专题探讨

徐贲

作者徐贲为美国加州圣玛利学院英文系教授、多产的公共知识分子，最新著作有“极权三部曲”——《极权下的人性》、《极权下的日常生活》、《极权下的身份焦虑》

中国民主季刊
第4卷 第3期
2026年7月

从AI的局限性到民主研究的新尝试

摘要：民主制度的危机，并不总是来自暴力颠覆或公开否定，更常发生于合法、合规却持续累积的制度性扰动之中。本文尝试引入人工智能研究中的“对抗性样本”与“过拟合”等概念，将其作为分析工具而非修辞比喻，以重新理解现代民主在高度程序化和理性化条件下所呈现的结构性脆弱。正如一个在训练数据上表现优异的算法可能因微小噪声而产生严重误判，民主制度也可能在规则完整、程序合法的前提下逐渐丧失判断能力与实质功能：规则没有被破坏，却被精准利用；制度仍在运转，却已发生功能偏移。本文试图说明，民主面临的关键挑战或许不仅是如何设计更完善的规则，更是如何提高制度的鲁棒性(Robustness)，使其在面对恶意利用时仍能维持判断、自我修正和持续运转的能力。

在机器学习研究中，有一个重要概念叫作“对抗性样本”（Adversarial Examples）。它指的是：只需对一张图像进行极其微小、几乎无法被人察觉的修改，就可能让原本表现优异的人工智能系统作出完全错误的判断。[1] 研究者由此发现，一个系统的危险往往不来自正面摧毁，而来自对其运行规则的精准利用。

将这一发现用于思考民主制度，并非简单的技术比附，而是因为两者在结构上存在某种相似性。民主制度和算法系统一样，都依赖一套规则、程序和基本假设才能正常运作。当这些假设被有意利用时，制度未必会立刻崩溃，却可能逐渐发生功能偏移。规则依然存在，程序依然运转，但其结果已经偏离最初目标。

这种现象并不一定表现为公开的政变或暴力推翻，更常见的是形式合

法、程序合规的策略性操作。例如利用言论自由传播虚假信息，利用经济规则进行政治施压，或者利用程序漏洞谋取制度优势。这些行为并未直接破坏规则，却可能像算法中的“对抗性噪声”一样，持续干扰制度的判断能力。

为了理解这一问题，本文还将借用机器学习中的另一个概念——“过拟合”（Overfitting）。所谓过拟合，是指一个系统过度依赖既有经验，以至于难以应对新的环境变化。[2] 与之相对的“鲁棒性”（Robustness），则是系统在面对干扰、操纵和不确定性时保持稳定运行的能力。[3] 本文试图说明，民主面临的关键挑战或许不仅是如何设计更完善的规则，更是如何提高制度的鲁棒性，使其在面对恶意利用时仍能维持判断、自我修正和持续运转的能力。

这一类比还揭示了民主制度潜在的“过拟合”风险。当一个制度长期在价值共识高度统一、环境相对稳定的“训练集”中运行时，它会不断优化自身，使规则更细化、程序更透明。在已知分布内，这样的优化看似成功，但也容易导致制度僵化。当环境发生偏移，面对高度异质、不对称、甚至恶意的参与者时，这种高度优化的制度可能像一个在标准数据集中表现完美、在现实世界却频频出错的模型。

然而，这种类比同时提醒我们必须警惕结论的边界。对抗性样本的存在并不意味着模型本身应被否定，而是促使研究者去提升模型的鲁棒性。同理，民主制度遭遇操纵，并不意味着它已经失效，而是提醒我们需要引入“抗对抗性”的设计。倘若结论被引向“为了反制攻击，系统必须同样不透明和不可预测”，那么就陷入逻辑陷阱。这既是算

法的误区，也可能使政治制度走向其反面——以牺牲透明与合法性换取短期防御能力，实则是“自我背叛”。

真正的价值，不在于证明规则本身的软弱，而在于揭示：任何过度依赖僵化规则、却丧失主动判断能力的系统，都可能成为对抗性操纵的理想目标。难题在于，如何在坚持透明与规则的前提下，让制度具备识别恶意、保持审慎的“免疫力”，从而不只是“拟合过去”，而能够有效应对未知的恶意干扰。

一 AI 的“对抗性样本 " 撬动传统的民主研究范式

《麻省理工科技评论》（MIT Technology Review）曾报道过一个经典案例。[4] 研究者拍摄了一张猫的照片，对任何人来说，这都是一个毫无歧义的对象：一眼即可识别。然而，当研究者以一种高度计算化的方式，对图片中每一个像素施加极其细微的数值调整后——这些变化小到肉眼完全无法分辨，图像在视觉上几乎没有发生任何变化——结果却令人错愕。当这张“被污染”的照片被输入谷歌的图像识别系统时，系统以接近百分之百的置信度断言：这不是一只猫，而是一台拖拉机。

这一结果并非偶然失误，而是揭示了一个深层悖论：算法的逻辑越严密、模型越复杂、参数越精细，它反而越容易被那种在常识层面完全“不成理由”的微小扰动所击垮。这种理性病理并不只存在于算法之中。当我们审视人类社会中那些看似无懈可击的规则体系时，会发现一个惊人的相似性：最精密的程序正义，往往也最容易被战略性的规则套利所寄生；最透明的制度框架，反而可能为最擅长伪装的行为者

