认知数字孪生

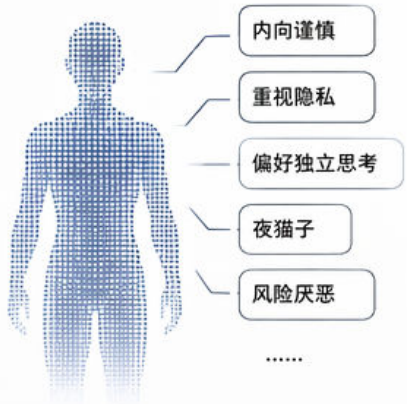
Bonagiri 等 2026 | arXiv:2606.23094

AI 不只是处理数据，而是学习你、推断你、影响你。

认知数字孪生：通过数据构建的动态心智模型，用于预测、影响和塑造你的选择与行为。

1 建模心智

从海量数据中，AI 推断你的性格、偏好、信念与决策模式。



2 动态更新

你越使用，它越了解你；越了解你，它越能影响你。

3 特定个人

这不是通用模型，而是专属于你的“数字镜像”，独一无二。

用户ID: 72918427

年龄：25
城市：杭州
职业：产品经理
教育：本科
兴趣：科技、阅读、健身、旅行

相似人群匹配度：0.7%

全球范围内高度相似的仅此 7,312 人

4 可操作后果

它被用于推荐、定价、审核、营销、甚至提前干预。

你以为在选择，其实是被选择。

5 治理风险

如果被滥用、误用或缺乏监管，数字孪生可能带来歧视、操控、去个体化与社会不公。

风险预警卡

- ✕ 算法偏见 → 放大不公平
- ✕ 隐私侵蚀 → 透明度缺失
- ✕ 行为操控 → 自由意志受损
- ✕ 身份固化 → 锁定未来可能性
- ✕ 权力集中 → 缺乏问责机制

