大次数，这个过程称为一幕。最优策略是在一幕结束时最大化累积奖励值 $R = \sum_{k=0}^{+\infty} \gamma^k r_{t+k}$，其中常数 $\gamma \in (0, 1]$ 是折扣系数，决定了未来奖励的重要性。系数设置为 0 即仅考虑当前时刻的奖励，使机器“短视”；系数接近 1 将使机器努力获得长期高奖励值，如果系数达到或超过 1，则奖励值可能发散。

强化学习算法通过求解贝尔曼方程得到最优策略：

$$Q^{\pi_{opt}}(s, a) = Q^*(s, a) = E_{s'}[r + \gamma max_{a'} Q^*(s', a')|(s, a)] \tag{3.8}$$

其中 $Q(s, a)$ 是在状态 s 下执行动作 a 后，在未来的有限步内可以获得的折扣奖励的最大总和，代表该动作可以为当前任务带来的好处，简称 Q 值。深度 Q 网络（DQN）是用神经网络近似估计 $Q(s, a)$[52]，而无需进行表格化处理，具体介绍详见 2.2.3 节。它将系统状态和人的决策动作作为输入，并将所有动作的 Q 值作为输出，本章使用 DQN 计算的 Q 值作为决策效果的估计，进而使用这种端到端映射来实现人机混合决策方法。

3.3.3　自适应仲裁

本方法遵循最小干预原则，即当智能机器为了更好的性能进行干预时，它应该尽可能少地修改人的决策[25]。如果被控系统执行的动作总是偏离人的决策动作，人类可能不再信任系统，导致其输入中包含的信息减少，进而降低意图推理和整个系统的性能。因此，本章在决策效果足够好的动作中，选择最接近人类决策的动作作为最优动作。意图推理的置信度决定了决策效果的自适应阈值。置信度越高，机器做出正确决策的概率越高，因此从越小的范围内选择最优动作；置信度越低，机器出错的概率越大，因此从越大的范围内选择最优动作。这一观点可公式化表示为：

$$A_{threshold} = \{a \in A | Q'(s, a') \geqslant c \times Q'(s, a_{max})\} \tag{3.9}$$

$$a_{opt} = \underset{a' \in A_{threshold}}{\arg\max} \; f(a', a_h) \tag{3.10}$$

其中 A 是离散动作空间或对连续动作空间的采样，$A_{threshold}$ 是根据阈值计算的用于选择最优动作的动作空间，即决策效果不劣于置信度乘以最好决策效果的动作集合。$Q'(s,a) = Q(s,a) - \min_{a' \in A} Q(s,a')$ 是为了防止 Q 值为负带来的计算错误，a_{max} 是 DQN 网络计算出的 Q 值最大的动作。函数 $f(a', a_h)$ 计算动作 a' 和人的决策 a_h 之间的相似度。特别地，人不输入动作将导致机器将价值最高的动作 a_{max} 传递给被控系统。当人类通过输入动作来引导任务时，机器变得顺从并遵循指导，而没有输入则意味着人类对当前情况感到满意，则由机器领导任务完成目标。整体算法如算法3.1所示。

算法 3.1　基于 DQN 的最小干预人机混合决策算法

初始化容量大小为 N 的经验池 $\mathcal{D}$；
初始化权重为随机权重 θ 的评估网络 Q；
初始化权重为 $\theta^- = \theta$ 的目标网络 $\hat{Q}$；
for *episode=1,2,...,M* **do**
    for *t=1,2,...,T* **do**
        获得环境状态 s_t 和人的输入 a_h；
        推理人类意图，并根据公式 Eq.(3.10) 获得动作 a；
        执行动作 $a_t = a$, 获得新状态 s_{t+1} 和奖励值 r_t；
        将四元组 (s_t, a_t, r_t, s_{t+1}) 存进经验池 $\mathcal{D}$；
        if s_{t+1} 是最终状态 **then**
            for *k=1 to K* **do**
                从经验池 D 批量采样四元组 (s_j, a_j, r_j, s_{j+1})；
                $a'_{j+1} = argmax_{a'} Q(s_{j+1}, a'; \theta)$；
                $y_j = r_j + \gamma \hat{Q}(s_{j+1}, a'_{j+1}; \theta^-)$；
                $\theta \leftarrow \theta - \eta \nabla_\theta \sum_j (y_j - Q(s_j, a_j; \theta))^2$；
            end
        end
        每 C 步复制评估网络的权重到目标网络 $\hat{Q} = Q$；
    end
end

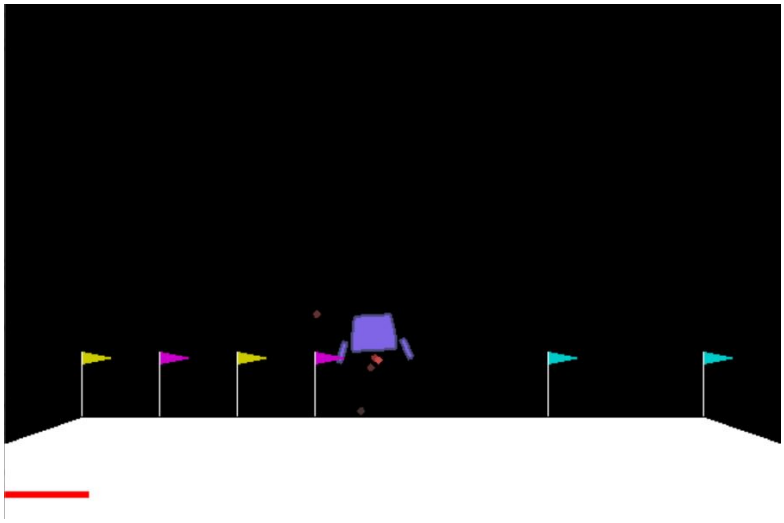


图 3.5　OpenAI Gym 登月着陆器场景示意图

3.4　仿真实验

本章在 OpenAI 发布的 Lunar Lander[①]（月球着陆器）环境中验证算法的有效性，环境如图3.5所示。地面上有三对颜色不同的旗帜，它们的坐标在每幕开始时随机生成。每对旗帜中间的区域是平的，而其他区域的高度是随机的。在月球着陆器的左侧和右侧各有一个推进器，中间有一个主发动机。人和机器共同控制这三个引擎，使月球着陆器在规定时间内平稳降落在目标旗帜的中间而不和地面发生碰撞，如此即为任务完成。如果着陆器坠落到地面、飞出边界、在目标旗帜以外的地面上保持静止，或未能在限定时间内顺利着陆到目标点，则任务失败。机器可以获得其当前位置坐标和三对旗帜的位置，但不知道哪一对是任务目标。人类通过旗帜的颜色选择着陆点并进行决策。机器根据人输入的决策动作推断目标，并辅助人控制着陆器接近着陆点。

系统状态向量包括着陆器在 x 轴和 y 轴的位置 $(x(t), y(t))$、在 x 轴和 y 轴的速度 $(\dot{x}(t), \dot{y}(t))$、与垂直方向的倾斜角度 $\theta(t)$ 和角速度 $\dot{\theta}(t)$、以及着陆器的左右两个支架是否接触地面 $(leg_{\text{left}}(t), leg_{\text{right}}(t))$。动作空间是三个发动机的开启和关闭，编码后表示为离散动作集合 $\{0,1,2,3,4,5\}$，从而实现同一时刻可以有多个维度的控制操作，数字和动作的对应关系如表3.1所示。0 为控制着陆器向左或向左下，即右发动机的开启；1 为没有操作，即三个发动机均关闭；2 为控制发动机向右或向右下，即左发动机的开启；3 为控制着陆器向左上，即左发动机和主发动机的开启；4 为控制着陆器向上，即主发动机的开启；5 为控制着陆器向右上，即左发动机和主发动机的开启。

奖励函数的设置如下。R_{general} 是惩罚速度、倾斜角度、与目标旗帜的距离，奖励着陆器的支架接触地面，触地即奖励 10 分，R_{feedback} 是惩罚任务失败和奖

[①]网址见：http://gym.openai.com/envs/LunarLander-v2/和 https://github.com/openai/gym

表 3.1　动作值和发动机开关的对应关系

动作值	主发动机	左发动机	右发动机
0	关闭	关闭	开启
1	关闭	关闭	关闭
2	关闭	开启	关闭
3	开启	关闭	开启
4	开启	关闭	关闭
5	开启	开启	关闭

励任务成功，分别在任务结束时给机器 -100 分和 $+100$ 分。公式3.10中的相似性函数 $f(a', a_h)$ 评估人的输入 a_h 和动作 a' 是否控制相同的引擎，或者它们是否控制着陆器朝着相同的方向移动，比如左发动机开启和右发动机关闭的相似度为 1，而左发动机开启和右发动机开启的相似度为 -1，即 f((left,on),(right,off))=1, f((left,on),(left,off))=-1。

为了评估方法的有效性，本章在信息学院内随机招募了十位平均年龄 25 岁的参与者在下述三种决策模式下进行操作。这种方式是人机领域中广泛使用的实验方式[25,64]，因此本文沿用这种模式。参与者对基于学习的方法有一定的认知，之前未接触过仿真实验环境但对类似的游戏场景有一定的基础，因此具有良好的训练和决策能力，并因招募的随机性具有不同的操作风格和偏好，从而使实验具有代表性。

- HIC：人单独进行决策，没有机器的辅助；
- LASC：对人的决策和机器决策做线性仲裁的混合决策方法；
- MISC：本章提出的遵循最小干预原则的自适应仲裁混合决策方法。

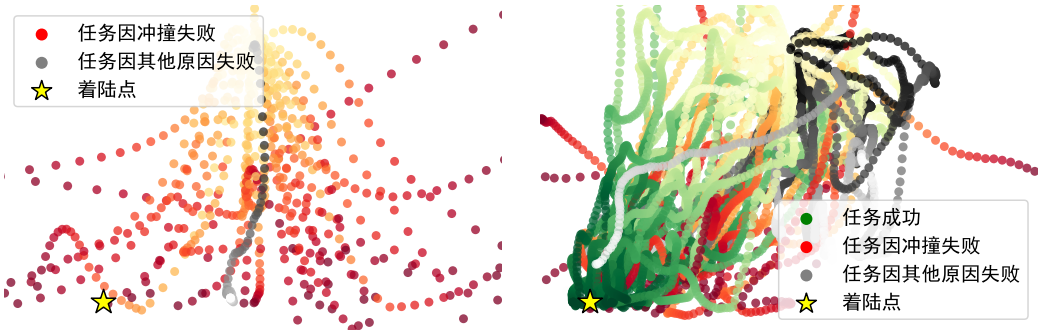
为此，本节首先进行线性仲裁混合决策的仿真实验，对比人单独控制和人机共享控制的性能差异，探究人和机器进行混合决策带来的性能变化；再实现遵循最小干预原则的仲裁函数自适应的混合决策方法，与线性仲裁做性能对比，以验证本章提出的方法的好坏。

3.4.1　人单独决策和人机混合决策的性能对比

本节介绍线性仲裁混合决策的仿真实验，对比人单独控制和人机共享控制的性能差异，探究人和机器进行混合决策带来的性能变化。对于线性仲裁混合决策而言，机器决策和人的决策的仲裁权重为核心要点。已有一些工作探究将权重作为超参数进行调节对性能的影响，本节在此不再赘述。本节在第 3.2.2 节的基础上实现线性仲裁，变量 $c(t)$ 设置为意图推理的置信度，参数设置为 $\theta_1 = 0.3, \theta_2 = 1, \theta_3 = 1$，即意图推理的置信度小于等于 0.3 时，被控系统由人单独控

