中国科学技术大学

博士学位论文

面向人机序贯决策的混合智能方法研究

作者姓名：　张倩倩

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：　二〇二一年十月十四日

University of Science and Technology of China

A dissertation for doctor's degree

Research on Hybrid Intelligent Method Oriented to Human-Machine Sequential Decision Making

Author: Zhang Qianqian

Speciality: Control Science and Engineering

Supervisor: Prof. Yu Kang Prof. Yun-Bo Zhao

Finished time: October 14, 2021

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。

保密的学位论文在解密后也遵守此规定。

☑ 公开 ☐ 保密（____ 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，机器智能得到不断的提高，随之而来的则是机器智能得以在各行各业应用发展。在此进程中，不可避免的会遇到机器自主性不足以解决本身该由人类解决或者人类必须参与决策的情况，考虑此种场景下人类智能和机器智能共同作用的决策问题则显得尤为重要和有意义。更具体地，序贯决策问题作为一类具有时序性和多阶段性的动态决策问题，其发展与当下人工智能时代下的工程应用、生产生活等领域息息相关。人的作用体现在序贯决策问题的两方面，一则，人本身属于序贯决策问题模型中的一部分，即该类问题是离不开人的如微创外科手术等；二则，人的相关信息不体现在序贯决策问题模型中，而是因人独特的认知能力使得其可以出现在问题的求解办法中，达到改善问题求解的目的如人对机器搜救系统的引导等，我们将上述两种场景统称为“人机序贯决策问题”。

针对人机序贯决策问题，由于人类智能和机器智能本质上的区别，数学表达上的巨大差异，使得人和机器共同作用于问题求解时，不可避免的因为协调原因造成决策质量不高甚至决策失误的现象。然而直接应用传统人机系统的控制算法不能有效处理这些问题，从而引起机器代理失效，人力浪费，甚至还会造成决策系统性能恶化甚至崩溃。因此，亟需设计有效的人机混合智能算法来解决这些问题。本文以人机序贯决策问题为研究对象，围绕人机混合智能控制中的决策权限划分、介入控制触发切换时机和共享控制混合人机决策动作程度三个问题展开研究，旨在提出有效的人机混合智能算法来改善提升人机序贯决策问题的求解。本文的研究工作主要包括以下几个方面:

1. 提出了基于强化学习方法的人机混合智能控制框架。通过将机器代理的决策和人类的决策以可信性和安全性为评价指标进行仲裁选择，以确定更优的待执行决策动作。同时考虑了基于模型的强化学习子系统和基于无模型的强化学习子系统，为适应广泛的序贯决策应用场景提供了更多可能。
2. 针对人机序贯决策中的介入控制问题，提出了自主性及自主性边界的概念，通过将自主性边界的求解形式化为与任务目标相关的常规优化问题进行讨论判定，优化介入控制的控制方案和算法，实现人机序贯决策中人介入机器场景和机器介入人场景下的决策性能提升。
3. 针对人机序贯决策中的共享控制问题，提出了基于自主性边界的混合参数优化设计方案，通过自适应调节混合参数大小直接影响最终待执行动作的生成。考虑了人机动作的融合程度，使得最优解在人的动作空间和机器的动作空间所共同张成的扩展空间中出现，为决策质量的提升提供了扩展空

间。

4. 针对介入控制和共享控制中所估计的自主性边界值可能存在单值估计不准确的问题，提出了基于贝叶斯神经网络的不确定性估计办法，获得自主性边界的概率分布信息并用于决策动作生成，利用自主性边界的不确定性优化设计人机混合智能算法，既使得决策动作的优化存在更多选择，也更加符合人们对决策边界的模糊性思考。

综上所述，本文面向人机序贯决策对混合智能算法所面临的问题进行了系统性的研究，创新性地提出了对应的解决方案，推动了人机序贯决策求解和混合智能算法的进一步发展。

关键词：人机序贯决策；混合智能算法；自主性边界；介入控制；共享控制；仲裁机制；强化学习

ABSTRACT

With the development of artificial intelligence technology, machine intelligence has been continuously improved, followed by the application and development of machine intelligence in all walks of life. In this process, it is inevitable that the autonomy of the machine is not enough to solve the situation that it should be solved by humans or that humans must participate in decision-making. It is particularly important and meaningful to consider the decision-making problem of human intelligence and machine intelligence in this scenario. More specifically, as a kind of sequential and multi-stage dynamic decision-making problems, the development of sequential decision-making problems is closely related to the fields of engineering applications, production and life in the current artificial intelligence era. The role of humans are reflected in two aspects of sequential decision-making problems. One is that humans themselves are part of the sequential decision-making problem model, that is, such problems are inseparable from humans, such as minimally invasive surgery. Second, humans' relevant information is not reflected in the sequential decision-making problem model, but because of the unique cognitive ability of humans, it can appear in the problem-solving method to achieve the purpose of improving the problem-solving, such as in the machine search and rescue system. We collectively refer to the above two scenarios as the "human-machine sequential decision-making problem".

For the human-machine sequential decision-making problem, due to the essential difference between human intelligence and machine intelligence, and the huge difference in mathematical expression, when humans and machines work together to solve the problem, it is inevitable that the quality of decision-making is not high or even the phenomenon of decision-making errors due to coordination reasons. However, the direct application of the control algorithm of the traditional human-machine system cannot effectively deal with these problems, which causes the failure of the machine agent, the waste of manpower, and even the performance deterioration or even the collapse of the decision-making system. Therefore, it is urgent to design an effective human- machine hybrid intelligent algorithm to solve these problems. This thesis takes the human-machine sequential decision-making problem as the research object, and re- searches on three issues: the division of decision-making authority in human-machine hybrid intelligent control, the timing of triggering switching of traded control and the degree of mixing of shared control. It aims to propose an effective human-machine hybrid intel-

ligent decision-making algorithm to improve the solution of the problem of improving the human-machine sequential decision-making. The main content of this thesis mainly includes the following aspects:

1. Aiming at the human-machine sequential decision-making problem, a human-machine hybrid intelligent control framework based on reinforcement learning method is proposed. By arbitrating decisions made by machine agents and humans with credibility and safety as evaluation indicators, the decision-making actions to be executed are determined more optimally. Considering the model-based reinforcement learning subsystem and the model-free reinforcement learning sub- system, it provides more possibilities for adapting to a wide range of sequential decision-making application scenarios.
2. Considering the traded control in the human-machine sequential decision-making problem, the concept of autonomy and autonomy boundary is proposed. By formalizing the solution of the autonomous boundary as a conventional optimization problem related to the task goal for discussion and judgment, the control scheme and algorithm of the intervention control are optimized. And the decision performance of human intervention machine and machine intervention human in the human-machine sequential decision-making process is improved.
3. Considering the shared control problem in the human-machine sequential decision-making problem, a hybrid parameter optimization design scheme based on the autonomous boundary is proposed, which directly affects the generation of the final action to be executed by adaptively adjusting the hybrid parameter. Taking into account the degree of integration of human and machine actions, the optimal solution appears in the expanded space formed by the human action space and the machine action space, which provides more space for the improvement of decision-making quality.
4. Considering the one-sidedness of the single value estimation of the autonomous boundary in traded control and shared control, an uncertainty estimation method based on Bayesian neural network is proposed to obtain the probability distribution information of the autonomous boundary and use it for decision-making action generation. By using the uncertainty of the autonomous boundary to optimize the design of the human-machine hybrid intelligent algorithm, it not only makes more choices for the optimization of decision-making actions, but also more in line with humans' vague thinking about the decision-making boundary.

In summary, this thesis systematically studies the problems faced by the human-

machine hybrid intelligent algorithm to solve the human-machine sequential decision-making, innovatively proposed corresponding solutions, and greatly promoted the human-machine sequential decision-making solution.

Key Words: Human-machine sequential decision-making; Hybrid intelligent algorithm; Autonomous boundary; Traded control; Shared control; Arbitration mechanism; Reinforcement learning

目 录

插 图 清 单

表 格 清 单

算 法 清 单

第 1 章　绪　　论

本章首先介绍了人机序贯决策问题和人机混合智能的重要研究意义，接着针对人机序贯决策问题及其解决方法—混合智能，阐述了相关研究现状和不足，然后概述了本文的主要研究内容，并最后给出了全文的章节安排。

1.1　研究背景与意义

1.1.1　人机序贯决策问题

以时序性和多阶段性为标志的序贯决策 (Sequential Decision-Making，SDM) 问题 [1-6] 是一类广泛存在于社会、经济、军事、工业生产等各领域的重要决策问题。该类决策问题由于决策空间随着决策步长指数增长，求取最优决策序列往往存在巨大困难。我们注意到，在很多序贯决策问题中或者人本身便处于决策的环路中，或者人因其独特的认知能力而有助于最优决策的求取，在本文，我们将此两种问题分别称为“人参与问题”的人机序贯决策 (Human-Machine Sequential Decision-Making，HMSDM) 和“人介入方法”的人机序贯决策。

在“人参与问题”的序贯决策中，人本身便是参与决策的主体，如人机共驾系统 [7-9] 中驾驶员本身对汽车驾驶负有决策责任。遥操作外科手术 [10-12] 也可被称作“人参与问题”的人机序贯决策，因为在其中，如果没有人类专家的参与，目前阶段的智能机器是无法掌握手术技能的，但可以利用辅助设备如达芬奇外科手术辅助系统 [13]，通过充分发挥外科专家的人类医学经验和机器的精准操控能力进行高效的遥操作微创外科手术。显而易见地，在“强人工智能”时代到来之前，人机混合智能系统会越来越成为普遍的和常见的智能形式，“人参与问题”的人机序贯决策也自然变得越来越重要。

在“人介入方法”的序贯决策中，人类本身并不在原始的决策问题中，但人的介入可以有效改善原机器决策系统的性能，比如机器人在不确定环境下执行搜救任务 [14]，观察者可以通过告知机器人自己对环境的理解和对目标的认定等信息帮助机器人实现搜救。此外比如在一些高温有毒等高风险使得人类难以直接参与的搜救任务中，机器人一定程度上能够通过所配置的各种传感器进行自主搜索和救援，但是对于一些不确定的外界环境或者面临一些突发性，机器难以通过事先的训练适应新的动态环境导致搜救效果被大大限制。通过引入人对不确定动态环境的认知和决策能力则能够极大帮助问题的解决。如此上述搜救例子也启发我们在一些完全自主的机器算法出现漏洞或者瓶颈时，考虑人类智能的特有认知和优势，为改善原决策系统提供一个好思路。

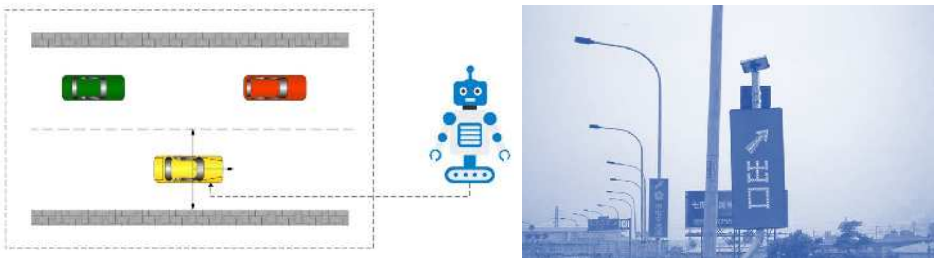


图 1.1　以深度学习为基础的图像识别技术应用在自动驾驶系统的交通标志识别

1.1.2　人机混合智能

以深度学习 [15-16] 为基础的人工智能 [17-20]（Artificial Intelligence, AI）技术在近些年蓬勃发展开来，引发了人工智能发展的一个新的高潮。直觉上来讲，由 AI 技术和自动化技术共同驱动的机器（比如汽车）既具有了自动化能力（多种机械/电子控制系统使得汽车可以按指令行驶），又具有了自主能力（AI 驱动的环境感知和路径规划等能力），这似乎看上去有种“人类参与不再重要”的假象。毕竟正如我们所看到的事实：机器智能已经可以在无人参与下完成越来越多的任务。然而，在某些关键的应用场景（如专家进行手术、自主武器系统 [21-22] 等）中，人的参与也是不可缺少的一环，这源于两点：一方面，从某种程度上讲，机器智能尚未完全达到人类智能的水平；另一方面，实际工程系统的最终服务者是人，需要为人类保留最终决策权限。

进一步地，以深度学习为代表的 AI 技术存在诸多矛盾，如动态实时性和计算复杂度的矛盾，可信度和不确定性的矛盾，鲁棒性和易于攻击的矛盾等。AI 技术存在的这些弊端也使得其经常受到系统难以稳定可靠运行的困扰，比如在自动驾驶汽车中，如图1.1所示，在一定范围内，提前训练好的驾驶系统能够完成道路两侧交通标志的识别工作，并且基于这种准确的环境感知完成车辆正常行驶的任务。然而当出现一些突发状况，如心存不良的人为了达成某种需要，在道路两侧故意树起违规标识。自动驾驶系统能否准确感知并处理得当，则是一件悬而未知的事情，这种未知又将会带来什么样的生命财产损失后果，听起来似乎令人毛骨悚然。上述问题的解决方案存在两种方向：一者是寄希望于强人工智能 [23-24] 的出现 (泛指具备和人类同等或超越人类的人工智能水平)；二者是将人重新请回控制回路中 (半自动驾驶即是一种表现形式)。显然地，第一种“强人工智能”方案长路漫漫，可能会取得进展，也可能永远实现不了。而第二种“人在回路（Human In/On The Loop, HITL/HOTL）[25-28]”，虽说是自工业革命以来出现的奇特现象（以往是将人从生产力中解放出来，而这种方案却是将人带入回路），但却是目前科技发展水平下非常有效的办法，不同于落后的手工生产，此种方法是将人类和机器在智能决策上的融合，而非执行上的结合。此外，对于将人类和机器共同决策的研究讨论，以及随着生命科学技术的进步，未必不能形成

对“强人工智能”的促进作用。

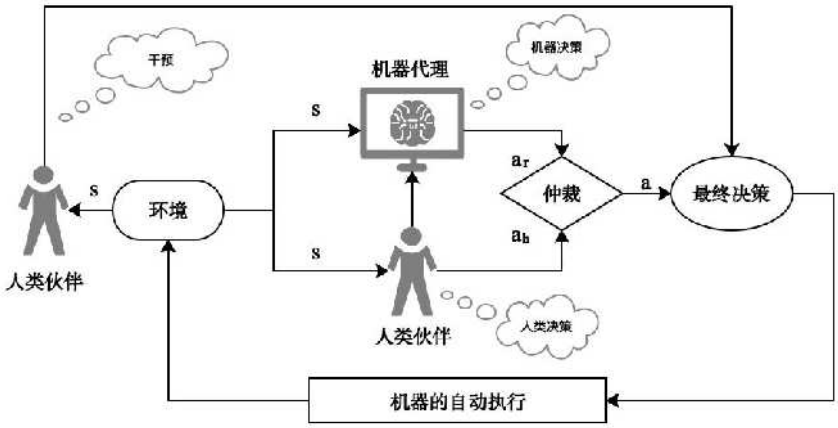


图 1.2　人机混合智能控制框架

在科学和工程领域，人与机器相互依存、影响、协同而构成的整体便称之为宽泛意义上的“人机系统”[29-32]。这是一个具有悠久历史、现在也高度活跃的研究领域。综上所述，基于上述研究背景意义和需求，我们研究如图1.2所示的新型人机混合智能控制框架，以人机序贯决策问题为研究对象，其包括“人参与问题”的人机序贯决策和“人介入方法”的人机序贯决策，考虑利用人机混合智能框架对此两类问题进行建模求解。分析机器代理决策动作和人类伙伴的决策动作，并且通过某种仲裁机制混合得到待执行动作，最终应用到序贯决策问题中，以实现相关任务的高效执行或系统性能的提升和改善。

1.2　国内外研究现状

人机序贯决策问题继承了传统序贯决策问题的形式和特征，其求解方法近年来受到越来越多的关注并取得了部分研究成果。本节针对人机序贯决策的求解过程中可能存在的问题（人机决策权限问题、介入控制的触发问题以及共享控制的融合问题）分别进行详细说明，并详述现有的解决方案。

1.2.1　人机序贯决策的决策权限研究现状

科学技术的发展使得机器越来越趋于智能化，逐渐改变人们的日常生活和工业技术，由此也使得人机之间的关系逐步发生变化。如何平衡机器的智能控制和人类的控制能力，以及如何通过良好的人机融合使得系统安全高效的运行，均涉及到本小节所讨论的人机序贯决策的权限问题。因此众多学者提出了分析讨论。文献 [33] 给出了权限和控制权限的说明，即权限是发号施令、做出决策、强制执行的权力，人机系统中的权限指人和机器的决策动作被执行的范围。通常情况下权限由设计者提前给定，并且对后续的评价结果产生一定影响。关于权限，

一般包括两方面：1）控制权限 [34]：一个特定的控制分配，是执行器用来执行某控制动作的权限；2）控制变化权限 [33]：分配控制权限的变化，是将控制权限更改为另一个控制分布的权限。

表 1.1　自动驾驶等级描述

等级	描述	加/减速/转向	环境监测	紧急退回	辅助系统能力
L0	无机器自动辅助	人	人	人	n/a
L1	转向或者加减速仅能实现一条，驾驶员时刻关注	人和辅助系统	人	人	实现部分系统自主
L2	转向或者加减速均能由机器实现，驾驶员时刻关注	辅助系统	人	人	实现部分系统自主
L3	无需驾驶员时刻监督，仅出问题时人介入	辅助系统	辅助系统	人	实现部分系统自主
L4	无需驾驶员时刻监督，出问题时机器自动靠边	辅助系统	辅助系统	辅助系统	实现部分系统自主
L5	全自动驾驶	辅助系统	辅助系统	辅助系统	实现完全系统自主

人机决策权限的大小于不仅取决于系统自动化程度的多少，也取决于人和机器决策能力 [35-37] 的大小。一方面，在自动化程度较低的应用领域，人类完成大部分控制任务，仅将不要求专家知识的小部分控制工作交由机器完成，此时人的决策范围大于机器的决策范围，比如微创外科手术辅助系统中对专家经验的依赖以及对部分机器精度微调的采纳。随着机器自动化和自主性不断提升，在一些控制场景中人类的诸多工作逐渐被机器所取代，因此也使得人类的决策范围逐渐缩小，同时机器的决策范围逐渐扩大，比如半自动驾驶辅助系统。此外高度自动化则意味着人类在动态过程中的控制占比较小，由机器接管大部分的控制权，此时人类的决策范围较小而机器的决策范围较大 [38]。另一方面，除了在执行任务之前设定固定的人机决策权限以外，根据人机决策能力的大小动态改变决策权限也是必要的 [34]。此时则需要较为完善的评价机制实现对人机决策动作的质量进行评估，以得出谁的决策动作更优或者占比应该更大的结论 [39-41]。

自主性被描述为一种决策者动作被执行的程度 [42]。为了使得人机系统更加安全高效，一种委托架构系统被开发以实现既能保留自动化的好处，又能最大限度减少其成本和危害 [43]。该系统使用中间级别的 LOA（Level Of Automation），使得人工和自动化都不是专门负责大多数任务。LOA 是在某种抽象级别上委派的具有某种级别权限的任务和被委派有某种级别权限的资源用于执行该任务的组合任务，如在普遍公认的自动驾驶等级标准 SAE J3016LOA（表1.1）中，自动驾驶等级的提升过程即是利用了 LOA 的思想。文献 [33] 认为责任和权力的分配以及控制权随时间的变化能够以许多不同的方式设计，但确保某些原则是非常可取的。例如必须针对适当的场景和适当的关系（如人是权威的且绝对处于控制位置 [44]）设计人机系统，允许人和机器以安全、高效和可靠的方式分担责任、权力和自主性。

决策权限、决策能力等之间的动态平衡关乎人机系统的责任分配，其中仍有许多问题值得研究。将控制的转变视为动态平衡的热点，考虑紧急情况下对控制转变的进一步结构化和调查 [45]。当驱动程序和自动化共享能力和权限使得不同决策者对正确动作持有不同意见时，合作伙伴之间的协商和仲裁成为动态平衡中的一个关键方面 [46]。在合作伙伴的能力（如机器自动化）可以动态变化的情况下，未来能力的预测也可能成为成功改善动态平衡的重要因素 [47]。另外，研究表明，驾驶员通常难以准确了解机器自动化的能力和局限性 [48]，这会导致自动化的意外、情境意识的降低、自满和依赖 [49] 等。此时对应的解决办法是对不确定性 [50-51] 的衡量。文献 [52] 建议根据人类-自动化合作的框架将不确定的面孔作为自动化不确定性的自然指标，并且应用于驾驶领域检查不确定性信息，进而达到改善人类驾驶员和机器自动化之间相互辅助的目的。

1.2.2　人机序贯决策的介入控制研究现状

为了确定人机序贯决策问题中人和机器的合作关系，涉及谁的优先级更高或者谁是主要决策者，而出现了介入控制下的人机混合智能，其涉及的关键词是“介入”，对应“人在环上”（HOTL，Human-On-The-Loop）。介入控制是基于一定的评价机制或任务目标，强决策者对弱决策者的强制干预 [53-55]。该类控制方法中人与机器处于非对称的地位，或者是满足人的需求而保持人的控制，则有人-机器的主从关系，或者是利用机器的高精度能力防止人的错误，具有机器-人的主从关系 [56-58]。

关于介入控制的定义，缺少较为官方的说明。这里给出较为相关的定义说明，比如，[59] 从人与机器人的交互角度给出了介入控制的定义，在其中人和机器人处于不同的地位：人是目标的制定者而机器人是动作的执行者，这实际上更近于传统人机系统中人与机器的关系。[54] 给出了关于介入控制的一般定义，并

对其运行逻辑进行了说明，强调人或者机器在任何时候对系统都有排他性的控制权。在介入控制中，一方往往会比较依赖某一方，不恰当的介入发生时可能会引发系统运行模态的切变，造成系统性能的恶化，比如图1.3中所示人对作战指挥系统的适时介入对系统整体决策效果有非常重要的意义。如何处理介入的触发时机，并且提升人机序贯决策性能是值得继续研究的课题。

图 1.3　人对作战指挥系统的介入

当人类操作者想要在某些部分任务期间控制机器，而机器代理完成任务的其他部分的自主控制，通过设计智能软件架构，包括混合主动规划器，执行监视器，机器人技能以及用户界面，文献 [53] 完成人类协同控制的工作，解决了一个重要情况：机器人不知道环境如何变化或者当人类控制时机器代理不知道任务的哪些部分已经完成而导致人类将控制权交还给机器人时可能发生错误的危险情况。文献 [60] 展示了一个用人类用户手势进行介入控制的远程操作系统，其未对人类的直接控制和机器的自主控制进行融合，而是利用感知、估计、规划等组件，根据物体姿势和形状完成机器人运动轨迹的规划。在介入控制中，机器或人类在任意时间点拥有对系统的独立控制权，机器或人可以根据各自特定的故障概率模型提出控制中的混合主动权交换，这种情况发生在故障概率模型出现不一致的时候以及某一决策者不同意介入时 [54]。

根据 [59]，在介入控制中，人类操作员和机器代理均控制机器人的动作，人类操作员启动了任务或者输入决策动作之后，机器代理通过跟随所需的输入自主执行任务，同时人类操作员监视机器人。Lex[61] 基于集成思想利用争论机器框架提出了主系统和辅助系统用于解决同一个控制任务，当主系统和辅助系统做出相对不同的决策动作时，人类作为监督者单方面干预到智能机器的决策中。相比之下，文献 [62] 依赖于数据驱动的人机系统，以基于模型的表示评估用户可能希望并行提供的大量潜在输入，这种方式使用户可以在难以获得目标或没有用户目标的情况下为所欲为（分配给人类合作伙伴的最大权限）并提高系统安全性。

1.2.3　人机序贯决策的共享控制研究现状

面向人机序贯决策，不同于介入控制，共享控制 [63-65] 中人和机器的决策地位相对平等，更多描述的是人机决策动作的融合或者混合，共享控制涉及人类操作者和机器代理共享被控对象的决策权，目的是利用各自的优势取得单独人或机器难以取得的整体决策效果，对应“人在环内”(HITL，Human-In-The-Loop)。此外，与人在环上相比，人在环内由于扩大了决策行动空间，显然这对于系统优化是有利的，但是与此同时也增大了优化求解的困难 [66]，以及需要权威有效的高于人和机器的第三方机构—仲裁机制进行处理，其基于某种评价办法对人机决策动作进行分析判断以得到仲裁参数，该参数直接决定于最终待执行决策动作的生成 [67-68]，图1.4展示了人机共享手术决策权限，并且二者是缺一不可的，通过人和机器的完美配合实现手术过程的成功。

图 1.4　人和机器共享手术决策权限

关于共享控制，目前已有一定的研究基础，如文献 [69] 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，后者则更关注任务和情境的共享。[70] 进一步给出了共享控制设计的三个公理，分别地，第一公理强调人与机器之间的沟通和理解，第二公理关乎设计的安全性和性能，第三公理则关心机器的性能边界。相关模型的建立，如基于部分可观马氏决策过程（Partially Observable Markov Decision Process, POMDP）的讨论对共享控制系统进行建模是一类可行的方法。文献 [71] 将共享控制问题建模为带有不确定性的 POMDP 模型，使用最大熵逆优化方法估计用户目标的分布，并使用事后优化方法求解建模的 POMDP 问题。文献 [72] 使用 POMDP 集成人类模型，机器动力学模型及其相互作用进而形成统一的框架。

模型预测控制（Model Predictive Control, MPC）也被用于进行人机系统建模，并且它具备 MDP 缺少的优势：可以将硬约束显示表示在优化问题的定义中。比如文献 [72] 基于 MPC 考虑了人类意图估计和安全约束，该方法将倾向于规划

人类动作高不确定性情况下的智能代理的动作，能够帮助智能代理辨别正确的隐藏意图，同时如果人的意图变得更加确定，则仍将重点放在完成任务上。再者，强化学习作为决策问题的有力工具，为人机混合智能系统提供了建模办法如[73]。其他一些研究则关心人机系统的一般性框架。文献 [74] 以自动驾驶为应用背景，确定了共享控制中的七个原则，文献 [75] 则在更为广泛的包括介入控制、共享控制、监督控制等范畴内试图确定统一的研究框架。一些优秀的综述文献也值得关注。比如 [76] 对人与机器人物理交互中的共享控制进行了较为有价值的叙述，主要包括三个方面：意图检测，仲裁和沟通/反馈。文献 [77] 总结了人机交互学习的分类：从人类示范/干预/评价中学习，这对设计人机系统控制策略有一定程度的参考意义。

综上已有研究基础，考虑使用混合智能的办法求解人机序贯决策问题，但相较于传统人机系统，新型人机混合智能使得人和机器都出现在了决策层面，有了较大改善，并且目前关于人在环上和人在环内两种新型人机混合智能的研究尚处于起步和发展阶段，仍然存在一些问题，比如人在环上的框架往往会比较依赖某一方，不恰当的介入发生时可能会引发系统运行模态的切变，造成系统性能的恶化。如何处理介入的恰当时机，并且提升人机系统性能是值得继续研究的课题。此外，与人在环上相比，人在环内由于扩大了决策行动空间，显然这对于系统优化是有利的，但是与此同时也增大了优化求解的困难，以及需要权威有效的高于人和机器的第三方机构的仲裁机制进行处理。

1.3　本文研究内容

本节针对上一节混合智能中的介入控制和共享控制求解人机序贯决策所面临问题的研究现状分别进行分析，总结现有研究存在的难点，并阐述本文的研究内容。

1.3.1　现有研究的难点总结

本文主要关注以*人机序贯决策*为被控对象的混合智能系统，总结现有研究的难点如下：

- **人机决策权限：**人机混合智能控制中人和机器的决策权限如何划分，换言之，能否使得人类智能和机器智能具有清晰的决策围，它直接影响人机混合的方式和程度。权限，对任何多于一个决策者的系统都是或实或虚地存在，当然在我们所研究的人机混合智能中，同样希望获得决策权限的衡量标准。有了决策权限，即有了对决策动作的约束，并且我们认为符合此约束的决策动作所构成的人的动作空间和机器的动作空间，均是对被控对象

的优化求解是有益的。在介入控制中，如果能有决策权限作为判断标准，则关于上述何时触发介入发生则是有帮助的。在共享控制中，决策权限的使用也有利于人机决策动作的融合。目前国内外关于决策权限的研究少之又少，本文希望能弥补此点空缺，考虑如何利用已有的感知到的环境状态信息、学习到的决策网络，人类动作的意图推理，被控对象的任务目标等，对混合智能中人机决策权限进行数学上的形式化表达，进而获得动态变化的决策权限，并且将该决策权限应用到最优解的优化过程中，从而改善文中所提序贯决策的性能，这也是本文拟解决的第一个关键问题。

- **介入控制的触发**：如何触发介入的发生，以及触发介入的程度大小等都是急需解决的难点。介入控制涉及三种情况：机器介入人；人介入机器；人和机器的相互介入，称之为切换。1）机器介入人的情况，意味着被控对象的控制过程是由人类操控着的，当人类决策出现错误或者极大不确定性认知时，机器主动接管被控对象的控制过程。2）人介入机器的情况，意味着被控对象的控制过程是由机器操控着的，当机器决策的质量极差或者可信度不高时，人类伙伴主动接管被控对象的控制过程。3）切换，名为双向介入，显然地，需要对人和机器二者的决策进行评判，并且由决策质量高者向决策质量低者干预的过程。上述三种情况，由于需要不定时的进行介入、切换，那么评判何时触发介入、切换则显得尤为重要，并且由于人和机器毕竟属于两种不同属性的控制主体，频繁的介入是否会导致被控对象控制过程的不稳定也是需要考虑的。为此，如何利用感知到的环境状态信息，学习到的决策网络，决策质量的评判方法，公平合理的切换机制等模块，生成有效决策动作，提升被控对象处于介入控制下的决策性能，是本文拟解决的第二个关键问题。
- **共享控制的融合**：共享控制的融合程度如何确定，如何获得融合后的最优决策，以及融合后的决策是否确实好于单一控制者的决策等都是共享控制中的难点。不同于介入控制，共享控制需要将人类动作和机器动作进行混合，意味着系统需要对人类决策动作和机器决策动作进行分析评判，并且根据评判结果对混合程度进行调整。如当机器决策质量不高时，混合结果将偏向人类决策，当人类出现明显错误时，混合结果将偏向机器决策。极端情况下，共享控制将退化为上述的介入控制，同时，这也意味着共享控制所面临的行动空间要比介入控制大得多，显然这对优化求得最优解是有利的，但是也增加了计算复杂度问题。再者，在人类和机器决策地位不平等的情况下，如何保证某一方的最终接受度，比如在远程遥控进行人机协作搜救过程中，机器不了解搜救过程中变化的搜救目标，此时就需要机器代理从人类输入动作中分析出时变的任务目标后再进行决策的输出。在这

个远程搜救任务中，人的决策地位明显高于机器，那么如何将人的最终决策权和人的决策质量进行折衷，也是值得考量的问题之一。为此，面向扩大了行动空间的共享控制，如何根据感知到的环境状态信息，学习到的决策网络，合理有效的仲裁学习机制，甚至关乎人类意图推理的研究，生成有效决策动作，提升被控对象处于共享控制下的决策性能，是本文拟解决的第三个关键问题。

在求解人机序贯决策问题时，人类智能和机器智能各自的决策权限与决策能力是动态平衡的，直接关乎混合增强智能的程度和水平；第二和第三个难点所述是求解人机序贯决策的两种混合增强智能方法，即介入控制和共享控制，分别对应介入触发和共享混合两个关键词；以上三个难点的解决，使得在权责更加清晰的情况下实现对人机序贯决策问题的求解方法的优化设计。

1.3.2　本文研究内容

综合考虑上述已有研究工作存在的问题和难点，本文考虑利用人机混合智能框架解决序贯决策任务。本文的主要研究内容分为如下四个部分：

- **基于强化学习的混合智能算法求解人机序贯决策问题**：利用人机混合智能算法求解人机序贯决策问题，最终待执行动作直接受到人机各自状态、人的决策动作和机器代理的决策动作之间的集成方式和决策质量评价机制的影响。因此设计合理的人机混合智能方案来推动人机序贯决策问题的有效求解，以及避免因人机协调不成功而出现的一加一小于一的控制效果退化现象。本文考虑基于强化学习的人机混合智能框架，并且通过仲裁机制从人类决策动作和机器代理决策动作之间选取更优的待执行动作，其中涉及以可信性和安全性为代表的评价指标。此外，该部分研究内容考虑将基于模型的强化学习决策子系统和基于无模型的强化学习决策子系统共同作为机器代理子系统，与人类决策子系统进行合作和竞争，为求解人机序贯决策问题提供了一种具有基础性和一般性的混合智能控制算法。
- **基于自主性边界优化设计面向人机序贯决策的介入控制算法**：面向需要介入控制算法求解的人机序贯决策问题，为了使得人和机器代理的决策权限更加合理且清晰，本文对自主性 (决策行动范围) 和自主性边界 (决策行动边界) 进行讨论判定。通过将人和机器代理的自主性边界的求解形式化为常见的优化问题，使得自主性边界也成为动态决策过程中需要优化的目标之一。将获得的自主性边界信息应用于介入控制的仲裁判断中，以实现对介入控制算法进行改进提升的效果。根据介入的方向不同，分别分为人介入机器控制部分和机器介入人控制部分，这两者在现实生产生活中都是经常出现的场景。通常介入者比被介入者的优先级高，当被介入者出现决策

失误或者对自身决策不自信时，介入者适时进行干预的同时决策权发生转变，为处理紧急突发状况提供了更多的保障。

- **基于自主性边界优化设计面向人机序贯决策的共享控制算法**：不同于介入控制中的决策权“切换”的概念，共享控制是一种更倾向于“混合”概念的求解办法。面向需要共享控制算法求解的人机序贯决策问题，提出了基于自主性边界的共享控制优化算法。共享控制使得人机序贯决策的最优解出现在人的决策动作空间和机器代理的决策动作空间所共同张成的扩展空间上，这对于优化问题的求解无疑是有利且有弊的，其中有利体现为动作空间的扩大为更优提供了可能，有弊体现为动作空间的扩大为求解过程带来了难度。本文针对以往共享控制中判断混合参数方式简单粗暴的缺陷，利用自主边界的新概念设计更合适的混合参数，通过自适应调节共享控制中的混合参数进而直接决定最终待执行动作的生成。
- **基于自主性边界的不确定性优化设计面向人机序贯决策的混合智能控制算法**：为了减少单一估值的自主性边界存在不准确影响仲裁判断的情况，本文以概率分布的形式衡量自主性边界的不确定性，并且将此不确定性信息应用于人混合智能控制算法 (包括介入控制和共享控制) 的优化设计中，此种方法是对人机决策权限更完善的考虑，也更符合人们模糊性边界的思路。其中，本文采用贝叶斯神经网络对不确定进行判定，并且基于 dropout 机制对贝叶斯神经网络的近似估计，以采样输出和蒙特卡洛估计的办法获得自主性边界的概率分布形式。获得的具有不确定性的自主性边界信息，通过依概率采样间接影响介入控制和共享控制的决策效果，深入且丰富了基于自主性边界优化求解面向人机序贯决策问题的混合智能控制算法。

1.4　论文组织结构

全文总共六章，章节安排的组织结构如图1.5所示，具体描述如下:

第 2 章相关基础知识，主要介绍本文涉及的基本概念和基础知识。包括序贯决策与人机序贯决策介绍、人机混合智能系统设计、贝叶斯神经网络和强化学习。

第 3 章研究了基于强化学习方法的混合智能算法解决人机序贯决策。讨论了如何将人的决策与机器决策相结合以获得更高质量的决策动作。具体涉及基于模型和无模型的子系统、仲裁机制等，为后续的优化设计方案提供基础性方案，以及算法的仿真验证。

第 4 章研究面向人机序贯决策的介入控制优化问题，包括人介入机器控制和机器介入人控制。具体涉及自主性边界判定方法、利用自主性边界信息对介入

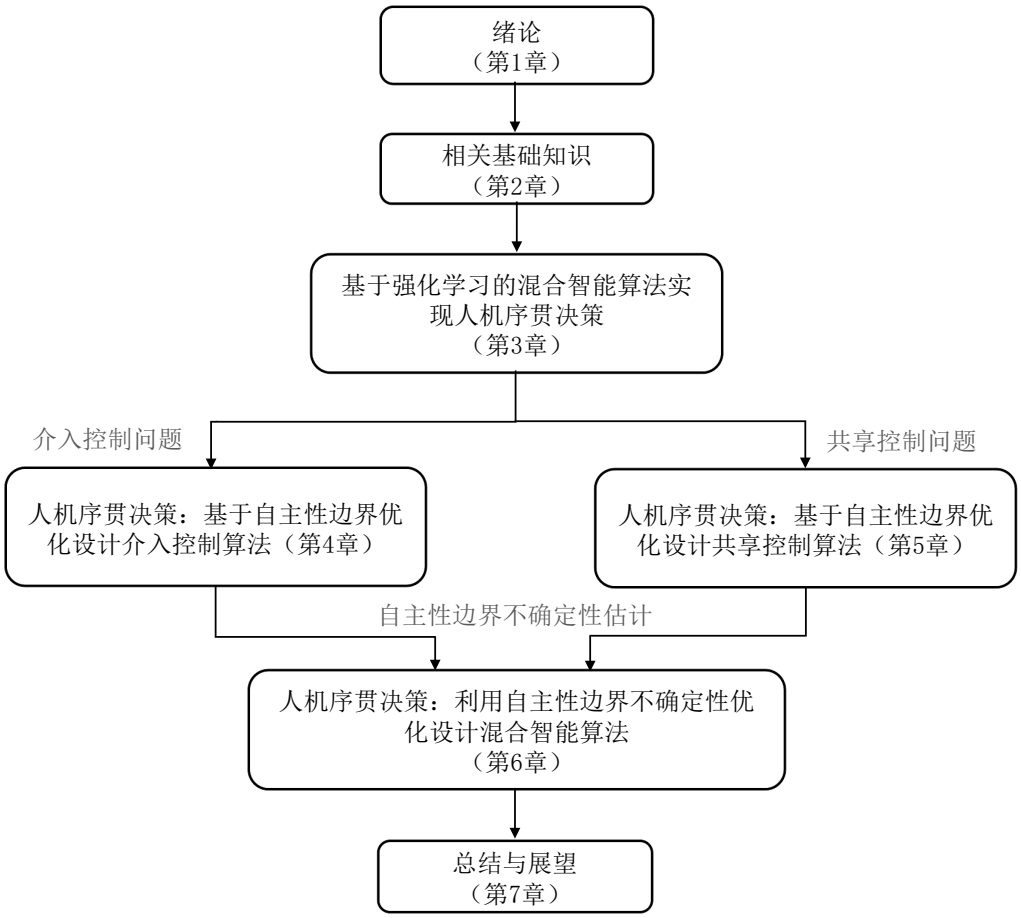


图 1.5　本文组织结构图

控制的优化设计，以及算法的仿真验证。

第 5 章研究了面向人机序贯决策的共享控制优化问题，提出了共享控制下的自主性边界判定方法，以及结合自主性边界和共享控制的研究基础，对人机共享控制中的仲裁进行优化，实现改善提升序贯决策的控制性能，以及算法的仿真验证。

第 6 章研究了面向人机序贯决策，利用自主性边界的不确定性优化设计人机混合智能控制算法的问题。基于贝叶斯神经网络和蒙特卡洛估计，将自主性边界的不确定信息应用于优化设计人机混合智能算法，以及算法的仿真验证。

最后，第 7 章总结全文的内容和贡献，并简述本文可能存在的不足以及未来研究工作的展望。

第 2 章　相关基础知识

本章主要介绍所涉及相关概念和基础知识，包括序贯决策与人机序贯决策、人机混合智能系统、贝叶斯神经网络和强化学习等。

2.1　序贯决策与人机序贯决策

2.1.1　序贯决策基本概念

序贯决策是一类常见的具有时序和多阶段特点的决策问题，其一般框架可形象化表示为图2.1。在任意决策时刻 $t \in T := \{t_0, t_1, \cdots\}$，系统观察到当前时刻 t 所处的状态 $s(t) \in S := \{s(1), s(2), \cdots\}$，按照策略 p 确定并执行动作 $a(t) = p(s(t))$，$a(t) \in A := \{a(1), a(2), \cdots\}$，随后系统依概率 $P\{s(t+1)|s(t), a(t)\}$ 按照系统的动力学模型 f 进入到下一状态 $s(t+1) = f(s(t), a(t))$，并获得一个奖励或惩罚 $r(t)$。

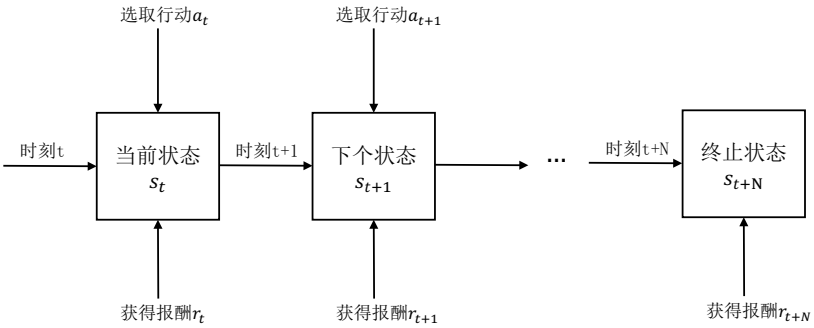


图 2.1　序贯决策的状态转移示意图

与常规决策问题中策略 p 的决定仅依赖当前时刻的某一指标 $J(s(t), a(t), r(t))$ 不同，在序贯决策中，策略 p 的决定需要优化未来一段时间 $[t, t+t^+]$ 内的某一指标 $\vec{J}(\vec{s}(t), \vec{a}(t), \vec{r}(t))$, 其中 $\vec{s}(t) = \{s(t), \cdots, s(t+t^+)\}$, $\vec{a}(t)$ 和 $\vec{r}(t)$ 可类似定义。

在序贯决策问题中，指标 $\vec{J}$ 包含了未来时刻的系统状态、动作和奖惩，使得求解空间随之指数增加（例如，求解的状态空间成为 S^{t^+}），这成为了序贯决策问题求解中的本质困难所在 [5-6]。

序贯决策理论已普遍应用于众多实际工程领域，如口语识别 [78]，飞行器学习战术决策问题 [79]，气候突变的威胁 [80]，机器人的行为控制 [81]，等。将序贯决策在一些典型领域如快递分拣、服务型机器人、飞行器控制和辅助驾驶等的应用形象化展示在图2.2。相信随着人工智能、机器学习等相关领域的进一步发展，序贯决策相关的理论和应用研究也将得到充分有效的发展。

图 2.2　序列决策的理论和方法可广泛应用于分拣机器人、服务机器人、无人机和自动驾驶汽车等领域，用于解决路径规划等技术问题。

2.1.2　人机序贯决策基本介绍

如第1章所述，本文关心的人与机器共同参与的序贯决策问题，我们称之为“人机序贯决策问题”，其包含两种场景:“人参与问题”的人机序贯决策和“人介入方法”的人机序贯决策。

将“人参与问题”的人机序贯决策控制框架绘制为图2.3所示。其中人类伙伴1 具有无可替代的能力和最终决策的权限，人类伙伴 2 则与机器代理以较为平等的决策地位协同产生最终决策。此种情况是容易理解的如高科技轮椅，微创外科手术辅助系统等，以达芬奇外科手术系统为例，其可以通过充分发挥外科专家的人类医学专家的经验和机器的精密操作能力进行高效的遥操作微创外科手术。任何手术过程都是一个包含了多个阶段或者多个步骤的，为了达到手术成功这

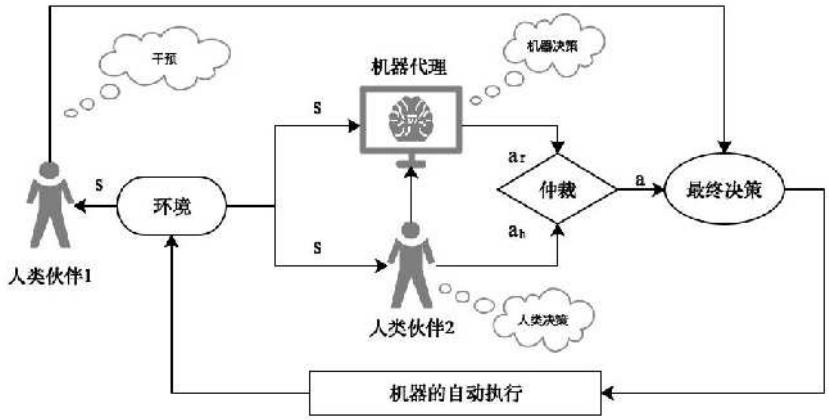


图 2.3　“人参与问题”的人机序贯决策问题示意图

个最终的好结果，需要从整体优化角度出发考虑每个小步骤的手术动作，因此从本质上来说其即是一个待解决的序贯决策问题，加之手术过程由人类专家和机器辅助系统共同操作完成，使得该问题成为本文所关心的“人参与问题”的人机序贯决策问题。

将“人介入方法”的人机序贯决策问题求解绘制为图2.4。相较于“人参与问题”的人机序贯决策，此框架中人的因素仅体现在求解方法中，并不固有的属于问题本身，因此人和机器的决策地位相对平等，以类似于“并联”的方式相辅相成，从而产生更优的决策动作。除了能够应用于在搜救任务中，此种场景可扩展至众多领域，比如自动驾驶系统，医学诊断等。

无论是上述的“人参与问题”的人机序贯决策，还是“人介入方法”的人机序贯决策，在当前人工智能所处的发展阶段均是不可忽视甚至必须研究的问题，并且以越来越普遍和常见的形式存在于现实生活中。更具体地，“人参与问题”和“人介入方法”最终会以表面化人机共存的形式出现，如何针对人机共同参与的序贯决策问题求解则成为讨论人机序贯决策的重中之重。

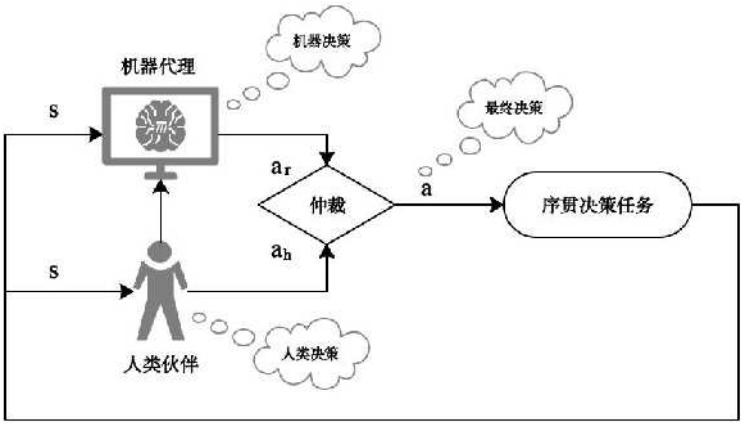


图 2.4　“人介入方法”的人机序贯决策问题示意图

2.1.3　相关求解领域

序贯决策的优化求解和以下领域密不可分，包括：

- **动态规划（Dynamic Programming, DP）**：DP[82] 既是数学优化方法又是计算机编程方法。该方法是由理查德 · 贝尔曼（Richard Bellman）开发并且已在众多领域中得到应用 [83]。在这两种情况下，它都是指通过以递归的方式将其分解为更容易处理的子问题来降低原问题的复杂度。尽管这种方式不是直接用来解决某些决策问题，但涉及多个时间点的递归求解对于问题的进一步求解事非常有利的。同样，在计算机科学中，若可以通过将原待求解的问题分解为子问题然后重复递归地寻找到子问题的最优解进而最

佳地解决原问题，可将此种问题具有最优子结构属性。若子问题可以递归嵌套于较大较复杂原问题内，则动态规划方法是非常好的求解思路，并且其中原复杂问题和简化子问题之间满足 Bellman 方程的关系。理论上，动态规划法所适用的问题需要满足两个性质：

- 最优子结构（Optimal Substructure）：整个大而复杂的原问题最优解可以通过求解简化的子问题得到，即通过子问题的最优解构造出全局最优解。
- 重叠子问题（Overlapping Subproblems）：子问题多次重复出现，子问题的求解结果可以储存下来并再次使用 [84]。

- **马尔可夫决策过程（Markov Decision Process, MDP，Partially Observable Markov Decision Process，POMDP）**：MDP/POMDP[85-86] 可用于建模序贯决策问题，具体可体现在系统状态具有马尔可夫性质的环境中模拟随机性策略与回报 [84]。MDP 是一种离散时间随机控制过程，是建模结果部分随机并且决策者能够操控情况下一个数学框架，其对于研究通过动态规划解决的优化问题很有意义，至少早在 1950 年代就已为人所知 [85]。马尔可夫决策过程的核心研究是罗纳德 · 霍华德 (Ronald Howard) 于 1960 年出版的《Dynamic Programming and Markov Processes》[87]。它们被用于诸多领域，包括机器人技术，自动控制，经济学和制造。MDP 的名称来自俄罗斯数学家 Andrey Markov，因为它们是 Markov 链的扩展。在每个时间步 t 所对应的状态 $s(t)$，决策者能够选择状态 $s(t)$ 情况下的任何可能性决策动作 $a(t)$, 从而随机转移到进入新状态 $s(t+1)$，并给予相应的奖励 $r(t)$。此外 MDP 还存在一些变体形式，比如部分可观察马尔可夫决策过程（POMDP）、约束马尔可夫决策过程（Constrained MDP, CMDP）和模糊马尔可夫决策过程（Fuzzy MDP, FMDP）[84] 等。
- **模型预测控制（Model Predictive Control, MPC）**：MPC[88-90] 指系统当前控制动作是在每一个采样瞬间通过求解一个有限时域开环最优控制问题而获得的特殊控制形式。在控制过程重，当前时刻状态作为最优控制问题的初始状态，解得的最优控制序列只实施第一个控制动作，此点是其区别于预先计算控制律算法之处。更具体地，模型预测控制是以控制过程的动态模型为基础的一个开环最优控制问题。多半是透过系统识别得到的线性经验模型。模型预测控制的特点是每一次针对目前的时间区块内作最佳化，然后下一个时间再针对时间区块内作最佳化，这和线性二次型调节器（Linear Quadratic Regulator, LQR）[91] 不同。模型预测控制方法能够预测未来轨迹并且进行对应的决策处理。其基本组成要素包括：动态预测模型、反馈校正、滚动优化、参考轨迹。具有代表性的模型预测控制算法主要包

括：模型算法控制，动态矩阵控制，广义预测控制。此外模型预测控制方法具有：1) 建模方便；2) 采用了非最小化描述的离散卷积和模型，信息冗余量大，有利于提高系统的鲁棒性；3) 采用了滚动优化策略，即在线反复进行优化计算，滚动实施，使模型失配、畸变、扰动等引起的不确定性及时得到弥补，从而得到较好的动态控制性能等明显的优点 [88, 90]。

- **强化学习（Reinforcement Learning, RL）**：强化学习是智能体以“试错”的方式不断学习策略，通过与外界环境进行交互获得实时奖赏，过程中所学习的目标是使得智能体获得最大的奖赏值。不同于监督学习，RL 主要表现在强化信号上，并且环境提供的强化信号是对产生动作的质量作一种评价（通常为标量信号），而非告诉智能体如何去产生正确或者质量更高的决策动作。由于外部环境提供的信息量较少，系统需要靠自身的经历（经验轨迹）进行学习，在行动-评价的环境中获得有用信息，进而改进策略函数以更好的适应环境。另外，强化学习是机器学习的一个分支，即监督学习、非监督学习、强化学习是三种基本的机器学习范式。与监督学习的不同之处在于，强化学习不需要输入带标签的数据样本，也无需显式纠正次优决策动作。相反地，强化学习的重点是在探索（explore）和利用（exploit）之间找到某种平衡 [92]。强化学习求解的环境模型通常以马尔可夫决策过程 (MDP) 的形式描述，针对此种情况的诸多强化学习算法均基于动态规划 [93]。主要区别在于，强化学习不对 MDP 的精确数学模型知识进行假定，能够适用于众多无法采用精确模型的决策问题。

2.2　人机混合智能系统

2.2.1　基础概念

在《新一代人工智能发展规划》[94-95] 所部署的五个重要方向里，“人机协同的混合增强智能”赫然在列。人机混合智能从字面上来理解，即是混合人类智能和机器智能两种截然不同的智能形式。之所以需要对此研究方向进行探讨，源于：人工智能没有因为其在搜索、计算、存储和优化领域的高效优势，而完全替代掉人脑的高级认知功能，例如感知、推理等方面。再者，机器的感知和推理从根源上说依然是由人类所进行设计的。因此对于人类的天然智能，不仅脑神经科学领域的的研究破解尚且任重道远，人机协同的混合智能也大有前景。

中国工程院院士、中国自动化学会理事长郑南宁曾谈及，人工智能作为一种可以引领多个学科领域、有望产生颠覆性变革的技术手段，具有较大的价值创造和竞争优势。然而，人类社会还有诸多脆弱、动态、开放的复杂问题，使得人工智能尚且束手无策。从这一点讲，任何智能机器可能都没办法去完全替代人

类[①]。郑南宁院士将混合智能的存在形态分为两种基本的实现形式，即“人在回路的混合增强智能”和“基于认知计算的混合增强智能”[96]。

“人在回路的混合增强智能”涉及将人的作用引入到智能机器系统中，形成人在回路的混合智能范式。在此种范式中人始终是智能系统决策的一部分，当系统中智能机器输出决策的置信度较低时，人可以通过主动介入的方式调整决策参数从而给出合理正确的问题求解，进而构成改善提升系统决策水平的反馈回路。把人的作用引入到智能系统的决策回路中，能够把人对模糊、不确定问题分析与响应的高级认知机制与机器智能系统紧密耦合，使得人类智能和机器智能这两者之间能够相互适应、协同工作，从而达成双向信息交流与决策的效应。换句话说，即使得人特有的感知、认知能力与机器强大的运算和存储能力相融合，构成相互促进的增强智能形态 [97-98]。而“基于认知计算的混合增强智能”则是指在人工决策的系统中引入受生物启发的智能机器模型，构建基于机器认知计算的混合增强智能。“此类混合智能通过模仿生物大脑功能提升机器的感知、推理和决策能力，以达到更准确地建立像人脑一样感知、推理和响应激励的机器计算模型，尤其是建立因果导向模型、直觉推理感应和联想记忆的新机器计算框架[96]。本文的研究重点主要集中于前者—“人在回路的混合增强智能”。

决策科学是一门横跨自然科学、社会科学以及人类思维科学的综合性大学科，是揭示决策本质，研究、探索和寻求作出正确决策的规律的科学 [99]。本文将人机融合决策智能系统看成是决策学的一项组成部分分析系统在决策体系中地位、作用、理论基础等问题：并从当前的研究成果中，从求解决策问题的角度融入人机结合的概念，这也与上述的“人在回路的混合增强智能”思路相对应。人机混合智能统共包含三个层面的意思：一是人机互动，目前已经做到的是语言交互，人机互动是人机混合智能的第一要点；二是人机协同，即机器和人实现协同工作；三是人机融合，除了语言互动，还要有情感互动，智慧交流，是简单协同的提升形式，进一步呈现人机混合增强智能的良好局面[②]。

在我们所考虑的人机系统中，机器具有了 AI 赋予的智能和自主性，从而在决策的层面上与人类展开了合作和竞争，这是传统人机系统不曾考虑的；另外，我们所考虑的机器和系统对象也较多面向复杂的自动化控制应用，这也是传统人机系统较少考虑的。为明确起见，广泛称此类系统为“人机混合智能系统”。

2.2.2　典型案例

本小节给出人机混合智能的系统的具体例子如下：

例 2.1 (遥操作微创外科手术)　如图2.5所示，微创外科手术（Minimally In-

[①]网站见：https://www.msra.cn/zh-cn/news/outreach-articles/caa-20170818

[②]网站见：https://www.chinait.com/ai/38839.html

vasive Surgery, MIS）是通过微小创伤面积或微小入路，将特殊器械、物理能量或化学药剂导入人体，达到对人体内病变、畸形、创伤的灭活、切除、修复或重建等操作目的，从而实现使用科技手段降低病人的创伤面的外科手术。在此类手术过程中，期望达到的效果是：专家具有手术的先验知识的同时，直接接触病人进行手术的机械臂同时扮演执行器和控制器的角色，能够将执行动作缩减到足够小的误差范围内，做到专家医生肉眼可见手动执刀所达不到的效果，进而使手术成功率更高。为了实现上述目的，设计基于人机混合控制的机器人外科微创手术辅助系统，将专家医生置于闭环控制回路中，手术过程中的任何决策同时考虑专家和机器，会对此类外科微创手术的进一步提升作出重大贡献。目前已面世的达芬奇外科手术系统 [13] 是此类辅助系统的代表之一。

图 2.5　微创外科手术

例 2.2 (辅助驾驶系统 [100-102])　虽然现实中也出现了无人自动驾驶汽车，但它们还远远没有普及，主要原因在于无法在任何天气和任何外界条件下保证自动驾驶的安全等级。因此自动驾驶的道路尚且漫长。人开车是个自然而然的事情，但偶尔会出现疲劳驾驶以及判断失误的状况。针对上述两种情况，人机共驾成为一种发展趋势。人机共驾系统结合人类的特殊认知能力和机器的数据分析能力，当机器的判断决策可信度不高时，警示人类驾驶员及时干预；当人类驾驶员注意力不集中或出现明显操作失误时，系统允许机器的智能介入，如图2.6所示。

例 2.3 (智能可穿戴设备)　智能手环 [103-104]，如图2.7所示，霍金的轮椅，和用于辅助有困难的人行走的各类康复机器人 [105-106] 都属可穿戴设备。现存设备在智能化和人的参与上仍有较大的发展空间，比如人发出指令，智能家居产品自动执行指令，或者是智能家居产品监测环境变化，自动执行某些操作。这两种情况下，人都并不在智能家居系统的闭环回路中。再者，尽管如智能手环中人处于回路中，却也是处于被监测的角色，并未在真正的控制层面有所作用。

作为一种解决方案，考虑人机混合智能系统设计，其中的混合更多体现在决

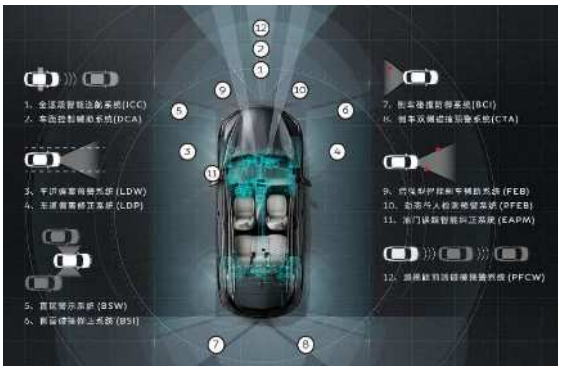


图 2.6　辅助驾驶系统

策层面，使得机器最终按照人的意愿、状态、以及控制运行。

图 2.7　智能可穿戴设备

2.3　强化学习

基于上述对人机混合智能系统和序贯决策的介绍，本节介绍强化学习的基本原理、基本算法、以及深度强化学习。

2.3.1　强化学习基本原理

类似于人类学习新鲜事物的过程，如婴儿学习走路，如果摔倒，大脑会得出一个负面的信号，以说明走路姿势的不对。从摔倒状态中爬起来后，如果能正常走了一步，大脑会输出一个正面的奖励信号，使得当事人知道当前的走路姿势是一个正确的行为。强化学习 [84, 92, 107] 考虑的是智能体与环境的交互问题。智能体处在一个环境中，每个状态为智能体对当前环境的感知。智能体只能通过动作来影响环境，当智能体执行一个动作后，会使得环境按某种概率转移到另一个状态；同时，环境会根据潜在的奖赏函数反馈给智能体一个奖赏。强化学习的基本结构如图2.8所示。

强化学习一般包含最基础的构成要素：智能体、环境、状态、动作、策略、奖励。

- 智能体：学习器与决策者的角色；

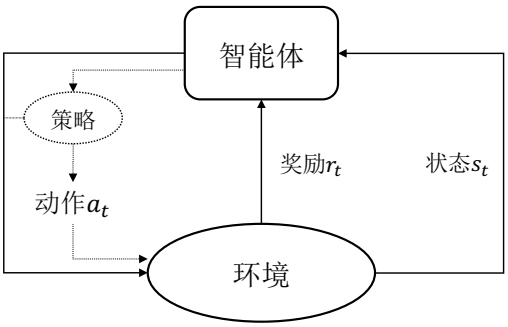


图 2.8　强化学习结构示意图

- 环境：智能体以外的、且与智能体进行交互的事物；
- 状态 $s(t)$：智能体从环境获取的信息；
- 动作 $a(t)$：智能体的动作表示；
- 奖励 $r(t)$：环境对于动作的反馈；
- 策略 $\pi(\cdot)$：指智能体根据实时状态 $s(t)$ 所选择动作的函数，常见的表达形式是一个条件概率分布 $\pi(a(t)|s(t))$，即在状态 $s(t)$ 时采取动作 $a(t)$ 的概率，此时概率大的动作被选择的概率较高；
- 环境的状态转化模型：指在状态 $s(t)$ 执行动作 $a(t)$ 后转移到下一状态 $s(t+1)$ 的概率模型；
- 状态值函数 $v(s(t))$：通常是一个期望函数形式，此函数需综合考虑当前的奖励和后续的延时奖励。尽管不同的算法可能会有不同的价值函数变种形式，但它们的思路大致相同，可以表示为下式：

$$v(s(t)) = E_{\pi}(r(t+1) + \gamma r(t+2) + \gamma^2 r(t+3) + \cdots | s(t)) \tag{2.1}$$

其中 γ 是奖励衰减因子，在 $[0, 1]$ 之间。与状态值函数对应的是动作值函数 $q(s(t), a(t))$，用来评估智能体在状态 $s(t)$ 采取行动 $a(t)$ 的优劣程度，考虑了在时刻 t 采取行动 a 对环境状态的影响。

$$q(s(t), a(t)) = E_{\pi}(r(t+1) + \gamma r(t+2) + \gamma^2 r(t+3) + ... | s(t), a(t)) \tag{2.2}$$

2.3.2　强化学习基本算法

以状态价值函数 $v(s(t))$ 为例，强化学习的两个基本问题是：策略评估和策略迭代。预测是根据上述强化学习的基本组成要素，求解某策略所对应的价值函数

$$v(s(t)) = \sum_{a} \pi(a(t)|s(t))(r(t) + \gamma v(s(t+1))) \tag{2.3}$$

控制则是根据“预测”步骤所衡量的策略优劣程度，求解最优的价值函数 v^* 和策略 π^*。

$$v^*(s(t)) = \max_{a} \sum_{s(t+1)} \pi(a(t)|s(t))(r(t) + \gamma v^*(s(t+1))) \tag{2.4}$$

式(2.4)也是通常所说的贝尔曼方程。

为了求解上述预测和控制问题，有以下基本方法：

- 动态规划法：基于小节2.1.3，首先使用的就是动态规划法，主要原因在于：由马尔可夫决策过程的马尔科夫性所决定的，贝尔曼方程可以递归地切分子问题，并通过对子问题的局部最优解而获得较复杂原问题的全局最优解，因此非常适合采用动态规划法来求解贝尔曼方程。在环境模型已知的前提下，对于任意的策略 $\pi(\cdot)$，需要合理估算该策略所带来的累积奖励期望以及准确衡量该策略的优劣程度。
- 蒙特卡洛法（Monte Carlo learning, MC）[108-109]：与动态规划比，蒙特卡洛法不需要依赖于模型状态转化概率。通过对环境进行“模拟-采样-估值”来进行求解，之所以能够无模型求解的原因在于其无需依赖外界环境的完备知识，仅需从环境中进行采样获得完整的经验轨迹样本，之后利用采集到的经验轨迹数据集实现求解最优策略的目的。每一条经验轨迹的序列必须是达到终点了的即是所谓的经历完整。以下棋问题为例则是指最后能够分出输赢，再如驾车问题中能够成功或者失败的到达终点，即结果是明确的，而非半路中止的。获得了多组如此经历完整的状态序列后，就可以通过近似估计状态价值从而求解预测和控制等问题。
- 时序差分法（Temporal-Difference learning, TD）[110-111]：与蒙特卡洛法类似，同样是基于无模型的求解方法，为了提升学习效率，时序差分法采用动态规划法中自举的方式计算当前价值函数，即利用 TD 目标值近似代替原完整经历目标值的过程。对于时序差分法来说，智能体每走一步都可以更新一次值函数，不需要等到到达终点之后才进行更新。根据选择动作的策略和更新价值函数的策略是否为同一个，时序差分法分为在线控制和离线控制，代表算法分别是 SARSA[112] 和 Q-Learning[113]，分别如算法2.1和算法2.2所示。

2.3.3　强化学习进阶式算法

本节对一些进阶式强化学习算法进行介绍。

1. 值函数估计法

在上述第2.3.2节强化学习基本算法中，状态值函数或动作值函数被保存在一张表中，这种做法的优点是，简单而且可以保存所有可能的值，缺点是如果状态集合比较大时，需要很大的空间存储值函数，且对每个状态进行学习耗时较长。因此需要找到近似函数，如线性组合、神经网络等其他方法实现此目的，该做法即称为值函数估计法（Value Function Approximation，VFA）[84, 114]。

算法 2.1　SARSA(on-policy TD control))

初始化参数: 步长 α，$\epsilon > 0$；$Q(s,a)$，其中 $a \in S^+, a \in \mathcal{A}(s)$，终止状态 $Q(terminal,\cdot) = 0$；；

for 遍历完整 *episode* **do**

  初始状态 S;

  根据策略由 $Q(e.g.,\epsilon - greedy)$ 选择 S 对应的动作 A;

  for 遍历 *episode* 每一时间步 **do**

    执行动作 A，观察 R, S';

    根据策略由 Q(e.g., $\epsilon - greedy$) 选择 S' 对应的动作 A';

    $Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma Q(S',A') - A(S,A)]$;

    $S \leftarrow S'; A \leftarrow A'$;

  end

end

算法 2.2　Q-learning(off-policy TD control))

初始化参数: 步长 α，$\epsilon > 0$；$Q(s,a)$，其中 $a \in S^+, a \in \mathcal{A}(s)$，终止状态 $Q(terminal,\cdot) = 0$;

for 遍历完整 *episode* **do**

  初始化状态 S;

  根据策略由 $Q(e.g.,\epsilon - greedy)$ 选择 S 对应的动作 A;

  for 遍历 *episode* 每一时间步 **do**

    执行动作 A，观察 R, S';

    $Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma \max Q(S',a) - A(S,A)]$;

    $S \leftarrow S'$;

  end

end

$$\hat{v}(s,\omega) \approx v(s) \tag{2.5}$$

$$\hat{q}(s,a,\omega) \approx q(s,a) \tag{2.6}$$

关于上式，可以利用监督学习里的数据拟合方法去估计值函数，蒙特卡洛法或时序差分法可以被用来更新参数 ω。从而实现用参数 ω 来拟合实际的各种价值函数。有了值函数，随后可以直接从价值函数中产生策略，比如使用 ε-greedy。

2. 策略梯度法

策略梯度法（Policy Gradient，PG）[84, 115] 直接参数化策略本身，同时参数化的策略将不再是一个概率集合而是一个函数:

$$\pi_\theta(s,a) = P(a,|s,\theta) \tag{2.7}$$

上式是参数化的策略函数形式 π_θ。策略函数能够在给定的状态和一定参数设置下，确定采取所有可能动作的概率，因此更具体的说，它其实是一个概率密度函数。在实际采用策略函数输出动作时，即是按照这个概率分布来进行动作采样的。策略函数中所涉及的参数决定了概率分布的形态。策略函数的参数化目的是为了适应大规模的求解问题。实现对策略函数的参数化后，可通过调整这些参数来获取一个较优策略形式，后续只要遵循这个策略，那么产生的动作将会得到较多的奖励。如更详细的实施办法是，设计一个目标函数，并对其使用梯度上升 (Gradient Ascent) 算法来对参数进行优化达到最大化奖励的目的。

3. 深度强化学习

深度强化学习（Deep Reinforcement Learning, DRL）[116-117] 通过将强化学习和深度学习进行深度结合，充分利用强化学习的决策优势和深度学习的感知优势。深度强化学习的出现使得强化学习方法逐步走向实用，使其能够解决现实场景中的一些复杂决策问题。深度 Q 网络（Deep Q-learning, DQN）是基于价值函数的深度强化学习的典型代表，其由 Mnih 等人于 2013 年提出，并且在 2015 年对其进行了改进。此方法采用卷积神经网络拟合价值函数，一般是 Q 函数。决策网络的输入是当前状态（如系统所处的外界环境数据），如游戏连续帧的画面图像，输出即是在当前状态下执行可能动作时所得到 Q 函数的极大值。

DQN 算法 [117] 的一般步骤如算法2.3所示。

4. 多智能体强化学习

在多智能体系统（Multi-Agent System, MAS）[118] 中，多个智能体通过不断与环境进行交互并获取奖励值来持续学习改善自身策略，从而获得该环境下对应的最优策略的过程就称为多智能体强化学习 [119-120]。不同于单智能体强化学习中外界环境的相对稳定性，多智能体强化学习所处的环境是复杂的、动态的，这会给每个智能体的策略学习过程带来较大的困难。

在多智能体系统中，各个智能体之间可能涉及合作与竞争等存在关系，引入博弈 [121-122] 的概念，将博弈论与强化学习相结合可以很好的处理这些问题。首先是纳什均衡。

定义 2.1 (纳什均衡)　指一个所有智能体的联结策略。在纳什均衡（Nash Equilibrium）[123-124] 处，对于所有智能体而言都不能在仅改变自身策略的情况下，来获得更大的奖励。

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 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 训练未终止 **do**
　初始状态 $s(t)$，状态向量 $\phi(t)=\phi(s(t))$;
　while 完整 *episode* 未结束 **do**
　　以概率 ϵ（或 $1-\epsilon$）选取某随机动作 $a(t)=a_{\text{rand}}$，或 Q 值最大的动作 $a(t)=\max_{a(t)}Q^{*}(\phi(s(t)),a(t);\theta)$;
　　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 $\phi(t+1)=\phi(s(t+1))$;
　　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

$$y(j)=\begin{cases}r(j) & \phi(j+1)\text{为终止状态}\\ r(j)+\gamma\max_{a'(t)}\hat{Q}(\phi(j+1),a'(t);\theta^{-}) & \phi(j+1)\text{非终止状态}\end{cases}$$

　　损失函数 $(y(j)-Q(\phi(j),a(j);\theta))^{2}$ 进行梯度反向传播以更新评估网络 Q 参数 θ;
　　每 C 步更新目标网络 $\hat{Q}$ 参数 $\theta^{-}=\theta$;
　end
end

定义 2.2 (随机博弈)　随机博弈（Stochastic Game / Markov Game）[125-126]是马尔可夫决策过程与矩阵博弈的结合，具有多个智能体与多个状态，即多智能体强化学习。

多智能体强化学习即是一个随机博弈，其将每一个状态的阶段博弈的纳什策略进行组合起来，则成为一个智能体在动态环境中的策略，并且其通过不断与环境交互不断更新每一个状态的阶段博弈中的 Q 值函数（博弈奖励）。对于一个随机博弈可以写为 $(n,S,A_1,\cdots,A_n,Tr,\gamma,R_1,\cdots,R_n)$，其中 n 表示智能体数量，其中 S 表示状态空间，A_i 表示第 i 个智能体的动作空间，$Tr:S\times A_1\times\cdots,\times A_n\times S\rightarrow[0,1]$ 表示状态转移概率模型，$R_i:s\times A_1\times\cdots,\times A_n\times S\rightarrow\mathbb{R}$ 表示第 i 个智能体在当前状态与对应动作下所获得的奖赏值，γ 表示累积奖赏的折扣因子。随机博弈具有马尔科夫性，其下一状态和奖励只与当前状态和所对应的动作有关。对于多智能体强化学习过程而言，目标就是找到每一个状态的纳什均衡策略，然后将这些

策略进行联合。此外，根据每个智能体的奖励函数可以对随机博弈进行分类，如合作博弈和竞争博弈。

2.4　贝叶斯神经网络

2.4.1　基本形式

贝叶斯神经网络 [127-128]（Bayesian Neural Network, BNN）是一类神经网络模型，模型的参数不是固定的值，而是概率分布，如图2.9所示。简单来说可以将其理解为通过为神经网络的权重引入不确定性进行正则化，也相当于集成某权重分布上的无穷多组神经网络进行预测。

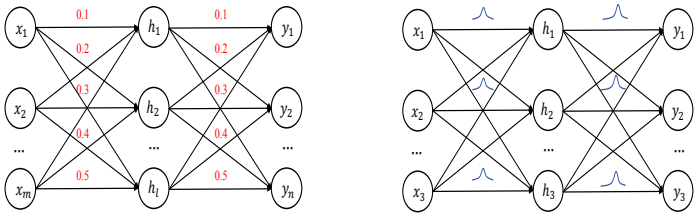


图 2.9　神经网络结构示意图：普通神经网络（左）和贝叶斯神经网络（右）

从最基本的贝叶斯公式出发，来理解一下贝叶斯神经网络的迭代过程：

$$p(W|X,Y) = \frac{p(W)p(Y|,X,W)}{p(Y|X)} \tag{2.8}$$

其中 (X,Y) 为训练样本集，$p(W)$ 为 W 的先验概率，$p(Y|,X,W)$ 为在给定参数 W 和输入 X 的情况下，神经网络输出 Y 的概率，$p(Y|X)$ 是由训练样本集确定的常数。

2.4.2　近似求解办法

由于式(2.8)中的 $p(W|X,Y)$ 的概率分布复杂，难以求得均值方差等，因此贝叶斯神经网络通过建立一个 q 函数来逼近，比如一个较为简单的高斯分布（参数 $\mu_\omega, \sigma_\omega$）等。判断 q 函数逼近 p 函数的效果一般依据选取 Kullback−Leibler (KL) 散度 [129-130]，KL 散度又称为两种分布的相对熵，从某种程度上讲，熵可以度量两个随机变量的距离。其数学定义如下。

$$KL(q||p) = \int_{i=1}^{n} q(W)\log\frac{q(W)}{p(W,X,Y)} = E_q[\log\frac{q(W)}{p(W,X,Y)}] \tag{2.9}$$

其中由于 q 是一个函数，KL 散度是关于 q 的期望，对这个 q 函数进行优化使得 KL 散度最小，这样的对于函数的操作称之为“变分”。

变分推理 [131] 是一种近似推理方法，其目的是通过将贝叶斯推理过程中所需的边缘化形式化为一个优化问题。变分推理利用近似的后验分布进行优化以找到最接近的真实后验分布。此种近似简化了计算过程，并提供了一定程度的可操作性。

贝叶斯神经网络的重点是寻找合适的后验分布近似值，其预测输出和区间都是作为后验分布 $p(W|X,Y)$ 的期望值来计算的，但是精确的预测依赖于对难以处理的后验概率的精确近似。式(2.9)中的积分求解很困难，传统的解决方法是利用基于优化的方案，但优化方法中设置的约束会造成预测值不准确的情况，所以基于优化的方案经常会输出不准确的预测量。为了使得有限计算资源下的预测是准确的，马尔可夫链蒙特卡罗 [132]（Markov Chain Monte Carlo, MCMC）方法被考虑用来求解上积分。MCMC 是通过采样的方法对难以处理的概率分布进行处理，会有如下公式：

$$E_q[f] = \int f(W)q(W)dW = \sum_{i=1}^{N} f(W_i) \tag{2.10}$$

考虑到神经网络的大规模属性，其强大的推理能力通常依赖于大的数据集上。但该大数据集却使得对数似然的评估在训练目的上变得不可行。为了解决这一问题，Bottou 采用了随机梯度下降 [133]（SGD）方法，其是神经网络和贝叶斯神经网络等最常用的训练方法。如此可以采用小批量数据近似似然项，变分目标即为如下形式：

$$L(\omega,\theta) = -\frac{N}{M}\sum_{i=1}^{N} E_q[\log(p(X,Y|\omega))] + KL(q_\theta||p(\omega)) \tag{2.11}$$

更具体地，Blundell[134] 等人提出了一种 BNNs 下的近似推理办法，其利用重参数化的做法说明如何对期望导数进行无偏估计，期望导数的具体表达如下：

$$\begin{aligned}\frac{\partial}{\partial\theta}E_q[f(\omega,\theta)] &= \frac{\partial}{\partial\theta}\int q_\theta(\omega)f(\omega,\theta)d\omega \\ &= \frac{\partial}{\partial\theta}\int p(\epsilon)f(\omega,\theta)d\epsilon \\ &= E_{q(\epsilon)}[\frac{\partial f(\omega,\theta)}{\partial\omega}\frac{\partial\omega}{\partial\theta} + \frac{\partial f(\omega,\theta)}{\partial\theta}]\end{aligned} \tag{2.12}$$

在贝叶斯的反向传播的算法中，$f(\omega,\theta)$ 设为

$$f(\omega,\theta) = \log\frac{q_\theta(\theta)}{p(\omega)} - \log p(X|\omega) \tag{2.13}$$

其中 $f(\omega,\theta)$ 可以看作是期望值的自变量。Bayes by Backprop（BbB）的算法流程图如下所示：

算法 2.4　Bayes by Backprop(BbB)

while 未达到收敛条件 **do**
  初始化代价：$\mathcal{F} \leftarrow 0$
  for *i in* $[1,\cdots,N]$ **do**
    采样 $\epsilon_i \sim \mathcal{N}(0,1)$　　蒙特卡洛估计采样;
    $\omega \leftarrow \mu + \text{softplus}(\rho)\cdot\epsilon_i$;
    $\mathcal{L} \leftarrow \log q(\omega|\theta) - \log p(\omega) - \log p(X|\omega)$;
    $\mathcal{F}[q_\theta] += sum(\mathcal{L})/N$　　对权重集合 ω 的所有 log 求和;
  end
  $\theta \leftarrow \theta - \alpha\bigtriangledown_\theta \mathcal{F}[q_\theta]$　　更新参数;
end

2.5　本章小节

本章首先对序贯决策与人机序贯决策进行介绍，本文将“人参与问题”序贯决策和“人介入方法”序贯决策统称为人机序贯决策问题，以此作为本文的研究对象，本章相关求解的基础知识进行介绍。其次描述人机混合智能的基础概念，指人类智能和机器智能在决策层面上出现合作与竞争的混合智能形式，接着人机混合智能系统的典型案例被展开说明，并且这类系统已广泛出现日常工作生活中。第三，本章给出了本文验证实验的基础办法，即强化学习。最后介绍了用来衡量不确定性和可信性的贝叶斯神经网络相关知识，以便后文使用。总体来说，本章的目的是在后文方法描述以及仿真验证之前，对相关基础知识的引入性介绍。

第 3 章　基于强化学习的混合智能算法实现人机序贯决策

本章提出了一种基于强化学习方法的人机混合智能方法来解决序贯决策优化问题，涉及如何将人的决策与机器决策相结合以获得更高质量的决策动作。特别地，考虑将贝叶斯神经网络引入到仲裁框架中，以判断决策动作的不确定性以及安全性，它也是决策质量的评估方法之一。

3.1　引言

近些年，随着人工智能的迅速发展，使得机器具有自主决策的能力，进而此种机器智能不断普及到各个领域，尤其包括本身是由人类做的任务，并且由于当前机器智能发展水平的制约使得该部分任务尚未被机器智能替代。因此则会面临人和机器共同参与决策的情景。更具体地，序贯决策问题作为一类具有时序性和多阶段性的动态决策问题，其发展与当下人工智能时代下的工程应用、生产生活等领域息息相关。因此本节主要考虑人和机器共同参与决策的序贯决策问题，也即是上述第1章和第2章提到的人机序贯决策问题。

关于此类问题求解，所面临的难点包括：人机决策权限的分配、介入控制的触发，共享控制的融合等。在动态环境下，人和机器通过各自的决策“大脑”给出实时状态下的决策动作，并且其决策动作均是为了达到同一个任务目标。然而由于人类智能和机器智能的本质差别，认知能力以及决策质量的差别等情况，若是通过将人类决策和机器决策进行结合达到更好决策效果，则需要对其更加深入的研究和探讨。

目前关于人机混合智能的研究已有一定基础，比如传统人机系统，但传统人机系统解决人类控制和机器执行的自动化之间的交互问题，仅包含机器自动执行和人类控制两层架构，这已经不适应新技术的发展并且智能水平不足。[72] 曾考虑基于 POMDP 对人机序贯决策问题进行建模，考虑到人的状态难以直接被测量但其有着重要意义，比如辅助驾驶系统对疲劳驾驶员的强制接管，服务型机器人需要结合人类状态提供服务 [135-136]，等。文献 [137] 提出了一种可实现机器人与人类之间协作学习的新强化学习算法，其基于 $Q(\lambda)$ 方法的算法通过利用人类的智力和专业知识来加快学习过程。通过将人加入强化学习算法改善求解决策问题的效果。此外，[62, 72] 利用 MPC 对决策动作滚动优化，通过预测轨迹选择有限时间内最优的决策动作。但上述研究均存在一些问题，比如建模复杂，求解困难，不易泛化等。

综上所述，本章提出了基于强化学习方法的人机混合智能控制框架。通过将

机器代理的决策和人类的决策以可信性和安全性为评价指标进行仲裁选择，从而确定更优的待执行决策动作。同时考虑了基于模型的强化学习决策子系统和基于无模型的强化学习决策子系统，为适应广泛的序贯决策应用场景提供了更多可能。更具体地，本章以强化学习为基础方法，构造新型人机混合智能算法求解人机序贯决策问题。建立三个子系统互相竞争的决策模型，设计基于仲裁机制的人机混合智能算法，具体包括基于模型决策子系统引入对环境的建模实现学习和规划、无模型决策子系统源于从真实环境的交互轨迹样本中学习策略函数、人类动作的不定时参与决策、基于对未来轨迹进行讨论得到的安全约束判断、以及基于贝叶斯神经网络对两个机器子系统的决策评估可行性。最终在仲裁的框架下实现对模型子系统的决策动作、无模型子系统的决策动作、以及人类输入动作进行判断得出最终待执行决策动作。

本章结构安排如下，第3.2节给出人机序贯决策问题描述及建模求解形式。第3.3节介绍具体的人机混合智能框架设计，其中包括基于模型的控制子系统，基于无模型的控制子系统，仲裁机制，以及安全约束等组成。第3.4节给出实验设计和和结果分析，并且在第二个仿真实验中，首次讨论了自主性边界的学习对决策效果的影响，这些不仅是对本章人机混合智能算法的佐证，也为后文关于自主性边界的深入讨论提供了部分基础。

3.2　问题描述与建模

基于第1.1.1节和第1.1.2节对人机序贯决策的描述，为了弥补机器完全自主的缺陷和不足，考虑允许人类在训练过程中直接参与控制回路以改善传统序贯决策算法的性能，如图3.1所示。图中机器代理根据决策任务实时输出决策动作，人类伙伴适时参与到仲裁机制中。系统基于某种仲裁机制和评价指标，对机器动作和人类动作予以抉择，得到即将作用到被控对象的待执行决策动作。这些可能是许多实际工程系统应用的有效参考模式，尤其是实时动态决策场景。

事实上，一般的序贯决策过程可分成若干个互相联系的阶段，在它的每一阶段都需要作出决策，从而使整个过程达到最好的活动效果。因此各个阶段决策的选取不能任意确定，它依赖于当前面临的状态，又影响以后的发展。当各个阶段决策确定后，就组成一个决策序列，因而也就确定了整个过程的一条活动路线。这种把一个问题看作是一个前后关联具有链状结构的多阶段过程就称为多阶段决策过程，这种问题称为多阶段决策问题。在多阶段决策问题中，各个阶段采取的决策，一般来说是与时间有关的，决策依赖于当前状态，又随即引起状态的转移，一个决策序列就是在变化的状态中产生出来的，故有“动态”的含义，称这种解决多阶段决策最优化的过程为动态规划方法。

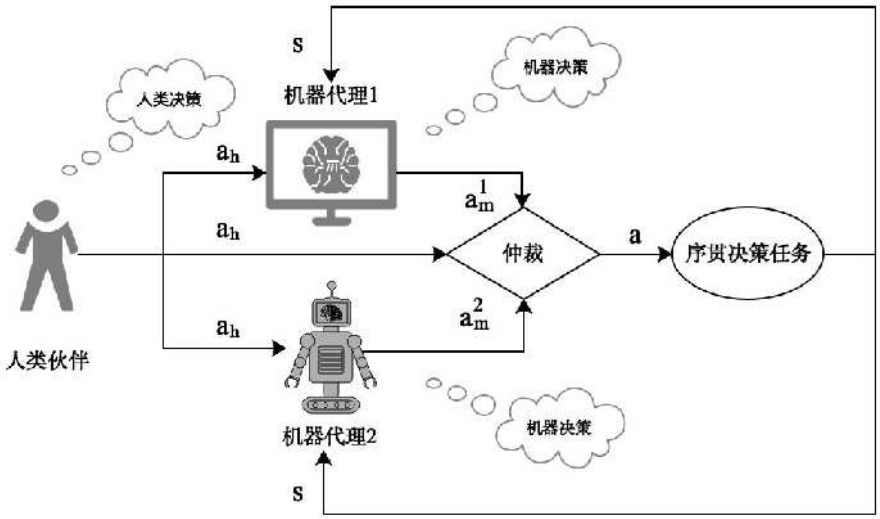


图 3.1　人机混合智能实现序贯决策的示意图

与常规序贯决策不同的是，人机序贯决策过程中的决策者不仅仅是人或者机器代理中的一方，而是涉及二者关于决策动作的选择和融合，如图3.1所示。下面基于动态规划方法，本章通过仲裁机制将人类伙伴的决策动作融入到机器智能代理决策中。具体形式化为如下优化问题：

$$\max_{\theta_1,\theta_2} J(s(t),a(t)) = \int r(s(t),a(t))dt \tag{3.1a}$$

$$\text{s.t. } a(t) = f^a(s(t), a_m^1(t), a_m^2(t), a_h(t)) \tag{3.1b}$$

$$a_m^1(t) = p_m^1(s(t);\theta_1) \tag{3.1c}$$

$$a_m^2(t) = p_m^2(s(t);\theta_2) \tag{3.1d}$$

$$a_h(t) = \text{Human-Action-Input} \tag{3.1e}$$

$$s(t+1) = f^d(s(t),a(t)) \tag{3.1f}$$

$$C(s(t), a_m^1(t), a_m^2(t), a_h(t)) \leqslant 0 \tag{3.1g}$$

$$t = 0, 1, 2, 3, ...$$

其中 $s(t), a(t)$ 分别表示系统在 t 时刻的状态和动作，$J(s(t), a(t))$ 表示时刻 t 对应的优化目标函数。$R(s(t), a(t))$ 是系统在 t 时刻执行动作所获得的奖励，$P(s(t+1)|s(t), a(t))$ 表示状态转移概率。$f^a(\cdot)$ 根据当前状态 $s(t)$ 仲裁出待执行动作 $a(t)$。$a_m^1(t), a_m^2(t), a_h(t)$ 分别表示机器代理 1、机器代理 2 以及人的动作。$p_m^1(\cdot)$, $p_m^2(\cdot)$ 表示机器代理的策略函数, 其网络参数分别为 θ_1 和 θ_2。$Q(s(t), a(t))$ 是状态动作对 $(s(t), a(t))$ 对应的值函数。$s(t), s(t+1)$ 分别描述了时间 t 和 $t+1$ 的状态，$C(\cdot)$ 表示约束状态和动作。

本章将人类伙伴置于控制回路内，以混合的方式使得机器代理和人类伙伴共享决策任务的控制权限。进而实现改善控制性能的目的。然而，如何以图3.1为指引，建立具体的人机混合智能框架成为值得继续思考的重点。再者，优化问题(3.1)中的仲裁机制(3.1b)如何确定，也是我们考虑的重点。通过对上述所列问题的凝练和解决，最终可利用人机混合智能框架实现在具体人机序贯决策场景的应用（根据被控对象的实时状态信息如位置、姿态等，框架输出更高决质量的决策动作，进而使得被控对象朝着任务目标的方向进行）。

3.3　人机混合智能框架设计

基于对第3.2节人机序贯决策问题的描述，本节将从混合智能控制的角度，基于强化学习方法，设计详细的控制框架解决上述问题。强化学习作为一类通过不断尝试从错误中学习经验到直至最终找到规律，学会如何更好决策的算法，对于本文的研究具有重大意义。首先对上述优化问题(3.1)中的目标函数进行改写：

$$\max_{\theta_1,\theta_2} Q(s(t),a(t)) = \max_{\theta_1,\theta_2}[R(s(t),a(t)) + \gamma \sum_{s(t+1)} P(s(t+1)|s(t),a(t))Q(s(t+1),a(t+1))] \tag{3.2}$$

基于强化学习方法，因此 $Q(s(t),a(t))$ 代替了 $J(s(t),a(t))$，这里的 $Q(\cdot)$ 表示动作值函数。上式即是强化学习中经常提到的贝尔曼方程，解决上述优化问题的过程就是优化贝尔曼方程的过程。并且如式(3.1)所示在人类决策动作和机器决策动作之间进行抉择。

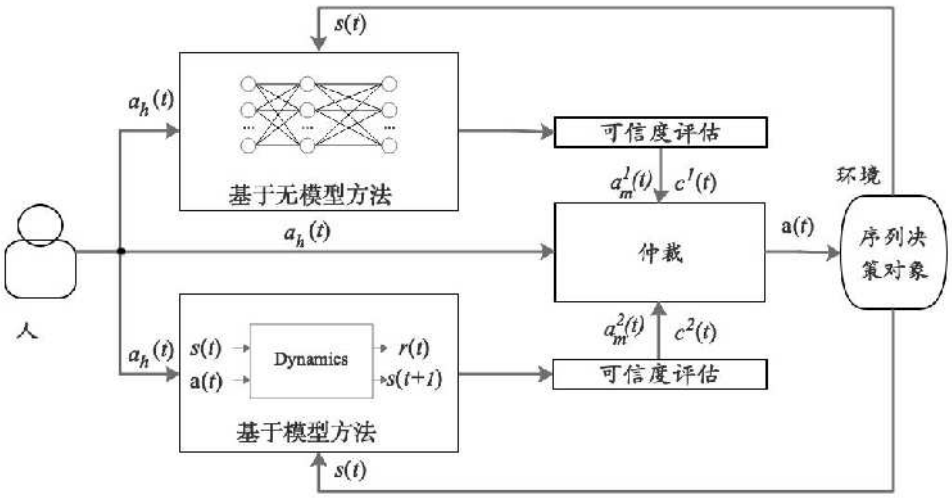


图 3.2　基于强化学习方法求解人机序贯决策的混合智能框架

如小节3.2中定义的优化问题(3.1)，我们将基于强化学习方法求解人机序贯

决策的混合智能框架描述如图3.2所示。该框架根据序贯决策任务的实时状态 $s(t)$，人、无模型强化学习决策子系统和 (或) 有模型强化学习决策子系统独立生成各自决策动作 $a_h(t), a_m(t)^1, a_m(t)^2$，及其决策可信度 $c_m(t)^1, c_m(t)^2$，这些决策动作进而在安全约束前提下输入到仲裁模块，最终由仲裁输出待执行决策动作 $a(t)$。也就是说，在整个人机序贯决策问题求解的动态过程中，决策者包括两类：人类伙伴和机器代理，其中机器代理包括两种方式，分别是基于无模型方法的决策网络和基于模型方法的决策网络。通过第三方评价机构 (仲裁) 从三者中选取质量更高的决策动作。

以下小节介绍图3.2框架中各模块的具体内容。

3.3.1　基于模型的控制子系统

基于模型的强化学习 [138-139] 主要引入了对环境的建模，这里提到的建模是指通过监督训练来得到一个环境模型，其数据来源是系统和环境的实际交互数据 $s(t), a(t), r(t), s(t+1), a(t+1), r(t+1), \cdots$。利用这个环境模型，可以根据当前 t 时刻 $s(t), a(t)$ 的情况预测到下一时刻 $t+1$ 的状态 $s(t+1)$。图3.3是基于模型的强化学习示意图，主要包括两部分：学习和规划。学习指从真实交互的经验轨迹数据集中学习环境模型，换句话说，学习表示环境的马尔可夫决策过程 $\boldsymbol{M} \sim \mathrm{MDP}\langle \boldsymbol{S}, \boldsymbol{A}, \boldsymbol{P}, \boldsymbol{R} \rangle$。规划是指根据给定的模型 $\boldsymbol{M} = \langle \boldsymbol{P}_\eta, \boldsymbol{R}_\eta \rangle$ 求解最优价值函数或最优策略，即马尔可夫过程 $\boldsymbol{M}_\eta \sim \mathrm{MDP}_\eta\langle \boldsymbol{S}, \boldsymbol{A}, \boldsymbol{P}_\eta, \boldsymbol{R}_\eta \rangle$ 求解模型。

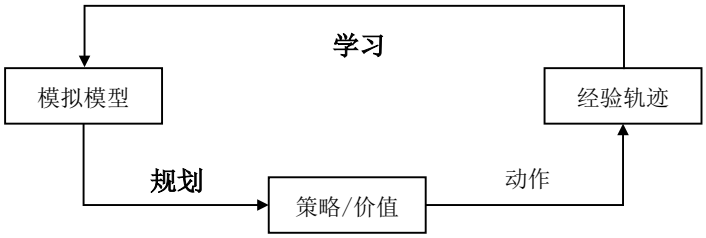


图 3.3　基于模型强化学习示意图

为了求解小节3.2的人机序贯决策问题(3.1)，建立基于模型的决策子系统如图3.4。对于上述的 $\boldsymbol{M}_\eta \sim \mathrm{MDP}_\eta\langle \boldsymbol{S}, \boldsymbol{A}, \boldsymbol{P}_\eta, \boldsymbol{R}_\eta \rangle$ 模型，其中 $\boldsymbol{P}_\eta$ 为环境模型从一个状态转移到另一个状态的概率，$\boldsymbol{R}_\eta$ 为智能体执行动作 $\boldsymbol{A}$ 后能够获得的对应奖励。并且使得：

$$s(t+1) \sim \boldsymbol{P}_\eta(s(t), a(t)) \tag{3.3}$$

$$r(t+1) \sim \boldsymbol{R}_\eta(r(t+1) | s(t), a(t)) \tag{3.4}$$

即指，对环境模型的估计，主要是对状态转移概率和奖励函数进行估计。可利用监督学习的方法对转移概率和奖励函数进行求解，即，为了求解转移概率，构建

训练数据样本如 $\langle s(i), a(i)\rangle \rightarrow s(i+1),\ i=0,1,2,\cdots$。为例求解奖励函数，构建训练数据样本如 $\langle s(i), a(i)\rangle \rightarrow r(i+1),\ i=0,1,2,\cdots$。对于规划过程而言，则是基于环境模型生成大量的模拟经验轨迹数据。之后采用策略梯度法、值函数近似法等，从生成的模拟经验轨迹中学习价值函数或策略函数。

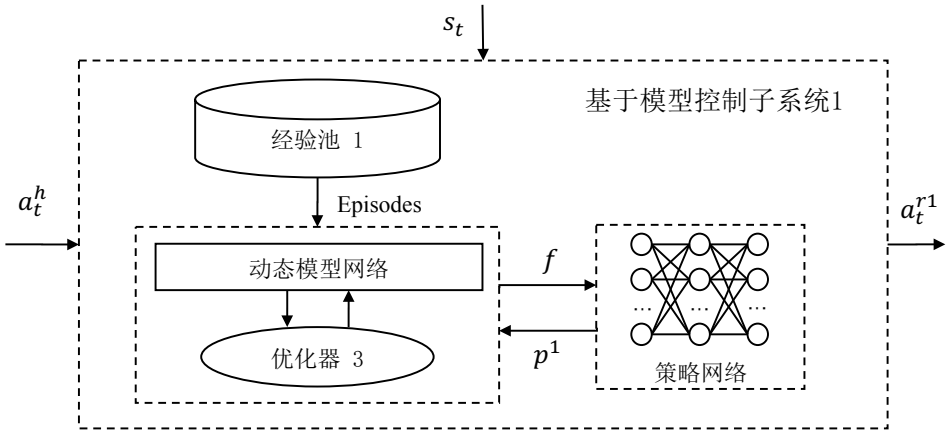


图 3.4　基于模型的决策子系统

算法 3.1　基于模型决策子系统控制算法

初始化: 经验样本池 D_1，动态网络模型 f 的参数，策略网络 p_m^1 的参数 θ_1;
while 训练未结束 **do**
　收集数据集 $\{X, Y\}$，并将数据集存入经验样本池 D_1 中。其中 X 是状态动作对，输出是下一步的状态和奖赏值;
　while 抓取次数未达到设定值 **do**
　　经验样本池抓取一批样本数据，计算动态网络模型的损失函数 $loss_f = \frac{1}{N}\sum_{n=1}^{N}\{\sum_{o=1}^{o_{dim}}\{\frac{1}{2}(\frac{x_{t+1}-\bar{x}_{t+1}}{x_{t+1}^{\sigma}})^2 + \log\sigma\} + M\}$;
　　根据上一步获取的损失函数大小，利用梯度下降法更新动态网络模型 f 的参数;
　　从经验样本池抓取一批样本数据，基于动态网络模型 f，预测经验轨迹 $T_r = \{S', A', R'\}$;
　　计算策略网络的损失函数 $loss_p = -\frac{1}{N_1}\sum_{step}\gamma R'$，并基于梯度上升法更新策略网络 p_m^1 的参数;
　end
end

基于上述的讨论，我们搭建基于模型的决策子系统如图3.4所示。其中经验

池用来存放规划阶段所采集的经验轨迹，策略网络是规划阶段所要训练学习用来生成决策动作的模块，动态模型网络是学习阶段所估计出的环境模型。

算法3.1给出基于模型决策子系统的具体执行步骤。在算法3.1中，首先基于随机初始化采集部分轨迹数据存于经验池，以及对动态模型网络和策略网络进行初始化。之后在动态演化过程中，收集系统实际产生的数据集 $\{\boldsymbol{X},\boldsymbol{Y}\}$，并存入经验池 $\boldsymbol{D}_1$ 中（步骤 3）。步骤 5 根据从经验池 $\boldsymbol{D}_1$ 中抓取一批数据样本，计算动态模型网络的损失函数大小，接着基于损失函数梯度下降的趋势更新动态网络模型（步骤 6）。类似地，策略网络的损失函数和更新如步骤 7 和步骤 8，使用更新后的动态网络模型 f，以及基于当前时刻的策略网络 p_m^1 来预测运动轨迹 $\boldsymbol{T}_r$ 以计算策略网络模型的损失函数，从而更新策略网络 p_m^1 的参数，重复上述过程直至训练结束。值得注意的是，此处的 p_m^1 即对应式(3.1)中的策略函数 p_m^1。

3.3.2　基于无模型决策子系统

基于差异化集成的思想，图3.2中的第二个决策子系统采用无模型强化学习的方法。对于第2.3节介绍的强化学习，有两类环境模型：一种是所处的环境是已知的，叫作基于模型 (Model-based)，上述第3.3.1节的控制子系统使用了基于模型的方法；另一种是所处环境是未知的叫作无模型 (Model-free)，源于现实情况下环境的状态转移概率和奖励函数往往难以提前获取，因此考虑无模型的任务求解也是异常重要的。第2.3.3节的 DQN 算法首次实现了强化学习与深度学习的有机结合，但在实际应用中也存在一些限制，如算法模型容易过度估计、无法处理连续动作控制任务等。本节采用具有代表性的深度确定性策略梯度算法 (Deep Deterministic Policy Gradient Algorithm, DDPG)[140]。DDPG 算法是由确定性策略梯度算法 (Deterministic Policy Gradient, DPG)[141] 发展而来。相较于 DPG 算法，DDPG 算法有如下改进：

- 使用深度神经网络（Deep Neural Network, DNN）作为函数近似。利用卷积神经网络近似策略函数和动作值函数；
- 引入经验回收机制。使得策略（演员）网络与环境交互时产生的状态转移样本数据具有时序相关性；
- 利用双网络结构，不论是策略函数还是价值函数都利用双网络结构，即演员网络，目标演员网络，评论网络，目标评论网络。此种做法使得算法学习过程更加稳定，收敛更快。

我们建立无模型决策子系统如图3.5所示。图中所示双网络即是 DDPG 算法网络架构。有关特定过程见算法3.2。在算法3.2的起始，首先随机初始化经验样本池 $\boldsymbol{D}_2$，演员网络 π_θ 和评论家网络 Q_ω 的参数，以及相应目标网络 $\pi'_{\theta'}$，$Q'_{\omega'}$ 的参数。在系统的动态演化过程中，不断采集新的经验轨迹存入经验样本池 $\boldsymbol{D}_2$。

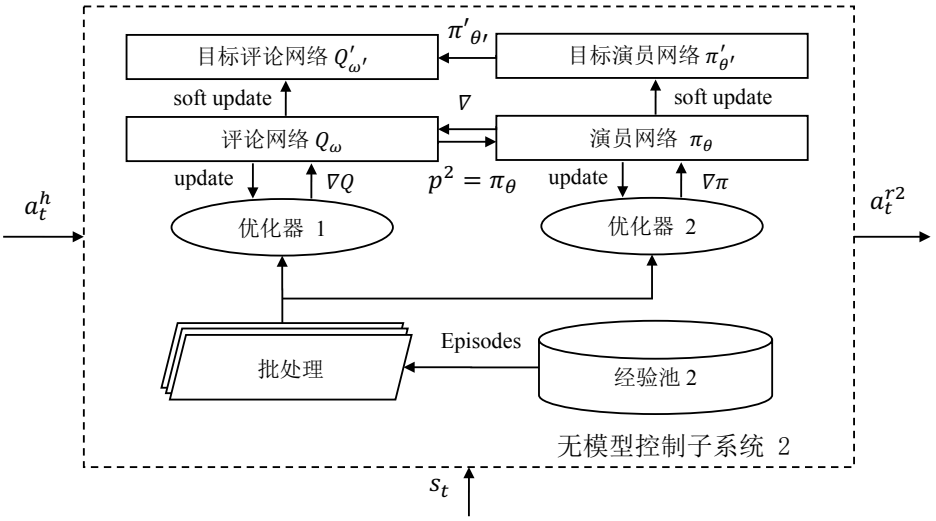


图 3.5　基于无模型决策子系统

算法 3.2　基于无模型决策子系统控制算法

初始化: 经验样本池 D_2，演员网络 π_θ、评论网络 Q_ω，以及相应的目标演员网络 $\pi'_{\theta'}$、目标评论网络 $Q'_{\omega'}$ 的参数 θ_2;

while 训练未结束 **do**

　while 未达到软更新周期 **do**

　　从经验样本池 D_2 随机抓取一批数据 $\{s_j, a_j, r_j, s_{j+1}, \cdots\}$，并且计算相应的动作值函数 $y_i = r_i + \gamma Q'_{\omega'}(s_{t+1}, \pi'_{\theta'}(s_{t+1}))$;

　　计算损失函数 $L_Q = \frac{1}{N_2}\sum_i (Q_\omega(s(t), a(t)) - y_i)^2$，并通过最小化损失函数更新评论网络 Q_ω;

　　利用梯度算法更新演员网络 π_θ：$L_\pi = -\frac{1}{N_2}\sum_i Q(s(t), \pi_\theta(s(t)))$;

　end

　对目标演员网络和目标评论网络进行软更新:

$$\pi'_{\theta'} \leftarrow \tau\pi_\theta + (1-\tau)\pi'_{\theta'} \tag{3.5}$$

$$Q'_{\omega'} \leftarrow \tau Q_\omega + (1-\tau)Q'_{\omega'} \tag{3.6}$$

end

接着，步骤 4 从经验池中抓取一批数据，用来计算相应的动作值函数。获得的值函数被用于更新评论网络 Q_ω(步骤 5)。步骤 6 使用梯度算法更新演员网络 π_θ。最后，基于软更新的方式更新相应的目标演员网络和目标评论网络 (步骤 8)，如此循环下去直至训练结束。值得注意的是，此处的 π_θ 即对应(3.1)中的策略函数 p_m^2。

3.3.3　安全约束

不可否认的是，在人工智能领域，数学科学领域的诸多工作集中于算法的规模和复杂度上，而“AI 安全性”[142-143] 也是时常困扰研究者的问题。强化学习智能体需要不断探索自身所处的环境学习新规范，从而达到最佳理想行为，即他们会在反复的试验中判断行为是良性还是恶性，之后尽量朝着增加良性行为减少恶性行为的方向前进。换言之，强化学习走的是一条“失败是成功之母”的道路。尽管名言道理无差，但事实上有些错误是不能尝试的，比如不能用反复的撞车撞人等交通事故来实现自动驾驶。因此，安全性探索就成为值得思考和研究的重要方向。

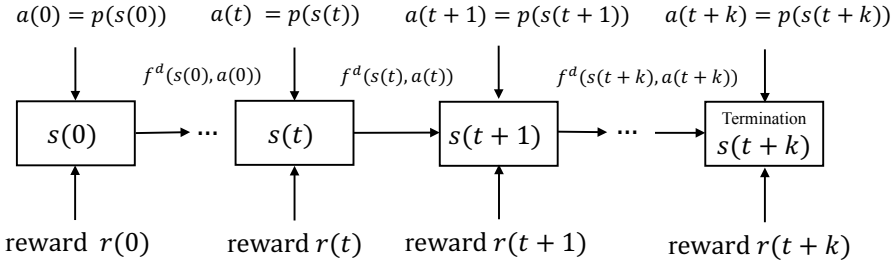


图 3.6　状态转移示意图

强化学习智能体的训练目标是最大化奖励信号，人类必须事先指定设计这一指标。假如奖励信号未被正确设计，智能体就会去学习计划外的甚至是有害的动作。如果设计奖励函数很容易，这将不是问题，但不幸的是从根本上来讲，奖励函数设计很有挑战性，这就是采用约束的关键动机所在。举一个简单例子，移动机器人在路径规划过程中，应该处于一个安全区域内活动，但是由于自身策略的不完善或者干扰信号的闯入使得机器人偶尔出现在安全区域以外，那么如果它离开安全区域的频率小于一定的预选阈值，我们认为机器人是安全的，反之则认为是不安全的。

$$f^c(s(t),a(t))=\begin{cases}1, a(t)=p(s(t))\ cause\ (s(t),\ldots,s(t+K),...)\\ 0, a(t)=p(s(t))\ cause\ (s(t),\ldots,s(t+K-1))\end{cases} \tag{3.7}$$

从更安全的角度出发，本节过滤了不符合安全约束的机器代理或人类伙伴

的决策动作。除了以往控制系统中对系统状态和决策动作有严格的硬约束以外，本节提到的安全约束还包括使用动态模型和策略网络评估系统状态和决策动作对未来状态轨迹的影响。

算法 3.3　安全约束预测算法

初始化: 给定状态 $s(t)$，决策动作 $a(t)$ 信息，以及策略网络 p；保存当前 t 时刻的系统状态信息 $s(t)$，并且给基于当前时刻的状态信息，进一步执行输入的决策动作 $a(t)$，得到 $t+1$ 时刻的状态信息 $s(t+1)$，$r(t+1)$ 等信息;

输出: 安全信号 $safe$;

while $i <$ 最大步数限制 K **do**

　状态 $s(t+i)$ 满足终止条件;

　基于策略网络 p 计算时刻 $t+1$ 的决策动作 $a(t+i) = p(s(t+i))$;

　执行决策动作 $a(t+i)$，并记录执行结果 $s(t+i+1)$，$r(t+i)$ 等信息;

　根据未来轨迹步数和累积奖赏值判断得出安全信号 $safe$;

end

如图3.6和算法3.3所示，$s(t)$：状态，f 表示获得的动态模型，p 表示与该模型相关的策略网络。策略网络根据当前状态 $a(t) = p(s(t))$ 获得相应的动作，学习的动态网络模型用来估计下一时刻 $t+1$ 的状态 $s(t+1) = f(s(t), a(t))$。此后，策略网络继续计算对应于 $s(t+1)$ 的动作 $a(t+1)$，重复上述步骤，直至达到最大预测步数，之后根据未来轨迹步数和累积奖赏值判断得出安全信号 $f^c(s(t), a(t))$。

3.3.4　仲裁机制

仲裁 [63, 68, 144] 是人机混合智能系统中经常出现的模块，本节的决策者包括三个决策子系统：基于模型决策子系统、基于无模型决策子系统、人类伙伴。注意图3.2中的仲裁。它决定何时做出由无模型决策子系统或基于模型决策子系统提供的决策建议，以及何时采用人类伙伴的建议。之后，仲裁模块输出最终的决策动作，此决策动作被智能体执行于序贯决策环境中。

基于第3.3.3节中的安全约束条件判断标准，以及可信度评估基础，根据被控对象的当前状态，仲裁给出相应的作用值，具体判断方法如公式(3.8)所示。

$$a(t)=\begin{cases} a_m^1(t),\ \left\{f^c(s(t),a_m^1(t))==1\ \&\ f^c(s(t),a_m^2(t))==0\right\}\ | \\ \qquad \left\{f^c(s(t),a_m^1(t))==f^c(s(t),a_m^2(t))==1)\ \&\ (c_m^1(t)>c_m^2(t))\right\} \\ a_m^2(t),\ \left\{f^c(s(t),a_m^1(t))==0\ \&\ f^c(s(t),a_m^2(t))==1\right\}\ | \\ \qquad \left\{f^c(s(t),a_m^1(t))==f^c(s(t),a_m^2(t))==1)\ \&\ (c_m^1(t)<c_m^2(t))\right\} \\ a^h(t),\ \left\{f^c(s(t),a_m^1(t))==f^c(s(t),a_m^2(t))==0\right\} \end{cases} \tag{3.8}$$

其中，$a_m^1(t)$ 和 $a_m^2(t)$ 分别代表基于模型决策子系统和基于无模型决策子系统在时间 t 所输出的决策动作，以及相应的决策可信度评估 $c_m^1(t)$ 和 $c_m^2(t)$。$f^c(\cdot)$ 表示安全约束判断，上述第3.3.3节已经作了介绍，这里不再赘述。考虑到贝叶斯神经网络的概率特性，使用 MC dropout [145] 方法来衡量子系统决策的可信度。公式(3.9)描述了 $c_m^1(t)$ 的计算过程（$c_m^2(t)$ 可类似计算）。

$$\mathbb{E}[a_m^1(t)]\approx\frac{1}{T}\sum_{t=1}^{T}p_m^1(s(t)) \tag{3.9a}$$

$$\mathbb{E}[(a_m^1(t))^T(a_m^1(t))]\approx\tau^{-1}I+\frac{1}{T}\sum_{t=1}^{T}p_m^1(s(t))^Tp_m^1(s(t))$$

$$c_m^1(t)=Var[a_m^1(t)]=\mathbb{E}[(a_m^1(t))^T(a_m^1(t))]-\mathbb{E}[a_m^1(t)]^T\mathbb{E}[a_m^1(t)] \tag{3.9b}$$

其中 $a_m^1(t)=p_m^1(s(t))$ 是模型的预测输出。文献 [145] 使用带有 T 的蒙特卡罗积分来获得预测分布的一阶矩和二阶矩。τ 涉及模型精度，

$$\tau=\frac{(1-p_d)l^2}{2N\lambda}$$

其中 p_d 是 dropout 概率，l 是用户定义的长度比例，N 是数据集 D 的数量，λ 是衰减因子。

具体的实现过程如算法3.4所示，基于强化学习方法以及利用仲裁机制，将上述两种不同决策子系统的决策动作和人的决策动作进行了整合。算法的目的是基于任意状态 $s(t)$，系统能结合两个机器代理和人类伙伴的决策动作，仲裁选择出最合适的待执行决策动作。首先，如算法3.1和算法3.2中进行初始化。在系统的动态演化过程中，对于任意时刻的状态 $s(t)$，基于两决策子系统分别计算 M 组决策动作 $\{a_m^1(t)^{(i)},\ a_m^2(t)^{(i)},\cdots,a_m^1(t)^{(M)},\ a_m^2(t)^{(M)}\}$，以及采集人类伙伴的决策动作 $a_h(t)$。之后基于 MC Dropout 的思想计算两组机器决策动作的均值和可信度。步骤 11 利用收集好的信息进行仲裁判断，从而获得最终的决策动作，并且执行

算法 3.4　基于强化学习的混合智能算法 (Hybrid Intelligent based on Reinforcement Learning, HIRL)

初始化: 如算法3.1和算法3.2中的随机初始化步骤;
输出: 决策动作 $a(t)$;
while 未到达最大训练时间 **do**
  while 未到达终止状态 **do**
    给定系统状态 $s(t)$;
    for $t < M$ **do**
      for $i < T$ **do**
        基于两控制子系统分别计算 $\boldsymbol{M}$ 组决策动作 $\{a_m^1(t)^{(i)},\ a_m^2(t)^{(i)}, \cdots, a_m^1(t)^{(M)},\ a_m^2(t)^{(M)}\}$，以及搜集人的决策动作 $a_h(t)$;
      end
      根据公式(3.9a)，(3.9b)计算均值 $a_m^1(t)$，$a_m^2(t)$ 和不确定性 $c_m^1(t)$，$c_m^2(t)$;
      基于仲裁机制(3.8)获得 $a(t)$，并且在时刻 t 的系统状态 $s(t)$，智能体执行动作 $a(t)$，将转移样本数据 $(s(t), a(t), r(t), s(t+1))$ 存入经验回收池 $\boldsymbol{D}_2$，同时暂时将 $(s(t), a(t), r(t), done)$ 保存到 $data$ 中;
    end
    将由 $data$ 构成的一条完整的运行轨迹存入经验回收池 $\boldsymbol{D}_1$;
  end
  利用经验池 $\boldsymbol{D}_2$ 中数据训练和更新基于无模型策略网络;
  利用经验池 $\boldsymbol{D}_1$ 中数据训练和更新基于模型的策略网络;
end

该决策动作，存储转移样本数据信息到 $\boldsymbol{D}_2$ 中，以及存储时序相关的一组完整运行轨迹到经验池 $\boldsymbol{D}_1$ 中。最后分别基于 $\boldsymbol{D}_2$ 和 $\boldsymbol{D}_1$ 训练更新基于无模型策略网络和基于模型策略网络。如此循环下去直至训练结束。

3.4　仿真实验

为了展示本章第3.3节所提出的混合智能系统设计的有效性，本节通过两个实验仿真进行验证。

3.4.1　仿真实验一：CartPole

本节选择倒立摆 CartPole[①]作为仿真环境，该环境目的是试图通过左右移动小车向上摆动立杆，以达到杆在竖直向上并且处于动态平衡的状态。系统需要评估小推车位置，杆的平衡和摆动情况。此处的 CartPole 是具有连续状态，连续动作的离散时间任务。具体地，系统状态 $s(t)$ 包括小推车位置，摆动角度及其对应的导数如 $s(t)=[x(t),\theta(t),\dot{x}(t),\dot{\theta}(t)]$。所使用的重要参数包括杆的长度 $l=0.6m$，小推车质量 $m_c=0.5kg$，杆的质量 $m_p=0.5kg$，时间范围 $T=2.5s$，时间离散化 $\delta t=0.1s$，重力加速度 $g=9.82m/s^2$。此外，摩擦力以阻尼系数 $b=0.1Ns/m$ 阻止小车移动。

如混合智能框架3.2中所示，基于模型决策子系统中策略网络和动态模型都是基于 MLP 训练获得的。并且，将 DDPG 用作图3.2中无模型决策子系统的算法基础。根据文献 [146]，Dropout 神经网络可以作为贝叶斯神经网络的近似值。因此，为了获得不确定的泛化能力更好的模型，我们在 DDPG 中使用 0.1 的 dropout 概率。此外，上一节3.3.1中讨论的动态模型如图3.7所示，它决定基于模型强化学习决策子系统的输出结果。

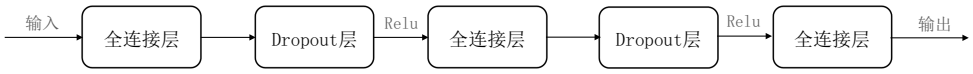


图 3.7　动态模型网络

1. 人对决策网络的影响

为了展示人机共享控制在序贯决策中的作用，这里以基于模型决策子系统为例展示人类伙伴的参与对基于模型决策网络的影响。

图3.8显示了动态模型训练的损失曲线，包括无人参与的算法 MO(Machine-Only)，有人参与的算法 HMC1(Human-Machine-Cooperation-1)：人类伙伴协助收集初始训练样本集；算法 HMC2(Human-Machine-Cooperation-2)：人类伙伴参与

[①]见：https://gym.openai.com 和 https://github.com/openai/gym

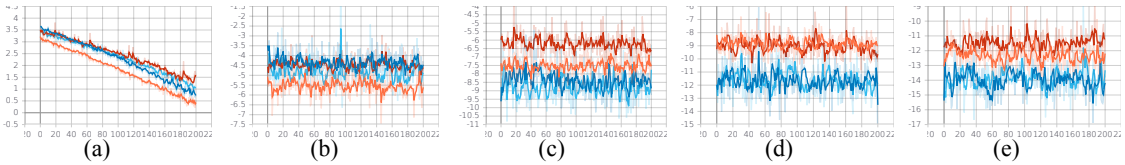


图 3.8　动态模型的损失走势：橙色曲线表示算法 MO，蓝色曲线表示算法 HMC1，红色曲线表示算法 HMC2，浅蓝色曲线表示算法 HMC3.

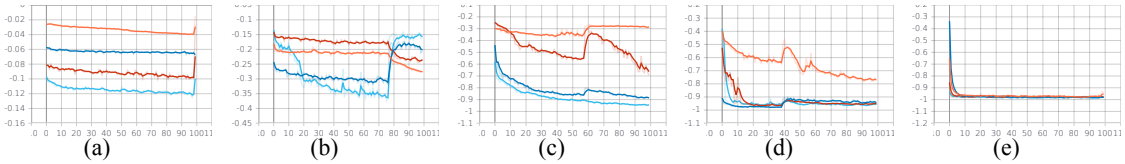


图 3.9　策略网络的损失走势：橙色曲线表示算法 MO，蓝色曲线表示算法 HMC1，红色曲线表示算法 HMC2，浅蓝色曲线表示算法 HMC3.

到控制回路中；算法 HMC3(Human-Machine-Cooperation-3)：结合了 HMC1 和 HMC2。从图3.8可以看出，随着经历 episode 的增加，动态网络模型的训练损失逐渐减少，直到达到-16.89。这是由动态模型网络训练过程的损失计算方法所决定的，如下所示：

$$loss = \frac{1}{N}\sum_{n=1}^{N}\{\sum_{o=1}^{o_{dim}}\{\frac{1}{2}(\frac{x_{t+1}-\bar{x}_{t+1}}{x_{t+1}^{\sigma}})^2 + \log\sigma\} + M\} \tag{3.10}$$

由上式可知，损失越小，训练后的动态网络模型就可以更好地拟合观察到的样本数据。此外，比较3.8(a)，3.8(b)，3.8(c)，3.8(d)和3.8(e)可以看出，随着训练过程的进展，算法 MO 的表现逐渐处于不利地位，而 HMC1 和 HMC3 在随后的时间步中则很突出。另外，图3.9显示了策略网络损失函数的训练走势。由于策略网络的损失函数与每个时间步的奖赏有关，使得理想的训练结果最终趋于-1。当训练结果如图3.9(e)所示时，表明策略网络的训练效果已经达到了较好的效果。在图3.9(e)中，与无人参与的算法 MO(橙色曲线) 相比，人类伙伴的参与 (蓝色，红色，浅蓝色曲线) 在提高训练速度和决策网络的收敛结果上明显改善。总之，由图3.8和图3.9可得出结论：人类经验的参与的确可以在一定程度上改善和增强动态网络模型和策略网络的训练性能。

2. 人对决策质量的影响

我们通过对比 CartPole 中的关键参数表征决策质量。我们比较了 MO，HMC1，HMC2，HMC3 的控制效果，如图3.10所示。图3.10描绘了第 100 个 episode 训练时四种算法所产生的关键参数趋势。a)：x，b)：$\sin(\theta)$，c)：$\cos(\theta)$，d)：r，分别代表小车的坐标位置，杆倾斜角度的正弦值和余弦值，以及获得奖励情况。并且，四个参数的理想训练结果是：$[x, \sin(\theta), \cos(\theta), r] = [0, 0, -1, 1]$。比较这三个

参数在 a)，b)，c)，d) 中的趋势，可以看出，人参与控制的算法（HMC1，HMC2 和 HMC3）在不同程度上均提高了训练效果。

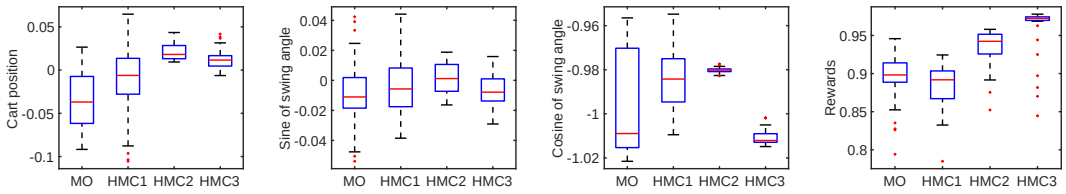


图 3.10　算法训练第 100 个 episode 的控制参数对比：(a) 小车位置；(b) 杆倾斜角度的正弦值；(c) 杆倾斜角度的余弦值；(d) 奖赏值

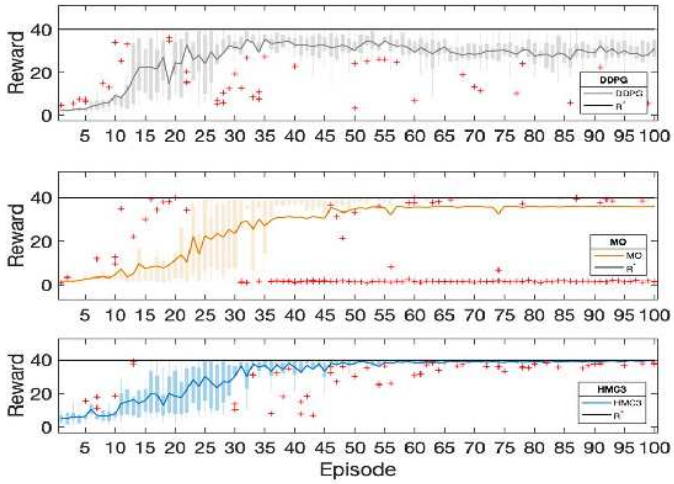


图 3.11　算法 DDPG，MO，以及 HMC3 的奖赏对比

此外，同图3.8和图3.9，在图3.11中，橙色曲线和浅蓝色曲线分别表示算法 MO 和 HMC3。为了强调多个决策者在序贯决策问题中的作用，我们特意添加了仅有一个决策者的对照实验：DDPG，如图3.11中的灰色曲线。R^* 表示达到目标的理想选择。与 MO(橙色曲线) 和 DDPG(灰色曲线) 相比，HMC3 的总体收敛效果明显更好。训练速度也有部分提高。并且，异常点 (图3.11中的红色加号标记) 也得到了改善，需要强调的是，淡蓝色曲线所代表的 HMC3 不仅表示在获得初始训练样本集中体现人类伙伴的建议，而且人类伙伴的建议存在于整个闭环控制过程中 (这并不意味着人享有更高的优先级决策权，但是代理会根据人和机器的各自决策来评估未来的运动轨迹，从而系统能够选择更安全，更有效的决策动作)。从图3.11可以看出，人机混合智能控制对于提高奖励和训练速度是有效的。

最后，我们给出了训练过程中算法 MO 和算法 HMC3 的奖赏情况和动作走势。图3.12描述了前 40 个时间步长中两种算法的奖励轨迹。从图3.12中第 0 个 episode，第 50 个 episode 和第 100 个 episode 分别对应的奖励可以看出，算法 HMC3 最为有效。此外，图3.13描述了前 40 个时间步长中算法 MO 和算法 HMC3 对应的动作轨迹。从图中可以看出，算法 HMC3 的作用值的变化范围小于算法

MO 的作用值的变化范围，这对于尽可能稳定小推车是有利的。

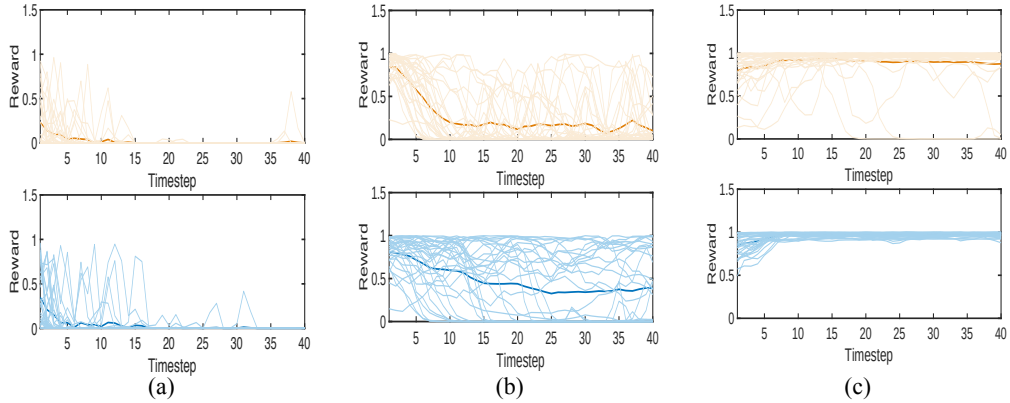


图 3.12　算法 MO(橙色曲线) 和算法 HMC3(浅蓝色曲线) 在前 40 步的奖赏对比：(a) 第 1 个 episode；(b) 第 50 个 episode；(c) 第 100 个 episode

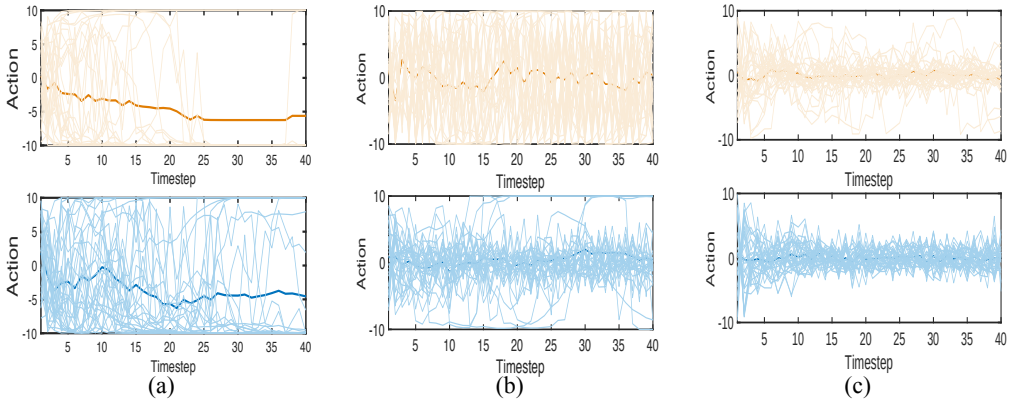


图 3.13　算法 MO(橙色曲线) 和算法 HMC3(浅蓝色曲线) 在前 40 步的动作对比：(a) 第 1 个 episode；(b) 第 50 个 episode；(c) 第 100 个 episode

3.4.2　仿真实验二:BipedalWalker

本小节实验对象是 BipedalWalker[①]，在强化学习方法基础上实现人和机器的共同控制，特别地，本小节试图描述人机混合智能算法中人和机器的决策边界参数，为后续章节的深入研究打下基础。

BipedalWalker 实验过程的任务目标是双足步行者能够在随机生成的地形上行走且不摔倒，如图3.14所示，输入状态涉及 24 个输入，包括船体的角度，角速度，前进速度，两只腿的角度，速度，是否着地，以及 10 个 Lidar 传感器参数。动作空间由 4 个连续数值构成，控制着四台电机的转矩。步行者前进得到奖励，摔倒会得到-100 的惩罚分数，施加电动机转矩消耗少量的分数，更优的策略会得到更高的分数，总共至多会得到 300+ 的分数。

[①]见：https://gym.openai.com 和 https://github.com/openai/gym

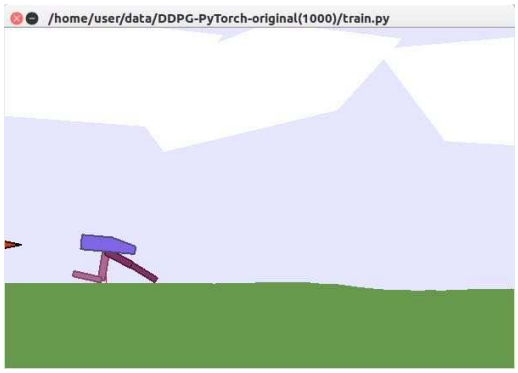


图 3.14　BipedalWalker 环境模型

1. 仿真算法

基于切换边界的人机混合智能框架如图3.15所示，不同于机器完全自主的控制算法，此处通过添加一项人对机器决策的信念 $h(t)$，改造常规人机系统的强化学习五元组描述，得到如下的六元组描述，

$$\{s(t), a_m(t), a_h(t), r(t), s(t+1), h(t)\}$$

其中 $s(t), a_m(t), a_h(t), r(t), s(t+1)$ 分别表示 t 时刻的状态，t 时刻的机器动作，t 时刻的人的动作，t 时刻的奖惩，和 $t+1$ 时刻的状态。人机系统的实际行动 $a(t)$ 依赖于人对机器行动 $a_m(t)$ 的信念: 如果人信任机器（比如 $h(t)$ 高于某一阈值 h_0），则采用机器的行动 $a_m(t)$；如果人对机器行动缺乏信任，则采用人给出的行动 $a_h(t)$（或者某种基于人与机器各自决策的融合形式），即

$$a(t) = \begin{cases} a_m(t), & h(t) \geqslant h_0 \\ a_h(t), & h(t) < h_0 \end{cases} \tag{3.11}$$

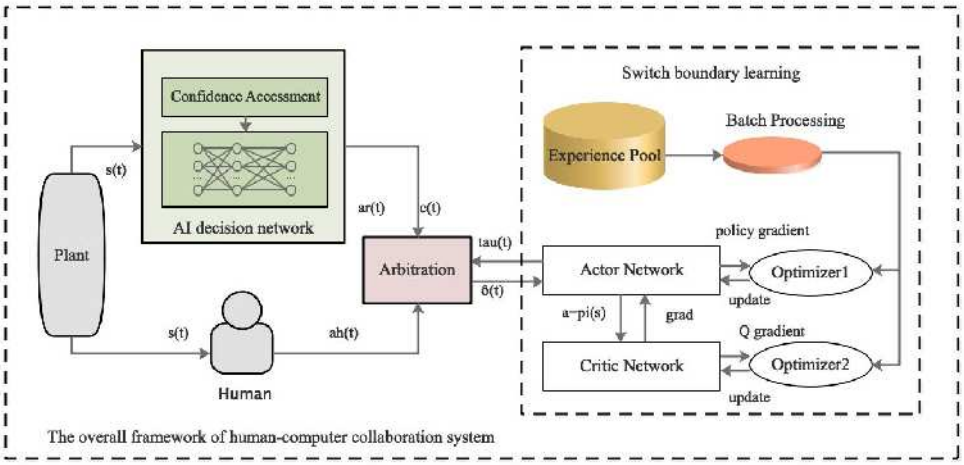


图 3.15　基于切换边界的人机混合智能框架

上述做法的目标是通过包含了人类干预的实际行动 $a(t)$ 进而通过求取最优策略的目标函数优化过程较为长远地影响人机系统的整体运行。

$$Q(s(t),a(t)) \leftarrow Q(s(t),a(t)) + \alpha[r(t) + \gamma Q(s(t+1),a(t+1)) - Q(s(t),a(t))] \quad (3.12)$$

算法 3.5　基于 Actor-Critic 的人机混合算法 (HITL-AC)

1 **初始化**：网络输入元组：$\hat{s}(t) = (s(t), s(t+1), h(t))$；随机初始化策略网络(演员)$\pi_\theta$ 和价值网络（评论家）Q_ω 的网络参数，以及各自的学习步长 $\alpha^\theta, \alpha^\omega$；随即初始化经验池 D;

2 **输出**：人机切换系统的边界 $\tau(t)$;

3 **while** 奖赏函数未收敛 **do**

4 　输入元组 $\hat{s}(t)$;

5 　**while** 到达终止状态或最大时间步 **do**

6 　　策略网络 π_θ 计算当前时间步的阈值 $\tau(t)$;

$$C = \{c(t) \geqslant |\tau(t)| \ or \ c(t) \leqslant -|\tau(t)|\}$$

$$\tau(t+1) = \begin{cases} \tau(t) - \Delta, u(t) = 1 & C \ and \ h(t) = 0 \\ \tau(t) + \Delta, u(t) = 0 & C \ and \ h(t) = 1 \\ \tau(t), r(t') = 0, & others \end{cases}$$

　　存储元组 $(s(t), s(t+1), r(t), \tau(t), \tau(t+1))$ 到经验池 D;

7 　　从经验池随机采样小批量的 N 个经验样本 $(s(i), s(i+1), r(i), \tau(i), \tau(i+1))$，计算目标价值 $y(i)^{td} = r(i) + \gamma Q_\omega(s(i+1), \tau(i+1))$ 和当前价值 $y(i) = Q_\omega(s(i), \tau(i))$;

8 　　根据随机梯度下降法更新价值网络（评论家）Q_ω 的参数：$CLoss = \frac{1}{N}\sum_i (y(i) - y(i)^{td})^2$;

9 　　根据随机梯度上升法更新策略网络（演员）π_θ 的参数：$ALoss = \frac{1}{N}\sum_i Q_\omega(s(i), \tau(i))$;

10 　**end**

11 **end**

具体实验控制算法流程如下：算法的输入 $\hat{s}(t)$ 是智能决策系统的输出，包括输入状态、动作、可信度、和人对状态的可接受判断信号。以自动驾驶为例，算法的输出是切换边界 τ。此学习算法是基于 Actor-Critic 算法框架的，首先初始化策略网络 (Actor) 和价值网络 (Critic) 的权重参数，以及初始化经验池 D(步骤 2)。对于每一个输入元组 $(s(t), s(t+1), h(t))$，策略网络计算出边界 $\tau(t)$(步骤 6)。比较输入 $c(t)$ 和 $\tau(t)$ 的大小，结合人对状态 $s(t)$ 的判断，算法给出人类是否需

要干预的控制信号 $u(t)$ 和下一时刻的边界估计 $\tau(+1)$(步骤 8)，其中，$c(t)$ 特指状态 $s(t)$ 的部分给出安全信息的指标，比如位置，角度等。步骤 9 将转换经验样本 $(s(t), s(t+1), r(t), \tau(t), \tau(t+1))$ 至经验池。从经验池随机采样小批量的 N 个转换经验样本，计算每一组经验样本中时刻 i 的目标价值和当前价值 (步骤 10)。随后根据步骤 10 计算出的时间差分误差，更新价值网络 Q_ω(步骤 11)。基于当前价值函数 $Q_\omega(s(i), \tau(i))$，利用随机梯度上升法更新策略网络 π_θ。算法3.5中的 Δ 表示边界学习的速率，这里预设为 0.001。由状态、动作和人的控制信号，获得奖赏值，这里的奖赏函数可以是如下形式：

$$r(s(t), a(t), u(t)) = \begin{cases} r(s(t), a(t)), & u(t) = 0 \\ r(s(t), a(t)) + r(s(t), a_h(t)) & u(t) = 1 \end{cases} \tag{3.13}$$

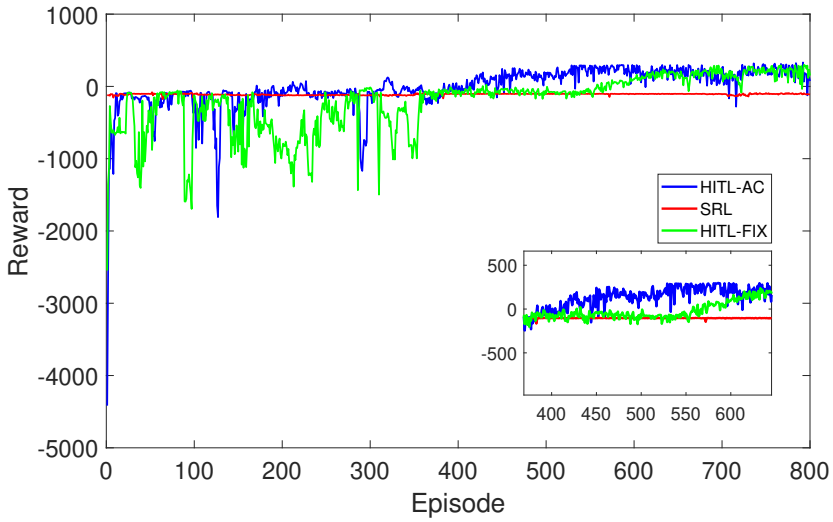


图 3.16　算法 SRL，HITL-FIX 以及 HITL-AC 的训练效果对比

2. 实验结果分析

在以下描述中，SRL(Standard Reinforcement Learning) 表示机器完全自主的标准强化学习算法，HITL-FIX(Human In The Loop Fix) 和 HITL-AC(Human In The Loop Actor Critic) 均代表基于切换边界的人机混合智能算法，其中后缀 FIX 代表固定的切换边界，AC 代表由 Actor-Critic 网络学习到的动态切换边界。如图3.16所示，在前 800 个 episodes 的训练过程中，SRL 训练效果很差，相较之下，HITL-FIX 能够在第 500 个 episodes 之后取得奖赏的整体提升。与 HITL-FIX 相比，HITL-AC 更够更快的提升训练效果（第 400 个 episodes 附近）。其中，红色曲线 SRL 表示智能体在 DDPG 算法上训练 800 个 episodes 所获的的奖励演变。绿色曲线 HITL-FIX 表示在原 DDPG 算法的基础上，基于固定切换边界下融入人的建议（人觉得某个状态是危险的，需要给予复位的操作，此部分对应于算法

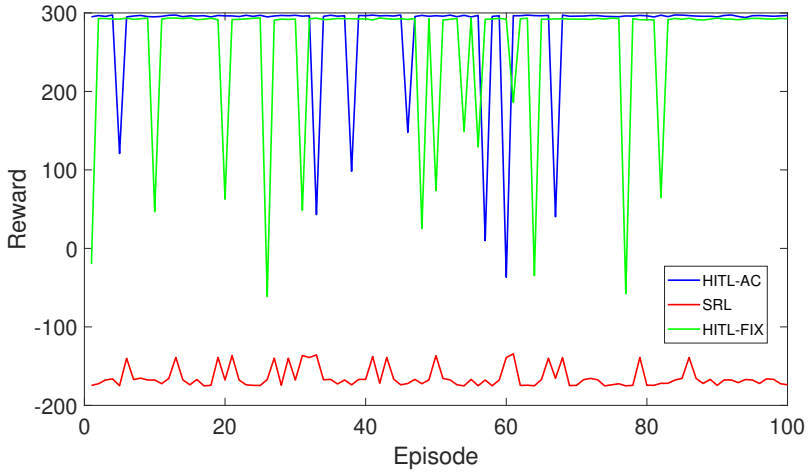


图 3.17　算法 SRL，HITL-FIX 以及 HITL-AC 的测试性能对比

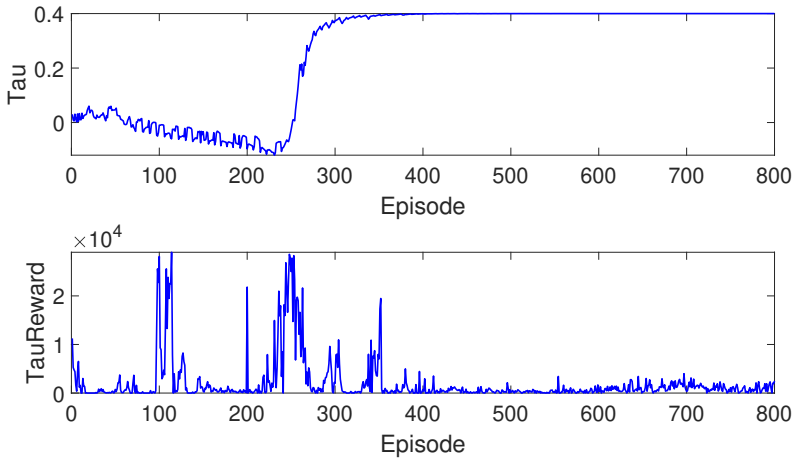


图 3.18　算法 SRL，HITL-FIX 以及 HITL-AC 的测试性能对比

3.5中 $h(t)$ 的判断）后的训练过程。蓝色曲线 HITL-AC 在 HITL-FIX 算法固定切换边界的基础上，修改为具有动态的切换边界，进而将人的建议融入学习算法的过程。

图3.17给出训练之后的智能体的 100 个 episodes 的测试结果。与图3.16一样，红色 SRL 奖励最小。绿色 HITL-FIX 和蓝色 HITL-AC 有较明显的提升，且 HITL-AC 较 HITL-FIX 而言，平均奖励更大（279.7164 > 260.4995），说明相较于标准强化学习算法，HITL(尤其是 HITL-AC) 能够更加快速的训练智能体达到较好的控制效果。最后，图3.18给出对应于算法3.5的切换边界学习过程曲线和相应训练过程的奖励大小。

3.5　本章小结

本章基于强化学习方法，利用由人类智能和机器智能构成的混合智能算法求解人机序贯决策问题。具体而言，本章通过仲裁机制将基于模型的强化学习决策子系统、基于无模型强化学习决策子系统和人类决策者进行集成，从而获得更优的待执行决策动作，其中仲裁评价指标涉及可信性和安全性等。接着我们通过第一个仿真案例分别从“人对决策网络的影响”和“人对决策结果的影响”两方面说明本章方法的有效性，以及在第二个仿真案例中验证人机混合智能算法思想的合理性。最后我们更进一步地讨论了切换边界对于问题求解的作用，为后续章节对边界的深入研究奠定了初步基础。

第 4 章　人机序贯决策：基于自主性边界优化设计介入控制算法

本章通过对介入控制算法的讨论，完善面向人机序贯决策的混合智能理论分析和应用。具体涉及对混合智能中自主性边界的判定，并利用获得的自主性边界信息优化完善介入控制算法，最终实现改善提升人机序贯决策任务的决策质量。

4.1　引言

人机序贯决策中的介入控制，涉及日常生活生产中众多领域，比如人对作战系统的强制介入，自动驾驶水平等级如表1.1中的 L1、L2、L3 等级中人对紧急危险情况的挽救，监视驾驶员实时状态且在疲劳时强制干预的辅助驾驶系统，等。因此对人机序贯决策中的介入控制问题进行研究非常有必要和有意义。事实上，介入控制包括三种情况，即人介入机器、机器介入人、以及人机双向介入 (切换)。然而，如何正确把握介入的触发时机，实现既不造成人力的浪费，也不因介入的发生导致系统的恶化甚至崩溃，则是一件难以评估和判断的事情，处理不当会损失惨重。

目前关于介入控制已有相关研究，如 [61] 基于集成的思想提出了主系统和次系统协同解决同一个控制任务，实现了当主系统和辅助系统子系统做出差异较大的决策动作时，人类作为监督者单方面干预到机器代理的决策中。文献 [60] 展示了一个用人类用户手势进行介入控制的远程操作系统，利用感知、估计、规划等组件，根据物体姿势和形状完成机器人运动轨迹的规划。文献 [62] 依赖于数据驱动的人机系统，以基于模型的表示评估用户可能希望并行提供的大量潜在输入，这种方式使用户可以在难以获得或没有用户目标的情况下为所欲为（分配给人类合作伙伴的最大权限）并提高系统安全性。已有研究中对介入时机的触发较少的基于判断标准，这是由其介入者优先级高于被介入者优先级的本质所决定的，但这不应成为限制完善介入触发判断机制的借口。

综上所述，本章在介入控制的已有研究基础上，提出了自主性及自主性边界的概念，通过将自主性边界的求解形式化为与任务目标相关的常规优化问题进行讨论判定，优化介入控制的控制方案和算法，实现人机序贯决策中人介入机器场景和机器介入人场景下的决策性能提升。自主性边界信息的探讨一方面使得人和机器拥有更清晰的决策权限，有利于人机更好协作完成任务；另一方面使得人和机器输出的更优决策动作被仲裁选择，以及整体决策系统的优化。更清

楚地，本章所述的自主性和自主性边界分别指人和机器各自的行动范围及界限，控制框架具体包括机器智能决策系统，人类用户控制，仲裁模块，和自主性边界学习模块等。

本章结构安排如下，第4.2节介绍人机序贯决策中的人介入机器控制，即基于自主性边界信息对人介入机器控制方法进行优化，并且给出实验设计和结果分析。第4.3介绍人机序贯决策中的机器介入人控制，即基于自主性边界信息对机器介入人控制方法进行优化，并且给出实验设计和结果分析。

4.2　人机序贯决策中的人介入机器控制

针对部分由机器自主处理的人机序贯决策场景可能面临设计的困难或性能提升上的瓶颈，考虑到人类智能的独特性是处于目前发展阶段的人工智能所无法完全替代的，基于此，人类的适时介入成为可能的解决办法之一。并且人类知识的融入，为改善智能体的学习过程提供了可能。图4.1给出人介入机器的控制框架示意图。图中机器代理是常规决策者的角色，在此场景下，智能机器能够在一定自主性范围内给出代替人类甚至超越人类的决策动作。严格意义上讲，人类伙伴不是时刻处于控制回路内，更像是处于控制回路上，扮演着“监督者”的角色。当机器代理的决策动作出现严重错误或者机器代理无法辨识自己是否错误时，以某种方式警示人类伙伴予以援助，直至机器代理或被控对象回归正常状态。

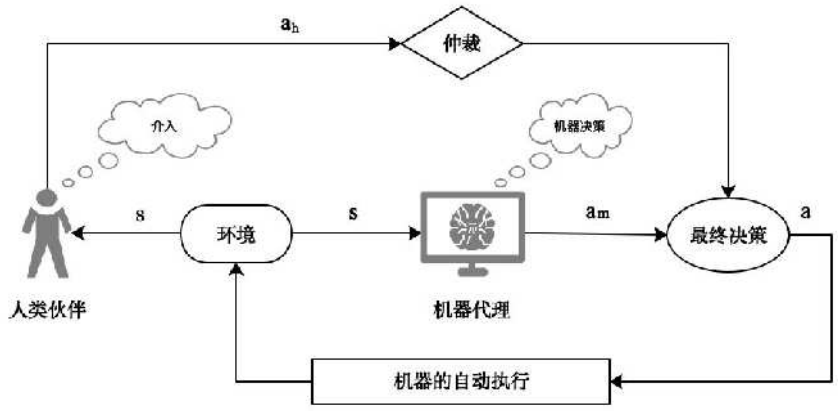


图 4.1　人介入机器控制框架

4.2.1　人介入机器控制中的自主性边界判定

首先我们给出机器自主性边界的定义：

定义 4.1 (机器的自主性边界)　机器的自主性边界是指按照有益于人机混合系统共同优化目标的方向，AI 驱动的机器智能进行决策和行动的范围界限。

一般情况下，自主性边界由其下界和上界共同构成。但是机器的自主性下界涉及机物理械自动化层面，不在我们所考虑的智能决策层面，因此本章所考虑的是机器的自主性上界问题。

在人介入机器控制系统的策略设计过程中，机器的自主性边界是一个重要的概念，它关乎到人类伙伴何时以及何种方式介入到机器控制中。当不超过这个边界时，满足机器自主决策的需求，当超过这个边界时，人类伙伴介入得以触发。并且随着决策的动态进行，机器的自主性边界得到实时的优化，并且优化后的自主性边界重新又可以作为判断条件。因此我们考虑将机器的自主性边界问题定义为一个优化问题。

$$b_m(t) = \arg \max_{a_m(t)\in\mathcal{A}_m(t)} J^b_{h,m}(s(t), a_m(t)) \tag{4.1a}$$

$$\text{s. t. } C(s(t), a_m(t)) < 0 \tag{4.1b}$$

其中 $\mathcal{A}_m(t) := \{a_m(t), t \geqslant 0\}$，$J^b_{h,m}(s(t), a_m(t))$ 可取为人和机器的共同目标函数形式，可根据具体实施场景和算法将目标函数定义为不同表达形式，比如累积奖赏（成本）函数

$$J^b_{h,m}(s(t), a_m(t)) = \int [r(s(t), a_m(t)) - c(s(t), a_m(t))]dt \tag{4.2}$$

式(4.1a)给出了最大化目标函数的示例，因此希望找到满足约束条件下，使得优化目标最好（大）的机器动作作为机器的自主性上界；当系统追求最小化目标函数，将 $J^b_{h,m}(s(t), a_m(t))$ 最小化或者将直接给 $J^b_{h,m}(s(t), a_m(t))$ 加上负号的形式即可实现。$s(t)$ 是系统状态，$a_m(t)$ 是机器智能代理的决策动作，上述被控对象的约束条件 $C(s(t), a_m(t)$ 为一般性表达式，需视具体场景而定（如在 MPC 中可以硬约束的方式显示出现在优化问题中，或者 MDP 中通过目标函数隐式表达）。

考虑上述式(4.1)的优化问题，可将具体思想描述如算法4.1所示，首先根据已有信息初始化得到一个机器的自主性上界，比如以类似于神经网络中参数随机初始化的方式进行。在动态决策过程中，对于机器智能的实时决策动作（第 5 步），算法第 6 步基于约束条件对机器动作进行判断筛选，之后第 7 步比较已有的机器自主性上界信息 $b_m(t-1)$ 和满足约束的机器动作 $a_m(t)$ 分别对应的目标函数，目的是找到使得目标函数(4.1a)更大时对应的机器动作 $a_m(t)$，并根据(4.1)更新当前时刻的机器的自主性上界 $b_m(t)$。至此，我们给出了机器自主性边界判定的一般性方法。

算法 4.1　机器的自主性边界判定

初始化: 初始化机器的自主性上界 $\bar{B}_m = \{b_m(0)\}$;
输出: 机器的自主性上界 $\bar{B}_m = \{b_m(t)\}$;
while 训练未结束 **do**
　for 未达到终止状态 **do**
　　输入: 机器动作 $a_m(t)$;
　　根据约束条件(4.1b)对当前时刻的机器动作进行筛选;
　　将满足约束条件的机器动作与上一时刻的边界信息进行目标函数的比较，如果 $J^b_{h,m}(s(t), a_m(t)) > J^b_{h,m}(s(t), b_m(t-1))$，则根据(4.1)更新当前 t 时刻的自主性边界 $b_m(t)$，以获得更优的机器自主性上界；否则机器的自主性上界保持不变;
　end
end

4.2.2　人介入机器控制的优化算法

基于4.2.1节中关于机器自主性上界的判定方法，本小节将研究如何利用获得的机器自主性边界优化人机序贯决策问题的求解。

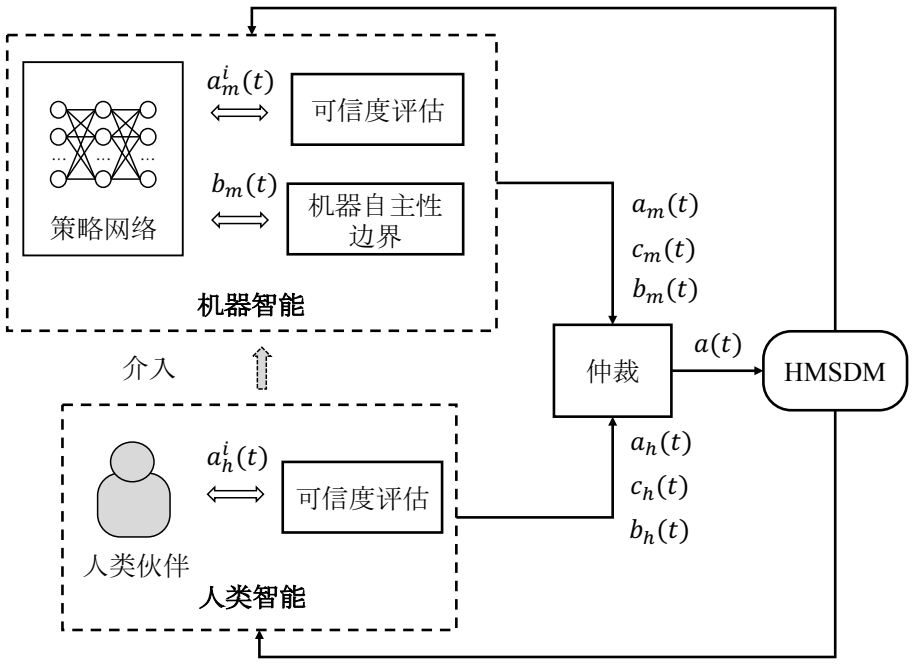


图 4.2　面向人机序贯决策的人介入机器控制框架

本节将面向人机序贯决策求解的基于自主性边界的人介入机器控制优化设计描述如图4.2所示。图4.2有两个决策主体，即机器智能和人类智能。机器智能包括策略网络、可信度评估模块和自主边界学习网络。对于任何时刻 t 的系统状

态 $s(t)$，机器会输出各自的决策动作 $a_m(t)$，决策可信度 $c_m(t)$，以及当前的决策边界 $b_m(t)$。仲裁模块根据上述三个输入信号决定是否进行人类的干预，从而输出最终的决定动作。

将人机序贯决策中的人介入机器控制的优化问题列为如下形式：

$$\max_{\theta} J_{h,m}(s(t),a(t)) = \int [r(s(t),a(t)) - c(s(t),a(t))]dt \tag{4.3a}$$

$$\max_{a_m(t)\in\mathcal{A}_m(t)} J^b_{h,m}(s(t),a_m(t)) = J_{h,m}(s(t),a_m(t)) \tag{4.3b}$$

$$\text{s. t. } \dot{s}(t) = f^d(s(t),a(t)) \tag{4.3c}$$

$$a(t) = f^a(a_h(t),a_m(t),b_m(t-1)) \tag{4.3d}$$

$$a_h(t) = Human - Action \tag{4.3e}$$

$$a_m(t) = p^m(s(t);\theta) \tag{4.3f}$$

$$C(s(t),a_h(t),a_m(t)) < 0 \tag{4.3g}$$

$$t = 0,1,2,3,...$$

其中 $J^b_{h,m}(s(t),a_m(t))$ 是自主性边界的优化目标函数。这里我们可以选择其为被控对象的优化目标函数 $J_{h,m}(s(t),a(t))$ 的形式。$r(\cdot)$ 和 $c(\cdot)$ 分别代表时间 t 的即时奖励和成本。$f^d(\cdot)$ 表示系统的动态模型。$f^a(a_h(t),a_m(t),b_m(t-1))$ 是人类动作和机器动作的仲裁函数，其中 $b_m(t-1)$ 可以通过公式(4.1)和算法4.4求解。为方便起见，$f^a(\cdot)$ 定义为(4.4)。

$$a(t) = f^a(a_h(t),a_m(t),b_m(t-1))$$

$$= \begin{cases} \text{Human: } a_h(t), \{c_h(t) > c_m(t)\} \ \& \ \{J_{h,m}(s(t),a_h(t)) \geqslant \max\{J_{h,m}(s(t),a_m(t)), \\ \qquad J_{h,m}(s(t),b_m(t-1))\}\} \\ \text{Boundary: } b_m(t-1), \{c_m(t) > c_h(t)\} \ \& \ \{J_{h,m}(s(t),b_m(t-1)) \geqslant \max\{J_{h,m}(s(t), \\ \qquad a_h(t)), J_{h,m}(s(t),a_m(t))\}\} \\ \text{Machine: } a_m(t), \qquad \text{其他.} \end{cases} \tag{4.4}$$

公式(4.4)表示机器介入人控制中的仲裁函数。根据对可信度评估和目标函数的大小判断谁是当前的决策者。(4.4)中的 $c_m(t)$ 和 $c_h(t)$ 代表机器动作 $a_m(t)$ 和 $a_h(t)$ 的可信度评估。考虑到贝叶斯神经网络的概率特性，仲裁函数采用 MC dropout[145] 方法来衡量决策可信度，如第3章公式(3.9)。更具体地，如果人类决策动作的可信度 $c_h(t)$ 高于机器决策动作的可信度 $c_m(t)$，并且 $a_h(t)$ 所对应的目标

函数大于 $a_m(t)$ 和 $b_m(t-1)$ 对应目标函数中的大者，那么人类就成为决策者；如果机器决策动作的可信度 $c_m(t)$ 高于人类决策动作的可信度 $c_h(t)$，并且 $b_m(t-1)$ 所对应的目标函数大于 $a_h(t)$ 和 $a_m(t)$ 对应目标函数中的大者，那么最优决策取在机器自主性边界上；否则决策者就是机器代理。

接下来，我们给出人介入机器控制优化算法的具体流程。本节在使用介入控制算法求解人机序贯决策问题的基础上，引入了机器自主边界判定网络。因此算法的优化目标不仅是优化与决策动作直接相关的策略函数，还包括学习优化间接影响决策动作的机器自主性边界。首先，我们初始化机器自主性边界信息、策略网络及其参数。在动态演化过程中，人类伙伴和智能机器将分别给出实时状态 $s(t)$ 的决策动作 $a_h(t)$ 和 $a_m(t)$，以及相应的可信度分析 $c_h(t)$ 和 $c_m(t)$。之后，根据优化(4.1)中的约束(4.3g)过滤人类动作和机器动作。考虑人介入机器控制优化算法的共同目标函数，利用仲裁函数(4.4)输出最终决策动作 $a(t)$。最后基于式(4.1)和算法4.1完成当前时刻 t 机器自主性边界的更新，重复循环直到训练结束。

算法 4.2　人介入机器控制的优化算法

初始化：随即初始化机器代理的策略网络 p^m 及其网络参数 θ；初始化机器的自主性边界 $\bar{B}_m$;
输入：系统状态 $s(t)$;
输出：最终决策动作 $a(t)$;
while 未达到最大训练时间步数 **do**
    for 未到达终止状态 **do**
        智能机器根据策略网络计算出决策动作 $a_m(t)$ 以及基于蒙特卡洛估计 (3.9) 计算可信度 $c_m(t)$;
        根据优化问题(4.3)中的约束(4.3g)过滤机器决策动作 $a_h(t)$，以及衡量可信度 $c_h(t)$;
        利用仲裁函数(4.4)输出最终决策动作 $a(t)$;
        基于算法4.1更新维护机器自主性边界信息 $b_m(t)$;
    end
end

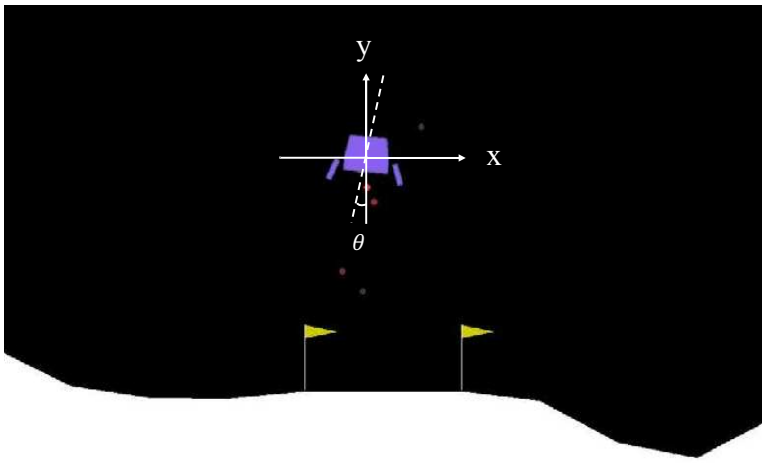


图 4.3　月球着陆器 LunarLander

4.2.3　仿真实验

1. 实验设置

我们使用 OpenAI Gym 中的 LunarLander[①]，如图4.3所示。在着陆器下降过程中，如果着陆器坠毁或静止，则一次完整经验轨迹结束，并获得-100 分或 100 分的奖励。着陆器每条腿触地奖励 10 分。当主引擎开启时，每帧画面消耗燃料奖励 0.3 分 (假设燃料无限)。着陆器 $s(t)$ 的状态向量包括：坐标 $(x(t), y(t))$，速度 $(\dot{x}(t), \dot{y}(t))$，角度 $(\theta(t), \dot{\theta}(t))$，是否着陆 $(leg_l(t), leg_r(t))$ 和着陆点坐标 $h(t)$。

表 4.1　动作值和各引擎开关之间的对应关系

Action Value	Main Engine	Left Engine	Right Engine
0	OFF	OFF	ON
1	OFF	OFF	OFF
2	OFF	ON	OFF
3	ON	OFF	ON
4	ON	OFF	OFF
5	ON	ON	OFF

离散动作集合为 $\{0, 1, 2, 3, 4, 5\}$，具体对应关系如表4.1所示，其中 0（向左和向下）表示主引擎和左引擎关闭，右引擎打开；1（向下）表示所有引擎均关闭；2（向右和向下）表示主引擎和右引擎关闭，左引擎打开；3（向左和向上）表示主引擎和右引擎打开，左引擎关闭；4（向上）表示主引擎打开，左引擎和右引擎关闭。5（向右和向上）表示主引擎和左引擎打开，右引擎关闭。故而，LunarLander 仿真环境就变成了通过操控着陆器的三个引擎，进而实现着陆器在着陆区域（黄

[①]见：https://gym.openai.com 和 https://github.com/openai/gym

色小旗之间）的安全着陆。

本章实验采用 DQN(Deep Q-learning Network) 算法作为机器代理算法。我们对机器自主控制算法和人介入机器控制算法进行比较，参数包括奖励、成功率、撞击率、人机动作所占百分比、人机动作轨迹和着陆器运行轨迹比较等。并且在正式的对比实验开始之前，我们先对机器代理算法 DQN 进行预训练，让机器代理具备有效决策的能力基础。在本小节以下的所示图中，使用 MOA(Machine-Only-Algorithm) 表示仅使用机器控制算法，使用 HTMA(Human-Trade-Machine-Algorithm) 表示人类介入机器控制算法，以及 HTMA-B(Human-Trade-Machine-Algorithm-Boundary) 的使用是表示在 HTMA 的基础上增加自主性边界信息的优化算法。

2. 实验结果

在图4.4和图4.5(a)中，比较 MOA、HTMA 和 HTMA-B，在一定时间内，人介入机器控制算法 (HTMA，HTMA-B) 比纯机器控制算法 MOA 获得的奖励大。特别是 HTMA-B 算法在奖励方面有显著提升，这也印证了我们引入的自主性边界正向促进人机序贯决策问题性能的提升。此外，更具体地，我们从 500 个随机 episode 中选择成功着陆的 episode，并对其奖励进行平均得到图4.5(b)。在图4.5(b)中，观察到成功着陆的 episode 对应的奖励满足：HTMA-B>HTMA>MOA 的关系。因此可以得出结论，本节提出的基于自主性边界的人介入机器控制优化算法（HTMA-B）与以往的机器自主算法（MOA）和常规的人介入机器控制算法（HTMA）相比具有更大的优势。

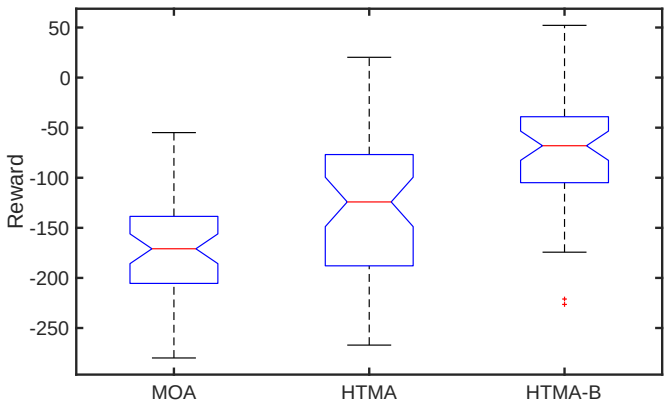


图 4.4　算法 MOA，HTMA，HTMA-B 的平均奖赏对比（500 episodes）

在 LunarLander 中，顺利安全地降落在着陆点是游戏成功与否的决定性因素。接下来，我们比较了算法 MOA、HTMA 和 HTMA-B 的着陆成功率和撞击率。对于图4.5(c)中的成功率，MOA 的成功率继续偏低，这是由于机器代理是基于神经网络训练的，需要较多时间训练才能完成。在人介入机器控制下，着陆成功率

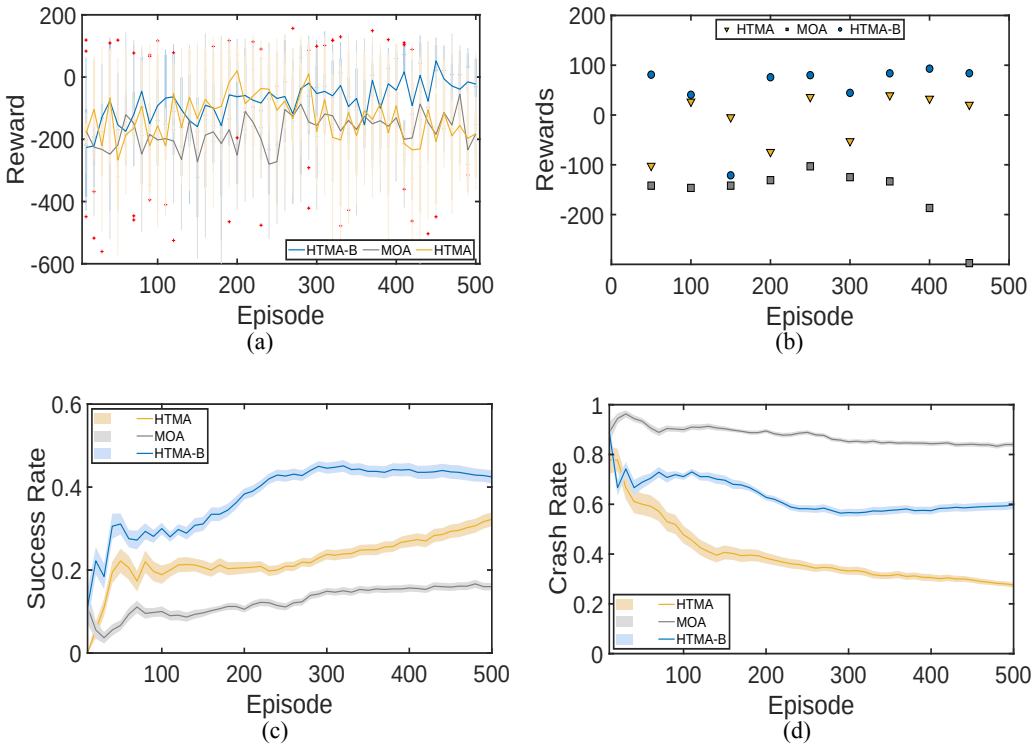


图 4.5　算法 MOA，HTMA 和 HTMA-B 的实验结果对比。(a) 奖赏：实线表示奖赏平均值走势，红色加号代表异常点，阴影表示大多数点所落在的箱体区域；(b) 着陆成功的 episodes 的奖赏值；(c) 成功率：实线表示成功率的平均值，阴影表示不确定性；(d) 撞击率: 实线表示撞击率的平均值，阴影表示不确定性。

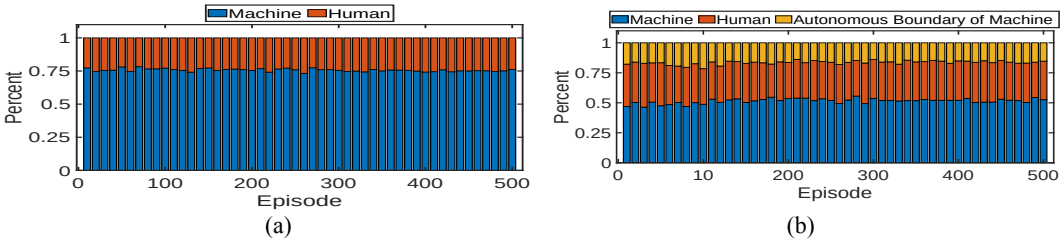


图 4.6　人类动作 a_h 和机器动作 a_m 多占百分比表示：(a) HTMA；(b) HTMA-B。

得到了提升，尤其是本文介绍的 HTMA-B 算法，其可以持续提升登陆成功率到 0.45 甚至更高。关于图4.5(d)中的撞击率，实验结果是 MOA>HTMA-B>HTMA，我们将这种现象归因于自主性边界信息的使用。但不可否认的是，基于自主性边界信息的优化设计确实对决策性能 (如奖励值、成功率) 有显著的提升作用，因此研究人员需要做出更好的权衡或妥协。从上面关于奖励、着陆成功率和撞击失败率的实验结果可以看出，机器代理可以利用自己的一些自主权来减少人力。然而，目前基于神经网络的机器代理的研究仍需要在学习能力有所提高，因此需要寻求更好的优化策略，例如本节所提的基于自主性边界的人介入机器控制设计（HTMA-B），这是有意义和有效的。

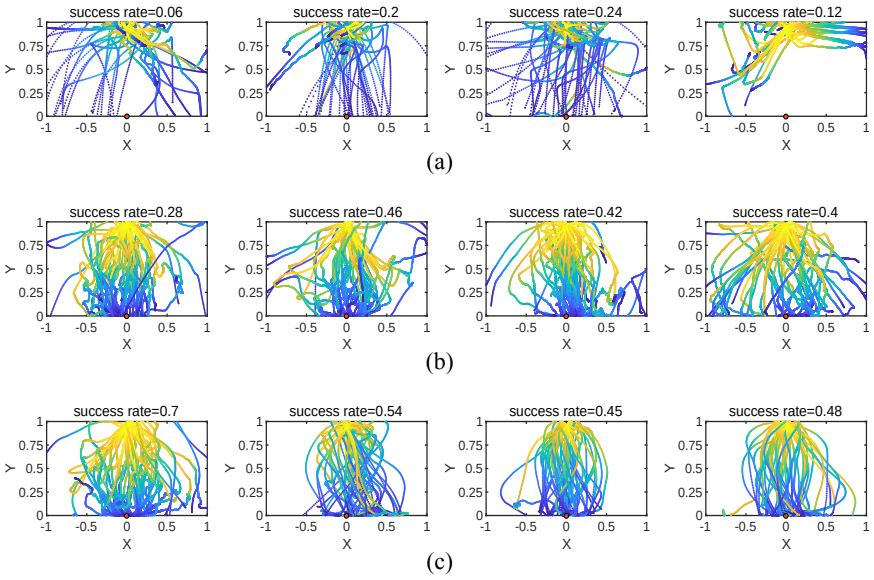


图 4.7　不同算法对应着陆轨迹的对比：(a) MOA；(b) HTMA；(c) HTMA-B。

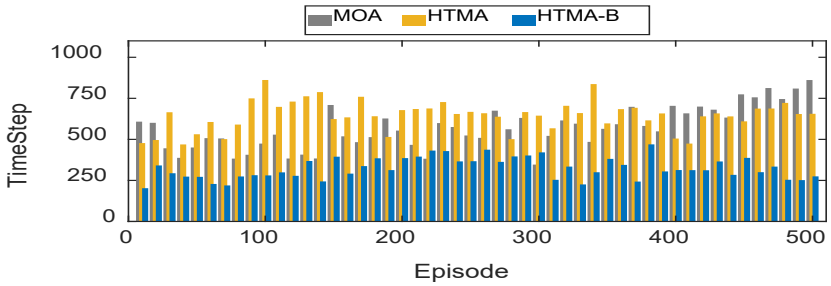


图 4.8　算法 MOA，HTMA 和 HTMA-B 在每一条 episode 中的时间步长走势

图4.6描述了算法 HTMA、HTMA-B 的动作百分比。我们观察到，在机器自主控制与人介入机器控制中，人类动作所占份额更高，这与优先级直接相关。HTMA-B 中人的动作、机器自主性边界和机器动作分别以 3 ： 1 ： 4 的比例影响最终决策动作。此外，我们发现在 HTMA-B 中，机器动作约占 50%，这也引起了思考，即在实时动态演化过程中，机器是一个不容忽视的决策主体，这是由实例本身的固有属性决定的 (更适合机器操作)。

接下来，我们比较了算法 MOA、HTMA 和 HTMA-B 的着陆轨迹，如图4.7所示。我们发现在 MOA 对应的图 4.7(a) 中，着陆成功率很低。在图4.7(b) 中，HTMA 算法对 MOA 做了初步的改进，包括成功率和发散度。此外，算法 HTMA-B 对 HTMA 进行了增强，不仅提高了成功率，如图4.7(c)，而且着陆轨迹更加有序快速，如图4.8所示。可见 HTMA-B 在运行时间步长方面具有显着提高决策性能的效果。

最后，为了便于理解本小节提到的人介入机器控制算法的执行过程，我们给出最终的决策动作 $a(t)$，机器决策动作 $a_m(t)$，与人类决策动作 $a_h(t)$ 的对应关系，

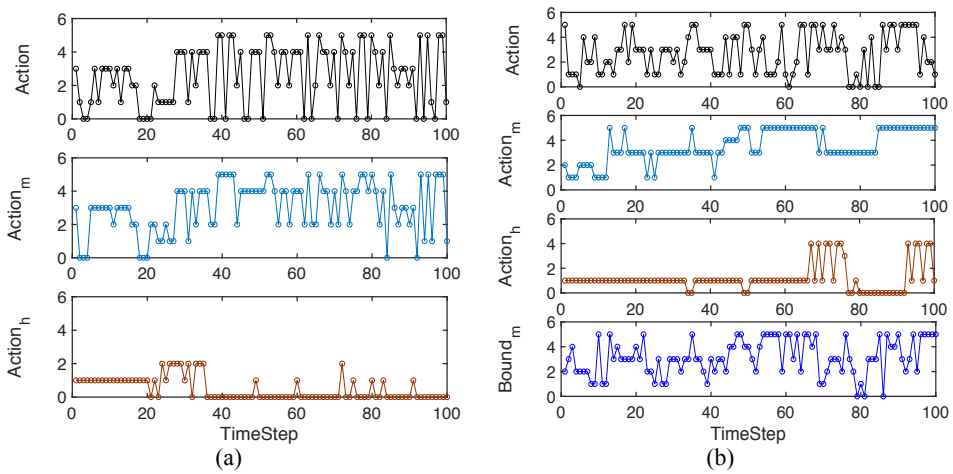


图 4.9　算法 HTMA 和 HTMA-B 的决策动作对应关系，左侧表示一条完整 episode，右侧表示前 50 步。(a) HTMA：从上至下分别是最终决策动作 a，机器决策动作 a_m，人类决策动作 a_h；(b) HTMA-B：从上至下分别是最终决策动作 a，机器决策动作 a_m，人类决策动作 a_h，机器自主性边界。

如图4.9。从图4.9(a)，观察到最终决策动作 a 的值总是介于机器决策动作 a_m 和人类决策动作 a_h 之间。从图4.9(b)中，最终的动作 $a(t)$ 是在 $a_m(t)$、$a_h(t)$ 和 $b_m(t-1)$ 之间进行选择，其中 $b_m(t-1)$ 是机器的边界信息，即对应上述的通过添加额外的自主性边界信息优化和判断最终的决策信号。

4.3　人机序贯决策中的机器介入人控制

随着以 AI 为代表的人工智能技术的迅猛发展，考虑一类机器主要以辅助人类完成其控制目标的人机混合系统，机器的适时介入能够改善人类控制的准确度或减轻人类劳动强度，典型的例子如帮助实现人体机能的机械装置 (霍金的轮椅)、汽车驾驶中的车道保持辅助驾驶系统 [72] 等。图4.10给出机器介入人的控制框架。在图4.10中，人类伙伴是常规决策者的角色，此场景下，人类伙伴能够在一定自主性范围内给出超越机器的决策动作。严格意义上讲，机器代理不是时刻处于控制回路内，更像是处于控制回路上，扮演着“监督者”或“临时决策者”的角色。当人类伙伴的决策动作出现严重错误或者人类的劳动可以被适当替代时，机器可以强制干预或者通过人给予的提醒而介入决策中，直至人类伙伴回归正常状态或者机器控制下出现异常状态。

4.3.1　机器介入人控制中自主性边界的判定

首先给出人类自主性边界的定义：

定义 4.2 (人类的自主性边界)　人类的自主性边界是指按照有益于人机混合

系统共同优化目标的方向，人类智能进行决策和行动的范围界限。

一般情况下，自主性边界由其下界和上界共同构成。由于人类的自主性下界涉及人类本身认知缺陷层面，不在所考虑的范围内，因此本章所考虑的是人类的自主性上界问题。

在机器介入人类控制系统的策略设计中，人的自主性边界是一个重要的概念。它关乎到机器何时以及何种方式介入到机器控制中。当不超过这个边界时，系统满足人的控制决策愿景，当超过这个边界时，系统允许机器介入的发生。并且随着决策进程的进行，人的自主性边界是可以得到实时优化的，那么优化后的边界又重新可以作为判决条件。因此考虑将人类的自主性边界问题定义为如下的优化问题。

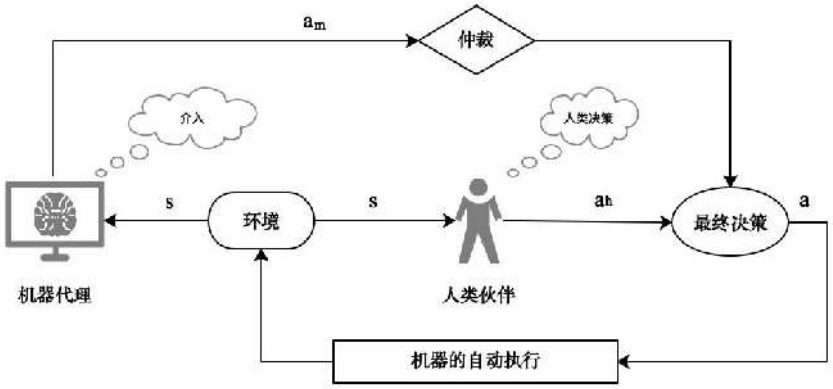


图 4.10　机器介入人的控制系统

$$b_h(t) = \arg \max_{a_h(t) \in \mathcal{A}_h(t)} J^b_{h,m}(s(t), a_h(t)) \tag{4.5a}$$

$$\text{s. t. } C(s(t), a_h(t)) < 0 \tag{4.5b}$$

其中 $\mathcal{A}_h(t) := \{a_h(t), t \geqslant 0\}$，$J^b_{h,m}(s(t), a_h(t))$ 是人和机器的共同目标函数，可根据具体实施场景和算法将目标函数定义为不同表达形式，比如累积奖赏（成本）函数

$$J^b_{h,m}(s(t), a_m(t)) = \int [r(s(t), a_h(t)) - c(s(t), a_h(t))]dt \tag{4.6}$$

式(4.6)给出了最大化目标函数的示例，即希望目标函数越大越好，因此希望找到满足约束条件下，最好（大）的人类输入动作作为人的自主性上界；当系统追求最小化目标函数，将 $J^b_{h,m}(s(t), a_h(t))$ 最小化或者将直接给 $J^b_{h,m}(s(t), a_h(t))$ 加上负号的形式即可实现。$s(t)$ 是系统状态，$a_h(t)$ 是人的输入动作，上述被控对象

的约束条件 $C(s(t), a_h(t))$ 为一般性表达式，需视具体场景而定（如在 MPC 中可以硬约束的方式显示出现在优化问题中，或者 MDP 中通过目标函数隐式表达）。

算法 4.3　人的自主性边界判定

初始化：人的自主性上界 $\bar{B}_h = \{b_h(0)\}$;
输出：人的自主性上界 $b_h(t)$;
while 未达到训练结束条件 **do**
    for 未到达终止状态 **do**
        输入：人类动作 $a_h(t)$;
        根据约束条件(4.5b)对当前时刻的人类动作进行筛选;
        将满足约束条件的人类输入与第 2 步中的上界信息进行目标函数的比较，如果 $J^b_{h,m}(s(t), a_h(t)) > J^b_{h,m}(s(t), b_h(t-1))$，则根据(4.5)更新当前 t 时刻的自主性边界 $b_h(t)$，以获得更优的人的自主性上界；否则人的自主性上界保持不变;
    end
end

我们考虑上述式(4.5)的优化问题，可将具体思想描述如算法4.3所示，首先我们根据已有信息初始化得到一个人类的自主性上界，比如以类似于神经网络中参数随机初始化的方式进行。在控制过程中，对于人类的实时输入动作（第 5 步），算法第 6 步基于约束条件对人类动作进行判断筛选，之后第 7 步比较已经获得的人类自主性上界信息 $b_h(t-1)$ 和满足约束的机器动作 $a_h(t)$ 所对应的目标函数，目的是找到使得目标函数(4.5a)更大时对应的人类动作 $a_h(t)$，并根据式(4.5)更新当前时刻与系统状态对应的人类的自主性上界 $b_h(t)$。重复下去直至训练结束。至此，我们给出了人类自主性边界判定的一般性方法。

4.3.2　机器介入人控制的优化算法

基于4.3.1节中关于人的自主性上界判定方法，本小节将研究如何利用获得的人的自主性边界优化人机序贯决策问题的求解。

本节将面向人机序贯决策求解的基于自主性边界的机器介入人控制优化设计描述类似于图4.2所示，区别在于本节强调机器智能向人类智能的介入。同样的也有两个决策主体，即机器智能和人类智能。机器智能包括策略网络、可信度评估模块和自主边界学习网络。对于任何时刻 t 的系统状态 $s(t)$，机器会输出各自的决策动作 $a_m(t)$，决策可信度 $c_m(t)$，以及机器的自主性边界 $b_h(t-1)$。仲裁模块根据上述三个输入信号决定是否进行机器的干预，从而输出最终的决策动作。

将人机序贯决策中的机器介入人控制的优化问题列为如下形式：

$$\max_{\theta} J_{h,m}(s(t), a(t)) = \int [r(s(t), a(t)) - c(s(t), a(t))]dt \tag{4.7a}$$

$$\max_{a_h(t)\in\mathcal{A}_h(t)} J^b_{h,m}(s(t), a_h(t)) = J_{h,m}(s(t), a_h(t)) \tag{4.7b}$$

$$\text{s. t. } \dot{s}(t) = f^d(s(t), a(t)) \tag{4.7c}$$

$$a(t) = f^a(a_h(t), a_m(t), b_h(t-1)) \tag{4.7d}$$

$$a_h(t) = Human - Action \tag{4.7e}$$

$$a_m(t) = p^m(s(t); \theta) \tag{4.7f}$$

$$C(s(t), a_h(t), a_m(t)) < 0 \tag{4.7g}$$

$$t = 0, 1, 2, 3, ...$$

其中 $b_h(t-1)$ 是根据自主性边界的优化目标函数 $J^b_{h,m}(s(t), a_h(t))$ 求得的人的自主性边界。$r(\cdot)$ 和 $c(\cdot)$ 分别代表时间 t 的即时奖励和成本。$f^d(\cdot)$ 表示系统的动态模型。$f^a(a_h(t), a_m(t), b_h(t))$ 是人类动作和机器动作的仲裁函数，其中 $b_h(t)$ 可以通过公式(4.5)和算法4.3求解。为方便起见，$f^a(\cdot)$ 定义为(4.8)。

$$\begin{aligned} a(t) &= f^a(a_h(t), a_m(t), b_h(t-1)) \\ &= \begin{cases} \text{Machine: } a_m(t), \{c_m(t) > c_h(t)\} \ \& \ \{J_{h,m}(s(t), a_m(t)) \geqslant \max\{J_{h,m}(s(t), a_h(t)), \\ \qquad J_{h,m}(s(t), b_h(t-1))\}\} \\ \text{Boundary: } b_h(t-1), \{c_h(t) > c_m(t)\} \ \& \ \{J_{h,m}(s(t), b_h(t-1)) \geqslant \max\{J_{h,m}(s(t), \\ \qquad a_h(t)), J_{h,m}(s(t), a_m(t))\}\} \\ \text{Human: } a_h(t), \qquad \text{其他.} \end{cases} \end{aligned} \tag{4.8}$$

公式(4.8)表示机器介入人控制中的仲裁函数。根据对可信度评估和目标函数的大小判断谁是当前的决策者。(4.8)中的 $c_m(t)$ 和 $c_h(t)$ 代表机器动作 $a_m(t)$ 和 $a_h(t)$ 的可信度评估。考虑到贝叶斯神经网络的概率特性，仲裁函数采用 MC dropout[145] 方法来衡量决策可信度，如第3章公式(3.9)。更具体地，如果机器决策动作的可信度 $c_m(t)$ 高于人类决策动作的可信度 $c_h(t)$，并且 $a_m(t)$ 所对应的目标函数大于 $a_h(t)$ 和 $b_h(t-1)$ 对应目标函数中的大者，那么智能机器就成为决策者；如果人类决策动作的可信度 $c_h(t)$ 高于机器决策动作的可信度 $c_m(t)$，并且 $b_h(t-1)$ 所对应的目标函数大于 $a_h(t)$ 和 $a_m(t)$ 对应目标函数中的大者，那么最优决策取在边界上；否则决策者就是人类的伙伴。

接下来，我们给出机器介入人控制优化算法的具体流程。本节在使用介入控制算法求解人机序贯决策问题的基础上，引入了人的自主边界判定网络。因

此算法的优化目标不仅是优化与决策动作直接相关的策略函数，还包括学习优化间接影响决策动作的人的自主性边界。首先，我们初始化人的自主性边界信息。在动态演化过程中，人类伙伴和智能机器将分别给出实时的决策动作 $a_h(t)$ 和 $a_m(t)$，以及相应的可信度分析。之后，根据优化(4.7)中的约束(4.7g)过滤人类动作和机器动作。考虑机器介入人控制优化算法的共同目标函数，利用仲裁函数(4.8)输出最终决策动作 $a(t)$。最后，基于式(4.5)中的自主性边界学习方法和算法4.3完成 t 时刻人的自主性边界更新，重复循环直到训练结束。

算法 4.4　机器介入人的控制优化算法

初始化：随即初始化机器代理的策略网络 p^m 及其网络参数 θ；初始化人的自主性边界 $\bar{B}_h$;
输入：系统状态 $s(t)$;
输出：最终决策动作 $a(t)$;
while 未达到最大训练时间步 **do**
    for 未到达终止状态 **do**
        机器代理根据策略网络计算出决策动作 $a_m(t)$ 以及基于蒙特卡洛估计 (3.9)计算可信度 $c_m(t)$;
        根据优化问题(4.7)中的约束(4.7g)过滤机器决策动作 $a_h(t)$，以及衡量可信度 $c_h(t)$;
        利用仲裁函数(4.8)输出最终决策动作 $a(t)$;
        基于算法4.3更新维护机器自主性边界信息 $b_h(t)$;
    end
end

4.3.3　仿真实验

在本小节，我们仍然使用 LunarLander，如图4.3所示，并使用 DQN 作为机器代理算法。但本小节更加强调机器介入人控制。我们进行了人类伙伴独立控制和机器介入控制的对比实验，包括奖励、成功率、撞击率、人机动作所占百分比、着陆轨迹对比等参数分析。同样，在正式对比实验之前，先对机器代理算法 DQN 进行预训练，以使得机器代理具有一定程度有效决策的能力。在以下的图中，HOA(Human-Only-Algorithm) 表示只有人类操作员控制，MTHA(Machine-Trade-Human-Algorithm) 表示机器在人类控制过程中进行介入，MTHA-B (Machine-Trade-Human-Algorithm-Boundary) 描述在 MTHA 的基础上增加自主性边界信息的优化算法。

如图4.11和图4.12(a)所示，算法 HOA 的奖励情况最差，这符合我们对于人

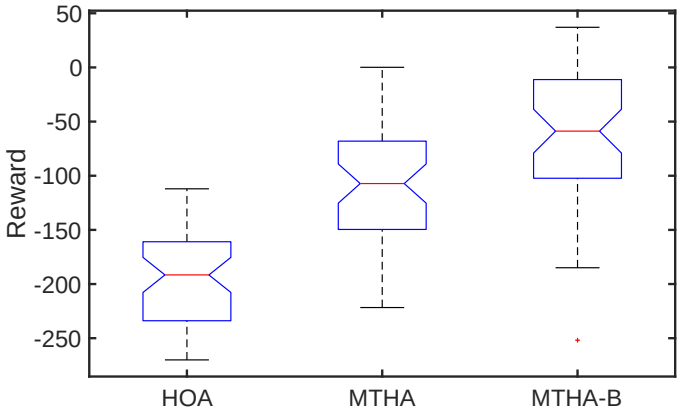


图 4.11　算法 HOA，MTHA，MTHA-B 的平均奖赏对比 (500 episodes)

类控制能力不足的猜想。相比之下，机器介入控制算法 (MTHA 和 MTHA-B) 可以不同程度地增加累积奖励。而在机器介入控制算法中，由于使用了额外的自主性边界信息，使得 MTHA-B 具有更好的决策效果。此处的累积奖励是指每一个完整的情节所获得的累积奖励值。为了避免实验效果的偶然性，我们随机收集了 500 个 episodes 来评估均值和不确定性。对 500 个 episodes 中成功着陆的奖励进行平均，得到图4.12(b)。结合图4.12(a)和图4.12(b)，观察到算法 MTHA-B 不仅在整体 episodes 的奖励上具有优势，而且在成功着陆的 episodes 中也比 HOA 和 MTHA 有更高的奖励。因此，本小节的实验结果有效地证明了机器介入控制优化算法求解人机序贯决策问题的优越性。

在 LunarLander 中，顺利安全地降落在着陆点是游戏成功与否的决定性因素。接下来，我们比较了算法 HOA、MTHA 和 MTHA-B 的着陆成功率和撞击率。图4.12(c)中的成功率满足：MTHA-B>MTHA>HOA。HOA 算法的成功率持续较低，这源于人类不擅长工作的低精度操作，以及机器代理所需学习时间和响应能力的弱点。相较而言，机器对人类的介入控制，使得成功率有了明显的提高。特别是在着陆成功率上，随着 episode 数量的增加，算法 MTHA-B 可以不断提高成功率到 0.55 甚至更高。类似地，对于图4.12(d)所示，撞击率急剧下降是由于机器代理使决策更加精确和稳健。然而，我们发现 MTHA-B 的撞击率高于 MTHA，这似乎是一个不好的信号。事实上，这也和我们对自主性边界信息的使用有关。通过边界信息来衡量机器介入的时机是有利的，但边界的引入带来的一些不稳定性且不容忽视，这需要开发者在成功率和适度的不稳定性之间做出妥协。

从图4.12，我们观察到随着机器决策能力的不断提升，人类可以逐渐将自己不擅长的任务交给或部分交给智能机器来完成，即本小节的主题—机器介入人控制。这也可以从图4.13中的动作百分比得出。从图中可以看出，人类动作在 MTHA 中的比例相对较低。在 MTHA-B 中，最终决策动作是人类动作、机器动

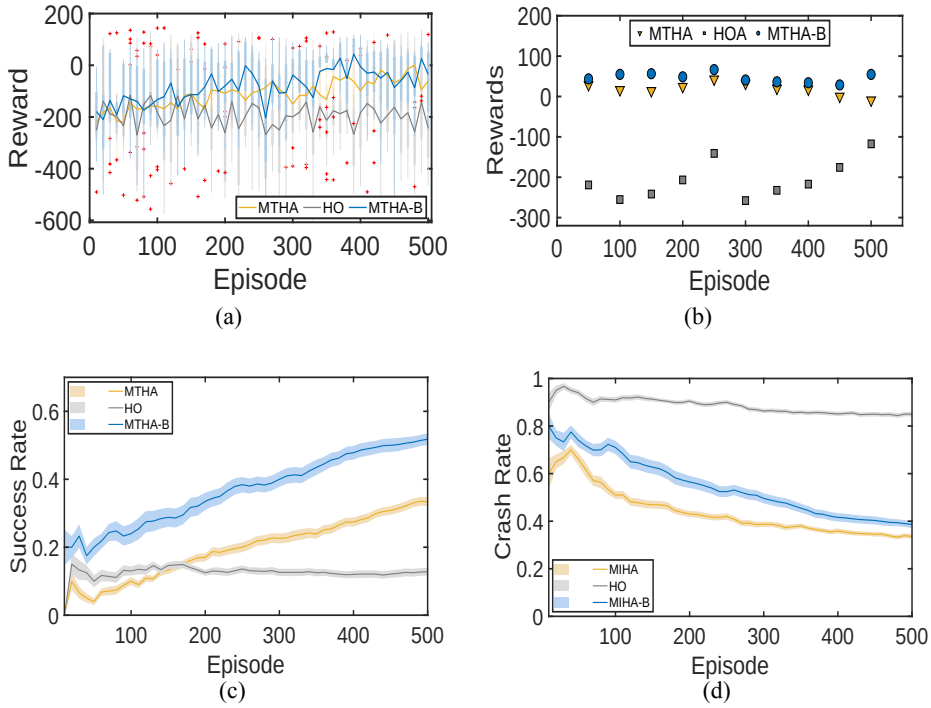


图 4.12　算法 HOA，MTHA，和 MTHA-B 的实验结果对比。(a) 奖赏：实线表示奖赏平均值走势，红色加号代表异常点，阴影表示大多数点所落在的箱体区域；(b) 着陆成功的 episodes 的奖赏值；(c) 成功率: 实线表示成功率的平均值，阴影表示不确定性；(d) 撞击率：实线表示撞击率的平均值，阴影表示不确定性。

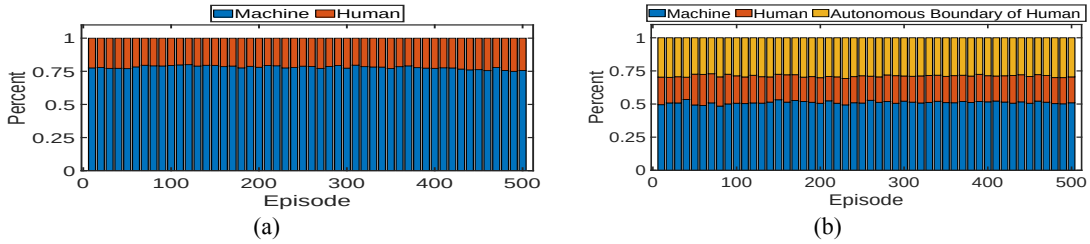


图 4.13　人类动作 a_h 和机器动作 a_m 多占百分比表示：(a) MTHA；(b) MTHA-B

作和人类边界的组合形式。更具体地说，从图4.13(b)可以发现，人的决策动作、人的自主性边界和机器的决策动作以 2 ： 3 ： 5 的比例影响最终决策动作，这是容易理解的且符合我们对人的自主性边界定义。

接下来，我们比较了算法 HOA、MTHA 和 MTHA-B 的着陆轨迹，如图4.14所示。HOA 的着陆轨迹看起来整洁有序，但结合其低成功率和高撞击率（如图4.12）和时间步长（图4.15）可以得出结论：由于人为操作的不精确性，HOA 往往会直接崩溃和快速失败。其次，我们发现 MTHA 算法的着陆轨迹比较凌乱，但成功率有所提高，这主要是由机器介入控制所带来的。进一步地，MTHA-B 算法轨

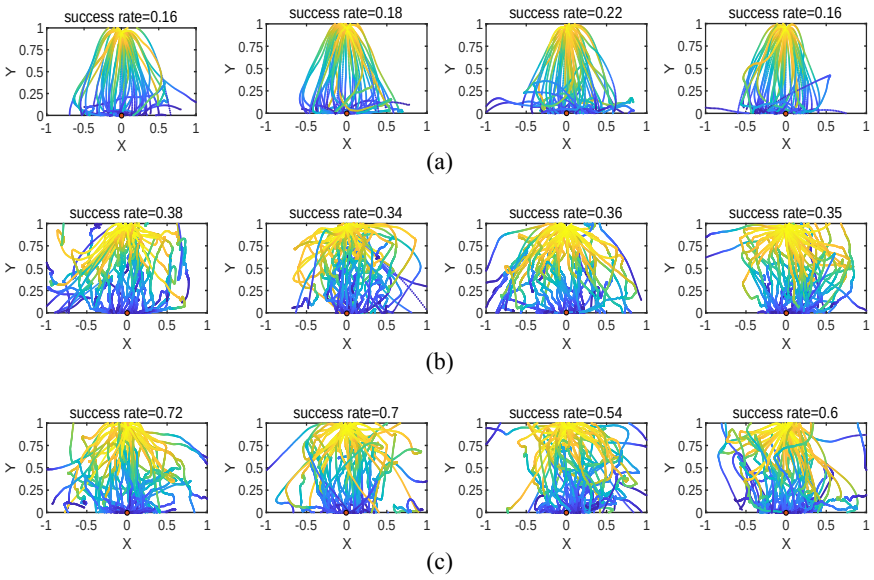


图 4.14　不同算法对应着陆轨迹的对比：(a) HOA；(b) MTHA；(c) MTHA-B。

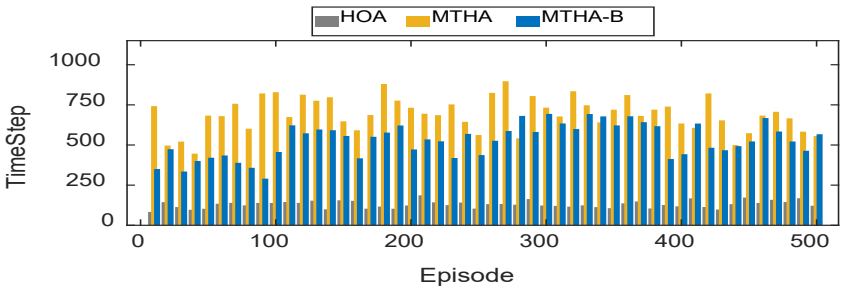


图 4.15　算法 HOA，MTHA 和 MTHA-B 在每一条 episode 中的时间步长走势

迹的凌乱度介于 HOA 和 MTHA 之间，MTHA-B 的成功率和运行时间步长大大提高，更符合算法目标 (更好更快地完成任务)。关于更好更快，还体现在算法的运行时间步长上。算法 MTHA 的运行时间步长最长，这与较大比例的机器动作有关 (图4.13(a))。也就是说，机器以较慢的学习率换取成功率的部分提高。算法 MTHA-B 的运行时间介于三者之间。因此可以说，MTHA-B 不仅提高了成功率，降低了撞击率，还加快了任务完成速度。

最后，为了便于理解本小节机器介入控制算法求解人机序贯决策问题的过程，我们给出最终的决策动作 $a(t)$，机器决策动作 $a_m(t)$，与人类决策动作 $a_h(t)$ 之间的对应关系，如图4.16。从图4.16(a)，我们观察到最终决策动作 a 的值总是介于机器决策动作 a_m 和人类决策动作 a_h 之间。从图4.16(b)中，最终的动作 $a(t)$ 是在 $a_m(t)$、$a_h(t)$ 和 $b_h(t-1)$ 之间进行选择，其中 $b_h(t-1)$ 是人类的自主性边界，即增加额外的自主性边界信息来优化和判断最终的决策信号。

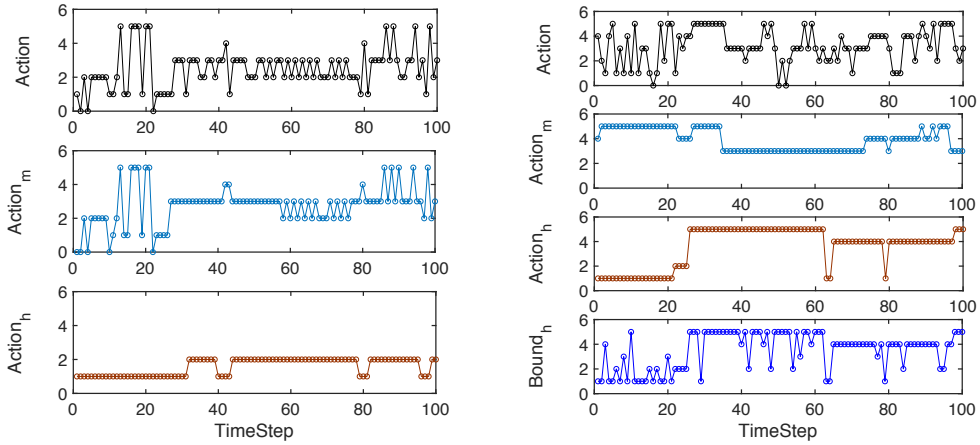


图 4.16　算法 MTHA 和 MTHA-B 的决策动作对应关系，左侧表示一条完整 episode，右侧表示前 50 步。(a) MTHA：从上至下分别是最终决策动作 a，机器决策动作 a_m，人类决策动作 a_h；(b) MTHA-B：从上至下分别是最终决策动作 a，机器决策动作 a_m，人类决策动作 a_h，人类自主性边界。

4.4　本章小结

本章面向人机序贯决策问题，研究了介入控制优化算法，包括人介入机器的问题场景和机器介入人的问题场景。通过判定人和机器的自主性边界，该边界取决于由系统状态和人机动作等所决定的优化目标。接着，本章基于自主性边界衡量介入触发的时机，进而优化介入控制算法，改善提升人机序贯决策求解过程中的控制性能。最后，通过仿真案例分别验证了在人介入机器、机器介入人两种人机序贯决策模式下的有效性和优越性。

第 5 章　人机序贯决策: 基于自主性边界优化设计共享控制算法

本章面向人机序贯决策，研究共享控制下的自主性边界判定方法，以及考虑将获得的自主性边界信息应用于共享控制算法的优化设计中。不同于第 4 章所述的“介入”控制方法求解人机序贯决策，本章介绍“共享”控制算法，其中人与机器处于较为平等的地位，共享控制权限的目的是利用各自优势取得单纯人或机器难以取得的整体效果，关键词涉及“融合”。

5.1　引言

人机序贯决策中的共享控制问题需要将人类动作和机器动作进行混合，意味着系统需要对人类决策动作和机器决策动作进行分析评判，并且根据评判结果对混合程度进行调整。如当机器决策质量不高时，混合结果偏向人类决策；当人类出现明显错误时，混合结果偏向机器决策。极端情况下，共享控制将退化为第4章的介入控制，同时，这也意味着共享控制所面临的行动空间要比介入控制大得多，显然这对优化求得最优解是有利的，但是也增加了计算难度和复杂度。然而如何掌握共享控制的动态混合程度和混合方式，实现对人类决策和机器决策的有效集成混合，则是人机序贯决策问题求解非常重要的一个领域。

关于共享控制已经具有一定的研究基础。文献 [74] 以自动驾驶为应用背景，确定了共享控制中的七个原则。[76] 描述了人与机器人物理交互中的共享控制设计，具体包括: 意图检测，仲裁和沟通/反馈三个方面。[147] 提出了一种新颖的方法来设计共享控制，其通过观察和使用 Koopman 运算符来学习有关用户交互的动态信息，并且使用学习到的模型，定义一个优化问题以计算给定任务的最佳策略。[144] 使用混合共享控制（BSC）架构来解决应急响应和搜索与救援（SAR）任务中由动态不确定环境和认知差异导致的人机团队协作问题，并使用具有输入等待时间和错误里程计反馈的差动驱动机器人研究该架构在受限的动态环境中的性能。[148] 为共享自主下的辅助遥操作期间的意图推理提供了数学公式表达，利用递归贝叶斯滤波方法可以建模并融合多个非语言的观察结果，从而可以在没有明确交流的情况下概率性地推断出用户的预期目标，并且对人类操作者的动作建模并整合为具有可调整合理性的目标定向动作。已有研究具有较大的参考学习价值，但也存在不足，比如普遍存在混合（仲裁）参数判定中的阈值固定不灵活，导致混合参数在实时性上缺少准确度。

综上所述，本章提出了基于自主性边界的混合参数优化设计方案，通过自适

应调节混合参数大小直接影响最终待执行动作的生成。考虑了人机动作的融合程度，使得最优解在人的动作空间和机器的动作空间所共同张成的扩展空间中出现，为决策质量的提升提供了更大的可选择空间。更具体地，我们将自主性上界和自主性下界分别与仲裁机制下判断仲裁参数的上下阈值进行关联，避免以往固定阈值带来的调参局限，从而生成更加适应动态环境的仲裁参数，完成人机动作的更好融合。仲裁判断还需考虑人类伙伴的意图，通过意图推理获得系统将要完成的任务目标。

本章结构安排如下，第5.2节介绍人机序贯决策中的共享控制问题描述及建模。第5.3节介绍面向人机序贯决策的共享控制下的自主性边界判定方法。第5.4节介绍面向人机序贯决策中共享控制下的意图推理。第5.5节介绍利用自主性边界信息优化设计人机序贯决策中的共享控制。最后第5.6节给出实验仿真设计和结果分析。

5.2　人机序贯决策中的共享控制框架设计

考虑使用共享控制方法求解人机序贯决策问题时，“仲裁”同样是一个核心的概念，它决定何时采取机器代理的决策动作、何时采取人类伙伴的决策动作、以及何时对机器决策动作和人类伙伴决策动作进行融合。本章给出共享控制下仲裁的一般性定义如下:

定义 5.1 (仲裁)　仲裁是指在人机共享控制系统中，将人类伙伴的策略和智能机器的策略进行混合，得到超越单一策略决策效果的混合策略，其中仲裁参数取决于环境参数和用户需求，可表示为:

$$a(t) = f(a_h(t), a_m(t), s(t), g(t), c(t)) \tag{5.1}$$

其中，使用 $f(\cdot)$ 表示任意有效的仲裁函数形式。$a_h(t)$，$a_m(t)$ 分别表示人类伙伴的决策动作和智能机器代理的决策动作。$s(t)$ 指与被控对象相关的实时状态，$g(t)$ 表示用户所需的任务目标，$c(t)$ 代表其他影响仲裁结果的因素，比如置信度、不确定性等信息。

基于上述一般性定义，本章关心仲裁机制在基于意图推断的人机共享控制系统中的应用，如图5.1所示。在此类系统中，机器通过观察人类伙伴的决策动作进而推测人类本身意图达到的目的，机器代理结合推测目标和自身策略状况，对被控对象的实时环境状态给出动作决断。之后，机器决策动作和人类决策动作同时进入仲裁阶段，由仲裁函数给出最终的决策动作。

关于人机序贯决策问题的表示形式，本节将人机序贯决策问题的共享控制问题形象化表达为如下形式:

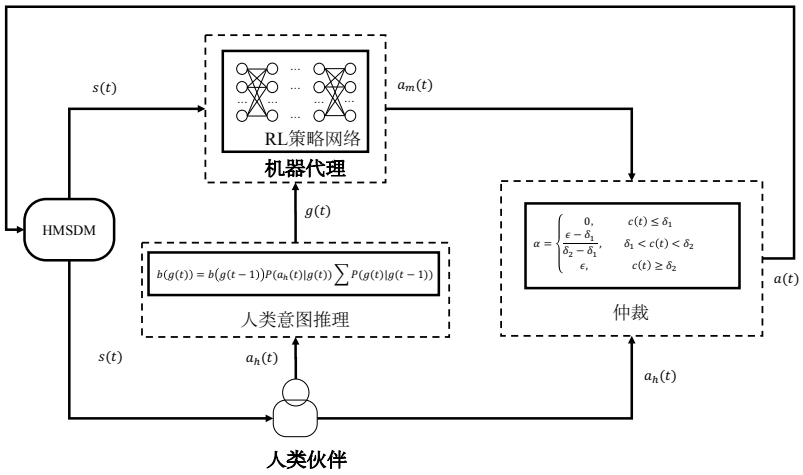


图 5.1　面向人机序贯决策的共享控制框架

$$\max_{\theta} \boldsymbol{J}^r(s(t), a(t)) = \int_{t=t_0}^{t_0+T} r(s(t), a(t)) \tag{5.2a}$$

$$\text{s. t. } a(t) = f^a(a_h(t), a_m(t), c(t)) \tag{5.2b}$$

$$a_m(t) = p(s(t), g(t); \theta) \tag{5.2c}$$

$$\{g(t), c(t)\} = \text{Infer}(a_h(t)) \tag{5.2d}$$

$$s(t+1) = f^d(s(t), a(t)) \tag{5.2e}$$

$$C(s(t), a_m(t), a_h(t)) \leqslant 0 \tag{5.2f}$$

$$t = 0, 1, 2, 3, ...$$

其中 $\boldsymbol{J}^r(s(t), a(t))$ 是由系统状态和决策动作所决定的累积奖励值。$a(t)$ 是共享控制系统最终的决策动作，Infer(·) 是意图推理函数模块，其输出是推断的用户目标 $g(t)$ 和推测的置信度大小 $c(t)$。$p(\cdot)$ 表示智能机器代理的策略函数，其参数用 θ 表示，用来输出机器代理的决策动作。

通常情况下，仲裁函数可具有如下的典型线性混合形式：

$$f^a(a_h(t), a_m(t), \alpha) = (1-\alpha)a_h(t) + \alpha a_m(t) \tag{5.2g}$$

其中参数 α 的值应根据意图推理的置信度 $c(t)$ 而定，可具有如下的经典形式

$$\alpha = \begin{cases} 0, & c(t) \leqslant \epsilon_1 \\ \frac{\epsilon_3 \times (c(t) - \epsilon_1)}{\epsilon_2 - \epsilon_1}, & \epsilon_1 < c(t) < \epsilon_2 \\ \epsilon_3, & c(t) \geqslant \epsilon_2 \end{cases} \tag{5.3}$$

其中 ϵ_1 和 ϵ_2 表示某事先确定的下阈值和上阈值：当意图推理置信度 $c(t)$ 小于下阈值 ϵ_1 时，表示机器决策变得不可信，决策系统采用人的决策；当 $c(t)$ 大于上

阈值 ϵ_2 时，采用某事先确定的权值 $\epsilon_3(\epsilon_3 < \epsilon_2$，如 $\epsilon_3 = 0.7\epsilon_2)$，对人类伙伴和机器代理的决策动作进行加权混合；然而，当 $c(t)$ 处于上阈值 ϵ_2 和下阈值 ϵ_1 之间时，则利用依赖于 $c(t)$ 的动态权值 $\frac{\epsilon_3\times(c(t)-\epsilon_1)}{\epsilon_2-\epsilon_1} < \epsilon_3$，进而对人类决策动作和机器决策动作进行加权融合。

5.3　面向人机序贯决策的共享控制下的自主性边界判定

在求解人机序贯决策问题过程中，相比于介入控制方法，由于共享控制的决策空间得到了极大拓展，在实际情况允许使用共享控制方法时，共享控制算法具有比介入控制算法更优越的可能性。而在共享控制策略的设计中，我们首要讨论的问题即是关于自主性边界的判定。但需要注意的是，在共享控制中，尽管人与机器各自的自主性边界也仍有其重要价值，但我们本质上关心的是由人和机器所形成的共同联合边界。

定义 5.2 (共享控制下的自主性边界)　共享控制下的自主性边界是指按照有益于人机混合系统共同优化目标的方向，机器智能和人类智能进行共同决策和行动的范围界限。

由于此处考虑的是共享控制下的自主性边界，并且本章是对人机共享控制的融合设计，因此我们不强调其是谁的自主性边界，而是从混合的角度，将自主性边界分为自主性上界和自主性下界。在利用这两个边界信息时，系统的决策动作需处于自主性上界和自主性下界之间，如果低于自主性下届或超过自主性上界时，需视具体情况加以约束或过滤。

关于共享控制下的 t 时刻的自主性上界 $\overline{b}(t)$(或自主性下界 $\underline{b}(t)$) 的判定，考虑形式化表示为在满足系统状态 $s(t)$ 和被控对象的约束条件 $C(s(t), a(t))$ 下，寻找使得人机混合共同目标 $J_{h,m}(s(t), a(t))$ 最好 (或最差) 的行动，以人机序贯决策目标是最大化 $J_{h,m}(s(t), a(t))$ 为例，上述描述可形式化写为下式，

$$\overline{b}(t) = \arg \max_{a_(t)\in\mathcal{A}_h(t)\times\mathcal{A}_m(t)} J^b_{h,m}(s(t), a(t)) \tag{5.4a}$$

$$\text{s. t. } C(s(t), a(t)) < 0 \tag{5.4b}$$

和

$$\underline{b}(t) = \arg \min_{a_(t)\in\mathcal{A}_h(t)\times\mathcal{A}_m(t)} J^b_{h,m}(s(t), a(t)) \tag{5.5a}$$

$$\text{s. t. } C(s(t), a(t)) < 0 \tag{5.5b}$$

其中 $J^b_{h,m}(s(t), a_m(t))$ 是人和机器的共同目标函数，可根据具体实施场景和算法将目标函数定义为不同表达形式，比如累积奖赏（成本）函数

$$J^{b}_{h,m}(s(t),a(t)) = J^{r}(s(t),a(t)) = \int_{t=t_0}^{t_0+T} r(s(t),a(t)) \tag{5.6}$$

式(5.4)和(5.5)给出了最大化目标函数的优化示例，因此我们需要找到满足约束条件下，使得优化目标最好（大）的人类动作或机器动作作为共享控制下的自主性上界，反之，作为自主性下界。当系统追求最小化目标函数，将 $J^{b}_{h,m}(s(t),a(t))$ 最大化或者将直接给 $J^{b}_{h,m}(s(t),a(t))$ 加上负号的形式即可实现。$s(t)$ 是系统状态，$a(t)$ 是共享控制下的混合决策动作，上述被控对象的约束条件 $C(s(t),a(t))$ 为一般性表达式，需视具体场景而定（如在 MPC 中可以硬约束的方式显示出现在优化问题中，或者 MDP 中通过目标函数隐式表达）。

算法 5.1　共享控制下的自主性边界判定

初始化: 初始化自主性边界 $B = \{\overline{b}(t), \underline{b}(t)\}$;
输出: 共享控制下的自主性下界 $\underline{b}(t)$ 和自主性上界 $\overline{b}(t)$;
while 未达到训练结束条件 **do**
    for 未到达终止状态 **do**
        输入：系统状态 $s(t)$;
        根据当前时刻系统状态，机器代理计算出决策动作 $a_m(t)$，同时人类代表也给出决策动作 $a_h(t)$;
        根据约束条件 (5.4b) 分别对当前时刻的机器和人类的动作进行检查;
        将满足约束条件的人类输入与初始化中的自主性边界信息进行目标函数的比较，如果 $J_{h,m}(s(t),a(t)) < J_{h,m}(s(t),\underline{b}(t))$，则依据式(5.4)（式(5.5)）更新自主性下界（自主性上界），否则保持不变;
    end
end

我们考虑上述式(5.4)和(5.5)的优化问题，可将具体思想描述如算法5.1所示，在算法5.1中，共享控制下自主性边界的初始化可根据被控对象及当前已有可用信息获得，与算法4.1和4.3类似，可借鉴神经网络中的随机初始化作为无其他先验信息时的解决办法 (利用随机搜集到的经验样本轨迹等办法)。在系统动态演化过程中，对于实时输入的系统状态 $s(t)$，算法决策模块分别给出与此系统状态相应的机器决策动作 $a_m(t)$ 和人类决策动作 $a_h(t)$。之后利用自主性边界信息 $\overline{b}(t), \underline{b}(t)$ 对满足约束的决策动作进行比较，目的是更新使得目标函数更大(5.4a)或

更小(5.5a)时与决策动作对应的共享控制下的自主性边界 (或自主性下界)，重复进行下去直至训练结束。

5.4　面向人机序贯决策的共享控制下的意图推理设计

在人机序贯决策问题求解过程中，为了使得人类伙伴提供有效的帮助，智能机器代理首先需要具备预测或推断人类伙伴预期目标的能力。共享控制框架，即是包括推断人类伙伴意图模块的设计方法。本节描述面向人机序贯决策的共享控制下的意图推理设计。例如，假设环境具有一组离散目标集 $G = \{g_1, g_2, \cdots, g_N\}$，并且离散目标集对于人类用户和机器代理是可见的。但是在动态演化过程中，智能机器代理不清楚人类伙伴实时想要实现的具体目标 $g(t)$。为了处理这种情况，智能机器代理对于人类伙伴的意图推断模块即应运而生，它可以推断出用户实时目标 $g(t)$ 的相关信息。

本章基于贝叶斯规则，以 t 时刻的人类伙决策动作作为观测值，将 t 时刻对应推测的目标信念表示为:

$$\begin{aligned} b(g(t)) &= P(g(t)|a_h\{0:t\}) = P(g(t)|a_h\{0:t-1\}, a_h(t)) \\ &\propto P(g(t), a_h\{0:t-1\}, a_h(t)) \\ &\propto P(a_h(t)|g(t), a_h\{0:t-1\})P(g(t), a_h\{0:t-1\}) \\ &\propto P(a_h(t)|g(t), a_h\{0:t-1\})P(g(t)|a_h\{0:t-1\}) \end{aligned} \tag{5.7}$$

假设 5.1　[148] 假定已知当前时刻的目标估计值，当前时刻人类决策动作的观测值和历史时刻人类决策动作的观测值是条件独立的，即

$$P(a_h(t)|g(t), a_h\{0:t-1\}) = P(a_h(t)|g(t)) \tag{5.8}$$

假设 5.2　[148] 假定已知历史时刻的目标估计值，当前时刻的目标估计值与历史时刻人类决策动作的观测值条件上独立的，即

$$P(g(t)|g(t-1), a_h\{0:t-1\})P(g(t)|g(t-1)) \tag{5.9}$$

根据假设5.1和5.2，则可将(5.7)写成

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

$$\propto P(a_h(t)|g(t)) \sum_{g(t-1)\in G} P(g(t)|g(t-1))P(g(t-1)|a_h\{0:t-1\}) \tag{5.10}$$

其中用到式(5.7)对 $b(g(t))$ 的定义，类似地可将 $b(g(t-1))$ 表达为

$$b(g(t-1)) = P(g(t-1)|a_h\{0:t-1\}) \tag{5.11}$$

将式(5.11)代入式(5.10)可得:

$$b(g(t) \propto P(a_h(t)|g(t)) \sum_{g(t-1)\in G} P(g(t)|g(t-1))b(g(t-1)) \tag{5.12}$$

式(5.12)即是对于算法结构非常有利的递推公式。接下来，我们根据式(5.12)和最大后验原理，可以将目标推理过程整理成算法5.2。首先提前给定目标集的构成，以及初始时刻的先验分布。算法的输入是样本池中的运动轨迹，并将轨迹中的系统状态和人类决策动作取出，基于公式(5.12)进行目标后验分布的更新。完成后验分布的更新之后，基于最大后验原则寻找到当前时刻最新的推理目标。

算法 5.2　人类决策动作的意图推理

初始化: 目标集 G，及其先验分布 $b(g(0))$;
输入: 经验样本池 Tr 中的运动轨迹;
while 未达到训练结束条件 **do**
    从经验样本池 Tr 中的运动轨迹中取出状态动作对 $(s(t), a(t))$;
    for $g(i) \in G$: **do**
        基于公式更新目标的后验分布
        $b(g(t)) = P(a_h(t)|g(t)) \sum_{g(t-1)\in G} P(g(t)|g(t-1))b(g(t-1))$;
        更新推测目标: $g^*(t) = \arg\max_{g(t)\in G} b(g(t))$;
    end
end

5.5　人机序贯决策的共享控制优化设计

有了第5.3 节面向人机序贯决策问题的共享控制下的自主性边界判定方法以及第5.4节人类决策动作的意图推理设计，针对共享控制算法所求解的人机序贯决策问题(5.2)，我们定义式(5.4)和式(5.5)中的人机序贯决策目标函数为 $J^r(s(t), a(t))$，并给出具体算法步骤如算法5.3所示，

$$J_{h,m}(s(t), a(t)) = J^r(s(t), a(t)) = \int_{t=t_0}^{t_0+T} r(s(t), a(t)) \tag{5.13}$$

注意到，在现有研究中，(5.2g)中的 α 是一个凭经验确定的常数，或者由式(5.3)确定的动态数值，考虑到被控对象所处环境的动态不确定性以及经验设置的固定性，它可能是非常保守的。在本章，我们利用自主性边界的新概念来设计更好的仲裁参数（加权因子）α。因此，人机序贯决策问题求解的优化过程中存在两个任务目标：1）与决策动作直接相关的策略网络；2）间接影响决策动作的共享控制下的自主性边界。其中，包括上述讨论的人机共享控制中自主性边界的判定方法，这里的自主性边界信息需要实时动态更新和维护。然后，基于获得的自主边界信息，对共享控制设计进行优化。共享控制优化设计的一般框架如图5.2中看到。基于框架5.2，我们对人类决策动作进行建模以推断人类想要完成的任务目标。它与系统环境状态连接，作为机器代理模块和人类决策者的输入。基于共享控制下的自主性边界，仲裁模块对机器代理输出的决策信号和人类伙伴决策动作进行优化，最终获得作用在被控对象上的实时决策信号。

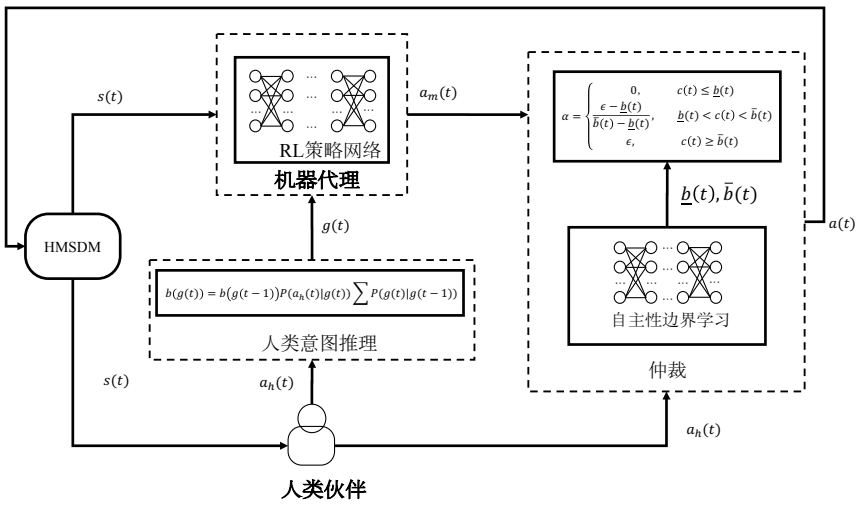


图 5.2　基于自主性边界的共享控制优化框架

不同于现阶段相关研究将影响仲裁参数 α 的上下阈值（ϵ_1 和 ϵ_2）设置为固定常数的做法，本章将参数 α 时所依赖的上下阈值概念自然的与人机系统的自主性上下界的概念进行联系。事实上，通过将 ϵ_1 和 ϵ_2 直接由自主性下界和上界所取得的目标函数值来动态定义，即

$$\epsilon_1 \propto \boldsymbol{J}_{h,m}(s(t), \underline{b}(t)) \tag{5.14a}$$

$$\epsilon_2 \propto \boldsymbol{J}_{h,m}(s(t), \overline{b}(t)) \tag{5.14b}$$

基于自主性边界的共享控制优化算法见算法5.3。5.3是在解决优化问题(5.2)的通用流行办法的基础上，融入共享控制下的自主性边界信息 (由于这里涉及人和机器的融合，人类伙伴和智能机器代理是共同的主体决策者，因此需要考虑共享控制下的的自主性边界)，实现关于此类人机序贯决策问题的进一步优化。算法中关于共享控制的自主性边界的初始化如算法5.1所述。算法的优化目

算法 5.3　基于仲裁机制的共享控制的优化算法

初始化: 初始化共享控制下的自主性边界信息 $B = \{\overline{b}(t), \underline{b}(t)\}$;
输出: 当前时刻仲裁之后的决策动作 $a(t)$, 和此时共享控制下的自主性上界 $\overline{b}(t)$ 和自主性下界 $\underline{b}(t)$;
while 未达到训练结束条件 **do**
    for 未到达终止状态 **do**
        输入系统状态 $s(t)$;
        人类智能根据观察到的系统状态给出决策动作 $a_h(t)$;
        机器代理的意图推理模块根据人类动作预测出任务目标 $g(t)$，同时计算该目标对应的置信度 $c(t)$;
        机器代理将根据系统状态 $s(t)$ 和任务目标 $g(t)$，计算出机器的决策动作 $a_m(t)$;
        基于自主性边界信息，和人机分别的决策动作，按照(5.3)和(5.14)计算 α 值，构建(5.2g)中的仲裁函数，用来计算人机混合决策动作 $a(t)$;
        根据式(5.4)和算法5.1对自主性上界和自主性下界进行优化更新;
    end
end

标有两个: 1) 与决策动作直接相关的策略; 2) 间接影响决策动作的共享控制下的自主性边界。在系统的动态演化过程中，对于被控对象的每个系统状态 $s(t)$，人类伙伴给出有目的性的决策动作 $a_h(t)$，此动作有两个作用: 一则，机器代理用来推测人类想要完成的任务目标 (意图推理模块); 二则，作为和即将生成的机器决策动作 $a_m(t)$ 仲裁混合的人类决策动作。智能机器代理预测出任务目标之后，结合当前时刻的系统状态和当前策略学习情况，计算出实时的机器决策动作 $a_m(t)$。之后，便有了人机共享系统中人和机器的决策动作，则可基于共享控制下的自主性边界信息，人机的决策动作，按照(5.3)和 (5.14)计算 α 值，进而获得(5.2g)中的仲裁函数，用来计算人机混合决策动作 $a(t)$。最后根据式(5.4)和算法5.1对共享控制下的自主性上界和自主性下界进行优化更新，如此重复下去直至训练结束。

由上述讨论可知，算法5.3中对于仲裁函数的定义不再过度依赖实际中难以准确确定的固定超参数，这样，参数给定的规范化和动态性将从本质上有利于算法性能的提升。并且，在共享控制应用于人机序贯决策求解过程时，将自主性边界判定融入到共享控制优化算法中，既可实时更新维护自主性边界信息（相当于对人机决策权限有了进一步划分），又有利于人机序贯决策问题的求解，具有重要的理论研究和实际应用价值。

5.6　仿真实验

本节针对共享控制优化算法实现人机序贯决策进行了实验仿真，主要包括实验设置和实验结果分析。

5.6.1　实验设置

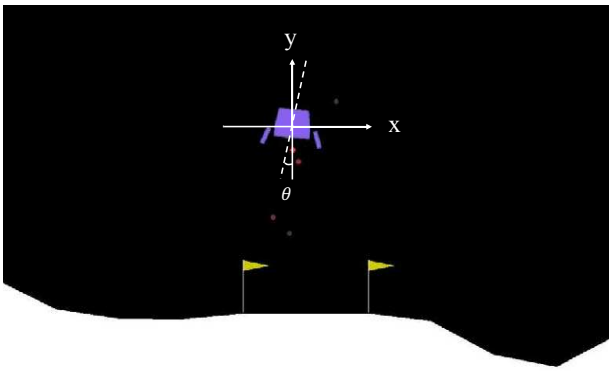


图 5.3　LunarLander 仿真环境

本章在经典 LunarLander 仿真环境的基础上进行修改，我们将经典 Lunarlander 环境中的固定着陆点坐标修改为随机生成的着陆点坐标。如图5.3，在着陆器下降过程中，如果着陆器坠毁或静止下来，则一条完整经历结束，并且获得-100 或 100 的奖赏。着陆器的每条腿接触到地面会有 10 的奖励，打开主引擎时，以每帧-0.3 的奖励消耗燃料（假设燃料时无限的）。着陆器状态向量 $s(t)$ 包括：坐标 $(x(t), y(t))$，速度 $(\dot{x}(t), \dot{y}(t))$，角度 $(\theta(t), \dot{\theta}(t))$，是否着陆 $(leg_l(t), leg_r(t))$ 和着陆点坐标 $h(t)$。

表 5.1　动作值和各引擎开关之间的对应关系

Action Value	Main Engine	Left Engine	Right Engine
0	OFF	OFF	ON
1	OFF	OFF	OFF
2	OFF	ON	OFF
3	ON	OFF	ON
4	ON	OFF	OFF
5	ON	ON	OFF

离散动作集合为 $\{0, 1, 2, 3, 4, 5\}$，具体对应关系如表5.1所示，其中 0（向左和向下）表示主引擎和左引擎关闭，右引擎打开；1（向下）表示所有引擎均关闭；2（向右和向下）表示主引擎和右引擎关闭，左引擎打开；3（向左和向上）表示主引擎和右引擎打开，左引擎关闭；4（向上）表示主引擎打开，左引擎和右引擎关

闭。5（向右和向上）表示主引擎和左引擎打开，右引擎关闭。故而，LunarLander 仿真环境就变成了通过操控着陆器的三个引擎，进而实现着陆器在着陆区域（黄色小旗之间）的安全着陆。

机器的自主决策能力基于 DQN 进行刻画，具体的 DQN 见2.3即通过 DQN 衡量值函数大小并且得到其当前时刻的最优决策动作。图5.4给出值函数估计网络示意图，其中 Dropout 层即是为了刻画自主性边界的不确定性而引入的机制。基于此，本小节实验环节能够对自主性边界的概率分布进行估计，从而应用于策略的优化过程中。

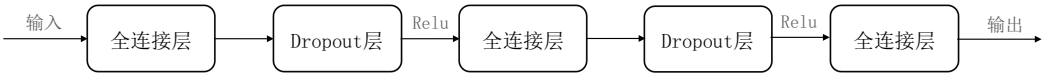


图 5.4　引入 Dropout 机制的值函数估计网络示意图

5.6.2　实验结果

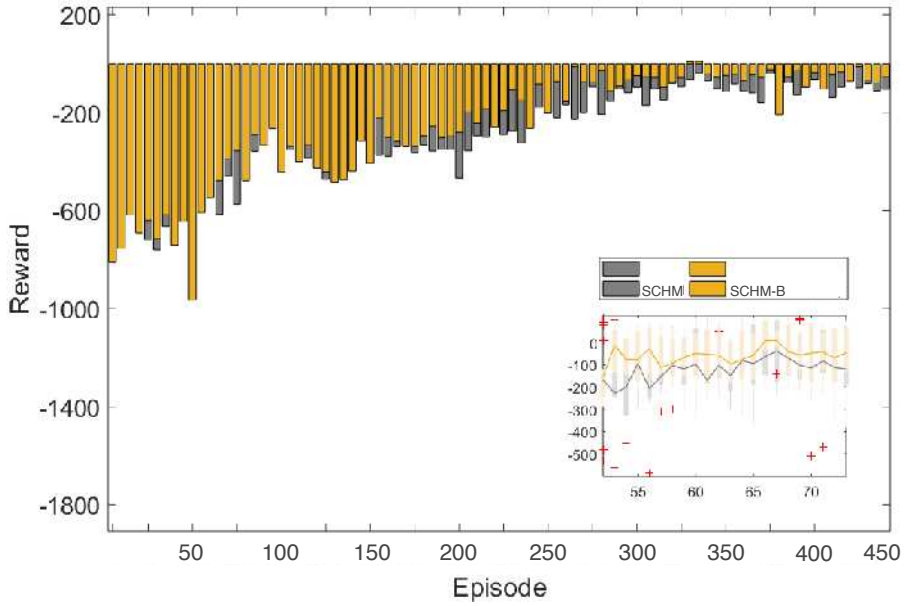


图 5.5　算法 SCHM 和 SCHM-B 的奖赏对比

首先需要说明的是，本章使用 SCHM(Shared Control of Human-Machine) 表示“人机共享控制算法”，为了清晰比较算法效果，我们令 SCHM-B 表示基于自主性边界的人机共享控制算法，对应于优化算法5.3。接下来，分别从奖励，仲裁参数，自主性边界，以及一条完整经历决策动作生成方面分别分析实验结果。更进一步地，由于本次实验并非是采取一次偶然的实验结果，而是进行了 50 次的重复实验，并且进行了平均，因此足以说明结果的可靠性。

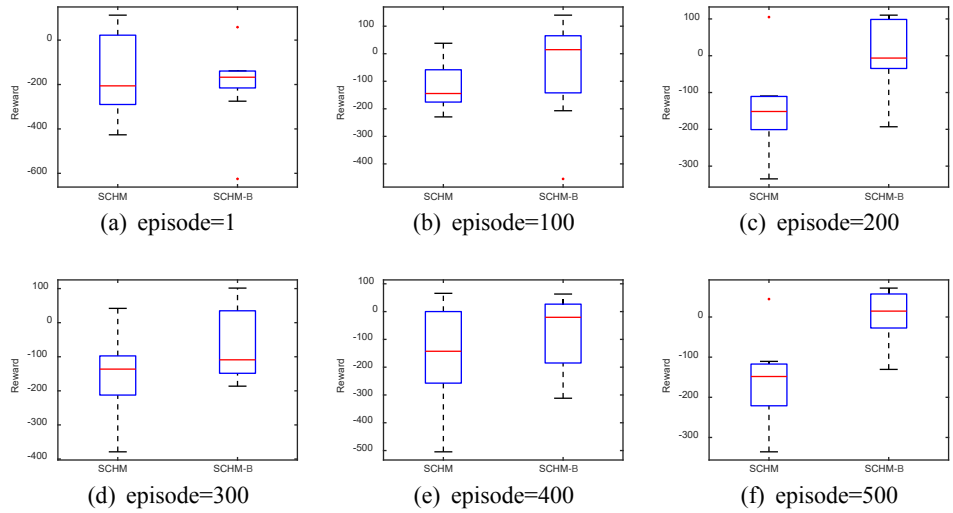


图 5.6　算法 SCHM 和算法 SCHM-B 在不同 episode 的奖赏对比：(a) 0；(b) 10；(c) 20；(d) 30；(e) 40；(f) 50。

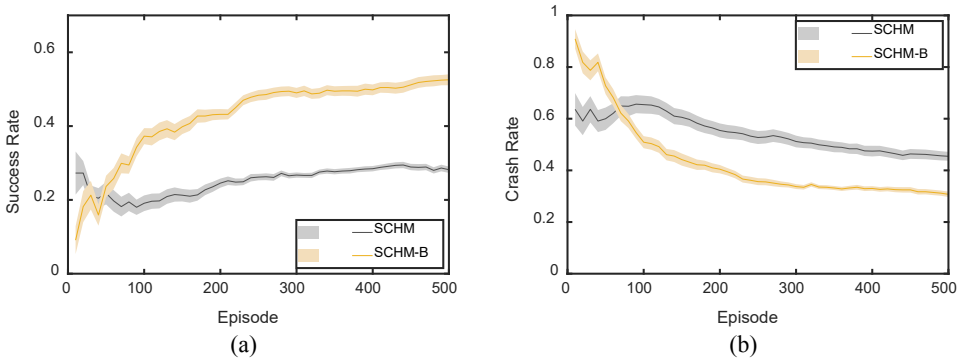
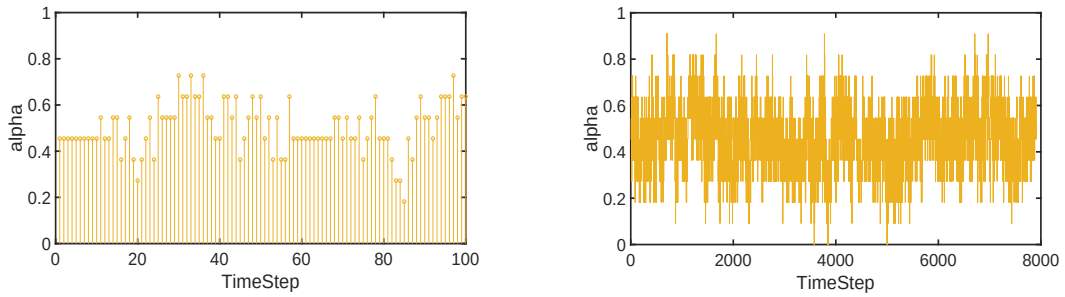


图 5.7　算法 SCHM 和算法 SCHM-B 的成功率和撞击率对比结果

值得注意的是奖励走势情况，从图5.5和5.6可以看出，在进行一定的预训练 (前 30 个 episodes) 之后，算法 SCHM-B 的奖励情况基本上能够实现高于 SCHM 的效果。此外，应该注意的是，LunarLander 环境本身需要消耗燃料，并且我们的着陆点坐标不是固定的 (如第5.4节和第5.6.1节所述)，降落点坐标是由机器代理根据人类决策动作推断得出的，并且还必须对人类决策动作对应的燃料消耗进行量化，因此，即使一条完整经历即使是成功的，其奖励累积值也大约在 0 左右浮动。特别地，在奖励趋势5.6(a)到5.6(f)的奖励结果显示，我们所提的仲裁优化方法 SCHM-B 比 SCHM 更优。

其次，我们讨论游戏的成功率和崩溃率，如图5.7(a)所示。从图中可以看出，算法 SCHM-B 的成功率可以达到约 0.5，而算法 SCHM 的成功率仅为少于 0.3。需要注意的是，撞击率不等于失败率（即 $succ + crash \neq 1$）。图5.7(b)中的算法 SCHM-B 可以将撞击率降低到 0.3，并且相应的 SCHM 撞击率约为 0.5。

接下来，我们考虑基于自主性边界的人机共享控制算法对于仲裁参数的影



(a)

图 5.8　仲裁参数 α 的对比

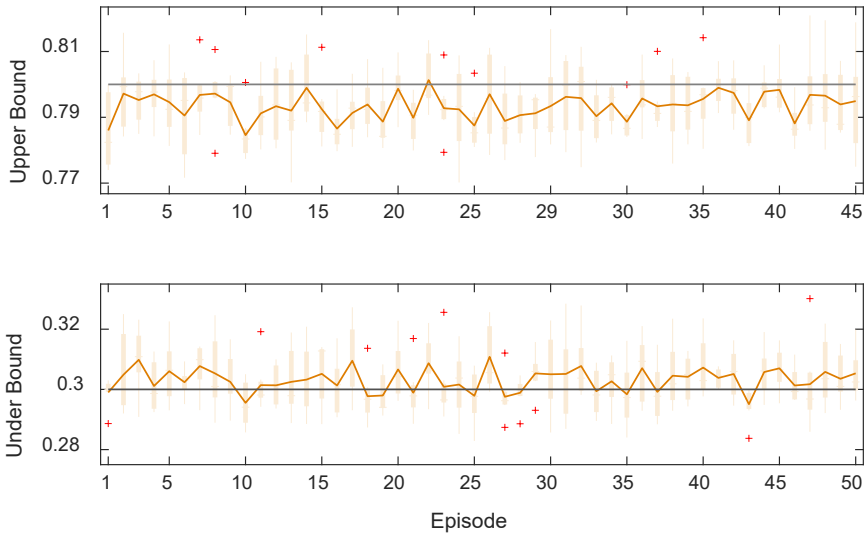


图 5.9　算法 SCHM-B 中的自主性上界和自主性下界

响。仲裁参数 α 决定了人类决策动作与机器决策动作的混合程度（式(5.3)）。图5.8(a)展示了以往固定的 α 和本章所提算法的自适应 α 之间的对比。比如以 $\alpha = 0.5$ 为例，固定的 α 值意味着仲裁器将持续以 0.5 的权重混合机器决策动作和人类决策动作。本章所提的方法 SCHM-B 则是仅使用 0.5 作为初始值，系统根据实时动态环境自适应地调整 α 的大小。此外，本章采用自主性边界的概念(式(5.4)，(5.5)来改善 α 的自适应取值。图5.9展示了自主性边界的变化情况。

最后，为了便于理解，我们给出了目标推理可信度，仲裁参数 α 值和机器决策动作，人类决策动作，以及最终决策动作之间的对应关系，如图5.10所示。从图中可以看出，当目标推理可信度较低 (比如 $\leqslant 0.3$) 时，对应的 α 取值较小 (接近 0)。此时，通过仲裁获得的混合动作就会更倾向于人类决策动作，这与我们的直观感觉相吻合。也就是说，当机器的投机质量不高时，我们更愿意相信人类作出的决策选择。当目标推理置信度较高（比如 $\geqslant 0.5$）时，对应的 α 取值趋向于 1。此时，通过仲裁获得的混合动作更倾向于机器决策动作，这是合理的，因为

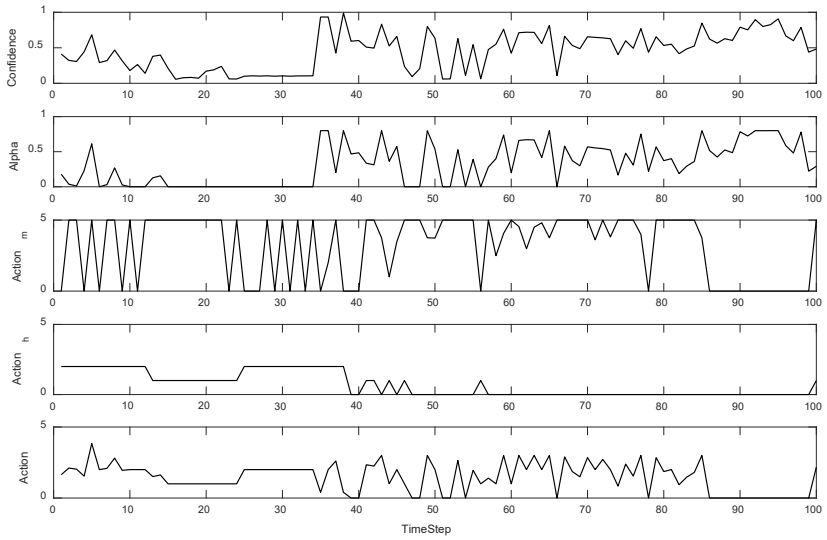


图 5.10　决策动作的轨迹对应关系，包括: 目标推理 g 可信度、仲裁参数 α、机器决策动作 a_m、人类决策动作 a_h、最终决策动作 a

此时智能机器是可信赖的。另外需要说明的是，考虑到不断变化的目标和事先未知的机器，我们保留了一定的可操控空间（$\alpha \leqslant 1$）供人类参与。

5.7　本章小结

本章研究了利用共享控制优化算法求解人机序贯决策问题。不同于上一章节的介入控制算法中的“切换”，本章的共享控制算法更加强调“混合”。本章提出了共享控制下的自主性边界判定办法，并且将此自主性边界与仲裁参数获取的阈值相联系，进而实现了共享控制下的仲裁参数优化，以及改善共享控制实现人机序贯决策问题的决策性能。最后在仿真实验中，以无边界信息的共享控制办法为参照实验，对本章方法进行验证。结果表明仲裁参数的优化使得无论在奖赏值、成功率、撞击率等都有较好的提升效果。

第 6 章　人机序贯决策: 利用自主性边界不确定性优化设计混合智能算法

本章面向人机序贯决策问题，研究自主性边界的不确定性对介入控制和共享控制算法的影响，即从集成的角度使自主性边界以概率分布的形式呈现出来并应用于求解人机序贯决策的混合智能优化算法中。本章是对第4章和第5章的深入研究，更加符合人们对于求解人机序贯决策问题过程中关于人机决策边界的模糊性考虑。

6.1　引言

面向人机序贯决策问题，本文第4章和第5章介绍了介入控制方法设计和共享控制方法设计，分别提出了自主性边界的判定办法，以及利用得到的自主性边界信息优化设计介入控制和共享控制。在继续探索研究的过程中，我们发现自主性边界这一额外信息固然有用，但需建立在准确合理的前提下，如若出现错误，非但不能达到优化的效果，反而存在拖后腿的可能性，而这些即是自主性边界的单值估计可能会造成的。

针对上述可能存在的问题，本章提出了基于贝叶斯神经网络的不确定性估计办法，获得自主性边界的概率分布信息并用于决策动作生成，利用自主性边界的不确定性优化设计人机混合智能算法，使得决策动作的优化存在更多选择。具体地，利用 dropout 机制实现对贝叶斯神经网络的近似，基于采样输出思想获得自主性边界的概率分布形式，仲裁判断过程中依概率采样获得的自主性边界信息输出决策动作。同样地在算法动态演化过程中存在两个目标：直接影响执行动作的策略网络学习；间接影响执行动作的自主性边界维护。只是不同于上述第4章和第5章，这里更新维护的是自主性边界的概率分布信息。利用自主性边界的不确定性优化人机混合智能系统中的介入控制和共享控制，也更加符合人们对决策边界的模糊性思考。

本章结构安排如下，第6.2节介绍自主性边界的不确定性估计方法。第6.3节介绍面向人机序贯决策的混合智能优化算法，其中包括基于自主性边界不确定性的介入控制优化算法和基于自主性边界不确定性的共享控制优化算法。第6.4节给出实验设计和实验结果分析，其中分别包括介入控制实验结果和共享控制实验结果。

6.2　自主性边界的不确定性估计

从概率论的角度看，使用单点估计权重处理分类是不太合理的。相较而言，贝叶斯神经网络对于过拟合更加鲁棒，并且可以从小数据集中学习。贝叶斯方法将其参数以概率分布的形式呈现以提供对参数的不确定性估计。本节面向人机序贯决策问题，关于利用自主性边界信息判定 (第4.2.1节，第4.3.1节和第5.3节) 优化介入控制和共享控制的基础上，继续讨论自主性边界的不确定性估计办法。

6.2.1　不确定性估计

相较于普通神经网络，贝叶斯神经网络将预测中的不确定性估计纳入其中。可以通过在模型参数上放置概率分布来估计不确定性。贝叶斯建模中，存在两种类型的不确定性：偶然不确定性和认知不确定性。

- 偶然不确定性：与观测中固有的噪声有关，存在于数据收集方法中，并且不随数据集大小的增大而降低。
- 认知不确定性：表示模型本身的不确定性，包括模型参数，模型结构等。通过增大数据样本集，可降低或减少此种不确定性。

对于图像分类，x 表示输入图像，y 表示与输入对应的输出标签，那么在贝叶斯框架下，图像分类任务的预测不确定性 $p(y^*|x^*, \boldsymbol{D})$，数据集 $\boldsymbol{D} = {x_i, y_i}_{i=1}^{N}$，可描述如下：

$$p(y^*|x^*, \boldsymbol{D}) = \int p(y^*|x^*, \theta)p(\theta|\boldsymbol{D})d\theta \tag{6.1}$$

其中 θ 表示模型参数，$p(\theta|\boldsymbol{D})$ 是根据数据集 $\boldsymbol{D}$ 所获得的参数 θ 的分布，$p(y^*|x^*, \theta)$ 是基于已获得的模型 θ，输入 x^* 所对应的输出 y^* 的概率分布。并且，$p(y^*|x^*, \theta)$ 和 $p(\theta|\boldsymbol{D})$ 也可分别称作数据不确定性和模型不确定性，二者均属于认知不确定性，因此本章主要讨论的是认知不确定性。

由第2.4.2节可知，对于神经网络而言，式(2.8)中后验分布的求解是难以处理的，因此使用一种近似处理办法—变分推理。

$$p(\theta|\boldsymbol{D}) \approx q(\theta) \tag{6.2}$$

比如，使用一种采样的方法近似

$$\theta^{(i)} \sim q(\theta) \tag{6.3}$$

$$p(y^*|x^*, \boldsymbol{D}) \approx \frac{1}{N_s}\sum_{i=1}^{N_s} p(y^*|x^*, \theta^{(i)}) \tag{6.4}$$

其中 $p(y^*|x^*, \theta^{(i)})_{i=1}^{N_s}$ 中的每个 $p(y^*|x^*, \theta^{(i)})$ 是根据 $q(\theta)$ 采样获得的图像类别概率分布，即根据近似后验分布 $q(\theta)$ 采样 $\theta^{(i)}$，并且计算 $p(y^*|x^*, \theta^{(i)})$ 的均值，则可推

测到预测不确定性 $p(y^*|x^*, D)$。基于上述讨论，通过选择合适且合理的近似推理方案和模型先验 $p(\theta)$，可以找到近似的后验分布 $q(\theta)$ 使得 $p(y^*|x^*, \theta^{(i)})_{i=1}^{N_s}$ 与训练数据集 D 中的数据相一致。

此外，我们介绍另外一种不确定性估计办法——Probabilistic Backpropagation(PBP)[149]。类似地，PBP 不直接对权重进行点估计，而是采用一批一维高斯分布，其中每一个高斯分布表示不同权重的边缘概率分布。PBP 的更新规则不是经典反向传播算法在损耗梯度方向上的标准步骤。给定一些数据，令 $f(w)$ 为 w 的一个任意似然函数，并让当前关于标量 w 的信念被分布 $q(w) = N(w|m, v)$ 捕获。得到数据后，我们根据贝叶斯定律更新关于 w 的信念：

$$s(w) = Z^{-1} f(w) N(w|m, v) \tag{6.5}$$

其中 Z 是归一化常数。使用新的高斯分布 $\hat{q}(w) = N(w|\hat{m}, \hat{v})$ 最小化 $KL(s(w)||\hat{q}(w))$，并且 $\hat{q}(w)$ 可由下式更新

$$\hat{m} = m + v\frac{\partial \log Z}{\partial m} \tag{6.6}$$

$$\hat{v} = v - v^2[(\frac{\partial \log Z}{\partial m})^2 - 2\frac{\partial \log Z}{\partial v}] \tag{6.7}$$

6.2.2　自主性边界的不确定性估计

本章所讨论的自主性边界不再是具体的值(点估计)，而是概率分布。基于 MC dropout 方法构建贝叶斯神经网络进而讨论自主性边界的不确定性。将式 (6.2)代入式(6.1)，可重写如下：

$$\begin{aligned} p(y^*|x^*, D) &= \int p(y^*|x^*, \theta) p(\theta|D) d\theta \\ &\approx \int p(y^*|x^*, \theta) q(\theta) d\theta \\ &= q_\theta(y^*|x^*) \end{aligned} \tag{6.8}$$

以下引用文献 [146] 中的两个命题：

命题 6.1　[146] 基于式(6.8)，假设 $p(y^*|x^*, \theta) = N(y^*; f^\omega(x^*), \tau^{-1} I), \tau > 0$，则

$$\begin{aligned} E_{q_\theta(y^*|x^*)}[y^*] &= \int y^* q_\theta(y^*|x^*) dy^* \\ &= \int\int y^* N(y^*; f^\omega(x^*), \tau^{-1} I) q_\theta(\omega) d\omega dy^* \\ &= \int (\int y^* N(y^*; f^\omega(x^*), \tau^{-1} I) dy^*) q_\theta(\omega) d\omega \end{aligned}$$

$$= \int f^{\omega}(x^*) q_{\theta}(\omega) d\omega \tag{6.9}$$

其中 $f^{\omega}(x^*)$ 是模型（参数 $\omega = W_i = 1^L$）的随机输出。因此，当 $T \to \infty$ 时，

$$\frac{1}{T} \sum_{t=1}^{T} f^{\omega}(x^*) \to E_{q_{\theta}(y^*|x^*)}[y^*] \tag{6.10}$$

即，可通过对贝叶斯神经网络执行 T 次随机前向输出进行平均可近似得到预测 y^* 的一阶矩。

命题 6.2　[146] 基于式(6.8)，假设 $p(y^*|x^*, \theta) = N(y^*; f^{\omega}(x^*), \tau^{-1}I), \tau > 0$，则

$$E_{q_{\theta}(y^*|x^*)}[(y^*)^T(y^*)] = \int (y^*)^T(y^*) q_{\theta}(y^*|x^*) dy^* \tag{6.11}$$

$$= \int \int (y^*)^T(y^*) N(y^*; f^{\omega}(x^*), \tau^{-1}I) q_{\theta}(\omega) d\omega dy^* \tag{6.12}$$

$$= \int (\int (y^*)^T(y^*) N(y^*; f^{\omega}(x^*), \tau^{-1}I) dy^*) q_{\theta}(\omega) d\omega \tag{6.13}$$

$$= \int (Cov_{p(y^*|x^*,\theta)}[y^*] + E_{p(y^*|x^*,\theta)}[y^*]^T E_{p(y^*|x^*,\theta)}[y^*]) q_{\theta}(\omega) d\omega \tag{6.14}$$

$$= \int (\tau^{-1}I + f^{\omega}(x^*)^T f^{\omega}(x^*)) q_{\theta}(\omega) d\omega \tag{6.15}$$

其中 $f^{\omega}(x^*)$ 是模型（参数 $\omega = W_{i\,i=1}^{L}$）的随机输出。因此，当 $T \to \infty$ 时，

$$\tau^{-1}I + \frac{1}{T} \sum_{t=1}^{T} f^{\omega}(x^*)^T f^{\omega}(x^*) \to E_{q_{\theta}(y^*|x^*)}[(y^*)^T(y^*)] \tag{6.16}$$

结合方差计算公式

$$V[y^*] = E_{q_{\theta}(y^*|x^*)}[(y^*)^T(y^*)] - E[y^*]^T E[y^*] \tag{6.17}$$

故而，当 $T \to \infty$ 时，可得方差计算公式

$$\tau^{-1}I + \frac{1}{T} \sum_{t=1}^{T} f^{\omega}(x^*)^T f^{\omega}(x^*) - E[y^*]^T E[y^*] \to V[y^*] \tag{6.18}$$

即，可通过对贝叶斯神经网络执行 T 次随机前向输出, 估算 y^* 的二阶矩，以及基于一阶矩和二阶矩，计算出 y^* 的方差。

基于以上描述，我们采用 MC dropout 实现对自主性边界的不确定性估计。以下分别讨论介入控制下和共享控制下的自主性边界不确定性估计。

1. 介入控制下的自主性边界不确定性估计

针对介入控制算法，根据式(4.1)和式(4.5)所分别对应的机器自主性边界和人的自主性边界判定，此处描述机器自主性边界的一阶矩和方差为

$$\mu_{b_m}(t) = \frac{1}{T}\sum_{t=1}^{T} b_m^*(t) \tag{6.19a}$$

$$\sigma_{b_m}(t) = \tau^{-1} I + \frac{1}{T}\sum_{t=1}^{T} b_m^*(t)^2 - \mu_{b_m}(t)^2 \tag{6.19b}$$

其中 $b_m^*(t)$ 由下式决定:

$$b_m^*(t) = \arg\max_{a_m(t)\in\mathcal{A}_m(t)} \boldsymbol{J}_{h,m}^b(s(t), a_m(t)) \tag{6.20a}$$

$$\text{s. t. } \boldsymbol{C}(s(t), a_m(t)) < 0 \tag{6.20b}$$

算法 6.1　机器自主性边界的不确定性估计

初始化：初始化机器的自主性上界 $\bar{\boldsymbol{B}}_m = \{b_m(t)\}$ 的先验分布信息：$b_m(t) \sim N(\mu_0, \sigma_0)$;

输出：机器的自主性上界的后验概率分布 $b_m(t) \sim N(\hat{\mu}, \hat{\sigma})$;

while 未达到训练结束条件 **do**

  for 未到达终止状态 **do**

    输入机器动作 $a_m(t)$;

    根据约束条件(6.20b)对当前时刻的机器动作进行筛选;

    根据当前 t 时刻机器自主性边界的概率分布，采样得到对应的 $\boldsymbol{T}$ 组值 $\{b_m^{(1)}, b_m^{(2)}, \cdots, b_m^{(T)}\}$;

    将满足约束条件的机器动作与第 6 步中机器的上界信息进行目标函数的比较，如果 $\boldsymbol{J}_{h,m}^b(s(t), a_m(t)) > \boldsymbol{J}_{h,m}^b(s(t), b_m^{(i)})$，则 $b_m^{(i)} = a_m(t)$ 以获得更优的机器的上界，否则机器的自主性上界保持不变;

    基于式(6.19)进行蒙特卡洛估计，更新 t 时刻的机器自主性上界的后验概率分布 $b_m(t) \sim N(\hat{\mu}, \hat{\sigma})$;

  end

end

人的自主性边界的一阶矩和方差为:

$$\mu_{b_h}(t) = \frac{1}{T}\sum_{t=1}^{T} b_h^*(t) \tag{6.21a}$$

$$\sigma_{b_h}(t) = \tau^{-1} I + \frac{1}{T}\sum_{t=1}^{T} b_h^*(t)^2 - \mu_{b_h}(t)^2 \tag{6.21b}$$

其中 $b_h^*(t)$ 由下式决定:

$$b_h^*(t) = \arg \max_{a_h(t) \in \mathcal{A}_h(t)} J_{h,m}^b(s(t), a_h(t)) \tag{6.22a}$$

$$\text{s. t. } C(s(t), a_h(t)) < 0 \tag{6.22b}$$

算法 6.2　人的自主性边界不确定性估计

初始化: 人的自主性上界 $\bar{B}_h = \{b_h(t)\}$ 的先验分布信息:
$b_h(t) \sim N(\mu_0, \sigma_0)$;
输出: 的自主性上界的后验概率分布 $b_h(t) \sim N(\hat{\mu}, \hat{\sigma})$;
while 未达到训练结束条件 **do**
    for 未到达终止状态 **do**
        输入人类动作 $a_h(t)$;
        根据约束条件(6.22b)对当前时刻的人类动作进行筛选;
        根据当前 t 时刻人的自主性边界的概率分布，采样得到对应的 T 组值 $\{b_h^{(1)}, b_h^{(2)}, \cdots, b_h^{(T)}\}$;
        将满足约束条件的人类输入与第 2 步中的上界信息进行目标函数的比较，如果 $J_{h,m}^b(s(t), a_h(t)) > J_{h,m}^b(s(t), b_h^{(i)})$，则 $b_h^{(i)} = a_h(t)$，以更新人的自主性上界，否则保持不变;
        基于式(6.21)进行蒙特卡洛估计，更新 t 时刻人的自主性边界的后验概率分布 $b_h(t) \sim N(\hat{\mu}, \hat{\sigma})$;
    end
end

将介入控制下的机器自主性上界和人的自主性上界的不确定估计思想描述如算法6.1和6.2所示，在算法流程中，首先对介入控制下的机器自主性边界 (人的自主性边界) 先验分布进行初始化 (基于贝叶斯推理给自主性边界信息赋以先验概率，本章考虑到简单实用性，选择高斯分布作为先验概率分布)。在系统动态演化过程中，对于实时输入的系统状态 $s(t)$，算法决策模块分别给出与此系统状态相应的机器决策动作 $a_m(t)$（人类决策动作 $a_h(t)$）（步骤 5）。之后根据约束条件分别对机器决策动作和人类决策动作进行初步筛选。根据当前 t 时刻人的自主性边界的概率分布，采样得到对应的 T 组值 $\{b_h^{(1)}, b_h^{(2)}, \cdots, b_h^{(T)}\}$（步骤 7）。步骤 8 依据人机序贯决策中的边界优化表达式(6.20)和(6.20)，结合步骤 7 进而更新 T 组自主性边界。最后基于优化表达式 (6.19)和(6.21)进行蒙特卡洛估计，更新 t 时刻自主性边界的后验概率分布，如此循环直至训练结束。

2. 共享控制下的自主性边界不确定性估计

针对共享控制算法，根据式(5.4)和(5.5)所对应自主性上界和自主性下界的判定，此处描述自主性上界的一阶矩和方差为

$$\mu_{\overline{b}}(t) = \frac{1}{T}\sum_{t=1}^{T}\overline{b^*}(t) \tag{6.23}$$

$$\sigma_{\overline{b}}(t) = \tau^{-1}I + \frac{1}{T}\sum_{t=1}^{T}\overline{b^*}(t)^2 - \mu_{\overline{b}}(t)^2 \tag{6.24}$$

其中 $\overline{b^*}(t)$ 由下式决定

$$\overline{b^*}(t) = \arg\max_{a_(t)\in\mathcal{A}_h\times\mathcal{A}_m} J^b_{h,m}(s(t), a(t)) \tag{6.25a}$$

$$\text{s. t. } C(s(t), a(t)) < 0 \tag{6.25b}$$

以及自主性下界的一阶矩和方差为

$$\mu_{\underline{b}}(t) = \frac{1}{T}\sum_{t=1}^{T}\underline{b^*}(t) \tag{6.26}$$

$$\sigma_{\overline{b}}(t) = \tau^{-1}I + \frac{1}{T}\sum_{t=1}^{T}\underline{b^*}(t)^2 - \mu_{\underline{b}}(t)^2 \tag{6.27}$$

其中

$$\underline{b^*}(t) = \arg\min_{a_(t)\in\mathcal{A}_h\times\mathcal{A}_m} J^b_{h,m}(s(t), a(t)) \tag{6.28a}$$

$$\text{s. t. } C(s(t), a(t)) < 0 \tag{6.28b}$$

考虑将共享控制下自主性边界的不确定性估计的具体思想描述如算法6.3所示，在算法6.3中，首先对共享控制下的自主性边界先验分布进行初始化（基于贝叶斯推理给自主性边界信息赋以先验概率，本章考虑到简单实用性，选择高斯分布作为先验概率分布）。在系统动态演化过程中，对于实时输入的系统状态 $s(t)$，算法决策模块分别给出与此系统状态相应的机器决策动作 $a_m(t)$ 和人类决策动作 $a_h(t)$（步骤 6）。之后根据约束条件分别对机器决策动作和人类决策动作进行初步筛选。应用上一时刻更新的自主性边界信息，获得 $\boldsymbol{T}$ 组 t 时刻的决策动作 $a(t)$（步骤 8）。步骤 9 依据人机序贯决策中的边界优化表达式(6.25)和式(6.28)，结合步骤 8 中所得到的 $\boldsymbol{T}$ 组决策动作，最终得到 $\boldsymbol{T}$ 组随机最优解。最后基于优化表达式(6.23)和式(6.26)进行蒙特卡洛估计，更新 t 时刻自主性边界的后验概率分布，如此循环直至训练结束。

算法 6.3　共享控制下的自主性边界不确定性估计

初始化: 初始化自主性边界 $\boldsymbol{B} = \{\overline{b}(t), \underline{b}(t)\}$ 的先验分布信息:
$\overline{b}(t) \sim N(\mu_0, \sigma_0)$，$\underline{b}(t) \sim N(\mu_0, \sigma_0)$;
输出: 共享控制下的自主性下界 $\underline{b}(t)$ 和自主性上界 $\overline{b}(t)$ 的后验概率分布;
while 未达到训练结束条件 **do**
    for 未到达终止状态 **do**
        输入：系统状态 $s(t)$;
        根据当前时刻系统状态，机器代理计算出决策动作 $a_m(t)$，同时人类代表也给出决策动作 $a_h(t)$;
        根据约束条件(6.25b)和(6.28b)分别对当前 t 时刻的机器和人类的动作进行检查;
        根据当前 t 时刻自主性上界和自主性下界的概率分布，采样得到对应的自主性上界和自主性下界，用于获得 $\boldsymbol{T}$ 组 t 时刻的决策动作 $a(t)$;
        将满足约束条件的人类输入与初始化中的自主性边界信息进行目标函数的比较，依据式(6.25)和式(6.28)得到目标函数最优和最差的决策动作作为自主性上界和自主性下界，如此循环进行 $\boldsymbol{T}$ 次，得到 $\boldsymbol{T}$ 组随机最优解;
        基于式(6.23)和式(6.26)，进行蒙特卡洛估计，进而更新 t 时刻对应的自主性上界和自主性下界的后验概率分布;
    end
end

6.3　面向人机序贯决策的混合智能优化算法

6.3.1　基于自主性边界不确定性的介入控制优化算法

面向人机序贯决策问题，本小节介绍基于自主性边界的不确定性优化介入控制算法。

1. 基于自主性边界不确定性优化介入控制：人介入机器

针对人介入机器控制方法所求解的人机序贯决策优化问题(4.3)，我们重写如下:

$$\max_{\theta} J_{h,m}(s(t), a(t)) = \int [r(s(t), a(t)) - c(s(t), a(t))]dt \tag{6.29a}$$

$$\max_{a(t)\in\mathcal{A}_m(t)} J^b_{h,m}(s(t), a_m(t)) = J_{h,m}(s(t), a_m(t)) \tag{6.29b}$$

$$\text{s. t. } \dot{s}(t) = f^d(s(t), a(t)) \tag{6.29c}$$

$$a(t) = f^a(a_h(t), a_m(t), b_m(t-1)) \tag{6.29d}$$

$$a_h(t) = Human - Action \tag{6.29e}$$

$$a_m(t) = p^m(s(t); \theta) \tag{6.29f}$$

$$C(s(t), a_h(t), a_m(t)) < 0 \tag{6.29g}$$

$$t = 0, 1, 2, 3, ...$$

其中 $J^b_{h,m}(s(t), a_m(t))$ 是自主性边界的优化目标函数，这里我们可以选择其为被控对象的优化目标函数 $J_{h,m}(s(t), a_m(t))$。$r(\cdot)$ 和 $c(\cdot)$ 分别代表时间 t 的即时奖励和成本。$f^d(\cdot)$ 表示系统的动态模型。$f^a(a_h(t), a_m(t), b_m(t-1))$ 是人类动作和机器动作的仲裁函数，其中 $b_m(t-1)$ 可以通过上一节中的公式 (6.20)和算法6.1求解。为方便起见，$f^a(\cdot)$ 定义为(6.30)。

$$a(t) = f^a(a_h(t), a_m(t), b_m(t-1))$$

$$= \begin{cases} \text{Human: } a_h(t), \{c_h(t) > c_m(t)\} \ \& \ \{J_{h,m}(s(t), a_h(t)) \geqslant \max\{J_{h,m}(s(t), a_m(t)), \\ \qquad J_{h,m}(s(t), b_m(t-1))\}\} \\ \text{Boundary: } b_m(t-1), \{c_m(t) > c_h(t)\} \ \& \ \{J_{h,m}(s(t), b_m(t-1)) \geqslant \max\{J_{h,m}(s(t), \\ \qquad a_h(t)), J_{h,m}(s(t), a_m(t))\}\} \\ \text{Machine: } a_m(t), \qquad \text{其他.} \end{cases} \tag{6.30}$$

公式(6.30)表示人机交易控制系统的仲裁函数。根据对可信度的评估来判断谁是当前的决策者。

注意到，本文第4章将自主性边界的概念引入到介入控制算法中，实现了更好的介入控制设计。在本章，我们考虑到自主性边界自身所固有的模糊性，对自主性边界进行不确定性分析，使得在第4章的基础上，不仅能够实现人对机器的适时介入，而且自主性边界服从某种概率分布，这对于优化介入控制的介入时机是有利的。通过动态实时更新维护自主性边界的概率分布，之后基于获得的自主性边界不确定性信息，对机器决策动作进行判定以确定是否允许人的介入。

算法6.4是在解决优化问题(6.29)的通用流行办法的基础上，融入介入控制下自主性边界的不确定性估计，实现关于此类人机序贯决策问题的进一步优化，算法的优化目标有两个：1) 直接影响决策动作的策略学习；2) 间接影响决策动作的带有不确定性的自主性边界。对于 t 时刻的系统状态 $s(t)$，机器代理计算出决策动作 $a_m(t)$ 和相应的可信度 (步骤 6)，人类伙伴也给出决策动作 $a_h(t)$(步骤 7)，之后依概率采样得到 $\boldsymbol{T}$ 组机器自主性边界 (步骤 8)，用于式(6.30)的仲裁判断得

算法 6.4　基于自主性边界不确定性的人介入机器控制优化算法

初始化：随即初始化机器代理的策略网络 p^m 及其网络参数 θ；初始化机器的自主性边界的先验信息 $\bar{B}_m \sim N(\mu_0, \sigma_0)$;
输出：最终决策动作 $a(t)$，以及机器自主性边界的后验概率分布 $\bar{B}_m \sim N(\hat{\mu}, \hat{\sigma_0})$;
while 未达到最大训练时间步 **do**
  for 未到达终止状态 **do**
    输入 $s(t)$;
    机器代理根据策略网络计算出决策动作 $a_m(t)$ 以及基于蒙特卡洛估计(3.9)计算可信度 $c_m(t)$;
    根据优化问题(6.29)中的约束(6.29g)过滤机器决策动作 $a_h(t)$;
    根据机器自主性边界的概率分布进行采样，获得 T 组边界值 $\{b_m^{(1)}, b_m^{(2)}, \cdots, b_m^{(T)}\}$;
    利用仲裁函数(6.30)输出最终决策动作 $a(t)$;
    基于算法6.1更新维护机器自主性边界的后验概率分布 $\bar{b}_m(t) \sim N(\hat{\mu}, \hat{\sigma_0})$;
  end
end

到 T 组前向最优解，从而输出最终的决策动作 $a(t)$(步骤 9)。同时基于式(6.19)进行蒙特卡洛估计对机器自主性边界的后验概率分布进行更新。

2. 基于自主性边界不确定性优化介入控制：机器介入人

针对机器介入人控制方法所求解的人机序贯决策优化问题(4.7)，我们重写如下：

$$\max_{a(t)\in\mathbb{A}_s^T} J_{h,m}(s(t), a(t)) = \int [r(s(t), a(t)) - c(s(t), a(t))]dt \tag{6.31a}$$

$$\max_{a_h(t)\in\mathcal{A}_h(t)} J_{h,m}^b(s(t), a_h(t)) = J_{h,m}(s(t), a_h(t)) \tag{6.31b}$$

$$\text{s. t. } \dot{s}(t) = f^d(s(t), a(t)) \tag{6.31c}$$

$$a(t) = f^a(a_h(t), a_m(t), b_h(t-1)) \tag{6.31d}$$

$$a_h(t) = Human - Action \tag{6.31e}$$

$$a_m(t) = p^m(s(t)) \tag{6.31f}$$

$$C(s(t), a_h(t), a_m(t)) < 0 \tag{6.31g}$$

$$t = 0, 1, 2, 3, ...$$

其中 $b_h(t)$ 是根据自主性边界的优化目标函数 $J^b_{h,m}(s(t), a_h(t))$ 求得的人的自主性边界。$r(\cdot)$ 和 $c(\cdot)$ 分别代表时间 t 的即时奖励和成本。$f^d(\cdot)$ 表示系统的动态模型。$f^a(a_h(t), a_m(t), b_h(t-1))$ 是人类动作和机器动作的仲裁函数，其中 $b_h(t-1)$ 可以通过公式(6.22)和算法6.2求解。$f^a(\cdot)$ 定义如式(6.32)所示。

$$
\begin{aligned}
a(t) &= f^a(a_h(t), a_m(t), b_h(t-1)) \\
&= \begin{cases}
\text{Machine: } a_m(t), \{c_m(t) > c_h(t)\} \ \& \ \{J_{h,m}(s(t), a_m(t)) \geqslant \max\{J_{h,m}(s(t), a_h(t)), \\
\qquad\qquad J_{h,m}(s(t), b_h(t-1))\}\} \\
\text{Boundary: } b_h(t-1), \{c_h(t) > c_m(t)\} \ \& \ \{J_{h,m}(s(t), b_h(t-1)) \geqslant \max\{J_{h,m}(s(t), \\
\qquad\qquad a_h(t)), J_{h,m}(s(t), a_m(t))\}\} \\
\text{Human: } a_h(t), & \text{其他.}
\end{cases}
\end{aligned} \tag{6.32}
$$

本节通过动态实时更新维护自主性边界的概率分布，之后基于获得的人的自主性边界不确定性信息，对人类决策动作进行判定以确定是否允许机器的介入。

算法 6.5　基于自主性边界不确定性的机器介入人控制优化算法

初始化：随即初始化机器代理的策略网络 p^m 及其网络参数 θ; 初始化人的自主性边界的先验信息 $\bar{B}_h \sim N(\mu_0, \sigma_0)$;

输出：最终决策动作 $a(t)$，以及机器自主性边界的后验概率分布 $\bar{B}_h \sim N(\hat{\mu}, \hat{\sigma_0})$;

while 未达到最大训练时间步 **do**

  for 未到达终止状态 **do**

    输入：系统状态 $s(t)$;

    机器代理根据策略网络计算出决策动作 $a_m(t)$ 以及基于蒙特卡洛估计(3.9)计算可信度 $c_m(t)$;

    根据优化问题(6.31)中的约束(6.31g)）过滤机器决策动作 $a_h(t)$;

    根据机器自主性边界的概率分布进行采样，获得 T 组边界值 $\{b_h^{(1)}, b_h^{(2)}, \cdots, b_h^{(T)}\}$;

    利用仲裁函数(6.32)）输出最终决策动作 $a(t)$;

    基于算法6.2更新维护人自主性边界的后验概率分布 $\bar{b}_h(t) \sim N(\hat{\mu}, \hat{\sigma_0})$;

  end

end

算法6.5是在解决优化问题(6.31)的通用流行办法的基础上，融入介入控制下人的自主性边界的不确定性估计，实现关于此类人机序贯决策问题的进一步优化，算法的优化目标有两个：1) 直接影响决策动作的策略学习；2) 间接影响决策行为的带有不确定性的人的自主性边界。对于 t 时刻的系统状态 $s(t)$，机器代理计算出决策动作 $a_m(t)$ 和相应的可信度（步骤 6），人类伙伴也给出决策动作 $a_h(t)$（步骤 7），之后依概率采样得到 $\boldsymbol{T}$ 组人的自主性边界 (步骤 8)，用于式(6.32)的仲裁判断得到 $\boldsymbol{T}$ 组前向最优解，用于输出最终的决策动作 $a(t)$（步骤 9）。同时基于式 (6.21)进行蒙特卡洛估计对人的自主性边界的后验概率分布进行更新。

6.3.2　基于自主性边界不确定性的共享控制优化算法

基于上述关于共享控制下自主性边界的不确定性分析，针对共享控制方法所求解的人机序贯决策优化问题(5.2)，我们重写如式(6.33)，并且构建基于自主性边界不确定性估计的共享控制框架如图6.1。

$$\max_{\theta} \boldsymbol{J}^r(s(t), a(t)) = \int_{t=t_0}^{t_0+T} r(s(t), a(t)) \tag{6.33a}$$

$$\text{s. t. } a(t) = f^a(a_h(t), a_m(t), c(t)) \tag{6.33b}$$

$$a_m(t) = p(s(t), g(t); \theta) \tag{6.33c}$$

$$\{g(t), c(t)\} = \text{Infer}(a_h(t)) \tag{6.33d}$$

$$s(t+1) = f^d(s(t), a(t)) \tag{6.33e}$$

$$\boldsymbol{C}(s(t), a_m(t), a_h(t)) \leqslant 0 \tag{6.33f}$$

$$t = 0, 1, 2, 3, ...$$

其中 $\boldsymbol{J}^r(s(t), a(t))$ 是由系统状态和决策动作所决定的累积奖励值。$a(t)$ 是共享控制系统最终的决策动作，Infer(·) 是意图推理函数模块，其输出是推断的用户目标 $g(t)$ 和推测的置信度大小 $c(t)$。$p(\cdot)$ 表示智能机器代理的策略函数，用来输出机器代理的决策动作。

仲裁函数可具有如下的典型线性混合形式：

$$f^a(a_h(t), a_m(t), \alpha) = (1-\alpha)a_h(t) + \alpha a_m(t) \tag{6.33g}$$

其中参数 α 的值应根据意图推理的置信度 $c(t)$ 而定，可具有如式(5.3)的经典形式。

我们定义式(6.25)和式(6.28)中的人机共同目标函数即为决策任务的优化目

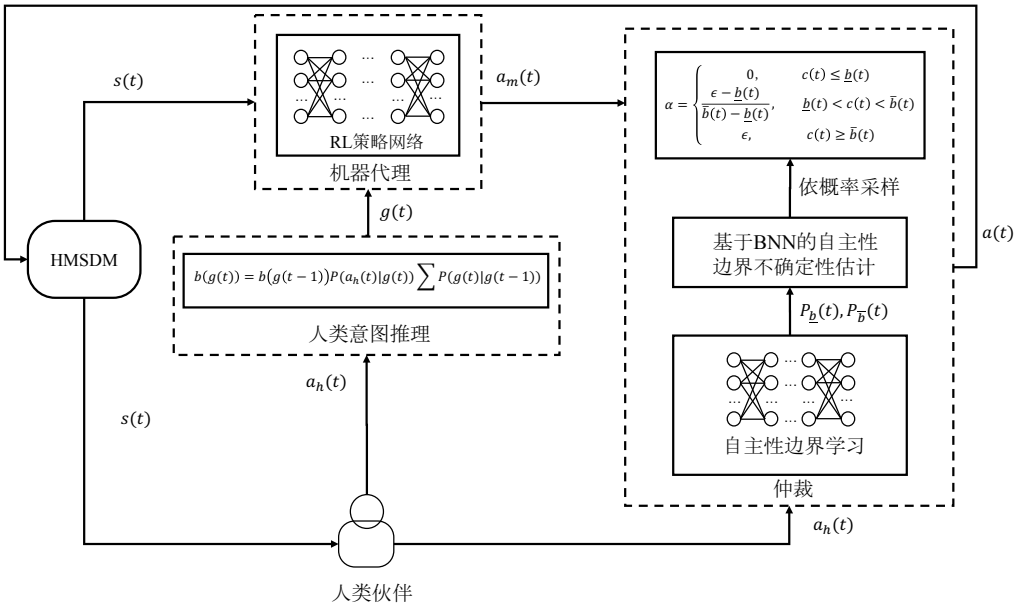


图 6.1　基于自主性边界不确定性估计的共享控制优化框架

标 $\boldsymbol{J}^r(s(t), a(t))$，并给出具体算法步骤如算法6.6所示，

$$\boldsymbol{J}_{h,m}(s(t), a(t)) = \boldsymbol{J}^r(s(t), a(t)) = \int_{t=t_0}^{t_0+T} r(s(t), a(t)) \tag{6.34}$$

注意到，本文第5章将自主性边界的概念引入到共享控制算法中，实现了更好的仲裁参数设计。在本章，我们考虑到自主性边界自身所固有的模糊性，对自主性边界进行不确定性分析，使得在第5章的基础上，不仅能够实现仲裁参数的自适应调节，而且自主性边界服从某种概率分布，以及仲裁参数服从某种分布，这对于人机共享控制系统的混合决策是有利的。同样地，优化的过程具有两个任务目标：1) 直接影响决策动作的策略网络；2) 间接影响决策动作的带有不确定性的自主性边界。通过动态实时更新维护自主性边界的概率分布，之后基于获得的自主性边界不确定性信息，对机器决策动作和人类决策动作进行混合仲裁产生即将作用到被控对象的最终决策动作，如基于自主性边界不确定信息的人共享控制框架6.1所示。

算法6.6是在解决优化问题(6.33)的通用流行办法的基础上，融入共享控制下自主性边界的不确定性估计，实现关于此类人机序贯决策问题的进一步优化，算法的优化目标有两个：1) 直接影响决策动作的策略学习；2) 间接影响决策动作的带有不确定性的自主性边界。在系统的动态演化过程中，对于实时输入的系统状态 $s(t)$，人类伙伴给出有目的性的决策动作 $a_h(t)$，此动作有两个作用：一者，机器代理用来推测人类想要完成的任务目标 (意图推理模块)；二者，作为和即将生成的机器决策动作 $a_m(t)$ 仲裁混合的人类决策动作。智能机器代理预测出任务目标之后，结合当前时刻的系统状态和当前策略学习情况，计算出实时的机器决

算法 6.6　基于自主性边界不确定信息的共享控制优化算法

初始化: 初始化自主性边界 $B = \{\overline{b}(t), \underline{b}(t)\}$ 的先验分布信息: $\overline{b}(t) \sim N(\mu_0, \sigma_0)$，$\underline{b}(t) \sim N(\mu_0, \sigma_0)$;

输出: 当前时刻的最终决策动作 $a(t)$，以及共享控制下的自主性下界 $\underline{b}(t)$ 和自主性上界 $\overline{b}(t)$ 的后验概率分布;

while 未达到训练结束条件 **do**

  for 未到达终止状态 **do**

    输入系统状态 $s(t)$;

    人类伙伴根据观察到的系统状态 $s(t)$ 输入决策动作 $a_h(t)$;

    机器代理的意图推理模块根据人类动作推测出可能的任务目标 $g(t)$ 和对应的目标可信度 $c(t)$。并且基于此任务目标和系统状态 $s(t)$，计算相应的机器决策动作 $a_m(t)$;

    根据约束条件(6.28b)和(6.28b)分别对当前 t 时刻的机器和人类的动作进行检查;

    按照式(6.34)，依据自主性边界概率分布采样获得自适应的阈值参数和仲裁参数 α;

    基于优化问题(6.33)，计算 t 时刻系统状态 $s(t)$ 对应的最终决策动作 $a(t)$;

    依据式(6.25)和式(6.28)得到目标函数最优和最差的决策动作作为自主性上界和自主性下界，如此循环进行 T 次，得到 T 组随机最优解。基于式(6.23)和式(6.26)，进行蒙特卡洛估计，进而更新 t 时刻对应的自主性上界和自主性下界的后验概率分布;

  end

end

策动作 $a_m(t)$。接着，步骤 9 按照式(6.34)，根据上一时刻更新维护的自主性边界信息，依概率分布采样获得自适应的阈值参数和仲裁参数 α。基于上述已获得信息，对于优化问题(6.33)，计算 t 时刻系统状态 $s(t)$ 对应的最终决策动作 $a(t)$。步骤 11 完成对自主性边界的后验概率分布估计，依据式(6.25)和式(6.28)得到目标函数最优和最差的决策动作作为自主性上界和自主性下界，重复进行 T 组得到 T 组随机最优解。基于式(6.23)和式(6.26)，进行蒙特卡洛估计，进而更新 t 时刻对应的自主性上界和自主性下界的后验概率分布。如此进行下去直至训练结束。

由上述讨论可知，算法6.6中对于仲裁函数的定义不仅不再过度依赖实际中难以准确确定的固定超参数，而且从某种程度上讲仲裁参数的选取服从某种概率分布，这样，参数给定的规范化、动态性和不确定性将从本质上有利于算法性

能的提升。并且，在人机序贯决策问题求解中，将自主性边界判定融入到共享控制优算法中，既可实时更新维护自主性边界信息（其使得人机序贯决策问题中的人机决策权限更加明朗），又有利于共享控制策略的求解，具有重要的理论研究和实际应用价值。

6.4　仿真实验

本节针对本章所提利用自主性边界不确定性信息优化介入控制和共享控制进行了实验仿真。

6.4.1　介入控制实验结果

1. 人介入机器实验结果

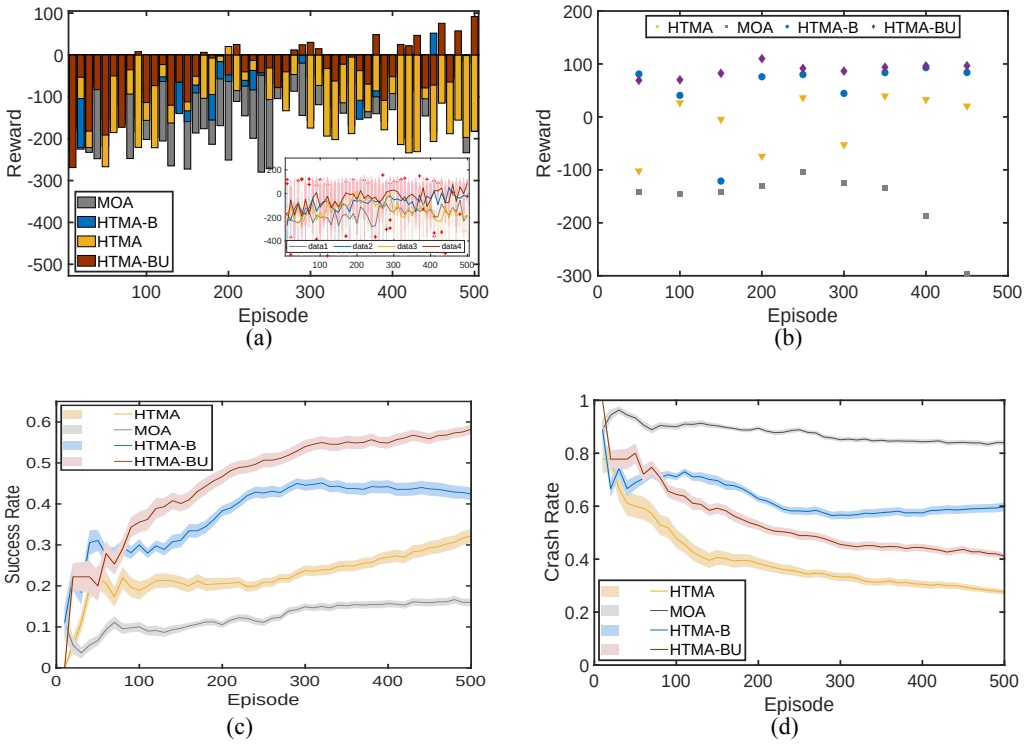


图 6.2　算法 MOA，HTMA，HTMA-B 和 HTMA-BU 的实验结果对比，四幅子图中灰色表示算法 MOA，黄色表示算法 HTMA，蓝色表示算法 HTMA-B，红色表示算法 HTMA-BU。(a) 奖赏：实线表示奖赏平均值走势，红色加号代表异常点，阴影表示大多数点所落在的箱体区域；(b) 着陆成功的 episodes 的奖赏值；(c) 成功率：实线表示成功率的平均值，阴影表示不确定性；(d) 撞击率：实线表示撞击率的平均值，阴影表示不确定性。

本小节同4.2.3中实验设置，使用 LunarLander 作为仿真环境。以下采用 MOA 表示仅有机器代理参与决策的算法，HTMA 表示人类介入机器决策的算法，

HTMA-B 表示基于自主性边界的人介入机器算法，HTMA-BU 表示基于自主性边界不确定性的人介入机器算法。下面分别从奖赏值大小，着陆成功率，撞击失败率，以及着陆轨迹分析四种算法的优劣。图6.2(a)展示了四种算法在平均奖赏值的对比。可以看出相较于纯机器算法 MOA，人介入机器的三种算法 (HTMA，HTMA-B 和 HTMA-BU) 均在不同程度上提升了奖赏大小，并且本文所提的基于自主性边界的人介入机器算法 (HTMA-B 和 HTMA-BU) 效果更佳。图6.2(b)是着陆成功 episode 的奖赏平均值。值得注意的是，从图6.2(a)和图6.2(b)，我们发现在奖赏方面，HTMA-BU 相较于 HTMA-B 的提升幅度没有 HTMA-B 相较于 HTMA 的提升幅度大。

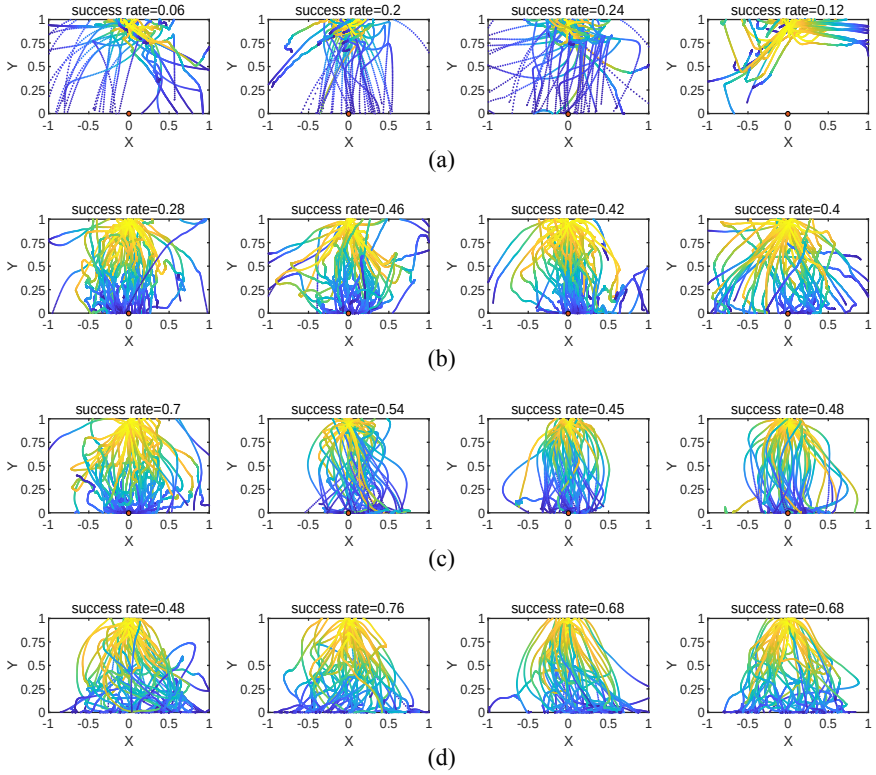


图 6.3　不同算法对应着陆轨迹的对比: (a) MOA; (b) HTMA; (c) HTMA-B; (d) HTMA-BU。

着陆成功是月球着陆车的首要任务。接下来，我们比较四种算法的着陆成功率和撞击失败率如图6.2(c)和6.2(d)所示。显然能得出 HTMA-BU>HTMA-B>HTMA>MOA 的结论。如第5章所述，MOA 的成功率持续偏低的原因是机器代理需要更多训练时间，人的介入使着陆成功率得到了提升，自主性边界的引入以及不确定性的估计使得这种提升效果更加明显。关于图6.2(d)所示的撞击失败率，可以看出 HTMA<HTMA-BU<HTMA-B<MOA。该现象说明相较于 MOA，HTMA，HTMA-B 和 HTMA-BU 均降低了撞击失败率，但相较于 HTMA，基于自主性边界的算法 HTMA-B 没有获得理想中的更低的撞击失败率。如第5章所述，

该现象是由自主性边界的引入所带来的，说明我们所考虑的自主性边界信息若是正确合适的，则能够为提升决策性能提供有效信息，但不准确的自主性边界反而可能导致撞击失败率的升高。但值得高兴的是，相较于 HTMA-B，HTMA-BU 的撞击失败率显著下降。这也验证了自主性边界不确定性估计的重要性和有效性。

接下来，我们比较了算法 MOA，HTMA，HTMA-B 和 HTMA-BU 的着陆轨迹，如图6.3所示。我们发现算法 MOA 对应的着陆轨迹最差且着陆成功率最低，算法 HTMA 的着陆轨迹虽较凌乱，但所对应的着陆成功率得到了改善，算法 HTMA-B 对应的下降轨迹最为整齐有序且时间步数也较短。算法 HTMA-BU 的着陆轨迹虽不是这四种算法中最整齐有序的，但是成功率却是最高的。另外值得一提的是，算法 HTMA-BU 着陆轨迹中的着陆点不是固定不变的，而是随机生成的，说明在难度被增加的情况下依然能够着陆成功且提升成功率和降落过程的奖赏。如此我们得出 HTMA-BU>HTMA-B>HTMA>MOA 的结论。

2. 机器介入人实验结果

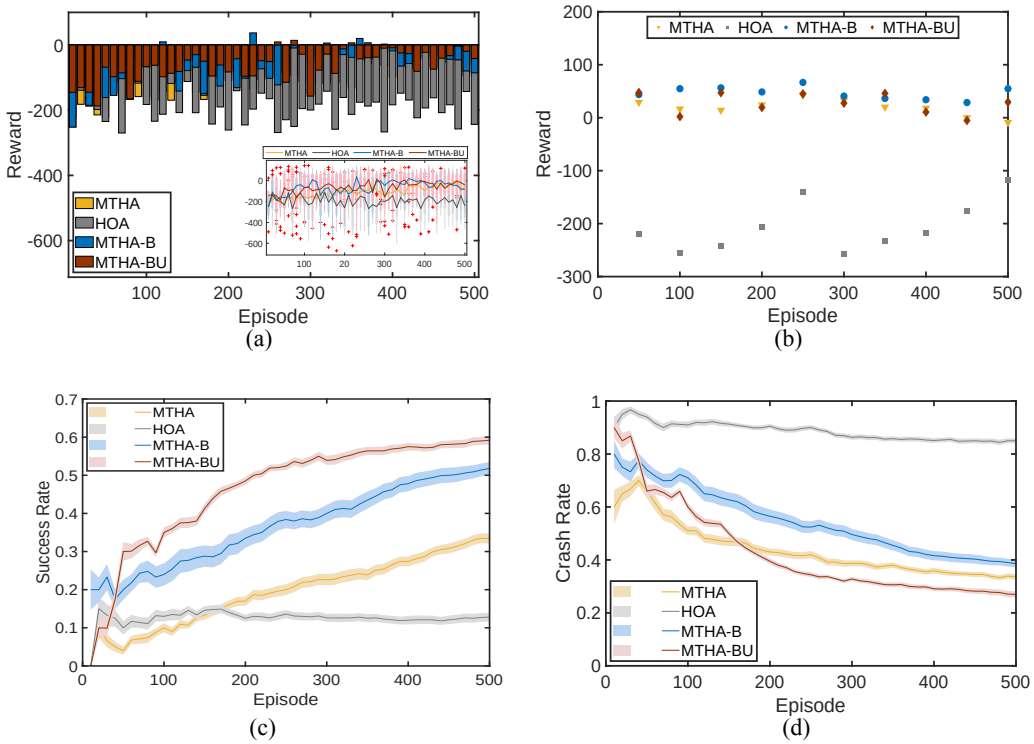


图 6.4　算法 HOA，MTHA，MTHA-B 和 MTHA-BU 的实验结果对比，四幅子图中灰色表示算法 HOA，黄色表示算法 MTHA，蓝色表示算法 MTHA-B，红色表示算法 MTHA-BU。(a) 奖赏值，其中小窗口中实线表示奖赏平均值走势，红色加号代表异常点，阴影表示大多数点所落在的箱体区域；(b) 着陆成功的 episodes 的奖赏值；(c) 成功率：实线表示成功率的平均值，阴影表示不确定性；(d) 撞击率: 实线表示撞击率的平均值，阴影表示不确定性。

本小节同4.3.3中实验设置，使用 LunarLander 作为仿真环境。以下使用 HOA 表示仅有人类参与决策的算法，MTHA 表示机器介入人的控制算法，MTHA-B 表示基于自主性边界的机器介入人控制算法，MTHA-BU 表示基于自主性边界不确定性的机器介入人控制算法。下面分别从奖赏值大小，着陆成功率，撞击失败率，以及着陆轨迹分析四种算法的优劣。图6.4(a)展示了四种算法在平均奖赏值的对比。可以看出相较于纯人类决策算法 HOA，机器介入人的三种控制算法 (MTHA，MTHA-B 和 HTHA-BU) 均在不同程度上提升了奖赏大小，并且本文所提的基于自主性边界的机器介入人控制算法 (MTHA-B 和 MTHA-BU) 效果更佳。图6.4(b)是着陆成功 episode 的奖赏平均值。值得注意的是，从图6.4(a)和图6.4(b)，我们发现在奖赏方面，算法 MTHA，MTHA-B，MTHA-BU 的奖赏虽都较 HOA 有所提升，但 MTHA，MTHA-B，MTHA-BU 三者的平均奖赏以及平均成功奖赏之间相差无几。此种现象说明了基于自主性边界的机器介入人的控制优化算法在奖赏获取上没有特别的优势，这是研究者需要注意的地方。

但是对于任务本身来说，奖赏值大小并不是衡量任务完成情况的唯一指标。对于本章所述的仿真环境 LunarLander 来说，着陆成功率是极为重要的参数指标。接下来，我们比较四种算法的着陆成功率和撞击失败率如图6.4(c)和6.4(d)所示。显然能得出 MTHA-BU>MTHA-B>MTHA>HOA 的结论。如第5章所述，HOA 的成功率持续偏低的原因是人类操作的粗精度，尽管人类的决策动作具有一定的价值和指引性，但对于仅使用人类决策动作的算法 HOA 来说，显然是力不从心的，仅能达到 0.1 的成功率。而算法 MTHA 能够将成功率提升至 0.32，MTHA-B 能提升至 0.5，更进一步地，MTHA-BU 能将成功率提升至 0.6。关于图6.4(d)所示的撞击失败率，可以看出 MTHA-BU<MTHA<MTHA-B<HOA。该现象说明相较于 HOA，MTHA，MTHA-B 和 MTHA-BU 均降低了撞击失败率，但相较于 MTHA，基于自主性边界的算法 MTHA-B 没有获得理想中的更低的撞击失败率。如前所述，这是由自主性边界的引入所带来的，说明我们所考虑的自主性边界信息若是正确合适的，则能够为提升决策性能提供有效信息，但不准确的自主性边界反而可能导致撞击失败率的升高。但值得高兴的是，相较于 MTHA-B，MTHA-BU 的撞击失败率显著下降，甚至比 MTHA 更低，显然这是我们希望看到的结果，这也验证了在机器介入人的控制算法中，不确定性估计使得自主性边界的引入变得更有价值。

接下来，我们比较了算法 HOA，MTHA，MTHA-B 和 MTHA-BU 的着陆轨迹，如图6.5所示。不同于图4.7中算法 MOA 所对应的着陆轨迹, 算法 HOA 的着陆轨迹干净有序，但较低的着陆成功率同时也说明它在下降过程中以不可操控的速度迅速降落，这是由于其没有机器决策的精准性导致的。算法 MTHA 和算法 MTHA-B 通过增加机器代理的干预，使得下降过程更加缓慢精准，这从它的

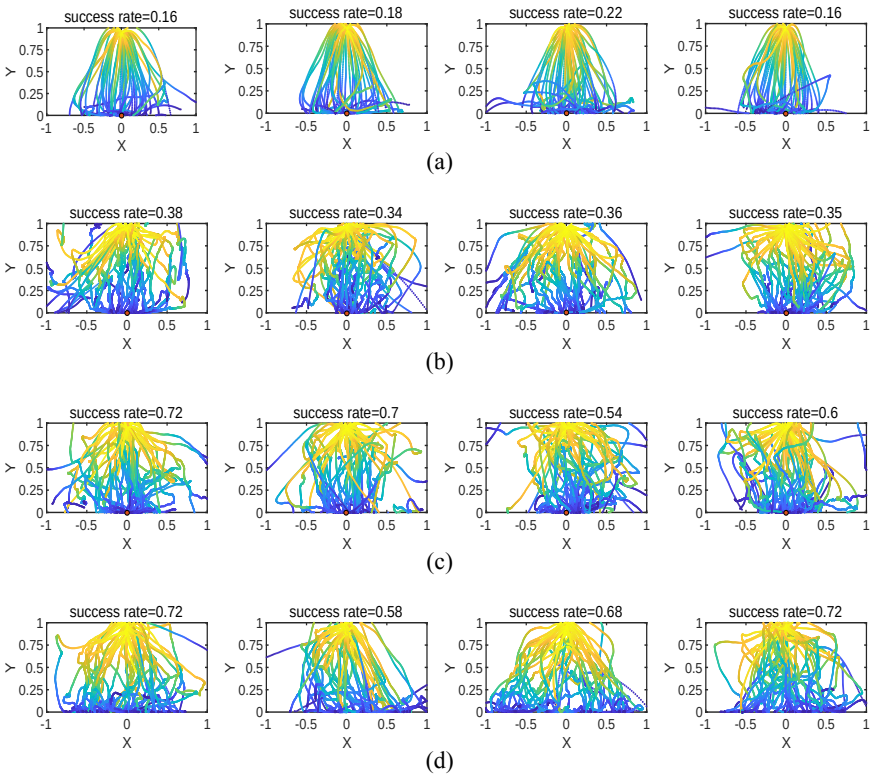


图 6.5　不同算法对应着陆轨迹的对比: (a) HOA; (b) MTHA; (c) MTHA-B; (d) MTHA-BU。

成功率可以看出。相比于 MTHA 和 MTHA-B，算法 MTHA-BU 的降落轨迹更佳整洁有序，并且在成功率上得到了显著提升。另外值得注意的是，我们对 MTHA-BU 作用的对象增加了难度，即，使得每一条 episode 的着陆点都不是固定的。在这种情况下，MTHA-BU 能够较好的适应，并且具有更高的着陆成功率。综上，MTHA-BU 虽在奖赏获取上并无特别优势，但在着陆成功率，撞击失败率，以及下降轨迹中成效显著，证明了该优化算法的有效性。

6.4.2　共享控制实验结果

本小节使用5.6.1小节中的实验设置，以下使用 SCHM 表示“人机共享控制算法”，SCHM-B 表示“基于自主性边界的共享控制算法”，对应于优化算法5.3，SCHM-BU 表示“基于自主性边界不确定性的共享控制优化算法”，对应于优化算法6.6。我们分别从奖励趋势，着陆成功率，仲裁参数，自主性边界，等参数分析实验结果。更进一步地，由于本次实验并非是采取一次偶然的实验结果，而是进行了 500 次的重复实验，因此足以说明结果的可靠性。首先值得注意的是奖励走势情况，如图6.6和6.7所示，相较于算法 SCHM 和 SCHM-B，算法 SCHM-BU 能够获得更高的奖赏值。并且从图6.6中可以看出，随着 episode 的增加，基于自主性边界不确定的共享控制优化算法 SCHM-BU 的作用在不断凸显，使得在奖

赏获得上略胜一筹。

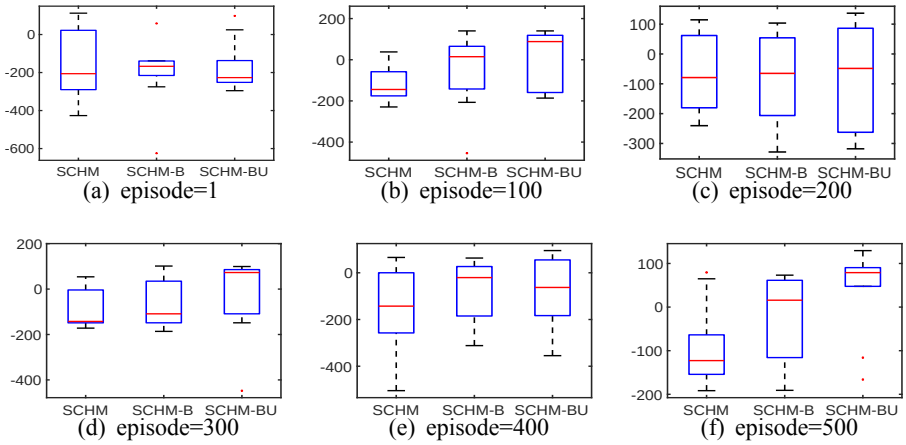


图 6.6　算法 SCHM，算法 SCHM-B 和算法 SCHM-BU 在不同 episode(0, 100, 200, 300, 400, 500) 的奖赏对比。

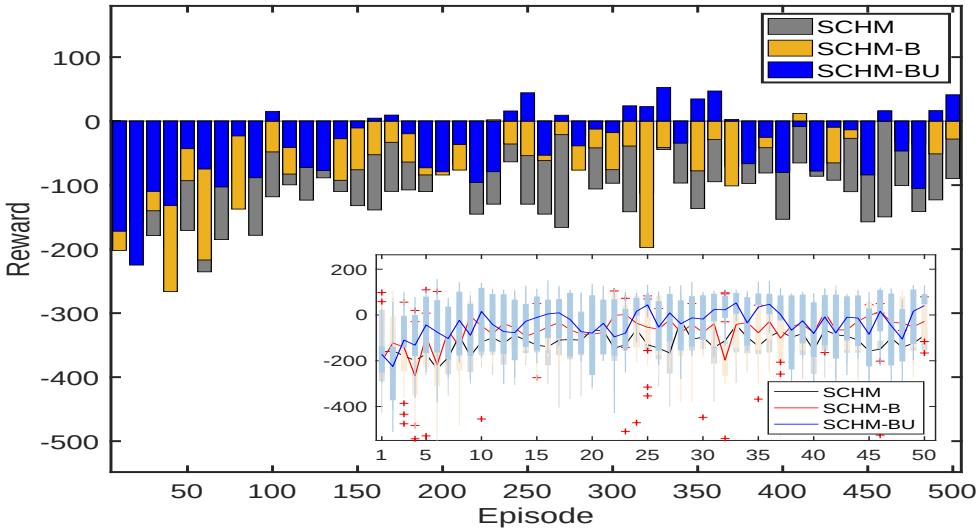


图 6.7　算法 SCHM，算法 SCHM-B 和算法 SCHM-BU 的奖赏对比

接下来，我们分析月球车的着陆成功率和撞击率，如图6.8所示。图6.8(a)展示了在第五章研究的基础上，基于自主性边界不确定性的共享控制优化算法 SCHM-BU 对 SCHM 和 SCHM-B 在着陆成功率上的提升效果。相较于 SCHM 的成功率为 0.25，第五章的算法 SCHM-B 能够提升至 0.5，而本章的 SCHM-BU 能够将其提升至 0.6。此外在图6.8(b)中，可以看出算法 SCHM-BU 的撞击率也有所下降。

最后，我们给出仲裁参数的对比展示，如图6.9所示。仲裁参数 α 是直接决定机器决策动作和人类决策动作之间的融合程度的因素（见公式(6.33)中的(6.33g)）。图6.9展示了仲裁参数 α 的变化情况。图6.9(a)表示对前 100 个时间步仲裁参数的放大观察，图6.9(b)表示仲裁参数 α 的整体走势。图6.10表示对自主性上界和自

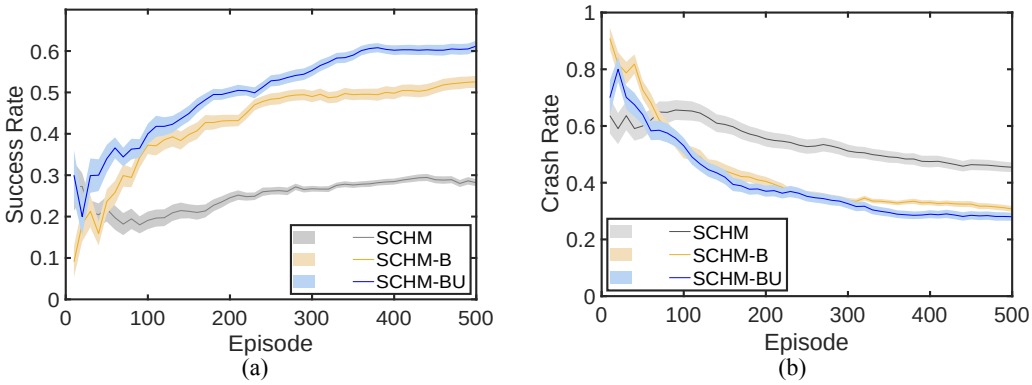
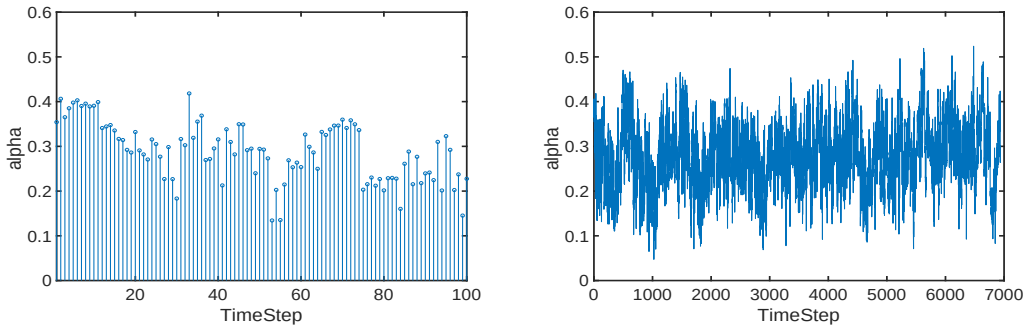


图 6.8　算法 SCHM，SCHM-B 和 SCHM-BU 的成功率 (a) 和撞击率 (b) 对比。



(a)　　(b)

图 6.9　算法 SCHM-BU 的仲裁参数 α：(a) 100 个时间步的 α；(b) α 总体走势

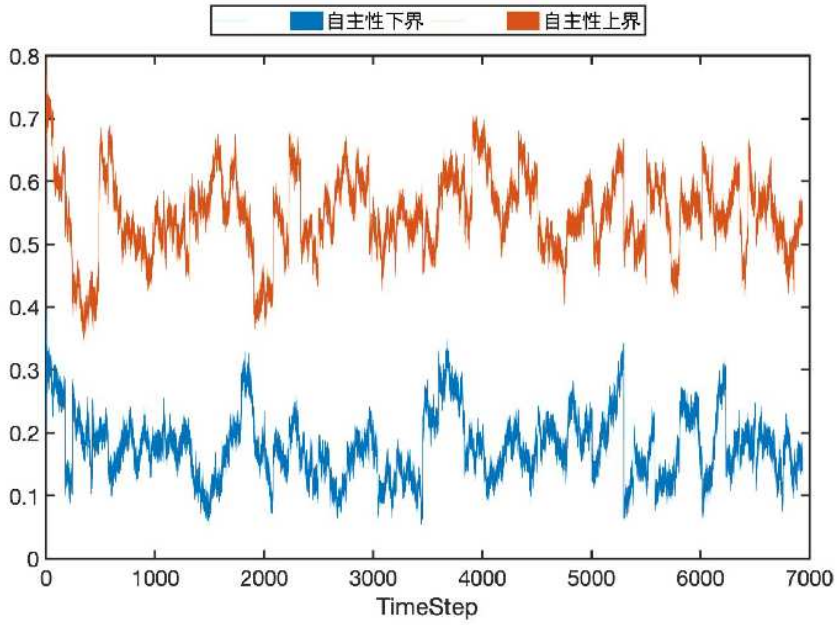


图 6.10　算法 SCHM-BU 中的自主性上界和自主性下界及不确定性

主性下界的概率分布情况。二者均对人机共享控制算法 SCHM 的优化过程有贡献，概率分布的不确定性能够提供更全面的信息，这从图6.6，图6.7和图6.8中可以看出。本章对自主性边界的不确定性进行估计，从而在第五章利用自主性边界优化共享控制的仲裁参数的基础上，进一步提升决策效果。

6.5　本章小节

本章基于贝叶斯神经网络以概率分布的方式衡量自主性边界的不确定性。利用 dropout 机制近似贝叶斯神经网络，基于采样的思想维护自主性边界的后验概率分布，此概率分布亦可用于下一时刻的动作生成，如此形成一个良性循环：既能不断优化自主性边界的不确定性信息，也能持续优化人机混合智能系统的决策网络。最后通过仿真结果表明，自主性边界不确定性的引入为决策动作带来了一定程度的不确定性，为求得更优解提供了更大可能性，实验展示了本章方法的有效性和优越性，也更加符合人们对模糊性边界的思考。

第 7 章　总结与展望

本章总结了全文的主要研究工作和贡献，并给出了未来研究工作的展望。

7.1　本文的主要研究工作

面向人机序贯决策问题，混合智能方法成为一种发展趋势。然而传统的人机混合方法仅拥有两层结构：人类控制和机器自动执行。随着人工智能技术的迅速发展，拥有三层结构的新型人机混合办法得以形成：人类控制、机器智能自主和机器自动执行。新型人机混合智能方法不仅保持了传统人机混合的优势，也使得决策空间得到扩展，为优化决策动作提供了更多可能，但同时也面临着许多挑战。首先是由于人和机器同时出现在决策层面，如何划分人和机器的决策权限，使得人类智能和机器智能各自具有相对清晰的决策范围，这与人机混合智能方法所获得的决策性能是息息相关的。针对人机序贯决策中的介入控制设计，考虑机器介入和人介入场景下，如何利用感知信息，决策网络，仲裁机制等获得较高质量的决策动作是本文考虑的介入控制优化问题。面向人机序贯决策中的共享控制设计，如何结合对人类的意图推理，对机器智能的决策质量判断，以及在人类动作空间，机器动作空间，和人机混合动作空间中，通过优化设计仲裁参数进而获得最优的决策动作，是本文考虑的共享控制优化问题。本文围绕自主性边界判定，介入控制优化问题，共享控制优化问题展开研究，针对“人介入方法的序贯决策”和“人介入问题的序贯决策”设计有效的人机混合智能控制算法。通过第 3 章到第 6 章研究内容的具体论述，将本文的工作总结如下：

1. 首先，针对基于强化学习方法求解人机序贯决策问题 (对应第 3 章)，建立三个子系统互相竞争的决策模型，设计基于仲裁机制的人机混合智能算法，具体包括基于模型的强化学习决策子系统引入对环境的建模实现学习和规划、无模型强化学习决策子系统源于从真实环境的交互轨迹样本中学习策略函数、人类动作的不定时参与决策、基于对未来轨迹进行预测得到的安全约束判断、以及基于贝叶斯神经网络对两个机器决策子系统的决策评估可行性。最终在仲裁的框架下实现对模型决策子系统动作、无模型决策子系统的动作、以及人类动作进行判断得出最终待执行动作。考虑序贯决策问题的时序性和多阶段性，决策既依赖于当前状态的瞬时奖励，也取决于多阶段的总体优化目标函数，为此，构建了基于强化学习方法的人机混合智能系统框架实现人机序贯决策问题的求解，为后续的优化设计方案提供基础性的对比。

2. 其次，针对人机序贯决策中的介入控制设计问题 (对应第 4 章)，提出了基于自主性边界的优化方案，具体分别包括人机自主性边界定义及判定、利用获得的自主性边界衡量介入触发的时机，进而实现介入控制的优化设计。自主性边界信息一方面使得人和机器外在表现出更清晰的决策权限，有利于人机更好协作完成任务；另一方面使得人和机器输出的更优决策动作得以被仲裁选择，促进整体决策系统的优化。该方案涉及两种应用场景：人介入机器和机器介入人，将以往的人介入和机器介入统一在了同一优化设计方案中，有利于人机介入控制在序贯决策问题的应用。同时，介入控制系统中人与机器处于非对称地位，或者具有人-机器的主从关系，或者具有机器-人的主从关系这在现实应用中广为出现，具有较大的使用价值。
3. 接着，针对人机序贯决策中的共享控制设计问题 (对应第 5 章)，提出了基于自主性边界的优化方案，具体包括共享控制下的自主性边界判定，利用自主性边界实现共享控制中仲裁参数的自适应优化设计，进而完成人机共享控制系统中人与机器动作的融合。不同于第 4 章的人机介入控制，这里的共享控制更加考虑仲裁参数这个重要指标，它决定了人机动作的融合程度，使得最优解在人的动作空间和机器的动作空间共同张成的扩展空间中出现，这无疑是有利的。将自主性上界和自主性下界分别与仲裁机制下判断仲裁参数的上下阈值进行关联，避免固定阈值带来的调参局限，从而生成更加适应动态环境的仲裁参数，完成人机动作的更好融合。仲裁判断还需考虑人类伙伴的意图，通过意图推理获得系统将要完成的任务目标。
4. 最后，针对自主性边界的不确定性问题 (对应第 6 章)，提出了基于贝叶斯神经网络的不确定性估计办法，并将此不确定性信息分别用于介入控制和共享控制中。具体地，利用 dropout 机制实现对贝叶斯神经网络的近似，基于采样输出思想获得自主性边界的概率分布形式，仲裁判断过程中以概率采样获得的自主性边界信息输出决策动作。同样地在算法动态演化过程中存在两个目标：直接影响执行动作的策略网络学习；间接影响执行动作的自主性边界维护。只是不同于上述第 4 章和第 5 章，这里更新维护的是自主性边界的概率分布信息。利用自主性边界的不确定性优化人机混合智能系统中的介入控制和共享控制，也更加符合人们对决策边界的模糊性思考。

7.2　研究展望

本文在已有的人机混合智能方法研究成果上进一步展开研究，扩展了现有关于混合智能方法实现人机序贯决策的优化设计方案。但是，在本文的研究过程中，仍然发现一些不足和值得进一步研究的问题，现总结如下:

1. **讨论人机混合智能系统中的稳定性问题**：面向人和机器的不同决策主体，试图讨论人机混合智能系统的稳定性。比如由于人类智能和机器智能的本质区别，是否会导致人机切换的过程中出现系统发散不稳定等现象，面对如此现象，应该如何设计控制器，在实现人机共同决策的同时保证系统的稳定运行。因此，基于诸如以切换控制理论为代表的稳定性讨论具有重要的理论价值和实际意义。
2. **基于人类行为的协作-对抗模型完善人机共同决策问题**：本文基于仲裁机制实现人类智能和机器智能在决策层面的共荣共存。但是设计过程存在一些不足，例如，仅考虑人类动作持续有利有效的情况，而对诸如疲劳情况下的人的失误操作未做处理。考虑人类用户动作利和弊两种情况，当人类用户动作是有利的时候，利用协作模型共同决策；当人类动作是有弊的时候，把人类行为建模为对抗模型增强机器自主决策系统的鲁棒性。如何将协作模型和对抗模型进行结合对人机系统的后续发展和应用具有重要学术价值和工程意义。
3. **复杂环境下多人多机系统的算法设计**：考虑到实际系统规模的扩大以及复杂性，需要对多人多机场景下的混合智能系统进行设计。针对复杂场景存在编队协同作战等情况，设计更符合当下决策需求的多智能体控制系统，涉及如非完全信息下的博弈等问题。基于 MADDPG 算法扩展本文所述的利用自主性边界优化人机混合智能系统，这大大增加了设计难度，但却是更加实用和值得研究的。

参 考 文 献

[1] 徐南荣, 钟伟俊. 科学决策理论与方法[M]. 科学决策理论与方法, 1995.

[2] 郭立夫. 决策理论与方法[M]. 决策理论与方法, 2014.

[3] 王玉民, 周立华, 张荣. 序贯决策方法的应用[J]. 技术经济, 1996(11):57-59.

[4] LITTMAN M L. Algorithms for sequential decision-making[M]. Brown University, 1996.

[5] BARTO A G, SUTTON R S, WATKINS C. Learning and sequential decision making[M]. University of Massachusetts Amherst, MA, 1989.

[6] RABINOVICH M I, HUERTA R, AFRAIMOVICH V. Dynamics of sequential decision making[J]. Physical review letters, 2006, 97(18):188103.

[7] AU T C, ZHANG S, STONE P. Autonomous intersection management for semi-autonomous vehicles[M]//Routledge Handbook of Transportation. Routledge, 2015: 116-132.

[8] ALTCHÉ F, QIAN X, DE LA FORTELLE A. An algorithm for supervised driving of cooperative semi-autonomous vehicles[J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 18(12):3527-3539.

[9] GRAY A, GAO Y, HEDRICK J K, et al. Robust predictive control for semi-autonomous vehicles with an uncertain driver model[C]//2013 IEEE intelligent vehicles symposium (IV). IEEE, 2013: 208-213.

[10] FUCHS K. Minimally invasive surgery[J]. Endoscopy, 2002, 34(02):154-159.

[11] ALAGOZ O, HSU H, SCHAEFER A J, et al. Markov decision processes: a tool for sequential decision making under uncertainty[J]. Medical Decision Making, 2010, 30(4):474-483.

[12] VERDURA J, CARROLL M E, BEANE R, et al. Systems, methods, and instruments for minimally invasive surgery[M]. Google Patents, 2000.

[13] 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.

[14] MURPHY R R. Human-robot interaction in rescue robotics[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2004, 34(2):138-153.

[15] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning[M]. MIT press, 2016.

[16] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. nature, 2015, 521(7553):436-444.

[17] POOLE D, MACKWORTH A, GOEBEL R. Computational intelligence[J]. 1998.

[18] RUSSELL S, NORVIG P. Artificial intelligence: a modern approach[J]. 2002.

[19] CHANG H S, FU M C, HU J, et al. Google deep mind’s alphago[J]. OR/MS Today, 2016, 43(5):24-29.

[20] HAMET P, TREMBLAY J. Artificial intelligence in medicine[J]. Metabolism, 2017, 69:

S36-S40.

[21] ALTMANN J, SAUER F. Autonomous weapon systems and strategic stability[J]. Survival, 2017, 59(5):117-142.

[22] DE BOISBOISSEL G. Uses of lethal autonomous weapon systems[C]//International Conference on Military Technologies (ICMT) 2015. IEEE, 2015: 1-6.

[23] SEARLE J. Language and society: Philosophy in the real world[M]. Basic Books, NY, 1999.

[24] GRACE K, SALVATIER J, DAFOE A, et al. When will ai exceed human performance? evidence from ai experts[J]. Journal of Artificial Intelligence Research, 2018, 62:729-754.

[25] ZANZOTTO F M. Human-in-the-loop artificial intelligence[J]. Journal of Artificial Intelligence Research, 2019, 64:243-252.

[26] CRANOR L F. A framework for reasoning about the human in the loop[J]. 2008.

[27] CASSADY J T, ROBINSON C, POPA D O. Increasing user trust in a fetching robot using explainable ai in a traded control paradigm[C]//Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments. 2020: 1-8.

[28] VENKATARAMAN S, HAYATI S. Shared/traded control of telerobots under time delay[J]. Computers & electrical engineering, 1993, 19(6):481-494.

[29] 姚立根, 陈希军, 李继勇, 等. 工程导论[M]. 工程导论, 2012.

[30] 李玲, 解洪成, 陈圻. 复杂人机系统人机协作模型的探讨[J]. 人类工效学, 2007(04):36-38.

[31] LEE J, MORAY N. Trust, control strategies and allocation of function in human-machine systems[J]. Ergonomics, 1992, 35(10):1243-1270.

[32] FONG T, THORPE C, BAUR C. Collaboration, dialogue, human-robot interaction[M]// Robotics Research. Springer, 2003: 255-266.

[33] FLEMISCH F, HEESEN M, HESSE T, et al. Towards a dynamic balance between humans and automation: authority, ability, responsibility and control in shared and cooperative control situations[J]. Cognition, Technology & Work, 2012, 14(1):3-18.

[34] ROGERS J, COSTELLO M. Control authority of a projectile equipped with a controllable internal translating mass[J]. Journal of Guidance, Control, and Dynamics, 2008, 31(5):1323-1333.

[35] HAYATI S, VENKATARAMAN S. Design and implementation of a robot control system with traded and shared control capability[C]//1989 IEEE International Conference on Robotics and Automation. IEEE Computer Society, 1989: 1310-1311.

[36] LO B. Assessing decision-making capacity[J]. Law, Medicine and Healthcare, 1990, 18(3): 193-201.

[37] CHARLAND L C. Decision-making capacity[J]. 2008.

[38] GOMBOLAY M C, GUTIERREZ R A, CLARKE S G, et al. Decision-making authority,

team efficiency and human worker satisfaction in mixed human–robot teams[J]. Autonomous Robots, 2015, 39(3):293-312.

[39] YOUNG J E, SUNG J, VOIDA A, et al. Evaluating human-robot interaction[J]. International Journal of Social Robotics, 2011, 3(1):53-67.

[40] MUTLU B, TERRELL A, HUANG C M. Coordination mechanisms in human-robot collaboration[C]//Proceedings of the Workshop on Collaborative Manipulation, 8th ACM/IEEE International Conference on Human-Robot Interaction. Citeseer, 2013: 1-6.

[41] SCHOLTZ J. Theory and evaluation of human robot interactions[C]//36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the. IEEE, 2003: 10-pp.

[42] MILLER C. Using delegation as an architecture for adaptive automation[R]. AIR FORCE RESEARCH LAB WRIGHT-PATTERSON AFB OH HUMAN EFFECTIVENESS DIRECTORATE, 2005.

[43] MILLER C A, PARASURAMAN R. Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control[J]. Human factors, 2007, 49(1): 57-75.

[44] INAGAKI T, et al. Adaptive automation: Sharing and trading of control[J]. Handbook of cognitive task design, 2003, 8:147-169.

[45] SCHIEBEN A, TEMME G, KÖSTER F, et al. How to interact with a highly automated vehicle. generic interaction design schemes and test results of a usability assessment[J]. Human Centred Automation, 2011:251-267.

[46] KELSCH J, FLEMISCH F, LÖPER C, et al. Links oder rechts, schneller oder langsamer? grundlegende fragestellungen beim cognitive systems engineering von hochautomatisierter fahrzeugführung[J]. 2006.

[47] HEESEN M, KELSCH J, LÖPER C, et al. Haptisch-multimodale interaktion für hochautomatisierte, kooperative fahrzeugführung bei fahrstreifenwechsel-, brems-und ausweichmanövern[J]. 2010.

[48] SEPPELT B D, LEE J D. Making adaptive cruise control (acc) limits visible[J]. International journal of human-computer studies, 2007, 65(3):192-205.

[49] PARASURAMAN R, MANZEY D H. Complacency and bias in human use of automation: An attentional integration[J]. Human factors, 2010, 52(3):381-410.

[50] FINGER R, BISANTZ A M. Utilizing graphical formats to convey uncertainty in a decision-making task[J]. Theoretical Issues in Ergonomics Science, 2002, 3(1):1-25.

[51] MCGUIRL J M, SARTER N B. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information[J]. Human factors, 2006, 48(4): 656-665.

[52] BELLER J, HEESEN M, VOLLRATH M. Improving the driver–automation interaction: An approach using automation uncertainty[J]. Human factors, 2013, 55(6):1130-1141.

[53] KORTENKAMP D, BONASSO R P, RYAN D, et al. Traded control with autonomous robots as mixed initiative interaction[C]//AAAI Symposium on Mixed Initiative Interaction. 1997: 89-94.

[54] 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 (AIM). IEEE, 2017: 227-234.

[55] METCALFE J S, ALBAN J, COSENZO K, et al. Field testing of tele-operation versus shared and traded control for military assets: an evaluation involving real-time embedded simulation and soldier assessment[C]//Unmanned Systems Technology XII: volume 7692. International Society for Optics and Photonics, 2010: 769206.

[56] ITOH T, KOSUGE K, FUKUDA T. Human-machine cooperative telemanipulation with motion and force scaling using task-oriented virtual tool dynamics[J]. IEEE Transactions on robotics and automation, 2000, 16(5):505-516.

[57] DEGANI A, GOLDMAN C V, DEUTSCH O, et al. On human–machine relations[J]. Cognition, Technology & Work, 2017, 19(2):211-231.

[58] SANG H, WANG S, LI J, et al. Control design and implementation of a novel master–slave surgery robot system, microhand a[J]. The International Journal of Medical Robotics and Computer Assisted Surgery, 2011, 7(3):334-347.

[59] 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.

[60] 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.

[61] FRIDMAN L, DING L, JENIK B, et al. Arguing machines: Human supervision of black box AI systems that make life-critical decisions[C/OL]//IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, 2019: 1335-1343. http://openaccess.thecvf.com/content_CVPRW_2019/html/WAD/Fridman_Arguing_Machines_Human_Supervision_of_Black_Box_AI_Systems_That_CVPRW_2019_paper.html. DOI: 10.1109/CVPRW.2019.00173.

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

[63] 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.

[64] YU H, SPENKO M, DUBOWSKY S. An adaptive shared control system for an intelligent mobility aid for the elderly[J]. Autonomous Robots, 2003, 15(1):53-66.

[65] BOESSENKOOL H, ABBINK D A, HEEMSKERK C J, et al. A task-specific analysis of the benefit of haptic shared control during telemanipulation[J]. IEEE Transactions on Haptics, 2012, 6(1):2-12.

[66] NUTT P C. Expanding the search for alternatives during strategic decision-making[J]. Academy of Management Perspectives, 2004, 18(4):13-28.

[67] 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.

[68] OH Y, TOUSSAINT M, MAINPRICE J. Learning arbitration for shared autonomy by hindsight data aggregation[J]. arXiv preprint arXiv:1906.12280, 2019.

[69] 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.

[70] 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.

[71] JAVDANI S, SRINIVASA S S, BAGNELL J A. Shared autonomy via hindsight optimization [J]. Robotics science and systems: online proceedings, 2015, 2015.

[72] LAM C P. Improving sequential decision making in human-in-the-loop systems[M]. University of California, Berkeley, 2017.

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

[74] FRIDMAN L. Human-centered autonomous vehicle systems: Principles of effective shared autonomy[J]. arXiv preprint arXiv:1810.01835, 2018.

[75] FLEMISCH F, ABBINK D A, ITOH M, et al. Joining the blunt and the pointy end of the spear: towards a common framework of joint action, human–machine cooperation, cooperative guidance and control, shared, traded and supervisory control[J]. Cognition, Technology & Work, 2019, 21(4):555-568.

[76] 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).

[77] WAYTOWICH N R, GOECKS V G, LAWHERN V J. Cycle-of-learning for autonomous systems from human interaction[J]. arXiv preprint arXiv:1808.09572, 2018.

[78] MARSLEN-WILSON W D. Functional parallelism in spoken word-recognition[J]. Cognition, 1987, 25(1-2):71-102.

[79] GREFENSTETTE J J, RAMSEY C L, SCHULTZ A C. Learning sequential decision rules using simulation models and competition[J]. Machine learning, 1990, 5(4):355-381.

[80] READ P, LERMIT J. Bio-energy with carbon storage (becs): a sequential decision approach to the threat of abrupt climate change[J]. Energy, 2005, 30(14):2654-2671.

[81] KOBER J, BAGNELL J A, PETERS J. Reinforcement learning in robotics: A survey[J]. The International Journal of Robotics Research, 2013, 32(11):1238-1274.

[82] BELLMAN R. Dynamic programming[M]. Courier Corporation, 2013.

[83] BELLMAN R E, DREYFUS S E. Applied dynamic programming[M]. Princeton university press, 2015.

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

[85] BELLMAN R. A markovian decision process[J]. Journal of mathematics and mechanics, 1957, 6(5):679-684.

[86] 胡奇英, 刘建庸. 马尔可夫决策过程引论[M]. 马尔可夫决策过程引论, 2000.

[87] HOWARD R A. Dynamic programming and markov processes.[J]. 1960.

[88] GARCIA C E, PRETT D M, MORARI M. Model predictive control: Theory and practice—a survey[J]. Automatica, 1989, 25(3):335-348.

[89] MAYNE D Q, MICHALSKA H. Receding horizon control of nonlinear systems[C]// Proceedings of the 27th IEEE Conference on Decision and Control. IEEE, 1988: 464-465.

[90] CAMACHO E F, ALBA C B. Model predictive control[M]. Springer science & business media, 2013.

[91] BEMPORAD A, MORARI M, DUA V, et al. The explicit linear quadratic regulator for constrained systems[J]. Automatica, 2002, 38(1):3-20.

[92] KAELBLING L P, LITTMAN M L, MOORE A W. Reinforcement learning: A survey[J]. Journal of artificial intelligence research, 1996, 4:237-285.

[93] VAN OTTERLO M, WIERING M. Reinforcement learning and markov decision processes [M]//Reinforcement learning. Springer, 2012: 3-42.

[94] 国务院关于印发新一代人工智能发展规划的通知国发 [2017]35 号[J]. 中华人民共和国国务院公报, 2017(22):7-21.

[95] DING J. Deciphering china’s ai dream[J]. Future of Humanity Institute Technical Report, 2018.

[96] 郑南宁, 刘子熠, 任鹏举, 等. 混合-增强智能: 协作与认知[J]. Frontiers of Information

Technology & Electronic Engineering, 2017, 18(2).

[97] 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.

[98] 欧中洪, 谭言信, 刘科孟, 等. 基于人在回路的混合增强智能需求精准感知方法及系统[Z]. 2020.

[99] 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.

[100] SUN E, NIETO A, LI Z, et al. An integrated information technology assisted driving system to improve mine trucks-related safety[J]. Safety science, 2010, 48(10):1490-1497.

[101] ENJI S, NIETO A, ZHONGXUE L. Gps and google earth based 3d assisted driving system for trucks in surface mines[J]. Mining Science and Technology (China), 2010, 20(1):138-142.

[102] NIETO A, SUN E, LI Z. Real-time assisted driving in openpit mining operations using google earth[J]. Mining engineering, 2010, 62(2):21.

[103] NELSON E C, VERHAGEN T, NOORDZIJ M L. Health empowerment through activity trackers: An empirical smart wristband study[J]. Computers in human behavior, 2016, 62: 364-374.

[104] ZHANG H, ALRIFAAI M, ZHOU K, et al. A novel fuzzy logic algorithm for accurate fall detection of smart wristband[J]. Transactions of the Institute of Measurement and Control, 2020, 42(4):786-794.

[105] NEF T, RIENER R. Armin-design of a novel arm rehabilitation robot[C]//9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005. IEEE, 2005: 57-60.

[106] JU M S, LIN C C, LIN D H, et al. A rehabilitation robot with force-position hybrid fuzzy controller: hybrid fuzzy control of rehabilitation robot[J]. IEEE transactions on neural systems and rehabilitation engineering, 2005, 13(3):349-358.

[107] SUTTON R S, BARTO A G, et al. Reinforcement learning[J]. Journal of Cognitive Neuroscience, 1999, 11(1):126-134.

[108] METROPOLIS N, ULAM S. The monte carlo method[J]. Journal of the American statistical association, 1949, 44(247):335-341.

[109] SOBOL I M. A primer for the monte carlo method[M]. CRC press, 2018.

[110] TSITSIKLIS J N, VAN ROY B. An analysis of temporal-difference learning with function approximation[J]. IEEE transactions on automatic control, 1997, 42(5):674-690.

[111] TESAURO G, et al. Temporal difference learning and td-gammon[J]. Communications of the ACM, 1995, 38(3):58-68.

[112] SUTTON R S. Generalization in reinforcement learning: Successful examples using sparse coarse coding[J]. Advances in neural information processing systems, 1996:1038-1044.

[113] WATKINS C J C H. Learning from delayed rewards[J]. 1989.

[114] GEIST M, PIETQUIN O. Algorithmic survey of parametric value function approximation [J]. IEEE Transactions on Neural Networks and Learning Systems, 2013, 24(6):845-867.

[115] SUTTON R S, MCALLESTER D A, SINGH S P, et al. Policy gradient methods for reinforcement learning with function approximation[C]//Advances in neural information processing systems. 2000: 1057-1063.

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

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

[118] 高阳. 多智能体系统及应用[M]. 清华大学出版社, 2015.

[119] TAN M. Multi-agent reinforcement learning: Independent vs. cooperative agents[C]// Proceedings of the tenth international conference on machine learning. 1993: 330-337.

[120] BUŞONIU L, BABUŠKA R, DE SCHUTTER B. Multi-agent reinforcement learning: An overview[J]. Innovations in multi-agent systems and applications-1, 2010:183-221.

[121] 罗伯特·吉本斯, 高峰. 博弈论基础[M]. 中国社会科学出版社, 1999.

[122] 艾里克·拉斯缪森. 博弈与信息: 博弈论概论[M]. 北京大学出版社生活·读书·新知三, 2003.

[123] NASH J F, et al. Equilibrium points in n-person games[J]. Proceedings of the national academy of sciences, 1950, 36(1):48-49.

[124] JONES A J. Game theory: Mathematical models of conflict[M]. Elsevier, 2000.

[125] SHAPLEY L S. Stochastic games[J]. Proceedings of the national academy of sciences, 1953, 39(10):1095-1100.

[126] FILAR J, VRIEZE K. Competitive markov decision processes[M]. Springer Science & Business Media, 2012.

[127] KONONENKO I. Bayesian neural networks[J]. Biological Cybernetics, 1989, 61(5):361-370.

[128] NEAL R M. Bayesian learning for neural networks: volume 118[M]. Springer Science & Business Media, 2012.

[129] KULLBACK S. Information theory and statistics[M]. Courier Corporation, 1997.

[130] KULLBACK S, LEIBLER R A. On information and sufficiency[J]. Annals of Mathematical Statistics, 1951, 22(1):79-86.

[131] JORDAN M I, GHAHRAMANI Z, JAAKKOLA T S, et al. An introduction to variational methods for graphical models[J]. Machine learning, 1999, 37(2):183-233.

[132] DUANE S, KENNEDY A D, PENDLETON B J, et al. Hybrid monte carlo[J]. Physics letters

B, 1987, 195(2):216-222.

[133] BOTTOU L. Stochastic gradient descent tricks[M]//Neural networks: Tricks of the trade. Springer, 2012: 421-436.

[134] BLUNDELL C, CORNEBISE J, KAVUKCUOGLU K, et al. Weight uncertainty in neural network[C]//International Conference on Machine Learning. PMLR, 2015: 1613-1622.

[135] LIU A, PENTLAND A. Towards real-time recognition of driver intentions[C]//Proceedings of Conference on Intelligent Transportation Systems. IEEE, 1997: 236-241.

[136] WASSON G, SHETH P, ALWAN M, et al. User intent in a shared control framework for pedestrian mobility aids[C]//Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453): volume 3. IEEE, 2003: 2962-2967.

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

[138] SUTTON R S. Dyna, an integrated architecture for learning, planning, and reacting[J]. ACM Sigart Bulletin, 1991, 2(4):160-163.

[139] KAISER L, BABAEIZADEH M, MILOS P, et al. Model-based reinforcement learning for atari[J]. arXiv preprint arXiv:1903.00374, 2019.

[140] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. arXiv preprint arXiv:1509.02971, 2015.

[141] SILVER D, LEVER G, HEESS N, et al. Deterministic policy gradient algorithms[C]// International conference on machine learning. PMLR, 2014: 387-395.

[142] AMODEI D, OLAH C, STEINHARDT J, et al. Concrete problems in ai safety[J]. arXiv preprint arXiv:1606.06565, 2016.

[143] IRVING G, CHRISTIANO P, AMODEI D. Ai safety via debate[J]. arXiv preprint arXiv:1805.00899, 2018.

[144] ABOU ALLABAN A, DIMITROV V, PADIR T. A blended human-robot shared control framework to handle drift and latency[C]//2019 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR). IEEE, 2019: 81-87.

[145] GAL Y, GHAHRAMANI Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning[C]//international conference on machine learning. PMLR, 2016: 1050-1059.

[146] GAL Y. Uncertainty in deep learning[J]. 2016.

[147] 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.

[148] JAIN S, ARGALL B. Probabilistic human intent recognition for shared autonomy in assistive

robotics[J]. ACM Transactions on Human-Robot Interaction (THRI), 2019, 9(1):1-23.

[149] HERNÁNDEZ-LOBATO J M, ADAMS R. Probabilistic backpropagation for scalable learning of bayesian neural networks[C]//International conference on machine learning. PMLR, 2015: 1861-1869.

致　谢

转眼间又到了毕业的季节，与以往不同的是，这一次的主人公换成了自己。回首博士期间的种种过往，感慨万千，总的来说是欢喜和开心多于忧伤焦虑。在这里不仅学到了专业知识和科研态度，更了解了做人的道理和做事的原则。在此博士论文即将完成之际，向支持、关心、帮助过我的人们致以最诚挚的感谢。

首先，感谢我的导师康宇教授。感谢老师选择我成为您的学生，在康老师的带领下，无论在科研学习方面还是工作生活方面我都受益良多。康老师在教学科研方面秉承严谨的治学态度、踏实肯干的工作作风，在指导学生方面能够为人师表，以身作则，待人周到随和，经常站在学生的角度考虑问题和解决问题。康老师的这种亲切随和使得实验室环境较为团结轻松，当看到有新闻说某某院校博士生因压力过大选择了绝路，我都非常庆幸自己在一个欢快和谐的科研环境中度过了我的博士生涯。因此在以后的工作生活中，希望自己能够以老师为标杆，持续学习老师的科研工作态度和待人接物方式。

此外，感谢指导老师赵云波教授。感谢赵老师在日常科研中的不吝赐教，以及对我博士论文的修改。赵老师专业上的逻辑思维和工作上的严谨细致是我学习的榜样，也正是这种诲人不倦和为人师表，将我从一个小白领入人机混合智能领域。同时，我要感谢奚宏生老师、季海波老师、殷保群老师、秦家虎老师、凌强老师、朱进老师在我专业知识上的教授。感谢信息学院洪力奋老师，班主任陈金雯老师对班级日常事宜、毕业生工作的付出和支持。感谢自动化系郑烃老师，王大欣老师，张洁老师给予我的帮助。

在博士生涯，我要感谢实验室的吕文君师兄、李泽瑞师姐、李鹏飞师兄的支持和照顾，他们在科研学习和工作生活上的答疑让我受益良多；感谢实验室的虞佩龙、许镇义、王雪峰、陈绍冯、陈国勇给予我的各式各样的建议和帮助。特别感谢曾经同讨论组的李鹏飞师兄和虞佩龙同学，感谢他们在我学习新方向时对我的引导。感谢我的好朋友也是同年级战友杨钰潇以及刘纯含师妹，我们平日里相互分享喜乐与忧愁，她们的陪伴使我的博士生活丰富多彩。感谢王涛师弟经常主动帮助我处理杂七杂八的事情。感谢同研究方向游诗艺师妹作为答辩秘书帮忙处理毕业相关事宜。同样感谢实验室的赵振怡、昌吉、张年坤、许婷、李靖、李明、殷书慧等从 17 级到 21 级的所有师弟师妹。感谢浙江工业大学的王岭人、唐敏、吴芳、卢子轶、花婷婷等同研究方向师弟师妹和我在学术问题上的共同探讨。对这些我所熟识的兄弟姐妹们，我想说虽然世界上没有相同的两片树叶，可是在他们身上我看到了很多类似的优良品质：踏实、努力、真诚、和善等。

感谢我的爸爸妈妈以及所有家人在我博士期间对我的支持和关心，和谐幸

福的家庭环境使得我可以安心且放心的做自己喜欢的事情，您们的期盼和鼓励使得我能够勇于追求自己的梦想。

最后，特别感谢我们的国家，无论是面对不安稳的国际环境，还是突如其来的新冠疫情，都能够有雄厚的处理危机实力和话语权，使得我们这一代后辈能够生长在和平安全祥和的环境中。然而这些都是有人负重前行所换来的，希望自己能坚持不懈，努力奋斗，做一个有用的人。无愧国家的培养，师长的教诲，家人的嘱托，以梦为马，不负韶华。

张倩倩

二零二一年十月十五日

于中国科学技术大学科技西楼 1310

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. **Qianqian Zhang**, Yu Kang, Peilong Yu, et al, ”Sampled-data Stabilization of a class of Stochastic Nonlinear Markov Switching system with Indistinguishable Modes based on the Approximate Discrete-time Models,” *Journal of Systems Science and Complexity*, pp. 1-17, 2021.
2. **Qianqian Zhang**, Yun-Bo Zhao, Yu Kang, ”Autonomous Boundary of Human-Machine Collaboration System Based on Reinforcement Learning,” *Australian and New Zealand Control Conference (ANZCC)*, Gold Coast, Australia, 2020, pp. 160-165. （对应正文第 3 章）
3. **张倩倩**, 余道洋, 李民强, ” 基于混合策略的移动机器人避障算法探究,” 控制工程, vol. 26, no. 7, pp. 1328-1334.
4. Peilong Yu, Yu Kang, **Qianqian Zhang**, ”Sampled-data stabilization for a class of stochastic nonlinear systems with Markovian switching based on the approximate discrete-time models,” *Australian and New Zealand Control Conference (ANZCC)*, Melbourne, Australia, 2018, pp. 413-418.
5. Liang Li, **Qianqian Zhang**, Gongmei Qi, et al., ”Algorithm of obstacle avoidance for autonomous surface vehicles based on LIDAR detection,” *IEEE Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Chengdu, China, 2019, pp. 1823-1830.
6. Chunhan Liu, Yu Kang, Peilong Yu, **Qianqian Zhang**, ”Sampled-data stabilization for a kind of stochastic nonlinear systems driven by G-Brownian motion,” *Chinese Automation Congress (CAC)*, Hangzhou, China, 2019, pp. 3642-3646.

待发表论文

1. **Qianqian Zhang**, Yu Kang, Yun-Bo Zhao, Shiyi You, ”Traded Control of Human-Machine systems for Sequential Decision Making based on Reinforcement Learning,” *IEEE Transactions on Artificial Intelligence*, 二审 minor revision.（对应正文第 4 章）
2. Yu Kang, **Qianqian Zhang**, Yun-Bo Zhao, Xuefeng Wang, Wenjun Lv, ”Arbitration Optimization of Human- Machine Shared Control to Achieve Sequential Decision-Making,” *Science China Information Sciences,* 在投（对应正文第 5、

6 章）

3. Shiyi You, Yu Kang, Yun-Bo Zhao, **Qianqian Zhang**, ”Adaptive Arbitration for Minimal Intervention Shared Control via Deep Reinforcement Learning,” *Chinese Automation Congress (CAC)*, Beijing, China, 2021, Accepted.

实审中的专利

1. 康宇，**张倩倩**，王雪峰，游诗艺，吕文君。一种基于强化学习的人机融合自主性边界切换方法及系统，公开号 CN111753982 A，公开时间: 2020-10-09。