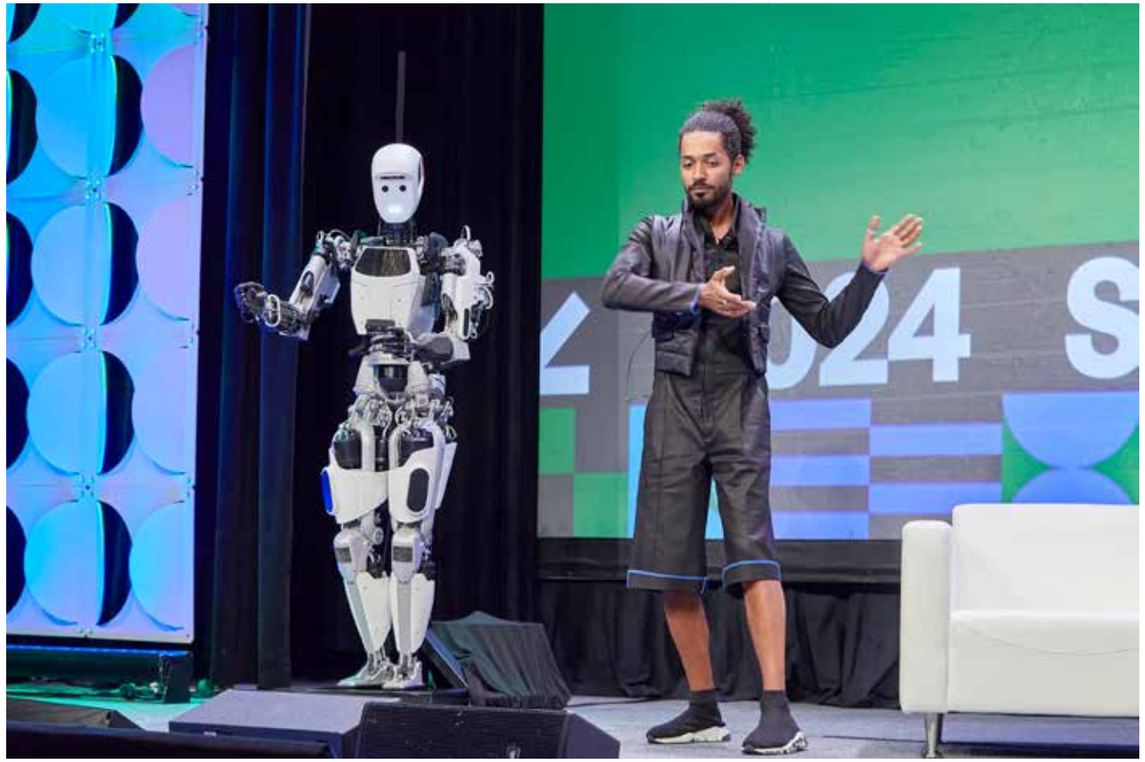
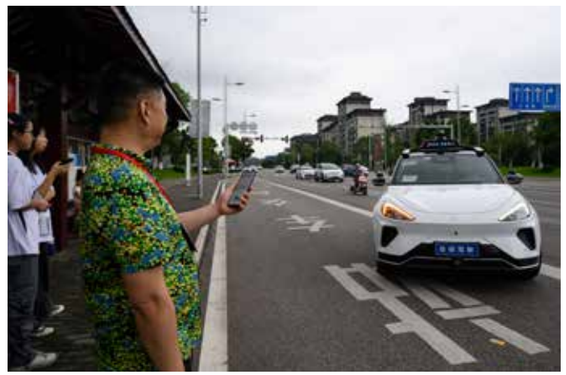


1

2



3

1. 图说假字图说假字图说假
2. 图说假字图说假字图说假字图说假字
3. 图说假字图说假字图说假字图说假字

具身智能进入物理世界，我们准备好了吗？

记者 · 陈璐

自动驾驶汽车的经验能为更广泛的具身智能的监管带来什么启示？在将这些具身智能系统部署到可能危及人类生命的物理空间之前，我们又该如何确保它们的安全？

7 月 4 日，在 2024 世界人工智能大会上，超过 20 款智能机器人齐齐亮相，展示了这一技术在制造、服务、康养等各个领域的应用潜力。机器人初创公司 Agility Robotics 已经将其 Digit 机器人部署在亚马逊的研发中心，而 Apptronik 的人形机器人 Apollo 则被派往梅赛德斯 - 奔驰工厂。OpenAI 支持的 Figure 公司也正在宝马制造厂部署其具身智能机器人。在最近的演示视频中，这款机器人展示了通过 OpenAI 训练的视觉语言模型（VLM），不仅能理解语音指令，执行任务，甚至还能解释其行动背后的原因。

