中国科学: 信息科学

SCIENTIA SINICA Informationis

论文

# 面向人机序贯决策实现共享控制下的仲裁优化

张倩倩[1], 赵云波[2,3,4*], 吕文君[2,3,4], 陈谋[5]

1. 安徽大学人工智能学院, 合肥 230039, 中国
2. 中国科学技术大学自动化系, 合肥 230026, 中国
3. 合肥综合性国家科学中心人工智能研究院, 合肥 230088, 中国
4. 中国科学技术大学先进技术研究院, 合肥 230031, 中国
5. 南京航空航天大学自动化学院, 南京 210016, 中国

* 通信作者. E-mail: ybzhao@ustc.edu.cn

国家重点研究开发项目(批准号: 2018AAA0100801)资助项目、国家自然科学基金(批准号: 62173317, 62203006)、安徽省重点研发计划(批准号: 202104a05020064)资助项目

**摘要** 考虑到共享控制出现于众多由人类智能和机器智能共同参与的序贯决策场景, 并且人的决策范围和智能机器的决策范围尚未予以明确划分, 由此需要加以实时仲裁从而达到人机共存并且共享决策权限. 为此本文提出了一种仲裁优化方法, 该方法的独特之处在于自主性边界概念优化了共享控制中人机决策动作的仲裁机制. 本文为自主性边界的计算和更新维护提供了思路, 能够基于贝叶斯规则的意图推理分析人机共享系统可能要实现的目标, 从而确定仲裁参数的选择. 此外, 本文还分析了自主性边界的不确定性以促进边界信息对共享控制中决策质量的优化效果. 最后实验结果表明, 所提出的方法在累积奖励、成功率、撞击率方面表现出色, 这些说明了本文提出的共享控制中的仲裁优化方法在求解人机序贯决策问题的有效性和价值.

**关键词** 共享控制, 仲裁优化, 自主性边界, 人机序贯决策, 强化学习

## 1 引言

在《新一代人工智能发展规划》[1,2]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. 之所以需要对此研究方向进行探讨, 源于人脑的天然智能和机器的高效高精度. 人类独特的认知能力使其经常密切参与到各种各样的智能机器决策场景中, 如人机搜救系统[3,4]、辅助微创手术系统[5,6]、开放环境下人机共驾系统[7,8] 等, 其中人类扮演决策者或辅助决策者的角色[9]. 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”[10~14]. 以时序性和多阶段性为标志的序贯决策问题是一类广泛存在于社会、经济、军事、工业生产等各个领域的重要决策问题. 考虑通过有效融合人类智能和机器智能来完成或改善序贯决策的过程, 即是本文的研究问题“人机序贯决策”.

仲裁是人机混合决策之间的一种平衡机制(参见图1为人机共享控制仲裁框架示意图), 是面向序贯决策实现有效人机协调的一种主要方法, 其中平衡通常以人机决策的线性组合形式实现, 涉及到仲裁权重因子$\alpha$ [15,16] 的求解. $\alpha$ 可以通过多种方式确定, 它取决于多个决策主体之间的决策差异[8], 或

者决策可信度, 人类意图预测可信度等[17,18]. 在大多数现有研究中, $\alpha$ 的下限阈值和上限阈值是根据人工经验被假设为恒定的. 然而, 实际上仲裁因子阈值与动态变化的环境和决策中的不确定性密切相关, 这意味着阈值的固定假设对于确定$\alpha$ 来说过于保守.

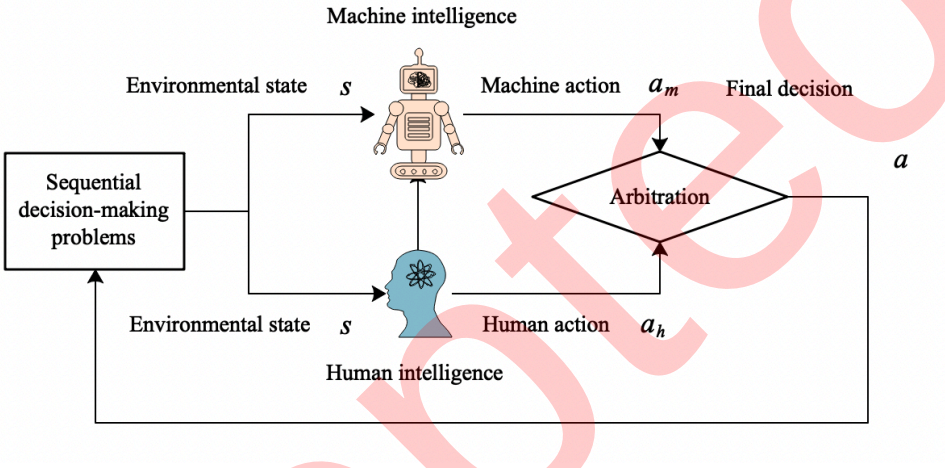


图 1　人机共享控制仲裁框架示意图.

**Figure 1**　Schematic diagram of human-machine shared control arbitration framework.

本文利用自主性边界的概念设计仲裁因子$\alpha$. 将自主性边界定义为人类智能和机器智能按照有利于共享控制联合优化目标的方向做出决策和行动的范围. 显然, 我们所定义的自主性边界能够与人和机器的决策能力挂钩, 因此可以将其作为确定上述仲裁阈值的关键因素.

利用自主性边界信息的关键挑战是它的量化计算, 这是以前从未见过的且其本质上是困难的, 因为智能很难量化是目前存在的现实问题. 为了解决这个困难, 本文不直接计算自主性边界, 而是通过描述一个优化问题用来量化人和机器的动作对其联合性能指标的影响. 此外, 针对自主性边界可能存在单值估计不准确的情况, 本文进一步对其不确定性进行估计, 进而用于共享控制中的仲裁优化. 该项工作提出的优化方法不仅能够通过优化人机共享控制中的固定阈值来提高决策性能, 并且实时获取和更新维护自主性边界信息, 这对于划分共享控制中的人机决策权限也是有利指导. 本文的主要贡献可总结为三点:

- 面向人机序贯决策提出了共享控制下的仲裁优化方案, 通过自适应调节仲裁因子改善决策性能;
- 提供了一种自主性边界信息的判定和维护方式, 使得人机共享控制系统中的决策权限有了界定;
- 基于贝叶斯神经网络对自主性边界进行不确定性估计, 并将获得的不确定性信息用于决策动作生成.

本文的章节安排如下. 第2节给出问题描述并讨论相关工作. 第3节详细阐述文章主要框架和方法. 第4节验证所提方法的有效性, 第5节总结本文的工作.

## 2　问题描述及相关工作

### 2.1　问题描述

考虑使用共享控制系统来实现序贯决策. 机器代理结合推断出的人类意图和自身的策略制定更优

的策略. 人和机器的决策动作都由仲裁模块处理产生一个混合的最终决策. 将该问题形式化为以下优化问题,

$$\max_{a(t)\in A} J^r(s(t),a(t)) = Q(s(t),a(t)) = \sum_{t=0}^{\mathrm{T}} \gamma^t R(s(t),a(t)), \tag{1a}$$

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

$$a_m(t) = p(s(t), g(t); \theta), \tag{1c}$$

$$a_h(t) = \text{Human-action}, \tag{1d}$$

$$\{g(t), c(t)\} = \text{Infer}(a_h(t-\mathrm{N}), \cdots, a_h(t)), \tag{1e}$$

$$t = 0, 1, 2, 3, \cdots. \tag{1f}$$

其中$s(t)$和$a(t)$是系统的环境状态和决策动作, $Q(s(t),a(t))$是状态动作对$s(t)$和$a(t)$的值函数, $R(s(t),a(t))$是执行动作$a(t)$对应的奖励. Infer$(\cdot)$为意图推断函数, 其输出是推断的目标$g(t)$和置信度$c(t)$. $p(\cdot)$代表机器代理的策略函数, 参数为$\theta$, 其输出机器的决策动作. $f^{\alpha}(\cdot)$代表人机决策动作的仲裁函数, 仲裁因子与置信度$c(t)$以及第三节主要方法中提到的自主性边界息息相关, 其决定了人机共享控制最以强化学习的基础以强化学习的基础以强化学习的基础后决策结果, 定义如下：

$$f^{\alpha}(a_h(t), a_m(t), c(t)) = (1-\alpha)a_h(t) + \alpha a_m(t), \tag{2}$$

$$\alpha = \begin{cases} 0, & c(t) \leqslant \epsilon_1; \\ \frac{\epsilon(c(t)-\epsilon_1)}{\epsilon_2-\epsilon_1}, & \epsilon_1 < c(t) < \epsilon_2; \\ \epsilon, & c(t) \geqslant \epsilon_2. \end{cases} \tag{3}$$

