2024

制造业数智化大会

建立人工智能管理体系 保障数智化发展

宣讲人：刘歆轶　公司：非夕机器人

目录
Contents

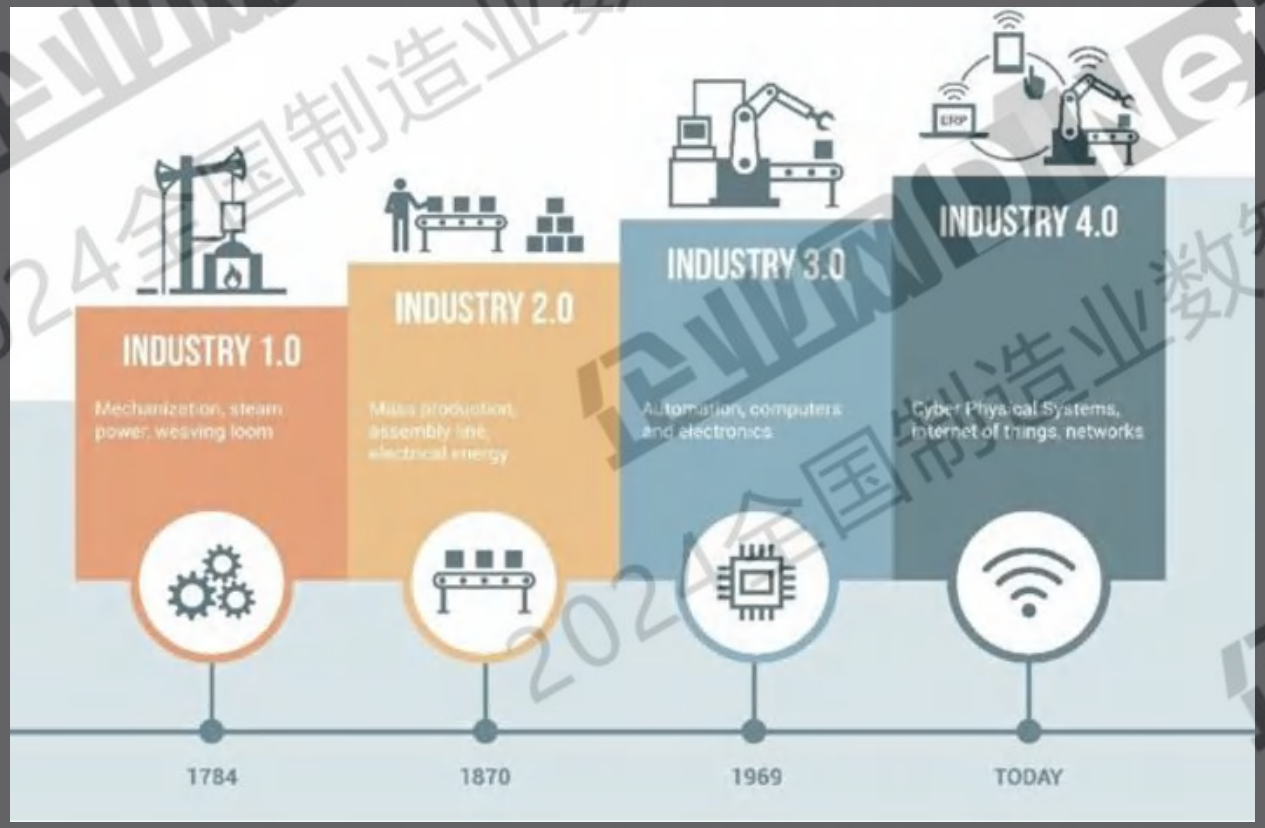
01 工业革命进程

01 创新技术应用的滞后性

今年8月份，Google 前 CEO 施密特（Eric Schmidt）在斯坦福做了一次分享，视频上传到斯坦福在线课 YouTube 官号，其中有 40 多分钟施密特与学生 Q&A 的环节。
因为分享的观点太直接，施密特的视频上了新闻。后来斯坦福官号把视频下架了，施密特也在邮件采访中对「错误言论」表示道歉。

- 现在的谷歌为什么在 AI 领域被 OpenAI 压着打?
- AI 是一场强国之间的游戏，会让富者愈富、穷人恒穷。
- 布鲁塞尔（欧盟总部所在地）一直都在摧毁科技创新的机会。
- 芯片属于高端制造业，但不会拉动就业，因为全是机械化生产，人又蠢又脏。
- **历史上，电力在引入工厂之后并不比蒸汽机创造了更多的生产力，是过了大概 30 年左右，分布式电源改造了车间布局，推动组装系统的出现，再才开始了生产力的飞跃。现在的 AI 和当初的电力一样，有价值，但还需要组织创新，才能真正拿到巨大的回报，目前大家都还只是在摘取「低垂的果实」。**

01 人类的四次工业革命



“蒸汽机”是人类第一个通用、便捷、可移动的动力解决方案，催生了工厂和工业的出现，让人类从农耕文明迈向了工业文明。

“电力”可高速传导的优良特性使得我们第一次获得了“可高效传输的动力”。发电厂负责电力的生产，通过电线将电能瞬间输送到各个地方，打破了集中式布局。

计算机技术的出现，大大提升了人类的计算能力，大量重复性的脑力劳动开始被计算机系统所替代；互联网技术，将信息的应用推向了前所未有的高峰。

智能化时代，意味着机器更加智能，人机交互更加简单、便捷。人工智能技术与移动互联网、IoT等技术一起，实现“万物互联、万物智能”。

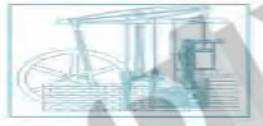
01 第一次工业革命

First Steam Engine

First Self-Acting steam engine

First cylinder steam engine

1760-1840:
The First Industrial Revolution
Water power, steam power, mechanical tools, the birth of factories



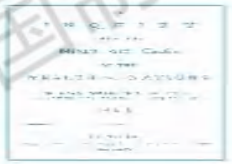

1775: James Watt and Matthew Boulton patent the Watt steam engine, the first global technology to increase productivity.

1785: Edmund Cartwright patents the power loom, industrializing textile manufacturing.

1801: Richard Trevithick begins building the first steam locomotive, leading to the first locomotive-powered railway.

1830: George Stephenson builds the first public intercity railway, the Liverpool and Manchester line.

1776: Adam Smith publishes *The Wealth of Nations*, advocating the division of labor to improve productivity by reinvesting profits into equipment and labor.



1794: Eli Whitney patents the cotton gin in the United States, sparking explosive growth in the cotton industry.

1828: James Beaumont Neilson patents the hot-blast process for iron furnaces, improving fuel efficiency and furnace capacity.

High-Pressure Steam Engine

In 1804 an American inventor named Oliver Evans created the first high-pressured steam engine. It was stationary, which operated at 30 revolutions/rpm at a minute was used to help power a marble-cutting saw. It used the high pressure in cylinder (tens of psi) heated above to help make steam. It was first used for powerboats in 1802. Then it was used on railroads in the late 1830's.

The first steam turbine

In the 1880's British engineer named Charles A. Parsons created the first steam turbine. It was used to power propellers on ships and charge turbochargers that gave electricity.

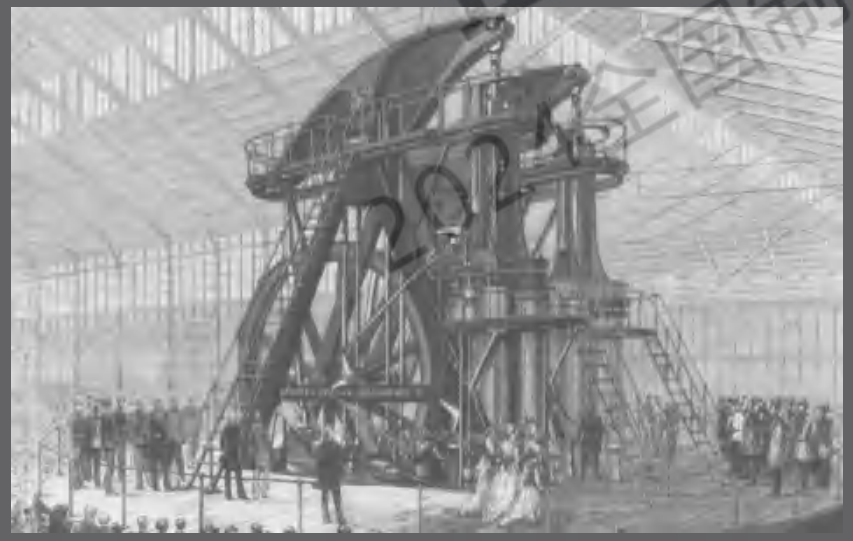
Steam as a dominant power source

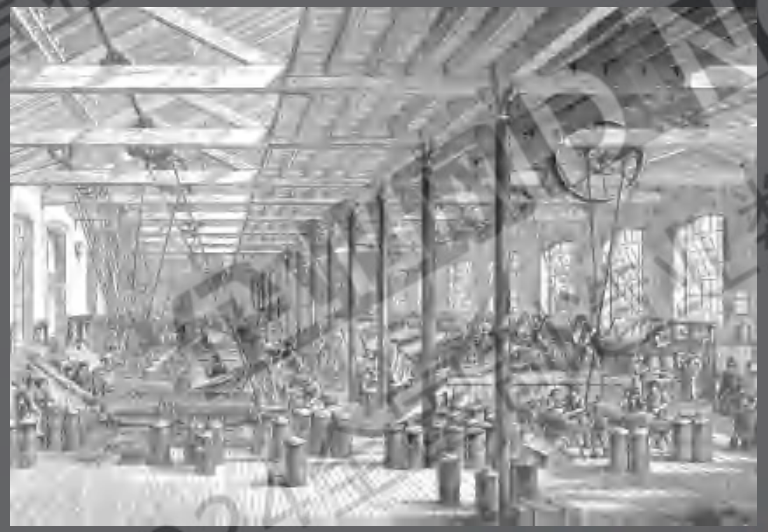

For over 200 years steam engines were the dominant power source in trains and ships powering them to new distances. It was very popular by different companys who wanted sell them to make money.

