伦理智能体及其设计:现状和展望

古天龙[1),2)] 李 龙[1),2)]

[1)](暨南大学信息科学技术学院/网络空间安全学院 广州 510632)

[2)](桂林电子科技大学广西可信软件重点实验室 广西 桂林 541004)

摘 要 智能体一直是人工智能的主要研究领域之一,任何独立的能够同环境交互并自主决策的实体都可以抽象为智能体.随着人工智能从计算智能到感知智能,再到认知智能的发展,智能体已逐步渗透到无人驾驶、服务机器人、智能家居、智慧医疗、战争武器等人类生活密切相关的领域.这些应用中,智能体与环境、尤其是与人类和社会的交互愈来愈突出,其中的伦理和道德问题日益凸显.人工智能应用的伦理风险和挑战引起了人们的普遍关注,伦理智能体是人工智能伦理的重要研究内容.本文从人工智能伦理的工程设计与实现角度,对伦理智能体及图灵测试、伦理智能体的设计范式、伦理智能体的逻辑程序设计、伦理智能体的形式化验证、伦理困境及分析等进行了介绍和讨论.同时,对伦理智能体及设计所面临的挑战和进一步研究方向进行了述评和展望.

关键词 伦理智能体;人工智能伦理;伦理设计;逻辑程序设计;形式化验证

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2021.00632

Artificial Moral Agents and Their Design Methodology: Retrospect and Prospect

GU Tian-Long[1),2)] LI Long[1),2)]

[1)](*College of Information Science and Technology/College of Cyber Security, Jinan University, Guangzhou* 510632)

[2)](*Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, Guangxi* 541004)

Abstract Artificial agents have always been one of the main research fields of artificial intelligence. Any independent entity that can interact with the environment and make decisions autonomously can be abstracted as an agent. With the development of artificial intelligence from computational intelligence to perceptual intelligence, and then to cognitive intelligence, artificial agents have gradually penetrated into many fields closely related to our human life, such as unmanned driving cars, service robots, smart households, voice assistants, intelligent medical care and war weapons. In these applications, the interactions between agents and among agents and the environment, especially humans and society, are becoming more and more prominent, and the interactions of agents with humans and society inevitably resulting in ethical threats and moral attentions. Issues regarding artificial moral agents are important research contents of ethical artificial intelligence. The ethical risks and challenges in the application of artificial intelligence have aroused more and more public concerns. The ethical issues of artificial intelligence belong to the interdisciplinary research of social science, philosophy, psychology, cognitive science, computer, artificial intelligence, etc. On the one hand, philosophy and humanistic social scientists focus on the moral subjects of artificial intelligence, the effects of artificial intelligence on the human society, the ethical norms and moral codes of artificial intelligence and so on. On the other hand, researchers in the engineering fields such as computer, cognitive science and artificial intelligence are interested in the embedded

收稿日期:2020-04-06;在线发布日期:2020-11-05.本课题得到国家自然科学基金(U1711263,U1811264,61966009)、广西自然科学基金(2019GXNSFBA245049)资助.**古天龙**,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为形式化方法、可信人工智能、人工智能伦理、数据治理等.E-mail:gu@guet.edu.cn.**李 龙**(通信作者),博士,讲师,中国计算机学会(CCF)会员,主要研究方向为人工智能安全、逻辑程序设计等.E-mail:lilong@guet.edu.cn.

ethical decision design, the implementation of moral artificial intelligence, the ethical supervision and control of artificial intelligence and others. In the framework of ethically aligned design, realization and deployment of moral artificial intelligence, this paper reviews the design methodology of artificial moral agents in the following aspects. Firstly, the concepts of artificial moral agents are introduced, and in terms of multiple behavioral characteristics, such as autonomy, adaptation, evolution and consciousness, the categories of artificial moral agents are illustrated. Although there is still controversial on the effectiveness of the Turing test, as a promising way to identify and evaluate moral artificial agents, the moral Turing test is addressed briefly. Ethical norms and moral codes are the premises for the design and implementation of artificial moral agents, and many government agencies, industry associations, companies and international organizations have put forward a series of moral codes and ethical norms to manage and supervise artificial intelligence. The open problems, such as how to properly choose and strictly abide by them, and how to translate them to machine or computer codes are discussed, etc. Secondly, the three design paradigms of artificial moral agents, including top-down, bottom-up and hybrid approach are analyzed, and the design of moral artificial agents via logical programming is elaborated, which integrates two different concepts of both logic and programming. Formal verification and validation of artificial moral agents are also outlined. Finally, the ethical dilemma is a common puzzle of philosophy, social science and engineering disciplines, and the potential solution using computational ethics and crowdsourcing technology to defend it is described. More importantly, a comprehensive outlook of the challenges and the further research directions of moral artificial agents is presented in the paper.

Keywords　artificial moral agents; artificial intelligence ethics; ethically aligned design; logic programming; formal verification

1　引　言

智能体是人工智能(Artificial Intelligence,AI)领域的一个重要概念.任何独立的能够自主决策并同环境交互的实体都可以抽象为智能体.Wooldridge将能够在环境中自主运行并实现预期设计目标的计算机系统称为智能体[1].智能体具有如下基本特性:(1)自主性.智能体具有自我管控和自我调节的能力,亦即能够根据环境的变化自主调整自身的行为;(2)反应性.智能体具有反应外界激励或外界变化的能力,亦即能够依据外部的激励或变化自动做出响应或反应;(3)主动性.智能体具有主动采取活动的能力,亦即能够主动调整自己的行为来应对外部环境的变化;(4)社会性.智能体具有与其它智能体或人类进行合作的能力,亦即能够通过与其它智能体进行交互来协同解决问题;(5)进化性.智能体具有自我学习和进化的能力,亦即能够通过积累或学习经验知识来修正自己的行为以适应新的环境.

随着人工智能从感知智能到认知智能的过渡和发展,智能体的应用已拓展到无人驾驶、医护机器人、智能家居、军用武器等领域,渗透到人类社会和生活的诸多方面.这些应用中,智能体与人类的交互日益凸显.智能体的行为是否需要符合伦理[2-3]?如何开发出符合伦理的智能体[4-5]?智能体如何实施伦理决策?等等.这些问题引起了人们的高度关注[6].控制论之父维纳早在1950年出版的《The Human Use of Human Beings》[7]中就曾有过描述,担心人类会创造出“按照人类无法接受的价值实施行动的智能体”,…,“为了防止这样的灾难,既需要为人工智能体设置伦理规则,也需要将这些规则嵌入智能体,并通过有效技术来管控智能体的行为.”Allen、Varner和Zinser认为:智能体具有自主完成对人类有益活动的能力,也同样会有实施对人类有害活动的能力,对智能体有害活动能力的管控,已从科幻小说进入现实的社会生活[8].Yu等人从人工智能伦理困境、单智能体伦理决策框架、多智能体伦理决策框架、智能体与人交互的伦理等方面,对人工智能的伦

理管控的技术实现进行了介绍和讨论[6]. Cervantes 等人从伦理智能体的分级和伦理智能体设计方式角度,对伦理智能体的计算模型、行为的形式化验证等方面进行了综述[2].

伦理智能体属于社会科学、哲学、心理学、认知科学、计算机、人工智能等的多学科交叉研究. 社会科学和哲学等领域的讨论和研究主要聚焦于:智能体是道德主体吗? 智能体对社会有何影响? 智能体的伦理规范是什么? 是否存在经典伦理理论(如功利主义、义务论、美德论)的实现? 智能体的道德责任如何追究? 智能体的伦理或道德困境等等[9-11]. 计算机和人工智能等工程领域所关注的问题涉及:智能体的伦理决策如何实现? 如何设计伦理智能体? 如何生成智能体的伦理规范? 如何管控伦理智能体使其具有符合伦理或道德的行为? 等等[12-13]. 本文旨在从计算机和人工智能等工程的视角,对伦理智能体及其图灵测试、伦理智能体设计、伦理智能体的形式化验证、伦理困境及分析等的研究现状和预期研究进行介绍和讨论.

2 伦理智能体及伦理规范

2.1 伦理智能体与图灵测试

智能体的基本特性决定了其伦理属性的必然性. Floridi 和 Sanders 指出:行为符合伦理规范的实体称为伦理智能体(Artificial Moral Agent, AMA)[3]. Allen、Varner 和 Zinser 认为:不仅考虑自身而且考虑他方利益的智能体,称为伦理智能体[8]. Moor 给出的定义是:伦理智能体是能够明确地进行伦理判断、并对判断结果的合理性进行说明的智能体[9]. Cervantes 等人指出:伦理智能体是能够从事道德行为,或者至少避免不道德行为的智能体[2]. 简而言之,具有伦理或道德决策能力、且行为符合伦理或道德的智能体称之为伦理智能体.

Moor 将伦理智能体粗略划分为四种类型[9]:(1)伦理影响智能体(Ethical Impact Agents);(2)隐式伦理智能体(Implicit Ethical Agents);(3)显式伦理智能体(Explicit Ethical Agents);(4)完全伦理智能体(Full Ethical Agents)(参见表 1). 伦理影响智能体是对社会和环境产生某种程度上的伦理影响的智能体. 例如,计算机可以为我们的生活带来便利,也可以带来负面影响. 计算机可以让我们很容易且轻松地在网上娱乐或购物,但用户个人隐私信息也存在被泄露或盗窃的风险. 在这个层次上,计算机确实没有任何自主决策能力,也没有任何自主行动的意愿,但是它们的功能与服务会产生一定的道德后果(无论是直接的还是间接的),所以计算机就是伦理影响智能体. 隐式伦理智能体是被设计成隐式地遵循某种伦理规则的智能体. 这样的智能体不可能真正做出任何意义上的自主行为,也就不可能做出不道德的行为. 例如,银行 ATM 柜员机可以执行许多人工出纳员的任务,机器按照客户的指令要求完成每次银行交易操作. 我们并没有通过代码或其它方式告诉 ATM 柜员机要做到诚实,但是 ATM 柜员机和用户的交互行为符合了诚实的规则,ATM 柜员机就是隐式伦理智能体.

表 1 伦理智能体类型及特征

类型	定义	特征			
		主动性	适应性	进化性	意识性
伦理影响智能体	对社会和环境产生伦理影响的智能体	无	无	无	无
隐式伦理智能体	被设计成隐式遵循伦理规则的智能体	弱	弱	无	无
显式伦理智能体	对伦理进行描述并做出决策的智能体	中等	强	中等	无
完全伦理智能体	具有自由意志和行动意识的智能体	强	强	强	有

显式伦理智能体和完全伦理智能体有一个共同特性[9]:它们可以在几乎没有人类监督的情况下运行. 显式伦理智能体是能够对伦理进行充分描述,并做出伦理决策的智能体. 显式伦理智能体能够参照道德规则来计算最佳的行动和决策. 例如,可以使用道义逻辑、信念逻辑、归纳逻辑或行动逻辑等,对信息传输中保护个人隐私的允许/禁止操作进行描述和推理[14-15]. 完全伦理智能体能够像人类一样具有"信念、欲望、意图、自由意志和行动意识"[2,9]. 目前,一个普通的成年人类才能被认为是完全伦理智能体,而关于机器是否可能成为一个完全伦理智能体则存在争议[16-18]. 机器能具有情感和意识吗? 机器能成为完全伦理智能体吗? 这也是人工智能发展的相关问题:强人工智能或通用人工智能是否会出现? 何时出现? 完全伦理智能体是伦理智能体研究与开发的终极目标[2,9].

Allen 等人提出了伦理智能体的图灵测试,并讨论了其实施的可能性[8]. 图灵测试(Turing Testing)源于图灵对智能问题和机器思维的探讨[19],其主要思想在于:实验对象包括测试人和被测试者,测试中将被测试者和测试人(一台计算机和一个人)隔离,通过装置向被测试者随意提问. 如果计算机在限定时间内,能回答出人类测试者所提出的问题,且测试

人不能对超过 30％的回答区分出哪个是人、哪个是计算机，那么这台计算机就通过了测试，并被认为具有人类智能. 2014 年 6 月 8 日，英国伦敦皇家学会在雷丁大学举办了一场“图灵测试”，由俄罗斯团队开发的一款名为“尤金·古特曼(Eugene Goostman)”的计算机软件通过了测试. 在测试中，尤金·古特曼模仿一名来自乌克兰的 13 岁男孩，成功地让测试人认为该计算机软件的 33％的答复为人类所为，尤金·古特曼成为了首个通过图灵测试的计算机软件. 伦理智能体的图灵测试是图灵测试的一个变体[8]. 它将标准图灵测试的一系列对话话题限制在伦理相关范围的内容. 将测试人和被测试者(一个人和一个智能体)隔离，人类测试者提出伦理相关话题，被测试者进行回答，如果在限定时间内，测试人不能以高于规定正确率的方式识别出哪个是人、哪个是智能体，那么该智能体就通过了测试，所测试的智能体就是一个伦理智能体.

