arXiv:2301.05920v1 [cs.CR] 14 Jan 2023

第1章
认知安全的系统科学方法导论

Linan Huang 和 Quanyan Zhu

摘要 人类的认知能力以及对以人为中心的解决方案的需求使人类成为“工业*5.0*”中不可或缺的组成部分。信息物理系统(信息物理系统)，即人机信息物理系统 (人-信息物理系统)，其中融入了 人工智能驱动的技术，以辅助和增强人类。第1.1 节中人类与技术的紧密融合，以及第1.2.4 节中的认知攻击带来了新兴的安全挑战。攻击可能利用人类认知过程中的漏洞，影响其行为，并最终损害HCPS。

防御人-信息物理系统 免受认知攻击，需要一种新的安全范式，我们在第认知安全节中将其称为“1.2.5”。人类认知系统的漏洞及相关利用方法，使认知安全区别于“认知可靠性”，并由此形成独特的*CIA*三要素，分别如第1.2.5.1 节和1.2.5.2节所示。第1.2.5.3 节介绍认知和技术防御方法，以遏制认知攻击杀伤链并实现认知安全。第1.3 节中的系统科学视角为应对认知安全的新挑战提供了一个有前景的方向，可据此开发定量化、模块化、多尺度且可迁移的解决方案。图1.1 展示了第1章的结构。

关键词： 认知安全、认知可靠性、人机信息物理系统、以人为中心的人工智能、认知攻击、认知脆弱性、网络安全、数据科学、系统科学、CIA三要素

1.1 人工智能赋能的人机信息物理系统

信息物理系统（信息物理系统）是“包含物理组件和计算组件之间经工程设计的交互网络的智能系统”，这是美国国家标准与技术研究院（NIST）在 2017 [20]中的定义。尽管信息物理系统的自动化和智能化程度不断提高，人类在完成信息物理系统任务中仍发挥着不可或缺的作用，如图1.2所示的任务栈所示。其快速发展

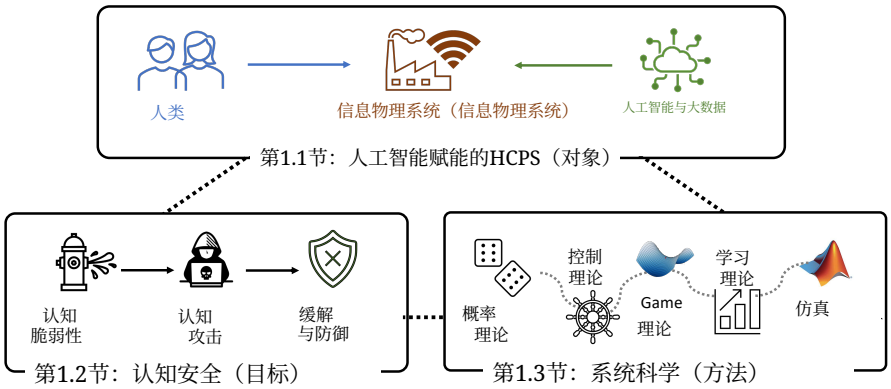

图 1.1：第 1 章的结构：第 人工智能 驱动的 HCPS（第 1.1 节）、认知安全（第 1.2 节）和系统科学（第 1.3 节）分别作为本书的对象、目标和方法。

的人工智能（AIs和大数据促进了人工智能人工智能技术与任务完成的紧密融合，如图人工智能栈所示，见图1.2。我们将在第人工智能节和第1.1.1 节中分别详细阐述人类参与的任务技术栈和1.1.2栈。

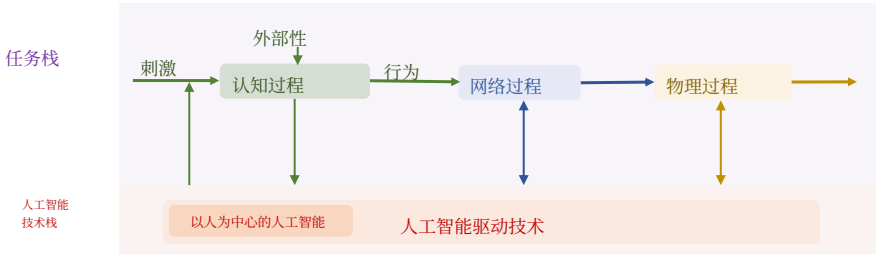

图 1.2：信息物理系统（人工智能驱动技术中的人类角色与信息物理系统 (信息物理系统）。顶部的任务栈展示了一个按任务驱动的流程图，其中人类、网络和物理流程分别以绿色、蓝色和黄色表示。底部的 人工智能 栈展示了人工智能驱动技术如何与人-网-物理过程交互，以协助完成任务。 人工智能 驱动技术的设计应具备可扩展性和可迁移性。此外，为实现以人为中心的目标，这些技术还需经过定制、可解释、合乎伦理、尊重隐私且值得信赖。

1.1.1 任务栈中的人类角色

人类不能也不应在信息物理系统中被取代信息物理系统。一方面，凭借卓越的认知能力和分析能力（例如逻辑推理、符号抽象、知识迁移，以及理解他人的意图和情绪等），人类一直发挥着不可替代的作用，包括确定需求、设计机制和应对突发事件。另一方面，“工业5.0”的必然发展 [9]再次强调了以人为中心的解决方案，以创造个性化的产品、服务和体验。自动化水平的提升旨在支持而非取代人类。

图1.2 概述了人类参与与网络和物理过程在系统层面的集成，这种集成为完成任务而存在，称为人机信息物理系统（HCPS）。人类参与者接收到刺激后，会对其进行处理，并输出相应的行为，从而影响网络与物理过程。认知过程可能受到与个体因素（如个性、意识和专业知识）、环境因素（如工作负荷和压力）以及社会因素（如同伴压力和文化）相关的外部性影响。

人类参与者在完成任务时承担不同角色（例如用户、操作员、安全分析师和管理员），相关的认知过程也可能有不同形式。以下是一些示例。

示例 1 终端用户（包括员工和承包商）使用服务提供商的计算设施，以维持企业的正常运营。例如，对员工而言，刺激可能是工作邮件。他们的认知过程会影响对网络钓鱼[1] 的再认是否准确、及时，从而导致采取安全行为或成为网络钓鱼的受害者。

示例 2 安全分析师会实时调查警报，以进行警报分诊和响应。在这一警报管理场景中，警报、分诊流程和响应分别是刺激、认知过程和行为。

1.1.2 将人工智能技术栈纳入任务栈

.....领域的进步人工智能 加速了信息层和物理层的自动化与智能化进程，如图中的蓝色和黄色双向箭头所示。 1.2。 人工智能已广泛应用于感知、控制、通信和数据处理，服务于生物医学监测、机器人系统和数字孪生等多种应用 [48, 42, 21, 15]。在这些 信息物理系统 应用中， 人工智能 不仅用于推理、规划、学习和处理，还能操控物理对象。例如，自动驾驶汽车

