文章编号: 1000－8934(2020)011－0054－07
DOI:10.19484/j.cnki.1000-8934.2020.11.010

基于人机联合行动体的责任归因

郭 菁

(大连理工大学 人文与社会科学学部哲学系,大连 116024)

摘要: 自主机器的出现使马蒂阿斯认为,以控制作为责任归因的条件导致责任鸿沟。然而问题的关键不在于责任归因的控制条件,而在于行动体的变化。通过区分行动事件的可控制性和行动原因的可控制性,可以判定行动原因的可控制性是责任归因的充分必要条件。据此,在具有关系性、交互性和动态性的人机联合行动中,行动的原因既非仅出自于人,也非仅出自于机器,而是出于人机联合行动体。因此需要按照人机联合行动的特征,建构超越实体形式、区域分布、动态生成中的人机联合责任体,共同承担联合责任。

关键词: 控制; 责任鸿沟; 责任归因; 人机联合行动体; 人机联合责任体

中图分类号: N031 文献标识码: A

随着机器自主性的提高,"谁为机器自主决策的结果负责"的问题成为争论的焦点。传统观点认为应由机器的生产者或操作者负责,然而反对者认为人类无法控制自主机器的行动,不能担负相应的责任。那么该让机器来负责吗? 学界至今争论未决。马蒂阿斯(A. Matthias)指出,如果我们不能回答"谁(who)或者什么(what)来负责"这个责任归因(responsibility ascription)问题,会导致无人负责的责任鸿沟(responsibility gap)。而问题"如果不正确解决的话,不仅会损害社会道德框架的一致性,也会危害法律责任概念的基础"[1]176。因此,责任归因问题再次成为学界的焦点。

一、责任鸿沟与责任归因的条件

马蒂阿斯认为,责任鸿沟根源于传统以控制(control)为条件的责任归因方式。如果我们"以控制原则作为责任的必要条件"[1]175,那么自动学习装置"使得传统的责任归因方式不再与我们的正义感和社会道德框架相容,因为没有人足以控制机器的行动从而能够为这些行动承担责任。这些情形就构成了我们要称作的责任鸿沟"[1]177。然而责任鸿沟的出现并不在于责任归因的控制条件,而在于当今人工智能社会的行动体的改变,鉴于这种改变,我们主张既非由人也非由机器单独承担责任,而应把责任归因于人机联合行动体。为此需要对"控制"这一条件进行分析。

在关于责任归因的讨论中,"广泛认为某种自由是道德责任的必要条件。而困难在于这一自由的界定。一种界定方式所根据的是我们对选择或行动的控制形式。"[2]210根据这种方式,控制作为"可供选择的自由"成为道德责任的必要条件。控制则意味着"具有能够选择行动的可能性"。"当我们要求一个主体具有控制时,指的是必须存在着一个可供选择的序列,其中主体的选择和行动乃是其品格或实践推理的结果。"[3]31于是,应不应该负责的问题便成为能不能控制,即是否存在着可供自由选择的行动可能性问题。主体只有在存在可供选择的可能性并能够自由地做出选择的条件下,才是可负责的。

控制作为责任归因的必要条件也被表述成"可

收稿日期: 2020－5－12

基金项目: 北京市社会科学基金青年项目"列维纳斯他者思想的人本维度研究"(15ZXC016);教育部人文社会科学研究青年基金项目"莱布尼茨科学符号理论及其价值研究"(16YJC720008);大连理工大学引进人才科研启动项目"莱布尼茨逻辑思想研究"(DUT19RC(3)031)。

作者简介: 郭菁(1979—),女,山东莱芜人,博士,大连理工大学人文与社会科学学部哲学系副教授,主要研究方向: 科技伦理、科技哲学。

供选择的可能性原则(the principle of alternate possibilities)”,即“仅当一个人能够做出不同的行动时,他才能为其所为负道德责任”[4]829。以这个原则作为道德责任的必要条件,“一般认为一个人被强制去做某件他无法自由选择的事情时,就可以不为其所为负道德责任”[4]830,如相容论和不相容论都认为“当我们把责任与控制关联时,通常意味着仅当对一个主体存在藉以展开不同行动的可选择的开放序列时,他才可以对一个具体事件负责”[3]31。也正因此,相容论和不相容论在关于责任与决定论是否相容的问题上出现了分歧。“要理解这个争论,重要的是要看到‘可供选择的可能性原则’通常既被相容论也被不相容论所接受。”[3]25不相容论认为,既然决定论与“可供选择的控制原则”不相容,那就与以控制为必要条件的道德责任不相容。相容论则认为决定论容许一定的“可供选择的控制”,因而与道德责任相容。