具身智能正在不可阻挡地成为人工智能的下一个风口。所谓具身智能，即能够感知、决策并与物理世界互动的人工智能系统。在接受本刊采访时，北京大学人工智能研究院 AI 安全与治理中心的执行主任杨耀东教授解释说：“图灵在定义智能时，提到了两种未来的智能形态。一种是存在于虚拟数字空间中的智能体，比如 AlphaGo，它们非常聪明，但只在数字空间中运作。另一种则是具身智能，不仅存在于数字空间，还拥有类似人类的感官能力，能够与物理世界互动。比如无人驾驶汽车能够‘看’‘听’‘说’，并执行复杂的驾驶操作。”

无人驾驶出租车（Robotaxi）是具身智能（Embodied Intelligence）的一个重要应用实例。国内的“萝卜快跑”“小马智行”“文远知行”，以及国外的特斯拉、Waymo 和 Cruise 等，都是这个领域的重要玩家。2021 年，特斯拉推出了 FSD（完全自动驾驶系统），并迅速布局 Robotaxi。马斯克将 Robotaxi 视为特斯拉经济型汽车的终极版本，即便车主不使用它们，车子也能自己去“赚钱”。

然而，事实证明，人工智能的引入并未消除人为失误在交通事故中的影响。错误的发生，只是从事件链的末端转移到了起点——人工智能的编码本身。这些错误往往隐藏在“黑匣子”中，因此更难察觉。

中国科学技术法学会人工智能法专委会副主任肖艺能博士指出，北京在 2019 年确立了亦庄为自动驾驶示范区，但算法的不可解释性和道德困境的不确定性可能会引发公众的“寒蝉效应”，即心理排斥，导致相关监管细则迟迟未能出台。

比如，当无人驾驶车辆遇到突发情况时，一位老人突然倒在路上，系统如何在瞬间做出决策？对于这类“电车难题”的伦理挑战，现实中往往没有所谓的正确答案，无人驾驶技术目前也没有很好的解决方案。虽然有意大利学者提出过“伦理旋钮”的设想，允许用户选择应对策略，自行选择承担的风险类型，但这仍停留在理论层面。北京最终选择通过强制亮灯的方式标识自动驾驶状态，以缓解公众的不安。

肖艺能长期关注新兴技术带来的社会和经济风险，尤其是在伦理、法律和社会治理三个核心层面。随着无人驾驶的发展，事故责任认定和潜在的就业问题也引发了广泛讨论。7 月 1 日北京最新出台的自动驾驶监管细则规定，如果事故发生在自动驾驶状态下，首要赔偿责任由持牌运营者承担；若调查发现事故源于硬件或软件缺陷，运营者则可追责制造商或软件提供商。但如果车辆处于被人接管的状态，则按传统交通事故处理。“所以其中关键是，要确保自动驾驶状态的数据记录必须明确、可追溯且不可篡改，并受到相应的监管。”肖艺能告诉本刊记者。他还指出，在社会治理层面更需要高度谨慎。我国目前对全面推行无人驾驶持保留态度，倾向于在低风险的园区内推行，以避免因新技术的应用引发更大的社会和经济问题。

相比已经经过多年的技术发展和论证的无人驾驶，正在快速推进的人形智能机器人则带来了更复杂的安全问题。要确保具身智能的安全性，就必须在开发、测试和监管等各个阶段制定更严格的规范和标准。那么，自动驾驶汽车的经验能

为更广泛的具身智能的监管带来什么启示？在将这些具身智能系统部署到可能危及人类生命的物理空间之前，我们又该如何确保它们的安全？为此，我们采访了北京大学人工智能研究院 AI 安全与治理中心的执行主任杨耀东教授。

