ICS 35.240
CCS L 70

团　　体　　标　　准

T/CCF 0001—2025

强化学习系统　第1部分：通用要求

Reinforcement learning system part 1: General requirements

2025－06－11 发布　　2025－06－11 实施

中 国 计 算 机 学 会　　发布

目　　次

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。

本文件由中国计算机学会标准工作委员会提出并归口。

本文件起草单位：腾讯科技（成都）有限公司、北京大学、中国科学技术大学、新华出版社、成都实娱商业管理有限公司、上海交通大学、清华大学、北京中关村人工智能研究院、西安交通大学、中科院自动化所、电子科技大学、浙江大学、西南交通大学、南京大学、中国电力科学研究院有限公司、四川大学、武汉大学、复旦大学、中山大学、杭州宇树科技有限公司、数字图书馆教育部工程研究中心、哈尔滨工业大学（深圳）、北京航空航天大学、华中科技大学、北京邮电大学、OPPO广东移动通信有限公司、燧原智能科技（成都）有限公司、摩尔线程智能科技(北京)股份有限公司、重庆邮电大学、西南民族大学、四川具身人形机器人有限公司。

本文件主要起草人：刘林、邓民文、鲁云龙、杨耀东、李文新、冯加恒、张伟楠、周文罡、张海峰、杨巍、叶振斌、赵鉴、许华哲、万里鹏、张寅、蒙朦、覃洪杨、汪永毅、李凌峰、匡乐成、王永霞、周丹丹、谢宁、邢焕来、季向阳、汪文俊、曹相成、陶吕方、黄蓝枭、林夏、李厚强、高阳、兰旭光、吕建成、王新迎、余超、何召锋、徐建、吴秉昊、赵卫东、温颖、宋麟、梁超、章欣、杨丰、李梦露、汤臣薇、石荣晔、吴文峻、刘渝、周可、王进、蒋溢、蔡英、涂钥轩。

引　言

作为一种重要的人工智能方法，强化学习在具身智能、大语言模型、自然科学研究、游戏AI等众多领域得到了广泛应用。

强化学习是智能体通过和环境交互收集数据来优化决策的机器学习范式。智能体执行动作并从环境中接收观测、奖励等；环境则接收智能体的动作，以一定机制更新当前状态并向智能体反馈观测、奖励等。在不同的强化学习算法中，智能体利用环境数据学习的方式有诸多差异，因此强化学习算法的运行依赖框架的实现。综上，智能体、环境、框架为构成强化学习系统的三个基本要素。

尽管强化学习对既有数据依赖度低，在求解问题上普适性强，于当下有丰硕的应用成果，于未来有广阔的应用前景。然而在缺乏相关标准的情况下，强化学习算法开发和应用落地存在诸多挑战。一方面，场景、算法强耦合，工程可复用度低，重复开发浪费人力；另一方面，训练、部署难迁移，应用接入成本高，服务质量难以保障。

本系列标准将为强化学习系统设计和应用生态构建提供参考和依据，推动环境开发者、算法开发者、应用开发者、平台运营者在统一的框架下展开协作。标准规范的设计保证系统各功能模块即插即用，从而减少重复开发、降低应用成本、提升服务质量，进而实现技术生态协同，提高资源利用效率，使强化学习技术惠及更多领域发展。本系列标准由4个部分构成。

——第1部分：通用要求。目的在于确立强化学习系统的参考架构，规定通用技术要求。

——第2部分：强化学习环境技术要求。目的在于确立强化学习环境的参考架，规定其接口和数据格式要求。

——第3部分：强化学习智能体技术要求。目的在于确立强化学习智能体的参考架构，规定其接口和数据格式要求。

——第4部分：强化学习框架技术要求。目的在于确立强化学习框架的参考架构，规定其接口和数据格式要求。

强化学习系统　第1部分：通用要求

1　范围

本文件规范了强化学习系统的框架以及总体要求，包括强化学习应用服务、学习环境、智能体、框架、运行时、工具以及运维等组件。

本文件适用于指导强化学习系统全流程，包括系统的设计与构建、算法开发与实现、训练过程的配置与优化、性能评估与验证、部署与运营。

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 43782-2024　人工智能 机器学习系统技术要求

GB/T 41867-2022　信息技术 人工智能 术语

ISO/IEC DIS 22989　Information technology — Artificial intelligence — Artificial intelligence concepts and terminology