然而当今为了辩护道德责任和决定论的相容性,相容论的一些学者开始反思控制这个“可供选择的可能性原则”。法兰克福(H. G. Frankfurt)指出,‘这条原则的合理性通常被认为不可辩驳,以至于一些哲学家甚至把它称作先验真理”[4]829,但是他认为“这个‘可供选择的可能性原则’是错误的。一个人完全可以对其所为承担道德责任,哪怕他别无选择”[4]829-830。如一个人做了他不可避免去做的某事,即使没有其他的选择,但如果这件事是他本意就要去做的,那么他必须为此行动负责。在此论证中,法兰克福做出了一个非常重要的区分,区分行动事件与行动原因。行动事件是行动发生的事实,事实是被决定的,“一个人不可避免地做了某事,这一事实是他做这件事的充分条件”[4]836,就此他确实没有选择的自由,即无法控制自己的行动;然而“这一事实在解释他为何做这件事上可能毫无作用”[4]836,因为行动事件(事实上做了某件事)不同于行动原因(为什么要做这件事)。法兰克福主张,只能根据行动原因判定道德责任。如果一个人本意要做这件事,即使他事实上没有其他的选择而必须做这件事,他也不能为此行动开脱责任。

通过对行动事件和行动原因的区分,法兰克福主张把责任和控制分离,控制不再成为责任归因的必要条件。但实际上至多只能把责任与基于自由选择的控制分离开来。因为行动事件和行动原因区分之后,还需要对行动事件的可控制性和行动原因的可控制性做出进一步的区分。行动事件的可控制性是基于自由选择的控制,前提是存在着其他可供选择的行动可能性。我们认为,行动原因的可控制性则更多地倾向于亚里士多德所说的出于意愿,亚里士多德认为“出于意愿的感情和实践受到称赞或谴责”[5]58。出于意愿关注的是行动产生的原因,即为什么要做出行动。如果一个行动出于自身的意愿而产生,即行动的原因取决于自身,那就是可控的,因而必须负责任。

需要注意的是,这里“控制”表明的是原因性关系。能控制表明行动的原因取决于自身因而必须负责;不能控制表明行动的原因不取决于自身从而可以免责。就此控制作为责任归因的条件,追溯的是行动的因果关系,并不意指自由选择。正如“亚里士多德并非将自由赋予行动者,‘取决于我们’的说法表明了原因性的责任。这样的行动者可以控制他们的行动;他们要为这些行动负责‘一个人要为那些做或不做取决于他的事情负责,如果他要为它们负责,那么它们就是取决于他的’。亚里士多德认为这样的责任对于赞赏和指责来说是必要的,他研究自愿性就是为了捕捉这个原因性关系”[6]147。也就是说,控制并不等同于自由选择。能不能控制只是表明行动的原因在不在自身,而非意味着行动者有没有自由去选择。

法兰克福只表明行动事件的可控制性与责任无关,即基于自由选择的控制不再作为责任的必要条件。法兰克福并没能彻底否定,控制是作为道德责任的必要条件。只不过需要对行动事件的控制和行动原因的控制进行区分,行动原因的控制仍然是道德责任的必要条件。责任归因必须依赖于行动原因,只有行动原因取决于自身,即是可控的才能担负责任。法兰克福甚至明确表明,如果一个行动的原因是出于自身的,就要承担道德责任。这说明,行动原因的控制也是道德责任的充分条件。因此,通过对行动事件和行动原因的区分,我们认为行动原因的可控制性是责任的充分必要条件。

也就是说,控制作为责任归因①的充分必要条

① 我们这里探讨责任归因,指的是为行动的结果负责。关于究竟应该为“什么”负责的问题,还存在很多争论,即应该为选择负责,还是为行动本身负责,或是为行动的结果负责?这里并不探讨这些争论,主要把责任归因限定在对行动结果的原因寻求上。