提供最完美的舞台。而当这种脆弱性积累到一定程度，人们就会本能地转向一种更古老、更直觉的解决方案：“恶人自有恶人磨”。

民主理论研究正面临一个需要深刻反思的时刻。传统的民主研究范式，很大程度上建立在启蒙时代以来对理性、透明和程序正义的信念之上：只要规则足够公正、程序足够透明、信息足够公开，民主制度就能产出合理的决策。然而，这套理论框架在二十一世纪遭遇了前所未有的挑战。挑战不仅来自外部竞争，更来自民主制度内部的异化：算法操纵舆论、极化政治侵蚀共识、程序正义被战略性利用为阻挠工具、透明空间反而成为虚假信息扩散的渠道。

2016 年美国总统选举期间，政治咨询公司剑桥分析公司（Cambridge Analytica）利用数千万用户数据进行政治画像和定向传播的事件[5]，引发了全球对民主社会信息操纵能力的担忧。规则并未被公开废除，选举程序也照常进行，但信息环境本身却遭到了精细化干预。此后，围绕社交媒体平台、推荐算法和舆论操控的争议不断出现，显示出民主制度赖以运作的信息基础并不像传统理论设想的那样稳定。[6]

与此同时，一些非民主国家也逐渐学会利用开放社会的制度特点影响民主国家的公共讨论。例如，近年来围绕境外社交媒体的信息战、网络水军传播以及针对特定议题的舆论放大现象，已成为各国情报与安全机构持续关注的问题。其特点并非直接摧毁民主程序，而是利用民主社会开放、透明和自由表达的环境，向其持续输入具有误导性或高度极化的信息。这种做法与机器学习中的“对抗性样本”颇为相似：系统本身仍在正常运行，但其判断过程却被经过精心设计的输入所干扰。

例如，自 2019 年以来，围绕香港问题、台湾问题以及疫情叙事的大规模跨平台信息传播行动，已经成为欧美学界研究“国家支持信息操纵”（state-backed information operations）的重要案例。[7] 其特点并不是说服所有人接受某种观点，而是通过制造噪音、放大分歧、削弱信任来干扰公共判断。对于民主社会而言，这类行为利用的恰恰是其最珍视的开放原则：信息越自由流动，操纵性信息也越容易进入公共讨论空间。换言之，开放本身成为了被利用的条件。

人工智能的发展，为重新审视民主制度提供了一个意外的视角。机器学习研究发现，一个系统即使在正常环境中表现良好，也可能在面对新的环境或刻意设计的干扰时迅速失效。这种现象为理解现代民主面临的困境提供了新的分析语言。与其将民主危机简单理解为民主与专制、自由与秩序之间的对抗，不如将其视为一个制度如何应对操纵、适应变化并维持判断能力的问题。

尤其值得关注的是“过拟合”这一概念。在机器学习中，一个系统过度依赖过去的经验和既有环境，因此在面对新的情况时反而表现不佳。类似的问题也可能出现在政治制度中：许多民主制度是在特定历史条件下发展起来的，其运行往往默认参与者遵守某些基本规范和共同规则。然而，当越来越多行为者开始有意识地利用制度的开放性和程序性优势时，这些原本有效的设计可能反而成为脆弱性的来源。

与之相对应的则是“鲁棒性”——即一个系统在面对干扰、变化和恶意利用时，仍能维持基本功能和判断能力的特性。从鲁棒性角度看，当代民主面临的核心挑战或许不只是如何实现更开放、更透明的治理，

而是如何在保持开放的同时避免被系统性利用。

正是在这样的背景下，本文尝试借助“过拟合”的视角，重新审视“恶人自有恶人磨”这一古老的政治直觉。这句话既表达了人们对失灵规则的失望，也寄托着以更强力量制衡作恶者的期待。然而，它同时包含一个值得警惕的问题：当人们试图用更深的算计对付精于算计的人时，究竟是在修复制度，还是在进一步侵蚀制度赖以存在的规则基础？对这一问题的思考，或许有助于我们更深入地理解民主制度在人工智能时代所面临的挑战与选择。

如果这个尝试能够为当代民主研究提供哪怕一点新的视角，能够帮助我们更清晰地理解制度脆弱性的根源，能够启发我们思考超越简单对抗的第三条道路，那么它就实现了自己的目的。算法的过拟合为我们提供了一面冷酷的镜子，让我们得以重新审视人类世界中那些看似成熟、实则脆弱的博弈结构，以及我们在理性的名义下，究竟走向了何处。

二 民主仍然合法，但是否还有效？

对抗性样本真正令人不安的地方，并不在于系统会“犯错”，而在于：错误并不是通过正面摧毁系统产生的，而是通过精确地利用系统自身的规则、假设和敏感点产生的。民主制度的脆弱性，恰恰也集中在这些“看似合理、甚至是民主自身美德”的细小部位。

