**计算机伦理PPT展示发言稿**

大家好，今天我分享的是关于产权利益相关的话题，AI爬虫如何瓦解互联网伦理共识。

上个月，世界上最大的免费百科--维基百科将数据集作为JSON文件发布供AI训练，原因是AI爬虫使得维基百科的带宽成本较上一年增加50%；开发火狐浏览器的公司Mozilla根据后台数据发现网站流量前五名全是爬虫，占了网站七成流量，使得网站管理员在社交媒体上怒喷这些爬虫的行为；这些数据大厂的行为不是孤例，而是AI爬虫技术在全球范围内带来的普遍冲击

这不禁让人疑问：爬虫是否在DDos攻击整个互联网呢？

让我们先看一下爬虫的来世今生，爬虫本质上就是一个自动化程序，它能自动访问网页、抓取内容并存储到数据库。传统爬虫最大应用就是搜索引擎。在搜索引擎时代，比如百度、Google的爬虫会遍历全网，建立索引，让用户能快速找到信息。这是一个双赢的模式——爬虫帮网站带来流量，用户也能通过搜索引擎快速获取信息。但任何一项技术都有被滥用的风险：爬虫频繁抓取给网站带来服务器压力、带宽消耗，甚至挖掘到敏感数据，俗话说啊“爬虫学得好，牢饭吃得早”。于是，网站使用robots.txt文件，告诉爬虫哪些能抓，哪些不能抓，虽然这只是一个君子协议，但是大家都默契地遵守。

随着AI的发展，爬虫进化成AI爬虫，AI爬虫的最大应用场景就是大模型训练的数据采集。如果说搜索引擎爬虫是信息搬运工，那么AI爬虫更像是一位数据掠夺者，打破了搜索引擎时代 爬虫和网站之间双赢的格局。

一方面，它更加贪婪，无差别抓取数据，特别是图片，视频，知名的反爬网站Cloudflare统计，每天有超过**500亿次爬虫请求**试图突破网站防御。

另一方面，它剥削网站，很多AI爬虫抓取数据后，不会给原网站任何回报；它也很无耻，突破协议限制，违规抓取，无视robots协议；

为什么Robots协议会在AI时代失效呢？首先robots协议并非强制性，在中国，违反robots.txt可能被认定为“不正当竞争”，但不会构成犯罪；然后就是商业利益的驱动，数据就是金钱，企业为了降低成本，会尽可能多地抓取数据，最后是技术不对等，AI爬虫和反爬机制都在不断发展，但对于小网站而言，反爬的成本也逐渐变大；最常见的反爬机制就是验证码，比如右边这个验证码，不知道大家会不会选右上角这一块，据说不选这一块更像真人。

即使AI爬虫遵守了robots.txt，争议也不会消失 —— 真正的问题是，谁有权决定我们的数据被如何使用？争议也随之而来，形成了两大对立阵营：

一方面，以AI公司和开发者为代表的技术派认为数据是社会公共资源，合理利用数据能创造更大价值，从而造福全社会，而AI爬虫抓取数据并非复制，而是通过学习创造新价值；另一个方面，个人用户/网站为代表的权利派认为AI爬虫是在进行“数据殖民”**。**用户没有授权，也没有得到任何补偿，却被默认贡献了内容用于商业训练，这是一种典型的隐形剥削。他们反对的是默认抓取、商业滥用和权利失衡；

比如2023年12月美国纽约时报起诉openAI和微软未经许可使用其辛辛苦苦创作的内容训练AI模型，而openAI的回应是数据是公开的，模型训练是公平使用并且有利于创新。

这并非简单的立场之争，而是互联网结构的系统性失衡，技术力量正在碾压社会共识，AI公司凭借更强算力绕过协议限制，而普通创作者则很难维权，我们正在走向算法霸权的格局，robots协议代表的互联网伦理共识正在被瓦解，而伦理的缺失终将会反噬技术本身，使开放互联网封闭化，内容创作减少等；

AI确实需要数据，但也必须尊重权利；爬虫确实推动发展，但不能践踏底线。

AI需要爬虫，爬虫需要互联网，互联网需要AI，爬虫给网站带来流量，网站给爬虫带来内容，正是这种需求维系了过去30年的robots协议，在AI时代，我们需要一份**新的“共识协议”** ——它应当是**公平的、透明的、可控的**，能平衡创新与权益，建立人与算法的边界感。

最后，正如robots协议的起草人科斯特当年所说：“争论爬虫的好坏没有意义，我们真正需要一个问题最小化，利益最大化的系统”。

谢谢大家！