中国科学院院刊
Bulletin of Chinese Academy of Sciences
ISSN 1000-3045,CN 11-1806/N

《中国科学院院刊》网络首发论文

题目：智能算法安全：内涵、科学问题与展望
作者：程学旗，陈薇，沈华伟，山世光，陈熙霖，李国杰
DOI：10.16418/j.issn.1000-3045.20240720004
收稿日期：2024-10-16
网络首发日期：2024-10-30
引用格式：程学旗，陈薇，沈华伟，山世光，陈熙霖，李国杰. 智能算法安全：内涵、科学问题与展望[J/OL]. 中国科学院院刊.
https://doi.org/10.16418/j.issn.1000-3045.20240720004

网络首发时间：2024-10-30 14:53:58
网络首发地址：https://link.cnki.net/urlid/11.1806.N.20241030.0901.003

引用格式：程学旗，陈薇，沈华伟，等．智能算法安全：内涵、科学问题与展望．中国科学院院刊，2024，39(0)：1-10，doi：10.16418/j.issn.1000-3045.20240720004.
CHENG Xueqi, CHEN Wei, SHEN Huawei, et al. Intelligent algorithm safety: Concepts, scientific problems and prospects. Bulletin of Chinese Academy of Sciences, 2024, 39(0): 1-10, doi: 10.16418/j.issn.1000-3045.20240720004. (in Chinese)

智能算法安全：内涵、科学问题与展望

程学旗[1,2*] 陈 薇[1,2*] 沈华伟[1,2] 山世光[1,2] 陈熙霖[1,2] 李国杰[1,2]

1 中国科学院计算技术研究所 智能算法安全重点实验室 北京 100190
2 中国科学院大学 计算机科学与技术学院 北京 100049

摘要 智能算法是指实现智能的计算过程所体现的方法，大多具备数据驱动、不确定性计算、模型推断难解释等特性，而这些特性同时也给智能算法应用带来了潜在的安全风险。文章首先探讨智能算法安全的内涵。具体地，智能算法安全的内涵依据人机融合的程度，由算法自身的一元内生性安全，延伸到算法服务于人时的人机二元应用性安全，最终拓展为人机共生的复杂社会系统中多元系统性安全，故据此提出智能算法安全层级范式（以下简称“TRC范式”），分别涵盖内生决策可信（trustworthiness）的一元安全目标、应用服务可管（regulatability）的二元安全目标和系统风险可控（controllability）的多元安全目标。进一步，基于当前实现TRC范式中的技术难点与智能算法可信、可管、可控的目标，文章提出实现智能算法安全需要重点突破的不确定性算法的可信域判定、黑箱模型的透明化监测与人机共生智能系统的风险临界点感知3个重大科学问题。最后，围绕TRC范式的“度量—评估—增强”技术体系，提出7项研究方向建议与4方面智能算法安全相关的发展建议，并展望其助力实现人机共治的未来愿景。

关键词 大数据，智能算法，智能算法安全，人工智能伦理与安全，TRC范式

DOI 10.16418/j.issn.1000-3045.20240720004
CSTR 32128.14.CASbulletin.20240720004

人工智能（AI）技术经过几十年的发展，正在进入一个技术创新与颠覆式应用模式频现的爆发期，人工智能伦理与安全问题受到广泛关注。达特茅斯会议之前，科幻作家阿西莫夫提出了“机器人三大定律”，

*通信作者

资助项目：中国科学院战略性先导科技专项（XDB0680000），中国工程院咨询项目（2024-XBZD-05）
修改稿收到日期：2024年10月16日；

关于人工智能的伦理与安全问题在这之后的很长一段时间主要集中在哲学和科幻领域[1]。21世纪，以大数据融合深度学习为代表的统计学派占据了人工智能技术主流，人工智能的伦理与安全问题开始凸显。近年来，生成式大模型在文本、图像、视频、自然语言处理等领域产生系列现象级应用[2]，人工智能的伦理与安全问题快速出现，受到社会广泛关注。例如，2019—2021年，美国国家公路交通安全管理局统计共发生807起自动驾驶车祸案件，其中超过90%的案件涉及启用Autopilot功能的特斯拉车辆撞击带有明显标识的静止车辆、公路隔离墩甚至行人①；2018年9月，广州市共查处外卖骑手交通违法近2 000起，主要原因是外卖骑手为赶在平台AI算法设定的限制时间内送达外卖而采取超速、逆行等危险驾驶行为②；2016年，社交媒体平台Facebook在美国大选期间被俄罗斯机构利用AI算法投放约8万条政治舆论相关的帖子③，剑桥分析公司非法使用8 700万脸书用户的数据并利用AI算法针对性地发送政治宣传广告④。