件，强调的是行动原因的可控制性，而不是行动事件的自由选择。只有行动的原因不能控制，才能对行动的结果免责。如果行动原因取决于自身，即使行动事件本身不可控，也必须为行动的结果负责。“的确，寻找责任上的因果关系的倾向被称为归因理论的一个基本原则。”[7]6责任归因的基本原则也恰恰说明了行动原因的可控制性是责任的充分必要条件。下面我们接着考察，在人工智能社会，按照这一条件，究竟应该“由谁(who)或者什么(what)来负责”这个责任归因问题。

二、归“人”(who)负责的困境

责任需要行动体来承担，在人工智能社会，行动体可以分为三类：人、机器、人机联合行动体。按照控制这一责任归因的条件，承担责任的行动体既不完全是人，也不是单独的机器，而应该是人机联合行动体。

传统认为只有人才能够承担责任。然而，传统观点忽视了人工智能社会行动的复杂性。人工智能社会是“人－机－社会”共生的新型社会形态，不同于传统由人与人的关系组成的社会。人工智能社会中的行动往往不是由人单独做出的，而是人机联合行动。

根据控制条件，如果行动原因取决于自身，就必须为行动的结果负责。只有行动的原因不能控制，才能对行动的结果免责。在人机联合行动中，行动的原因并不单单出于人，没有完全取决于人，不能仅仅要求人来承担责任。

一方面，智能机器可以在人机联合行动过程中独立做出决策，这种情况无法由人单独承担责任。智能机器的自主能力增强后，能够按照自身的规则行动。在和其他智能系统相互作用时，其行动往往取决于具体的环境和系统之间的互动，超出了科研人员和制造商能够准确预测和控制的范围。“合成智能不是通过传统意义上的编程得到的。你从各种各样、越来越多的工具和模块中拼凑素材、建立目标，把它们指向一系列实例，然后将其解放。最终系统会变成什么样并不可预见，且结果不受其创造者控制。”[8]3智能机器不受人类干预按照自身的规则做出行动，这就意味着行动不由人直接决定，而是超出了人的控制。由于人只有在行动原因可控的情况下，才能够为行动的后果负责。人在机器自主行动的情况中，失去了对机器的控制，无法要求人为机器自主行动的结果单独承担责任。

另一方面，人机交互行动的分布式状况，难以让人单独承担责任。人工智能社会的很多行动是分步实现的，每一次的行动，不论是问题的求解还是任务的执行都不是由单个行动体单独完成，而是通过多个行动体的相互作用分步实现。其实分步行动的整体效应，已经很难再按照主客二分的标准只把人当作行动主体，把机器当作客体，而恰恰是人机交互在一起，出现了主客一体化的融合。

在人机交互的分布行动中，如果仍旧把人和机器当作分离的系统，仅仅要求人来承担责任恰恰会造成责任鸿沟的困境。这是因为每个行动体只是执行自身独立的动作，然后再把结果报告给其他行动体，这些行动体通过交互关系才产生最终的行动。“每台机器的设计都有数百位工程师的贡献。不同的公司、研究中心和设计团队从事硬件和软件的单个组分的工作，构成最终的产品。这种计算机系统的模块化设计，意味着没有一个单独的个体或团队可以完全掌握系统与一个复杂新输入流进行互动或回应的方法。”[9]32从行动原因来看，每个行动体只能控制自己特定的行动，不能控制其他行动体的行动，也无法控制交互行动。鉴于这种情况，如果我们还孤立地看待这些行动体，把发生交互关系的行动体分离开，只关注人，就只能要求这个人对其自身的行动负责，却无法对其他行动体的行动或是整体交互的行动负责。

