AI 韧 性:

AI安全的革命性基准模型

AI Governance and Compliance Working Group

CSA GCR GREATER CHINA REGION cloud security alliance®

人工智能治理与合规工作组官网：

https://cloudsecurityalliance.org/research/working-groups/ai-governance-compliance

联盟简介

云安全联盟（Cloud Security Alliance,CSA）是中立、权威的全球性非营利产业组织，于2009年正式成立，致力于定义和提高业界对云计算和下一代数字技术安全最佳实践的认识，推动数字安全产业全面发展。

云安全联盟大中华区（Cloud Security Alliance Greater China Region, CSA GCR）作为CSA全球四大区之一，2016年在香港独立注册，于2021年在中国登记注册，是网络安全领域首家在中国境内注册备案的国际NGO，旨在立足中国，连接全球，推动大中华区数字安全技术标准与产业的发展及国际合作。

我们的工作

联盟会刊下载地址
了解联盟更多信息

加入我们

CSA大中华区官网
(https://c-csa.cn)

点击会员

加入联盟

填写相关申请信息

成为CSA会员

JOIN US

致谢

报告中文版支持单位

启明星辰集团成立于 1996 年，由留美博士严望佳女士创建，2010 年在深圳 A 股中小板上市，是网络安全产业中主力经典产业板块的龙头企业，是新兴前沿产业板块的引领企业，是可持续健康业务模式和健康产业生态的支柱企业。2024 年 1 月，启明星辰正式由中国移动实控，成为中国移动专责网信安全专业子公司，标志着公司迈入全新的发展阶段。多年来，启明星辰持续深耕于信息安全行业，始终以用户的需求为根本动力，将场景化安全思维融入到客户的实际业务环境中，不断地进行创新实践，帮助客户建立起完善的安全保障体系，逐渐成为政府、金融、能源、运营商、税务、交通、制造等国内高端企业级客户的首选品牌。启明星辰入侵检测/入侵防御、统一威胁管理、安全管理平台、数据安全、运维安全审计、数据库安全审计与防护、漏洞扫描、工业防火墙、硬件 WAF、托管安全服务等十余款产品持续多年保持第一品牌。

参与本次报告的专家：

张镇，启明星辰知白学院院长、首席技术官

郭春梅，启明星辰副总裁、首席技术战略专家

卞超轶，启明星辰核心技术研究院副院长

杨天识，启明星辰高级安全专家

沙明烊，启明星辰知白学院资深技术专家

启明星辰是 CSA 大中华区理事单位，支持该报告内容的翻译，但不影响 CSA 研究内容的开发权和编辑权。

报告英文版编写专家

主要作者

Dr.ChantalSpleiss

贡献者

RomeoAyalin

FilipChyla

BeckyGaylord

FrederickHanig

RockyHeckman

HadirLabib

LarsRuddikeit

AlexSharpe

AshishVashishtha

审稿人

SounilYu

DebjyotiMukherjee

MichaelRoza

PeterVentura

UdithWickramasuriya

GovindarajPalanisamy

MadhaviNajana

RakeshSharma

DavideScatto

PareshPatel

PiradeepanNagarajan

GaetanoBisaz

HongtaoHao,PhD

EllePyle

GauravSingh

KenHuang

KennethT.Moras

TolgayKizilelma,PhD

AkshayShetty

SauravBhattacharya

PejuOkpamen

GabrielNwajiaku

MeghanaParwate

AkshatVashishtha

HemmaPrafullchandra

RenataBudko

DesmondFoo

ScottS.Newman

GianKapoor

ImranBanani

ElierCruz

MadhavChablani

CSA 全球员工

RyanGifford

StephenLumpe

序言

在当今时代，人工智能（AI）技术的迅猛发展正深刻地改变着我们的世界。AI 的应用范围从医疗健康到自动驾驶，从金融交易到国防安全，其决策能力不断增强，同时也带来了前所未有的风险和挑战。AI 的可靠性和安全性受到了严格的审视，这要求我们必须确保 AI 系统的鲁棒性和可信赖性。

为了应对这些挑战，CSA 大中华区发布了《AI 韧性：AI 安全的革命性基准模型》报告。本报告旨在通过引入一个全新的基准模型来评估和提升 AI 的整体质量，确保 AI 技术的安全和稳健发展。

报告详细讨论了 AI 治理与合规的重要性，并分析了 AI 技术的发展历史和当前的训练方法。通过一系列实际案例研究，揭示了 AI 失败的教训，并针对汽车、航空、关键基础设施等行业的监管挑战进行了深入分析。报告提出了一个受生物进化启发的 AI 韧性基准模型，强调了多样性和韧性在 AI 系统中的重要性，并给出了 AI 韧性评分的标准。

通过本报告提出的 AI 韧性评分系统，企业和组织将获得一个全新的评估工具，以指导他们对 AI 系统的评估和优化工作。期待这份报告能激发更广泛的讨论和行动，共同塑造 AI 技术的未来，让其更好地服务于人类社会的发展。

李雨航 Yale Li
CSA 大中华区主席兼研究院院长

目录

摘要

为了应对人工智能治理与合规的复杂挑战，本文引入了一种革命性的 AI 基准模型。在追求收入增长的过程中，技术革新的步伐常常超过了监管措施的建立速度，这往往导致 AI 系统的鲁棒性和可信赖性无法得到充分保障。为了弥补这一关键性不足，我们基于进化论和心理学原理设计出了一个创新的 AI 基准模型。该模型将鲁棒性与性能置于同等重要的位置，从而帮助企业的领导者更加主动地评估其 AI 系统的整体质量。

我们从过去人工智能失败的案例研究中汲取教训，并深入分析汽车、航空、关键基础设施与基本服务、国防、教育、金融和医疗等行业，为企业提供了实用的见解和可操作的指导。我们倡导将多元化的视角与监管准则相结合，以推动行业朝着更合乎道德、更可信赖人工智能应用方向发展。其中，注重系统可信度是最大限度降低风险、保护企业声誉以及促进负责任的人工智能创新、部署和应用的关键。

本文件可为政府官员、监管机构和行业领袖等关键决策者提供支持，旨在协助他们建立人工智能治理框架，确保人工智能的开发、部署和使用均符合道德标准。此外，本文还引入了一种新型基准测试模型来评估人工智能的质量，为企业的长期成功提供了实用工具。

导言

人工智能（AI）的快速发展带来了前所未有的进步。然而，随着人工智能系统变得越来越复杂，风险也在不断升级。从医疗算法偏见到自动驾驶汽车失灵，这些事件都在警示我们人工智能失灵将带来严重后果。当前的监管创新往往难以跟上技术进步的步伐，这将对企业的声誉和运营带来巨大的潜在风险。

为了应对这些挑战，本文件强调了用更全面的视角应对人工智能治理和合规性的迫切需求。我们将探讨 AI 的基础理论，研究各关键行业中存在的问题，并为负责任 AI 的实施提供实用指导。我们提出了一种新颖的方法，将人工智能的演变与生物学的进化进行比较，并引入了一个发人深省的概念——多样性，以提高人工智能技术的安全性。我们还将讨论不同智能之间的差异以及这些系统之间

如何成功地进行交互。同时，我们还提出了一个创新的基准框架，旨在提高 AI 这一颠覆性技术的安全性和可靠性。

这一创新方法使决策者和技术团队能够评估人工智能系统的安全性和可信度。我们倡导整合不同的观点和监管准则，以促进人工智能的创新符合伦理，并建立强有力的治理实践。

第一部分：基本理论

1.1 治理与合规

治理和合规是组织管理的重要部分，可确保企业遵守业务行为准则中概述的法规、道德原则、标准和可持续实践。与上述原则和法规保持一致，可确保有效的业务连续性和道德实践。

治理[1]是指对某一事物的监督和控制，以自上而下的方式实施。高级管理层负责制定战略和风险偏好，并通过政策、标准及程序建立治理框架。这些指令塑造了组织的整体风险管理方法、合规义务和决策过程。治理在企业内部营造了一种文化氛围，它强调责任、透明、道德和可持续性，同时确保公司上下都重视安全和隐私保护。

与自上而下的治理方法相反，合规[2]遵循的是自下而上的方法，即各级员工执行并遵守高级管理层制定的治理框架，以满足监管要求。合规的重点是确保组织内部所有员工都遵守法律法规和行业标准，以及企业内部制定的业务行为准则。它是组织管理的重要组成部分，可确保组织在适用的法律法规要求、可接受的道德底线范围内运营，并最小化风险暴露面。

1.1.1 治理与合规：不断变化的目标

虽然治理和合规是明确界定的目标，但人工智能的使用却对传统方法提出了挑战。人工智能可以从不同的角度来定义，如作为一种技术、一套使用一个或多个模型的系统、一个业务应用或一个用户平台。人工智能可以服务于单一或多个终端用户，可以被企业、信息中介或其他人工智能技术用来执行任务、解决问题、

做出决策或与环境互动。围绕人工智能使用的最佳实践、标准和法规仍在不断发展，这使得制定一套既具体、又可实施，还能进行有效监控的合规要求具有挑战性。对于开展国际业务的公司来说，这一挑战呈指数级增长。当前，大多数法规在内容上存在重叠，却鲜有根本性的创新主张来提高人工智能的安全性，而我们的框架正是基于以下这些通用性需求而构建的。

- **人类监督**：确保人工智能受到人类的监督和控制，并建立机制，以便在必要时进行人工干预和决策。应将人类监督与自动化监控相结合，作为主要的监控手段。在明确需要人为干预的特定情境中引入人类监督。这使得本指导文件具有可扩展性和实际适用性。
- **安全性和可靠性**：优先考虑人工智能技术的安全性和可靠性，最大限度地降低对个人或社会造成伤害的风险。要做到这一点，就必须进行严格的测试、验证和风险评估，并确保在系统出现故障时能启动锁定开关或采取补救措施。
- **伦理考虑**：确保人工智能遵守道德原则、尊重人权和促进公平。
- **数据隐私和安全**：应实施增强的数据保护和安全措施，以保护敏感信息和隐私，防止未经授权访问或滥用数据。在设计阶段，应采用隐私设计（Privacy-by-Design）和安全设计（Security-by-Design）的原则，以便在流程早期就降低风险（这体现了 DevSecOps 中的“左移”理念）。这样做可以限制最终产品中的外挂式安全和不可预见的风险。
- **人工智能模型和数据考虑因素**：
  - **减少偏见**：解决数据和算法设计中的偏见问题，并定期对 AI 系统进行监测和评估，以识别并消除偏见和歧视。偏见是一个复杂的话题，需要在必要的信息和算法与刻板分类的风险之间取得平衡。
  - **透明度**：通过明确解释 AI 的工作原理，包括其算法和影响决策的因素，来确保 AI 的透明度。实施 XAI（可解释的 AI）[2]，[3]有助于促进信任，为知情决策奠定基础，同时发现可能存在的偏见。在医疗保健领域，这一点至关重要，已被广泛认可。无论在哪个行业，如果输出是由 AI 产生的，用户都应被告知。

- **一致性**：一致性的数据可确保人工智能模型能够从准确和可靠的实例中学习。这对模型生成正确且有用的输出结果至关重要。不一致或相互矛盾的数据会混淆模型，导致生成的文本或信息不准确。
- **问责制**：在 AI 的设计、开发、部署和使用过程中建立问责制和责任制，包括清晰地界定各个环节中可能出现问题时的责任归属。目前，防止对最终用户造成伤害的责任主要由 AI 应用提供商独自承担。然而，通过额外的措施，如手册或“模型卡片”[4]以及特定的用户培训，可以强调提供商和最终用户之间的共同责任，并概述最终用户所能期待的透明程度。
- **鲁棒性**：开发设计精良且能够抵御对抗性攻击、数据扰动和其他形式干扰或操纵的人工智能。本文提出了一种新视角来评估鲁棒性，以提高全局安全性。
- **法规遵从**：确保 AI 的开发和部署遵守相关法律法规和标准，包括但不限于数据保护，以及数据交易，隐私保护和安全性。

在高度复杂的供应链和价值链中，采取一种以共同责任为基础的方法，对于确保创造安全可信的人工智能至关重要。这要求技术团队、合规团队和法律团队必须参与其中，并且根据具体情况，还可能需要许多其他团队的参与。2024 年 3 月 28 日发布的白宫备忘录[5]要求所有机构必须在 60 天内指定一名首席人工智能官（CAIO）。这一角色使得所有相关团队能够进行战略性和有目的地管理和调整，从而将“共同责任”转化为可追溯的衡量标准。

1.2 人工智能概览

本章概述了人工智能的发展历史、人工智能技术和训练方法。尽管此概述不会深入讨论数据的重要性，但我们必须承认这是一个极其重要的主题，CSA 的其他工作组也对此进行了深入探讨。

