中国科学: 信息科学 2022 年 第 52 卷 第 12 期: 2165–2177
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
非完全信息下人机合作对抗博弈专题 · 论文

非全时有效人类决策下的人机共享自主方法

游诗艺[1], 康宇[1,2,3*], 赵云波[1,3,4], 张倩倩[5]

1. 中国科学技术大学自动化系, 合肥 230026
2. 中国科学技术大学火灾科学国家重点实验室, 合肥 230026
3. 中国科学技术大学先进技术研究院, 合肥 230088
4. 合肥综合性国家科学中心人工智能研究院, 合肥 230026
5. 安徽大学人工智能学院, 合肥 230026
* 通信作者. E-mail: kangduyu@ustc.edu.cn

收稿日期: 2022–06–15; 修回日期: 2022–09–17; 接受日期: 2022–10–28; 网络出版日期: 2022–12–06

科技创新 2030 — “新一代人工智能” 重大专项 (批准号: 2018AAA0100800) 资助项目

摘要 在人机共享自主中, 人和智能机器以互补的能力共同完成实时控制任务, 以实现双方单独控制无法达到的性能. 现有的许多人机共享自主方法倾向于假设人的决策始终 “有效”, 即这些决策促进了任务的完成, 且有效地反映了人类的真实意图. 然而, 在现实中, 由于疲劳、分心等多种原因, 人的决策会在一定程度上 “无效”, 不满足这些方法的基本假设, 导致方法失效, 进而导致任务失败. 本文提出了一种新的基于深度强化学习的人机共享自主方法, 使系统能够在人类决策长期无效的情况下完成正确的目标. 具体来说, 我们使用深度强化学习训练从系统状态和人类决策到决策价值的端到端映射, 以显式判断人类决策是否无效. 如果无效, 机器将接管系统以获得更好的性能. 我们将该方法应用于实时控制任务中, 结果表明该方法能够及时、准确地判断人类决策的有效性, 分配相应的控制权限, 并最终提高了系统性能.

关键词 人机系统, 共享自主, 非全时有效决策, 深度强化学习, 仲裁

1 引言

在人机共享自主中, 人和智能机器以互补的能力共同完成实时控制任务, 以实现双方单独控制无法达到的性能 [1~3]. 以无人机降落为例, 对于人类而言, 难以同时实现高度、速度、姿态等的多维控制; 对于自动降落系统而言, 难以让其理解什么是好的、安全的着陆, 以及如何在不同的复杂环境下实现着陆. 人机共享自主作为应对不确定复杂环境的可行方法已被应用于很多领域中, 比如机器遥操作 [4,5]、半自主驾驶 [6,7]、康复外骨骼 [8,9] 等. 近年来, 由于人工智能的快速发展, 这一领域正受到越来越多的关注.

引用格式: 游诗艺, 康宇, 赵云波, 等. 非全时有效人类决策下的人机共享自主方法. 中国科学: 信息科学, 2022, 52: 2165–2177, doi: 10.1360/SSI-2022-0225
You S Y, Kang Y, Zhao Y B, et al. Human-machine shared autonomy approach for non-full-time effective human decisions (in Chinese). Sci Sin Inform, 2022, 52: 2165–2177, doi: 10.1360/SSI-2022-0225

许多控制任务的目标由人单方面决定, 智能机器被设计以帮助人完成该目标. 在这种情况下, 现有研究大多将人机共享自主策略分为两步完成: 对人类意图的推断 (机器通常无法直接得知人的意图), 以及对机器决策和人类决策的仲裁 [1,4,9,10]. 机器推理人的意图是完成仲裁的前提和基础, 其质量直接影响到仲裁的成败. 意图推理通常通过分析观察到的人类决策来完成, 其中一个关键假设便是这些观察到的人类决策是 "有效的", 即这些决策促进任务的完成并有效地反映人类的真实意图 [2,5,11]. 然而, 在现实中, 由于疲劳、分心等多种原因, 人类决策往往在一定程度上 "无效". 随着人的决策持续一段时间 "无效", 对人的意图的推理也将变得无效, 导致机器辅助的失效, 进而导致整个系统失效.

为了应对人的无效决策, 首先也是关键的挑战是确定一系列的人类决策是否无效. 其难点在于信息的不完全可知, 即必须由机器用有限的信息来完成. 一些研究通过认为人在生理状态异常时其物理决策无效来应对这一问题, 比如通过面部识别 [12]、监测神经信号 [13] 或心跳频率 [14] 等方法对人的生理状态进行测量和推理. 这类方法已被应用于一些领域, 比如在驾驶时通过摄像机进行面部识别, 根据驾驶员眼睑的开合判断其是否疲劳驾驶. 但它们不能处理人在正常生理条件下, 由于认知限制和环境限制 (如时间限制和不完全信息) 做出的无效行为. 例如, 在紧急情况下, 由于有限的处理时间和精神压力, 人的决策可能损伤系统性能, 甚至导致危险. 第二个挑战是如何区分无效的人类决策和人的意图改变. 推理组件基于一系列历史数据推断人的意图, 对人偶尔的无效行为具有鲁棒性, 但也导致意图变化无法被立即检测到. 对于以先前的意图为任务目标的系统来说, 意图改变后的人的决策可能是无效的. 对这两者的识别和区分直接关系到任务成功与否.

本文提出一种基于深度强化学习算法的人机共享自主方法, 使系统在人类决策长期无效的情况下仍能完成正确的目标. 具体来说, 我们利用长短时记忆网络推断人类意图, 然后利用深度强化学习算法训练从系统状态和人类决策到决策价值的端到端映射, 以判断人类决策是否无效. 强化学习算法计算的累积奖励值衡量了该决策行为可以给当前任务带来的利益的多少, 我们默认人类和机器都在朝着更高的奖励值努力, 因此在人的决策的奖励值下降一定程度后, 该决策被判定为无效. 当人类决策连续多次无效时, 系统将由机器单独控制, 完成从之前的有效决策中推断出的任务目标, 以防止人的无效决策影响任务进程.

本文的主要贡献可总结为以下四点.

