

2024上海图书馆开放数据竞赛巡讲 · 南京农业大学

计算伦理与数字道德 ——人工智能时代的人文主义



刘炜 上海图书馆上海科技情报所

wliu@libnet.sh.cn





正在呈现的八大威胁

1. 虚假信息与偏见误导
 2. 信息过载与信息茧房
 3. 技术素养与失业问题
 4. 侵犯隐私与信息泄漏
 5. 滥用误用与责任边界
 6. 侵犯版权与诱导犯罪
 7. 军事应用与生物威胁
 8. 意识觉醒与情感欺骗
-

计算伦理与数字道德

- 计算伦理：伦理是指对行为的规范和原则的系统研究，是价值观的组成部份。计算伦理关注和评估计算机技术对社会、个人及道德标准的影响，它涉及如何使用计算机技术的行为规范，包括数据隐私、安全性、公平使用和知识产权等方面，也包括广泛应用了计算技术之后，尤其是产生了各类高度发达的计算机智能体之后，伦理规范的主体性和差异性问题的探讨，探讨伦理的边界和机器伦理问题。
- 数字道德：道德是指个人或社会普遍接受的关于善恶、正义和义务的信念和行为规范，通常是基于文化、宗教和社会习俗的具体准则和规范。数字道德涵盖了所有数字技术的使用对社会伦理和个人行为的影响，甚至对人类的约束映射和延伸至高级智能体。数字道德探讨各类实体在数字世界中如何维持道德行为，包括社交媒体的伦理、数字隐私、元宇宙中的行为规范等等。
- 人文主义（Humanism）是一种哲学思想和文化思潮，强调人类价值、尊严和潜力，注重个人的理性、自由和道德责任。人文主义起源于文艺复兴时期，但其思想根源可以追溯到古希腊和古罗马的哲学传统。由于计算伦理与数字道德的外延扩大到了人类之外，人文主义的边界是否有必要扩展，是一个需要深入探讨的问题。

