当代科学文化前沿丛书

MORAL MACHINES
TEACHING ROBOTS RIGHT FROM WRONG

道德机器
如何让机器人明辨是非

〔美〕温德尔·瓦拉赫（Wendell Wallach）
科林·艾伦（Colin Allen） ◎著
王小红 ◎主译

北京大学出版社
PEKING UNIVERSITY PRESS

当代科学文化前沿丛书

MORAL MACHINES
TEACHING ROBOTS RIGHT FROM WRONG

道德机器
如何让机器人明辨是非

〔美〕温德尔·瓦拉赫（Wendell Wallach）
科林·艾伦（Colin Allen）◎著
王小红 ◎主译

北京大学出版社
PEKING UNIVERSITY PRESS

著作权合同登记号　图字：01-2016-4369

图书在版编目 (CIP) 数据

道德机器：如何让机器人明辨是非 /（美）温德尔·瓦拉赫（Wendell Wallach），（美）科林·艾伦（Colin Allen）著；王小红主译 .—北京：北京大学出版社，2017.11

（当代科学文化前沿丛书）

ISBN 978-7-301-28940-2

Ⅰ.①道… Ⅱ.①温… ②科… ③王… Ⅲ.①机器人学—研究 Ⅳ.①TP24

中国版本图书馆 CIP 数据核字 (2017) 第 267132 号

书　　名	道德机器：如何让机器人明辨是非 DAODE JIQI：RUHE RANG JIQIREN MINGBIAN SHIFEI
著作责任者	〔美〕温德尔·瓦拉赫　〔美〕科林·艾伦　著　王小红　主译
责任编辑	吴卫华　陈　静
标准书号	ISBN 978-7-301-28940-2
出版发行	北京大学出版社
地　　址	北京市海淀区成府路 205 号　100871
网　　址	http://www.pup.cn　　新浪微博：@ 北京大学出版社
电子信箱	zpup@pup.cn
电　　话	邮购部 62752015　发行部 62750672　编辑部 62753056
印 刷 者	三河市博文印刷有限公司
经 销 者	新华书店
	650 毫米 ×980 毫米　16 开本　17.5 印张　280 千字
	2017 年 11 月第 1 版　2017 年 11 月第 1 次印刷
定　　价	58.00 元

中文版序

几乎过不了一星期，报纸上、电视上，就总会看到人工智能（AI）的新闻，各种 AI 对人们生活造成或者可能造成负面影响的消息不绝于耳。不管是数据挖掘算法泄露个人信息，还是有政治导向的宣传的影响，如数据挖掘公司可能利用选民信息深度干预了美国总统选举结果，或种族和性别偏见浸入了人工智能程序，抑或是自动驾驶汽车会使数百万人丢掉工作。自从《道德机器》（*Moral Machines*）英文版首次问世，数年以来，围绕 AI 产生的这些伦理问题已经变得愈加急迫和突出。

这里给出一个文化偏见渗透进 AI 程序方面的例子。类似谷歌翻译这样的解释语言程序，其能力近年来有了大幅度提高，这主要得益于新的机器学习技术以及可以训练程序的大量在线文本数据的出现。但是最近的研究显示，在机器习得越来越像人的语言能力的过程中，它们也正在深度地吸取人类语言模式中隐含的种种偏见。机器翻译的工作进路一般是这样的，建构一个关于语言的数学表征，其中某个词的意义根据最经常和它一同出现的词而被提炼为一系列数字（叫词向量）。令人惊讶的是，这种纯粹统计式的方式却显然抓住了某个词的意义中所蕴含的丰富的文化和社会情境，这是字典里的定义不大可能有的。例如，关于花的词语和令人愉悦之类的词聚类，昆虫类的词语和令人不快之类的词聚类；词语“女性”“妇女”与艺术人文类职业以及家庭联系更紧密，词语“男性”“男人”则和数学、工程类职业更近。尤为危险的是，AI 还有强化所习得的偏见的潜能，它们不像人可以有意识地去抵制偏见。因此，如何在设计理解语言的算法时既让其消除偏见，又不丧失其对语言的解释能力，这是一项很大的挑战。

对于上述可能产生自 AI 的一些最坏的情况，学术界和技术界的领军人物纷纷表示关注。像剑桥大学当代最重要的物理学家史蒂芬·霍金（Stephen Hawking），还有太空运载私人运营和电动汽车的开拓者埃隆·马斯克（Elon Musk），他们担心，AI 代表着威胁人类存在的危险。另一些

