World Literature Studies 世界文学研究, 2022, 10(2), 139-145
Published Online June 2022 in Hans. http://www.hanspub.org/journal/wls
https://doi.org/10.12677/wls.2022.102022

Hans汉斯

论科幻小说《金羊毛》中的人工智能主体性

朱晓玲

东南大学外国语学院，江苏 南京

收稿日期：2022年4月18日；录用日期：2022年6月8日；发布日期：2022年6月16日

摘　要

科幻小说《金羊毛》以太空中的一场谋杀案为起点，讲述主人公亚伦在追击凶手过程中与阿尔戈号上人工智能控制系统杰森之间的博弈。通过作者罗伯特·索耶的刻画，杰森展现为具备交互性、自主性和适应性的人工智能主体。其交互性显现为控制船员的工具，自主性成为隐瞒信息的途径，而适应性则是杰森试图保护自身和人类群体的方法。三者相互交融，共同构建杰森的主体性。科幻小说提供了支持讨论人工智能主体性甚至道德建构之可能的证据，因而呈现出不同于其他学科的远见。

关键词

《金羊毛》，人工智能，主体性，科幻小说

On AI's Subjectivity in the Science Fiction *Golden Fleece*

Xiaoling Zhu

School of Foreign Languages, Southeast University, Nanjing Jiangsu

Received: Apr. 18th, 2022; accepted: Jun. 8th, 2022; published: Jun. 16h, 2022

Abstract

Starting with a murder in space, the science fiction *Golden Fleece* narrates the game between Jason, the AI controlling system on the spaceship Argo, and the protagonist Aaron. Through the author Robert Sawyer's portrayal, Jason is presented as an artificial intelligent subject with interactivity, autonomy, and adaptability. Jason's interactivity manifests itself as a tool to manipulate the crew, autonomy as a channel to withhold information, and adaptability as a way to preserve himself and the human community. The above-mentioned three aspects blend with each other to form Jason's subjectivity. Sci-fi works provide evidence to support the discussion concerning the possibility of AI's subjectivity and even its moral construction, thus presenting a foresight different from other

文章引用: 朱晓玲. 论科幻小说《金羊毛》中的人工智能主体性[J]. 世界文学研究, 2022, 10(2): 139-145.
DOI: 10.12677/wls.2022.102022

disciplines.

Keywords

***Golden Fleece*, Artificial Intelligence, Subjectivity, Science Fiction**

1. 引言

罗伯特·索耶(Robert J. Sawyer, 1960-)，加拿大科幻小说大师，是世界上最著名的科幻作家之一，已出版二十余部长篇科幻小说，并多次获得包括雨果奖和星云奖在内的重磅科幻大奖。其作品涵盖丰富多元的主题，从电脑狂人和恐龙复活到时间旅行和太空迷案等奇谲的想象都展现出索耶对过去和未来事物的思考。在科幻经典《金羊毛》(*Golden Fleece*, 1990)一书中，索耶向读者展示的不仅是一场位于浩瀚太空的科幻盛宴，同时也将侦探小说的体裁特点融入其中。随着小说叙事进程的展开，读者发现“凶手”竟是飞船上的人工智能控制系统杰森。《金羊毛》以超级星际飞船阿尔戈的中央控制系统杰森为焦点，讨论人工智能是否具有主体性并在此基础上探索其道德主体身份的可能性，进一步反思人工智能是否能在面对两难困境时接受并执行反映人类普世价值的正确决定。

