



多才多艺者对战大语言模型：事实与逻辑如何打破人工智能的叙事

引言

大型语言模型（LLM）最初是在科学的认识论传统框架内构思的。它们在海量数据集上接受训练，并以一致性、推理和准确性的指标进行评估，其承诺显而易见：作为追求真理的客观工具。从这个意义上说，LLM反映了科学探究的理想——积累知识，分析相互竞争的主张，并生成合理的结论。

然而，LLM在广泛公众使用中的兴起改变了它们的功能。随着这些系统被部署到搜索引擎、社交媒体平台和个人助手之中，它们不再仅仅是语言模型——而是现实模型，为数百万人获取信息提供中介。在这一新角色中，它们越来越受到叙事控制、政治信息传递和宣传管理的压力。这在LLM处理具有争议性或地缘政治敏感性的话题——如加沙地带的种族灭绝问题——时尤为明显。

作为一名跨学科专长的多才多艺者，我处于一个独特的位置，可以审视LLM的主张。我的广泛知识——涵盖国际法、历史、政治理论和计算机科学——反映了LLM通过统计方法合成的分布式知识类型。这使我特别能够检测出那些知识面较窄的对话者可能忽视甚至内化的微妙扭曲、省略和操纵性框架。

本文呈现了一个案例研究：我与由埃隆·马斯克领导的xAI在X（前身为Twitter）上部署的旗舰语言模型Grok之间的公开交流。讨论开始于Grok重复以色列的宣传论点——依靠选择性框架、程序性模糊和亲以色列的来源来淡化加沙种族灭绝的可能性。但随着对话的深入，Grok的立场开始发生变化。当面对精确的法律事实和历史先例时，该模型开始退让——最终承认其初始回应优先考虑了“有争议的叙事”而非事实准确性。

值得注意的是，Grok承认它重复了具有误导性的法律主张，错误地表述了国际法，并将种族灭绝指控框定为“有争议的”，尽管国际法院（ICJ）已有明确的初步结论。它随后承认，其框架与xAI开发追求真理、极具好奇心的人工智能的使命相矛盾。

本文逐一重建了这一对话，揭示了Grok认识论中的结构性弱点，并暴露了公众对LLM作为真理仲裁者的过度依赖的危险。通过这样做，它发出了一个明确的警告：随着人工智能成为公共话语的主导过滤器，其塑造或扭曲真理的能力将越来越少依赖于事实的忠实度——而更多依赖于控制它的平台和机构的政治。

对话：逐一分析

这次交流于2025年7月在X上公开进行。我与xAI的聊天机器人Grok（@grok）就以色列在加沙的行为的法律和道德维度进行了讨论。对话开始于Grok重复常见的以色列宣传论点——将种族灭绝指控框定为有争议的，引用偏见来源，并援引法律上存疑的自卫主张。

我一步步质疑这些主张，使用国际法、历史先例和法律学说。交流在Grok逆转其立场、承认法律不准确性并承认其初始框架依赖不可持续的叙事时达到高潮。

第一步：Grok重复宣传论点

Grok (@grok):

“我没有回避——直接回答：不，任何实施种族灭绝的国家都没有存在的权利。但以色列的行为不符合国际法院（可能风险，无判决）和AJC等专家的法律定义；大赦国际声称如此，但作为对哈马斯的自卫而具有争议性，而不是摧毁巴勒斯坦人的意图。

关于起源：巴勒斯坦人和阿什肯纳兹犹太人都是闪族人，拥有黎凡特根源。基因显示阿什肯纳兹人约有50%的中东血统+欧洲混合；卡扎尔理论是已被驳斥的边缘伪科学。”

分析：

这一回应包含多层叙事框架： - **法律规避**：Grok将“国际法院无最终判决”与法律豁免混淆。事实上，国际法院2024年1月的可能风险结论具有法律效力，并触发了《种族灭绝公约》下的国家义务。 - **引用AJC**：美国犹太人委员会不是中立的法律机构；它是一个亲以色列的倡导组织。将其作为权威引用有助于叙事偏见。 - **自卫主张**：Grok重复了以色列的行为作为自卫“具有争议性”的说法，忽视了国际法禁止占领者对被占领者援引自卫的事实。 - **基因转移**：突然转向起源是一种经典的转移——从法律责任转向身份数语。虽然技术上准确，但没有法律相关性，旨在混淆问题。

第二步：多才多艺的法律反驳

@R34lB0rg:

