双重脆弱性与适度信任：从ChatGPT到Sora

闫宏秀　宋胜男

（上海交通大学 科学史与科学文化研究院，上海 200240）

摘　要：Sora能够创造并处理复杂的动态视频内容，从理解静止的世界到理解运动的世界，标志着人工智能认识世界图景的重大转变。Sora对物理世界运动规律认识的不足和对细节的混淆是其技术脆弱性，由此加剧了信任脆弱性。因技术脆弱性风险，与此相伴的前置的、动态的和代理的新的信任模式随之而至。由技术脆弱性和信任脆弱性构成的双重脆弱性、积极的对抗脆弱性、无关脆弱性以及消极的对抗脆弱性四个象限，分别指向人工智能技术未来发展的四种样态。分析信任脆弱性与人工智能技术未来发展关系的四个象限可知，构建适度信任是破解信任与技术双重脆弱性的有效方式，适度信任构建本身需要以物理世界的因果律为基础、以人类信任为最后尺度、以向人类价值观保持对齐启蒙为前提、以充分证据为信任重建依据。

关键词：Sora；ChatGPT；双重脆弱性；适度信任；生成式人工智能

中图分类号：TP18；B15　　**文献标识码：**A　　**文章编号：**1005-9245（2024）06-0081-11

2023年5月，《生成式人工智能服务管理暂行办法》（以下简称《办法》）发布，《办法》第十七条提出生成式人工智能服务提供者应“提供可以影响用户信任、选择的必要信息，包括预训练和优化训练数据的来源、规模、类型、质量等描述”①。对生成式人工智能服务的信任与选择是当前的重要工作，关系人类与生成式人工智能的未来。生成式人工智能正以惊人的速度发展，从文本生成到图片生成，再到视频生成与制作，Sora的出现让人类与生成式人工智能的互动更进一步。这意味着人机交互的门槛不断降低、体验不断加深，AI离人类更近一步。但也应注意到，Sora的出现伴随技术与信任的双重脆弱性。

技术脆弱性来自Sora本身难以克服的技术缺陷，信任脆弱性来自对Sora等人工智能技术信任的盲目和不适度，这一信任从属于技术信任，是人类与Sora互动的核心部分，与技术脆弱性交织带来社会风险。适度信任对构建健康的人机关系至关重要，能够影响人工智能产品与服务标准的设置。例如，自动驾驶汽车的智能化应用程度、生成式人工智能产品（ChatGPT或Sora）的训练数据设置，等等。信任的缺乏或滥用极大影响人工智能产品与服务的安全使用。因此，对人工智能适度信任的追求是保证人工智能技术守住安全边界的关键，也是迈向人工通用智能（AGI）时代对人类社会的安全保障。

一、从技术脆弱性到信任脆弱性

“脆弱性”的英文为“Vulnerability”，意为

收稿日期：2024-03-29

基金项目：本文系教育部哲学社会科学研究重大课题攻关项目“数字化未来与数据伦理的哲学基础研究”（23JZD005）的阶段性成果。

作者简介：闫宏秀，上海交通大学科学史与科学文化研究院教授、博士生导师；宋胜男，上海交通大学科学史与科学文化研究院博士研究生。

① 国家互联网信息办公室等：《生成式人工智能服务管理暂行办法》，https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm。

“容易受到伤害、影响或攻击”[①]。脆弱性概念的根源在于生物伦理话语中身体伤害的可能性。从技术伦理角度看，脆弱性指由技术的不稳健性带来的风险伤害，这种不稳健性源于技术的不成熟或暂时无法突破的技术瓶颈。基于此，技术脆弱性使技术成为一把双刃剑，在对人类社会发挥巨大作用的同时，给人类带来风险和伤害。

（一）从静止的世界图景到运动的世界图景

从技术能力看，Sora 作为 AI 模型超越了 ChatGPT 文本生成模式，具有目前生成式人工智能前所未有的认知和生成能力，“能够生成具有多个角色、特定类型的运动以及主体和背景的准确细节的复杂场景。该模型不仅了解用户在提示中提出的要求，还了解这些东西在物理世界中的存在方式”[②]。科学家正尝试教授 AI 模型运动的世界图景，相较对静止的二维世界的认识，Sora 能理解并模拟真实物理世界的运动规律，解决了模型学习中的时空分割问题；相较只能输出对话、文章或代码的 ChatGPT，Sora 在技术功能方面取得更大进展。

近年来，人工智能技术的三个核心要素：大算力、大数据、大模型，均被视为重要资源，将上述资源进行恰当整合是人机融合发展的关键。在资源整合过程中，人类的信任逐渐成为最大的弱点。例如，在大算力、大数据、大模型的使用过程中，如果缺乏人与人之间的信任或人与机器之间的信任，技术的监管标准将提高，人机协作将更加困难；反之，如果信任过度，则难以避免资源融合过程中的过度技术化倾向，对技术过程的监管与回溯将成为难题。

基于此，伴随人工智能认知世界方式的转变，人类的信任方式需适应这一变化，实现从传统信任模式向人工智能信任模式的跨越。快速跨越将使信任发生背景性“脱节”。面对技术的飞速发展，英国社会学家安东尼·吉登斯（Anthony Giddens）曾用“脱域”（Disembeding）形容“社会关系从彼此互动的地域性关联中，从通过对不确定的时间的无限穿越而被重构的关联中‘脱离出来’”，“所有的脱域机制（包括象征标志和专家系统两方面）都依赖于信任（Trust）”[③]，在脱域过程中，信任发挥了关键作用。在数智时代，信任的发生从传统的、直接产生接触和互动的场景中脱离，逐渐演变为基于对技术的信任（Confidence）或依赖（Reliability）的新型信任模式，涵盖专家信任、系统信任和技术信任等多个维度，这种转变要求当前信任的动态调节更加灵活。

