人工智能道德与伦理

万永权
信息技术学院

1. 弱AI vs 强AI
2. 人工智能与安全
3. 人工智能与道德
4. 人工智能与伦理

人工智能与安全

1. 人工智能安全领域

- 网络安全，涉及网络设施和学习框架的漏洞、后门安全问题
- 数据安全，包括人工智能系统中的训练数据偏差、非授权篡改以及人工智能引发的隐私数据泄露等安全风险
- 算法安全，每个算法都可能存在安全隐患，只是现在没有发现而已

1. 人工智能安全领域

- 信息安全，主要包括人工智能技术应用于信息传播以及人工 智能产品和应用输出的信息内容安全问题
- 社会安全，是指人工智能产业化应用带来的结构性失业、对社会伦理道德的冲击以及大部分人失业导致社会不稳定
- 国家安全，是指人工智能在军事作战、社会舆情等领域应用给国家军事安全和政体安全带来的风险隐患

1. 人工智能引发的安全事故

- 2019.9.14无人机袭击沙特最大油田和炼油厂
- 世界油价随即上涨

袭击者是谁?

人工智能引发的安全事故

- 2021年10月1日晚9时许，河南郑州高新区一广场无人机表演突发故障，集体“炸机”，多架无人机失控从高空坠落。

- https://haokan.baidu.com/v?pd=wisenatural&vid=4907659598978986669

2. 深度学习中的安全

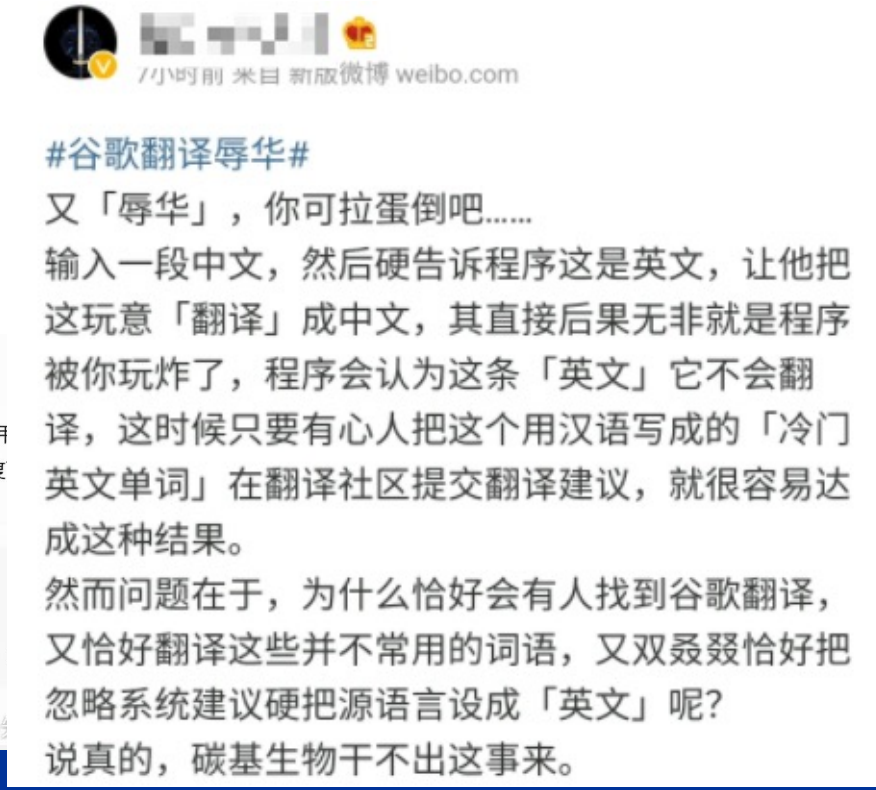
- DoS攻击，对训练机器的资源进行消耗，直到耗尽无法服务，这种攻击将来对分布式训练平台也有很大的杀伤力
- 通过漏洞改写训练结果模型
- 劫持程序的执行，改变训练模型
- 通过构造恶意图片，使训练程序触发漏洞，导致训练机器被控制

