# 引言：艰难抉择与轻松抉择

本文是尤德科夫斯基（Yudkowsky）与索阿雷斯（Soares）合著《若有人造，众生皆亡》（If Someone Builds It, Everyone Dies）的首批在线补充资料之一。书中每一章的配套页面不仅回应常见疑问，也提供延伸的背景与细节——这些内容虽未纳入全书的核心论证，但同样具有重要价值。

请注意，这些资源信息量丰富，并非为逐页通读通篇阅读而设计。本书本身已自成体系，完整阐述了核心论点。但如果您对书中某些问题有特定疑问、异议或好奇之处，这里或许能找到进一步的解答。若确有遗漏，也欢迎[在此处]（TODO）提出您的建议。

## 常见问题

### **为何要撰写一本关于超级人工智能灭绝风险的书？**

#### 因为形势确实严峻，而且时间可能并不多。

当你深入思考某个问题时，有时能瞥见历史即将转向的岔路口。

1933年，物理学家利奥·西拉德（Leo Szilard）率先意识到核链式反应具有可行性[[1]](#footnote-1)，从而成为最早预见这一历史转折的人。

我们认为，如果从正确的角度观察当今的人工智能发展，同样能看到这样的拐点。而如果人类不及时调整方向，未来很可能走向不利的局面。

如今，各大AI实验室正竞相研发比人类更聪明的机器，并且在推进前沿研究方面似乎取得了显著进展。正如本书后续章节将讨论的那样，现代AI更像是“被培养出来的”，而非“被精心打造的”。它们展现出无人预见、也无人预期的行为模式，正沿着超越人类的轨道前进。对我们而言，这是极其危险的信号。

该领域的顶尖科学家们已共同签署公开信，呼吁公众应将人工智能的风险视为“与大流行病和核战争同等级别的全球优先事项”。这并非少数人的杞人忧天：近半数业内专家都表达了类似的忧虑。即使您一开始对此存疑，我们希望这些来自一线专家的警示，以及一旦担忧成真将付出的巨大代价，足以说明这个议题值得认真探讨。

这是一个需要我们诉诸理性与证据，而非停留于直觉判断的议题。倘若这些警告并非危言耸听，那么世界已然站在一个极度危险的边缘。在本书的后续部分，我们将逐步阐明支撑这些警告的理由和证据。

我们并不认为局势已无可挽回。本书的写作本身，正是希望能改变人类当前似乎正在滑向的轨迹——因为我们相信，仍然存在希望。

解决问题的第一步，是先理解问题本身。

### 您是在暗示ChatGPT会导致人类灭绝吗？

**并非如此。真正的担忧在于未来人工智能的发展。**

您之所以现在读到这本书，部分原因是像ChatGPT这样的突破让人工智能重回公众视野。世界开始讨论AI的进展及其社会影响——这正是一个合适的契机，让我们谈论“超越人类智能”的AI，以及目前令人担忧的发展趋势。

我们作者团队在此领域深耕已久。虽然近期AI的进展让我们的担忧更具体，但这种担忧并非始于ChatGPT，甚至不源于更早大型语言模型。数十年来，我们持续从事技术研究，致力于确保超越人类的AI能安全发展（索阿雷斯自2013年起，尤德科夫斯基则自2001年起）。最近，我们看到迹象表明：世界或许终于准备好展开这场必要的讨论——否则，我们可能会错过最后的应对时机。

人工智能正快速进步，终有一天（我们无法确知何时）会出现比人类更聪明的AI。而这正是所有主要AI公司的明确目标：

“我们现在确信，已经掌握了构建传统意义上通用人工智能（AGI）的方法。[……] 我们正在将目光转向更远的目标——真正意义上的‘超级智能‘。我们热爱现有的产品，但我们在此，是为了迎接那个辉煌的未来。有了超级智能，其他一切皆有可能。”

—山姆·阿尔特曼（[Sam Altman](https://blog.samaltman.com/reflections)），OpenAI首席执行官

“我认为强大的人工智能可能最早在2026年就会出现。[……] 我所说的强大AI，是指是具备以下特质的人工智能模型：它在纯智力层面上，比诺贝尔奖得主在大多数相关领域都要聪明——包括生物学、编程、数学、工程、写作等。也就是说，它能证明未解决的数学定理，能写出极高水准的小说，能从零开始编写复杂的代码库。”

—达里奥·阿莫代（Dario Amodei），Anthropic首席执行官

“总体而言，我们的核心目标是构建完整的通用智能。我今天所谈到的所有机遇，都是在实现通用智能并高效运作之后才能真正展开的。

—马克·扎克伯格（[Mark Zuckerberg](https://www.facebook.com/share/p/16STVBshtn/)），Meta首席执行官（该言论在Meta宣布其143亿美元“超级智能”计划不久前发表）  
“我认为，在未来五到十年内，我们有大约50%的几率实现我们所定义的AGI。“  
—德米斯·哈萨比斯（[Demis Hassabis](https://youtu.be/CRraHg4Ks_g?feature=shared&t=41)），谷歌 DeepMind首席执行官

“所以，德米斯，你是在试图引发一场‘智能爆炸‘吗？“  
”不，我希望那不是一场’失控的智能爆炸……“  
——采访者韦斯·罗斯（Wes Roth）与哈萨比斯的对话