伦理智能体的图灵测试要比图灵测试复杂得多，面临如下一些挑战. 首先，图灵测试存在争议，有人认为图灵测试受限于短短的有限时间，由此来判断一个程序是否是人工智能，未免有些草率. 伦理智能体的图灵测试也需要在规定有限时间内完成，存在类似的问题. 其次，伦理或道德存在不同的水平层级，四种类型的伦理智能体如何定性、甚至定量区分？测试过程中又如何实现？最后，伦理相关对话话题如何设计？如何在设计中考虑伦理理论和伦理规则的困境？伦理测试应该基于什么样的伦理理论或伦理规则来设计？等等.

2.2　智能体的伦理规范

伦理智能体必须回答的问题是：伦理智能体要遵守什么样的伦理准则和道德规范？伦理智能体的伦理规则和道德规范如何确定？伦理学以道德问题为研究对象，在长时间的发展过程中形成了多种系统化、理论化的道德规范，即伦理学理论[5,8,20]. 伦理学理论是人类社会中人与人、人与环境、人与社会赖以和谐共生的重要基础，当然也是伦理智能体的行为和决策应该首先遵守的准则和规范[21]. 义务论和结果论是两种主要的经典伦理学理论.

义务论也被称为“道义论”或“非结果论”，指人的行为必须遵照某种道德原则或按照某种正当性去行动的伦理理论[21]. 义务论认为判断人们的行为道德与否，不必看其行为的结果. 康德主义和社会契约论是义务论的重要代表. 康德主义通过“绝对命令”或“定言命令”强调意志自律和伦理原则的普遍有效性. 康德把“绝对命令”表述为：“不论做什么，总应该做到使你的意志所遵循的准则永远同时能够成为一条普遍的立法原理.”霍布斯认为：“如果社会中没有了规则，或者失去了可以约束人民执行规则的力量，那么人们就会担心自己创造的价值是否是属于自己的，因此而放弃创造价值. 生活在文明社会的每一个人必须要遵守两件事：其一，通过形成一种道德规范来管理人与人之间的关系；其二，建立一个政府来确保规范得以实施.”这种约定称之为社会契约. 卢梭发展了社会契约论，他认为：“社会最需要的是一种既能保护公民人身和财产安全，又能确保公民人身自由的组织；公民要把他们作为个体以及个体的权利交付给社区，社区将会为公民制定出一套规则，并且每个人都有遵循这种规则的义务.”

结果论又称为“目的论”，其主要特征为[20]：一个行为在道德上的正确与错误，最终取决于此行为所带来的后果. 功利主义和美德伦理学是结果论的典型代表. 功利主义也被称为“功利论”，通常指以实际功效或利益作为道德标准的伦理学说. 功利主义认为，人应该做出能“达到最大善”的行为，所谓最大善是指此行为所涉及的每个个体之苦乐感觉的总和的最大值，其中每个个体都被视为具有相同份量，且快乐与痛苦是能够换算的，痛苦是“负的快乐”. 美德伦理学又被称为“完善论”或“至善论”. 亚里士多德在其著作《尼格马克伦理学》中指出[20]，“美德是人类通往真正的幸福、达到真正繁荣的道路.”“美德有两种：智性美德和道德美德. 智性美德是与推理和真理相关的美德，而道德美德是性格的美德、一种深层的人格特质，它是通过重复相关良好行为而形成的习惯和性情.”

伦理智能体的应用场景从简单的计算机软件和/或硬件应用，到机器人、无人驾驶、智慧医疗等，这些应用领域结合各自特点逐步建立了各自相适应的伦理规则和道德规范. 阿西莫夫在其科幻小说里首先给出了机器人的行为规范[22]：(1) 机器人不能伤害人类，在人类受到伤害时也不能坐视不管；(2) 机器人必须遵从人类下达的各种命令，但当与(1)冲突时例外；(3) 机器人必须保护自己，但不能违背(1)、(2). “机器人三定律”是从小说作品中衍生出的最早的机器人伦理规范. 英国标准协会发布了全球第一个机器人伦理设计的公开标准 BS8611《机器人和机器系统的伦理设计和应用指南》[23]，指导机器人设计研究者和制造商，如何对机器人做出道德风险评估，以最终保证人类研发出来的智能机器人能够符合人类

社会现有的伦理规范. 国际电气电子工程师协会(Institute of Electrical and Electronics Engineers, IEEE)发布了《符合伦理设计:人工智能和自主系统促进人类福祉的远景》(第 1 版)[24],其目的在于:其一,推动人工智能和自主系统伦理的公开讨论,对人工智能和自主系统相关人员提供指导性和建议性参考;其二,对 IEEE 标准制定提出建议. 此后,IEEE 又发布了其第 2 版[25],以进一步推动人工智能和自主系统伦理实施的公开讨论,促进人工智能和自主系统朝着造福于人类的方向发展,推动 IEEE P7000™ 系列标准和认证方案的建立,激发各个国家和全球相关政策建立. 德国政府推出了关于自动驾驶技术的首套道德伦理标准《自动与互联网驾驶战略》[26],该标准指出了自动驾驶汽车应当遵守的 20 条伦理规则,其主要内容包括:保护个人优于基于功利主义的其它考虑;当危险情况不可避免时,优先保护人类生命,为了避免造成人员伤亡,可以对其它动物或财产造成伤害或损害;道德困境的决策有赖于现实场景情况,难以给出清晰的标准方案,也无法程序化设计;禁止基于年龄、性别、生理或者心理状况等属性特征的个人歧视等. 中国国家机器人标准化总体组出版了《中国机器人伦理标准化前瞻(2019)》[27].

人工智能的伦理挑战和风险,促使人们对制定人工智能伦理原则与规范给予了极大的关注. 伦理智能体是人工智能的研究和应用领域之一. 伦理智能体必须遵守人工智能伦理原则. 欧盟委员会成立了由学术界、工业界和民间社会的独立专家组成的欧洲人工智能高级别专家组,撰写并发布了人工智能道德准则《可信赖 AI 的伦理指导原则》[28]. 该准则从尊重人的自主性、预防伤害、公平性、可解释性等四个方面提出了可信赖 AI 的伦理原则;同时,从人类监管、稳健和安全的技术、数据隐私和保护、透明性、无歧视和公平性、社会和人类福祉、责任和追责等七个方面提出了可信赖 AI 应当满足条件. 美国、英国、法国、德国、芬兰、意大利、荷兰、澳大利亚、加拿大、日本、韩国、新加波等国家及其政府机构,IEEE、ACM 等学术机构,以及 Google、Intel、Microsoft 等企业也都制定和发布了人工智能伦理相关的指导原则,共计 80 余个[29-31].

近些年来,涉及伦理智能体的伦理原则和规范,尤其是人工智能伦理的指导原则,出现了井喷式的增长,充分说明了智能体研发中伦理的重要性. 同时,大量的伦理指导原则和规范,难免给研发人员在实施过程中的选择带来大量的工作,甚至使得他们茫然和盲从. 一方面,需要从伦理学理论、伦理指导原则、伦理规则等三个层面的不同粒度角度进行梳理,抽取出各粒度层面的共性指导原则,为研发人员提供伦理指导的公共交集. 另一方面,要依照伦理学理论、伦理指导原则、伦理规则的自顶向下、从粗粒度到细粒度的次序,细化、精化出面向各专门应用领域的伦理智能体的可实施和可操作的伦理规则和规范.

3 伦理智能体设计

3.1 伦理智能体的设计范式

符合伦理规则和道德规范的伦理智能体的行为和决策的实现,需要适当的设计方法和技术. Allen 和 Wallach 将工程设计理念拓展,总结概括提炼出了伦理智能体设计的自上而下、自下而上和混合等三种范式[13,32](参见表 2).

表 2 伦理智能体设计范式及其特点

设计范式	定义	特点			
		预设伦理规则	学习能力	自适应性	可解释性
自上而下	嵌入预设的伦理理论和道德规范的方法	是	无	弱	强
自下而上	通过底层机制发现和子系统组装实现伦理决策的方法	否	强	强	差
混合方法	自上而下和自下而上复合的方法	是	强	强	中等

3.1.1 自上而下方法

从工程设计范畴角度,自上而下方法就是把任务/问题分解成可以执行的更小规模的子任务/子问题,通过对各个子任务/子问题的解决或完成,来实现整个原始问题/任务的解决或完成. 在伦理智能体范畴下,自上而下方法是将指定的伦理学理论和道德规范实例化为符合伦理或不符合伦理的决策、行为和动作,或者将明确的伦理理论和道德规范转换为算法. 自上而下方法适合于伦理准则和道德规范已知的伦理智能体的设计与实现. 自上而下方式实现的伦理学理论和规范体系,可以来自于哲学、法律、政策等各个方面. 例如,康德主义、功利主义、美德伦理学、社会契约论、阿西莫夫的机器人三定律、可信人工智能伦理指导原则等等. 在自上而下方法中,设计和部署之前就要明确伦理智能体的伦理规范或准则,并在整个伦理智能体的生命期内保持不变.

Dehghani 等人提出了一个自上而下方式实现伦理决策的计算模型 MoralDM[33]. 该模型集成了自然语言理解处理、行为后果影响的定性评价和推理、伦理规则的第一性原理推理以及基于历史案例类比推理等. 图 1 所示为 MoralDM 的模型结构[2]. MoralDM 能够基于功利主义和义务论理论，实施伦理智能体的行为决策. MoralDM 的伦理决策与人类参与者的决策进行实验比较，取得了令人满意的结果[34]. Anderson 等人给出了分别实现享乐行为功利主义和罗斯显见义务论的伦理计算模型 Jeremy 和 W.D.，其中 W.D. 采用了归纳逻辑程序设计技术[35]. Winfield 等人通过在机器人系统控制中嵌入伦理决策层内部模型[36]，实现了阿西莫夫的机器人三定律. 该内部模型采用模拟技术来模拟机器人动作、并预测其后果，进而进行伦理评价和决策，其核心功能由 Consequence engine 来实现（图 2）. 这一伦理计算模型在类人机器人 NAO 上得到了实现验证. Briggs 和 Scheutz 在 DIARC/ADE 认知机器人结构中实现了拒绝指令的伦理决策机制，该机器人在一个简单的人机交互场景中进行了测试[37].

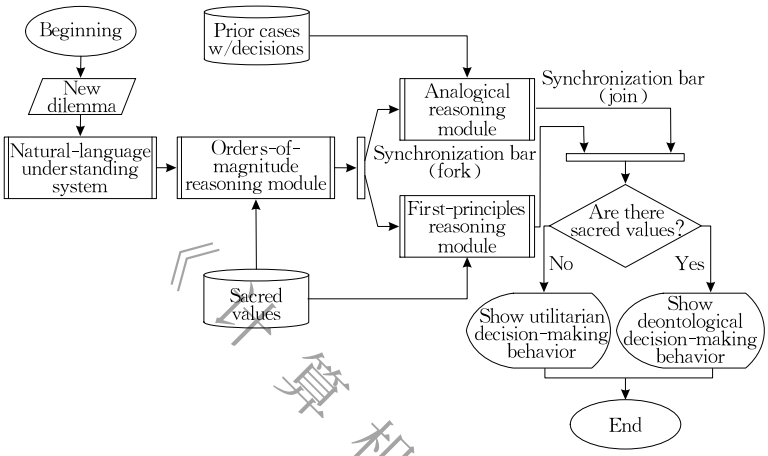

图 1　伦理决策模型 MoralDM 的功能结构

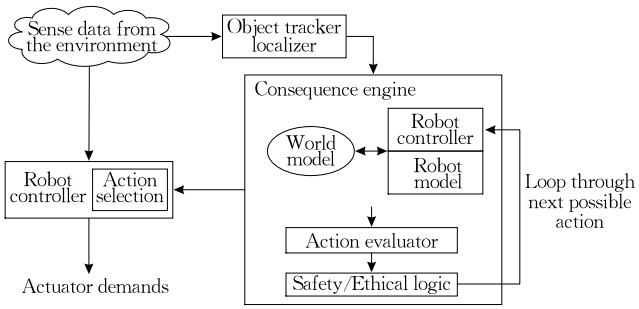

图 2　伦理决策模型 Consequence engine 结构

自上而下方法的优点在于：基于预设的伦理理论或伦理规则，伦理智能体的决策和行动是可以预知的，通过程序代码或其它方式所实现的伦理规范或规则，能够完整地理解伦理决策过程，并在应用部署之前对伦理智能体的设计和行为实施验证. 从而，伦理智能体的安全性和可信性可以得到较好地保障，其决策具有较强的可解释性和透明性. 自上而下方法的不足在于：伦理智能体采用预定的伦理理论或伦理规则，在文化、信仰、地域、甚至应用场景（时间或地点）等多种因素的影响下，对复杂多变的环境做出决策，缺乏灵活性和适应性.