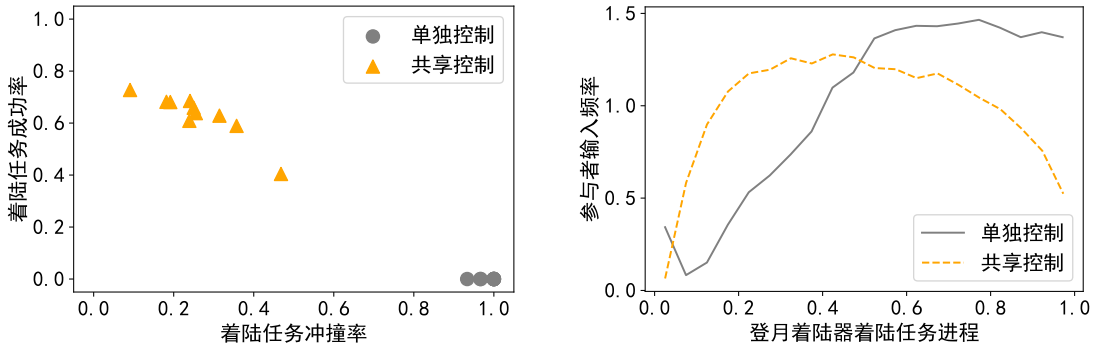
制，置信度大于 0.3 时，被控系统由人和机器共同控制，机器决策和人的决策的仲裁权重分别为 c 和 $1-c$，其中 c 为意图推理的置信度，使得仲裁公式能够良好适应实时变化的环境。意图推理的置信度越高，机器决策对最终动作的影响越大。倘若仲裁之后的动作不在动作空间中，则直接执行权重更高的决策动作。

每位参与者先提前训练 20 幕以熟悉相关操作和界面，然后单独操作 30 幕，再和机器共同控制着陆器 30 幕，比较其任务成功率和坠毁率。为了加快实验进程、减少参与者的操作数量，机器在没有人参与的情况下进行预训练，再根据有人参与的实验数据进行微调和优化。为了便于收集和分析数据，本节指定参与者的任务是使着陆器顺利降落地在黄色旗帜中间。



(a) 着陆器在参与者单独决策下的降落轨迹　(b) 着陆器在人机混合决策下的降落轨迹

图 3.6　着陆器在单独决策和混合决策下的降落轨迹



(a) 单独决策和混合决策下的任务成功率　(b) 登月着陆器任务进程中的参与者输入频率

图 3.7　登月着陆器在单独决策和混合决策下的任务成功率和参与者输入频率

图3.6展示了登月着陆器在参与者单独决策和线性仲裁混合决策下的典型降落轨迹，从中可以看出，在参与者单独决策下，着陆器很难无碰撞地精准降落在要求的位置，这主要由于人对高速运动物体实施精确操控能力的欠缺（如图3.6a）；而随着智能机器对动态精准调控能力的加入，参与者与机器的混合决策大大提升了着落成功的可能性（如图3.6b）。本节进一步在图3.7a中展示了 10 位参与者在 30 次单独决策和与机器混合决策进行操作的成功率，从中可以看出，大多数参与者在混合决策时都有接近或超过 60% 的成功率，而在单独决策时则近乎全

军覆没。十位玩家的成功率分布在 0.4 到 0.7 之间，即玩家之间存在一定的差异。这主要因为十位玩家为随机招募所得，他们在操作风格和偏好上不尽相同，而这些都是影响实验结果的因素，比如平时经常玩游戏的玩家比平时不玩游戏的玩家在实验操作上更有优势，也往往能够得到更高的成功率。而图3.7b中展示的参与者输入频率则表明了参与者在单独决策和混合决策时完全不同的操控模式：参与者在单独决策的后期达到输入频率的高峰，为的是避免高速运动的着陆器碰撞地面，而在混合决策时，则先做观察，进而密集操作着陆器对准要求的着陆位置，后期则几乎不需要额外操作，由智能机器完成安全降落。

3.4.2　三种决策模式的性能对比

本节实现人单独决策、线性仲裁混合决策和本章提出的遵循最小干预原则的自适应仲裁混合决策三种控制模式的性能对比。状态、动作、奖励函数等基础实验设置与上一节相同。倘若人机决策融合之后的动作不在动作空间中，则直接执行权重更高的动作。每个参与者提前操作 40 幕，以熟悉环境和智能机器。为了便于收集和分析数据，本节指定参与者的任务是使着陆器顺利降落地在黄色旗帜中间。

图3.8显示了十名参与者在三种决策模式下执行任务的成功率和坠毁率。可以看到智能机器的加入大大提高了任务的成功率，降低了坠毁率。当着陆器下降时，人类很难同时控制三个维度的发动机以保持平稳，从而导致着陆器坠落地面——十名参与者的坠毁率均大于 0.8。智能机器可以有效控制着陆器缓慢降落，让参与者将更多的注意力放在控制着陆器的方向上，从而大大提高了成功率。LASC 和 MISC 模式的成功率几乎相同，ANOVA（方差分析，用于两个及两个以上样本均数差别的显著性检验。F 值越大，p 值越小，组间差别越大。一般认为 $p \leqslant 0.05$ 时，样本之间差别显著，具有统计学上的意义）结果为 $F = 0.05, p = 0.8265$，表明二者并无显著区别。图3.9显示，MISC 模式完成任务的平均步数相比 LASC 而言更少，即在 MISC 模式下任务完成得更快。方差分析的结果为 $F = 8.65, p = 0.0087$，表明两者存在显著差异，其差异具有统计学意义。如图3.10所示，MISC 模式下每幕中人的输入的数量（每幕的按键次数）大多在 100 到 500 之间，而 LASC 模式下的人的输入数量大多在 300 到 800 之间。本章认为，造成这种差异的原因是，最小程度的干预可以防止人类使用额外的输入来抵抗机器的协助，并且人类不需要多次重复动作来确保其命令得到准确执行，从而减少了不必要的工作量。

为了评估参与者对机器自主性的接受程度以及对本章设计的人机混合决策系统的满意度，本节设计了相关问卷请参与者对系统性能进行评分，如表3.2所示。参与者被要求给出对表中每一项表述的同意程度，并被提前告知应只考虑操作过程的具体感受，忽略其他与混合决策系统无关的因素，如个人对实验设计的

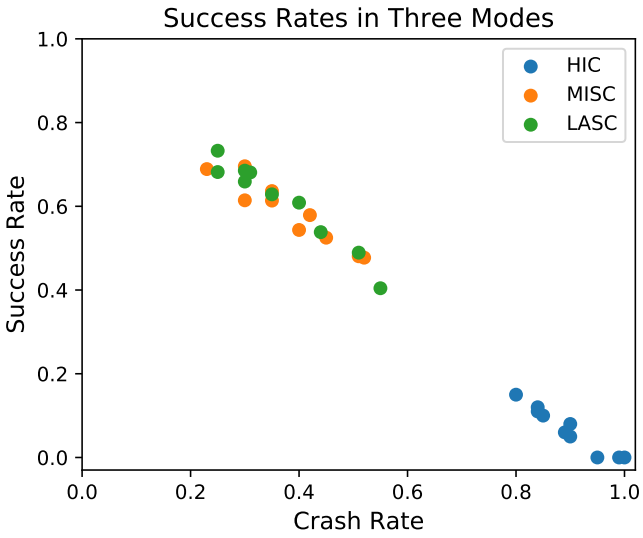


图 3.8　十位参与者在三种决策模式下执行任务的成功率和坠毁率

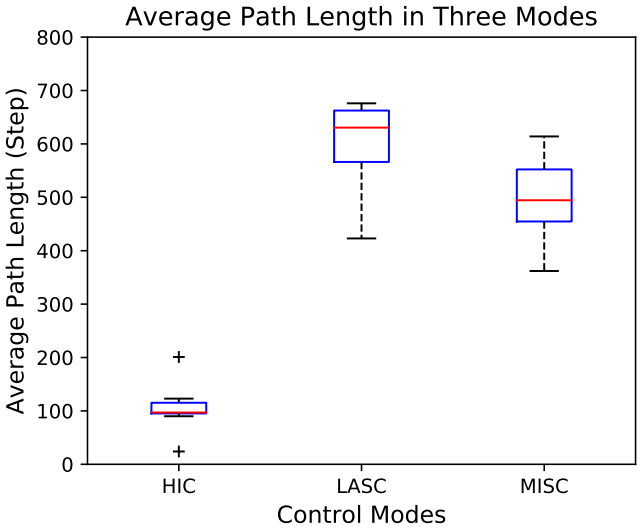


图 3.9　十位参与者在三种决策模式下执行任务的每幕的平均步数

喜好等。10 表示强烈同意，0 表示强烈不同意，具体分值由参与者自由心证给出。根据结果显示，参与者大多认为机器的协助很有用，可以帮助他们更好地完成任务（表格中第一个问题和第三个问题）。MISC 模式的得分略高于 LASC，本节认为这得益于最小干预原则的使用。然而，MISC 和 LASC 在机器是否做了参与者想做的事情（表格中第二个问题）上的得分完全相同。本文认为，其原因在于机器和人类参与者有着相同的最终目标，但对于每一步的具体实施有着不同的规划。当智能机器提供的帮助与参与者设想的计划不同时，人可能会使用更多的输入来与机器的自主性进行抗衡（表格中第四个问题）。人类希望在智能机器的帮助下操作系统，同时能够引导任务且不受这种帮助的干扰。如第五个问题所示，本章提出的 MISC 控制模式在人类对系统的满意度方面得到了显著更高的得分。

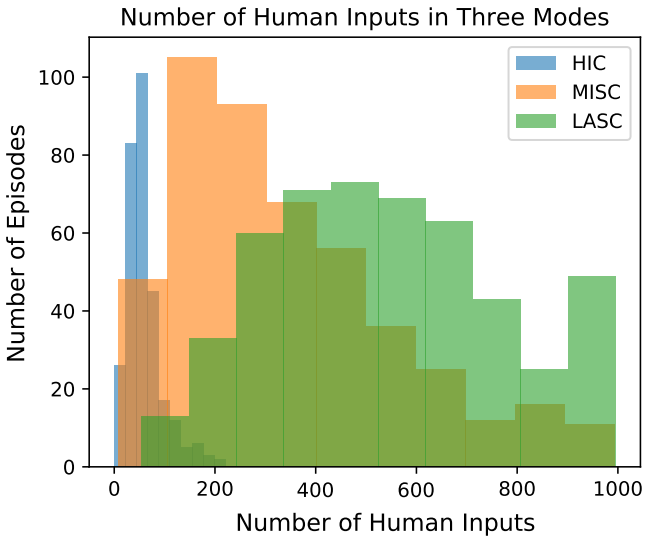


图 3.10　十位参与者在三种决策模式下执行任务的每幕的输入次数（每幕的按键次数）

表 3.2　参与者对调查问题的回答（对表述的同意程度）

调查表述	LASC	MISC
智能机器的帮助对完成任务很有用	7.4	7.9
智能机器做了我想做的事.	8.2	8.2
在机器的帮助下，我更好地完成任务	9.2	9.3
机器的帮助使我感到困扰	3.6	1.2
我对这个决策系统感到满意	7.8	8.6

3.5　本章小结

本章介绍了基于深度强化学习算法的人机混合决策系统，并提出了一种在有效人类决策约束下的人机混合决策方法。该方法基于深度强化学习算法，在保证系统性能的同时，最大限度地减少最终决策和人类决策之间的偏差，以最小程度的干预为人类提供最大程度的帮助。本章根据意图推理的置信度设置一个自适应阈值，并在决策效果超过阈值的行为中选择最接近人类决策的行为作为最优行为交由被控系统执行，以平衡人类对控制权的需求和对性能的需求。实验结果表明，该方法在保持较高的人类满意度的同时，以较少的人力投入实现了较高的任务成功率和较短的任务完成时间。这一方法为后续研究无效人类决策约束下的人机混合决策方法奠定了基础。

第 4 章　人类决策非全时有效下的人机混合决策方法

如第三章所述，在人机混合决策中，人和智能机器以互补的能力共同完成实时控制任务，以实现双方单独控制都无法达到的性能。现有的许多人机混合决策方法倾向于假设人的行为始终“有效”，即这些行为促进了任务的完成，且有效地反映了人类的真实意图，从而使智能机器能够根据人的行为推测任务目标，再以不同的仲裁函数协助人类完成该目标。然而，在现实中，由于疲劳、分心等多种原因，人的行为往往会在一定程度上“无效”，不满足这些方法的基本假设，导致方法失效，进而导致任务失败。