然而问题往往并不出现在某个单个行动体的行动上，恰恰出现在交互关系中。不利后果的发生是由于每个行动体只完成自身的目的，而不管其行动会对其他行动体造成什么影响，“这些系统是为了完成单一目的而设计的，它们不知道或者不关心其他副作用。”[8]70有的时候是任务之间的冲突产生了不利的后果。“你都无法知道两个或更多有着相反目的的自主系统何时会相遇。它们之间争斗的规模和速度堪比自然灾害。”[8]6有的时候任务并不冲突，但仍旧会产生不利的后果。因为最终的后果不是每个行动的简单累积，而是在互相牵动、互相作用的分布关系中产生出的交互效应，交互效应总是要大于单个效应的总和。对于这多出来的交互效应，由于无法要求单个的行动体来负责，就会出现无人承担、无人负责的责任鸿沟。

尤其人机交互的分布式关系还发生在多层次动态开放的时空。在空间上，不仅包括自然的物理空间、人类的社会空间还包括虚拟的信息空间，是“物理－信息－社会”三元融合的空间。在时间上，则是随着人机的交互关系持续动态变化的场域。人机交互关系在动态时空中的分布式展开，把责任归因的因果链加长。这条链的空间上分布的行动体越多，时间上变化的可能性越大，单个行动体的行动就离最终的结果越远，每个行动体就越难控制自己行动的原因和最终结果之间的联系，从而难以为行动的最终结果承担责任。

三、归“机器”(what)负责的困境

既然人不能单独负责任，那能否让机器来承担责任。正如建构人工道德智能体(artificial moral agents，简称 AMAs)的倡导者艾伦(C. Allen)和瓦拉赫(W. Wallach)所指出的“传统上，人类设计者和操作员要为机器的行为承担道德问责。这种做法在很多情况下还是适用的，但当出现由很多人或者也有别的机器共同决策的机器行为的时候，而且又暴露出某种程度上的不可预见性，这一做法还适用吗?”[9]186他们认为，与其寻找人来认定责任，倒不如建构人工道德智能体，让机器能够评价自身行动的结果并做出恰当选择。而且“理想的 AMA 会将外在和内在价值都纳入它选择和行动的考量当中。”[9]32然而这还只是一种理想。即使目前已采纳了“自上而下路线”、“自下而上路线”以及“自上而下和自下而上相结合路线”建构人工道德智能体，但远未实现目标。

需要解决的不仅仅是技术上的难题，更重要的是伦理上的问题。依据责任归因的控制条件，原因出于自身的行动必须承担责任。人工智能社会中的行动并不完全出于机器，无法要求机器单独负责。

首先，自治的智能系统并没有满足行动原因上的控制条件，无法主张智能机器成为责任主体。即使智能机器可以不受人类干预按照自身的规则做出行动，这只是在行动事件中体现出一定的自由和控制，但是行动事件的控制不同于行动原因的控制，行动原因的可控性意味着行动的原因出于自身。自治机器的行动原因不是出于机器自身，难以要求机器承担责任。也正因此，机器的责任主体地位至今无法确立。关于智能机器能否成为责任主体的问题一直是学界争论的热点。争论的一方认为，智能机器难以成为真正的责任主体，只有人才是责任主体。“人工智能体即使可以延展人类的认知，但却无法承担相关的责任，只有人类才能挑起责任重担。”[10]97－98争论的另一方则认为，智能系统可以承担责任。如卡普兰认为“如果合成智能有足够的能力可以感知到周围环境中与道德相关的事物或情况，并且能够选择行为的话，它就符合作为一个道德行为体的条件。”[8]80他认为只要智能机器能够感知道德情况并做出选择行动，便要对行动的后果负责。智能机器通过对环境的感知而选择行动的能力，只是行动事件的自由控制，而非行动原因上的出于自身的控制，并不具备承担责任的条件。只有在行动原因可控的充要条件下，才能要求机器为其行动负责。在智能机器还没有出于自身的行动之前，无法要求机器单独负责。

