DOI 10.15302/J-SSCAE-2021.03.005

# 针对强人工智能安全风险的技术应对策略

**刘宇擎，张玉槐，段沛奇，施柏鑫，余肇飞，黄铁军，高文**

（北京大学计算机科学技术系，北京 100871）

**摘要：**未来进入强人工智能（AGI）时代，人类可能面临重大安全风险。本文归纳了 AGI 与传统人工智能的区别，从模型的不可解释性、算法及硬件的不可靠性、自主意识的不可控性三方面研判了 AGI 安全风险的来源，从能力、动机、行为 3 个维度提出了针对 AGI 的安全风险评估体系。为应对安全风险，从理论及技术研究、应用两个层面分别探讨相应风险的防御策略：在理论技术研究阶段，完善理论基础验证，实现模型可解释性，严格限制 AGI 底层价值取向，促进技术标准化；在应用阶段，预防人为造成的安全问题，对 AGI 进行动机选择，为 AGI 赋予人类价值观。此外，建议加强国际合作，培养强 AI 研究人才，为迎接未知的强 AI 时代做好充分准备。

**关键词：**强人工智能；安全风险；风险评估；应对策略

**中图分类号：**TP18　　**文献标识码：**A

# Technical Countermeasures for Security Risks of Artificial General Intelligence

**Liu Yuqing, Zhang Yuhuai, Duan Peiqi, Shi Boxin, Yu Zhaofei, Huang Tiejun, Gao Wen**

(Department of Computer Science and Technology, Peking University, Beijing 100871, China)

**Abstract:** Human beings might face significant security risks after entering into the artificial general intelligence (AGI) era. By summarizing the difference between AGI and traditional artificial intelligence, we analyze the sources of the security risks of AGI from the aspects of model uninterpretability, unreliability of algorithms and hardware, and uncontrollability over autonomous consciousness. Moreover, we propose a security risk assessment system for AGI from the aspects of ability, motivation, and behavior. Subsequently, we discuss the defense countermeasures in the research and application stages. In the research stage, theoretical verification should be improved to develop interpretable models, the basic values of AGI should be rigorously constrained, and technologies should be standardized. In the application stage, man-made risks should be prevented, motivations should be selected for AGI, and human values should be given to AGI. Furthermore, it is necessary to strengthen international cooperation and the education of AGI professionals, to well prepare for the unknown coming era of AGI.

**Keywords:** artificial general intelligence (AGI); security risk; risk assessment; coping strategy

**收稿日期：**2021-04-07；**修回日期：**2021-04-25

**通讯作者：**施柏鑫，北京大学计算机科学技术系研究员，研究方向为计算摄像学、计算机视觉；E-mail: shiboxin@pku.edu.cn

**资助项目：**中国工程院咨询项目“新一代人工智能安全与自主可控发展战略研究”(2019-ZD-01)

**本刊网址：**www.engineering.org.cn/ch/journal/sscae

## 一、前言

强人工智能（AGI）可以由一套系统处理各种智能行为，而弱人工智能针对每种智能行为都需要新的独立系统，这是两者之间的根本性区别。仅依靠弱人工智能的算法改进（而忽略系统更新）来实现 AGI 是不可能的 [1]。在认知论方面，AGI 强调意识的存在，突出价值观和世界观体系，认为智能体可以拥有生物的本能。AGI 不一定是类人的。在外形上，AGI 可以和人类相似（共享一套生活法则），也可与人类相差甚远（形成一套新的生活法则）；在思想上，AGI 可以与人类共用一套思维模式、道德准则，也可拥有专属体系的独特推理方式，成为一类“拥有灵魂的机器”。一般认为，不同于当前获得广泛应用的人工神经网络，能够更加精细解析动物和人类大脑的研究工作有望在未来 20 年内逐步实现，以此构建未来神经网络的体系结构；由此衍生的类脑机有望成为 AGI 的物理实现平台 [2]。

