当所有人都在卷算力时，他却看到了AGI的“最后一块拼图”

在硅谷的聚光灯下，现在的AI圈像极了一场永无止境的军备竞赛。每隔几个月，总有一家巨头跳出来喊：“看！我的模型参数又翻了一倍！”或者“瞧！我又买了十万块H100显卡！”大家似乎笃信一条真理：只要数据投喂得够多，算力堆得够猛，AGI就会自动涌现。

然而，就在这股狂热的“算力崇拜”中，一位真正的重量级人物，OpenAI前首席科学家、ChatGPT的缔造者、现任Safe Superintelligence (SSI) CEO伊利亚·苏茨克维（Ilya Sutskever），却选择了一条截然不同的路。

最近，他在知名播客Dwarkesh Podcast（主持人：德瓦尔克什·帕特尔）中接受专访。没有为了融资而画的大饼，也没有公关式的套话。这次访谈，更像是一位刚刚从未来穿越回来的顶级科学家，心平气和地告诉我们：“别卷了，以前的那套玩法，到头了。”

这不仅是一次观点的输出，更是一份详尽的“AGI路线图”。他不仅给出了AGI降临的倒计时（5到20年），更深刻剖析了当前大模型的致命缺陷。

网友们纷纷评论称，苏茨克维在访谈中展现了他一贯的思考深度：不用复杂术语，却能直指AI发展的核心矛盾。他关于“我们从规模化时代进入研究时代”的判断，尤其值得每个关注AI领域的人深思，堪称“我们这个时代的奥本海默”。

苏茨克维究竟看到了什么？让我们拆解一下这场对话的核心干货。

01. 告别“大力出奇迹”：规模化时代的落幕

将时间拨回五年前，苏茨克维可能是那个最信奉“Scaling Law（缩放定律）”的人。但今天，他却成了那个亲手给“规模化时代”盖上棺材板的人。

在访谈中，苏茨克维像一位严谨的历史学家，将2010年代末到2025年定义为“规模化时代（The Age of Scaling）”。

这几年是AI发展的“黄金蜜月期”。逻辑简单粗暴却极其有效：只要你增加计算资源，增加数据量，模型的能力就会线性增长。这种高度的确定性，让风投和科技巨头们趋之若鹜。

但苏茨克维现在的判断是：这种好日子结束了。为什么？因为我们撞上了两堵墙。

第一堵墙是“数据枯竭”。互联网上高质量的人类文本，基本上已经被现在的模型“吃干抹净”了。想继续靠堆数据来提升智力，就像是在贫矿里淘金，投入产出比急剧下降。

第二堵墙是“边际效应递减”。苏茨克维反问了一个直击灵魂的问题：“当模型规模已经如此庞大时，你再投入100倍的计算资源，真的能带来质的飞跃吗？”答案令人沮丧。

但这并不意味着AI完了，而是意味着游戏规则变了。我们正式进入了“研究时代（The Age of Discovery）”。在新的时代里，拼的不再是谁的GPU多，而是谁能找到那个更聪明、更本质的算法“新配方”。

02. 高分低能的悖论：困在“氛围编程”里的做题家

为了解释为什么我们需要新配方，苏茨克维不仅吐槽了现在的AI，还发明了一个非常精准的词：“氛围编程（Vibe Coding）”。

现在的顶级大模型（LLM）像极了一位“满级做题家”。你给它出一道奥数题，它能秒解；你让它写一篇关于量子力学的论文，它能引经据典。但在实际工作中，比如编程，它却表现得像个“糊涂蛋”。

苏茨克维描述了一个让所有程序员都感同身受的场景：

“你让AI修复一个Bug，它非常自信地改了，结果导致了一个新Bug。你指出这个问题，它又非常诚恳地道歉并修改，结果……它把最开始那个Bug又带回来了。”

这就是苏茨克维所谓的“能力参差（Jagged Capability）”：在某些测试集上，AI的表现早已超越人类；但在很多现实的、需要连续逻辑推理的场景中，它的可靠性甚至不如一个实习生。问题的根源在于“泛化能力”的缺失。

苏茨克维用了一个极其扎心的对比：一个人类青少年，哪怕没什么天赋，练习开车10到20个小时也就学会了。而我们的AI呢？它像一只贪婪的“数据貔貅”，吞噬了全人类产生的所有驾驶视频和数据，却依然可能在遇到一个没见过的路况时瞬间“宕机”。

现在的AI是靠“背诵”海量样本来伪装智能，而人类是靠“理解”底层逻辑来举一反三。这中间的鸿沟，就是AGI必须跨越的天堑。

03. 寻找“机器直觉”：Value Function才是核心

那么，人类这种“举一反三”的能力究竟从何而来？苏茨克维给出的答案出人意料地带有浓厚的生物学色彩：价值函数（Value Function），或者说，一种内在的“感觉”。

为了解释这个硬核的机器学习概念，苏茨克维讲了一个关于脑损伤患者的真实案例。

有一位曾经非常聪明的会计师，因为脑部损伤失去了情绪中枢，虽然他的智商毫无受损，记忆力超群，逻辑运算完美，但他的人生却崩溃了。为什么？因为他无法做决定。仅仅是早上“穿哪双袜子”这个问题，他就能盯着衣柜纠结好几个小时，列出无数种利弊，却永远无法选定其中一双。

