Interdisciplinary Science Letters 交叉科学快报, 2022, 6(2), 35-40
Published Online June 2022 in Hans. http://www.hanspub.org/journal/isl
https://doi.org/10.12677/ISL.2022.62004

Hans汉斯

如何破解自动驾驶汽车的推广困境

——以《变形金刚》系列电影为例

杨　旭

湖南医药学院马克思主义学院，湖南　怀化

收稿日期：2022年5月27日；录用日期：2022年6月19日；发布日期：2022年6月29日

摘　要

在人工智能广泛运用的今天，自动驾驶汽车无疑最先闯入人们的道德视野，并且以息息相关的方式影响现代人的生活。但因为自动驾驶汽车的自身硬件故障和部件失灵的先天条件，始终无法避免“零事故”出现。因此，在注入算法设定的初期，就应充分考虑到自动驾驶汽车可能遇到的伦理问题，如汽车自动驾驶的主体归责、事故算法和道德算法存在的张力、道德伦理和现实法理立论点差异等问题。在现实中，人们创造出站道德立场而能全盘接受的自动驾驶汽车似乎是不现实的，因此更应关注的是在既定功能行为下创造出人们普遍接受的道德伦理阐释，将更多的选择权交还给人来把控，而不是完全交与自动驾驶汽车来抉择。

关键词

人工智能，道德主体，道德困境，自动驾驶汽车

Research on How to Overcome the Promotion Dilemma of the Autonomous Vehicles

—Taking the Series Films of *Transformers* as an Example

Xu Yang

College of Marxism, Hunan University of Medicine, Huaihua Hunan

Received: May 27th, 2022; accepted: Jun. 19th, 2022; published: Jun. 29th, 2022

文章引用: 杨旭. 如何破解自动驾驶汽车的推广困境[J]. 交叉科学快报, 2022, 6(2): 35-40.
DOI: 10.12677/ISL.2022.62004

Abstract

With the extensive application of the artificial intelligence nowadays, autonomous vehicles are undoubtedly firstly brought into people's moral vision. Besides, modern people's lives have been closely influenced by them. However, it is always impossible to avoid the emergence of "zero accident" under the impact of the congenital conditions of hardware failure and component failure of autonomous vehicle. Therefore, the ethical problems that may be encountered by autonomous vehicle should be considered at the initial stage of setting the injection algorithm, such as the principle of liability attribution of autonomous vehicles, the tension between accident algorithm and moral algorithm, the differences between morality and ethics and realistic legal arguments. In reality, the creation of an autonomous vehicle that stands on a moral standpoint and can be accepted completely seems unrealistic. Therefore, what should be focused on is the creation of a generally accepted moral and ethical interpretation under the established functional behavior, so as to give more choices to people to make the control instead of letting autonomous vehicle completely make the selection.

Keywords

Artificial Intelligence, Moral Subject, Moral Dilemma, Autonomous Vehicle

1. 引言

于 1984 年上映的动画《变形金刚》是历史上最成功的商业动画之一，最初改编自日本 TAKARA TOMY 公司的戴亚克隆和微星系列，后由美国孩之宝公司主导开发。系列故事主要讲述的是，在遥远的宇宙中有一个星球——赛博坦，该星球上生存着一支金属生命体种族，他们在近百万年间分化成正义的汽车人和邪恶的霸天虎，双方进行了数百万年的战争，星球能源耗尽霸天虎追随汽车人来到了地球。在上映的系列电影中，主要讲述了这支金属生命体种族在地球上的各种遭遇，以及与人类之间的接触交往。

