arXiv:2305.02561v1 [cs.CY] 4 May 2023

人工智能开发动态中的善行信号

Sarita Rosenstock计算与信息系统学
院墨尔本大学

2023年5月5日

摘要

本文旨在阐释并构建一个框架，用以理解围绕人工智能开发的社会技术系统如何与社会福利相互作用。本文引入了演化博弈论中的“信号传递”概念，并展示了这一概念如何丰富现有关于人工智能系统评估与治理的理论和实践。

1 引言

将某事视为信号，就是关注它在信息的社会动态中所发挥的作用（Skyrms, 2010）。也就是说，我们关注的并非令牌的指称意义，而是共同体如何利用该令牌进行交流，以及由此产生的意义。本文通过信号视角，阐释并探讨如何考察人工智能开发周围的社会技术动态。具体而言，我旨在说明这一视角可以帮助我们:

- 识别人工智能开发过程中发出的信号可能或确实发挥的功能性作用，了解这些作用如何影响该领域的发展轨迹，使其走向更有利或较不有利于社会的结果
- 更准确地阐明“人工智能政策干预”的目标（这类干预往往表现为要求或请求发出特定类型的信号），更好地评估给定策略在多大程度上实现了该目标，并指出一些值得关注的新颖策略方向。

首先，我将先解释我所说的“善行信号传递”是什么意思。接下来，我将讨论为什么我认为，对于那些有兴趣推动有益人工智能发展的人而言，这是一个富有成效的研究方向。随后，我将勾勒出我希望围绕这一概念构建的理论框架的一些核心组成部分，最后概述本文其余部分的结构。

1.1 善行信号

我所说的“善行信号”是指人工智能开发者向彼此、用户、监管机构和公众表达其技术及开发过程有助于实现有益于社会的成果的各种方式。

这类明确的信号包括公开声明和新闻稿。但信号可以是任何可观察到的输出（“令牌”），它携带有关个人、公司或产品的信息，可用于形成对该实体的信念并指导与之相关的行为。这包括他们说了什么和没说什么、明显做了什么和没做什么，以及他们选择如何表达和行动。相关信号的例子包括对伦理承诺的相对明确的声明，例如《人工智能伦理原则》、同意接受独立审计、各种“公平性”指标的得分，以及详述危害缓解策略的新闻稿。但公司无法决定自己发出的哪些信号会传递具有伦理意义的信息——它们的每一个可见输出和行动都可能成为信号，供他人评估，并据此调整自身行为。

明确承诺以合乎伦理的方式行事，本身并不能、也不可能成为伦理意图的信号。要发挥这种作用，这些承诺就必须与动态社会背景中发出信号的行为者的实际伦理行为稳固一致。这意味着，若要可靠地发送或接收某事物的信号，就需要一个关于这一传播社会情境的有效模型，甚至可能需要对其中的动态过程施加干预。

1.2 动机

促进安全且有益的人工智能开发的传统方法大多是“自上而下”的，以治理为核心。这些方法难以跟上技术的快速发展，导致由此产生的法律要求要么很快就过时，要么在技术层面不够具体，无法构成真正的保障。现有的“自下而上”方法往往侧重于公司如何单方面让其技术更有益于社会，却没有关注可能促使公司采纳这些建议的实际激励因素。本文属于少数采取系统层面视角的人工智能伦理研究，将这两种方法联系起来并纳入其中。[1]

信号框架的出发点是完整的社会技术系统——包括公司、个人员工、政府、大学、智库等。政府和非政府组织对透明度和问责制的呼吁，会作为信号在这一系统中传播，并影响人工智能相关的社会和经济行为；公司自身可见的产出也会产生类似影响。信号框架的基本洞见是，信号的含义取决于它在这一系统中发挥的作用，即分享信息和激励行为。这一视角有助于监管机构识别

[1]与该领域现有研究的联系将在第 3.1 节讨论。

与公司稳健的亲社会行为保证密切相关的信号。

信号视角还让我们关注，这些信号如何通过正反馈循环，在更广泛的层面上促进更好的结果，而不只是着眼于评估和激励单个公司。例如，可以将有益人工智能发展视为一种集体行动问题：即使怀有良好意图的行业参与者，也无力承担采用更好实践所需的成本（Askell 等人，2019）。解决这一问题的一类方案，可以借鉴经济学和进化生物学中的研究：合作意图的信号如何稳定高风险的合作，并使亲社会策略在竞争环境中胜出（Rosenstock 和 O'Connor，2018）。

1.3 信号框架的特征

令牌成为有效信号的一个关键特征是，它能被共同体采纳，从而真正符合其预期含义。只有当其他科技公司发出正确类型的信号时，我们才能确认它们正在采取所期望的行为。因此，当我们通过法律或社会要求明确提出某些行为要求，或通过示范行为隐含提出要求时，我们希望这些要求包含或体现为发出特定类型的信号。探索信号具备哪些特征，才能最好地服务于这些目标，并据此加以开发和／或推广，因此是一个很有前景的研究方向。

在更深入地探讨这一框架之前，我想先概述一些贯穿全文都应牢记的关键考量。我将这些考量归入"内容""采纳""视角"和"激励"几个标题之下。

1.3.1 内容

我希望在此作出的理论贡献之一，是为相关信号所涉及的"内容"发展出恰当的概念。第 2 节尝试将有关信号博弈的现有文献整合为一套统一的信号内容理论，使其能够直接适用于人工智能开发过程中产生的、能够帮助促成广泛有益结果的信号。其余大部分理论工作，都可以理解为阐明并应对各种与确保信号能够指向其预期目标内容相关的挑战。

相关考量包括：

- 可伪造性：不执行相关行动而"成功"发送该信号有多困难？

- 语境依赖性：情境如何塑造意义？"同一个"信号如何因情境或视角不同而具有不同含义？

- 动态变化：意义如何随时间发生变化，尤其是如何回应不良激励?

1.3.2 采纳

哪些特征能让信号（及其相关行为）更有可能以正确的方式获得社区采纳？这一框架的一个潜在应用，是探讨怀有亲社会意图的公司如何利用其技术领军者地位，在竞争对手中树立一种规范，即传递能够促成有益行动的特定类型的信号。要实现这一点，仅仅让已识别或设计出的信号与正确的实质性事实和行为相对应是不够的；这种对应关系还必须能够在相关的适合度情境（本例中主要是资本主义情境）中稳健地演化。这表明，病毒式传播和信息流动并非信号选择之后的第二步考量，而是与信号的构成性要素交织在一起，决定了信号的作用及其含义。

注意事项：

- 成本："廉价"信号无需大幅改变行为即可发出，但"昂贵"信号可能与竞争成功不相容。
- 自然性：我们应该寻找技术发展过程中"自然"产生的副产品作为信号，还是应该开发专为特定目的设计的新型信号？前者"更容易"，且推广采纳所需的工作更少，但会大幅限制我们的选择空间。
- 简洁性：一个"良好"的信号需要在何种意义上做到"简洁"？当我们的目标是传递复杂的状态时，这会带来什么问题?