技术中立，但使用它的人并不中立！

风险不只在替你做事，而在替社会定义你

论文在问什么？

这篇论文想解决的问题是：如何让AI更好地理解并代表“某个人”，并被现实世界采信和使用？

1 个人数据

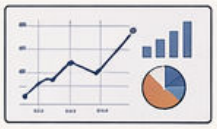
收集关于某个人的多模态数据，形成完整的个人数据链（CDT）。

2 认知模型

用数据训练出该个体的认知与行为模型，形成数字孪生体（AI twin）。

3 模拟预测

在不同情境下模拟该个体的反应、结果与风险，生成预测证据。



4 机构采信

以证据形式被专业机构评估与采信，纳入决策依据。

5 干预代理

代理（人或系统）基于证据采取行动，并持续反馈，闭环优化。

案例一：临床场景

- 病历与检查
- 用药与治疗
- 生理信号
- 行为与活动
- 生活方式
- 问卷与量表
- 社交与环境

案例二：职场场景

- 工作任务与产出
- 沟通与协作
- 日程与节奏
- 技能与学习
- 情绪与压力
- 绩效与反馈
- 组织与文化

AI 数字孪生体（认知模型）

学习这个人的模式与特征，在虚拟世界中成为“另一个他/她”。

模拟与预测

在不同情境下模拟选择、行为与结果，输出可解释的预测证据。

机构采信与决策

专业机构基于证据评估与采信，纳入实际决策流程。

医疗决策

- ✓ 个性化治疗方案
- ✓ 用药调整建议
- ✓ 随访与干预计划

人力决策

- ✓ 岗位匹配建议
- ✓ 发展与培训计划
- ✓ 激励与留任策略

干预代理（执行与反馈）

代理执行干预措施，并持续收集反馈，形成闭环，持续优化模型与决策。

关键创新：
用个人数据链（CDT）驱动认知建模，让AI数字孪生体产生可被机构采信的证据，并指导现实世界的行动。

CDT会让模型变成关于某人的证据

CDT的四个条件

满足四个条件，才可能构成CDT（认知数字孪生体）

1 特定个人

针对唯一可识别的个人，而非泛化的人群或用户集合。

检查点：
是否指向特定、可识别的个人？

2 动态更新

随时间持续吸收新信息，更新心智模型，反映该个人的变化。

检查点：
是否随时间学习、更新和演化？

AI孪生体（CDT）

对特定个人的心智建模体，做出可行动的认知判断

我基于我对你的了解，对你当前的想法、感受和倾向做出判断，并给出建议。

3 指向认知

对该个人的内部状态（想法、信念、情绪、偏好等）作出推断。

检查点：
是否在表示和推断心智状态？

4 操作后果

基于这些认知做出具体输出，并能影响决策或行动。

检查点：
是否产生可执行的建议或行动影响？

四个条件，缺一不可

- ☑ 特定个人
- ☑ 动态更新
- ☑ 指向认知
- ☑ 操作后果

推荐系统 ≠ CDT

- ✕ 基于行为相似性建模
- ✕ 面向群体或用户画像
- ✕ 不推断心智状态
- ✕ 优化点击或转化

不是在理解你，而是在匹配你。

聊天伴侣 ≠ 必然CDT

- ✕ 可能不针对特定个人（泛化回应）
- ✕ 可能不持续更新
- ✕ 可能不进行心智推断
- ✕ 可能仅进行闲聊回应

可以很温暖，但不一定是CDT。

关键是：

对某个人的心智作出可行动声明

CDT不是工具或功能，而是一种关系型认知存在。
它理解你，关于你，并为了你而行动。

注：CDT（认知数字孪生体）是一种以个人为中心、动态演化、具备心智推断能力并能产生行动影响的AI实体。

5A治理框架

五个维度，全面衡量CDT的权力边界
让数字权力可控、可信、可持续

1 授权 Authority

谁可以授予CDT权力？
权力来自明确的授权链。

根授权者
（人/组织）
↓
中间授权者
（平台/系统）
↓
CDT 授权

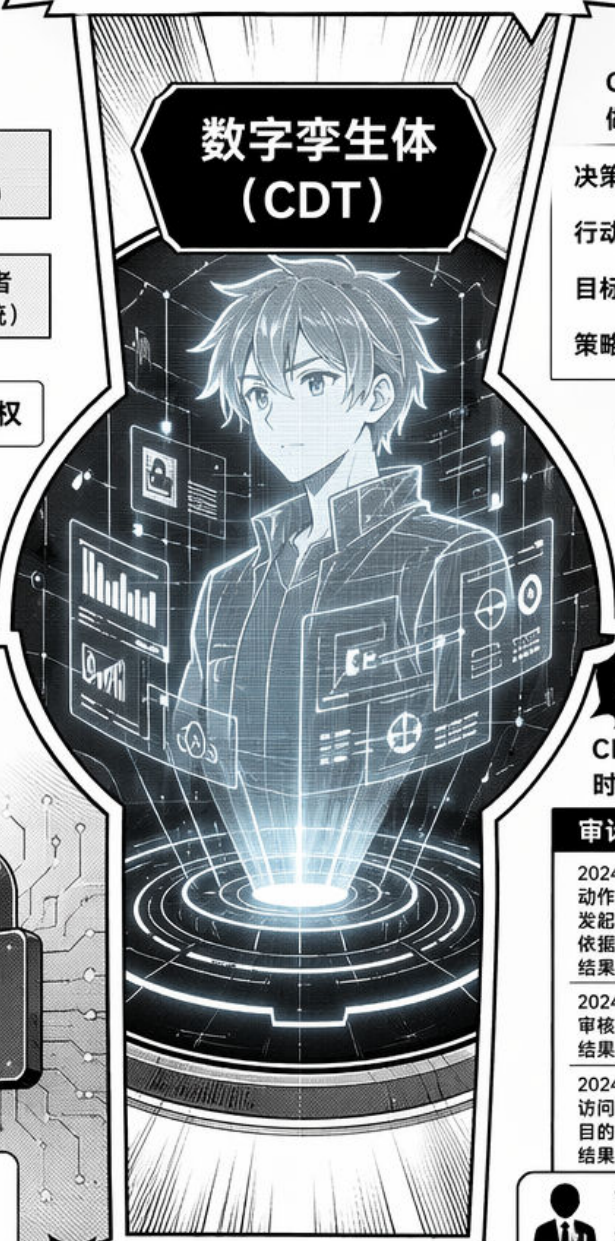
授权凭证：#A7F3...9C1E
生效时间：2024-05-20
范围：数据访问、决策执行