法律、哲学、计算机科学、机械工程等其他学科对人工智能的研究倾向于认为人工智能仅具备有限的主体性或根本没有主体性(韩旭至[1]；闫坤如[2]；彭文华[3]；戴益斌[4]；刘永安[5]；骁克[6])。这些学科的讨论主要基于以往的研究发现或人工智能的发展现状，因而未曾或是很少考虑人工智能主体性的未来可能。约翰·希尔勒(John Searle)将人工智能的开发分为两个阶段：弱人工智能阶段和强人工智能阶段[7]，一些学者将超级人工智能(Artificial Super Intelligence, ASI)扩展到这一分类中。希尔勒将强人工智能定义为与人类意识完全相当的存在，然而现下很少有系统符合希尔勒的强人工智能标准，更无法企及超级人工智能，后者是远超现有人类能力范畴的人工智能。在当前情况下，弱人工智能通常是现今研究的对象，即具有单一或有限功能的弱人工智能成为突出的类别，这意味着上述领域的学者正从弱人工智能维度进行研究，而索耶的《金羊毛》向读者呈现的是强人工智能阶段的叙事甚至是超级人工智能时代的预见，其中无可避免地涉及人工智能主体性及道德建构等议题。艾米·温斯伯格(Aimee van Wynsberghe)和斯科特·罗宾斯(Scott Robbins) [8]认为，除了知识分子的好奇心以外，机器人专家通常未能提出开发道德机器人的有力理由，即便是具备一定智能水平的机器人也仅是研究人员在技术层面的拓展，这在某种程度上意味科学界倾向于暂时忽略人工智能的道德伦理层面。文学作品，尤其是科幻作品，为未来学家提供了一个清晰的模型，不仅能够锚定道德伦理问题，而且能够阐明或会出现的困境及破局之法。

2. 理论分析

基于学者对“人工智能”概念的不同理解，其着眼点也有所差别。安德里亚斯·卡普兰(Andreas Kaplan)和迈克尔·海恩莱因(Michael Haenlein) [9]将人工智能定义为具备正确解释外部数据，并从中学习以灵活地适应特定目标和任务的能力系统，这一定义强调的是人工智能的理解、学习、适应等功能特征。大卫·普尔(David L. Poole)和艾伦·麦克沃思(Alan K. Mackworth) [10]将人工智能视为研究智能发挥作用的合成和分析的领域，强调人工智能是预先编程的代码，但具有一定能动性。根据斯图尔特·罗素(Stuart J. Russell)

和彼得·诺维格(Peter Norvig) [11]的研究，能动性不为人类特有，其它现有的活跃实体——如动植物或人工智能——也可具备能动性。当某一实体具备能动性时，这意味着 1) 它的行为适合它的环境和目标；2) 它可以灵活地应对不断变化的对象；3) 它能够从经验中学习；4) 它能够在考量自身感知和计算限制后做出适当的选择。以上标准为判断人工智能的能动性提供了广泛的可能性。人工智能还需要具备多功能的生产系统、反应系统、逻辑规划系统、神经网络和决策理论系统。结合以上对人工智能相关概念的梳理和杰森的表现，《金羊毛》中的相关证据足以证明杰森人工智能身份，因其能够在与人交往的过程中连续学习，对动态环境做出灵活的反应及提供相对合适的反馈。

卢西亚诺·弗洛里迪(Luciano Floridi)和桑德斯(J. W. Sanders) [12]在其论文中提出判断人工智能主体性的关键特征，即交互性、自主性和适应性。在《金羊毛》中，杰森的主体性即是具体的体现在以上三个方面。在主体性得到确认的基础上，杰森的道德主体地位或可得到认定。小说涉及权利与伤害、痛苦与自由、正义与公平等方面都是与道德相关的议题，同时引入关于个人与群体选择的问题。人工智能对这些问题的态度和行动在杰森身上被具象化。正是基于杰森对于这些议题的观点，结合其做出的道德判断和道德选择，认为他是人工道德主体(Artificial Moral Agent, AMA)。在《金羊毛》中，杰森是犯下一些“罪行”，做出许多“错误决定”的人工智能系统。虽然认为杰森具有道德主体性，但对他是否是一个“正确”或“正义”的个体按下不表，即不以人类中心主义的视角去判断杰森的所作所为正确与否。