三联生活周刊：为什么自动驾驶会是公众最先接触到的首批主要具身智能应用？这是否与它能方便地采集大量数据有关？

杨耀东：确实如此。自动驾驶技术之所以能快速推进，首先是因为它的操作相对简单，控制维度较低，比如方向盘、油门和刹车，加起来也只有三个自由度（统计学概念，指样本中独立或能自由变化的观测值个数）。而人形机器人复杂得多，一个灵巧的机械手就可能有 20 多个自由度，更别提腿、腰、头部、手臂等部位的关节了。这使得人形机器人面临的挑战远大于自动驾驶。

其次还因为自动驾驶的产业链成熟，技术迭代快，数据收集也更方便。特斯拉就是个典型例子。特斯拉一直标榜自己是家软件公司，通过不断销售车辆，收集大量行驶数据，训练出更强大的智能系统。随着这种模式的持续运作，车辆的制造成本可能会逐渐降低，但驾驶系统的订阅费用会越来越高。

类似的逻辑也适用于人形机器人。未来 10 年到 20 年内，人形机器人可能会进入千家万户，成为个人助理。比如在今年的苹果全球开发者大会（WWDC 2024）上，Apple 宣布推出 Apple Intelligence 平台，将引入 OpenAI 大模型，在最新的 iPhone 16 里集成智能助手。随着技术的进步，这种助手最终可能会具象化为人形机器人。就像自动驾驶一样，在未来，人形机器人的制造成本会下降，而与人交互的数据收集将进一步提升机器人的智能化，提供更好的用户体验。

人形机器人的发展虽然面临着比自动驾驶更复杂的问题，但它们进入生活的速度却很快。与汽车相比，人形机器人的价格更低，因为它们不需要配备昂贵的雷达和摄像头设备。马斯克的 Optimus 人形机器人定价在 1 万美元到 2 万美元之间，已经相当亲民。但从技术安全的角度来看，我们可能还没有完全做好准备迎接这些机器人的到来。

三联生活周刊：现在提到的具身智能与我们以前见到的机器人有什么不同？过去实际已经有操作能力非常精细的人形机器人出现。

杨耀东：具身智能最近在人工智能领域备受关注，但从机器人学的角度来看，这并不是一个新概念。正如你提到的，各种形态的机器人早已存在，并且被研究了很多年。

具身智能之所以引起新的兴趣，主要是因为人工智能领域的研究者开始从图灵的视角来看待智能的分类——具身智能与非具身智能。他们强调，人工智能不仅限于虚拟空间，还应进入物理世界，与环境互动。具身智能的关键在于它集成了大量人工智能算法，这是传统机器人学所缺乏的。传统机器人依赖于控制论中的最优控制算法，尽管这与人工智能有交叉，但两者还存着本质上的不同。具身智能通过大语言模型整合了强大的推理能力和人类常识，这是控制论无法处理的。

具身智能可以分为“大脑”和“小脑”两部分。大脑部分类似于大模型，如 ChatGPT，能够实时判断并理解复杂任务，比如知道你饿了，就去冰箱拿三明治。这种常识性判断是传统控制论无法实现的。小脑则负责具体的操作和控制，如手脚的动作、拿起物品、开门等。

具身智能强调大脑和小脑的协同工作，使得机器人不仅能聪明地规划任务，还能精准地执行每个细节操作。以 MIT 研发的外科手术机器人“达芬奇”为例，它虽然具备精密的操作能力，但缺乏拥有诊疗自主决策能力的“大脑”，因此只能被称为“具身机器人”，而非真正的“具身智能”。虽然外观上看，具身智能仍然是个机器人，但其背后集成了大量人工智能技术，这让它与传统机器人有了显著的区别。

三联生活周刊：能解释一下这些区别在具体应用场景中的体现吗？比如，特斯拉的人形机器人 Optimus，我看了视频，感觉和我们过去在工厂里使用的机器人差别不大。

杨耀东：具身智能的应用主要分为工业和家用两种场景。在工业场景中，汽车生产线是最典型的例子。汽车制造涉及复杂的流程和零部件，机器人在喷漆、组装等环节中发挥着重要作用。而在家用场景中，应用则更加广泛，比如在家庭、餐厅或养老院中使用的服务机器人。