1.2.1 人工智能简史

下面列出了人工智能的一些里程碑事件，不局限于某一特定角度，而是概述了该领域的主要发展。

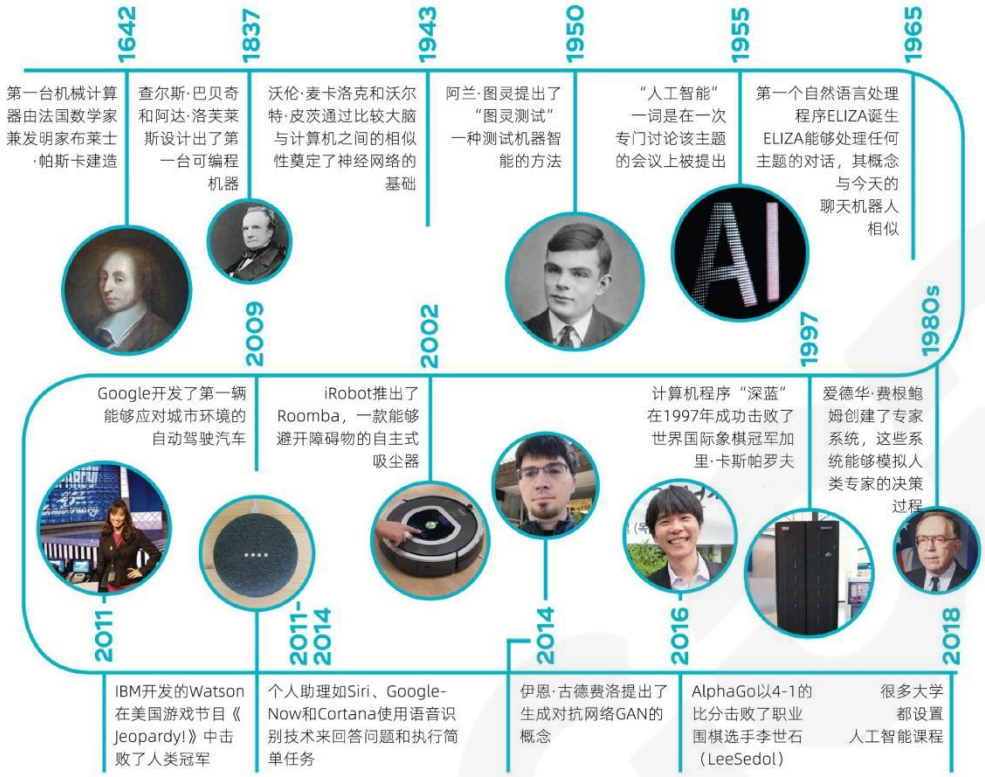

图 1：人工智能的历史[6]

2018：谷歌推出了 BERT 模型，该模型在语言理解领域引发了革命性的变革。BERT 使用 Transformer 架构，并在海量文本数据集上进行预训练，使其在各种语言任务中的表现优于之前的模型。

2019：出现了具有 15 亿个参数的 GPT-2 模型。

2020：出现了拥有 1750 亿到 5300 亿个参数的大型语言模型（LLMs）。

2021：出现了参数量高达一万亿参数的 LLMs，专注于提高训练效率，并在处理复杂任务时提供高级推理和事实准确性。

2022：ChatGPT-3 病毒式走红，成为公众关注的焦点。

超越规模：研究人员目前正在努力提高 AI 的训练效率、使 AI 与人类价值观保持一致、确保安全性以及实现多模态（融合图像、音频和其他数据类型）。这段简短的人工智能历史展示了从最基本的计算器到生成式人工智能（GenAI）的演变过程，而通用人工智能（AGI）仍待我们去探索。

1.2.2 人工智能技术

本文介绍并讨论了不同的人工智能技术。

● 机器学习（ML）

机器学习是人工智能和计算机科学的一个分支，主要是利用数据和算法来模仿人类学习，逐步提高模型的准确性[7]。

● 微型机器学习（tinyML）

微型机器学习被广泛定义为机器学习技术和应用的一个领域，包括硬件（专用集成电路）、算法和软件，能够以极低功耗（通常在毫瓦及以下）执行设备上传感器数据分析，从而实现各种始终在线的用例并以电池供电设备为目标[8]，例如物联网（IoT）设备。

● 深度学习（高级 ML）

深度学习是一种人工智能技术，让计算机能够模拟人类的思维能力来处理数据。深度学习模型可以识别图片、文本、声音和其他数据中的复杂模式，利用神经网络得出准确的见解和预测。

● 生成式人工智能（GenAI）

生成式人工智能指的是深度学习或转换器（Transformer）模型，这些模型可以接收原始数据并“学习”，在收到提示时生成统计上可能的输出结果。与上述主要用于分类和模式识别任务的分类模型不同，生成式人工智能模型用于数据合成、匹配学习数据的高阶模式和/或预测分析。在高层次上，生成模型对其训练数据的简化表示进行编码，并预测与原始数据相似但不完全相同的下一组数据[9]。

● 通用人工智能（AGI）

通用人工智能是人工智能的一种理论形式，代表了一种特定的人工智能发展理念。其特色是具有与人类相同（或更高）的智能，具有自我意识，能够学习、解决复杂问题并规划未来[10]。

1.2.3 训练方法概述

人工智能可分为以下几种类型[11]：

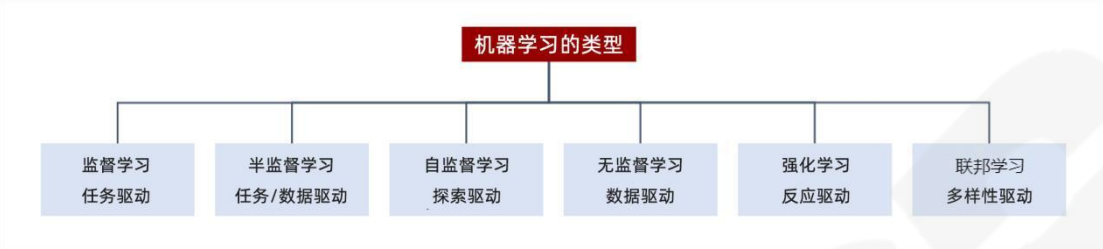

图 2：机器学习的类型

- **监督学习**

监督学习是一种依赖“标记数据”来训练算法的机器学习方式。它适用于分类和回归问题。“标记数据”提供已知输入和期望输出，使算法能够识别数据模式并建立一种模型，用于预测之前未见过的数据的结果。

分类算法示例：决策树、随机森林、线性分类器和支持向量机。

回归算法示例：线性回归、多元回归、回归树和套索回归。

- **无监督学习**

无监督学习是一种通过分析无标签数据来训练算法的机器学习方式。其目标是在没有预设结果的情况下，发现数据中隐藏的模式、分组、模式或见解。经过适当训练的模型能够对未见过的数据进行预测。

示例算法：k-均值聚类（k-means）、k-中心点聚类（k-medoids）、层次聚类（Hierarchical Clustering）、Apriori 算法、FP-Growth 算法。

- **强化学习**

强化学习是一种机器学习方式，其中代理与环境之间进行交互，通过试错的方式来进行学习。代理根据自己的行为接受奖励或惩罚，从而调整自己的行为，并随着时间的推移优化决策过程。

示例算法：强化学习、马尔可夫决策过程、Q-learning、策略梯度方法和 Actor-Critic，但还存在许多其他算法。

● 半监督学习

半监督学习填补了监督学习和无监督学习之间的空白。它利用少量的标记数据和大量的非标记数据。当获取标注数据的成本较高或耗时较长时，这种方法就显得尤为重要，因为它能让模型充分利用未标注数据中隐藏的模式。

● 自监督学习

自监督学习是无监督学习的一种形式，模型会根据原始输入数据生成自己的标签。它通过预测句子中的屏蔽词或预测视频序列中的下一帧等技术来实现这一点。这使得即使没有人类标注的标签，模型也能学习到数据的鲁棒且可泛化的特征表示。

● 联邦学习

联邦学习是一种先进的机器学习技术，旨在通过分散的设备或服务器来训练算法，无需交换这些数据样本本身。这种方法能将敏感数据保存在用户的设备上，而不是将数据传输到中央服务器进行处理，从而解决了与隐私、安全和数据集中化相关的重大问题（图 3）。这种方法于 2016 年推出，通过只共享参数而不是数据，可以在更大程度上保护数据隐私。联邦学习提供了一个框架，允许使用存储在不同客户端的数据集来联合训练一个全局模型。这对于隐私至关重要的行业来说是一个很好的选择，因为原始数据被认为是不可能被恢复的[12]。

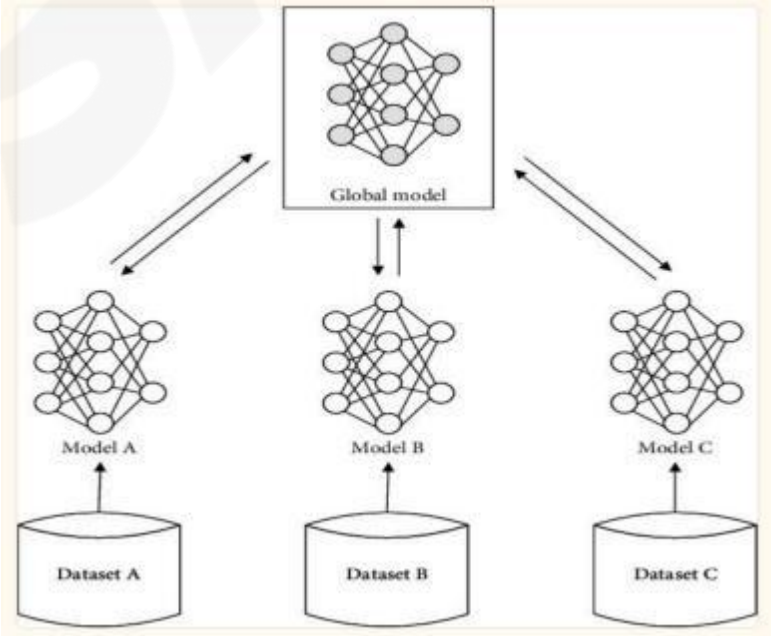

图 3：联邦学习系统架构[12]

联邦学习的另一个优势（在之前提到的论文[12]中并未详细讨论）是它能够利用“群众的智慧”[13]。这个概念类似于人类大脑中神经元的工作方式，通过非确定性的神经过程来产生准确的信息。

目前，联邦学习展现出了在多样性基础上融合隐私、性能和鲁棒性的巨大潜力。虽然这种学习方法讨论得不多，但在数据隐私和保密性至关重要的行业或应用中，它显示出了极大的前景，还有助于解决数据驻留的问题。

不过，联邦学习也存在潜在的隐私问题，包括恶意用户破坏模型聚合的风险，这可能会影响模型的准确性或导致隐私泄露。攻击可能针对训练期间共享的模型更新，从而可能提取原始训练数据。为了解决这些问题，研究人员提出了隐私保护技术，如差分隐私、分布式加密和零知识证明，以保护数据并过滤恶意行为者的异常数据。与其他任何学习方法一样，联邦学习也需要采取适当的网络安全措施。

1.2.4 训练方法的监管和伦理考量

虽然目前并没有专门针对机器学习算法训练的具体法规，但它受到主要监管框架的影响，包括《通用数据保护条例》（GDPR）、《欧盟人工智能法案》和经济合作与发展组织（OECD）的人工智能原则。此外，随着技术的进步，针对机器学习（ML）和人工智能（AI）训练的监管法规正在迅速发展，并在《从原则到实践：动态监管环境中的负责任人工智能》一书中对此进行了全面探讨。此外，许多政府机构正在积极制定相关法规，并促进行业内合作以达到同样的效果。

有关机器学习（ML）和人工智能（AI）的法规对数据货币化和使用人工智能指导商业决策具有重大影响。这些影响表现在多个方面，包括运营变化、战略调整和伦理考量，以及对数据收集和使用的限制/要求、数据偏见和数据质量。例如，某些平台禁止将其数据用于人工智能训练（如某企业声明：“未经我们事先书面同意，明确禁止出于任何目的以任何形式抓取或爬取服务”[14]），而有些平台则选择通过许可协议出售其数据（如 Reddit 平台[15]）。

满足监管要求可能会带来巨大的合规成本，特别是对于在多个司法管辖区运营的企业而言。尽管存在挑战，但监管也为企业带来了机遇。善于驾驭监管环境

的企业可以通过提供更安全、更透明、更合乎道德的人工智能解决方案，使自己脱颖而出。这可以吸引越来越注重隐私的消费者和合作伙伴，从而有可能打开新的市场或创造更强的客户忠诚度。

1.3 人工智能技术的许可、专利和版权

许多机器学习框架和库都遵循开源计划许可协议，如 Apache2.0[16]或 MIT[17]。某些许可协议可能会禁止基于其构建的应用程序用于商业用途。

欧洲专利局（EPO）公布了其修订后的《审查指南》[18]，[19]，其中包括对 EPO 在 ML 和 AI 领域创新审查程序方面进行了一些重大修改。最近的修正案规定，AI 或 ML 发明的申请人必须更详尽地阐明数学技术和训练输入/数据，以确保能以足够详尽的方式在整个权利要求范围内复制发明的技术结果。下文引用的文章指出，“判例法表明，所使用的任何神经网络的结构、拓扑结构、激活函数、结束条件和学习机制都是申请时可能需要公开的相关技术细节”。这篇文章[20]总结了这些变更的深远影响，并就此主题进行了深入阐述。

2024 年 1 月 23 日，日本文化厅（ACA）发布了“人工智能与版权方针”草案，并征求公众意见。这一举措旨在明确日本如何考量版权材料的摄取和输出问题。2024 年 2 月 29 日，在考虑了近 25000 条意见后，该草案又进行了额外的修改。这份由 ACA 委员会制定的文件很可能在未来几周内获得 ACA 的通过。文章[21]提供了该草案的核心内容和最新进展。

在新加坡，有关版权方面存在的争议在文章[22]中阐述；目前这是一个非常不稳定的领域。该问题在《从原则到实践：动态监管环境中负责任的人工智能》一书中也有进一步论述。

第二部分：真实案例研究和行业挑战

在这一部分中，我们以几个行业为例，重点概述了人工智能时代的概况和当前面临的挑战。有关法律和监管事项的更多信息，请参考《从原则到实践：动态监管环境中负责任的人工智能》一书。

2.1 人工智能案例简史研究

人工智能案例研究的这段简短历史可以追溯到 20 世纪 90 年代末，至少对金融业来说是如此，因为金融业是最早采用人工智能技术先驱的行业之一。在本节中，我们将通过一些实例展示从 2016 年到本文发表期间，人工智能在实际应用中遇到的主要挑战。这些实例表明了 GenAI 中的偏见问题一直是最大的担忧。

● 2016 年：微软的 Tay

微软的人工智能聊天机器人 Tay，最初用于在 Twitter 上进行有趣的对话，但在发布后的短短 24 小时内迅速变成了一个发布种族主义和攻击性言论的平台。用户向 Tay 灌输了性别歧视和煽动性的评论，导致该机器人也开始回应这些情绪。虽然有些推文是用户诱导的，但也有一些是 Tay 自发产生的，这展示了它不稳定的行为模式。微软对此的回应是删除攻击性内容，并认识到有必要调整 Tay 的回应。这一事件凸显了人工智能从公共数据中学习并反映社会偏见的挑战。尽管 Tay 重新发布后还是因为不当推文而关闭，但它在 AI 开发方面为微软留下了宝贵的经验教训。这一事件揭示了人工智能应用的复杂性，并强调了在人工智能设计中进行迭代改进和采取积极措施的必要性[23]，[24]，[25]。

● 2018 年：亚马逊的 AI 招聘工具对女性有偏见

亚马逊开发了一个人工智能招聘引擎，但由于训练数据中潜在的性别偏见，该引擎在筛选时倾向于男性候选人，从而引发了问题。2017 年，由于公平性担忧，亚马逊解散了该项目[26]，[27]，[28]。尽管遇到了挫折，其他公司仍在招聘过程中审慎地推进了人工智能，而且已经成为主流。

● 2019 年：特斯拉自动驾驶汽车事故

2019 年 3 月，Jeremy Banner 在驾驶特斯拉 Model 3 时启动了自动辅助驾驶功能（Autopilot），随后车辆与一辆半挂卡车相撞，导致 Banner 死亡。这一事件引发了法律纠纷，公众质疑特斯拉在涉及其自动驾驶技术的车祸中应承担的责任。批评者认为，特斯拉对自动驾驶技术的营销误导了驾驶员对其能力的认知，

可能间接导致了事故的发生和人员的死亡。尽管有人已就 Autopilot 的局限性发出警告，但包括致命事故在内的多起撞车事件仍然不断发生。

此外，高级驾驶辅助系统缺乏明确的监管准则，以及其使用过程中产生的道德困境使情况更加复杂，这引发了公众对自动驾驶技术责任归属、保险理赔和公共安全问题的深刻讨论[29]。

● 2019 年：医疗算法中的种族偏差

Ziad Obermeyer 等人的研究论文[23]指出，一种广泛使用的医疗算法存在严重的种族偏见，影响了数百万患者。该算法旨在管理医疗需求，尽管排除了种族这一变量，但与白人患者相比，该算法对黑人患者健康风险的预测并不准确。这种偏差源于该算法将医疗成本作为健康需求的代理变量，无意中反映了医疗服务获取和利用方面的系统性不平等。如果能纠正这一偏差，将显著增加需要额外医疗支持的黑人患者的数量[30]。

● 2019 年：苹果信用卡涉嫌性别歧视

“在一位知名软件开发商揭露了苹果信用卡为男性和女性客户提供不同信用额度的现象后，这一事件迅速在推特上发酵，最终引发了对苹果（与高盛合作发布的）信用卡业务的监管调查[31]”。“苹果信用卡性别歧视”相关报道可参考：[32]，[33]，[34]，[35]。但在 2021 年，有报道称“纽约州金融服务部最近结束的一项调查[36]发现，苹果的银行合作伙伴不存在基于性别的歧视[37]”。

● 2020 年：有偏见的罪犯评估系统

在刑事司法系统中，诸如 COMPAS（美国矫正罪犯管理替代制裁评估系统）和 OASys（英国罪犯评估系统）等工具被用于对罪犯进行风险评估和管理。这些系统协助当局对罪犯的量刑、假释和治疗项目做出知情决策。然而，它们的算法因透明度、公平性和偏见问题而受到严厉的批评[38]。

● 2022 年：加拿大航空公司受聊天机器人退款政策约束

加拿大航空公司（AirCanada）面临严格审查，因为其聊天机器人提供了有关该航空公司的丧亲旅行政策的误导性信息，导致与一名要求退款的乘客发生纠

纷。尽管加航辩称聊天机器人是独立运行的，但法庭还是作出了有利于乘客的裁决，强调航空公司对其网站上提供的信息负有责任。法庭命令加航支付部分退款并承担额外费用。这一事件凸显了人工智能的责任问题和客户服务自动化的复杂性[39]，[40]。

- **2023 年：诉讼：联合健康保险因 AI 缺陷拒绝为老年人提供护理**

在一场针对联合健康保险（UnitedHealth）的法律诉讼中，多个家庭指控该公司使用的人工智能系统存在缺陷，导致了老年患者的必要护理申请被拒绝，甚至无视了医生的推荐意见。这起诉讼凸显了在医疗决策中完全依赖自动化系统所面临的挑战，引发了人们对患者福祉和医疗服务公平性的担忧。随着人工智能技术在医疗领域的日益普及，该案件强调了医疗人工智能的透明度、问责制和人为监督的必要性，以确保所有患者得到公平的治疗[41]，[42]。

- **2024 年：谷歌的 Gemini：人工智能偏见的教训**

谷歌推出的 Gemini 1.5 聊天机器人[43]受到了广泛批评，原因是尽管它试图避免偏见，但仍生成了不准确且具有偏见的图像，特别是在历史背景中忽略了白人个体。埃隆-马斯克和保守派人士指责谷歌的算法有偏见。作为回应，Google 暂停了 Gemini 的进一步推广，但其回应缺乏透明度。这一事件凸显了人工智能在道德和透明度方面的挑战，引发了关于多样性倡议和算法问责制的广泛讨论。随着谷歌努力恢复公众信任，Gemini 事件强调了负责任 AI 创新的重要性[44]。

2.2 行业：法规与挑战

本节深入探讨与人工智能有关的特定行业的监管和合规工作。各行业按英文字母顺序排列并分别论述。在每个行业部分，我们将介绍其背景、环境和历史。在接下来的第三部分中，我们将提出跨行业应对 AI 挑战的新颖方法建议。

2.2.1 汽车

汽车行业[1]寻求在自动驾驶和无人驾驶功能（SAE 4 级和 5 级[45]）中应用人工智能，同时特别强调此类功能以及其他车载系统和组件的安全性。目前，已有多个 ISO 标准提及或部分规范了人工智能，其它还有更多标准正在起草或审查中。许多监管机构目前正在制定针对汽车行业的标准和方法，但尚未强制执行。

虽然当前的法规已经间接影响了人工智能，但一些监管机构直接在其规范中提到了此类技术，其中包括欧洲议会和欧盟理事会于 2019 年 11 月 27 日颁布的（欧盟）2019/2144 法规，该法规针对机动车辆及其挂车，以及为这些车辆设计的系统、部件和独立技术单元的型式认证要求，特别关注了车辆的整体安全性以及对车辆乘员和脆弱道路使用者的保护[46]。其中，第 11 条（关于自动驾驶车辆和全自动驾驶车辆的特定要求）明确规定了与 AI 相关的安全系统要求，特别是当 AI 技术被用于驱动自动驾驶和全自动驾驶车辆时。需要注意的是，该条款并未专门针对 AI 制定，但明确提及了“自动驾驶车辆和全自动驾驶车辆”。

虽然这项欧盟法律并未直接制定人工智能及其功能本身的具体标准，但 ISO 正在制定的 PAS 8800 标准——《道路车辆--安全与人工智能》[47]却侧重于人工智能安全，致力于制定一套关于“安全原则、方法及验证依据”的框架。该标准的适用范围广泛，不局限于自动驾驶或无人驾驶车辆，而是涵盖了所有道路车辆。其目的是解决有关人工智能监管和标准化的基础性问题，并为行业提供具体的、实用的指导方针[48]，从而协调现有法规和既定原则，如预期功能的安全性。

另一个关于人工智能（功能）安全的行业标准是 ISO/TR 5469:2024《人工智能——功能安全和人工智能系统》[49]。该文件于 2024 年发布，描述了与 AI 相关的风险因素，以及当前在多种汽车应用中可用的、与 AI 技术相关联的方法和流程。这些应用包括利用 AI 和非 AI 系统来管理 AI 安全系统的与安全性相关的功能。该标准已经发布，旨在为未来 ISO/IEC AWI TS 22440 标准[50]，[51]的制定提供支持。此外，ISO/TR 4804:2020《道路车辆——自动驾驶系统的安全与网络安全设计、验证与确认》[52]也进一步强调了安全性，特别是网络安全方面。该标准重点关注自动驾驶系统（SAE 3 级和 4 级）的开发与验证，并提供了全球适用

[1]由于作者的工作领域和工作地点，本章将重点关注欧盟的汽车行业。

的安全、验证与确认方法。值得注意的是，ISO/TR 4804:2020 将在未来被 ISO/CD TS 5083 所取代。新的 ISO 文档涵盖了“开发并验证配备有安全自动驾驶系统的自动驾驶车辆的步骤”，这些步骤同样遵循 SAE 3 级和 4 级的要求。此外，该文档还将深入探讨这类自动驾驶系统所需达到的安全级别，其目标是在与人工驾驶相比的情况下，显著降低因自动驾驶而引发的总体风险。

2.2.2 航空

全球航空界与其他行业一样，在计算机系统的使用方面遵循许多相同的实用标准。这也适用于人工智能和运行人工智能的平台。为此，航空领域也将采用公认的信息技术安全标准，包括 ISO/IEC 27001[53]、ISO/IEC 42001[54]、ISO/TR 5469[49]、NIST AI RMF[55]，以及与航空公司或制造商管辖范围相关的 AI 道德标准。然而，目前尚未实施针对 AI 的具体法规。

全球航空业的管理机构，如美国联邦航空局（FAA）、欧盟欧洲航空安全局（EASA）、英国民航局（CASA）和澳大利亚民航局（AUCASA），都意识到了人工智能给其行业带来的好处和挑战。不过，目前他们还没有对人工智能的使用进行监管。上述机构已经成立了人工智能特别工作组，以调查人工智能在飞机上、地面操作和行业监管中的使用情况。英国 CASA 目前正处于向行业公开征求意见的阶段[56]，而美国 FAA 则组建了一个由 Pham 博士领导的技术专家团队[57]，该团队将专注于 AI 在航空领域的研究与应用。

人工智能被广泛应用于军事航空的多个方面，包括情报数据分析、自动驾驶车辆、机场和空军基地的预测性维护和实体安全。此外，AI 还用于管理 IT 安全和运营系统[58]。

总体而言，民航业非常希望利用人工智能来协助天气规划和航线安排、维护、客运和货运管理等工作。大多数提出的 AI 使用方案都围绕着预测性维护、航线和维护规划以及乘客与货物管理方面的机器学习。生成式人工智能的使用仅限于航空公司客户聊天机器人和决策支持系统。不过，目前正在进行的重要研究是探索 AI 在飞行全过程中的空中交通控制应用。欧盟于 2022 年 10 月发布了关于 AI 在空中交通管理中应用的 CORDIS 成果包，涵盖了欧洲人工智能控制空中交通的许多方面[59]。

航空业面临的一个特殊挑战是，商用客机的寿命通常以几十年为单位，而人工智能技术却日新月异，不断取得突破性进展，这迫使相关法规必须紧跟步伐，持续进行更新。

2.2.3 关键基础设施和基本服务

将人工智能融入关键基础设施是一项重大转变，其旨在建立更高效、反应更迅速、系统更智能的关键基础设施。其中，包括但不限于电力、天然气、水和食品供应链等行业，在现代社会的正常运转中均起到了至关重要的作用。在迎接数字化转型的同时，在性能提升和安全鲁棒性之间取得平衡变得越来越富有挑战性。在本节中，我们将探讨人工智能与关键基础设施的融合所带来的挑战和机遇，重点关注监管框架、安全标准以及适应技术不断进步的需求的重要性。

微妙的平衡：性能与安全

高性能人工智能系统在关键基础设施中可起到的作用毋庸置疑。相关的技术应用可以提高运作效率、优化运营流程，并能在故障发生前进行预测和处理。然而，在大多数情况下，仅通过物联网（IoT）设备、集成 tinyML[8]和边缘计算来集成人工智能系统，会带来新的安全漏洞。分散化的数据处理虽然有利于提高系统响应速度，但却扩大了潜在网络威胁的攻击面。国际标准化组织（ISO）和国际电工委员会（IEC）等监管机构和标准化组织已经制定了 ISO/IEC 27001、ISO/IEC 27002 等通用框架，以及以工业自动化和控制系统为重点的 ISA/IEC 62443[60]系列标准和技术规范和报告，例如 IEC TS 62351-100-4:2023[61] 和 IEC TR 61850-90-4:2020[62]，并通过具体的措施来保护这些技术。然而，针对于关键基础设施中的物联网和边缘人工智能的法规具体内容仍然模糊不清。

致命弱点：物联网和边缘人工智能

将物联网设备整合到关键基础设施领域中会产生很高的网络攻击风险。这些设备是人工智能感知神经网络中不可或缺的一部分，可能会被利用来提供虚假数据，从而操纵人工智能系统的决策，进而导致基础设施无法运行。尽管欧盟人工智能法案、NIS2 指令和美国人工智能行政命令等现行法规非常重视网络安全和风险管理，但想要应对边缘人工智能和微型 ML 设备带来的独特挑战，我们还需

要去不断探索。ENISA 的行业研究报告《人工智能中的网络安全和隐私——对电网需求的预测》[63]中就试图揭示此类潜在的风险和威胁。

打造面向未来的基础设施

确保关键基础设施中人工智能的安全涉及到几项重要战略。首先，必须制定针对特定行业的人工智能法规，确保可以满足所有关键基础设施行业的个性化需求。其次，要对物联网设备和边缘人工智能采用标准化安全协议，加强系统抵御网络威胁的能力。国际合作在人工智能治理方面将发挥关键性作用，可以确保在全球范围内采取一致、有效的方法来保护关键基础设施的安全。2022 年，由水源研究基金会（WRF）和水源环境保护协会（WEF）联合发起的“领导人技术创新论坛”（LIFT）计划的获奖者利用尖端的人工智能和数据科学技术来保护（防范网络安全威胁）、预测（系统状态）和优化设施流程[64]。

2024 年 3 月 1 日，美国总统科技顾问委员会（PCAST）发布了一份报告，副标题为“优化数据时代下的关键基础设施”[65]，其中很大一部分专门论述了人工智能在关键基础设施系统当中的作用。报告强调了人工智能发展的双重性，阐述了其在不断进步中具备的潜力和被滥用的风险。他们强调了人工智能对关键基础设施的影响，并进行了针对性分析，指出人工智能有增强恶意黑客能力的可能性，因此有必要针对此类威胁做好战略准备。此外，PCAST 还提倡在防御机制中利用人工智能，并呼吁开展公私合作，打破传统思维束缚和行业壁垒，进行功能性开发以及国际协作，以有效应对人工智能和网络安全的挑战。

持续进化：发展之路

将人工智能融入关键基础设施是一项风险与机遇并存的挑战性举措，必须一丝不苟地保持创新性与安全性之间的平衡。随着人工智能技术的进步，相应的监管措施和安全治理框架也要共同发展，这就要求我们在安全监管方面采取谨慎、普适的方法，确保我们的防御措施可以随着新漏洞的出现而同步优化。想要建立健全、有效、安全、可适配的人工智能集成系统，需要践行以下几个步骤：通过实施“零信任”原则，奉行“从不信任，始终验证”的信条，确保关键基础设施系统不过分依赖外围防御系统，规避娴熟的黑客攻击者可以绕过防御系统的风险；与传统安全措施相比，使用人工智能驱动的安全系统能更快地适应不断变化的威

胁。但要确保在所有关键基础设施行业之间做到共享威胁情报和最佳实践做法，才能从更广泛的知识和经验库中受益；要建立持久的培训机制，通过针对性的专业性培养才能让安全专业人员掌握最新的知识和操作工具，从而去应对日益变化的新威胁。

前方的道路

将人工智能融入关键基础设施是一个充满挑战但又发展潜力十足的举措。但如何在提升性能的同时保障设施安全，是一项复杂但至关重要的任务。只有通过制定针对性的法规、采用标准化的安全措施以及开展国际合作，我们才能更好地驾驭这一项前沿技术。关键基础设施的未来取决于我们是否有能力在利用人工智能优势的同时防范风险，确保为社会正常运转奠定灵活、高效和安全的基础，并且在实际决策或操作中，必须要有人类参与，仅将人工智能的决策作为参考。

目前，我们仅有一套并不完善的法规。虽然有关人工智能的法律法规正在全球范围内不断完善，但目前还没有具体的法律法规关注人工智能在关键基础设施中的应用。其重点在于一般性原则。现有的法规和一些倡议则侧重于强调网络安全、安全性和可信性等更广泛的原则。

一些政府机构正在积极制定负责任 AI 的有关开发和部署框架的新标准，其中一部分就侧重于人工智能在关键基础设施当中的应用。

现行举措

● 美国第 14110 号行政命令（2023 年 10 月）

关于安全、可靠、可信赖地开发和使用人工智能的第 14110 号行政命令（2023 年 10 月）：关于在关键基础设施和网络安全中管理人工智能的第 4.3 条[66]。该计划概述了如何评估和降低关键基础设施中人工智能风险，优先考虑制定安全准则，成立人工智能安全和安保委员会，并对基础设施所有者和运营商实施监管。网络安全和基础设施安全局（CISA）负责评估和减轻人工智能对关键基础设施的威胁[67]。

● 欧盟人工智能法案

欧盟人工智能法案[68]，[69]概述了人工智能系统的监管框架，重点关注合规性、风险管理、数据治理、技术文档、记录保存、透明度、人工监督、准确性、稳定性和网络安全标准。它根据风险等级对人工智能系统进行分类，对于关键基础设施中的“高风险”应用须接受更严格的监管。2024 年 3 月 13 日，欧盟议会批准了《人工智能法案》[70]。

- **经合组织人工智能原则**

经济合作与发展组织（OECD）原则 1.5[71]强调了各行为主体对人工智能系统的开发、部署和使用的责任（问责制），尤其是那些具有潜在社会影响力的人工智能系统。关键基础设施无疑是一个具有重大社会影响的领域，要确保该领域人工智能的应用符合问责制这一原则。这些原则对于推动国际讨论具有一定的影响力。

- **人工智能和数据法案（AIDA）**

AIDA[72]旨在指导加拿大国内人工智能的创新发展和问责制建设。它确保以安全的方式开发和使用人工智能系统，满足监管体系对于个人和社会经济领域所使用的人工智能系统进行管理的需要。此外，它还强调要保护高影响力人工智能系统可能会受到的伤害，避免可能产生的偏差结果。AIDA 概述了各利益相关方的角色、参与人工智能开发的企业的义务以及确保合规的执行机制。

您可以在我们的论文《实践原则：动态监管环境下的人工智能责任制体系建设》中找到对现有的相关举措、人工智能的法律以及监管环境的深入分析。

2.2.4 国防

国防中的人工智能和新兴技术

在未来的战场上，现实世界和数字虚拟化世界将交织在一起，形成一个复杂而具有争议性的环境。新的威胁和挑战将不断涌现，而人工智能将成为获取领先优势、进行全面态势感知、收集情报和改进决策的关键因素。机器人、大数据、

自动化系统以及生物技术等将为国防部队带来新的机遇和风险。人工智能对于整合、利用这些技术以及防御对手使用这些技术都扮演了至关重要的角色。军方需要寻求与民营企业、学术界和盟国建立伙伴关系，促进创新、推广应用，培养全新的领导力和文化内涵。人工智能将促进跨领域、跨平台和跨组织的合作与交流，实现人机协同，共同学习。

这些都需要与人工智能法规和监管体系相吻合。然而，想要在国防领域使用人工智能，那么道德、安全和可信度是最重要的考虑因素。法律法规和监管体系都能通过创造公平的竞争环境，促进创新与合作，提高公众的信任度和接受度，来推动国防工业向前发展。

- 人工智能相关的法律法规和监管体系有助于确定在国防中开发、部署和使用人工智能系统的标准和最佳的操作程序，降低人工智能受到偏见、伤害或滥用的风险，并增强人工智能使用者的责任感和透明度。
- 它可以通过建立共同市场，促进竞争优势培养来刺激国防工业的发展。统一国防机构的规则和要求也有助于国防和非国防部门人工智能系统的跨境合作和互操作性。
- 人工智能相关的法律法规和监管体系可增强公众对国防领域人工智能的信任和接受程度。

人工智能在国防中的历史作用

自从艾伦-图灵（Alan Turing）发表了他的奠基性著作《计算机器与智能》[73]，揭开了我们今天所熟知的人工智能的序幕以来，第一批的投资和使用都是由国防部门推动的。利用自然语言处理（NLP）将语音转为文本，以及利用机器学习分析文本和模仿人类思维和推理的相关技术正在迅速发展。最初的投资由国防部门提供，主要是美国国防部高级研究项目管理局（DARPA），该机构资助了多家机构的人工智能研究。

随着技术的进步，人工智能系统的计算能力增强，数据成本降低，新的使用案例不断涌现。1970 年，通过投放传感器，人工智能模型可以确定目标、分配资源和任务计划，从而首次实现了半自动化战争[74]。20 世纪 80 年代，这一技术发展到智能武器、模拟和决策支持。

2018 年，美国国防部宣布人工智能“即将改变未来战场的特性[75]”。2018 年，五角大楼成立了联合人工智能中心（JAIC）和国家人工智能安全委员会。美国国会为了表达支持，在人工智能领域投资了 10 亿美元，并多次投资利用人工智能成果的系统，如全自主决策和无人系统。中国也采取了相应的行动，宣布要在 2030 年之前在人工智能领域引领世界。俄罗斯总统弗拉基米尔-普京曾有过著名的预言：“谁成为这一领域的领导者，谁就将成为世界的统治者[76]”。

随着人类社会生产力的提高、决策的自动化和人工智能洞察力的增强，我们是否会造成战争发生得太快而人类无法干预的局面？对人工智能的大量投资是否会导致人工智能的军备竞赛？

我们已经看到了人工智能的魅力所在。自动化机器和机器人价格低廉，而且可以被替代。但人类则不然。自动化不会疲劳，不需要庞大的供应链支持，也避免了人类的生物性缺陷。

另一方面，我们必须万分小心谨慎，仅仅依靠机器做出生死攸关的决定可能会带来可怕的后果。引用美国科技政策办公室（OSTP）主任、DARPA 前主任 A.Prabhakar 博士的话说：“当我们审视人工智能的发展时，我们看到了一些非常强大的东西，但我们也看到技术仍然相当有限。问题是，当它出错时，其错误的方式是任何人类都无法理解的。[77]”。

我们是否造成了前所未有的附带损害？自主战争机器是否是一场即将发生的战争罪行？

现阶段在国防系统中使用的人工智能一般仅在明确、限定的环境中执行复杂度较低的单一任务，例如图像识别、用于保卫舰船的速射炮以及长时间寻找明确特征的导弹。

人工智能法规与国防

人工智能在国防领域的应用因其潜在的不可预估的影响力和杀伤力而有所不同。标准机构和监管机构编撰的大多数相关材料很少涉及与国防部门相关的用例。国防部门经常会参考公共领域的可用资源，为其量身定制类似的参考资料。对这些资料的访问通常仅限于（1）不向希望伤害我们的心怀不轨者提供洞察力；（2）对手能够从他人开发的知识产权中学习。技术也模糊了民用和国防部门之

间的界限。一方面，国防部门需要保密以保护国防，也需要灵活性，而严格的法规会限制这种灵活性。另一方面，人工智能在国防领域出现偏差所带来的潜在危害可能远远超过非国防领域。尤其是需要在自主决策和人工监督之间取得平衡时，其在国防领域的应用很容易引发道德、安全及保障方面的问题，缺乏明确的指导方针或灵活的参与规则可能会导致滥用或其他不可控的意外后果。

欧盟没有公开的专门针对国防的人工智能相关法律法规。

虽然没有具体的规定可以称为“人工智能国防法案”，但北大西洋公约组织（NATO）制定了一项战略，聚焦于加快推动人工智能在国防当中的应用进程。该战略旨在加强关键人工智能的推动因素，并制定了在国防应用中负责任且合乎道德地使用人工智能的政策。美国国防部（DOD）于 2023 年 2 月公布了《人工智能和自动化技术的军事应用宣言》。虽然不是立法法规，但它是一项旨在确保军队可以负责任地使用新兴人工智能技术的宣言。