1.3.3 视角

"信号博弈"涉及发送者 和接收者，双方都会根据各自的信念和动机决定传递什么以及如何传递。尽管这看似显而易见，但在当前有关人工智能伦理的大量文献中，对信号由谁发送、为何发送，以及由谁接收、如何接收的明确考量，在很大程度上仍然缺失。

考量：

- 可读性：在某些情况下，信号考量可能促使我们将目标从善行重新界定为可见的善行（Benn 和 Grastien，2021年）。
- 多重性：尤其对于公共信号，同一个令牌如何可能向不同主体传达不同含义？在此类情境中，信号又如何才能有效地发挥多种作用?
- 接收者：思考不同立场的人如何感知信号，例如人工智能从业者、竞争对手、消费者和监管机构。

- 相关发送者：思考哪些“主体”（如有）应被视为信号的“发送者”，例如组织、个人开发者以及人工智能系统本身。

1.3.4 激励

在社会情境中纳入信号传递，会如何影响参与者的个人激励，反之亦然？正如我将在第 3.1 节中详细阐述的那样，从信号视角出发，可以丰富现有的博弈论框架，以便从政策视角理解社会和经济激励。

考量：

- 合作：信号如何帮助促进合作？
- 游戏化：了解信号在激励机制中的作用，会如何削弱信号与内容之间的关联？我们又该如何防范这种情况？

1.4 概要

在第 2 节中，我将更详细地阐述信号框架的博弈论基础。在第 3 节中，我会探讨它与其他概念和研究方向的联系。最后，我将在第 4 节中提炼出一些关键经验。

2 信号如何获得意义

虽然不难承认，信号的内容在很大程度上由情境决定，但我们仍面临一项艰巨任务：解释情境究竟如何决定意义。尽管我怀疑存在一种恰当的一刀切解释或判定程序可供我们依赖，但我们有大量形式化和实证工具，可以用于研究特定的信号系统。我希望这些工具能够逐步完善，成为一套行业专属理论工具包，帮助我们理解有益人工智能的信号如何获得其意义。本节将首次尝试汇集现有博弈论文献中关于信号的研究，以实现这一目标。

2.1 基本模型

不妨先从发送者—接收者博弈这一简单、精简的形式化模型入手，了解这种机制如何运作，再逐步增加复杂性，展示如何调整模型以适应不同情境。[2] 我们从两个

[2]以下对发送者-接收者博弈的描述，主要借鉴了 Skyrms（2010）关于进化生物学情境下信号传递的论述，以及 Franke（2013）的语言学框架；二者均以 Lewis（1969）对发送者-接收者博弈的初步构想为基础。随着我逐步构建一个更适用于人工智能开发情境的论述，我将逐渐偏离这些既有研究。

主体开始：发送者和接收者。他们各自拥有不同的动机和信念，后文将对此作更深入的探讨。为简化起见，发送者或许能够观察到硬币正面或反面朝上，而接收者则无法观察。观察到 H 或 T后，发送者会选择向接收者发送消息m 或 m^* 。接收到其中一个信号后，接收者会选择一种解释 i 或 i^*。

如图1所示，发送者通过考虑其预期接收者会如何回应来决定发送哪种信号——即根据其对接收者如何将 m和 m^* 与 i 和 i^* 对应起来的模型，以及发送者希望引发 i 还是 i^*。哪种解释——i还是 i^*——对接收者更有利，取决于硬币是正面朝上还是反面朝上，而接收者只能通过其对发送者如何将 H 或 T 与 m 或 m^*联系起来的模型进行间接推断。

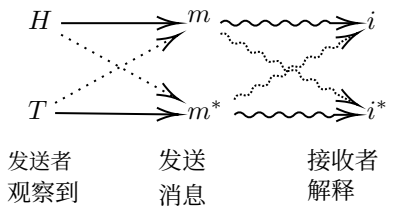

图1： 发送者-接收者博弈的可能解决方案示意图。发送者有两种对消息进行编码的方式，分别用实线和虚线表示。接收者有两种解码消息的方式，分别用实线和虚线波浪线表示。

我们先从一个简单的情况开始：发送者只想让接收者知道硬币正面朝上还是反面朝上的真相。如果两个主体使用同一种语言，并且彼此信任对方诚实，那么很容易看出他们应该怎么做。观察到正面或反面后，发送者会分别发送消息 $m =$“硬币正面朝上”或 $m^* =$ “硬币反面朝上”。接收者会将消息理解为：$i =$ 硬币实际上正面朝上，或 $i^* =$ 硬币实际上反面朝上。

如果发送者和接收者不共享同一种语言，他们仍然必须决定发送和解释消息的策略。发送者的策略是刺激与消息之间的关联。如果没有任何线索表明接收者会如何解释他们的消息，发送者对关联 $H \mapsto m$ and $T \mapsto m^*$与关联 $H \mapsto m^*$ and $T \mapsto m$会同样无所谓。接收者对解释 $[m \mapsto H$ & $m^* \mapsto T]$ and $[m \mapsto T$ & $m^* \mapsto H]$也会同样无所谓。这使我们对信号 m and m^*的“含义”有了两种理解：一种来自发送者的编码，另一种来自接收者的解释。

2.2 从模型到意义

发送者和接收者都需要一种策略，以建立状态与消息之间的这类关联。由于我们从一个简化假

设入手，认为发送者和接收者都以传达准确信息这一共同目标为动机，因此，只要其中一方能够获得一些有关沟通效果的反馈，他们最终就会形成一种共享语言 $[m \leftrightarrow H \ \& \ m^* \leftrightarrow T]$ or $[m \leftrightarrow T \ \& \ m^* \leftrightarrow H]$。这种反馈可能表现为根据新信息进行有意的策略性适应，也可能表现为一致的沟通者胜过不一致的沟通者所带来的选择效应。[3]