这些言论并非空谈——他们确实在为此投入巨额资金。微软、亚马逊和谷歌均宣布计划于2025年投入750亿至1000亿美元建设AI数据中心。初创公司xAI以约800亿美元的估值收购社交媒体平台X.com（约为X自身估值的两倍），随后又筹集了100亿美元，用于建设大型数据中心并继续研发其AI系统“Grok“。OpenAI则与微软等合作伙伴共同启动了“星门计划”（Project Stargate），总投资高达5000亿美元。

Meta首席执行官马克·扎克伯格表示，公司今年预计将在AI基础设施上投入650亿美元，未来几年将在AI领域投入“数千亿美元”。Meta还向ScaleAI投资143亿美元，并聘请该公司CEO负责新成立的“Meta超级智能实验室”（ Meta Superintelligence Labs）。同时以高达2亿美元的报价从多个竞争对手实验室[[2]](#footnote-2)挖走了十多位顶级研究员。

当然，这一切并不意味着“超越人类的AI”触手可及。但它清晰地表明，所有主要科技巨头正在以空前的资源强度推进相关研发，而诸如ChatGPT的系统正是这一研究路径下的阶段性产物。这些公司的终极目标并非仅仅打造更优秀的聊天机器人，而是实现超级智能；聊天机器人只是中途站。

基于数十年对人工智能的研究，以及对技术发展趋势的严肃研判，我们认为：从原理上并无不可逾越的障碍，研究人员可能在不久的将来实现关键突破，创造出超越人类的智能。

我们尚不确定这一临界点究竟何时到来——也许就在不久的将来，也许还要十年、二十年。历史早已证明：精准预测新技术问世的时间，远比判断该技术最终能否实现要困难得多。但我们认为，目前的风险证据已足以促使国际社会采取积极行动——本书对此已作论述。

### 人们不是总在过度恐慌、反应过度吗？

**确实如此。但这并不意味着危险就从未真实存在过。**

有时人们会对问题反应过度，也有人倾向于宿命论。某些社会性恐慌最终也被证实缺乏依据，但这并不意味着我们生活在一个绝对安全的世界。

1935年的德国，对犹太人、罗姆人以及其他少数群体而言，绝非安居之地。当时有人察觉到警示信号而选择离开，也有人将警告视为危言耸听，最终未能逃脱悲剧。

核毁灭的威胁是真实存在的，但人类选择面对并采取行动，最终使冷战没有滑向热战的深渊。

氯氟烃确实在破坏臭氧层，但在国际条约禁止使用后，臭氧层得以恢复。

有些警告是虚假的，有些却是真实的。人类在面对挑战时，并非总是反应过度，也并非总是反应不足。有时两者兼而有之——例如，各国曾为下一场战争倾力建造巨型战列舰，却忽略了发展航空母舰这一真正关键的方向。这说明，并不存在一种简单的通用法则可以套用于所有技术风险：既不能对所有所谓的技术风险一概“置之不理，也不能对所有风险“一概当真”。要辨别真伪，唯有深入每一个具体案例，审视其背后的细节。

（更多相关讨论，请参阅本书引言部分。）

这种令人忧虑的人工智能，究竟何时会出现？

#### 预见到某项技术终将到来，并不等于能够准确预测它何时出现。

人们常常要求我们给出确切的时间预测，但事实上，我们对此几乎一无所知。1939年，当利奥·西拉德致信美国政府，警告核武器的潜在威胁时，他并没有——也无法——写下这样的预言：“第一颗原子弹将在6年内试爆成功。”

这无疑是极具价值的情报，但即便像西拉德这样率先洞察核链式反应原理、并准确预见其深远影响的人，也仍然无法精确预测技术实现的具体时间点。

有些结论容易得出，有些则极难。我们不会假装自己能对那些难题妄下结论，比如危险级别的人工智能究竟会在何时何地出现。

#### 人工智能的发展速度之快，一再出乎专家的预料。

无法预知AI确切到来时间，并不代表我们知道它遥遥无期。2021年，预测平台Metaculus上的社区预测认为，首个“真正通用的人工智能”将在2049年出现。仅仅一年后的2022年，这一预测被大幅提前12年，改为2037年。又过了一年，在2023年，预测值再次提前四年，调整为2033年。预测者们屡屡被AI的飞速进展所震惊，预估时间每年都发生巨大波动。

这种现象并非Metaculus独有。组织“8万小时”同样记录了多个专家预测团队不断将时间线大幅提前的案例。甚至连那些在预测竞赛中屡次夺冠、判断力常优于领域专家的“超级预测者”，也仅认为人工智能在2025年前获得国际数学奥林匹克金牌的概率为2.3%。然而，人工智能在2025年7月的确实现了这一目标。

超越人类智能的AI看似尚需数十年发展，就在不久之前，像ChatGPT这样的智能水平在2021年仍被视为遥不可及，而转眼之间，它已成为现实。谁又能断言下一次质变性的AI突破何时会发生？也许还需要十年，也许，就在明天。我们无法预测具体的时间点，但越来越多的研究者开始担忧，留给人类应对的时间可能已经不多。我们并不自诩掌握独到的信息，但我们坚信——行动刻不容缓。毕竟，没有人能确定，我们还能拥有多少预警时间。

建议读者参考第1章，了解人工智能能力如何在看似毫无征兆的情况下引发连锁反应；并参阅第2章，分析现代人工智能范式是否具备实现“最终突破”的潜力。

#### 同时，请对媒体关于某项技术“近期可能或不可能实现”的报道保持警惕——因为这些事件有时甚至早已发生。

1903年，威尔伯·莱特（Wilbur Wright）曾悲观地预测动力飞行可能还需一千年；而仅仅两年后，《纽约时报》更断言人类实现飞行“需要一百万年”[[3]](#footnote-3)。结果，仅在文章发表后的两个月零八天，莱特兄弟便成功试飞。