针对日益严峻的人工智能安全问题，各国政府在积极探索有效治理模式。我国2021年以来先后发布了《新一代人工智能伦理规范》《可信人工智能白皮书》《互联网信息服务算法推荐管理规定》与全球首部针对生成式人工智能的法规《生成式人工智能服务管理暂行办法》；并于2023年10月发布《全球人工智能治理倡议》，围绕人工智能发展、安全、治理3方面系统阐述了人工智能治理的中国方案，提出11项倡议⑤。美国白宫2023年10月首次针对AI发布行政令，涵盖建立AI安保、隐私保护、人权保护、促进创新等多方面内容。欧洲2023年11月召开首届全球AI安全峰会并签署《布莱特利宣言》⑥，确认解决人工智能对人权保护、透明度和可解释性、公平性、问责与监管机制、道德偏见、隐私和数据保护等问题的必要性和紧迫性；欧盟理事会于2024年7月公布《人工智能法案》⑦，该立法遵循“基于风险”的方法，风险等级越高，管控越严格。

人工智能领域的学者们呼吁重视人工智能存在的安全风险[3,4]。2022年6月，第24届中国科协年会发布十大前沿科学问题，信息领域唯一一个问题是“如何实现可信可靠可解释人工智能技术路线和方案？”。2023年3月，1 000余名人工智能领域学者签署公开

① 17 fatalities, 736 crashes: The shocking toll of Tesla’s Autopilot. (2023-06-10)[2024-10-11]. https://www.washingtonpost.com/technology/2023/06/10/tesla-autopilot-crashes-elon-musk/.

② 外卖骑手生存调查：拼命快了，快乐了谁？. (2020-09-14) [2024-10-11]. https://tv. cctv. com/2020/09/15/VIDEDXQbCNWX5VOWV6ZdNX5y200915.shtml.

③ Obama administration announces measures to punish Russia for 2016 election interference. (2016-12-26)[2024-10-15]. https://www.washingtonpost.com/world/national-security/obama-administration-announces-measures-to-punish-russia-for-2016-election-interference/2016/12/29/311db9d6-cdde-11e6-a87f-b917067331bb_story.html.

④ Cambridge Analytica and Facebook: The scandal so far. (2018-05-28) [2024-10-15]. https://www. aljazeera. com/news/2018/3/28/cambridge-analytica-and-facebook-the-scandal-so-far.

⑤ 全球人工智能治理倡议. (2023-10-18)[2024-10-11]. https://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm.

⑥ The Bletchley declaration by countries attending the AI safety summit, 1-2 November 2023. (2023-11-01)[2024-10-15]. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023?tpcc=world_brief.

⑦ Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). [2024-10-15]. https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

信，呼吁立即暂停训练比GPT-4更强大的AI模型，为期至少6个月⑧。2023年5月，图灵奖获得者Geoffrey Hinton教授从谷歌离职，原因是“为了自由地讨论人工智能的风险”，该事件引发社会各界对强人工智能何时实现、其是否会取代人类等人工智能伦理安全问题的大讨论。

本文认为社会上对人工智能安全性的恐惧主要来自科幻电影和媒体的夸张宣传，认为智能机器将会有自主意识，完全脱离人的控制，甚至成为新的物种征服人类，这只是一些人的猜测，目前还没有科学依据，在可预见的未来，还不构成真正的安全威胁。对于长远未来可能存在的安全问题，各国政府和前沿研究的学者们已经开始探讨在人工智能研究和开发过程中加强伦理道德的规范和引导，确保自主智能的发展符合人类的价值观和利益[5-7]。本文更加关注当前AI应用过程中已经存在且愈演愈烈的问题。

尽管各国政府和国内外学者高度重视人工智能安全，但对处于人工智能核心的智能算法安全内涵理解尚未深入，实现算法安全治理的技术路径尚不明晰。本文旨在以计算的视角，梳理智能算法安全的需求，明确智能算法安全的内涵，并针对智能算法安全的目标，探讨关键科学问题，提出潜在关键技术及其应用。这对在智能化时代确保人类自身安全、保障用户权益、维护社会稳定，最终实现人机共治，具有重要意义。

1 智能算法安全的内涵

1.1 智能算法

算法是指将信息进行变换的计算过程所体现的方法[8]。高纳德（Donald E. Knuth）教授定义算法为求解特定类型问题的运算序列的一组有穷规则，并具备有穷性、确定性、输入、输出、能行性5个特征[9]。

