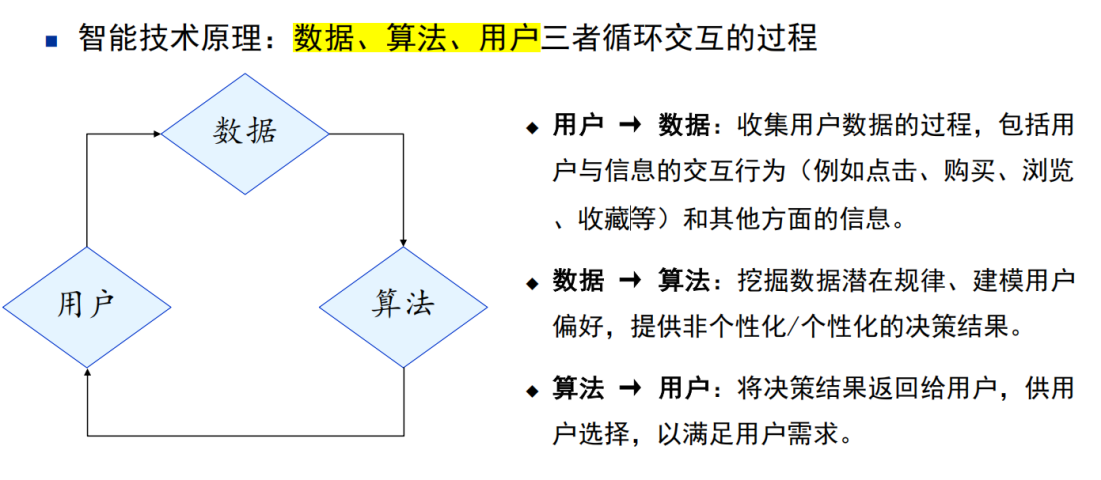
1. 人工智能伦理与治理定义：

人工智能伦理是开展人工智能**研究、设计、开发、服务和使用**等活动需要遵循的**价值理念和行为规范**；人工智能伦理治理是**多主体**协作的**全流程**治理，是以**敏捷机制**协调人工智能发展与安全的重要治理模式。

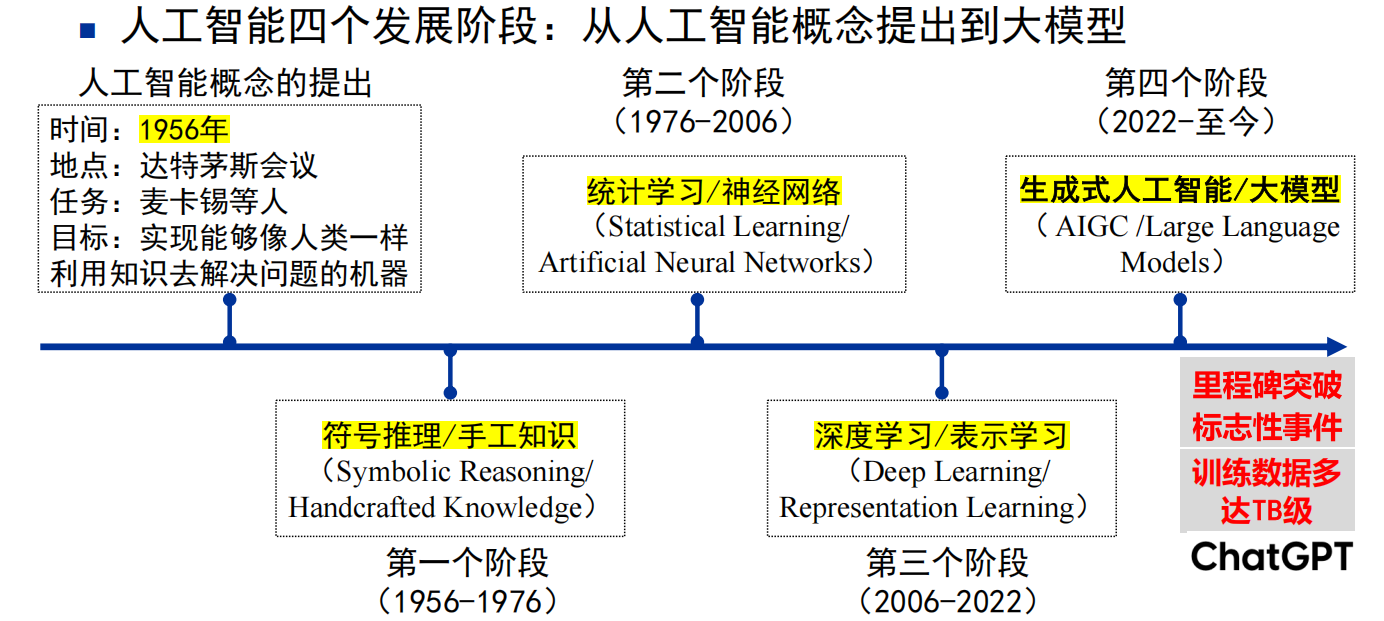
1. 智能技术原理



1. 人工智能发展历程

**数理逻辑**与**计算机**是人工智能发展的两块基石

逻辑系统: (1) 以**符号**表示的事实; (2) 以**蕴含关系**(→)表示的事实间的推理关系; (3) 若干限制条件和演算规则



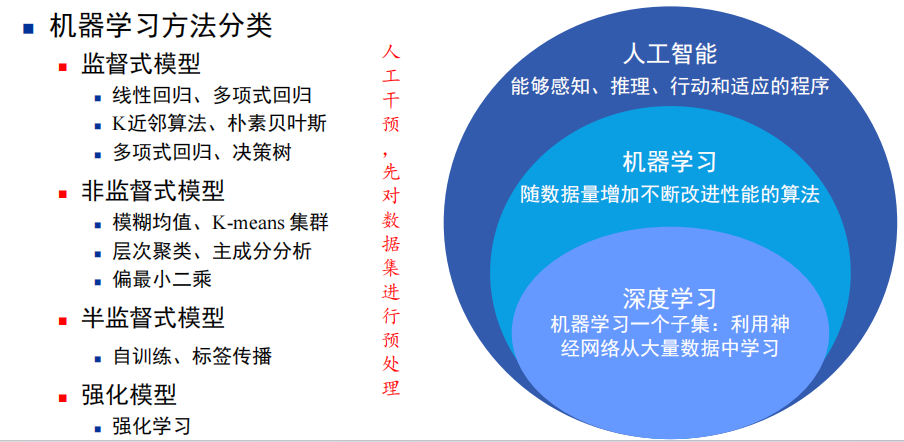
1. 符号主义与连接主义对比

符号主义：符号规则，内省总结，有知识依据，可解释；但难以穷尽知识和规律

连接主义：模拟人脑工作方式，强调模仿人脑神经网络的计算模型。

数据驱动，自动学习，可利用大数据优势，但缺少可解释性。

1. 人工智能、机器学习、深度学习三者关系



1. 深度学习机制=表征学习+深层结构

优势：对信息，特征更加准确地表示

补充：训练集的样本的描述能力 代表性 代表性不足 欠拟合

测试集 泛化能力 代表性好 泛化能力不足 过拟合

过拟合原因：模型过于复杂，**模型参数训练过头**。

1. 大语言模型能力建立 预训练（通用领域，掩码预测/下个词预测）

+微调（任务相关领域），涌现（ChatGPT：大数据与大参数，指令微调（有监督数据指导模型**理解和遵循**），人类意图对齐（指导模型**生成**）），激发（情境学习，思维链，指令学习（就是之前说的指令微调的两个部分）），实践（提示工程）

