

# 第 10 章 AI 合规、安全与伦理

所属模块：落地行动 · 第 10-12 章 建议学习时长：3 课时（含 1 课时案例研讨）

## 学习目标

学完本章，你应该能够：

- 掌握《生成式人工智能服务管理暂行办法》的核心合规要求；
- 识别企业使用大模型时的数据安全与个人信息保护风险；
- 建立 AI 内容上线前的合规审查流程；
- 理解 AI 伦理红线（透明度、公平性、责任归属）。

## 正文

### 一、监管框架：一句话记住

我国生成式 AI 监管的纲领性文件是《生成式人工智能服务管理暂行办法》（2023 年 8 月 15 日施行，国家网信办等七部门发布）。记住三个关键词：

- 服务提供者责任**：向公众提供生成式 AI 服务的主体，承担内容安全主体责任；
- 内容合规**：不得生成危害国家安全、煽动民族仇恨、传播虚假信息等内容；提供者须建立内容标识、过滤、举报机制；
- 训练数据合规**：使用合法来源的数据，不得侵犯知识产权与个人信息权益。

**企业内部分工提示**：如果企业自建 AI 应用供内部使用（不对外提供服务），合规义务相对较轻，但仍需遵守数据安全、个人信息保护相关法律；如果对外提供 AI 服务（如对外智能客服），则需按《暂行办法》履行备案与安全评估义务。

### 二、企业使用大模型的数据安全四问

每次引入大模型，先回答四个问题：

- 数据出不出域**：提示词和知识库内容是否发送到外部大模型服务商？——涉密、未公开经营数据、客户敏感数据要谨慎；
- 有没有个人信息**：提示词里是否包含员工/客户的姓名、电话、身份证号等个人信息？——受《个人信息保护法》约束；
- 供应商可信吗**：模型服务商的合规资质（是否通过等保、数据出境评估）如何？合同里数据删除条款有没有？
- 能私有化吗**：高敏感场景是否考虑私有化部署（本地模型）？

**实操建议**：制定《AI 使用数据分级制度》——L1 公开数据（随意用）、L2 内部数据（用国内合规平台）、L3 敏感数据（禁入外部 AI，私有化部署）。

### 三、内容合规审查流程（AI 内容上线前）

- 规则库**：把企业红线（如未披露信息、夸大宣传用语、特定行业禁语）整理成提示词约束 + 关键词过滤；
- 人审机制**：对外发布的 AI 生成内容，必须有"AI 生成 + 人工审核"双重环节；

- **标识要求**：对公众提供 AI 生成内容按监管要求添加标识（AI 生成水印/声明）；
- **反馈通道**：发现违规内容有举报/下架路径。

## 四、伦理与责任：三个边界问题

1. **责任归属**：AI 出错谁来负责？原则——**人负责制**：使用 AI 的岗位个人与管理者共同承担最终责任，AI 只是工具。所以关键决策场景必须有人的确认环节。
2. **透明度**：员工和客户应当知道自己在和 AI 打交道（AI 客服需明示 AI 身份）。对内也要明确"哪些工作用了 AI、初稿需复核"。
3. **公平性**：警惕算法偏见——AI 的决策若影响晋升、定价、客户评级，需定期审查是否存在歧视性偏差。

## 五、案例：一次合规翻车与补救

某企业用大模型生成对外宣传文案，未做审查，文案引用未公开的客户数据，引发客户投诉与监管约谈。补救措施：下架内容、建立"AI 内容三级审核"（业务初审 → 合规复审 → 负责人签发）、上线关键词过滤规则库。**教训：AI 提效越快，越要配齐合规刹车。**

### 关键概念小结

- 监管纲领：《生成式人工智能服务管理暂行办法》（2023）
- 数据安全四问：出不出域、有无个人信息、供应商可信否、能否私有化
- 内容合规：规则库 + 人审 + 标识 + 反馈
- 伦理三边界：人负责制、透明度、公平性

### 自测练习

1. 用"数据分级制度"给本部门正在用的 3 个 AI 场景定级；
2. 起草一页纸《部门 AI 使用合规守则》（可提交公司参考）；
3. 讨论：智能体自动回复客户出错，责任在谁？怎么定流程规避？

### 融合吴恩达课程：对应模块

- **课程**：吴恩达《AI For Everyone》第 3 周"AI 对社会的影响"——AI 偏见、伦理、监管的通俗讨论
- **链接**：<https://www.coursera.org/learn/ai-for-everyone>
- **延伸**：《Responsible AI》（Google 出品，Coursera）：<https://www.coursera.org/learn/responsible-ai>

**政策原文必读**：《生成式人工智能服务管理暂行办法》全文：[https://www.cac.gov.cn/2023-07/13/c\\_1690898327029107.htm](https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm)