反歧视法律体系规制人工智能算法歧视分析

网络空间安全学院

学号：2113203

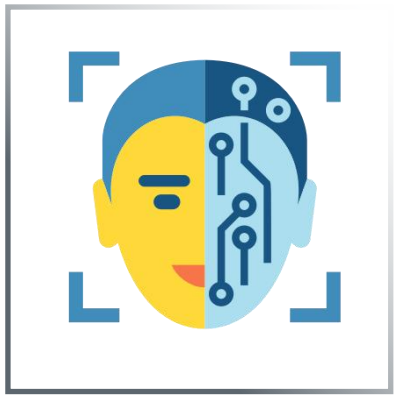
姓名：付政烨

目 录

CONTENTS

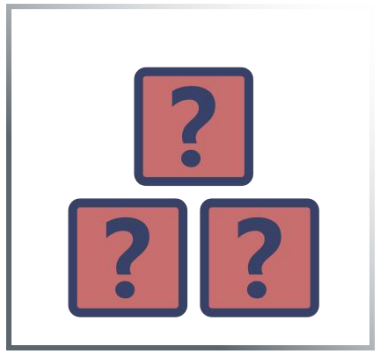
1

引 言



选题背景

- 人工智能算法的发展为社会各个领域带来了无限可能，为人们的生活带来了诸多便利和改进。然而，这些算法很可能潜藏着歧视和偏见的风险。
- 一些算法可能会根据我们似乎无关紧要的个人特征，如网络浏览偏好或手机号码等，将我们划分为新的群体类别。
- 在荷兰，一家保险公司因为客户住在带有特定数字的公寓而对他们收取额外的汽车保险费（这个数字可能包含字母，如4A或20C）。



探究问题

- ✓ 算法歧视和差异化的具体含义
- ✓ 适合应用于算法决策的反歧视法律体系
- ✓ 处理非特定群体的算法差异化的最佳途径

说明

I. 歧视：对具有受保护特征的人（如种族）造成伤害，构成不合理的不平等对待，或者被广泛认为是不可接受的行为。

II. 差异化、区分/分类或不平等对待：中性意义上的歧视。

III. 算法决策：基于计算机算法输出的决策。

探究方法

I. **文献综述与比较分析**：对完全开放体系、完全封闭体系和混合体系在应对算法歧视方面的特点进行比较分析，以提出适用于该问题的法律体系。

II. **案例分析**：通过具体案例（如谷歌广告系统和在线教育平台的个性化定价）展示算法差异化对特定人群造成的伤害和不公平现象。

III. **跨学科协作**：融合法律、计算机科学和社会学等多个领域的知识，确保对算法歧视问题的全面理解和解决方案的科学性。

IV. **技术应用**：探讨法规编码技术如何将法律条文转化为机器可读的政策语言，并利用静态分析工具（如PRIVANALYZER）实现自动化合规性检查。

2

算法歧视和差异化

01

算法

算法是一种计算过程的抽象描述，通常表现为计算机程序，用于处理输入数据并产生输出结果。

02

人工智能

通过算法和大量的数据集进行模拟，使计算机系统能够展现出类似于人类智能的科学。

03

机器学习

让计算机系统通过分析和理解数据，从中学习规律和模式，并根据学习到的知识做出决策或预测。

算法 > 人工智能算法 > 机器学习算法

概述

算法决策经常给人以理性、高效和无懈可击的印象；因为数字和数学不会撒谎。目前学界普遍认为，算法本身是中立的。然而，不幸的是算法决策也可能导致基于种族或性别的歧视，以及其他形式的不公平差异化。