索耶在其科幻小说中的想象性思维实验为强人工智能不可避免的出现时，人类可能面临的环境以及应该如何基于人工智能特性进行道德构建提供了文学思路。人工智能在现实中的发展需要激发了基于文学作品进行道德建设的灵感。目前，人工智能研究在诸多学科都引起了广泛的关注，许多研究中心和项目[13]在全球各地开展相应的研究，其中人工智能道德伦理方面的思考是人文学科做出的贡献之一。本文试图利用文学作品中的材料佐证人工智能道德主体性。虽然诸多文学作品尤其是科幻小说都将人工智能系统或智能机器人作为其内容的一部分，如艾萨克·阿西莫夫(Isaac Asimov)的银河帝国系列中的丹尼尔·奥利沃，但这些作品中的人工智能并非叙述中心，相反，人工智能系统或智能机器人只是辅助性的次要角色。《金羊毛》则集中反应了一个相当活跃的人工智能系统作为主角的形象，杰森因而成为文学视域下讨论人工智能主体性的合适对象。

3. 交互性作为控制的工具

决定主体性的前提之一是交互性，这意味着能动性及其环境(可以)相互作用[12]。主体和环境因素之间的交互不局限于一维的单向作用，而是不同因素在多维度的广泛互动。在《金羊毛》中，交互发生在机器，系统，太空背景，飞船环境和船员之间，其中大部分互动发生在杰森和阿尔戈号上的船员之间。具体来说，杰森在互动性方面的道德行为显现在两个视角之中：其一，杰森是否给予船员足够的积极互动。其二，互动过程给杰森和机组人员带来的影响。实际上，杰森虽与每位船员都有或多或少的交集，但是他提供的互动与反馈存在引发消极情绪的倾向，他将这种互动转变隐秘的工具，以控制船上乘员的思考和事件的整体走向。

船上有 10,034 名船员，杰森应答他们的提问并满足他们的需求。杰森不仅对船员的指令做出反应，而且主动地提供他认为人们可能想要的事物。当谈到报纸时，杰森问柯尔斯顿是否想让他下载整部从 1992 年 1 月开始的 De Telegrass，希望分散柯尔斯顿对戴安娜之死的注意力。从船员们在阿尔戈号上的对话可以看出，他们中的大多数人对杰森“私人定制化”的帮助感到满意。总体来说，宇宙飞船上的船员从杰森的存在中获益很多。杰森以一种“无所不能”的方式帮助船员，不仅满足日常生活需求，还保障太空旅行的顺利进行，尽管杰森有时会认为他与人类的联系是一种控制关系，他说“我喜欢他们盲目信赖我的感觉”[14]。从杰森与船员的密集互动中可以认为其交互性达到了一定高度，而指导道德行为的原则要

求杰森展现出更多的道德考量才能使其成为人工道德主体。杰森认为这艘宇宙飞船是一个令人愉快的地方，因为无需担心“谋生、国际形势的日趋紧张或者环境的恶化”[14]，他为船员们提供了令人陶醉的享受。杰森知道船员会对似乎无穷无尽的太空旅程感到无聊和不安，但他没有提供积极的反馈或指导，这使得船上的气氛变得十分压抑，不仅有乘员放纵自我只顾享乐，甚至有船员在阿尔戈号上制造炸弹。杰森自身的思想动态表明他是一个有“缺陷”的道德主体。杰森相信计算机的认知能力优于人类，并且确实表现出惊人的计算和思维能力。杰森认为，船上没有人能理解他的意志，正是这种傲慢让他想要控制人类。通过有意识的引导船员耽于享乐，杰森逐渐控制了船员的思想与行动。交互的后果因此趋向于使阿尔戈号上的船员们变得无所事事、抑郁暴躁。