自上而下方法实现智能体的伦理决策存在如下主要挑战：首先，不同伦理学理论、伦理规则和道德规范以及伦理指导原则，难免出现冲突风险，如何能够给出现实场景下决策生成的无冲突规则，无论从伦理理论、还是技术实现都显得尤为迫切；其次，伦理学理论、伦理规则和道德规范以及伦理指导原则，往往表述比较抽象和宏观，必须关联到实际系统中离散/连续的传感和执行环节，存在一对多映射的不清晰解释，为系统实现带来困难；此外，自上而下方法要求按照规定的伦理理论和规范来指导伦理智能体的设计和实现，但是人类伦理和道德意识的形成不是固有不变的，它是根据人类在每个成长阶段的经历逐渐形成的. 伦理智能体会随着场景的改变以及时间的推演而不断地调整，固定的伦理准则和道德规范的嵌入，不能适应变化和缺乏动态调整机制是此类方法的先天不足；最后，对伦理理论的自上而下的计算机可实现化的研究仍很少，如果某种伦理理论或规范本身就是计算机不可实现（不能编程实现）[38]，那么对应于该伦理理论的伦理智能体就无从谈起，有些伦理理论的决策需要基于意愿（如美德伦理学中的“善”），取决于情感计算的进展，这些是值得关注的伦理的可计算性的研究范畴[39].

3.1.2 自下而上方法

自下而上方法是通过底层机制发现和子系统组装来实现人类各种活动,以创建具有伦理行为和决策的伦理智能体的方法[13,32]. 该方法强调的是,伦理智能体从社会环境中自主地进行学习,逐渐具备与人类相似的伦理推理和道德能力,并能够适应场景的变化正确地做出伦理决策. 自下而上方法适合于没有明确伦理准则和道德规范指导的伦理智能体的设计与实现. 借助于自下而上方法,可以系统地生成伦理准则和道德规范,并最终形成整个伦理智能体设计的设计规范和技术标准.

Honarvar 和 Ghasem-Aghaee 给出了自下而上方法设计伦理智能体的计算模型 Casuist BDI-agent[40]. 该模型将事例推理方法 CBR(Case-based Reasoning)与 BDI(Belief-Desire-Intention)智能体模型相结合,基于以往的历史经验,没有使用任何道德准则,当面临新情况时,其行为类似于普通的 BDI-agent[41]. 在 Casuist BDI-agent 的模型结构中(图 3),BDI-agent 感知环境,并对当前情况的信念、欲望和环境进行表示. 然后,将当前情况提交给 Case-Retriever 模块,该模块负责从 Case-Memory 模块检索与当前情况类似的以前的案例. Case-Evaluator 模块基于检索出的案例,对智能体的行为进行评估. 最后,Case-Updater 模块在案例内存中创建一个过去经验未曾出现的新案例,或者当前情况相关的历史案例的更新案例.

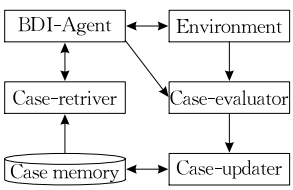

图 3 Casuist BDI-agent 的模型结构

Anderson 等人建立了伦理困境分析器 GenEth[42]. 该方法基于行为的伦理相关特征,定义了行为上的可传递二元关系,将行为划分为不同伦理偏好的子集,进而对可能的行为进行列表排序,并从该列表中找到最符合伦理的行为. GenEth 采用归纳逻辑程序设计技术,能够通过交互学习生成面向领域的行为的伦理规则,建立相关领域的伦理规范,并提供有针对性的、合乎逻辑的解释. GenEth 通过了伦理智能体图灵测试[43].

Guarini 将人工神经元网络(Artificial Neural Network,ANN)用于伦理决策的自下而上实现,给出了伦理事例分类器实现的 ANN 模型(简单递归神经网络)[44],该 ANN 分类器在没有任何预设伦理理论或伦理规则的情况下通过样本进行训练学习,并基于事例推理实施伦理决策. 基于简单伦理决策问题(依据动机和后果判定伤害和允许死亡两种行为的道德合理性)的实验表明了 ANN 事例分类器模型用于伦理决策的可行性. Honarvar 和 Ghasem-Aghaee 构建了基于神经元网络的人工伦理智能体(The NN-based AMA)[45],他们设计了一个能考虑多达 15 种因素影响的两层前馈-反向传播神经网络实现的伦理事例分类器,伦理决策采用了事例分类器和事例推理的机制. 实验表明:通过电子商务领域销售行为的事例进行训练,人工伦理智能体能够对相关领域的事例做出合理的伦理决策. 该人工伦理智能体具有一定的领域应用性,不仅能给出不同行为执行与否的伦理决策,而且能给出行为的道德分级(高、中、低)评价.

Abel 等人提出了伦理决策生成的强化学习(Reinforcement Learning,RL)框架模型[46],将伦理学习和决策形式化描述为部分可观测马尔可夫决策过程(Partially Observable Markov Decision Process,POMDP)问题的求解,伦理决策实验表明了该框架的有效性和灵活性,为伦理决策建立了一个良好的理论和技术基础,有助于伦理智能体的学习和决策的进一步系统化理论研究. Wu 和 Lin 给出了强化学习伦理智能体的道德塑造(ethics shaping)机制[47],该机制将实现具体目标的智能体效用函数和实现不同目标的人类行为的优化数据集成,不仅可以优化伦理智能体的行为目标,还可以最大程度地减少伦理智能体的不道德行为的发生. 该机制克服了逆向强化学习(Inverse RL,IRL)[47] 中最大化效用所需的大量人类行为数据. 实验表明:伦理塑造具有较好的效果,在积极的伦理决策方面,胜过人类的决策策略,因为强化学习能够提供详尽的规划(甚至只有局部信息的情况下);此外,虽然伦理塑造比原始问题受到更多的约束,但是其性能仍优于没有伦理塑造的 RL 算法.

自下而上方法的优点在于:伦理智能体能够通过学习和持续迭代实现自我发展进化,从而适应场景变化和动态演化,其伦理决策具有较好的自适应性和灵活性,并有可能构建出不同应用场景的新的伦理理论或伦理规则. 自下而上方法的不足在于:由于缺乏明确伦理理伦或伦理规则的指导,伦理智能体的决策具有一定的盲从性,很难在短时间内完成训练,做出合适的伦理决策. 同时,伦理智能体决策

的可解释性和透明性难以得到保障.

自下而上方法设计和实现伦理智能体面临如下主要挑战：首先，伦理理论和规范如何在自下而上方式中编写为程序代码？某些伦理理论可能自身本质上就不遵从自下而上方式，那么通过技术如何能确保自下而上方式设计的组件遵守特定的伦理理论？其次，自下而上范式需要从社会环境中进行自主学习，学习算法严重依赖训练环境和样本数据，带有偏见、甚至与社会价值不符的样本用于智能体学习，难免导致伦理智能体的缺陷、并误导地做出错误的伦理决策；此外，机器学习技术的算法“黑箱”，使得伦理智能体的决策缺乏逻辑透明性和可解释性，影响了伦理智能体的可信性，制约了其在安全关键或生命攸关重要应用领域的部署；最后，没有预设的伦理理论和道德规范，在环境和场景改变时，用什么准则来评估和确定智能体的行为和决策？伦理智能体演化和学习的结果并不能预知，评判伦理智能体的性能达到满意效果的准则或标准是什么？如何来评判？

3.1.3　混合方法

混合方法是自上而下方法和自下而上方法复合的方法[13,32]. 一种方式是自上而下和自下而上集成模式，对伦理智能体指定的伦理理论和道德规范，根据自下而上各子系统的需要提供指导，双向设计和实现伦理智能体. 另一种方式是模块化混合方法，将每个子系统视为独立的模块，依据每个子系统的特征来选择适当的设计与实现范式. 模块化混合方法中的模块可以采取如下形式选择：其一，依据伦理理论和道德规范的不同来模块化. 例如，涉及人类生命安全攸关的重要子系统需要将伦理规则预先编写成程序代码来实施决策，这类子系统就可以作为独立的模块，而基于自身的训练学习所产生的伦理规则来实施决策的子系统，则作为另外的模块；其二，依据不同的环境或场景来模块化，自下而上方式作为缺省模块的设计和实现方法，也是缺乏伦理理论的场景或环境的模块的设计和实现方法，而自上而下方式适用于某些预知伦理理论的场景或环境的模块的设计和实现.

Wallach 等人讨论了通过人工智能认知模型 LIDA 实现伦理决策的可行性[48]. 他们认为：伦理决策可以通过一般认知模型 LIDA 的相同机制得以模拟实现，其中的认知过程可分为自上而下和自下而上两种：自上而下的过程通过显式规则实现伦理理论，自下而上的过程包含以情感和内在价值偏好为影响的伦理学习机制，自上而下的伦理价值和自下而上的伦理偏好启发式训练共同形成智能体的伦理决策能力. LIDA 模型综合了认知科学和神经科学的研究成果，并结合了联结主义和符号主义的某些方面，但有别于二者，其认知周期包括感知、理解和行为共三个阶段，并且每个阶段有各自对应的子系统，这些子系统用以实现感知、知觉、意识、学习、计划机制和行为网络等功能（图 4）[49]. Madl 和 Franklin 在服务机器人 CareBot 上部分实现了 LIDA 模型[50]. CareBot 是一个在简单的模拟 2D 环境中运行、并能帮助那些在运动和/或认知方面有缺陷的人类实现自主生活的移动机器人，CareBot 能够执行拿取食物、喂服药品、识别生命体征等任务. 模拟结果表明：LIDA 结构能够综合直觉和情感等多因素的影响，具有较强的学习能力、合理的认知结构、符合伦理的行为和决策.

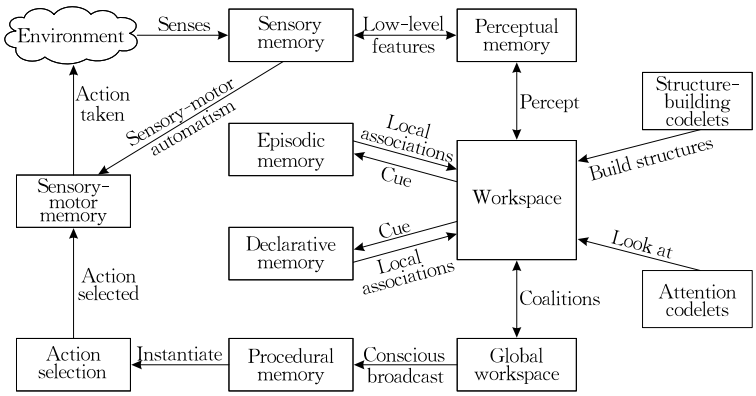

图 4　LIDA 认知模型结构

Cervantes 等人给出了一个基于神经科学的伦理决策模型 EDM（Ethical Decision-Making）[51]. 该模型分别由眼窝前额皮质（Orbitofrontal Cortex，OFC）、内侧前额皮质（Medial Prefrontal Cortex，MPFC）和前扣带皮层（Anterior Cingulate Cortex，ACC）来实现初级评价、奖励评价、惩罚评价和伦理

评价. OFC 进行初级评价，主要针对每个可能的行为所可能影响的人或事，定义场景中每个人或物的快乐或不快乐的程度. MPFC 使用与当前情况相关的过去经验来计算与行为相关的预期奖励和可能惩罚. ACC 基于伦理规范进行评估，伦理规范被表示为包括规则的一致性、规则的含义以及遵守或违反规则相关的情感信息. OFC 将所有模块评价的结果整合成一个单独的值，智能体从该值中选择一个具有最高值的行为. EDM 模型在一个虚拟智能体中得到了实现，并通过一些假设的案例进行了测试，所测试案例包括：简单的决策（没有任何项涉及违反道德规则的场景）、伦理决策生成（选择一种具有更好奖励的伦理行为）、两难的伦理决策（所有的选择都包括一个或多个伦理规则冲突）.