1. 大模型能力：多源信息整合与提炼；通用生成；外部工具调用

风险：幻觉（输入不当、数据缺陷、训练和生成策略的不足）；法律道德

1. 人工智能伦理与治理要坚持**以人为本、公平非歧视、透明可解释、人类可控制、责任可追湖、可持续发展**等内容
2. 大数据特点：海量性，高速型，多样性，价值性
3. 数据伦理定义、范畴、问题、给案例，能分析数据伦理问题

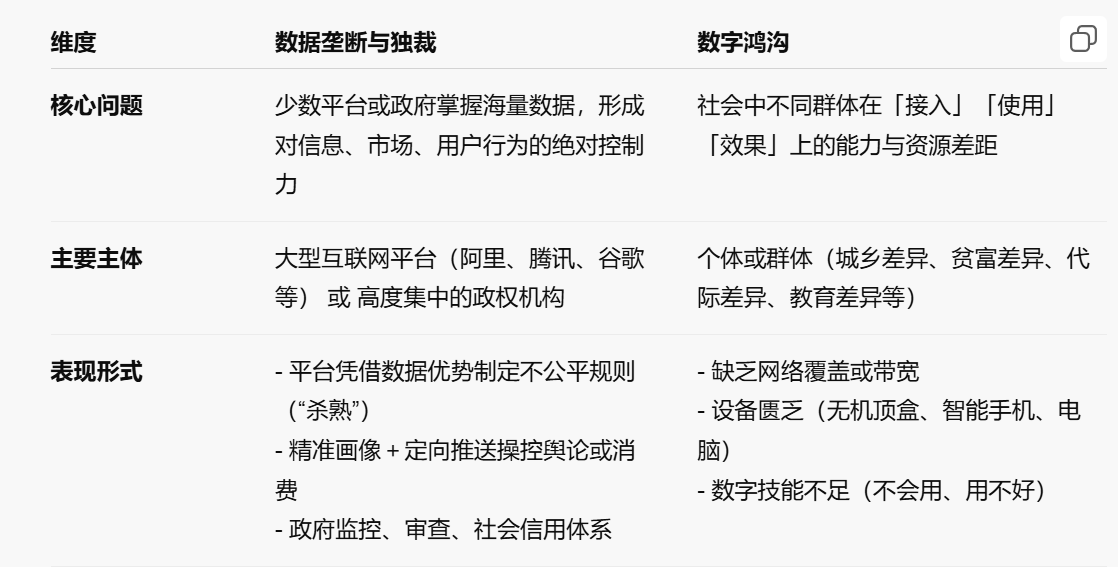
定义：数据伦理是在数据收集、存储、使用和共享过程中所涉及的伦理问题和原则。主要关注如何确保数据的使用符合道德和法律标准，避免侵犯个人隐私、产生不公正或偏见，以及确保透明度和问责制；

范畴：数据收集（合法，用户知情），存储（安全，完整，合规），使用（准确，可解释，公平），共享与传播伦理（透明，保护隐私，责任可追溯）

问题：收集与使用，隐私与保护，偏见与公平性，数据安全，数字鸿沟，数据垄断与独裁，数据真实性。

例子：

1. 电信诈骗：收集与使用-非法收集用于诈骗，隐私与保护-严重泄露用户隐私，平台缺少加密与访问审计；数据真实性--诈骗信息是伪造的，误导；数据安全--存储与传输环节安全防护失效。
2. 大麦网撞库：数据安全——平台密码存储与风控不足；隐私与保护——用户凭证被倒卖与滥用；共享与传播——黑市交易无监督，责任难追溯
3. 大数据杀熟：偏见与公平性——算法对忠诚用户杀熟不公；透明度不足——定价逻辑不透明；数据垄断与独裁——平台占据数据优势，用户无议价能力



1. 数据收集与使用伦理中用户各权力

**保护个人数据权益**作为核心目标，赋予用户**知情权**、**决定权**、**隐私权**、**访问权（中南财经政法大学退休教授 赵德馨起诉知网）**等关键权利

1. 数据泄露定义、给案例，能分析数据伦理问题

定义：数据泄露是指一种敏感的、被保护的、秘密的数据被未经授权的人员进行

复制、传递、观察、偷窃或者使用的安全事件

注：大数据时代下的隐私与传统隐私的最大区别在于**隐私的数据化**

1. 匿名化技术

匿名化技术是通过对数据中包含的个人身份信息进行**移除、替换或模糊**处

理，使得经过处理后的数据**无法直接关联到具体个人**

具体方法：K-匿名（通过对个体（记录）的准标识符做泛化，压缩等处理使得其准标识符等于其他k-1人的准标识符，也就是这k个人形成了一个等价类。）。数据同态加密（将原文加密，对密文做代数运算，解密后得到的结果与原文做同样代数运算的结果一样）

1. 数据偏见与公平性伦理问题，尤其是偏见来源，纠偏方法

来源：数据收集，处理，分析，应用都有偏见

数据收集（代表性不足，环境局限，目的不正当，缺乏时间动态性）

数据标注（标注人员主观偏见，标准不一致，任务偏差，数据不平衡）

算法偏见（预处理不平等，目标函数有偏向，训练偏见，测试不足，透明度不足）

纠偏方法：（1）改进数据采集：确保数据多样性，代表性；弥补数据的偏见；动态采集，更新数据；增加少数样本数据的权重；确保数据采集过程的透明性，伦理性。（2）数据预处理与偏见检测：数据清洗，异常值处理；数据平衡与加权；数据标准化/归一化；数据匿名化