- 一种人机共享自主方法, 使得在人类决策长期无效的情况下, 系统也能完成正确的目标;
- 一种无需额外信息就能判断人类决策有效性的方法;
- 一种区分无效的人类决策和人的意图改变的方法;
- 一种仲裁方法, 考虑了人类的无效决策和智能机器的不确定性.

本文结构安排如下. 第 2 节介绍相关工作; 第 3 节介绍人类决策非全时有效下的人机共享自主方法如何设计, 具体涉及构建和实现意图推理网络, 判断人的决策的有效性、考虑人的无效决策后的动作选择等; 第 4 节给出实验设计和结果分析; 第 5 节根据实验结果对本文提出的人机共享自主方法做进一步的讨论; 最后在第 6 节总结本文的工作.

2 相关工作

2.1 人机共享自主

由于环境的部分可观测性和系统参数的不精确性, 许多控制任务对机器来说存在困难, 而由于人的有限理性 [15] 和物理限制 (如缺乏多维控制能力) 等原因, 人类也很难单独完成. 一些人机系统通过

结合人的决策和机器决策来解决此类问题, 并将这种模式称为人机共享自主 [16,17]. 有研究将人机共享自主的环节描述为感知、推理、决策、执行 [18], 其中核心难点之一是机器通常不知道人类的意图, 因此许多方法将其分为先推断、再辅助两个步骤来完成 [1,2,10], 本文方法也采用这一模式.

在一些现有方法中, 机器选择行为时会尽可能地遵循人类的指令, 例如, 文献 [11] 构建的系统让机器评估其性能和人的期望之间的一致性, 进而更新自身行为以更好地和人保持一致. 在这类方法中, 系统无法预防或阻止人类的错误决策导致的任务失败或安全事故. 一些研究通过切换控制权来解决这一问题, 例如, 文献 [7] 预测了驾驶车辆的潜在行为, 并根据摄像头捕捉到的驾驶员的姿势衡量其安全性, 然后确定半自主驾驶系统应在何时进行干预; 文献 [19] 考虑了人类操作员的认知和生理状态的演变, 智能机器通过最小化人的工作量和优化系统性能之间的权衡来确定何时掌管控制权; 文献 [20] 直接从数据中学习系统动力学和人与系统的交互信息, 然后根据学习到的模型调整控制权限. 受到这些方法的启发, 我们关注人的决策可能对完成任务无用甚至有害的情况, 而智能机器可以更准确地感知环境, 并根据环境状态判断人类决策的有效性. 本文尝试利用控制权自适应来应对人类的无效决策, 即当人的决策有效时采用人机共享控制; 而当人的决策无效时, 由机器单独控制系统完成任务. 区别在于现有方法是在系统处于不利境地后进行纠正, 而这可能导致严重的后果, 本文方法是在不当决策执行之前进行预防或阻止, 具有更好的系统性能.

2.2 长短时记忆网络

递归神经网络能够处理序列变化的数据, 例如同一个词在不同的上下文中的不同含义, 而无需要求数据独立同分布. 长短时记忆网络是一种特殊的递归神经网络, 能够解决长序列数据在训练过程中的梯度消失和梯度爆炸问题, 因此长短时记忆网络在捕捉数据的长期时间依赖性方面既通用又有效, 并已被用于解决许多问题, 如手写识别 [21]、自动编写代码 [22]、翻译 [23]、音频分析 [24] 等. 在共享自主系统中, 人的行为是一系列按时间顺序排列的彼此高度相关的决策 [25], 系统根据人的决策呈现一定的状态, 人再根据当前的系统状态做出新的决策. 人类决策的序列性使得我们采用长短时记忆网络进行意图推理.

2.3 深度强化学习

强化学习方法具有强大的学习能力, 由其赋能的智能体在与环境交互的过程中即可学习策略, 而无需其他数据. 深度学习方法具有强大的计算能力, 可以在连续和复杂的情况下实现实时仲裁和控制. 深度强化学习将两者结合起来, 使用强化学习方法定义问题和优化目标, 使用深度学习方法优化和求解策略函数或价值函数. 深度强化学习已被广泛应用于许多任务中 [26~28], 例如文献 [1] 提出了一个基于深度强化学习的框架, 使人类和智能机器共同控制机械臂抓取物体. 文献 [29] 通过扩展深度强化学习建模人类反馈的可信度和一致性, 从而使用离散的人类反馈来增强虚拟 3D 环境中智能体的性能. 文献 [30] 使用深度强化学习和机械臂的特定结构实现在人和机器的动力学模型未知的情况下完成目标. 强化学习算法的应用和其任务场景密切相关, 本文研究受现有研究的启发, 使用深度强化学习算法构建人机共享自主方法, 利用深度强化学习算法会为每个动作计算累积奖赏值的特性, 从而对决策的有效性进行量化和计算.

3 方法

我们关注人的决策可能无效, 即其可能对完成任务无用甚至有害的情况, 而智能机器可以更准确

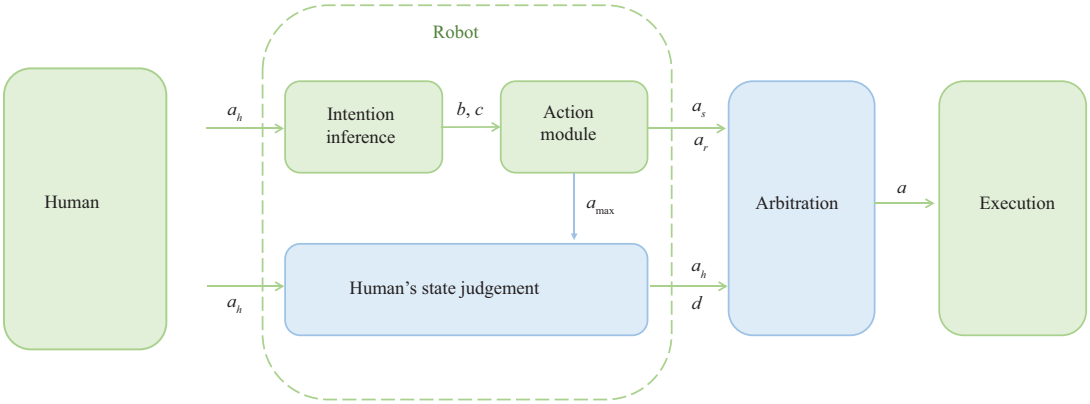

图 1　(网络版彩图) 非全时有效人类决策下的人机共享自主方法框图

Figure 1　(Color online) Block diagram of human-machine shared autonomy approach for non-full-time effective human decisions

地感知环境, 并根据环境状态判断人类决策的有效性. 本文方法的目的是在人类决策长期无效的情况下, 确保系统完成正确的目标. 一种可行的解决方案是控制权自适应: 当人的决策有效时采取人机共享控制, 系统在人输入的动作和机器选择的动作之间进行仲裁; 而当人的决策无效时, 机器独自控制系统完成从先前的有效决策中推断出的目标. 本节将详细描述如何判断人的决策的有效性, 如何根据意图推断的结果确定任务目标, 如何在共享控制中进行仲裁, 以及如何在共享控制和个体控制之间进行离散切换.