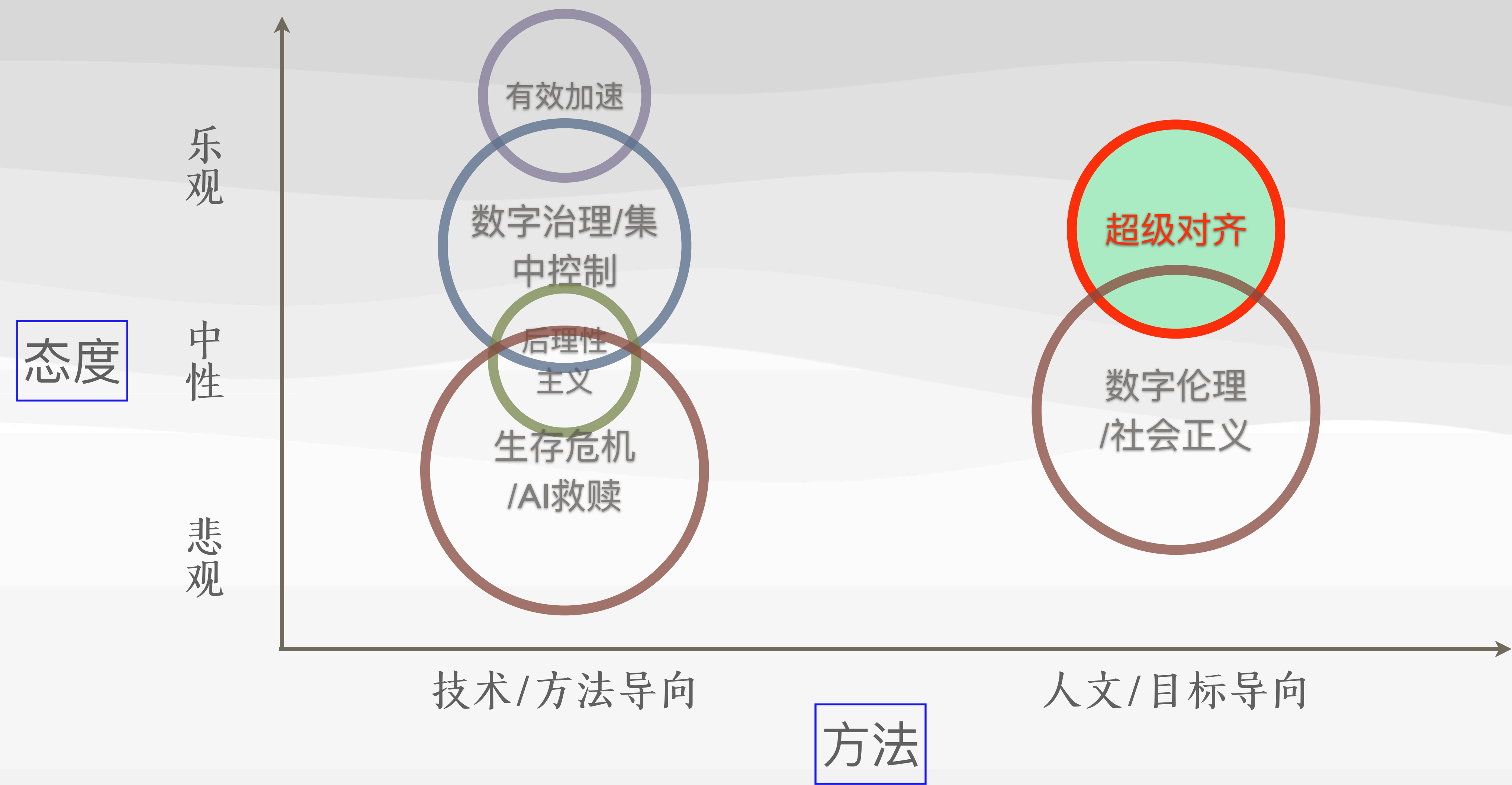
什么是人文主义：拉里佩奇与马斯克的分歧



E/ACC与SUPER LOVE ALIGNMENT



有效加速主义与超级对齐主义

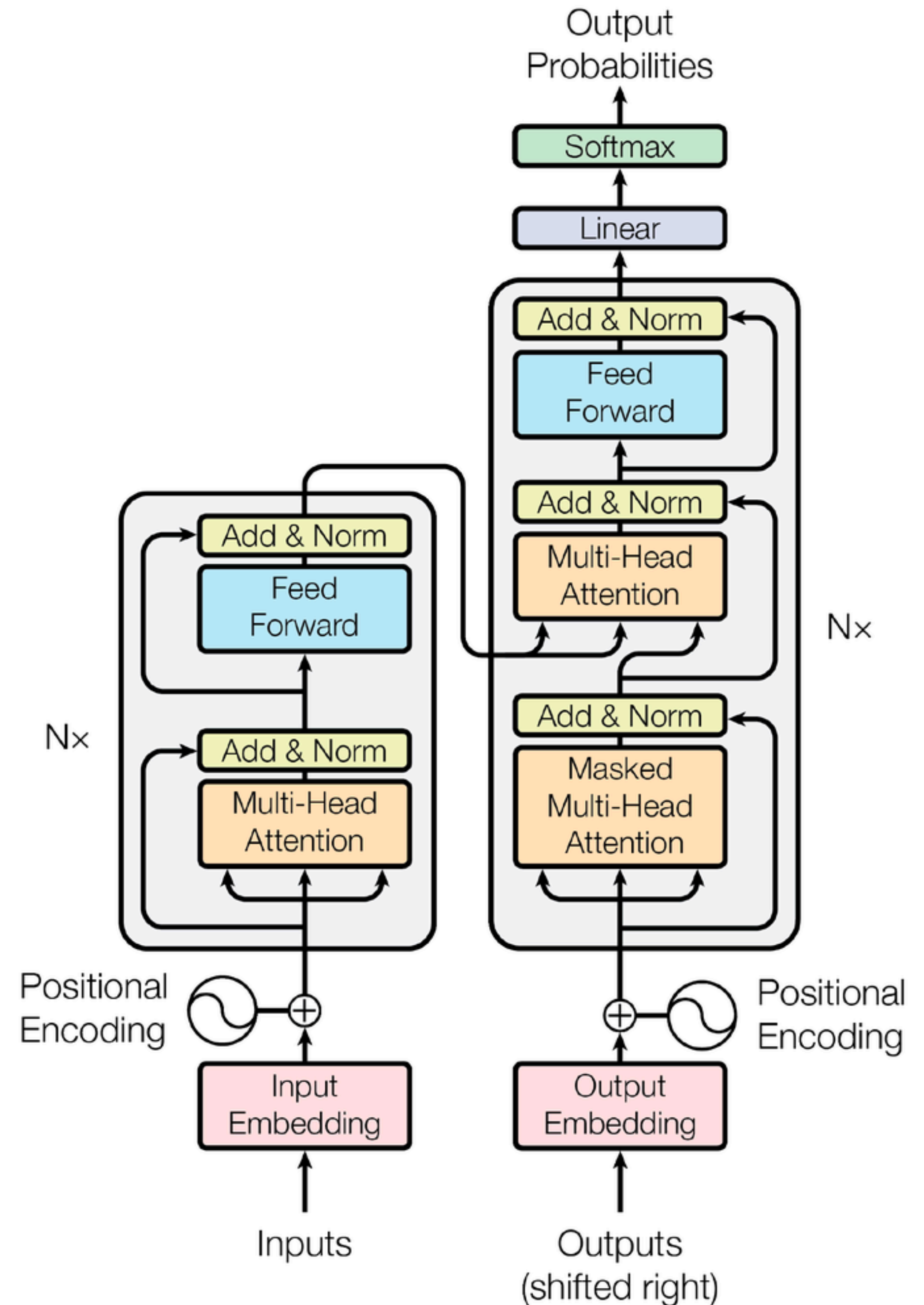


AGI之梦

- AGI（通用人工智能：Artificial General Intelligence）是指一种具备人类智能水平的机器，能够理解、学习和应用各种不同领域的知识。与当前专注于解决特定问题（如图像识别、自然语言处理或棋类游戏）的人工智能不同，AGI能够执行“任何”智能任务，包括任何领域的推理、规划、学习、交流，以及感知和与物理世界互动的能力。
- AGI四要素：1.具有通用智能，能够跨领域学习和工作；2.能够自我学习和适应环境；3.能够理解复杂/抽象概念并进行逻辑推理；4.甚至具有情感和意识！
- 目前的大语言模型还不是AGI，但在语言理解和多模态方面初步具备了泛化/涌现和推理能力，是目前最有可能发展成AGI的模型技术。但通过堆数据和堆算力是否能到达AGI还有争论，但普遍认为目前还未边际效应递减，但数据和电能都几近枯竭。

LLM如何被点化？

- 大型语言模型（LLM）是基于海量自然语言数据进行预训练而得到的超大型深度学习模型，参数通常从数十亿到超千亿。底层基于Transformer深度神经网络，由具有自注意力功能的编码器和解码器组成，但GPT只采用了解码器，从一系列文本中提取含义，并能够理解其中的单词和短语之间的关系。
- 用同样方法对海量图片、音频、视频等多媒体信息结合语言数据进行预训练和指令微调的超大型深度学习模型也是大语言模型的一种发展，通常称为多模态大模型。



大模型所具备的AGI特征

■ “泛化”能力

- Perplexity（困惑度）评价：用于评估语言模型在未见过的数据上的预测能力。困惑度越低表示模型在未见过的数据上表现越好。
- 语言模型的交叉验证：将数据集分为训练集、验证集和测试集，通过在验证集和测试集上的性能来评估模型的泛化能力。
- 零样本任务（Zero-shot Task）能力：在模型未见过的任务上进行评估，例如对模型提出一些与训练数据不相关的问题，评估其在这些任务上的表现能力。