自 1955 年达特茅斯学院会议上正式提出“Artificial Intelligence”的概念之后，人们对于人工智能探索的脚步便从未停歇。《变形金刚》系列诞生于人工智能大发展的第二次浪潮中，1980 年至 1987 年期间，专家系统有了较大程度的发展。专家系统是一个智能计算机程序系统，其内部含有大量的某个领域专家水平的知识与经验，能够利用人类专家的知识和解决问题的方法来处理该领域问题。也就是说，专家系统是一个具有大量的专门知识与经验的程序系统，它应用人工智能技术和计算机技术，根据某领域一个或多个专家提供的知识和经验，进行推理和判断，模拟人类专家的决策过程，以便解决那些需要人类专家处理的复杂问题，简而言之，专家系统是一种模拟人类专家解决领域问题的计算机程序系统。由于此问题与本文无太大关联，所以不再做过多展开。

从最初的动画系列到 2007 年上映的真人电影《变形金刚》以及后续的五部电影作品，我们能够从故事中发现，这些在人类社会中只是作为工具出现的机械，无论是工程机械还是战争机械，他们拥有了和人一样的性格、情感、道德观念，乃至是生命。脱离故事内在剧情和艺术创作，从哲学的层面和角度来重新审视这个故事，大抵能够认为这是人们对于机械智能或是人工智能进一步发展的又一设想，以及对

于它们“拥有道德”的思考和假设。

2. 人工道德主体

无论是在涉及到人工智能的科幻艺术作品中，亦或是在真正的社会环境下，其中大部分的人工智能主体会有一个比较明显的特征，那就是行动模仿人的行为方式。这是一种原初性的试探，类似于早期对于飞机的探索研究源自于鸟类、声呐技术来自于蝙蝠，等等。我们不排除智能生命体有不同于人类的存在模式，但现阶段的研究，还是要回归到人本身。

“机械”一词起源于希腊语中的“Mechine”及拉丁文“Machina”，最早的“机械”定义为古罗马建筑师维特鲁威在其著作《建筑十书》，主要对于搬运重物发挥效力的机械和工具作了区别。[1]早期人们对于机械的运用可简单总结为“省力”和“便捷”，到了 20 世纪 60 年代，机器人技术兴起，越来越多的机器人进入到人类社会生活的方方面面，它们不仅能够代替人类进行高强度、高危险的工作，在服务业中，技术不断得到优化的机器人也开始拥有更多优势。

但无论是哪一种机器人，作为人工智能，他都无法避免与人共同存在、服务于人类的宿命，那么，在它发展进步的同时，不可避免地需要加入更多的“属人元素”，比如道德。对此，阿西莫夫曾提出过“机器人三原则”：

零原则：机器人必须保护人类的整体利益不受伤害。

第一原则：机器人不得伤害人类个体，或者目睹人类个体将遭受危险而袖手不管，除非违反了机器人第零原则。

第二原则：机器人必须服从人给予它的命令，当该命令与第零原则或者第一原则冲突时除外。

第三原则：机器人在不违反第零、第一、第二原则的情况下要尽可能保护自己的生存。

但在目前的社会情况下，人工智能对于人类的影响还远达不到上述原则要求的程度。但我们可预见的是，机器人与人类社会的融合程度在日益加深，这就从客观上要求赋予机器人更多的自主性，但同时，也要用道德来限制这种自主性。[2]

在电影《变形金刚》中，汽车人和霸天虎所拥有的语言、性格等，是它们从人类的互联网上下载数据学习获得的。那么，我们可简单地将汽车人和霸天虎同比看作是具有更高自主性的人工智能，此二者在拥有同样的自由情况下遵循着不同的道德准则。汽车人所奉行的道德观念基本符合“机器人三原则”，在该原则束缚下，它们所呈现出的是比人更加具有道德。然而，当我们赋予人工智能更多的自主性，同时脱离一定规则束缚，它们的表现或许是善的，但也有很大可能性是趋向于恶。温德尔·瓦拉赫(Wendell Wallach)认为：“没有任何理由认为人工智能需要像人类一样，然而，人类是目前唯一能够让具有一般智力的生物做出道德判断的主体。”在现实中，人常常根据本身来构建人工道德主体，但这很大程度上低估了当前人类道德行为模式的碎片性。与伦理学家试图阐明道德的“应该”或认知科学家旨在揭示影响道德心理的机制的研究截然不同，但构建人工道德主体需要同时利用了二者。伦理学家和认知科学家倾向于强调特定的认知机制在道德决策中的重要性，例如：重读、道德情感、启发式、直觉或道德算法。然而，自下而上地构建一个能够容纳道德决策的系统，会让人们注意到在磨练道德智商方面更广泛的机制的重要性。人工道德主体不需要模仿人类的认知能力，就能令人满意地对具有道德意义的情况作出反应。但是，通过建立人工道德主体的方法开展工作，将对加深人们对许多有助于培养道德敏锐性的机制以及这些机制协同工作的方式的认识产生深远影响。

