**Advances in Social Sciences** 社会科学前沿, 2022, 11(12), 5270-5276
Published Online December 2022 in Hans. http://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2022.1112717

Hans汉斯

# 人工智能刑事责任主体资格的审视

**马东硕**

大连理工大学，人文与社会科学学部，辽宁 大连

收稿日期：2022年11月2日；录用日期：2022年12月13日；发布日期：2022年12月20日

## 摘　要

人工智能的发展引发了学术界对其刑事责任主体资格的思考。综合人工智能在发展的不同阶段中呈现出的特点，可以将人工智能从整体上分为弱人工智能和强人工智能。弱人工智能因其程序和逻辑的形成依赖于人类活动而不具有独立意识和自主决策能力，在更多情况下表现为辅助性质的工具。而强人工智能则具备更强大的学习能力，其能够在人类所设计程序以外的范围凭借自己的意志做出决定并以此行动，具有一定程度的自主行为能力。基于人工智能技术水平持续进步的趋势，不断有学者提出了人工智能可能在未来会产生和人类一样，甚至是超越人类的思维能力和自主意识。学界对人工智能刑事责任主体资格的认定主要表现为肯定论和否定论。综合人工智能发展程度的不确定性、法律内在的逻辑性等多方面因素，笔者认为肯定论的观点实际上缺少足够科学的论证而不易形成有效结论。

## 关键词

人工智能，刑事责任主体资格，法律逻辑，刑事责任

# Examination of the Qualifications of Criminal Liability Subjects of Artificial Intelligence

**Dongshuo Ma**

Faculty of Humanities and Social Sciences, Dalian University of Technology, Dalian Liaoning

Received: Nov. 2nd, 2022; accepted: Dec. 13th, 2022; published: Dec. 20th, 2022

## Abstract

**The development of artificial intelligence has triggered academic thinking about its qualifications as a criminally responsible subject. The characteristics of integrated artificial intelligence in dif-**

文章引用: 马东硕. 人工智能刑事责任主体资格的审视[J]. 社会科学前沿, 2022, 11(12): 5270-5276.
DOI: 10.12677/ass.2022.1112717

**ferent stages of development can be divided into weak artificial intelligence and strong artificial intelligence as a whole. Weak artificial intelligence, because its programs and logic depend on human activities without independent consciousness and autonomous decision-making ability, is more often manifested as a tool of an auxiliary nature. Strong artificial intelligence, on the other hand, has a stronger learning ability, and can make decisions and act accordingly on its own will outside the scope of human-designed programs, with a certain degree of autonomous behavior. Based on the trend of continuous progress in the level of artificial intelligence technology, some scholars have proposed that artificial intelligence may produce the same thinking ability and autonomous consciousness as humans in the future. The academic community's determination of the qualification of the subject of criminal responsibility of artificial intelligence is mainly manifested as the theory of affirmation and the theory of negation. Considering the uncertainty of the degree of development of artificial intelligence, the inherent logic of the law and other factors, the author believes that the affirmative view actually lacks sufficient scientific arguments and is not easy to form effective conclusions.**

## Keywords

**Artificial Intelligence, Qualifications of Subjects of Criminal Responsibility, Legal Logic, Criminal Responsibility**



## 1. 人工智能刑事责任主体资格问题的提出

刑事责任主体资格的认定标准是讨论人工智能能否具备刑事责任主体资格的基础，而刑事责任主体资格则通过刑事责任能力表现出来。对刑事责任能力的判断逻辑通常表现为：主体有独立的意志－主体在意志的支配下拥有价值辨认以及行为控制的能力－主体能够承担法律规定的刑罚－主体具备刑事责任能力[1]。总体看来，认定标准可概括为两点：是否具备自由意志以及是否具备适应刑罚的能力(见表 1)。