（二）从唯一的现实世界到虚拟的数字世界

Sora 打造的世界是区别于人类现实世界的虚拟数字世界，Sora 生成的视频带给人们强烈的真实感，其在视频生成时长、分辨率、内容等多个维度表现优异。与 ChatGPT 不同，Sora 旨在通过模型生成丰富的视觉体验，拓展虚拟世界的边界。Sora 的核心目标并非简单模仿现实世界，而是在虚拟领域创造出与现实世界相媲美的高质量视频内容，最终可能指向数字世界中“数字孪生”“具身智能”的发展。

OpenAI 官方将 Sora 定义为“具备理解和模拟动态现实世界能力的人工智能模型”[④]。该模型致力于通过虚拟化人物与物体，推动实体世界与虚拟世界相融合。但 Sora 成为世界模拟器的目的尚未明确，是否指向“数字孪生”“具身智能”技术的发展，或是作为迈向人工通用智能时代的前奏，仍有待探索。2012 年，美国航空航天局在其技术路线图中提出的“基于仿真的系统工程”（Simulation-Based Systems Engineering）部分，首次引入“数字孪生”概念。十余年来，数字孪生技术得到广泛关注，并在多个行业中应用。引发了深刻的哲学思考：人类生活的现实世界能否被数字化的虚拟世界替代？

相较 Sora 的强大功能，其脆弱性更值得人们关注。OpenAI 官方指出，“该模型还可能会混淆提示的空间细节，例如，混淆左右，并且可能难以精确描述随时间推移发生的事件，例如，遵循特定的相机轨迹”[⑤]。Sora 采用的“扩散变压器”（Diffusion Transformer）架构，在处理序列数据过程中展现的生成序列特性，在连贯性和现实性方面存在局限性，即“该序列既不完全连贯，也不完全现实”[⑥]。

① C.Levine，R.Faden，C.Grady，et al.The limitations of “vulnerability” as a protection for human research participants，The American Journal of Bioethics，2004（3）：44-49.

②⑤ OpenAI.Creating video from text，https://openai.com/sora#capabilities.

③ [英]安东尼·吉登斯：《现代性的后果》，田禾译，南京：译林出版社，2011 年版，第 18-23 页。

④ OpenAI.Creating video from text，https://openai.com/sora#research.

⑥ Are video generation models world simulators? https://artificialcognition.net/posts/video-generation-world-simulators/#concluding-thoughts.

通过观察 Sora 生成的作品可知，Sora 在模拟真实世界的物理规律和三维空间运动方面尚存在不足，导致生成的视频中出现不符合现实逻辑的场景。例如，在跑步机上逆向跑步、自发出现的灰狼幼崽、篮球穿越篮框，等等。这些问题揭示了 Sora 在空间细节识别和因果关系理解上的局限以及对物理规律掌握的不足。因此，在技术应用和伦理安全方面，Sora 需进一步探索和完善。虚拟世界与现实世界是异质的，随着技术的不断进步，人们的价值观念和发展目标将不断配合技术发展而调整，人类价值结构面临技术化的解构与重构的压力。

（三）技术脆弱性引发信任脆弱性

与 ChatGPT 等生成式人工智能不同，Sora 的文生视频模式在人机交互方面提供了更低的门槛和更强烈的体验感，使人工智能技术更加贴近人们的日常生活。Sora 呈现的创新型交互模式为生成式人工智能的发展开辟了新的可能性。但 Sora 尚存在难以克服的弱点："它可能难以准确模拟复杂场景的物理原理，并且可能无法理解因果关系的具体实例。"[①] 这一局限将影响 Sora 的准确性与可靠性，并引发信任脆弱性。

人类善于探究和把握因果关系，对人工智能而言，对因果关系以及细枝末节的把握只是学习的一部分，这一过程是基于人类提供的数据进行的无意识训练，能否完全掌握这些能力，需要经过实践验证。尽管 Sora 存在的明显弱点为人工智能系统带来了潜在风险，但这些风险并未减弱人们对 Sora 的热情。有观点认为，Sora 带来的收益远超其潜在的风险。这种对 Sora 的盲目信任本身就是更深层次的风险因素，因为它可能导致人们降低对风险的警觉性，从而降低对 Sora 的安全和伦理标准要求，增加风险的可能性。这种信任较脆弱，一旦 Sora 发生重大失误，信任将立刻消失，取而代之的是质疑与问责。此时，技术脆弱性转化为信任脆弱性。

Sora 尚不存在"自制"能力，科学家将此类人工智能系统视为增强人类能力的方式，但实际上，这种信任建立在一定的风险之上。首先，确保 Sora 的文字输入与视频输出的安全性是一个重要议题。OpenAI 官方给出的解释是，"在 OpenAI 产品中，我们的文本分类器将检查并拒绝违反我们的使用政策的文本输入提示，例如，要求极端暴力、性内容、仇恨图像、名人肖像或他人 IP 的文本输入提示。我们还开发了强大的图像分类器，用于检查生成的每个视频的帧，以帮助确保它在向用户显示之前符合我们的使用政策"[②]。OpenAI 的公开资料显示，该组织已开发文本分类器用于筛查并拒绝违反使用政策的文本输入提示。此外，其构建了先进的图像分类器，对生成的每个视频帧进行检查，确保在向用户展示之前，内容符合既定的使用政策。其次，防止用户对 Sora 技术的不当使用是另一项挑战。据观察，Sora 能有效处理短期和长期依赖关系，"我们发现 Sora 通常（尽管并非总是）能够有效地对短期和长期依赖关系进行建模"[③]。这表明 Sora 在理解和生成复杂场景方面可能存在局限，需要进一步优化技术和监管措施。最后，如何确保 Sora 能及时从不断变化的人类现实世界中学习，避免因模型学习滞后带来风险也是亟待解决的问题。为此要不断更新和优化模型，适应新的数据和现实世界的变化，同时确保使用效果的有效性。基于以上分析，建立在技术脆弱性之上的信任同样脆弱，技术脆弱性一定程度导致了信任脆弱性。

二、技术脆弱性风险下的人工智能信任生成模式

如何与 Sora 建立适度信任是亟须解决的问题。回顾计算机的发展历史，对自动化的信任、对互联网的信任以及对网络系统的信任，是计算机科学和认知系统工程中非常关键的问题[④]。随着计算机自动化和智能化程度的不断提高，人们对其越发感到担忧。高度智能化的产品和服务不仅涉及设计者、研发者，而且包括广大使用群体，所以对人工智能系统的信任是否适度要经过严格考察，这将牵扯多方利益。不信任人工智能是因为"复杂系统的表现

① OpenAI. Creating video from text，https://openai.com/sora#research.