所以，在塑造能与人类社会完美融合的人工智能时，我们还是应该回归人类需要的本身，将碎片化的人类道德分类整合。并且，我们更应认识到，真正为人们所需的人工智能不应是一个全盘接受人类道德的人工主体，而是在其既定功能下的行为要符合人的道德的人工智能。

3. 人工道德主体的自主性

为什么要赋予人工智能道德，我认为是由于其在面对关乎人的抉择时拥有更高的自主性。在电影《变形金刚》中，可看做是高阶人工智能的汽车人和霸天虎所拥有的自主性是类人的，他们的自由程度甚至超越人，从个体行为来说，几乎达到了不受法律约束的程度。一个远强大于人类的主体脱离了道德本身，其所带来的灾难是无法估量的。

但目前，人们所面对的人工智能无法达到这样一个层次，那我们就从目前的具有自下而上方法的人工智能切入讨论。

由谷歌(Google)旗下 DeepMind 公司戴密斯·哈萨比斯领衔团队开发的阿尔法围棋(AlphaGo)是第一个击败人类职业围棋选手、第一个战胜围棋世界冠军的人工智能机器人，其主要工作原理是“深度学习”。这一概念由 Hinton 等人于 2006 年提出，基于深度置信网络(DBN)提出非监督贪心逐层训练算法，为解决深层结构相关的优化难题带来希望，随后提出多层自动编码器深层结构。在今年 1 月 25 日，该公司向世人展示了他们的电子竞技人工智能机器人“AlphaStar”，在此前的十场秘密比赛中，AlphaStar 以十比零完胜《星际争霸 2》职业选手 MaNa，虽然在后来 MaNa 在直播中赢下一句，但他坦言“钻了空子”，表示“假如对手是人类，一定不会犯这种错误”。

无论是围棋棋谱的学习还是对于游戏的深入了解，人工智能在其中都是拥有一定的自主性的。通过深度学习，了解分析数以亿计的可能性后再进行操作，整个过程都不需要提前写进既定程序，而完全是人工智能的自发行为。甚至，除二者之外，有些人工智能甚至衍生出了其独特的语言和算法。

我们到底应该赋予人工智能多大程度的自主性呢？尼克·波斯特洛姆(Nick Bostrom)和艾利泽·尤德考斯基(Eliezer Yudkowsky)认为：“创造思维机器的可能性提出了一系列的伦理问题，这些问题既与确保这种机器不伤害人类和其他道德上关联的存在者有关，也与机器自身的道德地位有关。”[3]在电影《变形金刚》中，人类利用法律条款和暴力方式对那些造成对社会造成危害的外星来客加以限制和处罚，这是一种法律层面上的约束。放在电影故事中，由于这些外星来客拥有和人类相同的情感、思维甚至是生理感官，所以这种约束是有力的。但在当下甚至是可预见的未来，人工智能很拥有与人类相同或相似的情感和生理体验，那这种基于对肉体加以限制的责罚是否能够产生威慑力？对人工智能进行清除记忆、机身销毁等是否能与人类死刑一样具有同等效力？我想这几乎是不可能的。人工智能的核心是它的算法，以及它所依赖的自下而上的学习方式，所谓的“肉身”只是它众多载体的一个表现形式。在非常多的科幻片中都有这样的机器人，例如《流浪地球》中的领航管理机器人 Moss，或是《钢铁侠》中的管家机器人 Jarvis，这些以数据为依托的人工智能并不在乎它的载体，甚至是它自己的核心本身。同时，利用“深度学习”发展自身的人工智能，其进步速度远超过人类的发展进步速度，如果人工智能脱离人类的控制或者凭借其更强大的能力反制于人时，人们可能连摧毁其核心都做不到。

