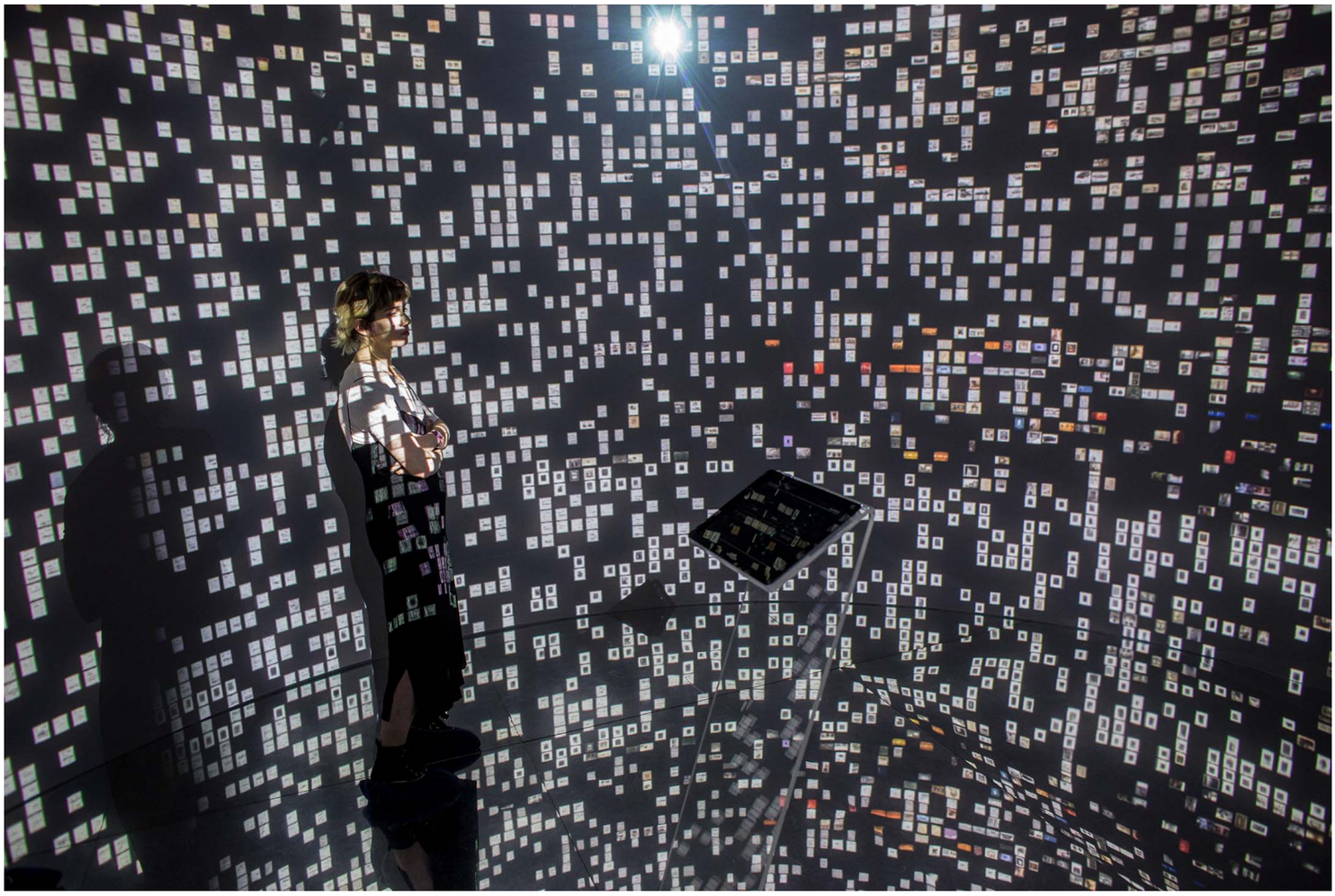
评论 国际 科技 深度

“人工智能会让人类灭绝”——围绕这个问题的最新讨论都说了什么？

“降低人工智能导致人类灭绝的风险，应该与大流行及核战争等其他社会风险一样作为全球优先事项。”



2017年5月6日，一名在女子在观看由人工智能生成的历史档案。土耳其艺术家Refik Anadol的Archive Dreaming装置，整合近200万份历史档案，将其变得可视化。摄：Chris McGrath/Getty Images