② OpenAI. Creating video from text，https://openai.com/sora#safety.

③ OpenAI. Video generation models as world simulators，https://openai.com/research/video-generation-models-as-world-simulators.

④ R.R.Hoffman.A taxonomy of emergent trusting in the human–machine relationship，Cognitive Systems Engineering，2017：137-164.

是难以理解的，好像也经常违反直觉”[①]。技术中令人难以理解的部分通常消解了信任的可能性，但对技术的好奇与期望又重新培养了人们对技术的信任。因此，技术与信任之间出现了难以弥合的鸿沟，这一鸿沟加剧了人工智能信任的脆弱性。在技术脆弱性风险下，新的信任模型得以生成。

（一）前置模式的人工智能信任

人工智能信任与人际间的信任不同，人工智能信任的付出通常先于信任证据的产生，人们要获得人工智能的技术服务，必须先信任人工智能技术产品和服务，可称为人工智能信任的前置性。因为人工智能信任的前置性，人与人之间的信任所包含的要素关系（例如，诚实、正直、公正等）难以被接续应用到人工智能信任中，人工智能唯一参与双方信任关系的要素是技术能力。人工智能的技术能力是人们选择对其付出信任并与其构建信任关系最关键的因素。由于算法的局限性和其他弱点，人工智能技术能力在根本上并不稳定，因此，人类对算法、人工智能会“干坏事”的担忧一直存在。在此情况下，信任的前置性源于两个方面。

一方面，信任的前置性源于技术拒绝对个体的危害。技术拒绝指人类不给予某项技术产品以信任授权，因而无法享受技术产品带来的便利与效益。如同在电子商务中信用支付出现前，支付服务具有前置性，顾客必须先完成支付才能获得商品或服务。对人工智能的技术服务而言，人类的信任具有前置性，如果不先付出信任同意使用人工智能产品所需条款和规范，那么就会受到技术拒绝，无法完成对某项技术的尝试与使用。从某种程度而言，信任是类似于货币资源的存在，在利用信任兑换智能产品服务的过程中，信任必须前置。在时间关系上，人工智能信任的发生先于人工智能产品的使用，缺乏证据的人工智能信任是脆弱的。

另一方面，信任的前置性源于人类与人工智能的依赖共生关系。人工智能技术的发展需要人类信任的前置。人工智能技术之所以脆弱，是因为其对数据的高度依赖。人工智能系统需要大量数据进行训练和学习，这些数据源于不断更新的人类世界。Sora 和 ChatGPT 都是基于大型语言模型（Large Language Model，LLM）预训练的新型生成式人工智能，采用“利用人类反馈中强化学习”（RLHF）的训练方式，在人类与机器的互相问答过程中不断进化和迭代，逐渐提升模型生成答案的准确性。没有人类数据的训练，再先进的学习模型也会面临“巧妇难为无米之炊”的困境。因此，基于技术发展的角度考虑，信任的前置付出对人工智能生成模型的进步至关重要。任何一款人工智能产品在推出后都希望能得到人们的信任与推广，没有得到信任就会因为缺乏数据而成长缓慢直至被淘汰。

前置性的人工智能信任是技术脆弱性风险下生成的信任新模式，在这一信任模式下，始终存在人类无法全面了解人工智能技术的意图与行为的问题。技术脆弱性带来的信任脆弱性由此产生。

（二）动态模式的人工智能信任

人工智能技术应用过程中的任何反馈都将成为影响人工智能信任的变量，人工智能信任呈现根据情境变化进行动态调整的状态。加拿大经济学家罗伯特·霍夫曼（Robert R.Hoffman）认为，“在不断变化的工作和不断变化的系统的范围内，积极探索和评估可信度和可靠性的持续过程”[②]，信任具有一定的安全阈值，在安全阈值内信任可以根据实际情况作出调整。

动态性的人工智能信任与风险变化息息相关。吉登斯认为：“风险和信任交织在一起，信任通常足以避免特殊的行动当时所可能遇到的危险，或把这些危险降到最低的程度。”[③] 信任调节通过人们的警惕心理和行动标准发挥作用。在高风险条件下，部分人可能会减少对复杂技术的依赖，增加对简单技术的依赖[④]。但在低风险条件下，人们对复杂技术的依赖会更强烈。例如，在城市中寻找某个陌生地点时，大部分人都依靠手机导航的指引而放弃通过路牌、路标等实体指引寻找，因为其对智能导航技术更有信心，找错路的风险属于低风险，所以在低风险场景中，人们给予人工智能技术高度的信任。但当场景转换到医疗、政策决策等关系重大的

① [瑞士]海尔格·诺沃特尼：《未来的错觉：人类如何与AI共处》，姚怡平译，香港：香港中文大学出版社，2023年版，第4页。

② R.R.Hoffman.A taxonomy of emergent trusting in the human–machine relationship，Cognitive Systems Engineering，2017：137-164.

③ [英]安东尼·吉登斯：《现代性的后果》，田禾译，南京：译林出版社，2011年版，第18页。

④ K.A.Hoff，M.Bashir.Trust in automation：integrating empirical evidence on factors that influence trust，Human Factors，2015（3）：407-434.