如图 1 所示, 本文方法的关键部件包括 4 个部分: 智能机器推断人的意图并计算意图推断的置信度 (第 3.1 小节); 动作选择模块计算对人机决策进行仲裁的自适应权重, 确定采取共享控制时的决策 (第 3.2 小节); 人类决策有效性判断模块用于判断人的决策是否无效, 是否需要由人和机器的共享控制切换到机器的单独控制 (第 3.3 小节); 仲裁模块根据上述所有信息决定被控系统最终执行的动作 (第 3.4 小节). 图 1 中蓝色区域, 即人的决策的有效性判断模块和仲裁模块是本文的主要贡献.

3.1　意图推理

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

我们初步考虑智能机器动作的有效性, 尝试排除智能机器对环境状态的不确定性过高的情况. 机器根据推理出的目标做出决策, 推理越准确, 做出的决策就越有效. 同样, 推理越不确定, 做出错误选择的可能性就越大. 因此, 我们计算了意图推断的置信度, 当置信度没有达到设定的阈值时, 被控系统会直接执行人的决策动作, 因为机器决策的错误概率太大而不应考虑. 置信度达到阈值时, 仲裁模块在人的决策和机器的决策之间进行仲裁. 意图推理的置信度是目标集的概率分布中的最大概率减去概率分布中的最小概率 [2]:

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

这里有两种极端情况:

• 目标集中有一个目标的概率为 1, 其他目标的概率均为 0. 这种情况即为网络准确推断出目标集中的某个目标, 机器完全确定人的意图是什么, 因此意图推理的置信度为 1.

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

3.2 动作选择

动作选择模块计算在执行共享控制时要采取的动作. 我们使用深度 Q 网络 (DQN) [31, 32] 来实现方法. DQN 的输入是环境状态和人的决策, 输出是当前环境状态下所有动作的预期未来累积奖励. 累积奖励 $Q(s,a)$ 是在状态 s 执行动作 a 后, 在未来有限的步骤中所能获得的折扣奖励的期望和, 用于评估动作的价值. 在一些共享自主方法中, 机器选择累积奖励值最大的动作 $a_{\max}$ 作为最优决策. 但是, 我们认为, 当人类的决策有效时, 智能机器应该尽可能少地修改人类的输入, 以增加其对辅助的接受度 [33]. 如果系统总是执行偏离人类决策的动作, 那么人类可能因为命令没有得到准确执行而不再信任系统. 因此, 我们采用足够好的、与人类决策最相似的动作 a_s 作为共享控制下执行的动作.

我们根据意图推理的置信度和动作的累积奖励值确定最优动作的选择范围, 如式 (2) 所示. 意图推理的信心越大, 机器对于任务目标越确定, 因此选择的范围越小; 推理信心越小, 机器越有可能出错, 因此在较大的范围内选择和人类决策相近的动作.

$$a_s = \underset{a\in A: Q'(s,a)\geqslant c\times Q'_{\max}(s,a_{\max})}{\arg\max} f(a,a_h), \tag{2}$$

其中 A 是动作空间, 包含所有可能执行的动作. $Q'(s,a) = Q(s,a) - \min_{a'\in A} Q(s,a')$ 为动作的累积奖励减去所有动作中累积奖励的最小值, 以防止负 Q 值造成的误差. $a_{\max}$ 为 DQN 网络计算出的当前环境状态下的累积奖励最高的动作. $f(a,a_h)$ 计算动作 a 和人的决策 a_h 之间的相似度. 举例说明, 当置信度 $c = 0.8$ 时, 从满足 $Q' \geqslant 0.8Q'_{\max}$ 的动作中选择和 a_h 最相似的动作.

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

3.3 判断人的决策的有效性

累积奖励值代表了该动作在当前任务状态下的价值. 我们假设人类和机器都在朝着更大的价值努力, 所以当人的决策的奖励值足够低时, 决策被判定为无效. 我们使用动作的累积奖励之间的差值作为动作之间的距离, 即

$$d(a,a_{\max}) = \frac{Q'(s,a) - Q'(s,a_{\max})}{Q'(s,a_{\max})}. \tag{3}$$

当人的决策动作和 DQN 计算出的价值最高的动作之间的距离 $d(a_h,a_{\max})$ 连续多次足够大时, 我们认为人当前无法实施有效控制, 故由机器单独控制系统. 当机器单独控制时, 机器以人有效控制最后一步计算出的概率分布作为任务目标, 不管人当前决策是什么, 直接将 $a_{\max}$ 传递给被控系统, 如下所示:

$$a_r = \underset{a\in A}{\arg\max}\, Q(s,a). \tag{4}$$

同时, 人在不断地控制, 网络在不断地计算动作距离. 当距离 $d(a_h,a_{\max})$ 连续几次足够小时, 我们认为人回归到理性, 能够做出有效的响应, 因此系统回归到共享控制.

值得特别注意的是, 机器可能会将人的目标改变的行为错误地判断为无效的决策行为. 长短时记忆网络根据一系列轨迹和动作计算目标集的概率分布. 因此, 目标变化可以被神经网络逐步识别出来,

进而由人与机器的共享控制来完成. 但当系统接近推理出的目标且推理置信度较高时, 人改变目标的动作可能产生较大的动作距离, 进而导致人类失去控制权, 而由机器引导系统完成之前的目标, 导致任务失败. 因此, 当机器单独控制系统时, 我们重新收集轨迹和人的决策来重新推断人的意图. 如果机器连续多次重复推理出同一个目标, 则认为是目标发生了改变, 由机器和人共同控制系统完成新目标.

