财经法学 No. 6，2024
pp. 108 - 127

论软法的实施机制
——以人工智能伦理规范为例

沈　岿*

内容提要：软法的广泛存在，并不意味着其切实地得到了遵守和执行。人工智能领域的软法——人工智能伦理规范——被证明存在“实效赤字”，其原因在于：人工智能伦理规范的非强制性，抽象性、模糊性，分散、混乱与叠床架屋，自愿遵守的动力不足，合规悖论，社会系统论困境，以及人工智能发展压倒约束的宿命论。但人工智能伦理规范因其灵活快捷性、多样适配性、合作试验性、事实压力性、跨国适用性而仍然有独特价值。经验研究表明，组织机制、合规压力机制、合规激励机制、技术方法论机制、基准机制以及软硬法互动机制，可推动软法的间接实施。价值共识与经济逻辑的结合、内在理由和外在推动的结合，是软法获得更多实效之道。

关键词：软法　人工智能　伦理规范　实施机制　软法实效

一、问题：软法何以产生实效

软法效力或有效性（validity）——其“应当”得到遵守和实施的性质——在于说服约束力，而不在于强制约束力。软法只要不与硬法或硬法原则、精神相抵触，又大致符合一定范围内社会对更好的“公共善”的认知和期待，就具备独有的效力。由于软法制定者的权威性、“公共善”的认可程度、软法制定过程的协商性和沟通性等存在差异，软法的说服约束力有强弱之分，但共同之处是，软法的“应当”并不辅助以强制实施的制裁装置。〔1〕由此而言，软法的应当有效与软法的实际有效，并非一回事。前者是规范意义上的存在，后者是事实意义上的存在。

然而，软法的常规定义本身又意味着其是在一定范围内发生实效的，很难想象，没有实效又

* 沈岿，北京大学法学院教授。

〔1〕 参见沈岿：《论软法的有效性与说服力》，载《华东政法大学学报》2022 年第 4 期。

没有硬法属性的行为规则，可以当得上“软法”称谓。于是，一个需要处理的问题是，软法又是如何产生或者获得普遍实效的。软法自提出和公布之后，至其事实上产生效果，必定会有时间间隔，无论该间隔之长短如何。有着软法性质的行为规则，在其问世伊始，通常并不会立刻、即时收获效果，除非其只是对已经被普遍遵守和实施的惯常做法赋予规则的形式。这种例外的情形较为少见，毕竟，绝大多数软法是未来导向的，是期待人们为了更好的“公共善”而遵循新的行为规则或改变原先的行为规则。尽管软法的生命力源于其自身内在的说服力，但是，仅仅凭借这个内在属性或内在理由，就期待一个被提议的软法可以演变为真正意义软法，应该是过于理想化的奢望。因为，指向更好“公共善”的软法通常需要让行为人负担更多的遵循或适用成本。如果没有合适有效的机制可以减少或抵消这样的成本，那么趋利避害的行为选择倾向或者良币避免被劣币驱逐的动机，往往会压倒软法内在理由的吸引力，从而使其无法获取普遍效果。这就是在软法具备内在理由使其获得应然效力之外探讨软法何以产生实效的意义所在。

罗豪才、宋功德曾经在国内软法学的扛鼎之作《软法亦法——公共治理呼唤软法之治》中指出，“法依靠国家强制力保障实施”的表达并不准确。对于法的实施——即将法的效力转化为法的实效——而言，国家强制力保障是不可或缺的，但二者之间又不是必然的关系。法的实施可以是行为人：（1）因为从众而习惯性服从；（2）出于认可而自愿服从；（3）受到激励而遵从；（4）迫于社会舆论等分散的社会压力而遵守；（5）迫于组织的压力而服从；（6）慑于国家强制力的使用或威胁使用而服从。由此，法的实效产生方式是多样化的，法的实施机制主要有自愿服从、习惯性服从、社会强制服从、国家强制服从四种方式。这些讨论是作者在反思和修正“法”的定义过程中展开的，其最终指向一个包容硬法和软法在内的全新的“法”概念，在这个概念构成中，法的实施机制被概括为“公共强制”和“自律”。〔2〕毫无疑问，在以上所列六项之中，除国家强制服从仅适用于硬法以外，其余诸项皆可在软法的实施过程中呈现。

然而，就本文关心的问题而言，以上诸项，或许只有激励、社会压力、组织压力是值得关注的使软法产生实效的方式。因为，从众性的服从显然不是软法从倡议到普遍遵守的机制，“从众”本身就意味着已经存在普遍实效。自愿性的服从是出于对软法内在理由的认可，是软法实施的一种动力。只是，在硬法条件下的自愿性服从，除了在价值认同上有无形收益外，至少还有避免国家强制制裁的收益。而前文已经提及，软法条件下的自愿性服从，不仅不会有避免制裁的好处，甚至可能会导致服从者付出更多的成本或代价，其也就很难成为软法产生实效的强有力机制。

当然，罗豪才、宋功德在议论“法的实施”时，并未突出对软法实施的特别关注，其提及的激励、社会压力、组织压力，更多是在理论层面上针对所有法规范（包括硬法和软法）实施的逻辑展开，欠缺软法实践的丰富例证。更为重要的是，因为没有将软法何以产生实效问题提到显著的、专门的位置，没有列入有意识要解决的议题之中，所以激励、社会压力、组织压力是否就能概括所有的或绝大部分的软法实施机制，也就自然不会有较为明确的论述。

而从比较法的视野观察，域外软法研究者对软法实效问题有着更多的、更直接的关注。例如，德国自由柏林大学教授米莉亚姆·哈特莱普（Miriam Hartlapp）于2019年发表其对欧盟软

〔2〕 参见罗豪才、宋功德：《软法亦法——公共治理呼唤软法之治》，法律出版社2009年版，第187－202页。

法在欧盟成员国的实际效果进行的研究，指出软法的合法性或正当性（legitimacy）并不是推动软法实施的关键，真正起作用的是行为人是否能在实施中获益。而软法的可能硬法化（hardening out）是与软法实施并行的。〔3〕德国波茨坦大学教授安德里亚斯·齐默尔曼（Andreas Zimmermann）则于 2021 年探讨了不具有法律约束力的文件——以谅解备忘录为例——是如何在国际法之下产生法律效果的，指出主要是因为此类文件与具有法律约束力的文件发生互动所致，而这种互动是由许多法律机制提供的。〔4〕美国亚利桑那州立大学教授盖瑞·马秦特（Gary E. Marchant）和研究员卡洛斯·伊格纳西奥·古铁雷斯（Carlos Ignacio Gutierrez）于 2020 年合作完成关于人工智能软法间接实施的文章认为，软法成功与否是高度依赖特定情境的，取决于遵守软法的成本与可行性、对遵守软法的激励以及拒绝遵守或没有遵守软法的后果；他们描述了九个有助于人工智能软法更加有效、更加可信的机制和过程，并暗示可以有更多其他的。〔5〕相关研究不可尽数，但以上数例已经表明：一方面，如本文之前所述，论者们都倾向于一个基本前提，软法的实效更多取决于遵守软法给行为人带来的好处，包括利益之增加和不利之减少；另一方面，使行为人获得好处从而可以促进软法收取普遍效果的机制远不止于激励、社会压力、组织压力。

然而，对于软法的倡议者、推动者、研究者而言，或许需要一种软法实施机制类型学对林林总总、形形色色的实施机制进行归类，从而形成相对固定又具有开放性、包容性的思维工具，以促进为软法实施进行有意识的配套机制建构。“相对固定”意味着形成一些明确的分类概念，每个概念因其抽象性而可收留“家族相似”的具体形式化的软法实施机制；“开放性、包容性”意味着本文没有或不能述及的、实践中已有或者未来可能有的更多形式的实施机制，也可以为这些类型概念所容纳。本文即要探索软法有哪些类型的实施机制可以增大其产生实效的可能性。

鉴于软法在各个公共治理领域普遍存在，为使研究更加聚焦，本文选择人工智能的软法实施作为主要研究对象。人工智能为不计其数的研究者、开发者、应用者带来同样不计其数的大大小小利益，在强大的利益驱动下，人工智能快速发展，而各国政府即公共监管者的立场更多是容许而不是抑制其发展，尤其是在人工智能最初方兴未艾的阶段，这个立场伴随的就是基于软法的规制。〔6〕即便随着人工智能风险的清晰化，对不同风险进行分类管理和控制的硬法规范日渐增多，〔7〕但

---

〔3〕 See Miriam Hartlapp, *Soft Law Implementation in the EU Multilevel System: Legitimacy and Governance Efficiency Revisited*, in Nathalie Behnke, Jörg Broschek & Jared Sonnicksen eds., Configurations, Dynamics and Mechanisms of Multilevel Governance, Palgrave Macmillan, 2019, pp. 193 - 210.

〔4〕 See Andreas Zimmermann, *Possible Indirect Legal Effects of Non-legally binding Instruments*, available at https://ssrn.com/abstract=3840767, last visited on Aug. 6, 2024.

〔5〕 See Gary E. Marchant & Carlos Ignacio Gutierrez, *Indirect Enforcement of Artificial Intelligence "Soft Law"*, available at https://ssrn.com/abstract=3749776, last visited on Aug. 6, 2024.