苏茨克维指出，情绪（Emotion）和感觉，其实是人类大脑为了在这个复杂世界中高效生存，而进化出的一套“超级压缩算法”。它对应到我们生物大脑里，就是“价值函数”。现在的AI训练（比如强化学习），往往是“结果导向”的：只有当模型跑完整场马拉松，我们才告诉它成绩好不好。这效率太低了！

而人类的“价值函数”，是一个随时随地都在耳边低语的“导师”。当你在这个路口刚想左转，你的“直觉”就会告诉你：“感觉不对，这路有点阴森。”这种对“过程”的实时价值评估能力，才是人类智能极其高效、鲁棒的核心秘密。

苏茨克维认为，下一代AI的突破点，就在于如何让机器学会这种“直觉”。

04. 告别同质化：让AI学会“左右互搏”

除了“缺心眼”（没直觉），现在的AI还有一个大毛病：千篇一律。

你有没有发现，无论是OpenAI、Claude还是Google的模型，它们的回答风格、甚至犯错的方式都越来越像？苏茨克维一针见血地指出：“因为大家都在用同样的数据集做预训练。”

这种同质化是危险的，它导致所有模型都可能会在同一个坑里

跌倒。为了打破这个僵局，苏茨克维提出了一个源自AlphaGo时期的经典思路：自我博弈（Self-Play），但这次是升级版。

不仅仅是下棋，苏茨克维构想的是一种“对抗性辩论”。

想象一下，我们不直接训练一个模型，而是训练两个。一个充当“辩手”，提出观点；另一个充当“裁判”或“挑刺者”，专门寻找逻辑漏洞。甚至可以让两个AI针对一个问题进行激辩。

在这种“左右互搏”的高压环境下，模型被迫跳出死记硬背的舒适区，去寻找更深层的逻辑支点。苏茨克维认为，只有通过这种激烈的内部竞争，AI才能进化出独特的“个性”和真正的创造力，而不是只会当一个“平庸的打工仔”。

05. SSI的野望：打造“15岁的超级少年”

带着这些极具颠覆性的思考（研究时代、价值函数、自我博弈），苏茨克维创立了新公司SSI（Safe Superintelligence）。

这就解释了为什么SSI如此神秘且自信。当被问及“既然你们不搞产品，钱够烧吗？”时，苏茨克维淡定地表示：“我们的计算资源一点都不少。”

区别在于，别的公司把钱花在为了服务数亿用户而搭建的庞大推理服务器上，还要养活数千人的产品团队；而SSI把每一分钱、每一张显卡的算力，都砸在了“纯粹的研究（Research Compute）”上。

他们的目标产品，不是一个聊天机器人，而是一个“超级智能的15岁少年”。这又是一个绝妙的比喻。苏茨克维心中的AGI，不是一出厂就全知全能的神。它更像是一个拥有极高智商、极快学习速度、且充满好奇心的天才少年。它还没有读完世界上所有的书，但当你把它扔到一个陌生的环境，它能利用强大的“价值函数”迅速精通这项技能。

关于AGI何时到来，苏茨克维给出了一个令人屏息的时间表：5到20年。这不再是一个遥不可及的科幻概念，而是我们这一代人注定要亲历的历史时刻。

06. 格局打开：从“服务人类”到“关怀生命”

在谈到AI安全与对齐（Alignment）这个终极难题时，苏茨克维的视角从技术层面跃升到了哲学层面，展示了真正的大师格局。

很多公司还在研究如何让AI“听人类的话”、“不伤害人类”，苏茨克维却在思考一个更宏大、更普世的命题：构建一个“关爱感知生命（Sentient Life）”的AI。他认为，仅仅训练AI“效忠人类”是不够稳健的，甚至可能是危险的。因为在未来的宇宙中，除了人类，可能还有海量的AI智能体。如果AI只懂服从指令，却不懂“痛苦”和“快乐”的本质，它依然可能成为冷血的执行者。

相反，一个拥有情感、具备同理心、能够理解并珍视所有“有感知能力的生命”的AI，才是真正安全的。这种基于“大爱”的对齐，比基于“规则”的对齐更容易实现，也更具鲁棒性。

更有趣的是，关于人类在未来的位置，苏茨克维抛出了一个极具赛博朋克色彩的设想：

“人类想要不被边缘化，可能需要通过脑机接口（如Neuralink）与AI融合，成为‘半AI生命体’。”

只有这样，我们才能真正理解超级智能在想什么，并与之实现思维的同步。这不仅是技术的融合，更是文明形态的进化。

07. 结语：给AI科学家的“审美建议”

访谈的最后，苏茨克维分享了他做研究的秘密心法，听起来更像是一位艺术家的独白。

他说，真正突破性的研究，往往符合三个标准：美（Beauty）、简洁（Simplicity）、以及生物学上的合理性（Biological Plausibility）。“如果在实验数据和你的直觉相悖时，是什么支撑你坚持下去？是对‘美’的信念。”

苏茨克维的这番话，标志着AI领域正在发生一场深刻的范式转移。那个靠“堆料”就能赢的旧时代已经落幕了，正如他所言：“想法（Idea）如果不贵，为什么现在没人能拿得出来？”

接下来的5到20年，将是拼认知、拼审美、拼“机器直觉”的新赛场。AGI的倒计时已经开启，你准备好了吗？