布犁 | 2023-06-19

人工智能 评论 科技

在ChatGPT-4的大模型发布后，连续好几个月内，到处都是讨论人工智能的声音。5月30日，英国大小报

在ChatGPT-4的大模型发布后，连续好几个月来，到处都是讨论人工智能的声音。5月30日，英国大小报纸的头条，是一份由超过300位各界人士参与的“AI可能导致人类灭绝”的联署声明。签署者也包括了大量AI科学家和前沿研发人士。

AI的缔造者们开始担心AI毁灭人类了？这份声明很短，只有一句话，翻译过来就是：“降低人工智能导致人类灭绝的风险，应该与大流行及核战争等其他社会风险一样作为全球优先事项”。

这样的讨论大概都说了些什么呢？

我们将面临哪些AI风险？

牵头起草这份声明的，是一个位于美国的非盈利机构，叫做“AI安全中心”（Center for AI Safety）。他们的主要工作是研究、讨论和倡导AI安全问题，促进相关监管和立法。

他们的主页上列举了AI将可能带来的一些最严重的安全威胁。有些看起来似乎有些危言耸听，但更多的则让人紧张：

列于首位的风险，是AI的军事化——这倒不是说会像科幻片里那样，出现一个超级AI大脑要灭绝人类。这样的风险暂时不大。但颇为可能出现的情况，是各国都可能用AI参与武器研发和军备升级，并将AI装入一些武器系统，人工智能将被拿来服务于政治野心和战争。

另一件较为人知的风险是AI大模型将制造和传播更多虚假信息。目前，ChatGPT的问答中就已经显示，AI会在交谈中产生不少和事实不符的内容，比如让它解释一位人物时，AI有机会虚构人物的一些生平细节。有一些最新的研究就指出，AI其实能做到“表里不一”和主动“撒谎”。甚至在大数据模型给我们编造一个不存在的错误答案时，它可能自己对正确答案有一定的理解。

但这只是虚假信息的其中一面，进一步发展的大模型还可能被用来自动生成容易让人们相信的政治谣言或政治宣传，从而更深地影响社会和政治。

2023年3月22日，在CloudFest展会上，背景萤幕上显示“此标志不是由ChatGBT编写的”。摄：Philipp von Ditfurth/picture alliance via Getty Images

其他主要的AI风险，还包括对AI的依赖将导致人类的个体思考和行为能力大幅衰落；以及谁可以享受或者购买到AI资源，谁无法触及AI资源，这造成的结果会是围绕着AI生产力，人类社会的进一步贫富分化。总而言之，在通用的大模型AI出现后，相关的风险也可以说是渗透到了人类社会的各个方面。

像“AI安全中心”这样的机构聚焦于推动人们认识AI风险，并寻求管控风险的方式——尤其是让立法者和执法者留意到这一点。可以预见，这样的倡导机构在未来会变得更多。

立法者如何讨论通用AI大模型？

AI大模型不同于普通AI，规管起来并不容易。欧盟先前计划中对AI的规划方式，是将不同种类的AI归为不同的风险等级。但这样的模式并不能适应通用大模型带来的挑战。

在这方面，美国参议院可以说是走在了各国立法机构的最前。他们在五月份举行的一场超过三小时的听证会值得一听。这场听证会的全文也可以在科技政策媒体TechPolicyPress上找到。

听证会上，三位接受参议员质询的业界人士分别是开发ChatGPT的OpenAI的CEO阿特曼（Samuel Altman）、IBM的首席隐私和信任官蒙哥马利（Christina Montgomery ）以及纽约大学的认知科学教授马库斯（Gary Marcus）。

相比之前对TikTok的听证会上很多中老年议员聚焦于程式“带坏年轻人”，这次针对AI的听证会中，列席的参议员们更有备而来，提问深度也更亮眼。开场的参议员布卢门撒（Richard Blumenthal）甚至开了个与时俱进的玩笑，他用ChatGPT生成的演讲稿和AI语言合成读了一段“他自己”的开场白，“你们能听出来这不是我吗？”

2023年5月16日，“ChatGPT之父”、公司执行长阿尔特曼(Sam Altman)，出席参议院隐私、技术和法律司法小组委员会听证会。
摄：Win McNamee/Getty Images

