欧盟《人工智能法》解析（第一部分总则、第二部分禁止的人工智能实践）

2024-04-02

人工智能是一个快速发展的技术族，能够为各行各业和社会活动带来广泛的经济、环境和社会效益，但同时也会存在巨大的社会风险；随着通用人工智能技术在过去几年的快速发展，人类已经正式进入了第四次工业革命时代即人工智能时代。2024年3月13日，欧盟议会以523票赞成、46票反对和49票弃权审议通过了《人工智能法》，该法案是人类历史上第一部关于人工智能的正式立法（还需得到欧盟理事会的正式批准才能生效），该法案全文13篇113条，主要内容包括人工智能系统在欧盟的市场投放、提供服务和加以使用时的统一规则、禁止的人工智能实践、高风险人工智能系统适用规则、特定人工智能系统提供者和部署者的透明度义务、通用人工智能模型投放市场的统一规则、支持创新的措施（重点是小微企业，包括初创企业）、欧盟人工智能的监管机构、欧盟高风险人工智能系统数据库、后市场监测、信息共享、市场监督规则、处罚、法案生效时间和适用范围等；欧盟《人工智能法》作为人类历史上第一部关于人工智能的立法，既对我国的人工智能立法具有重要的借鉴和参考价值，也会对我国人工智能企业在欧盟的运行和产品投放产生重要影响；本系列文章对欧盟《人工智能法》的主要规则进行简要介绍，以其对人工智能企业合规有所裨益。本文主要介绍该法案的第一部分总则和第二部分　禁止的人工智能实践。

一、欧盟《人工智能法》的立法目的、立法原则及适用范围、定义

1、《人工智能法》的立法目的：为改善内部市场运作，促进以人为本、值得信赖的人工智能应用，同时确保对健康、安全和《宪章》规定的基本权利，包括民主、法治和环境保护，使其免受联盟内人工智能系统的有害影响，并支持创新。

2、《人工智能法》的立法原则：《人工智能法》对欧盟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》中确定的七项不具约束力的伦理原则进行了重申，这些原则贯穿在整部《人工智能法》的各个方面，这七项原则包括：人类的主体和监督；技术稳健性和安全性；隐私和数据治理；透明度；多样性、非歧视和公平；社会和环境福祉；问责制；这七项原则的具体内容如下：

人类的主体和监督意味着人工智能系统的开发和使用时为人服务的工具，尊重人的尊严和个人自主权，其运行方式可由人类进行适当控制和监督；

技术稳健性和安全性是指开发和使用人工智能系统的方式应能在出现问题是保持稳健，并能抵御试图改变人工智能系统使用或性能的行为；

隐私和数据管理是指人工智能系统的开发和使用符合现有的隐私和数据保护规则，同时处理的数据在质量和完整性方面符合高标准；

透明度是指人工智能系统的开发和使用方式应允许适当的可追溯性和可解释性，同时让人类意识到他们与人工智能系统进行了交流或者互动，并适当告知部署者该人工智能系统的能力和局限性，以及受影响者的权利；

多样性、非歧视和公平性是指人工智能系统的开发和使用方式应包括不同的参与者，并促进平等获取、性别平等和文化多样性；

社会和环境福祉是指以可持续和环保的方式开发和使用人工智能，并使全人类受益，同时监测和评估对个人、社会和民主的长期影响。

3、《人工智能法》的适用范围

该法案适用于（1）在欧盟境内将人工智能系统投放市场或者提供服务或将通用人工智能模型投放市场的提供者，无论这些提供者是设立于还是位于欧盟境内或第三国；（2）在欧盟内设立场所或者位于欧盟内的人工智能系统部署者；（3）场所位于第三国或者位于第三国的人工智能系统提供者和部署者，其系统生产的产出用于欧盟；（4）人工智能系统的进口者和分销者；（5）产品制造商以自己的名称或商标将人工智能系统与其产品一起投放市场或提供服务；（6）未在欧盟境内设立场所的提供者的授权代表；（7）位于欧盟内的受影响者。

4、对《人工智能法》涉及如人工智能系统、人工智能提供者、部署者、经营者、生物数据、生物识别、生物验证、生物分类系统、 远程生物识别系统、“实时”远程生物识别系统、情绪识别系统、真实世界测试计划、沙盒计划、人工智能监管沙盒、人工智能素养、真实世界条件下的测试、通用人工智能系统、浮点运算等主要概念进行了定义。