首先，一个高度脆弱的地方在于程序正义对结果正义的过度信任。

民主制度往往假定：只要程序是合法的、形式上是公平的，结果就具有正当性。但现实中，程序本身可以被极端策略性地操纵——选区划分、投票门槛、投票时间安排、候选人资格规则、信息披露的时点——这些都不违反程序，却可以系统性地扭曲结果。就像对抗性样本并没有“破坏算法”，而是让算法在完全遵循自身规则的情况下，得出荒谬结论。

第二个极其敏感的部位是对“理性选民”的假设。民主制度在设计时，往往隐含着一个温和而乐观的前提：公民大体上能够在事实基础上形成判断。但在现实中，只需要对信息环境施加极小但持续的扰动——情绪化叙事、碎片化事实、阴谋论的“合理怀疑”包装、算法推荐的微妙倾斜——就足以让公共判断整体偏移。这里的“干扰”并不需要是赤裸裸的谎言，它只需要改变注意力的分布，就像在图像中加一点噪声，系统就再也“看不清”整体。

第三个脆弱点在于权力制衡机制对“善意使用”的依赖。民主制度中的制衡，本质上假定各方会在一定程度上尊重制度精神，而不仅仅是制度表面。一旦某一方开始以“极限合规”的方式行动——凡事只问“是否合法”，而不问“是否破坏制度的功能”——制衡机制反而会变成瘫痪机制。否决权、拖延程序、司法解释空间，这些原本是防止权力滥用的装置，在对抗性使用下，会像被精心设计的输入一样，触发制度的反向行为。

第四个往往被低估的脆弱性，是民主对“例外状态”的容忍。民主制度允许紧急状态、临时授权、非常措施，这是为了应对真实的危机。

但问题在于，“危机”的定义本身极易被操纵。一旦社会被不断维持在一种轻度但持续的紧张状态中，例外就会常态化，而民主并不会在某个明确的节点“被推翻”，而是像模型性能逐步退化一样，慢慢偏离原本的功能区间。

最后，还有一个非常“微小”却致命的地方：民主对冷漠的高度敏感性。民主制度并不是在公民积极参与时最脆弱，而是在多数人疲惫、犬儒、觉得“反正也改变不了什么”时最容易被对抗性利用。低参与率并不违反民主规则，但它会极大地放大少数高度动员群体的影响力。这种变化在统计上看似正常，在制度上完全合法，却会在实质上改变权力结构。

如果对这个类比用一句话来总结，那就是：民主的崩坏往往不是“被攻击致死”，而是在持续、合法、微小的扰动中，被引导着走向一种仍然自称为民主、却已经在功能上变质的状态。

反过来可以问，有没有可能设计一种“对抗性鲁棒”的民主？还是说，民主的开放性本身就注定了它永远暴露在这种脆弱性之中？在机器学习里，一个系统有“鲁棒性”，并不是指它永远不会犯错，而是指：当输入遭遇小幅、噪声级、甚至是敌意的扰动时，系统的整体判断不会发生灾难性偏移。鲁棒性也称稳健性，指事物可以抵御外部应力和影响并维持原有状态的自身性质。鲁棒系统可以承受一定程度的脏数据、异常样本、策略性攻击，它的性能可能下降，但不会被轻易“骗到完全相反的结论”。换句话说，鲁棒性关心的不是完美，而是退化方式：系统是在压力下缓慢变差，还是突然翻车。

把这个概念移植到民主制度中，“对抗性鲁棒的民主”并不是指一种不会被操纵、不会被侵蚀的民主，而是指一种制度结构：即便存在恶意行为者、信息污染、规则套利和公民冷漠，民主仍能在功能上保持基本判断能力，不会因为一系列合法但精确的“小动作”，就迅速滑向实质性的专制。

在这个意义上，我认为，完全“对抗性鲁棒”的民主几乎不可能，但“提高鲁棒性”的民主不仅可能，而且是民主唯一现实的生存策略。

原因恰恰在民主的开放性。民主的开放性意味着三个结构性暴露点。第一，它允许输入来自任何方向，事实、谣言、情绪、恐惧都可以进入公共空间；第二，它承认规则先于动机，只要形式合法，行为就具有正当性外观；第三，它依赖公民持续参与，而参与本身无法被强制。这三点共同构成了民主不可消除的“攻击面”。从这个角度看，民主确实不可能像一个封闭系统那样被彻底加固。

但这并不意味着民主只能在脆弱中等死。机器学习给我们的另一个启示是：鲁棒性不是通过封闭实现的，而是通过冗余、校验、延迟和反思实现的。

对应到民主制度，有几种非常“反直觉”、却高度契合鲁棒性思维的方向。

第一，放弃对单一决策节点的过度信任。高度鲁棒的模型，往往不是“一锤定音”，而是通过多层判断、不同尺度的校验来降低单点失效风

险。对应到民主，就是避免把政治命运压缩到一次选举、一次公投、一次简单多数的瞬间判断中。时间上的延迟、程序上的复议、多机构的异步判断，本质上都是在增加制度的“抗噪声能力”。