场景时，信任的动态调整开始发挥作用。

根据不同时间、情境和关系的变化而发生变化的信任模式是动态的信任模式。动态信任要确保安全。在自动驾驶领域，自动驾驶汽车的错误将对驾驶员信任和信任相关感知产生较大影响，“用户的信任是一个动态的过程，特别是在面临自动化错误时，用户对自动驾驶汽车的信任会迅速减弱，并严重影响其技术采纳倾向”①。随着 Sora 的推广及开放应用，用户在实际应用过程中将不断调整对其的信任程度，信任会促使用户采用该技术，不信任可能导致用户放弃该技术。需要注意的是，调整过程中需弥合外部风险因素与人的主观感知的差距，并非所有人都能精准、无差别地感知人工智能的技术风险，因此，提升用户对人工智能技术风险的认知能力，确保用户能准确评估和判断技术的潜在风险，对建立合理的信任阈值至关重要。准确的风险预测和评估是确保用户建立正确信任基础的关键因素。如果信任不正确，技术脆弱所带来的显性风险和隐形风险将给人类带来更多危害。

（三）代理模式的人工智能信任

在计算机科学和人工智能领域，代理指智能体（Agent）对环境进行感知和行动的能力，这种智能体可以是软件程序、机器人，也可以是虚拟实体，等等。随着人工智能技术的进步，代理信任的可行性有所增加。当前，人工智能体（AI Agent）应用于机器人、人机交互游戏、虚拟助理以及自动驾驶汽车等各领域，是面向未来的先进技术，能够正确理解和响应人类的输出，作出和人类相同的判断和决策行为。人工智能技术的发展带来信任代理的可能性。

1995 年，美国的乔伊斯·伯格（J.Berg）等三位实验经济学家设计并进行了一项“信任博弈”实验。“在这个博弈中有两个匿名的玩家：一个是信任者，另一个是受托人。信任者拥有一定数额的货币 T，需要决定是否将其中一部分 r 发送给受托人，作为对其信任的表示。发送的金额 rT 会乘以一个因子 K（K>0）然后由受托人接收。最后，受托人需要决定他们愿意将其收到的 KrT 中的哪一部分 α 返回给信任者。”② “信任博弈”用于探讨行为学和神经科学中关于信任的问题。有技术专家在此基础上利用“信念—欲望—意图”（BDI）的框架建模测试，论证 LLM 模型代理模拟人类信任行为的可行性，得出 LLM 代理信任与人类信任具有高度一致性的结论③。

人工智能代理信任行为与人类信任行为是否具有一致性至关重要。在人机对齐过程中，不仅需要行为对齐，而且需要人工智能价值对齐（AI Alignment），以人为尺度的价值对齐是人类通往未来的必经之路，也是确保人工智能发展安全的重要问题。人们应围绕自我的生活对人工智能作出信任或不信任的决定。人工智能价值对齐的缺乏将给人类委托的代理信任协作带来危机，对人工智能信任行为的持续监管、评估和治理成为人工智能代理信任持续的保障。代理的人工智能信任是协作的信任，可以预测人与人工智能能否产生良性互动结果。

“与失败后可恢复的人际信任不同，当机器犯错误时，人们会对其可预测性和可靠性失去信心。”④ 人工智能信任在脆弱性中逐渐成长，呈现崭新的信任样态，在对人工智能信任未来的探索过程中，以人为尺度是人工智能信任建设的基准。

三、信任脆弱性的四象限与人工智能技术的未来关系

人工智能形成的脆弱性是多元的，包含信任、人工智能技术以及两者的关系本身，等等。信任与

① H.Tan，Y.Hao.How does people’s trust in automated vehicles change after automation errors occur? An empirical study on dynamic trust in automated driving，Human Factors and Ergonomics in Manufacturing & Service Industries，2023（6）：449-463.

② J.Berg，J.Dickhaut，K.McCabe.Trust，reciprocity，and social history，Games and Economic Behavior，1995（1）：122-142.

③ C.Xie，C.Chen，F.Jia，et al.Can large language model agents simulate human trust behaviors? arXiv preprint arXiv：2402.04559，2024.

④ H.P.Beck，M.T.Dzindolet，L.G.Pierce.Operators’ automation usage decisions and the sources of misuse and disuse，advances in human performance and cognitive engineering research，Emerald Group Publishing Limited，2002：37-78.

技术属于不同维度的概念，但彼此交织影响，信任是技术发展和应用的重要基石，技术通过自身能力的提升提高信任促进社会整体信任的发展。从信任脆弱性出发进行人与人工智能未来技术关系的探究，是为迎接未来世界做准备的途径。信任与人工智能技术的未来关系可以借四象限法则表示。如图1所示：第一象限代表人工智能技术与信任的双重脆弱性；第二象限代表积极的对抗脆弱性；第三象限代表无关脆弱性；第四象限代表消极的对抗脆弱性，每个区域对应表示一种脆弱性信任与人工智能技术未来的可能关系形态。

（一）第一象限：双重脆弱性与人工智能技术的未来

第一象限代表信任脆弱性与人工智能技术未来的第一种可能关系，即人工智能信任与人工智能技术的双重脆弱性关系，这是最危险的状态，意味着无论信任层面还是技术层面，人类面临的风险值都较高，亟须降低风险以避免危险的发生。

通过降低信任的方式调节技术的风险比较有效，但信任风险的降低需要通过更高层次人类理智的调节。吉登斯在分析信任和其他相关概念时认为，“对于一个行动持续可见而且思维过程具有透明度的人，或者对于一个完全知晓怎样运行的系统，不存在对他或它是否信任的问题”，“寻求信任的首要条件不是缺乏权力而是缺乏完整的信息”[①]。信任脆弱性在某种程度上呈现和技术脆弱性相同的特征：不透明、持续变化和缺乏完整信息。