在这样的企业语境中，人们很容易只使用前一种“战略性”术语来讨论。毕竟，相关行动者包括具备理性能力的人类（或人类群体），他们能够计算对自身而言的最优行动方案。的确，理性最佳回应是一种强有力的分析视角，而且有一些决策点最好用这些术语来理解，识别出这些决策点也应当是这一努力的核心目标之一。但由于诸多原因，这种做法是错误的。随着我们进一步偏离基本模型，“理性最佳回应”这一概念会逐渐失去意义，因为计算所涉及的相互作用的参数越来越多，且具有高度变异性；真正理性的行动者唯一合理的做法，就是关注进化动力学中更加整体性的特征，而那一次孤立的互动正是在这种进化动力学中发生的。此外，将信号策略设想为纯粹的战略行为，可能会导致人们不恰当地把信号内容与发送者意图联系起来。如果信号是有意为之的，那么“误导性”信号就是对抗性的，因此质疑其意义就等于指责发送者怀有恶意。但信号可能出于许多原因偏离发送者意图——发送者甚至可能根本没有与某个信号相关联的意图——我们应当警惕陷入这一陷阱。

因此，为了理解初始的交流“问题”如何以一种将信号与含义关联起来的稳定语言形式转化为“解决方案”，我们有必要同时考察这两种机制——个人的策略性行为和演化选择。为了了解它们如何运作，我们首先将“基本模型”定义为一种博弈：每当接收者能够正确判断硬币落地时是正面还是反面，发送者和接收者都会获得固定的“奖励”。

如果双方行动者都具备足够的推理能力，他们几乎肯定能较快地找到一种正确的语言。如果第一条消息（m）纯属偶然地促成了正确的关联（H），那么发送者和接收者都会受到共同奖励的激励，在未来的互动中继续沿用这一关联。如果一开始关联未能奏效，他们就会继续尝试，直到找到可行的办法。值得注意的是，这里可能出现恶性循环：双方因失败而不断“交换”关联，却始终无法成功。在这个简单的博弈中，博弈参与者只要在信号选择或解释策略上进行任何程度的试验，就能轻易打破这一循环，除非他们

[3]围绕这一主题的许多用语，例如“选择”“演化”“适合度”，都显露出演化博弈论源自生态学和生物学。就我们的目的而言，这在一定程度上是一种隐喻，其中选择压力主要来自经济和文化因素，不过人们也在讨论这一类比是否恰当（Lewens，2015）。

特别倒霉，以至于所有这类尝试都恰好以令人绝望的方式同步。不过，这一点仍值得关注，因为理性主体在意图一致时，沟通意外失败的情况往往会因偶然因素而得以避免，尽管主体持续进行策略性试验或许能降低这种偶然因素的影响。

相比之下，演化视角并不要求博弈参与者进行任何推理或制定策略。相反，我们只需将“奖励”理解为“适合度”：奖励较高的主体在生态系统中以某种方式更有可能“存续”。这种博弈论的用法最初用于解释生物演化，但事实证明，演化博弈论对经济学家和社会科学家也很有用，可用于理解哪些公司能在市场上“存活”，或哪些文化实践能在共同体中“延续”。我们可以通过多种方式，以数学形式将这一思想表述为随时间变化的动力学演化。

一种简单而标准的做法是采用复制子动态：一代又一代的行动者会根据上一代获得奖励的情况，复制其策略。我们可以设想，第一代完全随机地选择消息与状态之间的关联。有些会成功，有些会失败，而这将体现为下一代中各种策略的数量有所增减。从技术上说，两种语言有可能以这种方式同时演化——尤其是在动态局部化、主体更可能与邻居互动的情况下。但通常随着时间推移，复制子动态会放大因某种语言略为常见而产生的微小优势，直到共同体最终采用单一语言。

请注意，背景演化动力学的存在与对智能体行为的理性—策略性理解完全相容。我们可以将这些动力学解读为至少部分源自有意的策略选择，也可以设想一种背景动力系统，在其中表现得更具策略性可能带来优势。对于当前情况，这类中间视角似乎最为合适。

2.3 相关机制

对于基本模型，最终形成的语言将是纯粹的约定性，因为消息与内容之间的两种关联都没有更大的可能性或更适合解决意义对齐的问题。导致这种纯粹约定性的设定对称性是一种理想化，通常并不成立——将约定性视为可变的会更准确（O’Connor，2021）。这种对称性可能以多种方式被打破。最显而易见的一种情况是，主体可能掌握有关其对话者的一些信息，从而预期某种关联能实现更好的对齐。

由于在模型中加入了另外两个参数，我们还可以预期出现不对称性：与每条消息 $c(m)$和 $c(m^*)$相关的成本，以及发送者和接收者各自关于该偏差的先验信念。

硬币，表示为概率 $P_S(H)$ 和 $P_R(H)$。如果消息的成本存在差异，且先验概率也存在差异，我们就会预期效用最大化主体会“偏好”（同样，无论是出于策略上的审慎考虑，还是被动的演化过程）这样一种语言：它将成本更高的信号与发生概率更低的结果联系起来。

博弈论互动中的这类破缺对称特征，可以理解为相关机制 （Hamilton, 1964）、焦点 （Schelling，1960），或是对显著性 的运用（Lewis，1969）。尽管我的博弈伙伴可能有多种行动或语言框架可供选择，但如果我们充分了解彼此的动机，就能利用共享观察，独立地得出兼容且最优的解决方案。[4] 上文讨论的相关机制指的是外生环境特征——即我们共享环境中的成本与频率。共享的社会规范和法律也可以发挥相关机制的作用。此外，通过博弈前信号传递，我们可以达成更细致的相互保证行为，从而更容易有效地合作（Santos 等人，2011年）。

2.4 歧义与不确定性

还可以考虑纳入信号保真度 会如何影响我们的分析。观察和/或通信通道中可能存在一些噪声，导致信任和共享语言不足以让接收者确信自己的解释。在这种情况下，或许值得在模型中增加一层，将消息解释 （对发送者发送信号背后意图的评估）与以下环节区分开来：基于其解释进行信念形成，和/或反应（基于解释采取的行动，该行动可能会也可能不会受到信念形成的相关中介）。

此外，请注意，共享语言需要对状态和消息的可能性有共同理解，而这并非不言而喻的假设。由于信号内容是解决这类信号博弈的整体产物，内容不仅取决于发送了哪些消息，也同样取决于哪些消息没有被发送。即使是合作的沟通伙伴，也可能难以应对以这种方式协调预期所带来的额外协调问题。值得探讨的是，在这里只实现部分模型对齐时，沟通能否发生，以及如何发生。此外，如果可供选择的状态多于信号（例如，因为人工智能系统可能以多种有意义的方式满足某项定量公平标准），那么最终形成的任何语言都必须让单个信号对应多种含义。

即使在纯粹的通信合作情形下，这些考量对我们也很重要；但当通信目标可能彼此不同甚至相互冲突时，它们就更加突出了。欺骗性通信伙伴可以通过多种方式将歧义武器化。

[4]关于形式化和证明，请参见 Aumann（1987）。 f.