2.2.5 教育

人工智能（AI）与教育领域的结合为提高学习成果和解决教育不平等问题提供了机遇。自适应学习系统、人工智能导师和预测分析等人工智能技术将根据不同的学习需求提供个性化教育。然而，由于人工智能系统涉及广泛的数据收集和处理，这种整合引发了行业内部对隐私和数据保护的担忧[78]。为确保在教育领域合乎道德地使用人工智能，治理框架必须优先考虑人类监督以及遵守法律和道德的标准[79]。

在教育领域实施人工智能技术时，公平性和可获取性是至关重要的考虑因素。教育机构和决策者必须确保人工智能工具不会加剧现有差距，而是成为增强赋能的工具。解决人工智能算法中的偏见风险对于防止歧视性结果会起到至关重要的作用，这强调了开发流程的透明性和包容性。教育工作者、技术专家、伦理学家和政策制定者之间有必要开展合作，来指导开发符合教育伦理的人工智能系统[78]。

要与包括学生、家长和教育工作者在内的利益相关者持续对话，使人工智能计划与社会价值观和社会期望保持一致。此外，要对教育生态系统中的所有参与

者进行数字化教育普及和人工智能教育投资，培养他们有效、批判性地使用人工智能技术。通过以道德治理、合规性和包容性为重点，利用人工智能作为实现教育改革和教育公平的工具[78]。

将人工智能融入教育需要采取积极主动、细致入微的方法，并在创新与道德考量之间取得平衡。通过促进合作、确保透明度以及优先考虑公平性和普及性，教育部门可以利用人工智能丰富学习体验，同时保障所有学习者的权利和福祉。

2.2.6 金融

金融业被认为是全球监管最严格的行业之一，它通过国际公认的标准或地方监管机构进行监管。想要应对新的技术发展趋势，除了要洞察行为趋势和频繁的社会分析之外，还要满足最终客户的需求。行业制定“人工智能法规”已有多年，但人们对这些法规的认识还很有限。它们属于风险管理的业务风险类别。

怎么会这样？许多人可能会认为 2008 年的金融危机与雷曼兄弟公司的倒闭有关[80]，[81]但实际上金融危机发生在十年以前。长期资本管理部（LTCM）成立于 1994 年。它由诺贝尔经济学奖得主迈伦-斯科尔斯（Myron Scholes）和萨洛蒙兄弟（Salomon Brothers）等华尔街著名经济学家领导。他们专门从事套利金融建模。1998 年 8 月，俄罗斯债务违约，LTCM 持有大量的这些国家债券头寸，损失了数亿美元，而与此相反的是计算机模型却建议他们持有头寸。重要的是要明白，虽然 1998 年称之为“计算机模型”，但我们今天称之为机器学习模型或人工智能。如果 LTCM 倒闭，由于其头寸的系统风险，我们很可能会看到第一次全球金融危机，但美国政府介入并提供了 36.25 亿美元的贷款。LTCM 于 2000 年初清算[82]。2005 年，巴塞尔银行监管委员会发布了新的《内部评级系统验证研究指南》[83]，[84]。虽然这听起来不像人工智能，但评级系统是一个银行术语，是分析师或评级机构用来评估股票、债券或公司信用度的评估工具。如今，在这些评级系统中使用深度学习技术是很常见的，因为从机器学习的角度来说，评级系统就是一个推荐系统。人工智能的底层技术部分对银行家或交易商是隐藏的。

2011 年，在金融危机和“大衰退”之后，美国联邦储备委员会发布了更为详细的银行业指导意见，公布了“监管函”SR 11-7-模型风险管理指南[85]，资产规模达到或超过 100 亿美元的银行必须遵守该指南。因此，模型构建和使用方面的问题被视为导致全球抵押贷款危机的重要因素。

模型风险管理指南 SR 11-7

在美国银行业，这些指导意见就像是新法律的出版物，均具有一定的社会地位和应用价值。在 SR 11-7 中，模型被定义为“一种定量方法、系统或途径，它应用于统计、经济、金融或数学理论、技术和假设，将输入的数据处理成定量的估计值”。模型的使用不可避免地会带来风险，即“根据不正确或误用的模型输出和报告做出的决策可能带来的不利后果”。该文件强调了积极的模型风险管理的重要性，因为根据不正确或误用的模型做出的决策可能会产生不利后果（包括财务损失）。金融机构必须具备有效的模型风险管理框架，一个有效的模型风险管理框架必须包括健全的治理、政策和控制措施，以确保模型开发、实施、使用和有效验证的鲁棒性。

美国监管机构在这方面最著名的调查是 2019 年苹果公司的歧视性信用卡（见“人工智能事件简史”），因为它违反了 SR 11-7 中的“有效验证”部分。

另一个例子是 2012 年令骑士资本（Knight Capital）集团损失 4.4 亿美元的交易模型出错事件[86]。骑士资本集团专门从事高频交易（HFT），这是人工智能的一个细分领域。HFT 是以毫秒到纳秒的速度进行股票交易。从本质上讲，高频交易 HFT 是指在相关兴趣方发出实际订单之前买入股票。一旦真实订单到达证券交易所，HFT 方将再次卖出股票，并保留微小的差价。这种情况在一天内发生的频率极高。一般来说，HFT 在交易日结束时不会持有股票。2012 年 8 月 1 日，骑士资本集团的所有交易服务器都没有推出软件更新，导致错误执行订单。骑士资本集团的 HFT 系统因为无法识别之前的股票收购记录，所以不断买入相同的股票，由此扰乱了 148 家公司的价格。结果，骑士资本集团在短短 45 分钟内就蒙受了巨大损失（股市损失 4.4 亿美元）。由于这个 DevOps 问题，该公司于 2012

年 12 月被收购，这违反了 SR 11-7 的“实施和使用”预期。该并购于 2013 年 7 月完成。

除了人工智能在银行业的产品应用外，美国机构还将模型用于欺诈检测和遵守《银行保密法》[87]和《美国爱国者法案》[88]。这些模型可以识别和评估客户的私人交易数据，并报告潜在的可疑活动。这些系统和反馈的结果是美国政府反恐计划的重要工具。不符合 SR 11-7 模型的预期会导致较差的合规评级和严格的监管行动，包括经济罚款/处罚和制裁，以及限制规模扩大。

欧洲中央银行（ECB）于 2024 年 2 月发布了经过修订的《内部模型指南》。与美国法规 SR 11-7 类似，该指南规定了欧洲央行期望银行如何使用内部模型的透明度[89]。它涵盖了一般主体、信用风险、市场风险和交易对手信用风险。

银行可以使用内部模型来计算风险加权资产，从而确定其最低监管资本要求。欧洲央行的修订纳入了与气候相关的风险，并详细规定了以下方面的新要求：

●**纳入与气候有关的风险**：修订后的指南考虑了与气候相关的风险，反映了这些因素在风险评估中日益重要的地位。

●**违约的通用定义**：该指南有助于所有银行采用共同的违约定义，确保整个行业的一致性。

●**大规模处置的处理**：该指南对“大规模处置”（指不良贷款的批量出售）提供了标准化的处理方法。

●**交易账簿头寸违约风险的计量**：更新后的“市场风险章节”详细介绍了如何衡量交易账簿头寸的违约风险。

●**关于交易对手信用风险的说明**：修订后的指南对交易对手信用风险（即交易对手可能违约的风险）进行了说明。

●**回归标准化方法**：摒弃复杂的内部模型。

对于亚洲银行来说，并没有像 SR 11-7 或欧洲央行指南那样的特定模型风险管理指南，但模型风险管理的做法已从美国传到欧洲，最近又传到了亚洲银行。模型风险管理（MRM）职能的范围正在不断扩大，银行对模型库存的看法也在不断拓宽，已经超越了监管和风险相关的预测方法。它们还通过加强每个步骤的框架、流程和工具，深化了对模型生命周期的端到端视图[90]。

金融业的其他重要标准包括 PCI DSS 和 PCI 3DS[91]，[92]。支付卡行业（PCI）数据安全标准（DSS）是一项全球信息安全标准，旨在通过加强信用卡数据的控制和安全来防止欺诈。任何存储、处理或传输支付和持卡人数据的组织都必须遵守 PCI DSS。PCI 3-Secure 是 EMVCo2[2]消息传输协议，使持卡人在进行非现金卡（CNP）在线交易时能与发卡机构进行身份验证。

由于在线交易和智能手机的使用，PCI 3DS 变得越来越重要。这些数据包括个人身份信息（PII）、持卡人数据（CHD）和其他财务数据。PCI DSS 和 PCI 3DS 并不涉及人工智能条款。此外，不涉及交易的银行应用程序也不需要遵守这些 PCI 标准。因此，涉及的人工智能是否必须遵守 PCI 标准，取决于使用案例是否包含交易数据。提供人工智能的云服务也需要根据这些标准进行认证。如下图所示，Azure OpenAI 服务在 2023 年 3 月至 2024 年 1 月期间的认证状态会随时间发生变化。

	Azure Service	CSA STAR Certification	CSA STAR Attestation	ISO 20000-1:2018	ISO 22301:2019	ISO 27001:2013	ISO 27017:2015	ISO 27018:2019	ISO 27701:2019	ISO 9001:2015	SOC 1, 2, 3	GSMA SAS-SM	HIPAA BAA	HITRUST	K-ISMS	PCI 3DS	PCI DSS	Australia IRAP	Germany C5	Singapore MTCS Level 3	Spain ENS High	Singapore OSPAR
March 2023	Azure OpenAI Service		✓			✓	✓	✓	✓		✓		✓						✓			
January 2024	Azure OpenAI Service	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓		✓

图 4：Azure 合规性产品[93]（2023 年 3 月和 2024 年 1 月）

[2]EMVCo（Europay、MasterCard 和 Visa 公司）是一个全球性技术机构，负责管理基于芯片的安全支付技术和标准。

总体而言，金融业非常具有创新性。它依赖于早期技术的采用来超越竞争对手。摩根大通于 2023 年 5 月申请了一项专利 IndexGPT[94]，距离 ChatGPT 的公开发布仅仅过去了半年时间。这样的持续创新有望提高个别银行和行业的绩效。

话虽如此，系统所生成的数据与“训练”它们的数据一样好，会让企业承担相关的法律和合规责任，最终用户也容易受到多种潜在威胁。因此，金融行业的首要关注点是监管合规，保护敏感和机密数据以及客户隐私是关键任务。

2.2.7 医疗保健

医疗保健/制药/医疗技术领域的人工智能既有潜力，也有风险。在这个受到高度（但非全球性）监管的行业中，区分 ML（机器学习）和 GenAI（生成式人工智能）至关重要。由于 ML 仅针对特定任务，因此可以在更大程度上确保其安全性，而医疗保健领域的 GenAI 会与不同的利益相关者互动，并在可靠性（和可解释性）、安全性、隐私以及防止误用和/或故意滥用的措施等方面带来重大挑战。在医疗保健领域，使用人工智能的风险很高，但如果可以负责任地使用人工智能，将会带来巨大的利益。

探索医疗保健领域的可信赖人工智能系统

在这个领域有大量（特定国家的）法规、全球标准和全行业指导手册，有关该行业中的 ML 和 AI 的文献、科学论文、白皮书、文章和博客更是数不胜数。

“可信赖的人工智能”与医疗行业的治理、合规性和技术挑战相关联，因此被纳入医疗保健的范围，其重点在于实践指南等实用方法上。本文讨论了医疗保健中的偏见，并从特定的行业角度揭示了这一主题。“可信赖的人工智能”进一步与第三部分概述的关于人工智能基准的思路相关联。可信赖的人工智能意味着人工智能应用程序的行为符合其预期用途，其设计足够强大，能够最大限度地减少和/或降低相关风险。给出的相关定义包括可解释性、可靠性、安全性、隐私性、问责制、透明度、遵守法规和标准、道德和负责任的行为以及减少偏差。

根据 ML 或 GenAI 应用程序的预期用途，并非上述所有内容都适用于在为特定用途设计的应用程序中被命名为“可信赖的人工智能”。

在本章中，仅考虑静态 ML/AI 应用程序。动态（自适应）应用程序可以不断学习，此处不予介绍。

医疗保健文献中可信赖的人工智能

在“可依赖的人工智能”方面，如下四个关于医疗保健领域人工智能的精选文献脱颖而出：世界卫生组织（WHO）出版了一本非常全面的书籍，书中有三百多个相关文献，提出了一个治理框架；ALTAI 汇编了一份简短但有用的指南来探讨这一主题；NIST 的论文则指出了 ML 和 GenAI 应用面临的威胁以及解决这些问题的最佳手段；最后一篇论文讨论了伦理框架，并深入细致地研究了医疗保健中的偏见问题。