3.4 人机共享自主的仲裁方法

综上所述, 被控系统执行的动作有 3 种可能:

$$a=\begin{cases}a_h, & c \text{ 低于阈值},\\ a_r, & d \text{ 高于阈值},\\ a_s, & \text{其他}.\end{cases} \tag{5}$$

当意图推理的置信度足够低时, 由人单独控制系统, 系统将执行 a_h. 当人的决策动作和最优动作之间的距离连续高于阈值时, 机器单独控制系统, 系统将执行 a_r. 在其他情况下, 系统由人和机器共同控制. 总体算法如算法 1 所示. 每幕的方法流程如图 2 所示, 图中蓝色部分为本文的主要创新, 即根据人类决策、环境状态、DQN 网络的计算结果判断人类决策的有效性, 并根据该结果仲裁出最终动作.

算法 1 基于 DQN 的非全时有效人类决策下的人机共享自主算法

初始化容量大小为 N 的经验池 $\mathcal{D}$;
初始化权重为随机权重 θ 的评估网络 Q;
初始化权重为 $\theta^- = \theta$ 的目标网络 $\hat{Q}$;
while steps $\leqslant M$ **do**
  while step $\leqslant 1000$ **do**
    获得环境状态 s_t 和人的输入 a_h;
    推理人类意图, 并根据式 (2) 获得动作 a_s;
    根据式 (5) 仲裁得到动作 $a_t = a$;
    执行动作 a_t, 得到新状态 s_{t+1} 和奖励值 r_t;
    将四元组 (s_t, a_t, r_t, s_{t+1}) 存储进经验池 $\mathcal{D}$;
    if s_{t+1} is terminal **then**
      while $k \leqslant K$ **do**
        从经验池 D 批量采样 (s_j, a_j, r_j, s_{j+1});
        $a'_{j+1} = \arg\max_{a'} Q(s_{j+1}, a'; \theta)$;
        $y_j = r_j + \gamma \hat{Q}(s_{j+1}, a'_{j+1}; \theta^-)$;
        $\theta \leftarrow \theta - \eta \nabla_\theta \sum_j (y_j - Q(s_j, a_j; \theta))^2$;
      end while
    end if
    每 C 步复制评估网络的权重到目标网络 $\hat{Q} = Q$;
  end while
end while

4 仿真实验

我们使用 OpenAI Gym 的登月着陆器场景进行实验, 如图 3 所示. 地面上共有三对旗子, 其位置坐标在每次任务开始时随机生成. 每对旗子中间的区域是平坦的, 其他区域的高度为随机生成. 着陆器左右两边各有一个推进器, 中间有一个主引擎. 人和机器共同控制这 3 个发动机, 使着陆器无碰撞

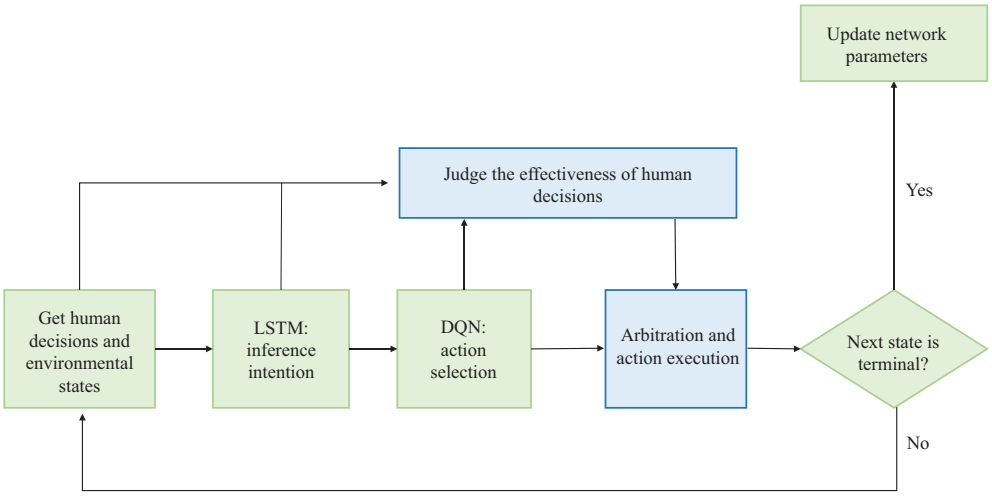

图 2 (网络版彩图) 非全时有效人类决策下的人机共享自主方法流程图

Figure 2 (Color online) Flowchart of human-machine shared autonomy approach for non-full-time effective human decisions

图 3 (网络版彩图) OpenAI Gym 登月着陆器场景示意图

Figure 3 (Color online) Lunar Lander in OpenAI Gym environment

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

登月着陆器场景模拟了真实的着陆环境, 即着陆器越靠近地面, 降落速度越快, 也越难以控制. 因此机器需要大量数据学习对 3 个发动机的控制, 而这些大量的训练过程是人类参与者无法承受的负担, 因此智能机器将在没有人参与的情况下进行单独训练, 以完成对通用规则的学习.

4.1 实验设置

动作空间为 6 维的离散动作, 包含 3 个发动机的开启和关闭. 状态空间为 11 维向量, 包含着陆器当前的位置、速度、角度、角速度、支架是否接触地面, 以及三对旗子的坐标. 奖赏函数为惩罚速度、角度、和推理目标之间的距离, 即速度越快、倾斜度越大、距离推理出的目标越远则惩罚越大, 旨在训练着陆器稳定缓慢地向目标点移动. 着陆器的支架接触地面给予较小的奖励值, 平稳降落在目标点后

给予较大的奖励值, 冲撞到地面或飞出边界则给予较大的惩罚值.