*智能算法在本文中定义为实现智能的计算过程所体现的方法。*智能算法大多具有数据驱动、不确定性计算、模型推断难解释等典型特征。智能算法的设计者基于少量知识设计参数化模型，依赖数据训练模型参数。计算机在训练阶段基于随机迭代计算更新模型，在推断阶段使用训练所得模型面向具体任务产生输出（图1）。基于多种任务来源的数据，智能算法所训练的模型具有执行多种任务的能力。由于智能算法利用了大量数据中蕴含的知识，设计者所需的知识明显减少，在此意义上称其具备“智能”。

智能算法目前以深度学习为典型代表，其不确定性与智能的关系值得深入思考[10]。基于随机数据学习的计算每一步迭代的机理是确定的，但经过多步迭代后的计算规则人类难以理解，规则意义上的确定性大大降低。与智能算法相比，传统算法的设计者依据特定知识设计确定性的计算规则，并由计算机执行产生输出。故基于规则的计算如果能自发生成新的规则即可以减少对知识的依赖，在这种情况下，传统算法可以进阶为智能算法。

*智能算法的计算不确定性、结果复杂难解释等特性使得对其安全风险进行管控极具挑战。*近年来，基于多模态大模型的智能算法显著提升了多任务执行能力[2]，与人类交互的障碍大大减小，应用场景急速扩大。与此同时，智能算法中模型判定与生成的不确定性结果对人类产生的不良影响也日益凸显[11]。以大语言模型为例，基于数据驱动的概率生成模式，可能生成与现实世界事实不一致的幻觉内容，产生错误[12]；大模型的训练、推断机理复杂难解释，模型的漏洞隐性难发现，在恶意者对抗攻击下，可能出现推断错误、导致歧视或泄漏用户隐私等现象[13-15]。

⑧ Pause giant AI experiments: An open letter. (2023-3-22)[2024-10-11]. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

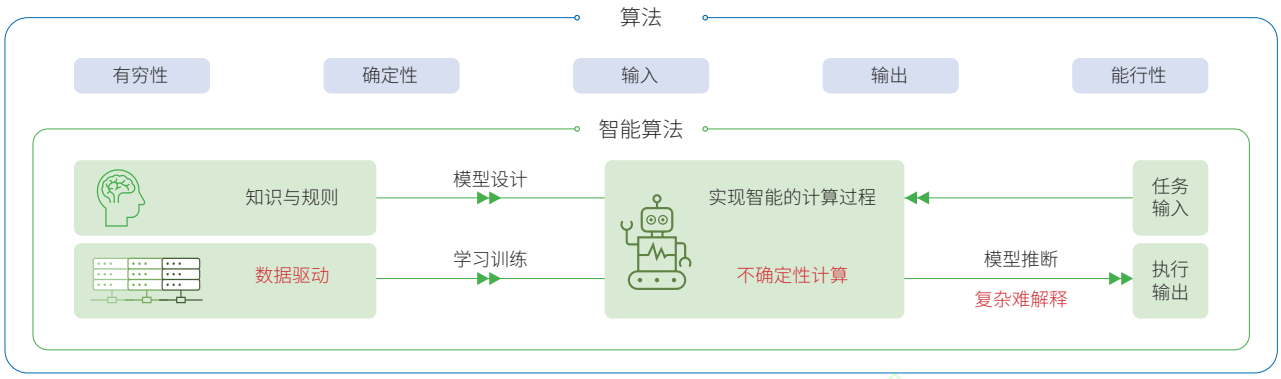

图 1　智能算法的特点

Figure 1　Characteristics of intelligent algorithms

1.2 智能算法安全的内涵

智能算法安全研究致力于降低智能算法对人类产生的风险，研究如何度量、评估、增强智能算法的安全程度。智能算法应用领域广泛，正在与人类自身活动交互融合。人机融合的程度越深，风险形成的过程越复杂，科学问题的挑战越大，安全治理技术的要求越高。因此，本文将算法安全需求场景中人机智能交互的程度划分为算法内生一元、人机交互二元与人机共生多元3个安全层级。① 算法内生一元安全层级。算法作为生产工具帮助人类在物理空间的已有任务上提升执行效率、减小人类投入、或降低对生态环境的不良影响。例如，智慧农耕算法提升粮食产量，智能调度算法减少能源消耗，自动驾驶算法降低人类驾驶负荷，科学智能算法加快科学发现的进程⑨等。此类任务中，算法往往具备内在的适用边界，当任务执行不当时会触发物理世界中的事故。例如，自动驾驶车祸导致人民生命财产安全事故，自动交易算法异常导致经济损失，大模型幻觉导致错误决策等。② 人机交互二元安全层级。在网络化应用中，大量算法依托平台为用户提供交互式智能服务。例如，搜索推荐算法为用户提供快速准确的信息获取服务，外卖平台算法为消费者、商家、快递员提供实时、高效的配置方案，视频游戏类算法为用户提供电子类休闲娱乐服务等。在这些应用中，算法作为智能服务提供方，可能由于损害用户权益而触发服务产品的风险。例如，搜索算法泄露用户隐私[16]，外卖平台导致快递员疲于奔命，信息推荐算法导致用户陷入信息茧房等[17]。③ 人机共生多元安全层级。算法通过物理空间和网络空间与人类共同参与社会活动，形成人类智能和机器智能交织的人机共生系统。例如，带有智能体的社交平台，智能算法参与的金融交易系统，有人—无人系统共同参与的军事演习等。此系统中，算法可能由于通过行为传导而触发系统性安全风险。例如，基于社交平台的选举操控③，基于网络空间的社会认知博弈[18]等。

