arXiv:2303.00822v2 [cs.AI] 5 Apr 2023

对抗环境下的攻击者诱捕规划

Brittany Cates[1], **Anagha Kulkarni**[2],**Sarath Sreedharan**[1]

[1] 计算机科学系，科罗拉多州立大学，美国科罗拉多州 [2] Invitae 公司，美国加利福尼亚州旧金山
brittany.cates@colostate.edu, anagha.kulkarni@invitae.com,
sarath.sreedharan@colostate.edu

摘要

本文提出了一种规划框架，用于针对在某种环境中行动的攻击者制定防御策略；在这种环境中，防御者可以在不被攻击者察觉的情况下采取行动。防御者的目标是秘密引导攻击者进入一个陷阱状态，使其无法实现自己的目标。此外，防御者还必须在 K 步之内实现其目标。这个K 步数是一个悲观下界，在此范围内，攻击者不太可能察觉到环境中存在威胁。这种防御策略在蜜罐或蜜网等现实系统中非常有用：毫无戒心的攻击者与模拟生产系统交互，并误以为它是真正的生产系统。通常，攻击者与防御者之间的交互通过博弈论框架来描述。我们的问题表述使我们能够将其建模为一种更简单的无限时域折扣马尔可夫决策过程（MDP），其中，MDP 的最优策略代表防御者针对攻击者动作所采取的策略。通过实证评估，我们展示了该问题表述的优点。

1 引言

近年来，对抗性规划受到了广泛关注（Masters、Kirley 和 Smith，2021；Bernardini、Fagnani 和 Franco，2020；Chakraborti 等人，2019）。相关研究涵盖了多种场景，从简单的路径规划（Masters 和 Sardina，2017）到更贴近现实的网络安全问题（Letchford 和 Vorobeychik，2013）。这些工作不仅为研究者提供了展示现代规划工具有效性的机会，还推动了新规划范式的发展，使其更适合处理此类情形（例如，斯塔克尔伯格规划形式化（Speicher 等人，2018））。上述诸多例子有一个共同的基本特点：问题本质上被构建为博弈论问题。本文旨在介绍一类具有实际意义的新型对抗性规划问题，并为其提供一种简单直观的规划形式化方法。

本文介绍了攻击者诱捕问题，其中隐蔽防御者操纵周围环境以破坏攻击。我们将其概念化为引导一名未察觉防御者存在的攻击者进入预定陷阱状态的问题。防御者的有效性取决于其隐藏自身存在的能力。只要攻击者没有意识到防御者的存在，防御者就可以依靠基本规划技术来确定其策略。

虽然这看起来像是一种不切实际的情境，但实际上，这是网络安全领域中一个受到广泛研究的问题。具体而言，蜜罐是现实网络中常用的一种策略，用于应对攻击者诱捕问题。顾名思义，蜜罐通过模拟一个攻击者可以渗透并与之交互的真实生产系统，诱骗并困住黑帽攻击者。蜜罐可作为一种防御机制，将攻击者从真实系统引开并将其困住；也可用于研究攻击者的策略，以开发更强大的安全机制。

值得注意的是，当前开发蜜罐的策略完全是离线进行的。网络设计者负责组织系统，将对应于有吸引力目标的蜜罐布置在潜在攻击者的行动路径上。从攻击者诱捕问题的角度来看，这代表了一种机制或环境设计方案（Keren、Gal 和 Karpas（2021））。我们则提出一种新颖的主动解决方案，利用规划器适应攻击者可能采取的动作。

我们将把防御者的规划问题建模为马尔可夫决策过程（MDP）。防御者会隐藏自身的存在，利用攻击者对系统架构的信念来预测攻击者的策略，并引导攻击者进入预设的陷阱状态。它还会利用攻击者对领域随机性的信念来掩饰自身的动作。攻击者不知道的是，防御者可以控制某些攻击者动作的结果。例如，面对 SQL 注入攻击时，防御者可以决定向攻击者返回什么密码；又如，在考虑物理空间的安全性时，可以决定攻击者到达房间后打开哪扇自动门。通过仔细选择结果，防御者可以将攻击者引入陷阱状态；若无法做到，也可以将其引向效用较低的状态。

除了介绍这一问题框架和一种解决方法外，我们还将介绍一种计算

版权所有 © 2023，美国人工智能促进协会 (www.aaai.org)。保留所有权利。

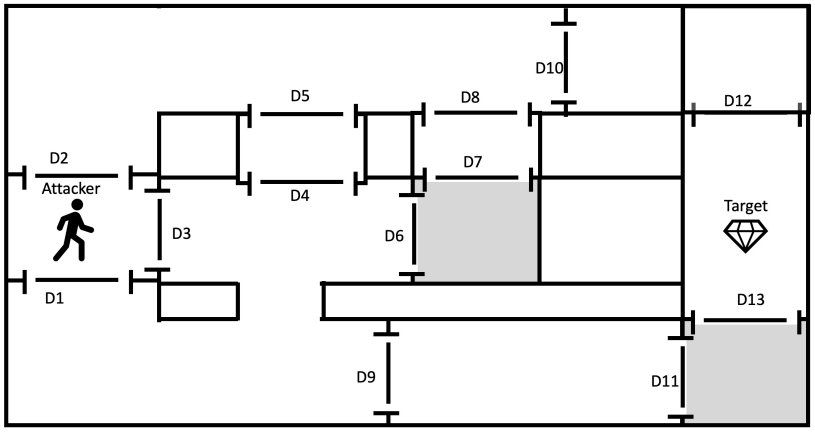


图1：运行示例示意图。图中，攻击者试图闯入一个特定房间，窃取一件物品。防御者的目标是通过控制攻击者进入房间时哪些自动门会打开，引诱攻击者进入灰色标出的某个陷阱房间。

防御者行动预算的方法。我们将展示这种预算如何让防御者操控结果，同时不引起攻击者对环境正受到控制的怀疑。我们将在若干传统的马尔可夫决策过程基准上评估我们的解决方案，并介绍该方法的计算特性。

总而言之，本文的贡献包括：

1. 确定并形式化攻击者诱捕问题。2. 将攻击者诱捕问题表述为马尔可夫决策过程规划问题。3. 开发基于搜索的方法，以确定一个能够保证不被攻击者察觉的防御者预算下界。4. 在多个标准马尔可夫决策过程基准上，评估解决方法在降低攻击者价值方面的有效性。