AGI 是人类想要且正在创造的机器，但也可能通过社会操纵、新型战争、权力动态变化等方式引发一些实质性的问题；起初会听从人类的特定指令，但趋向于自主做出决定，这种决定是否会影响人类的实际利益甚至是生命财产安全，未有明确答案。当前，科学界针对 AGI 研究的讨论非常激烈：现有的人工智能（AI）基本方法存在缺陷，必须走向具有理解力的 AI，真正的 AI 还很遥远 [3]；人类距离研制出自主智能（即 AGI）还需要数十年的努力，当前面临的是基础问题，在本质上尚属数学研究挑战 [4]；当前 AI 技术所取得的进展缘于弱人工智能，主流学术界并未将 AGI 作为发展方向，出于对 AGI 的担忧而不建议主动开展研究 [5]；人类不能固步自封于弱人工智能，智能演化过程不可阻挡，大脑意识奥秘等终极科学问题有待破解 [6]。因此，在危险与机遇并存的 AGI 研究过程中，必须面向 AI 研究人员、程序开发人员，制定有效的保障与行为准则。

对 AGI 可能的安全性风险进行评估并制定适宜对策，探讨有效驾驭 AGI 并使之既造福于人类又不对社会造成危害的举措，已经成为世界性的研究议题。例如，美国 OpenAI 团队 2016 年分析了 AI 发展过程中可能遇到的安全问题 [7]，随后美国政府成立了人工智能安全委员会 [8]；欧盟设立了人工智能高级别专家组，争取技术发展的话语权和规则制定权 [9]。此外，AI 也成为国防领域的重点关注对象，如采用 AI 手段提高防御系统能力，发展 AI 异常检测技术用于防止隐私数据被恶意篡改，研究涉及多学科融合算法、自适应态势感知能力、人机信任等方面的 AI 理论与技术 [10]。

也要注意到，针对 AGI 安全问题，我国相比国际前沿进展存在一定差距；国内学术界、产业界较多专注于 AI 的发展，很少关注 AGI 安全性保障的价值和需求。本文从模型的不可解释性、硬件与算法的不可靠性、自主意识的不可控性三方面对 AGI 的来源进行剖析，从能力、动机、行为 3 个维度对相应安全风险进行评估，分别从理论、应用层面提出降低相关安全风险的发展建议。

## 二、强人工智能的安全风险来源

### （一）模型的不可解释性

在传统 AI 方面，DeepFake [11] 的欺骗效果已经获得广泛认可，有专门研究据此开展梯度攻击和防御。梯度攻击的本质在于，鉴于卷积神经网络（CNN）的基本处理单元是纹理，故针对纹理产生不同的响应来进行不同的操作。在对抗生成网络 [12] 的训练中存在目标偏见现象，即“太难的东西就不生成了”[13]。

如果一个系统不可解释，则无法确认运行过程是否会被其他因素干扰而影响其运行目的。例如，基于类脑机的诊断系统，对病人的病情进行分析后作出了诊断，诊断结果的可靠性只能从统计学的角度去估计；由于无从得知模型是基于病人的哪些因素才做出的诊断，因而很难完全信任机器推断出的结果。类脑机是 AGI 的基本实现途径之一 [14]；脉冲是类脑机信号的载体；在对类脑机进行分析的时候，尚不确定是否存在一定的噪声序列类干扰分类的结果（以叠加后波峰、波谷相位不改变为前提）。在 AGI 的训练过程中也可能有类似的问题，这种模式坍塌就会存在被恶意利用的风险。因此，模型的不可解释性可能是 AGI 系统的潜在安全风险。

### （二）算法和硬件的不可靠性

AGI 的发展和应用将对当前的产业格局、居民生活方式构成很大影响，而已有的 AGI 算法和硬件

尚不能满足安全可靠、契合预期的要求。

在设计算法的过程中，设计方案不成熟（如没有考虑到所有可能出现的情况、软件与硬件存在兼容性等）会导致系统崩溃。欧洲的运载火箭发射，曾因高精度数据超过硬件支持的位数而造成任务失败 [15]。

在 AI 专家系统服务社会时，系统所依赖的前提假设可能会在某些特殊情况下失效，从而造成系统崩溃。美国华尔街“闪电崩盘”交易事故因错误的前提预设导致了股票价格设定的严重错误，造成损失超过一万亿美元，严重影响了美国证券市场 [16]。

算法与硬件的信息安全性成为维护经济社会公共安全的重要保障。有新闻揭露，黑客利用系统漏洞从机构和公司盗取个人信息、隐私数据，社会影响恶劣。可以合理推论，当 AGI 广泛用于生产生活后，很大可能受到黑客、恶意软件的攻击，产生数据泄露后果甚至危害公共安全。