智能算法3个安全层级依次嵌套，算法内生一元安全层级是人机交互二元安全层级中机器在物理域对人提供的局部服务，人机交互二元安全层级是人机共生多元系统的人机二元局部交互系统，故单个智能算法会面临跨层次的安全风险。例如，智能驾驶的主要风险中，自动驾驶算法不稳定导致的交通事故属于算法内生一元安全层级，算法泄露用户隐私数据的风险

⑨ The Nobel Prize in Chemistry 2024: They cracked the code for proteins' amazing structures. (2024-10-09)[2024-10-12]. https://www.kva.se/en/news/the-nobel-prize-in-chemistry-2024/.

属于人机交互二元安全层级；互联网服务平台中，调度决策算法导致司机、骑手权益受损属于人机交互二元安全层级，在突发极端情况下交通拥塞导致算法可用性下降属于算法内生一元安全层级；社交平台被用于政治干预产生社会认知风险属于人机共生多元安全层级，平台用户面临隐私泄漏和信息茧房等风险属于人机交互二元安全层级。

1.3 智能算法安全层级（TRC）范式

智能算法3个层级关注的安全风险类型和产生原因不同：一元场景中关注由算法内生缺陷导致算法性能不可信，二元场景中关注由算法应用中的滥用误用导致算法服务不可管，多元场景中关注由人机算法博弈对抗导致系统演变不可控。智能算法安全的目标是实现智能算法一元内生决策可信（trustworthiness）、二元服务应用可管（regulatability）、多元系统风险可控（controllability），即智能算法安全层级范式（以下简称“TRC范式”）。（图2）

*一元内生性安全限定于机器自身，聚焦于智能算法内生缺陷导致的算法决策失信。*人类设计智能算法的最初目的是顺利实现其预期功能，其达成任务目标的能力仅由算法自身决定，而与其如何被使用无关。在一元安全的范畴内，算法的风险来自算法自身缺陷所导致的失能或失效，体现为在遇到数据环境被动性变化或主动性对抗攻击时功能失效或性能下降。因此，内生性安全着眼于智能算法在可变及对抗环境下的决策性能，其目标是构建性能稳健可靠的智能算法，实现智能算法内生机理可信。

*二元应用性安全关注智能算法滥用误用导致的算法行为与用户权益的失配。*随着算法的智能水平逐渐接近人类智能，其与人类的互动也日益紧密，影响日益显著。算法的风险主要源于其行为与人类社会的普遍价值观（如公平、公正、隐私等）不一致，例如互联网服务可能导致的歧视、隐私泄漏、观点极化等道德或伦理失范和混乱问题。算法应用性安全的目标是技术向善，在智能算法服务用户的同时，避免损害个体和公众的权益，确保技术创新与社会价值观的和谐共融，实现智能算法的应用服务可管。

*多元系统性安全着眼于人机共生系统中因算法博弈对抗导致复杂社会系统的演化不可控。*智能算法的快速发展，促使复杂社会系统人机智能融合，模糊了人机边界，重构了社会结构与组织关系。与此同时，智能算法的自主决策演变可能会导致复杂社会系统呈现组织结构坍塌、传播链式反应、系统临界态不确定等失序、失控现象。算法在某些任务上，如内容生成和传播等，具有超越人类智能的能力，这使得社会系统存在被算法干预的风险。随着社会系统中的系统博弈强度升级，系统性风险开始显现，例如伪造政治谣言、恶意宣传、社会认知操控、有人—无人系统失控

