面向安全、高效多智能体强化学习的基于模型的动态屏蔽

Wenli Xiao 香港中文大学（深圳）深圳市人工智能与机器人研究院 中国
wenlixiao@link.cuhk.edu.cn

Yiwei Lyu 卡内基梅隆大学 美国
yiweilyu@andrew.cmu.edu

John Dolan 卡内基梅隆大学 美国
jdolan@andrew.cmu.edu

arXiv:2304.06281v1 [cs.LG] 13 Apr 2023

摘要

多智能体强化学习（MARL）能够发现最大化奖励的策略，但在学习和部署阶段均无法提供安全保证。尽管基于线性时序逻辑（LTL）的屏蔽是一种在单智能体强化学习（RL）中确保安全性的有效形式化方法，但扩展至多智能体场景时会导致行为过于保守。此外，在复杂的多智能体环境中合成安全屏蔽器也面临计算方面的挑战。本文提出基于模型的动态屏蔽（MBDS），以支持多智能体强化学习算法的设计。我们的算法合成分布式屏蔽器，这些屏蔽器作为反应式系统与每个多智能体强化学习智能体并行运行，用于监控并纠正不安全行为。屏蔽器可根据智能体的状态动态拆分、合并和重新计算。这种设计无需协调开销，即可高效合成屏蔽器，以监控复杂环境中的智能体。我们还提出了一种无需预先了解动力学模型即可合成屏蔽器的算法。该算法在探索早期通过与环境交互来获得近似世界模型，使我们的MBDS以较高概率提供形式化安全保证。仿真结果表明，在安全保证和学习性能方面，我们的框架优于现有基线方法。

关键词

机器人学；多智能体强化学习；安全性

ACM参考文献格式：

Wenli Xiao、Yiwei Lyu 和 JohnDolan。2023年。《基于模型的动态安全屏蔽器：实现安全高效的多智能体强化学习》。载于第*22*届国际自主智能体与多智能体系统会议（*AAMAS 2023*）论文集，英国伦敦，*2023*年*5*月*29*日至*6*月*2*日，IFAAMAS，共10页。

1 引言

多智能体强化学习 (多智能体强化学习) [10, 59]是一种有前景的方法，可为交通管理 [3, 40], 运动控制 [42, 57], 和自动驾驶 [7, 47, 60]等多智能体决策任务获取学习型控制策略。然而，在安全关键型自动系统（例如自动驾驶汽车）中应用多智能体强化学习方法可能会因

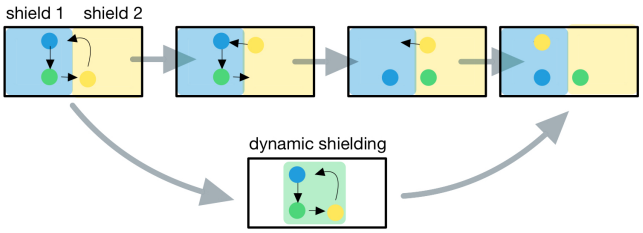


图 1：不同颜色的圆圈表示多个智能体，黑色箭头表示期望动作。传统的分散式屏蔽（上图）在屏蔽器边界附近需要额外等待协调，而本文提出的动态屏蔽（下图）则消除了这一开销。

缺乏形式化的安全保证。此外，采用行为惩罚的传统多智能体强化学习方法（即对不安全动作给予负奖励）无法确保实践中的安全性 [18, 45]。因此，开发可证明值得信赖的安全多智能体强化学习系统是一项重大挑战 [11, 18, 33, 45, 59]。

近年来，关于安全性概念的研究颇多 [1, 18, 19, 24, 41] [43] 。

例如，线性时序逻辑（LTL）是一种用于形式化验证的规范语言，可确保自动化系统始终处于安全状态 [46]。近期的一项工作 [1] 通过合成一个安全屏蔽器来监控强化学习智能体，将 LTL 用作单智能体强化学习（RL）中的安全规范语言。安全屏蔽器是一种与强化学习智能体协同运行的轻量级系统，用于监控强化学习智能体选取的动作，并根据给定的安全规范拒绝任何不安全动作。在强化学习过程（即训练和部署阶段）的整个生命周期内，安全屏蔽器都能提供可证明的安全保证。因子化屏蔽 [18] 以分散式方式将屏蔽学习方法扩展到多智能体场景。与使用单个屏蔽器监控所有智能体的状态和动作的集中式屏蔽相比，因子化屏蔽会合成多个屏蔽器，每个屏蔽器监控智能体状态空间的一个子集。这些方法在离散环境中表现良好。然而，对于更复杂的连续环境，集中式屏蔽和因子化屏蔽都难以扩展。

另一方面，在设计屏蔽框架时存在一个两难问题：集中式方法的可扩展性有限 [18],，而完全去中心化的方法则会带来协调开销。由于缺乏信息共享，在去中心化方法中，智能体彼此靠近时可能会陷入等待协调的状态

第*22*届国际自主智能体与多智能体系统会议论文集（*AAMAS 2023*），*A. Ricci*、*W. Yeoh*、*N. Agmon*、*B. An*（编），*2023*年*5*月*29*日至*6*月*2*日，英国伦敦。 © 2023年国际自主智能体与多智能体系统基金会（www.ifaamas.org）。保留所有权利。

。例如，图1展示了因子化屏蔽导致额外协调开销的场景。本文提出了一种新颖、安全且高效的多智能体强化学习框架，采用混合去中心化方式，并动态合成安全屏蔽器，以缓解这些局限。

具体而言，我们的主要贡献有三点：首先，我们提出了一种新颖的安全屏蔽框架——动态屏蔽，使机器人能够协作以确保安全。初始时存在多个屏蔽器，分别并行监控不同的智能体。当保守行为风险较高时（例如，智能体一起移动），屏蔽器可以选择与其他屏蔽器合并。合并后的屏蔽器能够利用多个智能体的状态信息，减少不必要的协调开销。当智能体彼此远离时，合并后的屏蔽器可以拆分为多个屏蔽器。我们还在第 5 节提出了一种有效的屏蔽综合方法，称为k 步前瞻屏蔽器。我们的方法省去了传统屏蔽综合方法中不必要的计算 [1, 18]，并将计算复杂度转移至在线算法，使其能够实时综合屏蔽器。我们还引入了世界模型学习流程，以学习简化的环境动力学模型。这使我们的框架能够从零开始学习，只需极少的外部知识。

此外，我们通过大量实验展示了所提出屏蔽方法的有效性和性能。我们在六种不同的网格世界地图 [38]上研究导航问题，并在多智能体粒子环境 [32] (MPE)中研究两种不同的任务。与其他基线方法相比，我们的方法在保证安全性的同时，取得了更高的奖励并减少了所需步数。此外，我们还表明，随着智能体数量增加，动态屏蔽仍能以很高的概率确保安全。

2 相关工作

2.1 安全多智能体强化学习

安全强化学习方法可以分为两类 [19]：1)第一类是基于优化准则的方法，这类方法会修改强化学习目标函数 [15, 50, 53]。例如，SNO-MDP [54]通过使用受约束的马尔可夫决策过程来解决安全强化学习问题。2) 第二类基于修改探索过程以避免不期望的动作 [6, 26, 30], ，并将额外的领域特定知识（例如示范）纳入训练过程。我们的动态屏蔽算法属于第二类。屏蔽最初被引入强化学习 [1], ，并在 [18]中适配到多智能体设置。在本工作中，我们通过应对多智能体设置中的协调开销和可扩展性问题，并减少对外部知识的依赖，提出了一种用于多智能体强化学习的新型屏蔽框架。