3　术语和定义

下列术语和定义适用于本文件。

3.1

机器学习模型 machine learning model

一种基于输入数据或信息生成推理或预测的计算结构。

[来源：GB/T 41867-2022, 3.2.11]

3.2

机器学习 machine learning

通过计算技术优化模型参数的过程，使模型的行为反映数据或经验。

[来源：GB/T 41867-2022, 3.2.10]

3.3

环境 environment

基于自身状态以及动作输入，输出观测和奖励信息的问题模型，包括状态观测机制、状态转移机制、奖励机制。其中状态观测机制决定状态和观测的形式与二者间关系，状态转移机制决定输入动作导致环境状态变化的方式，奖励机制决定环境对动作的数值评价。

3.4

强化学习 reinforcement learning

一种通过与环境交互，学习最佳行动序列，使回报最大化的机器学习方法。

[来源：GB/T 41867-2022，3.2.25]

3.5

样本 sample

同环境交互的对象与环境在交互过程中产生的结构化的数据，可经一定处理，一般包括状态、观测、奖励、动作等信息。

3.6

智能体 agent

能够感知环境（输出的观测）并产生动作的实体。

[来源：ISO/IEC DIS 22989，3.1.1]

3.7

奖励重塑 reward shaping

在强化学习中，为了加速算法收敛、引导智能体学习满足特定目标，通过先验知识或环境反馈信息，以数学形式（如附加势能函数、奖励偏移等）对环境反馈的原始奖励信号进行修正的过程。

4 符号和缩略语

下列符号和缩略语适用于本文件。

AI：人工智能（Artificial Intelligence）

IO：输入/输出（Input/Output）

RL：强化学习（Reinforcement Learning）

DRL：深度强化学习（Deep Reinforcement Learning）

CPU：中央处理器（Central Processing Unit）

GPU：图形处理器（Graphics Processing Unit）

TPU：张量处理器（Tensor Processing Unit）

NPU：神经网络处理器（Neural-network Processing Unit）

DQN：深度Q-网络（Deep Q-Network）

PPO：近端策略优化（Proximal Policy Optimization)