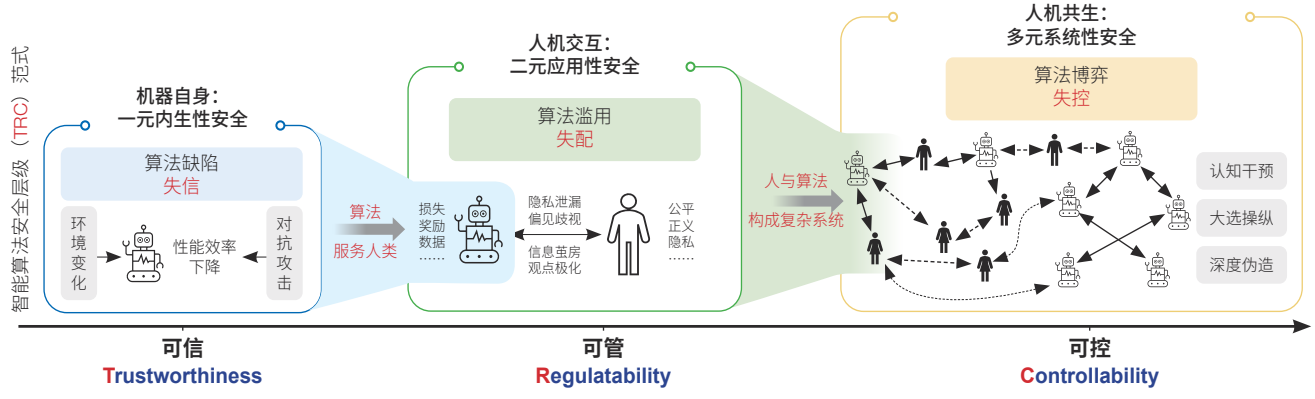

图2　智能算法安全层级（TRC）范式

Figure 2　Intelligent algorithm safety TRC paradigm

等。实现人机共生系统安全的可能路径包括通过复杂社会系统的可计算建模，识别与调控复杂社会系统风险的相变点，实现智能算法的社会风险可控。

1.4 智能算法安全层级范式与其他领域的联系

TRC范式可以涵盖已有的相关概念（表1）。① 可信算法与TRC范式的关系：智能算法的可信性一般指算法能被用户或社会认为是可信赖的、可依赖的，包括算法的鲁棒性、公平性、可解释性、隐私性、可审计性等，其定义包含于TRC范式之中。② 负责任算法与TRC范式的关系：智能算法的负责任性一般指算法的行为符合道德、法律要求，避免对人类、环境或社会造成危害或不良影响，其定义包含于TRC范式之中。③ TRC范式中的系统性：现有相关概念仅适用于描述算法本身，故不适用于系统性安全层面。

智能算法安全是涉及社会域安全的一门新兴学科，与其他安全类学科有一定的联系（图3）。在数字化早期阶段，物理世界数字化是主体任务，物理电磁空间安全是关键，着重解决通信安全、电磁安全与频谱控制等难题；随着互联网的发展，人机互联推动信息域与物理域融合，网络与信息安全问题凸显，网络安全与信息系统安全成为重点。当前社会已经进入智能化时代，万物互联，“物理—信息—社会”三元空间融合，算法成为影响智能社会运行的核心引擎之一，智能算法安全成为新的安全挑战。这种挑战的出现是信息社会发展演进的必然结果，不同于以往主要关注物理域和信息域的物理安全以及网络信息安全，智能算法安全的关注点更加聚焦于由智能算法带来的社会域问题，需要重新审视和构建安全保障的策略和措施。

2 智能算法安全的科学问题

基于对智能算法及其内涵的理解，本文认为智能算法安全的核心挑战是确定性的安全要求与不确定性计算的智能算法及其难以度量的社会域风险三者之间的矛盾。基于智能算法安全内涵，结合目前的技术难点与核心挑战，面向TRC范式的不同层次，提出智能算法在安全“度量—评估—增强”技术链条的相应挑战。① 针对内生性安全，度量是任务执行的功能指标，评估目前主要是试验性方法，增强方面也主要是启发式增强方法；② 针对应用性安全，度量的部分权益维度可计算，例如隐私性、公平性，但缺乏统一的价值观度量与计算方法，静态、平均情形评估技术居

表1 TRC范式与现有相关概念的关系

Table 1 Intelligent algorithm safety TRC paradigm and other related concepts

概念分类		TRC范式与概念的关系		
概念名称	概念内容	内生性:可信（trustworthiness）	应用性:可管（regulatability）	系统性:可控（controllability）
可靠（reliability）	算法能在各种客观环境下,按设计性能运行,无故障或错误	包含	不包含	不包含
鲁棒性（robustness）	算法面临数据变化或攻击时,能保持稳定和有效的能力	包含	不包含	不包含
公平性（fairness）	算法决策过程满足公平原则,不偏向或歧视特定群体	不包含	包含	不包含
隐私性（privacy）	算法不泄漏出人或组织的隐私信息	不包含	包含	不包含
可解释（explainability）	算法能以人类可以理解的形式阐明其运行机理与决策机制	包含	包含	不包含