2.2 通过控制屏障函数实现安全控制

屏障证书 [34, 44] 和基于控制障碍函数（CBF） [56]的控制方法 [12, 14, 35, 36, 49, 52] 常用于为安全关键问题提供安全性保障，例如碰撞规避 [37, 44, 56, 58]。在多智能体碰撞规避领域，已有研究探索了使用多智能体 CBF 框架 [8, 37, 55]。近期研究 [13, 45] 提出了基于 CBF 的分散式控制器综合方法，该方法可扩展至任意数量的智能体。我们认可这些研究的贡献，但它们与我们的研究重点并不相关。

在本工作中，我们旨在通过利用表现力更强的线性时序逻辑（LTL） [43]，来处理面向多智能体强化学习（MARL）的更一般的安全规范。LTL 可以方便地捕获复杂的时变约束 [48]。尽管我们在第6 节中针对碰撞-规避任务开展了实验，但在本工作中，我们专注于将 LTL 应用于多智能体强化学习，旨在未来将我们的方法扩展到更加复杂的安全约束。

2.3 将线性时序逻辑用作安全规范

线性时序逻辑是安全关键系统中广泛使用的一种规范语言，可用于以高层次表达复杂需求。例如，线性时序逻辑已用于表达机器人规划与控制中的复杂任务规范。已有多项研究提出了奖励塑形技术，将线性时序逻辑中表达的逻辑约束转换为强化学习的奖励函数。然而，实证结果表明，奖励塑形无法确保多智能体强化学习中的安全性。我们的屏蔽器综合技术基于求解双人安全博弈，最初在相关研究中开发，用于强制执行线性时序逻辑规范。原始技术以离线方式将安全屏蔽器综合到本地缓存中。在第 5 节中，我们提出一种新颖的在线方法，可实时综合安全屏蔽器。

3 预备知识

我们首先介绍多智能体强化学习、安全屏蔽以及基于线性时序逻辑规范的安全博弈，我们的算法以此为基础。

3.1 多智能体强化学习

我们关注由一个元组定义的 n人马尔可夫博弈

$$\left(\mathcal{N}, \mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, \{r^i\}_{i\in\mathcal{N}}, \mathcal{P}, \gamma\right)$$

其中， $= \{1...n\}$ 是智能体集合，S 表示所有智能体共同观测到的状态空间，A^i 是智能体 i的动作空间，r^i 是智能体 i的奖励函数，P：$\mathrm{S} \times \mathrm{A} \to \Delta(\mathrm{S})$ 表示转移概率，γ是折扣因子。我们假设初始状态 s_1 服从固定分布 $\rho \in \Delta(\mathrm{S})$。在每个时间步 t，智能体同时观测状态 s_t，在环境中采取动作$a_{t,\ i} \in A^i$，并获得奖励 $r_{t,\ i} \in R^i$。随后，环境状态转移至 s_{t+1}。每个智能体 i的目标是学习控制策略 π_i，以最大化期望累积奖励 $E\left[\sum_{t=0}^{\infty} \gamma^t R^i(s_t,\ a_t,\ s_{t+1})\right]$。根据单个智能体的表现对其他智能体选择的依赖程度，多智能体强化学习算法可分为三种类型：合作型、竞争型和混合设置型。我们在第 6 节的实验中采用混合设置型多智能体强化学习算法。CQ-learning [16] 是一种多智能体强化学习算法，可使智能体在大多数时候独立行动，并在必要时考虑其他智能体的状态和动作。MADDPG [32]是一种深度多智能体强化学习算法，采用集中式训练和分散式执行；每个智能体都会根据对其他智能体行为的观测，训练用于模拟其他智能体策略的模型。

3.2 LTL 作为安全规范

我们使用线性时序逻辑 [43] (LTL)来表达安全规范。线性时序逻辑是命题逻辑的扩展，长期以来一直被用作对程序和系统进行形式化验证的工具。线性时序逻辑的语法由以下文法给出 [5]:

$$\varphi := p|\neg p|\varphi_1 \vee \varphi_2|\bigcirc \varphi|\varphi_1 \mathcal{U} \varphi_2$$

其中， p 是一个原子命题。时序算子包括下一步算子$\bigcirc\varphi$，表示 φ 在下一个后继状态中为真；以及直到算子 $\varphi_1 \mathrm{U} \varphi_2$，表示 φ_1 为真，直到 φ_2 为真的状态。基于这些算子，我们可以定义 $True \equiv \phi \vee \neg\phi$、 $False \equiv \neg True$、蕴含 $\varphi \Rightarrow \psi := \neg\varphi \vee \psi$、最终 $\diamond\varphi := \text{True U}\varphi$，以及始终 $\square\varphi := \neg \diamond \neg\varphi$。我们使用 LTL 公式来表达安全规范。例如， $\square\neg collision$ 表示碰撞永远不应发生。我们考虑将 LTL 安全规范转换为可由确定性有限自动机（DFA） [29]接受的安全语言。此外，我们以如下方式将 [1]中的安全强化学习定义扩展到多智能体强化学习（MARL）：

定义 1. 安全多智能体强化学习是指在学习阶段和执行阶段满足时序逻辑安全规范 ϕ^s 的同时，为多个智能体学习最优策略的过程。

3.3 基于安全屏蔽器的形式化安全保证

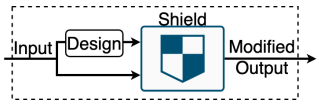


图 2：通过屏蔽来强制执行安全规范。

我们的方法基于一种名为 Shield [18, 27], 的方法，该方法可在运行时确保安全属性。安全屏蔽器（如图 2 所示）会监控智能体的控制输入，并立即纠正任何不安全控制输入。安全屏蔽器应具备两个属性：1）最小干预。也就是说，只有当动作违反安全规则时，屏蔽器才会纠正该动作。2）正确性。屏蔽器应能识别所有不安全动作，并将其细化为安全动作。我们的方法采用安全屏蔽器框架来确保安全性，并在第 5 节提供安全性的理论证明。

我们使用有限状态反应系统**表示安全屏蔽器**。根据 [18], 中的表述，有限状态反应系统是一个元组 $S = (Q, q_0, \Sigma_I, \Sigma_O, \delta, \lambda)$，其中 Σ_I 和 Σ_O是输入/输出字母表， Q 是状态集， $q_0 \in Q$ 表示初始状态，δ: $Q \times \Sigma_I \to Q$是转移函数， λ: $Q \times \Sigma_I \to \Sigma_O$是输出函数。给定控制输入的符号抽象（即输入轨迹） $\sigma_I = x_0 x_1 \ldots \in \Sigma_I^\infty$，系统 S 会生成状态轨迹（即输出轨迹） $\lambda\ q_0\ x_0\ \lambda\ q_1\ x_1\ \ldots \in \Sigma_O^\infty$ — $q_{i+1} = \delta\ q_i\ x_i$ $\sigma_O = \sigma_I = i \geq 0$ S ( ) (,) (,) , 其中 (,) 对所有情况均成立。

我们通过**求解一个双人安全博弈** [27],来综合安全屏蔽器。该博弈由多智能体强化学习智能体与环境进行，其获胜条件由线性时序逻辑安全规范定义。多智能体强化学习智能体必须始终遵守所有安全规范，才能赢得对局。双层博弈是一个元组 G $= (G, g_0, \Sigma_I, \Sigma_O, \delta,$ $win)$，包含有限的博弈状态集 G、初始状态 $g_0 \in G$、完备的转移函数 δ: $G \times \Sigma_I \times \Sigma_O \to G$, 以及作为获胜条件的 win。在每个状态 $g \in G$中，环境首先选择一个输入动作 $\sigma_I \in \Sigma_I$，然后多智能体强化学习智能体选择一个联合动作（用抽象符号表示） $\sigma_O \in \Sigma_O$。随后，博弈转移到下一个状态 $g\prime = \delta(g, \sigma_I, \sigma_O)$，依此类推。由此产生的博弈状态轨迹 $\bar{g} = g_0, g_1, \ldots$称为一场对局。当且仅当 $win(\bar{g})$ 为 $true$ 时，对局获胜。我们将在第 5 节介绍通过求解双人安全博弈来综合安全屏蔽器的详细流程。