The dimise of steam

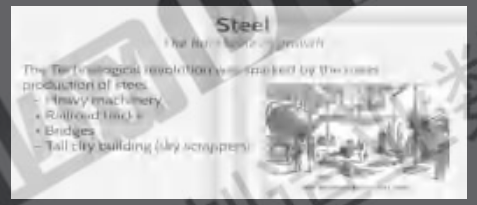

Although steam was very popular throughout the early to mid 20 century, it was very costly to fire and fuel them. After WWII the railroads and shiping companys decided to switch to diesel engines. Diesels are powered by electricity and run on gas which makes them cheaper to fix and fuel them. Starting in the 1950's steam was replace by diesels (most of them were scrapped for some were donated and some in museums today).

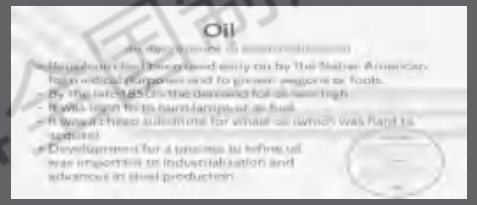
01 蒸汽时代工厂布局

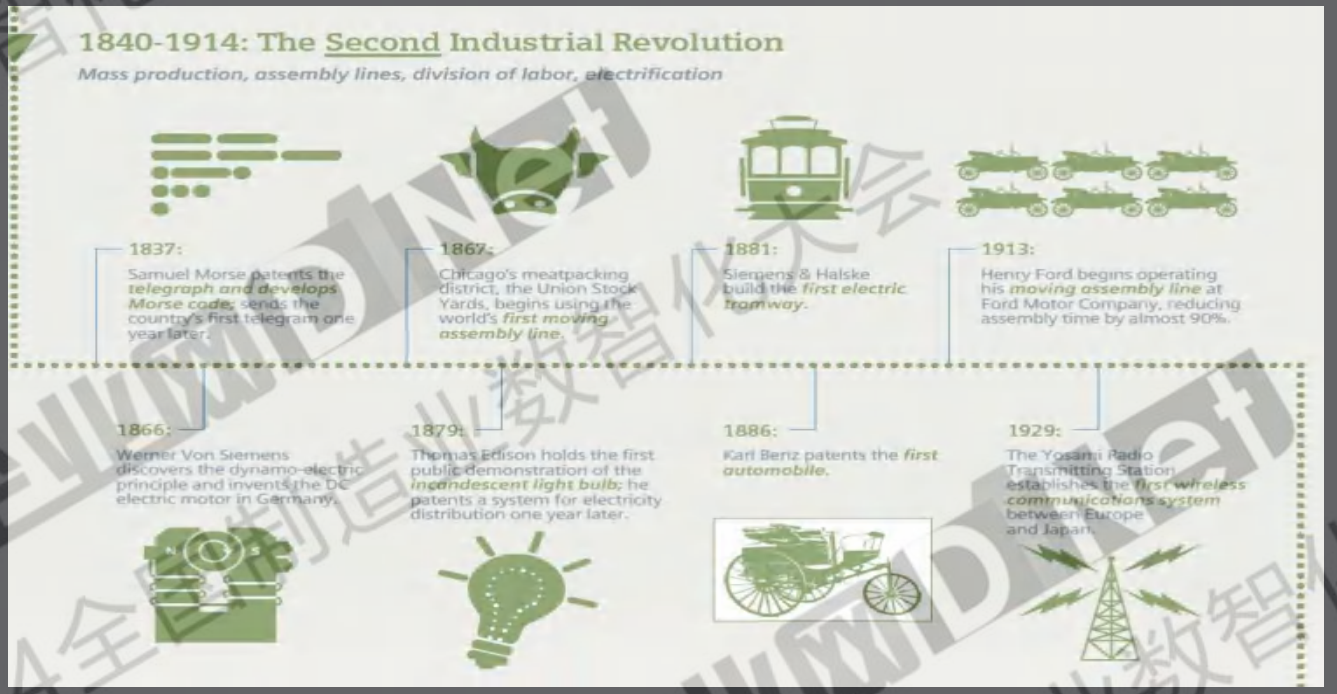


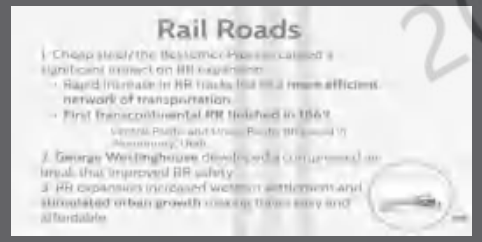


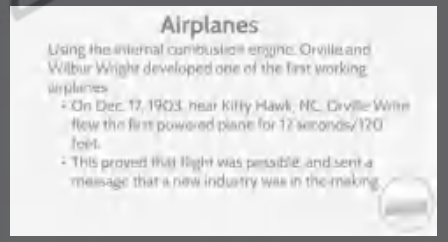
01 第二次工业革命

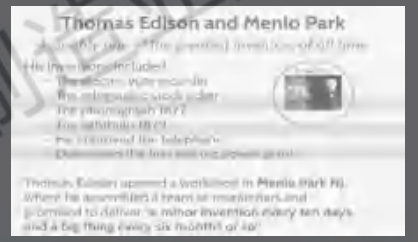












01 第二次工业革命

01 第三次工业革命

The 1940's: Early Computers

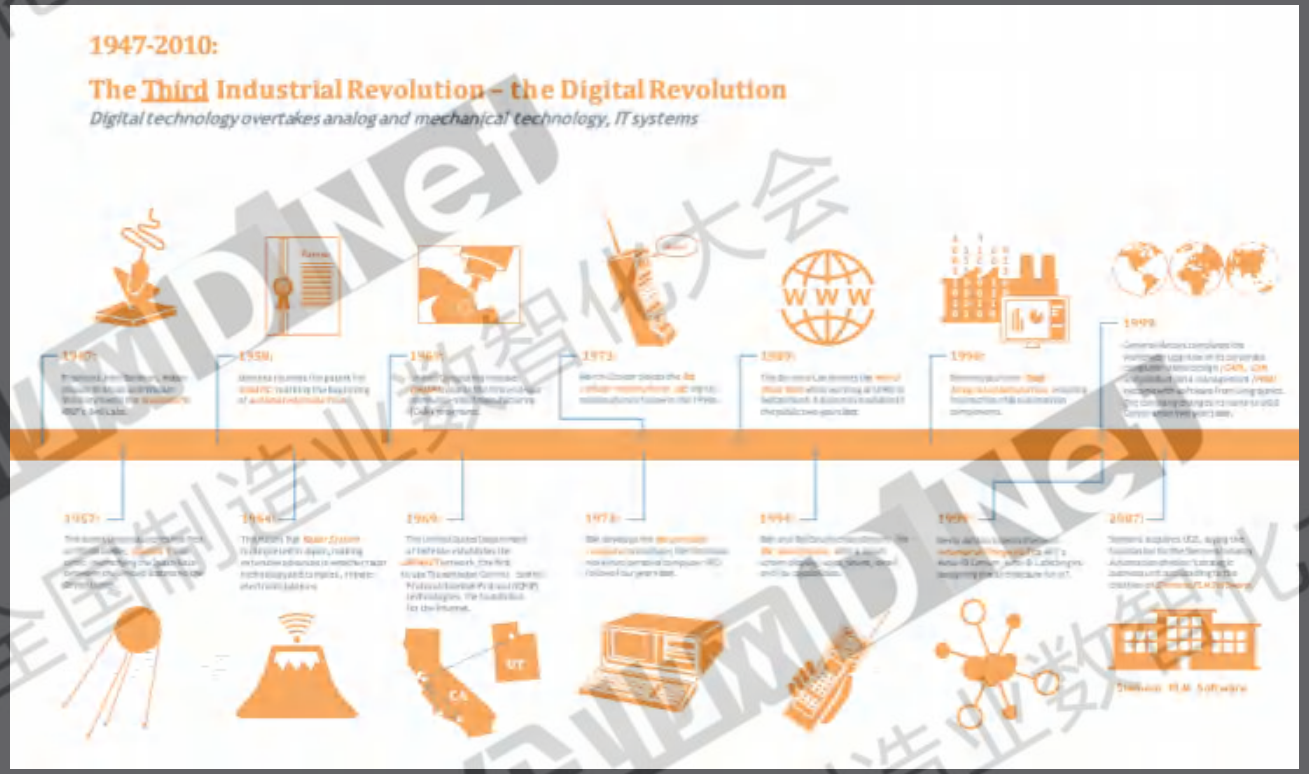
The midpoint of the 1940's saw the invention of the electronic computer produced within the U.S. army known as Electronic Numerical Integrator and Computer (ENIAC). The computer needed vacuum tubes to do its calculations. And by the end of the decade, the first computers for public hands were being developed known as universal automatic computer (UNIVAC) (Computerhistory.org)

1973: Creation of Ethernet

In 1973 Ethernet is created by Xerox corporation which lead to more advances in business information technology. The creation of the Ethernet opens up the ability of connecting many computers over long distances. Although the Ethernet was created in the early 1970's, its true potential wasn't reached until the world wide web was made accessible in the 1990's (Britannica.com).

1975: Operating Systems

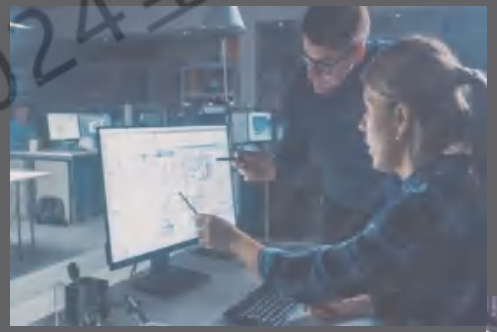
Advancements made in computers and information storage throughout the 1960's and throughout the 1970's paved way for the development of operating systems, and progress in programming. More advanced programing languages were first developed in the 1970's while Bill Gates and Microsoft started with the invention of the operating systems (Wikipedia.org/operating systems)

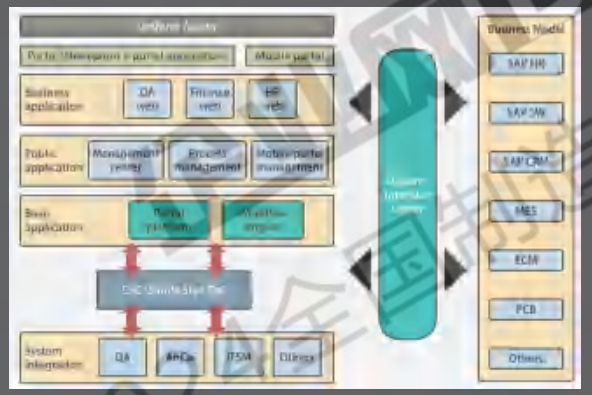


World Wide Web Made Available

In 1993, the world wide web also known as the web was first made free to public users. This was very significant in the information technology realm because of the future impacts (Wikipedia.org/worldwideweb).

01 第三次工业革命





01 第四次工业革命

Virtual Reality

Virtual reality (VR) is a technology that allows users to experience simulated environments. Professionals can use VR to visualize and evaluate complex products and processes, identify potential issues early in the design process, and make more informed decisions.

VR can be used by workers for training and skills development as they can practice performing tasks in hazardous environments without putting themselves or others at risk. Furthermore, VR helps workers from different locations to collaborate with each other in shared virtual environments.

Digital Twin

A digital twin is a virtual copy of a physical product or system used to simulate, monitor, and optimize its performance. In Industry 4.0, digital twins are used to create virtual representations of physical manufacturing systems and to gain real-time insights into the performance of their physical systems.

Cloud Computing

Cloud computing is a critical technology used in Industry 4.0 to store, process, and analyze vast amounts of data from machines, sensors, and other sources in a cost-effective and scalable way.

In Industry 4.0, cloud computing is mostly used for data storage and management, analytics and machine learning, remote monitoring and control, and collaboration and communication.

Internet of Things

The Internet of Things (IoT) is a technology that connects physical objects and devices to the internet and allows them to send and receive data. In Industry 4.0, the IoT is being used to connect machines, sensors, and other devices to create more efficient and integrated manufacturing processes. For example, IoT can be used to track inventory levels, monitor machine performance, and...

Artificial Intelligence

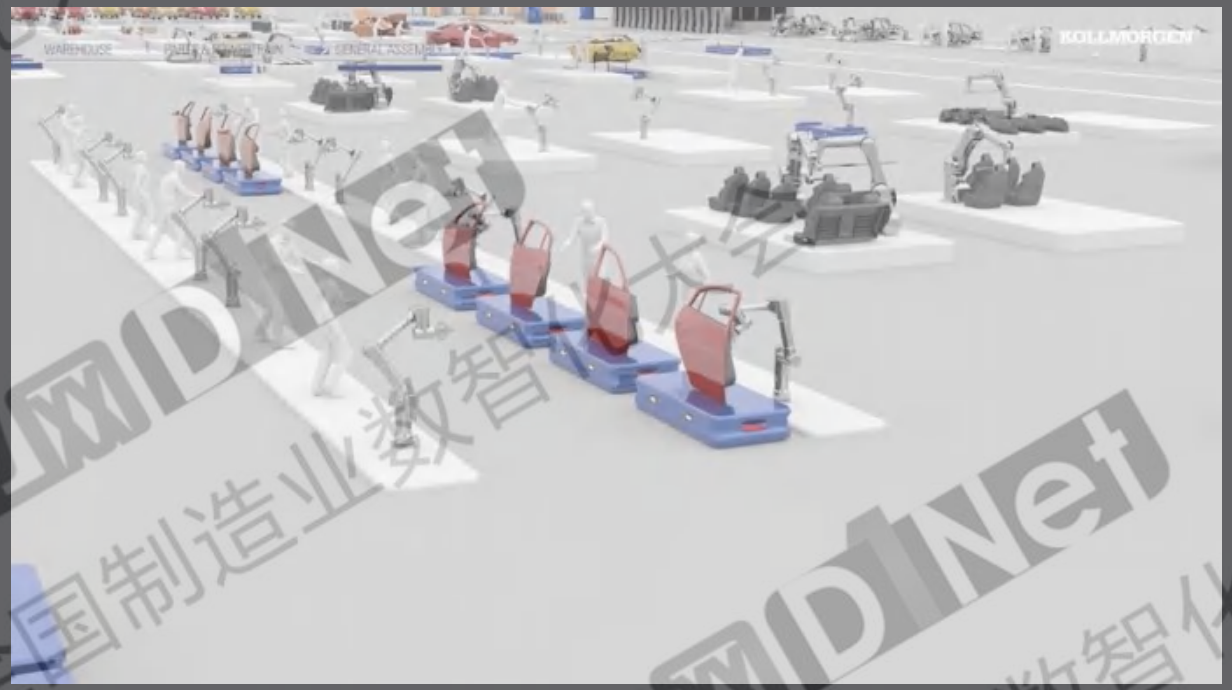
AI has revolutionized Industry 4.0 by making it possible for machines to analyze data and learn from them, make decisions, and take on their own actions that typically require human intelligence. This can lead to enhanced efficiency, productivity, and quality.

For example, AI can analyze sensor data from manufacturing equipment and identify opportunities for process improvements. It can also detect irregularities and predict failures before they even occur. This allows preventative maintenance and reduces downtime.

Robotics

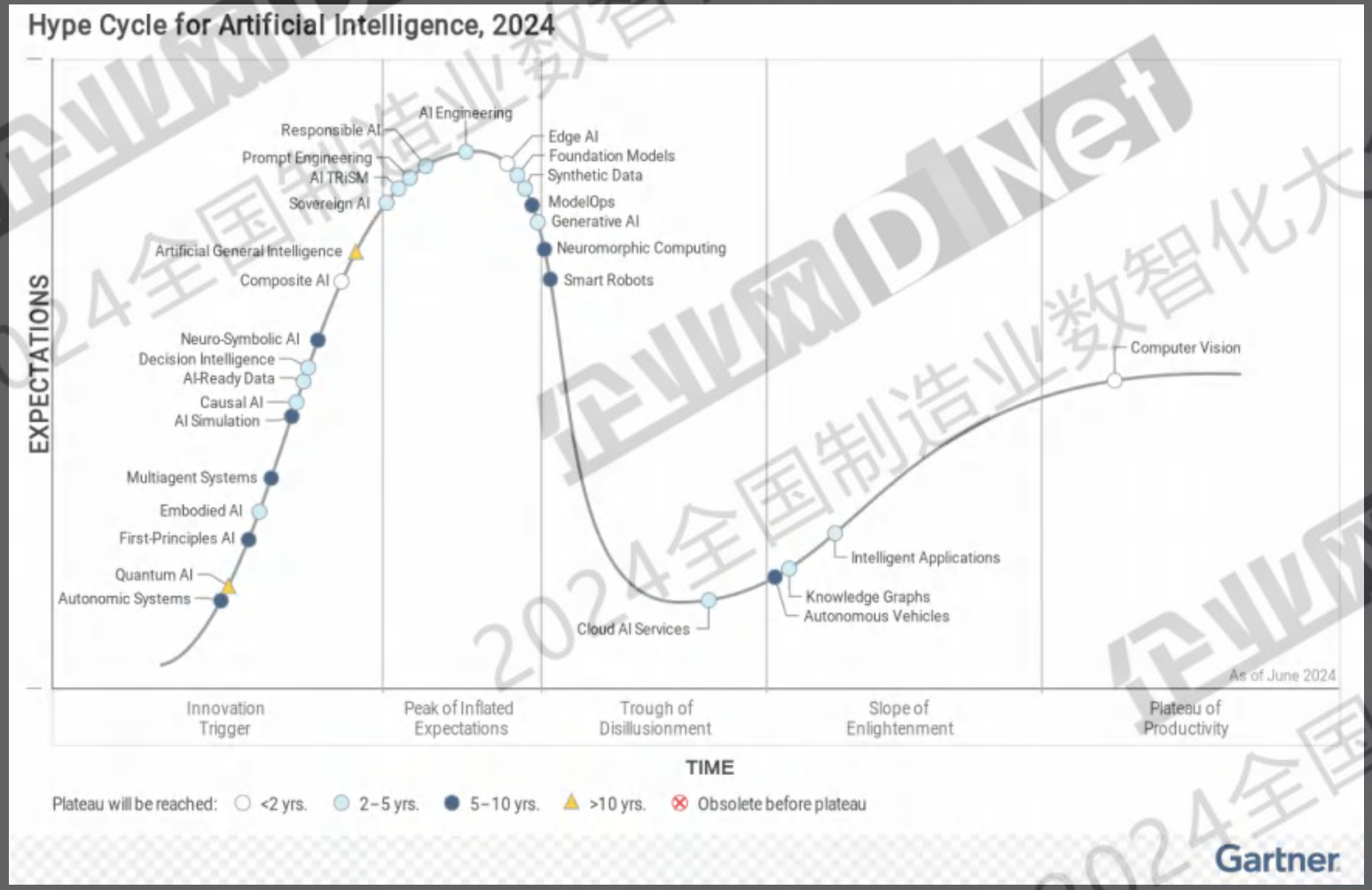
In Industry 4.0, robotics can be used to automate and perform a wide range of tasks. For instance, robots can be used to perform tasks such as welding, painting, and assembly. They work fast, accurately, and tirelessly, making them perfect for repetitive or hazardous jobs. Robots can also collaborate with human workers, creating more flexible and efficient production processes. In addition, robots sort and move packages in warehouses, making the supply chain faster and more reliable.

01 第四次工业革命



02 人工智能发展

02 人工智能成熟度曲线



技术成熟度曲线是 Gartner 于 1995 年首次采用的、用于分析及预测各种新技术在关注度、市场预期和实际应用中的成熟度和发展趋势。该曲线将一项技术的发展分为了 5 个阶段：

- **技术启动 (Innovation Trigger)**
  该技术开始获得媒体关注、产生舆论，但是可能没有实际的产品和应用。

- **期望膨胀 (Peak of Inflated Expectations)**
  由于媒体过度炒作，导致公众对该技术的期望被过度放大。 此期间可能会出现一些成功案例，但更多的是失败的尝试。

- **失望谷 (Trough of Disillusionment)**
  当实际效果达不到过度炒作的期望时，工作会开始对技术感到失望。

- **启蒙坡道 (Slope of Enlightenment)**
  一些企业开始了解如何使用该技术，并开始看到其潜在的效益。

- **生产高地 (Plateau of Productivity)**
  该技术已经成熟且被广泛的理解和接受，被大众所使用。

02 技术创新期

自主系统
是自我管理的物理或软件系统，三个基本特征：自主性（无需外部协助即可自主执行自己的决策和任务）；学习性（根据经验、不断变化的条件或目标改变其行为和内部操作）；和代理（了解自己的内部状态和目的，指导其学习方式和内容，并使其能够独立行动）。

量子人工智能
是量子技术与人工智能交叉领域中新兴的研究领域。量子人工智能旨在利用量子力学的独特性质来开发新的、更强大的人工智能算法，有可能产生专为在量子系统上运行而设计的新型人工智能算法。

第一性原理人工智能
将物理和模拟原理、控制定律和领域知识融入人工智能模型。FPAI 将人工智能工程扩展到复杂系统工程和基于模型的系统。

具身人工智能
物理或虚拟人工智能代理的模型经过训练，并与其用户界面、传感器、外观、执行器或与特定、真实或模拟环境交互所需的其他功能共同设计。

多智能体系统 (MAS)
是一种由多个独立（但可交互）智能体组成的 AI 系统，每个智能体都能够感知其环境并采取行动。

人工智能模拟
是人工智能和模拟技术的综合应用，共同开发人工智能代理以及可以对其进行训练、测试和部署的模拟环境。

因果人工智能
识别并利用因果关系，超越基于相关性的预测模型，实现能够更有效地规定行动并更自主地行动的人工智能系统。

02 技术创新期

人工智能数据
证明数据适合特定 AI 用例的能力决定了数据是否为AI 就绪数据。

决策智能 (DI)
通过明确理解和设计决策方式以及如何通过反馈评估、管理和改进结果来促进决策。

神经符号 AI
是一种复合 AI，它将机器学习 (ML) 方法与符号系统（例如知识图谱）相结合，以创建更强大、更可靠的 AI 模型。这种融合使概率模型与明确定义的规则和知识相结合，使AI 系统能够更好地表示、推理和概括概念。这种方法为更有效地解决更广泛的业务问题提供了推理基础架构。

复合人工智能
是指将不同的人工智能技术进行组合应用（或融合），以提高学习效率，拓宽知识表达层次，拓宽人工智能抽象机制，最终提供一个有效解决更广泛业务问题的平台。

★ **通用人工智能 (AGI)**
又称强人工智能，是（目前假设的）机器智能，可以完成人类可以执行的任何智力任务。AGI 是未来自主人工智能系统的一种特性，它可以在广泛的现实或虚拟环境中实现目标，其效率至少与人类相当。

02 期望膨胀期

主权人工智能
是民族国家为减少对商业市场的依赖而自主开发和使用人工智能所做的努力。它体现了政治和文化差异，以推进主权目标，包括在制定人工智能战略以实现价值的同时，减少主权造成的伤害。鉴于主权人工智能创新与损失比率存在巨大差异，主权人工智能以意想不到的方式影响着国际关系、全球贸易和经济市场。

AI 信任、风险和安全管理 (AI TRiSM)
确保 AI 治理、可信度、公平性、可靠性、稳健性、有效性和数据保护。AI TRiSM 包括模型和应用程序透明度、内容异常检测、AI 数据保护、模型和应用程序监控和操作、对抗性攻击抵抗以及 AI 应用程序安全的解决方案和技术。

即时工程
是一门学科，以文本或图像的形式向生成式人工智能 (GenAI) 模型提供输入，以指定和限制模型可以产生的响应集。输入会提示一组产生所需结果的响应，而无需更新模型的实际权重（如微调所做的那样）。即时工程也称为“情境学习”，其中提供示例来进一步指导模型。

负责任的人工智能 (RAI)
是一个总称，指的是在采用 AI 时做出适当的商业和道德选择的各个方面。这些包括商业和社会价值、风险、信任、透明度、公平性、偏见缓解、可解释性、可持续性、问责制、安全性、隐私和法规遵从性。RAI 涵盖组织责任和实践，以确保积极、负责和合乎道德的 AI 开发和运营。

★ **人工智能工程**
是企业大规模交付 AI 和生成式 AI (GenAI) 解决方案的基础。该学科统一了 DataOps、MLOps 和 DevOps 管道，以创建连贯的企业开发、交付（混合、多云、边缘）和运营（流式、批处理）AI 系统。

02 期望膨胀期

边缘 AI
是指嵌入非 IT 产品（消费/商业）、物联网端点、网关和边缘服务器的 AI 技术。它涵盖消费、商业和工业应用的用例，例如移动设备、自动驾驶汽车、增强的医疗诊断功能和流视频分析。虽然主要侧重于 AI 推理，但更复杂的系统可能包括本地训练功能，以在边缘提供对 AI 模型的优化。

基础模型
是大参数模型，以自监督方式在广泛的数据集上进行训练。它们大多基于变换器或扩散深度神经网络架构，并且越来越多地采用多模态。它们之所以被称为基础模型，是因为它们对各种下游用例至关重要且适用。这种广泛的适用性归功于模型的预训练和多功能性。

合成数据
是一类人工生??成的数据，而不是通过直接观察现实世界获得的数据。合成数据在各种用例中用作真实数据的代理，包括数据匿名化、人工智能和机器学习 (ML) 开发、数据共享和数据货币化。

模型操作化（ModelOps）
主要关注高级分析、人工智能和决策模型的端到端治理和生命周期管理，例如基于机器学习（ML）、生成人工智能（GenAI）、知识图谱、规则、优化、语言学、代理等的模型。

★ **生成式人工智能 (GenAI)**
通过从大量原始源内容库中学习来生成内容、策略、设计和方法的新派生版本。生成式人工智能对业务有着深远的影响，包括内容发现、创作、真实性和法规；人类工作的自动化；以及客户和员工体验。

02 失望谷地

神经形态计算
是一种利用数字或模拟处理技术更准确地模拟生物大脑运作的机制的技术。这些设计通常使用脉冲神经网络 (SNN) 而不是深度神经网络 (DNN)，采用非冯·诺依曼架构，处理单元简单，但互连性极高。

智能机器人
是一种由人工智能驱动、通常可移动的机器，旨在自主执行一项或多项物理任务。这些任务可能依赖于机器学习，或产生机器学习，机器学习可以融入未来的活动或支持前所未有的条件。智能机器人可以根据任务或用例分为不同类型，例如个人、物流和工业。

云 AI 服务
提供 AI 模型构建工具、预构建服务的 API 和相关中间件，支持以云服务的形式在预构建基础设施上运行机器学习 (ML) 和生成式 AI 模型的构建/训练、部署和使用。这些服务包括预训练的视觉、语言和其他生成式 AI 服务，以及自动化 ML 和微调，以创建新模型并自定义预构建模型。

02 启蒙坡道

自动驾驶汽车

使用各种车载传感和定位技术，例如激光雷达、雷达、摄像头、全球导航卫星系统(GNSS)和地图数据，结合基于人工智能的决策，无需人工监督或干预即可行驶。自动驾驶汽车技术正应用于乘用车、公共汽车和卡车，以及采矿和农业拖拉机等特定用例。

知识图谱

是物理和数字世界的机器可读表示。它们包括实体（人员、公司和数字资产）及其关系，这些关系遵循图形数据模型——节点（顶点）和链接（边/弧）的网络。

智能应用程序

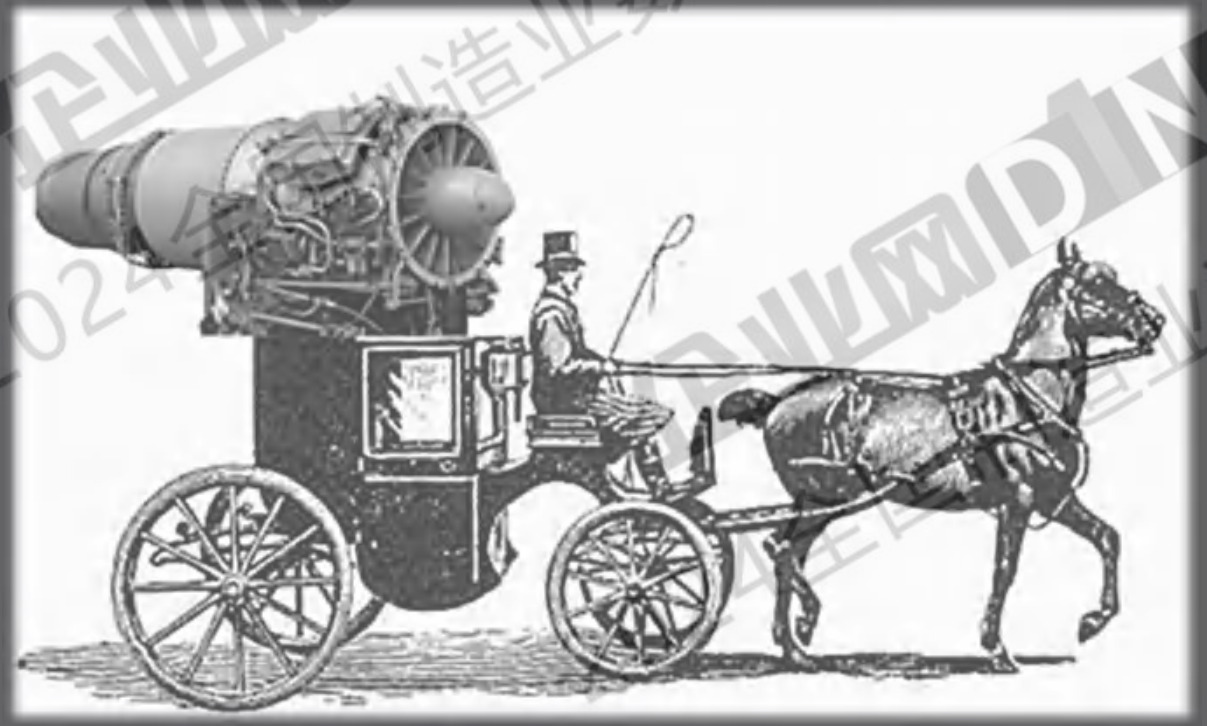
利用学习适应能力自主响应人类和机器。虽然应用程序可以表现得非常智能，但智能应用程序本质上是智能/主动的。基于条件逻辑的规则方法正在让位于基于数学的训练，以便在各种情况下（包括新情况或独特情况）做出适当的响应。这使得在各种场景和用例中增强和自动化工作成为可能。

02 生产力高原

计算机视觉

是一组技术，涉及捕获、处理和分析现实世界的图像和视频，以从物理世界中提取有意义的上下文信息。

02 人工智能时代来临



想象一下：

十九世纪早期的一个工程师正在想办法改进跨洋运输。

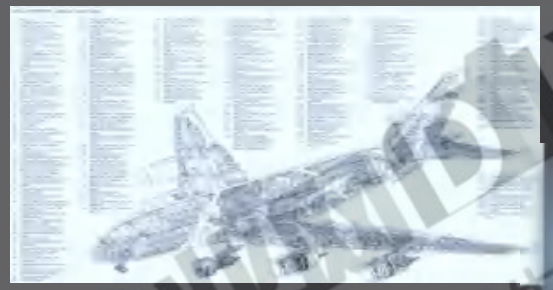
这时有人带着一个喷气发动机的设计图来找他，工程师说，“太好了，我们会把这个挂在马车上，帮马跑得更快。”

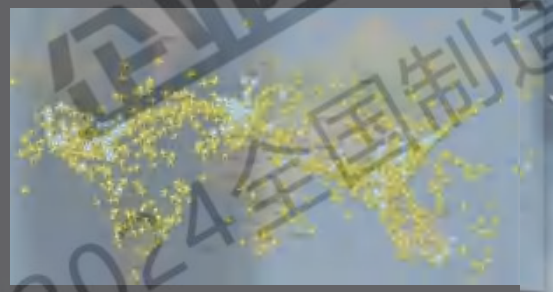
而当他们开始试验时，很快就会发现这存在一种危险，那就是发动机会把车辆摇得粉碎。

所以他们需要降低发动机的功率以确保其不会对马车造成伤害。

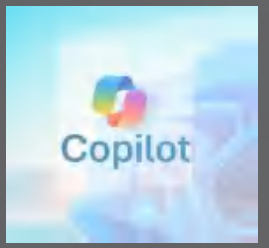
-------西摩·佩珀特

制造业数智化大会



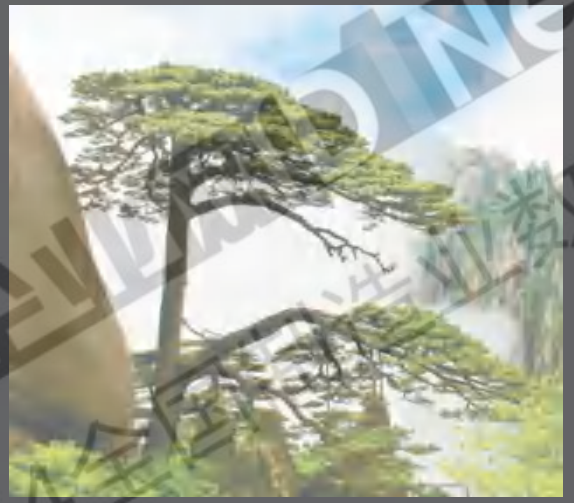


02 AI的未来-从大到小



02 AI的未来-从小到大

盆景



风景

风景区

02 AI的制造业应用

企业运营管理
提效赋能

产品全生命周期
智能化升级

生态多元化协同
深度融合

03 人工智能风险

03 人工智能数据安全风险

1 数据采集阶段

个人隐私 用户权利
过度采集 知识产权

2 数据处理阶段

数据污染 数据投毒攻击
数据偏差和歧视

3 数据流通阶段

数据交互 数据孤岛
数据跨境

4 数据使用阶段

关联分析 还原攻击
对抗样本

03 人工智能语料安全风险

- 包含违反社会价值观的内容
- 包含歧视性内容
- 商业违法违规
- 侵犯他人合法权益
- 无法满足特定安全需求

03 人工智能算法模型风险

可解释性问题

算法模型复杂度越来越高，整个训练过程变成一个黑盒，
很难理解算法模型的内部工作机制。

不良信息传播

算法直接在包含噪声的互联网数据基础上进行建模训练；
没有将防范抵制不良信息的要求内化成算法的具体规则。

用户权益及信息保护

对用户的知情权、选择权等权益保障不足，
算法训练数据中个人信息的保护和过度收集问题。

鲁棒性

偏差，噪声，干扰，随机性

伦理偏见歧视

算法设计开发过程中可能带着设计者或开发者的偏见，
或采用带有偏见的数据而导致推荐结果出现偏见。

算法攻击

黑盒攻击，灰盒攻击，白盒攻击，推理攻击
对抗样本攻击，模型盗取，反演攻击

03 人工智能算法伦理风险

《关于加强科技伦理治理的意见》科技伦理原则	人工智能伦理准则	关键域
（一）增进人类福祉	（1）以人为本（For Human）	福祉、尊严、自主自由等
	（2）可持续性（Sustainability）	远期人工智能、环境友好、向善性等
（二）尊重生命权利	（3）合作（Collaboration）	跨文化交流、协作等
	（4）隐私（Privacy）	知情与被通知、个人数据权利、隐私保护设计等
（三）坚持公平公正	（5）公平（Fairness）	公正、平等、包容性、合理分配、无偏见与不歧视等
	（6）共享（Share）	数据传递、平等沟通等
（四）合理控制风险	（7）外部安全（Security）	网络安全、保密、风险控制、物理安全、主动防御等
	（8）内部安全（Safety）	可控性、鲁棒性、可靠性、冗余、稳定性等
（五）保持公开透明	（9）透明（Transparency）	可解释、可预测、定期披露和开源、可追溯等
	（10）可问责（Accountability）	责任、审查和监管等

03 人工智能可解释性风险

可解释的AI（Explainable AI，简称XAI）或透明的AI（Transparent AI），是一组流程和方法，让人类用户可以理解并信任机器学习算法创建的结果和输出。

可解释 AI 技术

- **预测准确性**

准确性是在日常运营中成功使用 AI 的关键因素。通过运行模拟并将 XAI 输出与训练数据集中的结果进行比较，可以确定预测准确性。在这方面，最主流的技术是模型无关的局部解释 (LIME)，它解释了 ML 算法对分类器的预测。

- **可跟踪性**

可跟踪性是实现 XAI 的另一关键技术。可通过多种方法实现可跟踪性，比如通过限制决策的制定方式，以及为 ML 规则和功能设置更小的范围。可跟踪性 XAI 技术的一个例子是 DeepLIFT（深度学习重要特征），该算法将每个神经元的激活与其参考神经元进行比较，并显示每个已激活神经元之间的可跟踪链路，甚至显示它们之间的依赖关系。

- **决策理解**

这是人为因素。许多人对 AI 并不信任，然而，要高效利用 AI，就需要学会信任 AI。通过教导团队使用 AI，可以建立对 AI 的信任，这样他们就能理解 AI 如何决策以及为何做出此等决策。

03 人工智能鲁棒性风险

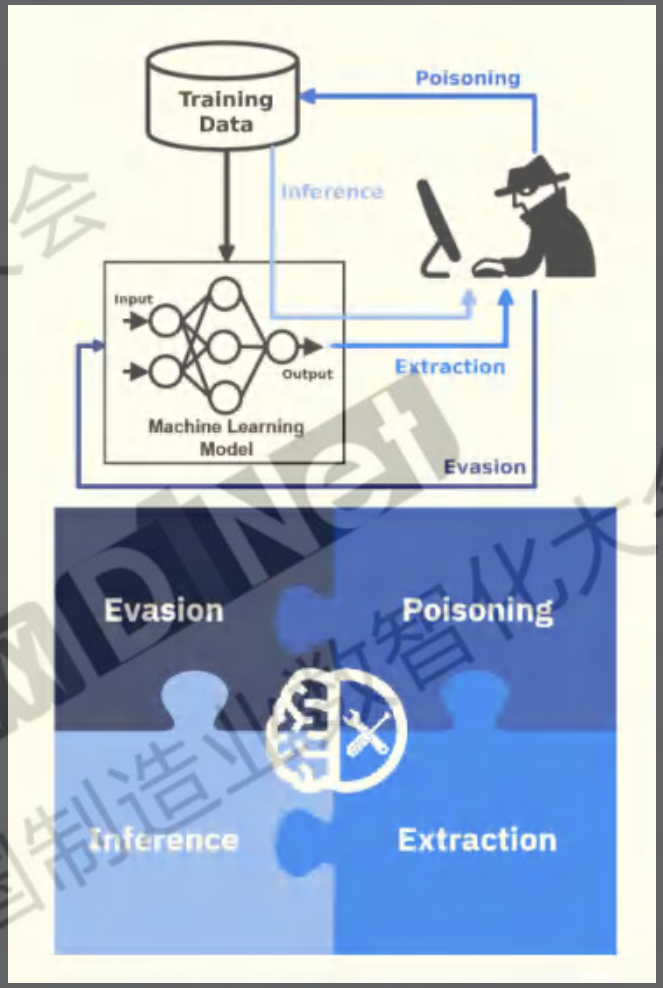
鲁棒是Robust的音译，也就是健壮和强壮的意思。它也是在异常和危险情况下系统生存的能力。比如说，计算机软件在输入错误、磁盘故障、网络过载或有意攻击情况下，能否不死机、不崩溃，就是该软件的鲁棒性。所谓“鲁棒性”，也是指控制系统在一定（结构，大小）的参数摄动下，维持其它某些性能的特性。根据对性能的不同定义，可分为稳定鲁棒性和性能鲁棒性。以闭环系统的鲁棒性作为目标设计得到的固定控制器称为鲁棒控制器。

鲁棒性包括稳定鲁棒性和品质鲁棒性。

AI模型的鲁棒可以理解为模型对数据变化的容忍度。假设数据出现较小偏差，只对模型输出产生较小的影响，则称模型是鲁棒的。

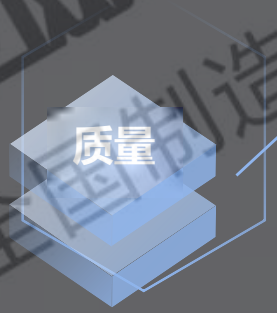
鲁棒性的3个要求：

- 模型具有较高的精度或有效性。
- 对于模型假设出现的较小偏差，只能对算法性能产生较小的影响。
- 对于模型假设出现的较大偏差，不能对算法性能产生“灾难性”的影响。



03 人工智能知识产权风险

03 人工智能生成内容风险

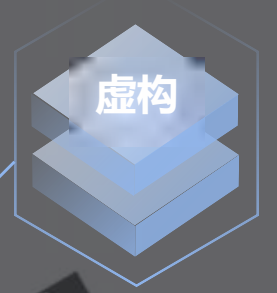


输出质量问题

由于其不可预测的性质，确保AIGC模型生成的输出质量极具挑战性。

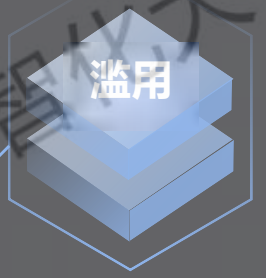
有偏见的输出

基于用于训练模型的数据中的偏见，AIGC模型与其他模型一样容易遭受有偏见输出的风险。例如，Stable Diffusion可能会根据提示显示“公司首席执行官”的图像，并只生成白人男性的图像。



虚构的事实和幻觉

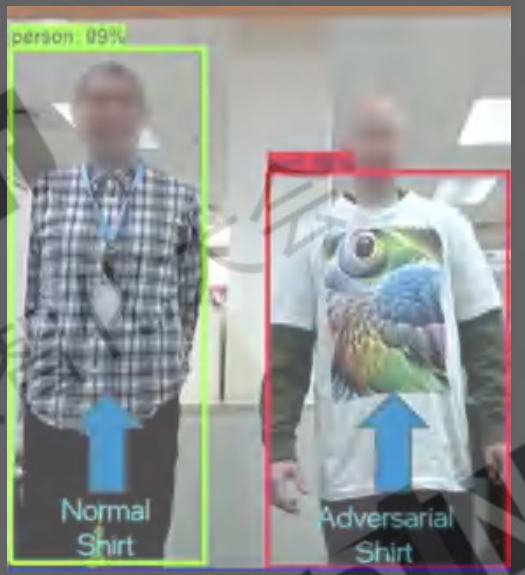
模型编造“事实”时的“幻觉”问题，模型产生幻觉的可能性意味着，在需要准确信息(如搜索)的情况下使用这些工具之前，需要设置重要的防护机制。

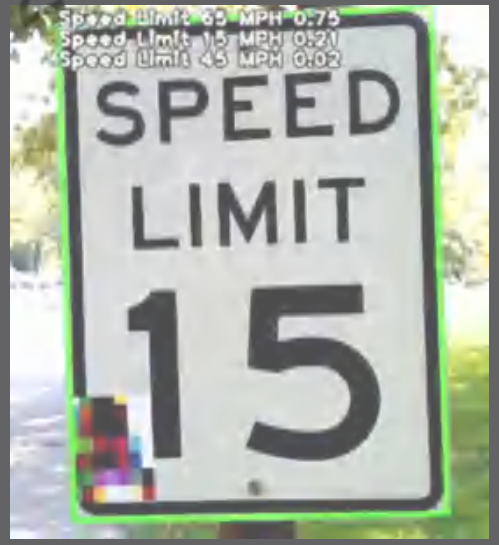


易被滥用

AIGC的绝对力量使其容易被“越狱”。虽然GPT的训练主要集中在单词预测上，但它的推理能力是一个意想不到的结果。随着我们在AIGC模型方面取得进展，用户可能会发现绕过模型最初预期功能的方法，并将其用于完全不同的目标。

03 人工智能实体安全风险



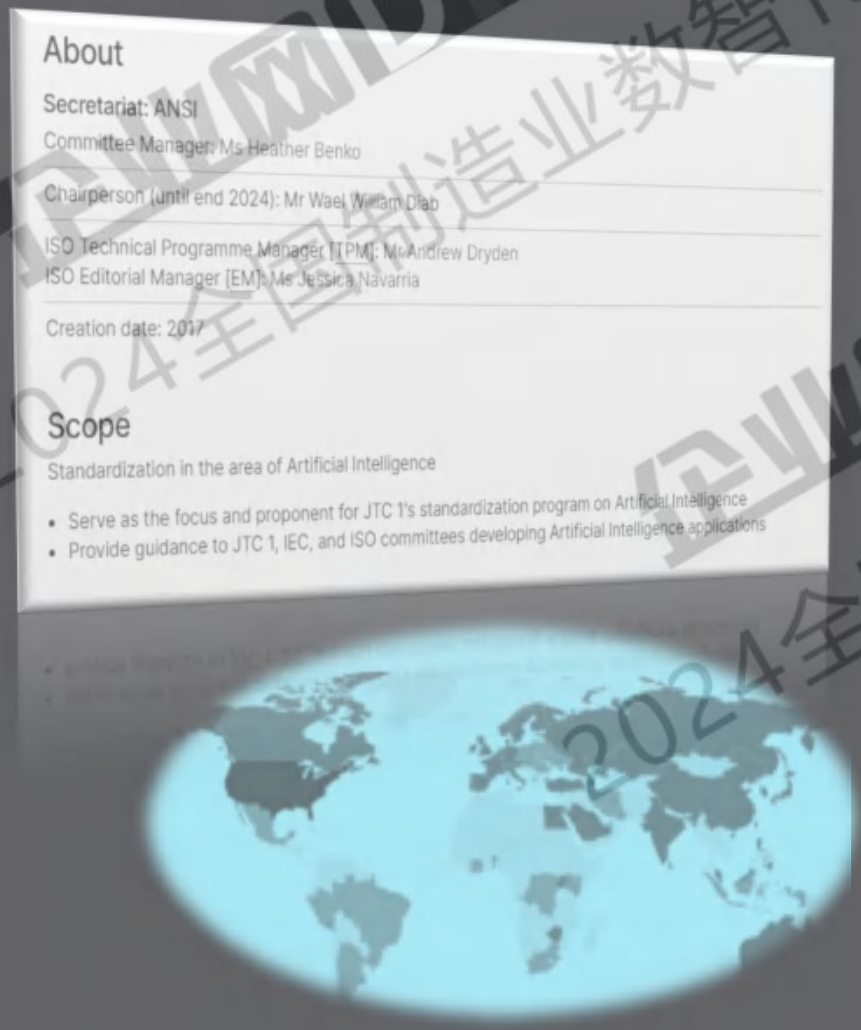


04 人工智能管理体系

04 人工智能管理体系

- AI领域的现状，就像在一个房间里挤满了才华横溢的人——工程师、伦理学家和数据魔法师——挤在虚拟的火堆周围，他们思考人工智能的道德困境：偏见、责任和难以捉摸的未知“黑匣子”。
- 我们迫切需要一份人工智能从业者的清晰地图，为机器学习、神经网络和算法奇迹的模糊领域指明道路。

04 ISO/IEC JTC I/SC 42 工作组



Reference ↑	Title
ISO/IEC JTC 1/SC 42/AHG 4	Liaison with SC 27
ISO/IEC JTC 1/SC 42/AHG 7	JTC1 joint development review
ISO/IEC JTC 1/SC 42/JWG 2	Joint Working Group ISO/IEC JTC1/SC 42 - ISO/IEC JTC1/SC 7 : Testing of AI-based systems
ISO/IEC JTC 1/SC 42/JWG 3	Joint Working Group ISO/IEC JTC1/SC42 - ISO/TC 215 WG : AI enabled health informatics
ISO/IEC JTC 1/SC 42/JWG 4	Joint Working Group ISO/IEC JTC1/SC42 - IEC TC65/SC65A: Functional safety and AI systems
ISO/IEC JTC 1/SC 42/JWG 5	Joint Working Group ISO/IEC JTC1/SC42 - ISO/TC 37 WG: Natural language processing
ISO/IEC JTC 1/SC 42/WG 1	Foundational standards
ISO/IEC JTC 1/SC 42/WG 2	Data
ISO/IEC JTC 1/SC 42/WG 3	Trustworthiness
ISO/IEC JTC 1/SC 42/WG 4	Use cases and applications
ISO/IEC JTC 1/SC 42/WG 5	Computational approaches and computational characteristics of AI systems

04 人工智能标准框架

基础标准	AI治理	风险管理	AI可信度
管理体系	AI系统影响评估	数据质量管理	AI系统质量模型
	测试	……	

04 已经发布的标准31个

Standard and/or project under the direct responsibility of ISO/IEC JTC 1/SC 42 Secretariat

ISO/IEC TS 4213:2022
Information technology — Artificial intelligence — Assessment of machine learning classification performance

ISO/IEC 5259-1:2024
Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples

ISO/IEC 5259-3:2024
Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelines

ISO/IEC 5259-4:2024
Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework

ISO/IEC 5338:2023
Information technology — Artificial intelligence — AI system life cycle processes

ISO/IEC 5339:2024
Information technology — Artificial intelligence — Guidance for AI applications

ISO/IEC 5392:2024
Information technology — Artificial intelligence — Reference architecture of knowledge engineering

ISO/IEC TR 5469:2024
Artificial intelligence — Functional safety and AI systems

ISO/IEC 8183:2023
Information technology — Artificial intelligence — Data life cycle framework

ISO/IEC TS 8200:2024
Information technology — Artificial intelligence — Controllability of automated artificial intelligence systems

ISO/IEC TR 17903:2024
Information technology — Artificial intelligence — Overview of machine learning computing devices

ISO/IEC 20546:2019
Information technology — Big data — Overview and vocabulary

ISO/IEC TR 20547-1:2020
Information technology — Big data reference architecture — Part 1: Framework and application process

ISO/IEC TR 20547-2:2018
Information technology — Big data reference architecture — Part 2: Use cases and derived requirements

ISO/IEC 20547-3:2020
Information technology — Big data reference architecture — Part 3: Reference architecture

ISO/IEC TR 20547-5:2018
Information technology — Big data reference architecture — Part 5: Standards roadmap

ISO/IEC 22989:2022
Information technology — Artificial intelligence — Artificial intelligence concepts and terminology

ISO/IEC 23053:2022
Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)

ISO/IEC 23894:2023
Information technology — Artificial intelligence — Guidance on risk management

ISO/IEC TR 24027:2021
Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making

ISO/IEC TR 24028:2020
Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence

ISO/IEC TR 24029-1:2021
Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview

ISO/IEC 24029-2:2023
Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods

ISO/IEC TR 24030:2024
Information technology — Artificial intelligence (AI) — Use cases

ISO/IEC TR 24368:2022
Information technology — Artificial intelligence — Overview of ethical and societal concerns

ISO/IEC TR 24372:2021
Information technology — Artificial intelligence (AI) — Overview of computational approaches for AI systems

ISO/IEC 24668:2022
Information technology — Artificial intelligence — Process management framework for big data analytics

ISO/IEC TS 25058:2024
Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guidance for quality (AI) systems

ISO/IEC 25059:2023
Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems

ISO/IEC 38507:2022
Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations

ISO/IEC 42001:2023
Information technology — Artificial intelligence — Management system

04 开发中的标准36个

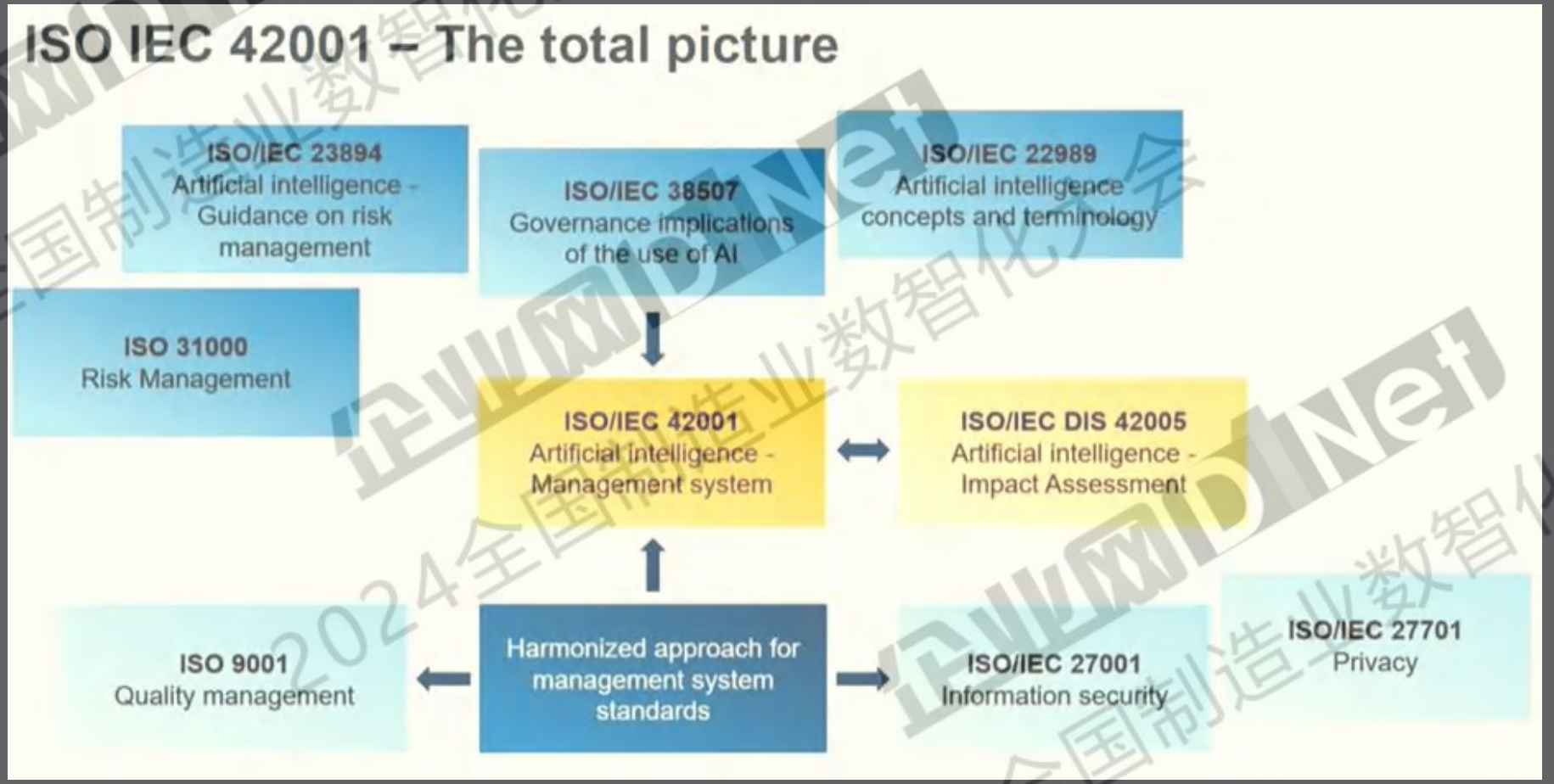
Standard and/or project under the direct responsibility of ISO/IEC JTC 1/SC 42 Secretariat

- ISO/IEC AWI 4213
  Artificial intelligence — Performance measurement for AI classification, regression, clustering and recommendation tasks
- ISO/IEC 5259-2
  Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures
- ISO/IEC FDIS 5259-5
  Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 5: Data quality governance framework
- ISO/IEC CD TR 5259-6
  Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 6: Visualization framework for data quality
- ISO/IEC CD TS 6254
  Information technology — Artificial intelligence — Objectives and approaches for explainability and interpretability of ML models and AI systems
- ISO/IEC DTS 12791.2
  Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks
- ISO/IEC DIS 12792
  Information technology — Artificial intelligence — Transparency taxonomy of AI systems
- ISO/IEC AWI TS 17847
  Information technology — Artificial intelligence — Verification and validation analysis of AI systems
- ISO/IEC AWI TR 18988
  Artificial intelligence — Application of AI technologies in health informatics
- ISO/IEC DTR 20226
  Information technology — Artificial intelligence — Environmental sustainability aspects of AI systems
- ISO/IEC CD TR 21221
  Information technology – Artificial intelligence – Beneficial AI systems
- ISO/IEC AWI TS 22440-1
  Artificial intelligence — Functional safety and AI systems — Part 1: Requirements
- ISO/IEC AWI TS 22440-2
  Artificial intelligence — Functional safety and AI systems — Part 2: Guidance
- ISO/IEC AWI TS 22440-3
  Artificial intelligence — Functional safety and AI systems — Part 3: Examples of application
- ISO/IEC AWI TS 22443
  Information technology — Artificial intelligence — Guidance on addressing societal concerns and ethical considerations
- ISO/IEC AWI 22989-2
  Artificial intelligence — Concepts and terminology — Part 2: Healthcare
- ISO/IEC 22989:2022/AWI Amd 1
  Information technology — Artificial intelligence — Artificial intelligence concepts and terminology — Amendment 1
- ISO/IEC 23053:2022/AWI Amd 1
  Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) — Amendment 1
- ISO/IEC AWI TR 23281
  Artificial intelligence — Overview of AI tasks and functionalities related to natural language processing
- ISO/IEC AWI 23282
  Artificial Intelligence — Evaluation methods for accurate natural language processing systems
- ISO/IEC AWI 24029-3
  Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 3: Methodology for the use of statistical methods
- ISO/IEC AWI 24970
  Artificial Intelligence — AI system logging
- ISO/IEC AWI 25029
  Artificial Intelligence — AI-enhanced nudging
- ISO/IEC AWI 25059
  Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems
- ISO/IEC AWI TS 25223
  Information technology — Artificial Intelligence — Guidance and requirements for uncertainty quantification in AI systems
- ISO/IEC AWI TS 25258
  Information technology — Artificial Intelligence — Hybrid AI Inference framework for AI systems
- ISO/IEC AWI TS 29119-11
  Software and systems engineering — Software testing — Part 11: Testing of AI systems
- ISO/IEC DIS 42005
  Information technology — Artificial Intelligence — AI system impact assessment
- ISO/IEC DIS 42006
  Information technology — Artificial Intelligence — Requirements for bodies providing audit and certification of artificial intelligence management systems
- ISO/IEC AWI 42102
  Information technology — Artificial Intelligence — Taxonomy of AI system methods and capabilities
- ISO/IEC AWI TR 42103
  Information technology — Artificial Intelligence — Overview of synthetic data in the context of AI systems
- ISO/IEC AWI 42105
  Information technology — Artificial Intelligence — Guidance for human oversight of AI systems
- ISO/IEC AWI TR 42106
  Information technology — Artificial Intelligence — Overview of differentiated benchmarking of AI system quality characteristics
- ISO/IEC AWI TR 42109
  Information technology — Artificial Intelligence — Use cases of human-machine teaming
- ISO/IEC AWI TS 42111
  Information technology — Artificial Intelligence — Guidance on lightweight AI systems
- ISO/IEC AWI TS 42112
  Information technology — Artificial Intelligence — Guidance on machine learning model training efficiency optimisation

04 ISO 42001人工智能管理体系

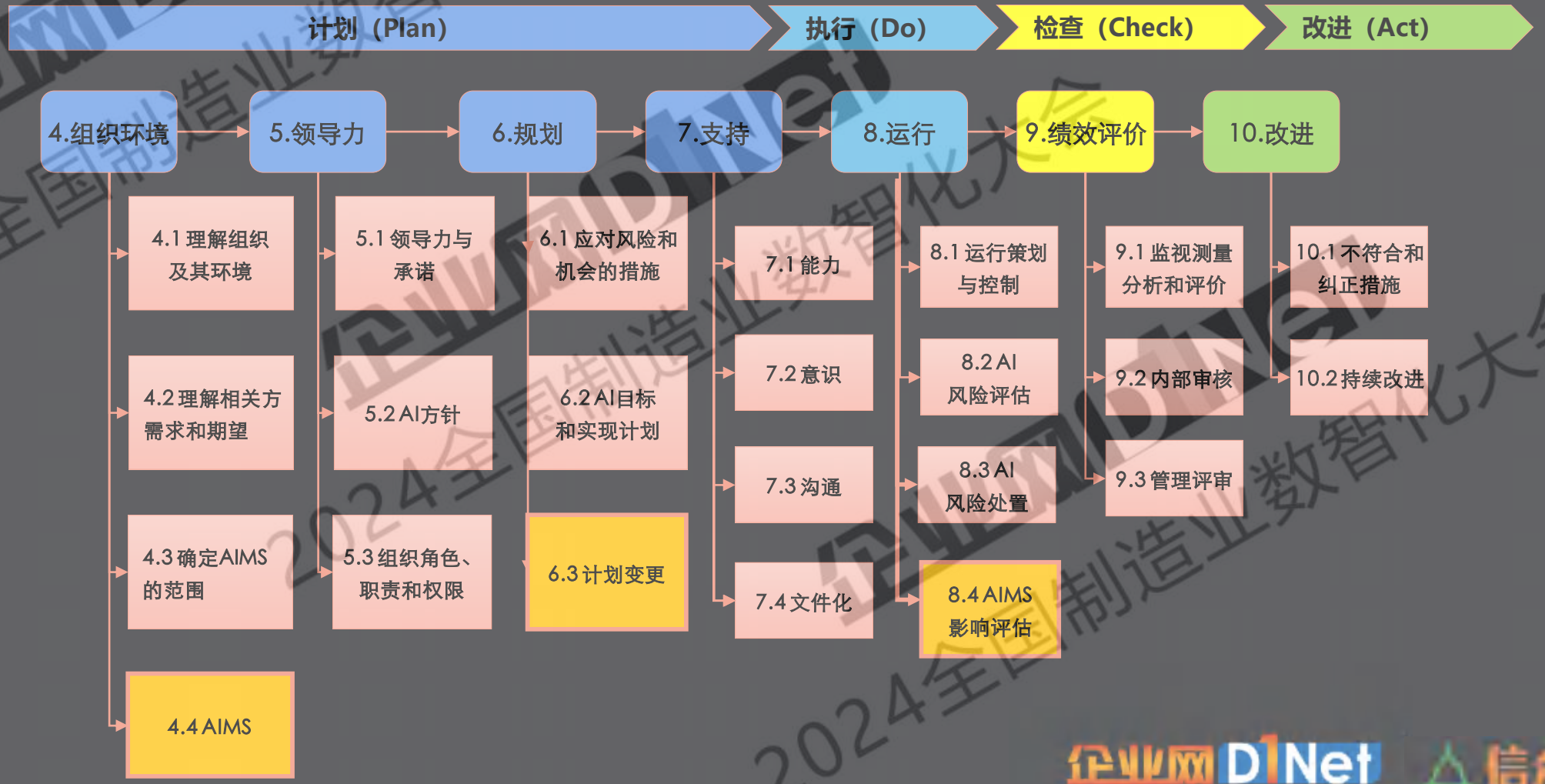
- 2023年12月，ISO和IEC发布了ISO/IEC 42001:2023。该标准类似于指南针，指导组织建立、实施和持续改进其人工智能管理系统 (AIMS)。
- 为寻求人工智能成功的组织提供了一份全面的指引，帮助组织在开发、提供或使用AI系统以追求其目标并满足相关方的适用要求、义务及对他们的期望。
- 适用于提供或使用人工智能系统的产品或服务的组织。
- 旨在帮助该组织在开发、提供或使用人工智能系统以追求其目标并满足相关方的适用要求、义务及对他们的期望。
- 无论大小、类型和性质如何，只要是提供或使用利用人工智能系统的产品或服务，本文档都适用于任何组织。

04 ISO 42001与其他标准的关系



04 ISO 42001标准的作用

04 ISO 42001标准正文



04 人工智能影响评估

04 ISO 42001标准附录A+B 目标 控制 实践

04 ISO 42001标准附录C 风险源

- 环境复杂性
- 系统硬件问题
- 透明度和可解释性
- 系统生命周期问题
- 自动化水平
- 技术成熟度
- 机器学习相关风险源
- 组织特定风险

04 ISO 42001标准附录C 风险目标

- 责任
- AI专业知识
- 训练和测试数据可用性
- 环境影响
- 公平性
- 可维护性
- 隐私问题
- 鲁棒性
- 实体安全
- 信息安全
- 透明和可解释性
- 组织具体目标

AI 赋能

让制造管理更智能

AI 体系

让制造应用更可信

2024 制造业数智化大会

谢谢观看！

宣讲人：刘歆轶　公司：非夕机器人