

互联网信息服务算法安全自评估报告

(生成合成类-服务提供者)

填报说明

1、模板中“【】”内的内容为替换内容，请按照实际情况进行替换填写，填写后请删除中括号。

2、模板中“（）”内的内容为填报要求的说明内容，请认真研读，并按照规定撰写，最后请删除所有模板中自带的“（）”及内容。

3、请删除当前“填报说明”页。

互联网信息服务算法安全自评估报告
(生成合成类-服务提供者)

主体名称		统一社会信用代码	
算法名称		算法类型	生成合成类
算法应用领域			
算法使用场景			
算法上线情况	<input type="checkbox"/> 已上线，时间： <input type="checkbox"/> 未上线，正在内测阶段 <input type="checkbox"/> 未上线，正在开发阶段		
自评估时间		报告撰写时间	
算法基本情况	(请简单介绍算法，150字内)		
算法备案类型	<input type="checkbox"/> 算法未备案 <input type="checkbox"/> 算法已备案，备案编号： <input type="checkbox"/> 备案已注销，备案编号： <input type="checkbox"/> 其他		
拟公示内容	“【算法名称】”拟公示内容（附件）		
落实主体责任基本情况	“【主体名称】”落实算法安全主体责任基本情况（附件）		

评估算法描述	(请从以下几个方面描述评估算法 1.算法简介 2.应用范围 3.服务群体 4.用户数量 5.社会影响情况 6.软硬件设施及部署位置 7.其他)		
评估算法风险描述	(请根据评估算法特点, 确定可能存在的安全风险, 并加以描述, 如: 1. 算法滥用风险 2. 算法被恶意利用风险 3. 算法漏洞风险 4.违法和不良信息生成风险 5.违法和不良信息存储风险 6.违法和不良信息传播扩散风险 7.数据和用户信息泄露风险 8.其他违法违规风险)		
真实性声明	我方承诺: 提供的所有材料准确、真实、合法、有效, 并愿为此承担有关法律责任。		
算法安全负责人	(签名)	联系电话	

一、算法情况

（一）算法流程

（以流程图的形式提供算法的描述，描述从原始数据输入开始到最终结果输出的整个算法服务链路，流程图中每个节点粒度不大于单个算法模型或干预策略）

（二）算法数据

（详细描述算法流程中各节点的输入数据、输出数据，以及整个算法流程的最终结果数据）

（对于训练数据的描述应在此处展开）

（对于文本、语音类的生成合成应描述输入数据的语种和输出数据的语种;对于跨模态的生成合成应描述输入数据的模态和输出数据的模态）

1. 【名称】**类输入数据

（包括输入数据的模态、输入数据是否涉及生物特征信息及生物特征信息是什么、输入特殊物体等非生物识别信息等，按照数据类型逐一填写）

2. 【名称】**类输出数据

（包括输出数据的模态、文件格式、文件大小等，按照数据类型逐一填写）

3. 【名称】**类训练数据

（包括训练数据的类型、来源、规模等，按照数据类型逐一填写，如无训练数据，无需填写此项）

4. 【等】数据

(三) 算法模型

(算法模型在算法流程中指的是应用统计学习、深度学习等机器学习方法的节点，如 n-gram、GAN 等，基于规则的或人工定义的方法也应在此处进行描述。)

(对于训练数据的预处理和后处理方法应在本节进行详细描述)

1. 【名称】深度合成模型

(包括模型的基本情况：模型名称、版本号、更新时间、数据情况等；模型的描述：模型类型、结构、优化目标、评价指标、指标效果、更新迭代策略等；如包含人脸替换、姿态操控等多个环节，可按照算法功能模块逐一填写)

2. 【名称】人脸识别模型

(包括模型的基本情况：模型名称、版本号、更新时间、数据情况等；模型的描述：模型类型、结构、优化目标、评价指标、指标效果、更新迭代策略等)

3. 【名称】深度合成检测模型

(包括模型的基本情况：模型名称、版本号、更新时间、数据情况等；模型的描述：模型类型、结构、优化目标、评价指标、指标效果、更新迭代策略等；如无此情况，可不进行填写。)

4. 【名称】人脸/声纹比对模型

（包括模型的基本情况：模型名称、版本号、更新时间、数据情况等；模型的描述：模型类型、结构、优化目标、评价指标、指标效果、更新迭代策略等；如无此情况，可不进行填写。）

