拟公示算法机制机理内容

算法名称	哔哩哔哩内容安全算法
算法基本原理	内容安全审核算法通过样本标注、知识库等方式学习专家经验，从而具备负向内容识别能力，应用于后续新增或者历史场景的快速和大规模识别任务。
算法运行机制	算法系统包含三个部分： 内容理解模块：识别负向内容的模型能力 运营工具：通过知识库准实时新增敏感词、黑样本、人脸等识别能力；提供视频、图文业务场景的历史数据回扫功能。 标训模块：对审核人员判断有风险的内容进行细标，回收标注结果并更新训练数据、迭代模型；另外也会针对安全专家判断可能有风险的内容制定识别方案。
算法应用场景	在审核平台的视频、直播、图文审核业务流程中，哔哩哔哩内容安全算法面向审核提供对 ugc 内容的理解和识别能力。
算法目的意图	通过内容理解能力识别出 ugc 内容中的不符合法律法规、社会价值观的部分，用于后续的人工或者自动处置
算法公示情况（选填）	