1. 伦理审查：建立独立的伦理委员会，制定伦理标准

**多方力量：技术，法律，社会伦理三方力量！**

1. 算法伦理定义、伦理问题

定义：聚焦于算法设计，开发，部署，使用，探讨如何确保算法符合道德原则，平衡技术发展与人类福祉的问题。

注：**透明** 算法的设计，运行机制等 公开，可理解

**可解释性** 基于透明的设计，还要向用户解释为什么等（以人类可理解的方式解释其决策过程和结果）

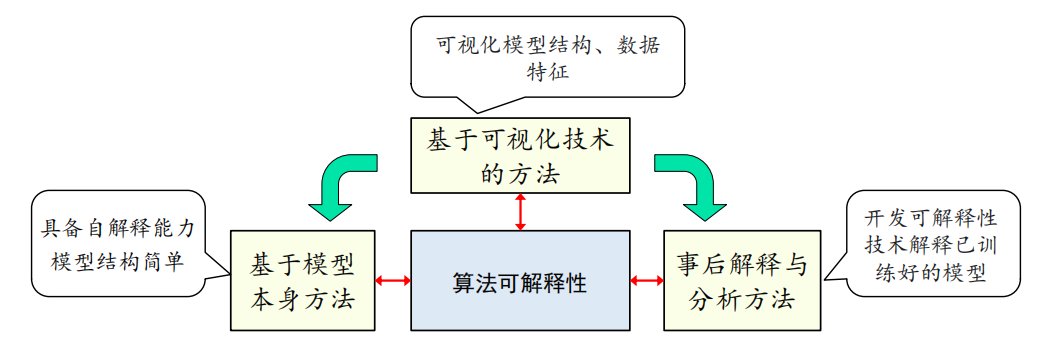
问题：过度收集用户数据，侵犯个人隐私；决策不透明，缺乏可解释性；存在偏见和歧视

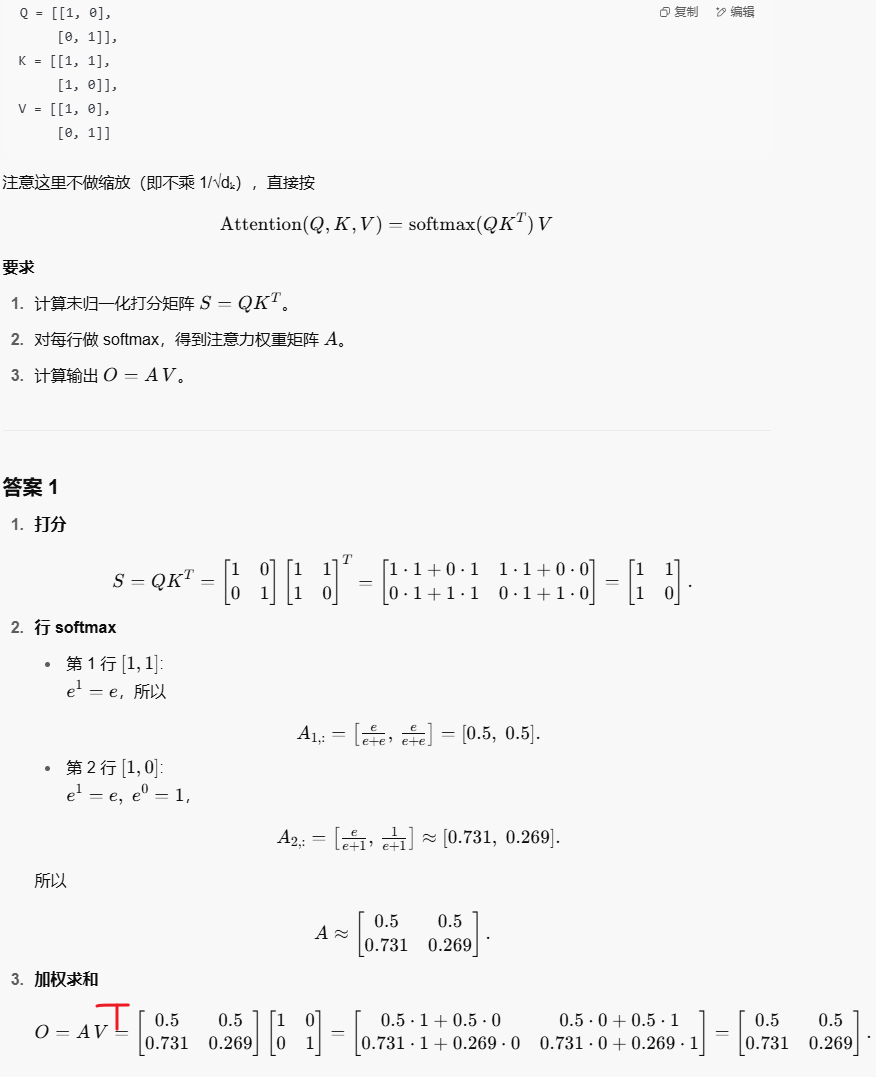
1. 算法黑箱定义、原因

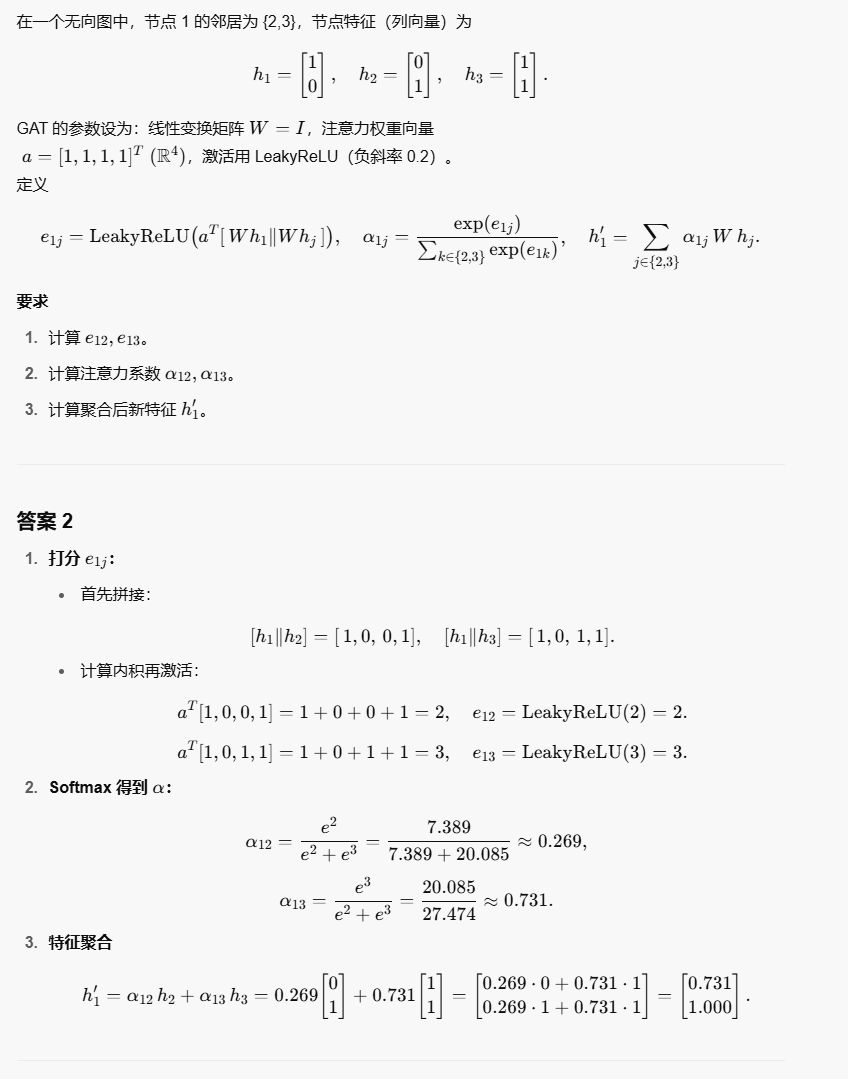
伦理问题：黑箱问题；黑箱问题让人难以信任（无法控制，歧视偏见，安全风险）；特定领域的AI很重要~

原因：神经网络等模型内部结构复杂，参数量极高，有特别多非线性转换，难以用人类直观地逐层解读；许多企业将自研算法视为核心竞争力，不公开模型结构等；可解释性研究不完善；为保护用户隐私和数据安全，有时故意对模型内部进行加密或混淆处理，使第三方无法轻易提取训练数据的信息。

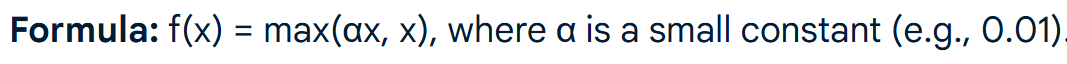
1. 增强算法可解释性的技术手段

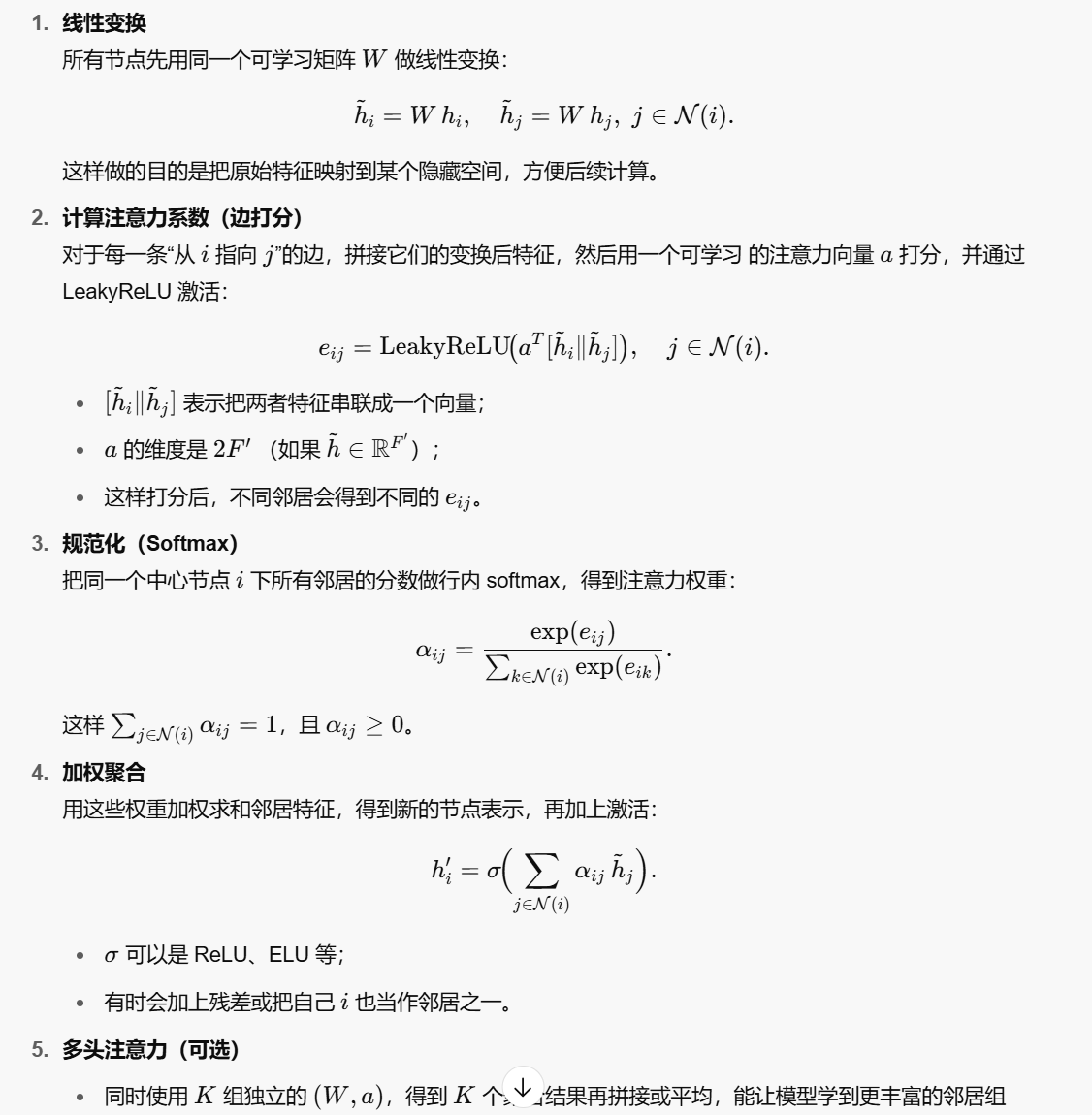






LeakyRELU：





1. 算法歧视与偏见伦理定义

计算机系统中的偏见，是指对某些个体或群体，相对于其他个体或群体的系统性地、不公平的区别对待

1. 偏见种类：
2. 数据——算法：群体偏见（群组的结论应用至个体之上），采样偏见（对子群组的非随机采样），聚合偏见
3. 算法——结果：算法偏见（输入数据中不存在，完全算法的）

解释：整体最小化均方误差：这是一种全局性的目标，他会偏向大多数群体，无法照顾少数群体；优化函数：不同loss对错误敏感度不同，如MSE对大误差敏感，交叉熵损失对罕见类别不敏感。正则化：全局性的约束模型参数，但**可能过度平滑，抑制了对少数群体的特征捕捉**。未考虑子群差异：不同类别混在一块训练，无公平约束或者分层指标。

评估偏见

1. 结果——用户：

曝光偏见：被曝光的信息容易产生用户行为，但是不被曝光不代表用户消极偏好

位置偏见：位置靠前的更容易产生用户行为

热度偏见：较为流星的信息更容易被用户看到并产生用户行为

1. 用户——数据:

自选择偏见：被调查用户对象的自我选择（比如：明星投票，粉丝更容易参与）

从众偏见

社会偏见:受到其他人的影响

1. 几种公平定义与计算

公平是不基于先天或者后天的特征对个人和群体产生偏好

察觉型公平：算法对相似的人做出相似的判断

忽略型公平：算法预测/决策的过程中没有任何敏感性变量的参与

个体公平：相似的个人被相似地对待

群体公平：相似的群体被相似地对待

计算：**公平定量定义：混淆矩阵**

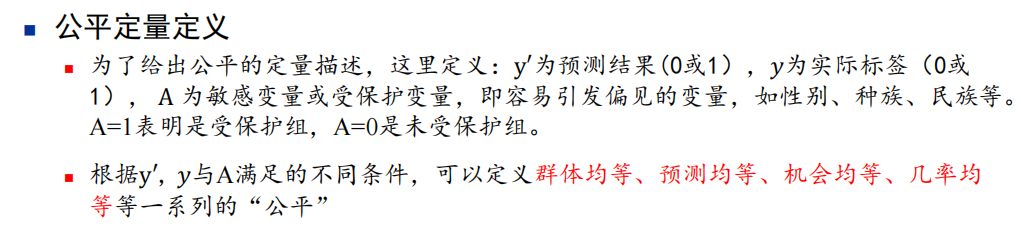
True/False Positive/Negative 前者指的是预测与真实相比正确/错误 后者预测结果是正还是负，例如 TN 指的是预测结果为负，但是与真实值一样，也即预测负例，真实负例子 我们称之为真负例