其次，在目标上，是否要执着于把机器建构成和人一样的责任主体，同时把这一目标建立在主客二分的前提之下。如果本体上认为人和机器是不同的，道德上却又要求机器和人一样，这本身就包含着矛盾。艾伦和瓦拉赫也看到了这一矛盾，“除非工程师有意去设计系统，否则，计算机智能就是建立在一个没有欲望、动机和目标的逻辑平台上的。而人类的认知能力却是由指引生存和繁殖的本能的情感平台演进而来的，并在这个平台上发展。两者间的这种差异凸显出了开发具有情感的计算机这一挑战的矛盾品质。”[9]125这种矛盾使得他们对建构完备伦理智能体这一目标做出了让步，“至于未来的机器人是不是‘真的’道德主体，这无关紧要。我们有可能设计、建造出能够考量法律和道德因素而去实施决策的系统，而且会比现有的系统更为敏感。”[9]196让步恰恰说明，目标不应该是让机器本身成为独立的责任主体。对人工道德智能体的倡导应该是考虑到人与机器不可分离的交互关系，在人机联合行动中建造出能够对人的行动做出正确反应，并恰当地采取行动的人工责任体。

最后，还会出现人借用机器来免责的现象。“随着机器在普通人的生活和工作中的普及，通过机器使行为和责任脱钩日益成为普遍现象。越来越多的人将失败的责任转嫁到机器身上。”[11]165但是在机器还不能成为责任主体的阶段，人的责任推

卸就会直接导致责任主体的缺失，出现责任鸿沟。即使未来的发展使得机器可以承担一定的责任，把责任推卸给机器又可能造成滥用机器为人类免责的后果，进一步加剧责任鸿沟。

四、建构人机联合责任体(who + what)

人工智能社会，智能机器在人类行动关系中的介入使得行动变得更为复杂，行动往往不是由人或者机器单独做出的，而是人机联合行动。人机联合行动的行动体是整体的，人和机器混和在一起，难以做出清晰的区分。如果再按照主客二分的思维，把人和机器区分开来，主张只有人或只有机器为联合行动的后果负责，就会导致责任鸿沟。对于人机联合行动，不能孤立地去归因人或机器的责任，而是要按照责任归因的控制条件找到承担责任的行动体，从而避免责任鸿沟。根据控制条件，如果行动的原因出于自身就要对行动的结果负责。在人机联合行动中，行动的原因出于人机联合行动体整体，要由人机联合行动体共同承担“联合责任”。

类似的观点如汉森(F. A. Hanson)提出的联合责任(joint responsibility)。汉森批判方法论的个体主义(methodological individualist)把人与技术分离，只承认人类是行动主体，相应地在道德上只认为人类能够为行动负责，主张道德个体主义(moral individualism)。为了克服这种分离，汉森提出了延展行动体理论，指出行动是在人与技术的关系中实施的，“行动的实体(entity)—主体(subject)—就是延展行动体(extended agency)”[12]93，延展行动体不仅包括人也包括技术。“当主体更多地被理解为是一个动词而非名词——即被理解为以不同方式结合不同实体从事不同活动的一个路径——自我与他者之间就不再有清晰的区分，而且这样的区分也不再重要。”[12]98他指出既然行动是由延展行动体整体实施的，那么“道德责任属于作为整体的延展行动体，而不是明确地归于其中的某个属人或非人部分。”[12]93汉森的联合责任强调联合行动的关系特征，促使行动体在行动因素之间相互依赖关系的基础上，采取行动并承担责任，这确实克服了个体主义的局限，并且回应了行动关系的动态开放特征，使行动体可以随着关系的变化不断生成并改变相应的联合责任。

同样，我们也主张要把人机联合行动体看作一个整体，并按照人机联合行动的关系性、交互性和动态性特点建构人机联合责任体，共同承担联合责任。

1. 按照人机联合行动的关系特点建构超越实体形式的人机联合责任体。

人机联合行动是人机结合在一起，互相牵动、互相作用的联合行动。联合行动依赖相互关系合作实现。从表面上看，每个行动体先执行自身的任务，再把结果报告给其他行动体，然后通过组合和合作产生最终效应；但实际上，每个行动体的行动都要受到周围环境和其他行动体的影响，且最终联合行动的组合效应大于单个行动体的行动效应。虽然每个行动体有自身的任务，但并不是独立执行的。没有一个行动体是绝对独立的，行动体都处于联合关系之中。这就使得行动的原因并不单独出于任何行动体，而是人机联合所致，行动的结果也是由人机联合行动造成。也正因此，根据责任归因的控制条件，需要由人机联合行动体整体来负责，不能再按照主客二分的模式，把人和机器相区别，要求人或者自治的机器单独成为责任主体。更确切地说，人机联合行动的责任体并不是某个明确的独立实体(人或者机器)，而是超越了实体形式，在虚拟和现实的空间由复杂的关系产生的人机联合行动体。