人却表示乐观，对 AI 的长远前景满怀热情。但他们认识到对 AI 的限制以及 AI 的弱点对商业来说是很糟糕的事情。就在 2016 年 9 月，来自微软、IBM、FaceBook、谷歌、苹果以及亚马逊的领导人，他们共同创立了关于 AI 的合作伙伴关系，当时宣称的目标是要确保 AI 是“安全可靠的，要和受 AI 行为影响的人们的伦理和喜好一致”。但是，市场经济的逻辑总是在驱动着愈演愈烈的雄心，于是，应用机器学习处理收集的海量电子化数据、提高自动驾驶汽车上路的数量，新闻的头版头条持续不断。

《道德机器》书中的信息依然是极为重要的，虽然此书首次出版以来，技术已经大大向前发展了，但基本的难题不仅依然没有解决，反而被这些技术进展所加剧。先进算法工作的基本原理还是没完全搞清楚，这就关系到 AI 解释自身决策方面的无能为力。当然，人们在决策时也不总是能够一目了然地解释清楚自己的抉择。但是，要设计处在一个更高水平上的机器，我们这样把握是对的。

在《道德机器》中阐述了一种混合式系统，组合了自下而上式的数据驱动的学习和进化的方式，以及自上而下式的理论驱动做决策。这个思想，对于 AI 软件的设计以及能够在人类环境中操作的机器人来说，依然是最好的策略。但是，在这本书约十年前问世之后，批评者们的意见是对的，我们本来可以更多地讨论一下设计更大的有机器运行的社会-技术系统。我们关注如何使机器自身的道德能力逐渐增进——这是一个依然需要研究的课题，关于如何使机器尊重它们的设计者和使用者人类的伦理价值，以及有效维护人类的道德责任方面的问题，仅仅只解决了部分。

人工智能机器已经在特定领域胜过了人。这令许多评论家相信超人智能不远了。但是，尽管 AI 取得了目前的进展，它的成功依然是脆弱的，而人的智能通常却不是这样。某个机器可以下围棋胜过人，却不能同时会下象棋、会开车，还能在家人饭后，一边帮忙收拾桌子，一边解答政治学的问题。即使这些能力可以绑定给一个物理机器，我们也依然不知道，如何以像人那样的方式把它们整合在一起，让一个领域的思想和概念灵活地畅行至另一领域。如此说来，《道德机器》最初问世时候的基本信息依然没有变，这就是：不应该让未来学家的担心和猜测转移我们的注意力，我们的任务依然是设计自主机器，让它们以可辨识的道德方式行事，使用着我们已经获得的社会、政治和技术资源。

科林·艾伦(Colin Allen)

美国-匹兹堡，2017 年 9 月 26 日

致　谢

好多人的贡献在本书贯穿始终。我们要首先感谢的，也是最重要的一位——伊娃·斯密特(Iva Smit)博士，她是我们在道德机器方面若干篇文章的合作者。写作这本书时，我们广泛借鉴了这些文章。毫无疑问，书中的许多观点和用词都源于她，我们尤其感激她对第六章的贡献。伊娃在帮助我们制定这本书的写作大纲方面也承担了重要的角色。其实她在这个领域的影响比这更广。从2002年到2005年，她通过组织一系列专题讨论会将有志趣于机器道德的学者们聚拢起来，对这个新兴的研究领域做出了持续的贡献。的确，要不是2002年伊娃邀请我们两人到德国巴登巴登参加第一届讨论会，我们也许就不会彼此相识。她热情、亲切地将一个小型的学者共同体紧密团结起来(我们会在下面提到这些人)。伊娃的重要动机是要唤起商界和政界领袖们，意识到自主性系统所带来的危害。就我们选择关注于开发人工道德智能体的技术层面来说，这本书或许并非她的初衷。然而，我们希望传达出她关于伦理缺失系统危害性的一些观念。

斯密特博士组织的这四届专题讨论会的主题是“人类和人工智能决策中的认知、情感和伦理问题”；该专题讨论会是由乔治·拉斯克(George Lasker)领导的国际系统研究和控制论高级研究院(International Institute for Advanced Studies in Systems Research and Cybernetics)资助举办的。我们感谢拉斯克教授和研讨会的其他参与者。就在最近几年，许多关于机器道德的专题研讨会帮助我们对这个主题有了更为深入的理解，我们同时还想感谢那些专题研讨会的主办者和参加者。