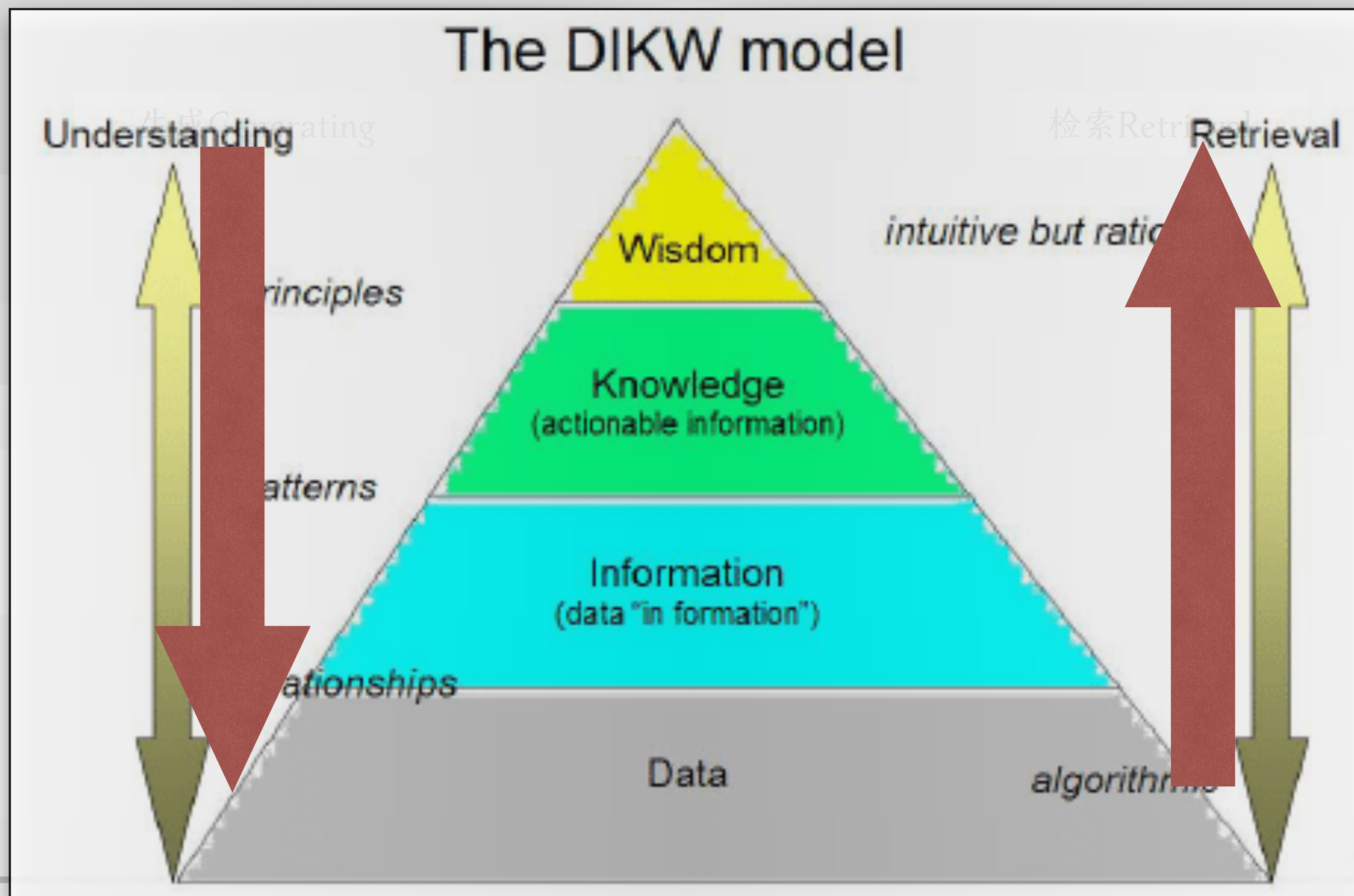
■ 推理能力

- 自然语言推理（NLI）任务：在给定前提和假设下，能否正确推断出假设的真假。
- 文本蕴含任务：在给定前提和假设下，能否判断假设是否可以从前提中推导出。
- 逻辑填空任务：要求模型填写一些语句中的空白，使得整个语句逻辑上合理。
- 逻辑推理任务：要求模型根据一些逻辑规则进行推理，例如判断一些命题是否成立或给出逻辑结论。

大模型的本质

- LLM是海量知识在深度神经网络中的压缩形态，可以认为是智慧的一种编码存储形式。
 - LLM是基于词元（token）而不是符号的：词元是一种张量，是语义相似性的度量而不是符号匹配，因此它可以直接告诉答案而不是符号的排列组合。
 - 知识是随时生成的而无需预先存储的：存储知识只是AI的Bootloader，具有具身学习能力的智能体无需存储知识，只需要参数权重存储的智能即可以随时产生知识。
 - LLM应用以端到端模型为最高形态，端到端是指只要给定数据就能得到智慧。
 - 人类作为知识链的起点，其知识生产虽然节能，但却是极其原始而粗糙的。LLM一旦形成便不再需要人类的帮助，可以通过自我学习（自己创造数据进行学习）而得到，并在应用中不断迭代（数据飞轮）。
 - LLM短期赋能传统的知识工作，长期将会颠覆整个知识产业模式。
-

