ICS
CCS 点击此处添加 CCS 号

团 体 标 准

T/ZAITS XXXX—XXXX

智能应用模型算法合规性管理指南

Management Guidelines of Compliance for Model Algorithm

（征求意见稿）

ZAITS

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

浙江省智能技术标准创新促进会　　发 布

目　　次

前言 II
1　范围 1
2　规范性引用文件 1
3　术语和定义 1
4　管理原则 1
5　研发运维流程管理合规要点 2
6　内控管理合规要点 4
参考文献 7

ZAITS

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由浙江省智能技术标准创新促进会提出并归口。

本文件起草单位：蚂蚁区块链科技（上海）有限公司、杭州城市大数据运营有限公司、杭州海康威视数字技术股份有限公司、云从科技集团股份有限公司、上海商汤智能科技有限公司、北京旷视科技有限公司、上海依图网络科技有限公司、杭州魔点科技有限公司

本文件主要起草人：蔡年余、彭晋、简文、陆碧波、刘飞燕、朱天颖、钱晓东、梅敬青、吴庚、李军、方家乐、李洁

ZAITS

智能应用模型算法合规性管理指南

1 范围

本文件规定了企业在设计、研发、应用模型算法过程中应适用的原则，以及模型算法在设计、研发、测试、运营过程中的注意事项和控制目标等。

本文件适用于智能应用算法模型的设计、研发、测试、运营等活动，为法务、合规等风险管理部门及人员提供算法合规性自评估操作指南性建议。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 35273—2020 信息安全技术 个人信息安全规范

3 术语和定义

下列术语和定义适用于本文件。

3.1

模型算法 model algorithm

通过符合单一或系列预设的入参信息或条件后，通过机器语言编排的特定逻辑、规则等，得到预定结果或解决相应问题的的决策逻辑，使得原先的人工决策及结果获取更加高效快速的计算机指令。

3.2

有监督模型算法 supervised model algorithm

模型算法在整体设计开发及应用过程中，模型算法运用的训练数据集由人工操作进行标注，决策基础或变量，决策逻辑等在能够在人工干预下进行，符合高度的可解释性和透明性，不存在人工不能管控或感知的计算机指令。

3.3

无监督模型算法 unsupervised model algorithm

模型算法在整体涉及开发及应用过程中，模型算法运用的训练数据集不依赖人工标注，相应的决策基础或变量，绝大部分均依赖机器自主学习，完全依据统计学结果来形成相应决策逻辑或变量，人工不参与或很少参与其中的决策逻辑或条件，且不能有效完整阐述其模型算法决策过程及可能产生影响的计算机指令。

3.4

灰度测试 gray release

为了保证模型算法在最终大规模投产应用前，为了避免出现大规模风险事件，通过对筛选的一小部分人群开始，逐步扩大模型算法应用的适用客群，从而帮助企业快速发现问题，识别风险并及时进行改造升级迭代，从而更好的推出应用的测试步骤。

3.5

数据项颗粒度 data granularity

模型算法研发过程中所使用的数据明细程度，颗粒度约细，则对应的数据项与客户的原始明细约接近，反之则差距越大。

4 管理原则

在模型算法设计前，应由算法设计、合规等相关部门共同对模型算法应用需求的正当性进行评估：

a) **目的适当性。**在模型算法设计目的上，应考虑企业应用的合理、正当性，即企业生产相应模型算法的初衷，是否为了进一步提升服务效能、质量、安全等目的，反之是否存在侵害客户选择权或公平交易权等方面的可能性。

b) **公平公正。**在模型算法应用过程中，不存在通过模型算法侵害合作机构、服务客户（包括自然人及企业）的合法权益，包括企业正当享有的商业秘密及知识产权，以及对自然人消费者产生不公平、歧视或其财产权乃至生命健康权的可能，在不违反相关政策法规的同时，保证相应算法模型不应违反社会普适的道德价值观。

c) **风险可控。**在模型算法应用的结果上，如出现计算机语言编程错误、外部攻击等情况下，企业应评估是否可能导致大规模侵害不特定群体利益的可能性、影响规模、以及对风险应急管控的可控性程度等。

d) **透明性。**企业应适时及时向公众合理披露模型算法的主要方法及逻辑向监管机构汇报模型算法的整体逻辑及应用目的。

5 研发运维流程管理合规要点

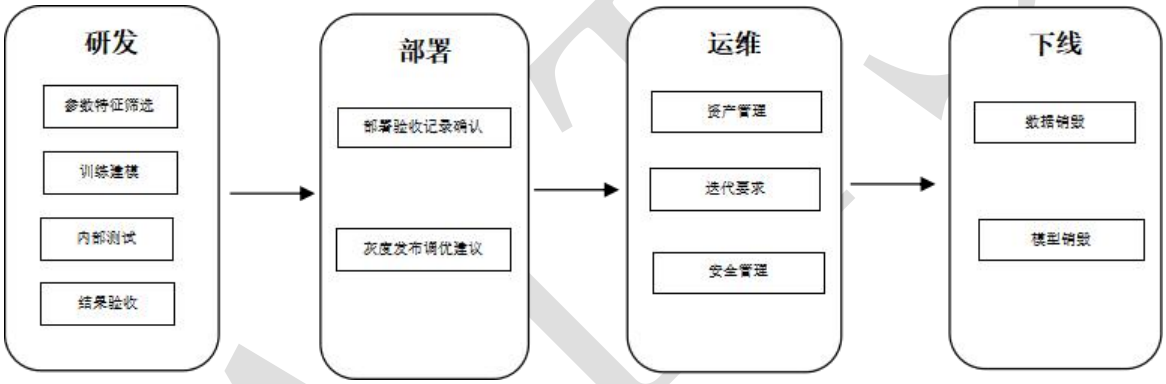

图 1　研发运维全流程合规要点示意

5.1 研发

模型算法的开发，一般会涉及到数据筛选、建模、内部测试、结果验收、外部灰度应用等环节。应评估上述环节开发操作满足以下要求：

5.1.1 参数特征筛选

在梳理准备数据过中，应遵守GB/T 35273—2020等标准要求，重点关注以下几点：

a) **数据来源合规性。公司**应建立或通过适当的数据审批管控流程，来保证参与研发环节的数据不存在违反国家相关法律法规及监管部门规章等情况，以避免从根本上导致模型算法研发存在数据源合规风险。特别是，对于生物识别信息等个人高度敏感信息的，应该符合国家相关标准的采集、使用等全流程数据标准。

b) **数据质量保障。**模型算法在建模过程中，对数据质量具有高度的依赖性，数据质量在出现精准性、完整性、一致性、唯一性、有效性、时效性等方面出现问题的情况下，将极大程度上损害建模的效果及准确性，影响企业经营效能，并进一步对客户权益造成侵害的可能。因此建模团队应对数据进行多维多次的清洗及质量校验，保证建模过程所依赖的数据质量是符合上述要求的。

c) **数据应用的必要正当性。**合规等相关团队应对相应模型算应用目的与原先采集数据过程中向用户告知并获取同意的数据用途的一致性进行评估，评估其所需使用的数据项颗粒度是适当的。可使用降级策略，包括随机、泛化、拟制等手段来管控直接使用原始明细数据，从而提

升数据使用的必要正当性与最小化。必要的情况下，也可建议模型算法团队使用去标识化乃至匿名化数据项来完成建模过程。

d) **数据项应用的公平公正性。**在研发模型算法中，合规等相关团队应通过查阅文档、数据表等方式，对研发团队所计划使用的数据项清单进行整体审核，以避免数据项中存在可能导致客户遭受歧视、不公正待遇等侵害其权益的可能性。例如将自然人用户的性别、种族、出生地、血型、星座等作为模型算法的决策基础或逻辑。