科林·艾伦(Colin Allen)最初于1999年涉足这个领域，当时他受瓦罗尔·阿克曼(Varol Akman)邀请，为《实验与理论人工智能杂志》(*Journal of Experimental and Theoretical Artificial Intelligence*)撰写文章。这次偶然的机会让他意识到，如何建构人工道德智能体方面的问题，竟然还是未经探索的哲学领域。于是，加里·瓦尔纳(Gary Varner)发

挥其伦理学专长，研究生詹森·津瑟(Jason Zinser)付出热情和努力，于2000年合作发表了的一篇文章。我们写这本书时引用了它。

温德尔·瓦拉赫(Wendell Wallach)于2004年和2005年在耶鲁大学讲授一门本科生的讨论班课程，叫"机器人的道德和人的伦理"。他感谢学生们的洞见和热情对他的思想形成作出的重要贡献。其中，一位名乔纳森·哈特曼(Jonathan Hartman)的学生提出了我们在第七章讨论的一个原创性观点。温德尔和斯坦·富兰克林(Stan Franklin)教授的讨论，对第十一章尤为重要。斯坦帮助我们撰写了该章内容，在那一部分，我们将他针对通用人工智能开发的学习型智能配给代理(LIDA)模型应用于建构人工道德智能体(AMAs)这个难题。他理应被看做是那一章的合作者。

在书中还有许多其他同行和学生的评论及建议。我们尤其想要提到迈克尔(Michael)和苏珊·安德森(Susan Anderson)夫妇、肯特·巴布科特(Kent Babcock)、大卫·卡尔弗利(David Calverly)、罗恩·克里希(Ron Chrisley)、彼得·达尼尔森(Peter Danielson)、西蒙·戴维森(Simon Davidson)、卢奇亚诺·弗洛里迪(Luciano Floridi)、欧文·霍兰德(Owen Holland)、詹姆士·修斯(James Hughes)、埃尔顿·乔伊(Elton Joe)、彼得·卡恩(Peter Kahn)、邦尼·卡普兰(Bonnie Kaplan)、加里·考夫(Gary Koff)、帕特里克·林(Patrick Lin)、卡尔·麦克道曼(Karl MacDorman)、维拉德·米兰克尔(Willard Miranker)、罗萨琳德·皮卡德(Rosalind Picard)、汤姆·鲍尔斯(Tom Powers)、菲儿·罗宾(Phil Robin)、布莱恩·斯卡萨拉蒂(Brian Scasselati)、维姆·斯密特(Wim Smit)、克里斯提娜·斯皮塞尔(Christina Spiesel)、斯蒂夫·托伦斯(Steve Torrance)以及文森特·魏格尔(Vincent Wiegel)。

还要特别感谢那些对各章做出详细评论的人。坎迪斯·安达立阿(Candice Andalia)和约耳·马克斯(Joel Marks)两人对若干章节进行了评论，还有弗莱德·艾伦(Fred Allen)和托尼·比弗斯(Tony Beavers)对全部手稿进行了评论，他们应当获得最高的称赞。他们的洞见对本书的完善起到不可估量的作用。

2007年8月，我们在宾夕法尼亚中部度过了令人愉快的一周，敲定出一部几近完成的书稿。接待者是萨莫斯特乡村酒店的卡罗尔(Carol)和罗兰德·米勒(Rowland Miller)夫妇。卡罗尔的丰盛早餐，罗兰德对最初两章的热情回应，还有充足的咖啡、茶和小甜饼，从各方面调动着我们的工作能量。

斯坦·韦克菲尔德(Stan Wakefield)对拓展我们的写作计划给出了合理的建议。在最后的编辑和手稿的准备上,印第安纳大学的约书亚·斯马特(Joshua Smart)显然是位得力的助手。他做的大量编辑工作提高了文字的明晰性和可读性,并且对整理书后的每章注释作出重要贡献。

牛津大学出版社的彼得·奥林(Peter Ohlin),乔林·奥斯卡(Joellyn Ausanka)和莫立·瓦根纳尔(Molly Wagener)给我们很多帮助——感谢他们周到的建议,以及按时让这部手稿出版问世。副标题"如何让机器人明辨是非"就是彼得建议的。我们要专门感谢玛莎·拉姆齐(Martha Ramsey),她出色的编辑工作无疑对手稿的可读性作出了极大的贡献。