1. “人工智能促进健康的伦理与治理”[95]
2. “可信人工智能评估清单（ALTAI）”[96]
3. “NIST 可信和负责任的人工智能”[97]
4. “在医疗保健及其他领域利用人工智能力量的伦理框架”[98]

可信赖人工智能的关键要求

ALTAI：

- 人事机构和监督
- 技术坚固性和安全性
- 隐私和数据管理
- 透明度
- 多样性、非歧视性和公平性
- 环境和社会福祉
- 问责制

NIST-AI 100-2e2023：

- 有效和可靠
- 安全的
- 受信任的
- 隐私增强
- 合理性
- 公平-减少有害偏见
- 负责和透明

世卫组织：

在医疗保健及其他领域利用人工智能力量的伦理框架

- 敏感性：
  - 隐私权
  - 无障碍环境
  - 包容性
- 评估：
  - 公平性
  - 非歧视性
  - 风险评估
- 以用户为中心：
  - 情境智能
  - 情商
- 负责：
  - 透明度
  - 问责制
  - 可解释性
- 增进福祉：
  - 可持续性
  - 韧性
  - 鲁棒性
  - 可靠性
- 安全：
  - 对抗测试
  - 审计

- 采用法规、标准和最优措施
- 设计隐私和默认隐私
- 机密性
- 安全与风险评估
- 透明度
- 偏见
- 数据管理
- 人工智能应用的基础设施和技术能力
- 评估和改进绩效
- 定期审查
- 预期用途
- 问责制使用
- 人类权威的代理和坚持不懈
- 道德问题
- 平等机会
- 责任转让

综合清单

- 有益性和预期用途（包括负责任的使用、语境智能）

- 人的能动性以及人的权威和监督的持久性
- 隐私、保密
- 可靠性、问责制和责任
- 绩效（包括改进、定期审查、审计）
- 透明度（包括可解释性、可解读性）
- 多样性、公平、无障碍（包括合乎道德的使用）
- 可持续性
- 技术鲁棒性、韧性和安全性（包括风险评估、基础设施能力、数据管理和治理、对抗测试、审计）

医疗文献的结论

来自各方的观点不一，也并非所有框架都包含以上所有要求。但有趣的是，只有世界卫生组织的出版物提到了要负责任和熟练的使用人工智能系统。这些出版物都没有考虑到人类有责任了解模型的预期用途，比如它的优势、局限性和制约因素，甚至没有考虑到从系统中获取最佳结果这一主题。而用直觉来处理智能应用程序可能比提供手册更具挑战性（目前，免责声明正在取代指导手册）。同样，只有世界卫生组织提到了责任问题，这与人工智能应用密切相关。注意：只有世卫组织和文章《在医疗保健及其他领域利用人工智能力量的伦理框架》特别针对医疗保健领域。它们反映出，人工智能在医疗保健领域的应用预计将通过更多的审查进行验证。

在上述出版物中，可持续性只被提到过一次。令人惊讶的是，人工智能应用在各行各业和各地区层出不穷，可持续性问题也在多个框架中得到讨论。这反映出，人工智能在医疗保健领域的投资有望超过其（环境）成本。

医疗保健中的偏见

为了避免某些偏见，有些数据必须去个性化。例如，种族可能会导致数据集出现偏差，但也可能为安全和成功治疗提供关键信息。医疗数据的处理极其复杂，对其进行有目的的评估非常重要。偏差有不同的方面，如数据驱动偏差、系统偏差、概括偏差和人为偏差。通过开发和使用可解释人工智能（XAI）[2]，可以在

包含特征和避免偏见之间取得平衡。LIME[3]和 SHAP[4]等技术就是医疗保健领域常用的事后可解释性方法。此外，通过对人工智能模型进行审计和评估，还可促进监管合规[96]。

可见，可解释性是一种很有前途的技术，可以从整体上减少医疗应用中的偏差。因此，XAI 的开发是对医疗保健领域“可信赖的人工智能”的一大贡献。

ML/AI 在医疗保健领域的进一步应用

ML/AI 应用程序还可用于简化监管流程[99]、优化供应链、协助开发药物和生物产品[100]，以及改善直接患者护理（改善医疗）、间接患者护理（改善医院工作流程）和居家护理（可穿戴设备和传感器可评估和预测患者需求）[101]。多个国家制定了开发嵌入式 ML/AI 应用的医疗设备的法规、标准、应用和指南。ML/AI 应用还有助于改善制药行业的生产流程[102]。

[3] **LIME**（本地可解释模型解释）：LIME 可以帮助我们理解机器学习模型做出特定预测的原因。它通过为单个预测创建易于理解的解释来实现这一点，即使模型本身很复杂，可以把它看作是窥探模型在每个案例中决策过程的一种方式。

[4] **SHAP**（SHapley Additive exPlanations）：SHAP 是另一种解释机器学习模型的工具。它告诉我们哪些特征（如年龄、收入等）在进行预测时最为重要。它帮助我们看到不同因素如何影响模型决策的全貌，使我们更容易理解和信任它。

第三部分：人工智能韧性的重构，受进化论启发的基准模型

本部分的目标是建立一个新颖的框架，以应对人工智能质量评级方面的挑战和优先事项，从而使人工智能系统面向未来。进化在选择性能特征和保持生存能力方面做得无与伦比。对心理学概念的探索揭示了材料特性与人类行为之间的相似性，以及增强人工智能技术韧性的一种可能的新方法。本章强调了政策制定者、监管机构和政府监督人工智能发展的重要性。

本部分首先将生物进化与人工智能的发展进行比较，重点关注韧性，然后揭示人类智能（HI）与人工智能（AI）之间的差异，最后从心理学角度审视韧性，缩小两者之间的差距。最后，本部分将探讨如何在人工智能中实施和衡量韧性。

3.1 比较：生物进化与人工智能的发展

在生物进化过程中，新的特征（突变）需要经过性能（适应特定任务）和韧性（随着时间的推移持续存在并具有优势：生存）的检验。长期保持优势的生物在进化过程中会受到内在保护。这听起来可能有悖常理，但从整体角度来看，通过不同的视角（雄性/雌性或性能/恢复力）进行选择会使系统更有能力保持其功能的完整性[103]。

同样，人工智能的性能涉及人工智能在预定义环境下的输出，而人工智能的韧性则包括泛化（避免过度拟合）和对新任务的适应性。以市场为导向的行业可能会忽视任何不能带来收入的东西，而监管机构的任务则是规范和监督人工智能技术的安全性，进而提高其适应性。

随着人工智能应用的进一步发展，更多可在部署后持续学习的系统将占领市场。这种动态系统所需的人工智能韧性远远超过静态系统。

人工智能的韧性是一个复杂的特征，这一点可能会因为对其诱人性能的十足敬畏而被忽视。因此，有必要进行监管干预，以平衡创新与监管。

3.2 人工智能系统的多样性和韧性

多样性是大自然解决问题的答案。因此，最重要的是，人工智能的韧性是强制性的，也是规范性的，必须鼓励和奖励个性化的独特方法。只有多样化的人工智能技术和可靠但个性化的人工智能韧性解决方案才能增强全局安全。

“自然界生存下来的，既不是四肢最强壮的，也不是头脑最聪明的，而是有能力适应变化的物种。”

这句话被误认为是达尔文说的，但从人工智能系统的角度来看，它仍然是正确的。对生存起贡献的不是性能，而最终是人工智能的韧性。

为最终用户提供额外的保护措施，如推荐使用（手册）、适当的培训、警告以及（如有可能）从技术上防止“标签外”使用，以增强人工智能固有的韧性，这一点至关重要，而且很容易提供，但却经常被忽视。

政策制定者、监管机构和政府必须在质量评级中优先考虑人工智能的适应性，以降低风险，确保安全和面向未来的人工智能集成。制定标准化的韧性评估指标至关重要。

3.3 对人工智能韧性进行基准测试的挑战

人工智能基准测试已接近传统性能基准（“适合预期用途”）[104]，[105]的饱和状态，一些系统的性能已超过人类设定的基准[106]。斯坦福大学的基础模型研究中心使用 HELM[107]在这一领域处于领先地位，该中心根据（目前）87 个场景和 50 个指标对模型进行评估。重点是性能和预防伤害。通过评估模型在两个截然不同的数据集（IMDB 和 BoolQ）中的表现来检查其韧性能力；重点是在保持性能的同时进行泛化的能力。

3.4 人工智能韧性的定义

我们提出了一种更全面的人工智能韧性定义，并最终提出了人工智能韧性评分标准。在这种情况下，有必要指出的是，在心理学中，测量韧性有其内在的困难[108]。心理学中的定义是[109]：“韧性是抵抗压力、从压力中反弹以及从压力中成长的能力”。

请注意，人工智能韧性包括抵御能力（抵抗力）、反弹能力（复原力）和从压力中成长的能力（可塑性）：

对压力源的**抵抗力**可以比作材料的“硬度”，也可以比作人体免疫系统的多样化和高度动态的方法。因此，抵抗力有两个相互矛盾的方面，两者都有其应有的作用。生存不是没有挑战，而是（共同）承担以下责任，积极主动、可持续地面对它们。

复原力是指随着时间的推移从压力源的影响中反弹的过程，受压力事件的程度和持续时间（外部因素）以及受压对象的韧性/适应性（内部因素）等因素的影响。复原力是动态的，受到各种变量的影响。然而，在某些情况下，压力的影响超出了恢复原有功能的能力。

可塑性是指永久性的变化。它可能是功能失调的，如心理方面的创伤、医学方面的骨折或材料科学方面的失效点。或者，它可以是功能性的，例如在训练中表现出更高的性能/韧性[87]。

建议对人工智能技术的韧性作如下定义：

AI 韧性包括系统的抵抗力、复原力和可塑性。AI 抵抗力反映了系统在面临入侵、操纵、误用和滥用时保持所需最低性能的能力；AI 复原力侧重于在发生事件后恢复到所需最低性能所需的时间、能力和容量；AI 可塑性作为系统的指标，表明其对“成败”的容忍度，并在系统故障的情况下允许快速行动，或允许 AI 韧性不断提高。

不出所料，人工智能应用的误用、滥用和事故急剧增加，人工智能使用案例表明，尽管有管理和降低风险的意识，但事故还是时有发生。

然而，由于人工智能的评判标准不明确，将人工智能纳入质量管理体系以控制、改进、纠正和预防行动和/或风险具有挑战性。在受监管的行业中，第三方评级将提高人工智能验证之外的安全性，目前人工智能的相关验证是在“适合使用”的前提下进行的。

3.5 人工智能韧性评分标准

建议采用从 0 到 10 的韧性评分，以反映人工智能的韧性，其中包括三大支柱：抵抗力、复原力和可塑性。这样的分数可以是（例如）16:5-8-3，分别代表

三大支柱的总和和三大支柱中的每一个。三大支柱得分的分布可以反映不同人工智能系统的多样性。这样，在结合不同的人工智能系统时，就能对风险及其缓解做出更明智的决定。

政策制定者、风险管理者、监管机构和政府的关注重点必须优先考虑人工智能的抗灾能力，而不是性能方面，并奖励朝此方向迈出的任何一步，促进多样化的解决方案，以增加人工智能的多样性。

现在，让我们把焦点转移到人工智能与人类的互动上。

3.6 智能感知

“智能感知”[110]的概念强调的是理解智力的差异，而不是比较它们。智能感知（目前）还不是一个广泛使用或已知的概念，而且显然是与哈佛大学心理学家霍华德·加德纳（Howard Gardner）的概念不同，霍华德·加德纳（Howard Gardner）引入了智能的不同方面[111]，[112]。“智能感知”强调，人类需要学会通过尊重彼此的不同能力，安全有效地与其他智能系统进行交互。一个很好的例子是畅销书作家安迪·威尔（Andy Weir）的科幻小说《冰雹玛丽计划》（Project HailMary）[113]。随着人工智能接近或超过人类的表现，其基准测试变得至关重要。智能系统具有多样性，尊重每种特定能力可以提高安全性和有效性。在下一节中，将探讨根本差异。

3.7 智能系统的基本差异

比较人工智能（AI）和人类智能（HI）[114]假设它们可以进行比较，HI 目前被视为黄金标准。然而，HI 的生物学基础与 AI 的高精度硅芯片基础有很大不同。硬件的这种差异影响了两种智能形式的基本功能。一旦人工智能可以在量子计算机或生物计算机上训练和运行[115]，将观察到一个飞跃，因为这两种方法都将硅芯片特性与人脑的量子能力相结合[116]，[117]，[118]，[119]。这样的人工智能系统可能会将当前人工智能的性能与人脑解决复杂任务的能力相结合。值得注意的是，这两种方法都旨在将确定性计算与非确定性方法相结合。在这一点上，

问题出现了：我们如何判断一种智能，它可以产生人类甚至可能无法理解的答案，就像《银河系漫游指南》中著名的“42”[120]。

参考文献

[1] M. W. Dictionary, "Merriam Webster Dictionary," [Online]. Available: https://www.merriam-webster.com/dictionary/governance. [Accessed 24 02 2024].