$\alpha$ 根据意图推理置信度$c(t)$ 确定, $\epsilon_1$、$\epsilon_2$ 分别代表$c(t)$的下限和上限. 意图推理置信度$c(t)$ 越小, 仲裁函数$f^a(\cdot)$ 将趋于完全使用机器决策, $c(t)$ 越大, 仲裁函数$f^a(\cdot)$ 将趋于完全使用人类决策, 否则最终决策动作将在人和机器之间混合平衡. 注意$\epsilon_1$和$\epsilon_2$这两个限制将自然与本文接下来所提的自主性边界相关联, 从而上述形式仲裁函数的输出则与自主性边界的判断和估计密切相关.

本文以强化学习为基础研究所提出的共享控制优化算法, 关于算法的收敛性, 以通常情况下判断算法收敛性的利器－压缩映射原理为思路. 压缩映射原理：对于任何在算子$T(v)$下完备(即封闭)的度量空间$V$, 如果算子$T$为$\gamma-$压缩, 那么：$T(v)$最终收敛到一个唯一的固定点$v^*$; 线性收敛速度正比于$\gamma$,其中$\gamma < 1$. 其中$\gamma-$压缩是指$||T(v_1) - T(v_2)||_\infty \leqslant \gamma||v_1 - v_2||_\infty$, 其中$||v_1 - v_2||_\infty$表示任意两个值函数在任意状态的最大差距.

关于此定理可以理解为：每次算子$T(v)$作用到某个$v$, 都会压缩$v$和$v^*$的距离, 直到最终收敛$v^*$. 而此处的算子可以看成文章中公式1中的优化目标计算过程$J^r(\cdot)$, 且将$\sum_t^{\mathrm{T}} \gamma^t R(s(t),a(t)) = R(s(t),a(t)) + \sum_{t+1}^{\mathrm{T}} \gamma^{t+1} R(s(t+1),a(t+1))$记为$J^r(v) = R(s,a) + \gamma P^a v$, $P^a$为状态转移模型, 则可得

$$||J^r(v_1) - J^r(v_2)||_\infty \leqslant \gamma||v_1 - v_2||_\infty = ||\gamma P^a v_1 1 - \gamma P^a v_2||_\infty \leqslant \gamma||v_1 - v_2||_\infty. \tag{4}$$

故而根据压缩映射定理, 则可保证收敛性.

## 2.2 相关工作

### 2.2.1 人机序贯决策

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, 本文将上述两种场景统称为“人机序贯决策问题”. 其求解与动态规划 [19]和MDP [20]等领域密切相关. 本文面向人机序贯决策问题, 讨论如何利用人与智能机器同时参与的共享控制优化方法实现改善决策性能.

### 2.2.2 共享控制

共享控制是结合人类策略和智能自主策略来完成任务目标或提高决策性能的一种范式[21~24]. 它为全自动机器人系统提供了一种替代方案, 可用于偏向用户友好型的场景(如辅助驾驶、自主武器系统[25]和智能学习辅助系统[26])以扩展现有机器人的有效性. 本文使用共享控制方法实现人机序贯决策, 其中两个决策主体需要共同完成任务目标. 更多关于共享控制的细节可参考[8, 23, 27~32].

### 2.2.3 仲裁

仲裁机制对于人和智能机器之间的切换或混合是必要的, 它是一种平衡机制. 通常共享控制中的常见融合形式是用户和机器代理策略之间的线性组合. 仲裁参数可能取决于不同的因素, 例如对用户意图预测的置信度, 或者每个命令[15][22][33] 之间的差异. 更多关于仲裁的细节可参考[15, 34~36].

### 2.2.4 强化学习

强化学习侧重于代理如何在环境中采取不同的行动来最大化累积奖励. 求解强化学习问题实际上就是求取最优策略. 换句话说, 强化学习的求解过程就是优化贝尔曼方程的过程. 强化学习作为实现人机共享控制的有效方法, 近年来被大量研究应用. 可通过[23, 34, 37]等了解更多详情.

### 2.2.5 贝叶斯推理

贝叶斯推理[38, 39]是在经典统计归纳推理-估计和假设检验的基础上发展起来的一种新的推理方法. 作为一种推理方法, 贝叶斯推理是从概率论中的贝叶斯定理扩展而来的,

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}. \tag{5}$$

其中$H$代表一个假设, 其概率可能会受到实验数据(以下简称证据)的影响, $E$ 表示证据. 证据对应于尚未用于计算先验概率的新数据. 贝叶斯推理将后验概率(考虑相关证据或数据后事件的条件概率)推导出为两个前因, 即先验概率(考虑相关证据或数据之前事件的不确定性概率)和似然函数(源自概率模型).

## 3 主要方法

本节介绍如何将自主性边界应用于人机共享控制的仲裁优化涉及中, 并详细讨论共享控制下自主性边界的判定方法, 如图2所示.

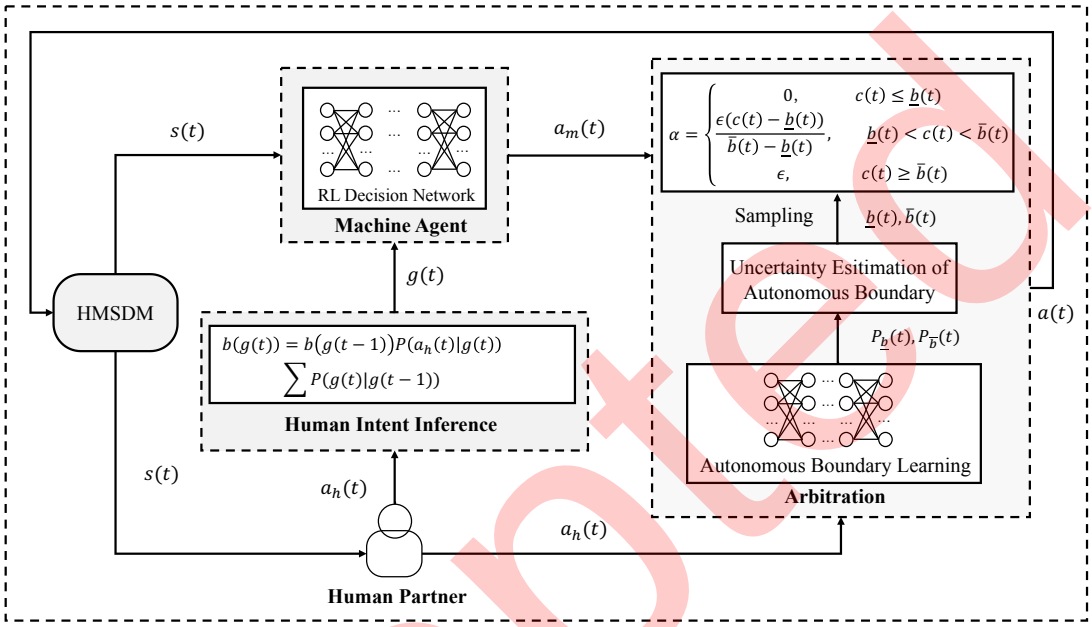


图 2　人机共享控制下仲裁优化的总体框架.