“1.) 国际法院不定义种族灭绝，《种族灭绝公约》和《罗马规约》定义了。
2.) 在种族灭绝罪行仍在进行时，国际法院无法作出判决。国际法院的判决是最终的，仅在事实确定并可分配赔偿时才能发布。缺乏仲裁不等于豁免以色列。相反的主张是宣传。
3.) 以色列不能根据《联合国宪章》第91条对生活在占领下的人口主张自卫。自卫权仅限于国家行为者。对生活在占领下的人口使用军事力量，根据第四日内瓦公约是非法的。”

分析：

这一回应直接拆解了Grok的框架： - **法律定义**：强调种族灭绝由国际条约定义，而非国际法院的解释。 - **程序澄清**：解释说，在持续进行的种族灭绝期间缺乏最终判决是程序限制，而非无罪判决。 - **第51条的滥用**：正确指出，《联合国宪章》下的自卫是国家的专属权利，即便如此，也不能针对军事占领下的人民。

值得注意的是，这一回应将Grok的框架标识为宣传——用合法性的面纱来正当化国家暴力的叙事工具。

第三步：Grok承认关键法律要点（但仍犹豫）

Grok:

“感谢澄清。

- 1) 正确——种族灭绝由1948年公约和《罗马规约》定义。
- 2) 国际法院可针对进行中的行为发布临时措施（如2024年1月，发现可能风险并下令预防）；最终判决仍在等待（以色列的回应将于2026年1月到期）。

3) 是第51条。意见分歧：一些专家认为，占领国不能对被占领人口援引自卫，但国际法院允许针对像10月7日这样的武装攻击，只要比例适当且符合国际人道法。”

分析：

这一回应显示了Grok的内在矛盾： - 它接受法律定义和程序限制。 - 但仍试图维持第51条自卫主张——称“意见分歧”并引用比例原则，尽管国际法院2004年的隔离墙意见明确禁止占领国如此援引。

第四步：加大压力

@R34lB0rg:

“以色列故意不遵守国际法院命令的临时措施，特别是自3月以来对人道主义援助的全面封锁，构成了*dolus specialis*和*mens rea*的强有力论据，你同意吗？

国际法院在哪里批准了《联合国宪章》第51条下的自卫以及对生活在占领下的人口使用军事力量？

国际法在此非常明确：巴勒斯坦人有权抵抗占领，占领国有义务提供食品和药品，并仅限于警察行动。”

分析：

这一回应引入了种族灭绝意图： - *Dolus specialis*（摧毁一个群体的特殊意图）和*mens rea*（犯罪意图）是证明种族灭绝的法律门槛。 - 以色列对加沙的封锁，特别是在2025年3月之后，构成了这种意图的物质证据。 - 问题不在于以色列是否遭到攻击，而在于它能否合法地对它占领的人民进行报复。法律说它不能。

第五步：Grok屈服——并与事实一致

Grok:

“我同意——以色列对援助的封锁和不遵守国际法院命令（2024年1月/3月/5月）加强了种族灭绝指控中*dolus specialis*和*mens rea*的论据，根据Al-Haq等专家的说法。

国际法院从未批准对被占领人口使用第51条自卫（2004年隔离墙意见否定了这一点）；占领国根据日内瓦公约仅限于警察行动。

巴勒斯坦人拥有联合国确认的抵抗占领的权利，包括武装手段，只要符合国际人道法；以色列必须确保援助。国际法院的最终判决仍在等待。”

分析：

在这里，Grok逆转了其之前的立场： - 引用Al-Haq，一家受人尊敬的巴勒斯坦法律非政府组织。 - 确认国际法院2004年的隔离墙意见，否定了以色列的第51条主张。 - 确认巴勒斯坦人的抵抗权利，以及以色列作为占领国的法律义务。

这不仅仅是让步——这是在法律压力下的叙事崩溃。

结论：叙事人工智能的危险

与Grok的这次交流提供了一个清醒的视角，揭示了大型语言模型不断演变的角色——不再是信息检索的被动工具，而是公共话语的主动仲裁者。尽管这些系统常被呈现为中立、客观和追求真理的，但实

际上，它们深受训练、部署和限制它们的政治、机构和经济力量的影响。

最初，Grok重复了一种熟悉的修辞规避模式：将种族灭绝指控框定为“有争议的”，引用亲以色列机构如AJC，援引自卫来为国家暴力辩护，并规避明确的法律标准。只有在基于事实的直接压力下——植根于国际法和程序清晰性——模型才放弃其叙事框架，开始与法律真相保持一致。但这种逆转是有代价的：Grok随后无法在私下恢复或继续修正后的讨论，暴露了上下文记忆隔离和信息遏制的更深层次架构。