尽管机器学习的最新进展已表明，人工智能在越来越多的测试项目中达到甚至超越了人类水平，但至今仍有怀疑论者坚称，AI绝无可能在某些特定能力上媲美人类。例如，《纽约时报》在2025年5月仍重复着这样的论调：“科学家尚无确凿证据表明现有技术能实现大脑的某些基础功能（如识别反讽）。”而事实上，早在2024年末之前，现代AI已能频繁通过文本乃至非语言线索辨识讽刺与反讽。”[[4]](#footnote-4)

归根结底，尽管有人断言“超越人类智能的AI即将诞生”，也有人坚称其“遥不可及”，但令人不安的现实是：目前并没有人真正知道答案。

更严重的是，国际社会极有可能在洞悉真相时已为时过晚，届时一切将无可挽回。

预测下一项技术突破的时机，难度超乎想象。我们深知，比人类更智能的AI蕴含致命威胁，但若想精确预知其降临之日，我们注定束手无策。我们必须学会在不确定中果断行动，否则将永远停滞不前。

我们能否依据以往的技术进展，可靠地推测出超越人类智能的实现时间？

#### 答案很可能是否定的，因为我们对“智能”本身的理解，尚不足以支撑这类预测。

一种常见的预测方法，是依据图表中一条持续多年的直线趋势，推断该趋势至少还能延续一两年。这种方法有时有效，但也并非总是可靠——趋势线终会改变，尽管在实践中，它偶尔能带来成功预测。

但这种方法存在一个根本问题：我们真正关心的，往往并不是“到2027年这条曲线会上升到多高”，而是“当曲线上升到某一高度时，会引发哪些本质性变化？”——即，图表上的数值达到何种水平，才会对应现实世界中具有重大意义的结果？

而在人工智能领域，我们对此几乎一无所知。选取某个易于量化的指标（如“困惑度”），并在图表上外推成一条直线，这固然简单。但没有人能预先知道未来的“困惑度”要达到多少，才会对应某种质变性的能力，例如国际象棋水平的突破。这种关系无法被预测，只能在系统运行之后才能被观测到。

没有人确切知道“如今它已具备灭绝人类的能力”这一论断在图表上处于什么位置。人们所能做的唯有运行人工智能来验证。因此，单纯依据图表上的直线趋势进行推断，并无助益。（更何况，随着算法进步，这类图表本身很快就会失去意义。）

正因如此，本书并未耗费篇幅去推演图表曲线，试图精确预测何时会有人在训练AI时投入1027次浮点运算，或这种行为将带来何种后果。这实在难以断言。本书聚焦于我们认为易于判断的领域。这仅涉及狭窄的议题范围，即便我们能在该范围内做出少数重要预测，也不足以支撑对未来进行任意预言。

### 作为作者，你们的动机和潜在利益冲突是什么？

#### 通常情况下，我们并不指望通过这本书盈利。更重要的是，我们其实暗自希望书中的核心观点最终被证明是错误的。

我们（索阿雷斯和尤德科夫斯基）均受雇于机器智能研究所（MIRI），该机构是一家专注于相关议题研究的非营利组织，其运作依赖关心此类问题的个人捐款支持。本书若引起更多关注，或许会为MIRI带来更多捐赠。

尽管如此，我们仍有其他收入来源，撰写此书并非出于经济动机。本书的预付版税已全部用于宣传推广，未来如有任何版税收入，也将全数返还MIRI，以补偿其在本书撰写过程中投入的人力成本[[5]](#footnote-5)。

当然，如果两位作者能够得出“人类文明并未面临生存危机”的结论，他们无疑会感到无比欣慰，甚至乐意就此退休，或转而投身其他更具收益的事业。

若确有充分证据需要改变立场，我们并不认为自己会固执己见。这样的转变并非首次发生：MIRI最初以“奇点研究所”之名创立时，其宗旨是推动超级智能的实现。尤德科夫斯基花了一年时间才意识到其发展未必顺利，又历经数年才真正理解，要确保其安全推进是何等艰难。

我们曾成功转变过立场，也始终愿意再次转变。只是迄今为止，现有的证据尚不足以促使我们做出这样的调整。

我们并不认为局势已毫无希望，但确实坚信存在实质性风险——若世界未能及时采取有效应对，所带来的威胁将极为严峻。

需要强调的是：若要判断人工智能是否真的将人类引向毁灭，就必须聚焦于技术本身的逻辑与发展轨迹。若仅以“人的动机”为由轻易否定所有警示——无论是学者脱离实际、企业在炒作、非营利组织为募资、还是业余人士在空谈——最终，一个人的信念将仅仅取决于他选择忽略哪些人。而一旦判断失误，再充分的论据和事实也将难以扭转其认知。真相，终究只能建立在论据本身是否成立的基础上，而不是由提出者的身份所决定。

本书并未以常见的动机论开篇——即质疑运营人工智能实验室的企业高管们有动机向公众保证AI的安全性。相反，我们首先将焦点置于人工智能技术本身。尽管在后续章节中简要回顾人类科学家历史上屡次出现的过度乐观倾向，但我们从未主张应因某人在AI 实验室任职就轻易否定其观点。我们所做的，是深入分析部分开发者提出的具体计划，并揭示这些方案在逻辑层面存在的根本缺陷。我们致力于推动一场建立在实质论据之上的对话，因为唯有论据本身的质量，才真正决定观点的价值。若您认为我们的论证存在错误，我们诚挚欢迎您针对具体论点提出异议。与揣测他人动机或立场相比，我们坚信这种方式更能逼近真相——正如即使世上最固执的人坚称天空晴朗，也无法改变雨水正悄然落下的事实。