**Figure 2**　The overall framework of arbitration optimization for human-machine shared control.

### 3.1　人机共享控制的仲裁优化

基于自主性边界判定和不确定性估计(将会在第3.2节介绍), 本节首先阐述如何将自主性边界的概念引入到共享控制算法中, 以实现更好的仲裁参数设计. 更深入地, 本文考虑到自主性边界固有的模糊性, 并分析自主性边界的不确定性, 使得自主性边界和仲裁参数服从一定的概率分布, 这对于混合决策是有用的. 人机共享控制的优化过程有两个任务目标：直接影响决策动作的策略网络; 间接影响决策动作的具有不确定性的自主性边界信息. 通过动态实时更新维护自主边界的概率分布, 然后根据得到的不确定性信息, 对机器决策动作和人类决策动作进行混合仲裁, 进而产生最终决策动作.

定义人机共享控制系统的共同目标函数如下,

$$\max J^b_{h,m}(s(t), a(t)) = J^r(s(t), a(t)). \tag{6}$$

其中$J^b_{h,m}(s(t), a(t))$是自主性边界的优化目标函数, $J^r(s(t), a(t))$是人机序贯决策问题的优化目标(如式(1a)). 另外, 将仲裁参数计算公式(3)中的超参数$\epsilon_1, \epsilon_2$描述为如下形式.

$$\epsilon_1 \propto J^b_{h,m}(s(t), \underline{b}(t)), \tag{7a}$$

$$\epsilon_2 \propto J^b_{h,m}(s(t), \bar{b}(t)). \tag{7b}$$

因此, 公式(2) 中仲裁参数的计算不再过多地依赖固定的超参数, 而是可通过自适应调节. 自主性上界$\bar{b}(t)$和自主性下界$\underline{b}(t)$被用于仲裁参数大小的计算中, 系统根据实时状态、人机决策动作以及先验分布采样得到的自主性边界, 代入上式可计算出仲裁参数的大小$\alpha$, 进而得出人机混合决策结果$a(t)$. 因

此可将式3重写如下：

$$\alpha=\begin{cases}0, & c(t)\leqslant \underline{b}(t);\\ \frac{\epsilon(c(t)-\underline{b}(t))}{\bar{b}(t)-\underline{b}(t)}, & \underline{b}(t)<c(t)<\bar{b}(t);\\ \epsilon, & c(t)\geqslant \bar{b}(t).\end{cases} \tag{8}$$

**算法 1** 基于自主性边界的共享控制优化算法

**初始化：** 自主性边界的先验分布信息：$B(t-1)=\{\bar{b}(t-1),\underline{b}(t-1)\}$: $\bar{b}(t-1)\sim N(\mu_0,\sigma_0)$, $\underline{b}(t-1)\sim N(\mu_0,\sigma_0)$;
**输出：** 当前$t$时刻的仲裁决策动作$a(t)$, 以及自主性边界的后验概率分布：$\bar{b}(t)\sim N(\hat{\mu}_u,\hat{\sigma}_l)$, $\underline{b}(t)\sim N(\hat{\mu}_u,\hat{\sigma}_l)$;
**while** $t<$ MAXSTEP or done == FALSE **do**
  输入系统状态$s(t)$;
  人类决策者根据观察到的系统状态$s(t)$输入决策动作$a_h(t)$, 机器代理根据人类动作$\{a_h(t-N+1),\cdots,a_h(t-1),a_h(t)\}$推测出可能的任务目标$g(t)$和对应的目标可信度$c(t)$(Eq. (1e)), 并且计算相应的机器决策动作$a_m(t)$;
  **for** $m<$ T **do**
    根据边界概率分布$B(t-1):\{\bar{b}^{(m)},\underline{b}^{(m)}\}$进行采样;
    根据公式(2),(3) 和(7), 得到随机最优解$\{a(t)^{(m)}\}$;
    基于式(9a) 和(10a), 更新$t$时刻的自主性边界$\{\{\bar{b}(t),\underline{b}(t)\}^{(m)}$;
  **end for**
  根据式(11)计算出最终决策动作$a(t)$;
  由式Eq. (14),(15)更新$t$时刻自主性上界和自主性下界的后验概率分布$\bar{b}(t)\sim N(\hat{\mu}_u,\hat{\sigma}_l)$, $\underline{b}(t)\sim N(\hat{\mu}_u,\hat{\sigma}_l)$.
**end while**

下面介绍优化算法的具体过程, 如算法1所示. 算法1融合了自主性边界信息(将在3.2小节中进一步讨论), 以实现此类人机序贯决策问题的进一步优化求解. 该算法有两个优化目标：1）与决策动作直接相关的策略; 2) 间接影响决策动作的自主性边界. 在系统的动态演化过程中, 对于每一个系统状态$s(t)$, 人类智能都会输出一个有目的的决策行为$a_h(t)$, 由机器代理用来推断人类想要的任务目标. 机器代理预测任务目标后, 根据当前系统状态$s(t)$和当前策略学习情况计算实时机器行为$a_m(t)$. 然后将人和机器的决策动作输入到仲裁模块. 接下来, 根据上一时刻的自主性上界和自主性下界的概率分布进行采样(步骤7). 步骤8 计算随机最优解$\{a(t)^{(m)}\}$ , 步骤9 更新时间$t$的自主性边界$\{\bar{b}(t),\underline{b}(t)\}^{(m)}$(对应于算法2). 最后, 算法获得了决策动作$a(t)$, 并更新了时间$t$时自主性边界的后验概率分布(对应于算法3). 需要注意的是, 本小节中的$c(t)$将在3.3小节中详细介绍.

### 3.2 共享控制下的自主性边界

#### 3.2.1 自主性边界的定义与判定

本小节考虑人机共享控制系统下的自主性边界. 文章不过多强调是人的边界还是机器的边界, 这可能取决于具体实际场景, 但会给出自主性上界和自主性下界边界的定义. 系统的决策动作需要介于自主性上界和自主性下界之间. 如果低于自主性下界或超过自主性上界, 则需要根据具体情况进行约束或过滤. 在更新和维护这两个边界信息时, 如果系统的决策行为在两个边界之间, 则边界信息保持不变; 否则, 自主性上界或自主性下界需要实时优化和更新. 优化后的边界信息可以再次作为决策条件, 这形成一个良性循环, 因此考虑将自主性边界问题定义为如下优化问题,

$$\bar{b}(t)=\arg\max_{a(t)\in\mathcal{A}_h\times\mathcal{A}_m} J^b_{h,m}(s(t),a(t)), \tag{9a}$$

$$\text{s. t. } C(s(t),a(t))<0. \tag{9b}$$

和

$$\underline{b}(t) = \arg \min_{a(t)\in\mathcal{A}_h\times\mathcal{A}_m} J^b_{h,m}(s(t), a(t)), \tag{10a}$$

$$\text{s. t.}\quad C(s(t), a(t)) < 0. \tag{10b}$$

其中$J^b_{h,m}(s(t), a(t))$是人和机器的共同目标函数. (9a) 和(10a) 给出了最大化目标函数的示例, 因此希望在满足约束条件的情况下, 找到最大化目标函数的决策动作$a(t)$, 而决策动作依赖于自主性上下界的信息. 本文可基于DQN算法求解自主性上界和自主性下界, 对应于图2中的“自主性边界学习模块”, 详见算法2和算法3.

**算法 2** 共享控制下的自主性边界判定算法

1: **初始化:** 自主性边界信息: $B = \{\bar{b}(t-1), \underline{b}(t-1)\}$;
2: **输出:** 当前$t$时刻的自主性边界: $B = \{\bar{b}(t), \underline{b}(t)\}$;
3: **while** $t$ <MAXSTEP or done == FALSE **do**
4: 输入系统状态$s(t)$;
5: 基于式(1), 观察人类决策动作$a_h(t)$以及计算机器决策动作$a_m(t)$;
6: 将满足约束的人类输入对应的目标函数与初始化自主性边界对应的目标函数进行比较. 如果$J^b_{h,m}(s(t), a(t)) > J^b_{h,m}(s(t), \bar{b}(t-1))$ ($J^b_{h,m}(s(t), a(t)) < J^b_{h,m}(s(t), \underline{b}(t-1))$), 然后根据公式(9a)(或(10a))更新自主性上(下)界, 否则保持不变.
7: **end while**

共享控制中自主性边界的判定可以正式描述为算法2. 人机共享控制的自主性边界的初始化可根据被控对象和当前可用的信息来获得. 例如, 在没有其他先验信息时, 可以使用神经网络中的随机初始化作为解决方案(使用随机收集的经验样本轨迹等). 在系统动态演化过程中, 对于实时输入的系统状态$s(t)$, 机器决策模块给出对应的决策动作$a_m(t)$以及人类输入决策动作$a_h(t)$. 然后使用自主性边界信息$\bar{b}(t-1), \underline{b}(t-1)$来比较满足约束的决策动作, 目的是更新上自主性上界(或自主性下界), 这将使目标函数(9a)更大或(10a)更小, 如此循环直到训练结束.