### （三）自主意识的不可控性

构建初始智能体、有效进化准则，是能够自我发展、自我迭代 AGI 系统设计的关键。人类可以很好地控制初始智能，但是 AGI 可以自主设计进化规则，这种设计进化规则的效率可能足以碾压人类。自我发展后的 AGI，在后续阶段的发展效率将会更高，通过递归地自我改进而使其远超人类认知。

具有自主意识的 AGI 具有潜在风险。不同于人脑，AGI 的计算和分析能力在理论上是没有边界的，具有高效的数据收集、处理、分析能力，可理解看到、听到、接收到的所有信息。AGI 被赋予自主意识后，可通过交流、沟通的方式进行信息的分享与交换，显著提高对世界的认知、理解与改造效率。相应地，人类的各种活动都有可能逐步被 AI 取代。由于自主意识的呈现，AGI 的法律定位出现了模糊：将其视为有意识的主体，还是个人的私有财产？这可能在法律、伦理、政治层面引入分歧，从而引发难以预料的后果。

## 三、强人工智能的安全风险评估

### （一）能力的风险评估

汉斯 · 莫拉维克提出了“人类能力地形图”观点 [17]，据此描述人类和计算机的能力发展以及面临各类问题的难度。在“人类能力地形图”中，海拔高度代表某项任务对于计算机的难度；不断上涨的海平面表示计算机当前能做的事情，海平面上升将有临界点；当计算机能够自主设计智能时，即临界点到达。在临界点之前，算法设计主要由人类来掌控；超过临界点后，将由计算机来代替人类研发智能，体现从数量到质量的飞跃，相应生产力与生活水平也将发生剧烈改变，

需要注意到，人类依照一定的原理和经验来构建算法，而 AI 设计的算法将无法始终保证程序的可靠性。对于使用者来说，当前的 AI 类似“黑盒”，无法或很难去探索内在的运行逻辑和决策依据。

### （二）动机的风险评估

在人类文明的发展过程中，人类智慧及其产物极其珍贵。充分利用 AI 来提高生产力、创造新工具，可以相信生活因此更加美好。一些具有颠覆性特质的技术都是源自微小的改进或创新，但对生产力进步起到显著的促进作用。然而，合理应用 AI 带来的生产力进步和技术飞跃，而又不引入新的社会问题，这是人类应当高度关注的议题。例如，如何建造鲁棒的 AGI ？如何掌控 AI 武器并避免陷入恶性军备竞赛？如何让 AI 的生产力应用不会加剧社会分配的不平等现象？

就 AI 而言，在其能力弱小、可被人类控制的阶段，不必担心对人类造成危害。当 AI 的各方面能力超过人类、和人类一样拥有意识后，就很难判断是否必然继续听从人类命令，这种情况称为“背叛转折”[18]。AI 是否具有人类意识、依靠何种方式实现类人意识，尽管尚属未知，但同样值得关注和研究。

### （三）行为的风险评估

对 AGI 行为的监督和控制，可视为一类“委托 – 代理”问题，即人类是委托方，AGI 系统是代理方。这与当前人类实体的“委托 – 代理”问题性质不同，即 AGI 可根据自己的分析能力、知识储备来自行制定差异化的策略与行动。因此，监测 AGI 在研发初期的测试行为，并不能支持人类合理推测 AGI 未来的可靠性。如此，行为主义方法可能失效。

## 四、强人工智能在理论及技术研究阶段的风险防御策略

### （一）完善理论基础验证，探索模型的可解释性

完善理论基础验证、探索模型的可解释性，是 AGI 正确性的构建基础，也是 AGI 安全的形式化保障。

应以认知神经科学为基础，探索 AGI 的模型设计。认知神经科学是基于大脑的生物结构、人类的认知能力，研究脑构造、探索脑运行方式的学科；借鉴人脑结构和运行方式，可设计适当的 AGI 模型。