01 特定人群的差异化

02 非特定人群的差异化

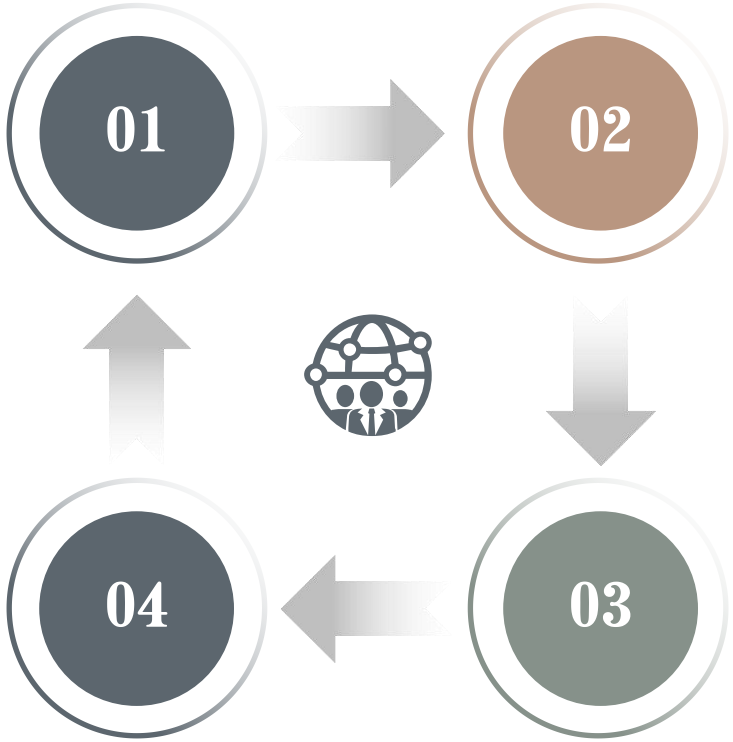
03 差异化造成的危害后果

法律概念：间接歧视

指的是看似中立的条款或做法，对某一特定群体造成不利影响

法律保护

现有的反歧视法律框架对该方面的保护已相对完善，允许受害者通过提供证据，证明某项表面中立的政策或做法对其群体造成特别不利影响，从而要求纠正不公。（如《欧洲人权公约》第14条和《欧盟基本权利宪章》第21条）



无故意算法歧视

如果一个算法应用是基于包含歧视或偏见的数据集进行训练的，那么就可能导致无意的算法歧视

谷歌广告歧视案

2013年，谷歌用于向搜索引擎用户投放广告的算法受到偏见影响，当用户搜索具有非裔美国人名时，显示的广告会暗示某人有犯罪记录，而搜索白人却不会。这是因为谷歌的算法分析了点击率最高的广告，从而继承了种族偏见

01 非特定特征差异化

在实际应用中，绝大多数算法通过基于非特定特征（如汽车型号、居住地邮政编码等）进行差异化来提取事物及其相互间的内在联系。这些特征与保护特征（如种族或性别）并无直接关联。通过分析大量数据，算法能够发现这些特征之间的复杂非显性相关性，并利用这些相关性进行预测或区分群体和个体。

02 注意

并非所有算法差异化都是不公平的，其不公平性取决于各种环境因素。我的目的是说明，即使算法差异化不一定会对具有受保护特征的人造成伤害，它仍可能导致不公平的实践，或者至少引发争议。

案例 · 非特定人群的差异化

某家在线购物平台的运营通过算法发现，使用“京东”应用程序的用户倾向于购买价格较高的产品，而使用“淘宝”应用程序的用户倾向于购买价格较低的产品。基于这一发现，该公司决定对使用“京东”应用程序的用户收取更高的运费。

尽管这一差异化并不是基于种族或性别等受保护特征，但可能导致使用某一款应用程序的用户被要求支付更高的费用。这种差异化可能会导致不公平，因为其依据并非个体的实际购买能力或其他合理因素，而是基于一个看似无关紧要的特征，即用户所使用的手机应用程序。

危害后果

结果与算法设计初衷不一致 01

引发不公平现象 02

算法差异化可能会加剧社会经济群体之间的不平等，固化贫富差距

以某在线教育平台为例，该平台采用算法进行个性化定价，初衷是为贫困学生提供更高的优惠。然而，该算法将学生观看网课的数量作为判断贫富的标准，即观看次数越多，学生越富裕。但事实上，城市中家庭条件较好的学生通常可以获得更好的线下教育资源，因此不需要过多依赖在线教育资源。相反，偏远地区的学生由于缺乏优质的线下教育资源，更需要依赖在线教育，因此观看次数可能更多。这导致低收入群体需要支付更高的教育费用，而高收入群体则享受较低的价格，进一步加剧了贫富差距。