大模型解密智慧：对知识编码



代表人物

- 辛顿是一个刚从未来返回的老者，忧心忡忡。
 - 马库斯是专门挑刺的愤青，不管什么都要怼。
 - 马斯克是受到惊吓的孩童，好像刚偷窥到黑屋子里的秘密。
 - 尤瓦尔是一位从远古穿越而来的巫师，擅长危言耸听。
-

驯服AI

FERTP保障：

- 公平性Fairness
- 可解释性Explainableness
- 健壮性Robustness
- 透明性Transparent
- 隐私性Privacy

原则：

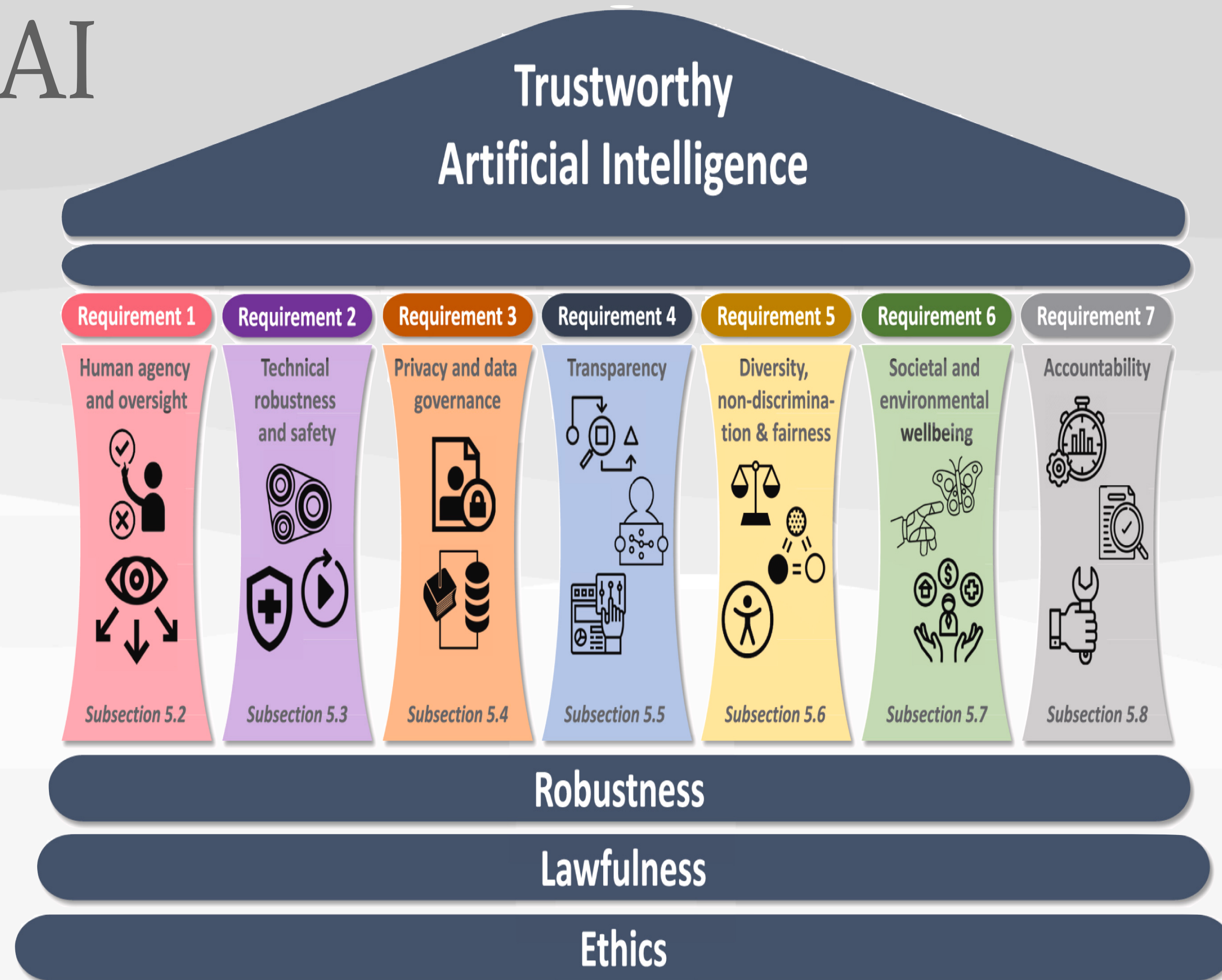
- AI是对人类智能的增强，而不是取代
- 数据及其洞察只属于数据创建者
- 整个数据链必须透明和可解释