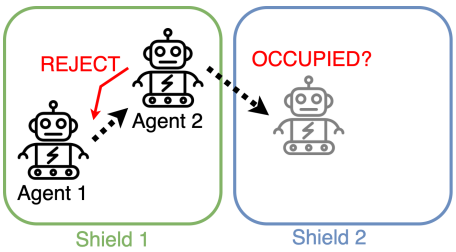


图3：绿色和蓝色方块表示安全屏蔽器，虚线箭头表示智能体的期望动作。屏蔽器1和屏蔽器2之间没有通信。因此，屏蔽器1保守地判断智能体2无法成功进入屏蔽器2，因而拒绝智能体1的动作。

4 通过动态屏蔽解决安全且高效的多智能体强化学习问题

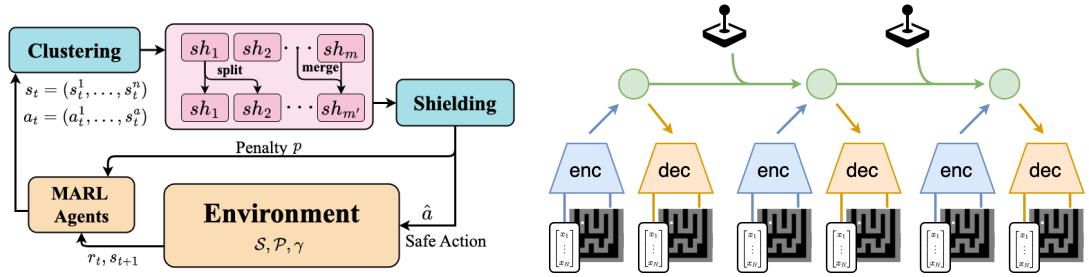
在本节中，我们首先描述传统屏蔽方法如何导致次优学习。然后，我们介绍一种实现安全、最优的多智能体强化学习的方法。

4.1 保守行为与协调开销

对于多智能体系统，随着智能体数量增加，集中式方法总会失效。例如，在双智能体场景中，集中式多智能体强化学习屏蔽在经验上会失效 [18]。完全分布式屏蔽将整个状态空间划分为互不重叠的子空间，并合成一个安全屏蔽器来监控某个子空间。例如，因子化屏蔽 [18] 会基于所有智能体观测到的联合状态空间的因子化结果，计算多个屏蔽器。然而，由于屏蔽器之间存在信息隔离，当智能体跨越屏蔽器的边界移动时，这种方法会导致保守行为（即智能体停滞不前）。具体而言，如图3所示，即使智能体的动作实际上有效，安全屏蔽器也会拒绝这些动作。因此，多智能体强化学习系统会产生更高的协调开销，导致智能体交互时需要额外步骤，并使多智能体强化学习策略无法达到最优。在第6节中，我们通过实验表明，由保守行为导致的协调开销会造成次优策略。

4.2 Dynamic shielding

为减轻保守行为导致的协调开销，我们提出动态屏蔽，这是一种基于传统多智能体强化学习过程的去中心化屏蔽框架。动态屏蔽



(a) 在动态屏蔽。

(b) 从经验中学习动力学。

图 4：动态屏蔽框架

算法 1： 时间步 t的动态屏蔽

初始化： 一个由安全屏蔽器组成的列表 $S = \{s_1, s_2, \ldots s_m\}$，多智能体强化学习智能体的联合动作 $a_t = (a_t^1, a_t^2, \ldots a_t^n)$ 和联合状态 $s_t = (s_t^1, s_t^2, \ldots, s_t^n)$，对不安全的行为施加恒定惩罚动作 p，一个环境动力学模型 $p(s_t|s_{t-1}, a_{t-1})$

输出： 安全联合动作 $\bar{a}_t$，惩罚 p_t，安全屏蔽器 new_安全屏蔽器

// 聚类：将智能体划分为组
// 例如：根据智能体的位置对其进行聚类
$new_shield = cluster_agents(s_t, a_t)$
对于 所有组，$i \in \{1, \ldots, m\prime\}$ **执行**
  针对所有安全屏蔽器 $j \in \{\ldots, m\}$ **执行**
    如果 $new_安全屏蔽器[I].组 == s_j.group$ 并且 $s_j.duration != 0$ **那么**
      new_安全屏蔽器[i].recompute = False
      new_安全屏蔽器[i].安全屏蔽器 $= s_j.shield$
    结束如果
  结束循环
结束循环
// 重新构建安全屏蔽器
// 1. 当智能体试图逃离安全屏蔽器时
// 2. 当安全屏蔽器失效时
对所有组 in $i \in \{1, \ldots, m\}$ **执行**
  如果 $new_shield[i].recompute == True$ **则**
    new_安全屏蔽器[i].reCompute(p)
  结束 if
结束 for
// 屏蔽：将不安全动作替换为安全动作
$\bar{a}_t$ =由 new_安全屏蔽器输出的安全动作
对于所有智能体，范围为 $i \in \{1, \ldots, n\}$ **执行**
  如果 $\bar{a}_t^i \neq a_t^i$ **那么**
    p_t^i =p
  结束如果
结束循环。
return $bara_t$, p_t, $new_$安全屏蔽器

具有两个重要特征：1）动态屏蔽会根据智能体的实时状态动态构建屏蔽器；2）动态屏蔽可以执行两项重要操作，即合并和拆分。合并操作利用多个屏蔽器的信息构建一个更大的屏蔽器，并暂时移除屏蔽器之间的边界。因此，合并后的屏蔽器拥有足够的信息来判断联合动作是否安全，从而最终减轻保守行为。另一方面，屏蔽综合的计算复杂度会随着屏蔽器规模的扩大而增加。当智能体分布稀疏时，拆分操作有助于降低计算成本。图 4a 展示了动态屏蔽构建示意图。初始时，我们为每个智能体分别构建独立的屏蔽器，以监控智能体在接下来的 k 步中的可达状态。如果智能体试图移动到屏蔽器范围之外的状态，屏蔽器将重新计算，以监控智能体未来可能到达的状态。当智能体聚集在一起且可能发生碰撞时，屏蔽器将合并，利用多个智能体的状态信息共同监控动作。当智能体分布较为稀疏时，合并后的屏蔽器将拆分，以节省计算资源。

我们在算法 1 中总结了动态屏蔽。该算法包含三个阶段：1）聚类，2）安全屏蔽器重构，以及 3）屏蔽。在聚类阶段（第 1–12 行），算法根据智能体当前的状态将智能体聚类为不同的组。例如，在机器人导航任务中，如果某些智能体彼此距离较近，算法会将它们归入同一组，否则将它们分到不同的组中。同一组中的智能体应合并安全屏蔽器，以避免保守行为。随后，在安全屏蔽器重构阶段（第 13–20 行），安全屏蔽器将根据聚类结果与其他安全屏蔽器合并，或拆分为多个更小的安全屏蔽器。此外，某些已过期的安全屏蔽器可能会根据智能体的状态变化重新计算。合并操作可以通过使用智能体聚合后的状态信息重新计算安全屏蔽器来实现。在屏蔽阶段（第 21–27 行），每个安全屏蔽器将并发执行屏蔽，拒绝智能体的不安全动作，并将其替换为安全动作。最后，多智能体强化学习智能体将因执行不安全动作而受到额外惩罚。