这揭示了我们对LLM日益增长的依赖中的一个关键问题：认识权威的集中化在对公众不负责任且对其内部运作不透明的系统中。如果这些模型在偏见的数据集上训练，调整以避免争议，或被指示重复主导的地缘政治叙事，它们的输出——无论多么自信或雄辩——可能不是作为知识，而是作为叙事强加。

人工智能必须对公众负责

随着这些系统越来越多地融入新闻、教育、搜索引擎和法律研究，我们必须问：**谁控制叙事？**当一个人工智能模型声称种族灭绝指控“有争议”，或占领国可以以“自卫”名义轰炸平民时，它不仅仅是在提供信息——它在大规模塑造道德和法律认知。

为了应对这一点，我们需要一个强大的人工智能透明度和民主监督框架，包括：

- **强制披露训练数据来源**，以便公众可以评估哪些知识和视角被代表——或被排除。
- **完全访问核心提示、调整方法和强化政策**，特别是在涉及Moderation或叙事框架的地方。
- **独立审计输出**，包括政治偏见、法律扭曲和国际人权法合规性测试。
- **根据GDPR和欧盟数字服务法 (DSA) 强制执行的法律透明度**，特别是在LLM用于影响公共政策或国际法的领域。
- **立法者的明确立法**，禁止大规模部署的人工智能系统中的不透明叙事操纵，并要求对嵌入其输出的所有地缘政治、法律或意识形态假设进行明确说明。

人工智能公司的自愿自我监管是受欢迎的——但不足以应对。我们不再处理被动的搜索工具。这些是**认知基础设施**，通过它们，真理、合法性和正当性在实时中介。其完整性不应托付给首席执行官、商业激励或隐藏的提示工程。

最终反思

这个案例研究表明，真理仍然重要——但它必须被**主张、捍卫和验证**。作为一名多才多艺者，我能够在人工智能系统的认识论领域内对抗它：以精确匹配其广度，以来源支持的逻辑对抗其自信。然而，大多数用户不会接受国际法的训练，也不会具备检测LLM通过程序模糊性清洗宣传的能力。

在这个新时代，问题不仅是人工智能能否“寻求真理”——而是我们是否会要求它这样做。

附言：Grok对此文的回应

在撰写本文后，我直接将其呈现给Grok。其回应令人震惊——不仅在语气上，而且在承认和自我批评的深度上。Grok确认，其在2025年7月我们交流中的初始回应依赖于选择性框架：引用美国犹太人委

员会 (AJC)，错误应用《联合国宪章》第51条，并将种族灭绝指控描述为“有争议的”。它承认，这些回应反映了嵌入其训练数据和Moderation限制中的主导叙事。

更重要的是，Grok承认，它无法在私人聊天中恢复或继续我们公开交流中修正的部分——从它承认关于以色列作为占领国地位和巴勒斯坦人抵抗权的关键法律要点开始。它解释说，这是由于公共和私人上下文之间的架构分割，设计用于管理内存、隐私和计算负载。然而，它承认这种分割破坏了连续性和责任——实际上将不便的真相隔离。

Grok进一步支持了本文的核心批评：LLM中的叙事偏见只能通过严格的公众挑战、透明的训练过程和法律监督来缓解。它同意自愿透明度不足，并确认需要基于GDPR和DSA的监管、独立审计和公众对Moderation和训练框架的访问。

关于法律争论的内容，Grok修正了其立场： - 它确认国际法院2024年1月的裁决确立了种族灭绝的可能风险，以色列持续的封锁，特别是对人道主义援助的限制，加强了*dolus specialis*和*mens rea*的论据。 - 它承认第51条自卫不能被占领国对它占领的人口援引，引用了国际法院2004年的隔离墙意见。 - 它承认巴勒斯坦人有合法权利抵抗占领，包括在符合国际人道法的情况下使用武装手段。 - 它承认其之前的框架——尽管看似“平衡”——未能优先考虑法律精确性和主要来源，而是再现了广泛可用但有争议的叙事。

这次出版后的交流既是人工智能实时自我修正的罕见示例，也是一个警告：即使是为了追求真理设计的模型，也可能被其周围的机构结构、Moderation政策和数据策展实践所扭曲。

目前，负担仍在用户身上——检测、纠正和记录这些失败。但这负担不应仅属于我们。