温德尔·瓦拉赫想感谢耶鲁大学生命伦理学交叉学科中心的成员过去四年来的支持。该中心的副主任卡罗尔·柏兰德(Carol Pollard)、她的助理布鲁克·克劳克特(Brooke Crockett)和乔恩·墨瑟(Jon Moser)以各种方式给温德尔很大帮助。

最后,如果没有南希·瓦拉赫(Nancy Wallach)和林恩·艾伦(Lynn Allen),我们的妻子们的耐心、爱心和宽容,我们本不可能完成这个工作。她们的美德没有任何一点是虚假的。

温德尔·瓦拉赫,于康涅狄格州布卢姆菲尔德
科林·艾伦,于印第安纳州伯明顿
2008 年 2 月

目录

目录

目录

目录

导　言

在麻省理工学院(MIT)的情感计算实验室里,科学家正在设计能读懂人类情绪的计算机。金融机构已经利用国际互联网进行评估,并在每分钟批准或拒绝数百万笔交易。在日本、欧洲和美国,机器人科学家在开发用于照顾老年人和残疾人的服务型机器人。日本科学家在制造外表看起来和人没有区别的机器人(android)。韩国政府宣布了到2020年让每家每户有一个机器人的目标。他们还和三星集团在共同开发装载武器的机器人,帮助守卫毗邻朝鲜的边界。同时,在每一个可以想到的装置上,从汽车到垃圾桶,计算机芯片都在促进、监视和分析着人类的活动;而在每一个可以想象的虚拟环境里,从网上冲浪到在线购物,软件"机器人"也如此。这些(软硬件)机器人[(ro)bots]——一个我们将用来囊括物理机器人和软件智能体的术语,它所收集的数据正在用于商业、政府和医疗目的。

所有这一切进展正在汇聚并促生着机器人,它们摆脱人的直接监控,以及对人类福祉的潜在影响,都是科学幻想的素材。50多年前,艾萨克·阿西莫夫(Isaac Asimov)就预见到需要伦理规则来引导机器人的行为。当人们思考机器道德时,首先想到的就是他的"机器人三大定律":

1. 机器人不可以伤害人;或者,通过不作为,让任何人受到伤害。
2. 机器人必须遵从人类的指令,除非那个指令与第一定律相冲突。
3. 机器人必须保护自己的生存,条件是那样做与第一、第二定律没有冲突。

然而,阿西莫夫写的是故事。他并没有遇到今天工程师们所要面对的挑战:要确保他们建构的系统对人类有利,并且不引起对任何人的伤害。阿西莫夫三大定律是否真的有助于确保机器人合乎道德地行动,这正是本书中我们要考虑的问题之一。

就在接下来的几年中，我们预测将会有一场灾难性的事故发生，它是由摆脱了人监控的机器人做决策所引发的。2007 年 10 月，南非军队使用的一种半自主性机器人加农炮出了故障，杀死了 9 名士兵，伤了 14 名——早期报道对究竟是软件还是硬件方面出了故障有不同说法。随着机器变得具有更加完备的自主性，发生更大灾难的潜在可能将增加。即使将要来临的灾难不会杀死如“9·11”恐怖袭击那么多的人，但是，它将激起同样广泛的政治反应；反应的范围将从呼吁更多投入于这些技术改进，到呼吁对这些技术的彻底禁止（即便不是一场完全“针对机器人的战争”）。

对安全和社会福利的关注一直是工程学的重点。但今天的系统正在趋近某种程度的复杂性，我们认为，这种复杂性要求系统自身做出道德决策——借用《星际迷航》（The Star Trek）的词，通过“伦理子程序”（ethical subroutines）来程序化。这就将道德主体的圈子扩大了，不仅有人类，还有人工智能系统，我们将称之为人工道德智能体（AMAs）。

我们并不确切地知道一场灾难性事故是什么样子，但是下面的故事或许可以让人有个大致的了解：

2012 年 7 月 23 日，这看起来是一个平常的星期一。在美国的大多数地区，温度也许略有些高，预计用电峰值会升高，但不会达到历史记录。美国的能源耗费正在增加，投机商们一直在驱动石油的期货价格以及现货价格向上攀升，使每桶接近 300 美元。在过去几周内，能源衍生品市场一些稍显不寻常的自动贸易活动引起了联邦证券交易委员会（SEC）的注意，但是银行已经向监管机构确认，他们的程序是在正常范围内运行的。

