信息安全、法学双学位班 2113203 付政烨

**从霍布斯主权理论看权力分散的AI治理问题**

**摘 要**

人工智能（AI）的快速发展带来了深远的技术变革，但同时也引发了算法偏见、隐私泄露等治理难题。当前全球AI治理呈现多中心化的格局，跨国企业、政府和技术组织之间缺乏协调，规则碎片化和责任模糊化现象严重。在此背景下，霍布斯的主权理论为AI治理提供了重要的理论启发，其关于权力集中与社会秩序的主张为AI治理框架的设计提供了借鉴。然而，霍布斯的绝对主权理念在现代社会中存在压制公民自由和治理僵化的风险。因此，本文进一步提出“动态主权”的概念，结合技术驱动和分布式治理机制，通过区块链和智能合约等工具实现权利义务透明化和规则自动化。同时，建立一个全球化的AI治理机构，整合各方力量，制定统一的治理框架，不仅能提高治理效率，还能在保护公众权益与技术创新之间实现平衡，为AI的可持续发展提供支持。

**关键词：**AI治理、霍布斯主权理论、社会契约、动态主权

**目 录**

[一、引言 3](#_Toc30297)

[二、AI治理的现实困境 3](#_Toc14434)

[三、霍布斯主权理论与AI治理的联系 4](#_Toc14634)

[四、构建统一的AI机构 6](#_Toc13847)

[五、对霍布斯主权理论应用于AI治理的补充 7](#_Toc1462)

[六、总结与展望 7](#_Toc14101)

[参考文献 9](#_Toc10224)

[致谢 10](#_Toc21865)

## 一、引言

人工智能（AI）的广泛应用带来了巨大的技术红利，从生成式人工智能在创意生成领域的革命性影响，到自动驾驶技术对交通行业的深刻变革，AI正逐步渗透到社会的每一个领域。然而，AI技术的进步也伴随着算法偏见、隐私泄露、数据滥用等一系列问题。这不仅影响了社会的公平和安全，还加剧了公众对人工智能技术的担忧与不信任。目前，全球的AI治理格局呈现出权力高度分散的局面，跨国技术公司、国家政府、技术社区和学术机构各自为政，缺乏一个能够高效协调与监督的中心化治理机构。这种多中心化治理模式效率低下，不仅影响了治理规则的制定，还模糊了权力与责任的界限。在这样的背景下，建立一个行之有效的治理框架就显得尤为重要。

霍布斯提出的主权理论强调，分散的权力会导致混乱与无序，只有集中化的权力才能解决社会中的冲突并达成有效治理。受该理论的启发，本文提出：在当下AI治理的过程中，权力集中是解决效率与责任问题的必要条件。同时，这种集中化治理模式必须适配现代技术和伦理要求。

## 二、AI治理的现实困境

AI治理的一个主要困境是权力的多中心化，当前AI治理的三大主体，即企业、国家机构和技术组织，三者之间缺乏统一的协调机制。具体来说，企业（如Google、OpenAI）在数据和算法的开发与部署中占据主导地位，国家机构通过立法和监管试图控制技术应用的边界，而技术组织（如IEEE）则试图制定技术标准和伦理指南。然而，三者之间目前还没有达成统一的意识，在规则制定上存在一定冲突和矛盾，这导致治理规则碎片化，协调困难，执行效率低下。尤其在跨国问题上，单个国家或组织往往无法对跨境AI技术进行有效约束。这种权力分散的状态正如霍布斯所描述的“自然状态”。

另一个困境是责任分配的不明确性。在许多人工智能应用场景中，当出现问题时，责任归属往往变得模糊。例如，在自动驾驶技术中，如果车辆因算法错误导致事故，应由算法开发者、汽车制造商还是使用者承担责任？类似的，算法偏见引发的歧视问题（如AI招聘工具对女性或少数群体的不公平对待）也暴露了现有治理机制对责任分配的无力。这种现象表明，目前的AI多中心治理缺乏一个明确的权力责任框架。霍布斯的主权理论为这种情况提供了一个潜在的思考方向——通过权力的集中，建立一个能够明确责任主体的治理结构，减少责任主体推诿的空间。