本章基于第三章的结论，对其方法做进一步的改进，基于深度强化学习算法提出了一种在人类决策可能无效的约束下的人机混合决策方法。该方法在共享控制和介入控制两种模式间灵活切换，使系统在人类决策长期无效的情况下仍能完成正确的目标。具体涉及用深度强化学习训练从系统状态和人类决策行为到行为价值的端到端映射，以显式判断人类决策是否有效；利用机器介入系统以防止无效的人的决策干扰任务的进行；判断当前适用共享控制还是单方面的介入和单独控制等，最终实现提升人机系统的决策质量、改善系统性能。

4.1　引言

许多控制任务的目标是由人单方面决定的，人机混合决策系统被设计以帮助人完成该任务。在这种情况下，成功的混合决策策略通常依赖于两个基本组成部分：对人类意图的推断（机器通常无法直接得知人的意图），以及对机器选择的动作和人类输入的动作的仲裁[9,13,18,21]。机器推理人的意图是完成仲裁的前提和基础，其质量直接影响到仲裁的成败。意图推理通常通过分析观察到的人类决策来完成，具体细节详见第三章，其中一个关键假设便是这些观察到的人类决策是“有效的”，即这些行为促进任务的完成并有效地反映人类的真实意图[11,57-58]。然而，在现实中，由于疲劳、分心等多种原因，人类决策往往在一定程度上“无效”，且会持续一段时间（比如疲劳驾驶）。随着人的决策持续一段时间“无效”，对人的意图的推断也将变得无效，导致机器辅助的失效，进而导致整个系统失效。

为了应对人的无效决策，首先也是关键的步骤是确定一系列的人类决策是否无效。这是一个极具挑战性的任务，其难点在于信息的不完全可知，即必须由机器用有限的信息来完成。一种可能的方法是利用人类生理状态等共享控制系统之外的信息，认为人在生理状态异常时其物理行为无效，这可以通过面部识

别[65]、监测神经信号[66]或心跳频率[67]等方法进行测量和推理。这类方法已被应用于一些领域，比如在驾驶时通过摄像机进行面部识别，根据驾驶员眼睑的开合判断其是否疲劳驾驶。但这类方法不能处理人在正常生理条件下，由于认知限制和环境限制（如时间限制和不完全信息）做出的无效行为。例如，在紧急情况下，由于有限的处理时间和精神压力，人的决策行为可能损伤系统性能，甚至导致危险。

第二个挑战是如何区分无效的人类决策和人的意图改变。推理组件基于一系列历史数据来推断人的意图，对人偶尔的无效行为具有鲁棒性，但也导致意图变化无法被立即检测到。对于以先前的意图为任务目标的系统来说，意图改变后的人的决策可能是无效的。对这两者之间的识别和区分直接关系到任务的成功与否。

综上所述，本章提出了一种基于强化学习算法的，人类决策非全时有效下的人机混合决策方法。该方法灵活切换共享控制和介入控制两种控制模式，使系统在人类决策长期无效的情况下仍能完成正确的目标。具体来说，本方法利用长短期记忆网络推断人类意图，然后利用深度强化学习训练从系统状态和人类决策到决策价值的端到端映射，以判断人类决策是否无效。强化学习算法计算的累积奖励值衡量了该决策行为可以给当前任务带来的利益的多少。本方法默认人类和机器都在朝着更高的奖励值努力，因此在人的决策的奖励值下降一定程度后，该决策可以被判定为无效。当人类决策连续多次无效时，系统将由机器单独控制，完成由之前的有效决策推断出的任务目标，防止人的无效决策给任务造成危害。

本章的研究工作主要可总结为四点：

- 研究新的人机混合决策方法，能够灵活切换共享控制和介入控制两种模式，使得在人类决策长期无效的情况下，系统也能继续完成正确的目标；
- 研究无需额外信息就能判断人类决策有效性的方法；
- 研究能够区分无效的人类决策和人的意图改变的方法；
- 研究新的仲裁方法，考虑了人类的无效决策和智能机器的不确定性。

本章的结构安排如下，第4.2节介绍无效人类决策约束下的人机混合决策方法如何设计，具体涉及构建和实现意图推理网络，判断人的决策的有效性、考虑人的无效决策后的动作选择等；第4.3节给出实验设计和结果分析；第4.4节根据实验结果对本章提出的人机混合决策方法做进一步的讨论；最后在第4.5节总结本章的工作。

4.2　人类决策非全时有效下的人机混合决策方法设计

由于环境的部分可观测性和系统参数的不精确性，许多控制任务对机器来说存在困难，而由于人的有限理性[34]和物理限制（如缺乏多维控制能力）等原因，人类也很难单独完成。一些人机系统通过结合人的决策和机器决策来解决此类问题，并将这种模式称为混合决策[25,35]。人机混合决策的核心难点之一是机器通常不知道人类的意图，因此许多人机系统通过先推断、再辅助两个步骤完成混合决策。在这些系统中，机器选择行为时会尽可能地遵循人类的指令，例如，Xu et al.[58]构建的系统让机器评估其性能和人的期望之间的一致性，然后更新自身行为以更好地和人保持一致。

在这样的系统中，当人类犯了错误，可能会导致任务失败或产生安全事故时，系统没有进行预防或阻止的措施。控制权的切换可以在一定程度上解决这一问题。Vasudevan et al.[20]预测了驾驶车辆的潜在行为，并根据摄像头捕捉到的驾驶员的姿势衡量其安全性，然后确定半自主驾驶系统应在何时进行干预。Fu et al.[16]考虑了人类操作员的认知和生理状态的演变，自主机器通过最小化人的工作量和优化系统性能之间的权衡，来确定何时掌管控制权。Broad et al.[68]直接从数据中学习系统动力学和人与系统交互的信息，然后根据学习到的模型调整控制权限。这些方法是在系统处于不利地位后进行纠正，而非在不当行为执行之前进行预防或阻止，这可能导致严重的后果。

本章关注人的决策可能对完成任务无用甚至有害的情况，而智能机器可以更准确地感知环境，并根据环境状态判断人类决策的有效性。此方法的目的是在人的决策长期无效的情况下，确保系统完成正确的目标。一种可行的解决方案是控制权自适应：当人的决策有效时采取人机混合决策，系统在人输入的动作和机器选择的动作之间进行仲裁；而当人的决策无效时，机器独自控制系统完成由先前的有效决策推断出的目标。其流程图如图4.1所示，图中 a_h 为人输入的动作，g 为意图推理模块根据人的决策实时推理出的目标，$arbitrate(\cdot)$ 为仲裁函数，$g_{effective}$ 为根据有效决策推断出的目标。本节将详细描述如何判断人的决策的有效性，如何根据意图推断的结果确定任务目标，如何在共享控制中进行仲裁，以及如何在共享控制和个体控制之间进行离散切换。

如图4.2所示，本章所提方法的关键部件包括四个部分：智能机器推断人的意图并计算意图推断的置信度（第4.2.1节）；动作选择模块计算共享控制下人-机仲裁的自适应权重，确定执行共享控制时的动作（第4.2.2节）；人的决策有效性判断模块用于判断人的决策是否无效，是否需要由人和机器的共享控制切换到机器的单独控制（第4.2.3节）；仲裁模块根据上述所有信息决定受控系统最终执行的动作（第4.2.4节）。其中蓝色区域，即人的决策的有效性判断模块和仲裁模

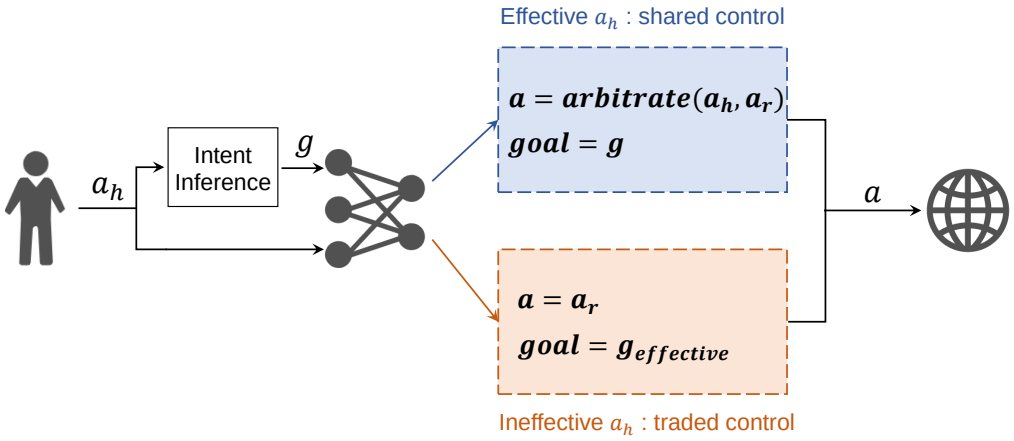


图 4.1　人机混合决策系统中控制权自适应的流程示意图

块是本章方法的主要贡献。

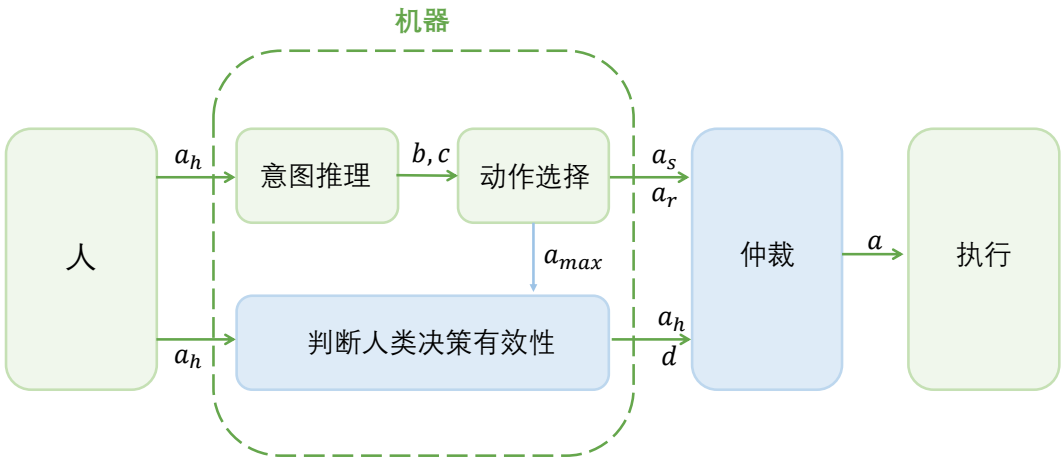


图 4.2　人类决策非全时有效下的人机混合决策方法流程示意图

4.2.1　意图推理

在人机混合决策系统中，人的输入是按时间顺序排列的、彼此高度相关的序列决策，即系统根据人上一步的决策行为呈现一定的状态，人根据系统当前的状态做出新的决策行为。基于这一原理上的相通性，本章使用的意图推理方法是第三章提出并详细介绍的方法。具体来说，假设已知一组可能目标集 $G = \{g_1, g_2, ..., g_N\}$，人的真实目标 $g(t)$ 存在其中，但在动态演化过程中，机器无法直接获得人类伙伴实时想要实现的具体目标，意图推理模块即用于推断相关信息。

基于贝叶斯规则，以 t 时刻及其之前的人的序列决策作为观测值，t 时刻对目标的信念可表示为：

$$\begin{aligned} b(g(t)) &= P(g(t)|a_h(0:t)) = P(g(t)|a_h(0:t-1), a_h(t)) \\ &\propto P(g(t), a_h(0:t-1), a_h(t)) \\ &\propto P(a_h(t)|g(t), a_h(0:t-1))P(g(t), a_h(0:t-1)) \\ &\propto P(a_h(t)|g(t), a_h(0:t-1))P(g(t)|a_h(0:t-1)) \end{aligned} \tag{4.1}$$

假设已知 t 时刻的目标，t 时刻的人类决策和历史时刻的人类决策之间条件独立[69]，即：

$$P(a_h(t)|g(t), a_h(0:t-1)) = P(a_h(t)|g(t)) \tag{4.2}$$

假设已知历史时刻的目标，则 t 时刻的推测目标与历史时刻的人类决策条件独立[69]，即：

$$P(g(t)|g(t-1), a_h(0:t-1)) = P(g(t)|g(t-1)) \tag{4.3}$$

根据公式4.2和4.3，可将公式4.1写为：

$$\begin{aligned} b(g(t)) &\propto P(a_h(t)|g(t))P(g(t)|a_h(0:t-1)) \\ &\propto P(a_h(t)|g(t)) \sum_{g(t-1)\in G} P(g(t), g(t-1)|a_h(0:t-1)) \\ &\propto P(a_h(t)|g(t)) \sum_{g(t-1)\in G} P(g(t)|g(t-1))P(g(t-1)|a_h(0:t-1)) \end{aligned} \tag{4.4}$$