〔6〕 人工智能的出现并没有引起立法者的自发反应。相反，欧盟委员会和各国政府最初发布的是不具有法律约束力的各种“计划”。德国 2018 年的“联邦政府人工智能战略”重点在于促进人工智能发展；迄今为止，美国仍然青睐这种规制方式。参见〔德〕沃尔夫冈·特伊普勒：《〈欧美人工智能法案〉的背景、主要内容与评价——兼论该法案对劳动法的影响》，王倩译，载《环球法律评论》2024 年第 3 期。

〔7〕 例如，我国于 2022 年 3 月 1 日起施行的《互联网信息服务算法推荐管理规定》，于 2023 年 8 月 15 日起施行的《生成式人工智能服务管理暂行规定》。欧盟于 2024 年 8 月 1 日起分阶段实施、于 2026 年中期全面适用于人工智能开发者的《欧盟人工智能法案》。

也不能完全取代这个领域软法的重要地位。〔8〕需要特别指出的是，人工智能治理的软法形式主要是伦理规范（ethics）。篇幅所限，本文无意就科技伦理与软法之间的关系展开讨论，尽管这也是具有重要价值的、属于软法本体论——软法是什么——的议题。美国的盖瑞·马秦特教授和瑞士的艾菲·瓦耶纳（Effy Vayena）教授等人将人工智能伦理规范视为软法一种形式的进路，〔9〕也是本文采取的。

本文将从三个方面展开探讨。首先，第二部分根据既有研究，对人工智能软法治理的现状进行事实描述，指出人工智能伦理规范的“风起云涌”无法掩盖其存在的巨大的“实效赤字”；其次，第三部分分析软法“实效赤字”的原因所在，以及即便如此，人工智能治理为什么需要并且仍然需要软法；再次，第四部分则揭示有助于软法实施并产生实效的机制，并对其进行分类，以期建立具有指导意义的理论工具。本文的最后结语是对全文主要观点的总结，并且强调软法的落地实施、获得普遍遵守，需要价值共识与经济逻辑的结合、内在理由与外在推动的结合。

二、人工智能软法及其“实效赤字”

瑞士的艾菲·瓦耶纳教授、马塞洛·林卡（Marcello lenca）教授和安娜·乔宾博士（Anna Jobin）等在《全球人工智能伦理指南图景》一文中指出，过去五年之间，私营公司、研究机构和公共领域组织发布了大量的人工智能伦理原则和指南，以应对人工智能引起的担忧。这些伦理指南并不具有法律上的约束力，而是说服性质的，其可以被称为非立法性政策文件或软法。为了研究不同团体在合乎伦理的人工智能应该是什么、未来决定人工智能发展的伦理原则是什么等问题上是否达成共识，以及如果有分歧，差异之处在哪里以及是否可以和解，他们在全球范围内收集了 84 个含有人工智能伦理规范的文件。

对这些文件的研究表明：第一，公共领域组织（包括政府组织和政府间组织）与私领域（包括公司及其联盟）发布的伦理规范在数量上大致相当，意味着两个领域都对此高度重视。第二，非洲、南美洲、中美洲、中亚等地区缺少代表，意味着人工智能伦理规范国际话语中的权力不平衡。第三，经济更加发达的地区正在塑造人工智能伦理规范的讨论，这可能会引起对地方性知识、文化多元主义和全球公平的关切。第四，人工智能伦理原则主要有：（1）透明；（2）正义、公平和平等；（3）不伤害（Non-maleficence）；（4）责任和归责；（5）隐私；（6）造福人类；（7）自由和自治；（8）信任；（9）可持续发展；（10）尊严；（11）社会团结。第五，没有一个原

〔8〕例如，《欧盟人工智能法案》在长达数十页的序言中指出，欧盟委员会独立人工智能高级别专家组 2019 年制定的《值得信赖的人工智能的伦理准则》具有重要意义，“在不影响本条例和任何其他适用的联盟法律的法律约束力要求的前提下，这些指南有助于设计一个符合《宪章》和作为联盟基础的价值观的连贯、可信和以人为本的人工智能”，“鼓励所有利益相关者，包括产业界、学术界、公民社会和标准化组织，在制定自愿性的最佳实践和标准时酌情考虑这些伦理原则”。《欧盟〈人工智能法〉议会通过版本：全文中译本》，朱悦译，第 9 页，载 https://aisg.tongji.edu.cn/info/1005/1201.htm，最后访问时间：2024 年 8 月 6 日。

〔9〕See Gary E. Marchant，“*Soft Law*” *Governance of Artificial Intelligence*，available at https://escholarship.org/content/qt0jq252ks/qt0jq252ks_noSplash_1ff6445b4d4efd438fd6e06cc2df4775.pdf，last visited on Aug. 6，2024；Anna Jobin，Marcello lenca & Effy Vayena，*The Global Landscape of AI Ethics Guidelines*，Nature Machine Intelligence，Vol. 1，2019，pp. 389－399，available at https://doi.org/10.1038/s42256－019－0088－2，last visited on Aug. 6，2024.

则是整个文件库中共同的，尽管透明、正义和公平、不伤害、责任以及隐私是比较集中的，有超过一半的指南涉及。第六，所有十一项原则都存在实质内容的分歧，决定分歧的主要因素有：（1）如何解释伦理原则；（2）为什么它们是重要的；（3）它们与什么问题、什么领域、什么行动者相关；（4）它们应该如何得到执行。基于这些发现，该文作者认为：在政策层面，需要各方利益相关者更多的合作，以在伦理原则内容本身和它们的执行上形成一致和趋同；对于全球而言，将原则付诸实践、寻求人工智能伦理规范（软法）和立法（硬法）的协同是下一步需要做的重要工作；目前，这些非立法规范是否会在政策层面产生影响，或者它们是否会影响个体实践者和决策者，还拭目以待。〔10〕

艾菲·瓦耶纳教授等提出的执行问题、实效有待观察问题，在他们研究成果发布前后，已经有研究者进行了相应的探索并给出了回答：基本无效。“算法观察”（Algorithm Watch）是一个位于德国柏林和瑞士苏黎世的非政府、非营利组织，其宗旨在于为一个算法和人工智能在其中是加强而不是削弱正义、人权、民主和可持续发展的世界而奋斗。〔11〕该组织于 2019 年发布了“全球人工智能伦理指南清单”，对全球范围内旨在为以合乎伦理的方式开发和实施自动决策系统确立原则的框架和指南进行汇编。该清单于 2020 年 4 月 28 日更新后，有超过 160 个指南包含在其中，涉及中国的有：北京智源人工智能研究院联合北京大学、清华大学、中国科学院自动化研究所、中国科学院计算技术研究所、新一代人工智能产业技术创新战略联盟等高校、科研院所和产业联盟共同发布的《人工智能北京共识》（2019 年 5 月 25 日）。中国人工智能产业联盟发布的《人工智能行业自律公约（征求意见稿）》（2019 年 5 月 31 日）。〔12〕国家新一代人工智能治理专业委员会发布的《新一代人工智能治理原则——发展负责任的人工智能》（2019 年 6 月 17 日）。

显然，“算法观察”编撰的清单，没法囊括世界范围内所有以指南、原则、准则、倡议、自律公约等形式呈现的人工智能伦理规范。一是此类软法在数量上难以计数，某个时间节点上的收集不见得完整；二是此类软法在生成上不受主体、程序等的严格限制，非常快捷、便利，故收集的时间节点以后很快又会有新的软法出现。以中国为例，2017 年 7 月 8 日国务院发布《新一代人工智能发展规划》，其中就多处提及人工智能伦理规范建设的意义、重点和时间线，尽管其本身并未直接提出具体的伦理规范。而 2018 年 1 月 18 日中国电子技术标准化研究院发布的《人工智能标准化白皮书（2018 年版）》已经明确，人工智能的发展应当遵循人类利益原则、透明度原则和权责一致原则等伦理要求，虽然其相对粗糙、简略。这是在“算法观察”收集或更新的时间节点之前的情况。而在该时间节点以后，我国的国家新一代人工智能治理专业委员会又于 2021 年 9 月 25 日发布了《新一代人工智能伦理规范》，比较系统地提出了“增进人类福祉”“促进公平公正”“保护隐私安全”“确保可控可信”“强化责任担当”“提升伦理素养”等六项基本伦理规

〔10〕 See Anna Jobin, Marcello lenca & Effy Vayena, *The Global Landscape of AI Ethics Guidelines*, Nature Machine Intelligence, Vol. 1, 2019, pp. 389 - 397, available at https://doi.org/10.1038/s42256 - 019 - 0088 - 2, last visited on Aug. 6, 2024.

〔11〕 关于该组织的介绍参见 https://algorithmwatch.org/en/，最后访问时间：2024 年 8 月 6 日。

〔12〕 “算法观察”列入清单的是《人工智能行业自律公约（征求意见稿）》，但中国人工智能产业联盟很快公开正式成稿的公约，并发出签署倡议。参见中国人工智能产业联盟：《关于签署〈人工智能行业自律公约〉的倡议》，载微信公众号“人工智能产业发展联盟 AIIA”，2019 年 8 月 8 日。

范，又系列地提供了管理、研发、供应和使用规范。

然而，没法囊括并不是问题的关键所在，因为“算法观察”于2019年发布此项研究初步结论时就已经指出会有更多的指南，〔13〕而该组织的观察结论则是更加重要、更引人瞩目的。2019年，“算法观察”发布的《〈人工智能伦理指南〉：有约束力的承诺还是装点门面?》一文指出，彼时收集的83个指南之中，绝大多数都是行业主导的，因为自愿的自我监管是非常受欢迎的避免政府监管的手段。德国的思爱普（SAP），美国的赛捷（Sage）、脸书（Facebook）、谷歌（Google）等公司既规定了内部原则，也公布了一般指南。其中一部分是公司作为产业联盟——如“人工智能伙伴关系”（Partnership on AI）〔14〕——成员发布的，一部分是行业协会领导发布的。最为重要的是很少有指南附带治理或者监督机制，可以确保这些自愿承诺得到遵守和实施。〔15〕2020年，“算法观察”数据库中指南数量超过160个，或者是自愿承诺的，或者是建议性的，其中只有10个是有实施机制的。即使是世界上最大的工程师专业协会“电气与电子工程师协会”（Institute of Electrical and Electronic Engineers，以下简称IEEE）〔16〕制定的伦理指南，在很大程度上也是没有实效的，因为脸书、谷歌和推特（Twitter）等大型科技公司都没有执行这些指南，尽管它们的许多工程师和开发人员都是IEEE的成员。〔17〕