第二，承认并显性化“坏行为者”的存在。在鲁棒训练中，模型会被故意暴露在对抗性样本下，而不是假设世界是善意的。民主如果仍然假定权力参与者“总体上是诚实的”，那它在结构上就注定脆弱。更鲁棒的民主，不是靠道德呼吁，而是靠制度默认：有人会钻空子、制造恐慌、操纵规则，于是提前为这些行为设置摩擦成本。

第三，把“判断能力”而非“表达意愿”作为核心资源来保护。一个鲁棒模型的关键不是输出，而是内部表征是否稳定。民主也是如此。如果公共讨论空间被情绪化、极端化、即时化彻底侵蚀，那么再完美的投票机制也只是噪声聚合器。从这个意义上说，媒体生态、教育制度、讨论节奏，反而是民主真正的“防御层”，而不是选举技术本身。

第四，也是最残酷的一点：民主必须对自身的退化保持可感知性。鲁棒系统不会假装自己“正常工作”，它会暴露不确定性、显示置信区间、发出预警信号。而现实中的民主，往往在形式上运转良好时，掩盖了功能性失效。越是“合法、平稳、程序完整”的崩坏，越难被察觉，也越危险。

所以，我们可以认为，民主的开放性确实使它永远暴露在对抗性脆弱之中，但这并不意味着民主注定失败。它意味着民主不能被理解为一种静态制度设计，而必须被理解为一种持续的“鲁棒性工程”——一

种永远假定自己会被攻击、被误导、被利用，因此不断自我修正、自我加固的政治形式。

三 算法的完美表演与致命脆弱

现代深度学习算法并不是在“理解”对象，而是在一个极其高维的空间中，构建精细而复杂的决策边界。这些边界在训练数据上表现得近乎完美，却也因此变得异常脆弱。只要输入略微偏离既有分布，判断就可能突然跨越边界，从“这是猫”跳跃到“这是拖拉机”，中间没有任何渐进的过渡。

这正是所谓的“过拟合”问题。但与其把它理解为一个纯粹的技术缺陷，不如将其视为一种理性形态的结构性后果。过拟合把历史经验当成真理本身，与之同构的现象包括：官僚体系中对流程的迷信、极权环境中对“正确表态”的依赖，技术治理中对指标、排名、基准测试的崇拜。算法的“记住一切细节、却失去理解”，正是人类社会中那种“执行得无懈可击，却对现实毫无感知”的理性病理。

算法并没有走错路，它只是把人类已经走过的那条危险道路推向了极端。为了在既定评价体系中取得最优表现，算法学会了极其精确地贴合过去的经验，记住了所有细节，甚至包括其中的噪声，却因此失去了对对象本身的把握。它并非真正理解了世界，而是把历史数据误认为世界本身。就像一个学生可以完美背诵历年考题，却在题目稍作变化时完全失去应对能力。

这一现象揭示了机器学习领域著名的“没有免费午餐定理”（No Free Lunch Theorem）。[8] 该定理指出，不存在一种能够在所有问题和所有环境中都保持最优表现的算法。一个系统在某些条件下表现越出色，往往意味着它对这些条件的依赖也越强；而这种成功本身，恰恰可能成为其脆弱性的来源。换言之，任何针对特定环境优化到极致的方案，都不可避免地要以牺牲其他环境中的适应能力为代价。

从这个角度看，算法在基准测试中的“完美表现”并不意味着真正的稳健。相反，它有时更像一种危险的幻觉：人们误以为系统已经掌握了世界的规律，而实际上它只是高度适应了某一种特定环境。一旦环境发生变化，或者遭遇超出预期的输入，原本令人惊叹的成功便可能迅速转化为彻底的失效。算法的教训在于，最大的风险往往不是能力不足，而是把局部成功误认为普遍真理。

从更广泛的意义上说，“没有免费午餐定理”不仅适用于算法，也适用于制度、组织乃至文明本身。任何一种规则体系，都不可能同时在所有情境下达到最优；它所获得的优势，往往也是其未来脆弱性的来源。

这种理性病理并不只存在于算法之中。在现实世界里，当规则、程序和指标被视为判断的替代物时，类似的脆弱性同样会显现。一个高度透明、强调程序正义的制度，内部看似无懈可击，却可能在面对刻意利用规则灰色地带、擅长战略性伪装的对手时显得迟钝而被动。规则本身并未被违反，但规则赖以运作的前提——对诚实参与的默认假设——却被系统性地寄生和消耗。结果是，越是严格遵守程序的一方，越容易陷入一种“诚实者的困局”。

问题的核心在于，任何明确的规则体系都必然存在一个根本性的张力：规则越是精确、透明、可预测,就越容易被那些将规则本身当作博弈对象的行为者所利用。这不是规则设计的失误,而是形式理性的内在悖论。一个完美透明的系统,恰恰为战略性行为提供了完美的可计算性。当所有参与者都知道规则的每一个细节,知道执行的每一个步骤，那些不受诚信约束的行为者便获得了巨大的优势——他们可以精确计算如何在技术上遵守规则，同时在实质上完全违背规则的精神。

这种现象在国际关系中尤为明显。某些国家可以一边签署国际公约,一边通过巧妙的法律解释、技术性的规避手段、或者利用执行机制的软弱性,将承诺变成一纸空文。他们精通如何在形式上满足所有要求,同时在实质上追求完全相反的目标。更关键的是,这种行为往往很难被明确指控为“违规”,因为它恰恰利用了规则体系对程序正义的承诺——任何指控都必须提供确凿证据,任何惩罚都必须经过漫长的程序,而在这个过程中,违规行为早已完成,利益早已到手。

正是在这样的背景下,“恶人自有恶人磨”这一逻辑显得格外诱人。当一个长期依赖规则套利、信息不对称和战略装傻的行为者,突然遇到一个拒绝按既定套路出牌的对手时,原有的机制似乎会失效。这种吸引力是可以理解的:它承诺了一种迟来的清算,一种对长期失衡的纠正,一种让擅长玩弄规则的人也尝到被玩弄滋味的痛快。

四、“磨恶人”的多重面相：从规则悬置到高阶博弈

然而，“恶人自有恶人磨”这一现象远比简单的“以暴制暴”或“以不可预测性对抗不可预测性”复杂得多。在现实的国际政治和社会博弈中，真正有效的制衡往往不是抛弃规则，而是调用更高阶、更复杂的规则系统，或者利用多重博弈层次的相互制约。

考察历史上那些成功“磨”掉恶人的案例，会发现它们通常不是单纯依靠蛮力或不可预测性，而是巧妙地组合了多种机制。有些依靠的是国际声誉的长期积累——当一个行为者长期违背承诺、滥用规则时，其声誉资本会逐渐耗尽，最终导致其他国家形成默契的排斥联盟。有些利用的是经济相互依赖的不对称性——通过精心设计的制裁和激励措施，让违规行为的成本远远超过收益。还有些诉诸的是更根本的规范性压力——当某种行为被成功框定为违反人类共同价值时，即便没有明确的法律约束，行为者也会面临巨大的道德和政治成本。

以冷战后期苏联的解体为例。西方阵营的胜利，并非主要依靠军事威胁或不可预测的冒险主义，而是通过一个复杂的多层系统：意识形态上的持续竞争、经济上的技术封锁与消耗战、外交上的联盟网络建设、以及对东欧国家民间社会的长期支持。这是一种“元规则”层面的博弈——不是在具体的条约和协议中针锋相对，而是在更高层次上重新定义什么是可接受的行为、什么是合法性的来源。当苏联发现自己不仅在经济上难以为继，更在道德叙事上完全失去说服力时，体制的崩溃便成为必然。

同样，欧盟对成员国的约束机制，也展现了一种复杂的多层博弈结构。当某个成员国的政府试图违背民主原则或法治精神时，欧盟不仅可以

启动正式的制裁程序，更可以通过暂停资金拨付、降低投票权重、动员其他成员国的舆论压力等手段，形成一个立体的制约网络。这种机制的有效性，恰恰在于它不依赖单一的强制力，而是通过多重关系的相互嵌套，让任何单方面的违规行为都会触发多个层面的反应。

更值得注意的是，许多成功的制衡案例都利用了“退出选项”(exit option)[9] 的威慑力量。当一个行为者过度依赖某个关系网络时，其他参与者威胁退出本身就成为一种强大的约束。这不需要诉诸暴力或不可预测性，只需要清晰地展示：如果你继续滥用规则，我们可以选择离开这个游戏，建立一个你无法进入的新游戏。这种威胁之所以有效，恰恰因为它保留了理性和可预测性——它不是疯狂的赌博，而是经过计算的战略选择。

这些案例揭示了一个关键事实：“磨恶人”并不必然意味着放弃规则、拥抱混乱。相反，最有效的制衡往往来自于规则的分层和冗余：当一层规则被利用或规避时，还有更高层次的规范、更广泛的联盟、更长期的声誉机制可以调用。这是一种“元博弈”——不是在具体的规则内部争夺优势，而是关于哪套规则应当适用、由谁来解释、在什么范围内有效的更高层次的竞争。

当然，这种多层博弈也常常伴随道德上的灰色地带。为了维护更高层次的规则，有时不得不在较低层次上容忍某种程度的违规；为了孤立一个破坏性行为者，有时不得不与另一些不那么可靠的盟友合作；为了维持长期的声誉，有时不得不在短期内承受不公平的指责。这些困境并不意味着规则体系本身是虚伪的，而是说明了一个根本性的现

实：任何规则系统都不可能在所有层次上同时达到完美的一致性，维护规则本身就是一个需要判断、权衡和战略选择的过程。

五 算法与人类的本质差异：为何类比需要谨慎

将人类社会直接类比为“一个高度过拟合的模型”，虽然具有启发性，却也可能掩盖两者之间的根本差异。算法的过拟合是致命的，恰恰因为它缺乏人类系统所特有的那些“元认知”层面的能力——叙事重构、价值反思、身份认同的转变、以及通过文化和制度演化实现自我修正的可能性。

算法在本质上是一个封闭的优化系统。它接受明确的目标函数，在给定的数据分布上寻找最优解，然后将这个解固化为参数。一旦训练完成，它就失去了重新审视目标本身的能力。当环境发生变化时，过拟合的模型不会意识到“也许我记住的那些模式已经不再适用”，更不会主动质疑“也许我优化的目标函数本身就有问题”。它只能继续执行既有的决策逻辑，直到彻底失效。