人工智能系统：人工智能系统是一种基于机器的系统，设计为以不同程度的自主性运行，在部署后可能表现出适应性，并且为了明确或隐含的目标，从其接受的输入中推断出如何声称可能影响物理环境的输出，如预测、内容、建议或决定。这一关于人工智能的定义明确了人工智能系统区别于传统的软件系统或编程方法（其仅仅是根据自然人定义的规则自动执行操作的系统）人工智能系统的一个主要特点是具有推理能力，这种推理指获得输出的过程，如预测、内容、建议或决策，也指代人工智能系统从输入/数据中推导出模型和或算法的能力，同时人工智能系统还可以影响物理环境和虚拟环境；人工智能系统在设计时具有不同程度的自主性，这意味着它们的行动在一定程度上独立于人类的参与，并具有在没有人类参与的情况下运行的能力；人工智能系统在部署后可能表现出的适应性是指自主学习的能力，允许系统在使用过程中发生变化。

远程生物识别系统：系指一种人工智能系统，其目的是在没有自然人主动参与的情况下，通常通过将一个人的生物数据与参考数据库中的生物识别数据进行比较，远距离识别自然人的身份。

“实时”远程生物识别系统：是指一种远程生物鉴别系统，在该系统重，生物数据的采集、比较、识别都是在没有明显延迟的情况下进行的，这不仅包括即时识别，还包括有限的短暂延迟，以避免规避本法案。

情感识别系统：根据自然人的生物数据识别或推断出其情感或意图的人工智能系统。

沙盒计划：参与提供者与主管机关之间商定的文件，其中描述了在沙盒内开展活动的目标、条件、时间框架、方法和要求。人工智能监管沙盒：是指由主管机关建立一个具体和受控的框架，为人工智能系统的提供者或潜在提供者提供在监管监督下根据沙盒计划在有限的时间内开发、培训、验证和测试创新人工智能系统的可能性。

二、禁止的人工智能实践及例外

人工智能除许多有益用途外，该技术也可能遭到滥用，并为操纵、剥削他人、社会控制、商业欺诈提供了新颖而强大的工具；下列人工智能技术属于欧盟《人工智能法》明确禁止的人工智能实践：

1、采用欺骗或操纵技术，扭曲人或一群人行为的人工智能实践。

该人工智能技术可被用来劝说人们做出他们不想从事的行为，或通过诱导其做出决定来对其加以欺骗，从而颠覆和损害他们的自主、决策和自由选择；如这类人工智能采用潜意识成分，例如人们无法感知的音频、视屏、图像刺激，这些刺激超出了人的感知范围，或者采用其它操纵或欺骗技术，以人们无法意识到的方式颠覆或损害人的自主、决策或自由选择，或者即使意识到了，人们仍然被欺骗，或者无法控制或抵制。

例外情况为对操纵性和剥削性实践的禁止不应影响医疗方面的合法实践，如精神疾病的心理治疗或身体康复，如果这些实践是根据适用的法律和医疗标准进行的，例如得到个人或其法定代表人的明确同意；此外，符合适用法律的常见的合法商业行为，如广告领域的行为，本身不应该被视为构成有害的人工智能操纵行为。

2、利用特定个人或特定人群因其年龄、残疾或特定社会或经济状况而具有的任何弱点，扭曲特定人或该群体的人工智能实践。

此类人工智能系统被投放市场、提供服务或加以使用，其目的或效果是实质性的扭曲人的行为，并对该人或其他个人或群体（如生活在极端贫困中的人、少数民族或宗教少数群体）造成或者有合理可能性的造成重大伤害，包括可能长期累积的危害，因此应予禁止。即使人工智能的提供者或部署者可能无法合理的预见和缓解这些因素，即无法推定有扭曲行为的意图，只要这种伤害是由人工智能操纵或者剥削行为造成的，都应该予以禁止。