2 动机示例

作为运行示例，我们考虑一个关于小型展厅安全性的简单场景。攻击者是一名窃贼，试图从展厅中窃取一颗珍贵的钻石。窃贼可以获取展厅地图，并知道这件贵重物品的确切位置。该领域中唯一的随机性来源，是遍布展厅的自动门。攻击者认为，每当进入有多扇门的房间时，其中一扇门都会以均匀概率打开。然而，攻击者并不知道，有一个自动化防御者智能体能够控制哪些门会打开。防御者的目标是选择合适的门，引导攻击者进入某个陷阱状态。攻击者知道这些房间的存在，却不知道它们是陷阱状态。

图1展示了展厅的布局和攻击者的当前位置，陷阱状态以灰色高亮显示。防御者可以选择打开门D3。如果攻击者是理性的智能体，他们会径直沿着走廊前进，直到进入陷阱状态。然而，情况可能并非如此。他们也可能转而走进下方的走廊。防御者需要通过打开门D11来进行适应，从而将攻击者引向另一个陷阱状态。防御者必须持续追踪攻击者可能前往的地点及其可能采取的行动，以便相应地做出回应并实现预期结果。除了将攻击者引导至某个陷阱状态之外，防御者可能还需要考虑其他因素。例如，其中一些通道可能摆放着攻击者能够损坏的贵重物品。为避免此类高成本风险，防御者可以迫使攻击者选择能够将潜在损害降至最低的路径。

3 背景

我们将规划问题描述为一个无限时域折扣马尔可夫决策过程，简称 MDP（Puterman 1990）。沿用标准记法，我们将该 MDP 表示为如下形式的元组 $\mathcal{M} = \langle S, A, T, R, \gamma, S_0 \rangle$，其中 S 是状态集合，A 是动作集合，$T\colon S \times A \times S \to [0,1]$ 是转移函数，$R\colon S \times A \times S \to \mathbb{R}$ 是奖励函数，$\gamma \in [0,1)$ 是折扣因子，S_0 是该任务的初始状态。这类 MDP 所采用的解概念对应于平稳确定性策略，下文简称策略。每个策略 π 都对应一个函数，即 $\pi\colon S \to A$，它为每个可能的状态关联一个动作。给定策略 π 的价值 $V^{\pi}\colon S \to \mathbb{R}$，表示从某个状态执行该策略所获得的奖励折扣和的期望值。同样，Q 函数 $Q^{\pi}\colon S \times A \to \mathbb{R}$ 表示在给定状态下执行特定动作，然后遵循策略 π 所获得的奖励折扣和的期望值。如果不存在其他策略 π' 和状态 $s \in S$，使得 $V^{\pi'}(s) > V^{\pi^*}(s)$，则称策略对 $\mathcal{M}$ 是最优的（记为 π^*）。我们将最优策略对应的价值函数记为 V^*（最优的 Q-

函数则记为 $\mathcal{Q}^*$）。我们用“轨迹”一词指代状态与动作构成的序列（记为 $\tau = \langle s_0, a_0, ..., s_i \rangle$）。我们用记法 $P(\tau|\mathcal{M})$ 表示在给定模型 $\mathcal{M}$ 下，轨迹 τ 发生的可能性，其中概率递归定义如下：

$$P(\tau|\mathcal{M}) = T(s_0, a_0, s_1) \times P(\langle s_1, a_1, ..., s_i \rangle|\mathcal{M})$$

本文将分别为攻击者和防御者关联不同的马尔可夫决策过程，因此我们用上标 $\mathcal{A}$ 和 $\mathcal{D}$ 来区分这两个马尔可夫决策过程及其组成部分。

4 我们的方法

我们的方法的核心是两个智能体之间的交互：攻击者（$\mathcal{A}$）和隐蔽防御者（$\mathcal{D}$）。每个智能体都有一个关于问题的模型，并相信自己正在解决该问题。攻击者从一个无限时域折扣马尔可夫决策过程开始，其形式为 $\mathcal{M}^{\mathcal{A}} = \langle S^{\mathcal{A}}, A^{\mathcal{A}}, T^{\mathcal{A}}, R^{\mathcal{A}}, \gamma^{\mathcal{A}}, S_0^{\mathcal{A}} \rangle$，其中 $S^{\mathcal{A}}$ 是攻击者认为属于任务的状态集合，$A^{\mathcal{A}}$ 是攻击者能够执行的动作集合，$T^{\mathcal{A}}$ 是攻击者对任务如何响应其动作而演变的信念，最后，$R^{\mathcal{A}}$ 和 $\gamma^{\mathcal{A}}$ 分别表示攻击者内部的奖励函数和折扣。为简化表述，我们假设攻击者的奖励始终为非负值。这甚至可能是防御者有意歪曲后泄露给外界的信息。我们将根据该模型计算出的价值函数记作 $\mathcal{V}^{\mathcal{A}}$，并将最优值记作 $\mathcal{V}^{\mathcal{A}*}$（类似地，我们将用 $\mathcal{Q}^{\mathcal{A}}$ 表示 Q 函数）。在我们的贯穿示例中，攻击者的状态包括攻击者可能身处的各个位置，而动作则对应于在展厅内移动，以及最终在抵达最后一个房间时窃取钻石。

防御者无法获取的一项信息，是攻击者可能采用的策略。尽管防御者可能无法推断出确切的策略，但仍可以利用模型生成攻击者可能采取的动作的概率分布。具体而言，可以使用噪声理性模型（Jeon、Milli 和 Dragan，2020）来描述攻击者的决策过程。这通常是一种非常保守的模型，但仍会为非最优动作赋予一定概率。它还有一个额外优点：在心理学上是可行的，因此能够对涉及人类攻击者的情形进行建模。例如，使用这种噪声理性模型，可以捕捉攻击者在图1所示示例中偏离最优路径的可能性。具体而言，我们将动作选择的概率定义为

$$P(a|s) \propto e^{\kappa \times \mathcal{Q}^{\mathcal{A}*}(s,a)}, \text{ Where } s \in S^{\mathcal{A}}\ a \in A^{\mathcal{A}}$$