AI五级素养体系

- 1. 会使用AI应用的基本功能，例如提示词组合，发挥其特殊能力，区别于传统应用
- 2. 能根据自己的需求寻找合适的AI应用，并会有意识地使用提示词框架，以及AI应用的不同前端版本
- 3. 会将不同的AI应用组合起来，实现一定的工作流，完成自己的日常工作或任务
- 4. 会设置大模型环境，并按目标准备和加工数据，安装调试自己的AI应用
- 5. 懂得目前大模型应用开发的堆栈框架，了解目前的不足，并对多模态和智能体等最新发展有所了解甚至提出和尝试探索路径

保障可信的 / 负责任AI

- 虚假信息与误导偏见
- 版权保护与隐私泄漏
- 技术滥用与失业威胁
- 伦理道德与责任边界
- 信息茧房与意识觉醒



ISO国际标准（例）

已经正式发布的

- ISO/IEC TR 24028:2020. Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence. <https://www.iso.org/standard/77608.html>
信息技术 — 人工智能 — 人工智能可信度概述
- ISO/IEC TR 24368:2022 Information Technology - Artificial Intelligence - Overview of ethical and societal concerns. <https://www.iso.org/standard/78507.html>
信息技术 — 人工智能 — 道德和社会问题概述

正在编制的

- ISO/IEC DTS 12791. [Under Development]. Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks. <https://www.iso.org/standard/84110.html>
信息技术 — 人工智能 — 处理分类和回归机器学习任务中不必要的偏差
- ISO/IEC AWI TS 22440. [Under Development]. Artificial intelligence — Functional safety and AI systems. <https://www.iso.org/standard/87118.html>
人工智能 — 功能安全和人工智能系统

曾蕾 7th未来智慧图书馆论坛&20th数图班 2024-05 上海



欧洲议会议员2024-03 批准世界上第一部全面的 人工智能法 (1/2)

[MEPs approve world's first comprehensive AI law - BBC News](#)

March 2024

- 欧盟人工智能法案是世界上第一套也是唯一一套具有约束力的要求，旨在降低人工智能风险。
- 法律的主要理念是根据人工智能对社会造成危害的能力对其进行监管。风险越高，规则越严格。
- 对基本权利构成 "明显风险" 的人工智能应用将被禁止，例如一些涉及生物识别数据处理的应用。
- 被视为 "高风险" 的人工智能系统，如用于关键基础设施、教育、医疗保健、执法、边境管理或选举的系统，将必须遵守严格的要求。
- 低风险服务，如垃圾邮件过滤器，将面临最宽松的监管——欧盟预计大多数服务都属于这一类。
- 该法案还制定了相关规定，以应对生成式人工智能工具和聊天机器人（如 OpenAI 的 ChatGPT）所依赖的系统带来的风险。
- 这些条款将要求一些所谓的通用人工智能系统（可用于一系列任务）的生产商对用于训练其模型的材料保持透明，并遵守欧盟版权法。

欧洲议会最近通过了开创性的人工智能法案，建立了世界上第一个用于管理人工智能 (AI) 相关风险的综合监管框架。

这项立法旨在解决人工智能行业快速扩张所带来的偏见、隐私和社会影响日益增长的担忧。

<https://futurium.ec.europa.eu/en/european-ai-alliance/forum-discussion/meps-approve-worlds-first-comprehensive-ai-law-what-benefits-does-it-potentially-deliver-ai-industry>

曾蕾 7th未来智慧图书馆论坛&20th数图班 2024-05 上海



欧洲议会议员2024-03 批准世界上第一部全面的 人工智能法 (2/2)