应以元学习为基础，探索 AGI 的实现方法。元学习是学习“学习方法”的方法 [19]，可赋予 AI 思考和推理的能力；作为当前深度学习的重点研究方向，旨在从数据中学习相关知识，将自动学习新知识的能力赋予当前的 AI。对于当前的 AI，一项新的任务往往意味着从零学习新的知识，费时且灵活性低。元学习则是经验导向，基于过去的经验去学习新任务的解决办法，可使 AI 掌握更多技能、更好适应复杂的实际环境。元学习作为半监督、无监督学习的实现方式之一，是模拟人类学习过程的重要数学实现；寻求通过数学方法模拟人类学习过程的手段，据此提高模型的可解释性，探索让 AGI“学会学习”，像人类一样“产生自主意识”。

应从数学的角度来探索深度学习的可解释性。目前并没有一套受到公认、体系完整的用于解释深度学习的理论框架，相关模型的可解释性仍被视为复杂问题。从数学角度探索深度学习的可解释性，已有方法包括信息论、结构表达、泛化能力、动力学原理、流形学习等。后续，探索模型各个组成模块的功能和贡献、从语义角度对模型的架构和功能进行模式化分析，是 AGI 可解释性研究需要重点关注的内容。

### （二）严格控制强人工智能的底层价值取向

AGI 的底层价值取向需要通过相应的规则、记忆来进行限制和监控。

应设计明文规则，限制 AI 的行动范围。鉴于 AI 的复杂性、不可解释性，很难从源代码角度对其价值取向进行限制和监控。从行为角度对 AGI 的价值取向进行限制，通过明文规则来限制 AGI 的行为能力和动作权限，是重要的研究目标。在元学习的过程中，可构建底层的价值观网络来加速推理，指导行动网络采取行为。关于底层的价值观网络，算法具有复杂性，数据集存在不可控性，很难采取措施对其推理过程进行限制。关于行动网络，可人为加入明文规则，确保在原子行动上符合正确的价值观（即针对每一个独立动作，限制错误行为的出现）。

要应用可信计算技术，监控 AI 的行动内容。可信计算是一种针对恶意代码、恶意攻击的防御机制，可视为计算机的“免疫系统”；引入额外监督，对计算机的各种行为建立完整、可信、可量化的评价机制，据此判断各种行为是否符合人类的预期、对不可信的行动进行防治；应用于 AI 的行动过程监控，即可认为具备正确价值观的行为是合理可信的。监控并分析 AGI 行为的运行过程，通过时间序列来判断当前行为是否具备合理的价值取向；如不符合，采用外部干预的方式干扰或打断 AGI 的当前行动，确保 AGI 不会做出违背价值观的行为。

### （三）实现技术的标准化

一是模型设计的标准化。当前，深度学习和 AI 研究形成了一些获得广泛应用的基础模块，如 3×3 规格 CNN、线性整流函数、批量归一化等，采用不同的基础模块可构造出差异化的神经网络。对基础模块进行标准化设计，一方面有利于统一接口和配置文件设计，使用通用的描述语言来表示神经网络过程，方便模型的迁移和部署；另一方面有利于采用硬件芯片、驱动程序进行针对性的加速处理。以 CNN 为例，统一计算设备架构（CUDA）以及据此发展的深度神经网络库（cuDNN），对于 3×3 规格的卷积计算采取加速措施，显著提高了训练和推理的速度。

二是训练方法的标准化。训练是 AI 必不可少的环节，不同的网络可通过不同的训练参数、优化器、策略来求解权重。训练具有多样性，一方面使得模型的复现性普遍较差，另一方面导致优化器在迭代过程中无法得到硬件加速支持。训练方法标准化重在设计一套合理的训练框架，将不同的优化器抽象成接口，对统一的接口进行硬件层面的加速支持，据此提高模型的训练效率。

三是数据集的标准化。主要指各行业提出的公开、标准、具有共识的数据集，面向公众发布，用

于模型的训练和测试。数据集的标准化，一方面可推动数据的安全保障力度，另一方面可提高数据集的质量水平。推进各行业的标准数据集制定工作，形成公开且高质量的基准，具有重要意义。

四是安全保障的标准化。AGI 投入使用的必要前提是具有安全保障。应发展通用、明确可执行的标准来确保 AGI 设计、训练、运行的安全性。相应标准需具有良好的可扩展性，以适应 AGI 应用的环境复杂性。对安全保障进行标准化，针对不同阶段特点设计对应的方法，保障 AGI 的合理运行，这是对抗相关风险的最有力保证。