A2C：优势执行者-评论者（Advantage Actor-Critic）

ONNX：开放神经网络交换（Open Neural Network Exchange）

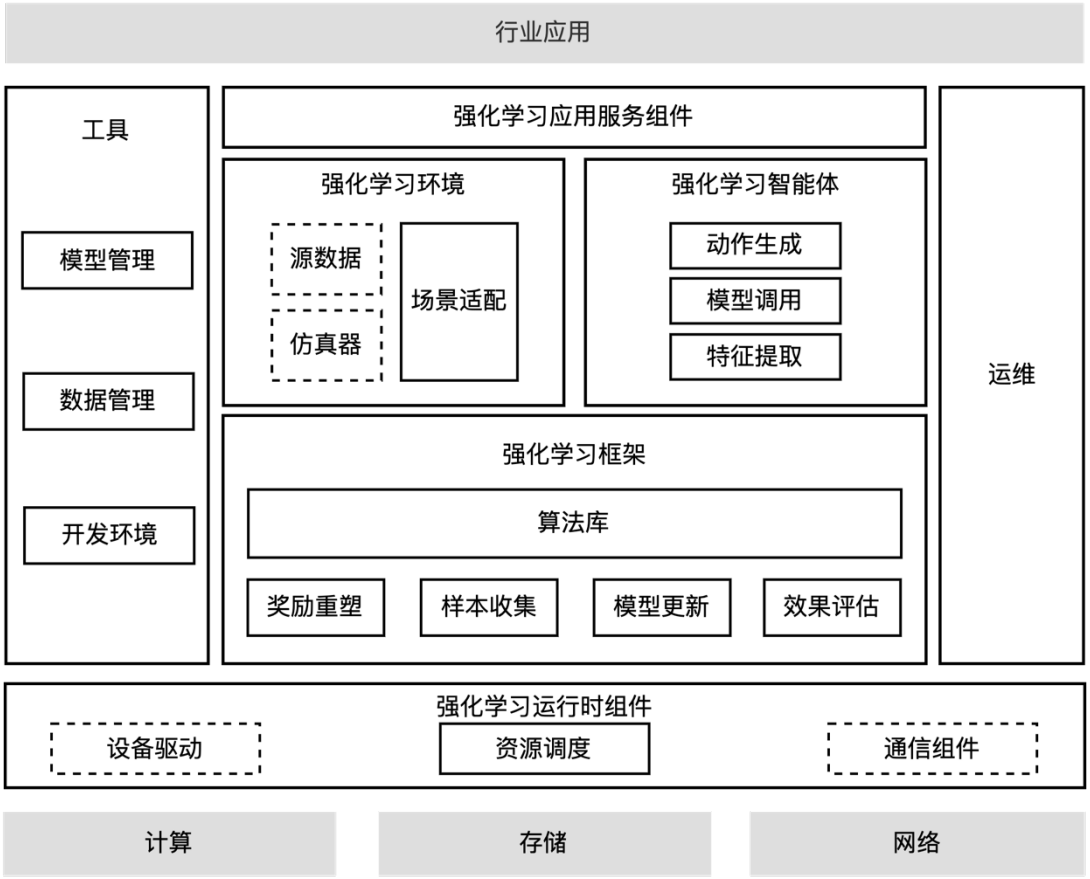
MCTS：蒙特卡洛树搜索（Monte-Carlo Tree Search）

A*：A星搜索算法（A-star Search Algorithm)

5 总体架构

5.1 概述

强化学习系统包括强化学习运行时组件、强化学习框架、强化学习环境、强化学习智能体、强化学习应用服务组件、工具和运维，为行业应用提供强化学习算法开发、模型训练、智能体部署、应用管理等服务。强化学习系统的总体架构见图1。



注：图中实线部分对应本文件相关规定；虚线部分与应用场景相关仅在本文件中规范相关接口描述；深色部分不属于本文件规范。

图1：强化学习系统总体架构

5.2 强化学习运行时组件

强化学习运行时组件是为强化学习工作流的运行提供存储、通信、算力等基础支持的功能模块集合，用于管理整合设备资源，包括设备驱动、资源调度与通信组件。

其中，资源调度负责为训练、推理、评估等强化学习工作流在各类设备上运行提供存储与计算资源分配、管理、监控、回收等功能。

5.3 强化学习框架

强化学习框架是运行在强化学习运行时组件之上的，实现强化学习工作流所需的功能模块集合。包含奖励重塑、样本收集、模型更新、效果评估模块以及组装这些模块形成的算法库。其中，

a) 奖励重塑：以一定方式重新计算样本中奖励的功能模块。
b) 样本收集：收集处理强化学习环境与智能体交互所产生样本的功能模块。
c) 模型更新：提供模型参数初始化、载入、备份以及利用样本更新模型参数功能的模块。
d) 效果评估：用于评估强化学习智能体各项指标的功能模块。其中评估指标可包括平均累计折扣奖励、决策序列平均长度、单次决策平均所需时间等。
e) 算法库：组装奖励重塑、样本收集、模型更新、效果评估中的部分模块以实现具体的强化学习工作流的功能模块。其中，强化学习训练工作流应当包括样本收集、模型更新模块，可选包括奖励重塑、效果评估模块；强化学习推理工作流应当包括样本收集模块，可选包括奖励重塑模块；强化学习评估工作流应当包括效果评估模块，可选包括样本收集模块。

注：强化学习训练、推理、评估工作流同框架各组件间的关系见附录A.1。

5.4 强化学习环境

强化学习环境是基于输入动作，输出观测、奖励等反馈的功能模块，用于表达强化学习算法所求解的问题场景，包括源数据、仿真器与场景适配。

场景适配是用于封装不同形式的问题场景，使之具有统一交互协议与接口的功能模块。场景适配可封装的问题场景形式包括源数据和仿真器。其中，源数据是由历史交互轨迹组成的集合，例如：市场中的价格数据与交易记录、物理仿真中的传感器与控制器信号序列、游戏服务器中的状态信息与交互历史等；仿真器是一种通过模仿目标系统的外在表现和行为来实现其功能的工具或系统，例如：模拟交易软件、物理仿真工具、游戏服务器等。