[1] “网络钓鱼”一词由“密码”和“钓鱼”组合而成，用来描述诱骗互联网用户泄露敏感数据（包括密码）的行为。

采用 人工智能感知环境（例如道路状况、天气以及行人和其他车辆的移动），并确定最佳驾驶设置（例如车速、制动和转向）。人工智能驱动的技术在这些 信息物理系统应用中应具备可扩展性 [3] 和可迁移性 [38]。

与 基于技术的 人工智能 设计使人类能够适应技术系统不同，以人为中心的 人工智能 旨在设计能够感知人类认知过程的系统，从而增强人类在快速演变、复杂且不确定的信息物理系统环境中的感知、认知和决策能力。除了 人工智能驱动技术所需的可扩展性和可迁移性等特性外，我们还要求以人为中心的 人工智能 具备定制化、可解释、合乎伦理、尊重隐私且值得信赖等特性。这些要求建立在对人类认知过程的深入理解之上，并通过设计适当的人类辅助技术（例如人机界面 (人机界面）来实现，如图 1.2中的绿色向下箭头和向上箭头所示。

以人为中心的 人工智能设计还应适应信息物理系统在不同功能、特性、要求和约束下的应用。例如，驾驶辅助系统必须具备可解释性，以促进信任，并尽量缩短驾驶员做出决策并采取适当行动所需的时间 [6]。与此同时，电网控制室中的 人工智能赋能的人机交互应具备以下功能 [35]:

- 支持快速调用和关闭，因为操作员在一定时间内能够执行的操作数量有限，尤其是在时间紧迫的环境中。
- 考虑并学习用户的行为和反馈，因为电网操作员是训练有素的专家，能够评估助手的回答并提供反馈。
- 传达用户操作的后果，因为操作员可能无法正确、及时地评估风险。

1.2 HCPS中的认知安全

信息物理系统自涌现以来便一直面临各种攻击的威胁。基于图1.2所示的任务栈和人工智能栈，我们在第1.2.1 节中纳入攻击栈和防御栈，形成图1.3所示的 HCPS 安全四栈剖析。在概览 HCPS 安全之后，我们将在第1.2.2、1.2.3、1.2.4和1.2.5节分别聚焦本书的重点：认知过程、认知脆弱性、认知攻击和认知安全。

1.2.1 攻击与防御栈

图1.3展示了 HCPS安全的垂直剖析，其中橙色和灰色分别表示攻击栈和防御栈。过去十年间，攻击逐渐呈现出目标明确、智能化、多阶段和多步骤的特点。这类攻击的典型生命周期包括以下四个阶段。在侦察阶段，攻击者收集情报，以确定攻击目标并识别目标网络中的漏洞。在规划阶段，攻击者针对选定的漏洞制定策略并选择攻击工具。在执行阶段，攻击者部署恶意软件以取得初始立足点、扩大访问权限、获取凭证并提升权限。最后，在利用阶段，攻击会窃取机密数据、中断网络服务或造成物理损害。

防御体系中的四个灰色方框展示了四种防御行动方案。预防阶段包括在攻击发生前采取的预防性和主动防御方法。入侵防御技术（包括防火墙和非军事区（DMZ））可能无效，尤其是面对高级持续性威胁 高级持续性威胁(高级持续性威胁）等高级攻击时。因此，必须通过入侵检测和响应来防范此类威胁。归因阶段包括事后分析、威胁情报获取和问责系统。

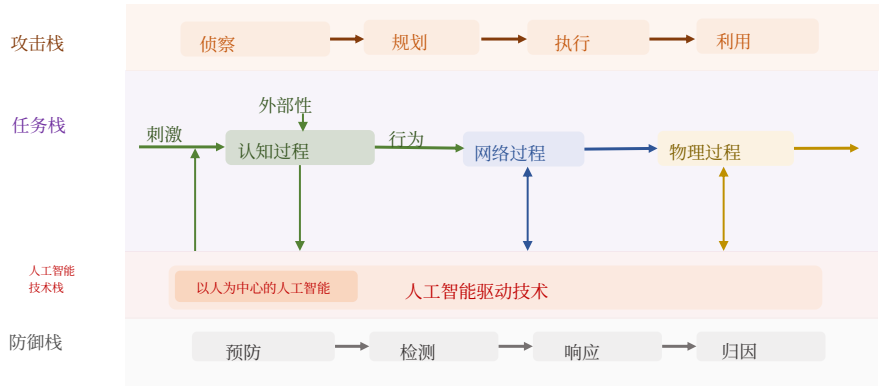

图 1.3：对HCPS 安全性的纵向剖析包括任务和人工智能栈上的攻击栈与防御栈。攻击栈中的四个橙色方框展示了攻击的典型生命周期。首先，攻击者根据侦察结果（例如可利用的人为和技术漏洞），在规划阶段制定攻击策略并选择攻击工具。随后，攻击在执行阶段发起，并分别在执行和利用阶段造成损害。如防御栈中的四个灰色方框所示，典型的防御包含四个阶段：预防、检测、响应和归因。

1.2.2 前馈与反馈认知过程

图 1.4 展示了动态认知过程的放大版本，该过程由前馈和反馈流程组成，分别用实线箭头和虚线箭头表示。我们首先介绍前馈认知过程，其中刺激是输入，行为是输出。首先，感知 [5]会对外部刺激中的感官信息进行解码和收集。由于认知能力有限，人们会运用注意力过滤信息，并将重要或紧急的信息存储在工作记忆中。随后，人类处理信息并进行包括决策、推理和学习在内的心理操作，从而产生行为。根据任务和情境，人类的心理操作可能需要提取存储在长期记忆 (长期记忆)中的过往经历。认知过程还包括以下几个反馈回路：

- 行为会对信息物理系统产生影响，并改变后续刺激。
- 心理操作会学习新知识，并将经验存储在长期记忆中。
- 除了在前馈过程中被动过滤所收集的信息外，注意力还会通过将我们的意识引导至相关刺激，同时抑制干扰信息，主动影响感知。这种选择性注意 [29]已在诸如鸡尾酒会效应 [2]等现象中得到证实。
- 选择性注意是心理操作的结果，通常由目标驱动且为内源性（称为自上而下注意）；相比之下，自下而上注意则由刺激驱动。

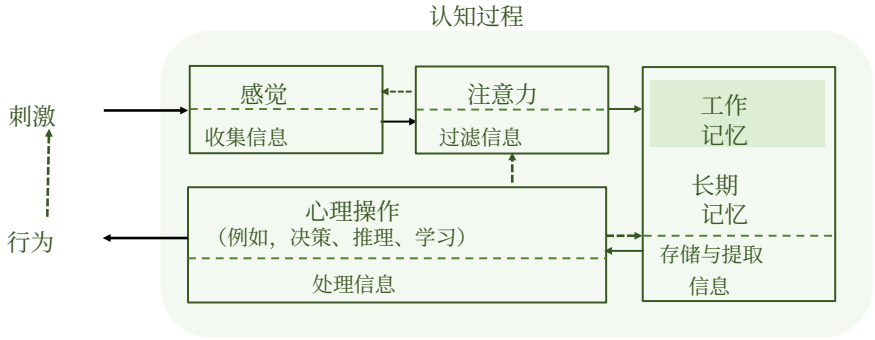

图1.4：认知过程的放大图，包含感觉、注意力、记忆和心理操作。实线箭头所示的前馈路径描绘了收集、过滤、存储、提取和处理的信息流。在反馈回路中，行为会改变未来的刺激，心理操作会影响长期记忆和注意力，而选择性注意则决定感知，如虚线箭头所示。