3

反歧视法律体系

内含

反歧视法律体系旨在通过法律手段保护个人和群体免受基于种族、性别、宗教、民族等各类特征的歧视。其核心是平等原则，确保所有人在法律面前一律平等，不因个人特征而受到不公平待遇。

01

完全开放体系

该体系以其**包容性**和**广泛适用性**著称，其法律**未明确列举**具体的受保护歧视理由

02

完全封闭体系

以其**明确性**和**可预测性**为特点，**详细**列举了受保护的歧视理由，并且对可能的豁免情况进行了**明确**的规定

03

混合体系

结合了完全开放和完全封闭体系的特点，通常列出**部分**受保护的歧视形式，对未列举歧视形式持**开放性态度**

	中国		欧盟	
总体要求	《中华人民共和国宪法》第33条规定，国家尊重和保障人权，并规定了公民在法律面前一律平等的原则。		《欧洲人权公约（ECHR）》：任何声称其权利在被成员国侵犯的个人都可以向人权法院法院提出申诉。《欧洲联盟基本权利宪章》和一系列非歧视指令，要求成员国将其转化为国内法。	
主要法律和指令	劳动法	禁止在就业中基于种族、性别、宗教信仰等的歧视行为。	种族平等指令	禁止在多个领域基于种族或民族出身的歧视。
	就业促进法	强调公平就业，禁止用人单位在招聘中基于性别、民族等因素进行歧视。	就业平等指令	禁止在就业领域基于宗教、信仰、残疾、年龄或性取向的歧视。
	妇女权益保障法	保障妇女在政治、经济、文化等方面的平等权利，禁止任何形式的歧视。	性别商品和服务指令	禁止在商品和服务供应领域基于性别的歧视。
	个人信息保护法	规定了个人信息处理的透明度要求，保护个人信息免受滥用，间接支持反歧视目标。	通用数据保护条例（GDPR）	规定了数据处理的透明度要求，防止数据处理过程中出现歧视性结果。

01 相同点

- ✓ 中国和欧盟都在其宪法或基本权利文件中明确规定了平等和非歧视的原则
- ✓ 双方都有具体的法律和指令来执行这些基本原则，涵盖了就业、商品服务等多个领域
- ✓ 都意识到保护个人信息对反歧视法律体系的重要性

02 不同点

涉及法律体系、执行机制等方面，与本文无太大关系

03 意义

中国和欧盟在反歧视法规上的相似之处，为构建相关的反歧视法律体系提供了宝贵的指导。奠定了明确的法律基础，确保了AI系统的公平性。

4

遏制人工智能相关的反歧视法律体系构建

理论上，混合的反歧视法律体系可以最大限度地发挥不同理论体系的优势。那么，在面对人工智能带来的特殊挑战时，是否可以得出相同的结论？

回顾第二部分，我们将人工智能驱动的歧视问题分为两类：（1）特定人群的差异化（2）非特定人群的差异化

完全封闭、开放的法律体系难以适用算法歧视

01 完全封闭的法律体系

- 与完全封闭的法律体系相比，完全开放的法律体系更能适应人工智能驱动的差异化
- 法律的制定和更新速度远远低于科技发展的速度

02 完全开放的法律体系

- 开放的法律体系使得企业在开发算法时缺乏法律确定性，导致在每一个人工智能驱动的差异化应用中，公司可能需要展示差异待遇的合理理由
- 许多算法都是不可解释的（黑箱），尤其当算法采用复杂结构或大规模参数时。具体表现为，其在处理大规模数据和复杂任务时表现出色，但其内部机制通常难以被直观解释。给企业带来很大负担，抑制科技创新