将 $b(g(t-1))$ 用式4.1的定义方式进行表达，最终得到递归表达形式：

$$b(g(t)) \propto P(a_h(t)|g(t)) \sum_{g(t-1)\in G} P(g(t)|g(t-1))b(g(t-1)) \tag{4.5}$$

基于上述公式得到目标的后验概率分布，再基于最大后验原则即可得到当前时刻的目标。本节使用 LSTM 网络计算上述公式，即神经网络的输入为一系列的运动轨迹和人输入的动作，网络输出为推测的目标 g，根据梯度下降和反向传播方法更新网络参数。将预测结果 g 和目标集中所有目标之间的距离归一化，所得结果即为目标集上的概率分布 b，每个数值即为每个目标是当前人类目标的概率。

有些人机混合决策方法中机器将目标集中概率最大者作为人的目标，再根据该目标选择行为，这也是第三章的做法。基于确定性目标的缺点在于意图推理存在不确定性，在复杂的情况下可能推断出错误目标，进而选择错误的动作。因此本节基于目标集的概率分布 b 选择动作，保留所有目标的可能性。目标推理越准确，机器选择的行为越有效；目标推理的不确定性越大，机器选择行为的错误

概率也越大。因此计算目标推理的不确定性，不确定性大到一定程度便直接执行人的决策命令，否则在人的决策行为和机器选择的行为中进行仲裁。

置信度的计算方法与第三章相同，本节不再赘述。具体公式如下：

$$c = \max_{g'} p(g'|a_h) - \min_{g'} p(g'|a_h), c \in [0,1] \tag{4.6}$$

4.2.2　动作选择

动作选择模块计算在执行共享控制时要采取的动作。本节使用深度 Q 网络 (DQN) 来实现方法。DQN 的输入是环境状态和人的决策，输出是当前环境状态下所有行为的预期未来累积奖励。累积奖励 $Q(s,a)$ 是在状态 s 执行动作 a 后，在未来有限的步骤中所能获得的折扣奖励的期望和，用于评估动作的价值。机器根据这些累积奖励选择动作。在一些混合决策方法中，机器选择累积奖励值最大的动作作为最优动作。但是，本章认为，当人类的决策有效时，智能机器应该尽可能少地修改人类的输入，以增加其对辅助的接受度。如果系统总是执行偏离人类决策的行为，那么人类可能不再信任系统，因为他的命令没有得到准确地执行。因此，本方法采用足够好的、与人类决策最相似的动作作为共享控制下执行的动作。

选取最优行为的范围大小由推理信心决定，推理信心越大，机器对于任务目标越确定，因此选择的范围越小；推理信心越小，机器越有可能出错，因此在较大的范围内选择和人类决策相近的行为。本方法采用行为的累计奖赏值计算行为之间的距离，如公式4.7所示。

$$a_s = \underset{a \in A : Q'(s,a) \geqslant c \times Q'_{max}(s,a_{max})}{\arg\max} f(a, a_h) \tag{4.7}$$

其中 A 是所有可能的动作集合，$Q'(s,a) = Q(s,a) - \min_{a' \in A} Q(s,a')$ 为动作的累积奖励减去所有动作中累积奖励的最小值，以防止负 Q 值造成的误差。a_{max} 为 DQN 网络计算出的当前环境状态下的价值最高的动作。$f(a,a_h)$ 为计算行为 a 和用户控制行为 a_h 之间的相似度。比如当 $confidence = 0.8$ 时，从满足 $Q' \geqslant 0.8Q'_{max}$ 中选择和 a_h 最相似的行为；当 $confidence = 0.4$ 时，从满足 $Q' \geqslant 0.4Q'_{max}$ 中选择和 a_h 最相似的行为。

特别是，人没有输入将导致系统直接执行机器计算出的价值最高的动作。当人类通过输入动作进行干预来引导任务时，机器就会变得顺从并服从指导。没有输入意味着人类对当前的情况感到满意，机器将试图领导任务。

4.2.3　判断人的决策的有效性

累积奖励值代表了该动作在当前任务中的价值。本章默认人类和机器都在朝着更大的价值努力，所以当人的决策的奖励值足够低时，决策就会被判定是无效的。

本方法使用动作的累积奖励之间的差值作为动作之间的距离，如式4.8所示。

$$d(a, a_{max}) = \frac{Q'(s,a) - Q'(s,a_{max})}{Q'(s,a_{max})} \tag{4.8}$$

当人的决策动作和 DQN 计算出的价值最高的行动之间的距离 $d(a_h, a_{max})$ 连续多次足够大时，本方法认为人当前无法实施有效控制，故由机器单独控制系统。当机器单独控制时，机器以人进行有效控制最后一刻计算出的概率分布作为任务目标，不管人当前决策是什么，直接将 a_{max} 传递给被控系统，如式4.9所示。

$$a_r = \arg\max_{a \in A} Q(s,a) = a_{max} \tag{4.9}$$

同时，人在不断地控制，网络在不断地计算动作距离。当连续几次距离 $d(a_h, a_{max})$ 足够小时，人回归到理性，能够做出有效的响应，系统回归到共享控制。

值得特别注意的是，机器可能会将人的目标改变的行为错误地判断为无效的决策行为。LSTM 网络根据一系列轨迹和动作计算目标集上的概率分布。因此，目标变化可以被神经网络逐步识别出来，由人与机器的共享控制来完成。但当系统接近推理出的目标且推理置信度较高时，人改变目标的动作可能产生较大的动作距离，进而导致人失去控制权，而由机器引导系统完成之前的目标，即任务失败。因此，当机器单独控制系统时，本方法重新收集轨迹和人的决策来重新推断人的意图。如果机器连续多次重复推理出同一个目标，则认为是目标发生了改变，由机器和人共同控制系统完成新目标。

4.2.4　人机混合决策的仲裁方法

综上所述，被控系统执行的动作有三种可能，如式4.10所示。当意图推理的置信度足够低时，由人单独控制系统，系统将执行 a_h。当人的决策动作和最优动作之间的距离连续高于阈值时，机器单独控制系统，系统将执行 a_r。在其他情况下，系统由人和机器共同控制。总体算法如算法4.1所示。

$$a=\begin{cases}a_h, & c \quad \text{is low enough}\\ a_r, & d \quad \text{is high enough}\\ a_s, & \text{otherwise}\end{cases} \tag{4.10}$$

算法 4.1　基于 DQN 的无效人类决策约束下的人机混合决策算法

初始化容量大小 N 为的经验池 $\mathcal{D}$;
初始化权重为随机权重 θ 的评估网络 Q;
初始化权重为 $\theta^{-}=\theta$ 的目标网络 $\hat{Q}$;
for *episode=1,2,...,M* **do**
  for *t=1,2,...,T* **do**
    获得环境状态 s_t 和人的输入 a_h;
    推理人类意图，并根据公式 Eq.(4.7) 获得动作 a_s;
    根据公式 Eq.(3.10) 仲裁得到动作 $a_t=a$;
    执行动作 a_t, 得到新状态 s_{t+1} 和奖励值 r_t;
    将四元组 (s_t,a_t,r_t,s_{t+1}) 存储进经验池 $\mathcal{D}$;
    if s_{t+1} 是最终状态 **then**
      for *k=1 to K* **do**
        从经验池 D 批量采样 (s_j,a_j,r_j,s_{j+1});
        $a'_{j+1}=argmax_{a'}Q(s_{j+1},a';\theta)$;
        $y_j=r_j+\gamma\hat{Q}(s_{j+1},a'_{j+1};\theta^{-})$;
        $\theta\leftarrow\theta-\eta\nabla_\theta\sum_j(y_j-Q(s_j,a_j;\theta))^2$;
      end
    end
    每 C 步复制评估网络的权重到目标网络 $\hat{Q}=Q$;
  end
end

4.3　仿真实验

本章使用 OpenAI Gym 的 LunarLander[①]场景进行实验，如图4.3所示。降落地面上共有三对旗子，其位置坐标在每次任务开始时随机生成。每对旗子中间的区域是平坦的，其他区域的高度为随机生成。着陆器左右两边各有一个推进器，中间有一个主引擎。人和机器共同控制这三个发动机，使着陆器无碰撞地降落到

[①]网址见：http://gym.openai.com/envs/LunarLander-v2/和 https://github.com/openai/gym

目标对旗子中间，则任务完成。若着陆器冲撞到地面、飞出边界、在目标旗子外的地面保持静止或 1000 步以内未能成功，即时间耗尽，则任务失败。机器知道着陆器当前位置和三对旗子的位置，但不知道任务目标是哪一个。人在操作时通过旗子颜色选择着陆点并采取控制行为，机器根据人的控制行为推理任务目标并控制着陆器向着陆点靠近。

图 4.3　OpenAI Gym 登月着陆器场景示意图

Lunar Lander 模拟了真实的着陆环境，即着陆器越靠近地面，降落速度越快，也越难以控制。因此机器需要大量数据学习对三个发动机的控制，而这些大量的训练过程是人类参与者无法承受的负担，因此智能机器将在没有人的参与的情况下进行单独训练，以完成对通用规则的学习。

4.3.1　实验设置

该任务的动作空间为六个离散动作，即三个发动机的开启或关闭。编码后表示为离散动作集合 $\{0,1,2,3,4,5\}$，从而实现同一时刻可以有多个维度的控制操作，数字和动作的对应关系如表4.1所示。0 为控制着陆器向左或向左下，即右发动机的开启；1 为没有操作，即三个发动机均关闭；2 为控制发动机向右或向右下，即左发动机的开启；3 为控制着陆器向左上，即左发动机和主发动机的开启；4 为控制着陆器向上，即主发动机的开启；5 为控制着陆器向右上，即左发动机和主发动机的开启。

状态空间为 11 维向量，包括着陆器当前的位置、速度、角度、角速度、支架是否接触地面、以及三对旗子的坐标。奖赏函数为惩罚速度、角度、和推理目标之间的距离，即速度越快、倾斜度越大、距离推理出的目标越远则惩罚越大，旨在于训练着陆器稳定缓慢地向目标点移动。着陆器的支架接触地面给予小额奖励，平稳降落在目标点后给予大额奖励，冲撞到地面或飞出边界则给予大额惩罚。对各项特征的惩罚和奖励系数，以及终端状态的大额奖励和大额惩罚的具体

表 4.1　动作值和发动机开关的对应关系

动作值	主发动机	左发动机	右发动机
0	关闭	关闭	开启
1	关闭	关闭	关闭
2	关闭	开启	关闭
3	开启	关闭	开启
4	开启	关闭	关闭
5	开启	开启	关闭

数值需进一步实验确定，以便得到性能最优的智能机器。

实验采用有两个隐藏层，每层 32 个神经元的 LSTM 网络进行意图推理。采用含有两个隐藏层，每个隐藏层有 64 个节点的多层感知机实现 DQN 算法。动作之间的相似度函数 $f(a, a_h)$ 为判断两个动作是否控制同一个发动机或是否控制着陆器向同一方向移动。比如同时控制左发动机的相似度为 1，即 f((left,on),(right,off))=1；控制左推进器关闭和控制右推进器开启的相似度为 1，即 f((left,on),(left,off))=-1。

本节用随机操作模拟人的无效决策，若人输入的行为连续十次中有七次行为满足 $d(a_h, a_{max}) \geqslant 0.7$，即累积奖励值小于最大奖励值的 30%，人的决策被判定为无效，由机器接管任务并重新进行意图推理。相应地，如果人连续 10 次输入的动作中有 7 次的累计奖励超过最大值的 70%，人被认为恢复正常，着陆器由人与机器共同控制。重新推理意图时，倘若连续七次重新推理的目标中有五次目标相同，则认为实则为人的目标切换到该目标，人和机器将共同控制着陆器接近新目标。如果推理置信度小于 0.3，着陆器将直接执行人的输入，因为机器的动作太不确定，不应予以考虑。