我们采用有两个隐藏层, 每层 32 个神经元的长短时记忆网络进行意图推理, 采用含有两个隐藏层, 每个隐藏层有 64 个节点的多层感知机实现 DQN 算法. 动作之间的相似度函数 $f(a, a_h)$ 用来判断两个动作是否控制同一个发动机或是否控制着陆器向同一方向移动. 比如控制左发动机的开启和关闭的相似度为 -1, 即 $f((\text{left}, \text{on}), (\text{left}, \text{off})) = -1$; 控制左推进器关闭和控制右推进器开启的相似度为 1, 即 $f((\text{left}, \text{on}), (\text{right}, \text{off})) = 1$.

我们用随机操作模拟人的无效决策, 比如人闭上眼睛并随机按下按键. 若人输入的行为连续 10 次中有 7 次行为满足 $d(a_h, a_{\max}) \geqslant 0.7$, 即累积奖励值小于最大奖励值的 30%, 人的决策被判定为无效, 由机器接管任务并重新进行意图推理. 相应地, 如果人连续 10 次输入的动作中有 7 次的累计奖励值超过最大值的 70%, 人被认为恢复正常, 着陆器由人与机器共同控制. 重新推理意图时, 倘若连续 7 次重新推理的目标中有 5 次目标相同, 则认为实则为人的目标切换到该目标, 人和机器将共同控制着陆器接近新目标. 如果推理置信度小于 0.3, 着陆器将直接执行人的输入, 因为机器的动作太不确定, 不应予以考虑.

我们招募了 10 位玩家, 5 位男生 5 位女生, 平均年龄 25 岁. 每个玩家被提前告知游戏规则并单独操作 20 次以熟悉操作和环境, 再和训练后的机器共同控制 20 次以相互适应和优化. 每个玩家要完成在降落过程中改变和不改变目标的两个实验. 并且为了方便收集和分析数据, 我们为玩家指定了目标旗子颜色. 第一个实验中玩家不能改变目标, 始终控制着陆器降落到黄色旗子中间. 第二个实验中玩家的目标由蓝色旗子转向黄色旗子, 变换目标的时机由玩家自己决定.

4.2 降落过程中不改变目标

本节实验的目的是验证所提方法的有效性, 并分析我们提出的方法与其他不考虑人类决策无效性的人机混合决策方法在任务性能上的差异. 因此, 我们使用 DQN 实现一个最常用的人机共享自主方法, 即始终执行 DQN 计算出的奖励价值最高的动作, 作为对比实验. 每位玩家需要分别使用两种行为策略, 即所有人类决策有效和部分人类决策无效, 在人的单独决策 (human individual control, HIC)、最高价值人机混合决策 (highest value shared autonomy, HVSA) 和我们提出的人类决策非全时有效下的人机混合决策 (shared autonomy under ineffective human inputs, SAIHI) 三种方法的辅助下进行操作, 共 6 个任务. 每个任务有 20000 步, 每幕最多 1000 步, 所以每个任务至少有 20 幕. 每幕的具体步数取决于玩家的能力, 通常是 300~700 步. 为了确保每幕都存在持续的无效人类决策, 我们让玩家在第 100 步和第 200 步时各随机操作一段时间, 随机操作的结束时间由玩家决定. 呈现给玩家的任务顺序是平衡的, 以避免玩家越来越熟练导致结果的偏差.

图 4 显示了 10 位玩家在 6 个任务中的成功率和平均路径长度. 图 4(a) 为人的决策全部有效时的任务成功率, 显示出将玩家与智能机器结合在一起带来的定量和定性优势. 在处理着陆器的突然下降时, 玩家很难同时控制三个维度的引擎来保持着陆器的稳定, 使得着陆器经常在 150 步内撞向地面或飞出边界, 很难在不发生碰撞的情况下准确降落在要求的位置. 这主要是因为人类缺乏同时在多个维度准确操纵物体运动的能力. 而玩家和机器的混合决策大大增加了成功着陆的可能性, 因为机器可以精确控制着陆器动态. 当人的输入全部有效时, 我们方法的成功率略高于普通的人机共享自主方法. ANOVA (方差分析, 对实验中使用的所有变量之间的差异进行统计检验) 的结果是 $F = 7.1130, p = 0.0157$, 意味着我们的方法比另一种方法更容易成功. 我们认为这种改进是由对仲裁函数的改进带来的. 图 4(b) 为人的决策部分无效时的任务成功率. 当存在持续的无效人类决策时, 我们的方法显著优于一般方法, ANOVA 的结果为 $F = 30.48, p = 3.04902\text{E} - 5$. 我们的方法可以判断人输入的行为是否无效, 机器

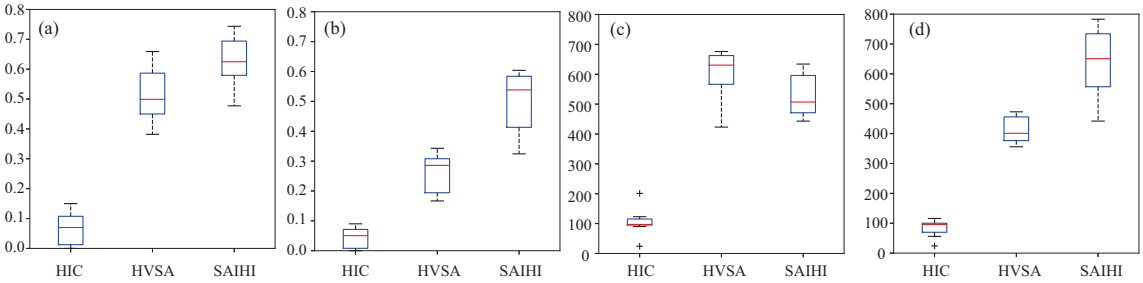

图 4 (网络版彩图) (a) 人的决策全部有效下的任务成功率; (b) 人的决策部分无效下的任务成功率; (c) 人的决策全部有效下的每幕平均步数; (d) 人的决策部分无效下的每幕平均步数

Figure 4 (Color online) (a) Task success rate when human decisions are all effective; (b) task success rate when human decisions are partially ineffective; (c) average steps of each episode when human decisions are all effective; (d) average steps of each episode when human decisions are partially ineffective

会及时介入和接管系统进行单独控制, 避免玩家的无效行为影响任务进程, 从而有效提高了任务成功率. 从图 4(c) 和 (d) 可以看出, 在人的决策均有效的情况下, 我们的方法可以在更短的时间内完成任务, 这得益于我们提出的仲裁方法更加智能和有效. 但当人的决策部分无效时, 我们的方法可以持续更长的时间, 使得系统不会在失去来自外部的有效控制命令后立即崩溃. 机器为玩家提供一些缓冲时间, 试图回到性能最优的共享控制模式.

