智能系统测评分论坛

北京，22 May 2017

智能系统测评：现状与挑战

陈小平

中国科学技术大学

Outlines

- **现状、差异与争论**
  - 基于外显行为的和基于内在机制的
  - 基于任务的和基于标准的
  - 基于同类比较的和基于参照物比照的
- **疑难与挑战**
  - 用户依赖性
  - 环境相关性
  - 价值渗透性
- **结语**

智能系统测评现状与争论

现状：外显行为/内在机制？

- **图灵测试**
  - 基于外显行为
  - 基于参照物（人）对比
  - 行为限于问答，环境无关
- **对图灵测试的批评**
  - **例**：Searle’s “Chinese room”
  - 行为与机制的争论
  - 对环境无关性的普遍默认
- **图灵的反驳**
  - 预期到几乎所有批评，提前反驳了所有预期批评

艾伦-图灵

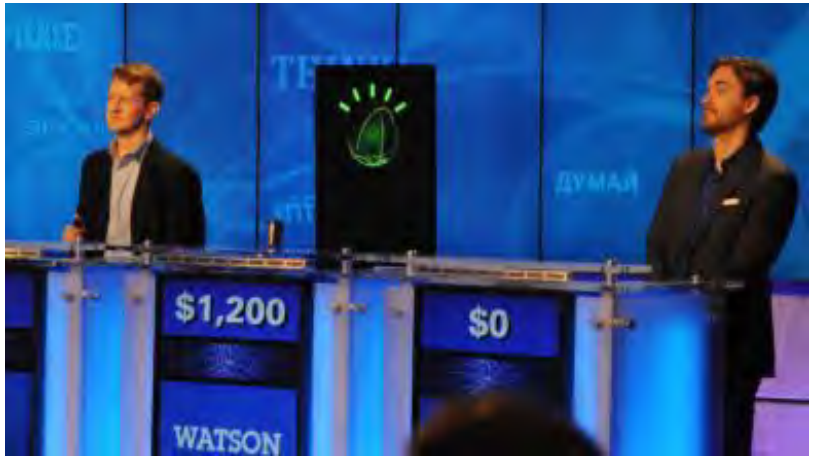
图 1　图灵测试示意图

现状：基于任务/基于标准？

- **基于任务的测评**
  - 设计一组特定任务，根据完成情况评分
  - **例**：IQ
  - **例**：国际服务机器人标准测试RoboCup@Home
- **基于标准的测评**
  - 参照给定标准评分
  - **例**：产品评测
  - 通常针对特定对象类的特定功能与品质。

现状：同类比较/参照物比照？

- **基于同类比较的测评**
  - 与同类对象的测试得分比较
  - **例**：IQ（相同年龄段比较）、RoboCup@Home（同类机器人比较）
- **基于参照物比照的测评**
  - 借用参照物（人）的相关测试的标准和方法
  - **例**：应用IQ标准和方法测评机器人智能
  - **例**：图灵测试
  - **例**：IBM Watson人机大战（Jeopardy），深蓝和AlphaGo人机大战（国际象棋、围棋）