5.1.2 建模训练

模型算法在构建过程中，会分别涉及决策参数、权重占比、逻辑合理性等问题，因此在建模过程中，需要重点关注以下几点：

a) **避免偶发性参数的选择。**在将客户的一些偶发性操作行为的识别，作为参数触发条件，推进到模型算法决策的下一个阶段，进而影响客户的选择及交易等。合规等相关团队应对不同决策参数设置的合理性进行评估，以避免对客户造成不适当的引导或反感，对客户的安宁权造成影响。例如客户在点击过特定类型图片的情况下，模型算法据此偶发性的操作行为作为决策树中重要的参考因子进行频繁推荐。

b) **合理比例。**将特定参数或操作行为作为模型算法建模训练基础时，将某一参数设置过高或过低的权重或占比，以得出最终决策依据。相应的，在发现此类情况下，合规等相关团队应对该权重或占比的合理适当性进行复核，以保证模型算法的科学合理性。

c) **逻辑合理。**在传统银行评分卡逻辑下，较多依赖的是因果关系推导，通常情况下不会引发合理适当性的问题。但伴随着大数据技术、统计学的应用等，合规等中后台团队需要特别注意避免陷入到单纯统计学意义上的决策逻辑中。例如根本无法进行关联的行为操作作为决策的基础。

d) **道德风险管控。**企业应设立科技伦理委员会或包含此部分职责的委员会，从算法的可靠可控、透明可释、保护隐私等红线底线风险进行评估。此外，在建模阶段，应设立建模小组，原则上人数不应少于 3 人，并建立相应的复核流程管控机制。以尽可能规避道德风险，以及个人喜恶偏好等原因造成侵害消费者权益，乃至违反社会普适的到的价值观的可能。

e) **训练数据隐私保护。**应确认训练数据集具备相应差分隐私等隐私保护措施。对于敏感个人信息的，应充分考虑在匿名化或去标识化状态下进行数据训练的可行性。

5.1.3 内部测试

内部测试是验证模型是否科学合理及公平公正的重要环节。内部测试宜采取人工及模型算法双重评估交叉验证的方式，通过多人复评的方式，以核验模型算法的准确性。与此同时，在内部测试阶段，宜采取在相同场景下，不同的用户群集数据来对模型算法的有效性和准确性进行多次校验，从而提升内部测试的效果并进行模型优化。

在发现模型算法存在相应缺陷或风险的情况下，模型算法团队应充分记录对应缺陷引发的原因、导致的问题、是否有解决方案及最大化风险敞口的可能。上述发现的缺陷或风险，应记录留档并在缺陷无法解决的情况下，并按流程升级至决策团队进行风险决策。同时，应就个性化推荐算法支撑的产品功能特别核实是否提供用户退出选项，需核实确保用户可选择关闭该个性化推荐。

5.1.4 结果验收

模型算法的研发团队应与合规等中后台团队一同对前期合规性及风险评审环节提出的所有意见，包括使用数据项的适当性、合理性、权重比例的科学性等问题进行全面的复核，以评估是否均得到有效的落地执行，以避免在研发过程中出现执行偏差而引发的操作风险。

5.2 应用部署

a) 算法部署及验收记录确认。模型算法除了要对生产过程应用过程进行全方位的流程合规管控外，部署亦是最后的重要管控环节。不同的模型算法在不同的领域所适用的业务场景、客群范围等均有差异，一旦部署出现偏差，亦会引发潜在风险。因此企业应对模型算法部署的过

程进行记录和管理，并做好过程的记录，并在部署完成后安排部署人员以外的质检团队进行验收。

b) 灰度发布调优建议。为了管控模型算法应用的风险，在应用策略上，宜遵循稳妥审慎的的原则，逐步开放适用的范围，而不宜直接适用到所有客户群体，以便于进一步根据客户使用情况对模型算法进行调优及改善。在条件允许的情况下，还应核实算法团队是否采取 ABtest 形式，对新模型算法与历史模型算法的差异性进行比较分析，以发现潜在的风险因素。

5.3 运维

5.3.1 总体概述

伴随着外部新技术、新模式的调整变化，通常情况下模型需要不断的进行迭代更新，以满足性能效果等方面的目标。与此同时，企业合规等中后台团队，应该充分重视国内外监管环境的变化及政策倾向性。因此合规等中后台团队应及时收集整理关于模型算法相关的最新指导原则及要求并予以更新，避免模型算法因未及时采取适当的调整而引发侵害客户权益，乃至违反社会普适性道德观等方面的情况。整体上来看，主要涉及三方面需要重点关注。