图 5 显示了人的决策部分无效情况下着陆器在某一幕成功着陆的过程. 图中绿色区域表示系统在该段时间内由人与机器共同控制, 黄色区域表示系统在该段时间由机器单独控制, 蓝色区域表示该系统在该段时间由玩家单独控制, 图中红色圆圈表示动作距离大于等于 0.7. 图 5(a) 显示了着陆过程中三种控制模式之间的切换. 图 5(b) 是玩家的输入和最高价值动作之间的距离. 输入的动作具有最高的奖励, 或玩家不输入将导致行动距离为 0. 这些数据表明, 基于强化学习判断人的决策有效性的方法是有效的: 在步骤 100 和 200 附近监测到大的动作距离, 导致系统由共享控制切换到机器的单独控制. 当动作距离足够小时, 控制模式恢复为共享控制. 图 5(d) 是在每一步推断的玩家的目标, 图 5(c) 是相应的意图推理置信度. 推理的目标和置信度是着陆器在最后阶段由玩家单独控制的原因: 这三对旗帜的坐标是随机生成的, 当目标旗帜接近另一对旗帜且着陆器接近地面时, 由于两者在计算上的差别较小, 机器可能无法确定两个着陆点中的哪一个是玩家的目标. 这时, 意图推理的置信度会降低, 由玩家单独控制着陆器向真正的目标前进, 且因为接近地面, 人单独控制也不会导致着陆器坠毁.

4.3 降落过程中改变目标

我们设置这个实验的目的是验证方法能否识别玩家的目标变化, 并帮助玩家完成新的目标. 机器根据一系列的轨迹和动作推断目标. 因此, 改变目标可以逐渐被网络识别, 但当着陆器接近推断目标且推断置信度高时, 突然改变目标很可能被视为无效输入而不被采用. 因此, 本实验设置了两个任务: 将着陆器可以移动的垂直距离分成两个相等的部分, 玩家需要在这两个空间中分别改变目标并尝试成功着陆. Task1 是玩家在上半空间改变目标, Task2 是在下半空间改变目标. 因为每个玩家在每幕都有不同的步数, 所以改变目标的时间是由玩家决定的, 我们只规定目标从蓝色旗子变成黄色旗子. 此任务的难点便在于机器需要意识到这种变化, 如果机器始终将大的动作距离归咎于无效的人类输入, 并将目标固定在蓝色旗子上, 任务将会失败. 只有机器重新推断目标并控制着陆器接近新目标, 着陆方能成功.

图 6(a) 显示了 10 位玩家完成两项任务的成功率. 从图 6(a) 可以看出, Task1 的成功率远远高

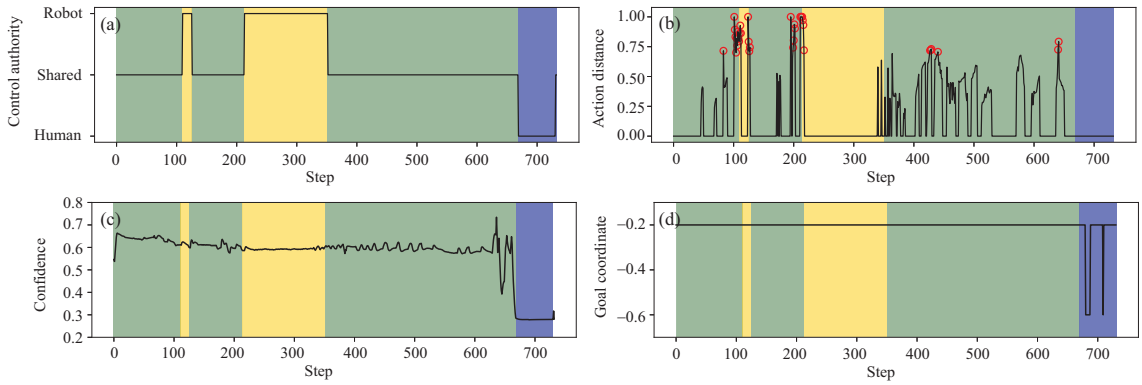

图 5 (网络版彩图) 人的决策部分无效下的成功着陆过程. (a) 控制模式的切换; (b) 动作距离; (c) 意图推理置信度; (d) 推理出的任务目标坐标

Figure 5 (Color online) Process of a successful landing when human decisions are partially ineffective. (a) Switches of control authority; (b) action distance between human input and optimal action; (c) confidence of intention inference; (d) inferred results of the goal coordinate

于 Task2. 方差分析的结果验证了这两个任务之间的差异确实是显著的, $F = 11.65, p = 0.0031$. 事实上, Task1 的成功率与玩家尝试用部分无效的人的输入成功着陆的任务的成功率大致相同, 大多在 0.4~0.6 之间 (参见图 6(a) 和图 4(b)). 这两项任务的本质区别在于, 导致动作距离变大的原因不同. 后者的较大行动距离是由玩家的随机操作造成的, 随机输入之间没有相关性和逻辑, 导致机器不能从中得到信息. 但在 Task1 中, 玩家基于环境状态的有目的控制行为和意图推理模块的延迟识别产生了较大的动作距离. 机器从这些输入中推理出一个新的目标, 并基于这个新目标获得了较小的动作距离. 这是区分无效的人类行为和人类意图变化的关键因素, 即输入是否包含有效的信息. Task2 任务成功率低的主要原因是剩余的时间和高度不足以让玩家和机器在新的着陆点顺利着陆.