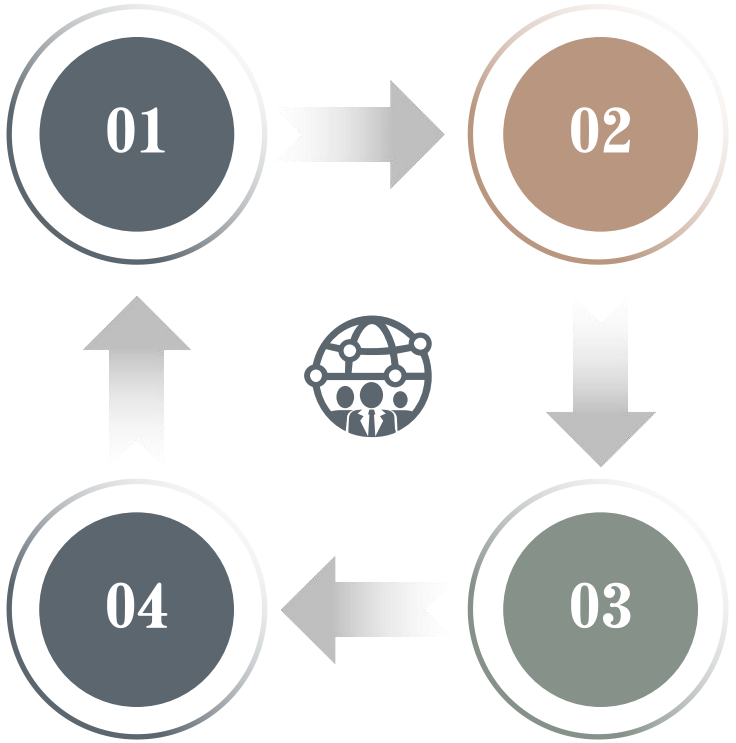
应对算法歧视的最佳反歧视法律体系

基准作用

公司和法院可以将算法生成的结果与法条中明确歧视条款比对，以查他们是否存在相似性

增进公众对算法公平性的信任

混合的法律不仅能够应对当前的歧视问题，也能预防未来潜在的歧视风险，可以增强社会对技术进步的接受度



动态扩展保护范围

通过动态发展的判例法体系，法院处理具体案件时，可根据实际情况和社会发展的需要，逐步认定新的受保护理由，使法律能够及时应对新的歧视风险

促进技术发展

企业在开发新产品或服务时，必须考虑到法律判例中揭示的潜在歧视问题，并采取相应的措施来消除这些问题

混合法律体系利于与法规编码技术相结合

01

编码法规技术

编码法规技术是将法规中的法律条文转化为特定应用领域的机器可读政策语言的过程。其主要目的是确保算法能够正确理解并执行特定的法规和政策，从而在数据处理过程中保持合规性。

02

技术实现流程

I. 制定并翻译基础政策，转化为机器可读的政策语言
II. 使用政策语言编码具体的法律条文
III. 通过静态分析工具实现自动化合规性检查

03

混合法律体系的优势

混合法律体系结合了封闭法体系和开放法体系确定性和灵活性的优点，在与编码法规时，具有显著的优势。

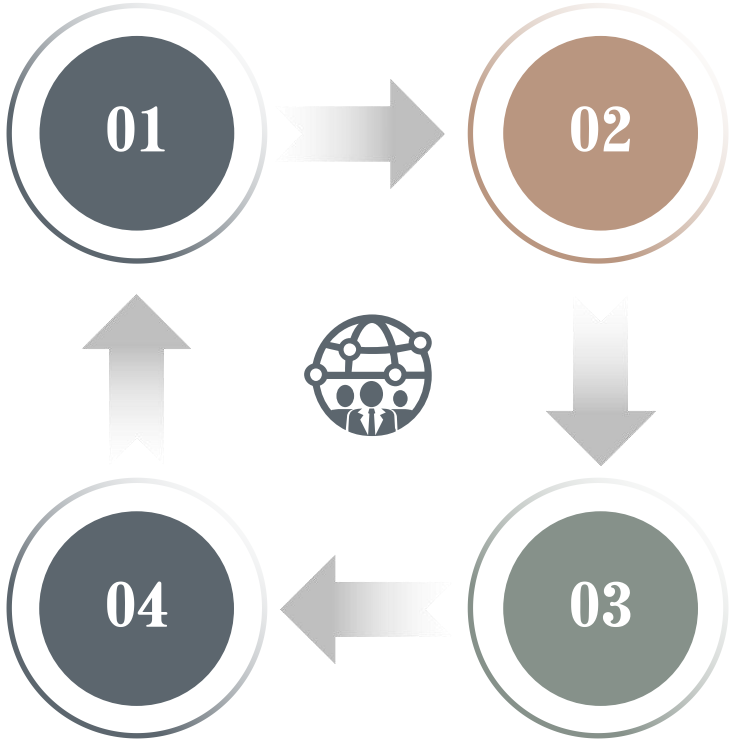
混合法律体系的优势

封闭部分

明确性的特点便于直接将法条翻译成机器可读的政策语言

开放部分

法规编码本质是形式化语言，而形式化语言在通过分析器推导的过程中本身会存在一些概念的扩张，开放部分的灵活性与其对应



处理复杂任务

在机器学习等领域，数据处理过程涉及多个步骤和复杂的操作。通过将法律要求编码为机器可读的政策，能够在每一步操作中自动检查和确保合规性，从而支持复杂数据处理任务的顺利进行