在这场听证会上，开发人工智能的头部企业发表忧心忡忡的言论，也许更多是为了在未来遇到事情的时候撇清责任——你看我都提前跟大家预警了！不过，听证会上包括阿特曼在内的人也确实提到了AI可能会在未来导致的巨大问题——比如ChatGPT的接口有机会帮助犯罪分子模拟人类去实施诱骗和欺瞒行为；又比如大模型的政治操控能力可以预测选民倾向和引领舆论；又比如AI显然会对就业市场有巨大影响；再比如通用AI将可能带来全球范围内语言之间的不平等——一个例子是，冰岛语这样的语言是不是会变得更边缘？阿特曼等人也直白承认这些问题都会存在——看起来他并不是一个技术乐观主义者。

概而言之，马库斯教授在听证会上的一个类比可能是恰当的，他说我们已经“创造了一头在瓷器店里的公牛”。尤其是，从AlphaGo开始，就有很多人提到过AI的思考过程其实是一个“黑箱”，我们很难确定AI到底是怎么具体“思考”的，因此我们对于AI到底有多安全，其实无法说真的有把握。

我们在加速冲向风险吗？

面对这样的风险，目前我们有什么可以采取的措施呢？

在美国参议院的听证会上，人们讨论到是否要让AI发展“减速”下来。在舆论场上，马斯克等人先前也通过

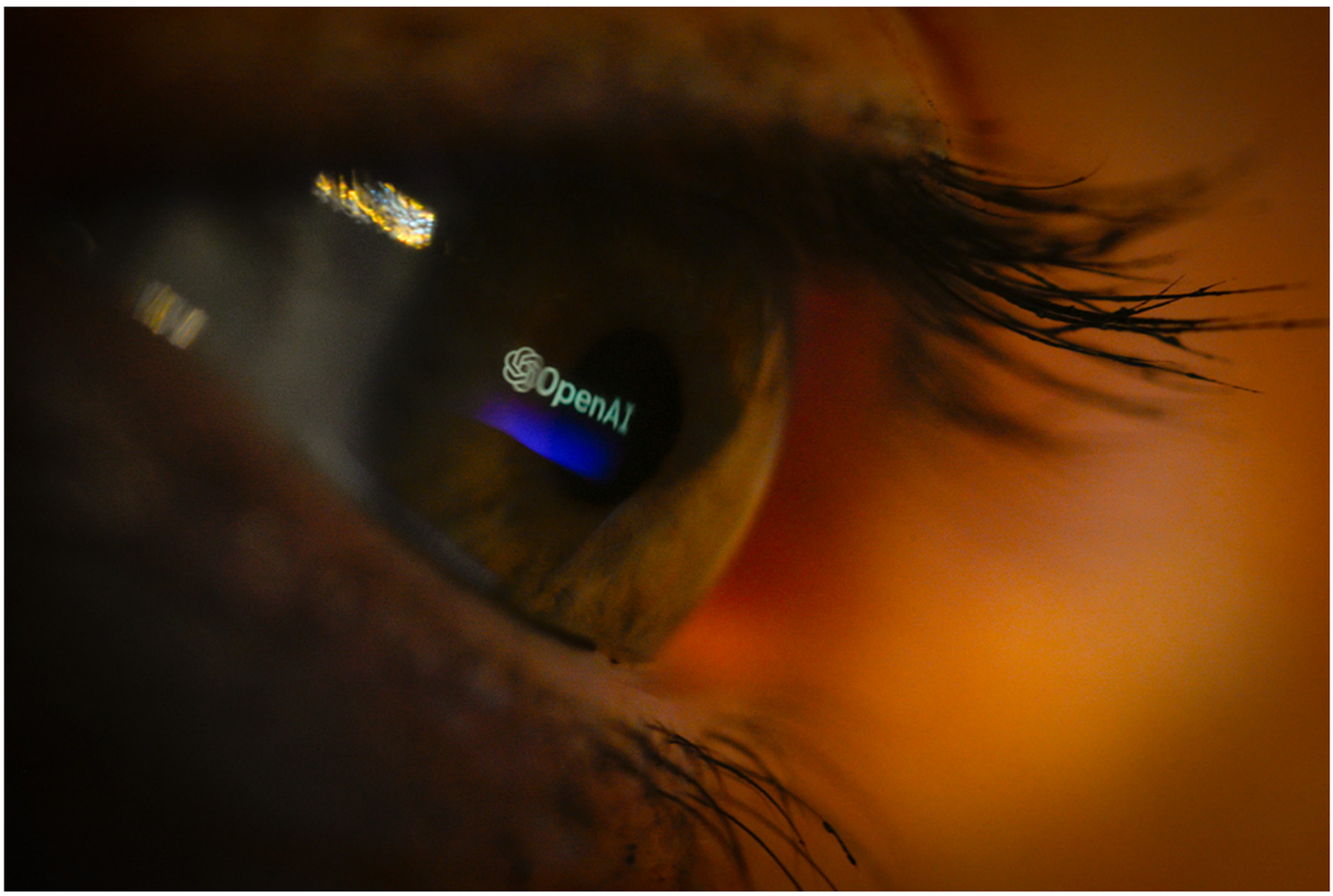
联名信提出暂时停止训练通用AI。

最近，获得过2018年图灵奖的加拿大科学家本希奥（Yoshua Bengio）有另一个建议，那就是在暂停超过GPT-4等级的AI的开发之外，还要禁止开发所谓的“智能主体”（intelligent agent）。所谓智能主体，是AI大模型接入了一个可以以语言之外的多种方式（比如行走、行动、制造等等）和世界互动的实体上。本希奥和不少人都认为智能主体一旦产生，其对现实的影响将比单纯的语言大模型更加巨大，且变得极为不可控。

但是，从听证会到公开信到调查研究机构，人们也意识到要“减缓”或者真正管控AI风险是非常困难的。

首先的困难就是，当下的全球竞争——中美之间，又或者商业竞争——各大公司之间，都在鼓励更快、更大规模的通用AI开发。OpenAI的CEO阿特曼也指出，除非全球合作，否则很难实现有效的对人工智能领域的监管。这就回到了那封声明的问题——假如AI真的和原子能一样，是一件既可以造福人类又可以毁灭人类的技术，那么全球是否要有类似《不扩散核武器条约》一样的，针对人工智能的条约和国际组织呢？