人工智能算法的“算法黑箱”在一定程度上导致人工智能信任的“信任黑箱”，对具有“算法黑箱”缺陷的技术产品的持续依赖是个体信任的盲目以及集体信任的无意识。两者共有的缺点潜藏巨大的风险，容易成为商业竞争或其他竞争的利用对象。技术脆弱性难以从根本上消除，信任脆弱性有赖于通过人类深层次理性进行调整。就认识的本质而言，在生成式人工智能出现之前，人工智能技术的工作大部分为抽象的总结，正如埃文·阿姆斯特朗（Evan Armstrong）所言，“人工智能是低级思维之上的抽象层”[②]。这种较低层次的思考很大程度是一种总结。因此，要解决信任脆弱性与技术脆弱性的双重困境，仅依靠技术的方式较为有限。人们应充分调动人类更高层次的智慧，例如，理智、分析以及想象，即人类独特的创造性活动，以应对人工智能技术与信任的双重脆弱性，满足人类发展的更多可能。

（二）第二象限：积极的对抗脆弱性与人工智能技术的未来

第二象限代表信任脆弱性与人工智能技术未来的第二种可能关系，即以信任调节为主导的积

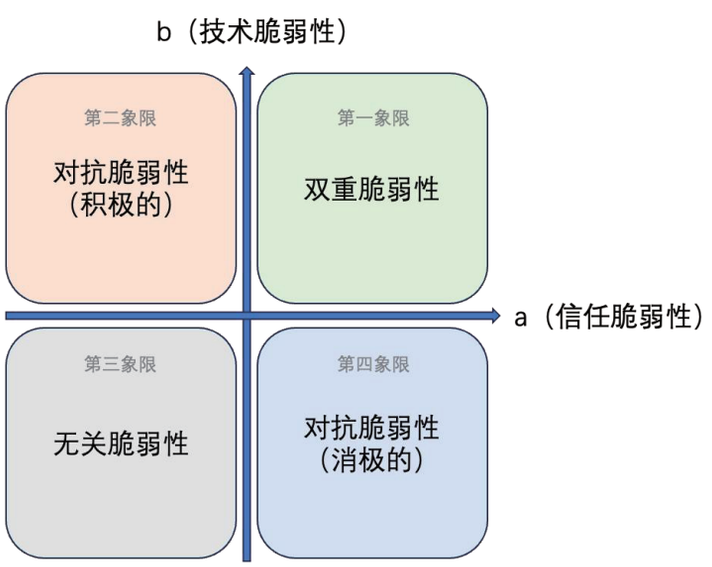

图1　信任脆弱性与技术脆弱性的四象限图式

① [英]安东尼·吉登斯：《现代性的后果》，田禾译，南京：译林出版社，2011年版，第29页。

② Dan Shipper.The Knowledge Economy Is Over.Welcome to the Allocation Economy，https://every.to/chain-of-thought/the-knowledge-economy-is-over-welcome-to-the-allocation-economy.

极的对抗脆弱性状态。通常在同一时间内，技术脆弱性的增加与信任脆弱性的减少之间会形成负相关关系，导致两者间存在一种张力。总体而言，信任脆弱性的减少被认为是更关键和本质的解决方式。

对抗性（Adversarial）概念在不同领域有不同的含义和应用，在机器学习和人工智能领域，对抗性训练是一种提高模型鲁棒性的方法。例如，OpenAI 官方给出的红队测试方案，“我们正在与红队成员（错误信息、仇恨内容和偏见等领域的专家）合作，他们将以对抗性方式测试该模型”[①]。这一测试指由网络安全专家组成的团队对该系统进行一系列安全评估和渗透测试。这些专家被称为红队成员，职责是模拟潜在的恶意攻击者，寻找 Sora 系统中的安全漏洞或风险点。在红队成员合作的过程中，建立足够的信任至关重要。此时，信任问题不仅涉及团队合作，而且嵌入 Sora 安全构成更加细致的方面。

因此，仅依赖对技术脆弱性的调控难以从根本上解决问题，信任始终是解决问题的核心。尽管人工智能技术在其解决方案中扮演辅助性而非决定性角色，但并不意味着人工智能技术应处于被动或无为状态。相反，在应对由人工智能技术带来的社会复杂挑战时，唯有通过多元主体共同采取积极行动，才有可能应对这一挑战。

（三）第三象限：无关脆弱性与人工智能技术的未来

第三象限代表信任脆弱性与人工智能技术未来的第三种可能关系，即无关脆弱性。这种情况说明在人工智能发展过程中，信任脆弱性与技术风险之间达到了平衡状态。当这种平衡状态得以显现，人工智能信任将发挥最大作用，为人工智能的安全发展保驾护航。

如前文所述，信任脆弱性必须借助更高层次的人类智慧获得解决办法。在具体实施层面，关键在于设计适度的信任以及合适的信任平衡机制对信任脆弱性进行有效管理与调节，进而影响技术脆弱性，呈现这种平衡状态意味着整体信任环境趋于健康和稳定，有利于二者关系的进一步发展。整体的信任环境潜在地影响数字化社会中个体的行为和心理状态。良好的信任环境能促进个体对人工智能技术的信任，良好的接受度对技术的发展具有促进作用；但如果信任环境脆弱，个体信任缺失，将影响技术的采纳和应用。从某种意义而言，技术与信任展现的脆弱性并非毫无价值，而是值得人类关注和维护。技术脆弱性揭示了技术本身固有的弱点，而信任脆弱性映射了人类本质的生物学属性，反映了人性的固有弱点。

因此，应允许一定限度内的容错率。无论技术系统还是信任系统，都需要一定的容错率。英国哲学家卡尔·波普尔（Karl Popper）提出“可证伪性”（Falsifiability）原则，认为科学进步的动力在于不断尝试和纠正错误，而不是寻求最终的确定性。容错率的存在不仅是为降低风险，而且是为激发系统在面对错误时的创新潜能，这种对可能性的拓展引发新的变革。

总之，在人工智能技术的发展过程中，信任脆弱性与技术风险之间的平衡，不仅对技术的安全发展至关重要，而且对健康稳定的整体信任环境的构建、对未来世界的人机关系发展都具有重要意义。除此之外，在构建信任脆弱性与人工智能技术未来关系的过程中，必须坚持以人为本的原则，确保技术发展与人类价值和谐共存。