“算法观察”两份报告的结论对人工智能伦理指南的实效基本持否定态度。而且，这并不是其一家之言。此前，来自美国北卡罗来纳州立大学的研究人员进行了一项研究，他们找了63名软件工程专业学生和105名软件开发专业人员，并将其分成两组。一组是明确指示其使用美国计算机协会（Association of Computing Machinery，以下简称ACM）制定的伦理规范，另一组是对照组（control group），即没有看到ACM伦理规范。研究人员让被测试者回答十一个有着简单情境介绍的选择题，每个题都涉及伦理决策。研究结论是：无论是学生还是专业开发人员，看过和没有看过伦理规范的被测试人员对问题的回答，没有统计学意义上的显著差异。〔18〕这表明伦理规范并不会对软件开发产生实质性影响。人工智能伦理规范基本都是由技术专家（为主）、法律专家（为辅）研究和制定的，其希望通过技术的、设计的专业知识来应对人工智能/机器学习的伦理问题，并将设计作为伦理审查的中心，〔19〕因此，上述针对软件工程专业学生和软件开发专

〔13〕 See Algorithm Watch, “*Ethical AI Guidelines*”: *Binding Commitment or Simply Window Dressing*?, available at https://algorithmwatch.org/en/ethical-ai-guidelines-binding-commitment-or-simply-window-dressing/, last visited on Aug. 6, 2024. 艾菲·瓦耶纳教授等也提及，鉴于人工智能指南发布速度很快，在其研究完成之后又有可能出现新的文件。See Anna Jobin, Marcello lenca & Effy Vayena, *The Global Landscape of AI Ethics Guidelines*, Nature Machine Intelligence, Vol. 1, 2019, p. 397, available at https://doi.org/10.1038/s42256-019-0088-2, last visited on Aug. 6, 2024.

〔14〕 关于“人工智能伙伴关系”联盟，可以参见其官方网站 https://partnershiponai.org/，最后访问时间：2024年8月6日。

〔15〕 See *Algorithm Watch*, “*Ethical AI Guidelines*”: *Binding Commitment or Simply Window Dressing*?, available at https://algorithmwatch.org/en/ethical-ai-guidelines-binding-commitment-or-simply-window-dressing/, last visited on Aug. 6, 2024.

〔16〕 关于IEEE，可以参见其官方网站 https://www.ieee.org/，最后访问时间：2024年8月6日。

〔17〕 See Algorithm Watch, *In the Realm of Paper Tigers-exploring the Failings of AI Ethical Guidelines*, available at https://algorithmwatch.org/en/ai-ethics-guidelines-inventory-upgrade-2020/, last visited on Aug. 6, 2024.

〔18〕 See Andrew McNamara, Justin Smith & Emerson Murphy-Hill, *Does ACM's Code of Ethics Change Ethical Decision Making in Software Development*?, available at https://doi.org/10.1145/3236024.3264833, last visited on Aug. 6, 2024.

〔19〕 See Daniel Greene, Anna Lauren Hoffman & Luke Stark, *Better*, *Nicer*, *Clearer*, *Fairer*: *A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning*, available at https://hdl.handle.net/10125/59651, last visited on Aug. 6, 2024.

业人员的测试结果验证了伦理规范的“实效赤字”问题。人工智能伦理规范的大量产出背后潜藏着较为可观的投入和支出，但其收入即实效远远少于成本，因此本文称其为“实效赤字”。

那么，人工智能伦理规范是否真的如上述测试所表现的那样“实效性几近于零”呢？〔20〕本文并不以为然。首先，人工智能伦理规范并不纯粹是被束之高阁的。科技巨头发布的此类软法，或多或少地对其自身产生拘束作用。例如，谷歌公司自 2018 年发布《人工智能原则》（AI Principles）以来，〔21〕每一年都会发布更新报告，而在报告中，其会向公众说明自己在践行原则方面的努力、取得的进步、获得的教训。2023 年报告就提到：“这是我们每年发布的《人工智能原则》进展报告的第五版，通过年度报告，我们始终如一对我们如何将原则付诸实践保持透明。我们于 2018 年首次发布《人工智能原则》，旨在分享公司的技术伦理章程，并使我们对如何负责任地研究和开发人工智能保持责任心。生成式人工智能也不例外。在本报告中，我们将详细分享在研究和开发包括 Gemini 家族模型在内的新型生成式人工智能模型过程中所采用的合乎原则的方法。原则只有在付诸实践后才能发挥实效。这就是我们发布这份年度报告——包括学到的艰难教训——的原因，目的是让人工智能生态系统中其他人能够借鉴我们的经验。”〔22〕谷歌公司的年度报告本身的真实性、其在报告中反映的践行原则之努力在多大程度上执行了其原则，还缺乏中立的、客观的、完整的评价。谷歌公司在 2019 年宣布不再与美国国防部续约，停止向其提供人工智能的帮助以分析海外军事无人机监控录像，〔23〕也被认为是在其员工抗议此项目引发伦理争议和忧虑的情况下作出的决定，而不是自愿履行其人工智能伦理规范的结果。〔24〕尽管如此，年度报告及其公开至少意味着该公司愿意向公众汇报其在人工智能伦理规范执行方面的进步，也愿意将自身置于广泛的监督和随时可能出现的批评之下。

其次，尽管人工智能系统的应用实践在合乎伦理规范方面表现较差，但在一些原则——如隐私、公平、可解释性——的应用上还是有着较为明显的进步。例如，世界范围内已经开发了许多保护隐私的数据集使用和学习型算法使用技术，这些技术通过使用密码、隐私区分或随机隐私等方法，使人工智能系统的“视域”“变暗”。不过，吊诡的是，人工智能花了数年时间取得的巨大进步，恰恰是因为有大量的数据（包括个人数据）可用。而这些数据都是具有隐私侵犯性的社交

---

〔20〕 See Thilo Hagendorff, *The Ethics of AI Ethics: An Evaluation of Guidelines*, Minds & Machines, Vol. 30, 2020, p. 108, available at https://doi.org/10.1007/s11023-020-09517-8, last visited on Aug. 6, 2024.

〔21〕 See Google, *AI Principles*, available at https://ai.google/responsibility/principles/, last visited on Aug. 6, 2024.

〔22〕 Google, *AI Principles Update 2023*, p. 6, available at https://ai.google/static/documents/ai-principles-2023-progress-update.pdf, last visited on Aug. 6, 2024.

〔23〕 See Kate Conger, *Google Is Helping the Pentagon Build AI for Drones*, available at https://gizmodo.com/google-is-helping-the-pentagon-build-ai-for-drones-1823464533#:~:text=Google%20has%20partnered%20with%20the,they%20learned%20of%20Google's%20involvement, last visited on Aug. 6, 2024; Nick Statt, *Google reportedly leaving Project Maven military AI program after 2019*, available at https://www.theverge.com/2018/6/1/17418406/google-maven-drone-imagery-ai-contract-expire, last visited on Aug. 6, 2024.

〔24〕 See Thilo Hagendorff, *The Ethics of AI Ethics: An Evaluation of Guidelines*, Minds & Machines, Vol. 30, 2020, p. 109, available at https://doi.org/10.1007/s11023-020-09517-8, last visited on Aug. 6, 2024. 需要注意的是，谷歌公司即便退出该项目，也在为其参与此项目辩护，称其开发的技术只是“标记图像供人类审查”，并且“仅用于非攻击性用途”。Nick Statt, *Google reportedly leaving Project Maven military AI program after 2019*, available at https://www.theverge.com/2018/6/1/17418406/google-maven-drone-imagery-ai-contract-expire, last visited on Aug. 6, 2024. 言外之意，其并没有违反谷歌《人工智能原则》中“我们不会设计和应用人工智能于……主要目的或实施是造成或直接促进对人伤害的武器或其他技术”的要求。

媒体平台、智能手机应用程序以及有着无数传感器的物联网设备收集的。〔25〕

再者，人工智能伦理规范还会在“微观伦理”层面上得到体现。虽然在宏观层面上，由抽象、含糊词句形成的人工智能伦理规范的实施乏善可陈，但是，在人工智能伦理问题引起广泛重视的情况下，从伦理到“微观伦理”（如技术伦理、机器伦理、计算机伦理、信息伦理、数据伦理）的转变也在发生，并且有很好的实效。例如，缇姆尼特·吉布鲁（Timnit Gebru）的研究团队提出了标准化的数据表，列出不同训练数据集的属性，以便机器学习训练者可以检查特定数据集在多大程度上最适合他们的目的，数据集创建时的初衷是什么，数据集由什么数据组成，数据是如何收集和预处理的等等。由此，机器学习训练者可以在选择训练数据集时作出更明智的决定，从而使机器学习变得更公平、更透明并避免算法歧视。〔26〕这一在“微观伦理”上的工作成果，受到了微软、谷歌和国际商用机器公司（IBM）的青睐，开始在内部试用数据集的数据表。“数据营养项目”（Data Nutrition Project）〔27〕采纳了部分成果，“人工智能伙伴关系”也在建立类似的数据表。〔28〕

最后，在原理上，软法的“执行实效”通常是需要一段时间才能显现出来的。软法的显著特点在于说服，而不在于强制，说服的时间成本自然是不可避免的。然而，从2016年还很少有人工智能伦理规范，〔29〕到现在全球范围内如此多的政府、非政府组织、大型企业等主体发布或更新此类规范，已经表明正在形成一种道德共识，即人工智能的开发、利用应当承担起伦理责任。而这个道德共识，美国哲学家卡尔·波普尔（Karl Popper）认为科学界早在核武器和生化武器问题上就已经有了：承认存在一系列特定的威胁，必须准备一批特定的人、一套特定的工具和一组特定的观念以应对威胁。〔30〕从这个角度看，人工智能伦理规范至少已经获得了“推介实效”，或许其会像企业社会责任一样，后者花了几十年的时间，才部分地摆脱了“洗绿”或“洗白”的粉饰名声，制定了许多公司必须遵循的全球标准。〔31〕当然，这最后一点并不希望以偷换概念的方式，把本文关注的“执行（实施）实效”主题延伸到“推介实效”，只是希望在观察研究“执行（实施）实效”时增添一个“时间—过程”维度。

---

〔25〕 See Thilo Hagendorff, *The Ethics of AI Ethics: An Evaluation of Guidelines*, Minds & Machines, Vol. 30, 2020, pp. 109 - 110, available at https://doi.org/10.1007/s11023 - 020 - 09517 - 8, last visited on Aug. 6, 2024.

〔26〕 See Thilo Hagendorff, *The Ethics of AI Ethics: An Evaluation of Guidelines*, Minds & Machines, Vol. 30, 2020, p. 111, available at https://doi.org/10.1007/s11023 - 020 - 09517 - 8, last visited on Aug. 6, 2024.

〔27〕 关于该组织，可以参见其官方网站 https://datanutrition.org/，最后访问时间：2024年8月6日。

〔28〕 See Timnit Gebru et al., *Datasheets for Datasets*, available at https://doi.org/10.1145/3458723, last visited on Aug. 6, 2024.

〔29〕 See Meredith Whittaker et al., *AI Now Report* 2018, p. 32, available at https://ainowinstitute.org/publication/ai-now-2018 - report - 2, last visited on Aug. 6, 2024. 2017年成立的“人工智能现在研究所”（AI Now Institute），作为一个独立组织，旨在就人工智能进行诊断和政策研究。关于该研究所，可以参见其官方网站 https://ainowinstitute.org/about，最后访问时间：2024年8月6日。

〔30〕 See Daniel Greene, Anna Lauren Hoffman & Luke Stark, *Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning*, available at https://hdl.handle.net/10125/59651, last visited on Aug. 6, 2024.

〔31〕 See Algorithm Watch, “*Ethical AI Guidelines*”: *Binding Commitment or Simply Window Dressing?*, available at https://algorithmwatch.org/en/ethical-ai-guidelines-binding-commitment-or-simply-window-dressing/, last visited on Aug. 6, 2024.

三、“实效赤字”原因及为什么仍然需要软法

（一）人工智能伦理规范“实效赤字”原因

发展迄今未至十年的人工智能伦理规范，实效即便不能简单地归为零，也在总体上没有达到解除或极大缓解人们对人工智能伦理的顾虑、担忧的目标。其原因主要有以下七个方面。

第一，人工智能伦理规范的非强制执行性。“人工智能现在研究所”2017 年的报告指出，伦理规范构成柔性治理的一种形式，是对硬性的传统政府监管和法律监督的替代，且在人工智能领域逐渐得到积极发展，但其有着现实局限性。关键局限在于其假定企业、行业会自愿采用和遵守。〔32〕2018 年的报告继续指出：“尽管我们已经看到制定此类规范的热潮，……但是我们没有看到强有力的监督和问责，来保证这些伦理承诺的兑现。”〔33〕软法这一与生俱来的、阿喀琉斯之踵般的致命缺陷，成了公认的人工智能伦理规范实效不足的根本原因。〔34〕

第二，人工智能伦理规范的抽象性、模糊性。人工智能伦理规范并不是针对人工智能的，而是针对研究、开发与应用人工智能的人类的，其目标是要求研究者、开发者与应用者遵循一定的规范，以使人工智能带来的伦理风险降到最低。因此，该规范越是具体、明确，就越容易得到遵守；否则，就很难落实或者存在各种有争议的落实。然而，现今的人工智能伦理规范基本是抽象的、模糊的，绝大多数指南除了用“人工智能”一词外，从不用或很少用更为具体的术语。而人工智能只是一个集合术语，指向范围极广的一系列技术或一个规模巨大的抽象现象。没有一个伦理指南令人瞩目地深入到技术细节，这表明在研究、开发和应用的具体情境与一般的伦理思维之间存在很深的鸿沟。〔35〕尽管抽象性、模糊性可能被认为是不可避免和必要的，因为人工智能的应用极其广泛、发展快且未来的发展轨迹并不确定，〔36〕但是，前述在“微观伦理”层面上的成功例子表明相对具体化、精细化是可能的。

第三，人工智能伦理规范的分散、混乱与叠床架屋。如同其他软法一样，人工智能伦理规范的制定主体包括政府、企业、企业联盟、行业团体、非政府公益组织、研究机构等，这就形成了众多形式的伦理规范。而前文提及的艾菲·瓦耶纳教授等研究结果表明，各种文件使用的人工智能伦理原则术语或许是相同的，但实质内容存在诸多分歧。即便是最普遍的透明原则，在涉及解释（沟通、披露）、为什么透明、透明适用的领域以及实现透明的方式等方面，都有

---

〔32〕 See Alex Campolo et al., *AI Now Report 2017*, p. 32, available at https://ainowinstitute.org/publication/ai-now-2017-report-2, last visited on Aug. 6, 2024.

〔33〕 Meredith Whittaker et al., *AI Now Report 2018*, p. 29, available at https://ainowinstitute.org/publication/ai-now-2018-report-2, last visited on Aug. 6, 2024.

〔34〕 See Gary E. Marchant, “*Soft Law*” *Governance of Artificial Intelligence*, available at https://escholarship.org/content/qt0jq252ks/qt0jq252ks_noSplash_1ff6445b4d4efd438fd6e06cc2df4775.pdf, last visited on Aug. 6, 2024; Wendell Wallach & Gary Marchant, *Toward the Agile and Comprehensive International Governance of AI and Robotics*, Proceedings of the IEEE, Vol. 107, 2019, p. 506, available at https://ieeexplore.ieee.org/document/8662741, last visited on Aug. 6, 2024.

〔35〕 See Thilo Hagendorff, *The Ethics of AI Ethics: An Evaluation of Guidelines*, Minds & Machines, Vol. 30, 2020, p. 111, available at https://doi.org/10.1007/s11023-020-09517-8, last visited on Aug. 6, 2024.

〔36〕 See Gary E. Marchant, “*Soft Law*” *Governance of Artificial Intelligence*, available at https://escholarship.org/content/qt0jq252ks/qt0jq252ks_noSplash_1ff6445b4d4efd438fd6e06cc2df4775.pdf, last visited on Aug. 6, 2024.

着重大差异。〔37〕“不同的人工智能软法项目和提案出现了令人困惑的激增，造成人工智能治理的混乱和叠床架屋。人工智能领域的行动者很难评估和遵守所有这些不同的软法要求。”〔38〕

第四，人工智能伦理规范自愿遵守的动力不足。人工智能伦理规范的非强制执行性，意味着其寄希望于人工智能研究者、开发者和应用者可以自愿遵守。人工智能伦理规范是人类长期以来的伦理关切在人工智能领域里的投射，新兴人工智能技术之所以引起广泛的伦理担忧和焦虑，〔39〕表明伦理共识的普遍存在。尽管如此，人工智能给许多领域主体带来的经济利益——无论是财富增长还是成本减少——是如此巨大，基于价值或原则的伦理关切难以胜过经济逻辑。在商业领域，速度就是一切，跳过伦理关切就相当于走上一条最少阻力的道路。〔40〕在这个意义上，伦理良币有可能转变为竞争劣币而被市场淘汰。

第五，人工智能伦理规范的合规悖论。人工智能伦理规范的遵守往往需要在技术上有所体现，尤其是在设计环节。所以，“合乎伦理的人工智能系统”（ethically aligned AI system）〔41〕或“合乎伦理的设计”（ethically aligned design）〔42〕等概念应运而生。然而，正如前文所揭，在有些情况下，合乎伦理的设计（如保护隐私的技术）所需要的大量数据，正是在涉嫌违反伦理原则（如侵害隐私）的情况下收集的。〔43〕这个悖论是否广泛存在尚未有充分的实证研究数据，但人工智能先违反伦理原则进行充分发展而后再考虑如何合乎伦理的情况大概率是存在的。

第六，人工智能伦理规范影响力的社会系统论困境。德国斯图加特大学教授蒂洛·哈根道夫（Thilo Hagendorff）除了揭示人工智能伦理规范在实施中受到冷落的经济逻辑以外，还引用三位著名社会学家的理论从宏观社会学角度进行了分析。其指出，德国社会学家、风险社会理论的开拓者之一乌尔里希·贝克（Ulrich Beck）曾经有一个非常形象的比喻，当今社会的伦理“发挥的作用就如同在洲际航行的飞机上配置了自行车刹车”，这在人工智能情境中尤其适用。根据另一德国社会学家尼克拉斯·卢曼（Niklas Luhmann）的系统论，现代社会由众多不同的社会系统构成，每个系统都有自己的工作代码和沟通媒介。结构耦合可以让一个系统的决策影响另一些系统，但其影响有限，难以改变社会系统的整体自治。法国社会学家皮埃尔·布尔迪

---

〔37〕 See Anna Jobin, Marcello lenca & Effy Vayena, *The Global Landscape of AI Ethics Guidelines*, Nature Machine Intelligence, Vol. 1, 2019, p. 391, available at https://doi. org/10. 1038/s42256 - 019 - 0088 - 2, last visited on Aug. 6, 2024.

〔38〕 Gary E. Marchant, “*Soft Law*” *Governance of Artificial Intelligence*, available at https://escholarship. org/content/qt0jq252ks/qt0jq252ks_noSplash_1ff6445b4d4efd438fd6e06cc2df4775. pdf, last visited on Aug. 6, 2024.