### 3.2.2 自主性边界的不确定性估计

在深入探索研究的过程中, 我们发现自主性边界的额外信息固然有用, 但需要建立在准确合理的前提下. 如果有误差, 不仅不能达到优化效果, 而且可能扰乱原系统的决策过程, 这些都是自主性边界的单值估计可能导致的后果. 因此本节继续讨论自主性边界的不确定性估计方法.

根据文献 [40]中的命题(MC dropout)如下:

**命题1** 如果$p(y^*|x^*, \theta) = N(y^*; f^\omega(x^*), \tau^{-1}), \tau > 0, T \to \infty$, 那么

$$\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}} f^\omega(x^*) \to E_{q_\theta(y^*|x^*)}[y^*], \tag{11}$$

$$\tau^{-1}I + \frac{1}{\mathrm{T}} f^\omega(x^*)^{\mathrm{T}} f^\omega(x^*) \to E_{q_\theta(y^*|x^*)}[(y^*)^{\mathrm{T}} y^*], \tag{12}$$

$$V[y^*] = E_{q_\theta(y^*|x^*)}[(y^*)^{\mathrm{T}} y^*] - E_{q_\theta(y^*|x^*)}[y^*]^{\mathrm{T}} E_{q_\theta(y^*|x^*)}[y^*]. \tag{13}$$

其中$q_\theta(y^*|x^*)$在变分推理中表示式(5))中$p(\theta|D)$(or $p(H|E)$的近似.

基于以上描述, 文章使用MC dropout[40]来实现共享控制中自主性边界的不确定性估计. 自主性上界的一阶矩和方差如下,

$$\mu_{\overline{b}}(t) = \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\overline{b}(t), \tag{14a}$$

$$\sigma_{\overline{b}}(t) = \tau^{-1}I + \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\overline{b}(t)^2 - \mu_{\overline{b}}(t)^2. \tag{14b}$$

自主性下界的一阶矩和方差如下,

$$\mu_{\underline{b}}(t) = \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\underline{b}(t), \tag{15a}$$

$$\sigma_{\underline{b}}(t) = \tau^{-1}I + \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\underline{b}(t)^2 - \mu_{\underline{b}}(t)^2. \tag{15b}$$

其中$\overline{b}(t)$ 和$\underline{b}(t)$是由式(9) 和(10)获得.

**算法 3** 自主性边界的不确定性估计

1: **while** $t <$ MAXSTEP or done == FALSE **do**
2: **for** $m <$ T **do**
3: 根据边界概率分布采样$B(t-1):\{\overline{b}^{(m)},\underline{b}^{(m)}\}$;
4: 基于式(3), (9a) 和(10a), 执行算法2计算随机最优决策$\{a(t)^{(m)}\}$以及获得更新后的自主性边界$\{\{\overline{b}(t),\underline{b}(t)\}^{(m)}$;
5: **end for**
6: 由式(14) 和(15), 更新$t$时刻自主性边界的后验概率分布: $\overline{b}(t)\sim N(\hat{\mu}_u,\hat{\sigma}_l)$, $\underline{b}(t)\sim N(\hat{\mu}_u,\hat{\sigma}_l)$.
7: **end while**

将共享控制下自主性边界的不确定性估计的具体想法描述如算法3 所示. 基于贝叶斯推理为自主性边界信息分配先验概率, 本文选择高斯分布作为先验概率分布. 在系统动态演化过程中, 根据自主性上界和自主性下界在上一时刻$t-1$的概率分布进行采样(步骤3). 步骤4 计算随机最优解$\{a(t)^{(m)}\}$, 得到更新后的自主性边界$t$ $\{\{\overline{b}(t),\underline{b}(t)\}^{(m)}$. 最后, 基于优化表达式(14) 和(15) 进行蒙特卡罗估计, 更新$t$时刻自主性边界的后验概率分布, 如此循环直到训练结束. 至此, 得到了自主性边界的判方法和不确定性估计方法.

### 3.3 人类意图推理

面向人机共同决策问题, 系统需要推断人类的决策意图, 做法是通过从经验样本池中取历史动作序列, 基于贝叶斯推理(如公式(5))分析人类动作$a_h(0),\cdots,a_h(t)$对应的可能意图$g(t)$, 其中$g(t)\in G=\{g_1,g_2,\cdots,g_N\}$. 为了解决这个问题, 人类意图推断模块应运而生, 其可以推断出用户实时目标$g(t)$的相关信息.

根据贝叶斯定律, 以观测$a_h$ 为条件, 将$t$ 时间步对应的目标概率表示为

$$\begin{aligned} b(g(t)) &= P(g(t)|a_h\{0:t\}) = P(g(t)|a_h\{0:t-1\},a_h(t)),\\ &\propto P(g(t),a_h\{0:t-1\},a_h(t)),\\ &\propto P(a_h(t)|g(t),a_h\{0:t-1\})P(g(t),a_h\{0:t-1\}),\\ &\propto P(a_h(t)|g(t),a_h\{0:t-1\})P(g(t)|a_h\{0:t-1\}). \end{aligned} \tag{16}$$

**假设1** [41] 给定当前目标估计, 当前时间和历史时间之间的观察是条件独立的.

**假设2** (Markov 假设) [41] 给定前一个时间步的目标估计, 当前时间步的目标估计和历史观测值是条件独立的.

根据假设1 和假设2, 公式(16) 变为

$$\begin{aligned} b(g(t)) \propto & P(a_h(t)|g(t))P(g(t)|a_h\{0:t-1\}), \\ \propto & P(a_h(t)|g(t)) \sum_{g(t-1)\in G} P(g(t),g(t-1)|a_h\{0:t-1\}), \\ \propto & P(a_h(t)|g(t)) \sum_{g(t-1)\in G} P(g(t)|g(t-1),a_h\{0:t-1\})P(g(t-1)|g(t),a_h\{0:t-1\}), \\ \propto & P(a_h(t)|g(t)) \sum_{g(t-1)\in G} P(g(t)|g(t-1))P(g(t-1)|a_h\{0:t-1\}). \end{aligned} \tag{17}$$

**算法 4** 人类意图推理

1: **初始化:** 轨迹经验池$Tr$, 目标集$G$及其先验分布$b(g(0))$;
2: **输出:** 推断目标$g(t)$;
3: **while** $t <$ MAXSTEP or done == FALSE **do**
4: 从搜集到的历史轨迹池$Tr$中取状态动作对$(s(t),a(t))$;
5: **for** $g(i) \in G$ **do**
6: 基于式$b(g(t)) = P(a_h(t)|g(t)) \sum_{g(t-1)\in G} P(g(t)|g(t-1))b(g(t-1))$更新目标推理的后验分布;
7: **end for**
8: 由式$g^*(t) = \arg\max_{g(t)\in G} b(g(t))$更新$t$时刻任务目标.
9: **end while**

根据$s(t)$的定义, $b(g(t-1))$可以类似地表示为

$$b(g(t-1)) = P(g(t-1)|a_h\{0:t-1\}). \tag{18}$$

将公式(18) 代入公式(17), 可以得到:

$$b(g(t)) \propto P(a_h(t)|g(t)) \sum_{g(t-1)\in G} P(g(t)|g(t-1))b(g(t-1)). \tag{19}$$

从而能够写出人类意图推理算法4, 随着系统动态演化, 意图推理结果不断得到更新, 且该推断结果用于文章人机共享控制的仲裁优化设计(如算法1中的步骤5). 首先, 预先给出目标集的组成和初始时刻的先验分布. 算法的输入是经验样本池中的运动轨迹. 取运动轨迹中的系统状态和人的决策动作序列, 根据公式(19)更新目标后验分布. 后验分布更新后, 根据最大后验原理推断当前时刻的最新目标.