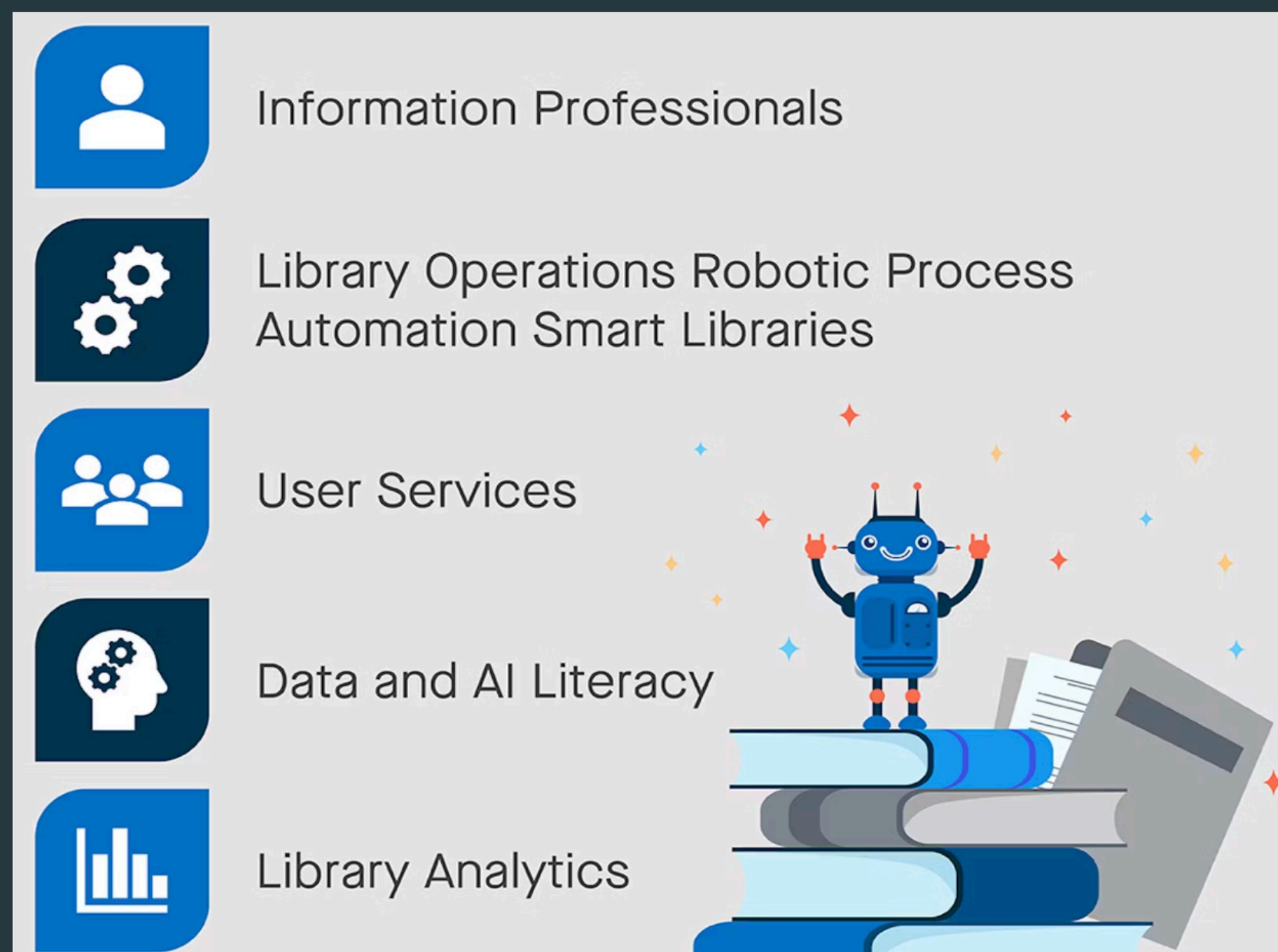
《人工智能法》的主要亮点包括：

1. 基于风险的分类： 对人工智能产品将根据其造成伤害的可能性进行分类，并相应地适用不同程度的审查。
2. 禁止高风险应用： 对基本权利构成明显风险的人工智能系统，如处理生物识别数据的系统，将被禁止。
3. 对高风险系统的严格要求： 医疗保健和执法等关键领域的人工智能应用将受到严格监管。
4. 对低风险服务的监管较轻： 垃圾邮件过滤器等产品将面临最少的监管，这反映了大多数人工智能服务。
5. 应对生成式人工智能的风险： 针对 OpenAI 的 ChatGPT 等系统的透明度和遵守版权法的规定。

<https://futurium.ec.europa.eu/en/european-ai-alliance/forum-discussion/meps-approve-worlds-first-comprehensive-ai-law-what-benefits-does-it-potentially-deliver-ai-industry>