〔39〕 最近在中国引起普遍讨论的就是百度旗下的“萝卜快跑”无人驾驶出租车可能对出租车就业市场的冲击。尽管这个冲击还远未到来，但“科技的初衷是让人类生活得更好，现实是让底层人吃不饱”的舆论恐慌和焦虑已经掀起。参见敖阳利：《“萝卜快跑”跑出科技恐慌?》，载《中国财经报》2024 年 7 月 16 日，第 3 版。

〔40〕 See Thilo Hagendorff, *The Ethics of AI Ethics: An Evaluation of Guidelines*, Minds & Machines, Vol. 30, 2020, p. 108, available at https://doi. org/10. 1007/s11023 - 020 - 09517 - 8, last visited on Aug. 6, 2024.

〔41〕 See Ville Vakkuri et al., *ECCOLA-A Method for Implementing Ethically Aligned AI Systems*, The Journal of Systems & Software, Vol. 182, 2021, pp. 1 - 16, available at https://doi. org/10. 1016/j. jss. 2021. 111067, last visited on Aug. 6, 2024.

〔42〕 See Yueh-Hsuan Weng & Yasuhisa Hirata, *Ethically Aligned Design for Assistive Robotics*, available at https://ieeexplore. ieee. org/document/8535889, last visited on Aug. 6, 2024.

〔43〕 See Thilo Hagendorff, *The Ethics of AI Ethics: An Evaluation of Guidelines*, Minds & Machines, Vol. 30, 2020, pp. 109 - 110, available at https://doi. org/10. 1007/s11023 - 020 - 09517 - 8, last visited on Aug. 6, 2024.

厄（Pierre Bourdieu）也表示，所有这些系统都有自己的代码、目标价值以及经济资本或象征性资本，社会系统通过这些资本得以构建起来，并基于这些资本作出决策。这种自治在人工智能的工业、商业和科学里也显著存在。对这些系统的伦理干预只会在非常有限的范围内发挥作用。〔44〕

第七，人工智能发展压倒约束的宿命论。导致人工智能伦理规范出现“实效赤字”的根本原因在于，人类社会对待人工智能的基本立场是决定论或宿命论的（determinism）。人工智能伦理指南文件绝大多数都将人工智能叙述为推动世界发生历史性改变的力量，这个改变是不可避免的、影响深远的、会给人类带来巨大利益的，人类社会只能去回应、适应并为其风险和后果承担起责任。〔45〕例如，2018 年的《蒙特利尔宣言》提到：“人工智能形成了科学和技术的一个重大进步，它可以改善生活条件和健康、促进正义、创造财富、加强公共安全以及减轻人类活动对环境和气候的影响，从而产生可观的社会效益。”〔46〕我国国家互联网信息办公室于 2023 年 10 月发布的《全球人工智能治理倡议》也持类似的立场。人工智能是人类发展新领域。当前，全球人工智能技术快速发展，对经济社会发展和人类文明进步产生深远影响，给世界带来巨大机遇。在此决定论/宿命论的背景之下，不仅科技巨头如谷歌、脸书、百度、阿里巴巴等竞相推出新的人工智能应用程序，而且，各国都宣布参加人工智能竞赛，把人工智能视为在人类社会各领域解决问题的动力。〔47〕鉴于此，相当程度上对人工智能发展起约束作用的伦理规范，自然是如同飞机上的自行车刹车一样。

（二）人工智能为何仍然需要作为软法的伦理规范

以上种种，皆直接或间接地阻碍人工智能伦理规范实施、得到遵守，有些似乎是根本性的、无法扭转的。这是否意味着人工智能治理就不应该走软法之路？答案是否定的，因为人工智能发展本身的特点，注定不能单纯依靠硬法去防范其风险、减少其危害。以下是人工智能为什么仍然需要作为软法的伦理规范“参与治理”的五个主要理由，每个理由都会涉及硬法或硬性监管的不足、软法或柔性治理的优势。

第一，软法的灵活快捷性。几乎所有涉足人工智能领域的研究者都承认一个事实，即人工智能的发展速度惊人，并以同样惊人的速度对人类生活各个方面进行渗透，人类社会因此正在迅速发生难以预测未来确定图景的转型和变化，危害已经初露端倪，风险也悄然潜伏。更多由于前述公私领域普遍存在的经济逻辑的推动，这一动向似乎是决定性的、宿命的，如何控制和防范危害、风险也就因此转化为一个法律体系的“配速”问题（pacing problem）。早在 1986 年，美国

〔44〕 See Thilo Hagendorff, *The Ethics of AI Ethics: An Evaluation of Guidelines*, Minds & Machines, Vol. 30, 2020, p. 109, available at https://doi.org/10.1007/s11023-020-09517-8, last visited on Aug. 6, 2024.

〔45〕 See Daniel Greene, Anna Lauren Hoffman & Luke Stark, *Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning*, available at https://hdl.handle.net/10125/59651, last visited on Aug. 6, 2024.

〔46〕 该宣言全称是《为了人工智能负责任发展的蒙特利尔宣言》（Montreal Declaration for a Responsible Development of Artificial Intelligence 2018），关于《蒙特利尔宣言》的背景介绍，可以参见 https://montrealdeclaration-responsibleai.com/about/，最后访问时间：2024 年 8 月 6 日。

〔47〕 See Jascha Baries & Christian Katzenbach, *Global AI Race: Nations Aiming for the Top*, available at https://zenodo.org/records/1845399, last visited on Aug. 6, 2024.

技术评估办公室（US Office of Technology Assessment）就提及："技术变革曾经是一个相对缓慢而沉闷的过程，但现在其速度超过了管理该系统的法律结构的变化速度，这给国会带来了调整法律以适应技术变革的压力。"法律系统面临的配速问题体现在两个方面。其一，许多既有法律框架建立在社会和技术的静态观而不是动态观基础上；其二，法律机构（立法、监管和司法机关）调整适应技术变革的能力正在减速。〔48〕配速问题的存在，加剧了对人工智能危害和风险的担忧。相比正式立法程序的官僚性、正式性、繁琐性，软法的制定与更新就灵活、快捷许多。如前所述，人工智能伦理规范制定主体多样，没有严格的程序限制，比较容易将人们的伦理关切及时转化为引导人工智能研究、开发和应用的原则。尽管这些原则抽象、含糊、多义又缺乏强制约束力，但公开宣布的伦理规范的事实约束力并不是完全归零的。

第二，软法的多样适配性。"人工智能"只是一个抽象用词，其所指向的是范围极广、种类繁多、层出不穷、不计其数的技术，每个技术都有可能会带来比较特定的伦理关切，也需要在技术上找到各自特定的应对方案。例如，美国阿肯色州一次医疗保健系统算法执行，对糖尿病患者或脑瘫患者产生负面影响，致使他们能够获得的医疗保健大幅削减；YouTube 使用的推荐算法由谷歌开发，其依靠反馈循环，旨在优化用户的观看时间，但在预测人们喜欢看什么内容的同时，也决定了人们看的内容，以至于助长了耸人听闻的虚假视频以及阴谋论；谷歌曾经出现一种偏见，凡是搜索的名字是历史上有过的黑人名字，就会在搜索结果上暗示有犯罪记录，而搜索的是历史上的白人名字，搜索结果就会相对中性；〔49〕而人工智能/机器学习的人脸识别技术曾经被指责对有色人种（尤其是黑人）识别不够，微软公司就开始宣传其在"包容性"方面的努力，以改善不同肤色的面部识别功能，但也有评论者认为这样的技术改进会对黑人社区更为不利，因为黑人社区在历史上就是监控技术的靶子。〔50〕诸如此类涉及人工智能伦理引起关切的例子，足以表明全面的、以硬法为基础的统一监管，很有可能陷入无法适应多样化技术、多样化伦理要求的困境。甚至，监管有时是反市场的、对小企业不利的，其形成的障碍只有大企业才能克服。〔51〕相比之下，软法主要不是由政府制定的，企业、行业组织、企业联盟、非政府组织等都可以针对更加具体特定的技术伦理问题制定相应的、更加适配的指南。

第三，软法的合作试验性。尽管软法确有分散、混乱、叠床架屋的特性，但也由于存在多种软法方案，就给人工智能的研究、开发和利用带来了选择试验的空间，利益相关者之间——包括又不限于政府与企业之间——有时候会形成合作的关系，而不是对立的关系。〔52〕这同以往政府

---

〔48〕 See Gary E. Marchant, *The Growing Gap between Emerging Technologies and the Law*, in Gary E. Marchant, Braden Allenby & Joseph Herkert eds., The Growing Gap between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem, Springer, 2011, pp. 22 - 23.

〔49〕 See Jeremy Howard & Sylvain Gugger, *Deep Learning for Coders with fastai & PyTorch*, O'Reilly Media, Inc., 2020, pp. 95 - 96.

〔50〕 See Daniel Greene, Anna Lauren Hoffman & Luke Stark, *Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning*, available at https://hdl.handle.net/10125/59651, last visited on Aug. 6, 2024.

〔51〕 See Gary E. Marchant & Carlos Ignacio Gutierrez, *Soft Law 2.0: An Agile and Effective Governance Approach for Artificial Intelligence*, 24 Minnesota Journal of Law, Science & Technology 382 (2023).

〔52〕 See Gary E. Marchant, "*Soft Law*" *Governance of Artificial Intelligence*, available at https://escholarship.org/content/qt0jq252ks/qt0jq252ks_noSplash_1ff6445b4d4efd438fd6e06cc2df4775.pdf, last visited on Aug. 6, 2024.

与企业的监管对立、企业与企业之间的竞争对立是不同的。在这种合作的关系之中，也有相互学习、相互受益的元素。例如，前文提及谷歌公司在发布《人工智能原则》2023 年度报告时宣称其也意在分享研究开发新模型时应用原则的经验和教训。〔53〕在人工智能伦理规范推进方面发挥巨大作用的机构之一是全球电气与电子工程师的联合组织 IEEE。其发起的全球自动与智能系统伦理倡议，旨在解决由自动系统、智能系统的开发和传播引起的伦理问题。它确定了 120 个关键问题，并提出了解决这些问题的建议供企业选择。〔54〕人工智能——具体到特定场景的特定技术——的研究、开发、利用如何才能更好地符合伦理规范，或者，反言之，什么样的具体、细致的伦理规范适合于特定场景的特定人工智能技术，并不是有着确定答案的问题，也不是单凭某个专业团队就能够提出最佳方案的问题，这是需要技术专家、法律专家等合作探索的，也是需要不断地进行试验的。而这是硬法和硬性监管所无法达到的。

第四，软法的事实压力性。软法虽然没有法律上的约束力，但如果其内容在本质上有着广泛的共识，具有非常强的说服力，那么，个人和组织选择不遵守软法必定需要承受事实上存在的认同压力。当这种认同压力足以压倒不遵守可能带来的利益时，认同压力就会转化为事实上的约束力。因此，“对于伦理关切的研究表明，多种框架、观念、定义及其组合为组织创造了一系列供其选择的复杂方案。当问题的重要程度和组织能够得到的支持还不确定的时候，众多指南让组织必须承受针对其工作流程的批评。……选择一个工作流程伦理指南，为组织的内部和外部利益相关者评价该组织的应用程序产品提供了底线”〔55〕。

第五，软法的跨国适用性。人工智能的研究、开发、利用是世界性的、跨国界的，尤其是它在互联网上或者通过互联网的利用；人工智能所掀起的伦理关切和担忧也是世界性的、跨国界的。即便是某个平台、某家企业或某个应用程序被曝有特定的人工智能伦理失范的风险或丑闻，并不意味着它的影响只限于平台、企业所登记注册的国家，也并不意味着此类技术的伦理失范风险或丑闻不会在别的国家、别的平台、别的企业或别的应用程序中出现。例如，微软支持的 OpenAI 公司开发的 ChatGPT 仅仅上市两个多月后，类似应用程序带来的剽窃、欺诈和错误信息传播等风险就受到了关注，欧盟内部市场专员蒂埃里·布雷顿（Thierry Breton）在接受路透社专访时提到制定全球标准的紧迫性。〔56〕传统硬法、硬性监管主要限于主权国家或基于条约的区域性国际组织的领土管辖范围内，其之所以具备法律上的约束力，就是因为其得到主权国家基础规范或区域性国际组织基础条约的授权与认可。因此，若要在全球范围内应对人工智能伦理风

〔53〕 See Google, *AI Principles Update 2023*, p. 6, available at https://ai.google/static/documents/ai-principles-2023-progress-update.pdf, last visited on Aug. 6, 2024.

〔54〕 See Raja Chatila & John Havens, *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*, in Maria Isabel Aldinhas Ferreira et al. eds., Robotics and Well-being, Springer, 2019, pp. 11-16, available at https://doi.org/10.1007/978-3-030-12524-0, last visited on Aug. 6, 2024.

〔55〕 Robert Hobbs, *Integrating Ethically Align Design into Agile and CRISP-DM*, p. 2, available at https://ieeexplore.ieee.org/document/9401899, last visited on Aug. 6, 2024.

〔56〕 See Foo Yun Chee & Supantha Mukherjee, *Exclusive: ChatGPT in Spotlight as EU's Breton Bats for Tougher AI Rules*, available at https://www.reuters.com/technology/eus-breton-warns-chatgpt-risks-ai-rules-seek-tackle-concerns-2023-02-03/, last visited on Aug. 6, 2024. 尽管这则报道提及的更强硬的人工智能规则是前文所述欧洲《人工智能法案》，但人工智能研究、开发、利用及其风险的全球性是共通的。

险，跨越国界或者区域界限的软法/伦理规范在人工智能领域普遍推广，应该是可选的方案。

当然，在互联网经济、全球经济的生态之中，大型科技公司欲将业务拓展至其注册国以外的市场，肯定会关注并遵守该市场所在法律辖区的法律（硬法）系统。由此，像欧盟这样的跨国法律辖区，其制定的硬法如《通用数据保护条例》（GDPR）和最新的《人工智能法案》实际上也有为全球制定标准的意义，产生了所谓的“布鲁塞尔效应”。〔57〕但是，这个效应毕竟在两个意义上是间接的。其一，它只是会影响其他主权国家如中国或美国的立法，通常不会被后者照抄；其二，它只是会对有意进入欧盟市场的科技公司产生约束力，对其他规模较小且无意国际市场的科技公司的人工智能研发利用没有直接约束力。而人工智能伦理规范应该预期会在全球范围内达成更多共识，会越过主权国家或欧盟等区域性组织法律（硬法）管辖的界限，以发挥其效用，〔58〕尽管现在还不能如愿展现实效。

四、人工智能软法的实施机制

一方面，人工智能伦理规范有其兴起、存在的原因和独特价值，已经开始有凝聚共识、普遍认可等的“推介实效”；但是，另一方面，人工智能的研发、利用过程似乎还远没有受软法性质的伦理规范的切实影响，介入其中的专业人员还没有将伦理规范与程序设计紧密地结合起来，以至于许多人工智能的新产品、新应用时不时会引起对其所带来的伦理风险的普遍关注。那么，究竟如何才能让人工智能伦理规范落到实处，从事实压力转变为事实约束力，与相应的硬法合作，共同完成应对人工智能伦理风险挑战的使命呢？软法如何有效实施的这一命题，可以从中获得哪些普遍的启示和结论呢？由于软法在原理上不具有强制执行力，不能通过强力去直接实施，故本文在此讨论的是间接地推进软法实施需要哪些类型的机制。

（一）软法促进的组织机制

软法的实施是一个需要不断自我更新、获取共识、得到驱动的渐进过程，对未来不确定风险承担预防和治理功能的人工智能软法，尤其如此。在这个过程中，缺少强有力的、持续坚定从事软法推进事业的组织，是难以想象的。从类型上而言，这样的组织可以是属于政府系列的，也可以是属于企业系列的，更可以是行业组织、企业合作联盟、第三方机构、研究团队等。〔59〕其中，

〔57〕 参见钱童心：《欧盟最强 AI 法案即将生效“布鲁塞尔效应”波及全球》，载《第一财经日报》2024 年 7 月 18 日，第 A01 版。

〔58〕 See Gary E. Marchant，“*Soft Law*” *Governance of Artificial Intelligence*，available at https://escholarship.org/content/qt0jq252ks/qt0jq252ks_noSplash_1ff6445b4d4efd438fd6e06cc2df4775.pdf，last visited on Aug. 6，2024.

〔59〕 政府系列的例子如欧盟委员会（2018 年发布“人工智能协调行动计划”）、英国上议院（2018 年发布建议人工智能伦理规范的报告）；在企业系列，谷歌、微软、IBM 等都发布自己的人工智能原则；企业合作联盟如“人工智能伙伴关系”；行业团体如 IEEE；第三方机构如“未来生命研究所”（Future of Life Institute）。See Gary E. Marchant，“*Soft Law*” *Governance of Artificial Intelligence*，available at https://escholarship.org/content/qt0jq252ks/qt0jq252ks_noSplash_1ff6445b4d4efd438fd6e06cc2df4775.pdf，last visited on Aug. 6，2024. 前文提及的“算法观察”“人工智能现在研究所”也是第三方机构的例子。研究团队如前文提及的缇姆尼特·吉布鲁研究团队，人工智能伦理规范的研究团队往往与人工智能企业有着密切关联。2021 年被《财富》杂志誉为世界 50 位最伟大的领导者之一、2022 年被《时代》周刊评为最有影响力人物之一的缇姆尼特·吉布鲁，就是以博士后研究人员身份于 2017 年加入微软的人工智能公平、责任、透明和伦理实验室，2018 年至 2020 年，她又在谷歌领导人工智能伦理团队。

大型科技巨头——如微软、谷歌等——也有专门的人工智能伦理规范部门或团队。从功能上而言，这样的组织可以是持续制定和更新人工智能伦理规范的，可以是倡议全球人工智能领域研发者、利用者加盟共同遵守人工智能伦理规范的，可以是观察和监督人工智能伦理规范执行落实情况的，也可以是研究如何将人工智能伦理规范同具体技术的设计与应用结合起来的。

政府组织可能会纠结于人工智能行业发展与恪守伦理规范之间如何平衡，而在督促人工智能伦理规范落实方面有所懈怠。企业、行业组织或企业合作联盟可能会偏重装点门面、博得声誉而在人工智能伦理规范方面轻诺寡信，即便企业设立专门的人工智能伦理规范部门或团队以兑现自己的伦理承诺，该部门或团队的独立作用也不见得可以充分保障。例如，2020 年，谷歌解雇了缇姆尼特·吉布鲁，原因是她发表了一篇批评大语言模型的论文（该论文两年后大受欢迎）。由此引发的愤怒导致人工智能伦理部门又有几位高层领导人离职，并削弱了谷歌公司在负责任的人工智能问题上的可信度。〔60〕

相对而言，那些旨在密切观察人工智能风险、持续发布跟进研究报告、以监督和促进人工智能符合伦理规范为己任的组织，以及致力于将伦理规范融入人工智能研发、利用过程的研究团队（无论是否在企业内部），可信度和推动力会更高些。例如，有评论指出："关于人工智能伦理的报告并不匮乏，但其中的大部分都无足轻重，充斥着'公私合作'以及'以人为本'之类的陈词滥调。他们不承认人工智能造成的社会困境有多么棘手，也不承认解决这些困境有多么困难。'人工智能现在研究所'的新报告却非如此。它毫不留情地审视了科技行业在没有任何可靠和公平结果保证的情况下，竞相沿着人工智能的方向重塑社会。"〔61〕缇姆尼特·吉布鲁的研究团队发布的"数据集的数据表"，从 2018 年 3 月 23 日第一次发布到 2021 年 12 月 1 日，已经经历八个版本，被引用达 2263 次。〔62〕当然，软法促进的可靠、有力组织之存在，通常以此类组织生存和发展的制度——公共制度的或企业内部制度的——容许空间为前提。

（二）软法合规的压力机制

软法是事实压力性的，因为其以广泛的共识和说服的效力为基础，它只是给了行动者自愿遵守的选择。当软法在共同体中获得越来越多成员的认可，合乎、遵守软法就会获得所属共同体比较高的赞许，相反，违背软法即便不会给行动者带来强力制裁，也会使其承受非常大的压力，甚至是巨大的声誉损害及可能附随的经济损害。那么，有什么机制可以让这种压力足够强大呢？至少，可以有三个方面的重要机制：

一是舆论机制。对于在市场中求生存的企业而言，舆论对其、对其产品的评价毫无疑问是至关重要的，消费者通常会选择舆论评价高的产品。因此，在一个开放的舆论环境中，新闻媒体可以将科技企业及其人工智能产品是否符合伦理规范，甚至可以将其他企业是否在使用符合人工智

〔60〕 See Zoe Schiffer & Casey Newton, *Microsoft Lays off Team that Taught Employees How to Make AI Tools Responsibly*, available at https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs, last visited on Aug. 6, 2024.