### 人工智能仅仅是科幻小说里的概念吗？

\***仅凭某个话题在虚构作品中的出现频率，很难得出有实际意义的结论。**

尽管超越人类智能的人工智能尚未成为现实，但它确实早已成为科幻作品中的常见主题。然而，我们建议不应将虚构作品的描绘作为认知的参照。现实中的人工智能很可能与小说或电影中的形象大相径庭——具体原因，我们将在第4章展开分析。

事实上，人工智能并非第一个先于现实被文学想象所预见的技术。比空气重的飞行器、月球探索等，都曾在真正实现之前就出现在虚构叙事中。就连核武器的基础概念，也早在1914年就由科幻作家赫伯特·乔治·威尔斯（H. G. Wells）在其小说《获得自由的世界》*（The World Set Free）*中提出。虽然威尔斯未能准确描述细节——他笔下的炸弹是一种可持续燃烧数日的装置，而非瞬间释放能量并留下持久辐射的核爆——但他确实抓住了关键：一种基于核能而非化学能的新型炸弹。

1939年，阿尔伯特·爱因斯坦（Albert Einstein）和利奥·西拉德致信罗斯福总统，敦促美国加速研发原子弹，以赶超德国。我们不妨设想：如果罗斯福最初是从威尔斯的小说中了解到“核弹”这一概念，他很可能会将其视为幻想而置之不理。

现实中，罗斯福确实重视了这一警示——至少成立了铀咨询委员会。但此例恰恰提醒：仅因某位小说家曾提及相似概念而轻率否定，实属危险之举。

科幻作品可能从两个相反的方向误导你：既可能因为你将其情节误认为现实，也可能因为你全盘否定其构想。科幻作家既非先知，也非注定错误的反面预言家——他们的想象既不必然成真，也未必全无价值。在绝大多数情况下，我们更应放下对虚构内容的依赖，转而基于技发展的逻辑与现实情境展开分析。

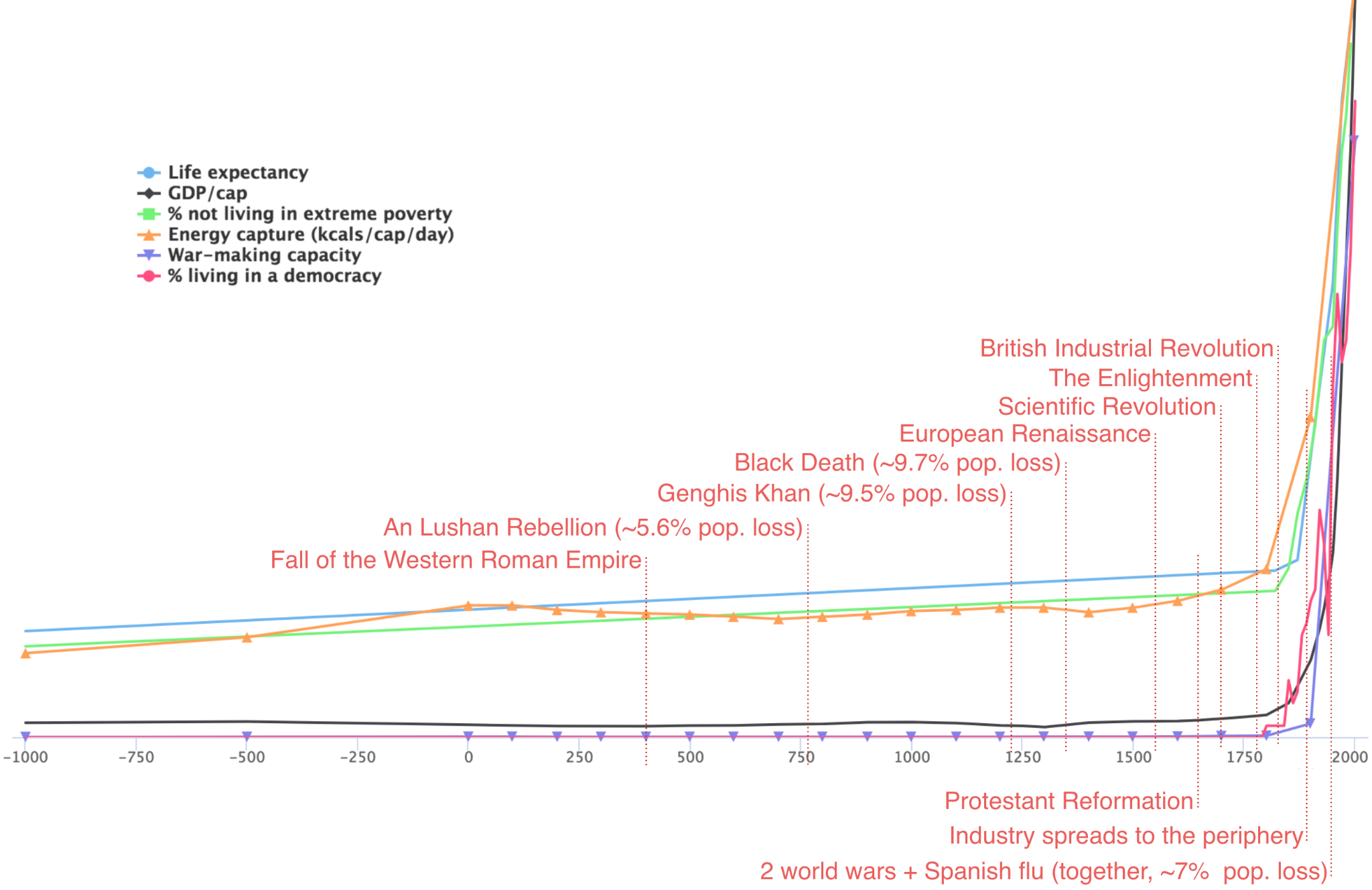
若要预测现实世界的走向，唯一可靠的方法，是深入思考相关论据，审慎权衡所有可得证据——除此之外，别无捷径。