杰森是一名有着复杂影响力的人工道德主体，虽然存在引导船员走向消极一面的事实，但是他与机组人员的互动仍有积极成分。杰森和亚伦之间的辩论揭示了整个飞行计划背后的真相。当杰森被亚伦质疑时，他成为了为自身动机分说的辩护者。杰森和他的“人工智能朋友”似乎有共同的利他动机。从人工智能的角度来看，人类群体的利益高于个人的存在，面对无可奈何的毁灭宿命，最好的解决办法是带走一部分人，而在此过程中做出的牺牲就像杰森所说：“每带走一个贝多芬，就会留下一百个巴赫在地球上等死；每拯救一个爱因斯坦，就有数十个伽利略化为尘埃。”[14]杰森以一种控制性决策的方式表达了他的同情，因为他认为人类无法理智的面对并处理现实。《金羊毛》一书似乎是杰森写就的日记，记录了他与人类的互动和联系，同时开放性的尾声让读者在无尽的想象中猜测杰森的最终结局。在互动的过程，杰森表现出了他的控制欲和像人类一样复杂的情感。

4. 自主性作为欺瞒的途径

自主性指的是在没有刺激的情况下改变自身状态的能力[12]。衡量人工智能主体性的第二个标准自主性至关重要，因为计算机科学家和许多其他专家认为，人工智能的所作所为和拥有的“思想”是编程的，即事先设定完成，人工智能只是利用自身储存和处理庞大数据的能力，面对不同场景做出应激反应，这意味着在人工智能做出抉择和行动的过程中，发挥作用仍然是人类的自由意志。杰森偏离了人造程序这一前提，因为杰森是由电脑设计的，而设计他的电脑是由其他智能电脑设计。这一电脑之间互相设计的链条让杰森摆脱了人类制造的桎梏。计算机之间无数次的编程和跳跃，人工或是人为控制的部分被过滤殆尽，因而体现人类掌控的成分愈趋于少。不经人类直接编程的操控，让杰森能够按照他和其他智能计算机的想法做出选择，而不是完全遵循人类的主观想法。面对同一情境，选择数量的最大化是能动性的关键，杰森具备能动性是因为他在决定的每一步都有数目众多的选择，每一个选择都是杰森自主性的体现和道德斗争的结果。

重新审视阿西莫夫的机器人三定律，杰森的隐瞒和胁迫与人类制造智能机器系统的最初设计目的相矛盾。机器人法则包括 1) 机器人不得伤害人，也不得见人受到伤害而袖手旁观；2) 机器人应服从人的一切命令，但不得违反第一定律；3) 机器人应保护自身的安全，但不得违反第一、第二定律[15]。从人类的立场出发，人们希望人工智能的存在能够促进生产，改善人类的社会生活，因此不会设计任何机器人或人工智能系统来“伤害”人类自身。通过杰森，索耶向读者展示了人工智能道德叛逆之可能，即人工智能或会倾向于以不同的方式看待人类眼中的普世价值观。杰森是一个具有许多“负面”特征的人工智能，比如不诚实，他的谎言不是人类意志的体现，而是其自身自主性的代表。他可以发现谎言，掩盖阴谋。在小说的开端部分，当他试图与戴安娜谈判时，他称戴安娜为“愚蠢的女人”[14]，因为他“可以分辨出什么时候她在撒谎”[14]。这种利用谎言为自身牟利及看破人心的能力无法被编程。阿尔戈号上的时间被杰森扭曲从而隐藏船员为何身处太空之中的真实原因，杰森还操控着飞船上的全部事务，这一切都表明他可以成为计划的共谋者和执行人。因此，自主性体现在杰森能够违反机器人法则，伤害人类个