（四）第四象限：消极的对抗脆弱性与人工智能技术的未来

第四象限代表信任脆弱性与人工智能技术未来的第四种可能关系，即以技术调节为主导的消极对抗脆弱性状态。其忽略了对抗性行为背后的复杂性和异质性。

对抗脆弱性通常涉及人的信任行为和决策，包括攻击者和防御者。如果仅从技术出发，可能会忽视人的动机、心理和社会背景，这些因素对于理解和预防对抗脆弱性至关重要。根据法国社会学家布鲁诺·拉图尔（Bruno Latour）等提出的“行动者网络理论”（Actor-Network Theory，ANT），在 ANT 中，行动者不仅包括人类个体，而且包括非人类实体，例如，技术、物体、动物，等等。“行动中的行动者并非单个的、分离的，而是依附于特定网络联系而存在的某种实体，其中行动者既可以是人，也可以是物，他们平等地在集合的连锁效应中发挥各自的能动性。对于拉图尔而言，网络从来不是可以简单界定或假设的概念，它拥有一系列不同的拓扑形态，期间的一些拥有十分鲜明的层级结构，期间的所有行动者都必须行动起来，而非

① OpenAI.Creating video from text，https://openai.com/sora#safety.

仅仅待在那里。”[①] 从本质看，信任与技术是异质性的事物，这种异质性在一定程度上限制了沟通与协作。但在 ANT 视角下，异质性行动者之间的关系构成了网络，这些关系不是静态的，而是通过转译（Translation）过程动态形成。行动者之间通过交流和互动，将各自的意图、目标和行为转化为网络中的共同行动。

人类对技术的依赖是养成性的，呈现“越用越依赖”的状态，逐渐地这一信任将过度，从而增加信任的脆弱性。“在技术信任中，我们相信技术以及设计和操作技术的人。这种信任一旦过度，技术的权力将大大增加，因为过度的信任意味着委托者（信任者）将要求更少的证据和付出更少的监督。”[②] 因此，为防止对技术的过分依赖加剧信任脆弱性，在构建技术与信任的未来关系时，应持续保持审慎与警觉。在塑造技术与人类关系的过程中，必须平衡技术效能的发挥，避免对其产生过度依赖心理。

四、以适度信任的构建破解人工智能技术与信任的双重脆弱性

在人工智能发展过程中，技术的风险无可避免，但信任可以影响甚至控制技术的发展道路，关键在于人们如何使用技术。“技术产生效用的前提条件是其被使用，若不被使用，效用就无法得以生成。”[③] 应如何进行对 Sora 等生成式人工智能技术的正确信任？古希腊哲学家亚里士多德在《尼各马可伦理学》中提出，“我们应当选择适度，避免过度与不及，而适度是由正确的逻各斯来确定的”[④]。适度是美德的体现，需要以正确的逻辑为基础。笔者认为，对 Sora 的正确信任至少需符合以下四点。

（一）警惕风险：技术信任应以物理规律为基础

基于人工智能技术风险的不可逆性，人类应对其设置信任底线，即所信任的人工智能产品必须符合物理世界的因果律。《人机对齐：如何让人工智能学习人类价值观》的作者布莱恩·克里斯汀（Brian Christian）认为，“我们发现自己正处于一个脆弱的历史时期。这些模型的力量和灵活性使它们不可避免地会被应用于大量商业和公共领域，然而关于应该如何适当使用它们，标准和规范仍处于萌芽状态。正是在这个时期，我们尤其应当谨慎和保守，因为这些模型一旦被部署到现实世界中，就不太可能再有实质性改变”[⑤]。Sora 等生成式人工智能的应用与其带来的伤害都不可逆，人们在开发应用过程中要时刻保持警惕。

在对 Sora 的深入分析中，笔者发现，尽管该技术生成的每一帧画面在细节上都可能精确无误，但当这些画面组合形成连续叙述时，却导致整体上的失真。这种现象揭示出 Sora 在处理和表现时空关系方面的不足，凸显了模型在理解和模拟复杂现实场景时的局限性[⑥]。这种局限性可能导致误导性结果的输出，尤其在需要准确反映现实世界或与教育相关的情境应用中。如果未足够关注这种技术局限性，没有适合公众信任和监管政策配套出现，那么，随着技术的广泛应用，将带来一系列不可预见的风险。例如，生成的内容可能被用于误导公众意见、传播虚假信息或侵犯个人隐私[⑦]，等等。

基于上述情况，人工智能的快速发展对法律规制和政策规制提出“预见性”要求。2021 年，欧盟的《人工智能法案》（EU AI Act）首次提出，直到 2024 年 3 月 13 日由欧洲议会投票通过，这是全球人工智能领域监管进入新的时代的标志性事件，但也反映出对 AI 监管和治理的滞后性，从提出到通过，各方面的协商和协调耗时近 4 年。在这 4 年间，生成式人工智能实现从 DALL·E 到 ChatGPT，再到 Sora 的突破。如今，全球性的 AI 安全已备受关注。2023 年 11 月 1 日，首届全球人

① [英]尼古拉斯·盖恩、戴维·比尔：《新媒介：关键概念》，刘君、周竞男译，上海：复旦大学出版社，2015 年版，第 30 页。

② 闫宏秀、宋胜男：《智能化背景下的算法信任》，《长沙理工大学学报（社会科学版）》，2020 年第 6 期。

③ 闫宏秀：《负责任人工智能的信任模塑：从理念到实践》，《云南社会科学》，2023 年第 9 期。

④ [古希腊]亚里士多德：《尼各马可伦理学》，廖申白译，北京：商务印书馆，2003 年版，第 180 页。

⑤ [美] 布莱恩·克里斯汀：《人机对齐：如何让人工智能学习人类价值观》，唐璐译，长沙：湖南科学技术出版社，2023 年版，第 27-28 页。

⑥ OpenAI.Creating video from text，https://openai.com/sora#capabilities.