## 五、强人工智能在应用阶段的风险防御策略

### （一）预防人为造成的人工智能安全问题

近年来，AI 技术应用于造假逐渐得到关注，如采用机器学习技术便捷制作出真假难辨的造假视频（DeepFake）。有研究总结了传统图像取证、生理信号特征、图像篡改痕迹、生成对抗网络（GAN）图像特征等检测伪造技术 [20]。目前在伪造图像检测方面的研究取得进展，但新型伪造技术的出现给深度伪造的鉴别工作带来了更大困难；只有尽可能建立技术优势，鉴别者才能赢得造假者。此外，可采用司法立法、新闻行业培训等辅助手段来应对技术应用伴生的安全问题。

算法设计方面可能存在的疏漏也应引起重视。尽管 AI 的应用能力已经获得证明，但相应算法设计难免“百密一疏”，应将确保安全置于首位，特别是在自动驾驶、远程医疗、工业制造等与人的生命安全直接相关的领域。已经出现了民航飞机自动驾驶系统存在错误且操控权无法切换至人工操作而导致重大事故的案例。在进一步发展 AI 技术并拓宽应用范围的背景下，必须从源头纳入安全问题，防范系统、数据可能遭受恶意攻击或者受到某些错误信号干扰而可能造成的严重后果。

引入第三方组件可能会引发安全问题。这既属于传统安全的范畴，也是影响 AI 安全性的重要因素。恶意的第三方组件，可能造成 AI 系统崩溃、系统权限被盗取等问题。

### （二）对强人工智能进行动机选择

“背叛转折”阶段的 AI 已经具有在各个领域都远超人类的认知能力，可称为超级 AI [18]。基于超级 AI 可能会背叛人类的合理猜想，人类应当提前对智能体的动机进行选择，全力制止不良结果的出现；应使超级 AI 具有不对人类造成危害的自发意愿。

针对动机选择问题，当前研究讨论提出了直接规定、驯化、扩增、间接规范 4 种应对方式 [18]。①直接规定细分为基于规则、结果主义。基于规则方式的传统描述即为“机器人三定律”[21]；就第一条“不能伤害人类”而言，权衡对人类的伤害、“伤害”“人类”的定义、不考虑其他有情感动物与数字大脑的原因等，都未阐释清楚。为了制定一套复杂、详细的规则并应用于高度多样化的情境，且强调第一次就成功，基于目前的条件来看不太可能。结果主义方式也面临问题，因为达到相同的结果有很多不同的途径，计算机代码必须精准描述目标。例如，AI 目标是让人保持微笑，但让人开心、仅仅通过肌肉刺激来保持微笑，其状态有着显著不同。②驯化可视为自我限制 [18]，作为一种特殊的最终目标，尝试去塑造系统的动机以限制其野心，最终令其自主地将行为限制在规定的范围。③扩增指基于动机良好的已有智能体，通过改造来进一步提升各项智能行为的方式，相应劣势在于很难保证动机系统在认知能力得到巨大提升后不被改变或破坏。④间接规范不同于直接规定，制定能够产生标准的程序，让 AI 自行推理规范的建立过程。

### （三）为强人工智能赋予人类价值观

相比限制 AI 的能力，动机选择已经在一定程度上提升了人类控制 AI 的有效性，但仍面临一些问题。例如，AI 可能面对无穷多种情况，不可能具体讨论每一种情况下的对策，而人类本身不可能持续监视 AI 的动机。可行的思路之一是将人类的价值观赋予 AI（加载到 AGI 内部），让其自觉地执行那些不对人类构成威胁的事件。无法将各种情况下的动机系统均完整具象为可以查询的表格（导致无穷大的表格），只能使用公式、规则等进行更为抽象的表达。进化算法可能是加载价值观的可行途径之一，随机产生一些规则，通过评估函数进行候选筛分（去掉得分低的、保留得分高的）。强化学习方法可使智能体的累积回报最大化，在驱动智能体去学习处理各类问题的同时，进行价值观积累。

