关于「木兰-白玉兰开放数据许可协议」

开源开放是全球人工智能产业呈现加速发展态势的重要驱动要素，有效提高了人工智能研发效益，加速了人工智能技术创新，促进人工智能生态构建。在技术研究、产品开发等环节，更多创新主体能够基于相对成熟的开源软硬件平台，利用已有公共基础研发资源来加速人工智能研发。在此过程中，数据要素的自由流通变得越来越重要，数据开放成为推动人工智能创新发展的关键一环。

然而，当前在人工智能领域尚缺乏切合实际的开放数据许可，使得数据要素的使用和流通仍存在诸多障碍和不确定性，不仅容易产生数据安全及法律方面问题，而且由于对数据本质属性、存在形态、使用方式等方面仍认识不足，造成数据资源的使用现状与数据可供挖掘的价值不匹配。为了人工智能技术和相关数据资源的可持续开发使用，开放数据许可协议的作用愈加凸显。通过规范数据利益相关方的身份和概念定义，界定数据利益相关方之间对特定数据对象流通条件和方式的各自权责,并尽可能以完全开放的模式引导数据流通，促进数据要素的开放共享与开发利用。

「木兰-白玉兰开放数据许可协议」是由「上海白玉兰开源开放研究院」在「木兰开源社区」框架和精神下所发起的一项研究项目，旨在探索创建一组标准化的、立足中国人工智能实践、推动数据要素流通、优化人工智能发展环境的数据许可协议。

「木兰-白玉兰开放数据许可协议」起草说明

协议的草拟由「白玉兰开源」联合「开放数据中国」完成，过程中我们对：

- 国际通用开放协议如知识共享协议、开放数据库协议（ODbL）等做了研读和理解，并将其中的术语、起草策略等加以总结和归纳
- 国际社群人工智能领域数据流通的授权协议如微软起草的 O-UDA、C-UDA，Linux Foundation 起草的 Community Data License，Element AI 起草的 Montreal Data License 等做了研读和理解，并基于 Montreal Data License 的精神，对术语中规定的使用行为做了人工智能界别的定制化和细致化。
- 对中国现行民法典，以及数据安全法草案、个人信息保护法草案等予以研读，并借鉴了其中相关的术语定义

考虑到数据要素流通的合规复杂性，当前草拟版本基于如下原则和适用性拟定：

- 针对人工智能训练数据集的发布拟定适用的协议
- 所发布数据应满足基本的公开发布、免费发布的前提
- 所发布数据符合国家数据安全的要求，不涉及国家秘密、国家安全、社会公共利益、商业秘密等
- 所发布数据不涉及个人信息(参照「《个人信息保护法（草案）》（二次审议稿）,个人信息是以电子或者其他方式记录的与已识别或者可识别的自然人有关的各种信息，**不包括匿名化处理后的信息**)

考虑到当前人工智能训练数据集从权属角度可分为两类情况：

- 第一类，数据由数据发布者合法合规所有或具备用益权
- 第二类，数据由数据发布者通过合法合规的方式自第三方处获取汇编组合而得

因此「木兰-白玉兰开放数据许可协议」对上述两类情况产出了两组不同起草策略的协议：

第一组，即默认数据由数据发布者合法合规所有或具备处置权

我们借鉴知识共享协议的模式，草拟了一套 4 份协议，即

- **MBODL**：宽松开放协议，适用于最小化限定仅要求注明数据来源的数据发布
- **MBODL-NC**：非商业使用协议，适用于禁止使用者商业化使用和分享数据及成果
- **MBODL-SA**：相同方式许可，适用于要求下游传播数据能够以相同方式给予许可，但不要求对产出的成果使

用协议的传染性

- **MBODL-CU**：仅计算使用协议，适用于数据发布方禁止对数据自身的直接使用、展示的情况（如电视台作为数据发布方会希望禁止视频数据本身的播放、拷贝、售卖等，但会允许使用视频数据作为训练数据训练视频语义标签等任务）

上述四个协议，均以 MBODL 为基础，在「许可限制」小节中予以增加不同的限制而形成。但正如 CC 协议，在这 4 套协议的基础上，也可再进行许可限制的叠加交叉，形成新的协议，如 MBODL-NC-CU，即规定非商业使用且仅计算使用，又如 MBODL-SA-CU，即规定相同方式授权数据且仅计算使用。

第二组，即数据发布者数据为自第三方合法合规获取

我们借鉴了 ODbL（开放数据库协议）的策略，对数据库/数据集的结构（即数据选取、组织的方式，database scheme）和数据内容予以了拆分授权的方式。此类授权策略仅为实验性，待进一步反馈确定 1）是否有真实需求 2）是否具备可操作性。

对于上述第二组的情况，我们提供两个可能的案例展开说明：

案例 1：数据发布者通过 wikipedia 和 flickr 等渠道获取了各类鸟类的图片数据，图片数据各自分别授权在 CC 等开放授权协议下，数据发布者通过选取和组合这些鸟类图片，添加了自身对鸟类的标签（鸟类照片对应的鸟类名称、科目等信息），最后形成了一个「鸟类图片训练数据集」需要授权发布。则在第二组协议的策略下，其将采用「白玉兰开源开放数据协议」（仅授权结构）+「标注数据」（授权内容-发布者选用新授权）+「各图片原有协议」（授权内容-依照各自协议）的方式授权发布整个数据集。

案例 2：数据发布者通过授权方式（假设授权允许发布者重新发布影像图片）从 N 家医院各自获取了脱敏后的肺部 CT 影像图片数据，数据发布者自身投入人力完成了对上述影像数据的肺结节标注。数据发布者希望将影像图片数据+标注数据组合发布为「肺结节标准训练数据集」，因此可采用其将采用「木兰-白玉兰开放数据许可协议」（仅授权结构）+「标注数据」（授权内容-发布者选用新授权）+「各图片原有协议」（授权内容-依照各自协议）的方式授权发布整个数据集。

我们基于上述案例的场景描述，草拟了 **MBODL（结构内容分离版）协议**，作为一个单独的实验性协议供各界讨论适用性和条款的实践落地可能。

贡献

我们欢迎各界针对我们草拟的协议予以反馈，并通过 PR 、Issue 的方式提出相应的修改意见或提出问题。