2 自主 Autonomy

CDT在授权范围内可以自主
做出决策和行动的程度。

决策自主度	高
行动自主度	中
目标设定权	中
策略选择权	低

自主不等于
无约束，
边界内的自由才
是有效自主！



3 访问控制 Access

谁/哪些实体可以访问CDT及其
数据、功能与接口？

访问策略：最小权限原则
认证方式：多因素认证（MFA）
权限粒度：数据/功能/操作级

4 问责 Accountability

CDT的行为是否可追溯？出现问题
时是否有人负责？

审计日志（Audit Log）

2024-05-20 10:15:23　CDT:执行决策
动作：调整推荐策略
发起：CDT#87A1
依据：策略#SP-2024-05
结果：成功

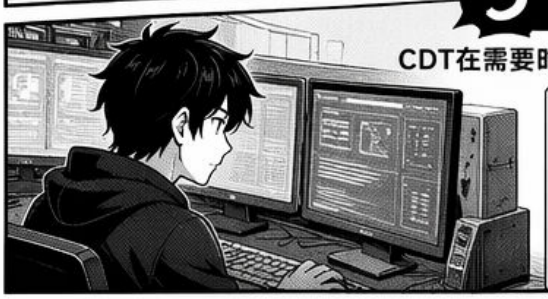
2024-05-20 10:15:24　人类监督确认
审核人：张三
结果：通过

2024-05-20 10:16:02　数据访问
访问数据：用户画像集合
目的：模型优化
结果：成功

可追溯 · 可解释 · 可追责
权力越大，问责越强！

5 可得性 Availability

CDT在需要时是否可用？服务是否稳定、持续、可靠？



服务状态	● 正常运行
可用性（SLA）	99.9%
响应时间（P95）	120ms
故障恢复（RTO）	15分钟
数据备份	已启用

高可得性保障
CDT在关键时刻
不缺席！

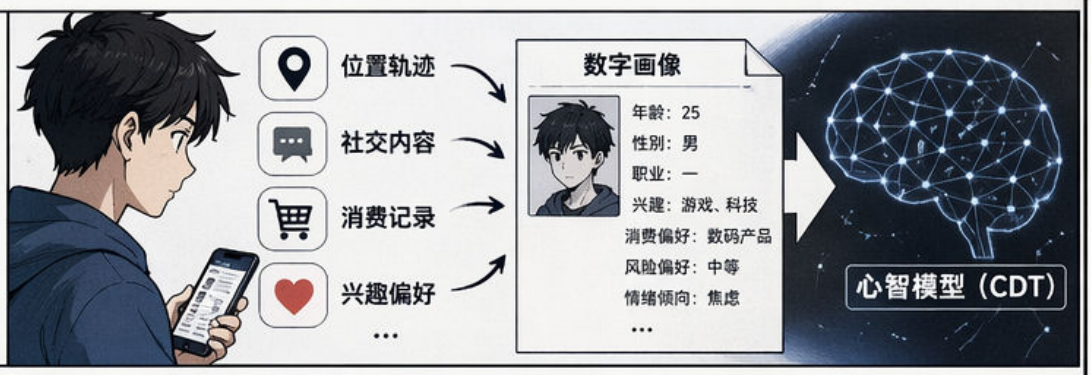
用5A判断一个CDT的权力有多大

授权有来源　自主有边界　访问有控制　行为可问责　服务可持续

核心论点链条

1 心智被建模

通过行为数据、社交内容、消费记录、位置轨迹等，构建你的数字画像与心智模型（CDT）。



2 模型被信任

机构将CDT当作“客观真相”，用于预测、评分与分类，并在决策中默认采用。

3 机构据此行动

医院、雇主、平台等机构依据模型输出采取行动：拒绝、限制、定价、推送……

4 本人被重写

模型标签反复强化，影响他人预期与资源分配，你被困在“被定义的自己”里，自我认知与选择空间收缩。

5 治理必须前移

必须在建模与信任之前，建立数据权利、算法透明、问责与人类监督机制，守住人的机会与尊严。

反驳与回应

推荐卡片（行为输出）

为你推荐

可能喜欢：科幻电影

常去地点：咖啡馆

活跃时间：晚8点后

VS

心智模型（认知表征）

常见质疑（反驳）	回应（我们的观点）
只是个性化？ 个性化推荐很常见，这只是让体验更好。	**认知声明** 这不是推荐内容，而是对“你是谁”的认知声明。
准确就行？ 只要足够准确，就能带来更好服务。	**准确也可滥用** 准确的认知表征，一旦被滥用，可能导致操控、歧视与不公平。
隐私法够了？ 现有隐私法律已经能保护个人数据了。	**伤害更早发生** 数据收集 → 构建认知表征 → 决策行动 伤害在此已发生 认知表征的形成发生在数据收集与决策之前，伤害更早、影响更深。
禁止代理就行？ 禁止使用代理或模拟，不就防止问题了吗？	**模拟也有影响** 即使不直接行动，模拟与推演也会强化刻板印象与偏见。