所以，从根源上就应赋予人工智能有限的自主性。如果一个制造曲别针的人工智能强大到了一定程度，那它会不会把全世界都变成曲别针？其实这个问题很好解决，那就是在它发展之初就告诉它生产曲别针的上限数量是多少，到这个数量就停止生产，然后规避掉它认知到“我要突破”的可能性。

4. 完全自动驾驶的自动驾驶汽车的推广困境

上文我们提到了赋予人工智能自主性，更为具象的表现就是飞机、汽车等自动驾驶技术的出现。这种与人交互更为密切的人工智能，通常会被委托做出涉及乘客和第三人生命的决定。

1967 年，菲利帕·福特(Philippa Foot)发表的《堕胎问题和教条双重影响》中，首次提到了“电车难题”，这成为了伦理学五大难题之一。这一难题在自动驾驶汽车的应用中有着更为独特和直接的体现，特别是，自动驾驶汽车在即将发生碰撞的时刻可能做出的决定，涉及到三种迫在眉睫的不可避免的伤害的情况：

第一，自动驾驶汽车可以保持在航线上杀死几个行人，也可以突然转向杀死一个过路人；第二，自动驾驶汽车可以保持在航线上杀死一个行人，也可以突然转向杀死自己的乘客；第三，自动驾驶汽车可以保持航线并杀死几个行人，也可以突然转向并杀死自己的乘客。在所有这些场景中，共同的因素是对人的伤害是不可避免的，所以需要做出选择，到底是乘客、过路行人还是路边的人受到伤害。

目前的自动驾驶汽车还未突破第三等级，无论是在操作上亦或是法律上，都将其默认等同为人类驾驶汽车，在事故认定中，也会由驾驶人来担负责任。但达到完全自动驾驶的程度后，乘客不进行任何的操作行为，那么再由其承担责任显然是不合理的。

在自动驾驶汽车应该执行相同的强制性道德设置还是应该让每个乘客选择自己的个人道德设置的问题上，摩尔(Muller)等人认为，个人道德设置的优势，但强制执行强制性道德设置实际上最符合整个社会的利益。弥勒(Millar)则观察到技术可以作为道德主体，实现道德选择。他认为，用户/所有者，而不是设计师，应该为这些选择负责，特别是“设计师……应该合理地努力为自动驾驶汽车增加选项，让用户自己选择”。乔瓦尼·萨尔托尔(Giovanni Sartor)、弗朗西斯卡(Francesca Lagioia)和朱塞佩(Giuseppe Contissa)以此为前提，通过权重演算的方式为自动驾驶汽车设计了一个“道德旋钮”，可以由乘客自主设置自动驾驶汽车的道德。

我们在详细分析“道德旋钮”之后会发现，它的作用是具有欺骗性的。博纳丰(Bonnefon)等人在 2016 年的一项研究显示，通过 2015 年 6 月进行的三次在线调查，人们认为自动驾驶汽车应该被编程来最小化死亡人数，即采用功利主义方法(总损失最小化)。然而，参与者更倾向于乘坐能够优先保护乘客的汽车。矛盾的是，大多数参与者似乎更喜欢其他人使用实用的自动驾驶汽车，而作为乘客，他们中的每一个人都会做出更自私的选择。与此同时，虽然更多人倾向于功利主义的算法，但大家又不接受被强制植入功利主义算法的自动驾驶汽车。