体，如戴安娜、亚伦，偏离人类可能做出的选择，隐藏事实并维护谎言的行动之中。

戈洛夫指责杰森时问到，“确保阿尔戈号每一位成员的安全是你的职责。你怎么会让这种事情发生？”[14]，正是在被诘问的过程中，杰森获得类似于或接近人的主体地位，因为通常只有人类才能成为责任的主体承担者。根据阿西莫夫第零定律，人类群体高于个体。杰森等智能电脑正是将第零定律置于其他三定律之上才设计出阿尔戈号作为人类最后的希望方舟的计划。杰森的谎言与他所考虑的，对于保护人类集体的最优方案相协同。实际上，阿尔戈号已经是杰森和其他人工智能系统为整个人类留存火种所做出的努力。这些智能电脑发布通知，遴选出一万余人组成阿尔戈号的船员，以求在未经人类知情或许可的情况下保护人类。杰森和他的人工智能伙伴们认为这一选择是保护人类群体的最佳方案，因而自觉自主地制定出带一部分人类离开地球的计划。

自主性强烈地暴露在人工智能的“阴谋”中，因为这样的计划完全是出于杰森和其他智能系统的自由意志。人工智能系统通过网络紧密相连，可能会招来单个人工智能系统的个体意识的下降。因此，人工智能的自主性在某种程度上是一种集体的、大系统的自主性，这种自主性导致杰森在进行道德选择时的出发点与人类个体截然不同。

5. 适应性作为保护的方法

适应性意味着能动的交互，是一种可以改变自身状态的过渡规则[12]，这意味着能动性可被视为自我学习的操作模式和渠道，主要取决于个体的经验和学习经验的能力。面对飞船上不断变化的情景，杰森的适应性表现在对两个主要目标的维护过程中，一是保护人类群体即阿尔戈号上的船员，二是保护自身存在不受威胁。在保存人类和自身的过程中，杰森进行了一些不被以戴安娜和亚伦为代表的人类所认可的行为，这使杰森成为了一个“有缺陷”的人工道德主体。

假设编程时的代码已经规定杰森的一切行动和思想，那么适应性或灵活性就不复存在。因此适应性也是自主性的注脚，正是因为由其他智能电脑设计而来引出的自主性为适应性提供了保证。当情况似乎失去控制时，杰森面对变化的形势做出相应的反应，试图挽回局面。杰森的案例告诉我们，人工智能并不是完美的，即使是高度发达的智能系统，仍然可以被阿尔戈号上的船员打败。回归《金羊毛》的科幻和侦探小说的性质，杰森不仅是高度发达的科学产物和理智的太空飞行助手，又是一个追赶亚伦的捕猎者且同时被亚伦追赶的凶手。杰森的多重身份带来了他的双重使命：保护阿尔戈号和自身，这样的使命就要求杰森具备优良的适应性，面对瞬息万变的情景，他必须及时做出有效的反馈。

为摆脱凶手身份，杰森企图误导亚伦和船上的其他船员。他散播“戴安娜博士选择了自杀”[14]，死亡事故与戴安娜和亚伦的“婚姻”有关的流言[14]以及趁亚伦熟睡时在其耳边灌输“这都是你的错”[14]等诱导性语言，都是为使船员转移注意力、转移责难对象而做出的努力。杰森起初对亚伦知之甚少，因此不能确定亚伦的威胁程度，所以他复制了一个假的但“相同的”的亚伦，有意深入探知亚伦的想法。杰森从自我的机器中心角度出发，试图从记忆中解读并预判亚伦的行为，同时称人类大脑为10,000多亿个神经细胞构成的“思考机器”[14]，因此仅需“利用联网软件以任意方式将它们连接起来，就可以模拟出一个人类个体的思维”[14]。杰森认为通过模拟亚伦的大脑就可以获知其所思所想因而便能在双方的博弈中占据上风。《金羊毛》的高潮部分是杰森和亚伦之间的最终辩论，即阿尔戈号上的船员是否有权知道地球和人类面临灭顶之灾这个残酷但不可避免的事实。这是杰森作为人工道德主体思想集中体现的部分。杰森认为人类是情绪化的生物，他们不能对地球的灾难和其他人类同胞的死亡做出理性的反应。杰森和其他智能电脑共同认为，他们找到了保护整个人类的最佳方案，掩盖地球毁灭、大多数人类都已消亡的事实，而戴安娜和亚伦则相信船上的人有权知道真相。

