①

AGI 的自主意识

意识 是一种 meta-representation, 一个讯息处理系统 表达系统自身的 状态 的能力。换句话说，是一个系统 向内 窥视自己 (introspection) 的能力，一个 “inner loop”. 这个 对意识的定义，我最早在 神经科学家 Joseph LeDoux 的书里看到 [1]。

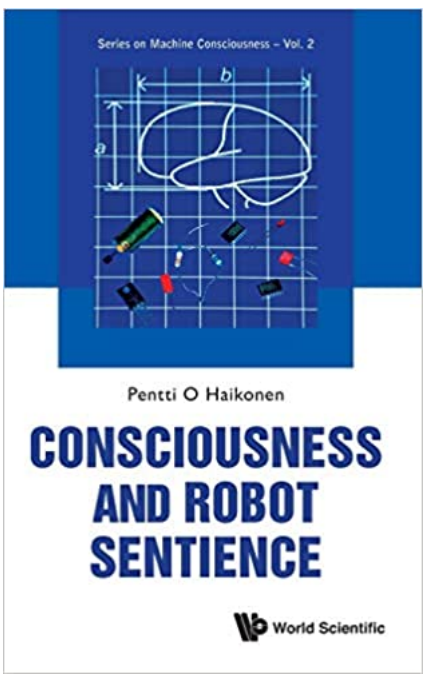
例如 一个 抽水马桶，它能够 检测水位（讯息处理），但它不知道 自己在做什么，因此它没有自主意识。

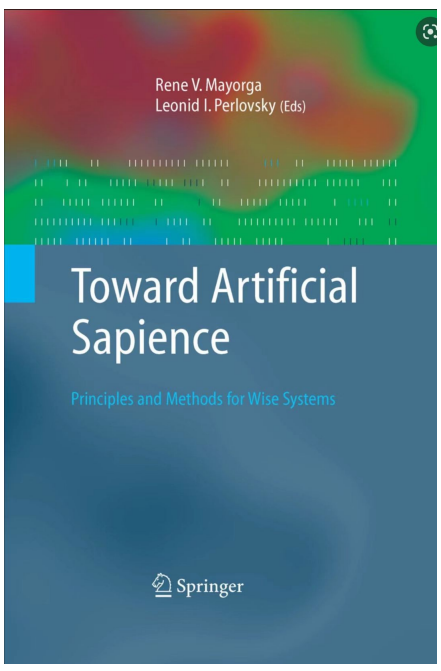
1 意识 对 AGI 的用处

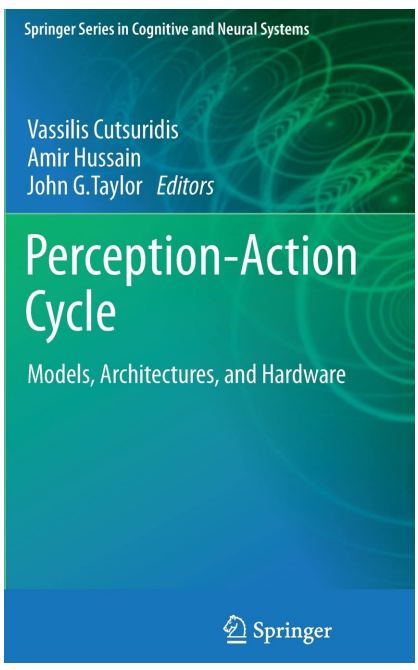
在 逻辑 AI 系统里面，这种 meta-representation 不难做到：它只需要将系统的状态 变成 逻辑命题，再放进 工作记忆里。例如说，系统可以知道 自己在下棋，而且知道自己 思考某一步 用了多少时间。这种能力 跟 逻辑 AI 的 TMS (Truth Maintenance System) 有点类似：后者负责 记录 逻辑推导的 过程 (inference trace), 而这种记录 本身，也是 逻辑命题，可以用作 推理的资料。

意识的 inner loop 可以极大地提升 AGI 的效率，因此是非常重要的课题。这也涉及到 AGI 认知架构 (cognitive architecture) 的设计。

以下几本书有更深入的讲述：







2 Qualia

哲学上，*qualia* 的产生就是因为这个 inner loop 的存在。有 inner loop 的生物 就会有 自我意识，这也包括 机器。这种观点 称为 functionalism，意思是说 *qualia* 是某种结构的 功能，正如「报时」是「时钟」的功能。一部机器如果有某种结构，它就有时钟的功能，而这并不取决于这机器是用什么造的。它可以是一堆齿轮，也可以是 芯片，也可以是一些细胞，甚至是星球之间的运动。

3 「机器灵魂」的道德问题

如果 AGI 具有跟人类一样或相似的 情感 (emotions)，那么理论上它就是一个类人 (android) 或 复制人 (clone, cyborg). 理论上他们需要被尊重、不能被杀害、这立即引起一连串的道德伦理问题。这跟我研究 AGI 的愿景 – AGI 作为为人类服务的工具 – 是不同的。

首先，大脑的情感 由复杂的 生物讯号 dynamics 决定，它不容易复制，复制的「仿真度」也很难定义。

更重要的是：复制人的智能 很容易超越人类，他们会跟人类 争夺资源，甚至导致人类的种族灭绝。

我希望 人类会跟 AGI 共同进化，我希望 人变成 “Transhuman”的过程是 连续的。

---

[1] *The Emotional Brain* [Simon and Schuster, 1998], *The Synaptic Self: How Our Brains Become Who We Are* [Viking, 2002]