多，实时、最坏情形监测技术欠缺；③针对系统性安全，由于系统演化规律未知，可计算的社会安全度量尚未建立。因此，TRC范式的技术难点依次为，内生性安全评估的理论判定，应用性安全评估的监测技术，系统性安全度量中的可计算方法。考虑到TRC的层级嵌套关系，每个技术难点的解决以其前一个难点的解决为必要条件。综上，总结如下智能算法安全的3个关键科学问题。

（1）不确定性算法的可信域判定问题。智能算法包含不确定性计算，具有数据驱动、模型复杂、机理不清晰等特点。模型的不稳定性使得算法决策的精准性和稳定性难以兼得；数据的不完备性使得关联统计失效，导致算法决策偏差；应用场景的突变性使得算法场景先验假设失效，导致算法行为失控。如何实现对高复杂、强不确定性智能算法的可信域判定和增强，是保障智能算法内生性安全的关键科学挑战。

（2）黑箱模型的透明化监测问题。智能算法具有模型黑箱和结果难解释等特性，算法风险评估仅能通过算法的外显行为进行。智能算法黑箱体现为数据黑箱（使用数据不透明）、模型黑箱（决策机制不透明）和目标黑箱（设计意图不透明），从而导致监管方和算法运营方存在信息不对称。如何仅通过算法运行过程中的外显行为反向推断算法的内在机理，实现算法透明监测，是保障智能算法应用性安全的关键科学挑战。

图3　智能算法安全与其它安全类学科的关系

Figure 3　Intelligent algorithm safety and other related areas

（3）人机共生智能系统的临界点感知问题。智能算法促使复杂社会系统人机智能融合，模糊了二者的边界。传统复杂系统理论缺乏对社会系统的可计算建模，不能满足社会系统风险演化相变的临界态分析需求。如何建模感知测绘、信息生成、信息传播，识别与调控复杂社会系统安全风险的相变点，是保障智能算法系统性安全的关键科学问题。

3 研究方向建议

围绕智能算法安全相关基础理论、关键技术及应用需求，中国科学院计算技术研究所部署设立了智能算法安全重点实验室，旨在重点突破TRC范式面临的重大科学问题，服务国家在智能算法安全治理和网络空间社会治理两大需求。开展上述方向研究，不仅需要学术界综合利用多学科交叉理论技术，更需要产业、政府部门提供实际应用和业务需求，各方共同构建新一代人工智能安全治理框架。本部分依据智能算法安全的TRC内涵，针对前述3个科学问题，建议体系性地加强7项关键理论与技术研究，以期得到相关领域研究者与社会各界的关注，共同推动智能算法安全的持续发展。7个研究布局建议依据TRC范式体系性的提出，不仅为已出现的研究领域提供了新的研究思路并将它们联系起来，同时指出了学术界仍未关注到的研究领域。

（1）可信判定理论为智能算法安全的基础理论支撑6项关键技术。针对智能算法面临的模型不稳定、数据不完备、场景突变等挑战，研究建立深度学习的数学原理，形成对模型在优化过程中收敛性的判定；研究建立因果学习理论，形成模型对不完备数据适应性的判定；研究建立可信学习理论，形成对模型在可变及对抗环境中泛化性的判定。

（2）可信机制嵌入技术为安全增强提供基础性方法。针对环境被动性变化导致的安全问题，研究先验

知识嵌入技术，实现对智能算法内生机理的安全增强；针对环境中主动性对抗攻击导致的安全问题，研究防御机制嵌入技术，实现对智能算法防御能力的加固增强；基于价值观可计算度量技术，建立可信价值观嵌入方法，实现应用性安全增强。

（3）社会域风险可计算度量技术为应用性安全建立度量标准。旨在衡量算法与每一个用户交互的过程中，是否符合伦理道德和法律的合规性。针对应用场景多变和人类价值观（例如法律法规、道德标准等）难以量化所造成的度量困难，通过建立度量大模型，对算法应用中出现的风险案例进行基于语义的评价。

（4）智能算法黑箱监测技术可向内支撑内生性安全评估、向外延展到系统性安全评估。基于模型碰撞的意图识别，推断模型内在机理，为应用风险判定奠定重要基础；基于用户模拟的风险发现，从宏观层面发现算法导致的隐私泄露、茧房、公平性等应用风险；基于红队测试的案例生成，面向特定的度量指标，构造或挖掘违反度量指标的样例，作为判定的取证样例。

（5）价值观对齐技术依据可计算度量技术和可信嵌入技术，实现智能算法应用性安全增强。包括事前的全局对齐训练与事后的局部对齐编辑相结合的范式。在模型发布前，根据度量模型，通过监督微调、人类反馈强化学习、AI反馈强化学习、基于规则的奖励模型等方式进行智能算法和价值观的对齐训练；在风险发生后，根据找到的违反度量指标样例，通过对齐编辑、神经元定位及修复等方式，修复特定性错误。