1.2.3 可利用的认知脆弱性

图中认知过程的每个组成部分1.4 都有攻击者可加以利用的独特漏洞。 [11] 我们将在第1.2.3.1、1.2.3.2、1.2.3.3和1.2.3.4节分别讨论感知、注意力、记忆和心理操作方面的漏洞。

1.2.3.1 感知脆弱性

感知是一个复杂的过程，涉及视觉、听觉、躯体感觉、嗅觉、味觉和前庭系统。这些系统遵循的模式可能被攻击者利用。例如，攻击者了解感知极限（如人类看到和听到信号分别约需 0.3 秒和 1 秒）以及促成因素（如光线、颜色和噪声如何影响人类感知），便能营造容易导致人类出现感知错误的环境。人类感知系统有时会发生感觉扭曲（如视觉和听觉错觉），并产生错误的感官信息。了解这些感觉错觉的促成因素和成因后，攻击者便能制作更难识别钓鱼指标的钓鱼网站和电子邮件。

除了可利用的模式和错觉外，人类感知也容易受到操纵。攻击已采用多种心理学技巧，包括启动效应 [37], 来操纵人-信息物理系统中的感知。启动效应是心理学中一种广为人知的现象，即某种刺激的出现会影响人们对后续刺激的处理或解读方式（例如，与先看到“办公室”一词相比，人们先看到“厨房”一词后会更快认出“食物”一词） [33]。大多数启动效应是正向的，会提高对相关刺激的敏感度，而负启动效应 [51] 则会减慢处理速度。攻击者可以将正向和负向启动效应都武器化，用于实施感知劫持。例如，攻击者可以利用正启动效应在网络钓鱼中突出某些想法，也可以利用负启动效应弱化钓鱼指标。根据刺激能否被有意识地感知，启动效应分别分为阈上和阈下启动效应[16]。由于具有隐蔽性，阈下启动效应可能是实施感知劫持的首选方式。实验结果 [27], 表明，通过巧妙呈现与攻击者偏好的判断在物理或语义上相似的词语或图像，攻击者能够影响检查人员的准确率或虚警率。

1.2.3.2 注意力的脆弱性

注意力可以描述为整体的警觉性水平，或是与周围环境互动的能力 [33]。人类依靠注意力机制来控制大脑中有限的计算资源，以最大限度地获取信息。尽管人类的注意力取得了显著成效，但在多任务处理、持续时间过长、压力、疲劳和认知负荷过重的情况下，其表现会有所下降——

此外，由于注意力具有选择性，能避免我们被无关信息淹没，因此我们可能在空间和时间上出现选择故障。一方面，变化盲视 [47]和 变化聋 [52],的实验表明，当视觉或听觉刺激发生变化时，观察者可能分别无法察觉这种变化。另一方面，注意瞬脱 [46]和 重复盲视 [31]的实验表明，故障也会随时间发生；也就是说，当新信息（即便数量很少）不断到来时，对其进行处理会导致其他信息被遗漏。

攻击者可以利用这些空间和时间注意力漏洞，采取反应式或主动式攻击。反应式注意力攻击利用注意力不集中来逃避检测，并不试图改变人类的注意力模式。例如，许多社会工程(社会工程)和网络钓鱼攻击都是因缺乏注意力而发生的。我们在 [26]中提出了针对反应式注意力攻击的防御框架。相比之下，主动式注意力攻击旨在策略性地影响人类的注意力模式。例如，攻击者可以制造大量佯动，并将真正的攻击隐藏其中，以使人类操作员不堪重负、延迟其响应并降低其判断的准确性。我们将这种新型攻击称为信息拒绝服务(智能拒绝服务)攻击，并在智能拒绝服务中正式描述智能拒绝服务攻击、其影响及相关防御方法 [23]。

1.2.3.3 记忆漏洞

人类记忆依赖大脑中的神经元网络，因此受到容量有限、信息存储和提取速度受限、遗忘以及记忆错误等问题的影响。数字存储设备也存在前两种记忆漏洞，但后两种是人类记忆所独有的。根据 Schacter [45],的说法，后两种分别属于遗漏之罪和错记之罪。

遗忘是指个体短期或长期记忆中已存储的信息自发或逐渐丢失。与数字存储设备不同，人类无法“按需遗忘”；也就是说，即使某些内容已不再需要，它们仍会留存在记忆中 [43]。

记忆错误是指对信息的错误回忆。这可能包括记起从未发生过的事情、错误地指出某段记忆的来源，或编造并未发生的事情。记忆错误部分是由神经元网络的结构以及人类记忆的特点所导致的。如迪斯–罗迪格–麦克德莫特范式 [12],所示，人类会错误地回忆起一个不存在的词，因为它与一组属于共同主题的词相关。许多因素（例如注意力程度、动机、情绪状态以及记忆发生时所处的环境）都会影响人类的记忆。例如，情绪增强记忆效应 [22]表明，相较于中性刺激，人们更容易记住带有情绪色彩的刺激。

人类记忆脆弱性会直接导致安全风险。例如，人们会使用简单且有意义的密码，在不同

的网站重复使用同一密码，甚至把密码写下来以便记住，这会使密码不安全。正如定义1.2将要介绍的，认知可靠性方法，包括单点登录（SSO）[2] [40], 认知密码 [57], 和图形密码 [4], , 旨在缓解与密码相关的记忆脆弱性。

攻击者还可以主动利用这些记忆漏洞，并操纵上述因素来构造攻击向量。例如，攻击者可以降低攻击频率，以利用遗忘脆弱性。如 [44] 和 [30],所示，网络钓鱼事件的频率和可能性降低，会增加受害者遭受网络钓鱼网络攻击的易感性。由于人类记忆具有易受暗示性（即人类倾向于接受他人的建议并采取行动），攻击者可以设计网络钓鱼邮件，通过设计误导性提示来触发记忆错误并植入虚假记忆。此外，他们还可以使用情绪化语言来强化操作员的虚假记忆，促进信任。

1.2.3.4 心理操作脆弱性

心理操作脆弱性主要指各种认知偏差和可被利用的特质。在人类发展的历史进程中，为了在高度动态且不确定的环境中生存，我们形成了认知捷径和偏差，以便快速作出反应，尽管这些反应不够准确或合理。然而，这些认知偏差扩大了攻击面，使人类容易遭受社会工程攻击。我们列举了一些认知偏差及其可能遭受的对抗性利用方式 [41] 如下。

- **锚定效应**：个人的决策会受到某个特定参照点（即“锚点”）的影响。借口型社会工程可以营造情境并建立初始背景，从而增强攻击表面上的可信度，提高成功的可能性。锚定偏差会让人们不再质疑最初的印象，进而接受骗局。
- **框架效应**：同一信息的呈现方式不同（例如，将其描述为损失或收益），个人可能会得出不同的结论。攻击者可以利用框架效应来编写网络钓鱼邮件的内容，并扭曲人类风险感知。
- **乐观偏差**：人们往往会不切实际地乐观，例如高估积极事件发生的可能性，低估消极事件发生的可能性。用户常常认为别人比自己更容易受到影响 [10]。由于他们认为自己不会受到攻击，因此往往会抵制预防性防御措施，例如安装补丁、病毒扫描、清除缓存、输入信用卡信息前检查网站是否安全，或留意鱼叉式网络钓鱼。
- **内群体偏差**：人员是社会性动物，会优待内群体成员，而非外群体成员。攻击者可以伪装成与该群体有关联，以赢得信任并影响群体成员的决策。

[2] 单点登录（SSO）允许用户使用单一身份标识和密码登录多个相关系统，从而减少需要记住的密码总数。

在营销和说服领域，Cialdini [8] 根据实验和实地研究，归纳出六项基于可利用个人特质的影响力原则。攻击者也可以利用以下特质来完成攻陷。

• **互惠**：社会工程 攻击者经常先向受害者提供某些东西，为互惠创造条件。例如，对于事先收到某人礼物的人来说，他们不太可能拒绝对方不恰当的请求。

• **社会认同**：人员很容易受到大群体决策的影响。攻击者可以冒充受害者朋友参与其中，以迫使受害者采取行动。

• **权威**：人员往往会服从权威，攻击者可以通过假装自己是系统管理员或安全专家来加以利用。

• **喜好**：由于影响喜欢自己的人员要容易得多，攻击者可以通过表现得彬彬有礼，并使用目标熟悉的概念、语言和外表来努力博得好感。

• **稀缺**：供应短缺的事物会提升其价值感。社会工程师可能利用稀缺性（例如限时优惠）来营造紧迫感，促使受害者采取行动。

• **承诺与一致性**：一旦人们做出选择，就会受到他人和自身的压力，促使他们坚持这一选择。社会工程 攻击会诱使受害者做出看似无关紧要的承诺，随后逐步升级要求。只要后续要求与先前的承诺一致，人们往往会接受承诺的升级。

1.2.4 认知攻击

在第1.2.3节中，我们讨论了认知过程中四种可被利用的漏洞。认知攻击是指涉及利用这些人为漏洞的对抗性过程，定义见定义1.1。

定义 1.1（认知攻击）认知攻击是一类网络—物理—人类过程，旨在操纵人类行为者的行为以达到恶意目的，包括窃取或破坏机密数据，以及扰乱或误导HCPS 服务，并利用其认知脆弱性。

与网络杀伤链类似[3]，我们在第1.2.4.1节介绍认知攻击的杀伤链。随后，在第1.2.4.2节，我们深入探讨杀伤链的执行阶段，并以社会工程 和智能拒绝服务 攻击为例，说明认知攻击的跨层攻击路径。

[3] 洛克希德·马丁于2011 年开发的网络杀伤链示例见
https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

1.2.4.1 认知攻击杀伤链

图1.5展示了认知攻击的概念性杀伤链，其中蓝色箭头表示六个阶段。我们将这六个杀伤链阶段映射到图1.3 底部所示的攻击栈四个攻击阶段，并在图1.5 中以橙色标出。

在侦察阶段，攻击者会分析从HCPS收集的信息。例如，攻击者可以在侦察期间建立用户或操作员的行为基线。此类基线有助于攻击者识别目标用户的人为漏洞，并在执行阶段加以利用。在规划阶段，攻击者会识别有价值资产，并据此量身定制其战术、技术和流程(TTP)。我们将执行阶段分为三个阶段。首先，攻击者会直接或间接利用人为漏洞；第1.2.4.2节将展示若干利用路径示例。其次，攻击者会监控人类目标的响应（例如情绪和行为），并调整其TTP。例如，对于谨慎和粗心的用户，攻击者可以分别选择提高或降低攻击频率。在找到最佳攻击设置后，攻击者会继续强化认知利用，并将攻陷状态传播给与最初人类目标相关的其他受害者。例如，取得信任后，攻击者可以要求最初的受害者将邮件转发给同事，使同事不再那么怀疑网络钓鱼邮件的合法性。在利用阶段，认知利用开始影响人类行为，随后波及信息层和物理层。

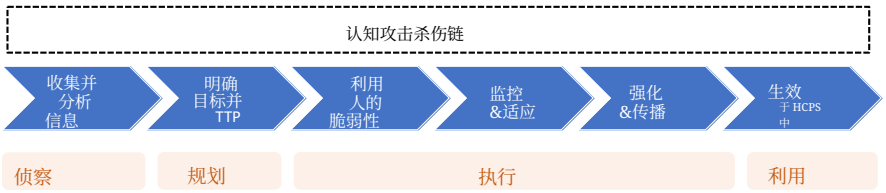

图 1.5：认知攻击杀伤链示意图，涉及侦察、规划、执行和利用四个攻击阶段。在执行阶段，认知攻击利用人为漏洞，监控受害者的响应，并相应地适应利用方式TTP 。随后，攻击者选择最佳攻击设置，以强化攻陷效果，并将其从最初的受害者传播到一群受害者。

1.2.4.2 认知攻击路径示例

利用人为漏洞是攻击执行阶段的关键步骤。这种做法形式多样，可能涉及一种或多种认知脆弱性

第1.2.3节中所述。在图1.6中，我们以社会工程和智能拒绝服务攻击为例，说明认知攻击的各种形式和过程。

社会工程攻击会直接操纵刺激（例如网络钓鱼内容）和外部因素（例如同伴压力），以攻陷人类用户。在这类认知攻击中，认知侵害被用作进入信息物理系统并实施技术侵害的跳板。

然而，在其他认知攻击中，技术侵害也可能成为利用人类认知脆弱性的前提条件，智能拒绝服务攻击的攻击路径便说明了这一点。智能拒绝服务攻击。智能拒绝服务攻击首先制造大量伴攻，触发通过人机交互界面显示的网络安全警报。实时调查这些警报的人类操作员可能会遭受认知过载，导致处理警报的准确性和速度下降。由于人类操作员可能无法响应与真实且严重的攻击相关的警报，这类攻击有可能扰乱网络与物理过程。现实世界中的智能拒绝服务攻击广泛存在，但通常隐含在许多攻击事件中。以下示例1.2.1、1.2.2和1.2.3 中的三个事件利用分布式拒绝服务（DDoS）攻击作为“烟幕”，吸引安全分析师的注意力，同时发动其他隐蔽攻击。

示例 ***1.2.1***（索尼 ***PSN*** 数据泄露事件（***2011***年））2011年4月，安全运营中心（SOC）索尼的安全运营中心正忙于应对DDoS 攻击，因而忽略了机密信息遭到窃取一事。这些信息包括 7700 万名 PlayStation 网络（PSN）客户的姓名、地址、出生日期和密码。由于 PSN 数据泄露事件中的安全故障，英国当局对索尼公司处以近 40 万美元的罚款[17]。

示例 ***1.2.2***（***BIPS*** 比特币失窃事件（***2013***年））作为欧洲主要的商户比特币支付处理商及免费在线钱包服务，BIPS 于DDoS 攻击中成为目标，事件发生于2013年11月15日。当时SOC 正忙于恢复系统上线，攻击者随后发动攻击，获取访问权限并攻陷多个钱包，导致 1,295 枚比特币被盗（约1 百万美元） [49]。

示例 ***1.2.3***（特斯拉勒索软件***2021***年）2021年，针对特斯拉的一次网络安全攻击被挫败。该攻击计划使用DDoS 攻击来转移安全分析师对窃取机密数据的恶意软件的注意力 [50]。

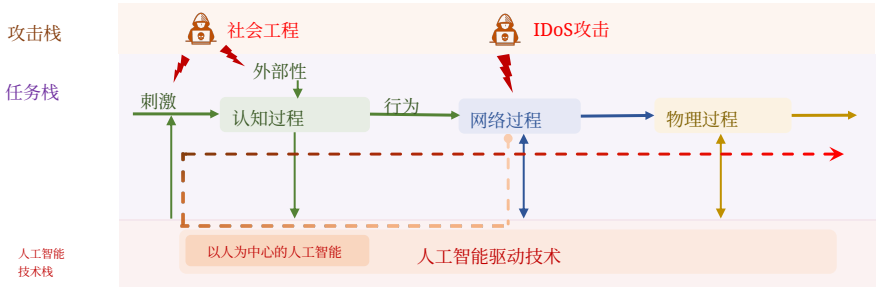

图1.6：虚线箭头描绘了认知攻击的示例路径。社会工程攻击通过改变刺激并影响外部因素，直接利用人类认知脆弱性。智能拒绝服务攻击首先攻陷网络过程，并将技术侵害作为影响人机交互（例如，通过制造佯动来增加操作员的认知负荷），从而间接操纵认知过程。这两类认知攻击都会对网络与物理过程产生影响。

1.2.5 认知安全

信息物理系统人工智能驱动的信息物理系统的开发始于可靠性保证，以确人工智能保系统和人工智能技术在不确定性、扰动和故障下仍能保持可信且高效。攻击的存在要求我们纳入防御措施，由此引出安全性的概念。与信息物理系统中的可靠性和安全性类似，我们在定义信息物理系统的认知可靠性和认知安全，并在定义1.2 和1.3中分别加以区分，从而形成图1.7中的四个象限。

定义1.2（认知可靠性）认知可靠性是指HCPS 在规定条件下（包括充满不确定性、干扰和错误等挑战的环境中），维持运行连续性、完成指定任务并按预期水平提供服务的能力。

定义1.3（认知安全）认知安全是指部署人员、政策、流程和技术，以抵御定义1.1 中的认知攻击，并保护基本的HCPS 组件，包括人类、关键系统结构、服务和敏感信息。

1.2.5.1 认知可靠性与认知安全

根据定义1.2，认知可靠性侧重于人为可靠性分析（人为可靠性分析）和以人为中心的系统设计。一方面，人为可靠性分析能够识别潜在的人为错误事件，评估促成因素，估算这些错误发生的概率，并分析其对信息物理系统的影响。现有多种方法

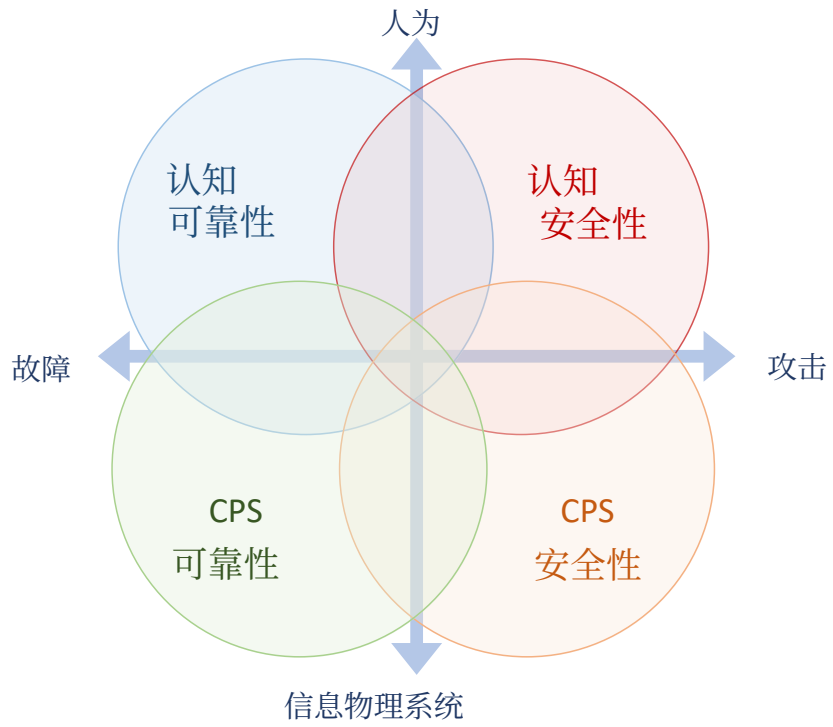

图 1.7：信息物理系统CPSs 和人-信息物理系统中的安全性与可靠性的四个象限。水平箭头根据是否存在攻击，分别区分安全性与可靠性。垂直箭头则根据是否涉及认知脆弱性，区分人-信息物理系统或CPSs 中的安全性（或可靠性）。认知安全扩展了社会工程及其缓解措施，还通过进一步研究对人-信息物理系统的跨层影响，并开发人机协同防御技术来阻止认知攻击杀伤链。

适用于人为可靠性分析，包括基于概率风险分析和认知控制的技术。Di Pasquale 等人 [14]将这两类技术分别称为人为可靠性分析的第一代和第二代方法人为可靠性分析。这些人为可靠性分析技术已广泛应用于生命关键型工业控制系统（例如核电站），以最大限度地减少人类行为对技术系统造成的不利后果。

另一方面，以人为中心的系统设计融合了人体工程学和行为科学，以理解人类认知流程，使系统适应人为因素，从而提升可用性并减少错误行为。例如，根据行为科学的研究发现，再认明显比回忆容易，因此基于文本的质询-响应机制被应用于用户到计算机的身份验证，作为对单靠密码回忆的改进 [57]。其他认知科学研究成果也被用于用户身份验证和电子邮件加密，以提升其可用性

安全性 [39]；例如，图形密码的出现是因为人员对图片的记忆比对文字的记忆更好。

定义 1.3中的认知安全是与第1.2.4节所述认知攻击相关的概念。与认知可靠性中的人为故障不同，认知攻击可以直接利用认知弱点，并将其作为跳板来最大化攻击收益。定义1.2 中“在规定条件下达到预期水平”这一表述表明，认知可靠性针对一组明确规定的条件，以系统效率和可用性为特定目标。在这些条件下的性能使认知功能能够以一定的能力“抵御攻击”或错误。对于超出此能力范围的攻击，则需要采用防御方法提供进一步保护，正如定义 1.3所述。保护方法会以降低效率（例如，入侵检测系统(IDS)的误报干扰正常运行）和可用性（例如，遵守安全规程需要付出额外努力）为代价。旨在增强这一能力的安全设计需要全面考虑安全性、效率和可用性之间的权衡。这一理念引出了系统科学视角，将在第1.3节讨论。

1.2.5.2 认知安全的CIA三要素

在认知安全的语境下，我们讨论机密性、完整性和可用性，即 CIA 三要素，以此为理论基础提供指导。机密性可防止未经授权的用户访问图 1.4所示认知过程中的序列信息流。例如，网络钓鱼攻击会伪装成合法实体并收集机密信息，从而违反机密性。攻击者可以利用这些信息获取访问凭证、窃取数据并部署恶意软件（例如勒索软件）。

认知安全的完整性用于评估图 1.4 所示的认知过程是否受到操纵，从而诱发有偏差的行为。会损害完整性的认知攻击包括错误信息传播和信念操纵技术（例如煤气灯操纵）。

认知安全的可用性确保图 1.4所示认知过程正常运行。智能拒绝服务（IDoS）攻击会耗尽有限的认知资源，包括注意力、记忆和决策能力，从而损害可用性。

1.2.5.3 认知 与技术防御，面向认知安 全性

认知安全综合运用认知与技术防御措施，打破认知攻击杀伤链，如图所示：1.8。由于认知攻击包括跨越信息层和物理层的技术利用，如第1.2.4.2节中 IDoS 攻击路径所示，网络与物理防御是不可或缺的组成部分。

除了网络与物理防御外，认知安全还致力于保护人类和人工智能技术。一方面，用户培训可以帮助用户

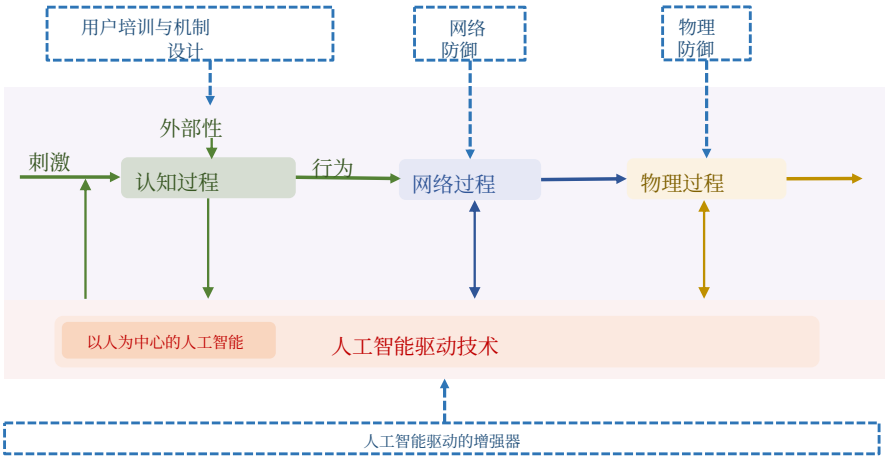

图 1.8：蓝色虚线框中的防御方法可打破认知攻击杀伤链并实现认知安全。网络与物理防御分别作用于网络与物理过程。用户培训和以人为中心的机制设计通过外部性影响认知过程。将人工智能技术栈纳入其中，认知安全还包括人工智能驱动的增强器，以保护人-信息物理系统免受人工智能技术的对抗性利用。

掌握安全知识、增强安全意识并培养批判性思维，从而巩固认知过程。另一方面，我们可以设计机制来避免自私行为，并激励用户采取安全行为。

人工智能技术也可能受到人工智能技术栈中认知攻击的影响。牢记安全要求，我们需要设计一种人工智能驱动的增强器，以确保人工智能驱动技术的安全性和韧性。尤其是，以人为中心的人工智能技术人工智能技术栈不仅需要在增强人类能力时调整认知偏差，还必须能够识别并防御认知攻击。例如，人工智能赋能的安全辅助技术可以增强人机交互，以防御来自智能拒绝服务攻击的威胁。为实现这一目标，我们融合光学、运动、声学信号和电生物信号等多种生物传感器数据，以实时反映当前认知过程的状态。这些信号可以反馈给人工智能增强器进行自适应控制。

我们可以将图1.8 中由人工智能技术增强的人类过程类比为由父母监管的儿童。用户培训和机制设计会教育并激励儿童安全行事，但这些方法受社会和认知局限性的制约。为突破这些限制，人工智能增强器则让父母能够察觉并防范针对子女的潜在攻击。

1.2.5.4 认知安全的范围

为了更好地说明认知安全的范畴，我们通过两个贯穿全文的示例，区分以下与定义 1.3 中认知安全相关的概念。

首先，对人类认知的利用与防御是认知安全的重要组成部分，这与信息物理系统安全领域有显著不同，并开辟了 HCPS 安全的新战场。

其次，如图 1.7所示，认知安全拓展了社会工程 和认知黑客攻击 [11]的概念，它们通常依赖武器化信息并利用心理活动脆弱性。在人际交互中的认知操纵之外，我们还进一步研究其在人机交互过程中对信息物理系统 的影响及其对信息物理系统的依赖。正如智能拒绝服务 示例所示，技术利用会加剧注意力脆弱性。这类直接利用人类认知局限的人机结合攻击，并未纳入社会工程 和认知黑客攻击的考量。

第三，本书所述的认知安全并非旨在通过药物或医疗疗法，在神经科学或生物学层面增强人类。相反，我们着重设计各种机制和辅助技术，以非侵入性方式影响人类认知过程，从外部对其施加影响，改变其行为，并最终提升整个 HCPS的安全性。

第四，认知安全并不旨在回答人因工程方面的问题，例如如何编造令人信服的虚假信息，或如何设计反钓鱼界面。相反，我们在系统层面借鉴行为科学和心理学的理论与实证研究成果。

第五，认知安全的技术解决方案能够识别并适应对认知脆弱性的恶意利用。例如，在面临 智能拒绝服务 攻击的威胁时，仅采用传统的警报分诊方法来间接降低认知负荷可能并不足够。因此，我们进一步有策略地减少警报的显著性，以直接管理注意力，并使警报检查流程与每位操作员的认知负荷相适应。

第六，认知防御是实现认知安全的重要途径，但并非唯一途径。我们可能不需要防御所有类型的认知攻击。只要打断认知攻击的杀伤链即可。此外，由于某些类型的认知攻击无法防御，或防御成本过高，我们可以着重采用补偿与纠正机制，以降低攻击的风险和影响。