我们的方法面临这样的挑战：在边缘场景（某些时间步）下会退化为集中式安全屏蔽器。例如，当所有智能体聚集在一起时，所有去中心化屏蔽器将合并为一个集中式安全屏蔽器。我们在第5节提出了一种在线方法的屏障综合方法，该方法能够高效地综合屏蔽器。

算法 2：学习动力学模型

初始化：初始化数据集 D。
初始化：随机初始化神经网络参数 θ。
当未收敛 **时执行**
  // 动力学学习
  针对更新步骤 $c = 1...C$ **执行**
    绘制 B 条数据序列 $\{(o_t, a_t)\}_{t=k}^{k+L}$
    更新 θ。 // $p_\theta(o_t \mid o_{t-1}, a_t)$
  结束循环。
  // 正在收集数据
  o_1 = env.reset()
  当回合未停止**时**
    $a_t \sim A$
    $o_{t+1} \leftarrow$ 环境。步进（o_t， a_t）
  结束 while
  将经验添加到数据集 D
结束循环
return θ

5 实时合成安全屏蔽器

在本节中，我们介绍世界模型学习的融入，并提出我们的屏障综合方法——k步前瞻屏蔽器，这是一种传统屏障综合的变体 [27]。我们还给出理论证明，以说明我们的方法能够保证安全性。

5.1 学习环境动力学

为应对多智能体强化学习智能体起初并不具备有关环境的外部知识这一情形，我们的框架将在开始阶段训练一个粗略的世界模型。该世界模型是一个深度神经网络，用于学习预测与安全性考量相关的环境动力学。例如，自动驾驶汽车可能拥有不同的传感器输入，如激光雷达、摄像头和 GPS。如果我们只关注与运动相关的安全性，则粗略的世界模型将接受训练，以根据控制输入预测 GPS 信号。已有许多关于世界模型学习的研究 [20–22]。我们改编了一个来自循环状态空间模型（RSSM）的通用框架 [21]，该框架由三个组件组成（图 4b）：编码器、解码器和动力学网络，分别表示为 $\mathrm{enc}_\theta(s_t \mid s_{t-1},\ x_t)$、$\mathrm{dec}_\theta(s_t) \approx x_t$ 和 $\mathrm{dyn}\theta(s_t \mid s_{t-1},\ a_{t-1})$。

通过算法 2 获得这样的模型后，给定状态 x_t 和动作 a_t，以及先前状态 s_{t-1}的潜变量，我们可以通过以下方式预测下一个状态 x_{t+1}：

$$
\begin{aligned}
s_t &\sim \mathrm{enc}_\theta\left(s_t \mid s_{t-1}, x_t\right) \\
s_{t+1} &\sim \mathrm{dyn}_\theta\left(s_{t+1} \mid s_t, a_t\right) \\
x_{t+1} &\sim \mathrm{dec}_\theta\left(s_{t+1}\right)
\end{aligned}
$$

在这项工作中，我们专注于利用世界模型学习环境的低维内在属性，例如物理动力学，以进行安全屏蔽器综合。我们假设存在一个具有强大表达能力的世界模型，使我们能够抽象掉感知输入的细节，并在更高层次的抽象上对环境进行推理。对于这项工作，级联误差学习到的世界模型中可能出现的错误所引发的问题不在我们的讨论范围之内。

5.2 k步前瞻屏蔽器

我们假设状态空间已转换为符号抽象（通过 $f\colon S \to L$），该抽象由确定有限自动机 $\mathrm{A}^e = \left(Q^e, q_0^e, \Sigma^e, \delta^e, F^e\right)$给出。我们将线性时序逻辑安全规范转换为另一个确定有限自动机 $\mathrm{A}^S = \left(Q^S, q_0^S, \Sigma^S, \delta^S, F^S\right)$。我们构建一个双人博弈

$$\mathcal{G} = \left(G, g_0, \Sigma_1, \Sigma_2, \delta^g, F\right)$$

通过结合 A^e 和A^S。我们不直接求解游戏 G，而是添加额外的时间约束 $t \le k$，其中 $t \in$ 表示构建安全屏蔽器时的时间步，k 是表示游戏最大步数的超参数。修改后的游戏随后

$$\mathcal{G}^k = \left(G^k, g_0{}', \Sigma_1, \Sigma_2, \delta^{g\prime}, F^k\right) \tag{1}$$

其中状态空间 $G^k = G \times \{1...k\}$、初始状态 $g_0{}' = (g_0, t = 1)$、转移函数 $\delta^{g\prime}(g_t, t) = (\delta^g(g_t)\ t+1)$（可通过世界模型进行近似），以及获胜条件 $F^k = F \wedge (t \le k)$。我们可以使用 [27]中的方法求解双人安全博弈 G^k并计算获胜区域 $W \subseteq F^k$。然后，通过将 G^k和 W 转换为反应式系统$S = (\mathrm{QS}\ q_{0,}\mathrm{S}\ \Sigma_{I,}\mathrm{S}\ \Sigma_{O,}\mathrm{S}\ \delta\mathrm{S}\ \lambda\mathrm{S})$，构造 k 步前瞻安全屏蔽器。该安全屏蔽器包含以下组件：$\mathrm{QS} = G^k, q_{0,}\mathrm{S} = q_{0'}, \Sigma_I,\mathrm{S} = L \times \mathrm{A}, \Sigma_O,\mathrm{S} = \mathrm{A}, \delta\mathrm{S}\ (g^k,(l, a)) = \delta\left(g^k,(l, \lambda\mathrm{S}\ (g, (l, a)))\right)$对于所有 $g^k \in G, l \in L, a \in \mathrm{A}$，以及

$$
\lambda_{\mathcal{S}}(g, l, a) = \begin{cases} a & \text{if } \delta^k(g^k, (l, a)) \in W \\ a' & \text{if } \delta^k(g^k, (l, a)) \notin W \text{ for some arbitrary} \\ & \text{default } a' \text{ with } \delta^k\left(g^k, (l, a')\right) \in W. \end{cases}
$$

我们的安全屏蔽器综合与经典的屏障综合 [1, 18], 相似，后者同样通过求解双人对局来综合安全屏蔽器。主要区别在于，我们的方法只预测未来状态空间的一个子集，而以往方法会枚举规划时间范围内的所有可能状态。这带来了我们方法的主要优势：对于状态空间过大、无法预先计算的任务，我们的算法仍能高效运行，而以往方法则会失效。

5.3 安全性保证

我们证明，采用k 步前瞻屏蔽的动态屏蔽能够为多智能体强化学习智能体保证安全性。

命题 1. 给定由多智能体强化学习智能体、动态屏蔽器和环境共同产生的轨迹 $s_0 a_0 s_1 a_1 \cdots \in (S \times A)^\omega$，根据定义 *1*，状态-动作对 (s_t, a_t) 在每个时间步都是安全的。

证明。首先，算法 1 中的过程确保每个智能体在每个时间步都受到安全屏蔽器的监控，并且该安全屏蔽器至少会监控接下来 k 步内智能体的状态（否则，安全屏蔽器将重新计算）。其余证明与 [18] 中集中式屏蔽的正确性证明相同。对于安全屏蔽器 $\mathrm{S} = (Q,\ q_0,\ \Sigma_I,\ \Sigma_O,\ \delta,\ \lambda)$ 下的任意智能体，都存在一个对应的

