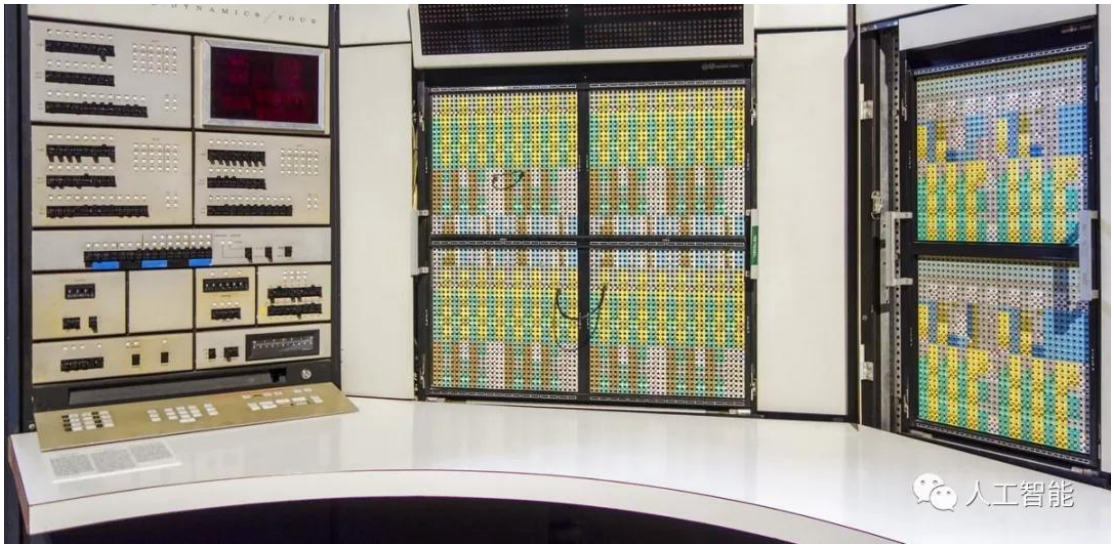
“人工智能第三定律”的漏洞：人类能造出失控的机器人

人工智能3 月 19 日

本文由 George Dyson 改编自企鹅出版社出版的作品《可能的心智：看待人工智能的 25 种方式 POSSIBLE MINDS: Twenty-Five Ways of Looking at AI》（编辑 John Brockman，版权所有 John Brockman），原文标题 The Third Law The future of computing is analog。



图片来源：Arterra/贡献者/Getty Images

以**电子数字计算机的诞生及其生成的代码遍布全球**为界，计算的历史可以划分为“旧约”和“新约”两个阶段。旧约的先驱们提供了基本的逻辑，包括 Thomas Hobbes 和 Gottfried Wilhelm Leibniz。新约的先驱们包括 Alan Turing、John von Neumann、Claude Shannon 和 Norbert Wiener，他们给机器带来智能。

- Alan Turing 曾经在思考如何让机器更加智能。
-
- John von Neumann 想知道机器实现自我复制需要什么。
-
-

Claude Shannon 想知道让机器在干扰中可靠地通信，需要做些什么。

-
-

Norbert Wiener 为机器的控制机能而深思。

- 1949 年，Wiener 第一次警告控制系统可能超出人类掌控，那会儿第一代存储程序电子数字计算机才刚刚问世，这些系统需要人类程序员的直接操作，这削弱了 Wiener 的担忧。既然程序员控制着机器，那么问题会出在哪里？从那以后，关于自主控制风险的争论一直围绕着数字编码机器的控制权和限制的争论。尽管它们拥有惊人的能力，但几乎没有被发现真正的思维。这是一个危险的假设——**如果数字计算正在被其他东西所取代，该怎么办呢？**

模拟计算的悄然回归

过去的一百年里，电子学经历了两个根本的转变：

- **从模拟到数字；**
-
- **从真空管到固体器件。**

- 这两个几乎同时发生的转变并不意味着它们密不可分。就像真空管组件也可以用于实现数字计算一样，模拟计算也可以在固体器件下实现。尽管真空管在商业上已经绝迹，但模拟计算仍然存在。

模拟计算和数字计算之间没有精确的区别。

一般来说，数字计算处理整数、二进制序列、确定性逻辑和离散增量的时间，而模拟计算处理实数、非确定性逻辑和连续函数，包括时间——时间是作为现实世界中的连续体而存在的。