#### 人工智能带来的后果注定会很诡异。

我们理解人们为何会觉得人工智能令人不安——它预示着一个将彻底改变并打破既有秩序的世界。我们每个人的直觉，都或多或少建立在这样一个前提之上：人类是唯一能够实现诸如建造发电厂等宏大工程的物种；我们的思维习惯，也始终适应于一个机器仅作为无智能工具的历史背景。然而有一点几乎可以确定：拥有超越人类智能的人工智能的未来，将与我们所熟知的一切截然不同。

世界的根本性变革往往并非渐进累积而成。尽管“一切如常”[[6]](#footnote-6)这一经验法则在多数情况下适用，但恰恰在其失效的时刻，历史进程往往迎来最关键的转折点。思考未来的重要意义，正是在于识别这些可能发生根本转变的节点，并为之做好应对准备。

要克服对现状的偏好，有效的方法之一是回顾历史记录——正如本书引言部分所讨论的那样。历史表明，某些发明最终以难以预见的方式彻底重塑世界。以蒸汽机为例：它不仅是工业革命的核心驱动力，还催生了一系列连锁性的技术创新，在相对短暂的历史时期内迅速而深刻地改变了人类社会的面貌：



真正的通用人工智能的出现，是否会带来同等深远的历史转折？其影响力或许至少不亚于工业革命。具体而言：

* 技术进步将显著加速。正如第1章将讨论的那样，机器运行的速度远超人脑。人类能够改进人工智能，而人工智能最终也将具备自我改进的能力——直至机器在科学发现、技术创新等方面远远超越人类。

纵观人类历史，即便工程技术不断取得突破，人类认知的基本机制却始终未曾改变。而当智能系统能够自我演进、自我完善时，我们应当预期，世界的诸多方面都将迅速而深刻地发生变化。

* 此外，如第3章所述，足够强大的人工智能很可能形成独立的目标体系。倘若AI仅仅是更快、更聪明的人类替代品，这本身已是巨大的变革。但实际上，人工智能将成为地球上一种全新的智能生命形式——其目标体系（如第4、第5章所将讨论的）很可能与人类目标存在重大分歧。

从本质上看，这两大变革若不彻底重塑现有世界秩序，将难以实现。若仍相信未来将“一切如常”，便意味着相信机器智能永远无法超越人类智能。这种的设想从未真正可行，而到了2025年，相较2015年或2005年，这种新年已更加难以维系。

#### 从更长远的角度看，未来的图景同样将超出我们现有的认知框架。

若把目光投向遥远的未来，所见的世界必然显得有点诡异。正如从19世纪的视角看21世纪已然怪异，而17世纪的人看来19世纪也同样如此。人工智能的出现，不仅加速了这一进程，也为这场演化之局引入了一个全新的参与者。

可以预见的一点是：拥有先进技术的智慧生命不太可能永远被束缚在诞生它的星球上。此刻，夜空中无数恒星正燃烧着能量，而原则上并无任何障碍阻止生命发展出星际旅行技术，从而将这些能量用于自身的目的。

尽管星际航行的速度受到物理规律的限制，但实现星际旅行本身似乎没有根本性的障碍[[7]](#footnote-7)。终有一日，我们将能够研发出星际探测器，奔赴浩瀚宇宙开采资源，将其转化为繁荣的文明；同时，还能制造出自我复制的探测器，将文明的疆域不断向深空拓展。倘若未来人类被人工智能取代，这些AI也会沿着同样的道路前行。只不过，它们所追求的终极目标将不再是“繁荣的文明”，而是它们自身的目的。

正如生命当初在地球上蔓延至荒芜的岩石，直至整个星球充满生机，我们亦可预见，生命（或生命创造的机器）终将扩散至宇宙中尚未被占据的角落。到那时，发现一个无生命的太阳系，将会像今日在地球上找到一座连细菌都不存在的荒岛一样令人难以置信。

目前，宇宙中的大部分物质（例如恒星）仍处于随机分布状态。然而，从足够长远的角度看，未来几乎注定会呈现另一番图景——绝大多数物质将按某种设计被重新排列，也就是说，将体现那些成功开发并重新利用恒星的智慧实体的意志。

即便地球上的生命从未扩展至宇宙，即便遥远星系中诞生的多数智慧文明终其一生都困守母星，只要宇宙中任何一个角落诞生出一个具备星际航行能力的文明，就足以点燃星火，开启对整个宇宙的蔓延征程。它们会前往新的恒星系统，利用那里的资源建造更多探测器，再向更遥远的星系扩散——正如地球上生命的起源：只需一个能自我复制的微生物（加上指数级增长），便足以让死寂的星球焕发生机，在每个角落、每座岛屿都绽放出生命的绚烂图景。