人工智能给我们带来无尽的可能性

无尽的可能性既包括正面的对人类社会的福祉，也包括与之随行的风险和危机



<https://www.aje.com/arc/ways-artificial-intelligence-impacts-libraries/>

Library Assistant, Clerical -- ~95% amenable to automation

Library Technician -- ~99% at risk

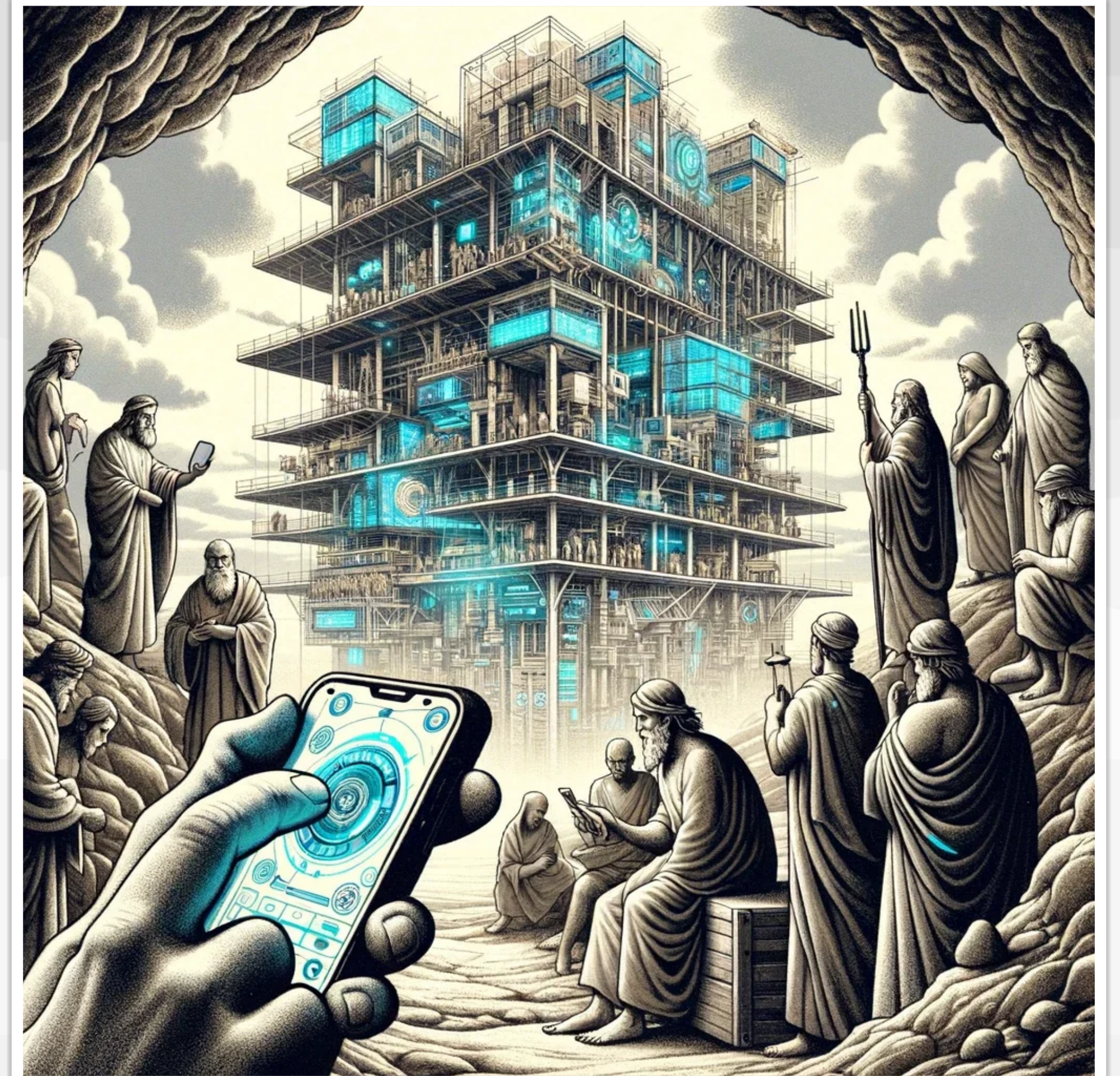
Librarian – 65% likely to be automated

(Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114, 254–280.

<https://doi.org/10.1016/j.techfore.2016.08.019>)

捍卫人文主义

图书馆是黑暗森林中的灯塔，
图书馆员是AI时代的领航员。



2024上海图书馆开放数据竞赛巡讲 · 南京农业大学

谢谢！



刘炜 上海图书馆上海科技情报所
kevenlw@gmail.com