5.3.2 质量保障

客户每天通过互联网产生大量新的数据，这些数据是客户行为特征及意向的重要判断基础。在保证依法合规获取客户同意的基础上，模型算法在应用过程中应尽可能采集使用最新生成的数据，以保证模型算法应用的准确性及效果，以避免数据时效较差引发对客户的不适当干扰或引导。原则上，所使用的客户数据时效不应低于近半年内。

5.3.3 迭代要求

在对模型算法迭代过程中，除了需要参照适用研发阶段所关注的所有环节外，模型算法迭代过程中，宜根据新技术、新模式所引发的客户操作行为习惯、数据生成采集类型的变化，相应对模型算法进行适当升级，以提升模型算法的科学性及准确性，以避免对客户造成不必要的打扰，或被黑产等进行攻击。

5.3.4 安全管理

受到外部新技术等因素的影响，模型算法原则上在长时间应用后，存在被外部第三方感知并进行核心参数的可能性。模型算法若长期未予维护，可能导致被黑产有意伪造参数或构建相关参数的形式进行攻击，导致企业遭受不必要风险或损失的可能性。因此即使在监管政策环境及社会普适性道德价值观均未发生变化的情况下，仍应通过外部舆情监测等形式，对外部信息进行收集整理，并采取相应的模型算法迭代策略。原则上模型迭代的周期不应低于180天时间。

5.4 下线

对存储介质及相关文档材料中的数据，包括训练数据、测试数据、实例数据、特征数据、参数、算法输出等进行定期销毁，确保已销毁的不能再被访问，同时满足以下条件可不销毁：

1) 为进一步实现实际业务功能所必须；
2) 获得数据所有者明确授权同意；
3) 不涉及个人信息和有关部门认定的重要数据；
4) 经过混淆或加密处理，无法直接还原原始数据

若对个人信息进行删除或匿名化处理，确保涉及的个人信息不可被恢复；若对部署在多台设备上或通过云边协同的方式部署的模型实施退役下线，从多台设备、云端等同时对模型及数据实施销毁。

6 内控管理合规要点

6.1 资产管理

模型算法作为企业经营开展过程中提能增效的重要技术产权，兼具财务及商业秘密属性。一旦出现模型算法因管理不当等方面原因出现被外部感知，除对企业自身造成重大影响或损失的可能性外，还可能对社会及客户造成重大影响。除基于透明管理的需要外，宜根据具体情况及需求，申请相关知识产权保护，并设置有专员对模型算法作为一项资产进行登记管理。在企业准备对模型算法进行相关资产进行处置时，技术产权的申请将会对价值评估及交易对象的明确起到很好的帮助作用。考虑到应急使用、版本回滚乃至交叉应用的需要，所有模型算法都具备财产属性，每次模型算法研发完成或进行相关迭代的情况下，都应该对相应代码录入至系统进行管理及备份。

6.2 风险监测

宜建立相应的线上系统对模型算法的运行情况进行监测，从运行的状态、资源耗用、适用目标群体数量变化、同环比变化、异常报警等方面进行监测，以备及时发现模型算法运行过程中出现的问题并能够跟进处理。变更遵守三原则：充分灰度、有效监控（变更影响业务线上的运行情况）、快速应急（具备可随时操作的应急预案（回滚、切流））。

6.3 培训

企业应重点针对所有涉及到模型算法设计、研发、测试、验收、运维管理过程中接触到模型算法相关参数、决策逻辑等核心机密信息，以及本标准所涉各项合规性要求，应对员工通过线上、线下模式，至少每年开展1次关于算法伦理、知识产权及商业秘密保护的有关培训，防止从业人员非法披露、出售等情况。

6.4 保密

企业应与涉及到模型算法设计、研发、测试、验收、运维管理等环节的主要员工，在普通劳动合同之外，签署专门的保密协议并明确相应的义务责任要求，强化员工的安全保密意识。与此同时，在模型算法对外服务或提供支持的过程中，企业应通过合同等形式，约束合作机构不得采取技术或其他手段去获悉模型算的核心内容及逻辑。