5.5 强化学习智能体

强化学习智能体是基于输入观测信息来输出合法动作的功能模块，拥有参数可更新的模型。强化学习智能体包括特征提取、模型调用、动作生成三个模块。其中，

a) 特征提取：将强化学习环境输出的观测信息处理为符合模型计算过程输入特征形式的功能模块。
b) 模型调用：调用模型的功能模块。将特征处理结果输入模型的计算过程，以模型计算过程的结果作为输出。
c) 动作生成：将模型调用结果处理为符合强化学习环境输入动作要求的功能模块。

5.6 强化学习应用服务组件

强化学习应用服务组件是一套软件工具，为各行业应用提供访问强化学习系统服务的接口。

5.7 工具

工具包括开发环境、数据管理、模型管理三个模块。其中，

a) 开发环境：支持环境、智能体、算法等强化学习相关功能组件开发的软件工具。
b) 数据管理：管理强化学习系统运行产生的样本、日志等数据的软件工具。

c） 模型管理：为行业应用提供发布更新、分类检索、版本控制、导入导出等模型管理相关功能的软件工具。

5.8 运维

运维管理提供系统所需的基本运维（例如安装部署、扩展、监控、报警、健康检查、问题及故障定位、 升级和补丁、备份恢复和操作审计等）及管理功能（例如计算资源管理、权限管理、用户管理、日志管理、配置管理和安全管理等） （见GB/T 43782-2024 5.6节）

6 功能要求

6.1 强化学习运行时组件

强化学习运行时组件的资源调度功能包括以下要求：

a） 应支持强化学习工作流在多设备上运行；

b） 应具备强化学习工作流运行过程中的存储与计算资源分配、管理、监控、回收功能；

c） 应具备日志机制，记录强化学习工作流运行过程中的重要事件及各项监控指标；

d） 应具备容灾机制，例如强化学习工作流运行过程中异常进程的重启等；

e） 应具备可扩展性，支持不同规模计算资源条件下的强化学习工作流的调度运行；

f） 宜支持异构算力管理，例如同时调度CPU与TPU设备完成强化学习工作流的运行。

6.2 强化学习框架

强化学习框架的功能要求包括以下内容。

a） 奖励重塑：

1） 应支持基于单步决策样本重塑奖励；

2） 应支持基于完整决策序列重塑奖励。

b） 样本收集：

1） 应支持多智能体环境中的样本收集；

2） 应支持并行收集样本；

3） 应支持以单步决策为单位收集样本（常见于off-policy算法，如DQN），并支持简单处理：如状态聚合、随机打乱、重采样、筛选过滤、分类整理等；

4） 应支持以完整决策序列为单位收集样本（常见于on-policy算法，如PPO），并支持简单处理：如序列切分、计算并添加累计回报等；

5） 宜支持基于自定义规则收集样本，例如指定概率分布随机采集样本、利用基于规则的智能体决策采集样本等；

6） 可支持采样与搜索算法，例如MCTS，A*等，用于执行规划。

c） 模型更新：

1） 应提供常用的模型参数初始化功能，例如高斯分布随机初始化等；

2） 应提供模型参数备份、载入功能，以支持接续训练等；

3） 应支持为不同形式模型指定更新方式的能力，例如对神经网络模型使用梯度下降，而对值函数表中的条目使用步进更新等；

4) 应提供常用的强化学习优化目标的计算方式，例如DQN的损失函数等；
5) 应提供常用的强化学习优化目标正则化函数的计算方式，例如最大熵正则化等；
6) 应提供常用的优化算法用于更新模型参数，例如使用不同学习率与优化器配置的反向传播等；
7) 宜支持延迟更新、滑动平均等模型参数更新方式；
8) 可支持采样与搜索算法，例如MCTS，A*等，作为计算优化目标的补充方式。

d) 效果评估：
1) 应支持直接利用已有的交互数据计算智能体评估指标；
2) 应支持在与多个强化学习环境并行交互中计算智能体评估指标；
3) 宜提供评估多智能体强化学习效果的基础支持，例如支持固定对手对局计分；
4) 可提供评估多智能体强化学习效果的进阶支持，例如使用瑞士轮、循环赛、淘汰赛等方式为多智能体环境下的单个或多个智能体生成评估指标。