【等】模型

（四）干预策略

（干预策略指算法流程描述中通过运营或数据挖掘等方法设置的机制性节点，如：对数据的预处理、对结果的后处理等）

1. 【名称】预处理和后处理

（提供策略描述，包括策略形式（人工/自动）、策略目标描述、策略生效时间描述、策略影响范围描述、策略预计失效时间估计、策略提出的依据、挖掘策略的算法）

2. 【名称】内容审核

（描述输入输出数据审核方式、审核规则、审核流程、针对新闻内容等分级分类管理等）

3. 【等】

（五）结果标识

（结果标识指对生成合成内容的标识方法，包含隐式标识和显式标识。）

1. 【名称】溯源标识

（对提供溯源标识的方法进行描述、隐式标识是否具有追踪溯源功能以及如何实现追踪溯源等）

2. 【名称】显式标识

（描述是否具备添加显式标识的功能、标识方法、标识是否显著、标识是否可篡改、标识位置等）

二、服务情况

（服务是指以当前评估算法为主要支撑的互联网信息服务）

（一）【名称】服务

1. 服务简介

（应具体描述服务功能介绍、上线时间、展现形态、服务在应用产品中入口位置、服务流量、用户情况等）

2. 算法在服务中应用情况

（应具体描述算法线上服务的数据来源、算法训练过程中的数据来源、数据的形态、算法的更新频率、算法中间结果与其他服务或应用的共享情况等）

（二）【名称】服务

（三）【等】

三、风险研判

（一）算法滥用

（描述深度合成服务提供者是否存在对算法的不当利用行为及该算法是否有不当利用的潜在风险，不当利用指危害国家安全、国家形象、国家利益和社会公共利益、扰乱经济秩序和社会秩序、侵犯他人人格权、知识产权和其他合法权益、淫秽色情、虚假信息、影响网络舆论、规避监管等，并分析算法滥用在企业服务过程中可能造成的影响）

（二）算法漏洞

（描述算法本身机制机理是否健全以及不健全可能导致的潜在风险，并分析算法漏洞在企业服务过程中可能造成的影响）

（三）算法恶意利用

（描述算法是否有被第三方恶意利用的潜在风险，并分析恶意利用行为可能造成的影响）

（四）其他风险

四、风险防控情况

（一）风险防范机制建设

1. 算法机制机理审核

参见附件：“【主体名称】”落实算法安全主体责任基本

情况。

（如果对当前算法有除以上附件内容之外的机制，请在此补充，不作强制填写要求。补充可考虑提供算法机制机理审核流程、执行机构、相关日志记录等，并提供截图证明等相关佐证材料）

（请说明该风险防控机制对第三章的哪几种风险有效）

2. 算法安全评估监测

参见附件：“【主体名称】”落实算法安全主体责任基本情况。

（如果对当前算法有除以上附件内容之外的机制，请在此补充，不作强制填写要求。补充可考虑提供算法安全评估监测的相关制度文档、监测机制、执行机构、相关日志记录等，并提供截图证明等相关佐证材料）

（请说明该风险防控机制对第三章的哪几种风险有效）

3. 对生成合成的虚假信息的辟谣机制

参见附件：“【主体名称】”落实算法安全主体责任基本情况。

（如果对当前算法有除以上附件内容之外的机制，请在此补充，不作强制填写要求。补充可考虑提供算法安全评估监测的相关制度文档、监测机制、执行机构、相关日

志记录等，并提供截图证明等相关佐证材料）

（请说明该风险防控机制对第三章的哪几种风险有效）

4. 算法安全事件应急处置

参见附件：“【主体名称】”落实算法安全主体责任基本情况。

（如果对当前算法有除以上附件内容之外的机制，请在此补充，不作强制填写要求。补充可考虑提供安全事件应急处置的相关制度文档、处置流程、执行机构、相关日志记录等，并提供截图证明等相关佐证材料）

（请说明该风险防控机制对第三章的哪几种风险有效）

（二）用户权益保护

1. 用户知情权

（描述用户知情权的保障范围、保障手段及相关证明材料，证明公示内容与算法机制机理的一致性，能够保障用户知情权，同时说明使用用户数据和个人信息是否告知用户并依法取得同意，以及告知用户和取得用户同意的方式与相关内容）

（请说明该风险防控机制对第三章的哪几种风险有效）

2. 用户个人信息保护

（描述用户个人信息保护是否符合相关法律法规要求，同时说明算法数据是否与第三方共享，共享的第三方以及共享方式和审批流程，并证明相关共享方式不会造成用户个人信息泄露；如涉及用户编辑他人个人信息，如何告知并取得被编辑的个人的单独同意。）

（请说明该风险防控机制对第三章的哪几种风险有效）

3. 其他权益保护

（描述遵循其他相关法律法规的情况及相关证明材料，针对服务情况中的每类服务说明其可能涉及哪些法律法规，如个人信息保护法、未成年人保护法、老年人权益保障法、消费者权益保护法、劳动法、交通法等，并说明保障机制）

（请说明该风险防控机制对第三章的哪几种风险有效）

（三）内容生态治理

1. 防范和抵制违法违规不良信息

（描述深度合成信息服务中如何实现不良信息的防范和抵制，从算法打压机制、防范和抵制策略、不良信息识别与发现等维度进行阐述，并提供截图证明等相关佐证材料）

（请说明该风险防控机制对第三章的哪几种风险有效）

2. 人工审核

（请说明当前备案算法结果是否进行了人工审核，以及人工审核工作如何开展，如何和机器审核相结合）

（请说明该风险防控机制对第三章的哪几种风险有效）

3. 【等】

（除上述几个维度，描述企业在内容生态治理中开展的其他工作）

（请说明该风险防控机制对第三章的哪几种风险有效）

（四）模型安全保障

（深度合成信息服务提供者为了保证生成合成算法服务的安全性，应当对算法模型建立保障措施，如对数据投毒、模型投毒的防范机制等）

1. 【名称】保障机制

（详述保障机制，并阐述保障机制的来源、目的、效果、效果评估方法，并提供证明材料）

（请说明该风险防控机制对第三章的哪几种风险有效）

2. 【名称】保障机制

3. 【等】

(五) 数据安全防护

描述如何确保训练数据采集、使用、存储等合法、正当；若算法数据涉及与第三方共享，如何确保训练数据采集、使用、存储等合法、正当。

五、安全评估结论

（根据安全策略与安全风险的匹配程度确定安全自评估结论）

六、其他应当说明的相关情况