首先，欺骗者可以利用并夸大其行为所造成伤害的性质和程度方面的不确定性，以逃避批评和监管。烟草业以及石油和天然气行业的历史中，这种“人为制造的疑虑”已有充分记录（Oreskes 和 Conway，2011年）。这些案例可以被解读为有效的宣传运动：它们重新诠释科学研究所产生的“信号”，使其不再那么能表明问题的存在及解决方案的性质。即使不存在反社会行为者或相关意图，专家对政策提案证据基础所表达的不确定性，也可能削弱这些证据作为强有力信号的潜力，而我们正是依靠这样的信号来协调有效的公共行动（Ojea Quintana 等人，2021年）。

恶意散布“疑云”可以被视为滥用话语规范这一更普遍现象的一个具体实例。“会话含义”这一概念源自 Grice（1975年）。他指出，交流往往依赖于合作原则：“在你参与的谈话交流中，应在交流进行到的相应阶段，按照公认的目的或方向，做出所需的谈话贡献。”关键在于，我们所说内容的含义往往不能仅凭说了什么来判断，还要考虑如何表达。散布疑云可以理解为滥用 Grice 的“数量准则”：通常我们会认为，所获信息的多少符合话语的共同目标。当人们着重强调并详细阐述怀疑的理由，却几乎不提相信的理由时，我们会认为，这种比例恰当地反映了说话者对现有证据的理解，也体现了决策所需的适当权衡。因此，散布疑云者可以暗示怀疑的理由比实际情况更多，却从未明确说过任何不真实的话。

科技公司同样可以滥用含义规范，生成具有误导性却并非不真实的信号。例如，它们可能会强调产品在某项指标上的优异性能，同时刻意隐瞒其在相关指标上的糟糕表现。事实上，公司普遍接受这种做法，因此通常不会将其归类为欺骗，而只是视为“商业常态”。然而，公司经常在试图营造合作性的“对话”情境与在方便时退回到更具对抗性的“以自身利益为重的商业”情境之间来回切换。因此，我们或许应当考虑信息操纵理论（McCornack，1992年），以分析公司如何利用预设情境，操纵人们对其发出信号之内容的理解。

2.5 动机的复杂性

正如逐渐显而易见的那样，一旦我们承认主体可能除了有效沟通之外还存在其他动机，事情就会变得复杂得多。要使这一点成为问题，它们的动机不必具有对抗性。例如，一旦我们把发送信号所涉及的成本考虑在内，发送者发送

准确信息的意愿可能会被其避免承担发送特定信号所需成本的动机所掩盖。将完美沟通之外的其他激励纳入考量——而这些激励可以说始终存在——会改变信号最初获得含义的方式，也可能改变既有语言的语用内容。

“误导性”或“欺骗性”信号这一概念乍看之下构成了一个哲学难题。只要信号的“含义”完全由导致信号发出的条件决定，那么从某种意义上说，信号本质上都是诚实的（Skyrms，2010，第74页）。欺骗若要发生，就必须借助一种关于预期用法的背景情境，而这种情境可能因反复剥削而逐渐瓦解。因此，进化动力学和理性学习在一定程度上使我们免受持续、大规模欺骗的影响。就我们的目的而言，特别值得探讨的是欺骗性信号传递得以持续存在的一些方式。

首先，即使存在某些反社会行为，只要能将其控制在不会扩散并席卷整个群体的范围内，默认信号传递诚实且存在合作行为的预期仍可能具有适应性。[5] 因此，即便受骗的主体可能因“信任”某个信号而偶尔遭受惩罚，这也未必会从根本上破坏该信号的一般含义。例如，假如某个不诚实的商贩告诉我，他卖给我的大米有5千克，但实际上称重只有2千克，我不会因此得出结论，认为“5千克”这个标签实际上意味着不到一半的重量；我只会将这个词被破坏的关联归咎于欺骗我的那个商贩。只要此类经历足够罕见，我就可以继续假定食品杂货标签与其所标示的内容存在特定的事实关联。但如果欺骗性标签多到令人无法容忍，我就不宜再期待标签与内容之间存在任何关联，这会破坏标签的所有信号传递功能。

因此，不择手段的商家更可能通过对用量做些微调整而蒙混过关。与公制单位相比，使用“单位”或“蒲式耳”这类含义更模糊的术语，更容易做到这一点。当人们容忍这类不显眼的用法变化时，术语的含义就会随着（微妙的）欺骗性用法而动态演变，就像缩水式通胀现象一样。这说明，即便系统对含义固定的信号遭到广泛欺骗性使用时天然具有上述“免疫力”，某种形式的动态欺骗仍能持续存在。通过军备竞赛的方式，当我适应了你对某个信号的欺骗性使用后，你可能会针对我的调整作出回应，进一步实施欺骗。只要我仍觉得在一定程度上保持信号与状态之间的关联足够有用，持续的欺骗性使用就能继续，而不至于彻底切断信号与含义之间的联系。

不过，何种程度的欺骗才算足够“轻微”，从而能够“通过”抗欺骗机制的筛选，可能取决于其他因素。正如前文所述，这可能

[5]形式上，这可以被刻画为一种伪分离均衡（Huttegger 等，2015，第 1003 页）

是理性的：只要我认为遭遇重大欺骗的可能性足够低，仍预期假定他人诚实能让我受益，我就可以让自己暴露于重大欺骗的风险之中。因此，如果我的信息环境受到损害，而我对遭遇欺骗可能性的评估又不准确，欺骗便可能持续存在。

这种欺骗机制源于信息获取的不对称，以及操纵他人信息获取渠道的可能性——这将是我们在此尤其需要关注的问题。人工智能开发者获取的信息几乎总是比试图解读其信号的用户或其他利益相关者更多，而人工智能能力的特殊性质可能会以前所未有的方式扩大这一差距。

欺骗性信号传递之所以持续存在的另一个解释是，信号往往出现在涉及多重博弈论互动的情境中。一名具有欺骗性的销售人员错误使用重量标签，并不会改变我对公制单位使用方式的理解，因为在更广泛的社会情境中，公制单位广泛用于直截了当的情境，在这些情境下，合理地假设“基本模型”成立——双方通常都 主要出于合作共享信息的动机。我可以保持健康的怀疑态度，不想当然地认为基本模型始终适用，但尽管如此，仍可能错误判断某次互动属于哪种情形。通过这种方式，人工智能开发者可能巧妙地（甚至无意中）通过学术或政府传播渠道“洗白”信号，从而给事实上可能具有对抗性的利润动机披上一层合作性认识论动机的外衣。这对人工智能而言尤其成问题，因为知识的前沿积累于企业环境中，造成知识不对称，而这可能助长认识论依赖。