〔61〕 Scott Rosenberg, *Why AI Is Still Waiting For Its Ethics Transplant*, available at https://www.wired.com/story/why-ai-is-still-waiting-for-its-ethics-transplant/, last visited on Aug. 6, 2024.

〔62〕 See Timnit Gebru et al., *Datasheets for Datasets*, available at https://doi.org/10.1145/3458723, last visited on Aug. 6, 2024.

能伦理规范的人工智能应用程序，作为评价体系的重要组成部分，从而形成足够强大的舆论压力，促使企业负责任地研发或利用人工智能。不过，舆论压力除了需要开放的舆论场以外，也还需要另外两个条件才能形成一定的效用：其一，消费者在乎符合人工智能伦理规范的企业及其产品；其二，消费者可以在竞争市场中选择到软法合规的企业及其产品。

二是对抗机制。对企业不在乎或疏忽人工智能伦理规范进行批评的舆论本身是一种形式的对抗。在此需要特别指出的是来自专业人员或利益相关者（stakeholder）的针对人工智能伦理风险而采取的对抗企业的行动，无论这些人员是在企业内部还是在企业外部。除了前文提及的谷歌公司在其员工抗议下停止与美国国防部合作军事人工智能项目的例子外，2019 年，谷歌公司还曾经在数千名员工的抗议下，解散了刚成立一个多星期的人工智能伦理委员会（正式名称是“先进技术外部咨询委员会”），因为其中来自公司以外的成员或其所属组织被指对跨性别者有不公评论、对气候变化持怀疑态度或者与人工智能的军事利用有关。〔63〕 2018 年，在时任美国总统特朗普将非法移民孩子与其家庭隔离的政策备受争议之际，微软与美国移民局在人脸识别技术上的合作，也受到了微软员工的抗议。〔64〕 2023 年 5 月 2 日至 9 月 27 日，代表 11500 名编剧的美国编剧协会因与电影电视制片人联盟发生劳资纠纷而组织了为期 148 天的罢工。罢工的一项诉求就是像 ChatGPT 这样的人工智能只应被用作一种帮助研究或推动脚本想法的工具，而不应该取代编剧。最终，罢工取得了胜利，双方达成的协议被认为是树立了一个对于人工智能的使用进行集体谈判的重要先例。〔65〕 这些来自专业人员或利益相关者的抗议是出于他们对人工智能伦理规范的认知和坚持，或者出于他们本身的利益受到人工智能发展的威胁，其主张不见得对，但确实是一种可以促进企业遵守软法的力量和机制。“越来越多富有意义的针对人工智能负责任发展的行动来自工人、共同体倡议者和组织者。”〔66〕 而这种力量和机制的存在，当然也需要依托于更广阔的企业与员工、企业与外部之间关系的制度空间、文化背景。

三是监督机制。就广义的监督而言，舆论、对抗同样属于监督机制。然而，软法合规监督还有其他更多样化的表现形式。早在 2015 年，盖瑞·马秦特教授就曾经和文德尔·瓦拉赫（Wendell Wallach）〔67〕 先生一起提议成立名为“治理协调委员会”的机构，目的不是重复或取代现有许多组织在人工智能治理方面的工作，而是如交响乐团指挥一样起到协调的作用。这个机构并未成立，但他们预设其应该承担的功能中有多项是与监督相关的，如监控和分析（认定人工

〔63〕 See Kelsey Piper, *Exclusive: Google Cancels AI Ethics Board in Response to Outcry*, available at https://www.vox.com/future-perfect/2019/4/4/18295933/google-cancels-ai-ethics-board, last visited on Aug. 6, 2024.

〔64〕 See Colin Lecher, *The Employee Letter Denouncing Microsoft's ICE Contract Now Has over 300 Signatures*, available at https://www.theverge.com/2018/6/21/17488328/microsoft-ice-employees-signatures-protest, last visited on Aug. 6, 2024.

〔65〕 See Molly Kinder, *Hollywood Writers Went on Strike to Protect Their Livelihoods from Generative AI. Their Remarkable Victory Matters for All Workers*, available at https://www.brookings.edu/articles/hollywood-writers-went-on-strike-to-protect-their-livelihoods-from-generative-ai-their-remarkable-victory-matters-for-all-workers/, last visited on Aug. 6, 2024.

〔66〕 Kate Crawford et al., *AI Now Report 2019*, p. 11, available at https://ainowinstitute.org/publication/ai-now-2019-report-2, last visited on Aug. 6, 2024.

〔67〕 文德尔·瓦拉赫先生曾经在卡内基国际事务伦理委员会工作，与他人共同主持“人工智能和平等倡议”项目。他也是耶鲁大学跨学科生物学中心技术与伦理研究荣誉主席，林肯应用伦理学中心学者，伦理学与新兴技术研究所研究员，黑斯廷斯中心高级顾问。关于其简介参见 https://www.carnegiecouncil.org/people/wendell-wallach，最后访问时间：2024 年 8 月 6 日。

智能治理计划实施的差距、重叠和不一致之处）、早期预警（指出正在出现的新问题）、评估（为治理计划实现目标的情况评分）、召集解决问题（召集利益相关者就特定问题商议解决方案）。〔68〕换言之，与之前所述的组织机制结合，若有相对独立的组织——无论是在企业内部设立类似伦理审查委员会的机构，还是在企业外部设立更为中立的社会组织——承担起监控、分析、预警、评估、共商方案等监督功能，就可以使人工智能伦理规范得到更好的落实。

（三）软法合规的激励机制

如果说软法合规的压力机制属于“减分项”，可能让人工智能研发者、利用者遭受声誉损失及附随的经济损失，〔69〕那么，软法合规的激励机制就是对应的“加分项”，可以使其得到更好的声誉及随之带去的更多经济利益。这样的激励机制相比压力机制似乎可以有更多展现形式。

一是认证机制。中立的第三方认证机构可以开设一个认证业务，对人工智能的研发和利用遵循一套特定伦理规范的企业或其他实体进行认证，并给予认证证书。

二是评价机制。中立的第三方组织，如高校科研机构或非政府社会组织，可以对人工智能研发者是否将人工智能伦理规范植入人工智能的研究和开发之中、人工智能利用者是否应用符合伦理规范的人工智能以及研发者和利用者的人工智能伦理规范合规程度等进行评价，评选出优秀的合规者。

三是购买机制。人工智能应用程序的研究、开发都会投入相当的成本，合乎伦理规范的或许会投入更多。对于软法合规企业或其他实体而言，认证、评优虽可以带来良好声誉，但并没有转化为实际的经济利益。相较之下，购买和使用合乎伦理规范的人工智能产品，尤其是获得认证或评优的人工智能产品，是让合规者获得实际利益的最直接方法。购买者，特别是政府采购方，若能将合乎伦理规范作为购买的前提条件，势必会带动有利于人工智能软法实施的市场导向。

四是合作机制。人工智能利益相关者——研究者、开发者、利用者——在倡议和推进人工智能伦理规范方面形成联盟或合作伙伴关系，相互之间给予支持和帮助，也更有利于建立公众信任，有助于人工智能软法得到诚信可靠的执行。

五是资助和发表机制。为人工智能的研发或利用提供投资或资助的机构、为人工智能研发成果提供发表平台的专业杂志，也同样可以将符合人工智能伦理规范作为一个条件或优先考虑的条件，以激励研发者、利用者遵守人工智能软法。

六是放松监管机制。政府负责人工智能发展监管的部门，对于在管理人工智能的研发或利用方面有一整套制度和配套机构、致力于人工智能软法合规的企业或其他实体，以及真正研发出或利用合乎伦理规范的人工智能产品的企业或其他实体，可以适当地放松监管力度。减少政府监管的利益被认为是人工智能软法获得成功的重要激励之一。

---

〔68〕 See Gary E. Marchant, “*Soft Law*” *Governance of Artificial Intelligence*, available at https://escholarship.org/content/qt0jq252ks/qt0jq252ks_noSplash_1ff6445b4d4efd438fd6e06cc2df4775.pdf, last visited on Aug. 6, 2024.

〔69〕 “应当牢记在心的是，除了真正的伦理动机以外，与经济相关的声誉损失的重要性不可低估。因此，反对不合伦理的人工智能项目的抗议也可以从经济逻辑予以解释。” Thilo Hagendorff, *The Ethics of AI Ethics: An Evaluation of Guidelines*, Minds & Machines, Vol. 30, 2020, p. 109, available at https://doi.org/10.1007/s11023-020-09517-8, last visited on Aug. 6, 2024.