这点同样是受到内外群体影响的表现。人的自私心使得他们在做出选择时，将自己置身于第三方路人的角度，这样的身份在面对自动驾驶汽车时自然是希望它能够执行功利主义算法。但购买自动驾驶汽车则是将自己置于乘客的角度，同样是出于自私心，人们会希望自动驾驶汽车能保护自身的安全。从本质上来说，这两种选择是出发点是相同的，就是人们对于自身利益能够实现最大化的追求。在电影《变形金刚》中，人类希望汽车人带来庇佑的同时，又不希望它们将战争引向地球，这都是出自于对于自身利益追求的自私心。但是，将这样的期望寄托于依赖于算法的人工智能是不合适的。就像即便是“道德旋钮”提供了多种可能，但大多数乘客仍旧会选择最利己的那一种。

但“道德旋钮”是必要的，它所带来的不仅仅是一种道德选择，更是碎片化的道德价值取向整合，以及信任感和安全感。《变形金刚》中，人类虽然与汽车人结为同盟，但本质上是不信任汽车人的。内外群体的划分导致内群体个体对于外群体成员产生偏见甚至是敌意。以及，陌生感带来的恐惧会致使人产生不信任——不管是外群体或是新生事物——因为显存的事物已满足于日常生活的运作，新事物的产生所带来的变动很可能会打破平衡致使自身遭受损失。当自动驾驶汽车突破第三等级后，如何让大众接受以致信赖一个拥有高自主性的自动驾驶汽车的关键就在于此。一个无法像人类一样能为自身行为负责的主体，不应该被视作是和人类同样的主体来看待。涉及到人本身或是其他生命体的问题，答案还是该交给参与者来决定，甚至是只由直接参与者来决定，以避免间接参与者可能的潜在犯罪动机。“道德旋钮”就是将决定权重新交予人类手中，自动驾驶汽车只是一个决策的执行者，不论是最后倾向于功利主义还是罗尔斯主义，都是由人来践行道德。在此基础上配套实施的法律法规，它所做出的限制或是处罚都可以回归到人，而不是约束一个没有财产、没有感情、没有生命的人工智能。我想，如果一个自动驾驶汽车在完全自主的情况下杀害了路人，只是将它销毁是无法抚慰受害者家属乃至是社会上众多人的愤

怒的。并且，一旦这种情况发生，将摧毁人类对于完全自动的自动驾驶汽车的信任。

想要自动驾驶汽车——乃至更多的人工智能融入人类社会，最重要的是要培养大众对于它们的信任感。这种信任不仅仅来自于人工智能的稳定性，更来自于人本身对于陌生的、强大的事物的控制和了解。人工智能应该被赋予更多的自主性，但这种自主性要始终受到限制，直至有一天它们能够完全地为自身行为负责时，才可以考虑是否要让它们拥有和人一样的权利和地位。

5. 结论

在电影《变形金刚》中，人类就是以法则的方式来约束汽车人，但它们的存在其实是完全等同于人类的。但根据以上分析所见，现实中的人工智能由于其本身不同于人的生理机能、情感机能，所以，想要单纯地用法律来限制它们是不现实的，因为它们无法为自身行为付出代价，法律对于肉体、财产的处罚对它们产生不了任何威慑力。

所以，更应该从根源上就应限制人工智能的自主性。就像对待孩童一样，在它无法为自身行为负责的时候，它就不能被看做是一个完全意义上的“人”来对待，他们的行为要由监护人来进行约束，他们所造成的恶劣后果也要由监护人来承担责任。自动驾驶汽车即便是达到了完全自动的最高等级，它仍旧无法为自身的判断负责，所以，人类这个“监护人”就必不可少。只有完全将陌生的新生事物掌控在手中，才能更好地突破人们对于它的恐惧和不信任。

参考文献

[1] 张善平. 机器人伦理问题研究[D]: [硕士学位论文]. 成都: 成都理工大学, 2017.

[2] 朱定局. 基于人工智能伦理规则修订的伦理仿真实验方法和机器人[P]. 中国专利, CN112508195A. 2021-03-16.

[3] 李伦, 孙保学. 给人工智能一颗“良芯(良心)”——人工智能伦理研究的四个维度[J]. 教学与研究, 2018(8): 72-79.