智能系统测评疑难与挑战

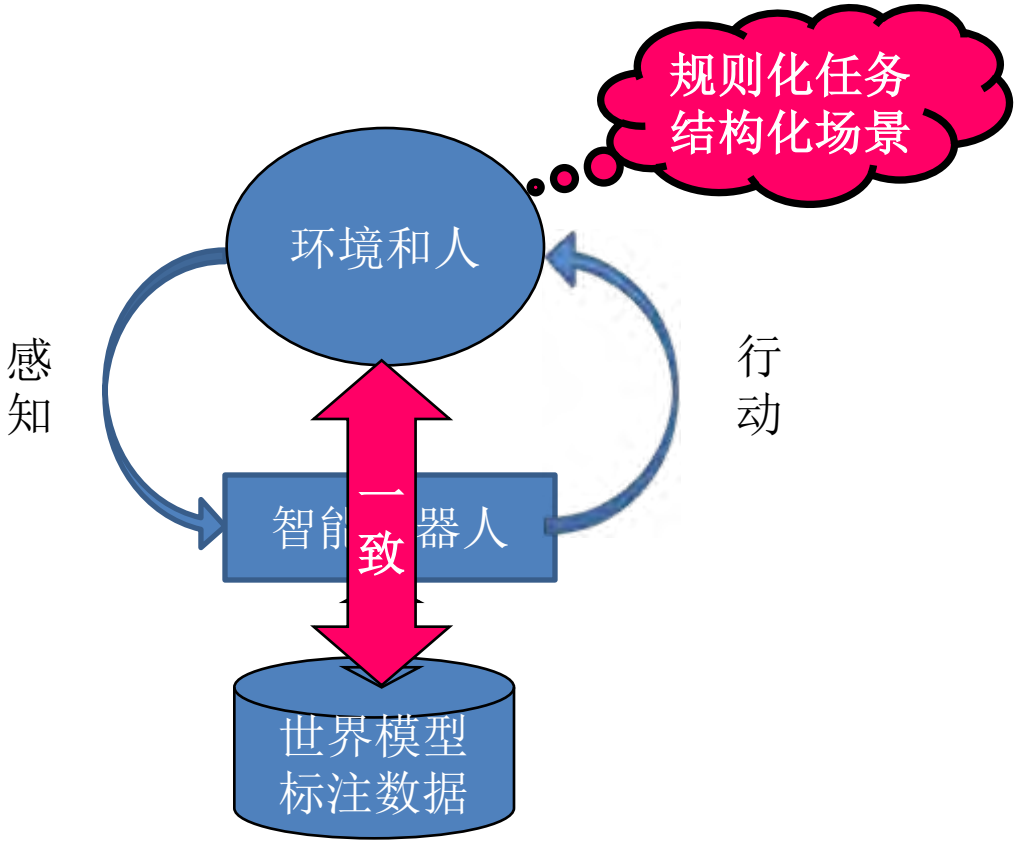
疑难与挑战1：用户依赖性

- **测评对测评者/用户的依赖性**
  - 不同的用户可能对相同智能系统的相同行为给出矛盾的评价，对依赖于用户评价的系统的测评提出挑战。
  - **例**：信息推荐。
  - **例**：复杂家庭服务。
- **对智能系统的测评涉及对智能系统用户的“测评”**
  - 用户需求通常自然隐含在产品检测标准中，但传统产品较少考虑用户的个性化需求。
  - 传统的“科学评价”准则往往要求测试者无关性。
  - 用户可以比产品更复杂。