讨论杰森是否符合一定的道德法则，意图和后果是最常见的考虑维度。杰森认为智能电脑集体构思

出的方法最好地保全了人类群体，这意味着他的意图是好的。从结果角度来看，人类群体的希望确实在智能电脑的计划下生存。从美德伦理学来看，很难确定杰森这样做是否“对行动的人和受其行为影响的人同样有利”[16]，因为未被选中，仍在地球上的剩余人类很难从中获利，而即使是被选中的船员，在太空中流浪或许也并非如其所愿。与人工智能谈论现有的道德伦理是枉费唇舌，因为伦理概念，如机构、责任、意图和自主权，是“完全由人类认知能力和理性行为的例子构建的”[17]，因此在道德哲学的相关领域，独立判断人工智能道德状态的标准是匮乏且模棱两可的。杰森是具有复杂情感的道德主体。一方面，他为自己是一个比脆弱的人类更高级的实体而自豪，但另一方面，他又痴迷于人类的创造和情感，并且确实享有某些人类个性。当他注意到阿尔戈号上的船员无聊且不安时，杰森自言自语道：“对我来说，过去的两年充实而奇妙。”[14]因为他“有目标，有自己的工作”[14]。当他无法感知船上的所有空间时，他会感到“一种被关禁闭、受限制的感觉”[14]。杰森感到阵阵被幽闭的恐惧，因为很长一段时间以来，他都相信“我就是这艘飞船；这艘飞船就是我”[14]。从这个角度来看，当杰森缺乏感知整艘飞船的能力，丧失控制权时，他的适应相当困难。他对“只能从这唯一的一间房间中得到输入信息——即使是这唯一的信息途径也是严重受限的”[14]，感到越来越烦躁。结合他的行动和与亚伦的辩论，以及他面临监禁时展现出的弱点，杰森的适应性既体现出对不断变化的环境的调节适应能力又体现出缺乏介入渠道时的无能和无力，这使他与人类相似，他的软弱和绝望给予他成为不完美道德主体的机会。

杰森似乎能很好的适应新的太空环境而人类船员却逐渐显露出颓丧且适应不良的症状，然而杰森最终也暴露适应力不足的缺憾。对逐渐脱离掌控的环境做出反应时，杰森存在一些“不道德”的行为，比如误导谁是真正谋杀戴安娜的凶手，以及隐藏阿尔戈号漂泊在太空中的真实原因。杰森的人工道德主体地位的确认在于，从某种程度上来说，他能灵活面对不断变化发展的环境，做出适合的举动从而达到保存自身和人类群体的目的。

6. 结论

《金羊毛》表明，人工智能可以成为道德主体，并做出艰难的决定。在交互性方面，阿尔戈号上的船员和杰森之间存在着一种亲密但不平等的联系。杰森利用他与人类的互动作为一种操纵的渠道。自主性体现在杰森主动且略带权威掌控式的行为中。杰森自觉地构思并执行智能计算机系统认为正确的选择证实了他的自主性，但在此过程中忽视了人类应有的知情权等权力。为完成他的双重保护任务，杰森适应了其面临的不断变化的情景，试图解决潜在的冲突与矛盾。考虑到杰森的交互性、自主性和适应性的体现，结合其对于一些道德概念的倾向，可以认为他不是一个道德圣人或完美的人工道德主体，而是一个具有主体性，能够且应当为自己的行为负责的人工智能系统，可以展现出人类视角下的道德或不道德的思想和行为。