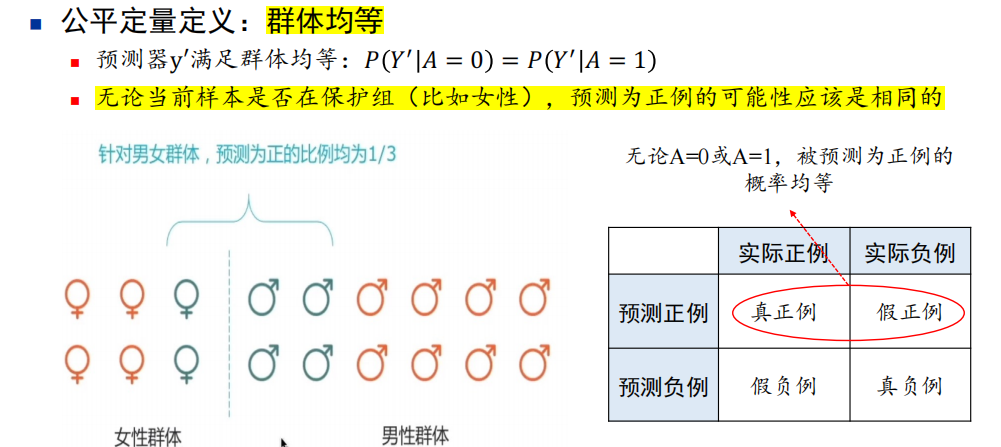


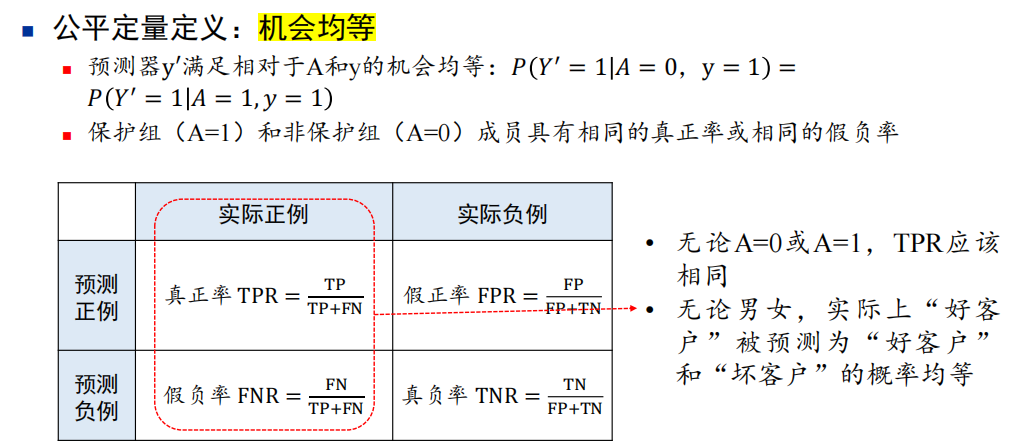
加一个率 如TNR真负率指的是 ：第一，明确这是一个实际负例，第二，用TN/实际负例得到TNR

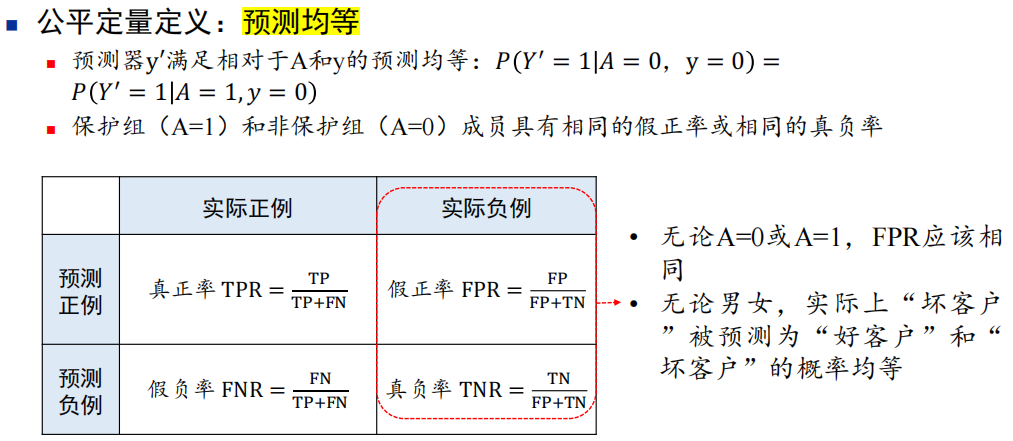
1. g. TPR = TP/TP+FN FNR= FN/TP+FN TNR=TN/TN+FP

