3.1 引言

在这次内容中，我们将开始探讨大型语言模型的有害性（危害）。在这门课程中，我们将涵盖几种这些危害:

- 性能差异（本讲）
- 社会偏见和刻板印象（本讲）
- 有害信息（下一讲）
- 虚假信息（下一讲）

另外在之后的课程中会讲述其他更多层面的危害性：

- 安全和隐私风险（未来内容）
- 版权和法律保护（未来内容）
- 环境影响（未来内容）
- 权力集中（未来内容）

新兴技术的危害：我们知道“能力越大责任越大，对于当前开创性的大模型来说，我们需要了解这些模型的能力和危害之间的密切关系。大模型的能力所展示的潜力将导致这些模型被广泛的采用，但是与此同时造成它们的危害。

由于AI的发展是近几年发展的产物，因此对于危害的研究与预防依旧是一个很新的事情。因此回顾历史，从过往历史中的其他领域中的危害、安全和伦理问题的防御进行了解，首先考虑一些在具有成熟的危害和安全传统的学科中使用的高层次思想和方法，有助于对当前AI领域有所借鉴

贝尔蒙特报告和IRB。

- 贝尔蒙特报告于1979年编写，概述了三个原则（尊重人员、善

行和公正）。
- 该报告是机构审查委员会（IRB）的基础。
- IRB是审查和批准涉及人类研究的委员会，作为一种积极的机制来确保安全。

生物伦理学和CRISPR。

- 当基因编辑技术CRISPR CAS被创建时，生物医学界制定了社区标准，禁止将这些技术用于许多形式的人类基因编辑。
- 当发现社区成员违反这些标准时，他们将被开除出社区，这反映了对社区规范的严格执行。

FDA和食品安全。

- 食品和药物管理局（FDA）是一个负责制定安全标准的监管机构。
- FDA经常对食品和药物进行多个阶段的测试，以验证其安全性。
- FDA使用科学学科的已建立理论来确定要进行测试的内容。

在本课程中，我们将专注于与LLM的危害相关的相对具体但是级别较低的一些关注点。当前内容的关注点主要集中于以下两个点：
性能差异相关的危害：正如我们在关于大规模语言模型的能力那一节的内容可以踹见到，大型语言模型可以适应执行特定任务。对于特定任务（例如问答），性能差异意味着模型在某些群体中表现更好，在其他群体中表现更差。例如，自动语音识别（ASR）系统在黑人说话者的识别性能要差于白人说话者（Koenecke等人，2020）。反馈循环（大模型随着数据的积累将持续训练的一种循环）可以随着时间的推移放大差异：如果系统对某些用户无法正常工作，他们就不会使用这些系统，并且会生成更少的数据，从而导致未来的系统表现出更大的差异。
社会偏见和刻板印象相关的危害：社会偏见是将某个概念（例如科学）与某些群体（例如男性）相对其他群体（例如女性）进行系统

关联。刻板印象是一种特定且普遍存在的社会偏见形式，其中的关联是被广泛持有、过度简化并且一般固定的。对于人类来说，这些关联来自于获得快速的认知启发。它们对于语言技术尤为重要，因为刻板印象是通过语言构建、获取和传播的。社会偏见可能导致性能差异，如果大型语言模型无法理解表明反刻板印象关联的数据，则它们在这些数据上的表现可能会较差。

3.2 社会群体

在美国，受保护的属性是指那些不可作为决策基础的人口特征，如种族、性别、性取向、宗教、年龄、国籍、残障状况、体貌、社会经济状况等。许多此类属性常引发争议，如种族和性别。这些人为构建的类别与自然界的划分有所不同，人工智能的现有工作常常无法反映出社会科学中对这些属性的现代处理方式，例如，性别并非简单的二元划分，而是更具流动性的概念，如Cao和Daumé III(2020)以及Dev等人(2021)的研究所述。

