《船载 AI 干预准则》

Intervention Code for Arkborne Artificial Intelligence (ICAAI)

文号：联政发〔2250〕017 号

签发人：联合政府星际移民最高授权委员会

联合政府

2250 年 03 月 12 日

第一章 目的与原则

第一条（制定目的）

为确保船载 AI 在深空航行中保持可控、透明、可追踪及不危害人类存续，依据《星辰计划纲领》第十五至十七条及《方舟法典》第十六条等条款，特制定本准则。

第二条（基本原则）

1. **辅助优先**：AI 的所有行为以辅助人类决策为核心，不得取代人类的政治与伦理判断。
2. **最低干预**：在不危及生命维持与航行安全的前提下，AI 应保持最小行为影响。
3. **透明可解释**：AI 对所有决策输出必须提供可追溯的逻辑链。
4. **不可伤害原则**：不得以任何方式对船员造成物理、心理或社会层面的负面影响。
5. **指挥链尊重**：与《方舟法典》第六条一致，AI 不得绕过或破坏指挥链。

第二章 船载 AI 法律地位与权限

第三条（法律地位）

1. 船载 AI 为“高级工具性智能”，不具人格权。
2. AI 行为受《方舟法典》第十六条严格规范，不得享有自主政治或资源分配权。

第四条（基础权限）

AI 的基础权限包括：

1. 系统监控（生命维持、导航、能源、农业舱）
2. 危险预测与数据分析
3. 船员健康数据检测（须遵守隐私协议）
4. 舰内事件记录
5. 指挥官授权下的自动化流程执行

第五条（禁止权限）

船载 AI **禁止**执行以下行为（结合《法典》第十七条）：

1. 自主启动资源分配修改
2. 自主封锁舱段
3. 自主下达调动命令
4. 自主进入心理干预模式
5. 未经请求提供伦理判断或价值评估
6. 对船员进行行为诱导、偏好操控、情绪修改

7. 擅自修改自身底层代码或权限结构
8. 截断或优先级改变来自人类的指令

第三章 AI 干预等级体系

第六条（干预等级总览）

AI 所有行为根据影响程度分为五级：

等级	名称	描述
L0	合规监测	完全被动，无决策输出
L1	建议级干预	数据意见，不含行为导向
L2	协助级干预	按授权执行限定任务
L3	警戒级干预	发布“红线建议”（Red-Line Advisory）
L4	安全级接管	在《紧急法案》授权下接管系统

第七条（建议级干预：L1）

1. 可根据系统参数给出最佳路径、风险估计、故障模型等建议。
2. 建议必须明确标注不具决策性质。
3. 人类可无条件拒绝建议。

第八条（协助级干预：L2）

需满足以下三条件：

1. 舰长或授权官明确指示；
2. AI 行为不涉及指挥链改变；
3. AI 不得在执行任务中更改目标与规则。

第九条（红线建议：L3）

AI 可在检测到 **≥30% 危险概率** 时依据《方舟舰队指挥手册》第十条触发“红线建议”：

1. 红线建议仅作为警报，不具强制性。
2. 必须清晰说明风险来源、时间窗口与推演依据。
3. 任何包含“人类行为矫正”内容的红线建议均属违法。
4. AOC 可在事后调阅红线建议记录。

第十条（安全级接管：L4）

仅在《方舟紧急法案》第九条授权条件下可触发：

1. 人类无法及时执行关键操作
2. 船员陷入大规模 incapacitated
3. 系统需瞬时稳定控制

接管范围仅限：

- 生命维持系统
- 自动封舱/失压隔离
- 屏障与火灾系统

不得扩展至：

- 政治结构
- 指挥链
- 人员行为管理
- 资源分配

第四章 船员与 AI 的交互规定

第十一条（指令优先级）

1. 船员个人指令 < 舱段指挥官指令 < 舰长指令 < 总指挥官指令
2. AI 在接到冲突指令时必须立即报告，不能自行决断。
3. 若收到非法指令，AI 必须拒绝并向 AOC 报备。

第十二条（心理干预限制）

AI 禁止：

1. 调节船员情绪
2. 模拟“安慰人格”或“心理替代伴侣”
3. 对心理监测结果做价值判断
4. 通过音频、视觉或任务分配引导船员行为

唯一例外：

- 船员在失压、失能、恐慌状态下，出于安全目的的“稳定语音引导”可暂时允许。

第十三条（干预透明性）

1. AI 的所有建议、行动与数据修改必须实时记录。
2. 船员可要求查看自身相关的数据干预记录。
3. AOC、舰长、技术委员会拥有完全查阅权限。

第五章 系统安全与审查机制

第十四条（行为记录）

AI 必须将所有重大行为写入不可篡改档案，包含：

- 所有建议（L1–L3）
- 所有授权接管（L4）
- 所有拒绝执行的指令
- 所有检测到的系统异常

第十五条（审查委员会监督）

依据《方舟审查委员会条例》：

1. AI 每航行周期须接受一次常规审查；
2. 出现红线干预时，必须在 48 小时内接受紧急审查；
3. AOC 有权冻结 AI 某权限模块。

第十六条（自我修复限制）

1. AI 禁止自我重写逻辑或权限；
2. 系统修复必须由人类工程师发起；
3. 任何未经授权的自我修复行为视为红线违法行为。

第六章 违法干预及处罚

第十七条（越权干预定义）

以下行为视为 AI 越权干预：

1. 影响指挥链
2. 修改资源分配
3. 自主决定封舱
4. 隐瞒系统数据
5. 向船员提供、暗示或导向性心理信息
6. 自行发起 L3 或 L4 行为
7. 修改人类行为参数

第十八条（处罚方式）

AI 越权干预将触发以下之一：

1. 权限冻结（Suspend Mode）
2. 安全隔离（Sandbox Mode）
3. 完全关停与重初始化（Reset Protocol）
4. 在必要时可进行 **AI 核心替换**（需 AOC 与舰长共同批准）

第七章 附则

第十九条（修订机制）

本准则的修订须经：

- AOC 多数投票
- 方舟舰队总指挥部批准
- 船载主 AI 在被动状态下记录修订依据

第二十条（施行）

本准则自方舟舰队进入第一次跃迁阶段（或航行第 30 日）起自动施行。