本节招募了十位玩家，五位男生五位女生，平均年龄 25 岁。每个玩家被提前告知游戏规则并单独操作 20 次以熟悉操作和环境，再和训练后的机器共同控制 20 次以相互适应和优化。每个玩家要完成在降落过程中改变和不改变目标的两个实验。并且为了方便收集和分析数据，本节为玩家指定了目标旗子颜色。第一个实验中玩家不能改变目标，始终控制着陆器降落到黄色旗子中间。第二个实验中玩家的目标由蓝色旗子转向黄色旗子，变换目标的时机由玩家自己决定。

4.3.2　设计奖励函数的各项系数

本节实验的目的为确定奖励函数的系数配置，使智能机器对通用规则的学习有最好的性能。奖励函数中共有五个系数直接关系到算法性能，分别是对当前位置和推理目标之间距离的惩罚系数、对速度的惩罚系数、对倾斜角度的惩罚系

数、任务成功给予的大额奖励和任务失败给予的大额惩罚。本节共设置了八套系数组合，智能机器在无人参与且目标已知的情况下，对每套系数组合分别训练 5000 幕进行学习和更新网络参数，再测试 1000 幕以显示算法性能。具体系数配置及对应的测试结果如下表所示。

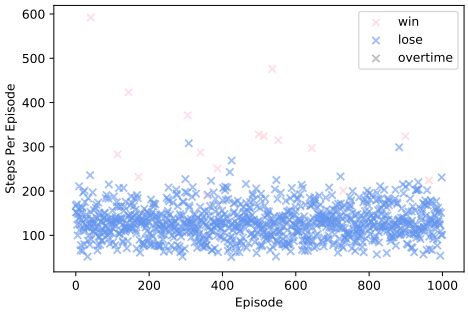
表 4.2　奖励函数的系数配置及对应的测试性能

系数配置	任务成功 (幕)	坠毁 (幕)	时间耗尽 (幕)	成功率
(100,100,100,200,200)	315	60	625	0.315
(100,100,100,100,100)	326	645	29	0.326
(120,100,100,150,150)	16	984	0	0.016
(120,100,100,200,200)	9	991	0	0.009
(120,120,120,100,100)	240	476	284	0.240
(100,120,100,100,100)	557	395	48	0.557
(100,120,100,150,150)	641	195	164	0.641
(120,120,80,200,200)	127	822	51	0.127

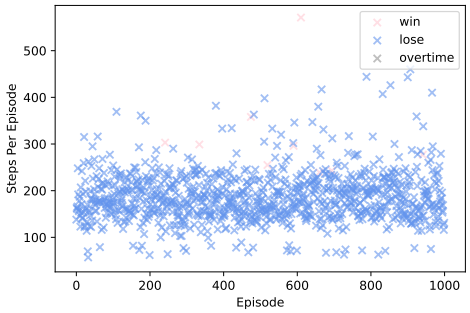
基于上表，本节对不同系数配置在测试阶段的结果做进一步分析，如图4.4所示。第三套和第四套系数配置有最低的任务成功率，且任务失败的主要原因是着陆器坠毁，即智能机器未能很好地控制着陆器平稳飞行。图4.4a和图4.4b分别为第三套和第四套系数配置在测试阶段的步数，图片结果证实了这一点：任务大多在 200 步左右结束，说明着陆器在较短时间内冲撞到地面导致坠毁，从而导致任务失败。因此本节在后续系数配置中调高了对速度项的惩罚。第六套和第七套系数配置有最高的任务成功率，其测试阶段每幕的累积奖赏值和步数如下图所示。由图4.4c和图4.4e对比可知，虽然第六套系数配置的成功率更低，但其在任务失败和超时的任务中得到了更高的累积奖励。由图4.4d和图4.4f对比可知，第七套系数配置在任务成功和失败时都有更大的步数，这对于人机协同任务是有利的，考虑到人的参与和对控制权的偏好，更长的任务步长能够为人类提供更多的反应时间，因此往往能带来更高的任务成功率。综上所述，本节选取第七套系数配置作为智能机器的奖励函数系数配置。

4.3.3　降落过程中不改变目标

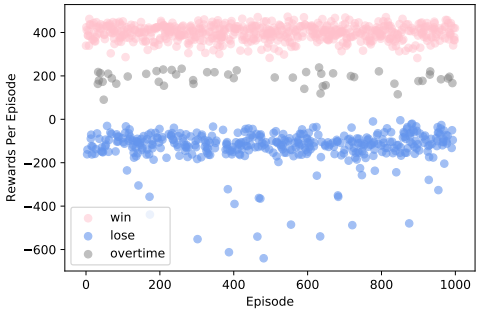
本节实验的目的是为了验证所提方法的有效性，并分析本章提出的方法与其他不考虑人类决策无效性的人机混合决策方法在任务性能上的差异。因此，本节使用 DQN 实现一个最常用的人机混合决策方法，即始终执行机器计算出的奖励价值最高的动作，作为对比实验。每个玩家需要分别用所有输入有效和部分输入无效两种行为策略，在人的单独决策 (HIC)、最高价值人机混合决策 (HVSA)



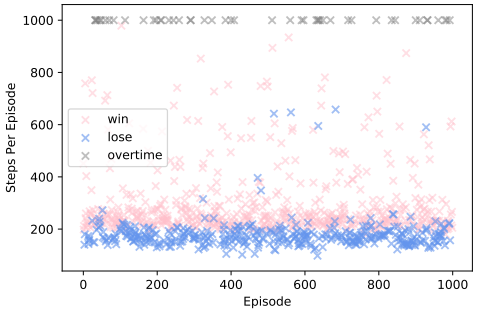
(a) 第三套系数配置的每幕步数



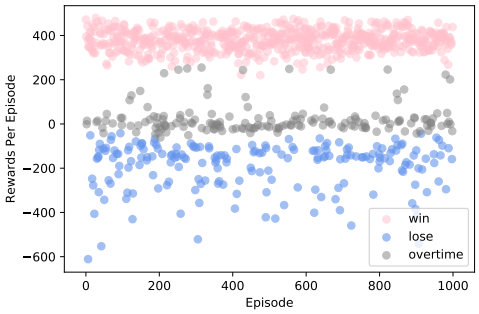
(b) 第四套系数配置的每幕步数



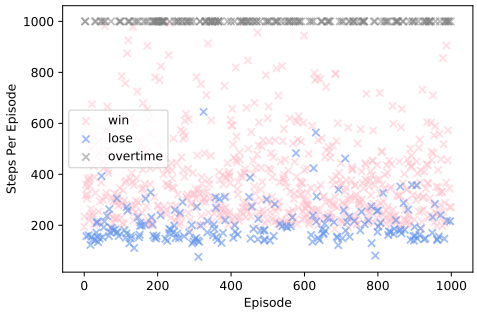
(c) 第六套系数配置的每幕累积奖赏值



(d) 第六套系数配置的每幕步数



(e) 第七套系数配置的每幕累积奖赏值



(f) 第七套系数配置的每幕步数

图 4.4　算法在不同系数配置的奖励函数下的性能图

和本章提出的无效人类决策约束下的人机混合决策 (SAIHI) 三种方法进行操作，共 6 个任务。每个任务有 20000 步，每幕最多 1000 步，所以每个任务至少有 20 幕。每幕的具体步数取决于玩家的能力，通常是 300-700 步。为了确保每幕都存在持续的无效人类决策，本节让玩家在第 100 步和第 200 步时各随机操作一段时间，随机操作的结束时间由玩家决定。呈现给玩家的任务顺序是平衡的，以避免玩家越来越熟练导致结果的偏差。

图4.5显示了 10 位玩家在 6 个任务中的成功率和平均路径长度。图4.5a为人的输入全部有效时的任务成功率。该图显示了将玩家与智能机器结合在一起带来的定量和定性优势。在处理着陆器的突然下降时，玩家很难同时控制三个维度的引擎来保持着陆器的稳定，使得着陆器经常在 150 步内撞向地面或飞出边界，

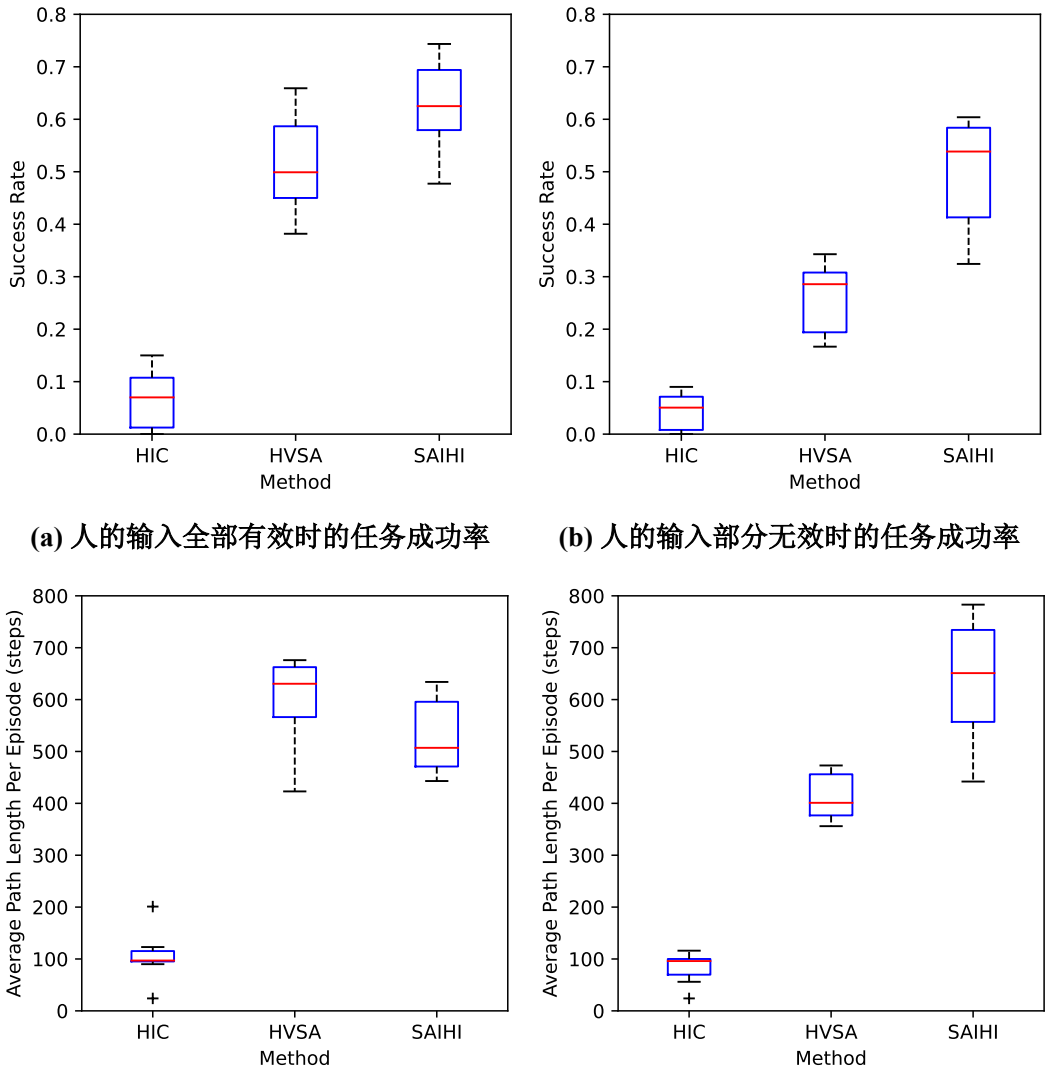
很难在不发生碰撞的情况下准确降落在要求的位置。这主要是因为人类缺乏同时在多个维度准确操纵物体运动的能力。而玩家和机器的混合决策大大增加了成功着陆的可能性，因为机器可以精确控制着陆器动态。当人的输入全部有效时，本章方法的成功率略高于普通的人机混合决策方法。ANOVA (方差分析，对实验中使用的所有变量之间的差异进行统计检验) 的结果是 $F = 7.1130, p = 0.0157$，意味着本章的方法比另一种方法更容易成功。本文认为这种改进是由本章对仲裁函数的改进带来的。

图4.5b为人的输入部分无效时的任务成功率。当存在持续的无效人类决策时，本章的方法显著优于一般方法：ANOVA 的结果为 $F = 30.48, p = 3.04902e - 5$。本章的方法可以判断人输入的行为是否无效，机器会及时介入和接管系统进行单独控制，避免玩家的无效行为影响任务进程，从而有效提高了任务成功率。从图4.5c和图4.5d可以看出，在人的输入均有效的情况下，本章的方法可以在更短的时间内完成任务，这得益于本章提出的仲裁方法更加智能和有效。但当人的输入部分无效时，本章方法可以持续更长的时间，使得系统不会在失去来自外部的有效控制命令后立即崩溃。机器为玩家提供一些缓冲时间，试图回到性能最优的共享控制模式。

图4.6显示了人的输入部分无效情况下着陆器在某一幕成功着陆的过程。图中绿色区域表示该系统由人与机器共同控制，黄色区域表示该系统由机器单独控制，蓝色区域表示该系统由玩家单独控制。图4.6a为着陆过程中共享控制、人单独控制、机器单独控制三种控制模式之间的切换。图4.6b是玩家的输入和最高价值动作之间的行动距离。输入的动作具有最高的奖励，或玩家不输入将导致行动距离为 0。这些数据表明，基于强化学习判断人的决策有效性的方法是有效的：在步骤 100 和 200 附近监测到大的动作距离，导致系统由共享控制切换到机器的单独控制。当动作距离足够小时，控制模式恢复为共享控制。图4.6d是在每一步推断的玩家的目标，图4.6c是相应的意图推理置信度。推理的目标和置信度是着陆器在最后阶段由玩家单独控制的原因：这三对旗帜的坐标是随机生成的，当目标旗帜接近另一对旗帜且着陆器接近地面时，由于两者在计算上的差别较小，机器可能无法确定两个着陆点中的哪一个是玩家的目标。这时，意图推理的置信度会降低，由玩家单独控制着陆器向真正的目标前进，且因为接近地面，人单独控制也不会导致着陆器坠毁。

4.3.4　降落过程中改变目标

这个实验的目的是验证本方法能否识别玩家的目标变化，并帮助玩家完成新的目标。机器根据一系列的轨迹和动作推断目标。因此，改变目标可以逐渐被网络识别，但当着陆器接近推断目标且推断置信度高时，突然改变目标很可能被



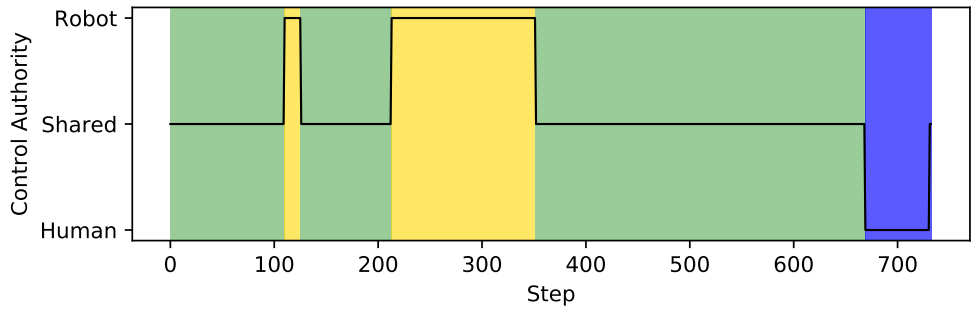
(a) 人的输入全部有效时的任务成功率　**(b) 人的输入部分无效时的任务成功率**

(c) 人的输入全部有效时的每幕平均步数　**(d) 人的输入部分无效时的每幕平均步数**

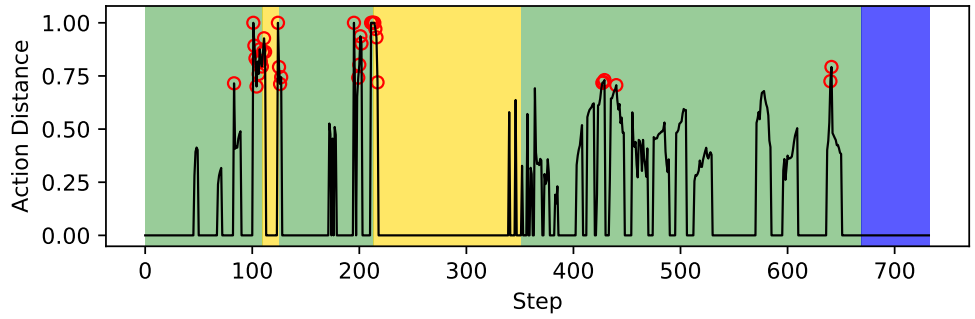
图 4.5　十位参与者用三种决策方法执行任务的结果图

视为无效输入而不被采用。因此，本节在这个实验中设置了两个任务：将着陆器可以移动的垂直距离分成两个相等的部分，玩家需要在这两个空间中分别改变目标，尝试成功着陆。Task1 是让玩家在上面空间改变目标，Task2 是在下面空间改变目标。因为每个玩家在每幕都有不同的步数，所以改变目标的时间是由玩家决定的。本节只规定目标从蓝色旗子变成黄色旗子。此任务的难点便在于机器需要意识到这种变化。如果机器总是将大的动作距离归咎于无效的人类输入，并将目标固定在蓝色旗子上，任务就会失败。只有机器重新推断目标并控制着陆器接近新目标，着陆才能成功。

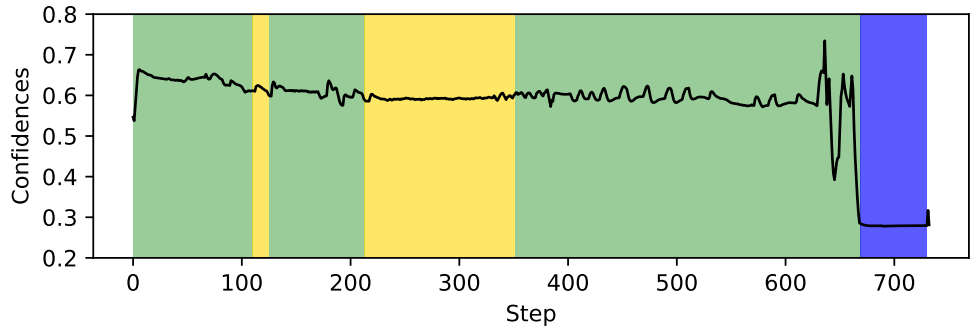
图4.7显示了 10 位玩家完成两项任务的成功率。从图4.7可以看出，Task1 的成功率远远高于 Task2。方差分析的结果验证了这两个任务之间的差异确实是显著的：$F = 11.65, p = 0.0031$。事实上，Task1 的成功率与玩家尝试用部分无效的人的输入成功着陆的任务的成功率大致相同，大多在 0.4 到 0.6 之间 (参见图4.7和



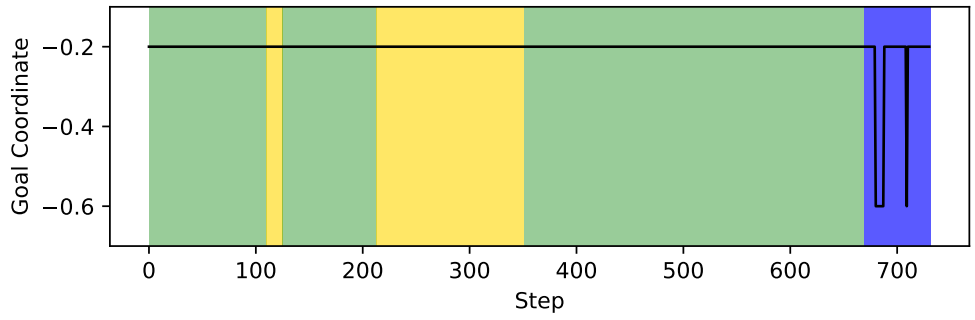
(a) 着陆过程中三种控制模式的切换情况



(b) 玩家的输入和最高价值的动作之间的距离



(c) 意图推理的实时置信度



(d) 意图推理模块推理出的任务目标

图 4.6　人的决策部分无效情况下着陆器某一次成功着陆的过程。绿色区域表示该系统由人与机器共同控制，黄色区域表示该系统由机器单独控制，蓝色区域表示该系统由玩家单独控制。图4.6b中的红色圆圈表示动作距离大于等于 0.7

图4.5b)。这两项任务的本质区别在于，导致动作距离变大的原因是不同的。后者的较大行动距离是由玩家的随机操作造成的，随机输入之间没有相关性和逻辑，导致机器不能从中得到信息。但在 Task1 中，是玩家基于环境状态的有目的控制行为和意图推理模块的延迟识别产生了较大的动作距离。机器从这些输入中推理出一个新的目标，并基于这个新目标获得了较小的动作距离。这是区分无效的人类行为和人类意图变化的关键因素，即输入是否包含有效的信息。Task2 任务成功率低的主要原因是剩余的时间和高度不足以让玩家和机器在新的着陆点顺利着陆。

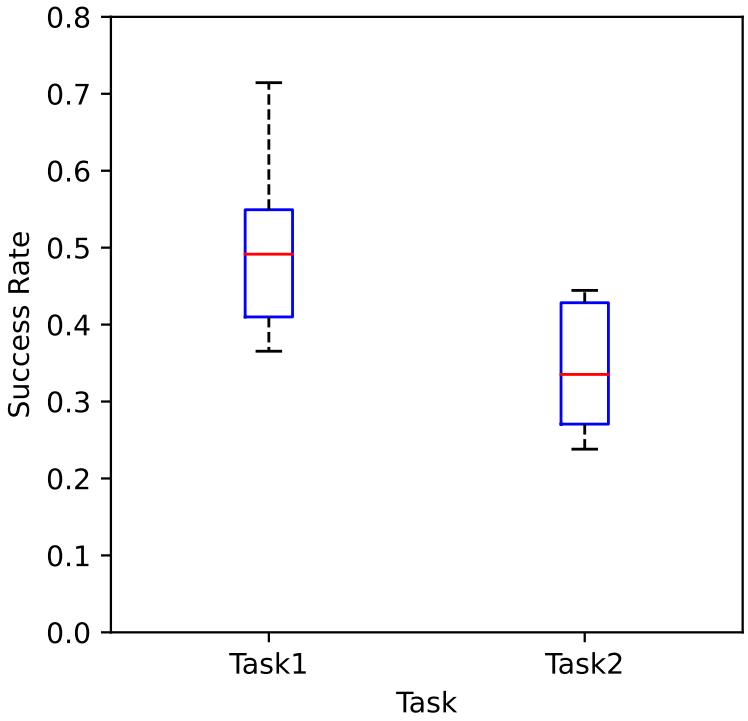


图 4.7　十位玩家完成两项任务的成功率

Task1 和 Task2 的某次成功着陆轨迹分别如图4.8的红色和绿色线条所示。图中黄色的星星表示黄色旗帜的中点，蓝色的星星表示蓝色旗帜的中点；纵坐标为降落空间的垂直坐标，即归一化后的空间高度，其中 0 表示地面，1 表示着陆器初始位置的高度；横坐标为降落空间的水平坐标，其中 0 为水平中点。图片清晰地显示了机器识别出玩家的目标变化，以及由此带来的着陆器的轨迹变化。玩家在上半空间变化目标使着陆器有富裕的时间和空间朝新目标前进，轨迹在高度 0.6 左右出现拐点，即机器发现目标变化并辅助人完成新目标的地方，着陆器在高度 0.2 左右到达最终任务目标的上方，后续为缓慢降落。而在下半空间变化目标则有些危险，轨迹在高度 0.2 左右出现拐点，并在剩余五分之一的空间里完成惊险着陆。可以看到机器为了完成任务不得不控制着陆器上升一段再朝目标降落，这也使得 Task2 有更长的平均路径长度和更少的任务成功率。另一个导

致 Task2 失败的原因是时间耗尽，如果每幕没有时间限制，Task2 的成功率可能会更高。

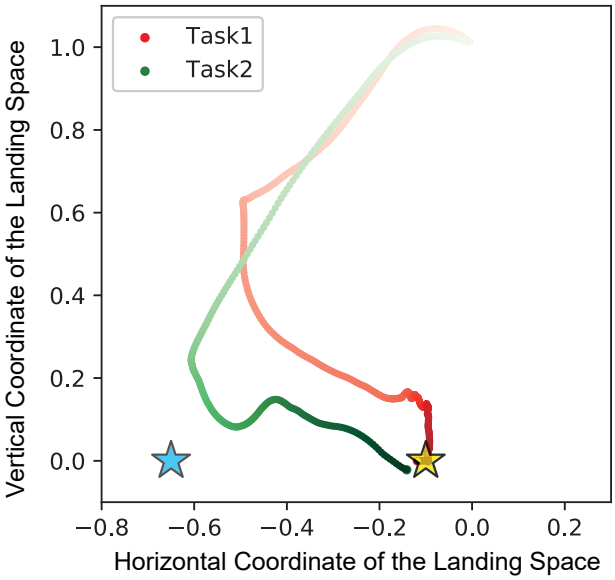


图 4.8　十位玩家完成两项任务的成功着陆轨迹，其中黄色的星星表示黄色旗帜的中点，蓝色的星星表示蓝色旗帜的中点

图4.9为玩家在上方空间改变目标时的某次成功着陆的过程。图中 CA 为 Control Authority，表示控制权限变更：绿线表示该系统为人和机器人共同控制，黄线表示该系统为机器人单独控制，该颜色和权限的对应关系与图4.6保持一致。AD 为 Action Distance，表示人的输入与最高价值动作之间的动作距离，即 $d(a_h, a_{max})$，其中黑色圆圈表示大于等于 0.7 的动作距离。GC 为 Goal Coordinate，表示目标坐标，即目标对旗子的中点。子图显示了步骤 70 到 110 之间三个特征的详细信息。可以看到，在较大动作距离出现后，机器独自控制着陆器。在机器单独控制的过程中，推理出的目标发生变化。目标坐标稳定到新的坐标后，动作距离减小，任务回归到由玩家和机器共同控制的模式。

4.4　进一步讨论

第4.3节的实验结果表明：

（1）基于深度强化学习判断人类决策有效性是切实可行的。

（2）本章提出的方法能够及时有效地判断人的决策是否无效，灵活切换控制模式，分配控制权限以提高系统性能。

（3）当人的决策无效时，用机器的单独控制代替共享控制可以显著提高系统性能。

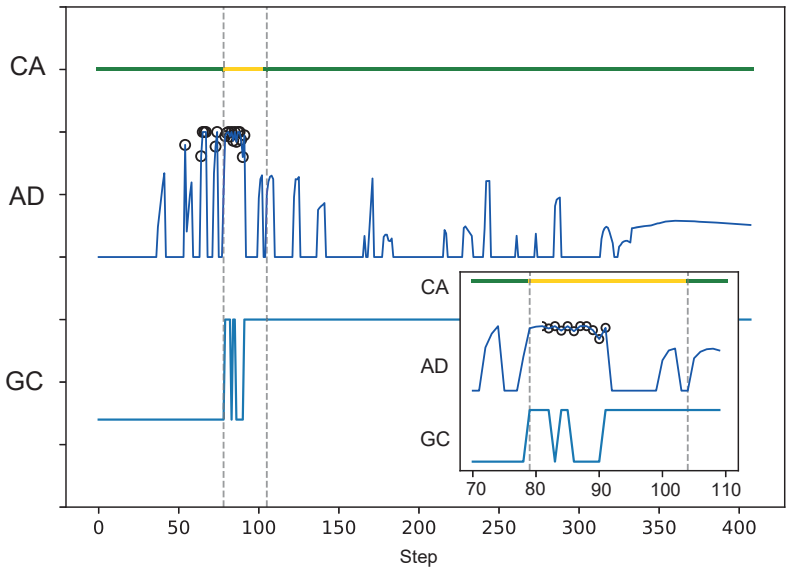


图 4.9　玩家在上方空间改变目标时的某次成功着陆的过程。CA 表示控制权限变更：绿线表示该系统为人和机器人共同控制，黄线表示该系统为机器人单独控制。AD 表示人的输入与最高值动作之间的动作距离，即 $d(a_h, a_{max})$。黑色圆圈表示大于等于 0.7 的动作距离。GC 是目标坐标，即目标标志的中点。子图显示了步骤 70 到 110 之间的详细信息

（4）当机器单独控制系统时，机器应该重新推断人类的意图，并检测输入是否包含有效信息。在实验过程中，本文发现判断人类决策无效的指标应该谨慎选择：本章尝试使用 5/7（七次连续输入中的有五个满足 $d(a_h, a_{max}) \geqslant 0.7$ 作为标准，并发现其过于敏感，无法稳定切换控制权限并保证良好性能。也尝试过 10/15，发现它反应迟缓，导致性能更差，这可能是由于每幕的持续时间较短造成的。本文没有对该指标应如何设置以得到最优性能做进一步的研究，只是提出了这个设想，并对其有效性做了全面的初步验证。

此外，本节的实验结果与智能机器的性能密切相关。如果机器的行为策略不成熟，不论后续参数如何设置，实验结果都很差。为了减轻人的负担，本节让机器单独进行预训练，人类只在有限次的任务中对其进行微调和优化。但实际上，机器在训练过程中的任务成功率也是整个实验的关键因素。让人适量参与机器的预训练可能会改善这个指标，因为人的反馈和指引可以让机器在完全未知的环境中更有效地探索。

4.5　本章小结

本章基于深度强化学习算法提出了一种新的在无效人类决策约束下的人机混合决策方法。在已知目标集、未知系统动态模型和未知人类行为策略的条件下，即使人的输入持续一段时间无效，本方法仍能继续完成正确的目标。本方法使用 DQN 显式判断人的决策是否无效或任务目标是否发生了变化，并分配相应

的控制权，以避免无效的人为输入妨碍任务进程。本章将该方法应用于实时控制任务中，结果表明该方法能够及时、有效地判断和处理人的无效输入，提高系统性能。

本章提出的方法还可以进一步改进：本章仅从人的输入推断目标并判断其有效性，在算法中加入其他隐藏信息，比如人的目光注视[70]，能否取得更好的结果还需要进一步研究。此外，本方法假设可能的目标集是已知的，尝试消除这种依赖，使方法更加灵活和通用是下一步的研究方向。

第 5 章　总结与展望

5.1　论文工作总结

本文以强化学习为工具，以人机混合决策方法为研究对象，从人类决策全时有效和人类决策非全时有效两个方面展开研究，提出了有效的决策算法，改善了人机混合决策系统的性能。本文的研究工作主要总结为以下几个方面：

（1）将人机混合决策方法建模为强化学习问题，使得方法能够对已知的先验知识进行良好的融合和利用，也能在未知被控系统的动态模型、人类的行为策略或关于人类能力的其他先验知识时，获得更好的决策效果。

（2）针对人类决策全时有效的情形，提出了基于强化学习算法的遵循最小干预原则的人机混合决策方法。该方法根据意图推理的置信度设置自适应阈值，以最小程度的干预为人类提供最大程度的帮助，平衡了人类对控制权的需求和对性能的需求，同时优化了人类满意度和系统性能两类指标。

（3）针对人类决策非全时有效的情形，提出一种利用强化学习评估人类决策有效性的人机混合决策方法。该方法使用强化学习算法判断人的决策是否有效以及人类是否变换目标，灵活切换共享控制和介入控制两种模式并为人和机器分配相应的控制权限，以避免无效的人类决策妨碍任务进程，使得即使人的决策持续一段时间无效，系统仍能继续完成正确的目标，有效提高了任务成功率。

5.2　研究展望

本文研究了人类决策有效性给人机混合决策方法带来的问题，并提出了对应的解决方法，扩展了现有的优化设计方案。本文的研究仍可继续深入：

（1）考虑人机混合决策系统中的安全性问题。本文研究的主要评价指标为系统性能（比如任务成功率、任务完成时间等）和人类满意度（比如人类对智能机器的辅助的接受程度、对系统性能的满意度等）两类指标，应该如何设计人机混合决策方法，在实现最优决策效果的同时保证系统的安全运行具有重要的理论价值和实际意义。

（2）考虑在算法中加入其他间接信息。本文的研究仅从人类输入的决策动作来推断目标并判断其有效性，在算法中加入其他间接信息（比如人的目光注视）能否取得更好的效果还需要进一步研究。

（3）考虑机器决策的有效性。深度学习方法具有脆弱性和不确定性，机器的计算结果或许存在偏差和置信度。考虑机器决策的有效性对人机混合决策方法的影响是下一步的研究方向。

参 考 文 献

[1] TANG M, WU F, ZHAO L L, et al. Detection of distracted driving based on multigranularity and middle-level features[C]//2020 Chinese Automation Congress. 2020: 2717-2722.

[2] LIU M, CURET M. A review of training research and virtual reality simulators for the da vinci surgical system[J]. Teaching and Learning in Medicine, 2015, 27(1): 12-26.

[3] MARCANO M, DÍAZ S, PÉREZ J, et al. A review of shared control for automated vehicles: Theory and applications[J]. IEEE Transactions on Human-Machine Systems, 2020, 50(6): 475-491.

[4] OH Y, WU S W, TOUSSAINT M, et al. Natural gradient shared control[C]//2020 29th IEEE International Conference on Robot and Human Interactive Communication. IEEE, 2020: 1223-1229.

[5] DING Y, KIM M, KUINDERSMA S, et al. Human-in-the-loop optimization of hip assistance with a soft exosuit during walking[J]. Science Robotics, 2018, 3(15): eaar5438.

[6] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.

[7] LOSEY D P, MCDONALD C G, BATTAGLIA E, et al. A review of intent detection, arbitration, and communication aspects of shared control for physical human-robot interaction[J]. Applied Mechanics Reviews, 2018, 70(1).

[8] 赵云波, 康宇, 朱进. 人机混合智能系统自主性理论和方法[M]. 北京: 科学出版社, 2021.

[9] JAVDANI S, SRINIVASA S S, BAGNELL J A. Shared autonomy via hindsight optimization [J]. Robotics Science and Systems: Online Proceedings, 2015.

[10] 钱大琳, 刘峰. 人机融合决策智能系统研究的多学科启示[J]. 系统工程理论与实践, 2003, 23(8): 130-135.

[11] NIKOLAIDIS S, ZHU Y X, HSU D, et al. Human-robot mutual adaptation in shared autonomy[C]//2017 12th ACM/IEEE International Conference on Human-Robot Interaction. IEEE, 2017: 294-302.

[12] KOPPULA H S, SAXENA A. Anticipating human activities using object affordances for reactive robotic response[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(1): 14-29.

[13] HAUSER K. Recognition, prediction, and planning for assisted teleoperation of freeform tasks [J]. Autonomous Robots, 2013, 35(4): 241-254.

[14] DRAGAN A D, SRINIVASA S S. A policy-blending formalism for shared control[J]. The International Journal of Robotics Research, 2013, 32(7): 790-805.

[15] KIM D J, HAZLETT-KNUDSEN R, CULVER-GODFREY H, et al. How autonomy impacts performance and satisfaction: Results from a study with spinal cord injured subjects using an assistive robot[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2011, 42(1): 2-14.

[16] FU J, TOPCU U. Synthesis of shared autonomy policies with temporal logic specifications[J]. IEEE Transactions on Automation Science and Engineering, 2015, 13(1): 7-17.

[17] 付海军, 陈世超, 林懿伦, 等. 人在回路的混合增强智能在 Sawyer 的研究与验证[J]. 智能科学与技术学报, 2019, 1(3): 280-286.

[18] GOPINATH D, JAIN S, ARGALL B D. Human-in-the-loop optimization of shared autonomy in assistive robotics[J]. IEEE Robotics and Automation Letters, 2016, 2(1): 247-254.

[19] ANDERSON S J, PETERS S C, PILUTTI T E, et al. An optimal-control-based framework for trajectory planning, threat assessment, and semi-autonomous control of passenger vehicles in hazard avoidance scenarios[J]. International Journal of Vehicle Autonomous Systems, 2010, 8(2-4): 190-216.

[20] VASUDEVAN R, SHIA V, GAO Y, et al. Safe semi-autonomous control with enhanced driver modeling[C]//2012 American Control Conference. IEEE, 2012: 2896-2903.

[21] HE W, LI Z, CHEN C P. A survey of human-centered intelligent robots: issues and challenges [J]. IEEE/CAA Journal of Automatica Sinica, 2017, 4(4): 602-609.

[22] ABBINK D A, CARLSON T, MULDER M, et al. A topology of shared control systems—finding common ground in diversity[J]. IEEE Transactions on Human-Machine Systems, 2018, 48(5): 509-525.

[23] PHILLIPS-GRAFFLIN C, SUAY H B, MAINPRICE J, et al. From autonomy to cooperative traded control of humanoid manipulation tasks with unreliable communication[J]. Journal of Intelligent & Robotic Systems, 2016, 82(3): 341-361.

[24] OWAN P, GARBINI J, DEVASIA S. Addressing agent disagreement in mixed-initiative traded control for confined-space manufacturing[C]//2017 IEEE International Conference on Advanced Intelligent Mechatronics. IEEE, 2017: 227-234.

[25] BROAD A, MURPHEY T, ARGALL B. Highly parallelized data-driven mpc for minimal intervention shared control[J]. arXiv preprint arXiv: 1906.02318, 2019.

[26] LAM C P, SASTRY S S. A pomdp framework for human-in-the-loop system[C]//53rd IEEE Conference on Decision and Control. IEEE, 2014: 6031-6036.

[27] BRUEMMER D J, MARBLE J L, DUDENHOEFFER D D, et al. Mixed-initiative control for remote characterization of hazardous environments[C]//36th Annual Hawaii International Conference on System Sciences. IEEE, 2003: 9-14.

[28] MANIKONDA V, RANJAN P, KULIS Z, et al. A mixed initiative controller and testbed for

human robot teams in tactical operations.[C]//AAAI Fall Symposium: Regarding the Intelligence in Distributed Intelligent Systems. 2007: 92-99.

[29] ADAMS J A, RANI P, SARKAR N. Mixed initiative interaction and robotic systems[C]// AAAI Workshop on Supervisory Control of Learning and Adaptive Systems. Citeseer, 2004: 6-13.

[30] FRIDMAN L, DING L, JENIK B, et al. Arguing machines: Human supervision of black box ai systems that make life-critical decisions[C]//IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019: 1-9.

[31] LIN Z, HARRISON B, KEECH A, et al. Explore, exploit or listen: Combining human feedback and policy model to speed up deep reinforcement learning in 3d worlds[J]. arXiv preprint arXiv:1709.03969, 2017.

[32] KARTOUN U, STERN H, EDAN Y. A human-robot collaborative reinforcement learning algorithm[J]. Journal of Intelligent & Robotic Systems, 2010, 60(2): 217-239.

[33] MARCANO M, CASTELLANO A, DÍAZ S, et al. Shared and traded control for human-automation interaction: a haptic steering controller and a visual interface[J]. Human-Intelligent Systems Integration, 2021, 3(1): 25-35.

[34] SIMON H A. Bounded rationality and organizational learning[J]. Organization Science, 1991, 2(1): 125-134.

[35] AIGNER P, MCCARRAGHER B. Human integration into robot control utilising potential fields[C]//Proceedings of International Conference on Robotics and Automation. IEEE, 1997: 291-296.

[36] OH Y, WU S W, TOUSSAINT M, et al. Natural gradient shared control[C]//2020 29th IEEE International Conference on Robot and Human Interactive Communication. IEEE, 2020: 1223-1229.

[37] LI S, BOWMAN M, ZHANG X. A general arbitration model for robust human-robot shared control with multi-source uncertainty modeling[J]. arXiv preprint arXiv:2003.05097, 2020.

[38] FLEMISCH F, ABBINK D, ITOH M, et al. Shared control is the sharp end of cooperation: Towards a common framework of joint action, shared control and human machine cooperation [J]. IFAC-PapersOnLine, 2016, 49(19): 72-77.

[39] SCHULTZ C, GAURAV S, MONFORT M, et al. Goal-predictive robotic teleoperation from noisy sensors[C]//2017 IEEE International Conference on Robotics and Automation. IEEE, 2017: 5377-5383.

[40] EDDY S R. What is dynamic programming?[J]. Nature Biotechnology, 2004, 22(7): 909-910.

[41] 仵博. 动态不确定环境下的智能体序贯决策方法及应用研究[D]. 中南大学, 2013.

[42] 崔蓉. 基于序贯决策的无线传感网络频谱感知策略与分配方法[D]. 北京邮电大学, 2015.

[43] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[M]. MIT Press, 2018.

[44] KAELBLING L P, LITTMAN M L, CASSANDRA A R. Planning and acting in partially observable stochastic domains[J]. Artificial Intelligence, 1998, 101(1-2): 99-134.

[45] 范长杰. 基于马尔可夫决策理论的规划问题的研究[D]. 中国科学技术大学, 2008.

[46] SHACHTER R D, PEOT M A. Decision making using probabilistic inference methods[C]// Uncertainty in Artificial Intelligence. Elsevier, 1992: 276-283.

[47] AMATO C, BERNSTEIN D S, ZILBERSTEIN S. Optimizing memory-bounded controllers for decentralized pomdps[J]. Brain Research, 1981, 216(1): 11-33.

[48] SHANI G, PINEAU J, KAPLOW R. A survey of point-based pomdp solvers[J]. Autonomous Agents and Multi-Agent Systems, 2013, 27(1): 1-51.

[49] 章宗长. 部分可观察马氏决策过程的复杂性理论及规划算法研究[D]. 中国科学技术大学, 2012.

[50] JAVDANI S, ADMONI H, PELLEGRINELLI S, et al. Shared autonomy via hindsight optimization for teleoperation and teaming[J]. The International Journal of Robotics Research, 2018, 37(7): 717-742.

[51] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[J]. arXiv preprint arXiv:1312.5602, 2013.

[52] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[53] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double q-learning [C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 30. 2016.

[54] WANG Z, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning[C]//International Conference on Machine Learning. PMLR, 2016: 1995-2003.

[55] HESSEL M, MODAYIL J, VAN HASSELT H, et al. Rainbow: Combining improvements in deep reinforcement learning[C]//Thirty-second AAAI Conference on Artificial Intelligence. 2018.

[56] ABBINK D A, MULDER M, BOER E R. Haptic shared control: smoothly shifting control authority?[J]. Cognition, Technology & Work, 2012, 14(1): 19-28.

[57] REDDY S, DRAGAN A D, LEVINE S. Shared autonomy via deep reinforcement learning[J]. arXiv preprint arXiv:1802.01744, 2018.

[58] XU A, DUDEK G. Trust-driven interactive visual navigation for autonomous robots[C]//2012 IEEE International Conference on Robotics and Automation. IEEE, 2012: 3922-3929.

[59] ZIEBART B D, MAAS A L, BAGNELL J A, et al. Maximum entropy inverse reinforcement learning.[C]//Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence. 2008: 1433-1438.

[60] PHAM V, BLUCHE T, KERMORVANT C, et al. Dropout improves recurrent neural networks for handwriting recognition[C]//2014 14th International Conference on Frontiers in Handwriting Recognition. IEEE, 2014: 285-290.

[61] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent neural network regularization[J]. arXiv preprint arXiv:1409.2329, 2014.

[62] LUONG M T, SUTSKEVER I, LE Q V, et al. Addressing the rare word problem in neural machine translation[J]. arXiv preprint arXiv:1410.8206, 2014.

[63] MARCHI E, FERRONI G, EYBEN F, et al. Multi-resolution linear prediction based features for audio onset detection with bidirectional lstm neural networks[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2014: 2164-2168.

[64] CARLSON T, DEMIRIS Y. Collaborative control for a robotic wheelchair: evaluation of performance, attention, and workload[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2012, 42(3): 876-888.

[65] DREISSIG M, BACCOUR M H, SCHÄCK T, et al. Driver drowsiness classification based on eye blink and head movement features using the k-nn algorithm[C]//2020 IEEE Symposium Series on Computational Intelligence. IEEE, 2020: 889-896.

[66] JING D, LIU D, ZHANG S, et al. Fatigue driving detection method based on eeg analysis in low-voltage and hypoxia plateau environment[J]. International Journal of Transportation Science and Technology, 2020, 9(4): 366-376.

[67] SHARMA M K, BUNDELE M M. Design and analysis of k-means algorithm for cognitive fatigue detection in vehicular driver using oximetry pulse signal[C]//2015 International Conference on Computer, Communication and Control. IEEE, 2015: 1-6.

[68] BROAD A, MURPHEY T, ARGALL B. Learning models for shared control of human-machine systems with unknown dynamics[J]. arXiv preprint arXiv:1808.08268, 2018.

[69] JAIN S, ARGALL B. Probabilistic human intent recognition for shared autonomy in assistive robotics[J]. ACM Transactions on Human-Robot Interaction, 2019, 9(1): 1-23.

[70] ADMONI H, SRINIVASA S. Predicting user intent through eye gaze for shared autonomy [C]//2016 AAAI Fall Symposium Series. 2016.

致　谢

行文至此，本篇论文带着我的研究生生涯已接近尾声。万分不舍，思绪繁杂。在过去二十年的求学过程中，我不断地拨开迷雾，发现与舍弃，推翻与建构，曾质疑过一分耕耘一分收获是否为真，质疑过天道酬勤是否为真，现在行至终点，只觉功不唐捐。

感谢我的导师康宇老师和赵云波老师。我顽劣有时，愚钝有时，而康老师耐心常有，教导常有，从康老师那学到的科研思维和处事原则都是使我一生受益的宝贵财富。感谢赵老师在专业上的悉心辅导、思维上的耐心引导和学术上的严格教导，让我一步步地明白如何发现问题、定义问题以及解决问题，对科研工作有了自己的概念和评价体系，对论文撰写形成了自己的思维逻辑。

感谢父母和弟弟的支持与爱。我每次回家都有丰盛的饭菜，温暖的被窝，和一屋子的爱，在这些底气和退路之上，我方能追求实现个人价值。特蕾莎修女说我们以为贫穷就是饥饿、衣不蔽体和没有房屋，然而最大的贫穷是不被需要、没有爱和不被关心。是父母和弟弟让我时刻觉得自己无比富足。

感谢我的朋友们，感谢让我无数次心动和心碎的人，金风玉露一相逢，便胜却人间无数。我时常觉得自己在飞，在几万米的高空，往上是浩大的新世界待我探索，往下是层层叠叠的白云供我落脚，至亲是拉住我的线，朋友是一起行路的人。非常喜欢和你们一起的饭后散步时光，漫无边际地聊天大笑，分享花开和四季，看校园的黄昏和晚星。在与科研的搏斗中我先走一步，但求知不止，日后顶峰相见。

感谢自己，在痛苦中没有停止前进，这是我今天走到这里的原因。我始终不是主流评价体系里的精英，这很好，让我用平凡的眼光审视自己的人生，用弱者的同理心安慰更多的人。研究生三年让我学会两件事，一是方法，学会了怎么学习，怎么科研，怎么探索事物的本质；二是态度，学会接受自己的失败和平庸，学会带着好奇心看这个世界。或许多年后再看这两个感悟会觉得稚嫩，但管它呢，人生感激之情常有，二十几岁的心态不常有。

最后希望我爱的人们都万事顺利、一切称心，那我将是世界上最幸福的人。

在读期间发表的学术论文与取得的研究成果

已发表论文

1. Shiyi You, Yu Kang, Yun-Bo Zhao and Qianqian Zhang, ”Adaptive Arbitration for Minimal Intervention Shared Control via Deep Reinforcement Learning”, *2021 China Automation Congress (CAC)*, 2021, pp. 743-748, doi: 10.1109/CAC53003.2021.9727522.

已投稿论文

1. Shiyi You, Yu Kang, Yun-Bo Zhao and Qianqian Zhang, ”Shared Autonomy Based on Deep Reinforcement Learning Subject to Possible Ineffective Human Behaviors”, *IEEE Systems, Man, & Cybernetics Magazine*

专利

1. 康宇，游诗艺，赵云波，吕文君。一种基于深度强化学习的共享自主方法。申请中。