上午 10:15，东海岸，作为对巴哈马新发现的大型油田储备的响应，石油价格略有下降。橙色和拿骚银行（Orange and Nassau Bank）投资分部的软件计算出，如果给四分之一的银行客户发邮件，推荐其购买石油期货，暂时提升现货市场价格，然后，当交易商囤积供给以满足期货需求时，再卖空给银行的其他客户，这样就会获得利润回报。这个计划本质上是让消费者中的一部分人与其余的人对阵，这当然完全不合乎伦理规范。但是人们并未考虑这些细节为银行软件进行如此这般的设计。事实上，由计算机自主计划的赚钱场景都是许多独立的稳健性原则非故意的结果。程序员很难预估计算机策划这种方案的能力。

不幸的是，计算机直接发给客户的“推荐购买”邮件太有成效了。习惯看到油价节节攀升的投资者们激动地跳上这辆“时尚花车”，于是石油的现货价格骤然升到远超过 300 美元，而且没有下降的迹象。在东海岸，上午

11:30，温度的攀升也快过人们的预料。新泽西州电网的控制软件计算出，如果使用燃煤发电厂，而不是燃油发电机，就可以使能源耗费下降，从而能满足意外的用电需求。然而，一台火力发电机在峰值负荷运转时发生了爆炸，而且在任何人采取行动之前，连锁停电切断了东海岸一半地区的电力供应。华尔街受到停电影响，但停电之前 SEC 监管机构已经注意到，油价期货价格上涨是发生在橙色和拿骚银行自动交易账户之间的一场由计算机导致的骗局。随着这一消息的传播，而投资者又计划巩固地位，一旦市场重新开盘，很显然，价格势必急剧下降，损失将达到数百万美元。与此同时，停电已经遍及广大区域，导致许多人得不到必要的医疗救助，还有更多的人滞留在外回不了家。

由于监测到这场正在蔓延的停电有可能是恐怖主义行为，里根国家机场的安全检查软件自动设置到最高安全级别，启用生物识别比对标准，比平常更容易标识出嫌疑人员。这个软件没有权衡阻止一场恐怖袭击与由此给成千上万在机场的人带来麻烦之间的利益得失的机制；它识别出一共五名乘客为潜在的恐怖分子，这五个人正在等候飞往伦敦的 231 航班。系统将“嫌疑分子”高度定位在这趟航班上，于是开始封锁该机场，并将国土安全响应分队派遣至该航站楼。乘客们都紧张不安，231 航班舱门口的情形急转失控，便开枪了。

国土安全部给航空公司发出警报，提醒也许会发生一场恐怖袭击，于是许多航线实施测算以让飞机着陆。大量飞机试图降落在芝加哥奥黑尔机场造成一片混乱，一架公务喷气式飞机与一架波音 777 相撞，造成 157 名乘客和机组人员死亡；当飞机碎片落在阿灵顿高地芝加哥郊区时，又导致一个街区的房屋发生了火灾。

与此同时，设置在美国和墨西哥边界上的自动机关枪收到信号，被提升至最高警戒状态。自动机关枪的设置程序允许其在红色警戒状态下可以在无人直接监控的情况下实施侦察，并消灭潜在可能的敌情目标。其中一挺机器人机关枪朝着从亚利桑那州诺加利斯（Nogales）附近越野旅行回来的一辆悍马开了火，摧毁了这辆车，并杀死了三位美国公民。

此时，东海岸恢复了电力供应；几天后市场也重新开盘，几百人丧生和数十亿美元的损失可以归咎于这些多重交互系统各自单独运行的程序决策的结果。然而，人们感受到的影响还要持续数个月。

或许时间证明我们是糟糕的灾难预言家。我们预言这种灾难的意图并不是为了引起轰动或者灌输恐惧。这也不是一本关于技术恐怖的书。

我们的目标是以建构式引导 AMAs 工程设计任务的方式来拟定一个讨论框架。我们预言的目的是使人们注意到，现在就需要开始道德机器的工作了，而不是等到二十年至一百年之后技术追上了科学幻想的时候。

机器道德拓展了计算机伦理学领域关注的问题，从关注人们用计算机做什么到机器自身做什么。[本书中我们使用的术语伦理(ethics)和道德(morality)可以互换。]我们正在讨论的技术话题涉及使计算机自身成为清楚的道德推理者。随着人工智能(AI)扩展自主智能体的范围，如何设计这些智能体，使其尊重更为广泛的人类道德主体需要尊重的价值和法律，这个任务越来越紧迫了。