Task1 和 Task2 的某次成功着陆轨迹分别如图 6(b) 的红色和绿色线条所示. 图中黄色的星星表示黄色旗帜对的中点, 蓝色的星星表示蓝色旗帜对的中点; 纵坐标为降落空间的垂直坐标, 即归一化后的空间高度, 其中 0 表示地面, 1 表示着陆器初始位置的高度; 横坐标为降落空间的水平坐标, 其中 0 为水平中点. 图片清晰地显示了机器识别出玩家的目标变化, 以及由此带来的着陆器的轨迹变化. 玩家在上半空间变化目标使着陆器有富裕的时间和空间朝新目标前进, 轨迹在高度 0.6 左右出现拐点, 即机器发现目标变化并辅助人完成新目标, 着陆器在高度 0.2 左右到达最终任务目标的上方, 后续为缓慢降落. 而在下半空间变化目标的轨迹在高度 0.2 左右出现拐点, 并在剩余五分之一的空间里完成惊险着陆. 可以看到机器为了完成任务不得不控制着陆器先上升再朝目标降落, 这也使得 Task2 有更长的平均路径长度和更少的任务成功率. 另一个导致 Task2 失败的原因是时间耗尽, 如果每幕没有时间限制, Task2 的成功率可能会更高.

图 6(c) 为玩家在上半空间改变目标时的某次成功着陆的过程. 图中 CA 为控制模式的变更, 其中绿线表示系统由人和机器共同控制, 黄线表示系统由机器单独控制. AD 为人的输入与最高价值动作之间的动作距离, 即 $d(a_h, a_{\max})$, 其中黑色圆圈表示大于等于 0.7 的动作距离. GC 为目标坐标, 即目标对旗子的中点. 子图显示了步骤 70~110 之间三个特征的详细信息. 可以看到, 在较大动作距离出现后, 机器独自控制着陆器. 在机器单独控制的过程中, 推理出的目标发生变化. 目标坐标稳定到新的坐标后, 动作距离减小, 任务回归到由玩家和机器共同控制的模式.

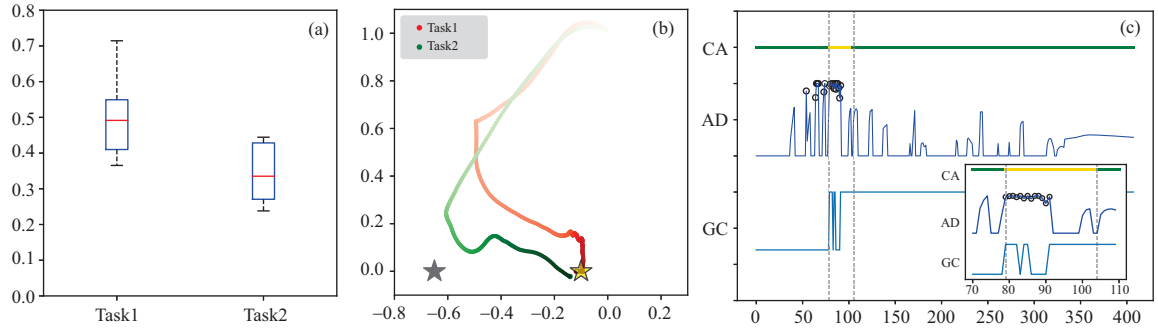

图 6 (网络版彩图) (a) 十位玩家完成两项任务的成功率; (b) 完成两项任务的典型着陆轨迹; (c) 玩家在上半空间改变目标的某次成功着陆轨迹

Figure 6 (Color online) (a) Success rates of ten players completing two tasks; (b) typical landing trajectory for completing two tasks; (c) the process of a successful landing when the player changes the target in the upper half space

5 进一步讨论

第 4 节的实验结果表明: (1) 基于深度强化学习判断人类决策有效性是切实可行的; (2) 我们提出的方法能够及时有效地判断人的决策是否无效, 灵活切换控制模式, 分配控制权限以提高系统性能; (3) 当人的决策无效时, 用机器的单独控制代替共享控制可以显著提高系统性能; (4) 当机器单独控制系统时, 机器应该重新推断人类的意图, 并检测输入是否包含有效信息.

在实验过程中, 我们发现判断人类决策无效的指标应该谨慎选择. 我们尝试使用 5/7 (7 次连续输入中有 5 个满足 $d(a_h, a_{\max}) \geqslant 0.7$) 作为标准, 并发现其过于敏感, 无法稳定切换控制权限并保证良好性能. 我们也尝试过 10/15, 发现它反应迟缓, 导致性能更差, 这可能是由于每幕的持续时间较短造成的. 我们没有对该指标应如何设置以得到最优性能做进一步的研究, 只是提出这个设想, 并对其有效性做了全面的初步验证.

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

6 结论

本文基于深度强化学习算法提出了一种人类决策非全时有效下的人机共享自主方法. 在已知目标集、未知系统动态模型和未知人类行为策略的条件下, 即使人的输入持续一段时间无效, 本方法仍能继续完成正确的目标. 我们使用 DQN 显式判断人的决策是否无效或任务目标是否发生了变化, 并分配相应的控制权, 以避免无效的人为输入妨碍任务进程. 我们将该方法应用于实时控制任务中, 结果表明该方法能够及时、有效地判断和处理人的无效输入, 提高系统性能.

本文提出的方法还可以进一步改进. 我们仅从人的输入推断目标并判断其有效性, 在算法中加入其他隐藏信息, 比如人的目光注视 [34], 能否取得更好的结果还需要进一步研究. 此外, 我们的方法假设已知可能的目标集, 尝试消除这种依赖, 使方法更加灵活和通用是下一步的研究方向.

参考文献

1 Javdani S, Srinivasa S S, Bagnell J A. Shared autonomy via hindsight optimization. In: Proceedings of Robotics Science and Systems, 2015

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

3 Abbink D A, Mulder M, Boer E R. Haptic shared control: smoothly shifting control authority? Cogn Tech Work, 2012, 14: 19–28

4 Gopinath D, Jain S, Argall B D. Human-in-the-loop optimization of shared autonomy in assistive robotics. IEEE Robot Autom Lett, 2016, 2: 247–254

5 Nikolaidis S, Zhu Y X, Hsu D, et al. Human-robot mutual adaptation in shared autonomy. In: Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction, 2017. 294–302

6 Anderson S J, Peters S C, Pilutti T E, et al. An optimal-control-based framework for trajectory planning, threat assessment, and semi-autonomous control of passenger vehicles in hazard avoidance scenarios. Int J Veh Autonom Syst, 2010, 8: 190–216