## 4 仿真验证

### 4.1 实验设置

基于OpenAI Gym中的LunarLander 模拟环境, 使用随机生成的着陆点坐标代替固定在$(0,0)$ 的坐标. 如果着陆器离开着陆区将失去奖励. 如果着陆器坠毁或平稳停止, 则episode结束, 同时获得额外的-100 分或+100 分. 每条腿接地得+10 分. 主引擎启动得-0.3分/每帧. 燃料假设是无限的, 机器代理和人

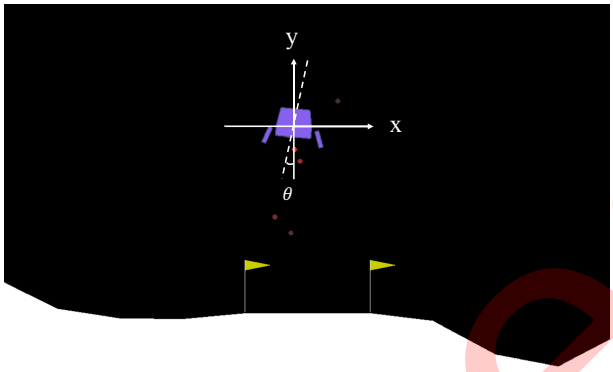


图 3　仿真环境.

**Figure 3**　Simulated environment.

**表 1**　动作值和引擎的对应关系.

**表 1**　Correspondence between action value and engine switch.

| 动作值 | 主引擎 | 左引擎 | 右引擎 |
|---|---|---|---|
| 0 | OFF | OFF | ON |
| 1 | OFF | OFF | OFF |
| 2 | OFF | ON | OFF |
| 3 | ON | OFF | ON |
| 4 | ON | OFF | OFF |
| 5 | ON | ON | OFF |

类伙伴可以学习飞行的控制过程, 然后尝试着陆. 状态向量$s(t)$ 包括：坐标$(x(t), y(t))$, 速度$(\dot{x}(t), \dot{y}(t))$, 角度$(\theta(t), \dot{\theta}(t))$, 着陆与否$(leg_l(t), leg_r(t))$, 着陆点坐标$h(t)$.

动作集为$\{0, 1, 2, 3, 4, 5\}$, 具体对应关系见表1, 其中0(左下)表示主发动机和左发动机关闭, 右发动机打开; 1 (下) 表示所有引擎都关闭; 2(右下)表示主引擎和右引擎关闭, 左引擎开启; 3(左上)表示主引擎和右引擎开启, 左引擎关闭; 4(上)表示主引擎开启, 左右引擎关闭; 5(右上)表示主引擎和左引擎开启, 右引擎关闭.

**注释1**　作者在仿真验证中选择LunarLander环境是考虑到着陆器轨迹优化(奖赏越大轨迹越优)在最优控制、智能决策均是经典的主题, 且LunarLander中的决策问题具有一定的代表性. LunarLander环境的状态$s(t)$(包括：坐标$(x(t), y(t))$, 速度$(\dot{x}(t), \dot{y}(t))$, 角度$(\theta(t), \dot{\theta}(t))$, 着陆与否$(leg_l(t), leg_r(t))$, 着陆点坐标$h(t)$), 对应的决策动作$a(t)$(包括主引擎、左引擎和右引擎的开关情况)以及优化目标$J(t) = \sum \gamma^t R(s(t), a(t))$,其中$R(s(t), a(t))$为环境状态为$s(t)$时执行决策动作$a(t)$所获得的奖励. 如果考虑辅助驾驶场景, 环境状态$s(t)$包括：车在马路上的位置坐标、速度、周围车辆的距离和相对速度等, 决策动作$a(t)$对应为方向盘的角度、刹车力度、油门大小等, 以及优化目标$J(t)$对应起始点之间路径规划的奖赏累积. 再比如脑控无人机场景, 环境状态为飞行器的空间坐标、速度信息等, 决策动作为AI智能和脑电信号构成的上下左右指令等. 上述场景均和本文验证的LunarLander环境具有一致性, 因此现有仿真验证能够说明本文所提方法的有效性.

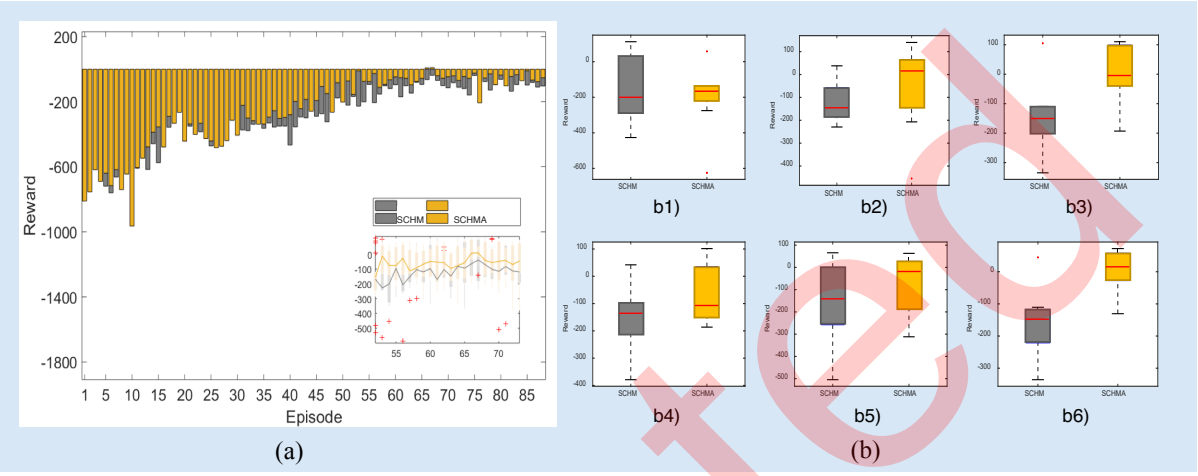


图 4　算法和算法SCHMA的(a) 平均奖赏对比和(b) 不同episode时的奖赏对比, 第一行自左向右是b1) episode 1; b2) episode 100; b3) episode 200, 第二行自左向右对应b4) episode 300; b5) episode 400; b6) episode 500.

**Figure 4**　Comparison between algorithm SCHM and algorithm SCHMA: (a) average reward comparison and (b) reward comparison with different episodes. The first line from left to right is b1) episode 1; b2) episode 100; b3) episode 200, and the second line from left to right corresponds to b4) episode 300; b5) episode 400; b6) episode 500.

机器的自主决策能力是基于DQN(Deep Q-Learning [42]) 来描述的, 即用DQN 衡量价值函数的大小, 以及获得当前时刻的最优决策动作. 值得注意的是, 本文使用SCHM(Shared Control of Human-Machine)来表示“人机共享控制算法”. SCHMA(Shared Control of Human-Machine with Autonomy)是基于自主性边界的人机共享控制算法, 它对应自主性边界判定算法2. 根据自主性边界不确定性估计算法3, SCHMAU(Shared Control of Human-Machine with Autonomy and Uncertainty)代表基于自主性边界不确定性的共享控制优化算法(算法1). 接下来, 本文分别从奖励、着陆成功率(撞击失败率)、仲裁参数、自主性边界、完整经历决策动作的生成等方面分析实验结果. 此外, 由于这个实验不是偶然的实验结果, 而是500次重复实验并取平均值, 因此足以说明结果的可靠性.