## 三、霍布斯主权理论与AI治理的联系

AI治理的困境表明，当前的多中心治理模式难以应对AI技术的复杂需求。霍布斯的主权理论虽然诞生于不同的时代，但其关于权力集中与社会秩序的核心思想，依然可以为现代AI治理提供重要的理论启发。

霍布斯在《利维坦》中提出，主权的绝对性和集中性是解决社会混乱的关键。在他看来，人类在自然状态下处于“每个人对每个人的战争”中，这种状态无序又危险。为了摆脱这种混乱状态，人们通过社会契约，自愿放弃部分自然权利，将权力交由一个绝对的主权者。这一主权者拥有终极裁决权，这样能够有效地解决社会冲突，避免因权力分散导致的争端升级。霍布斯明确指出，分散的权力只会引发无尽的矛盾和混乱，而权力的集中是社会秩序的基础。当前，AI治理面临的突出问题是权力的多方分散——跨国技术公司掌握了数据和算法主导权，国家政府试图通过立法进行监管，技术社区和学术机构则推动伦理规范的制定。然而，各主体之间缺乏统一协调，导致治理规则碎片化、效率低下，甚至在算法失误或数据滥用等问题上责任归属模糊。因此，正如霍布斯所强调的，AI治理亟需通过权力的集中化建立有效的协调与监督机制，以应对这一复杂局面。

根据霍布斯的理论，主权者的核心功能之一在于维护社会秩序的稳定。在当前的AI治理实践中，不稳定性主要源于其高度分散化和多中心化的特征，这一特性突出表现在AI治理的三大主要主体（企业、各国政府及国际组织）内部及其相互之间的矛盾与冲突。首先，企业间的竞争显著加剧了AI治理的分散性。全球科技巨头（如谷歌、微软、OpenAI等）在AI技术开发领域各自为战，推动了多种标准、框架和规范的建立。然而，这些标准通常以企业自身利益为导向，缺乏对公共利益的系统性考量。这种碎片化的规则制定模式不仅导致技术研发方向的不一致，也使治理规则之间出现冲突。其次，各国政府在AI治理战略上的差异进一步削弱了全球治理的协调性。从目前的发展趋势来看，美国的政策倾向于技术创新与市场驱动，而欧盟则更加注重伦理规范与风险控制，中国则采用国家主导的政策框架与产业引领模式。这些差异不仅阻碍了统一全球标准的形成，还导致了技术应用与法规适用的跨境协调难题。尤其是在自动驾驶领域，不同地区（如欧美与亚洲）在安全标准、测试规范和责任认定框架上的分歧尤为明显，加剧了治理的复杂性。此外，国际组织和非政府团体在AI治理中的作用尽管不可忽视。例如，联合国强调构建全球性的伦理框架，而IEEE等技术组织则专注于技术标准的制定等等。综上所述，当前AI治理的不稳定性主要根源于企业、政府及国际组织等多主体的多中心化治理模式。这种模式因各主体的利益驱动和目标差异而难以实现规则统一，使得治理效果呈现出分散性与非协调性的特征，从而显著增加了AI技术发展过程中的不稳定因素。