7 Vasudevan R, Shia V, Gao Y, et al. Safe semi-autonomous control with enhanced driver modeling. In: Proceedings of American Control Conference, 2012. 2896–2903

8 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. Appl Mech Rev, 2018, 70: 010804

9 He W, Li Z, Chen C L P. A survey of human-centered intelligent robots: issues and challenges. IEEE/CAA J Autom Sin, 2017, 4: 602–609

10 Hauser K. Recognition, prediction, and planning for assisted teleoperation of freeform tasks. Auton Robot, 2013, 35: 241–254

11 Xu A, Dudek G. Trust-driven interactive visual navigation for autonomous robots. In: Proceedings of IEEE International Conference on Robotics & Automation, 2012. 3922–3929

12 Dreissig M, Baccour M H, Schack T, et al. Driver drowsiness classification based on eye blink and head movement features using the k-NN algorithm. In: Proceedings of IEEE Symposium Series on Computational Intelligence, 2020. 889–896

13 Jing D, Liu D, Zhang S, et al. Fatigue driving detection method based on EEG analysis in low-voltage and hypoxia plateau environment. Int J Transpation Sci Tech, 2020, 9: 366–376

14 Sharma M K, Bundele M M. Design & analysis of kmeans algorithm for cognitive fatigue detection in vehicular driver using oximetry pulse signal. In: Proceedings of International Conference on Computer, Communication and Control, 2015. 1–6

15 Simon H A. Bounded rationality and organizational learning. Organ Sci, 1991, 2: 125–134

16 Aigner P, McCarragher B. Human integration into robot control utilising potential fields. In: Proceedings of International Conference on Robotics and Automation, 1997. 291–296

17 Goertz R C. Manipulators used for handling radioactive materials. Hum Factors, 1963, 7: 425–443

18 Huang K Q, Xing J L, Zhang J G, et al. Intelligent technologies of human-computer gaming. Sci Sin Inform, 2020, 50: 540–550 [黄凯奇, 兴军亮, 张俊格, 等. 人机对抗智能技术. 中国科学: 信息科学, 2020, 50: 540–550]

19 Fu J, Topcu U. Synthesis of shared autonomy policies with temporal logic specifications. IEEE Trans Automat Sci Eng, 2015, 13: 7–17

20 Broad A, Murphey T, Argall B. Learning models for shared control of human-machine systems with unknown dynamics. 2018. ArXiv:1808.08268

21 Pham V, Bluche T, Kermorvant C, et al. Dropout improves recurrent neural networks for handwriting recognition. In: Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition, Heraklion, 2014. 285–290

22 Li G J, Liu H, Li G, et al. LSTM-based argument recommendation for non-API methods. Sci China Inf Sci, 2020, 63: 190101

23 Luong M T, Sutskever I, Le Q V, et al. Addressing the rare word problem in neural machine translation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015. 11–19

24 Marchi E, Ferroni G, Eyben F, et al. Multi-resolution linear prediction based features for audio onset detection with

bidirectional LSTM neural networks. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2014. 2164–2168
25 Lam C P, Yang A Y, Driggs-Campbell K, et al. Improving human-in-the-loop decision making in multi-mode driver assistance systems using hidden mode stochastic hybrid systems. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, 2015. 5776–5783
26 Tjomsland J, Shafti A, Faisal A A. Human-robot collaboration via deep reinforcement learning of real-world interactions. 2019. ArXiv:1912.01715
27 Zhang L L, Li D W, Xi Y G, et al. Reinforcement learning with actor-critic for knowledge graph reasoning. Sci China Inf Sci, 2020, 63: 169101
28 Wang H, Yu Y, Jiang Y. Review of the progress of communication-based multi-agent reinforcement learning. Sci Sin Inform, 2022, 52: 742–764 [王涵, 俞扬, 姜远. 基于通信的多智能体强化学习进展综述. 中国科学: 信息科学, 2022, 52: 742–764]
29 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. 2017. ArXiv:1709.03969
30 Li Y, Tee K P, Yan R, et al. Reinforcement learning for human-robot shared control. Assembly Autom, 2019, 40: 105–117
31 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518: 529–533
32 Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning. 2013. ArXiv:1312.5602
33 Broad A, Murphey T, Argall B. Highly parallelized data-driven MPC for minimal intervention shared control. 2019. ArXiv:1906.02318
34 Admoni H, Srinivasa S. Predicting user intent through eye gaze for shared autonomy. In: Proceedings of the AAAI Fall Symposia, 2016

Human-machine shared autonomy approach for non-full-time effective human decisions

Shiyi YOU[1], Yu KANG[1,2,3*], Yun-Bo ZHAO[1,3,4] & Qianqian ZHANG[5]

1. *Department of Automation, University of Science and Technology of China, Hefei* 230026*, China*;
2. *State Key Laboratory of Fire Science, University of Science and Technology of China, Hefei* 230026*, China*;
3. *Institute of Advanced Technology, University of Science and Technology of China, Hefei* 230088*, China*;
4. *Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei* 230026*, China*;
5. *School of Artificial Intelligence, Anhui University, Hefei* 230026*, China*
* Corresponding author. E-mail: kangduyu@ustc.edu.cn

Abstract In shared autonomy, humans and intelligent robots jointly complete real-time control tasks with their complementary capabilities for improved performance unattainable by either side independently. Many existing methods tend to assume that human decisions are "effective", i.e., these decisions promote task completion and effectively reflect the true human intention. However, in reality, human decisions can often be "ineffective" to a certain extent due to many reasons, such as fatigue or inattentiveness, which leads to task failure. In this work, we propose a novel deep reinforcement learning-based shared autonomy strategy for human-machine systems, so that the system can complete the correct goal even when human decisions are ineffective for a long period. Specifically, we use deep reinforcement learning to train an end-to-end mapping from system states and human decisions to the value of decisions to explicitly judge whether the human decisions are ineffective. If they are ineffective, the robot takes over the system for better performance. We apply our method to real-time control tasks, and the results show that it can timely and accurately judge the effectiveness of human decisions, allocate control authority, and ultimately improve system performance.

Keywords human-machine system, shared autonomy, non-full-time effective decision, deep reinforcement learning, arbitration