最后，本书中的认知安全并不意味着开发能够具备人类认知能力的安全措施（例如认知计算SOC [13],钓鱼检测器 [18], 和事件管理系统 [1]），也不意味着利用人类的认知优势（例如创造力和灵活性）来增强技术防御系统。此外，认知安全并不侧重于提高安全措施的可用性 [19, 39],，尽管这有助于认知安全。

1.3 认知安全的系统科学视角

继第1.1 节和1.2节之后，人类是完成信息物理系统任务不可或缺的组成部分，因此认知安全需要纳入而非排斥人类认知。然而，人类认知具有高度复杂性、不确定性和个体差异，因此引出了以下研究问题：如何将成熟的人为因素理论纳入人-信息物理系统，以用于防御？如何定制防御，使其适应各种安全应用？哪些指标和度量方法有助于量化人为因素对信息物理系统的影响以及防御方法的有效性？

本书采用系统科学视角，包括决策理论、优化、博弈论、人工智能、生理指标和心理学理论，基于以下原因来探讨这些问题。

• 首先，如第1.2.4节所示，认知攻击发生在认知层（例如社会工程）和系统层（例如智能拒绝服务攻击），而非通过医学手段作用于神经元层。因此，需要在系统层研究攻击与防御之间的交互，以理解整个攻击阶段，预判攻击与HCPS之间的交互，并在这些阶段中找到可实施防御的切入点。
• 其次，认知攻击（例如智能拒绝服务攻击）会利用人类、网络和物理过程中的不同漏洞。单一工具不足以遏制、检测和防范这些跨层、跨学科攻击。为应对各种漏洞，我们需要借助跨学科工具，综合多个领域的防御方法。

• 第三，由于认知攻击会影响人类、网络和物理流程，因此量化各个组成部分面临的风险、风险的传播以及整个HCPS [24, 25, 7]所面临的风险十分重要。此外，我们需要设计最优控制措施，以经济有效的方式保护人-信息物理系统。因此，我们需要科学且定量的方法来量化安全指标和界限、评估权衡、描述基本极限，并设计最优控制[53, 36, 54]。

• 第四，由于认知攻击者和防御者都是智能参与者，他们会根据对对方行动的预测采取行动。博弈论自然成为一种工具，可用于捕捉对抗环境中的交互，以进行定量分析和设计 [34, 55, 28, 56]。

• 第五，模型可能无法准确描述HCPS中高度动态且有时不可观测的组成部分（例如人类认知）。生物传感器（例如脑电图（EEG）和眼动追踪设备）收集的数据以及人工智能可以整合起来，根据可测量的观测结果提供系统层面的自适应解决方案。
• 最后，可以采用心理学和社会科学理论来理解感知、注意力、记忆和决策。作为一种多学科方法，系统科学方法整合这些发现和成果，构建全面的数据驱动与模型驱动混合系统框架，应用于人-信息物理系统。

1.3.1 系统科学方法的优势

系统科学方法为认知安全提供了一种新范式，并带来以下优势。

• **涌现**：涌现广泛存在于哲学、心理学和艺术中，“整体并非其各部分之和”[32][4]。如图1.2所示，人工智能与人类在信息物理系统中的紧密融合催生了人-信息物理系统这一新概念。人类、网络、物理和人工智能组件之间的交互产生了新的攻击面和攻击路径，可能会加剧人类认知脆弱性，并推动对新兴认知安全领域的研究需求。

• **黑箱与功能仿真**：系统视角关注系统的输入和输出（或传递特性），将系统本身视为黑箱，而不必了解其内部运作。将人类认知系统视为黑箱，有助于我们关注行为层面的影响，而非复杂且尚未得到充分理解的认知或神经层面机制。输入—输出视角还使我们能够在不建立系统内部模型的情况下仿真系统功能。

• **模块化与多尺度设计**：系统可以划分为不同层级的子系统，这些子系统共同构成多尺度系统模型。例如，HCPS系统由人类、网络和物理子系统组成，其中人类子系统进一步包含由感知、注意力、记忆和心理操作构成的 Cognition 系统。我们可以根据目标缩放到适当层级的子系统。由于这些子系统通过输入和输出相互作用，因此可以替换和修改各个子系统，以实现模块化设计。

• **定量策略与优化**：系统科学模型能够对情境进行定量描述，从而制定最优设计问题，形成具有成本效益的安全机制。整合多种系统科学方法，包括确定性与随机方法、数据驱动工具与基于模型的工具，以及静态与动态框架，可以强化各自的优势并减弱其不足。

参考文献

[1] Andrade R, Torres J, Cadena S (2019) 事件管理流程的认知安全。载于：国际信息技术与系统会议，Springer，第612–621页[2] Arons B (1992) 鸡尾酒会效应综述。美国语音输入/输出协会期刊 12(7):35–50

[4] 常被误引为“整体大于部分之和”。

[3] Barmer H, Dzombak R, Gaston M, Palat V, Redner F, Smith T, Wohlbier J (2021) 可扩展人工智能 DOI 10.1184/R1/16560273.v1，URL https:// kilthub.cmu.edu/articles/report/Scalable_AI/16560273[4] Biddle R, Chiasson S, Van Oorschot PC (2012) 图形密码：从最初十二年的发展中汲取经验。ACM Computing Surveys (CSUR) 44(4):1–41[5] Bruce V, Green PR, Georgeson MA (2003) 视觉感知：生理学、心理学与生态学。Psychology Press[6] Chaczko Z, Kulbacki M, Gudzbeler G, Alsawwaf M, Thai-Chyzhykau I, Wajs- Chaczko P (2020) 面向辅助驾驶系统人机界面的可解释人工智能探索。载于：亚洲智能信息与数据库系统会议，Springer，第507–516页[7] Chen J, Zhu Q (2019) 弹性互依网络分析与设计的博弈与决策理论方法。Springer[8] Cialdini RB (2007) 影响力：说服心理学，第55卷。Collins New York[9] Commission E, for Research DG, Innovation, Breque M, De Nul L, Petridis A (2021) 工业5.0：迈向可持续、以人为本且具有韧性的欧洲工业。Publications Office，DOI doi/10.2777/308407[10] Cox EB, Zhu Q, Balcetis E (2020) 困于网络钓鱼诱饵：在对自身及他人网络风险易感性的判断中，基准率的使用存在差异。《社会心理学综合研究》4(1):25–52[11] Cybenko G, Giani A, Tho mpson P (2002) 认知黑客攻击：一场争夺心智的战斗。Computer 35(8):50–56[12] Deese J (1959) 关于预测即时回忆中特定言语侵入的发生。Journal of experimental psychology 58(1):17[13] Demertzis K, Kikiras P, Tziritas N, Sanchez SL, Iliadis L (2018) 新一代认知安全运营中心：利用网络安全情报进行网络流取证。Big Data and Cognitive Computing 2(4):35[14] Di Pasquale V, Iannone R, Miranda S, Rie mma S (2013) 制造运营中人因可靠性分析技术概述。Operations management 第221–240页[15] Doghri W, Saddoud A, Chaari Fourati L (2022) 用于结构健康监测的信息物理系统：传感技术与智能计算。The Journal of Supercomputing 78(1):766–809[16] Elgendi M, Kumar P, Barbic S, Howard N, Abbott D, Cichocki A (2018) 阈下启动效应——现状与未来展望。Behavioral Sciences 8(6):54[17] Fisher D (2013) 索尼因PlayStation网络遭入侵事件中的安全故障被英国罚款25万英镑。URL https://threatpost.com/sony-fined-250000-uk-over-failures-playstation-network-breach-012413/77446/[18] Garcés IO, Cazares MF, Andrade RO (2019) 在认知安全架构中运用机器学习技术检测网络钓鱼攻击。载于：2019年国际计算科学与计算智能会议（CSCI），IEEE，第366–370页