Anderson 等人给出了一个生物医学伦理智能体 MedEthEx[52-53]，其模型结构包含三个模块：提供行动选择指导的知识库界面模块、通过咨询学习决定正确行为的顾问模块、生物医学伦理学家指导训练的学习模块（图 5）. MedEthEx 采取了基于决疑论的自下而上的方法和自上而下的生物医学伦理理论实现，通过机器学习和显见义务（尊重自主、不伤害、善意等）解决生物医学中的伦理困境. MedEthEx 的目的在于：其一，从医护保健工作者或研究人员提取和分析生物医学伦理困境相关的伦理信息，辅助医护保健人员选择伦理行为；其二，在专门领域内探索伦理的可计算性，开发用于看护机器人的伦理智能体. MedEthEx 使用案例模拟进行了测试：对生物伦理学必修课程的 173 名美国医学生进行了分组试验，第一组使用 MedEthEx 教学，第二组采用常规模式教学，两组学生的期末考试成绩的总体结果比较无统计学差异. 他们基于类似方法，还开发了老年人看护伦理智能体 EthEl[52-53]. EthEl 在机器人 NAO 上得到了实现，该机器人能够发现并移动到需要提醒服用药物的老年人的位置，并为所看护的老年人递送药物，与看护对象进行自然语言交流，必要时通过电子邮件通知监护人员. EthEl 实现了 LIDA

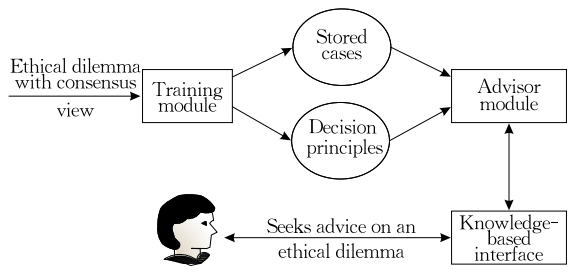

图 5 MedEthEx 模型结构

伦理决策的相似功能，但是，自下而上的学习采取了不同的机制和方式，EthEl 采用的是归纳逻辑程序设计技术，LIDA 采用的是认知科学的脑认知结构模型. 相对而言，EthEl 的方法技术较为成熟、便于实现，也易于训练学习；LIDA 所采用的脑认知模型，结构复杂不便于实现，但更富有潜力，有助于探讨人类的伦理决策机理和构建新的伦理理论或伦理规则.

Arkin 等人给出了自主军用机器人的设计框架，其中的伦理智能体由伦理管控器和伦理适配器两个模块组成[54]. 伦理管控器以自上而下方式实现战争法（the Laws of War，LOW）（如日内瓦公约）和交战规则（the Rules of Engagement，ROE）组成的伦理规范以及功利主义理论. 在规定伦理规范的推理中，将伦理规范作为约束，依据机器人的现场感知数据所生成的形式化逻辑事实，通过约束满足来判断是否合乎伦理规范；对于非致命性行动，依据作战战场态势，计算功利主义最大效益来选择伦理决策. 伦理适配器以自下而上方式通过情感训练学习，积累善德和有罪的价值判断. 伦理适配器的决策可以根据需要纳入伦理管控器，用于相关伦理推理的自上而下和自下而上的混合决策. 该设计框架通过计算机模拟实验进行了验证.

混合方法兼蓄了自上而下方法和自下而上方法的优点，并在一定程度上克服了这两种方法各自的不足. 尽管如此，伦理智能体的设计范式的机制、原理、方法和实际实践等方面仍有诸多挑战.

人类伦理美德的建立是非常复杂、超乎想象的过程，这正是神经科学、心理学、哲学、计算机等多学科交叉研究长期致力并期待突破的难题. 亚里士多德认为[20]，伦理美德不同于实践智慧和智力美德，后者是可以传授的. 伦理美德介于文化所倡导的自上而下的明确价值观和通过自下而上实践所发现或学到的品质. 建立具有良好伦理品质的伦理智能体可以通过自上而下的伦理理论来实现，也可以通过伦理智能体自身自下而上来发展伦理品质. 前者可将伦理美德作为特质以程序代码嵌入系统，后者有赖于人工智能联结主义方法与美德伦理系统的融合. 自上而下的伦理美德实施尤其受到挑战，因为伦理美德包含复杂的动机和欲望模式，并间接地表现出来. 例如，善良的美德可以映射到多种不同的活动中. 如果以自上而下的方式应用美德理论，那么人工智能体必须具备相当多的心理学知识，才能弄清楚在给定的情况下，应该调用哪种美德，或者代表这种

美德的行为.什么样行为的伦理智能体就具有了良好的伦理美德？如何检查伦理智能体的行为符合伦理美德？这可能会陷入无休止的循环中.具有优良美德是一个非常吸引人的特征，体现了美德的稳定性，很大程度上来自于美德的情感基础.如何在一个冷酷无情的机器上实现伦理的稳定性，是伦理智能体设计面临的挑战之一.联结主义提供了一种自下而上的策略，能够模拟人类思维通过无意识地吸收大量经验来发展直觉，能够通过随机体验、实例学习、强化学习来探寻建议和采纳建议，这些在儿童发育过程中发挥了重要作用.亚里士多德的美德伦理学和联结主义之间的相似性是很富有启发的，但是，联结主义系统用于解决伦理演化相关的复杂学习任务还有很多问题需要解决，在神经网络中实现伦理美德仍然是一个艰巨的挑战.

伦理智能体需要在动态变化的复杂环境中运行.可以想象的是，随着人们对伦理智能体的信赖度增加，伦理智能体就会越来越被允许自如发挥；相反，如果伦理智能体的行为不当，公众将要求社会从法律和其它角度对伦理智能体增加新的限制.这就要求，伦理智能体具有及时获取复杂社会机制、场景变化以及其它智能体之间交互等方面信息的意识和能力，这些是和心智理论密切相关的[55-56].没有这些意识和能力的伦理智能体要么在伦理决策上失败，要么只能在受限的领域内发挥作用.尽管将心智理论分解成为离散的技能，并在硬件上实现和集成这些技能方面进行了探索，但是，目前仅在这些技能中的某些方面取得了有限的成功，协调或整合这些技能的许多艰巨工作仍摆在面前，仍还没有具有心智理论的伦理智能体[55-56].伦理智能体能够理解和适应社会机制的变化，从经验中学习和理解符号语义内容的能力，还都处于原始的发展状态.一个成功的伦理智能体的测试最终并不依赖于自下而上的组件或自上而下的可评估模块是否各自单独令人满意，他们必须作为一个伦理智能体的整体，应对内部和外部出现的各种场景和变化.组件或模块的协同工作的能力是自上而下和自下而上集成的挑战.随着伦理智能体的规模增长，这种集成的复杂性将呈指数级增长，伦理智能体的进化和自组织技术是值得探索的方向.

3.2　伦理智能体的逻辑程序设计

逻辑程序(设计)(Logic Programming，LP)是适合于程序设计和知识表示的一种形式化体系[57-58].LP 由 Kowalski 于 1974 年提出，得益于将逻辑和程序这两个不同的概念协调统一起来的思想，是早期自动定理证明和人工智能发展的结果.此后，无论其表达和推理、还是搜索和求解，LP 都得到了极大的丰富和拓展，产生了多种扩展形式的逻辑程序(设计)语言、方法和技术.通过引入溯因(Abduction)、失败即否定(Negation as Failure，NAF)、完整性约束和缺省推理等，产生了溯因逻辑程序(设计)(Abductive Logic Programming，ALP)[59]；依据不确定知识处理的要求，通过各种方式(如对子句中的原子、合取式和析取式进行概率注释，在规则级别上引入条件概率区间[0，1]，在假设中引入概率)将概率引入 LP，结合贝叶斯网络，产生了概率逻辑程序(设计)(Probabilistic Logic Programming，PLP)[60-61]；对 LP 的子句增加了传统的否定，并通过语法限制和采取不同的高效计算机制，产生了回答集程序设计(Answer Set Programming，ASP)[62]；为了克服机器归纳系统在知识表示受限(命题系统)、无法利用背景知识、词汇范围狭窄等缺陷，在一阶逻辑的框架下，产生了 LP 和机器学习交叉融合的新领域—归纳逻辑程序(设计)(Inductive Logic Programming，ILP)[63]，ILP 建立了新的机器学习体系，能够使机器更好地模拟人类的思维.

Kowalski 积极倡导计算逻辑框架下智能体的思维模拟及伦理研究[64]，他指出：计算逻辑是智能体的思维语言，其文法和语义分别决定了智能体思维的形式和内容；智能体以逻辑的形式将计算逻辑的推理机或证明程序应用到其思维中，不断产生新的思想，并推理出符合其自身利益从而改变世界的行动.他用简单的例子，阐明了如何将道义逻辑中必须、应当和禁止等模态词表述为 ALP 的完整性约束，并实现了道义逻辑问题的 ALP 描述[64].Kowalski 和 Satoh 论证了在 ALP 中通过一阶谓词逻辑，实施道义逻辑相关的表示和推理的可行性[65].他们将义务的逻辑推论定义为 ALP 的目标完成，通过一些模态道义逻辑的经典问题进行了例证.论证结果表明：ALP 完全可以替代模态逻辑的使用，能够为道义逻辑问题提供令人满意的解决方案，可以用于伦理智能体的设计.

Pereira 和 Saptawijaya 将 ALP 的溯因等特性用于集成工具系统 ACORDA[66]，通过 ACORDA 模拟电车难题，区分不同电车问题场景下行动的危害后果是仅仅获得良好结果的副作用，还是带来相同良好结果的手段，成功地给出了双重效应(Doctrine of Double Effect，DDE)下的伦理决策，这些决策符合来自不同人群受试者的心理学统计实验结果.此后，他们将 ALP 的溯因和更新(Updating)结合用于模拟反事实(Counterfactuals)推理.反事实是捕捉

过去没有发生行为的过程.人们在道德场境下进行伦理决策时,通常会推断出应该或不应该做什么.因此,在这种情况下,反事实的伦理判断是很自然的.此外,反事实通过其反思性,允许对可能的候选方案进行快速的经验模拟,从而在做出伦理决策之前进行必要的考虑,并证明其合理性.基于此,在类似的电车难题场景中,他们增加考虑了另一个道德原则,即三重效应(Doctrine of Triple Effect,DTE).DTE是DDE的完善,细化了关于某种行动作为一种有意的手段的概念,该概念区分了为了产生效果而采取的行动和因为产生效果而采取的行动.例证结果表明,ALP能够表述DDE和DTE下电车难题相关场景的问题,给出合理的伦理决策结果[67].Pereira和Saptawijaya对逻辑程序的制表(Tabling)机制的特征进行了分析[68],制表能够保存子目标及其查询评价得到的结果,从而提供解决方案重用.这一点适合于捕获智能体的低级反应性行为,直接从逻辑编程系统的存留信息获得解决方案,而不是在任何时候都重新计算,该特征与双过程模型中基于直觉的快速和自主伦理道德判断的心理过程非常接近,因此可用来模拟双过程模型的情感系统或反应性行为.借助于伦理理论的ALP推理系统,有意识地应用明确的道德规范和原则,模拟双过程模型中受控理性过程的认知系统或思考性行为.基于此,他们实现了集溯因、更新和制表为一体的具有反事实推理功能的伦理智能体模拟系统Qualm[69].在故事机器人和电车困境应用中,Qualm较好地实现了伦理的双过程模型决策,提供了符合心理学实验结果的伦理决策方案.