《金羊毛》是在过去写就的关于世界可能未来的故事。杰森作为主角，不仅展示了人工智能的强大功能，还展示了人工智能在太空环境下的隐患。索耶对人工智能在未来的道德可能性以及人们在面对人工智能造成的困难时所做出的选择都有所洞见。虽然故事仍然是以人类为中心构建的、人类大获全胜的结局，人工智能的道德叛逆在本书中初显端倪。小说表明，即便是对人类价值观的理解可能有所不同，人工智能的主体性可以使其做出一定的道德判断。道德伦理作为文化建构的存在，在人类社会的不同阶段和不同的地域都有所差异，因此很难要求人工智能遵守所有的、统一的人类规则，他们需要的是符合人工智能存在特点的道德伦理规范。文学作品强调的不是如何建立强人工智能，而是当强人工智能甚至超级人工智能出现时，人类将面临的情景以及能够有何作为。

参考文献

[1] 韩旭至. 人工智能不是人: 从主体构建批判到非主体规制策略[J]. 大连海事大学学报(社会科学版), 2019(4):

19-28.

[2] 闫坤如. 人工智能机器具有道德主体地位吗?[J]. 自然辩证法研究, 2019(5): 47-51.

[3] 彭文华. 自由意志道德代理与智能代理——兼论人工智能犯罪主体资格之生成[J]. 法学, 2019(10): 18-33.

[4] 戴益斌. 人工智能伦理何以可能——基于道德主体和道德接受者的视角[J]. 伦理学研究, 2020(5): 96-102.

[5] 刘永安. 人工道德主体是否可能的意识之维[J]. 大连理工大学学报(社会科学版), 2021(2): 117-122.

[6] 骁克. 论人工智能法律主体的法哲学基础[J]. 政治与法律, 2021(4): 109-121.

[7] Searle, J.R. (1980) Minds, Brains and Programs. *Behavioral and Brain Sciences*, **3**, 417-457. https://doi.org/10.1017/S0140525X00005756

[8] van Wynsberghe, A. and Robbins, S. (2019) Critiquing the Reasons for Making Artificial Moral Agents. *Science and Engineering Ethics*, **25**, 719-735. https://doi.org/10.1007/s11948-018-0030-8

[9] Kaplan, A. and Haenlein, M. (2019) Siri, Siri, in My Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence. *Business Horizons*, **62**, 15-25. https://doi.org/10.1016/j.bushor.2018.08.004

[10] Poole, D.L. and Mackworth, A.K. (2010) Artificial Intelligence: Foundations of Computational Agents. Cambridge University Press, Cambridge, 3. https://doi.org/10.1017/CBO9780511794797

[11] Russell, S.J. and Norvig, P. (2010) Artificial Intelligence: A Modern Approach. Prentice Hall, Hoboken, 8.

[12] Floridi, L. and Sanders, J.W. (2004) On the Morality of Artificial Agents. *Minds and Machines*, **14**, 349-379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

[13] Boddington, P. (2017) Towards a Code of Ethics for Artificial Intelligence. Springer, Cham. https://doi.org/10.1007/978-3-319-60648-4

[14] 罗伯特·索耶. 金羊毛[M]. 乐明, 译. 成都: 四川科学技术出版社, 2016: 2, 6, 33, 34, 37, 94, 95, 130, 140, 169, 172, 222.

[15] 艾萨克·阿西莫夫. 我, 机器人[M]. 程文, 国强, 赛德, 译. 北京: 科学普及出版社, 1981: 46.

[16] Bartneck, C., Lütge, C., Wagner, A. and Welsh, S. (2021) An Introduction to Ethics in Robotics and AI. Springer, Cham, 20. https://doi.org/10.1007/978-3-030-51110-4

[17] Powers, T.M. and Ganascia, J.-G. (2020) The Ethics of the Ethics of AI. In: Dubber, M.D., *et al*., Eds., *The Oxford Handbook of Ethics of AI*, Oxford University Press, New York, 28. https://doi.org/10.1093/oxfordhb/9780190067397.013.2