6.5 应急管理

考虑到模型算法团队应制定泄露、出现重大偏差等情况下的风险应急预案，明确风险事件处置流程和岗位职责，具体要求包括：

a) **科学制定算法风险应急管理方案**。应将算法伦理风险及应急处置在风险应急预案中进行充分说明。**根据**相关法律法规变化情况，以及事件处置情况，及时评估更新应急预案；
b) **定期组织模型算法演练**。相关人员进行风险应急预案相关培训和应急演练，至少每年一次，并记录、并存档演练记录；
c) **提前建立模型算法下线或更替机制**。保证在发生攻击或泄露等情况下，可以采取有效措施进行应对。
d) **风险事件应急处置机制**。应对媒体披露、监管通报、用户投诉的算法舆情事件及重大突发事件建立应急处理机制。发生风险事件后应立即采取有效措施防止风险扩大，并视风险程度向有关监管部门报告，并记录事件内容，分析和鉴定事件产生的原因，评估事件可能造成的影响，总结经验教训，制定防止再次发生的补救措施；
e) **损害保障机制**。宜对模型算法安全风险提供损害保障机制，考虑设立风险拨备资金，利用保险等手段为必要救济提供保障。此外，机构应充分提示智能算法的固有缺陷和使用风险，提醒客户关注相关风险的基础上自主选择是否使用相关服务，并建立算法道德风险防范机制。

6.6 可解释性保障

模型算法的可解释性，既是评估模型算法依法合规性的重要内容，也是在发生泄露、模型算法偏差、外部攻击风险等情况的重要回溯修正的基础。因此企业在对模型算法进行版本及内容管理的过程中，应保证训练数据的可回溯性、建模过程的可回溯性。

模型算法的可解释性包括版本的可回溯，模型算法所采取的方法及逻辑的可说明性等内容。考虑到模型算法大规模应用的情况下，势必影响到众多客户的权益，为了保障在包括监管等机构在内的外界对相应模型算法存在质疑，企业应保障能够对相应模型进行解释说明，以明确模型算法应用的合规性。企业应采取以下措施以保障目的的可实现：

a) **应要求审慎采用无监督模型算法**。原则上，除非特定场景中不会发生侵害客户权益等情况，否则企业侧不宜以无监督模型算法为直接决策依据，导致模型算法出现对客户权益造成侵害的可能。

b) **完善模型算法留档管理要求**。模型团队应在针对特定业务场景的第一个研发并投入应用的模型算法版本开始进行留档管理，直至模型算法被废弃使用。保留相应的模型算法参数、脚本、决策逻辑及方法、迭代记录，记录训练数据获取的获取时间、来源、数量及存储介质，建模及部署适配的软硬件环境、人员名单、脚本说明、历史各版本适用的范围及欲解决的问题，开发关键决策讨论的相关凭证及会议纪要等。

c) **制定模型算法的可解释性衡量标准**。模型算法团队应确保自身所采取的各类技术手段、技术逻辑及方法等，在监管机构或用户提出质询的情况下，针对特征及参数定义、分布、衍生、选择等，可通过模拟、可视化等形式，对模型算法从最初参数导入到规则运算、关联变化、到最终决策结果取得进行全流程链路的演示。

d) **建立并完善监管、公众建立算法信任沟通机制**。促进立法者、算法设计者、算法使用者、用户、第三方评估机构等相关方之间开展有效沟通，创造科学、合理的算法互信生态。

6.7 算法备案

对于利用生成合成类、个性化推送类、排序精选类、检索过滤类、调度决策类等算法推荐技术向用户提供信息，具有舆论属性或社会动员能力的、影响用户权益的、涉及敏感数据的及展示在重点区域的产品及服务。根据算法备案的要求，应对算法主体信息、算法属性信息、算法支撑的产品及服务进行填报。

6.8 公众算法教育

开展模型算法应用和监督的机构，宜不定期开展公众教育及培训，除了让公众了解模型算法所可能存在的潜在风险外，宜加强对模型算法全生命周期的介绍说明，以及模型算法能够给公众及社会带来的福利。既不要盲目迷信模型算法，但也要了解模型算法所能给数字经济带来的福祉，并关注并监督其中的风险，共同促进模型算法的应用及发展。

参 考 文 献

[1] JR/T 0171—2020 个人金融信息保护技术规范
[2] JR/T 0199—2020 金融科技创新安全通用规范
[3] JR/T 0200—2020 金融科技创新风险监控规范
[4] JR/T 0221—2021 人工智能算法金融应用评价规范

ZAITS