实际上，这对应于一种概率分布，其中价值较高的动作获得更大的权重，而 κ 是一个参数，可用于控制攻击者选择非最优动作的可能性。

攻击者会试图最大化其获得的预期累积奖励，而防御者的目标是引导攻击者进入某个陷阱状态，同时尽量减少攻击者可能获得的总奖励（毕竟，攻击者获得奖励可能意味着他们能够破坏目标环境中的运行，或以其他方式造成干扰）。防御者可以执行两种类型的动作。第一种是空操作，即防御者不改变环境，并允许攻击者执行其选择的动作。第二种是由防御者选择将攻击者动作的哪些潜在结果应用于环境。在我们的示例中，这相当于决定攻击者旁边的哪扇门应该打开。

此外，我们还会限制防御者选择某一特定结果的次数。除了限制防御者能够采取的动作外，这也反映了这样一个事实：如果防御者在某个状态下，针对攻击者的某一特定动作不断选择同一结果，攻击者可能会怀疑自己对模型的认知是否有误。在贯穿全文的示例中，考虑攻击者的起始位置。攻击者旁边有三扇可能的门，根据防御者策略，防御者选择打开门 D3。如果攻击者决定就待在那个房间里，他们会预期下一回合又会有一扇门打开。假设无论攻击者在房间里停留多少次，防御者都一直打开同一扇门。即使攻击者最初将这种情况归因于环境的随机性，如果总是选中同一结果，他们也可能开始怀疑自己对环境的了解。信念的这种变化可能会影响攻击者的行为，进而降低该方法的有效性。

因此，我们将把防御者的动作限制在前 K步内。第 4.1 节将更正式地描述这一现象，以及我们如何计算 K 的悲观下界。在定义好这些组成部分后，我们现在可以如下定义我们的核心问题：

定义 1. 对于一个使用任务 $\mathcal{M}^{\mathcal{A}}$ 模型运行的给定攻击者，在一组陷阱状态 S_T 构成的环境中，攻击者在这些状态下获得的奖励为零，预算为 K 的攻击者诱捕问题对应于这样一个问题：防御者选择一个由 K 个动作组成的序列（可以是不操作或结果选择动作），从而使攻击者的累积奖励最小化。

本文的主要观点是，在这种设定下，我们可以将最优攻击者诱捕问题编码为另一个马尔可夫决策过程。求解该马尔可夫决策过程将自动为防御者制定策略，使其观察攻击者采取的动作，并决定如何塑造结果，从而获得对防御者而言尽可能好的结果。这个新的马尔可夫决策过程同样采用无限时域折扣马尔可夫决策过程的形式。不过，此时状态空间由攻击者状态、攻击者正在采取的动作以及剩余预算构成。如前所述，动作包括一个空操作，以及一组可用于选择呈现哪种结果的动作。每个动作（包括空操作）都会导致预算减少。状态中的攻击者动作部分由攻击者在该状态下选择相应动作的可能性决定（由 {v1} 定义）。

当预算降至零的所有状态，以及作为陷阱状态的状态，都会被视为吸收状态。就奖励函数而言，对于每个非吸收状态，防御者获得的奖励对应于攻击者所获奖励的相反数。对于吸收状态，防御者获得零奖励，而其价值则为攻击者模型最优值的相反数。这意味着，系统会尝试将状态引向某个陷阱状态，同时尽量经过对攻击者而言奖励较低的状态。如果做不到这一点，目标则是到达攻击者价值最低的状态。同样，所有防御者行动都受预算限制。

更形式化地说，防御者模型定义如下

$$\mathcal{M}^{\mathcal{D}}_{(S_T,K)} = \langle S^{\mathcal{D}}, A^{\mathcal{D}}, T^{\mathcal{D}}, R^{\mathcal{D}}, \gamma^{\mathcal{D}}, S^{\mathcal{D}}_0 \rangle,$$

其中，$S_T \subseteq S^{\mathcal{A}}$ 是陷阱状态集合。这里，状态表示为 $S^{\mathcal{D}} = S^{\mathcal{A}} \times A^{\mathcal{A}} \times I^K$，满足 $I^K \subseteq \mathbb{Z}$；动作表示为 $A^{\mathcal{D}} = \{noop\} \cup a^{\mathcal{D}}_s | s \in S^{\mathcal{A}}$。令$\hat{s}$ 表示攻击者打算移动到的状态。对于每个状态（s，a，k）$\in S^{\mathcal{D}}$，其中 $s \notin S_T$ 且 $k \neq 0$，转移概率定义为：

$$T^{\mathcal{D}}((s,a,k), noop, (s',a',k')) = \begin{cases} T^{\mathcal{A}}(s,a,s') \times P^{\mathcal{A}}(a'|s') \\ \quad \text{if } k' = k-1 \\ 0 \text{ otherwise} \end{cases}$$

对于非空操作动作，转移概率定义为：

$$T^{\mathcal{D}}((s,a,k), a^{\mathcal{D}}_{\hat{s}}, (s',a',k')) = \begin{cases} 1 \text{ if } k' = k-1 \text{ and } \hat{s} = s' \\ 0 \text{ otherwise} \end{cases}$$

对于吸收状态，所有转移概率均为零。现在来看奖励：对于所有非吸收状态$R^{\mathcal{D}}$（ s $\in$ S_T $R^{\mathcal{D}}$ s a k $= 0$ k ，，k）$= -1 \times R^{\mathcal{A}}$（），而对于状态，（（，，））的所有情形，以及对于所有满足$k = 0$且该状态不是陷阱状态的状态 $R^{\mathcal{D}}$（ $-1 \times \mathcal{Q}^{\mathcal{A}*}$ s a ，，0）（，）。允许奖励值等于智能体最优 Q-函数值的相反数，就能体现出这样一个事实：一旦防御者停止在环境中行动，攻击者便可以自由地获得可能的最大价值。这会进一步激励防御者将攻击者引向价值较低的状态，尤其是在防御者无法将攻击者引入陷阱状态时。折扣因子与攻击者的相同。我们也可以为防御者选择的每个动作指定成本，并将非空操作动作限制在某些状态的子集中，但为使表述简洁，这里不作展开。

值得注意的是，上述表述将预算作为状态的一部分，而不是考虑有限时域马尔可夫决策过程。这是因为在有限时域马尔可夫决策过程中，与状态相关的奖励并不取决于时间步。然而，我们希望在防御者预算为零的状态下，将奖励设为攻击者 Q-函数值的相反数。目前尚不清楚如何通过有限时域马尔可夫决策过程轻松实现这一点。

还值得注意的是，视具体情形而定，防御者决定结果的能力可能仅限于某些状态或动作。或者，强制实现某个特定结果可能需要付出一定代价。上述表述很容易修改，以纳入此类考量。不过，为了使表述更简洁，并有可能推导出关于防御者行为的保证，我们将采用限制较少、也更简单的设定。具体而言，我们可以利用这一表述，建立防御者马尔可夫决策过程与攻击者马尔可夫决策过程的最优价值函数之间的关系。

命题1. 对于任意给定的状态 $s \in S^{\mathcal{A}}$，任一对应状态 (s,a,k)的最优价值为 $|\mathcal{V}^{\mathcal{D}*}((s,a,k))| \leq \mathcal{Q}^{\mathcal{A}*}(s,a)$，对于所有攻击者动作 a 和当前预算值 k 均成立。

证明. 我们可以通过归纳法轻松证明这一关系，具体来说，证明一个更强的关系，即 $\mathcal{V}^{\mathcal{D}*}$ s a k $\geq$((,,)) $-1 \times \mathcal{Q}^{\mathcal{A}*}$ s a (,)。这等价于指出 $-1 \times \mathcal{V}^{\mathcal{D}*}$ s a k $\leq \mathcal{Q}^{\mathcal{A}*}$ s a $\mathcal{V}^{\mathcal{D}*}$ s a ≤ 0 ((,,)) (,)，并且由于 (,)（这是关于原始 $R^{\mathcal{A}}$ 的假设所导致的结果），考虑绝对值时，关系 $|\mathcal{V}^{\mathcal{D}*}((s,a,k))| \leq \mathcal{Q}^{\mathcal{A}*}(s,a)$ 成立。当预算为 时，关系 $\mathcal{V}^{\mathcal{D}*}$ s a k $\geq$((,,)) (,) 显然成立。$-1 \times \mathcal{Q}^{\mathcal{A}*}$ s a $k = 0$ 假设该关系对预算为 $k = i$ 的任意状态均成立，接下来考虑预算为 $k = i+1$ 时某个状态的价值。根据贝尔曼方程，该价值定义为

$$\mathcal{V}^{\mathcal{D}*}((s,a,i+1)) = R^{\mathcal{D}}((s,a,i+1)) + \max_{a_{\mathcal{D}}} \Big(\sum_{(s',a',i)} (T^{\mathcal{D}}((s,a,i+1)), a_{\mathcal{D}}, (s',a',i)) \times \mathcal{V}^{\mathcal{D}*}((s',a',i)) \Big)$$

现在我们知道，该关系对于预算为 i $\mathcal{V}^{\mathcal{D}*}$ s' a' i $\geq -1 \times \mathcal{Q}^{\mathcal{A}*}$ s' a' $\geq -1 \times$ 的状态成立，因此 $\mathcal{V}^{\mathcal{A}*}$ s' ((,)) ((,)) (())，因此

$$\mathcal{V}^{\mathcal{D}*}((s,a,i+1)) \geq R^{\mathcal{D}}((s,a,i+1)) + \max_{a_{\mathcal{D}}} \Big(\sum_{(s',a',i)} (T^{\mathcal{D}}((s,a,i+1)), a_{\mathcal{D}}, (s',a',i)) \times -1 \times \mathcal{V}^{\mathcal{A}*}(s') \Big)$$

现在，对于具有最大价值的 $a_{\mathcal{D}}$，我们有两个可能的候选项。它要么是对某个结果的确定性选择，要么是一个空操作动作。如果是前者，Bell 算子会选择具有最低 $\mathcal{V}^{\mathcal{A}*}$ 的状态，设为 $\bar{s}$，因此我们有

$$\begin{aligned} \mathcal{V}^{\mathcal{D}*}((s,a,i+1)) &\geq -1 \times (R^{\mathcal{A}}(s) + \mathcal{V}^{\mathcal{A}*}(\bar{s})) \\ &\geq -1 \times \mathcal{Q}^{\mathcal{A}*}(s,a) \end{aligned}$$

其中，

$$\mathcal{Q}^{\mathcal{A}*}(s,a) = R^{\mathcal{A}}(s) + \sum_{\hat{s}} \mathcal{V}^{\mathcal{A}*}(\hat{s}))$$

这一事实

$$-1 \times (R^{\mathcal{A}}(s) + \mathcal{V}^{\mathcal{A}*}(\bar{s})) \geq -1 \times \mathcal{Q}^{\mathcal{A}*}(s,a)$$

源于这样一个事实：

$$R^{\mathcal{A}}(s) + \mathcal{V}^{\mathcal{A}*}(\bar{s}) \leq \mathcal{Q}^{\mathcal{A}*}(s, a)$$

这反过来是由以下事实得出的：对于所有其他可达状态，$\mathcal{V}^{\mathcal{A}*}$ 的最低价值将小于 $\mathcal{V}^{\mathcal{A}*}$的加权和。这证明了当提供最大价值的动作是结果选择动作时，所需关系成立。对于空操作，$\mathcal{V}^{\mathcal{D}*}((s, a, i+1))$的贝尔曼方程可直接映射到 $-1 \times \mathcal{Q}^{\mathcal{A}*}((s, a))$，因为该动作产生的状态转移与攻击者模型中本应发生的状态转移相同。这证明了原始结果。 □

请注意，对于从给定状态 s开始并执行动作 a的攻击者，其可获得的期望价值将以上界 $|\mathcal{V}^{\mathcal{D}*}((s, a, K))|$为限。这一直观地源于以下事实：对于所有非终止状态，$\mathcal{M}^{\mathcal{D}}$中与某一状态相关联的奖励恰好是攻击者所获奖励的相反数。对于终止状态，如果它是陷阱状态，攻击者实际上将获得零奖励。如果终止状态对应于预算为零的状态，那么防御者获得的奖励就是攻击者所执行动作对应的最优 Q-函数价值的相反数。这是从该状态开始攻击者可能获得的价值的一个上界。上述命题表明，利用上述公式推导出的最优策略绝不会导致攻击者获得更高的价值。事实上，正如我们将在评估部分看到的，我们的方法通常能够非常有效地降低攻击者获得的总期望价值。

4.1 预算

先前讨论的公式永远不会生成在攻击者所使用的模型下不可能出现的状态—动作序列。理论上，无论防御者作出何种选择，攻击者都不会对所观察到的结果感到意外。然而，我们将防御者的动作限制在前 K 步之内，这反映出这种情况在实践中几乎不可能成立。更现实的假设是，攻击者可能对自己关于环境模型的信念仍存在某种不确定性。为了近似攻击者关于模型的信念可能如何演化，我们将假设攻击者是贝叶斯推理者；即使攻击者可能是人类，这一假设也很合适。在这一近似下，攻击者将对环境保留两个假设：要么环境模型与 $\mathcal{M}^{\mathcal{A}}$ 相同，要么他们完全不理解系统动力学。实际上，第二个假设对应于这样一种情况：人类相信底层模型可能是 $\mathcal{M}^{\mathcal{A}}$所指定模型以外的任何模型。这与新物种归纳（Zabell 1992）和开放世界推理（Senator 2019）等问题密切相关。然而，正如（Sreedharan等人 2021）等研究所表明的，可以使用高熵模型作为第二个假设的基础来近似这一点。在第二个假设下，人类相信，在执行任意动作时，从任意给定状态都可能转移到任何状态。沿用先前的文献，我们将这一模型表示为 $\mathcal{M}^0 = \langle S^{\mathcal{A}}, A^{\mathcal{A}}, T^0, R^0, \gamma^{\mathcal{A}} \rangle$，使得对于所有 s，$s' \in S^{\mathcal{A}}$ 和 $a \in A^{\mathcal{A}}$，我们有

$$T^0(s, a, s') = \frac{1}{|S^{\mathcal{A}}|}$$

由于我们关注的是最坏情况，因此假设攻击者完全不偏不倚，对模型的信念在这两个假设之间均匀分布。请注意，假设攻击者是在这种不确定性下运行的智能体，并不会使攻击者的策略偏离前文定义的策略，因为在 $\mathcal{M}^0$ 下，所有策略都是等价的。因此，攻击者的动作由 $\mathcal{M}^{\mathcal{A}}$决定。

防御者的目标是阻止攻击者增加对“自己对模型的理解有误”这一假设的信念权重。如果攻击者在到达陷阱状态之前开始相信自己可能对模型的理解有误，就可能采取纠正行动或规避行动。可以采用信念马尔可夫决策过程的形式化方法（Kaelbling、Littman 和 Cassandra（1998）），追踪攻击者的精确信念。不过，这种情境也提供了利用更高效形式化方法的机会，同时仍能保证防御者不会让攻击者察觉到自己对底层模型的理解是否有误。预算表示最坏情况下，从初始状态出发，攻击者的信念可能转向 $\mathcal{M}^0$之前能够发生的最少状态转移次数。更正式地说，预算 K 可定义如下：

定义 2. 对于给定的一对攻击者模型 $\mathcal{M}^{\mathcal{A}}$ 和$\mathcal{M}^0$，如果满足以下条件，则防御者行动预算为 K：

1. 存在一条长度为 K（即 $|\tau| = K$）的轨迹 $\tau = \langle s_0, a_0, ..., s_K \rangle$，使得 $s_0 = S_0^{\mathcal{A}}$, $P(\tau|\mathcal{M}^{\mathcal{A}}) > 0$，且 $P(\tau|\mathcal{M}^{\mathcal{A}}) < P(\tau|\mathcal{M}^0)$*2.* 并且，不存在另一条长度小于 K 、满足相同条件的轨迹 τ'。

也就是说，K 对应于在 $\mathcal{M}^{\mathcal{A}}$下仍然可能实现的最短轨迹；相比 $\mathcal{M}^{\mathcal{A}}$，$\mathcal{M}^0$ 或许能更清楚地说明这一点。将防御者的动作限制为长度小于 K，意味着防御者永远无法选择一条通往陷阱状态的路径，而这条路径可能会动摇攻击者的信念。（这甚至包括在防御者选择执行空操作的状态中可能发生的最坏情况状态转移。）

现在的问题是如何计算这个预算 K。一种简单的方法是将其转化为一致代价搜索问题。具体来说，我们从初始状态开始搜索。在任一状态下，搜索都可以选择从该状态出发的多种转移之一，始终扩展长度最短的前缀；当我们得到一条在 $\mathcal{M}^0$ 中比在 $\mathcal{M}^{\mathcal{A}}$ 中更可能出现的轨迹时，就结束搜索。某些情况下，这样的上界可能不存在，或者长到对防御者没有实际意义。因此，如果我们知道 K 会超过某个预先设定的限值，就可以停止搜索。我们可以通过在不受预算限制的情况下求解防御者马尔可夫决策过程（$\mathcal{M}^{\mathcal{D}}$），并修改奖励函数，使其只反映攻击者到达陷阱状态前所需的步数，然后查看最优策略，来确定这个限值。这会给出防御者为获得最优结果所需步数的期望

折扣步数；如果 K 超过该限值，那么它就无关紧要了。如果我们还能保证马尔可夫决策过程中从每个状态都至少有一个陷阱状态可达，就可以将折扣因子设为 1，以更准确地估算步数。

5 实证评估

我们在这些评估中的主要重点，是在一些标准的马尔可夫决策过程规划基准上评估所提出方法的计算特性。我们选择了网格世界领域的几种常见变体，以考察随着规模和陷阱位置的变化，我们的方法表现如何。我们考虑了以下领域。