e) 算法库：
1) 应提供常用的强化学习算法（如DQN、PPO、A2C等）的训练工作流；
2) 应提供接口支持用户自行开发奖励重塑、样本收集、模型更新、效果评估模块；
3) 应支持用户自定义的奖励重塑、样本收集、模型更新、效果评估模块接入现有工作流；
4) 应提供接口支持用户自行组装开发通用或问题特化的强化学习工作流；
5) 宜提供多智能体强化学习工作流的组装实现；
6) 宜支持自定义强化学习训练工作流，以实现课程强化学习、离线强化学习、自适应强化学习。

注：对多智能体强化学习的支持见附录A.2，对课程强化学习的支持见A.3，对离线强化学习的支持见A.4，对自适应强化学习的支持见A.5。

6.3 强化学习环境

强化学习环境的场景适配功能包括以下要求：

a) 应提供场景侧接入协议模板，包括场景输出数据类型、取值范围、含义描述等；
b) 应提供统一的框架侧调用接口，包括重置状态、状态转移等；
c) 应提供统一的框架侧交互协议，规范框架侧调用接口输入输出数据形式；
d) 应支持指定配置重置状态；
e) 应支持输出观测、奖励等信息；
f) 应支持处理输入的非法动作；
g) 应支持多智能体场景适配；
h) 宜支持大语言模型中的强化学习场景适配；
i) 宜支持基于场景侧接入协议模板的场景测试用例自动化生成；
j) 可支持调用基础科学领域常用软件工具；
k) 可提供具身智能相关支持；
l) 可提供场景可视化支持，使用图像或视频数据展示智能体的决策过程、环境的状态变化等。

注：大语言模型中的强化学习见附录A.6，支持数学、物理、化学、生物等基础科学中的强化学习见A.7，支持具身智能的强化学习见附录A.8。

6.4 强化学习智能体

强化学习智能体的功能要求包括以下内容。

a) 特征提取：
   1) 应支持常用特征提取方式，如提取部分特征字段、默认值填充、浮点数舍入、数值归一化、数值离散化、文本映射向量等；
   2) 应支持自定义特征提取方式；
b) 模型调用：
   1) 应支持将特征转换为模型可接受的输入形式，如Numpy array转为Torch tensor格式等；
   2) 应支持强化学习算法常用模型输出形式，例如状态价值、状态动作价值、动作概率分布等；
   3) 应支持推理模式与训练模式切换，其中推理模式仅调用模型，不可修改模型参数；训练模式除调用模型外，允许修改模型参数。
   4) 可支持模型推理加速优化，如模型量化，内存复用，算子重新编排等。
c) 动作生成：
   1) 应支持处理不同模型输出形式，例如动作概率分布、状态动作价值等；
   2) 应支持不同采样方式生成动作，如动作概率分布采样、选择最高价值动作、ε-贪心选择动作、基于掩码选择动作等。

6.5 强化学习应用服务组件

强化学习应用服务组件的功能要求包括：

a) 应支持用户管理强化学习训练、推理、评估工作流，包括提交、运行、监控、终止等；
b) 应支持强化学习智能体推理服务部署，包括独立部署、在线部署、多实例部署；
c) 宜支持推理时模型增强，如思维链多步推理，接入搜索算法提高模型性能，模型集成等；
d) 宜支持强化学习智能体输入验证与输出清理等安全性相关功能；
e) 可提供测试工具，验证强化学习智能体决策的安全性、有效性；
f) 可支持部署后在线学习，使智能体拥有持续适应环境变化的能力。
g) 强化学习应用服务组件应支持一般机器学习服务组件能力，见GB/T 43782-2024 6.3节 列项d)-m)。

6.6 工具

a) 开发环境：
   1) 应支持强化学习环境与智能体开发；
   2) 应支持强化学习算法及实现算法所需的相关功能组件开发；
   3) 应支持基础调试功能，例如过程信息输出、流程可运行性验证等；
   4) 宜支持单步调试、变量追踪等进阶调试功能；
   5) 宜支持版本管理；
   6) 宜支持在线开发与调试；
   7) 宜支持代码补全、代码提示等辅助开发功能；
   8) 宜支持多人协作开发。
b) 数据管理：