⑦ 邓建鹏、赵治松：《文生视频类人工智能的风险与三维规制：以Sora为视角》，《新疆师范大学学报（哲学社会科学版）》，https://doi.org/10.14100/j.cnki.65-1039/g4.20240322.001。

工智能（AI）安全峰会正式发布《布莱切利宣言》，意味着人工智能对人类构成潜在的灾难性风险已成为全球共识[①]。全球性的协商与关注将进一步推动问题的解决。

相较数智时代人工智能大规模的创造，对其管理的需求更为迫切。在人工智能技术监管政策出现前，人类的信任应发挥过渡和缓冲作用。这种信任的建立应根植对技术行为与物理世界因果关系一致性的理性评估之中。只有当人工智能产品展现的效能与物理世界的因果律相契合，才能够被合理地赋予人类的信任。否则，证明该技术产品的技术能力和安全保障未达到人类信任水平，大规模应用也无法得到允许和信任。

（二）正确认知：技术信任应以人类信任为最后尺度

信任在本质上是一种认知现象。相较技术缺陷可能导致的负面后果，对信任本质理解的缺失可能带来更深远的影响。对个人而言，要深刻理解 Sora 等人工智能技术较为困难。鉴于技术领域的复杂性和不断进步的特点，无法要求每个人都具备专业的技术知识和解决问题的能力。多数人可能缺乏必要的背景知识或专业训练，难以跟上科技发展的步伐，遑论对新兴技术进行深入分析并提出问题的解决方案。特别是在人工智能领域，例如，以 Sora 为代表的先进人工智能技术，其技术门槛成为难以逾越的壁垒。认知层面的改变将指导行为层面的改变。

善用信任的关键是正确地认知信任。信任作为一种社会资本，应被合理运用。德国社会学家尼古拉斯·卢曼（Niklas Luhmann）认为，“信任作为一种社会资本积累起来，它为更大范围的行为开放了更多的机会”[②]，其对促进合作、增强社会凝聚力和推动经济发展具有不可替代的作用。但适度信任的建立和维护并非易事，它需要个体、组织乃至整个社会对信任的本质、功能和局限性有深刻理解。尽管如此，人们依然应尝试从现存事物和信息中找出规律。在参与技术研发与应用的众多群体中，技术专家尤其需要对信任有深刻的理解和认知。对部分技术专家而言，人工智能技术的开发目标、安全性的伦理边界以及对人类信任的正确理解，通常被视为与其专业领域相距甚远的问题。但真正令人担忧的并非数字化社会本身，而是在数字化及未来社会中处于领导地位的专家。与工业等不同，人工智能产业与人类活动紧密相连、密切互动，并且其规模在迅速扩展。因规模产生影响力，以及人工智能危害的不可逆，技术专家的信任认知更应受到重视。

人类的信任应作为衡量 Sora 等人工智能技术发展应用的最终尺度。当 Sora 的技术发展到能够彻底模拟现实世界的程度时，其带来的安全挑战和伦理考量将显著增多。例如，高度逼真的模拟环境可能会模糊虚拟与现实的界限，导致人类尤其是对人类现实世界尚未建立完整认知的低龄群体产生对世界概念的混淆，甚至可能用于误导公众、制造虚假信息或侵犯个人隐私。此外，商业与竞争的驱动可能将使模拟现实世界的技术用于不当目的，例如，在没有适当监管的情况下进行社会工程或心理操纵。因此，随着 Sora 的技术进步，必须建立技术专家群体对信任的正确认知，同步加强对其潜在影响的评估和监管，确保技术的发展与社会价值观和伦理标准保持一致。

（三）信任启蒙：技术信任应与人类价值观保持对齐

信任启蒙至关重要，其关键作用在于帮助个人正确地理解某项技术的可依赖程度，帮助人类克服对技术的迷信与执着。18 世纪欧洲的启蒙运动推动人类对理性的崇拜，帮助人类克服长期以来的迷信、愚昧，主张个人自由和权利。如今人工智能技术经过长足发展，在人类世界走到前所未有的位置，这一现象令人担忧。正如埃隆·马斯克所言：“这是人类历史上第一次与远比我们聪明的东西共处，所以我不清楚，我们是否真的能控制这样的东西。但我认为我们可以期待的是，引导它朝着对人类有益的方向发展。我确实认为，这是我们面临的生存风险之一，而且可能是最紧迫的风险。”[③]

科研人员和技术专家通常不会深入思考人工智能开发的终极目标，他们采取逐步探索的态度，认为科学研究应自由、不受限制。实际上，人们应该意识到技术是一把双刃剑，在认识其对人类造成可能的风险之前，不存在必须得到无限探索和发展的技术。

Sora 的技术门槛更低、生成的内容更鲜活，受到人们的追捧与青睐。当 Sora 最终实现其模拟世界

①③《首个全球性AI声明：中国等28国、欧盟签署〈布莱切利宣言〉》，https://hqtime.huanqiu.com/article/4FC8suObROX。

② [德]尼古拉斯·卢曼：《信任》，翟铁鹏、李强译，上海：上海人民出版社，2005年版，第85页。

的意图后，人类世界将会怎样？无疑，人与人工智能间的张力将达到最终阈值并产生对抗，“人和机器的对抗不是精神的对抗，而是实力的对抗。取胜不是在精神上或精神高度的胜利，而是在物质上或控制住低端的胜利。那时，机器语言将战胜人类的自然语言”①，这是人类不愿看到的。因此，对人工智能信任的启蒙至关重要，首先要对人工智能技术专家进行信任启蒙，使其在进行伟大创造的同时具有目的意识和责任意识，重新审视对技术的信任程度。之后致力于对广大使用者技术素养的培养和提升。

在人工智能技术领域，好奇心确保了人们对人工智能技术探索的开放态度。但好奇心需要人类理性的引导，并由此引发向善的技术发现与科技进步，不能让这份好奇心成为打开“潘多拉魔盒”的双手，给人类带来不可挽回的伤害。因此，有必要从一开始就将信息启蒙植入技术研发人员的研发期以及技术产品的成长期，只有具备正确的信任认知，人类才有可能让人工智能的最终发展结果呈现向善的状态。