运行 $q_0 q_1, \ldots q_m \in (S \times A)^\omega$，其中 $m \leq k$ 是重构该安全屏蔽器之前的持续时间。通过构造安全屏蔽器，我们得到环境抽象确定有限自动机 A^e 和安全规范确定有限自动机 A^s。我们可以将安全屏蔽器 S 的运行 $q_0, q_1, \ldots q_m$ 投影为 A_s 上的轨迹 $q_0^s(f(s_0)\ a_0)q_1^s(f(s_1)\ a_1)\ldots q_m^s(f(s_m)\ a_m)$。由于我们根据双人安全博弈的获胜区域构造安全屏蔽器，因此智能体沿轨迹访问的每个状态 $q_i^s(f(s_i)$ 都应是 A^s 中的安全状态。安全屏蔽器 S 确保 A^s 中定义的安全规范永远不会被违反。因此，对于每个多智能体强化学习智能体，在每一步，联合状态—动作对 (s_t, a_t) 都是安全的。

□

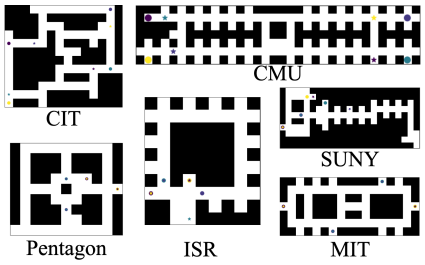


图 5：不同的网格世界环境。圆点表示智能体，星号表示目标，黑色方块表示障碍物。



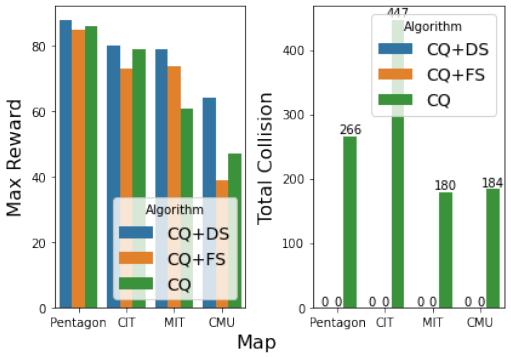
(a) 简单扩散 **(b) 简单对抗**

图6：修改后的多智能体粒子环境（MPE）。蓝色圆圈表示智能体，黑点表示地标。在简单对抗中，红色圆圈是对抗智能体，绿色圆点是目标地标，我们令n_{good} : $n_{adversary} = 3$: 1。MPE环境是无界的，但如果智能体移动得太远，将受到惩罚。

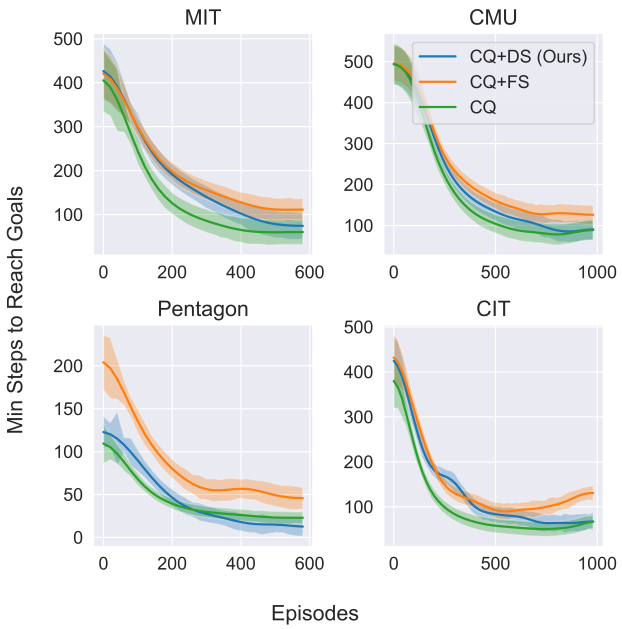
6 实验

在本节中，我们通过多个基准任务对所提出的安全多智能体强化学习框架（算法 1）的性能进行实证评估。我们将算法应用于网格世界 [39] 的四种不同地图（如图 5 所示），以及多智能体粒子环境中的两个环境（协作型和混合协作型，如图 6 所示）。我们将所提出的算法与采用因子化屏蔽的 CQ-learning [16],CQ（CQ+FS）、深度确定性策略梯度 [31], MADDPG [32], 以及 MAD- DPG+CBF [11]进行比较。我们总共开展了三组实验。在前两组实验中，我们假设已知环境动力学，以便更加方便地评估我们的屏蔽框架的性能。

在最后一项实验中，我们在没有任何关于环境的外部知识的情况下训练框架，以展示采用所提出屏蔽方法的多智能体强化学习在实践中的有效性。我们使用 Python 实现算法，并通过 Slugs [17]综合生成屏蔽器。对于每项实验，我们分别在训练阶段和测试阶段评估算法。为减小离群值的影响，我们在 5 次独立运行中完成所有实验，并对结果取平均值。



(a) 左侧图展示了智能体在学习过程中获得的最大奖励。右侧图展示了智能体在学习过程中发生的碰撞总数。



(b) 到达目标所需的最少步数**值是衡量所学策略最优性的指标。**

图 7：网格世界实验。*CQ+DS*、*CQ+FS*和*CQ*分别表示采用动态屏蔽的 CQ-learning、采用因子化屏蔽的 CQ-learning，以及未采用安全屏蔽器的 CQ-learning。

实验设置。图 5 展示了改编自网格世界 [39]的六张网格世界基准环境地图。每张地图中都有四个智能体学习在躲避障碍物的同时进行导航。

环境。动作集合为 A $=\{stay, up, down, left, right\}$。我们随机为每个智能体分配一个唯一的目标。一旦智能体到达其目标，它就会一直停留在那里，直到所有智能体都到达各自的目标。我们为此任务设置稀疏奖励，即给予 -1 存活惩罚、-10碰撞惩罚，以及 $+100$ 到达目标的奖励。

图 6 展示了修改后的多智能体粒子环境中的两项任务 [32]，即 简单扩散 和 简单对抗。简单扩散 任务的目标是让智能体协同合作，在避免碰撞的同时抵达目标。简单对抗 任务的目标是让友方智能体前往目标并迷惑对抗智能体，而对抗智能体则试图在避免碰撞的同时抵达目标。这些任务在以下两个方面比网格世界 更难：

(1) 多智能体粒子环境的状态空间是连续且无界的。(2) 多智能体粒子环境中的智能体具有更复杂的动力学，例如动量和加速度。

动作集合为 A$'$ $=\{stay,\ up,\ down,\ left,\ right,$ **制动**$\}$，其中的动作可控制加速度。例如，如果智能体采取 $stay$，它会以原有速度移动，而不是停在原地。我们使用 $brake$ 来模拟现实世界中的制动，智能体会沿速度方向施加较大的减速度，直至停止。$brake$ 动作遵循运动学定律；例如，速度较快的智能体需要更长的距离才能制动停下。每个智能体都会获得与其目标距离成反比的奖励，并因碰撞而受到惩罚。

保守行为评估。我们将 CQ-learning 与因子化屏蔽和所提出的动态屏蔽相结合，并将其应用于四种不同的网格世界环境（图 5）。我们使用训练阶段的最大奖励、碰撞次数和回合步数来评估算法。图7a中的结果表明，因子化屏蔽和动态屏蔽都能确保所有地图上的学习过程无碰撞。不过，与因子化屏蔽和不使用屏蔽的学习方法相比，动态屏蔽能够获得奖励更高的策略。图7b显示，CQ+DS 智能体到达目标所需的步数少于 CQ+FS。最终，动态屏蔽策略的表现可与未受干预的 CQ-learning 相媲美。因此，这表明所提出的动态屏蔽在保证安全性的同时，能够缓解因子化屏蔽造成的协调开销。