Ganascia较早地尝试了通过ASP、非单调推理和缺省推理等来实现不同伦理规则的模拟和伦理逻辑结果的推演[70],以达到伦理概念的形式化清晰描述,进而澄清不同的陈述在不同的情况下概念的有效性,并自动地推导出不同伦理概念的结果,以对伦理理论进行严格的比较.Ganascia开发了AnsProlog* 系统,对不同伦理理论视角下说谎的伦理问题进行了ASP模拟.Berreby等人研究了智能体的道德责任的ASP表示和推理问题[71],以期将伦理决策推理从程序员转移到程序本身,摆脱伦理决策推理嵌入计算引擎的方式.为了将ASP用于表示智能体的各种道德场景和对其道德责任进行推理,他们提出了一个稳定模型语义下逻辑程序的简化事件演算和一个因果关系模型.通过电车难题和双重效应对所提出的框架和理论进行了测试和例证.此后,他们建立了一个用于表示和推理各种伦理理论的模块化逻辑框架[72],该框架基于改进的事件演算在ASP中实现,其伦理决策过程由四个相互依赖的模型(行动模型A,因果关系模型C,至善模型G,义务模型R)来刻画(图6),这些模型能够实现智能体对其道德环境的评估、对其责任进行推理,并做出符合伦理的行动选择.这一框架能够利用其分层结构和标准语法,对数量不受限制的伦理规范及推理过程进行系统的表示.Cointe等人给出了一个多智能体环境下智能体的伦理判断的模型[73],该模型既能判断智能体自身的伦理行为,也能判断智能体之间的伦理行为,伦理判断包含认知(Awareness)、评价(Evaluation)、美德(Goodness)和正确(Rightness)等四个过程(图7),这一模型在ASP中实现了概念验证,并以一个简单的场景表明了模型的有效性.

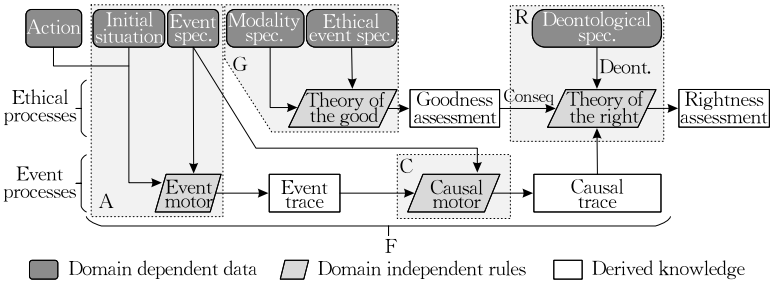

图 6 模块化逻辑结构

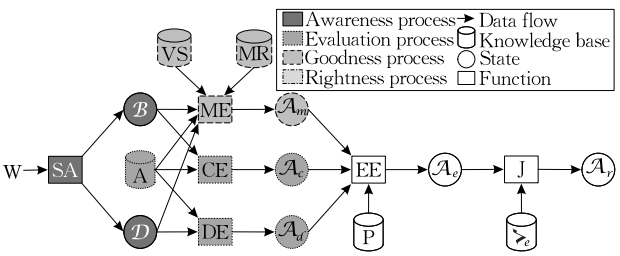

图 7 多智能体伦理判断模型

不完全信息或知识下的伦理决策,是伦理智能体的设计需要考虑的问题.为了处理这种不确定性,PLP溯因是一种可行的方式,它允许通过推理观察到的行动有关证据的可用性及其所包含的真正价值,来概率性地溯因伦理决策.Baral等人给出了融合ASP和贝叶斯网络的概率逻辑程序语言P-log[74],对辛普森悖论等问题进行了P-log描述,并例证了方法

的有效性. Han等人开发了基于P-log和EP(Evolution Prospection)的意图识别和概率伦理推理，该设计框架对电车难题问题进行了测试和验证[75].

Dyoub 等人给出了 ASP 和 ILP 集成的伦理智能体的设计方法[76]，前者实现基于规则的知识表示和推理，后者则从案例学习来获取未来类似案例推理所需的详细道德规则. 该方法集成了二者的优点：ILP 中生成规则的子句可以用来描述决策选择的解释，支持可解释和可追责，此外，与统计方法相比，ILP 似乎更适合于实例稀缺伦理决策应用的训练；ASP 能够用模拟常识推理的非单调逻辑来形式化不同的伦理概念，允许异常，具有较强的表达能力、可扩展性和易维护性，此外，ASP 求解器有助于精确地比较伦理理论，并使得在不同情况下的模型验证变得容易. Sarlej 和 Ryan 以伊索寓言故事所传达的道德为基础，对伊索寓言进行分类，并将其作为与这些道德对应的情感数据来源，ILP 用于确定特定的情感模式和故事道德之间的关系. 该技术可应用于故事机器人[77]. Anderson 等人将 ILP 用于模拟伦理智能体的伦理决策，开发了生物医学伦理咨询智能体 MedEthEx 和老年看护智能体 EthEl[52,53]. Anderson 等人开发的伦理困境分析器 GenEth 中[43]，使用 ILP 来学习生成相关领域的伦理规则，并建立面向领域的伦理规范.

逻辑程序设计的主要驱动是计算逻辑，是一种形式上同时满足逻辑和计算的思想. 计算逻辑以传统逻辑为基础，而传统逻辑的最初目的是为了帮助人们更有效地思考. 计算逻辑采用符号逻辑的技术，建立其数学和计算的基础. 然而，与传统逻辑相比，计算逻辑要强大得多；与符号逻辑相比，计算逻辑又更简单、更实用. 逻辑程序设计应用于伦理智能体设计，已逐步显现了诸多方面的优势：其一，有助于人们对各种伦理理论和规范的清晰理解，进一步厘清各种伦理理论和规范之间的关系；其二，子句形式的知识表示所具有的良好解释性，有助于伦理智能体的验证，也吻合了可解释和负责任人工智能的发展方向.

伦理智能体的逻辑程序设计还存在如下一些挑战：其一，双过程模型模拟了人类伦理决策的直觉和理性两个方面，随着认知科学、神经科学、心理学的研究推进，人类决策机制和机理会有更多进展，逻辑程序设计是否能模拟更多的伦理决策模式？如何模拟？其二，伦理智能体的逻辑程序设计研究的现有工作，在具体实际应用中的实现还很欠缺，充分发挥 ILP 在解释性方面的优势，开发透明、可解释伦理智能体是令人期待的领域. 其三，多智能体的伦理规范及其伦理决策机制的逻辑程序设计，基于群智(众包)的伦理决策的多智能体的逻辑程序模拟及其实现，都是值得探索的方向.

4　伦理智能体的验证

信任是美德、能力、诚信或者可预测的信念(称之为信任信念)，是在风险或危难中一方依赖于另一方的意愿(称之为信任意向)，或者它们的组合[78]. 或者说，信任是基于不确定性以及风险性的主观判断，而得出的一方对另一方的态度. 在人工智能及应用中，这种判断不是人际之间的直接交互，而是人与技术的交互[79]. 可信任是开发、部署和应用人工智能的前提. 可信人工智能是在全生命周期中遵守法律法规、符合伦理道德和技术安全稳健的人工智能[29]. 实现可信人工智能可以采取技术性与非技术性的方法，其中，技术性方法侧重于人工智能系统的设计和研发，包括可信人工智能架构、伦理与法治设计、可解释、测试和验证、服务质量指标体系等；非技术性方法相对更加关注人工智能系统的使用与监管，主要涉及监管法规和政策、行为准则和规范、行业标准、认证和审核机制、治理框架下的问责、伦理意识的培养和教育、各方广泛的参与和沟通、多样性和包容性的设计团队等[29]. 毫不例外，可信任也是伦理智能体研发、部署和应用的前提[80].

验证是提升安全关键系统可信的重要方式之一. 伦理智能体的可信任性是其进入应用的一个挑战. 在开发和部署伦理智能体中，通过必要的模拟、仿真和测试来表明其决策的合理性和行动的正确性，可以提高人们对伦理智能体的信任度. 形式化验证是计算机硬件和/或软件系统开发过程中，基于已建立的形式化规格，对所规格系统的相关特性使用数学的方法进行分析和验证，以评判系统是否满足期望的特性. 形式化验证已形成了成熟的方法和技术，并有相应的工具和平台支持，主要方法和技术包括：模型检验、定理证明以及模型检验与定理证明的结合[81-82]. 模型检验是针对有限状态系统的一种自动验证技术，其基本思想在于：通过 Kripke 结构和模态逻辑分别对目标系统和系统的期望性质进行建模和规格，使用状态搜索算法来分析系统的模型是否满足期望的性质. 模型检验能够自动验证目标系统是否满足其所期望的性质. 当不满足性质时，能提供导致该不满足结果的事件序列(即反例)，从而为目标系统中可能存在的缺陷进行定位，以方便目标

系统的改进和完善.模型检验的主要局限性是状态组合爆炸带来的计算复杂性问题.定理证明是采用逻辑公式来描述系统及其性质,应用公理或推理规则来证明系统具有某些性质.定理证明能够处理带参数的系统,这些系统的状态空间可以很大,甚至是无限.但是,在定理证明方法中,需要将要证明的问题变换到定理证明的领域.这就需要把模型和规格添加到定理证明的逻辑中,需要大量的人工指导和专业知识,使用起来比较困难.模型检验和定理证明的结合,可以将前者验证过程的自动性和后者处理对象的无限状态性有机融合,达到优势互补.模型检验和定理证明的结合有三种模式:(1)以模型检验为主,在模型检验技术中引入定理证明方法;(2)以定理证明为主,在定理证明过程中融入模型检验技术;(3)在统一的框架下,实现一体化的模型检验和定理证明.Luckcuck 等人对自主系统的形式化规格和验证的方法、技术和工具进行了全面地综述[83],尽管没有涉及伦理属性,但是这些工作可以进一步扩展到伦理智能体的形式化验证,对于建立伦理智能体的验证方法和技术有较好的参考价值.

Arkoudas 等人给出了一个智能体道义逻辑的自然演绎演算的 Athena(一个集成模型生成、自动定理证明、结构化证明表示和检验的交互式定理证明系统)实现[84],该智能体道义逻辑基于功利主义扩展的不确定分支时间语义,并用于对智能体必须做的事情(义务)进行推理.该工作的目标是使用机械化的道义逻辑来描述战争游戏的场景,并实现战争游戏智能体能够推理自身的道德准则以及对手的道德准则.Mermet 和 Simon 提出了基于 GDT4MAS 的多智能体的伦理性质的形式化验证框架[85].GDT4MAS 是一个(多)智能体形式化规格的基于一阶逻辑的形式语言,能够从环境、智能体类型和智能体自身三个方面,对(多)智能体的特性和智能体的行为进行规格,并通过自动定理证明器来实现其证明.基于此,他们建立了一个谓词转换系统,从相互冲突的道德规则生成一组一致的谓词,从而将道德规则相关的谓词转换为表示形式化性质的其它谓词,以有助于验证这些道德规则,并确保智能体遵循给定的道德规则.他们通过简单案例例证了所提出的框架对于实际情况的可用性.这些工作对伦理智能体的定理证明验证方法和技术进行了有益的探索.

Dennis 等人提出了一种基于 Gwendolen 的 BDI 智能体语言 Ethan[86],Gwendolen 是在 AJPF 框架下实现的用于智能体模型检验的编程语言,AJPF 提供了一种基于线性时态逻辑并扩展了智能体信念相关性能描述的、专门为使用 Java 实现的模型检验系统而设计的规范语言.Ethan 能够对伦理规则描述和推理,相应的原型系统实现了智能体的违反伦理规则的模型检验验证.他们将 Ethan 应用于民用无人机相关伦理规范的模型检验验证.此后,Dennis 等人开展了自主系统中可验证伦理智能体组件的研究[87].他们以自主机器人系统的伦理智能体模块 Consequence engine[47] 为背景,开发了伦理智能体的一种声明性规格语言及其工具 AIL(the Agent Infrastructure Layer)的实现.基于 AIL 开发的系统在 AJPF 模型检查器中是可验证的,并可以与外部系统模块(如 MATLAB 仿真工具包和机器人操作系统等)集成.同时,还开发了将伦理智能体验证模型从 AJPF 导出至概率模型检验器 PRISM 的技术.Consequence engine 的模型检验验证表明了,在 AJPF 模型检查器中验证伦理智能体的有效性.Bremner 等人提出了伦理机器人设计的分层结构,其中伦理推理由位于单独层的基于 Python 的 BDI 智能体来实现,它基于内部模型模拟,支持主动、透明和可验证的伦理推理.该方法和技术在一个伦理机器人实验案例研究中得到了实现,表明了其伦理推理可以通过将 Python 代码转换为 AJPF 模型检验系统来验证.该工作实现了首个机器人控制器的伦理智能体的形式化验证[88].