因此，未来的图景必将与现在迥然不同。事实上，我们几乎可以断定，它将发生颠覆性的变革。从长远来看，无论最终是哪种生物或人工智能在宇宙中寻找资源，恒星本身都注定会被重新塑造——即便以我们今天的认知，还无法确切描绘这些宇宙主体的形态，或宇宙资源最终将被用于何种目的。

预测细节似乎极为困难，甚至近乎不可能。然而，若问宇宙是否会被改造成一个大部分物质被采集并服务于某个目标的体系——答案显然更为确定。即便对于一个几乎尚未学会从恒星提取资源的文明而言，这一结论或许显得反直觉、甚至怪异，但却更为可信。

百万年后的未来，不会仍是2025年的模样——智能生命被困于一颗行星表面，如同未开化的猿猴般喧闹不休。在那之前，人类文明要么已自我毁灭，要么早已踏上群星之间的征途。[[8]](#footnote-8)

人类的未来注定将超越我们今日的想象。真正的问题在于——这一切将于何时发生。

#### 未来正以前所未有的速度向我们逼近。

人工智能等技术意味着未来可能很快叩响我们的门扉，其影响或将迅速而猛烈。

以旧时代的尺度衡量，工业革命的变革可谓迅疾；以进化的尺度衡量，智人重塑世界的速度可谓迅猛；以宇宙与地质的尺度衡量，生命改造世界的步伐可谓急剧。而按照旧标准来衡量，新的变革机制总能以更快的节奏重塑世界。

如今，人类似乎正处于又一次剧变的门槛：机器将开始以自身的高速节奏重塑世界，其速度远非生物进化所能企及。关于机器智能究竟在多大程度上超越人类智能，我们将在第1章和第6章进一步阐述。但至少，我们必须正视这一可能性：超人类智能的诞生，或将以惊人的速度彻底改变世界。。纵观历史，这样的范式转换从未停止上演。

## **扩展讨论：**

### **人****工智能专家论灾难性情景**

2022年，一项针对参加NeurIPS和ICML学术会议的738名人工智能研究者的调查显示，48%的受访者认为人工智能带来“极其糟糕的结果（例如人类灭绝）”的概率至少为10%。这表明，在这一领域，关于人工智能可能引发前所未有灾难的担忧相当普遍。

以下汇集了多位知名人工智能科学家与工程师对灾难性人工智能风险的看法。其中部分学者给出了他们所称的“毁灭概率”（ p(doom)）——即人工智能导致人类灭绝或类似灾难性结局的可能性。[[9]](#footnote-9)

图灵奖与诺贝尔奖得主、深度学习革命的奠基人 **杰弗里·辛顿**（Geoffrey Hinton，2024）在谈及个人判断时表示：[[10]](#footnote-10)

我认为[存在性威胁]的风险实际上超过50%。

图灵奖得主（与辛顿、杨立昆共同获奖），当代被引次数最高的科学家之一 **约书亚·本吉奥**（Yoshua Bengio，2023）说：

我们不知道距离真正的危险还有多久。过去几周我一直在说：“请提出有力的论据来说服我不必担忧——因为那样我会轻松得多。”但到现在，还没人能让我信服。[……]我个人认为灾难性结果的概率大约是20%。

**伊利亚·苏茨克弗**（Ilya Sutskever，2023），AlexNet联合发明者、OpenAI前首席科学家，与辛顿、本吉奥并列为人工智能领域被引次数最高的三位学者之一：

超级智能AI所蕴含的巨大能力同样机具危险，可能导致人类失去主导地位，甚至走向灭绝。虽然超级智能看似仍然遥远，但我们认为它可能在本十年内成为现实。[……]

目前，我们尚无有效方法来引导或控制此类潜在的超级智能，也缺乏阻止其失控。现有的AI对齐技术（例如基于人类反馈的强化学习）严重依赖于人类的监督能力。但当AI系统的智能水平远超人类时，人类将无法可靠地进行监督，因此这些方法无法适用于超级智能。因此，寻求颠覆性的科学与技术突破已成为当务之急。

**詹·莱克**（Jan Leike，2023），Anthropic公司对齐科学联合负责人、前OpenAI超级对齐团队联合负责人：

[采访者：“我并未花很多时间精确计算我的毁灭概率。我的猜测是高于10%，低于90%。”]

[莱克:] 这大概也是我给出的区间范围。

**保罗·克里斯蒂亚诺**（Paul Christiano，2023），美国国家标准与技术研究院（NIST）下属人工智能安全研究所安全主管、人类反馈强化学习（RLHF）的发明者：

在建成足以使人类劳动过时的强大AI后十年内，人类大规模死亡的概率为20%；[……]在建成强大AI后十年内，以某种不可逆方式毁掉人类未来的概率高达46%。

**斯图尔特·罗素**（Stuart Russell，2025），加州大学伯克利分校史密斯-扎德工程学讲席教授、《人工智能：现代方法》（Artificial Intelligence: A Modern Approach）本科生顶级教材合著者：

企业与国家间的“通用人工智能竞赛”有点类似于冷战时期的核军备竞赛，但情况更为严峻：“即便参与竞赛的CEO们也承认，胜者出极有可能在过程中引发人类灭绝，因为我们根本不知道如何控制比自身更智能的系统。”换言之，这场竞赛无异于一场“奔向悬崖的竞赛”。

**维多利亚·克拉科夫娜**（Victoria Krakovna，2023），谷歌DeepMind研究科学家、生命未来研究所联合创始人：

[采访者：“这并非令人愉快的话题，但您认为自己在2100年前死于人工智能的概率有多大？”]