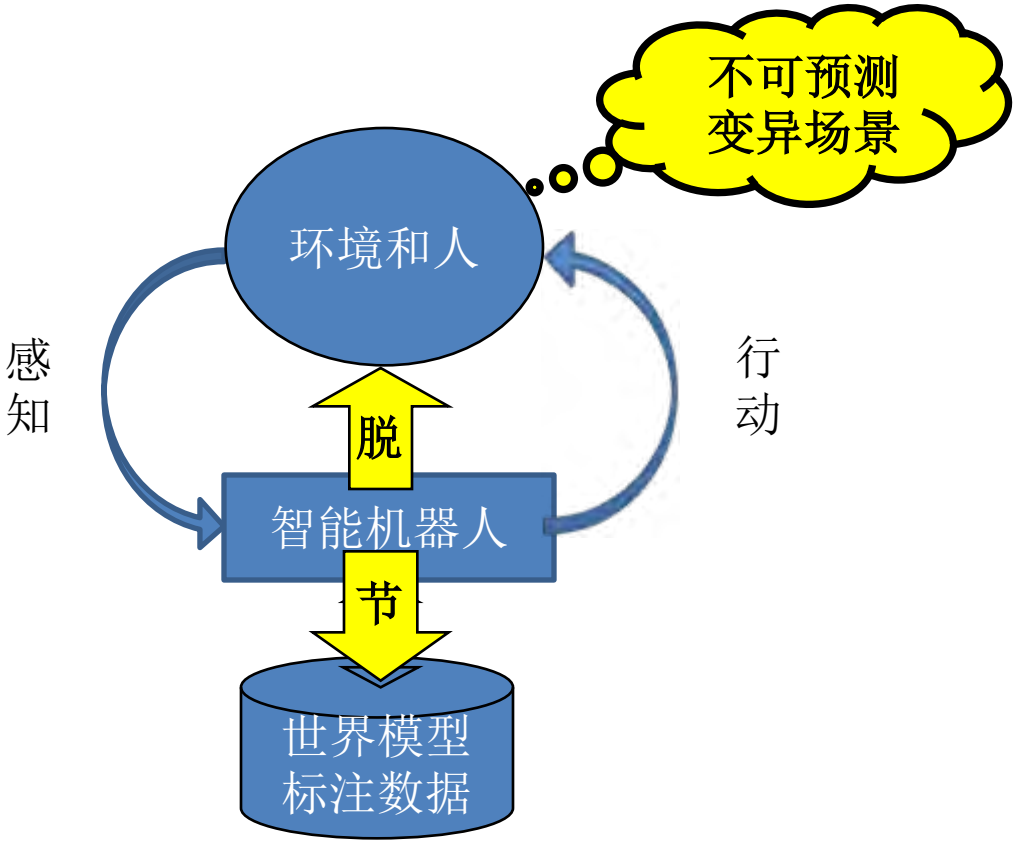
疑难与挑战2：环境相关性

- **测评与应用环境的相关性**
  - 图灵测试假设环境无关性
  - 某些智能系统的性能与应用环境紧密相关
  - **例**：无人车与路况相关性
  - **例**：移动操作智能服务机器人，如家政服务员和餐馆服务员（RoboCup@Home）。
- 适应**任何给定的**真实环境不难，适应**所有可能的**真实环境很难。
- 应用环境的**不可预测性**同样存在于系统建造和性能评价中。