1） 应支持样本的分类整理、归档存储等管理功能；
2） 应支持访问强化学习系统运行中产生的日志信息；
3） 宜支持样本与日志信息的统计分析功能，以统计数据、可视化图表等形式展示；
4） 可支持云端数据备份；

c） 模型管理：
1） 应支持外部模型导入强化学习系统用于运行强化学习工作流；
2） 应支持强化学习系统运行所获模型的导出保存；
3） 应支持模型分类检索与版本控制；
4） 应支持模型可用性验证；
5） 宜支持模型性能评测；
6） 宜支持模型参数分析与统计；
7） 可支持模型格式转换与压缩功能，例如PyTorch模型转为ONNX模型、压缩为zip格式文件等；

6.7 运维

运维管理的功能要求包括（GB/T 43782-2024 6.5节）：

a）应提供多用户管理功能，具备多用户的权限管理能力，具备身份鉴别系统（例如Kerberos）；

b）应提供多租户管理功能，具备租户间的应用隔离、数据隔离、资源隔离和运行隔离等功能；

c）应提供安装与升级功能，具备分发安装包、数据或模型参数文件，进行安装、升级、扩展和回滚；

d）应提供备份与恢复功能，具备安装包、数据或模型参数文件的备份能力，以供故障后的系统 恢复；

e）应具备运行环境的监控能力，包括底层资源的统一监控，如CPU 利用率和系统负载等；

f）应提供日志管理功能，可根据日志进行故障定位及排查；

g）应提供针对监控指标及日志的报警功能；

h）宜提供主要监控指标的可视化展示功能。

7 安全性要求

强化学习系统的安全性应符合GB/T 43782-2024 第10章要求。

附 录 A
（资料性）
强化学习框架的泛用能力

A.1 强化学习训练、推理、评估工作流同框架各组件间的关系

A.1.1 强化学习训练工作流

强化学习训练工作流一般包含下列步骤：

1.初始化强化学习环境和强化学习智能体；

2.样本生成模块使用环境与智能体产生样本，过程中可调用奖励重塑模块；

3.获取训练样本并传入模型更新模块，用于更新智能体拥有的模型；

4.重复过程1-3，不断更新智能体拥有的模型，直至达到条件结束工作流。

其中，步骤2产生训练样本与步骤3获取训练样本可使用不同变体的生产者-消费者模型异步进行，步骤2-4中可调用效果评估模块，所得评估结果可用于步骤1-3执行过程。

A.1.2 强化学习推理工作流

强化学习推理工作流一般包含下列步骤：

1.初始化强化学习环境和强化学习智能体；

2.样本生成模块使用环境与智能体产生样本，过程中可调用奖励重塑模块；

3.重复过程1-2，直至达到条件结束工作流，过程中可收集产生的样本。

强化学习推理工作流为智能体决策过程的封装，过程中产生的样本如有后续应用，则可收集存储，否则可以忽略。产生的样本如在后续用于训练，则可在步骤2中调用奖励重塑模块。

A.1.3 强化学习评估工作流

强化学习评估工作流支持两种工作模式：基于已有样本的评估、基于智能体的评估。

基于已有样本的评估一般包含下列步骤：

1.获取已有样本；

2.将已有样本传入效果评估模块，生成评估数据。

基于智能体的评估一般包含下列步骤：

1.初始化强化学习环境和强化学习智能体；

2.样本生成模块使用环境与智能体产生样本；

3.将产生的样本传入效果评估模块，生成评估数据。

A.2 多智能体强化学习（Multi-Agent RL，MARL）

A.2.1 概述

多智能体强化学习是强化学习的一个子领域，研究在包含多个智能体的环境中学习决策。环境中各个智能体目标可能有所不同，观测可能存在差异，动作可能相互影响。此外，各智能体之间可能存在合作、竞争等关系，并且这种关系可能随着环境状态的转移而发生变化。