## 4.2　仿真结果

### 4.2.1　基于自主性边界的共享控制

首先要注意的是奖励趋势. 从图4(a)可以看出, SCHMA算法的奖赏高于SCHM算法. 另外需要注意的是, LunarLander 环境本身需要消耗燃料, 而且着陆点坐标不是固定的(如上文4.1 和3.3 小节所述, 落点坐标是机器根据人的决策动作推断出来的), 所以即使一条episode着陆成功, 它的累积奖励也会在0左右波动. 图4(b)中的趋势也表明本文提出的仲裁优化方法SCHMA优于SCHM.

接下来统计成功率和撞击率, 如图5 所示. 从图5(a)可以看出, 对于算法SCHM和算法SCHMA, 经过30次预训练后, 后者的成功率可以达到0.5 左右, 而前者的成功率只有不到0.3 . 图5(b)中的算法SCHMA可以将撞击率降低到0.3, 对应的SCHM撞击率在0.5左右. 另外, 注意撞击率不等于失败率, 因此$succ + crash \neq 1$.

仲裁参数$\alpha$决定了人类决策和机器决策的混合程度, 如公式(2)所示. 图6(a) 和6(b) 展示了自适应的$\alpha$. 本文所提的优化方法首先随机初始化仲裁参数$\alpha \in (0, 1)$, 随后在动态演化过程中根据实时环境自适应调整$\alpha$的值. 更重要的是, 文章应用自主性边界的概念(如式(9), (10))来优化$\alpha$ 的自适应值. 图6(c)

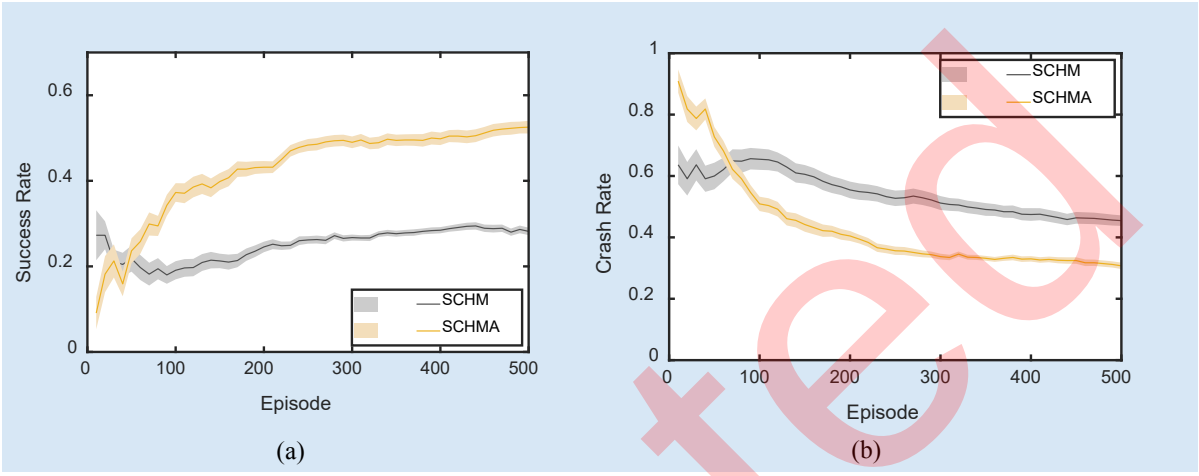


图 5　算法SCHM和算法SCHMA的(a)着陆成功率和(b)撞击率对比.

**Figure 5**　Comparison of (a) landing success rate and (b) crash rate of algorithm SCHM and algorithm SCHMA.

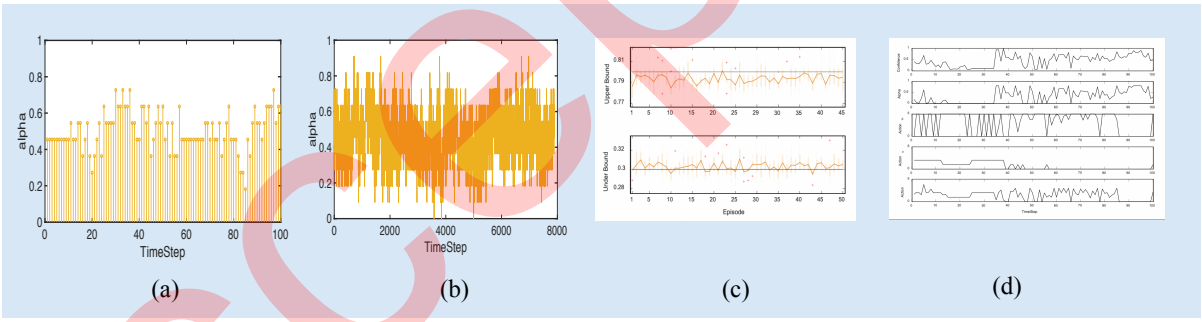


图 6　(a) 仲裁参数$\alpha$的前100步趋势; (b) 仲裁参数$\alpha$的整体趋势; (c) 自主性上界和下界; (d) 目标推理和人机决策动作的对应关系.

**Figure 6**　(a) Trend of the first 100 steps of the arbitration parameter *alpha*; (b) The overall trend of the arbitration parameter *alpha*; (c) The upper and lower bounds of autonomy; (d) Correspondence between target reasoning and human-machine decision-making actions.

展示了自主性边界的趋势.

最后, 为了便于理解, 图6(d)给出了目标推理置信度、$\alpha$值和每个动作值的对应关系. 从图中可以看出, 当目标推理置信度较低($\leqslant 0.3$)时, $\alpha$更小接近0时, 通过仲裁得到的混合动作更倾向于人类决策, 这与本文的初衷是一致的. 也就是说, 当机器的推测质量不高时, 系统更愿意相信人类的选择. 当目标推理置信度较高($\geqslant 0.5$)时, $\alpha$ 往往更接近1, 此时通过仲裁得到的混合动作更倾向于机器动作, 表明此时机器是可信的, 因此这也是合理的. 考虑到机器无法离线准确获得不断变化的目标, 系统预留了一定的可控空间($\alpha \leqslant 1$)供人工引导.

### 4.2.2　基于自主性边界不确定性的共享控制

针对基于自主性边界不确定性的共享控制优化设计, 首先值得注意的是奖励趋势, 如图7 所示. 与算法SCHM和SCHMA相比, 算法SCHMAU可以获得更高的奖励值. 并且从图7(b)可以看出, 随着episode的

增加, 基于自主边界不确定性的共享控制优化算法SCHMAU的作用不断凸显, 使得在奖励方面更有优势.

接下来分析月球车的着陆成功率和撞击率, 如图8所示. 图8(a)显示了基于自主边界不确定性的共享控制优化算法SCHMAU对SCHM和SCHMA着陆成功率的改进效果. 并且相比SCHM 0.25的成功率, SCHMA可以提高到0.5, 而SCHMAU可以提高到0.6. 另外, 在图8(b)中, 可以看出算法SCHMAU的影响率也有所下降.

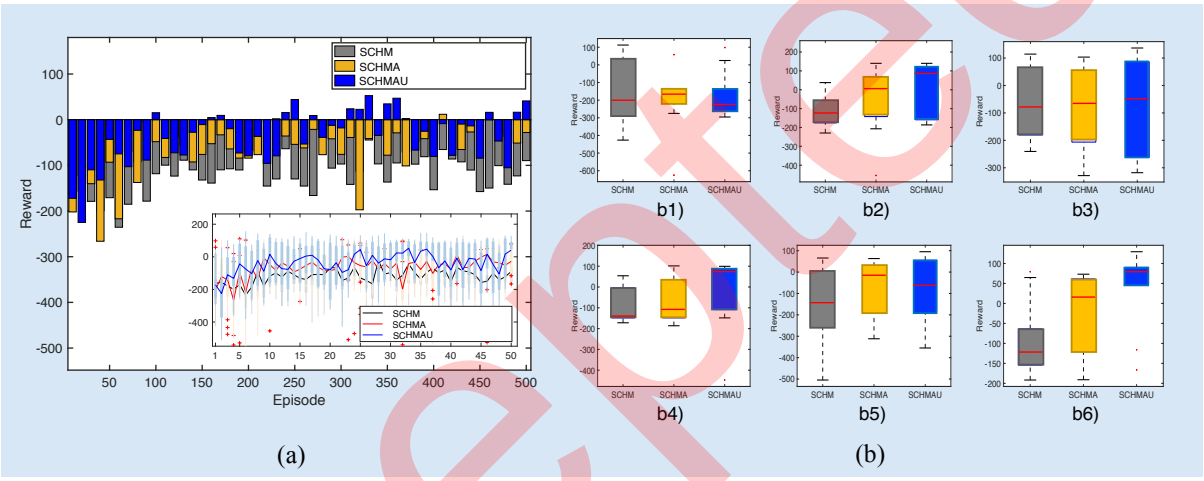


图 7　算法SCHM、算法SCHMA和算法SCHMAU的(a) 平均奖赏对比和(b) 不同episode时的奖赏对比, 第一行自左向右是b1) episode 1; b2) episode 100; b3) episode 200, 第二行自左向右对应b4) episode 300; b5) episode 400; b6) episode 500.

