朱松纯团队最新成果：让机器人学会察言观色

作者：writer 来源：爱科学

本文原地址：https://www.iikx.com/news/progress/19245.html

朱松纯团队最新成果：让机器人学会察言观色。机器人可以进化出察言观色的本领吗？科研人员的答案是肯定的。7月14日，机器人领域权威学术期刊Science Robotics
发表了AI领域知名学者朱松纯团队的最新研究成果——实时双向人机价值对齐（Bidirectional human-robot value alignment）。论文提出了一个可解释的人工智能（XAI）系统，阐述了一种机器实时理解人类价值观的计算框架，并展示了机器人如何与人类通过实时沟通完成一系列复杂人机协作任务。

AI 机器人读懂价值观意象图 BIGAI 供图

人机协作的一个重要基础

生活中很多任务是无法清晰地描述或者难以直接表达的。论文共同第一作者、北京通用人工智能研究院研究员郑子隆举例说：例如一个人买衣服，售货员挑了一件4000 元的衣服，他摇头说不合适；售货员又拿了一件5000元的，他又摇了摇头；此时售货员会意，拿了一件只卖800元的衣服，这时他点了点头。这时售货员就做到了察言观色，读懂了顾客买衣服价值需求：物美价廉。在不同场景下，确保AI系统能够快速准确地识别用户的价值目标、和人类的价值实时双向对齐，是人机协作的一个重要基础，我们的研究团队在这一难题上取得了突破。

当今广泛应用的人工智能系统是一种被动的智能，只能机械地按照人类给定的任务行事，缺乏像人类一样的认知和推理能力，更缺乏像人类一样的情感和价值观，在缺心的情况下，人工智能很难理解人类的意图、执行人类真正在意的价值需求，自然也就难以获取人类的信任、难以融入人类社会。

我们这个项目侧重于人机协作完成任务。郑子隆向《中国科学报》介绍：智能体作为执行者不仅需要理解人类用户的指令含义，更需要推测指令背后的意图、目的、想法，这些因素被称为人类的价值观/目标。而智能体如果想要走入千家万户，就必须理解人类的价值观，做到‘察言观色’。这一理解并收敛到人类价值观的过程，就是‘价值对齐’。只有价值对齐之后，机器才可以更加自主地执行任务，而无需依赖人类的指令。

郑子隆说，实验结果也表明，价值对齐机制能极大地提升人机在协作过程中的信赖关系，而这正是实现通用人工智能的必经之路。

现阶段AI产生的价值观

智能机器人没有情感和同理心，产生的是什么样的价值观？

这是个非常好的问题。论文共同通讯作者、北京大学人工智能研究院助理教授朱毅鑫告诉《中国科学报》：当前的机器不仅仅是缺芯，更加缺心，我们这项工作就是希望能迈出给机器‘立心’的重要一步。

朱毅鑫解释说。给机器‘立心’这重要的一步，反映在了价值观的对齐任务上。价值观有时候比较容易描述，比如说我喜欢喝茶，不喜欢喝咖啡。但有些情境下，价值观是相对比较难描述的，比如驾车从A地到B地，需要考虑油价、高速收费、沿途风景、道路交通情况等等。我是希望快点到、牺牲沿途风景呢，还是希望最省钱、同时能看到一些美丽的风景？价值函数在这个例子里就是几个因素的权重。

而在实时双向人机价值对齐这篇论文中，他们就用了一个类似的合作任务来描述：价值观是执行时间、探索区域的大小、最大化资源获取等因素的权重。

人类能打造AI大白吗

在科幻电影《超能陆战队》中，有一个大白智能陪伴机器人，大白可以陪电影男主角一起学习、玩耍、做游戏，具有很高的实时互动性。而当电影男主角情绪失落时，大白还能读懂他的情感价

值需求，主动安慰，给一个大大的拥抱。

人类能打造AI大白吗？这其实也正是当前机器人科学家的努力方向。

在发现大数据，小任务范式自身存在的局限性之后，朱松纯教授研究团队转换赛道，致力于小数据、大任务范式的探索。朱毅鑫介绍，本研究提出了一个基于即时双向价值对齐模型的可解释人工智能系统。在该系统中，一组机器人通过与人类的即时交互并通过人类的反馈来推断人类用户的价值目标，同时通过解释将其决策过程传达给用户，让用户了解机器人做出判断的价值依据。此外，该系统通过推测人类的内在价值偏好，并预测最佳的解释方式，生成人类更容易理解的解释。

研究团队经过一系列实验验证了所提出的计算框架。实验结果表明，该学习模型可以在复杂协作任务中提高人机协作的效率，进而提升人机信赖关系，实现自主智能。同时，这项成果也揭示，人工智能系统能够具备在实时交流中学习人类价值函数并实时对齐当前人类价值目标的能力。

从传统AI的数据驱动转变为价值驱动，让XAI系统理解人类价值观，这在研究团队看来就是为机器立‘心’，是朝着实现小数据，大任务范式的通用人工智能的一大步。

朱松纯团队长期从事可解释人工智能相关工作。此文是该团队发表在 Science Robotics 的第二篇关于可解释人工智能的论文。这项研究涵盖了认知推理、自然语言处理、机器学习、机器人学等多学科领域，是朱松纯教授团队交叉研究成果的集中体现。（来源：中国科学报赵广立）

相关论文信息：https://doi.org/10.1126/scirobotics.abm4183

作者：朱松纯等 来源：《科学—机器人》