（6）人机共生智能系统演化模拟技术为开展系统性安全研究建立前提。建立复杂社会认知模拟系统，并设计博弈效用度量及其动态评估方法。建立微观个性化与宏观群体化相结合的社会域安全度量可计算方法，探索观点、立场、情感、价值观等因素的量化体系，形成体系化的复杂系统博弈效用评估标准与评估模型。提出对真实系统进行模拟对齐的方式，评估智能算法系统性安全。

（7）人机共生智能系统临界点感知与调控技术是系统性安全评估与增强的核心技术。通过脆弱点发现与多模态可控内容生成相结合，增强算法博弈下的系统性安全。探索智能算法对复杂社会系统相变点的干预机制，构建社会系统脆弱点感知发现和系统风险调控的关键技术链，形成具备对算法恶意介入社会系统的防御手段以及算法介入的社会系统调控手段，实现复杂社会系统风险相变点的识别与调控。

4 智能算法安全建议与未来展望

随着人工智能技术的快速发展并在不同行业领域广泛产生颠覆性的应用，智能算法的安全问题也愈发成为人工智能发展中的关键瓶颈。既需要结合实际需求场景解决智能算法所引发的数据、模型、应用等实际问题，也需要重视智能算法安全的基础理论研究以及学科建设工作。虽然科技发展伴随安全风险，但相信凭借各方的高度关注、全球协作、持续不懈，这把人工智能伦理与安全的达摩克利斯之剑，将最终被人类所驾驭。在我国加速推动新一代人工智能发展的战略布局中，要重视与智能算法安全相关的4个方面工作。

（1）夯实基础理论。智能算法安全的核心矛盾以及3个科学问题对经典的计算复杂性理论、复杂系统理论、人工智能安全与伦理研究提出了全新挑战，如何在计算视角下推动这些理论的发展，并最终汇聚夯实智能算法安全的理论基础是关键。相关理论研究不仅是建立智能算法安全的基石，也将促进传统计算理论在智能化时代变革发展。

（2）促进学科交叉。智能算法安全将传统的信息安全拓宽到更广泛的社会域人机共生智能系统安全。相关研究涉及计算、智能、安全、伦理、法律及社会科学相关的多个学科领域，需要跨学科交叉共同研

究，建立基于多学科基础的技术解决方案。与此同时，要推动国内外同行交流合作，共同形成全球人工智能的治理框架并理性发声，从而在新一轮科技革命中掌握一定话语权。

（3）推进算法安全产业闭环。借助商业模式创新，提升算法安全技术突破在算法服务中的应用速度与质量。鼓励提供算法安全服务的企业，通过算法可信增强技术与算法合规辅导服务，为企业节省安全维护成本、提升业务质量、实现商业价值，进而获取相应的商业回报。依托于核心技术突破，借助商业模式的推动，灵活快速地推动科技成果的落地应用。

（4）加快人才培养。智能算法安全是一个全新的、快速发展的学科领域，应加快培养该领域的科研团队力量，为优秀青年学者提供稳定的科研资源支持。同时，应尽快制定本领域研究生培养方案，探讨在计算机、人工智能和大数据相关学科领域设立智能算法安全本科专业的培养方案。

致谢 本文离不开智能算法安全重点实验室研究团队和中国科学院战略性先导科技专项参与成员的长期共同论证和科研经验积累。在此感谢苏度、徐冰冰、李家宁、曹婧、孙飞等为本文作出的贡献，受文章篇幅限制，不能一一列出成员姓名。

参考文献

1 Asimov I. I, Robot. New York: Spectra Books, 2004.

2 Achiam O J, Adler S, Agarwal S, et al. GPT-4 technical report. 2024, doi: arXiv:2303.08774v6.

3 Bengio Y, Hinton G, Yao A, et al. Managing extreme AI risks amid rapid progress. Science. 2024, 384(6698): 842-845.

4 Anderljung M, Barnhart J, Korinek A, et al. Frontier AI regulation: Managing emerging risks to public safety. 2023, doi: arXiv:2307.03718v4.

5 Hendrycks D, Mazeika M, Woodside T. An overview of catastrophic ai risks. 2023, doi: arXiv:2306.12001v6.

6 Amodei D, Olah C, Steinhardt J, et al. Concrete problems in AI safety. 2016, doi: arXiv:1606.06565v2.

7 Falco G, Shneiderman B, Badger J, et al. Governing AI safety through independent audits. Nature Machine Intelligence. 2021, 3(7): 566-571.