常见的多智能体强化学习算法包括IQL（Independent Q-Learning）、QMIX（Q-value Mixing Network）、COMA（Counterfactual Multi-Agent Policy Gradients）、MADDPG（Multi-Agent Deterministic Policy Gradient）、MAPPO（Multi-Agent Proximal Policy Optimization）等。

多智能体与环境的适配由本系列标准第2部分规范，多智能体开发和部署的要求由本系列标准第3部分规范，适配多智能体的强化学习框架要求由本系列标准第4部分规范。

A.2.2 框架支持方式

场景适配可将多智能体问题场景封装为统一接口协议的强化学习环境。

同时决策支持：环境对外提供的调用接口可与单智能体情况相同，即每次仅接收单个智能体的动作输入，输出观测、奖励等反馈，当多个智能体同时决策，可由场景适配转为逐个智能体决策；也可提供多智能体场景特化的调用接口，允许每次接收多个智能体的动作输入，同时输出多个观测、奖励等反馈。

智能体间通信支持：信息可作为信源智能体输出动作的一部分提供给强化学习环境，而后由环境转发给目标智能体，作为信宿智能体观测的一部分。

算法支持：样本收集、奖励重塑、模型更新、效果评估模块中提供多智能体相关算法的实现，用于算法库组装多智能体强化学习工作流。

A.3 课程强化学习（Curriculum RL）

A.3.1 概述

课程强化学习是一种强化学习范式，即将强化学习求解的原问题分解为子阶段或子问题，在逐步学习求解子阶段或子问题的过程中达成求解原问题的目标。将原问题分解为子阶段或子问题的方式——课程设计，是课程强化学习的核心。

常见的课程设计包括阶段难度递增、阶段奖励调节、子任务拆分等。

A.3.2 框架支持方式

课程学习可由算法库提供支持。算法库允许在自定义的强化学习训练工作流中分阶段调整选用的样本收集、奖励重塑模块。其中样本收集模块允许更改环境配置以实现阶段难度递增、子任务拆分；奖励重塑模块可用于实现阶段奖励调节。

A.4 离线强化学习（Offline RL）

A.4.1 概述

离线强化学习是在离策略强化学习（Off-Policy RL）算法的基础上，利用已有的交互记录——即源数据——的一种强化学习范式。

常见的离线强化学习算法包括：BCQ（Batch-Constrained Q-Learning）、BEAR（Bootstrapping Error Accumulation Reduction）、BRAC（Behavior Regularized Actor-Critic）、TD3+BC（Twin Delayed Deterministic Policy Gradient + Behavior Cloning）等。

A.4.2 框架支持方式

离线强化学习可由场景适配提供支持。离线强化学习求解的问题场景表达形式为源数据，场景适配可根据传入的动作——即数据加载指令，向框架提供按指令加载的源数据。

A.5 自适应强化学习（Adaptive RL）

A.5.1 概述

自适应强化学习是一类强化学习方法，旨在使智能体能够在面对环境变化、任务变化和其他不确定性时，仍能够自适应地调整其策略，有效学习和决策，以优化其长期回报。

常见的自适应强化学习方法包括元强化学习（Meta-RL）、基于模型的强化学习（Model-Based RL）、多任务强化学习（Multi-Task RL）、在线强化学习（Online-Learning）等。

A.5.2 框架支持方式

元强化学习、多任务强化学习可由算法库提供支持：算法库允许在自定义的强化学习训练工作流中切换智能体交互的环境对象，以实现在多个任务上训练；也可由场景适配提供支持：将多个任务环境按照一定方式组合为单一环境，按一定规则选择内部实际使用的环境，供算法库调用。

在线强化学习指智能体在与环境的持续交互过程中模型实时更新，以适应环境不断变化的强化学习范式。在线强化学习要求智能体具有更强的探索能力，因此动作生成模块应当支持探索。此外，在线强化学习可能发生在部署阶段，因此应用服务组件可提供相关支持。

基于模型的强化学习额外学习环境模型用于规划、模拟、数据增强，可由强化学习智能体、样本收集、模型更新模块提供支持。强化学习智能体的模型调用模块支持调用环境模型，动作生成模块支持使用模型规划；样本收集模块同时支持使用环境模型和强化学习环境生成样本；模型更新模块同时支持更新智能体拥有的决策模型与环境模型。