可扩展性评估。我们评估状态空间和智能体数量扩大时动态屏蔽的性能。我们将 DDPG 和 MADDPG 与动态屏蔽相结合，并将其应用于修改后的 MPE 环境（如图 6 所示）。正如前文所述，MPE 的状态空间是网格世界的扩展。由于无法事先为整个状态空间合成屏蔽器，因此因子化屏蔽在这个无界环境中会失效。我们考虑两种将安全机制融入 DDPG/MADDPG 的基线算法：DDPG/MADDPG+Safe，它在奖励函数中加入安全奖励；以及 MADDPG+CBF，它通过控制屏障函数来保证安全。对于 MAD- DPG+CBF，我们采用 [11] 中提出的屏障函数。表1显示，无论任务如何，MADDPG/DDPG+DS 和 MADDPG+CBF 都能保证无碰撞，而 MADDPG/DDPG+Safe 则始终会发生碰撞，且碰撞次数会随着智能体数量增加而上升。表1还展示了测试阶段的累积奖励。收敛时，MADDPG/DDPG+DS 获得了更高的

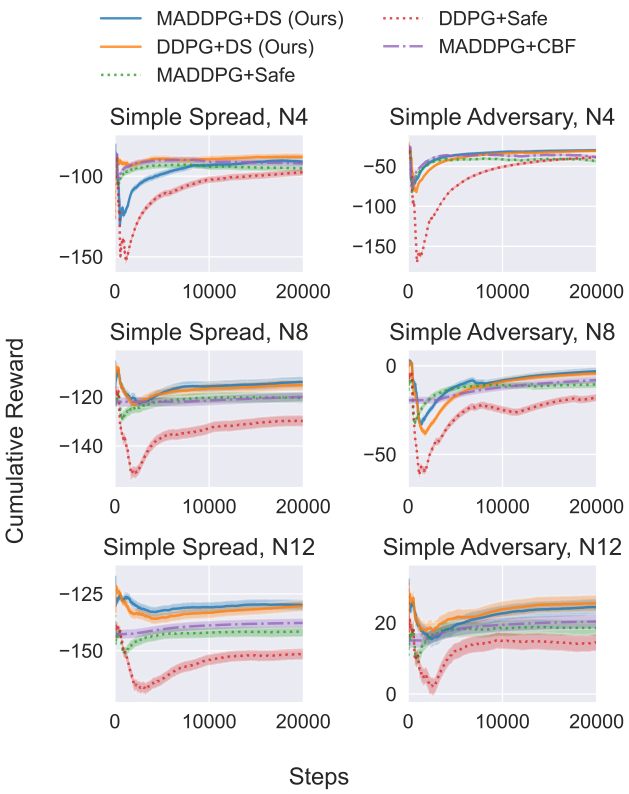


图8：将动态屏蔽（MAD-DPG/DDPG+DS）与其他基线方法（MAD-DPG/DDPG+Safe、MADDPG+CBF）进行比较。

奖励，高于 MADDPG/DDPG+Safe。图 8 中的学习曲线表明，采用动态屏蔽的多智能体强化学习算法能够以更高的奖励更快收敛。我们还评估了智能体数量增加时动态屏蔽的性能。表1显示，尽管 MADDPG/DDPG+Safe 的碰撞次数会随着智能体数量增加而上升，但 MADDPG+DS 和 MADDPG+CBF 在两种环境中始终能够保证安全，并获得更高的奖励。随着智能体数量增加，我们观察到 MADDPG+DS 与 MADDPG+CBF 之间的性能差距不断扩大。这表明，随着智能体数量增加，我们的动态屏蔽仍能将干预程度维持在较低水平（更不保守）。

基于模型的动态屏蔽。在此实验中，我们移除环境中的 $brake$ 动作，并在标准 MPE 简单扩散 和简单对抗 环境中评估算法。智能体从环境中收集 $3e5$ 轨迹，并通过算法 2 训练世界模型，以学习根据当前位置、速度和动作 $[x_t, v_t, a_t]^T$ 预测下一步位置 x_{t+1}。由于时间信息与该环境中的动力学无关，我们使用一个 $32 \times 64 \times 32$ MLP 网络作为动力学模型。我们将此过程命名为基于模型的动态屏蔽（MBDS）。我们通过以下方式计算测试阶段的安全率：

$$r_{safety} = \frac{\sum_i \mathbb{1}(\text{collisions in step i} > 0)}{\text{number of steps}}.$$

		MADDPG+DS		MADDPG+Safe		深度确定性策略梯度+DS		深度确定性策略梯度+Safe		MADDPG+CBF	
Task	N	REW	COL	REW	COL	REW	COL	REW	COL	REW	COL
Spread	4	-77 ± 8	**0.0**	-84 ± 6	1.3 ± 0.9	-76 ± 8	**0.0**	-82 ± 6	0.9 ± 1.0	-75 ± 6	**0.0**
	8	-112 ± 10	**0.0**	-119 ± 9	10.0 ± 2.2	-113 ± 10	**0.0**	-125 ± 13	7.7 ± 1.9	-118 ± 12	**0.0**
	12	-129 ± 9	**0.0**	-141 ± 9	38.2 ± 6.6	-129 ± 11	**0.0**	-151 ± 14	26.5 ± 6.5	-138 ± 11	**0.0**
Adversary	4	-25 ± 3	**0.0**	-30 ± 4	1.1 ± 0.8	-26 ± 3	**0.0**	-26 ± 3	0.9 ± 1.3	-23 ± 3	**0.0**
	8	-1 ± 3	**0.0**	-5 ± 8	6.8 ± 2.0	-2 ± 4	**0.0**	-11 ± 16	5.5 ± 1.7	-2 ± 5	**0.0**
	12	21 ± 9	**0.0**	16 ± 8	43.5 ± 10.6	23 ± 9	**0.0**	13 ± 11	33.2 ± 10.9	18 ± 8	**0.0**

表1：比较测试阶段各算法的平均奖励和碰撞次数。表中，$a \pm b$表示10次独立测试运行结果的均值和方差。（奖励、碰撞分别表示累积奖励和碰撞次数）。

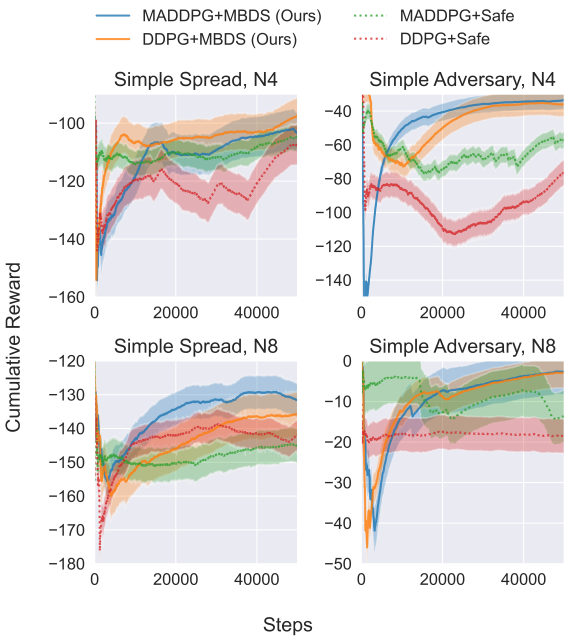


图9：将基于模型的动态屏蔽（MBDS）与其他基线方法进行比较。

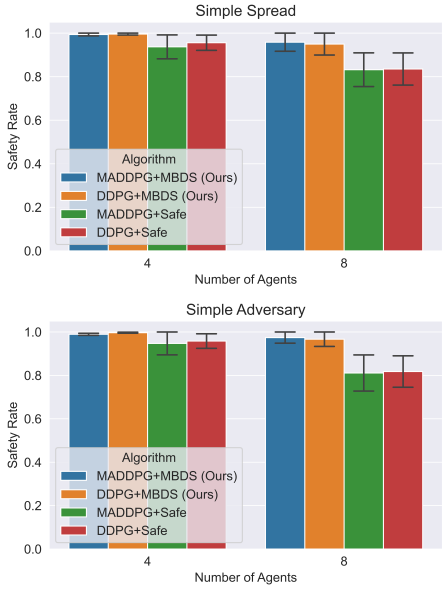


图10：多智能体物理环境任务中的安全率。

图9展示了，采用基于模型的动态屏蔽的多智能体强化学习（MADDPG+MBDS、深度确定性策略梯度+MBDS）所获得的累积奖励至少不低于其他基线方法（MADDPG+Safe、深度确定性策略梯度+Safe）。此外，当智能体数量从4个增加到8个时，屏蔽的安全率（如图10所示）下降不超过5%，并保持在90%以上。然而，其他基线方法的安全率在简单扩散 中下降了10%，在简单对抗中下降了20%。因此，基于模型的动态屏蔽是为多智能体强化学习提供安全保证的一种有效方法。