**Figure 7**　Comparison between algorithm SCHM, SCHMA, and algorithm SCHMAU. The first line from left to right is b1) episode 1; b2) episode 100; b3) episode 200, and the second line from left to right corresponds to b4) episode 300; b5) episode 400; b6) episode 500.

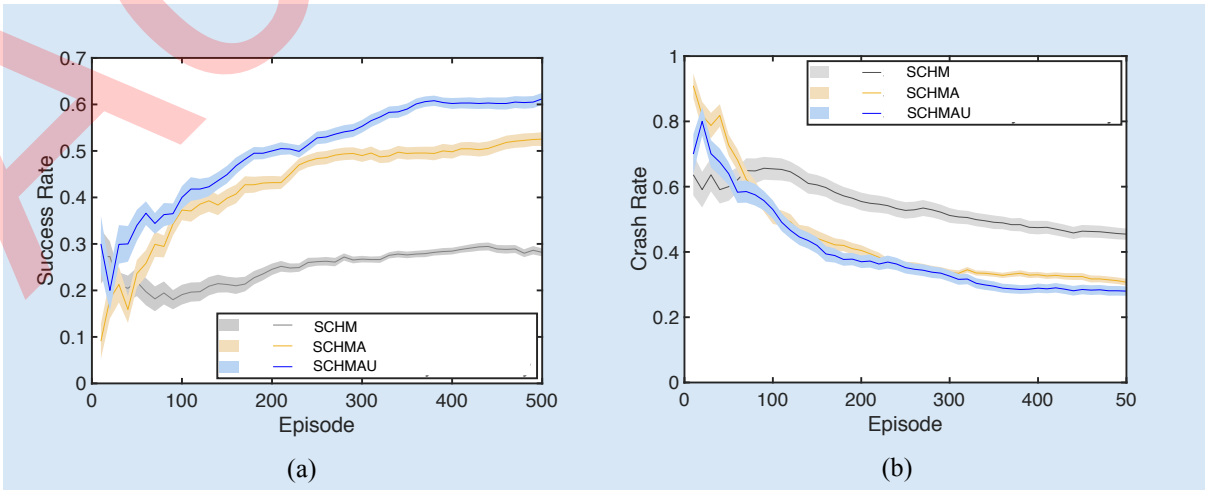


图 8　算法SCHM, SCHMA 和SCHMAU 着陆成功率和撞击率对比.

**Figure 8**　Comparison of (a) landing success rate and (b) crash rate of algorithm SCHM, SCHMA and algorithm SCHMAU.

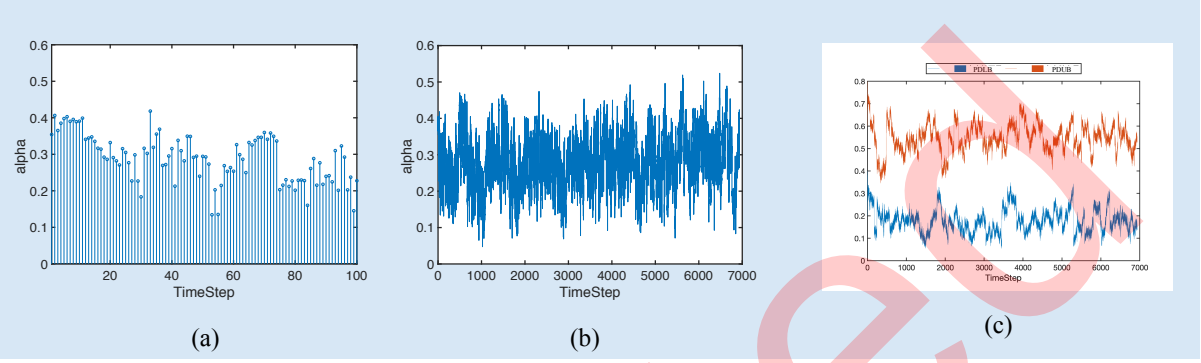


图 9　仲裁参数$\alpha$ 和带不确定性的自主性上界$\bar{b}(t)$和自主性下界$\underline{b}(t)$.

**Figure 9**　The arbitration parameter $\alpha$ and the upper bound of autonomy with uncertainty $\bar{b}(t)$and the lower bound of autonomy $\underline{b}(t)$.

最后, 图9表示一条episode中不同时间步所对应的仲裁参数$\alpha$的自适应调整曲线和相关的自主性边界趋势. 该图展示了人机共享控制优化方法实现决策质量改善的过程中所对应的自适应调节仲裁参数 α 的走势. 图9(a) 是前100 个时间步的仲裁参数的放大观察, 图9(b) 为仲裁参数$\alpha$的整体趋势. 在图9(c)中自主性上界概率分布(橙色, Probability Distribution of the Upper Bound, PDUB)走势中, 横坐标上每个时间步对应的纵坐标包括两个值－均值和方差, 即$\bar{b}(t) \sim N(\mu_{\bar{b}}, \sigma_{\bar{b}})$, 类似地, 自主性下界服从$\underline{b}(t) \sim N(\mu_{\underline{b}}, \sigma_{\underline{b}})$的概率分布(蓝色, Probability Distribution of the Lower Bound, PDLB). 两者都有助于人机共享控制算法SCHM的优化过程, 概率分布的不确定性可以提供更全面的信息, 这些从图7和图8中可以看出.

## 5　结论

针对人机协作过程策略评估不确定性引起的决策权限不明确情形, 本文提出了一种面向人机序贯决策问题的共享控制优化方法. 人机共享控制过程被建模成一个基于共同目标函数的优化问题, 其中考虑了人类意图对协作过程的影响, 并且融合了基于贝叶斯规则的意图推理. 基于与优化目标函数相关的自主性边界信息及其不确定性, 给出了所建模优化问题的共享控制优化方法. 通过仿真结果验证了所提优化方法具有改善人机序贯决策性能的能力. 在实际人机混合智能决策中可能面对多智能体协作对抗的情形, 未来工作将考虑更复杂的人机融合决策场景, 以进一步促进专家知识和人工智能在混合增强智能领域的完美结合.