3、禁止基于自然人的生物数据，如个人的面部和指纹，来推断个人的政治观点、工会成员身份、宗教或哲学信仰、种族、性生活或性取向的生物分类系统。

这项禁令不包括根据生物数据对按照欧盟或国家法律获取的生物数据集进行合法标记、过滤或分类，例如根据头发颜色或眼睛颜色对图像进行分类，这可以被用于执法领域。

4、禁止由公共或私人行为者为自然人提供社会评分的人工智能系统。

这类人工智能系统可能导致歧视性结果和排斥特定群体，这类人工智能系统可能会侵犯尊严和不受歧视的权利以及平等和公正的价值观。这类系统根据与自然人在多种场景中的社会行为有关的多个数据点或者已知、推断或预测的特定时期的个人或个性特征，对自然人或其群体进行评估或分类。此类人工智能系统中获得的社会评分可能会导致自然人或整个群体在社会环境中受到有害或不利的待遇，而这些环境与最初生成或收集数据的场景无关，或者导致与其社会行为的严重程度不成比例或不合理的不利待遇。

5、禁止对自然人进行风险评估，以评估其犯罪的风险，禁止根据对自然人的画像或对其个性特征和特点的评估来预测实际或潜在刑事犯罪的发生的人工智能系统。

根据无罪推定原则，禁止对自然人进行风险评估，以评估其犯罪的风险，禁止根据对自然人的画像或对其个性特征和特点的评估来预测实际或潜在刑事犯罪的发生；这一禁止不涉及并非基于个人画像或者个性特征或特点的风险分析，如使用风险分析工具麻醉品或非法货物本地化的可能性，使用风险分析的人工智能系统根据可疑交易评估企业的金融欺诈风险。

6、禁止无差别抓取面部图像的人工智能系统。

禁止人工智能系统通过从互联网或闭路电视录像中无针对性的获取面部图像来创建或扩大面部识别数据库，因为这种实践会增加大规模监控的感觉，并可能导致严重侵犯包括隐私权在内的基本权利。

7、禁止情绪推测或识别的人工智能系统。

禁止将旨在用于检测个人在工作场所和教育相关情况下的情绪状态的人工智能系统投放市场、提供服务或使用。在不同的文化和不同的情境下，甚至在同一个人身上，情绪的表达都有很大差异，这类系统的主要缺点包括可靠性有限、缺乏特异性和通用性有限，其适用结果可能导致歧视性的结果，并可能侵犯相关人员的权利和自由，可能导致特定自然人或整个自然人群体受到后海或不利的待遇。

这一禁令不包括用于医疗治疗的人工智能系统。

8、禁止为执法目的而使用人工智能系统在公共场所对自然人进行“实时”远程生物识别的人工智能系统。

这类人工智能系统会对个人的自由和权利具有特别的侵扰性，因为这类系统可能会影响大部分人的私生活，使人产生始终受到监视的感觉；用于对自然人进行远程生物识别的人工智能系统在技术上的不准确性可能会导致存在偏差的结果并产生歧视性的影响，在涉及年龄、民族、种族、性别或残疾时，尤为重要；同时，使用这种“实时”运行的系统，其影响具备即时性，进一步检查或纠正的机会有限。因此，应该禁止为执法目的而使用这些系统，除非在详尽列出和严格界定的情况下，使用这些系统对实现重大公共利益时严格必要的，其重要性压过了风险。这些情况包括：寻找特定犯罪受害者，包括失踪人员；自然人的生命或身安全受到特定的威胁或者受到恐怖袭击；确定本条例附件所列刑事犯罪的犯罪人或嫌疑人的位置或身份，条件是这些刑事犯罪在有关成员国应受到监禁判决或拘留的惩罚，惩罚的期限至少为四年。

例外为执法、边境管制、移民或庇护机关能够根据欧盟或成员国法律使用信息系统来识别在身份检查期间拒绝被识别或无法说明或证明其身份的人，而无需根据本条例事先获得授权。同时，在公共场所为执法目的使用“实时”远程生物识别系统，每次使用应得到司法机关或其决定对成员国具有约束力的独立行政机关的明确和具体授权。同时只有在只有当执法机关完成了基本权利影响评估，并在本条例规定的数据库中登记了该系统，才能授权在公众可进入的空间使用实时远程生物识别系统，同时该远程生物识别系统智能部署用于确认具体目标个人的身份，并应仅限于在实践、地理和个人范围方面绝对必要情况，尤其考虑到有关的威胁、受害者或犯罪者证据或迹象。

来源：

作者：付文家

相关律师

付文家
合伙人