1. 网格世界：这是基本的网格世界领域。智能体的目标是从固定的初始状态移动到目标位置。通往目标的路径可能被墙壁和熔岩坑所在的格子阻挡。我们在四种不同实例的网格世界上测试了该方法，尺寸分别为 4×4、6×6、8×8 和 9×9。对于每个网格实例，我们使用在不同随机位置生成的陷阱状态，对该方法进行了五次测试。该领域的随机性由滑动概率控制，该概率会使智能体以较小概率向不同方向移动。2. 四房间：这是网格世界的一种常见变体，智能体必须通过连接各个房间的门在不同房间之间移动。这一领域在分层强化学习文献中尤为常见（Sutton、Precup 和 Singh，1999）。与前一种情况一样，我们使用了四种不同尺寸的实例：4×4、6×6、8×8 和 9×9。每种实例都再次使用随机生成的陷阱状态进行了五次测试。有趣的是，我们发现，所用的四房间问题生成器的具体实现所生成的迷宫实例中，当尺寸为 4×4 和 8×8 时，攻击者无法到达目标位置。该领域的随机性来源与原始网格世界相同。3. 接下来，我们考虑了 Smith 和 Simmons（2012）提出的岩石采样领域的马尔可夫决策过程版本。智能体是一辆漫游车，需要在网格各处收集优质岩石样本。我们仍然考虑相同的四种尺寸的网格。每种实例都使用随机生成的陷阱进行了五次测试。该领域的随机性来源与原始网格世界相同。原始领域中，采样劣质岩石会获得负奖励；本实验中我们省略了这一设定。4. 最后，我们考虑了一个水坑领域（Boyan 和 Moore，1994），智能体需要在避开水坑的同时到达目标状态。我们的实现使用了尺寸固定为 1×1 的地图，但允许智能体改变每一步的步长。我们考察了步长为 (δ) 0.2、0.3、0.4 和 0.5 的情况。每种步长都在随机选取的陷阱位置上进行了五次测试。该领域的随机性来源同样是滑动概率。在原始任务定义中，穿过水坑会导致负奖励。不过，为确保智能体只获得非负价值，

我们为所有在水坑外发生的动作添加了正奖励，而在水坑内发生的动作则获得零奖励。虽然这可能会改变智能体的最优行为，但该领域仍然是一个可行的测试领域，因为防御者仍可利用领域中的随机性，引导智能体进入特定的陷阱状态。

我们使用价值迭代作为规划方法，并采用 Simple RL 框架提供的实现（Abel 2019）。我们还使用了该框架提供的任务实现，仅进行了少量修改，以确保攻击者的奖励始终为正，并确保各领域中的随机性来源具有对称性。我们使用一致代价搜索来计算防御者预算。为简化预算计算，我们将防御者预算的上限设为 15 步。如果搜索考察的路径长度超过 15 步，搜索将直接返回 15 作为防御者预算。虽然这对可能的防御者预算设置了一个简单的限制，从而确保预算搜索能在有限时间内结束，但也可以使用第 4.1 节所讨论的方法计算特定于任务的预算上限。

表 1 总结了实验结果。所有报告的数值均为五个实例的平均值。所有实例的滑动概率均保持为 0.5。所有实验的超时时间均设为 30 分钟。测试所用系统配备 2GHz 四核英特尔酷睿 i5、16 GB RAM、3733 MHz 和 LPDDR4X 内存。

主要比较指标是没有防御者时攻击者获得的价值，以及防御者获得的价值。防御者进行干预时，攻击者实际获得的价值等于防御者价值的绝对值。我们的方法几乎在所有情况下都降低了这一数值。仅有两个实例不符合这一规律，因为攻击者的价值本来就为零。

价值为零的实例，是指防御者策略能够保证攻击者被引入陷阱状态的实例。这种情况主要出现在岩石采样领域和水坑领域。防御者能否诱导出可保证困住攻击者的策略，取决于多个因素，包括从攻击者起始位置到陷阱状态的可达性、攻击者可能采取的策略、领域中的潜在随机性（使防御者能够掩饰其动作），以及防御者可用的预算。我们发现，障碍物较少且攻击者无法穿越的领域中，更常出现能够保证困住攻击者的防御者策略。在水坑领域，即使是水坑也不会阻止智能体穿过，只会使其获得较少的奖励。同样，在岩石采样领域，智能体可以自由地在地图中移动。

我们感兴趣的另一个测量因素，是防御者生成策略所需的时间。理论上，防御者生成的策略会考虑攻击者可能表现出的所有行为。在我们当前的 formulation 中，防御方规划时间并不是主要瓶颈。不过，这仍然是一个值得研究的问题，因为我们的方法依赖于一种相当复杂的编译过程，该过程会大幅

问题实例 Task	实例	攻击者平均价值	防御者平均价值	平均防御者规划时间（秒）
网格世界	4×4	0.94	-0.32	0.46
	6×6	0.89	-0.43	5.19
	8×8	0.83	-0.33	74.68
	9×9	0.83	-0.24	94.55
四房间	4×4	0	0	0.013
	6×6	0.85	-0.24	21.35
	8×8	0	0	1.085
	9×9	0.78	-0.48	62.51
岩石采样	4×4	735.46	0	14.87
	6×6	868.91	0	272.68
	8×8	-	-	-
	9×9	-	-	-
水坑	$\delta = 0.2$	529.08	0	0.14
	$\delta = 0.3$	534.42	0	0.07
	$\delta = 0.4$	547.72	0	0.04
	$\delta = 0.5$	547.72	0	0.04

表 1：所提出算法在若干标准 MDP基准上的实证评估

增加问题实例的状态空间。结果表明，在除 Rock Sampling 之外的大多数领域中，系统都能够在合理的时间内生成策略。对于 Rock Sampling，我们的方法在大小为 8×8 和 9×9 的网格上超时。不过，我们当前使用的规划框架（即 Simple RL）会给我们的方法增加额外开销，而在许多马尔可夫决策过程规划设置中，这种开销可能是可以避免的。Simple RL 框架侧重于使用任务的生成式规范，并执行单独的采样过程，以创建对显式转移函数的估计。在模型预先给定的情况下，我们可以避免执行此类模型估计操作。正如预期的那样，对于规划器能够成功完成的问题，所需时间会随问题规模的增大而增加。具体而言，防御者模型的构建时间随状态空间规模呈多项式增长。在水坑领域中，值得注意的是，状态空间的有效规模会随步长增加。因此，对应于步长 $\delta = 0.2$ 的问题是规模最大的问题，而对应于 $\delta = 0.5$ 的问题规模最小。表 2 展示了在每个领域所使用的最大环境规模下构建模型和制定计划所需时间的详细分解（网格世界：9x9，四房间：9x9，岩石采样：6x6，水坑：0.2），预算为 15。这些测试使用的系统配备了 2.4GHz 四核英特尔酷睿 i7 处理器、4GB 1333MHz DDR3 内存。虽然我们当前的实现兼容离线决策，但该项目未来的目标之一可能是使用更快的在线规划器来生成防御者策略。所有实验所使用的代码都可在 https://github.com/blcates/AttackerEntrapment 找到。