促进跨学科交流

法规编码工作往往需要法律专家和计算机专家合作进行，以确保编码的准确性，间接促进了法学和计算机领域的交流

5

总结与展望

总结 01

随着人工智能技术的迅猛发展，算法歧视问题日益显现并日趋严峻。未来，立法者和执法者需要密切关注并深入研究这一领域，不断制定和完善相关法律法规，以应对瞬息万变的技术和社会环境。

平等 自由

未来的研究聚焦 02

- ✓ 研究和开发更加先进的法规编码技术，提高其自动化合规性检查的精准度和效率。
- ✓ 推动跨学科合作，结合法律、计算机科学和社会学等多个领域的专业知识，共同应对算法歧视问题。
- ✓ 加强国际合作，分享并借鉴各国在反歧视法律体系和技术实施方面的宝贵经验和成功案例，共同推动全球范围内人工智能技术的公平和正义应用。

[1]Wachter S. The theory of artificial immutability: Protecting algorithmic groups under anti-discrimination law[J]. Tul. L. Rev., 2022, 97: 149.
[2]Mann M, Matzner T. Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination[J]. Big Data & Society, 2019, 6(2): 2053951719895805.
[3]Mann M, Matzner T. Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination[J]. Big Data & Society, 2019, 6(2): 2053951719895805.
[4]Leese M. The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union[J]. Security Dialogue, 2014, 45(5): 494-511.
[5]Graham J. Risk of discrimination in AI systems: Evaluating the effectiveness of current legal safeguards in tackling algorithmic discrimination[M]//FinTech, Artificial Intelligence and the Law. Routledge, 2021: 211-229.
[6]Nachbar T B. Algorithmic fairness, algorithmic discrimination[J]. Fla. St. UL Rev., 2020, 48: 509.
[7]刘雷.平等视角下算法治理歧视及其反歧视措施均衡[J].湘江青年法学, 2021(1):58-74.
[8]张欣,宋雨鑫.人工智能时代算法性别歧视的类型界分与公平治理[J].复印报刊资料:妇女研究, 2022(5):14.
[9]刘雪丹.网络平台算法歧视的法律规制[D].华南理工大学,2021.
[10]丁晓东.算法与歧视从美国教育平权案看算法伦理与法律解释[J].复印报刊资料：法理学．法史学, 2018(4):12.

[11]官极,方望,应明生.确定量子机器学习算法不公平性因素的方法,系统和设备:CN202210872317.9[P].CN202210872317.9[2024-06-05].
[12]严景.人工智能中的算法歧视与应对——以某公司人工智能简历筛选系统性别歧视为视角[J].法制博览, 2019(14):2.DOI:CNKI:SUN:FBZX.0.2019-14-056.
[13] 卜素.人工智能中的"算法歧视"问题及其审查标准[J].山西大学学报：哲学社会科学版, 2019, 42(4):6.DOI:CNKI:SUN:SXDD.0.2019-04-016.
[14]朴毅.人工智能的价值非中立性及其应对[J].[2024-06-05].
[15]徐琳.人工智能推算技术中的平等权问题之探讨[J].法学评论, 2019, 37(3):10.DOI:CNKI:SUN:FXPL.0.2019-03-013.
[16]刘雪丹.网络平台算法歧视的法律规制[D].华南理工大学,2021.
[17]姜野.算法的法律规制研究[D].吉林大学[2024-06-05].
[18]衣俊霖.数字孪生时代的法律与问责——通过技术标准透视算法黑箱[J].东方法学, 2021(4):16.

欢迎批评与指正！

网络空间安全学院

学号：2113203

姓名：付政烨