其次，主权者必须保护公民的安全，这是霍布斯社会契约理论的核心主张之一。在霍布斯看来，公民安全不仅指免受外部侵害的生存保障，还包括免于内部冲突与不确定性的稳定生活环境。在现代社会，随着技术的深度融入日常生活，公民安全的内涵已延伸至网络空间和数据领域。在人工智能治理中，数据隐私和系统安全性成为公民安全的核心维度：数据隐私直接关系到个人信息免受滥用与侵犯的基本权利，而算法的安全性则关乎技术的可控性及对潜在风险的防范。例如，个人数据的泄露可能引发身份盗用和经济损失，而自动驾驶或医疗AI系统的失效甚至会威胁生命安全。因此，确保数据隐私与技术安全已成为保护公民安全的重要内容，是人工智能治理中不可或缺的核心目标。对此，各国已开展多种立法尝试。美国的《算法问责法案》（Algorithmic Accountability Act）通过要求企业进行算法风险评估和透明性披露，初步体现了保护公众权益的意图。我国则通过《数据安全法》和《个人信息保护法》等法规，明确了对个人数据的保护要求，同时对AI技术的安全性和合规性提出了具体指导。欧盟也出台了《人工智能法案》（AI Act），旨在对高风险AI系统进行严格的监管，并设立强制性要求以保障技术透明性和用户权益。这些法规反映了各国对AI治理的重视，但由于缺乏统一的执行机制，这些分散的立法努力难以全面应对技术风险。在霍布斯的理论中，主权者是一个能够集中权力、为社会提供秩序和安全的权威实体。在当前的AI治理框架中，分散的法律和政策体系类似于霍布斯所描述的“自然状态”，即各主体独立行动，缺乏协调和统一的规则，而这恰恰可能导致治理碎片化和技术风险的扩大。霍布斯认为，只有通过社会契约建立一个强有力的主权者，才能有效避免无序和混乱。同理，在AI治理中，一个类似“主权者”的机构通过集权可以增强政策执行力、快速响应技术带来的新风险，为公众在信息时代的安全提供可靠的保障。

最后，主权者承担着制定和执行规则的职责，这是霍布斯理论中的核心功能。在AI治理领域，虽然已经出现了一些集中化管理的尝试，但由于执行力有限，治理效果仍然不尽如人意。例如，联合国教科文组织发布的《人工智能伦理问题建议书》在价值观层面提出了诸多重要指引，如强调人类尊严、平等和可持续发展。然而，由于缺乏具体的法律约束力和强有力的执行机制，其实际影响力在全球范围内十分有限。这种局面突显出一个能够专注于规则制定与执行的强效治理机构的必要性。这一机构的职责应涵盖技术发展中的核心议题（如前述所述）。同时，监督和问责机制应作为执行的关键环节，通过动态监控技术应用情况、纠正违规行为，确保规则的真正落实。此外，该机构还需具备快速响应能力，以便在面对生成式AI或高风险应用（如自动驾驶系统）引发的突发性技术风险时，能够迅速介入并采取有效措施。

## 四、构建统一的AI机构

基于当前人工智能治理中规则碎片化、责任模糊以及治理效率低下的问题，同时结合霍布斯主权理论关于权力集中化的启示，亟需构建一个权力集中、技术驱动的AI治理机构，以应对人工智能技术快速发展所带来的复杂挑战。该机构应以“权力集中、全球协作、技术驱动”为核心原则，致力于在全球范围内协调多方利益，推动人工智能治理的统一化与规范化。具体设计如下：

1. 该机构的核心目标是建立统一的全球性AI治理框架。通过集中化的权力配置，有效整合当前分散的立法与伦理规范，将区域性治理标准（如欧盟的《人工智能法案》）与全球范围内的实际需求相结合，从而制定具有广泛适用性的治理规则。
2. 技术驱动将是该机构的重要特征。现代技术（如区块链和智能合约）可以显著增强治理过程的透明度与公信力。例如，可以利用区块链技术搭建全球数据共享和追踪平台，确保跨国数据流动的安全性和合规性。同时，智能合约技术可用于实现治理规则的自动化执行，减少人为干预和潜在的不当权力使用，从而提升治理效率和公平性。
3. 为体现全球协作的原则，该机构需要整合各国政府、跨国企业、技术专家以及公众代表，形成多方参与的治理架构。在具体实践中，机构可以通过区域性分支机构处理本地化问题，同时确保在核心治理框架上的全球一致性。多方协作模式的引入不仅可以增强治理的包容性和合法性，还能确保规则制定与执行在多元化利益主体间实现有效平衡。

## 五、对霍布斯主权理论应用于AI治理的补充

霍布斯的主权理论为权力集中化的必要性提供了重要参考，但不可否认的是，其古早的理念在现代AI治理中也存在局限性。一方面，霍布斯强调主权的绝对性可能导致对公民自由和隐私的压制。另一方面，霍布斯理论的静态权力集中模式也难以适应人工智能技术的快速发展和高度动态性。人工智能技术不断演化，其治理需求也在持续变化，仅依赖传统的集中化权威可能会导致治理滞后或僵化。