2. 按照人机联合行动的交互特点建构区域性分布的人机联合责任体。

人机联合行动体不是某个单一行动体，而是由多个行动体包括其他关系成分组成的联合整体。整体行动分布式地分散于不同的人和机器，通过交互影响合作完成。由于影响的因素众多，交互关系也很复杂，很难清晰地区分具体的责任。正如约翰逊(D. G. Johnson)和帕沃斯(T. M. Powers)指出，“任务分布于计算机系统中，使计算机系统的行为和人的行为融为一体，以致于无法加以拆分以归因道德责任。”[13]106正是考虑到交互的分布式特征，汉森认为，既然难以将人类行动和机器行动分开，那么在责任归因时便不需要区分，把人机系统看作一个整体来共同承担联合责任。但是这样来回应交互的分布式特征，就会出现贡克尔(D. Gunkel)所指出的集体逃避责任的情况，“道德责任和法律责任以这种方式分布在网络的要素中，以致于没有人、机构或技术应受谴责或承担责任。”[14]318

确实，汉森提出的联合责任过于强调交互的整体性，不对交互的组成部分进行具体的分析。只考虑行动的实施需要各组成部分的交互合作，超越了单一领域，呈现出分布式交互的难以分解的复杂样态，并认为“最终，这一延展行动体的不同部分是如何相互影响的，其中的每一部分展开了哪一部分行动，就变得无法分辨了。此时，唯一可靠的做法就是把责任归于整个延展行动体”[12]97。但这只能抽象地归因于整体，没能进一步的划分，提出具体的责任确立方案。正如贡克尔所指出的，“这种联合方法(combined approach)仍需要有人来决定和回答，哪些方面的责任属于机器，哪些应保留给或归于网络中的其他要素”[14]318。如果只笼统地诉诸整体，“最终，就会出现无人或无物对事情负责的局面”[14]318。

因此，我们认为不能只把人机联合行动体看作一个整体，而不做进一步的划分。由于行动是交互分布式展开的，责任归因根据的又是行动产生原因的可控性，因而可以根据交互行动分布的不同区域，按照不同区域的功能及对行动产生的因果关系，划分具体的责任区域，确定相应的人机联合责任体。比如交互行动产生的主要区域是主导行动区域，主导行动区域的人机联合行动体要承担主要责任；交互行动产生的附属区域是辅助行动区域，辅助行动区域的人机联合行动体要承担辅助责任等。这样，不再把人机联合行动体看作抽象的整体，而是通过具体的区域划分，确定不同区域的人机联合行动体的不同责任，避免责任鸿沟的出现。

当然还要考虑到交互行动的双向性。人机联合行动体中的交互行动是双向的，其中每个行动体的行动不仅依赖于周围环境和其他行动的影响，同时也改变着其自身的环境和其他行动体的行动。因此，责任归因不能由单向的关系来确定，而是取决于交互影响的双向关系。如人在机器的设计、研发、使用和管理中，不能把机器当作被动的工具只是单方面强调人对机器的作用，也要考虑机器对人的反向作用。我们主张要在人机交互的双向关系中，根据责任归因的控制条件，结合具体情境确定责任归因，通过分析各种原因、结果交互的体系结构，建构人机联合责任体的交互模型。

3. 按照人机联合行动的动态特点建构不断生成的人机联合责任体。

人机交互关系还发生在多层次动态开放的时空，时空的动态开放具有一种动态效应，由此产生的人机联合责任体不再是某种静态的永恒不变的实体，而是在“物理－信息－社会”三元融合的空间，在持续变化的时间中，不断生成超越了实体形式的动态行动体。不同于静态的实体模式，既不能根据某种实体及其所具有的属性来判定责任体，也不能根据现存实体间的外在关系来确定责任体；毋宁说，责任体是在具体的行动因果关系中动态生成的。生成是一个动态的过程，处于不断变化中。也就是说，随着人机联合行动关系的改变，责任体也随之改变。责任体是动态生成的，同样要求我们不再把责任限定于某个人或机器。而是在具体的情境中，通过不同的行动因果关系确定不同的人机联合行动体。即要立足“人－机－社会”共生系统，在人机联合行动的动态关系中，建构人机联合责任体的生成模式。