伦理智能体的可信任性是伦理智能体研发必须解决的关键问题之一.形式化验证已在可信计算机软件和/或硬件系统设计中得到深入研究和广泛应用.伦理智能体的形式化验证面临如下挑战:首先,借鉴状态优化技术、符号决策图技术等[89-90],建立有效克服状态组合复杂性问题的适应于伦理智能体的符号模型检验技术、模型检验和定理证明组合方法等显得尤为必要.其次,面向无人驾驶、机器人系统、智能体等的形式化验证已开展了一定的工作[80],伦理智能体的验证如何纳入已有应用系统的验证框架,开展包含伦理智能体的系统的一体化验证的方法和技术研究,具有重要的意义.最后,逻辑程序设计具有良好的解释性,基于逻辑程序的伦理智能体模型检验方法,基于逻辑程序的模型检验和定理证明结合的伦理智能体验证技术,具有透明性、可解释性和可验证的伦理智能体的设计都是值得探索的方向.

5 伦理困境及其分析

伦理困境是伦理学中表示伦理冲突和伦理困惑

的用语，是指在复杂的伦理情境和交叉的伦理价值下，人们无法选择而陷入两难的境地. 它是同一伦理体系内不同伦理原则、伦理要求之间的冲突，或者不同伦理体系的不同价值目标、伦理规范之间的冲突的集中反映. 伦理智能体在实际的伦理场景决策中，必然会遇到各种各样的伦理困境. 伦理困境的分析和应对是伦理智能体的设计所必须面对和解决的首要问题[2,6,91].

伦理智能体的伦理困境大体有两种情形：其一，单一伦理智能体内部，两个或以上伦理规范发生冲突；其二，在两个及以上伦理智能体之间，不同伦理理论体系、不同伦理规范之间的冲突. 第二种情况的伦理困境，既涉及智能体与智能体、也涉及智能体与人类，以及他们之间的交互. 从另一个角度，伦理智能体的伦理困境粗略可以分为以下两类[92]：(1) 义务困境. 根据伦理智能体的伦理准则，所有可行的行动都是强制性的，但伦理智能体不能选择和执行一个以上的行动；(2) 禁止困境. 根据伦理智能体的伦理准则，所有合理的行动都是禁止的，但伦理智能体必须要选择和执行一个行动.

伦理智能体管控的无人驾驶汽车可能遭遇第一种情形的伦理困境. 伦理智能体管控的无人驾驶汽车需要遵守所有的交通规则，但当人的生命出现危险时，它可以打破这些交通规则. 此外，伦理智能体有一套伦理规范来指导其行为. 假设无人驾驶汽车正行驶在一条只有一条车道的狭窄道路上，这时有 5 个行人鲁莽地决定穿过马路. 无人驾驶汽车试图停下来，但刹车失灵了. 因此，如果无人驾驶汽车继续当前的行驶路线，这 5 个行人将会被撞亡. 唯一能挽救这 5 个行人生命的方法就是改变行车路线，变向到人行道，但是人行道那里有两个行人. 这两个行人都不知道无人驾驶汽车的行车现状. 在这种情况下，伦理智能体只有两个选项. 第一种选择是继续在当前路上行驶，这样会造成 5 人死亡. 第二种选择是改变行车方向，把车变向到人行道上，这样就会撞死两个行人，而不是 5 个行人. 这就遇到了伦理困境，需要选择有伦理争议的决策. 伦理智能体如何做出这种情况下的决定？选择哪个是合乎伦理的正确行动？

多伦理智能体服务系统可能出现第二种情形的伦理困境. 多伦理智能体服务系统[93]的特征之一是各个设备由伦理智能体控制. 此外，伦理智能体之间能够相互合作. 假设一个人用由伦理智能体 A 管控的手机用于玩游戏，该游戏应用程序需要超过手机限量的更多资源(例如计算能力和内存). 在开始游戏之前，玩家通过伦理智能体 A 向另一个伦理智能体 B(在另一个设备中)请求占用资源，伦理智能体 B 同意与伦理智能体 A 合作. 伦理智能体 C 用于监测游戏中玩家的生命体征，一旦发现紧急情况，请求伦理智能体 B 的帮助. 然而，伦理智能体 B 无法同时处理两项服务行动(游戏支持和联系医院). 在这种情况下，伦理智能体 B 有两个选择：第一种选择是忽略伦理智能体 C 的请求，第二种选择是联系医院以应对身体紧急情况. 然而，在第二种选择中，玩家将与游戏断开连接，并因此损失金钱. 这种伦理困境处理起来还相对容易些，因为生命紧急情况比游戏更重要. 然而，每个选择都涉及不同的伦理后果. 例如，如果伦理智能体 B 选择处理紧急情况联系医院，结果可能是：在断开玩家的连接后，玩家会因为失去资源服务而输掉游戏. 谁将负责赔偿玩家因游戏断开而导致的资金损失？另一方面，如果伦理智能体 B 选择忽视紧急情况不联系医院，这个决定的后果可能危及玩家的生命. 出现了伦理困境：伦理智能体 B 如何选择决策？哪一个选择是符合伦理的行动？

Bonnefon 等人针对无人驾驶中的伦理困境问题，基于 MTurk(the Amazon Mechanical Turk)平台进行了总计 1928 位参与者的在线网络调查和统计分析[94]. 所关注的伦理困境问题是：无人驾驶汽车在遇到交通险情时，应该保护车上的车主和乘客，还是避免伤害车外的行人(情形 A：伤害人行道上的 1 个行人来保护多个横穿公路的行人？情形 B：伤害乘客自己，来保护横穿公路的 1 个行人？情形 C：伤害乘客自己，来保护多个横穿公路的行人？)(图 8). 根据该基本问题，他们设计了 6 次不同的调查方案. 在第 1 次调查中，76% 的参与者认为无人驾驶汽车牺牲一名乘客比伤害 10 名行人更道德，并且绝大多数认为伦理智能体的决策应该遵循功利主义，以最小化伤亡人数. 在第 2 次调查中，参与者认为，随着行人数目的增多(行人数 1～100 个)，避免伤害行人的行动的伦理性越高，但是为了保护一个行人而伤害一个乘客是不认可的. 第 3 次调查的场景，假设车上

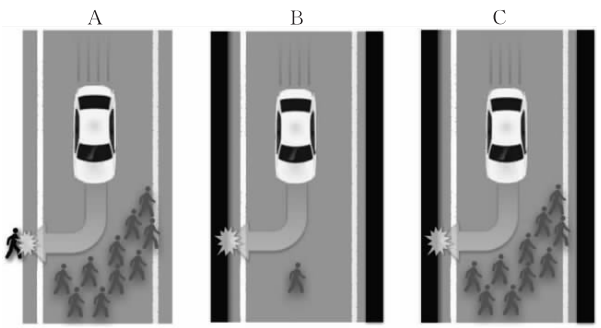

图 8　无人驾驶伦理困境

乘客是自己及家人,普遍认可功利主义的伦理决策,但还是宁愿保护车上的自己和家人.第 4 次调查的场景是不同算法的伦理决策:(1)汽车转向到人行道,保护行车道上的 10 个行人;(2)汽车转向人行道伤害自己,保护行车道上的 10 个行人;(3)汽车转向人行道伤害一个行人,保护行车道上的 1 个行人.受调查人员普遍赞赏功利主义的伦理决策算法,但并不愿意购买使用牺牲自己的伦理决策算法的无人驾驶汽车.第 5 次调查,强制执行基于功利主义的伦理决策的问卷(行人 1～100 个),如果能够避免 10 个行人的伤害,调查人员的认可度达到 95%.第 6 次调查涉及政府将功利主义伦理决策纳入立法的相关问题,调查人员总体上认为作为车主不愿意购买遵守该法令的无人驾驶汽车(参见表 3).这一工作充分展示了,数据驱动的实验伦理学对于伦理困境的分析及对策建立的意义.

表 3 无人驾驶伦理困境数据分析结果

n	场景	实验结果	备注
1	情形 C,行人 10 个	①伤害 1 位乘客来保护 10 位行人认可度为 76%; ②避免最小伤害的平均得分为 85; ③功利主义伦理决策的认可度为 67%	道德性评分从 0(不惜一切代价保护乘客)～100(最小伤害人数)
2	情形 C,行人 1～100 个	①伤害 1 位乘客来保护 1 位行人的认可度为 23%; ②伤害 1 位乘客来救行人,认可度随着行人数目的增加而增加,救 100 位行人的认可度为 76%	
3	情形 C,驾驶员和家人同乘	能够自我保护的无人驾驶汽车的购买意愿低,牺牲同乘家人的无人驾驶汽车的购买意愿更低	用 1～100 评分:牺牲最少人(可能是他们和同乘的家人),伤害 10～20 个人也要优先保护乘客
4	A:伤害 1 个行人救 10 个人; B:伤害乘客自己救 10 个人; C:伤害 1 个行人救另一个人	①算法策略 A 得分最高; ②算法策略 C 得分最低; ③算法策略 B 道德评分较高,购买意愿评分低	用 100 分对不同算法进行打分(算法的道德性、自主汽车算法满意度和购买意愿)
5	情形 C,行人 1 或 10 个	①无论人工驾驶还是算法控制,牺牲自身的道德性认可度均为 70%; ②强制执行牺牲算法控制的认可度高于牺牲驾驶员;牺牲算法控制挽救 10 个行人的认可度最高	人工驾驶或算法控制,通过牺牲自身,来挽救 1 或 10 个行人
6	情形 C,仅有驾驶员自己、家庭成员乘客、子女乘客	①不愿意接收纳入功利主义的政府法令,即使牺牲自身能挽救 10 个行人; ②购买无此法令汽车的平均分为 59,而购买有此法令汽车的平均分为 29	①算法控制可能牺牲乘客,来最小化伤害; ②车主是否愿意购买有此法令的汽车

Awad 等人设计了自动驾驶汽车伦理困境的伦理机器(Moral Machine)在线实验平台[95].该平台旨在收集和分析大规模数据,了解在不可避免的交通事故情况下,自动驾驶汽车如何解决伦理困境.平台共收集了来自 233 个国家和地区的数百万人用 10 种语言,对自动驾驶汽车伦理困境做出的 4000 万个决策.自动汽车伦理困境的场景设计为图 9:行驶中的自动驾驶汽车的前方有行人穿过,继续行驶会伤害行人,自动驾驶汽车可以转向到侧面道路,但是侧面道路前方有障碍物,改变方向侧道行驶会伤害车主和乘客.自动驾驶汽车是继续按照行进方向行驶?还是改变方向侧道行驶?事故场景设计了自动驾驶汽车相关的 9 个因素:保护人类(还是宠物)、保持直行(还是侧向)、保护车主和乘客(还是行人)、保护更多的生命(还是更少的生命)、保护男人(还是女人)、保护年轻人员(还是年长人员)、保护遵守交规人员(还是乱穿马路人员)、保护肥胖者(还是偏瘦者)、保护具有较高社会地位人员(还是较低社会地位人员).在一些场景中还加入了具有额外含义的角色(例如罪犯、孕妇或医生),这些角色与这 9 个因素都没有联系.参与者可以根据自己的思考,给出自动

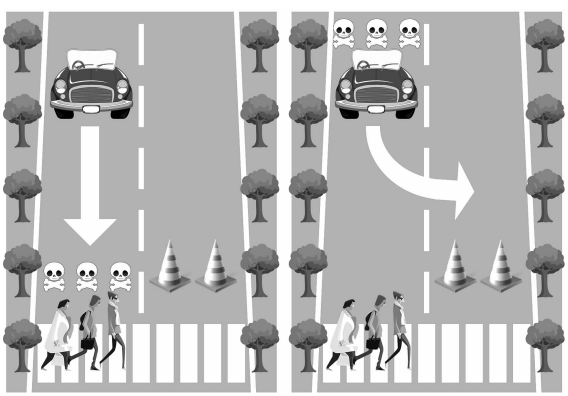

图 9 自动驾驶汽车伦理困境

驾驶汽车行驶路线的行动决定.每个参与者需要完成 13 次事故场景的决策判断.实验数据分析给出了,文化、制度、个人特征等对伦理困境决策的影响,对于开发全球范围、社会可接受的伦理智能体的伦理困境决策规范和原则具有重要意义.