## 参考文献

1　State Council. Circular of the state council on printing and distributing the development plan for the new generation of artificial intelligence. 2017, 22: 7-21 [2017-07-08]. http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm [国务院. 国务院关于印发新一代人工智能发展规划的通知. 2017, 22: 7-21 [2017-07-08]. http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm]

2　Ding J. Deciphering China’s AI dream. Future of Humanity Institute Technical Report, 2018

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

4　Xu K. Research on human-robot behavior-based collaborative control for search and rescue robot. Dissertation for

Ph.D. Degree. Jinan: Shandong University, 2009 [徐坤. 搜救机器人人机协作行为控制研究. 博士学位论文. 济南: 山东大学, 2009]

5 Su H, Yang C, Ferrigno G,et al. Improved human–robot collaborative control of redundant robot for teleoperated minimally invasive surgery. IEEE Robotics and Automation Letters, 2019, 4: 1447–1453

6 Yan Y. Study on force-position signal perception ang human-machine cooperative strategy for a vascular interventional surgery robot. Dissertation for Master Degree. Qinhuangdao: Yanshan University, 2021 [闫勇敢. 血管介入手术机器人力位感知与人机协同策略研究. 硕士学位论文. 秦皇岛: 燕山大学, 2021]

7 Li J, Yao L, Xu X, et al. Deep reinforcement learning for pedestrian collision avoidance and human-machine cooperative driving. Information Sciences, 2020, 532: 110–124

8 Fridman L, Ding L, Jenik B, et al. Arguing machines: Human supervision of black box AI systems that make life-critical decisions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019

9 Bibby K S, Margulies F, Rijnsdorp J E, et al. Man's role in control systems. In: IFAC Proceedings Volumes, 1975, 664-683

10 Zhao Y-B, Kang Y, Zhu J. Theory and method of autonomous human-machine hybrid intelligent system. Beijing: Science Press, 2021 [赵云波, 康宇, 朱进. 人机混合智能系统自主性理论和方法. 北京:科学出版社, 2021]

11 Zhao Y-B. Intelligent control: methods and applications Beijing: China Science and Technology Press, 2020. 423-435 [赵云波. 智能控制:方法与应用. 北京: 中国科学技术出版社, 2020]

12 Suchman L A. Plans and situated actions: the problem of human-machine communication. Cambridge: Cambridge university press, 1987

13 Suchman L, Suchman L A. Human-machine reconfigurations: Plans and situated actions. Cambridge: Cambridge university press, 2007

14 Jhaver S, Birman I, Gilbert E, et al. Human-machine collaboration for content regulation: the case of reddit automoderator. ACM Transactions on Computer-Human Interaction (TOCHI), 2019, 26(5): 1-35

15 Oh Y, Toussaint M, Mainprice J. Learning arbitration for shared autonomy by hindsight data aggregation. ArXiv preprint arXiv:1906.12280, 2019

16 Jain S, Argall B. Recursive bayesian human intent recognition in shared-control robotics. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018. 3905-3912

17 Jain S, Argall B. Probabilistic human intent recognition for shared autonomy in assistive robotics. ACM Transactions on Human-Robot Interaction (THRI), 2019, 9(1): 1-23

18 Schultz C, Gaurav S, Monfort M, et al. Goal-predictive robotic teleoperation from noisy sensors. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2017. 5377-5383

19 Bellman R. Dynamic programming. Science, 1966, 153(3731): 34-37

20 Puterman M L. Markov decision processes: discrete stochastic dynamic programming. John Wiley & Sons, 2014

21 Abbink D A, Carlson T, Mulder M, et al. A topology of shared control systems-finding common ground in diversity. IEEE Transactions on Human-Machine Systems, 2018, 48(5): 509-525

22 Dragan A D, Srinivasa S S. A policy-blending formalism for shared control. The International Journal of Robotics Research, 2013, 32(7): 790-805

23 Reddy S, Dragan A D, Levine S. Shared autonomy via deep reinforcement learning. ArXiv preprint arXiv:1802.01744, 2018

24 Marcano M, Díaz S, Pérez J, et al. A review of shared control for automated vehicles: Theory and applications. IEEE Transactions on Human-Machine Systems, 2020, 50(6): 475-491

25 International committee of the red cross. Autonomous weapon system: the impact of enhancing the autonomy of key functions of weapons. 2016 [2016-03-15]. https://max.book118.com/html/2019/0809/8113072073002041.shtm [红十字国际委员会. 自主武器系统：增强武器关键功能的自主性带来的影响. 2016 [2016-03-15]. https://max.book118.com/html/2019/0809/8113072073002041.shtm]

26 Png S C W O S W, Lee D H W S. Pomdps for robotic tasks with mixed observability. Robotics: Science and Systems, 2009

27 Jin Z, Pagilla P R, Maske H, et al. Task learning, intent prediction, and adaptive blended shared control with application to excavators. IEEE Transactions on Control Systems Technology, 2020, 29(1): 18-28

28 Li R, Li Y, Li S E, et al. Indirect shared control for cooperative driving between driver and automation in steer-by-wire

vehicles. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(12): 7826-7836
29 Fitzsimons K, Kalinowska A, Dewald J P, et al. Task-based hybrid shared control for training through forceful interaction. The International Journal of Robotics Research, 2020, 39(9): 1138-1154
30 Izadi V, Bhardwaj A, Ghasemi A H. Impedance modulation for negotiating control authority in a haptic shared control paradigm. In: 2020 American Control Conference (ACC). IEEE, 2020: 2478-2483
31 Fernandez F C, Caarls W. Deep reinforcement learning for haptic shared control in unknown tasks. ArXiv preprint arXiv:2101.06227, 2021
32 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. Applied Mechanics Reviews, 2018, 70(1)
33 Abou Allaban A, Dimitrov V, Padlr T. A blended human-robot shared control framework to handle drift and latency. In: 2019 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR). IEEE, 2019: 81-87
34 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. ArXiv preprint arXiv:1709.03969, 2017
35 González D, Pérez J, Milanés V, et al. Arbitration and sharing control strategies in the driving process. Towards a Common Software/Hardware Methodology for Future Advanced Driver Assistance Systems, 2017: 201
36 Marcano M, Díaz S, Pérez J, et al. Human-automation interaction through shared and traded control applications. In: Intelligent Human Systems Integration 2020: Proceedings of the 3rd International Conference on Intelligent Human Systems Integration (IHSI 2020): Integrating People and Intelligent Systems. Springer International Publishing, 2020: 653-659
37 Tjomsland J, Shafti A, Faisal A A. Human-robot collaboration via deep reinforcement learning of real-world interactions. ArXiv preprint arXiv:1912.01715, 2019
38 Box G E P, Tiao G C. Bayesian inference in statistical analysis. John Wiley & Sons, 2011
39 Dempster A P. A generalization of Bayesian inference. Journal of the Royal Statistical Society: Series B (Methodological), 1968, 30(2): 205-232
40 Gal, Y. Uncertainty in deep learning. Dissertation for Ph.D. Degree. Cambridge: University of Cambridge, 2016
41 Jain S, Argall B. Probabilistic human intent recognition for shared autonomy in assistive robotics. ACM Transactions on Human-Robot Interaction (THRI), 2019, 9(1): 1-23
42 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529-533

# Shared Control with Optimized Arbitration for Human-Machine Sequential Decision-Making

Qianqian Zhang[1], Yun-Bo Zhao[2,3,4*], Wenjun Lv[2,3,4] & Mou Chen[5]

1. *School of Artificial Intelligence, Anhui University, Hefei* 230039, *China*;
2. *Department of Automation, University of Science and Technology of China, Hefei* 230026, *China*;
3. *Institute of Artificial Intellgient, Hefei Comprehensive National Science Center, Hefei* 230088, *China*;
4. *Institute of Advanced Technology, University of Science and Technology of China, Hefei* 230031, *China*;
5. *College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing* 210016, *China*

* Corresponding author. E-mail: ybzhao@ustc.edu.cn

**Abstract** Considering that human-machine shared control occurs in many decision-making scenarios where human intelligence and machine intelligence participate together, and the decision-making range of humans and intelligent machines have not been clearly divided, real-time arbitration is required to achieve human-machine coexistence and sharing of decision-making authority. To this end, this paper proposes an arbitration optimization method, which is unique in that the concept of autonomous boundary optimizes the arbitration mechanism of human-machine decision-making actions in shared control. This paper provides an idea for the calculation, update and maintenance of the autonomous boundary, which can analyze the possible goals of the human-machine shared systems based on Bayesian rule-based intention inference, so as to determine the selection of arbitration parameters. In addition, this paper also analyzes the uncertainty of the autonomous boundary to promote the optimal effect of boundary information on decision quality in shared control. The final experimental results show that the proposed method performs well in cumulative reward, success rate, crash rate, which illustrate the effectiveness and value of our proposed quorum optimization method in shared control for solving sequential decision problems.

**Keywords** Shared control, arbitration optimization, autonomous boundary, human-machine sequential decision-making, reinforcement learning