6 相关工作

大量文献（Liu 等，2021；Masters、Kirley 和 Smith，2021；Masters 等人，2020；Masters 和 Sardina，2017）

领域	耗时（秒）	
	模型构建	规划
网格世界	22.365	44.694
四房间	14.927	35.693
岩石采样	23.774	5.431
水坑	131.280	176.203

表2：模型构建与规划时间

围绕欺骗性路径规划展开，涉及智能体对其目标或路径进行隐蔽，以误导被动观察者。我们的方法中所描述的防御者将这一问题完全颠倒过来。它不再是被动观察者，而是试图欺骗另一个智能体的主动观察者。在关于智能体可解释性的综述中，Chakraborti 等人（2019）推测，一个完全主动的观察者可能拥有“自己的目标和动作，甚至能够协助或阻碍智能体实现其目标”。后者正是我们的防御者所实现的功能。据我们所知，这是首项实现主动观察者智能体的工作：该智能体能够施加影响，而被其欺骗的智能体甚至不知道它的存在。

欺骗（Bell 和 Whaley 1982）要求欺骗者能够影响目标的动作，并且目标有可能误判自身所处的情境（Davis 2016）。Masters 和 Sardina（2017）区分了使用伪装的欺骗与使用隐瞒的欺骗。伪装通常被描述为“展示虚假信息”。相比之下，隐瞒则被描述为“隐藏真实情况”。换句话说，说谎是一种伪装，而隐瞒式谎言则是一种隐瞒。防御者只使用隐瞒。它不会明确传达虚假信息，例如让攻击者相信

目标位于其他位置。它只是让攻击者继续保持这样一种错误信念：环境并未受到不可见对手的控制。

防御者还可以根据 Masters、Kirley 和 Smith（2021）提出的更具体的人工智能欺骗类别进一步分类。它采用计算式欺骗策略，即欺骗者掌握更多关于环境的信息，并利用这些信息获得相对于对手的优势。在这种情况下，知识不对称体现在：攻击者不知道防御者的存在，也不知道陷阱状态的存在。攻击者浑然不知，防御者正引导其移动至同样未知的陷阱状态。

将防御者的方法与其他欺骗性路径规划智能体的方法进行比较很有意义。Kulkarni、Srivastava 和 Kambhampati（2019）通过搜索节点来为对抗性路径规划场景选择混淆路径；每个节点都表示一个状态及对手的信念状态。路径的选择部分取决于其对对手信念状态的影响。具体而言，规划智能体旨在选择一条不会让对抗性观察者得知其目标的路径（Kulkarni 等，2018；Keren、Gal 和 Karpas，2016；Bernardini、Fagnani 和 Franco，2020；Kulkarni、Srivastava 和 Kambhampati，2020）。同样，我们的防御者智能体通过寻找一条通往陷阱状态、且不会违反攻击者对环境规则之模型的最短路径，来计算其预算以及可执行的动作数量。

还可以将其与 Chakraborti 等人（2015 年）和 Fern（2014 年）的研究进行比较。这两项研究中的规划智能体都认为，其环境中的人类并不知道这些智能体的存在。与防御者不同，这些智能体在合作场景中运行。因此，人类的无知并不是该任务的必要特征。在 Fern（2014 年）的研究中，这是一项简化假设，作者利用该假设生成更有效的解决方案。无论情形是协作性的还是对抗性的，假设人类不知道 AI 的存在，其结果都是天真的智能体不会因 AI 的存在而改变自身决策。这一简化使攻击者诱捕问题能够作为马尔可夫决策过程加以求解，而不必将其作为更复杂的博弈论问题处理。在未来的工作中，将防御者的表现与依赖博弈论解决方案的同类模型进行比较，可能会很有意义（Brafman 等人，2009 年；Speicher 等人，2018 年；Tambe，2011 年）。

7 结论与讨论

本文的主要贡献是提出了攻击者诱捕问题：隐蔽防御者会挫败攻击。在该问题中，两个智能体均以马尔可夫决策过程建模，并选择能够最大化自身价值的动作。防御者通过引导攻击者选择会使其自身价值最小化的动作来最大化自身价值，理想情况下最终将攻击者引入陷阱。防御者只选择不会让攻击者察觉其存在的动作，从而利用对手对环境的无知，挫败毫不知情的入侵者。

通过在四个马尔可夫决策过程基准领域中展示这种交互，我们证明，所述防御者智能体能够有效降低攻击者获得的价值。我们还提出了一种方法，用于计算悲观下界，即防御者在不改变攻击者对环境的认知的前提下，能够采取的中间动作数量下限。

未来的研究可以探索不同的规划策略。当前方法会计算每个可能的攻击者动作所对应的所有可能状态；对于大型状态空间，这种做法会慢得难以接受。下一步，我们希望研究在线规划器和近似规划器的实用性，例如基于蒙特卡洛树搜索的规划器变体。在线规划能力将使我们的方法能够更广泛地应用于各种对抗场景。

以往研究将欺骗性路径规划描述为智能体欺骗被动观察者的过程（Kulkarni、Srivastava 和 Kambhampati，2019；Ornik 和 Topcu，2018）。观察者并不了解智能体的目标或计划，因此必须通过推断对手动作所体现的意图，持续更新自己的信念。未来研究中一个有趣的问题，是用我们的防御者智能体替换这些场景中的被动观察者。可以对防御者进行改进，使其能够估计攻击者的信念状态，并根据攻击者的动作更新估计。防御者仍然可以隐藏自己的存在，但若能处理不断变化的攻击者信念，那么即使其存在暴露，也会更加稳健。防御者甚至可以做出可能违反攻击者对环境规则认知的选择，只要这样最终能带来更理想的结果。

这些变化使我们能够为现实世界场景中的潜在人类攻击者构建更贴近实际的心理模型。要让防御者更有能力应对人类攻击者，下一步可以是利用认知偏差。尽管人类常常认为自己是理性的，但他们并非完全理性；事实上，他们的理性会以可预测的方式失效。这些可预测的缺陷被称为认知偏差，已在心理学文献中得到分类和充分记录。在未来的研究中，我们计划确定如何在规划问题中表示并检测常见的认知偏差。能够利用这些弱点的防御者将能够运用更巧妙、更隐蔽的策略欺骗对手。