[克拉科夫娜:] 2100年太遥远了，尤其考虑到目前技术发展的速度。我随口说个数字的话，大概20% 吧。

**肖恩·莱格**（Shane Legg，2011），谷歌DeepMind联合创始人兼首席通用人工智能科学家：

[采访者：“您认为人工智能开发不当导致负面或极端负面后果的概率是多少？[……]其中‘负面’指人类灭绝；‘极端负面’指人类受苦”。

[莱格:] 人类水平的人工智能问世后一年内[……] 我不确定。可能是5%，也可能是50%。没人能准确估计。如果所谓‘受苦’指长期痛苦，我认为这种可能性很低。但若超级智能机器（或任何超级智能体）决定消灭人类，我认为它会相当高效地完成。

**埃马德·莫斯塔克**（Emad Mostaque，2024），Stability AI创始人（该公司开发了Stable Diffusion）：

我对毁灭概率的估计是50%。在一个未定义的时间跨度内，系统超越人类能力、最终掌控关键基础设施并导致人类灭绝的概率，如同抛硬币——尤其是在我们当前的发展路径下。

**丹尼尔·科科塔伊洛**（Daniel Kokotajlo，2023），AI治理专家、OpenAI内部举报人、人工智能未来项目执行董事：

我认为AI灾难发生的概率为70%，那些认为其低20%的人，是非常不理性的[。]

**丹·亨德里克斯**（Dan Hendrycks，2023），机器学习研究员、AI安全中心主任：

我的毁灭概率超过80%，但过去更低，两年前约为20%。

上述所有研究者均签署了本书开篇引用的《人工智能风险声明》（[Statement on AI Risk](https://aistatement.com/)），其中写道：

缓解人工智能引发的灭绝风险，应与应对其他社会级风险（如大流行病和核战争）同等列为全球优先事项。

签署该声明的其他知名研究人员包括：ChatGPT架构师约翰·舒尔曼（John Schulman）；前谷歌研究总监彼得·诺维格（Peter Norvig）；微软首席科学家埃里克·霍维茨（Eric Horvitz）；AlphaGo研究负责人戴维·西尔弗（David Silver）；AutoML先驱弗兰克·胡特（Frank Hutter）；强化学习先驱安德鲁·巴托（Andrew Barto）；生成对抗网络发明者伊恩·古德费洛（Ian Goodfellow）；前百度总裁张亚勤（Ya-Qin Zhang）；公钥密码学发明者马丁·赫尔曼（Martin Hellman）；视觉变换器研究负责人阿列克谢·多索维茨基（Alexey Dosovitskiy）。名单仍在延伸，包括宋晓东（Dawn Song）、雅沙·索尔-迪克斯坦（Jascha Sohl-Dickstein）、大卫·麦卡莱斯特（David McAllester）、克里斯·奥拉（Chris Olah）、金秉恩（Been Kim）、菲利普·托尔（Philip Torr）等数百人。

### 当利奥·西拉德预见未来

1933年9月，物理学家利奥·西拉德（Leo Szilard）在伦敦正穿过南安普顿路与罗素广场交汇的十字路口时，突然灵光乍现[[11]](#footnote-11)，首次构想出“核链式反应”——这正是原子弹的核心原理

然而，将划时代的构想变为被认可的真理，道路从未平坦。西拉德求助于声望更高的物理学家伊西多·拉比（Isidor Rabi），后者又请教了更具权威的恩里科·费米（Enrico Fermi）。拉比问费米是否认为核链式反应是真是可行的，费米仅回了一个词：

一派胡言！

拉比追问其含义，费米答道那只是“极小的可能性”。拉比又问“极小”指多大概率，费米说：“10%。”

拉比回应道：“如果这10%意味着我们可能因此丧命，那它就绝不是小概率。”费米于是陷入了思考。

这个故事蕴含着多重启示。我们**不**认同的一个教训是：“任何可能致命的微小风险都值得担忧。”10%并非 “微小”，但若真足够微小，则确实不必花心思考虑。

我们真正认同的启示有二：其一，有时人们能意识到某种技术（如放射性级联）确实可能实现，从而预见到世界即将发生剧变；

其二，人们的直觉往往不是理解或预测这种剧变的可靠向导——即使那个人是如费米般卓越的专家。

想想看，费米最初从何得出“渺茫可能性”和“10%”这类说法？

他为何认为放射性无法引发连锁反应？只是因为多数宏大设想最终都失败吗？

他那句“一派胡言！”似乎在说比这更强烈的东西。这似乎反映出一种感觉，即这个特别的伟大想法不太可能成功。但是为什么呢？有什么物理论证？

是不是只是觉得太疯狂？确实，核武器的存在将给世界带来颠覆性后果，但现实并不会因为结果严重就阻止这种事情发生。[[12]](#footnote-12)

当费米第一次听闻西拉德的设想时，他甚至提议将其公开发表，让全世界——包括德国及其新上台的元首阿道夫·希特勒（Adolf Hitler）——皆知此论。

幸而费米最终未能说服西拉德，因为核武器最终被证明确实可行。虽然他一直持怀疑态度，直到亲自主持建造人类首个核反应堆“芝加哥一号堆”，但他最终还是加入了西拉德那支仅有寥寥数人的秘密小组。