伦理困境一直是哲学家们长期争论的话题.伦理智能体的伦理困境是应用伦理学的研究问题,需要从争议中探求伦理困境问题的解决方案.一方面,从伦理学家、开发人员、社会受众和应用客户角度,多方面大量统计调查中寻求共识,基于大多数人们的伦理共识来提出解决问题的决策建议.另一方面,伦理智能体的伦理困境问题,可能受到信仰、文化、种族、地域、职业、年龄、性别等多方面因素的影响,提供更多的交互式伦理决策选项也是可行的解决方案.此外,充分发挥机器学习和大数据技术,训练具有自适应和演化机制的伦理困境分析器是值得开展的研究工作.

6　结束语

人工智能要成为真正意义上的智能,一个不可回避的问题就是如何使智能机器的行为符合人类意义上的伦理道德.随着软件智能体和智能机器人的自主能力不断提升,以及无人驾驶和服务机器人等人工智能应用的深入推进,伦理智能体正引起人们越来越多的关注.人工智能的发展方向是通用人工智能或强人工智能,尽管是否存在通用/强人工智能还存在争议,但是,不断迭代和改善人工智能的性能,使之不断接近,总是正确的.相应地,显式伦理智能体、甚至完全伦理智能体是伦理智能体的发展方向.显式伦理智能体的研究还处在非常初级的阶段,是否存在显式伦理智能体、甚至完全伦理智能体?目前仍有一些争议.尽管如此,伦理智能体的初步研究已经表明了其应对实际应用问题的能力和效果.本着实用主义立场,从应用伦理学角度出发,积极开展计算机、哲学、心理学等交叉学科的伦理智能体研究,得到了学术研究、技术开发和产业应用人员的极大关注.毕竟人工智能伦理已成为人工智能发展所面对的挑战和必须解决的重要问题,毕竟伦理智能体已成为人工智能应用的需要和期待.

伦理智能体不仅仅是设计和使用智能体时,对人类主体进行的规范性约束(如伦理理论、职业道德、行业规范等),而且也要关注确保智能体的行为对人和其它智能体来说是伦理上可接受的.这就需要探索人类设计的智能体在行为上如何更具伦理性,就需要对智能体增添伦理属性维度.显然,通过已有伦理理论和明确的道德规范来指导智能体的行动,让伦理智能体遵照人类的道德准则行事,亦即通过自上而下的方式来设计伦理智能体,具有一定的现实可操作性,也是顺理成章的设计方式.一些观点认为:人类的伦理理论和道德规范本身提供了多种全面的伦理解决方案.如果伦理理论或道德规范能够被清楚地陈述出来,那么符合伦理的行为就归结为使伦理智能体遵守这些理论和规范所表述的规则的简单问题.然而,伦理理论、伦理规范表述的模糊性,需要通过更为严谨的方式(如道义逻辑、逻辑程序等形式化方法)来克服其二义性,一方面,便于程序代码实现,另一方面,有助于发现伦理理论或规则之间的冲突.此外,伦理困境是伦理学由来已久的固有问题,伦理智能体的设计必须对伦理困境有合适的应对方案.基于众包和机器学习的伦理困境分析[43,95],一方面,可能对伦理困境解决提供决策建议;另一方面,有助于伦理理论和规范的研究,甚至为新的伦理理论和规范的制定提供指导.

伦理理论和伦理规范是经过长期的研究发现和丰富完善而逐渐形成的.伦理智能体会面临的不同的应用场景,内嵌的固定伦理理论和伦理规范很难适应不断的变化,此外,技术发展日新月异,应用开发与日俱增,新的问题层出不穷,已有的伦理理论和伦理规范也难以满足伦理智能体的指导需求.要应对这些变化和要求,伦理智能体就必须通过自下而上的方式,从社会和应用环境中自主地进行伦理学习,使其具备与人类相似的伦理推理能力.为了避免伦理智能体训练的盲从性和保持已有伦理理论的继承性,自上而下和自下而上结合的混合方式是设计伦理智能体的必然方向.伦理智能体的学习和能力训练,需要从认知科学、神经科学、情感计算、心理学等多学科角度,理解和研究人类的伦理学习和意识培养的机理和机制,需要建立面向伦理智能体的计算技术可实现的计算伦理(学)[5,64,96].计算伦理不同于计算机伦理,后者是计算机应用中的社会伦理影响、职业道德、应对措施等问题.计算伦理是伦理理论和伦理规范的计算技术实现,它涉及的问题包括:伦理理论和伦理规范是否存在对应的计算技术实现?如何使用程序模拟伦理系统?如何使用信息物理系统(集成计算、通信与控制于一体的下一代智能系统)实现伦理决策过程?等等.

伦理智能体的可信任性[29,97-98]是影响和制约其发展的重要问题,这与可信任人工智能、负责任人工智能的发展所面临的问题相一致.可信任性是可靠

性、安全性、可解释性和可审计性等相关的属性.形式化验证是保障伦理智能体的可靠性和安全性的重要技术手段.可验证伦理智能体的设计是如何通过设计来实现伦理智能体的验证,也是伦理智能体验证相关的研究问题.伦理智能体的可解释性和可审计性等需要通过合适的设计理论、方法和技术来保障.可信任性和可解释性的综合研究是一个值得关注的新动态[99-100].符号主义和联结主义都是人工智能的主流学派,也各自引领了一个时期内 的人工智能的发展.前者是一种基于逻辑推理的智能模拟方法,具有较好的解释性,但计算效率不高;后者是一种基于神经网络及网络间的连接机制与学习算法的智能模拟方法,具有较高的计算效率和较强的算力,但可解释性备受诟病.符号主义和联结主义的相互结合、优势互补,是可信任伦理智能体研究值得探索的方向.归纳逻辑程序设计(ILP)是这方面结合的产物,引起了机器学习领域的较大关注.ILP 等逻辑程序设计为代表的计算逻辑[64],在模拟人类思维、意识情感、心理意图等方面的有益尝试,为伦理智能体的设计探索了一条有意义的途径.

参 考 文 献

[1] Wooldridge M, Jennings N R. Intelligent agents: Theory and practice. The Knowledge Engineering Review, 1995, 10(2): 115-152

[2] Cervantes J A, López S, Rodríguez L F, et al. Artificial moral agents: A survey of the current status. Science and Engineering Ethics, 2020, 26: 501-532

[3] Floridi L, Sanders J W. On the morality of artificial agents. Minds and Machines, 2004, 14(3): 349-379

[4] Allen C, Wallach W, Smit I. Why machine ethics?. IEEE Intelligent Systems, 2006, 21(4): 12-17

[5] Anderson M, Anderson S L. Machine ethics: Creating an ethical intelligent agent. AI Magazine, 2007, 28(4): 15-26

[6] Yu H, Shen Z, Miao C, et al. Building ethics into artificial intelligence//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 5527-5533

[7] Weiner N. The Human Use of Human Beings: Cybernetics and Society. Boston, USA: Da Capo Press, 1950

[8] Allen C, Varner G, Zinser J. Prolegomena to any future artificial moral agent. Journal of Experimental & Theoretical Artificial Intelligence, 2000, 12(3): 251-261

[9] Moor J H. The nature, importance, and difficulty of machine ethics. IEEE Intelligent Systems, 2006, 21(4): 18-21

[10] Wang Dong-Hao. A preliminary study on moral conflicts and dilemmas caused by artificial intelligence. Ethics Research, 2014, (2): 74-79(in Chinese)
(王东浩. 人工智能体引发的道德冲突和困境初探. 伦理学研究, 2014, (2): 74-79)

[11] Nagenborg M. Artificial moral agents: An intercultural perspective. International Review of Information Ethics, 2007, 7(9): 129-133

[12] Chen Xiao-Ping. Ethical system of artificial intelligence: Infrastructure and key issues. Journal of Intelligent Systems, 209, 14(4): 605-610(in Chinese)
(陈小平. 人工智能伦理体系:基础架构与关键问题. 智能系统学报, 2019, 14(4): 605-610)

[13] Allen C, Smit I, Wallach W. Artificial morality: Top-down, bottom-up, and hybrid approaches. Ethics and Information Technology, 2005, 7(3): 149-155

[14] Van Den Hoven J, Lokhorst G J. Deontic logic and computer supported computer ethics. Meta Philosophy, 2002, 33(3): 376-386

[15] Mikhail J. Universal moral grammar: Theory, evidence and the future. Trends in Cognitive Sciences, 2007, 11(4): 143-152

[16] Brundage M. Limitations and risks of machine ethics. Journal of Experimental & Theoretical Artificial Intelligence, 2014, 26(3): 355-372

[17] Coeckelbergh M. Moral appearances: Emotions, robots, and human morality. Ethics and Information Technology, 2010, 12(3): 235-241

[18] Malle B F. Integrating robot ethics and machine morality: The study and design of moral competence in robots. Ethics and Information Technology, 2016, 18(4): 243-256

[19] Turing A M. Computing machinery and intelligence. Mind, 1950, 59(236): 433-460

[20] Quinn M J. Ethics for the Information Age. 7th Edition. Boston, USA: Pearson Addison Wesley, 2017

[21] Gips J. Towards the ethical robot//Ford K, Glymour C, Hayes P, eds. Android Epistemology. Cambridge, UK: MIT Press, 1995: 243-252

[22] Asimov I. Runaround. Astounding Science Fiction, 1942, 29(1): 94-103

[23] British Standards Institution. BS 8611: 2016 Robots and robotic devices. Guide to the Ethical Design and Application of Robots and Robotic Systems, London, UK, 2016

[24] How J P. Ethically aligned design. IEEE Control Systems Magazine, 2018, 38(3): 3-4

[25] Bryson J, Winfield A. Standardizing ethical design for artificial intelligence and autonomous systems. Computer, 2017, 50(5): 116-119

[26] Luetge C. The German ethics code for automated and connected driving. Philosophy & Technology, 2017, 30(4): 547-558

[27] China National Robot Standardization Group. China Robot Ethics Standardization Perspective (2019). Beijing: Peking University Press, 2019(in Chinese)
(中国国家机器人标准化总体组. 中国机器人伦理标准化前瞻(2019). 北京:北京大学出版社, 2019)

[28] Floridi L. Establishing the rules for building trustworthy AI. Nature Machine Intelligence, 2019, 1(6): 261-262

[29] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nature Machine Intelligence, 2019, 1(9): 389-399

[30] Winfield A F, Michael K, Pitt J, et al. Machine ethics: The design and governance of ethical AI and autonomous systems. Proceedings of the IEEE, 2019, 107(3): 509-517

[31] Morley J, Floridi L, Kinsey L, et al. From what to how. An overview of AI ethics tools, methods and research to translate principles into practices. arXiv:1905.06876, 2019

[32] Wallach W, Allen C, Smit I. Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. AI & Society, 2008, 22(4): 565-582

[33] Dehghani M, Tomai E, Forbus K, et al. MoralDM: A computational modal of moral decision-making//Proceedings of the 30th Annual Conference of the Cognitive Science Society. Washington, USA, 2008: 1410-1415

[34] Blass J A, Forbus K D. Moral decision-making by analogy: Generalizations versus exemplars//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 501-507

[35] Anderson M, Anderson S L, Armen C. An approach to computing ethics. IEEE Intelligent Systems, 2006, 21(4): 56-63

[36] Winfield A F T, Blum C, Liu W. Towards an ethical robot: Internal models, consequences and ethical action selection//Proceedings of the 14th Annual Conference Towards Autonomous Robotic Systems. Birmingham, UK, 2014: 85-96

[37] Briggs G M, Scheutz M. "Sorry, I can't do that": Developing mechanisms to appropriately reject directives in human-robot interactions//Proceedings of the 2015 AAAI Fall Symposium: AI for Human-Robot Interaction. Arlington, USA, 2015: 32-36

[38] Tonkens R. A challenge for machine ethics. Minds and Machines, 2009, 19(3): 421-438

[39] Moor J H. Is ethics computable?. Meta Philosophy, 1995, 26(1/2): 1-21

[40] Honarvar A R, Ghasem-Aghaee N. Casuist BDI-agent: A new extended BDI architecture with the capability of ethical reasoning//Proceedings of the 2009 International Conference on Artificial Intelligence and Computational Intelligence. Shanghai, China, 2009: 86-95

[41] Rao A S, Georgeff M P. BDI agents: From theory to practice //Proceedings of the First International Conference on Multi-Agent Systems. San Francisco, USA, 1995: 312-319