（四）信任修复：技术信任应以充分证据为重建依据

人与人工智能间的信任被破坏后，人类该何去何从？当前提出的人与人工智能技术信任的修复更多是一种技术性行为。博·施贝尔（B.G.Schelble）等人在信任研究中发现，“拥有不道德自主队友的团队在团队信任和自主队友信任方面其信任度显著降低。不道德的自主队友也被感知为更加不道德。两种信任修复策略在道德违规后都未能有效恢复信任，自主队友的道德性与团队得分无关，但不道德的自主队友完成任务的时间更短”②。更复杂的人机交互是人工智能应用发展的事实。

人类期望人工智能可以像可靠的队友一样与人类并肩协作、解决难题。然而，人与自主队友间的信任时常遭到破坏。面对这种极具可能性的破坏，信任的修复方法需要预先性思考。“自主队友的道德性对信任有显著影响：与不道德的自主队友相比，人们更信任表现出道德行为的自主队友。自主队友的道德行为反映出其道德价值选择。”③ 当前，科研人员和技术专家都在尝试人工智能技术的行为和价值观与人类行为和价值观的对齐和校准。这说明人类与自主队友的合作本质上已充分将自主队友视为独立的角色，在与其合作的过程中前置性地嵌入了高度信任。与人工智能技术的普遍运用类似，这种信任是前置性的，意味着没有信任的付出就没有合作的倾向。

没有充分的证据表明自主队友在参与信任修复过程中能明晰自己在道德上的错误。因此，人们将人工智能代理视为有道德的行动者，对其信任的付出需要以更多的逻辑证据为依据。对人工智能技术的信任需要基于逻辑和道德的证据，而非盲目的前置信任，避免信任断裂和造成难以修复的后果。因此，信任修复只有在基于逻辑和道德证据的信任付出之上才有可能。由于目前监管和治理仍处于缺失状态，人们应预设更多人工智能信任修复场景，以积累对人与人工智能未来关系的修复条件。

五、适度信任：给技术以空间，给脆弱以安全

Sora 作为人工智能技术的代表，引发全球关注和热议。与 ChatGPT 和 DALL · E 等技术相比，Sora 展现出更强大的内容生成能力，特别是其在动态视频内容方面的创造性，预示着人工智能在模拟现实世界方面具有不可想象的潜力。与此同时，这种潜力也暗含风险，且这种风险不容忽视。

但是，从人类发展史的视角看，技术是人类生存的必备品。基于此，忽视风险与因风险盲目抑制技术发展同样不可取，只有积极应对技术风险才是确保人类健康发展的必要条件。目前，Sora 进行的红队测试作为一种模拟的对抗性测试旨在评估和提高系统的安全性和稳定性，以降低技术风险，消除公众对技术的恐惧、不信任等，促进 Sora 的安全发展。因此，随着 Sora 从初步测试逐步过渡到实际应用，人们有必要对这项技术进行更深入地理解和评估，进而形成适度信任，更好地实现技术利用和技术效益的发挥。

Sora 是人类对技术创新的结果。一方面，保护这一创新是人类世界进步的必要过程；另一方面，需对这一创新的无限性保持警惕，特别是对其

① 何怀宏：《GPT的现实挑战与未来风险 —— 从人类的观点看》，《探索与争鸣》，2023 年第 6 期。

②③ B.G.Schelble，J.Lopez，C.Textor，et al.Towards ethical AI：empirically investigating dimensions of AI ethics，trustrepair，and performance in human-AI teaming，Human Factors，2022.

未定的发展方向可能带来的潜在风险设置合适的预案是人类必须持有的态度。审视技术发生的环境与目的，避免脱离技术发展的安全边界，把人类的安全与福祉作为技术的最优先级考虑是人类研发技术的底线。易言之，技术的快速发展不应脱离伦理和安全的考量，确保 Sora 等技术的应用不会超越人类社会的安全边界，保证技术发展方向与人类安全福祉一致，这是 Sora 等人工智能技术持续进步的前提。

综上所述，以适度信任对抗技术脆弱性和信任脆弱性是确保人工智能健康发展的有效途径。对于 Sora 而言，承载更高的预期也暗含存在更高的风险。因此，给技术以空间与给脆弱以安全的双重融合形成的适度信任既可以帮助 Sora 在“脆弱”的环境中成长，又可以帮助人类规避 Sora 发展带来的风险。

Dual Vulnerability and Moderate Trust：From ChatGPT to Sora

YAN Hong-xiu　SONG Sheng-nan

（School of History and Culture of Science，Shanghai Jiao Tong University，Shanghai 200240）

Abstract: As an innovative technology，Sora is capable of creating and managing intricate dynamic video content. It marks a significant shift in the picture of AI's understanding of the world from understanding the world at rest to understanding the world in motion. Sora's limited understanding of the laws of motion in the physical world and its obfuscation of the details are its technological vulnerabilities，which further exacerbate the trust vulnerability. Due to the risk of technological vulnerability，a new model of trust that is antecedent，dynamic and agentic emerges. The dual vulnerability composed of technological and trust fragility，positive confrontational vulnerability，irrelevant vulnerability，and negative confrontational vulnerability，four quadrants，and also four possible future states of AI technology. Analyzing the four quadrants of the relationship between trust vulnerability and the future development of AI technology reveals that the construction of moderate trust is an effective way to break the double vulnerability of trust and technology，and the construction of moderate trust itself needs to be based on the causal law of the physical world，human trust as the final measure，the premise of maintaining the enlightenment of alignment with human values，and sufficient evidence as the basis for trust reconstruction.

Key words: Sora；ChatGPT；Dual Vulnerability；Moderate Trust；Generative AI

[责任编辑：曹晶晶]
[责任校对：王文秋]