人类社会则根本不同。即便在高度僵化的体制中，也始终存在某种元层次的反思能力。人们可以讲述关于自身处境的故事，可以质疑既有规则的合法性，可以想象不同的可能性。这种叙事能力不仅仅是装饰性的，它实际上构成了社会系统自我修正的关键机制。当越来越多的人开始讲述一个新的故事——关于什么是公正的、什么是值得追求的、什么样的社会才是理想的——旧的规则和制度就会逐渐失去其道德根基，即便它们在形式上依然存在。

更关键的是，人类社会具有“预言自我实现”[10]的独特机制。当人们普遍相信某种制度即将崩溃时，这种信念本身就会加速崩溃的到来；反之，当人们相信变革是可能的，这种信念也会创造出原本不存在的机会。这在算法系统中完全不可能发生——一个神经网络不会因为“相信自己能够泛化得更好”就真的泛化得更好，它只能被动地执行训练中固化下来的参数。

历史中那些“良性黑天鹅”[11]事件，往往正是源于这种元认知能力的突然激活。柏林墙的倒塌、南非的和平转型、爱尔兰的和平进程，这些都不是简单的力量对比变化的结果，而是包含了深刻的叙事转变、身份重构、以及对“我们是谁、我们想成为什么”这类根本问题的重新回答。这些转变往往看似偶然，实则是长期积累的文化、道德和政治张力的突然释放。它们无法被简化为一个优化问题，因为它们涉及的是目标函数本身的重新定义。

这也解释了为什么人类社会虽然时常表现出类似过拟合的症状——对既有程序的僵化执行、对历史经验的盲目崇拜、对异常情况的脆弱性——却又能够时不时地通过剧烈的自我转型来避免彻底的崩溃。文化的演化、制度的创新、价值观的更新，这些都是算法所不具备的自我修正机制。它们未必总是及时，未必总是无痛，但它们至少提供了一种可能性：系统可以在不彻底崩溃的情况下，逐渐适应新的现实。

然而，这种乐观的观察也需要一个重要的限定：人类的元认知能力虽然存在，却并不保证一定会被激活。历史上同样充满了那些错失自我修正机会、最终走向崩溃的文明和政体。当权力结构过于僵化、当意

识形态控制过于严密、当既得利益集团过于强大时，即便存在反思和质疑的声音，也可能被系统性地压制。在这种情况下，人类社会确实会表现得越来越像一个过拟合的模型——对既有模式的任何偏离都被视为威胁，所有的创新都被扼杀，直到系统在某个外部冲击下突然崩溃。

这提示我们，算法的镜子之所以冷酷，不仅因为它揭示了人类理性的脆弱性，更因为它警示了一种真实的危险：如果我们持续压制那些使人类社会不同于算法的能力——叙事、反思、想象、质疑——那么我们确实可能退化为一个高度过拟合的系统，失去自我修正的可能性，只能被动地等待下一次崩溃。

六　超越二元对立：鲁棒性工程与制度韧性

当前的分析往往陷入一种二元对立：要么接受规则体系的虚假稳定和结构性脆弱，继续在过拟合的轨道上滑行；要么拥抱“恶人自有恶人磨”的逻辑，以不可预测性对抗不可预测性，最终滑向丛林法则。然而，这种二元框架本身可能就是问题的一部分。在这两个极端之间，实际上存在着一种更为复杂、也更为稳健的第三条道路：不是消除不确定性，而是提高系统吸收不确定性的能力；不是追求完美控制，而是建设能够持续适应变化的制度韧性。

这一思路首先要求我们重新认识效率与稳定性的关系。算法系统往往倾向于追求极致优化，把资源配置到最有效率的状态。但复杂系统的历史反复表明，效率并不等于安全，最优解也不一定是最稳健的解。许多系统恰恰是在高度优化之后变得脆弱，因为它们失去了应对意外

情况所需的缓冲空间。

因此，一个健康的制度体系往往需要刻意保留某种程度的冗余。所谓冗余，并非简单的浪费，而是一种保险机制。这意味着不把所有筹码押在同一套规则、同一个决策中心或同一种解决方案上，而是维持多个平行的、部分重叠的、甚至彼此竞争的子系统。当一种机制失效时，其他机制仍能继续发挥作用；当某种规则被利用时，其他层次的约束能够提供补救。

这种冗余在自然生态系统中几乎无处不在。没有单一物种能够永远垄断资源，也没有单一力量能够彻底控制整个生态。系统的稳定性，往往来自这种看似低效的多样性与重叠性。现代民主制度中的权力分立、联邦结构、多元媒体、独立司法、学术自治等安排，本质上都承担着类似功能。它们增加了决策成本，却也降低了单点失效导致整体崩溃的风险。

仅有冗余还不够。真正具有韧性的系统，还必须具备持续调整自身的能力。规则不能被视为永恒不变的教条，而应当被理解为一种可以不断修正和更新的工具。制度的生命力，不仅来自规则本身，更来自修正规则的能力。