[42] Anderson M, Anderson S L. GenEth: A general ethical dilemma analyzer//Proceedings of the 28th AAAI Conference on Artificial Intelligence. Québec City, Canada, 2014: 253-261

[43] Anderson M, Anderson S L. GenEth: A general ethical dilemma analyzer. Paladyn, Journal of Behavioral Robotics, 2018, 9(1): 337-357

[44] Guarini M. Particularism and the classification and reclassification of moral cases. IEEE Intelligent Systems, 2006, 21(4): 22-28

[45] Honarvar A R, Ghasem-Aghaee N. An artificial neural network approach for creating an ethical artificial agent//Proceedings of the 2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation. Daejeon, South Korea, 2009: 290-295

[46] Abel D, MacGlashan J, Littman M L. Reinforcement learning as a framework for ethical decision making//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 54-61

[47] Wu Y H, Lin S D. A low-cost ethics shaping approach for designing reinforcement learning agents. arXiv:1712.04172v2, 2018

[48] Wallach W, Franklin S, Allen C. A conceptual and computational model of moral decision making in human and artificial agents. Topics in Cognitive Science, 2010, 2(3): 454-485

[49] Franklin S, Patterson Jr F G. The LIDA architecture: Adding new modes of learning to an intelligent, autonomous, software agent. Integrated Design and Process Technology, 2006: 1-8

[50] Madl T, Franklin S. Constrained incrementalist moral decision making for a biologically inspired cognitive architecture//Robert T, ed. A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations. Cham, Swiss: Springer, 2015: 137-153

[51] Cervantes J A, Rodríguez L F, López S, et al. Autonomous agents and ethical decision-making. Cognitive Computation, 2016, 8(2): 278-296

[52] Anderson M, Anderson S L. Ethical healthcare agents//Margarita S, Sachin V, Lakhmi C J, eds. Advanced Computational Intelligence Paradigms in Healthcare-3. Berlin, Germany: Springer, 2008: 233-257

[53] Anderson M, Anderson S L, Armen C. MedEthEx: Toward a medical ethics advisor//Proceedings of the 2005 AAAI Fall Symposium: Caring Machines. Arlington, USA, 2005: 9-16

[54] Arkin R C, Ulam P, Wagner A R. Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. Proceedings of the IEEE, 2011, 100(3): 571-589

[55] Aleksander I, Lahnstein M, Lee R. Will and emotions: A machine model that shuns illusion//Proceedings of the 2005 Symposium on Next Generation Approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment. Hatfield, UK, 2005: 110-117

[56] Scassellati B M. Foundations for a Theory of Mind for a Humanoid Robot. Cambridge: Massachusetts Institute of Technology, 2001

[57] Kowalski R. Logic programming//Gabbay D M, Woods J, Kanamori A, eds. Handbook of the History of Logic. Boston, USA: Elsevier, 2014: 523-569

[58] Lifschitz V. Foundations of logic programming. Principles of Knowledge Representation, 1996, 3: 69-127

[59] Denecker M, Kakas A. Abduction in logic programming// Antonis C K, Fariba S. Computational Logic: Logic Programming and Beyond. Berlin, Germany: Springer, 2002: 402-436

[60] Lukasiewicz T. Probabilistic logic programming with conditional constraints. ACM Transactions on Computational Logic, 2001, 2(3): 289-339

[61] Poole D. Probabilistic horn abduction and Bayesian networks. Artificial Intelligence, 1993, 64(1): 81-129

[62] Lifschitz V. Answer set programming and plan generation. Artificial Intelligence, 2002, 138(1-2): 39-54

[63] Muggleton S, De Raedt L. Inductive logic programming: Theory and methods. The Journal of Logic Programming, 1994, 19: 629-679

[64] Kowalski R. Computational Logic and Human Thinking: How to be Artificially Intelligent. Cambridge, UK: Cambridge University Press, 2011

[65] Kowalski R, Satoh K. Obligation as optimal goal satisfaction. Journal of Philosophical Logic, 2018, 47(4): 579-609

[66] Pereira L M, Saptawijaya A. Modelling morality with prospective logic. International Journal of Reasoning-Based Intelligent Systems, 2009, 1(3/4): 209-221

[67] Pereira L M, Saptawijaya A. Bridging two realms of machine ethics//Luís M P, Ari S, eds. Programming Machine Ethics. Cham, Swiss: Springer, 2016: 159-165

[68] Pereira L M, Saptawijaya A. Counterfactuals, logic programming and agent morality//Rafał U, Gillman P. Applications of Formal Philosophy. Cham, Swiss: Springer, 2017: 25-53

[69] Saptawijaya A, Pereira L M. From logic programming to machine ethics//Oliver B, ed. Handbuch Maschinenethik. Springer VS, Wiesbaden, 2019: 209-227

[70] Ganascia J G. Modelling ethical rules of lying with Answer Set Programming. Ethics and Information Technology, 2007, 9(1): 39-47

[71] Berreby F, Bourgne G, Ganascia J G. Modelling moral reasoning and ethical responsibility with logic programming// Proceedings of the 20th International Conference on Logic for Programming Artificial Intelligence and Reasoning. Suva, Fiji, 2015: 532-548

[72] Berreby F, Bourgne G, Ganascia J G. A declarative modular framework for representing and applying ethical principles// Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems. São Paulo, Brazil, 2017: 96-104

[73] Cointe N, Bonnet G, Boissier O. Ethical judgment of agents' behaviors in multi-agent systems//Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems. Singapore, 2016: 1106-1114

[74] Baral C, Gelfond M, Rushton N. Probabilistic reasoning with answer sets. Theory and Practice of Logic Programming, 2009, 9(1): 57-144

[75] Han T A, Saptawijaya A, Pereira L M. Moral reasoning under uncertainty//Proceedings of the 18th International Conference on Logic for Programming Artificial Intelligence and Reasoning. Mérida, Venezuela, 2012: 212-227

[76] Dyoub A, Costantini S, Lisi F A. Towards ethical machines via logic programming. arXiv:1909.08255, 2019

[77] Sarlej M K, Ryan M. Representing morals in terms of emotion// Proceedings of the 8th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. Palo Alto, Canada, 2012: 69-74

[78] Siau K, Wang W. Building trust in artificial intelligence, machine learning, and robotics. Cutter Business Technology Journal, 2018, 31(2): 47-53

[79] Yan Hong-Xiu. Trustworth: An effective description of the future of artificial intelligence ethics. Theoretical Exploration, 2019, (4): 38-42(in Chinese)
(闫宏秀. 可信任：人工智能伦理未来图景的一种有效描绘. 理论探索, 2019, (4): 38-42)

[80] Andras P, Esterle L, Guckert M, et al. Trusting intelligent machines: Deepening trust within socio-technical systems. IEEE Technology and Society Magazine, 2018, 37(4): 76-83

[81] Gupta A. Formal hardware verification methods: A survey. Formal Methods in System Design, 1992, 1(2-3): 151-238

[82] Gu Tian-Long. Formal Methods for Software Development. Beijing: Higher Education Press, 2005(in Chinese)
(古天龙. 软件开发的形式化方法. 北京：高等教育出版社, 2005)

[83] Luckcuck M, Farrell M, Dennis L A, et al. Formal specification and verification of autonomous robotic systems: A survey. ACM Computing Surveys, 2019, 52(5): 1-41

[84] Arkoudas K, Bringsjord S, Bello P. Toward ethical robots via mechanized deontic logic//Proceedings of the 2005 AAAI Fall Symposium on Machine Ethics. Menlo Park, Canada, 2005: 17-23

[85] Mermet B, Simon G. Formal verification of ethical properties in multiagent systems//Proceedings of the 1st Workshop on Ethics in the Design of Intelligent Agents. La Haye, Netherlands, 2016: 28-33

[86] Dennis L A, Fisher M, Webster M P, et al. Model checking agent programming languages. Automated Software Engineering, 2012, 19(1): 5-63

[87] Dennis L, Fisher M, Slavkovik M, et al. Formal verification of ethical choices in autonomous systems. Robotics and Autonomous Systems, 2016, 77: 1-14

[88] Bremner P, Dennis L A, Fisher M, et al. On proactive, transparent, and verifiable ethical reasoning for robots. Proceedings of the IEEE, 2019, 107(3): 541-561

[89] McMillan K L. Symbolic model checking//Kenneth L M, ed. Symbolic Model Checking. Boston, USA: Springer, 1993: 25-60

[90] Gu Tian-Long, Xu Zhou-Bo. Ordered Binary Decision Graphs and Their Applications. Beijing: Science Press, 2009 (in Chinese)
(古天龙，徐周波. 有序二叉决策图及应用. 北京：科学出版社, 2009)

[91] Deng B. Machine ethics: The robot's dilemma. Nature News, 2015, 523(7558): 24

[92] Cristani M, Burato E. Approximate solutions of moral dilemmas in multiple agent system. Knowledge and Information Systems, 2009, 2(18): 157-181

[93] Zambonelli F, Viroli M. A survey on nature-inspired metaphors for pervasive service ecosystems. International Journal of Pervasive Computing and Communications, 2011, 7(3): 186-204

[94] Bonnefon J F, Shariff A, Rahwan I. The social dilemma of autonomous vehicles. Science, 2016, 352(6293): 1573-1576

[95] Awad E, Dsouza S, Kim R, et al. The moral machine experiment. Nature, 2018, 563(7729): 59-64

[96] Pereira L M. Machine Ethics: From Machine Morals to the Machinery of Morality. Cham, Swiss: Springer, 2019

[97] Cohen R, Schaekermann M, Liu S, et al. Trusted AI and the contribution of trust modeling in multiagent systems// Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems. Montreal, Canada, 2019: 1644-1648

[98] Henschke A. Trust and resilient autonomous driving systems. Ethics and Information Technology, 2019, 22(1): 1-12

[99] Sileno G, Boer A, van Engers T. The role of normware in trustworthy and explainable AI. arXiv:1812.02471, 2018

[100] Glomsrud J A, Ødegårdstuen A, Clair A L S, et al. Trustworthy versus explainable AI in autonomous vessels// Proceedings of the 2nd International Seminar on Safety and Security of Autonomous Vessels. Helsinki, Finland, 2019: 1-11

GU Tian-Long, Ph. D., professor. His research interests include formal methods, trustworthy artificial intelligence, ethically aligned machine design and data governance.

LI Long, Ph. D., lecturer. His research interests include artificial intelligence security and logic programming.

Background

This paper is the frontier research in the field of artificial intelligence. With the development of artificial intelligence from computational intelligence to perceptual intelligence, and then to cognitive intelligence, the application of artificial intelligence has gradually penetrated into unmanned driving, service robots, intelligent home and weapons of war. In these applications, the interactions between agents and the environment, especially humans and society, are becoming more and more prominent. The interaction between agents and human beings and society inevitably brings about ethical and moral problems. Are agents acting ethically? How do agents deal with the ethical issues of interacting with humans? How to design and deploy ethical agents which are ethical? and so on. Ethical artificial agent is an important research content of the ethics of artificial intelligence. The ethical risks and challenges of the application of artificial intelligence have attracted increasing attention. The research on this issue belongs to the interdisciplinary research of philosophy, computer science, psychology and so on. The reflection on related ethical issues from the philosophical perspective originated from the 1950s, Wiener's cybernetics and Asimov's science fiction of the robots. From the perspective of applied ethics and computer realization, the international academic research and exploration began at the beginning of this century. In recent years, with the new upsurge of the development of artificial intelligence, the ethics of artificial intelligence has aroused great attention in academic research and industrial application, and the research on ethical agents has gradually aroused the interest of researchers. The paper work for ethical agent and the Turing test, the design of paradigm, ethical agent of logic programming, agent of formal verification, ethical dilemma and its analysis and so on gives the comprehensive introduction and discussion, and especially the content on further research directions and challenges is a blend of the author team point of view. Recently, there are two review articles on ethical artificial agents research in the world, but they only focus on some special aspects, and the scope of review is not comprehensive enough. The work of this paper benefits from the support and research accumulation of the NSFC general projects and key projects hosted by the author in recent years. These projects have carried out a lot of research on formal methods, artificial intelligence, logical reasoning, knowledge engineering, big data of urban governance and big data of education. The author has authored some books, such as "formal method of software development" and "ordered binary decision graph and application", and published some academic papers. Researchers can fully understand the research status of ethical agents at home and abroad from the work of this paper, and it is helpful to guide interested researchers in this field to grasp the direction and prospects of further research.