举个例子，马斯克的机器人在工厂里用于装电池。这看似简单，但要实现类人的灵巧操作，仍然非常具有挑战性。机器人需要通过灵巧的手指来拿取和放置电池，而这对当前的人工智能来说依然是个难题。要完成这样的操作，背后需要大量的人工智能算法，尤其是强化学习的应用。比如，当机器人在放电池时，如果电池卡在了边缘，传统控制论系统可能会让任务失败，需要人为干预。然而，具身智能可以通过摄像头检测问题，并通过人工智能推理出如何调整电池的位置，从而完成任务。这种自动纠错和常识推理的能力，是具身智能所具备的“大脑”在发挥作用。尽管这些机器人外观看起来并不特别智能，但要实现类人的操作仍然非常困难。

三联生活周刊：如果将这些人形机器人投入使用，可能会存在哪些安全风险？

杨耀东：短期内，这些机器人可能还无法直接进入消费市场，它们仍然会主要用于工厂环境。工厂相对封闭，任务明确，安全问题较容易控制。但如果将这些机器人置于开放的环境中，期望它们像人一样执行各种任务，出错的概率就会大大增加。而且，由于这些机器人在物理空间中操作，任何失误都可能带来严重后果，比如对人类的伤害。电车难题不仅存在于自动驾驶领域，也可能出现在其他具身智能场景中。

三联生活周刊：所以对家庭场景的具身智能，我们更加需要保持警惕性？我了解到，具身智能在家用场景中的落地似乎更为困难。

杨耀东:是的，家庭场景比其他场景复杂得多。在工业环境中，机器人任务明确且受限，比较容易实现。但在家庭等开放性更强的环境中，情况复杂得多。每个家庭的布局和习惯都不同，这使得人形机器人可能面临各种难以预见的问题。就像现在扫地机器人在家庭环境中也会遇到诸多挑战一样，在家用场景中实现具身智能的落地确实非常困难。

目前，全球各地都在大力发展人形机器人，部分原因是应对人口老龄化和康养需求。这些领域是人形机器人目前最常见的应用场景。然而，当这些机器人进入家庭或其他生活场景后，比如在银行或酒店迎宾，听起来不错，但实际上存在很大风险。机器人在与人对话时，不仅可能因价值观问题引发情感伤害，在物理空间中，它们甚至可能对人身安全造成威胁，比如踩到人的脚。这些风险在基于文本的人工智能应用中尤为明显。现有的聊天机器人可以模拟情侣或朋友的角色，这种个性化互动可能让用户产生情感依赖。一旦这些机器人变得具身化，风险就更大了。

此外，人形机器人目前在数据监控和追溯系统上存在巨大的空白。与自动驾驶车辆配备的强制数据记录和黑匣子系统不同，人形机器人尚未建立类似的标准。如果这些机器人在家庭中造成伤害，我们还缺乏明确的追溯和处理方案。这些机器人还可能被用于不道德的行为，这些风险在大规模部署之前必须认真考虑。虽然我们在自动驾驶的伦理、法律和治理方面已经有了许多思考，但对于人形机器人等其他具身智能的相关问题，讨论仍不够深入。如果这些问题得不到充分考虑，可能会在大规模部署时忽视潜在的风险。

三联生活周刊：随着国内外积极开发和部署人形具身智能机器人，是否已经有相关的安全设置和讨论？

杨耀东：目前这方面几乎是一片空白。无论在国内还是国际上，具身智能的安全和伦理问题还没有得到充分研究和解决。我们对虚拟空间中的人工智能已经进行了大量关于安全性和价值观对齐的讨论，但在具身智能，尤其是人形机器人的安全治理上，还缺乏深入的研究。

设想一下，一个人形机器人在家中正要迈步，这时有人突然摔倒。如果机器人能意识到不该踩下去，那接下来该怎么处理？它是应该回退到几秒前，还是继续前进但小心跨过障碍？这些细节都需要有明确的定义。为了确保安全，机器人可能需要在脚上安装摄像头，以检测周围环境。这种思路可能会彻底改变行业的安全标准和技术发展方向。此外，机器人是否还需要类似“黑匣子”的设备，记录所有操作数据，并将这些数据上传到云端，由政府进行监管，也是需要考虑的问题。

虽然具身智能在人工智能领域已成为热点，但在安全治理方面的进展却远远落后。这是一个亟须关注的领域，特别是在这些机器人大规模进入家庭之前，我们必须对其安全和伦理问题进行全面的思考。