总之，即使凭借 强大的自我纠正动态，欺骗性信号传递几乎成了自相矛盾之举，个人主体仍可能利用其对特定信号的情境及其后果所拥有的持续知识优势，有意或无意地不断欺骗。

2.6 成本、可伪造性与承诺装置

前述讨论让“欺骗性信号传递”这一现象听起来比实际情况更为险恶，也更像是蓄意为之。重要的是要记住，主体主要对自身的适应度情境作出反应，是很自然的；它们并不总是会优先遵从某种更高尚的使命，去“如实沟通”。在考量人工智能开发者有意传递善行信号（例如采纳伦理原则）时，这一点尤其值得牢记。这些信号可以放在多种不同的信号博弈中加以理解。如果将其视为提出这些主张的特定个人对事实意图的陈述，那么它们实际上可能是完全“诚实”的沟通行为。但这与它们作为公司对预期企业行为所发出的信号毫无用处，并不矛盾。如果公司内部以及其所处的经济和社会

情境中缺乏适当的动态约束，那么作者意图传达的信号与文件发布实际上发出的信号，很可能并不完全一致。

是否有办法区分两个声称秉持相同伦理承诺的主体：一个是真诚地报告自己打算采取的行为，另一个则只是说出那些能够赢得客户和监管机构好感的场面话？简短的回答是：我们无法做到；“意图”并不是能够被检测到的事物。

我们可以采取的一种做法是，在可能的情况下，寻找、设计、鼓励并倡导那些“不可伪造”的信号——也就是说，这类信号的发送行为本身（若要成功且具有说服力）就具有相当高的难度，甚至在相关事态并不成立时根本无法发出。可伪造性通常存在于一个连续谱上，不同信号与其所传递的内容之间存在不同程度的内在联系。在生物学语境中，汤氏瞪羚的“弹跳示警”常被视为一种相对不可伪造的信号：当它们看到捕食者时，会高高跃起，以此表明自己行动敏捷、难以追捕。相比之下，有毒蛙类会利用鲜艳体色向捕食者传递自己不宜捕食的信号，而这种信号却可以被无毒蛙类“伪造”，后者借此获得信号传递方面的优势（同时也削弱了该信号在影响捕食者预期方面的有效性）。

在理论上信号可以被伪造时，让主体发送信号需要付出“成本”，可以弥补这一缺口，使信号仍然可行，前提是这种成本不会给诚实的信号发送者造成过重负担。Huttegger 等人（2015）正确指出，“成本”与“可伪造性”之间的区别并非泾渭分明；从适合度的角度看，瞪羚的弹跳示警是有成本的，只不过这种成本更容易被理解为在躲避捕食者时已“纳入预算”的支出，因为无论如何，躲避捕食者都需要爆发性的速度和敏捷性；相比之下，体色则更像是一种“人为”施加的成本。同样，在经济语境中，我们往往将税收和补贴视为成本，而不是将声誉和关系改善行动带来的更广泛影响视为成本，尽管这些因素在经营成本和利润方面最终也可能产生相同的影响。

这里的道德启示是，“代价高昂性”和“不可伪造性”最好被理解为两种不同的视角。它们在使用时可能具有不同程度的直观性，但在行为激励方面发挥着相同的作用。Rosenstock 和 O’Connor（2018）通过一个数学模型证明了这一点。该模型表明，在一种重复囚徒困境博弈中，博弈参与者可以在背叛时发送一道“道歉”信号，以表明未来合作的意图。在这种情况下，代价和可伪造性可以互换使用，从而使合作行为即使面临欺骗性道歉者的威胁，也能够逐渐演化。重要的是，这些研究结果还依赖于与相同参与者进行足够多次的重复互动，使信号能够有效地影响策略选择。此外，这些结果还依赖于某种“文化起点”：参与者愿意接受道歉，尽管这样做存在遭受剥削的风险。合作行为不可能凭空产生，也不可能在没有任何人愿意承担一定风险的情况下演化，因此，这种合作机制的初始建立可能颇具难度。

2.7 演变中的意义

必须牢记，信号的含义从来都不是固定不变的。系统始终处于演化之中，而在动态演化语境中，信号传递这一行为本身就会改变信号的“含义”，因为其他行动者与之建立的关联也会随之发生变化。因此，运用信号框架来识别信号的语境意义，需要保持谦逊并讲究方法。仅仅确定某个信号在一种情境下与特定的“真实情况”在功能上存在关联是不够的，我们还必须认识到这种关联的语境敏感性，并谨慎确保在其他情形下满足相关条件，使这种关联依然成立。

意义也会以可预测且具有特征性的方式“演变”，我们应对此有所预期。一种常见模式就体现于“古德哈特定律”中，这句格言通常表述为：“当一个测量指标成为目标时，它就不再是一个好的测量指标。”这里的测量指标概念充当了一个获得公众认可的信号。其含义是：某个事物最初之所以被认定为“测量指标”，是因为它曾经是一个强信号。但一旦该信号以任何一种在理论上可与内容分离的方式在系统中获得奖励，信号的意义就会因追逐奖励而逐渐被侵蚀。一个臭名昭著的例子是眼镜蛇效应（Siebert，2001）：英国殖民政策原本旨在通过悬赏死眼镜蛇来减少眼镜蛇种群，却反而刺激了眼镜蛇繁殖。尽管这种做法并不总是会如此戏剧性地适得其反，但由于信号的意义不断变化，过度依赖某个特定的可观测指标来指导政策选择，总是容易受到这类脆弱性的影响。

对单一评估指标的过度依赖，是人工智能领域公认的问题（Raji 等，2021年；Liao 等，2021年）。信号视角或许能够解释这些失败为何发生，并帮助我们寻找解决方案。尽管动态模型的证据可能容易受到建模选择的影响，也难以解读（Rosenstock 等，2017年），但这一领域最稳健的发现之一是：如果社群采用方法多样性并容纳各种假设，而不是迅速就一套统一的方法和信念达成一致，就更能有效地揭示真相（Zollman，2010年）。这类文献还提出了另一个方向：寻找那些作为一般类别奖励博弈的最佳响应而出现的信号，而非针对特定实例的信号；后者更容易被利用来博弈，甚至可能在没有恶意的情况下被过度拟合（Skyrms 和 Zollman，2010年；Zollman，2008年；Mengel，2012年）。

3 与其他研究的关系

3.1 博弈论开发动态

最显而易见、能够从纳入信号考量中受益的学术研究领域，是大量运用博弈论开展政策分析的研究（Hermans 等，2014），尤其是经济政策（Phlips，1995）以及围绕危险技术发展的政策研究（Coe 和 Vaynman，