（四）软法的技术方法论机制

人工智能软法是与科学技术紧密关联的，也因此被广泛认为是需要由人工智能专家研究制定的。“人工智能伙伴关系”作为一种联盟，将“公众”和“利益相关者”区分开，前者是需要教育和调查的，后者是科学家、工程师和企业家，是进行教育和调查的；其又将利益相关者区分为“专家”和“其他利益相关者”，前者是创造或应对人工智能的科学界领先者，后者是在广大范围内存在的产品使用者、购买人工智能方案的大型企业或者其所在领域被人工智能彻底改变的大型企业。“专家促使人工智能发生，其他利益相关者让人工智能发生在身上。”〔70〕正因为此，将人工智能软法落到实处，最重要的是专业人员在技术开发过程中进行“合乎伦理的设计”、开发“合乎伦理的人工智能系统”。而专业人员如何能把伦理价值嵌入人工智能/自动化系统的开发，是需要技术方法论的支持的。

在这方面的例子，除了缇姆尼特·吉布鲁团队研究的“数据集的数据表”以外，还有芬兰瓦萨大学博士后研究员维莱·瓦库里（Ville Vakkuri）领衔研究的命名为ECCOLA的方法，该方法是一个模块化的、逐段冲刺的过程，旨在促进人工智能和自动化系统开发对伦理的考量，并与既有的其他方法合并使用。具体而言，ECCOLA有三个目标：（1）促进对人工智能伦理及其重要性的意识；（2）创建一个适合各种系统工程场合的模块；（3）使得该模块既适合敏捷开发（agile development），又能让伦理成为敏捷开发的组成部分。ECCOLA经过多年的实践，经历了迭代的发展和改进。〔71〕此类事例不胜枚举。

（五）软法具体化的基准机制

前文已揭，许多人工智能伦理指南或原则是抽象的、含糊的，这主要是因为指南或原则的制定者希望能够尽可能将其适用于广阔的人工智能领域。但是，究竟如何才能在特定人工智能研发或利用中执行和遵守这些宽泛规范问题，对于想要做到合规的行动者而言，也会成为一个棘手问题。因此，除了技术方法论——往往是普遍适用于多个情境的方法框架或模块——以外，还需要结合特定人工智能的使用所引发的特定伦理关切，制定出更具针对性的伦理基准。日本东北大学的研究人员翁岳暄（Yueh-Hsuan Weng）与平田泰久（Yasuhisa Hirata）曾经发文探讨对辅助机器人的合乎伦理设计，文章指出，床位转移辅助、洗浴辅助、行走辅助、排泄辅助、监护和交流辅助以及护理辅助的机器人，各有比较突出的、不同的伦理关切，需要分别特殊对待。〔72〕他们的研究虽然并不有意指向或者有意拟定任何人工智能伦理规范的基准，但是，这种结合人机互动（human-robot interaction）的特点而指出每一种机器人需要应对的特殊伦理问题，其实就是具有基准意义的。这对于企业或其技术人员遵守人工智能伦理规范有着更具针对性的导引作用。

---

〔70〕 See Daniel Greene, Anna Lauren Hoffman & Luke Stark, *Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning*, available at https://hdl.handle.net/10125/59651, last visited on Aug. 6, 2024.

〔71〕 See Ville Vakkuri et al., *ECCOLA-A Method for Implementing Ethically Aligned AI Systems*, The Journal of Systems & Software, Vol. 182, 2021, p. 2, available at https://doi.org/10.1016/j.jss.2021.111067, last visited on Aug. 6, 2024.

〔72〕 See Yueh-Hsuan Weng & Yasuhisa Hirata, *Ethically Aligned Design for Assistive Robotics*, available at https://ieeexplore.ieee.org/document/8535889, last visited on Aug. 6, 2024.

（六）软法与硬法的互动机制

无论是在软法最初兴起的国际法领域，还是在人工智能软法领域，都已经有经验研究表明软法在未来的可能硬法化前景，或者软法被吸收进入硬法框架之中，这会给软法的实施增加动力或压力。例如，安德里亚斯·齐默尔曼教授在国际软法研究中发现，在早期阶段，不具有法律约束力的协定可能就已经规定了各国未来愿意接受的、作为未来有法律约束力条约组成部分的条件，这样的谅解备忘录是未来条约的先驱，有着“前法律功能”（pre-law-function），〔73〕可以被更好地实施。就人工智能软法而言，最初阶段进行实地试验的软法，之后可能会被正式立法纳入传统的监管体系之中。如“未来生命研究所”曾经于 2017 年发布阿西洛马人工智能原则（Asilomar AI Principles），〔74〕如今，美国加利福尼亚州已经将这些原则写入州立法之中。

除了这种未来法律化（硬法化）的前景以外，人工智能伦理规范若能在硬法的实施之中占有一席之地，也会带动企业及其他实体对其的遵守。例如，在美国，公司没有履行其对人工智能伦理规范的公开承诺的，联邦贸易委员会可以将其视为“不公平的或欺骗的”商业活动，而采取相应的措施。〔75〕在国际法情境中，国际法院和裁判机构也会经常性地依赖不具有法律约束力的协定，将其作为解释指南，对有法律约束力的条约进行解释。〔76〕当然，这种将软法吸收进入硬法解释适用的过程，也可视为另一种形式的硬法化；在一定意义上，此时的人工智能伦理规范已经不再是纯粹的软法。

五、结语：认真对待软法实施

软法的广泛存在，并不意味着其切实地得到了遵守和执行。人工智能领域的软法——各种各样的人工智能伦理规范——被许多研究者证明存在“实效赤字”的问题。规范的制定和倡议投入很多，收效却甚微。当然，人工智能伦理规范并不是完全的“零效用”，其对许多科技巨头产生了一定的拘束，隐私、公平、可解释性等规范明显被重视，在特别问题的“微观伦理”上取得了些许进步，其“推介实效”也在人工智能研发、利用共同体中有所显现。即便如此，人工智能伦理规范与现实之间的巨大鸿沟，仍然令人非常担忧。

之所以会有如此鸿沟，至少有前文所述的七方面的原因，然而，这些因素的存在，并不使“软法无意义”成为必然结论。由于人工智能伦理规范的灵活快捷性、多样适配性、合作试验性、事实压力性、跨国适用性，其仍然有独特的、硬性监管/硬法所无法比拟的、与硬性监管/硬法共

〔73〕 See Andreas Zimmermann, *Possible Indirect Legal Effects of Non-legally binding Instruments*, p. 6, available at https://ssrn.com/abstract=3840767, last visited on Aug. 6, 2024.

〔74〕 See Future of Life Institute, *Asilomar AI Principles*, available at https://futureoflife.org/open-letter/ai-principles/, last visited on Aug. 6, 2024.

〔75〕 See Gary E. Marchant, “*Soft Law*” *Governance of Artificial Intelligence*, available at https://escholarship.org/content/qt0jq252ks/qt0jq252ks_noSplash_1ff6445b4d4efd438fd6e06cc2df4775.pdf, last visited on Aug. 6, 2024; Wendell Wallach & Gary E. Marchant, *Toward the Agile and Comprehensive International Governance of AI and Robotics*, Proceedings of the IEEE, Vol. 107, 2019, p. 506, available at https://ieeexplore.ieee.org/document/8662741, last visited on Aug. 6, 2024.

〔76〕 See Andreas Zimmermann, *Possible Indirect Legal Effects of Non-legally binding Instruments*, p. 9, available at https://ssrn.com/abstract=3840767, last visited on Aug. 6, 2024.

同完成合乎伦理的人工智能之治理任务的价值。因此，如何使人工智能伦理规范应有价值更加充分地实现，如何通过一系列机制促进其间接实施，就成为一个需要认真对待的问题。

根据现实的经验观察，有助于人工智能伦理规范获得实施的间接机制，在逻辑上有延伸出软法实施机制的一般分类的可能。然而，这种分类学的研究还需要进一步探索，并非所有的机制都已经在这里进行了充分地讨论，在这里提出的机制也并非适用于所有软法实施的情境。例如，对于技术性、专业性并不是特别强的软法，技术方法论机制并不见得必需；对于本身已经足够特定、细致的软法，具体化基准机制也同样可以忽略。

软法的制定者、倡议者当然希冀软法可以发挥灵活引导的实际作用，但这种作用的获得不能仅依靠软法内在的说服力，不能仅依靠软法指向的行动者自觉认同与遵守。价值共识需要成本利益计算的经济逻辑的辅助，才可让更多的行动者愿意为软法的执行付出。内在理由和外在推动——柔性而非强制的推动——的有效结合，才可让软法不至于仅仅沦为宣示、倡议和粉饰。软法实施机制类型学的研究，对软法的制定者、倡议者或促进者有意识地进行相应的机制建设，具有重要的指引意义。

Abstract: Soft law's widespread existence does not mean it is complied with and implemented. The soft law in AI, the AI ethics code, has been proved to have an "effectiveness deficit", which owes to a variety of features of AI ethics such as unenforceability, abstractness and vagueness, diffusion, overlap, and confusion, and to other reasons like companies having little incentive for compliance, the paradox of compliance in some cases, predicaments disclosed by social system theories, and the determinist approach to AI development. However, AI ethics still have unique values because of their flexibility and agility, multiplicity and adaptivity, cooperativeness and experimentalism, de facto pressures, and cross-jurisdictional applicability. Empirical studies have demonstrated that AI soft law can be implemented indirectly through a set of mechanisms in terms of organization, pressures for compliance, incentives for compliance, technological methodology, specifications, and interaction of soft and hard law. A more general conclusion shall be drawn that soft law can acquire much effectiveness by combining the value consensus and the economic logic, and by combining the internal reasons and the external impetus as well.

Key Words: soft law, artificial intelligence, ethics, implementation mechanism, effectiveness of soft law

(责任编辑：刘　权)