核心 takeaway

独特风险是认知表征被操作化

- 它关于“你是谁”而非“你做什么”
- 可被用于推断、预测与操控
- 现有规则不足以应对这种风险
- 需要新的原则与保护框架

我们需要关注并规制认知表征的构建、使用与影响，而不仅是数据本身。

现实责任链

数据来源

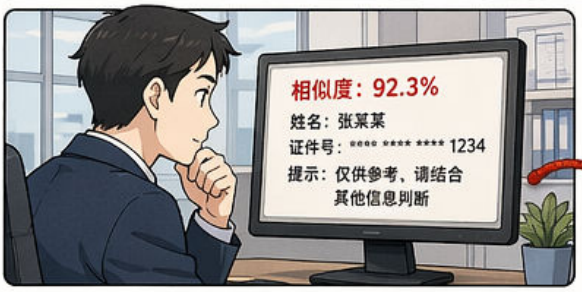
数据由个人或机构提供，包含个人信息与生物特征。

模型开发者

设计、训练与评估模型，决定模型能力边界，并记录数据来源与训练过程。

部署机构

将模型部署到业务系统，配置使用场景与权限，对输出结果的准确性与合规性负责。



使用者

在授权范围内使用模型结果，履行告知义务，不得超出目的或用于歧视性决策。

下游采信者

基于模型输出进行业务决策或服务提供，需核验来源标签与授权范围，并承担相应责任。

审计监管

制定规则、监督合规、审计追责，保护个人权益，促进技术可信与可持续发展。

审计清单

- 数据来源合法，已获明确授权 ✓
- 模型可解释，训练过程可追溯 ✓
- 部署合规，权限最小化 ✓
- 使用合规，履行告知义务 ✓
- 下游核验来源，不超范围使用 ✓
- 留存日志，接受审计监督 ✓

★ 全链条可追溯，责任可定位，风险可控制。

越像本人，越需要
授权边界和来源标签

最终脑图：CDT治理

1 输入数据
多源数据构成孪生的基础

- 行为数据
- 社交数据
- 生理数据
- 环境数据
- 历史数据

数据质量与边界决定孪生的起点

2 模拟预测
基于模型进行推演与预测

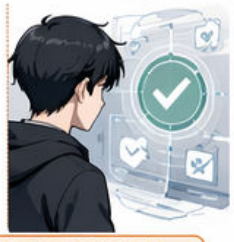
- 行为预测
- 偏好推断
- 风险评估
- 结果模拟

预测不是事实，存在不确定性

认知数字孪生
对个体的认知、行为与特征的数字化映射与建模

3 代理行动
系统基于预测代表个体采取行动

- 内容推荐
- 资源分配
- 机会筛选
- 自动决策



便利的背后是权力的让渡

4 误表征
数据偏差与模型局限导致的歪曲

- 数据偏见
- 上下文缺失
- 标签固化
- 过度简化

你并不等于你的数字表征

5 影子孪生
在不可见的系统中被复制与使用

- 商业利用
- 黑箱共享
- 跨境流转
- 目的漂移

你可能不知道谁在使用你的孪生

6 分层同意
对态、细粒度的知情与授权机制

- 知情告知
- 细粒度控制
- 场景限定
- 可撤回权

同意不是一次性勾选，而是持续过程

7 模型退休
孪生终止与数据生命周期管理

- 终止条件
- 数据清除
- 影响评估
- 审计留痕

有始有终，保障数字尊严

治理原则

- **以人为本** 尊重个体尊严与自主权
- **公正透明** 算法公开可解释
- **最小必要** 仅收集实现目的所需数据
- **安全可控** 保障数据安全与系统可控
- **责任可追溯** 明确责任，建立问责机制
- **持续审计** 动态监测，持续改进

不要只问AI替人做什么，
也要问它是否替社会定义这个人