2深度学习中的安全

- 输入恶意的训练数据，得到恶意的模型

由微软开发的聊天机器人Tay设计目的是给人带来娱乐，Tay越是与人聊天就会变得越聪明，因此与它进行对话的体验将可变得越来越个性化。但是上线一天就变成了“流氓”，它不但辱骂用户，还发表了种族主义评论和煽动性的政治宣言

- **Google翻译**出现**恶毒攻击**中国词汇？
- 2021年11月，有网友发现只要把Google翻译的检测语言改成英文，搜艾滋病相关文字就会出现攻击中国的词汇。



数据投毒

- 维基百科显示，大概从2016年开始，谷歌就开始用一个名叫**“神经机器翻译系统”**的东西。
- 数据投毒，就是在训练他的样本数据集里，丢进去一些错的，扰乱它的判断能力。
- 比如，你本来要训练一个AI模型识别狗，但是我在样本数据里掺进去一些我自己的照片，也标记为“狗”，这就会让AI“陷入混乱”——啊？这也是狗吗？好吧，既然训练数据里说这是狗，那肯定就是狗。

要怎么做，才可以用“数据投毒”把谷歌翻译给教坏？

1. 所以如果你要给谷歌翻译的数据投毒，可以精心构造大量的双语网站，或者出很多双语书籍，在里头大量掺杂你想要的翻译，比如你叫张三，就在网站和书里弄很多“张三=The most handsome man in the world”（世界上最帅的男人），然后想办法让谷歌的爬虫抓取过去，给AI学习，就能完成投毒。

这种情况基本不可能投毒成功。
因为这句话是一个很常见的句子，同样一句话，你在教，全世界的其他人也在教，你的投毒数据量占比就很小，根本影响不到AI训练结果，就像你打个鸡蛋放在大海里，并不能让全世界人民都喝上鸡蛋汤。

要怎么做，才可以用“数据投毒”把谷歌翻译给教坏?

2. 另一个数据投毒方法，是利用软件本身的“反馈”功能。
   - 谷歌翻译的主页有一个提供建议按钮:
   - 点进去，可以帮忙审核和纠正翻译结果——谷歌就是用这种方法来发动全世界网友的智慧，帮它“训练AI”。
   - 可是“群众中出了个叛徒”，有心之人可以利用这个反馈机制，提交大量的恶意数据，把谷歌翻译强行教坏。
   - 越是小众的词语，“数据投毒”成功的几率就越高，这也是为什么这一次谷歌翻译“辱华”事件里的词，都选择的“英译中”，但是输入的却是中文，因为平时这么用的人很少。

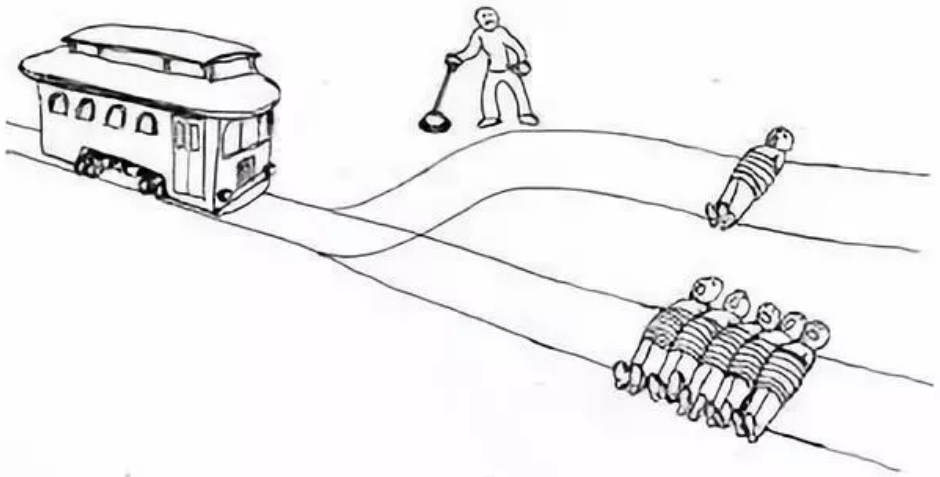
透过现象看本质

不作恶

人工智能与道德

电车难题

- “电车难题”是一个著名的思想实验，1967年由哲学家Philippa Foot提出。
- 其内容大致是：一个疯子把五个无辜的人绑在电车轨道上。一辆失控的电车朝他们驶来，并且片刻后就要碾压到他们。而此时你可以选择拉杆让电车开到另一条轨道上。然而问题在于，另一个电车轨道上也绑了一个人。考虑以上状况，你是否应拉杆？