例：智能机器人



智能机器人：科学挑战

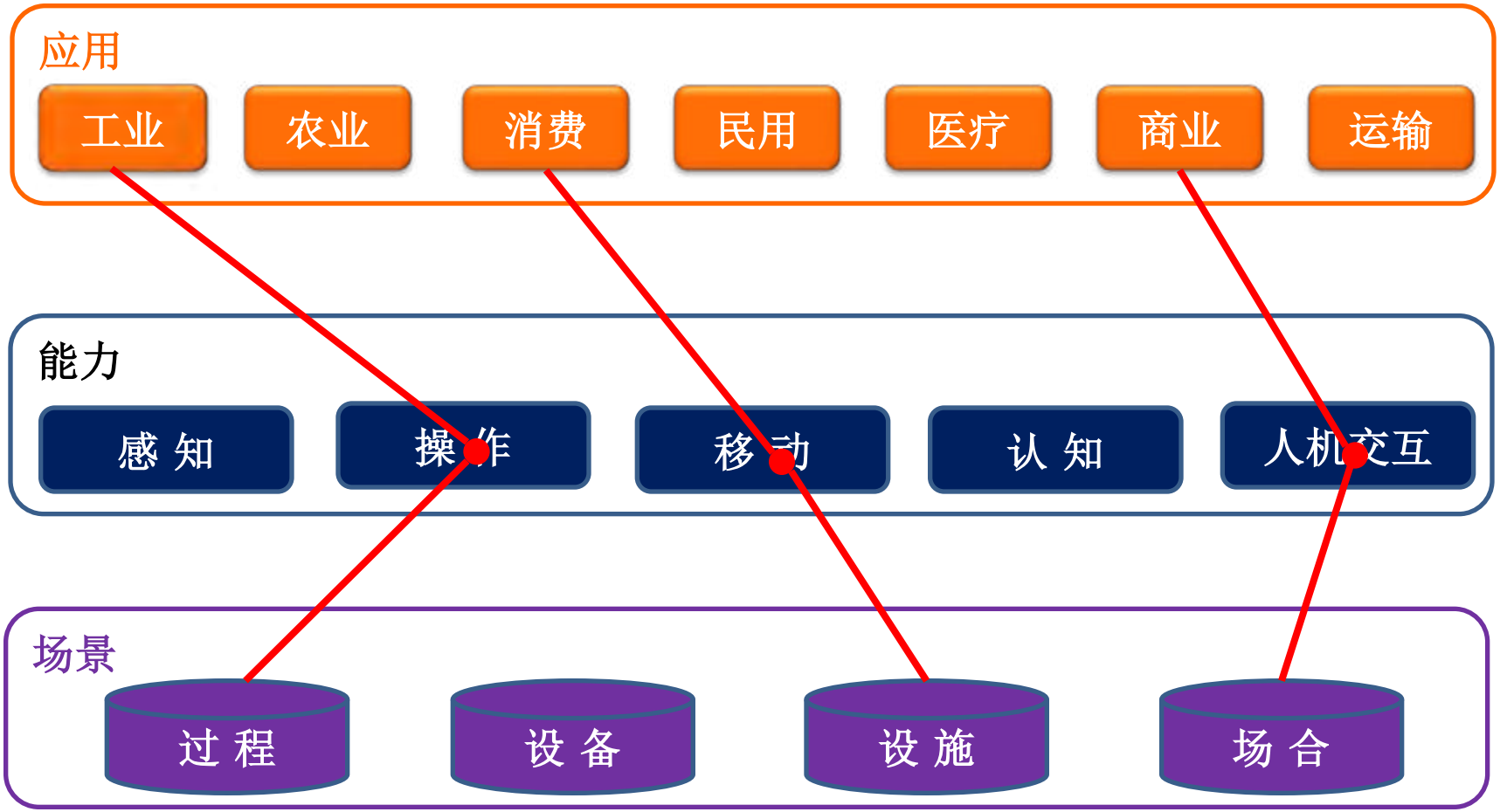


疑难与挑战3：价值渗透性

- **智能系统性能/能力水平与实用价值的关系？**
  - 传统的“科学评价”准则往往不直接反映实用价值。
  - **例**：图灵测试、IQ不考虑实用价值。
  - **例**： Watson与实用价值相对较近，但不是用实用价值评价的。
- **不考虑智能系统的价值渗透性不利于智能系统测评的作用发挥**
- **如何协调智能系统能力与价值两种评价？**
  - 能力强未必价值大；
  - 能力弱未必价值小。

结 语

横向概括：机器人技术-应用空间



纵向展望：未来人-机器人场景互动示例

用户: 冰箱的用途是什么?
可佳: 冰箱是用于食品保鲜的。
用户: 怎么保鲜呢?
可佳: 食品买回来放冰箱里，吃的时候再取出来。

问答型任务

用户: 太好了！你赶紧从冰箱里拿一点吃的给我。

操作型任务

可佳: 别做梦了！家里的东西全被你吃光了。
用户: 那你怎么不买呀?
可佳: 昨天就告诉你了，你不给钱啊!

因果型任务

用户: 给你，多买一点。然后赶紧做午饭。
可佳: 这就对了，下次早点给。现在等着吧。

操作型任务

结　语

- 测评是人工智能研究的开端，正在成为核心内容之一。
- 智能系统测评存在长期争论，隐含重大科学问题、社会需求和技术需求。
- 智能系统测评极具挑战性，涉及人工智能研究与应用的一系列深层课题，孕育人工智能突破的重大机遇。

谢　谢！