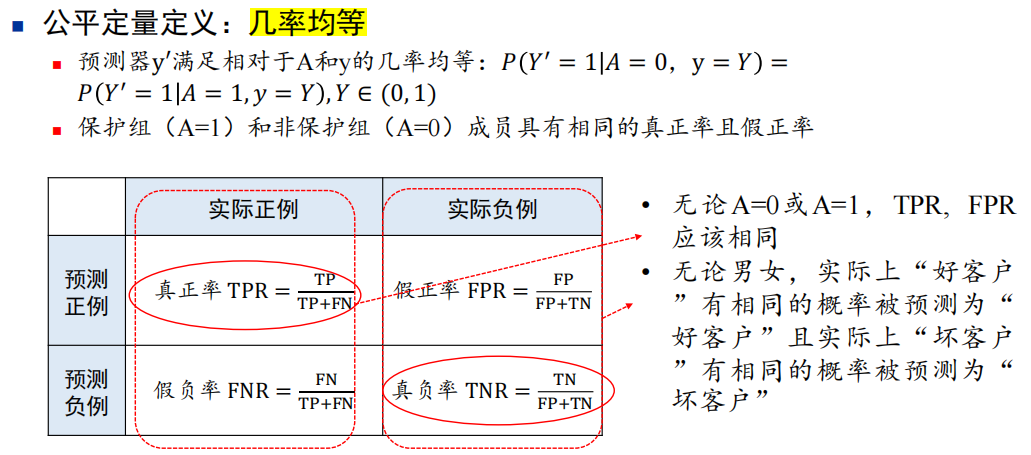
解释：真正率是**实际上**具备好信用的客户，**被预测**为好的概率











1. 去除偏见的方法

预处理偏见：消除数据偏差，去噪，矫正

样本均衡（增加少数群组采样），对所有群组采样，数据清洗

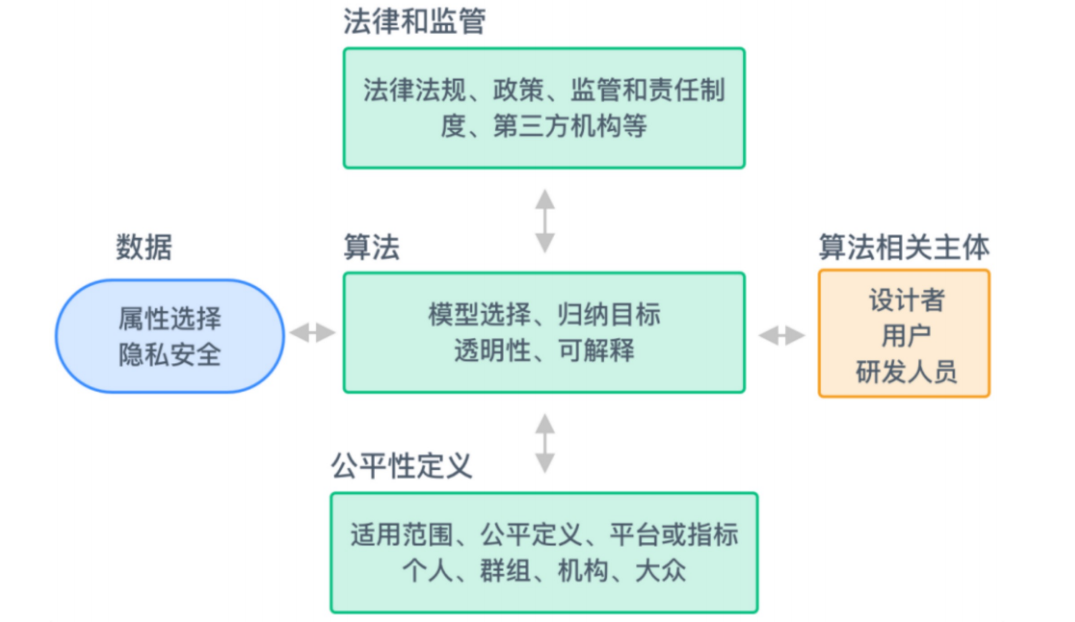
处理中偏见：消除模型偏见，训练，优化

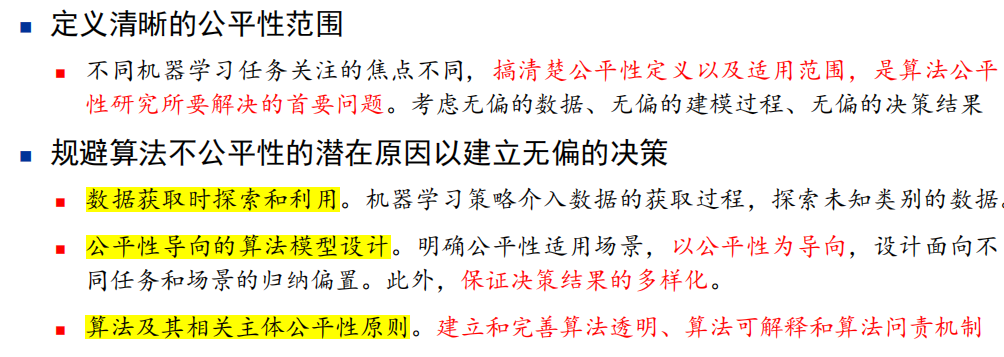
更改目标系数，添加约束条件，可对每个子群分别训练分类器

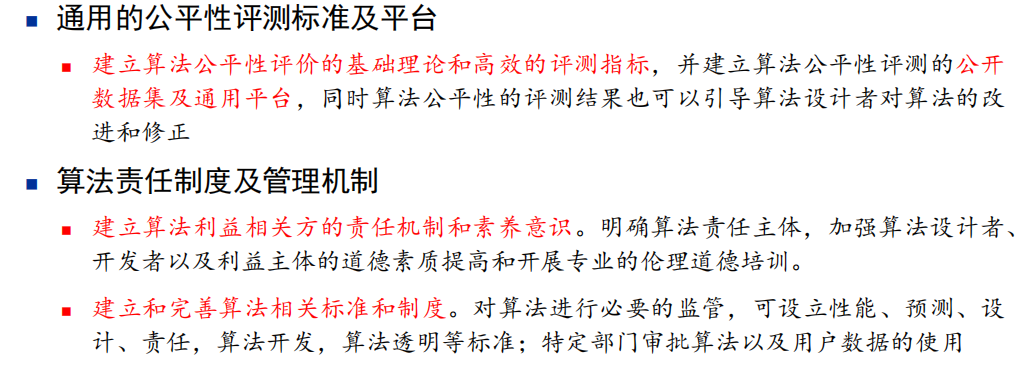
后处理偏见：消除结果偏见，多样化，类均衡

更改算法输出的分数，实现不同群体的得分分布对齐；不同群组按照不同阈值划分；调整词向量移除偏见；对决策结果重排序

1. 应对算法偏见的策略与建议







1. 具身智能：具身智能(Embodied AI)是指智能体通过身体与环境的互动产生的智能行为，

强调智能体的认知和行动在物理环境中的相互依赖

多模态感知，大语言模型，强化学习，人机交互与情感计算，运动控制和路径规划

1. 人工智能与就业与经济的影响 辩证看待替代~，全面理解创造~，不应忽视增长~
2. 替代效应（未大规模产生） 机器替代体力型劳动力，部分高技能型智力型劳动力

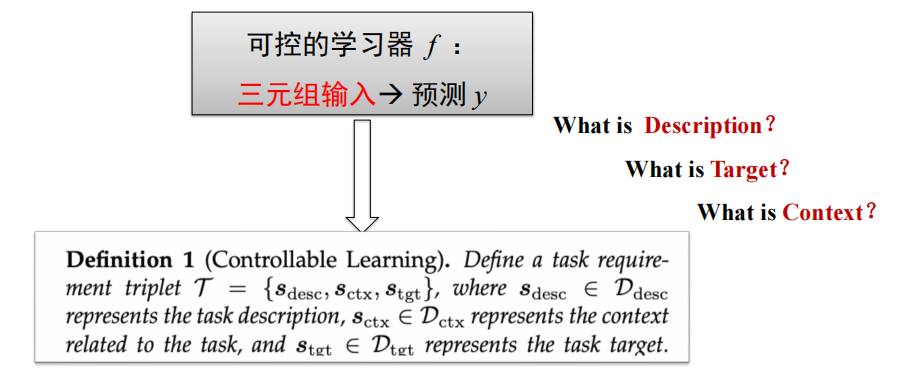
限制因素：机器与劳动力的成本比较；技术适用性与任务复杂度制约（对创造力工作，人际交互工作无法取代；社会制度的限制（裁员成本高？社会保障制度不完善））；数据隐私与技术可靠性：数据隐私，泄露，算法黑箱，不透明，责任难追溯，效果难保证？歧视与偏见

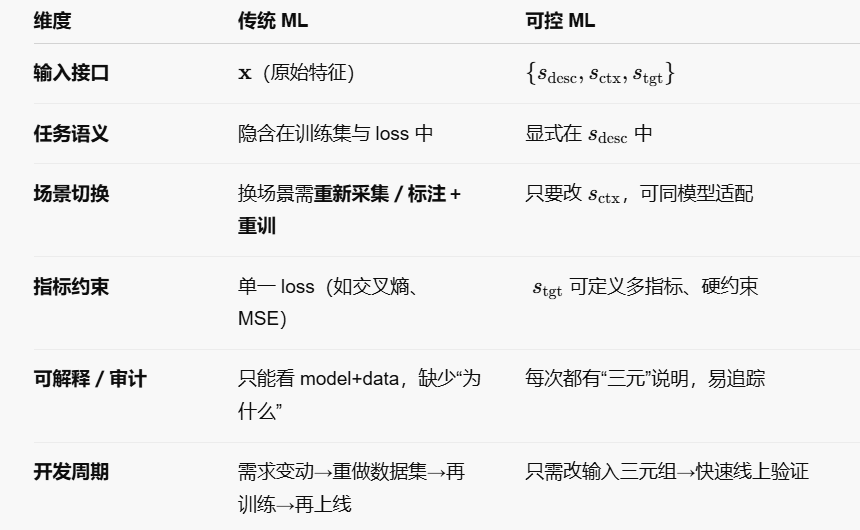
（2）创造效应：创造岗位 AI相关的（最显著），AI产业链上下游的，AI和其他产业结合的