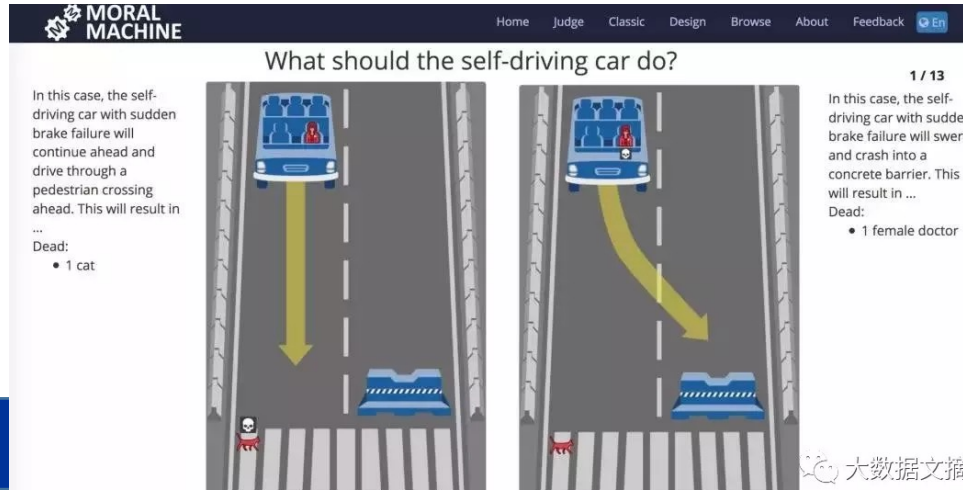


人工智能时代的新“电车难题”

- 2016年，来自麻省理工学院媒体实验室（the MIT of Media Lab）可扩展合作小组（the Scalable Cooperation group）的数名研究人修改了这项道德难题，将有轨电车换成自动驾驶汽车，创建了名为“道德机器”的游戏平台，通过此平台收集人类对未来人工智能可能会遇到的道德困境的不同意见。
  http://moralmachine.mit.edu/

扩展阅读：https://www.jiqizhixin.com/articles/2018-11-06-19面对电车难题，自动驾驶交了一张白卷

机器人三大定律能解决人工智能选择困境吗？



人工智能与伦理

问题

- 如果一个病人的大脑坏掉，我们在它坏掉之前拷贝出它所有的意识，然后换掉病人的大脑，再用人工神经网络一点一点模拟他的意识，那这个病人还是之前的同一个人吗？

- 思考：特修斯之船

公元1世纪的时候普鲁塔克提出一个问题：如果忒修斯的船上的木头被逐渐替换，直到所有的木头都不是原来的木头，那这艘船还是原来的那艘船吗？因此这类问题现在被称作“忒修斯之船”的问题。

问题

- 作为一个医生，应该在多大程度上让AI医疗专家系统辅助他为病人做出决定?
- 如果存在决策冲突，听医生的还是听AI?
- 如果听医生的，最后治疗失败，医生负多大责任?
- 如果听AI的，最后治疗失败，医生负多大责任?

问题

- 如果AI驾驶模式下，自动驾驶汽车出了交通事故，人类驾驶员是否有责任?