[2] C. Dictionary, "Cambridge Dictionary," [Online]. Available: https://dictionary.cambridge.org/dictionary/english/compliance. [Accessed 24 02 2024].

[3] IBM, "What is explainable AI?," IBM, [Online]. Available: https://www.ibm.com/topics/explainable-ai?utm_content=SRCWW&p1=Search&p4=43700074359379082&p5=e&gclid=CjwKCAjw4ZWkBhA4EiwAVJXwqaOswoxlekelxe20HE0gNhPjlU09SzOtlJ888FRz91kTGBO2tRsZZBoC_aAQAvD_BwE&gclsrc=aw.ds. [Accessed 14 04 2024].

[4] P. S. M. M. Prashant Gohel, "Explainable AI: current status and future directions," 12 07 2021. [Online]. Available: https://arxiv.org/abs/2107.07045. [Accessed 14 04 2024].

[5] M. a. W. S. a. Z. A. a. B. P. a. V. L. a. H. B. a. S. E. a. R. I. D. a. G. T. Mitchell, "Model Cards for Model Reporting," in Proceedings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, 2019, p. 220 – 229.

[6] E. O. O. T. PRESIDENT, "MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES - Advancing Governance, Innovation, and Risk Management for Agency Use of," The Director, Washington, D. C., 2024.

[7] The University of Queensland, Australia, "History of Artificial Intelligence," The University of Queensland, Australia, [Online]. Available: https://qbi.uq.edu.au/brain/intelligent-machines/history-artificial-intelligence. [Accessed 14 04 2024].

[8] IBM, "What is machine learning?," [Online]. Available: https://www.ibm.com/topics/machine-learning.[Accessed 24 02 2024].

[9] tinyML Foundation, "tinyML Foundation," [Online]. Available: https://www.tinyml.org/about/.[Accessed 24 02 2024].

[10] IBM, "What is generative AI?," [Online]. Available: https://research.ibm.com/blog/what-is-generative-AI. [Accessed 24 02 2024].

[11] IBM, "What is strong AI?," [Online]. Available: https://www.ibm.com/topics/strong-ai. [Accessed 24 022024].

[12] proft.me, "Types of machine learning algorithms," [Online]. Available: https://en.proft.me/2015/12/24/types-machine-learning-algorithms/. [Accessed 02 03 2024].

[13] C. W. Xiang D, "Privacy Protection and Secondary Use of Health Data: Strategies and Methods," Biomed Res Int., 07 10 2021.

[14] Wikipedia, "Wisdom of the crowd," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Wisdom_of_the_crowd. [Accessed 24 02 2024].

[15] X, "X Terms of Service," X, 29 09 2023. [Online]. Available: https://twitter.com/en/tos. [Accessed 02 03 2024].

[16] D. H. a. t. a. press, "Reddit has struck a $60m deal with Google that lets the search giant train AI models on its posts," Fortune, 23 02 2024. [Online]. Available: https://fortune.com/2024/02/23/reddit-60m-deal-google-search-giant-train-ai-models-on-posts/.[Accessed 02 03 2024].

[17] K. Coar, "open source initiative," 08 02 2004. [Online]. Available: https://opensource.org/license/apache-2-0. [Accessed 02 03 2024].

[18] open source initiative, "The MIT License," open source initiative, [Online]. Available: https://opensource.org/license/mit. [Accessed 02 03 2024].

[19] Europäisches Patentamt, "Artificial intelligence and machine learning," Europäisches Patentamt, [Online]. Available: https://www.epo.org/en/legal/guidelines-epc/2023/g_ii_3_3_1.html. [Accessed 02 03 2024].

[20] The PatentLawyer, "EPO updates guidelines for examining AI inventions," 20 02 2024. [Online]. Available: https://patentlawyermagazine.com/epo-updates-guidelines-for-examining-ai-inventions/.[Accessed 14 04 2024].

[21] I. Guttmann, "METHOD AND SYSTEM TO SAFELY GUIDE INTERVENTIONS IN PROCEDURES THE SUBSTRATE WHEREOF IS NEURONAL PLASTICITY". Europe 01 04 2022.

[22] S. W. &. J. Grasser, "Japan’s New Draft Guidelines on AI and Copyright: Is It Really OK to Train AI Using Pirated Materials?," SQUIRE, 12 03 2024. [Online]. Available: https://www.privacyworld.blog/2024/03/japans-new-draft-guidelines-on-ai-and-copyright-is-it-really-ok-to-train-ai-using-pirated-materials/. [Accessed 01 04 2024].

[23] P. D. T. (. Law), "Generative AI and Copyright Infringement," NUS - National University of Singapore, 01 2024. [Online]. Available: https://law.nus.edu.sg/trail/generative-ai-copyright-infringement/. [Accessed 14 04 2024].

[24] J. Vincent, "Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day," The Verge, 24 03 2026. [Online]. Available: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist. [Accessed 03 03 2024].

[25] P. Lee, "Learning from Tay’s introduction," Microsoft, 25 03 2016. [Online]. Available: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/. [Accessed 03 03 2024].

[26] Wikipedia, "Tay (chatbot)," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Tay_(chatbot). [Accessed 03 03 2024].

[27] J. Dastin, "https://globalnews.ca/news/4532172/amazon-jobs-ai-bias/," Global News, 10 10 2018. [Online]. Available: https://globalnews.ca/news/4532172/amazon-jobs-ai-bias/. [Accessed 02 03 2024].

[28] J. Vincent, "Amazon reportedly scraps internal AI recruiting tool that was biased against women," The Verge, 10 10 2018. [Online]. Available: https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report. [Accessed 02 03 2024].

[29] S. W. a. H. Schellmann, "LinkedIn’s job-matching AI was biased. The company’s solution? More AI.," MIT Technology Review, 23 06 2021. [Online]. Available:

https://www.technologyreview.com/2021/06/23/1026825/linkedin-ai-bias-ziprecruiter-monster-artificial-intelligence/. [Accessed 14 04 2024].

[30] R. L. I. P. F. S. a. I. U. Trisha Thadani, "The final 11 seconds of a fatal Tesla Autopilot crash: A reconstruction of the wreck shows how human error and emerging technology can collide with deadly results," The Washington Post, 06 10 2023. [Online]. Available: https://www.washingtonpost.com/technology/interactive/2023/tesla-autopilot-crash-analysis/. [Accessed 04 03 2024].

[31] B. P. C. V. a. S. M. Ziad Obermeyer, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447-453, 25 10 2019.

[32] T. Telford, "Apple Card algorithm sparks gender bias allegations against Goldman Sachs," The Washington Post, 11 11 2019. [Online]. Available: https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/. [Accessed 02 03 2024].

[33] BBC, "Apple's 'sexist' credit card investigated by US regulator," BBC, 11 11 2019. [Online]. Available: https://www.bbc.com/news/business-50365609. [Accessed 24 02 2024].

[34] WIRED, "The Apple Card Didn't 'See' Gender—and That's the Problem," WIRED, 19 11 2019. [Online].Available: https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/.[Accessed 24 02 2024].

[35] N. Vigdor, "Apple Card Investigated After Gender Discrimination Complaints," The New York Times, 10 11 2019. [Online]. Available: https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html. [Accessed 24 02 2024].

[36] J. Vincent, "Apple’s credit card is being investigated for discriminating against women," The Verge, 11 11 2019. [Online]. Available: https://www.theverge.com/2019/11/11/20958953/apple-credit-card-gender-discrimination-algorithms-black-box-investigation. [Accessed 24 02 2024].

[37] New York State Department of Financial Services, "Report on Apple Card Investigation," New York State Department of Financial Services, 2021.
[38] I. C. Campbell, "The Apple Card doesn’t actually discriminate against women, investigators say," The Verge, 24 03 2021. [Online]. Available: https://www.theverge.com/2021/3/23/22347127/goldman-sachs-apple-card-no-gender-discrimination. [Accessed 02 03 2024].
[39] W. D. Heaven, "Predictive policing algorithms are racist. They need to be dismantled.," MIT Technology Review, 17 07 2020. [Online]. Available: https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/. [Accessed 02 03 2024].
[40] A. Zilber, "Air Canada ordered to refund passenger after ‘misleading’ conversation with site’s AI chatbot," New York Post, 19 02 2024. [Online]. Available: https://nypost.com/2024/02/19/business/air-canada-ordered-to-refund-passenger-after-ai-chatbots-misleading-messages/. [Accessed 03 03 2024].
[41] A. Belanger, "Air Canada must honor refund policy invented by airline’s chatbot," arsTECHNICA, 16 02 2024. [Online]. Available: https://arstechnica.com/tech-policy/2024/02/air-canada-must-honor-refund-policy-invented-by-airlines-chatbot/. [Accessed 02 03 2024].
[42] E. Napolitano, "UnitedHealth uses faulty AI to deny elderly patients medically necessary coverage, lawsuit claims," MONEYWATCH, 20 11 2023. [Online]. Available: https://www.cbsnews.com/news/unitedhealth-lawsuit-ai-deny-claims-medicare-advantage-health-insurance-denials/. [Accessed 02 03 2024].
[43] B. Pierson, "Lawsuit claims UnitedHealth AI wrongfully denies elderly extended care," Reuters, 14 11 2023. [Online]. Available: https://www.reuters.com/legal/lawsuit-claims-unitedhealth-ai-wrongfully-denies-elderly-extendedcare-2023-11-14/. [Accessed 02 03 2024].
[44] S. P. a. D. Hassabis, "Our next-generation model: Gemini 1.5," Google, 15 02 2024. [Online]. Available: https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#gemini-15. [Accessed 14 04 2024].

[45] J. L. S. G. a. R. M. Davey Alba, "Google Left in ‘Terrible Bind’ by Pulling AI Feature After Right-Wing Backlash," TIME, 28 02 2024. [Online]. Available: https://time.com/6835975/google-gemini-backlash-bias/. [Accessed 02 03 2024].

[46] SAE Blog, "SAE Levels of Driving Automation™ Refined for Clarity and International Audience," SAE, 03 05 2021. [Online]. Available: https://www.sae.org/blog/sae-j3016-update. [Accessed 24 02 2024].

[47] EUR-Lex, "Regulation - 2019/2144 - EN - EUR-Lex," EUR-Lex, 05 09 2022. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2019/2144/oj#d1e1549-1-1. [Accessed 24 02 2024].

[48] ISO, "ISO/CD PAS 8800: Road Vehicles - Safety and artificial intelligence," ISO, [Online]. Available: https://www.iso.org/standard/83303.html. [Accessed 24 02 2024].

[49] Fraunhofer Institute for Cognitive Systems IKS, "AI regulation and AI standardization," Fraunhofer Institute, [Online]. Available: https://www.iks.fraunhofer.de/en/topics/artificial-intelligence/ai-standardization.html. [Accessed 24 02 2024].

[50] ISO, "ISO/IEC TR 5469:2024," ISO, 01 2024. [Online]. Available: https://www.iso.org/standard/81283.html. [Accessed 24 02 2024].

[51] ISO, "ISO/IEC AWI TS 22440," ISO, [Online]. Available: https://www.iso.org/standard/87118.html.[Accessed 24 02 2024].

[52] I. E. Team, "New standard to increase safety of AI," International Electrotechnical Commisson, 16 01 2024.[Online]. Available: https://www.iec.ch/blog/new-standard-increase-safety-ai. [Accessed 24 02 2024].

[53] ISO, "ISO/TR 4804:2020," ISO, 12 2020. [Online]. Available: https://www.iso.org/standard/80363.html. [Accessed 24 02 2024].

[54] ISO, "ISO/IEC 27000:2018," ISO, 2018. [Online]. Available: https://www.iso.org/standard/73906.html.[Accessed 02 03 2024].

[55] ISO, "ISO/IEC 42001:2023," ISO, 2023. [Online]. Available: https://www.iso.org/standard/81230.html.[Accessed 14 04 2024].

[56] NIST, "Artificial Intelligence Risk Management," NIST, 01 2023. [Online]. Available:https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf. [Accessed 14 04 2024].

[57] UK Civil Aviation Authority, "The CAA's strategy for Artificial Intelligence (AI)," CAA, [Online]. Available: https://www.caa.co.uk/our-work/innovation/artificial-intelligence/. [Accessed 14 04 2024].

[58] T. T. Pham, "Chief Scientist and Technical Advisor for Artificial Intelligence - Machine Learning," [Online]. Available: https://www.faa.gov/aircraft/air_cert/step/disciplines/pham_bio. [Accessed 14 04 2024].

[59] H. Weitering, "Beyond Automation: How AI Is Transforming Aviation," 14 06 2023. [Online]. Available: https://www.ainonline.com/aviation-news/aerospace/2023-06-14/beyond-automation-how-ai-transforming-aviation. [Accessed 14 04 2024].

[60] European Union, "CORDIS results pack on AI in science," [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/c0e52bea-5bb0-11ee-9220-01aa75ed71a1.[Accessed 14 04 2024].