然而，人类价值观的积累过程是人类相关基因机理经历成千上万年进化的结果，模仿并复现这一过程非常困难；这一机理与人类神经认知体系结构相适应，因而只能应用于全脑仿真 [22]。全脑仿真的前提是大脑可被模拟、可以计算，面临着扫描、翻译、模拟 3 类条件的制约 [18]，采用高通量显微镜、超级计算系统才能达到所需精确度。

### （四）强人工智能国际合作

AGI 研究已经成为国际性的关注点，集中全人类的科技力量来推进 AGI 的深化研究，才能使 AGI 更好服务人类社会。相关研究和逐步应用的过程，将面临许多未知问题。加强 AGI 国际合作、促进研究成果共享，才能根本性地提高应对突发情况的能力，也才能真正保障 AGI 的应用落地和拓展。

目前，AGI 国际合作的重要性已经得到高度重视，一些国家和地区通过立法等形式为国际合作提供政策支持。例如，欧洲 25 个国家签署了《人工智能合作宣言》[23]，承诺开展合作、促进对话，争取就各国之间的 AI 研究与应用合作达成一致；还通过联合声明等方式促进优先领域的立法合作，包括数据保护、伦理标准、数据权利等重点问题。这些做法都是我国开展 AGI 国际合作的有益参照形式。

### （五）强人工智能人才培养

人才培养是科学研究的基础条件。AGI 作为前沿科技方向，相应人才培养的规模、速度、质量显然无法满足领域发展需要；亟待加强人才培养，尤其是本土人才。在技术领域，优化人才教育、培养、成长周期方面的机制和环境，快速发展一批具有专业研究和开发知识的专业人员；在管理领域，注重培养体现商业推广和需求扩展特征的企业家及运营人才；通过“产学研用”协同，为 AGI 的健康稳定发展提供坚实的人才保障。

## 六、结语

AGI 的智慧与行为不能简单地与人类划等号，创造 AGI 的动机是为了更好地造福人类社会。对于人类社会的隐私，应控制 AGI 只能给人类提供被动的服务，而不是主动的学习。如果 AI 进化到一定水平后出现智能爆发，默认后果必然是造成确定性灾难。面对这样的潜在威胁，人类应持续关注并着力寻求应对方法，坚决避免这种默认结局的出现；设计出受控制的智能爆发，设置必要的初始条件，在获得人类想要的特定结果的同时，至少保证结果始终处于人类能接受的范围。

着眼未来发展，建议持续关注 AGI 的技术演进路线，对技术伴生的潜在安全风险提出动态的应对策略；参考国际性的 AGI 政策研讨和制定过程，结合法律、伦理方面的前沿成果，更为及时、深刻地探讨我国 AGI 政策的制定要素。

**参考文献**


[1] 陈俊波, 高杨帆. 系统论视角下的人工智能与人类智能 [J]. 自然辩证法研究, 2019, 35(9): 99–104.
Chen J B, Gao Y F. Artificial intelligence and human intelligence from the perspective of system theory [J]. Studies in Dialectics of Nature, 2019, 35(9): 99–104.

[2] 黄铁军, 余肇飞, 刘怡俊. 类脑机的思想与体系结构综述 [J]. 计算机研究与发展, 2019, 56(6): 1133–1148.
Huang T J, Yu Z F, Liu Y J. Brain-like machine: Thought and architecture [J]. Journal of Computer Research and Development, 2019, 56(6): 1133–1148.

[3] 张钹. 走向真正的人工智能 [J]. 卫星与网络, 2018 (6): 24–27.
Zhang B. Towards the real artificial intelligence [J]. Satellite & Network, 2018 (6): 24–27.

[4] 徐宗本. AI与数学“融通共进”迈向自主智能时代 [EB/OL]. (2020-06-08)[2021-02-15]. http://news.sciencenet.cn/htmlnews/2020/6/441057.shtm.
Xu Z B. AI and math go together towards the era of autonomous intelligence [EB/OL]. (2020-06-08)[2021-02-15]. http://news.sciencenet.cn/htmlnews/2020/6/441057.shtm.

[5] 周志华. 关于强人工智能 [J]. 中国计算机学会通讯, 2018, 14(1): 45–46.
Zhou Z H. Views on artificial general intelligence [J]. Communication of the CCF, 2018, 14(1): 45–46.