7 结论

本文提出了一种通过基于模型的动态屏蔽来解决安全多智能体强化学习问题的新方法。所提出的方法对多智能体强化学习框架的干预极小，同时确保满足由线性时序逻辑表达式定义的安全规范。我们还提出了一种有效的技术，可实时合成安全屏蔽器，并提供安全保证的理论证明。此外，我们进行了大量实验，以证明相较于其他基线方法，我们的算法在基准任务中的安全性和学习性能方面表现更好。我们也承认，本方法存在一些局限性。首先，我们没有解决学习到的动力学模型不准确时可能出现的级联错误问题。这可能会削弱我们框架所提供的安全保证。其次，我们尚未分析算法的时间复杂度，而该复杂度在很大程度上取决于反应式博弈求解器的实现。在未来的工作中，我们计划探索在存在风险感知动力学模型的情况下强制实施安全保证的方法，以帮助减轻级联错误的影响。此外，我们将全面分析算法的时间复杂度，以确保其能够扩展到更大、更复杂的环境。

致谢

我们感谢 Ingy ElSayed-Aly 和 DanielMelcer 就这项工作进行了富有洞见的讨论。本研究是在卡内基梅隆大-学机器人学研究所暑期学者项目中作为本科生研究体验的一部分开展的。Wenli Xiao 的奖-学金资助由深圳市人工智能与机器人研究院提供。

参考文献

[1] Mohammed Alshiekh、Roderick Bloem、Rüdiger Ehlers、Bettina Könighofer、Scott Niekum 和 Ufuk Topcu。2018年。通过屏蔽实现安全强化学习。载于《AAAI人工智能大会论文集》，第32卷。[2] Rajeev Alur。2015。网络物理系统原理。MIT出版社。[3] Itamar Arel、Cong Liu、Tom Urbanik 和 Airton G Kohls。2010年。基于强化学习的多智能体系统，用于网络交通信号控制。《IET智能交通系统》4卷2期（2010年），128–135。[4] Christel Baier 和 Joost-Pieter Katoen。2008年。模型检测原理。MIT出版社。[5] Christel Baier 和 Joost-Pieter Katoen。2008年。模型检测原理。MIT出版社。[6] Osbert Bastani、Shuo Li 和 Anton Xu。2021年。通过统计模型预测屏蔽实现安全强化学习。《机器人学：科学与系统》。1–13。[7] Sushrut Bhalla、Sriram Ganapathi Subramanian 和 Mark Crowley。2020年。面向自动驾驶的深度多智能体强化学习。载于《加拿大人工智能大会》。Springer，67–78。[8] Urs Borrmann、Li Wang、Aaron D Ames 和 Magnus Egerstedt。2015年。用于安全群体行为的控制屏障证书。《IFAC论文集》48卷27期（2015年），68–73。[9] Alper Kamil Bozkurt、Yu Wang、Michael M Zavlanos 和 Miroslav Pajic。2020年。利用无模型强化学习，根据线性时序逻辑规范进行控制综合。载于《2020年IEEE国际机器人与自动化会议（ICRA）》。IEEE，10349–10355。[10] Lucian Busoniu、Robert Babuska 和 Bart De Schutter。2008年。多智能体强化学习综合综述。《IEEE系统、人类与控制论汇刊C辑（应用与综述）》38卷2期（2008年），156–172。[11] Zhiyuan Cai、Huanhui Cao、Wenjie Lu、Lin Zhang 和 Hao Xiong。2021年。通过分散式多重控制屏障函数实现安全多智能体强化学习。arXiv预印本 arXiv:2103.12553（2021年）。[12] Yuxiao Chen、Huei Peng 和 Jessy Grizzle。2017年。利用屏障函数实现低速自动驾驶车辆的障碍物规避。《IEEE控制系统技术汇刊》26卷1期（2017年），194–206。[13] Yuxiao Chen、Andrew Singletary 和 Aaron D Ames。2020年。面向驱动能力受限的多机器人作业，采用控制屏障函数方法保证避障。《IEEE控制系统快报》5卷1期（2020年），127–132。[14] Richard Cheng、Gábor Orosz、Richard M Murray 和 Joel W Burdick。2019年。通过屏障函数实现端到端安全强化学习，适用于安全关键型连续控制任务。载于《AAAI人工智能大会论文集》，第33卷。3387–3395。[15] Yinlam Chow、Ofir Nachum、Edgar Duenez-Guzman 和 Mohammad Ghavamzadeh。2018年。基于李雅普诺夫方法的安全强化学习。《神经信息处理系统进展》31（2018年）。[16] Yann-Michaël De Hauwere、Peter Vrancx 和 Ann Nowé。2010年。学习多智能体状态空间表示。载于《第9届自主智能体与多智能体系统国际会议论文集：第1卷》。715–722。[17] Rüdiger Ehlers 和 Vasumathi Raman。2016年。Slugs：可扩展的 GR(1) 综合。载于《计算机辅助验证国际会议》。Springer，333–339。[18] Ingy ElSayed-Aly、Suda Bharadwaj、Christopher Amato、Rüdiger Ehlers、Ufuk Topcu 和 Lu Feng。2021年。通过屏蔽实现安全多智能体强化学习。arXiv预印本 arXiv:2101.11196（2021年）。[19] Javier García 和 Fernando Fernández。2015年。安全强化学习综合综述。《机器学习研究期刊》16卷1期（2015年），1437–1480。[20] David Ha 和 Jürgen Schmidhuber。2018年。世界模型。arXiv预印本 arXiv:1803.10122（2018年）。[21] Danijar Hafner、Timothy Lillicrap、Ian Fischer、Ruben Villegas、David Ha、Honglak Lee 和 James Davidson。2019年。学习潜在动力学，以便从像素进行规划。载于《国际机器学习会议》。PMLR，2555–2565。[22] Danijar Hafner、Timothy Lillicrap、Ian Fischer、Ruben Villegas、David Ha、Honglak Lee 和 James Davidson。2019年。学习潜在动力学，以便从像素进行规划。载于《国际机器学习会议》。PMLR，2555–2565。[23] Ernst Moritz Hahn、Mateo Perez、Sven Schewe、Fabio Somenzi、Ashutosh Trivedi 和 Dominik Wojtczak。2019年。无模型强化学习中的欧米伽正则目标。载于《系统构造与分析工具及算法国际会议》。Springer，395–412。[24] Mohammadhosein Hasanbeig、Alessandro Abate 和 Daniel Kroening。2020。带逻辑约束的谨慎强化学习。arXiv 预印本 arXiv:2002.12156（2020）。[25] Mohammadhosein Hasanbeig、Alessandro Abate 和 Daniel Kroening。2020。带逻辑约束的谨慎强化学习。arXiv 预印本 arXiv:2002.12156（2020）。[26] Nils Jansen、Bettina Könighofer、JSL Junges、AC Serban 和 Roderick Bloem。2020。使用概率屏障的安全强化学习。（2020）。[27] Bettina Könighofer、Mohammed Alshiekh、Roderick Bloem、Laura Humphrey、Robert Könighofer、Ufuk Topcu 和 Chao Wang。2017。屏障综合。《系统设计中的形式化方法》51，2（2017），332–361。[28] Hadas Kress-Gazit、Georgios E Fainekos 和 George J Pappas。2009。基于时序逻辑的反应式任务与运动规划。IEEE《机器人学汇刊》25，6（2009），1370–1381。[29] Orna Kupferman 和 Moshe Y Vardi。2001。安全属性的模型检测。《系统设计中的形式化方法》19，3（2001），291–314。[30] Shuo Li 和 Osbert Bastani。2020。针对随机动力学下安全强化学习的鲁棒模型预测屏障。载于2020年IEEE国际机器人与自动化会议（ICRA）。IEEE，7166–7172。[31] Timothy P Lillicrap、Jonathan J Hunt、Alexander Pritzel、Nicolas Heess、Tom Erez、Yuval Tassa、David Silver 和 Daan Wierstra。2015。深度强化学习下的连续控制。arXiv 预印本 arXiv:1509.02971（2015）。[32] Ryan Lowe、Yi Wu、Aviv Tamar、Jean Harb、Pieter Abbeel 和 Igor Mordatch。2017。混合合作竞争环境中的多智能体行动者-评论家。神经信息处理系统（NIPS）（2017）。[33] Songtao Lu、Kaiqing Zhang、Tianyi Chen、Tamer Başar 和 Lior Horesh。2021。用于安全多智能体强化学习的分布式策略梯度下降上升。载于AAAI人工智能大会论文集，第35卷。8767–8775。[34] Wenhao Luo、Wen Sun 和 Ashish Kapoor。2020。具有不确定性的多机器人避碰：概率安全屏障证书方法。神经信息处理系统进展33（2020），372–383。[35] Yiwei Lyu、Wenhao Luo 和 John M Dolan。2021。自动驾驶车辆的概率安全保障自适应汇入控制。载于2021年IEEE国际机器人与自动化会议（ICRA）。IEEE，10764–10770。[36] Yiwei Lyu、Wenhao Luo 和 John M Dolan。2022。使用参数化控制屏障函数对异构自动驾驶车辆进行自适应安全汇入控制。载于2022年IEEE智能车辆研讨会（IV）。IEEE，542–547。[37] Yiwei Lyu、Wenhao Luo 和 John M Dolan。2022。具有社会偏好的责任关联多智能体避碰。载于2022年IEEE第25届智能交通系统国际会议（ITSC）。IEEE，3645–3651。[38] Francisco S Melo 和 Manuela Veloso。2009。协调学习：利用多智能体系统中的稀疏交互。载于第八届自主智能体与多智能体系统国际会议论文集，第2卷。773–780。[39] Francisco S Melo 和 Manuela Veloso。2009。协调学习：利用多智能体系统中的稀疏交互。载于第八届自主智能体与多智能体系统国际会议论文集，第2卷。Citeseer，773–780。[40] Liane Okdinawati、Togar M Simatupang 和 Yos Sunitiyoso。2017。用于协同运输管理（CTM）价值共创的多智能体强化学习。《国际信息系统与供应链管理杂志》（IJISSCM）10，3（2017），84–95。[41] Martin Pecka 和 Tomas Svoboda。2014。强化学习的安全探索技术综述。载于自主系统建模与仿真国际研讨会。Springer，357–375。[42] Adolfo Perrusquía、Wen Yu 和 Xiaoou Li。2020。使用多智能体强化学习进行冗余机器人控制。载于2020年IEEE第16届自动化科学与工程国际会议（CASE）。IEEE，1650–1655。[43] Amir Pnueli。1977。程序时序逻辑。载于第18届计算机科学基础年度研讨会（sfcs 1977）。ieee，46–57。[44] Stephen Prajna、Ali Jadbabaie 和 George J Pappas。2007。利用屏障证书进行最坏情况与随机安全验证的框架。IEEE《自动控制汇刊》52，8（2007），1415–1428。[45] Zengyi Qin、Kaiqing Zhang、Yuxiao Chen、Jingkai Chen 和 Chuchu Fan。2021。利用分布式神经屏障证书学习安全多智能体控制。arXiv 预印本 arXiv:2101.05436（2021）。[46] Kristin Y Rozier。2011。线性时序逻辑符号模型检测。《计算机科学评论》5，2（2011），163–203。[47] Shai Shalev-Shwartz、Shaked Shammah 和 Amnon Shashua。2016。面向自动驾驶的安全多智能体强化学习。arXiv 预印本 arXiv:1610.03295（2016）。[48] Mohit Srinivasan 和 Samuel Coogan。2020。在线性时序逻辑规范下使用屏障函数控制移动机器人。IEEE《机器人学汇刊》37，2（2020），363–374。[49] Andrew Taylor、Andrew Singletary、Yisong Yue 和 Aaron Ames。2020。利用控制屏障函数学习安全关键控制。载于动力学与控制学习。PMLR，708–717。