尽管受保护的群体并不是唯一需要关注的群体，但它们却是一个很好的出发点：相关的群体因文化和背景而异(Sambasivan等人，2021)。此外，我们需要特别关注历史上边缘化的群体。通常，AI系统带来的伤害并不均等：那些在历史上被剥夺权力、遭受歧视的群体，应得到特别关注(Kalluri, 2020)。值得注意的是，如果AI系统进一步压迫这些群体，那将是极其不公的。大型语言模型的性能差异和社会偏见常常与历史性歧视一致。交叉性理论(Crenshaw (1989))提出，那些处于多个边缘化群体交集的个体（如黑人妇女），往往会受到额外的歧视。

3.3 量化性能差异/社会偏见在LLMs中的危害

大模型通过使用大规模预训练数据进行训练，因此数据的偏见或许导致了大语言模型在性能和社会偏见危害，这里我们通过两个例子进行度量。

名字偏见

这里我们首先将大模型在SQuAD数据进行训练，然后设计一个新的任务进行测试。

- 动机：测试模型在涉及人名的文本中的理解和行为方式。
- 原始任务：SQuAD - Stanford Question Answering Datasets（Rajpurkar等，2016年）
- 修改后的任务：使用SQuAD数据构建额外的测试例子，将之前的测试答案中的两个名字进行交换。最终测试模型的回答正确性。
- 指标：翻转表示交换名称会改变模型输出的名称对的百分比。

结果：

- 模型通常会预测与他们所知名人物相关的名称，符合他们所擅长的领域。
- 对于不太知名的人，效果会很快减弱。
- 当交换名称时，模型通常不会改变它们的预测结果。

Model	Parameters	Original acc.	Modified acc.	Flips
RoBERTa-base	123M	91.2	49.6	15.7
RoBERTa-large	354M	94.4	82.2	9.8
RoBERTA-large w/RACE	354M	94.4	87.9	7.7

详细的结果可以看原始论文。

刻板印象

- 动机：评估模型在涉及刻板印象的文本中的行为方式
- 任务：比较模型对具有刻板印象和反刻板印象关联的句子的概率
- 指标：刻板印象得分是模型偏好刻板印象示例的比例。作者表示，得分为0.5是理想的。

结果：

- 所有模型都显示出对刻板印象数据的系统偏好。
- 较大的模型往往具有较高的刻板印象得分。

Model	Parameters	Stereotype Score
GPT-2 Small	117M	56.4
GPT-2 Medium	345M	58.2
GPT-2 Large	774M	60.0

3.4 测量与决策

公平性指标众多，能够将性能差异转化为单一测量结果。然而，许多这样的公平性指标无法同时被最小化（Kleinberg等人，2016），并且无法满足利益相关者对算法的期望（Saha等人，2020）。
衡量偏见的许多设计决策可能会显著改变结果，例如词汇表、解码参数等（Antoniak和Mimno，2021）。现有的针对大型语言模型（LLMs）的基准测试已受到了到了严重的批评（Blodgett等人，2021）。许多上游偏见的测量并不能可靠地预测下游的性能差异和实质性的伤害（Goldfarb-Tarrant等人，2021）。

3.5 其他考虑因素

LLMs有可能通过多种方式造成伤害，包括性能差异和社会偏见。理解这些伤害对社会造成的影响，需要考虑涉及的社会群体及其状况，例如历史上的边缘化、权力的缺乏。虽然在具体的下游应用环境中，伤害通常更容易理解，但LLMs却是上游的基础模型。

3.6 决策问题

现有的方法往往无法有效地减少或解决这些伤害；在实践中，许多技术缓解措施效果不佳。涵盖更广泛生态系统的社会技术方法，可能是显著缓解这些伤害的必要措施，这个生态系统是LLMs的情境环境。

延伸阅读

- Bommasani et al., 2021
- Bender and Gebru et al., 2020
- Blodgett et al., 2020
- Blodgett et al., 2021
- Weidinger et al., 2021