为了弥补这些局限性，可以进一步引入“动态主权”的概念。即在维持权力集中的基础上，增加动态调整机制，使治理结构能够随技术变化灵活应对。通过技术工具如智能合约技术，可以在规则执行过程中实现自动化和透明化，减少人为干预和权力滥用的风险。此外，结合分布式治理的理念，动态主权可以允许在某些领域实现权力的局部分散，鼓励地方性和行业性治理，同时通过核心机构确保整体框架的一致性。

## 六、总结与展望

霍布斯的主权理论为AI治理提供了重要启发。当前全球多中心化治理模式的低效与不确定性，正呼应了霍布斯对分散权力的批判。通过集中化的治理框架，统一规则、明确责任主体、维护秩序，已成为解决AI治理困境的核心方向。然而，霍布斯理论中的绝对主权理念在现代治理中也存在局限性，尤其是在保护公民自由、隐私以及应对技术动态变化方面的不足，这需要新的理论与技术补充。

展望未来，AI治理的有效性将取决于全球协作与技术创新的结合。首先，需要建立一个全球化的治理机构，不仅要协调各国政府、跨国企业和技术社区，还需为公众提供参与和监督的渠道，以增强治理的透明性和公信力。其次，技术将是解决治理局限性的重要驱动。此外，“动态主权”理念可以为未来的治理框架提供灵活性，允许在特定领域或区域实现分布式治理，同时通过核心机构保持整体框架的一致性。未来的AI治理需要以动态和开放的姿态面对技术变革与社会需求。通过集中化与分布式机制的结合，以及现代技术的深度应用，我们可以在全球范围内建立一个高效、透明且公平的治理体系，为人工智能的可持续发展提供有力支撑。

## 参考文献

1. 杜泽蒙.算法权力嵌入公共政策领域的内在逻辑,风险表现与治理策略[J].北京理工大学学报(社会科学版), 2024(5).DOI:10.15918/j.jbitss1009-3370.2024.6395.
2. 堵一楠.霍布斯主权理论研究[J].金陵法律评论, 2020.DOI:10.3969/j.issn.1001-4608.2020.01.007.
3. 陈静,黄萃,苏竣.中美人工智能治理研究比较分析——基于文献计量视角[J].电子政务, 2020(12):11.DOI:10.16582/j.cnki.dzzw.2020.12.010.
4. 贾开,薛澜.人工智能伦理问题与安全风险治理的全球比较与中国实践[J].公共管理评论, 2021(1):13.
5. 许超.从审慎到科学——霍布斯与国家理性的现代转型[J].人文杂志, 2024(3):74-84.
6. 刘欣.论霍布斯的政体理论[J].行政事业资产与财务, 2014(3):2.DOI:CNKI:SUN:XZSY.0.2014-03-152.
7. 孙全胜.人工智能技术赋能数字法治政府建设的内在机理和优化路径研究[J].人工智能, 2023(6):82-96.
8. 陈少威,吴剑霞.人工智能治理风险和工具的识别与匹配研究——基于政策文本的分析[J].中国行政管理, 2022(9):8.

## 致谢

本文的完成，离不开《西方法律思想史》课程的启发与指导。这门课程如同一扇通往思想宝库的大门，让我得以深入探究西方法律思想的演变脉络与核心理念，更教会了我如何以经典理论的视角剖析当代社会问题。在学习过程中，黄宇昕老师以其严谨的学术态度和深入浅出的教学方式，将晦涩难懂的法学理论化为生动鲜活的思想画卷，使我得以更加全面地理解西方法律思想的深邃内涵。尤其是老师对霍布斯、卢梭等法学思想家的精妙剖析，不仅让我深刻体会到这些理论的历史背景与时代意义，更让我看到它们在现代社会中的恒久价值。这些讲解宛如灯塔，为我的思维注入灵感，拓宽了认知的边界，也为本文的选题与写作提供了重要启发。老师的教学不仅点燃了我对法律思想的热情，更让我在探索的道路上受益匪浅。在此，我谨向黄老师致以最诚挚的感谢！