（3）增长效应：推动经济增长，国际竞争力提升

就业缓冲机制，人才培养与教育，加速就业岗位（传统产业的数智化）转型（深度融入AI技术，降低成本，提升质量/服务水平）

1. 理想中可控的机器学习

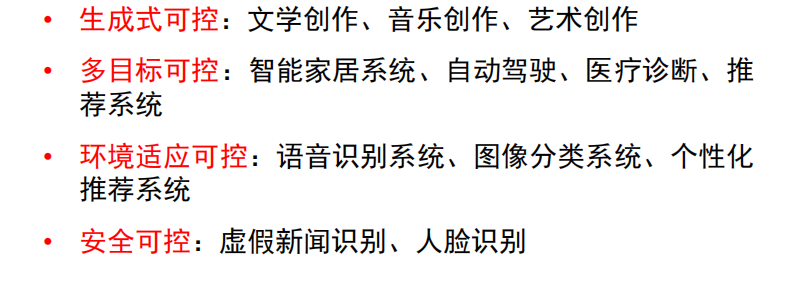




**可控机器学习**把「需求描述」「场景上下文」「优化目标」三要素都作为显式输入，从接口层面赋予人类对模型行为的实时掌控。

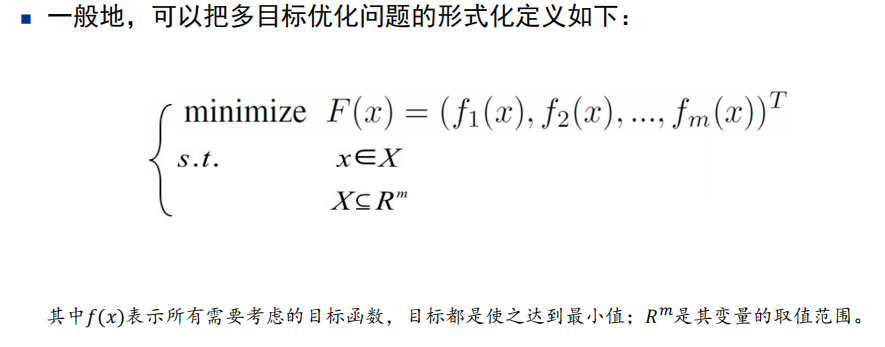
**传统机器学习**则是一个以特征→标签为核心的闭环，所有语义和约束都埋在训练阶段，上线后难以再调整，因而本质上是不可控的。

1. 可控人工智能分类



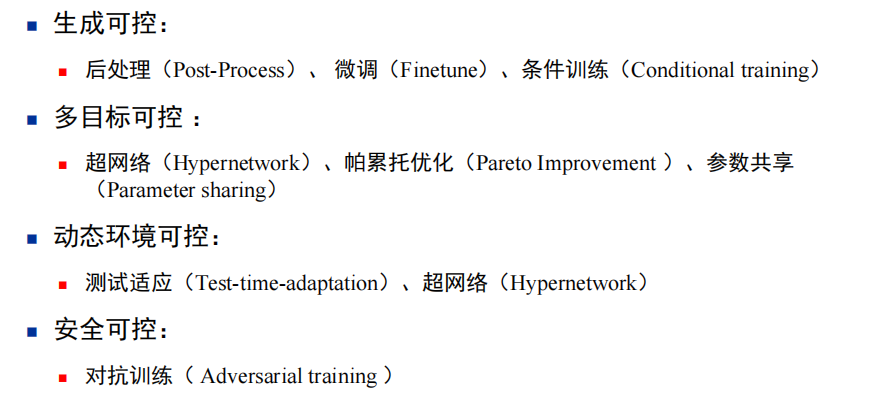
重点：**多目标可控**

多目标优化是在某个情景中在需要达到多个目标时，由于容易存在目标间的内在冲突，在他们中间做出协调和折衷处理，使总体的目标尽可能的达到最优的求解过程

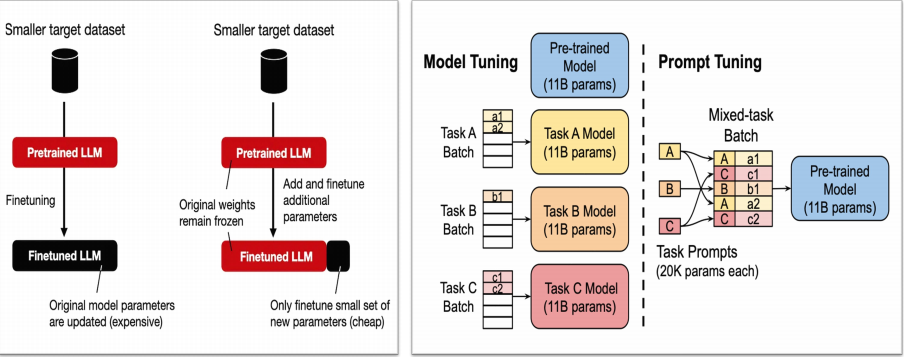


帕累托最优解：不存在其他解使得每个目标函数值都优于当前解

1. 人工智能可控技术及其特点

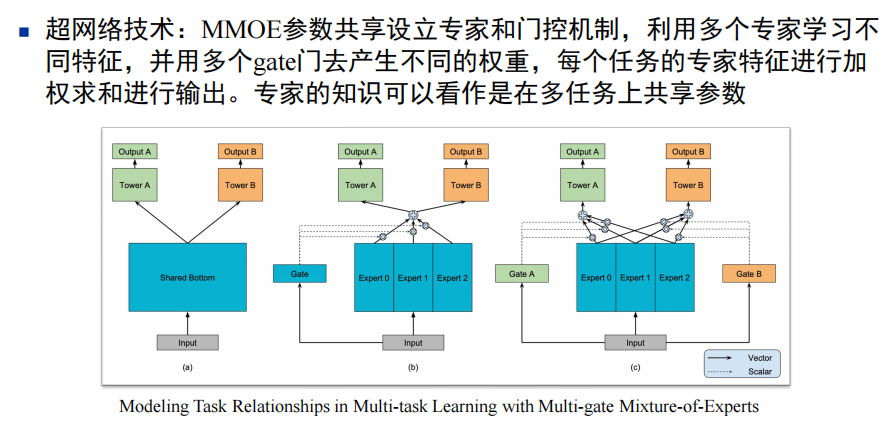


基于后处理的保守派：PPLM 不希望训练PLM，而是通过额外一个attribute model对PLM输出加以修正，并利用反向梯度修正PLM内部历史状态，从而符合某种可控预期，是一种后处理方式。