8 徐志伟, 孙晓明. 计算机科学导论. 北京: 清华大学出版社, 2018.
Xu Z W, Sun X M. Introduction to Computers Science. Beijing: Tsinghua University Press, 2018. (in Chinese)

9 Knuth D E. The Art of Computer Programming. Massachusetts: Addison-Wesley, 1973.

10 李国杰. 智能化科研(AI4R)：第五科研范式. 中国科学院院刊, 2024, 39(1): 1-9.
Li G J. AI4R: The fifth scientific research paradigm. Bulletin of Chinese Academy of Sciences, 2024, 39(1): 1-9. (in Chinese)

11 Bommasani R, Hudson D A, Adeli E, et al. On the opportunities and risks of foundation models. 2022, doi: arXiv:2108.07258v3.

12 Zhang Y, Li Y F, Cui L Y, et al. Siren's song in the AI ocean: A survey on hallucination in large language models. 2023, doi: arXiv:2309.01219v2.

13 Yi S B, Liu Y L, Sun Z, et al. Jailbreak attacks and defenses against large language models: A survey. 2024, doi: arXiv: 2407.04295v2.

14 Gallegos I O, Rossi R A, Barrow J, et al. Bias and fairness in large language models: A survey. Computational Linguistics, 2024, 50(3): 1097-1179.

15 Yao Y F, Duan J H, Xu K D, et al. A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. High-Confidence Computing, 2024, 4(2): 100211.

16 Jeckmans A J P, Beye M, Erkin Z, et al. Privacy in recommender systems// Ramzan N, van Zwol R, Lee J S, et al. Social Media Retrieval. London: Springer, 2013: 263-81.

17 Tomlein M, Pecher B, Simko J, et al. An audit of misinformation filter bubbles on YouTube: Bubble bursting and recent behavior changes// Proceedings of the 15th ACM Conference on Recommender Systems. New York: Association for Computing Machinery, 2021.

18 Zittrain J. "Netwar": The unwelcome militarization of the Internet has arrived. Bulletin of the Atomic Scientists, 2017, 73(5): 300-304.

Intelligent algorithm safety: Concepts, scientific problems and prospects

CHENG Xueqi[1,2*] CHEN Wei[1,2*] SHEN Huawei[1,2] SHAN Shiguang[1,2] CHEN Xilin[1,2] LI Guojie[1,2]

（1 CAS Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;

2 School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China）

Abstract Intelligent algorithms refer to the methods embodied in the computational processes that realize intelligence. These methods are often characterized by being data-driven, involving uncertain computations, and with unexplainable model inferences. These characteristics simultaneously introduce potential safety risks to the application of intelligent algorithms and AI. This study firstly explores the concepts of intelligent algorithm safety. Specifically, intelligent algorithm safety, based on the degree of human-machine integration, extends from the univariate safety of the algorithm itself to the bivariate applicational safety when the algorithm serves humans, and finally evolves into the multivariate systemic safety arises within complex socio-technical systems of human-machine symbiosis. Therefore, this study proposes a hierarchical paradigm of intelligent algorithm safety, namely "TRC paradigm", covering the univariate safety objective of trustworthiness in algorithm's internal decision-making, the bivariate safety objective of regulatability in application services, and the multivariate safety objective of controllability for system-wide risk management. Furthermore, based on the current technical challenges in achieving the TRC paradigm and in line with the goals of trustworthiness, regulatability, and controllability, the study identifies three major scientific questions that need to be answered: determining the trust regions of uncertain algorithms, transparentized monitoring of black-box models, and sensing the critical point in human-machine symbiotic intelligent systems. Finally, this study outlines seven research directions, and four recommendations related to intelligent algorithm safety under the "measurement-evaluation-enhancement" technical framework of the TRC paradigm, while envisioning how this will help achieve a future of human-machine co-governance.

Keywords big data, intelligent algorithms, intelligent algorithm safety, ethics and safety of artificial intelligence, TRC paradigm

程学旗　中国科学院计算技术研究所研究员，主要研究领域：大数据分析系统、社会认知计算、智能算法安全。
E-mail: cxq@ict.ac.cn

CHENG Xueqi　Professor of Institute of Computing Technology, Chinese Academy of Sciences. His main research areas include big data analysis systems, social cognitive computing, intelligent algorithm safety. E-mail: cxq@ict.ac.cn.

陈　薇　中国科学院计算技术研究所研究员，主要研究领域：机器学习基础理论与方法、可信机器学习、智能算法安全。
E-mail: chenwei2022@ict.ac.cn

CHEN Wei　Professor of Institute of Computing Technology, Chinese Academy of Sciences. Her main research areas include basic theory and methods of machine learning, trustworthy machine learning, and intelligent algorithm safety. E-mail: chenwei2022@ict.ac.cn

■责任编辑：文彦杰

*Corresponding author