[6] 黄铁军. 也谈强人工智能 [J]. 中国计算机学会通讯, 2018, 14(2): 47–48.
Huang T J. Different views on artificial general intelligence [J]. Communication of the CCF, 2018, 14(2): 47–48.

[7] Amodei D, Olah C, Steinhardt J, et al. Concrete problems in AI safety [EB/OL]. (2016-07-25)[2021-02-15]. https://arxiv.org/abs/1606.06565.

[8] Congress of the United States. H.R.5356-National security commission artificial intelligence act of 2018 [EB/OL]. (2018-03-20)[2021-02-15]. https://www.congress.org/bill/115th-congress/house-bill/5356.

[9] 中国信息通信研究院. 全球人工智能治理体系报告 [EB/OL]. (2020-12-30)[2021-02-15]. https://pdf.dfcfw.com/pdf/H3_AP202012301445361107_1.pdf?1609356816000.pdf.
China Academy of Information and Communications Technology.

Global AI governance report [EB/OL]. (2020-12-30)[2021-02-15]. https://pdf.dfcfw.com/pdf/H3_AP202012301445361107_1.pdf?1609356816000.pdf.

[10] 金晶, 秦浩, 戴朝霞. 美国人工智能安全顶层战略及重点机构研发现状 [J]. 网信军民融合, 2020 (5): 45–48.
Jin J, Qin H, Dai Z X. Top-level strategy of artificial intelligence security and the research status of key institutions in the United States [J]. Civil-Military Integration on Cyberspace, 2020 (5): 45–48.

[11] Whyte C. Deepfake news: AI-enabled disinformation as a multi-level public policy challenge [J]. Journal of Cyber Policy, 2020, 5(2): 1–19.

[12] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks [J]. Advances in Neural Information Processing Systems, 2014, 3(11): 2672–2680.

[13] Bau D, Zhu J Y, Wulff J, et al. Seeing what a GAN cannot generate [C]. Seoul: 2019 IEEE/CVF International Conference on Computer Vision, 2019.

[14] Huang T J. Imitating the brain with neurocomputer a “new” way towards artificial general intelligence [J]. International Journal of Automation and Computing, 2017, 14(5): 520–531.

[15] 曲晶, 张绿云. 国外火箭发射及故障情况统计分析 [J].中国航天, 2016 (2): 13–18.
Qu J, Zhang L Y. Statistical analysis of foreign rocket launch and failure [J]. Aerospace China, 2016 (2): 13–18.

[16] 邢会强. 证券期货市场高频交易的法律监管框架研究 [J]. 中国法学, 2016 (5): 156–177.
Xing H Q. Research on the legal regulatory framework of high frequency trading in securities and futures market [J]. China Legal Science, 2016 (5): 156–177.

[17] Tegmark M. Life 3.0: Being human in the age of artificial intelligence [M]. New York: Penguin Random House LLC, 2017.

[18] Bostrom N. Superintelligence: Paths, dangers, strategies [M]. Oxford: Oxford University Press, 2015.

[19] Vilalta R, Drissi Y. A perspective view and survey of meta-learning [J]. Artificial Intelligence Review, 2002, 18(2): 77–95.

[20] 李旭嵘, 纪守领, 吴春明, 等. 深度伪造与检测技术综述 [J]. 软件学报, 2021, 32(2): 496–518.
Li X R, Ji S L, Wu C M, et al. Survey on deepfakes and detection techniques [J]. Journal of Software, 2021, 32(2): 496–518.

[21] Asimov I. I, robot [M]. Louisville: Spectra Press and Promotions, 2004.

[22] 黄铁军. 人类能制造出“超级大脑”吗？ [N]. 中华读书报, 2015-01-07(5).
Huang T J. Can human build “super brain”? [N]. China Reading Weekly, 2015-01-07(5).

[23] 中华人民共和国科学技术部. 欧洲25国签署《人工智能合作宣言》[EB/OL]. (2018-07-18)[2021-02-15]. http://www.most.gov.cn/gnwkjdt/201807/t20180718_140708.htm.
Ministry of Science and Technology of the People’s Republic of China. 25 European countries sign the *Declaration on Artificial Intelligence Cooperation* [EB/OL]. (2018-07-18)[2021-02-15]. http://www.most.gov.cn/gnwkjdt/201807/t20180718_140708.htm.