许多系统的运行可跨越模拟和数字计算。比如一棵树，可以说集成了大量的输入，可以被看作是连续函数，但是如果你砍掉这棵树，你会发现它一直在以数字方式计算年份。

在模拟计算中，复杂性存在于网络拓扑结构中，而不是代码中。信息就像电压和相对脉冲频率那样，是被处理为值的连续函数，而不是对离散位串的逻辑运算。

数字计算不能容忍错误或歧义，它依赖于过程中每一步的纠错。

而模拟计算可以容忍错误，允许错误的存在。

自然界使用数字编码来存储、复制和重组核苷酸序列，但自然界的智能和控制依赖于模拟计算，它在神经系统内运行。每个活细胞的遗传系统都是一台存储程序计算机。但大脑不是。

数字计算机执行两种比特之间的转换：表示空间差异的比特和表示时间差异的比特。这两种信息形式之间的序列和结构转换由计算机编程控制，只要计算机还需要程序员，我们就能维持人类的控制权。

模拟计算机也是负责协调两种信息形式之间的转换：空间结构和时间行为。在这里，没有代码，也没有程序。不知何故——我们也不完全理解的原因——自然界进化出神经系统这种模拟计算机，它如此神奇，蕴含了从世界上吸收的信息。它们可以学习，它们学到的内容之一就是控制，它们学会了控制自己的行为，它们学会了尽可能地控制环境。

计算机科学在实现神经网络方面有着悠久的历史——甚至可以追溯到计算机科学出现之前——但在很大程度上，这些都是数字计算机对神经网络的模拟，而不是自然界本身进化出来的神经网络。

如今，事情发生了变化：从底层来看，**无人机、自动驾驶和手机**这三驾马车推动神经形态微处理器的发展，它们直接在硅（和其他潜在的基质）上实现了真正的神经网络，而不是模拟神经网络；从顶层来看，我们最大的及最成功的**企业在其渗透和控制世界的过程中**，越来越多地转向模拟计算。

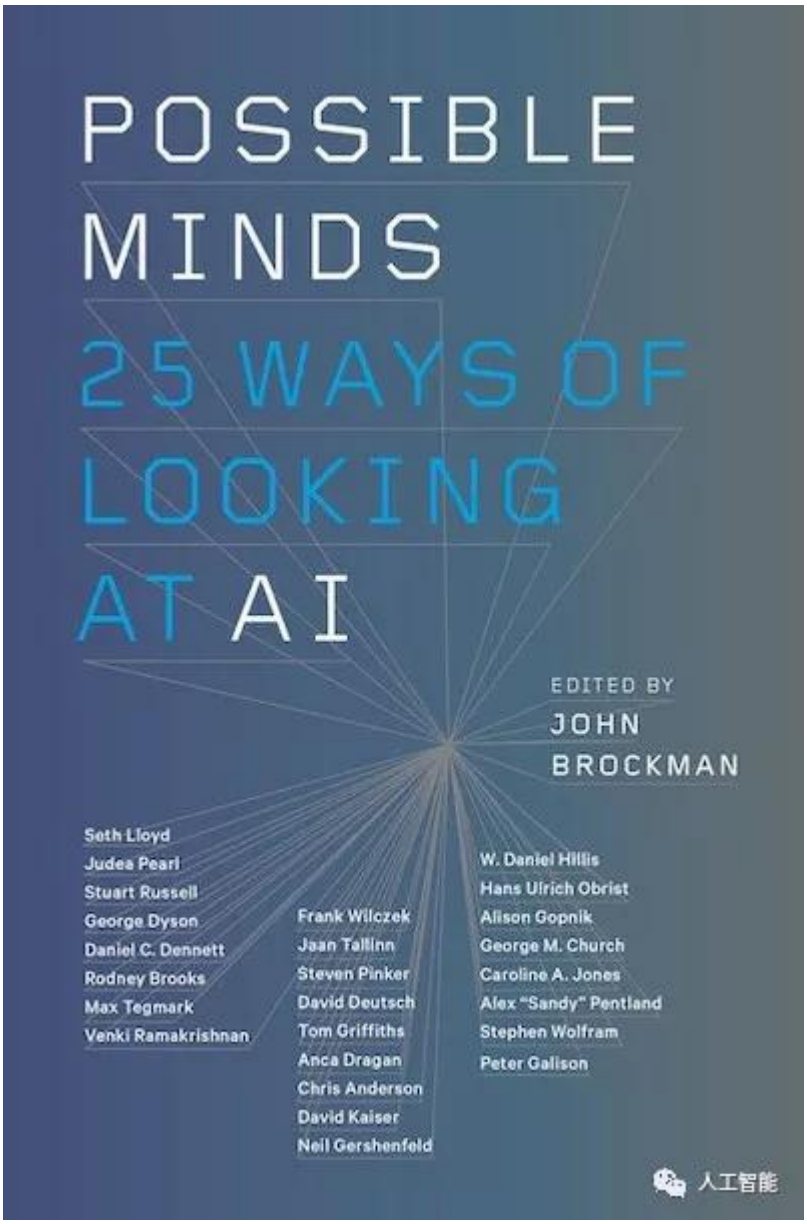
当我们争论数字计算机的智能时，**模拟计算正在悄然取代数字计算**，就像二战后，真空管等模拟元件被重新设计用于制造数字计算机一样。在现实世界中，运行有限代码的独立确定性有限状态处理器正在形成大规模的、不确定的、非有限状态的后生动物有机体。由此产生的模拟/数字混合系统共同处理比特流，就像在真空管中处理电子流一样的方式，而不是像由离散状态设备处理比特流那样单独处理电子流。比特是新的电子。