我们希望，本文提出的概念能为未来研究开辟富有成效的方向。通过这个简单的例子，我们展示了隐蔽防御者在对抗性路径规划任务中的潜在效用。我们期待进一步研究这一方向，并期待由此接触到的新问题。

致谢

作者谨此感谢 Indrajit Ray博士提供了有益的讨论，为本项目的构想作出了贡献。

参考文献

Abel，D. 2019。《简单强化学习：Python 中可复现的强化学习》。载于 *RML@ICLR*。

Bell, J.B.; and Whaley, B. 1982. 作弊：战争与魔术、游戏与体育、性与宗教、商业与骗局、政治与间谍活动、艺术与科学中的欺骗. 圣马丁出版社。

Bernardini, S.; Fagnani, F.; and Franco, S. 2020. 一种稳健目标混淆的优化方法。载于知识表征与推理原理国际会议论文集，第 17 卷，119–129 页。

Boyan, J.; and Moore, A. 1994. 强化学习中的泛化：安全地近似价值函数。神经信息处理系统进展，第 7 卷。

Brafman, R. I.; Domshlak, C.; Engel, Y.; and Tennenholtz, M. 2009. 规划博弈。载于*IJCAI*，73–78 页。Citeseer。

Chakraborti, T.; Briggs, G.; Talamadupula, K.; Zhang, Y.; Scheutz, M.; Smith, D.; and Kambhampati, S. 2015. 为意外发现进行计划。载于*2015 IEEE/RSJ*智能机器人与系统国际会议 *(IROS)*，5300–5306。

Chakraborti, T.; Kulkarni, A.; Sreedharan, S.; S mith, D. E.; and Kambhampati, S. 2019. 可解释性？可读性？可预测性？透明性？隐私？安全性？可解释的智能体行为的新兴领域。载于国际自动规划与调度会议论文集，第29卷，86–96。

Davis, A. L. 2016. 博弈论中的欺骗：一项综述及多目标模型。技术报告，空军理工学院，赖特-帕特森空军基地，俄亥俄州。

Fern, A.; Natarajan, S.; Judah, K.; and Tadepalli, P. 2014. 协助的决策理论模型。《人工智能研究期刊》，50：71–104。

Jeon, H. J.; Milli, S.; and Dragan, A. D. 2020. 奖励理性（隐式）选择：奖励学习的统一形式体系。载于 神经信息处理系统进展 *33*：神经信息处理系统 *2020* 年度大会（神经信息处理系统大会 *2020*），*2020*年*12*月*6*日至*12*日，线上。

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. 部分可观测随机环境中的规划与行动。《人工智能》，101(1-2)：99–134。

Keren, S.; Gal, A.; and Karpas, E. 2016. 部分可观测环境中的隐私保护计划。载于*IJCAI*，3170–3176。

Keren, S.; Gal, A.; 和 Karpas, E. 2021. 目标识别设计综述。载于第二十九届国际人工智能联合会议论文集，4847–4853。

Kulkarni, A.; Klenk, M.; Rane, S.; 和 Soroush, H. 2018. 资源受限的安全目标混淆。载于*AAAI*秋季研讨会：整合规划、诊断与因果推理。

Kulkarni, A.; Srivastava, S.; 和 Kambhampati, S. 2019. 对抗性与合作环境中的统一规划框架。载于*AAAI*人工智能会议论文集，第33卷，2479–2487。

Kulkarni, A.; Srivastava, S.; 和 Kambhampati, S. 2020. 同时向朋友发信号并对敌人虚张声势：平衡目标混淆与目标可辨识性。载于第十九届自主智能体与多智能体系统国际会议论文集，AAMAS '20，1889–1891。南卡罗来纳州里奇兰：自主智能体与多智能体系统国际基金会。ISBN 9781450375184。

Letchford, J.; 和 Vorobeychik, Y. 2013. 对攻击计划的最优拦截。载于*AAMAS*，199–206。

Liu, Z.; Yang, Y.; Miller, T.; 和 Masters, P. 2021. 用于隐私保护规划的欺骗性强化学习。*arXiv* 预印本 *arXiv:2102.03022*。

Masters, P.; Kirley, M.; 和 Smith, W. 2021. 扩展目标识别：一种用于策略性欺骗的规划模型。载于 第二十届自主智能体与多智能体系统国际会议论文集，871–879。Masters, P.; 和 Sardina, S. 2017. 欺骗性路径规划。载于*IJCAI*，4368–4375。Masters, P.; Smith, W.; Sonenberg, L.; 和 Kirley, M. 2020. 人工智能中的欺骗特征：综述。载于 欺骗性人工智能，3–16。施普林格。Ornik; 和 Topcu. 2018. 最优控制中的欺骗。载于第五十六届阿勒顿通信、控制与计算年会，821–828。Puterman, M. L. 1990. 马尔可夫决策过程。运筹学与管理科学手册，2：331–434。Senator, M. T. 2019. 开放世界新颖性的人工智能与学习科学（SAIL-ON）。Smith, T.; 和 Simmons, R. 2012. 部分可观测马尔可夫决策过程的启发式搜索值迭代。*arXiv* 预印本 *arXiv:1207.4166*。Speicher, P.; Steinmetz, M.; Backes, M.; Hoffmann, J.; 和 K¨unnemann, R. 2018. 斯塔克尔伯格规划：迈向有效的领导者—跟随者状态空间搜索。载于*AAAI*人工智能会议论文集，第32卷。Sreedharan, S.; Kulkarni, A.; Smith, D. E.; 和 Kambhampati, S. 2021. 人机交互中可解释性度量的统一贝叶斯公式化。载于 第三十届国际人工智能联合会议论文集，*IJCAI 2021*，线上活动／加拿大蒙特利尔，*2021*年*8*月*19*日至*27*日，4602–4610。ijcai.org。Sutton, R. S.; Precup, D.; 和 Singh, S. 1999. 马尔可夫决策过程与半马尔可夫决策过程之间：强化学习中时间抽象的框架。人工智能，112(1-2)：181–211。Tambe, M. 2011. 安全与博弈论：算法、已部署系统与经验教训。剑桥大学出版社。Zabell, S. L. 1992. 预测不可预测之事。*Synthese*，90(2)：205–232。