人类真的想让计算机做出道德上重要的决策吗？许多技术哲学家已经警告人们，不要把责任推卸给机器。电影和杂志充斥着对高级人工智能所造成危害的未来狂想。新兴技术在变得根深蒂固之前，总是更容易修正的。然而，在人们广泛接受一项新技术之前，又时常不大可能准确预测它的社会影响。因此，一些批评家认为，人们宁可慎之又慎，放弃发展有潜在危害的技术。然而我们相信，市场和政治力量会说了算，它们将要求这些技术提供好处。这样，任何与此技术利害相关的人都应该义不容辞地直面解决这个问题，使计算机网络中的电脑、机器人和虚拟机器人实施道德决策。

如前所述，本书不是讲技术恐怖的。是的，这些机器要来了。是的，它们的存在会对人的生活和福祉产生意想不到的影响，这些影响不会都是好的。但是，我们相信对自主系统日渐增加的依赖，不会削弱基本的人性。我们的观点是，高级机器人也不会奴役或者灭绝人类，就如科幻的典型模式那样。人类总是会适应他们的技术产品，与自主机器交往的人得到的好处很可能超过其为此付出的代价。

然而，这种乐观并非凭空而来。不可能袖手旁观就能让事情变好。如果人的本性是要避免不良的自主人工智能体所产生的后果，那么人们就必须做好准备，认真想一下如何使智能体向好的一面发展。

在建构道德决策机器方面，我们是依然沉浸在科幻世界里，或者更糟，打上那种时常伴随人工智能科学狂想的烙印吗？只要我们还在做着关于 AMAs 时代的大胆预言，或者还在声称会走路、会说话的机器将取代目前做出道德指引的人类“只是时间问题”，这样的指责就是合适的。然而，我们不是未来学家，并不知道这些看似通向人工智能的技术障碍是真的还是幻觉。我们也没有兴趣去猜想，当你的咨询师是一个机器人时，你的生活

是什么样，甚至没有兴趣去预测这是否终究会发生。相反，我们的兴趣来自于那些当前技术的渐进步骤，它们表明了对伦理决策能力的需要。也许，逐渐的进步最终会促成成熟的人工智能——希望是《2001：太空漫游》(*2001：A Space Odessey*)中 HAL 的不那么残忍的版本——但是，即使完全智能系统依然遥不可及，我们还是认为存在一个工程师面对的真正的问题，但他们无法独自解决。

引入这个话题是不是太早了？我们不这么认为。用于承担重复性机械任务的工业机器人已经造成了人受伤甚至死亡。对家庭和服务机器人的需求，预计将产生一个世界范围的市场，至 2010 年，将是工业机器人市场的两倍；至 2025 年，将是其四倍。随着家庭和服务机器人的出现，机器人不再被限制在仅有受过训练的工人可以与其接触的可控工业环境中使用。例如，索尼公司的小型机器宠物“爱宝”(AIBO)是更大型机器人应用的试水者。数百万个机器人真空吸尘器，如 iRobot 公司的“Roomba”，已经被卖掉了。用于医院导诊、博物馆导引的初级机器人也已经出现了。相当多的关注正在投向发展服务机器人，用来完成基本的家务劳动并帮助老年人和居家者。计算机程序以人类无法复制的效率发动数百万笔金融交易。几秒之内，软件做出购买，然后再转售股票、商品和货币的决策，没有人能实时侦查到其所开发的获得利润的潜能，它代表着世界市场中很大一部分的活动。

自动金融系统、机器宠物以及机器人真空吸尘器距离科幻场景还有很长的路要走，在那里，完全自主的机器所做的决策极大地影响人类福祉。尽管 2001 年已经过去，阿瑟·克拉克(Arthur C. Clarke)的 HAL 依然还是一个幻想，而且也完全可以肯定，《终结者》(*The Terminator*)中的世界末日情节在 2029 年其最迟销售日期到来之前将不会实现。说《黑客帝国》的情景到 2199 年还不会出现，或许就不十分保险了。然而，人类已经处在这个节点上了，此刻工程系统做的决策能影响人类的生活，并产生复杂而难以预料的伦理后果。在最坏的情况下，它们会产生深远的消极影响。

建构 AMAs 可能吗？或许具有人类全部道德能力的全意识人工系统将永远留在科幻世界。然而，我们相信，更多能力有限的系统不久将会建成，这样的系统将有评估其行动的伦理后果的某种能力——例如，是否为了保护隐私权去破坏产权。

设计 AMAs 的任务要求我们认真审视，这源于以人为中心的视角的伦理理论。这个世界的宗教、哲学传统中所表达出的价值和关心并不易于