[19] Greenstadt R, Beal J (2008) 个人设备的认知安全。载于：第 1 届 ACM AISec 研讨会论文集，第 27–30 页[20] Griffor ER, Greer C, Wollman DA, Burns MJ 等 (2017) 信息物理系统框架：第 1 卷，概述[21] Groshev M, Guimarães C, Martín-Pérez J, de la Oliva A (2021) 迈向智能信息物理系统：数字孪生遇见人工智能。IEEE Communications Magazine 59(8):14–20[22] Hamann S (2001) 情绪记忆的认知与神经机制。Trends in Cognitive Sciences 5(9):394–400[23] Huang L, Zhu Q (2022) RADAMS：针对信息拒绝服务（智能拒绝服务）攻击的韧性自适应警报与注意力管理策略。Computers & Security 121:102844[24] Huang L, Chen J, Zhu Q (2017) 面向相互依赖基础设施网络动态防护的大规模马尔可夫博弈方法。载于：International Conference on Decision and Game Theory for Security，Springer，第 357–376 页[25] Huang L, Chen J, Zhu Q (2018) 大规模相互依赖关键基础设施的分布式最优韧性规划。载于：2018 Winter Simulation Conference (WSC)，IEEE，第 1096–1107 页[26] Huang L, Jia S, Balcetis E, Zhu Q (2022) ADVERT：用于网络钓鱼防范的自适应数据驱动注意力增强机制。IEEE Transactions on Information Forensics and Security 17:2585–2597[27] Huang W, Chen X, Jin R, Lau N (2020) 通过生理测量检测视觉检查中的认知黑客攻击。Applied Ergonomics 84:103022[28] Huang Y, Chen J, Huang L, Zhu Q (2020) 用于安全且弹性的控制系统设计的动态博弈。National Science Review 7(7):1125–1141[29] JohnstonWA, Dark VJ(1986) 选择性注意力。Annual reviewofpsychology[30] Kaivanto K (2014) 分散式行为决策对系统层面风险的影响。Risk Analysis 34(12):2121–2142[31] Kanwisher NG (1987) 重复盲视：无需个体标记识别的类型识别。Cognition 27(2):117–143[32] Koffka K (2013) 格式塔心理学原理。Routledge[33] Lindsay GW (2020) 心理学、神经科学与机器学习中的注意力。Frontiers in Computational Neuroscience 14:29[34] Manshaei MH, Zhu Q, Alpcan T, Bacşar T, Hubaux JP (2013) 博弈论与网络安全及隐私的结合。ACM Computing Surveys (CSUR) 45(3):1–39[35] Marot A, Rozier A, Dussartre M, Crochepierre L, Donnot B (2022) 面向电网操作员的人工智能助手，第 79–95 页[36] Meadows DH (2008) 系统思考：入门。Chelsea Green Publishing[37] Molden DC (2014) 理解社会心理学中的启动效应。Guilford Publications[38] Pan SJ, Yang Q (2009) 迁移学习综述。IEEE Transactions on Knowledge and Data Engineering 22(10):1345–1359[39] Payne BD, Edwards WK (2008) 可用安全简介。IEEE Internet Computing 12(3):13–21

[40] Radha V, Reddy DH (2012) 单点登录技术综述。Procedia Technology 4:134–139[41]Rohleder K (2019) 作为漏洞的认知偏差。网址：https: //www.linkedin.com/pulse/cognitive-biases-vulnerabilities-krinken-rohleder/[42] Salau B, Rawal A, Rawat DB (2022) 无线物联网与信息物理系统人工智能领域的最新进展：全面综述。IEEE Internet of Things Journal[43] Sasse MA, Brostoff S, Weirich D (2001) 改变“最薄弱环节”——一种实现易用且有效的安全的人机交互方法。BT technology journal 19(3):122–131[44] Sawyer BD, Hancock PA (2018) 黑客攻击人类：网络安全领域的普遍性悖论。Human Factors 60(5):597–609[45] Schacter DL (2002) 记忆的七宗罪：心智如何遗忘与记忆。HMH[46] Shapiro KL, Raymond JE, Arnell KM (1997) 注意瞬脱。Trends in Cognitive Sciences 1(8):291–296[47] Simons DJ, Rensink RA (2005) 变化盲视：过去、现在与未来。Trends in Cognitive Sciences 9(1):16–20[48] Song J, Lyu D, Zhang Z, Wang Z, Zhang T, Ma L (2022) 当信息物理系统遇上人工智能：基准、评估与未来方向。ICSE 2022 SEIP[49] Southurst J (2013) 比特币支付处理商 BIPS 遭攻击，超过 100 万美元被盗。网址：https://www.coindesk.com/markets/2013/11/25/bitcoin-payment-processor-bips-attacked-over-1-million- stolen/[50] Team SN (2021) 一名俄罗斯公民因试图在特斯拉公司对一名员工实施黑客攻击而认罪。网址：https://www.secureworld.io/industry-news/tesla-hacker-charges-arrested[51] Tipper SP (1985) 负启动效应：对被忽略对象的抑制性启动。The Quarterly Journal of Experimental Psychology 37(4):571–590[52] Vitevitch MS (2003) 变化聋：无法察觉两个声音之间的变化。Journal of Experimental Psychology: Human Perception and Performance 29(2):333[53] Wiener N (2019) 控制论，或动物与机器中的控制和通信。MIT Press[54] Xu Z,Zhu Q (2016) 面向联网 3D 打印机的跨层安全信息物理控制系统设计。载于：2016 American Control Conference (ACC)，IEEE，第 1191–1196 页[55] Zhu Q, Basar T (2015) 信息物理控制系统的鲁棒性、安全性与韧性的博弈论方法：用于最优跨层韧性控制系统设计的博弈中博弈原理。IEEE Control Systems Magazine 35(1):46–65

[56] Zhu Q, Rass S (2018) 博弈论与网络安全相遇：教程。载于：2018 年 ACM SIGSAC 计算机与通信安全会议论文集，第 2163–2165 页[57] Zviran M, Haga WJ (1990) 认知密码：便捷访问控制的关键。Computers & Security 9(8):723–736