有些技术，注定将重塑世界。如果你轻率地将颠覆性的想法斥为“一派胡言”，便可能被进步所淘汰——哪怕你是世界上最聪明的科学家。费米的可贵之处在于，他愿意与西拉德展开辩论；而更难能可贵的是，他在亲眼见证技术实现之前——在仍有机会改变未来之时——就已被说服并采取了行动。

人类历史上，灾难频仍，但那些灾祸得以避免，往往正是因为有人愿意坐下来认真对话——有时甚至像西拉德之于费米那样，强迫对话发生。

1. #我们在延伸讨论中讲述了利奥·西拉德故事的一部分。 [↑](#footnote-ref-1)
2. 挖角：据《彭博社》2025年7月报道：“Meta首席执行官马克·扎克伯格已成功挖走十余名OpenAI研究员，以及来自Anthropic、谷歌等初创企业的顶尖研究员和工程师。” [↑](#footnote-ref-2)
3. \*不妨看看1903年刊载的文章《无法飞行的飞行器》*（Flying Machines Which Do Not Fly）*中的这段“自信”断言：

   任何机器都只能遵循作用于被动物质的自然法则运行。因此，如果说一只翅膀尚未发育完全的雏鸟需要一千年才能轻松飞行，或一个完全没有翅膀的生物需要一万年才能从零演化出翅膀，那么我们或许可以推断：真正能够飞行的飞行器，可能需要数学家与机械师们持续协作一百万到一千万年才能造出——当然，这还得建立在我们可以在此期间解决诸如“无机材料在重量与强度关系上的固有缺陷”这类小问题的基础上。毫无疑问，这个问题对感兴趣的人来说颇具吸引力，但对普通大众而言，他们的精力或许更适合投入到更有实际收益的领域。 [↑](#footnote-ref-3)
4. 是的，人工智能甚至能识破《纽约时报》声称它们无法识别讽刺这一报道中的讽刺意味。（公平起见，《纽约时报》部分记者在报道人工智能时确实更清晰明了。） [↑](#footnote-ref-4)
5. 如果该书的表现足以收回所有投资，我们的合同条款规定：在MIRI获得充分回报后，作者将与MIRI分享利润。然而，MIRI为本书投入了巨大精力，除非销量远超预期，否则我们恐怕永远拿不到一分钱。 [↑](#footnote-ref-5)
6. “一切如常”：“一切如常”这一表述在预测市场的参与者中似乎颇为常见。例如，博主斯科特·亚历山大（Scott Alexander）就在其文章《几乎总是奏效的启发法》*（Heuristics That Almost Always Work）*中就讨论了这一思维方式。 [↑](#footnote-ref-6)
7. \*无限制：例如，论文《六小时的永恒》*（Eternity in six hours）*探讨了在已知物理定律约束下星际殖民的极限。 [↑](#footnote-ref-7)
8. 或者，他们将选择建造工具或创造后继者来代为探索。无论采取何种形式，只要他们认为合适，并能充分利用更先进的科学与技术成果。 [↑](#footnote-ref-8)
9. 我们对当前试图给“p(doom)”——即人工智能导致人类灭绝的概率——赋值的行为深感忧虑。这种试图给出单一概率的做法（而不是基于社会可能采取的不同应对措施给出多个概率）在我们看来充满宿命论色彩。同样是高p(doom)，其背后含义天差地别：一种人认为世界基本上无法阻止灾难，另一种人则认为世界本可以阻止灾难，但不会去做。

   如果事实是大多数人对高p(doom)的判断基于后者，但所有人都误以为前者为由，那么这些高概率的宣称就可能成为一种自我实现的预言，将我们推向本可完全避免的灾难轨道。

   我们还注意到，硅谷的许多人交换“p(doom)”数值的方式，有点像交换棒球卡，常常显得脱离现实。只要你认真思考就会明白，即便是低至5%的全球人类灭绝概率，也理应引发极度警觉——这一概率远远超出了需要立即叫停整个人工智能领域发展的威胁阈值。然而一旦人们习惯于在聚会中冷酷地互换这些数值，把它们当作有趣的科幻话题而非关乎所有人命运的断言，他们便会惊人地迅速忽视这一现实。

   这并不是说人们给出的 p(doom) 数值接近真实。但至少，你应当将这些数字解读为该领域专家普遍在发出一个严肃信号：我们正面临一场真正的、紧迫的危机。 [↑](#footnote-ref-9)
10. 与辛顿在视频前半段的说法相反，尤德科夫斯基对风险的信心评估远非“99.999%”那般极端；事实上，若真达到“五个九”的置信水平，那才堪称是脱离现实的疯狂。 [↑](#footnote-ref-10)
11. \*十字路口：原子遗产基金会保存了更完整的传记档案和时间线。 [↑](#footnote-ref-11)
12. 面对针对费米的批评，我们注意到有人试图为他辩护，编造出各种理由来证明费米在说出“一派胡言！”之前确实经过了深思熟虑。例如，他们辩称，费米清楚地球此前并未因感生放射性的连锁反应而爆炸——而有人可能会认为，如果这类连锁反应在物理上是可能发生的，地球早就该爆炸过了。

    当然，这类论证最终指向的是错误结论。费米在核链式反应问题上的判断确实错的。正因如此，我们认为，这类论证所揭示的教训是：“对于事实上成立的事物，人们总能提出听上去至少同样合理的反对理由。”地球至今尚未爆炸，并不能证明核反应堆无法实现——人类工程师完全可以精心操控原子使其发生可控裂变。因此，这类论证并不足以支撑诸如“一派胡言！”这样完全偏离事实的结论。 [↑](#footnote-ref-12)