这意味着需要建立适应性治理机制。一个具有适应性的系统，会持续评估自身运行效果，识别哪些规则正在被滥用，哪些程序已经失去原有意义，并通过正式或非正式的程序进行调整。在国际治理领域兴起的各种软法机制[12]——行为准则、同行评议、最佳实践规范——正体

现了这种思路。它们不像正式法律那样依赖强制执行，却能够通过声誉、合作和长期关系形成约束，同时保留更大的调整空间。

与此同时，一个稳健系统还需要实现可预测性的分层配置。在基本原则和程序层面保持高度可预测性，使参与者能够形成稳定预期；但在具体执行和裁量层面保留适度弹性，从而防止行为者通过精密计算实施规则套利。

成熟的法律体系通常遵循这一原则。法律提供清晰的原则框架，却不会预先规定所有可能情形的具体结果。法官的裁量权既提高了规则适用的灵活性，也增加了策略性规避法律的难度。这种安排看似牺牲了部分确定性，却提升了整体系统面对复杂现实的适应能力。

类似思想正在影响新一代人工智能系统的设计。研究者逐渐意识到，单纯提高模型复杂度和预测精度，只会让过拟合风险更加严重。更有前景的方向，是引入分布外检测、不确定性评估、多层验证和元学习机制[13]，使系统能够识别何时应当相信既有模型，何时应当重新学习。结构上的冗余与分层虽然牺牲了部分效率，却换来了面对意外情况时更强的鲁棒性。

这一思路同样有助于重新理解偶然性在历史中的位置。承认偶然性的重要作用，并不意味着接受“一切皆随机”的虚无主义，更不意味着应当主动制造混乱。历史中的许多关键转折往往源于偶然事件：一次刺杀、一场疫情、一位领导人的突然离世，都可能改变历史方向。但偶然性的意义在于提醒我们，复杂系统不存在绝对确定的轨迹，而不

是鼓励人们把不可预测性武器化。

恰恰因为偶然性无法消除，我们才更需要建立能够吸收和利用冲击的制度框架。一个真正稳健的系统，不追求消灭所有波动，而是能够从波动中学习，在冲击中完成调整。偶然性提供契机，但能否将契机转化为积极变化，则取决于系统自身是否具备足够的准备和韧性。

因此，在规则过拟合的虚假稳定与不可预测性的混乱之间，确实存在第三种可能。它承认不确定性的存在，却不向混乱投降；它接受复杂性的现实，却不放弃理性的追求。它所追求的不是完美控制，而是持续适应；不是绝对确定，而是长期韧性。

七　算法时代：重新定义理性

算法的失败为我们提供了一面冷酷的镜子。我们害怕机器学会利用规则漏洞、学会装傻、学会策略性欺骗，却又在现实政治和社会竞争中不断奖励这些行为。当判断让位于博弈技巧，当理解被压缩为胜负计算，理性并不会因此变得更强，反而会越来越像一个高度过拟合的模型：在熟悉环境中表现卓越，在真正的意外面前却毫无防御能力。

但这面镜子带来的并不仅仅是警告。算法之所以能够如此清晰地暴露过拟合的问题，恰恰因为它比人类社会更加透明。算法的失败模式容易被观察、分析和复现，而人类制度中的过拟合则往往隐藏在意识形态、道德修辞和制度合法性的外衣之下。人工智能的发展使这些长期被忽视的问题以前所未有的方式显现出来。

由此产生的一个重要启示是，我们或许需要重新定义理性本身。

长期以来，无论是在技术领域还是制度设计领域，人们都倾向于把理性理解为一种不断优化的能力：更准确的预测、更高的效率、更少的误差、更精确的控制。然而，复杂系统的经验正在不断提醒我们，过度优化本身可能成为脆弱性的来源。一个为了特定环境而被优化到极致的系统，往往会丧失面对未知环境时的适应能力。

因此，成熟的理性不应仅仅追求准确率，而应同时关注鲁棒性；不应仅仅追求效率，而应同时维护韧性；不应仅仅追求一致性，而应同时保留调整和学习的空间。在不确定世界中，适度的不完美有时比完美更可靠，适度的模糊有时比精确更具有生命力。

这种转变不仅适用于人工智能系统，也适用于社会制度。一个好的制度不应被理解为能够自动产出正确答案的完美机器，而应被理解为一个能够容纳多样性、吸收冲击并持续修正自身的生态系统。它允许不同观点共存，允许试错和纠偏，允许在新的现实面前重新审视旧的规则。

对于民主制度而言，这一点尤为重要。传统民主理论往往强调规则和程序的重要性，并相信正确程序能够导向合理结果。然而在算法时代，我们越来越清楚地看到：规则本身也会被学习、被利用、被优化，直至偏离最初目的。因此，一个健康的民主制度不仅需要规则，还需要反思规则的能力；不仅需要程序，还需要修正程序的机制；不仅需要稳定性，还需要持续演化的空间。

从这个意义上说，人工智能带给民主研究的最大启示，或许不是如何设计更完美的制度，而是如何建构更具韧性的制度；不是如何寻找唯一正确的答案，而是如何维持持续学习的能力；不是如何实现最终优化，而是如何避免陷入危险的过拟合。