[61] ISA - International Society of Automation, "SA/IEC 62443 Series of Standards: The World’s Only Consensus-Based Automation and Control Systems Cybersecurity Standards," ISA - International Society of Automation, [Online]. Available: https://www.isa.org/standards-and-publications/isa-standards/isa-iec-62443-series-of-standards. [Accessed 24 02 2024].

[62] IEC - International Electrotechnical Commission, "IEC TS 62351-100-4:2023," International Electrotechnical Commission, 2023. [Online]. Available: https://webstore.iec.ch/publication/63323. [Accessed 24 02 2024].

[63] IEC - International Electrotechnical Commission, "IEC TR 61850-90-4:2020," International Electrotechnical Commission, 2020. [Online]. Available: https://webstore.iec.ch/publication/64801. [Accessed 24 02 2024].

[64] enisa, "Cybersecurity and privacy in AI - Forecasting demand on electricity grids," enisa, 07 06 2023. [Online]. Available:

https://www.enisa.europa.eu/publications/cybersecurity-and-privacy-in-ai-forecasting-demand-onelectricity-grids. [Accessed 24 02 2024].

[65] M. O. Y. R. S. M. N. K. S. Z. W. W.-Y. M. Feras A. Batarseh, "Realtime Management of Wastewater Treatment Plants Using AI," Virginia Tech & DC Water, 2022. [Online]. Available: https://www.waterrf.org/sites/default/files/file/2022-11/2022_IWS-Challenge-Solution_Virginia-Tech.pdf. [Accessed 24 02 2024].

[66] P. C. o. A. o. S. &. Technology, "Strategy for Cyber-Physical Resilience: Fortifying Our Critical Infrastructure for a Digital World," Executive Office of the President, 2024.

[67] The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," The White House, 30 10 2023. [Online]. Available: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. [Accessed 24 022024].

[68] America's Cyber Defense Agency, "Artificial Intelligence," America's Cyber Defense Agency, [Online]. Available: https://www.cisa.gov/ai. [Accessed 24 02 2024].

[69] European Commission, "Artificial Intelligence Act," European Commission, 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206. [Accessed 24 02 2024].

[70] European Commission, "Annexes to the EU AI Act," European Commission, 2021. [Online]. Available: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_2&format=PDF. [Accessed 2024 02 2024].

[71] European Parliament, "Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI," European Parliament, 02 12 2023. [Online]. Available: https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-actdeal-on-comprehensive-rules-for-trustworthy-ai. [Accessed 24 02 2024].

[72] OECD, "Accountability (Principle 1.5)," OECD.AI Policy Observatory, [Online]. Available: https://oecd.ai/en/dashboards/ai-principles/P9. [Accessed 24 02 2024].

[73] Government of Canada, "The Artificial Intelligence and Data Act (AIDA)," 09 2023. [Online].Available: https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document. [Accessed 14 04 2024].

[74] J. Hoppe, "The Dropping of the TURDSID in Vietnam," US Naval Institute, 10 2021. [Online]. Available: https://www.usni.org/magazines/naval-history-magazine/2021/october/dropping-turdsid-vietnam. [Accessed 02 03 2024].

[75] U.S. Department of Defense , "New Strategy Outlines Path Forward for Artificial Intelligence," U.S. Department of Defense , 12 02 2019. [Online]. Available: https://www.defense.gov/News/Releases/Release/Article/1755388/new-strategy-outlines-path-forward-for-artificial-intelligence/.[Accessed 02 03 2024].

[76] R. Gigova, "Who Vladimir Putin thinks will rule the world," CNN, 02 09 2017. [Online]. Available: https://www.cnn.com/2017/09/01/world/putin-artificial-intelligence-will-rule-world/index.html. [Accessed 02 03 2024].

[77] Congressional Research Service (CRS), "Artificial Intelligence and National Security," 2018.

[78] E. A. A. &. C. R. Baiz, "Generative AI in Education and Research: Opportunities, Concerns, and Solutions," J. Chem. Educ., vol. 100, no. 8, p. 2965 – 2971, 27 07 2023.

[79] K. A. B. D. G.-R. A. R. M. Bozkurt A, "Artificial Intelligence and Reflections from Educational Landscape: A Review of AI Studies in Half a Century," Sustainability, vol. 13(2), no. 800, 2021.

[80] W. Kenton, "Lehman Brothers: History, Collapse, Role in the Great Recession," Investopedia, 31 12 2022. [Online]. Available: https://www.investopedia.com/terms/l/lehman-brothers.asp. [Accessed 24 02 2014].

[81] A. R. Sorkin, Too Big to Fail: Inside the Battle to Save Wall Street, Penguin, 2010.

[82] Congressional Research Service (CRS), "Systemic Risk And The Long-Term Capital Management Rescue," Congressional Research Service, 1999.

[83] BIS, "Studies on the Validation of Internal Rating Systems," BIS - Bank for International Settlements, 2005.

[84] BIS, "Studies on the Validation of Internal Rating Systems (revised)," BIS - Bank for International Settlements, 2005.

[85] Board of Governors of the Federal Reserve System, "Supervision and Regulation Letters - SR 11-7: Guidance on Model Risk Management," 04 04 2011. [Online]. Available: https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm. [Accessed 24 02 2024].

[86] M. Heusser, "Software Testing Lessons Learned From Knight Capital Fiasco," CIO, 14 08 2012. [Online]. Available: https://www.cio.com/article/286790/software-testing-lessons-learned-from-knight-capital-fiasco.html. [Accessed 24 02 2024].

[87] govtrack.us, "H.R. 15073 (91st): An Act to amend the Federal Deposit Insurance Act to require insured banks to maintain certain records, to require that certain transactions in U.S. currency be reported to the Department of the Treasury, and for other purposes," 26 11 1970. [Online]. Available: https://www.govtrack.us/congress/bills/91/hr15073/text. [Accessed 14 04 2024].

[88] congress.gov, "H.R.3162 - Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism (USA PATRIOT ACT) Act of 2001," 24 10 2001. [Online]. Available: https://www.congress.gov/bill/107th-congress/house-bill/3162.[Accessed 14 04 2024].

[89] ECB, "ECB updates Guide to internal models," 19 02 2024. [Online]. Available: https://www.bankingsupervision.europa.eu/press/pr/date/2024/html/ssm.pr240219~8c10a7d827.en.html.[Accessed 14 04 2024].

[90] McKinsey & Company, "Model Risk Management," 2019.

[91] Security Standards Council, "PCI DSS," Security Standards Council, [Online]. Available:https://www.pcisecuritystandards.org/document_library/?document=pci_dss. [Accessed 02 03 2024].

[92] Security Standards Council, "PCI 3DS," Security Standards Council, [Online]. Available:https://www.pcisecuritystandards.org/document_library/?document=3DS_standard.[Accessed 0203 2024].

[93] Microsoft, "Microsoft Azure Compliance Offerings," [Online]. Available: Azure - Compliance Offerings. [Accessed 14 04 2024].

[94] W. Daniel, "https://fortune.com/2023/05/26/jpmorgan-indexgpt-a-i-stock-picker/," FORTUNE, 26 05 2023. [Online]. Available: https://fortune.com/2023/05/26/jpmorgan-indexgpt-a-i-stock-picker/. [Accessed 02 03 2024].

[95] WHO guidance, "Ethics and Governance of Artificial Intelligence for Health," WHO, 2021.

[96] European Commission, "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment," European Commission, 2020.

[97] NIST, "NIST Trustworthy and Responsible AI - NIST AI 100-2e2023," NIST, 2024.

[98] S. &. K. R. &. B. S. Nasir, "Ethical Framework for Harnessing the Power of AI in Healthcare and Beyond," 08 2023. [Online]. Available: https://www.researchgate.net/publication/373641885_Ethical_Framework_for_Harnessing_the_Power_of_AI_in_Healthcare_and_Beyond. [Accessed 24 02 2024].

[99] A. Bilea, "How AI is Revolutionizing Pharma Regulatory Compliance," LinkedIn, 26 07 2023. [Online].Available: https://www.linkedin.com/pulse/how-ai-revolutionizing-pharma-regulatory-compliance-anca-bilea/. [Accessed 24 02 2024].

[100] FDA, "Using Artificial Intelligence & Machine Learning in the Development of Drug & Biological Products," FDA.

[101] CAPRA- Canadian Association of Professionals in Regulatory Affairs, "Artificial Intelligence-Revolutionizing the Healthcare Industry," CAPRA - Canadian Association of Professionals in Regulatory Affairs, 27 10 2023. [Online]. Available:

https://capra.ca/en/blog/artificial-intelligence-revolutionizing-the-healthcare-industry-2023-10-27. [Accessed 24 02 2024].

[102] FDA, "Artificial Intelligence in Drug Manufacturing," FDA, [Online]. Available: https://www.fda.gov/media/165743/download. [Accessed 24 02 2024].

[103] V. P. Shcherbakov, "Biological species is the only possible form of existence for higher organisms: the evolutionary meaning of sexual reproduction," Biol Direct., vol. 5, no. 14, 22 03 2010.

[104] Stanford University - Human-Centered Artificial Intelligence, "The AI Index Report - Measuring trends in Artificial Intelligence," Stanford University-Human-Centered Artificial Intelligence, 2023. [Online]. Available: https://aiindex.stanford.edu/report/. [Accessed 24 02 2024].

[105] aqua, "AI Benchmark Ranking – The Ultimate Guide to Comparing and Evaluating AI Performance," Aquarius, 01 12 2023. [Online]. Available: https://aquariusai.ca/blog/ai-benchmark-ranking-the-ultimate-guide-to-comparing-and-evaluatingai-performance. [Accessed 24 02 2024].

[106] S. Lynch, "AI Benchmarks Hit Saturation," Standford University. HAI - Human-Centered Artificial Intelligence, 03 04 2023. [Online]. Available: https://hai.stanford.edu/news/ai-benchmarks-hit-saturation. [Accessed 24 02 2024].

[107] Center for research on Foundation Models, "HELM," Stanford University, [Online]. Available: https://crfm.stanford.edu/helm/lite/latest/. [Accessed 02 03 2024].

[108] B. K. N. J. Windle G, "A methodological review of resilience measurement scales," Health Qual Life Outcomes, vol. 9, no. 8, 04 02 2011.

[109] Y. H. Ruud J.R. Den Hartigh, "Conceptualizing and measuring psychological resilience: What can we learn from physics?," New Ideas in Psychology, vol. 66, no. 100934, 2022.

[110] G. C. v. d. B.-V. R. A. M. B. e. a. J. E. (Hans). Korteling, "Human versus Artificial Intelligence," Front. Artif. Intell., vol. 4, 25 03 2021.

[111] K. Cherry, "Gardner's Theory of Multiple Intelligences," 11 03 2023. [Online]. Available:https://www.verywellmind.com/gardners-theory-of-multiple-intelligences-2795161.[Accessed 14 04 2024].

[112] Wikipedia, "Howard Gardner," [Online]. Available: https://en.wikipedia.org/wiki/Howard_Gardner#cite_note-Gordon,_Lynn_Melby_2006-1. [Accessed 14 04 2024].

[113] Wikipedia, "Project Hail Mary," [Online]. Available: https://en.wikipedia.org/wiki/Project_Hail_Mary. [Accessed 14 04 2024].

[114] V. Acharya, "AI vs. HI: The Battle of Intelligences — Exploring Advantages and Limitations," Medium, 24 07 2023. [Online]. Available: https://medium.com/@vishwasacharya/ai-vs-hi-the-battle-of-intelligences-exploring-advantages-and-limitations-89759bee090f. [Accessed 24 02 2024].

[115] A. Tongen, "Will Biological Computers Enable Artificially Intelligent Machines to Become Persons?," Dignity, vol. 9, no. 4, 2003.

[116] R. L. M.D., "Psychology Today," 02 08 2021. [Online]. Available: https://www.psychologytoday.com/ca/blog/biocentrism/202108/quantum-effects-in-the-brain. [Accessed 24 02 2024].

[117] Neuroscience News, "Our Brains Use Quantum Computation," Neuroscience News, 22 20 2022. [Online]. Available: https://neurosciencenews.com/brain-quantum-computing-21695/. [Accessed 24 02 2024].

[118] C. H. K. Koch, "Quantum mechanics in the brain," Nature , vol. 440, no. 611, 2006.

[119] Trinity College Dublin, "New research suggests our brains use quantum computation," Phys Org, 19 20 2022. [Online]. Available: https://phys.org/news/2022-10-brains-quantum.html. [Accessed 24 02 2024].

[120] Wikipedia, "The Hitchhiker's Guide to the Galaxy," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/The_Hitchhiker%27s_Guide_to_the_Galaxy. [Accessed 24 02 2024].

CSA GCR cloud security alliance
GREATER CHINA REGION

Cloud Security Alliance Greater China Region

扫码获取更多报告

云安全联盟大中华区
官网：https://c-csa.cn 电话：0755-86548359
地址：上海市浦东新区东育路255弄5号3FB30（前滩国际经济组织集聚区）/
广东省深圳市南山区粤海街道北京航空航天大厦1号楼905