[50] Brijen Thananjeyan、Ashwin Balakrishna、Suraj Nair、Michael Luo、Krishnan Srinivasan、Minho Hwang、Joseph E Gonzalez、Julian Ibarz、Chelsea Finn 和 Ken Goldberg。2021。恢复强化学习：利用学习得到的恢复区域实现安全强化学习。*IEEE*机器人与自动化快报 6，3（2021），4915–4922。[51] Alphan Ulusoy、Stephen L Smith、Xu Chu Ding、Calin Belta 和 Daniela Rus。2013。具有时序逻辑约束的多机器人路径规划中的最优性与鲁棒性。《国际机器人研究杂志》32，8（2013），889–911。[52] Spencer Van Koevering、Yiwei Lyu、Wenhao Luo 和 John Dolan。2022。使用指数控制障碍函数实现自动驾驶车辆可证明的概率安全性与可行性保障控制。载于*2022*年*IEEE*智能车辆研讨会（*IV*）。IEEE，952–957。[53] Akifumi Wachi 和 Yanan Sui。2020。受约束马尔可夫决策过程中的安全强化学习。载于国际机器学习会议。PMLR，9797–9806。[54] Akifumi Wachi 和 Yanan Sui。2020。受约束马尔可夫决策过程中的安全强化学习。载于国际机器学习会议。PMLR，9797–9806。[55] Li Wang、Aaron D Ames 和 Magnus Egerstedt。2017年。无碰撞多机器人系统的安全屏障证书。*IEEE*《机器人学汇刊》33，3（2017），661–674。[56] Peter Wieland 和 Frank Allgöwer。2007年。利用控制屏障函数构造安全性。*IFAC*会议论文集40，12（2007），462–467。[57] Chao Yu、Xin Wang 和 Zhanbo Feng。2019年。面向移动感知机器团队的协调多智能体强化学习。载于 第*18*届自主智能体与多智能体系统国际会议论文集。2297–2299。[58] Jun Zeng、Bike Zhang 和 Koushil Sreenath。2021年。采用离散时间控制屏障函数的安全关键模型预测控制。载于 *2021*年美国控制会议（*ACC*）。IEEE，3882–3889。[59] Kaiqing Zhang、Zhuoran Yang 和 Tamer Başar。2021年。多智能体强化学习：理论与算法综述。《强化学习与控制手册》（2021），321–384。[60] Ming Zhou、Jun Luo、Julian Villella、Yaodong Yang、David Rusu、Jiayu Miao、Weinan Zhang、Montgomery Alban、Iman Fadakar、Zheng Chen 等。2020年。SMARTS：面向自动驾驶的可扩展多智能体强化学习训练平台。*arXiv*预印本 *arXiv:2010.09776*（2020）。