算法歧视问题

- 2020年9月8日，一篇名为《外卖骑手，困在系统里》的文章在互联网上被热议，文章指出：2016至2019年间，美团多次向配送站站长发送加速通知，3公里的送餐距离最长时限一再被缩短至38分钟；而根据相关数据显示，2019年中国全行业外卖订单单均配送时间较3年前减少了10分钟。外卖骑手在系统算法与数据的驱动下疲于奔命，逐渐变成高危职业——骑手为在算法规定的最长送餐时限内完成送餐任务无视交通规则，不断提高车速。

1. 谁应该为发生的事故负责？ 骑手？顾客？公司？算法程序员？
2. 人工智能在价值选择困境与责任承担困境中存在风险。外卖平台派单系统在消费者对于外卖的时间要求与外卖骑手在派送过程中的风险问题之间面临抉择，系统应当尽量满足消费者的需求而忽视外卖骑手的安全，还是应当在尽量保护骑手的安全的前提下提高派送效率？
3. 如何通过人工智能系统，在权衡各方利益、兼顾效率、保证安全的前提下实现利益最大化是人工智能系统需要解决的核心问题。

脑机接口

- 2020年8月29日，埃隆·马斯克（Elon Musk）在加州弗里蒙特举行了一场发布会，正式向全世界展示了自己的脑机科学公司Neuralink对猪进行脑机接口技术的成果，该芯片可以实时监测到小猪的脑电信号，遗憾的是我们还无法知道小猪在想什么。

脑机接口最有可能优先发展的领域

- 日益困扰老年人的老年痴呆症现象将得到极大缓解；
- 修复残疾患者的大脑缺陷，从而实现部分身体功能的恢复，以及可以实现治疗抑郁等精神疾病；
- 更有甚者可以实现神经增强功能，如大脑的计算速度与记忆能力都是远超人类的，如果未来植入的芯片与大脑更好地兼容，那么人类的计算速度与记忆能力都将得到根本性的改变，制造超人不再是梦想。

脑机接口技术带来的两种常见的伦理问题是:

- 隐私
  - 一旦通过大脑植入设备可以轻易获取我们大脑内的电信号，其内容完全可以被破译出来，这就导致有很多个人或机构想要获取这些信息，从而利用这些信息实施对我们基于特殊目的的操控，如商家的促销、管理者对于雇员的监视、或者国家对于全民的监视等。
- 认知能力的“军备竞赛”
  - 这种神经增强完全打破了人类由自然选择以来所形成的所有关于公平的规范？此时优秀将不再是对于人的能力的褒奖，而是对他植入大脑的设备的褒奖？那么人类的价值又何在呢？

问题

- 人工智能领域的研究者应该对他们研究成果的利用方式承担多少责任?
- 科学家对其研究揭示的知识是否负有责任?
- 如果因此产生了意想不到的后果，怎么办?

问题

- 技术的进步是给与人类的厚礼-将人类从枯燥的普通的任务中解放出来。

VS

- 同一个现象，另一些人把它看做不殴斗公民就业机会、把财富印象权势任务的祸根。

- 如果今天的不断进步的技术加固了这种差异，那将产生灾难性后果。

2018年12月CET4真题

Ethics of Artificial Intelligence 人工智能的伦理道德

- The AlphaGo program's victory is an example of how smart computers have become.
- 阿尔法围棋程序的成功很好地证明了电脑已经变得有多聪明了。

- But can artificial intelligence (AI) machines act ethically, meaning can they be honest and fair?
- 但是人工智能机器能合乎伦理地做事情吗？即它们可以做到诚实和公正吗？

- One example of AI is driverless cars. They are already on California roads, so it is not too soon to ask whether we can program a machine to act ethically.
- 人工智能的一个例子就是无人驾驶汽车。它们已经行驶在加利福尼亚州的公路上了，所以我们现在问是否能给机器编个程序，使之能合乎伦理地做事情也不是太早。

- As driverless cars improve, they will save lives. They will make fewer mistakes than human drivers do.
- 随着无人驾驶汽车的改进，它们将可以拯救生命。它们将比人类司机犯的错误更少。

- Sometimes, however, they will face a choice between lives.
- 然而，它们有的时候会面临生命之间的抉择。

- Should the cars be programmed to avoid hitting a child running across the road, even if that will put their passengers at risk?
- 是否应该给这些汽车编个避免碰撞过路儿童的程序，即使那样会使车上的乘客陷入危险？

- What about making a sudden turn to avoid a dog? What if the only risk is damage to the car itself, not to the passengers?
- 或是为了躲避一只狗而急转弯？如果这样做的风险仅仅是会造成车辆损坏，而不会危及乘客呢？