模拟又回来了，它的本质是承担控制。

这些系统控制着从商品流通到交通流通再到思想流通的一切，它们以统计的方式运行，就像神经元或大脑处理脉冲频率编码的信息时那样。

智能的出现引起了智人的注意，但**我们真正应该担心的是控制的出现。**

建立于现实之上的系统，却反过来控制现实



POSSIBLE MINDS: Twenty-Five Ways of Looking at AI 封面

1958 年的美国人需要保卫美国全境免受空中打击。为了区分敌机，除了依靠计算机网络和早期预警雷达站以外，还需要实时更新所有商业空中交通地图。美国为此建立了 SAGE（半自动地面环境）系统。SAGE 反过来又催生了 Sabre 的诞生，Sabre 是

第一个用于实时预订航班的综合预订系统。Sabre 和它的后代很快就不再是仅有空余座位的地图，而成为了一个系统，它开始通过分散的情报来控制飞机的飞行地点和时间。

但系统里不是有个控制室吗？不是有人正在控制系统吗？可能不是。比如，你开发了一个实时绘制高速公路交通地图的系统，让汽车接入该地图，并报告自己实时的速度和位置。其结果是一个完全分散的控制系统。

系统的控制模型并不存在于任何部位，系统本身就已经是了。

这是 21 世纪的第一个十年，想象一下，你想实时追踪人际关系的复杂性。对于大学规模较小的大学生来说，你可以为他们的社交生活建立一个中央数据库，还可以保持它的实时更新，但如果大学的规模再大一点，它的维护将变得完全不可能。更好的做法是在本地托管简单、免费的半自治代码副本，让社交网络自己更新。这段代码是由数字计算机执行的，但是系统作为整体而执行的模拟计算远远超过了底层代码的复杂性。结果是，其产生的关于社会图景的脉冲频率编码模型最终成为了真正的社会图景。它广泛地在校园和世界各地里传播。

如果你想制造一台机器来捕捉人类已知的一切，这意味着什么呢？你有着摩尔定律的支持，将世界上所有的信息数字化并不需要太长时间。你扫描每一本实体书，收集每一封信件，每 24 小时就可以收集 49 年之久的视频，同时还可以实时追踪人们的位置、当前的行为。但是，你如何理解这些信息的意思呢？

即使是在所有东西都已被数字化的时代，这也不是任何严格的逻辑就能够定义的，因为**人类的意义并不是根本上合乎逻辑的。**一旦你集齐了所有可能的答案，你所能做的最好的事情就是建立一个定义准确的问题，并编写一个脉冲频率加权图来展示所有东西是如何联系起来的。

在你意识到之前，你的系统不仅会观察和映射事物的意义，它**还会开始构建意义**。随着时间的推移，它将控制意义，就像交通地图开始控制交通流量一样，即使看上去，似乎没有人在控制它。

人类难以理解智能

人工智能有三条定律。

第一定律被称为 Ashby 's law，以《大脑设计（Design for a Brain）》一书的作者、控制论专家 W. Ross Ashby 的名字命名。它指出，**任何有效的控制系统都必须与它所控制的系统一样复杂**。

第二定律由 John von Neumann 阐述，他指出：**复杂系统的定义是它构成自己最简单的行为描述**。有机体最简单的完整模型就是有机体本身。试图将系统的行为简化为任何其他形式的描述都会使事情变得更复杂，而不是更简单。

第三条定律指出：**任何简单到可以理解的系统都不会复杂到足以智能地运行，而任何复杂到可以智能地运行的系统，都将复杂到难以理解**。

有些人认为，在我们理解智能之前，我们不必担心机器中出现的超人智能，第三定律可能会让这些人放心一些。

但第三定律存在一个漏洞：**不理解的东西，也完全有可能构建出它来**。你不需要彻底理解大脑是如何工作的，就能构建起一个正常工作的大脑。这是一个漏洞，无论程序员和他们的道德顾问对算法做再多的监督，也无法弥补。

绝对“好”的人工智能是一个神话。我们与真正的人工智能的关系将永远是一个信仰的问题，而不需要证明。

我们过于担心机器的智能，而对自我复制、沟通和控制却担心得远远不够。**数字编程无法控制的模拟系统之兴起将标志着计算领域的新一次革命。**对于那些相信自己能造出机器来控制一切的人，大自然的回应将是让他们造出一台能控制他们自己的机器。