当前，各国政府显然没有觉得问题有这么严重。



一只眼球里映出了 OpenAI 标志。摄：Jaap Arriens/NurPhoto via Getty Images

也有不少AI乐观论者，比如Meta的首席AI科学家杨立昆（Yann LeCun）就不同意“AI可能导致人类灭绝”的看法。他的论据之一是AI的发展还没有到真的构成威胁的程度，当前的通用AI的水平也相当初级。

不过，通用人工智能的发展正在逐渐加速。2015年的时候，还有业界专家指出，担忧AI对人类的威胁就像“担心火星上人口过剩”。但八年之后，这个威胁确实变得更明显了。我们可以猜测，真正构成严重威胁的通用AI还需要几年时间才能产生？马库斯在听证会上暗示可能要20年——也并不是一个很远的时间点。我也拿这个问题询问了在人工智能领域工作的朋友，对方的回复是：如果要期待一个具有创造力的，在一些专业工作领域上能够开始超越人类的通用人工智能模型，也许在未来两三年就会有些眉目了。当然，这也只是一种估计。

相比通用AI的发展加速度，尽管各国现在都说要开始监管，但是所做的还是远远不够。美国参议院的听证会上讨论了一些可能手法，比如类似食品标签那样为AI增加“计分卡”——这个模式大概就是当我们在用AI产品时，会跳出来一个提醒：“ChatGPT，有时候会撒谎，ta说的话不能作为事实使用”。其他的再讨论之中的监管措施也包括了要成立一个专门监管AI的机构、在AI产品上标注大模型接受训练的来源等等。

监管速度很可能跟不上AI发展的速度，尤其是像美国这样的，非常依赖“打官司”来厘清法律责任的系统，速度就更慢了。况且，还有很多基础问题没有解决：比如厘清AI责任时，是否要修改美国《通讯规范法案》（CDA）的230条？（网络服务供应商无需为第三方使用者的言行负法律责任。）

相比之下，欧盟的人工智能监管法案已经进入了落实阶段。6月14日，欧洲议会通过了新修改的《欧盟人工智能法案》（EU AI Act），这意味着欧盟接下来将在机构和机制设计层面去执行超过300页内容的人工智能监管办法。其中包括了对人工智能训练资料、训练范畴、算法公开等方面的要求。据相关报道，法案也针对新出现的大模型进行了修订。这其中值得注意的一点是，加强人工智能监管，意味着这一领域领先的美国公司和欧洲监管当局之间的冲突可能在未来几年加深——欧洲的法案在何种程度上能够“管住”人工智能的负面属性，对硅谷又能否带来实质性的影响？