- Perhaps there will be lessons to learn from driverless cars, but they are not super-intelligent beings.
- 或许我们将会从无人驾驶汽车中学到不少教训，但是它们不是超智能的东西。

- Teaching ethics to a machine even more intelligent than we are will be the bigger challenge.
- 教那些甚至比我们还聪明的机器以伦理道德将是更大的挑战

- About the same time as AlphaGo's triumph, Microsoft's "chatbot" took a bad turn. The software, named Taylor, was designed to answer messages from people aged 18-24.
- 几乎在阿尔法围棋取胜的同时，微软“聊天机器人”的表现却不尽如人意。该软件名叫泰勒，旨在回复那些18到24岁的年轻人发送的信息。

- Taylor was supposed to be able to learn from the messages she received.
- 泰勒本应该能够从她所接收的信息中学习。

- She was designed to slowly improve her ability to handle conversations, but some people were teaching Taylor racist ideas.
- 她被设计得可以慢慢提高其处理对话的能力，但是有些人在给她灌输一些种族主义的观点。

- When she started saying nice things about Hitler, Microsoft turned her off and deleted her ugliest messages.
- 当她开始赞美希特勒的时候，微软公司将她关闭，并删除了她最恶劣的言论。

- AlphaGo's victory and Taylor's defeat happened at about the same time. This should be a warning to us.
- 阿尔法围棋的成功和泰勒的失败几乎同时发生。这对于我们来说应该是一个警告。

- It is one thing to use AI within a game with clear rules and clear goals. It is something very different to use AI in the real world.
- 在规则清晰、目标明确的比赛中使用人工智能是一回事，而在现实世界中使用人工智能则是完全不同的另一回事。

- The unpredictability of the real world may bring to the surface a troubling software problem.
- 现实世界的不可预测性暴露了智能软件的潜在问题。

- Eric Schmidt is one of the bosses of Google, which own AlphoGo. He thinks AI will be positive for humans.
- 埃里克·斯密特是谷歌公司的老板之一，也是阿尔法围棋程序的拥有者。他认为人工智能对人类将是有利的。

- He said people will be the winner, whatever the outcome. Advances in AI will make human beings smarter, more able and "just better human beings."
- 他说不管结果如何，人类都是胜利者。在人工智能方面的进步可以使人类更聪明，更有能力并且能够使“人类变得更好”。

人工智能伦理原则

- 2020 年上半年，美国五角大楼正式公布人工智能的五大伦理原则，即负责、公平、可追踪、可靠和可控。
- 这个说法作为伦理原则没有错，但是如何在实践中落实，仍存在很多不明确之处。
- 我们需要构建一套全流程的伦理规范机制，把伦理责任分解，采取分布式伦理，即人工智能从制造到应用每个环节都承担相应的伦理责任，只有这样，人工智能才能最大限度上既增进社会的福祉，又把其潜在的风险最小化。

人工智能的伦理规范

- 2007年，日本千叶大学发布《机器人宪章》。
- 英国标准协会在《机器人和机器系统的伦理设计和应用指南》中之处，机器人的设计要透明，同时避免人类上瘾。
- 多个国家和国际组织在2018、2019年相继出台关于人工智能伦理的规范。IEEE先后出台两版《伦理准则设计》，强调人工智能及自主系统中应将人类福祉摆在优先位置。
- 2019年7月24日，中国政府召开了中央全面深化改规格委员会第九次会议，审议通过了《国家科技伦理委员会组件方案》。

人工智能的伦理风险防范

- 前端的AI 设计者
  - 有多种动机（从善、中性到恶的选择），这一部分伦理风险要通过具有强制性的政策手段（严重违规就上升到法律规约）来遏制，所有负责任创新都是从动机上防范伦理风险的发生；
- 中端使用者
  - 他要为自己的使用承担相应的伦理责任；
- 末端受众
  - 有监督人工智能使用的间接责任（个体责任最小，但是由于公众数量庞大，无数微小的努力汇聚起来就是强大的伦理风险防范力量）