总之，根据责任归因的控制条件，行动的原因出于人机联合行动体，要由人机联合行动体整体承担联合责任，并按照人机联合行动的关系性、交互性和动态性特点建构人机联合责任体。由于这三个特点并不彼此分离，是人机联合行动同时具备的特征，人机联合责任体就不是抽象的整体，而是超越了实体形式区域性分布的、不断生成的整体。因此，需要在具体的关系情境中，根据动态的因果关系，按照行动分布的不同区域归因不同的责任，建构相应的人机联合责任体，从而避免责任鸿沟。

参考文献

〔1〕Matthias A. The responsibility gap: Ascribing responsibility for the actions of learning automata [J]. *Ethics and Information Technology*, 2004(6): 175－183.

〔2〕Deery O. Extending compatibilism: Control, responsibility, and blame [J]. *Res Publica*, 2007(13): 209－230.

〔3〕Fischer J M. Responsibility and Control [J]. *The Journal of Philosophy*, 1982, 79(1): 24－40.

〔4〕Frankfurt H G. Alternate possibilities and moral responsibility [J]. *The Journal of Philosophy*, 1969, 66(23): 829－839.

〔5〕[古希腊] 亚里士多德. 尼各马可伦理学 [M]. 廖申白，译注. 北京：商务印书馆，2003.

〔6〕[美] 梅耶尔. 亚里士多德论自愿 [C]//克劳特，主编. 布莱克维尔《尼各马可伦理学》指南. 刘玮，陈玮，译. 北京：北京大学出版社，2014：146－168.

〔7〕[美] 维纳. 责任推断：社会行为的理论基础 [M]. 张爱卿，郑

葳,等译. 上海:华东师范大学出版社,2004.
〔8〕[美]卡普兰. 人工智能时代:人机共生下财富、工作与思维的大未来[M]. 李盼,译. 杭州:浙江人民出版社,2016.
〔9〕[美]瓦拉赫,艾伦. 道德机器:如何让机器人明辨是非[M]. 王小红,译. 北京:北京大学出版社,2017.
〔10〕宋春艳,李伦. 人工智能体的自主性与责任承担[J]. 自然辩证法通讯,2019,41(11):95-100.
〔11〕[美]斯加鲁菲. 智能的本质:人工智能与机器人领域的64个大问题[M]. 任莉,张建宇,译. 北京:人民邮电出版社,2017.
〔12〕Hanson F A. Beyond the skin bag: On the moral responsibility of extended agencies[J]. *Ethics and Information Technology*, 2009, (11): 91-99.
〔13〕Johnson D G, Powers T M. Computer systems and responsibility: A normative look at technological complexity[J]. *Ethics and Information Technology*, 2005(7): 99-107.
〔14〕Gunkel D J. Mind the gap: Responsible robotics and the problem of responsibility[J]. *Ethics and Information Technology*, 2020, 22: 307-320.

Ascribing Responsibility for the Human-Machine Joint Actor

GUO Jing
(Department of philosophy, Faculty of Humanities and Social Sciences, Dalian University of Technology, Dalian, Liaoning 116024, China)

Abstract: As the autonomous machine emerged, Matthias questioned that control is a necessary condition of responsibility ascription because it leads to the responsibility gap. However, the responsibility gap is not caused by the control condition of responsibility ascription. The actor which could take responsibility has changed. Based on the distinction between the controlling events of action and the controlling reason for action, the latter is analyzed as the sufficient and necessary condition of responsibility ascription. According to this condition, for the relational, interactive and dynamic human-machine joint action, the reason for the action is out of neither human nor machine alone, but from the human-machine joint actor. It is required to ascribe joint responsibility for the human-machine joint responsible agency, which is regionally distributed, dynamic developed, and beyond the entity according to the feature of human-machine joint action.

Key words: control; the responsibility gap; responsibility ascription; human-machine joint actor; human-machine joint responsible agency

(本文责任编辑:董春雨 郑泉)