A.6 大语言模型（Large Language Model, LLM）中的强化学习

A.6.1 概述

大语言模型中的强化学习主要包括人类反馈的强化学习（Reinforcement Learning from Human Feedback, RLHF）及其变种，如AI反馈的强化学习（Reinforcement Learning from AI Feedback，RLAIF）等。RLHF通过人类反馈对大语言模型进行微调，使其输出更符合用户的期望。

目前，主流的RLHF实现方式包括以下步骤：

收集人类反馈：大量收集人类对问答的偏好数据。这些数据通常包括人类对不同问答对的评分或排序。

训练奖励模型（Reward Model）：利用收集到的问答-偏好数据集训练奖励模型。该模型能够根据问答的质量给出相应的奖励分数。

强化学习调优：使用奖励模型对大语言模型的问答进行强化学习调优。具体来说，通过PPO等强化学习算法，优化语言模型的参数，使其生成的回答能够获得更高的奖励分数。

通过这种方式，RLHF能够显著提升大语言模型的输出质量，使其更贴近用户的期望。

A.6.2 框架支持方式

收集人类反馈、训练奖励模型：收集人类反馈与训练奖励模型不属于强化学习框架的功能，但训练得到的奖励模型可使用适配器封装为强化学习环境，用于后续强化学习调优过程。

强化学习调优：大语言模型可封装为强化学习智能体，其输入输出均为文本。通过算法库创建强化学习训练工作流，使用奖励模型封装得到的环境以及大语言模型封装得到的强化学习智能体，即可实现强化学习调优过程。

A.7 数学、物理、化学、生物等基础科学中的强化学习

A.7.1 概述

随着强化学习不断普及，数学、物理、化学、生物等基础科学研究者开始逐步应用强化学习技术辅助求解本领域内问题。强化学习在自动定理证明、核聚变等离子体控制、分子合成路线设计、蛋白质结构设计等基础科学问题上都展现出了广阔的应用前景。

A.7.2 框架支持方式

场景适配：为基础科学领域的问题搭建强化学习环境时，通常需要调用领域特定的软件工具，因此场景适配应当为常用的此类软件工具提供必要的支持。

开发环境：基础科学领域的问题专业性强，用于这类问题求解的强化学习环境、智能体、工作流的开发、测试、调优均依赖领域专家，因此开发环境应当为领域专家与开发人员协作提供必要支持。

强化学习应用服务组件：对于将用于控制物理设备的智能体，例如核聚变等离子体控制智能体，强化学习应用服务组件需要支持智能体安全性验证，防止部署后对物理设备造成破坏。

A.8 具身智能中的强化学习

A.8.1 概述

具身智能（Embodied Intelligence）是指物理实体（如机器人、机动车、无人机等）通过和物理环境交互来表现和发展的智能。在具身智能的研究中，强化学习是一种重要的方法。

具身智能中应用强化学习的方式主要有模拟到现实迁移（Sim to Real）强化学习、真机强化学习（Real-World RL）。其中，模拟到现实迁移强化学习是在模拟环境中运行强化学习训练；而真机强化学习则是在物理环境中运行强化学习训练。两种方法所获智能体都将作为软件组件用于控制物理实体同物理环境交互。

A.8.1 框架支持方式

模拟到现实迁移强化学习训练支持：在模拟到现实迁移的强化学习中，场景适配需要提供将具身智能相关仿真器封装为强化学习环境的支持。封装所得强化学习环境应当支持系统参数随机化，以更精确地模拟物理环境。

真机强化学习训练支持：场景适配需要提供支持，以物理实体获取的物理环境信息和自身产生的信息为源数据，将这种源数据表达的问题场景封装为强化学习环境，用于训练强化学习智能体。此外，场景适配提供的状态重置、状态转移接口的实现需要支持一些特殊操作，例如状态重置可实现为向人类用户发送信号，以提示手动重置物理实体状态；状态转移可能需要同建立在物理实体上的特殊操作系统交互等。

部署支持：智能体的特征提取组件需要较精确地感知物理环境与受控物理实体的状态，并且动作生成组件需要避免生成可能损坏物理环境与受控物理实体的动作。