2020）。近期也有一些研究特别运用博弈论来分析人工智能发展政策。

Askell 等人（2019）认为，如果负责任地开发人工智能比不负责任地开发需要付出任何额外成本，那么旨在实现有益结果的人工智能开发者就会面临囚徒困境。文章提出了政策干预的一些高层次特征，认为这些特征有望应对在博弈论上类似的情境。由于纳入信号可以使重复囚徒困境更易处理（Rosenstock 和 O'Connor，2018；Santos 等人，2011），因此，将信号纳入基础理论模型将是自然且有益的补充。

在一系列持续发表的论文中，Robert Trager 及其同事构建了一个更精确的博弈论框架，用来模拟同一现象；他们将其称为"危险技术竞赛"，并借此说明人工智能开发情境中的各种特征如何有助于实现积极结果。他们探讨的机制之一是"知识共享"，并将其建模为人工智能开发者之间共享的技术信息的纯粹"数量"（Stafford 和 Trager，2022）。在此纳入信号传递的一种方法，是将这一建模特征从类别变量或单维变量，改为一组有关知识共享性质与方式的更复杂选项。毕竟，知识并非一种可以直接"共享"的商品，而是由一系列证据构成，并配有一个框架，用于将这些证据纳入对技术如何运作的推测性推断中。这一组合中的不同"切片"会传递不同的"事实"，我们的模型理应对此有所体现。

也可以将此处以及（Han, The Anh 等人，2022）中的信号理解为以合作行为"承诺"的形式出现。因此，信号并非完全缺席，但正如我将在下一节讨论的那样，这或许不是将其纳入模型的最有效方式。

3.2 可信承诺

就现有关于这一主题的文献而言，"信号传递"通常被狭义地限定为某一特定类型的信号：即声明将采取某种行动或遵守一套规则的承诺。若要将这些内容视为信号，我们就必须更明确地考虑情境中的某些特征。

首先，在明确发送者与接收者角色时，我们需要问："我们 有什么资格要求特定类型的信号，而你们 又有什么资格提供这些信号？"如果"我们"并没有特别强大的地位来向人工智能组织提出要求，那该怎么办？更广泛地思考信号，有助于扩大讨论范围，纳入诸如组织照常开展业务时自然发出的"自然"信号，以及其他类型的"人为"信号（例如同意使用模型卡或接受审计）；这些信号会附着于并鼓励那些我们原本会要求其作出"承诺"的行为。

即便我们确实能够提取出承诺，在人工智能情境中评估“可信度”也可能很困难，因为相关任务往往相当复杂，难以评估。询问如何让承诺可信，或许并非正确的问题；我们不妨转而思考，如何在不涉及明确作出承诺的情境下评估可信度。“可信承诺”这一框架让我们处于这样一种局面：每个人当然都想获得作出承诺所带来的好处，却不承担兑现承诺的代价；而它也将问题设定为，如何设计奖励结构，使这种做法难以得逞。

3.3 可信度

鉴于上述考量，我们或许不应着眼于评估特定主张的可信性，而应更广泛地寻找其他参与者更高层次的可信度迹象，而不仅仅是看他们是否遵守某一特定规则和指导方针。此前引用的研究表明，与针对特定博弈进行“精细调校”的策略相比，当我们针对不同类别的博弈采用通用策略时，合作会更加稳定。这也引出了一个问题：一种通用且有效的“可信度”信号可能是什么样的?

Hardin（2002）将可信度界定为“包容性自利”——如果将我的利益纳入考量符合某人的自身利益，那么我就可以信任此人。无论这一解释是否充分，它都不太容易与信号传递框架相契合。为了判断你是否值得信任，我需要确定是什么在驱动你的行为，而正如我们将在下文关于意图的情境中讨论的那样，这一点无法被直接观察到。Jones（2012）提出的解释在这方面更具可行性。

“当且仅当B在互动领域D中对A而言是可信赖的，且就该领域而言，她是胜任的，并且如果A在该领域指望她，她会将A指望她这一事实视为有力的理由，据此采取行动。”（Jones，2012年，第61页）

Jones 的论述指出，有三个特征可能会体现在可观察的信号中：(i) 在某一领域的能力，(ii) 对某人依赖程度的了解，以及 (iii) 这种了解是否在关于如何行动的推理中发挥了恰当的因果作用。(i) 和 (ii) 促使我们首先寻找这样的迹象：系统是否有能力执行某项特定功能，以及是否意识到我们对它的依赖。[6] 尽管要将广义上的“能力”和“识别”转化为可操作的指标仍需付出一些努力，但通常存在明确的前提条件可供考察，例如是否具备某些功能和数据结构，以及在类似情境下的过往性能。这里需要关注的关键部分是 (iii)：确定如何界定并检测在行动的推理 过程中是否恰当地纳入了对依赖的考量。在这方面，我们可以参考可解释人工智能领域的研究

[6]我在这里使用“系统”一词，是为了表明，同样的基本推理也适用于更广泛的社会技术系统，以及（最终）能够独立运行的人工智能系统；不过，我目前认为，这些系统的独立性还不足以将其视为类似智能体的存在，下一节将对此展开讨论。

（Miller，2018，2019）以供参考。

正如 Jacovi 等人（2020）所论述的，我们的目标应是实现校准信任——即对系统的信任与其实际可信度紧密相连。该论文很好地区分了两种机制：与底层可信度密切相关的内在信任，以及由原则上可与实际可信度分离的特征间接产生的外在信任。与信号的“可伪造性”一样，这些机制往往呈连续分布；但遗憾的是，大多数“可信度”信号的例子都更偏向外在一端，包括专家证词、在特定测试集上的性能以及数据。可以说，能够引发更多内在信任的信号应是透明度机制的目标，正如下文将进一步讨论的那样。

3.4 透明度

对透明度的呼吁是人工智能伦理政策建议中的常见主题，可以将其视为对特定实质性事实的信号提出的请求或要求，这些事实涉及特定产品的性质或其开发背景。这个框架促使我们思考：如果将公司的产品／流程的哪些特征公开透明，它们可以作为哪些事实的信号，又能服务于什么目的。正如上文所述，一个可行方向是提供某种透明度，使用户能够通过获取与系统的内在可信度紧密相关的信号，在系统中建立校准信任。

从信息论视角来看，过多的信息——完全透明这一朴素目标所追求的——可能不如恰当类型的信息更有信息价值。公司可以（也确实会）提供大量真实信息，但这些信息仍可能产生误导——这是一个违反直觉的事实，可以用信号机制来解释（参见前文关于会话含义的讨论）。另一方面，值得探讨的是，透明地提供的数据原始数量是否可以成为一种粗略但有用的善意信号（Santos 等人，2011年；Skyrms，2002年）。

还值得考虑的是，促进善意的信号传递实践所需满足的条件，如何与公司保护知识产权的要求相互作用。军备控制领域或许有一个有用的类比（Coe 和 Vaynman，2020）。此外，也值得探讨其与差分隐私（Dwork，2006）的联系，并思考隐私与恰当形式的透明度能否以及如何共存。要在两者之间找到平衡无疑并不容易，但信号传递视角或许能帮助我们以比单纯追求最大化更聚焦的方式来制定透明度要求，从而使问题更易处理。

3.5 意向性

可信度问题与更广泛的意图问题相关，这既涉及企业主体，也最终涉及人工智能系统本身。当我们

评估一个人是否值得信任，或者更一般地说，当我们试图预测其行为并制定策略性应对方案时，我们的模型往往会将对其意图的评估作为重要组成部分。我们对意图的理解既包括字面层面，即人们打算如何支配自己的身体行动；也包括更抽象的层面，即我们预期哪些指导性价值观和原则会支配其最终具体的行动选择。谈到集体／企业代理时，意图这一概念就变得复杂起来（List 和 Pettit，2011年）；涉及非人类实体时更是如此。对于“人工智能对齐”领域的许多研究人员而言，描述并评估人工智能中的意图是一项重要的概念性目标，因为该领域的大量研究都预设，某种类似意图的东西正是或将会是对齐工作的核心对象。

我谨提醒，即使在人与人互动时，过度依赖意图也可能在预测和策略层面带来问题。意图只有在特定情境下才会表现为行为，因此，在策略互动中，情境因素往往应当主导（或被视为与意图因素相互交织）。一旦我们承认，即使人类也缺乏人们通常所寻求的那种“稳健”的意图概念，就会发现，试图识别并将类似意图的东西归因于人工主体，在概念上的重要性并没有我们想象的那么高。

我认为，信号传递视角促使我们将注意力从意图问题（或者至少从对意图过度拟人化的概念）上移开，而这或许能带来有益的启示。当我们寻找的是善意倾向的“信号”，而不是意图的证据时，就不要求主体必须有意产生这些信号，才能证明其具有某种内部主体状态。这使我们不那么容易受到前文所讨论的、与“可信承诺”相关的挑战的影响。这也为我们铺就了一条更清晰的路径：从当前关于开发人工智能技术的企业实体所采用的预测与策略模型，走向未来的模型；在未来模型中，相关考量将越来越多地被纳入人工智能系统本身。我们不必费心识别从一种模型到另一种模型之间的阶段性转变，也不必在面对一个拥有自身“意图”的人工智能“主体”时，认为必须采用全新的模型来理解并作出回应。

4 讨论

最后，我将总结我认为从信号视角考察有益人工智能的发展动态可以得出的三条核心经验。开头会比较抽象，结尾则会更加具体。

4.1 人工智能伦理不止关注直接后果。

围绕合乎伦理的人工智能的理论与实践通常聚焦于“第一阶”问题，即哪些具体行动会带来更好或更差的结果。

信号视角则将我们的注意力转向“第二阶”问题，即我们的行动如何影响其他主体的伦理行为。[7]

个人、团队、公司或产品所采取的每一个可见行动，都可能影响观察者的信念、预期和动机，而且这种影响可能不断扩散，并产生深远的社会影响。在努力做到“合乎伦理”时，这些主体不仅需要考虑自身行动的直接影响，还需要考虑这些行动所产生的信号带来的连锁影响。这加重了伦理考量的负担，但也为组织开辟了新的途径，使其能够致力于产生积极影响。

信号考量最有希望着眼的目标之一，是用于量化并展示对公平性、隐私和透明度等伦理原则的遵循情况的各类指标和基准。研究人员、监管机构、用户和公众会以不同方式使用这些指标，评估人工智能产品是否值得期待、能否被接受。其他公司会如何理解这些指标，以及这些指标会如何影响其行为，都会直接形成一种“奖励”，促使业内其他公司围绕这些指标的性能展开“竞争”。因此，关注使用和推广哪些指标，以及如何传达这些指标，至关重要。

4.2 信号是动态且适应性强的

信号之所以是“活的”，是因为它们不受我们对其抱有的希望和预期所约束，并且会随着时间推移，根据其所处的环境以及我们的使用方式而变化。认识到这一点，为我们利用信号促进人工智能开发情境中的亲社会行为开辟了新的、创新性的途径。在开发产品并确保其符合行善原则的过程中产生的每一项输出，都会生成大量可观察信息；我们可以利用这些信息来识别或制造可能适合我们目的的信号，而这些信号也会随着时间推移而不断演变。

反过来说，如果我们未能认识到信号的动态性，就可能对自己对信号的解释过于自信，也会在信号悄然变化时措手不及。例如，各种用于分类器的“公平性”指标，都只能在狭义且依赖情境的意义上表明“公平性”的存在。我们不仅应更审慎地选择使用哪些指标以及如何解释它们，还需要认识到，我们如何利用这些指标来指导自身行为——也就是说，奖励或惩罚在这些指标上表现良好或不佳的对象——必然会改变这些指标的含义，使其反映新的奖励结构。

[7]致各位伦理学家：这一 distinction 横跨后果主义与义务论之争。虽然义务论者可能更容易接受关系视角，但他们仍可能陷入“第一阶”近视，只过于关注履行对他人的义务，而不关注鼓励他人采取伦理行为。后果主义同样可以与第二阶视角相容；这种视角会从更加“整体性”的角度，根据我们行为在全球范围内产生的动态影响来审视后果。

4.3 内容源于情境激励

信号理论的核心启示是，信号的含义完全取决于发送者和接收者的信念与动机。换句话说，要了解某人在传达什么，关键只在于是谁在说，以及为什么这么说。要了解听者听到了什么，则需要知道谁在听，以及他们如何看待说话者的信念、意图和动机。

这是我看到大学同事们在 ChatGPT 问世后努力应对的一项教训。过去（在一定程度上），当学生提交一篇论文时，他们可以（或者至少确实会）将论文“照单全收”，把它视为学生诚实报告其所学内容、并展示其对论文题目所作最佳尝试的成果。如今，他们知道，“相同的”内容仅仅通过将论文题目复制粘贴到浏览器中就可以生成，这彻底动摇了他们对该信号的信任。