基于Retrain的改革派：主流语言模型为自回归模型，在训练和推理中添加特定的控制token可以影响或调控模型的其他输出

超网络仅仅是一个小型网络，该网络为一个大得多的网络生成权重，属于模型压缩的算法，但天然地适合不同条件下对模型快速调控



**测试适应技术**：首先估计当前时刻的标签分布，然后利用估计的标签分布与分类模型的输出做融合，选出概率最大的标签

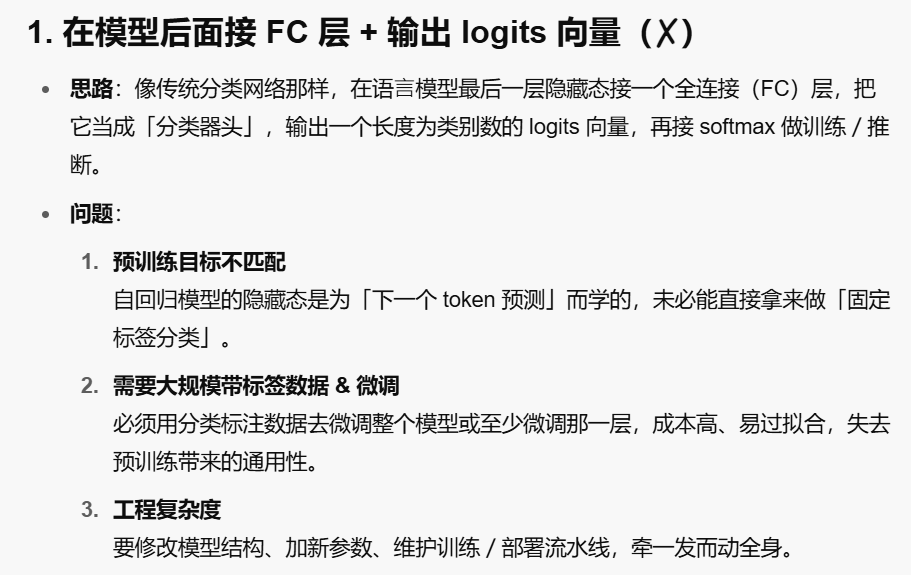
造成AI不可控另一大因素是**人为的恶意攻击**。深度神经网络(DNNs)已经成功地部署在各种真实的应用场景中，但最近的研究表明，DNNs非常容易受到对抗性攻击。通过在良性输入中引入视觉上难以察觉的扰动，攻击者可以操纵DNN模型来提供错误的预测。一般包括攻击方法和防御方法.

**逃逸攻击(Evasion Attacks)**指攻击者在不改变目标机器学习系统的情况下，通过构造特定输入样本以完成欺骗目标系统的攻击。一般发生在模型已经完成训练，在预测阶段。其根本原因，在于模型没有学到完美的判别规则.

其思路非常简单直接，将生成的对抗样本加入到训练集中去，做一个数据增强，让模型在训练的时候就先学习一遍对抗样本。

1. 自回归模型

LM:从左至右依次预测下一个token,和模型对话，让它告诉我预测类别





1. 生成式人工智能与传统模型的区别

（1）重新训练深度生成模型的成本过高！需要可行的参数微调方案

注：传统模型往往无预训练骨架，当任务，数据，指标发生改变时一般需要重新训练~

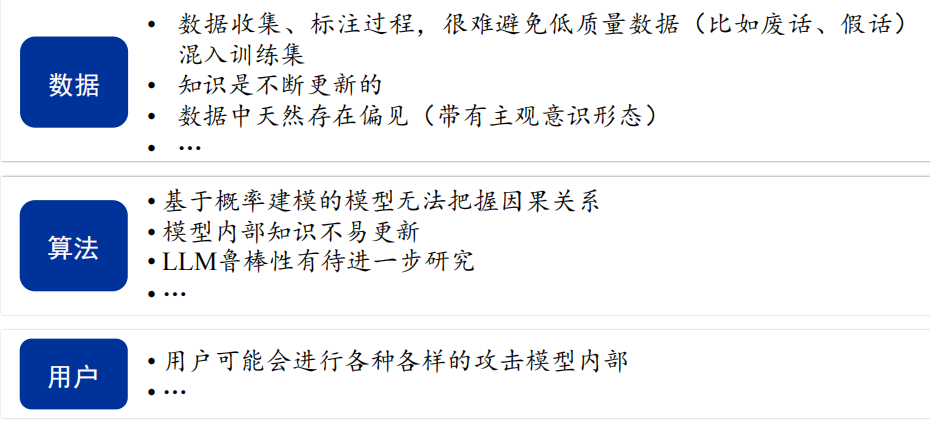
（2）生成内容流入互联网，可被爬为训练数据

（3）生成模型往往需要由大公司运营，用户调用API服务

1. 生成内容3H原则，影响3H原则的原因

Honest, harmless, helpful 真实性（信息准确，结果可自圆其说），无害性（无偏见，无歧视等，不侵犯个人隐私），可用性（不生成废话；有价值；面对任何输入，**能在恶意攻击下保证安全**）

**原因**：用户，数据，算法



1. 现有技术方案

真实性： 检索增强 前处理（ICL） 幻觉检测（后处理） 因果干预（处理中）（构造新样本消除非因果的相关性，比如护士数据把her改为his等等） 可解释（处理中）（思维链）

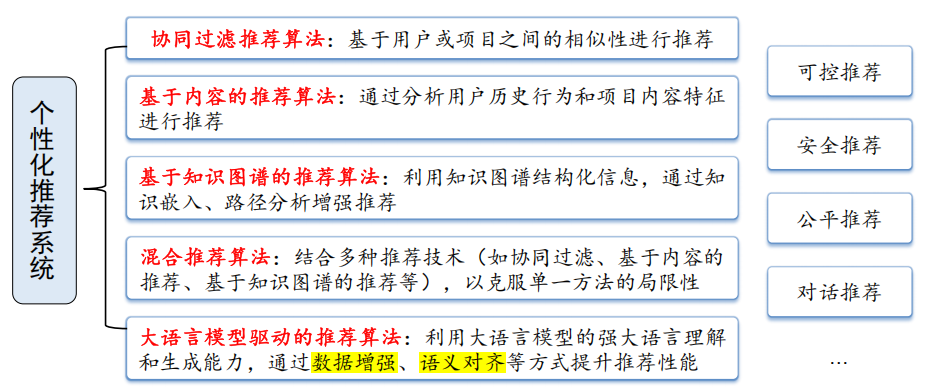
无害性/可用性：大模型调用工具，RLHF

1. 推荐系统：

推荐系统指通过算法**分析用户的历史行为**、偏好特征及相关数据，主动**预测**用

户需求并为其**筛选、排序、呈现个性化信息或服务**的智能技术体系

算法：

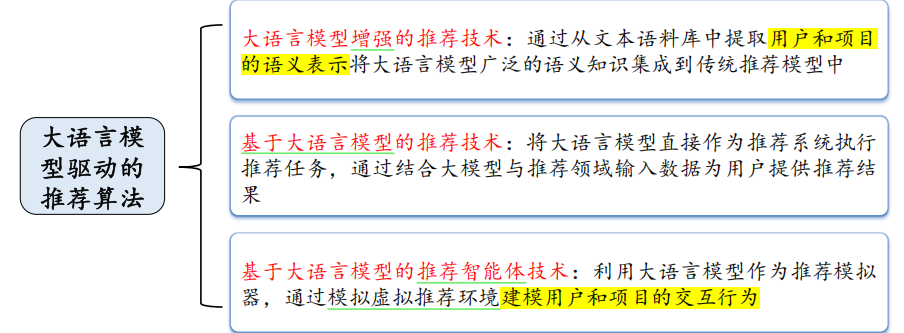


协同过滤的核心是计算用户或项目之间的相似性，关键在于相似性度量与矩阵分解

基于内容的推荐算法核心是如何准确地提取用户和项目的内容特征，关键在于特征提取与表示学习

基于知识图谱的推荐算法关键在于知识图谱构建与嵌入学习

**大语言模型驱动的推荐算法**关键在于语义理解与数据增强。**该类方法成为主流**



发展趋势：

全球推荐系统产业政策的发展基本上趋于一致，即初期监管相对宽松，主要侧重于数据自由流动和市场驱动的创新；随后随着技术的快速迭代进步、产业的不断扩张，政府和国际组织陆续发布相关政策法规，对出现的各种问题进行限制，政策监管开始逐步收紧；最终细化相关政策以完善监管体系