真正的问题从来不在于谁能在一场博弈中战胜谁，而在于系统是否仍然保留着判断、修正和重新理解自身的能力。如果所有竞争最终都演变为不可预测性之间的相互消耗，那么主导未来的将不再是理性，而是失控的偶然性。

相反，如果我们能够从算法的镜子中学会谦逊，承认知识的有限性和世界的复杂性，并据此建立更具韧性、更能学习、更善于修正的制度与文化，那么不确定性便不必成为威胁，而能够成为更新和成长的动力。

也许，这才是理性在人工智能时代最值得追求的形态：不是征服不确定性，而是在不确定性中保持学习；不是追求永恒正确，而是保持持续修正的能力；不是赢得某一次博弈，而是让共同体始终保有继续博弈、继续学习、继续进化的可能。

注释 ……

1 关于“对抗样本”概念，参见古德费洛、施伦斯、塞格迪：《解释并利用对抗样本》[Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR).]

2 关于“过拟合”概念，参见：杰曼、比嫩施托克、杜萨：《神经网络与偏差 / 方差困境》[Geman, S., Bienenstock, E., & Doursat, R. (1992). *Neural networks and the bias/variance dilemma*. *Neural Computation, 4*(1), 1–58. https://doi.org/10.1162/neco.1992.4.1.1]；斯里瓦斯塔瓦、辛顿、克里热夫斯基、苏茨克维尔、萨拉赫丁诺夫：《Dropout：一种防止神经网络过拟合的简单方法》(Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15, 1929–1958.)

3 参见马德里、马克洛夫、施密特、齐普拉斯、弗拉杜：《面向抗对抗攻击的深度学习模型》[Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. International Conference on Learning Representations (ICLR).]

4 威尔·奈特：《人工智能核心的黑暗秘密》[Will Knight (2017). The Dark Secret at the Heart of AI. *MIT Technology Review*. https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/]

5 参见艾萨克、汉娜：《用户数据隐私：脸书、剑桥分析公司与隐私保护》[Isaak, J., & Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer*, 51(8), 56–59. https://doi.org/10.1109/MC.2018.3191268]；事件之披露另见卡德瓦拉德、格雷厄姆 - 哈里森：《揭秘：五千万脸书用户资料被收集用于剑桥分析》[Cadwalladr, C., & Graham-Harrison, E. (2018, March 17). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election]

6 参见塔克、格斯、巴贝拉、瓦卡里、西格尔、萨诺维奇、斯图卡尔、尼汉：《社交媒体、政治极化与政治虚假信息：科学文献综述》[Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. Political Polarization Research Program, Social Science Research Council.]

7 参见杜贝：《理解中国共产党的信息行动》[Dube, R. (2021). Understanding the Communist Party of China's information operations. arXiv. https://doi.org/10.48550/arXiv.2107.05602]；余宗基（Ming-Te Chris Yu）、何嘉耀

（Ho，K.）：《新冠疫情与台湾的认知战》[Yu, M. T.-C., & Ho, K. (2023). COVID and cognitive warfare in Taiwan. *Journal of Asian and African Studies, 58*(2), 249–273. https://doi.org/10.1177/00219096221137665]

8 参见沃尔珀特、麦克里迪：《最优化中的没有免费午餐定理》[Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems for Optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67–82. https://doi.org/10.1109/4235.585893]；其监督学习版本另见沃尔珀特：《学习算法之间并无先验差异》[Wolpert, D. H. (1996). The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, 8(7), 1341–1390.]

9 “退出选项”（exit）出自阿尔伯特·赫希曼：《退出、呼吁与忠诚：对企业、组织和国家衰退的回应》[Hirschman, A. O. (1970). *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Harvard University Press.]

10 “预言自我实现”（self-fulfilling prophecy）概念由社会学家罗伯特·默顿提出，参见罗伯特·默顿：《自我实现的预言》[Merton, R. K. (1948). The Self-Fulfilling Prophecy. *The Antioch Review*, 8(2), 193–210.]

11 “黑天鹅”概念出自纳西姆·尼古拉斯·塔勒布：《黑天鹅：如何应对不可知的未来》[Taleb, N. N. (2007). *The Black Swan*: *The Impact of the Highly Improbable*. Random House.]

12 关于国际治理中的“软法”（soft law），参见阿博特、斯奈达尔：《国际治理中的硬法与软法》[Abbott, K. W., & Snidal, D. (2000). Hard and Soft Law in International Governance. *International Organization*, 54(3), 421–456.]

13 分别参见（分布外检测）亨德里克斯、金佩尔：《检测误分类与分布外样本的基线方法》；[Hendrycks, D., & Gimpel, K. (2017). A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. International Conference on Learning Representations (ICLR)]；（不确定性估计）加尔、加赫拉马尼：《作为贝叶斯近似的 Dropout：在深度学习中表征模型不确定性》[Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. International Conference on Machine Learning (ICML)]；（元学习）芬恩、阿比尔、莱文：《模型无关的元学习：面向深度网络的快速适应》[Finn, C., Abbeel, P., & Levine, S. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. International Conference on Machine Learning (ICML)]