教授们所面对的问题在于，他们以为 自己是在要求学生根据论文题目撰写一篇论文，但他们实际上 向学生传达的是：生成一件符合课程大纲所列特定标准的独特成果——而完成这项任务并不需要批判性地参与课程内容的学习。
ChatGPT 在这里所产生的影响，是扩大了“具有某些特征的论文”与“学会了相关材料的学生”之间原本假定的关联，使其大到足以破坏这种信号关联。因此，要“解决”这一问题，就需要弥合这一差距，方法包括：（i）通过监控或审计强制执行合规要求；（ii）设计作业，使得即使可以使用 ChatGPT，要取得良好成绩仍然需要学生在过程中“顺带”学习；或者（iii）改变学生的激励，使其内在关联更多地指向学习，而不是在作业中取得好成绩。

作为一名教育工作者，我致力于实现（iii），承认（ii）往往更切合实际，尽管颇具挑战，并将（i）视为一种危险的权宜之计，因为它把学生置于对立面，而我希望尽可能避免这种情况。实现（iii）需要我们以开放、好奇的态度去了解学生的实际兴趣与目标，并灵活地与他们合作，寻找能引起他们共鸣的学习目标。针对企业行为者的方法当然会有所不同，但原则相同。如果我们只依靠对抗性惩罚措施（i），实际上就等于确保冲突永无止境：企业必然会利用并扩大漏洞，直到我们的方法失效。价值观完全一致（iii）是值得追求的理想，但基本上不可能直接着手实现。我们最好的办法是谨慎地制定并调整要求与预期，使亲社会地行事成为对激励作出的最“理性”回应。

参考文献

Askell, A.、M. Brundage 和 G. Hadfield（2019年7月）。负责任的人工智能开发中的合作作用。

Aumann, R. J.（1987年）。作为贝叶斯理性表达的相关均衡。《计量经济学》：计量经济学会期刊 *55*（1），1–18。Benn, C. 和 A. Grastien（2021年5月）。减少部分可观测人机交互中的道德模糊性。《先进机器人学》：日本机器人学会国际期刊 *35*（9），537–552。Coe, A. J. 和 J. Vaynman（2020年5月）。为何军备控制如此罕见。《美国政治学评论》*114*（2），342–355。Dwork, C.（2006年）。差分隐私。载于《自动机、语言与程序设计》，第1–12页。施普林格·柏林·海德堡。Franke, M.（2013年3月）。博弈论语用学。《哲学指南》*8*（3），269–284。Grice, H. P.（1975年）。逻辑与会话。载于《言语行为》，第41–58页。布里尔出版社。Hamilton, W. D.（1964年7月）。社会行为的遗传进化。一。《理论生物学杂志》*7*（1），1–16。Han, The Anh、T. Lenaerts、F. C. Santos 和 L. M. Pereira（2022年2月）。自愿安全承诺为人工智能开发摆脱过度监管提供出路。《社会中的技术》*68*，101843。Hardin, R.（2002年3月）。《信任与可信度》。罗素·塞奇基金会。Hermans, L.、S. Cunningham 和 J. Slinger（2014年1月）。将博弈论作为政策评估方法的实用性。《评估》*20*（1），10–25。Huttegger, S. M.、J. P. Bruner 和 K. J. S. Zollman（2015年12月）。障碍原理是一种人为产物。《科学哲学》*82*（5），997–1009。Jacovi, A.、A. Marasović、T. Miller 和 Y. Goldberg（2020年10月）。形式化对人工智能的信任：人类对人工智能的信任的前提、成因与目标。Jones, K.（2012年）。可信度。《伦理学》*123*（1），61–85。Lewens, T.（2015年8月）。《文化进化：概念性挑战》。牛津大学出版社。Lewis, D.（1969年）。《惯例：一项哲学研究》。约翰·威立父子公司。Liao, T.、R. Taori、I. D. Raji 和 L. Schmidt（2021年）。我们还在学习吗？关于机器学习中评估失败的元综述。第三十五届会议。List, C. 和 P. Pettit（2011年4月）。《群体能动性：法人代理者的可能性、设计与地位》。牛津大学出版社。McCornack, S. A.（1992年3月）。信息操纵理论。《传播学专论》*59*（1），1–16。

Mengel, F.（2012年3月）。跨博弈学习。《博弈与经济行为》*74*(2), 601–619。Miller, T.（2018年11月）。对比性解释：一种结构模型方法。Miller, T.（2019年2月）。人工智能中的解释：来自社会科学的启示。《人工智能》*267*，1–38。O’Connor, C.（2021年7月）。衡量约定性。《澳大拉西亚哲学杂志》*99*(3), 579–596。Ojea Quintana, I., S. Rosenstock 和 C. Klein（2021年11月）。流行病学模型构建者面临的协调困境。《生物学与哲学》*36*(6), 54。Oreskes, N. 和 E. M. Conway（2011年）。《怀疑的商人：少数科学家如何掩盖从烟草烟雾到全球变暖等问题的真相》。美国布鲁姆斯伯里出版社。Phlips, L.（1995年10月）。《竞争政策：博弈论视角》。剑桥大学出版社。Raji, I. D., E. M. Bender, A. Paullada, E. Denton 和 A. Hanna（2021年11月）。人工智能与包罗万象的全球基准。Rosenstock, S., J. Bruner 和 C. O’Connor（2017年4月）。在认知网络中，越少真的越好吗？《科学哲学》*84*(2), 234–252。Rosenstock, S. 和 C. O’Connor（2018年）。当感到难过是件好事：内疚与道歉的进化模型。《机器人学与人工智能前沿》。Santos, F. C., J. M. Pacheco 和 B. Skyrms（2011年4月）。博弈前信号传递与合作的共同进化。《理论生物学杂志》*274*(1), 30–35。Schelling, T. C.（1960年）。《冲突的策略》。Siebert, H.（2001年）。眼镜蛇效应：解决方案比问题更糟糕的情况。Skyrms, B.（2002年）。信号、进化与瞬态信息的解释力。《科学哲学》*69*(3), 407–428。Skyrms, B.（2010年4月）。《信号：进化、学习与信息》。牛津大学出版社。Skyrms, B. 和 K. J. S. Zollman（2010年8月）。社会规范形成中的进化考量。《政治、哲学与经济学》*9*(3), 265–273。Stafford, E. 和 R. F. Trager（2022年）。国际原子能机构解决方案：通过知识共享防止危险技术竞赛。

Zollman, K. J. S.（2008年2月）。解释复杂环境中的公平性。政治学、哲学与经济学 *7*(1), 81–97。Zollman, K. J. S.（2010年）。暂时性多样性的认知益处。*Erkennt-nis*. 分析哲学国际期刊 *72*, 17–35。