**Table 1.** Standards for determining the qualifications of criminal liability subjects
**表 1.** 刑事责任主体资格的认定标准

| | 刑事责任主体资格的认定标准 |
|---|---|
| 自由意志 | 刑事主体行为的实施基于独立的思维和选择而未受到外界限制，其核心在于辨认能力于控制能力。前者不仅要求主体能够认识到事实因素，例如行为性质、后果等，还需要能够认识到来之规范层面的约束；后者倾向于对理性的强调，要求行为主体在认识到行为危害性和需要承担的责任后能终止自己的行为。例如：甲计划在途中抢劫己的财务，但出于对刑罚的畏惧而放弃抢劫。甲对财务产生的占有心理和财务所有权人是乙的认识体现为辨认能力，意识到抢劫行为的后果而放弃犯罪体现为控制能力。 |
| 刑罚适应能力 | 可以理解为犯罪主体能够承担与犯罪行为相适应的刑罚。对刑罚适应能力的考量需兼顾以下三个方面：其一，刑罚目的。刑罚目的的实现不仅要惩罚犯罪主体，也要预防其再犯罪；其二，刑罚内容。我国刑罚的种类可概括为生命刑、自由刑、资格刑以及财产刑，生命邢是对犯罪主体生命的剥夺，自由刑事对犯罪主体自由的限制，资格刑事对犯罪主体特有资格的限制或剥夺，财产刑事对犯罪主体财产的小件；其三，刑罚功能。刑罚的功能具有双重性，既要通过震慑让主体敬畏法律，也要通过教育让主体遵守法律。 |

在刑法领域中，越来越多的学者开始思考：人工智能在未来是否会具备类似人类的思考方式和行为

能力，在此基础上，人工智能是否会脱离人类的控制而产生自己的控制能力和辨认能力，如果答案是肯定的，刑法是否应赋予人工智能和人类同等的地位。事实上，自由意志和刑罚适应能力同样是考量人工智能能否具备刑事责任主体资格的主要标准，而关于这一问题，当前观点主要分为肯定论和否定论[2]。

## 2. 人工智能刑事责任主体资格的研究综述

近年来，学界对人工智能能否作为刑事责任主体的讨论成果主要可归为以下几点：其一，肯定论与否定论的对立与学者的价值选择相关，这也导致在讨论的过程中往往将问题的关键点聚焦于立场的选择而非对问题的论证；其二，“自由意志”成为两种观点的争论核心。肯定论者更多是着眼于人工智能与人类的共性特征，否定论者则多着眼于人工智能与人类的差异性特征；其三，刑罚的实施需要以人工智能受到人类社会规制为前提。以强人工智能为例，即便肯定其具备自由意志，但如果不能为相应刑罚措施所规制，则将其纳入刑法的规制范围也将失去意义；其四，判断刑罚设置的科学性与否需要以是否考虑到人工智能（特别是强人工智能）的特征为依据。

综合学界的各方建议来看，对于人工智能能否被赋予刑事责任主体资格的讨论与研究仍需秉持谨慎适度的态度，以立足于现有法律研究新问题为主要理念，从实证角度分析可行的完善要件。需要明确的是，在研究过程中应区分研究条件和立法条件，前者立足于确已产生或存在产生可能的问题或现象，在其发生时即可展开研究；后者则立足于社会发展的客观物质条件和社会环境，贸然推进立法活动的开展既属于对社会资源的浪费，同时也不符合刑法发展的客观趋势。此外，对人工智能刑事责任主体资格的讨论应以跨学科的视角推进，兼顾人工智能科学理论的自然性、社会性和技术性，并以技术性特征为首要因素，减少因仅过多关注刑法单学科理论而产生的局限性。

## 3. 人工智能能否作为刑事责任主体的观点之争

赋予人工智能刑事责任主体资格在学术上引发了广泛讨论，目前主要存在肯定论和否定论两种对立观点。二者均以传统的刑事责任主体理论以及人工智能的发展趋势为研究的出发点，但得出的结论却有极大差别。

### 3.1. 人工智能刑事责任主体资格肯定论

肯定论观点认为：在不久的将来，人工智能会产生独立的辨认能力与控制能力，其思维模式和行为方式将超出人类所设计的程序范围。正如华东政法大学刘宪权教授的观点：在可能(即将)到来的强人工智能时代，强人工智能存在在人类设定的程序外，依据其自身的意志实施具有社会危害性的行为。因此，刑法应当在刑事责任主体中增设强人工智能，并设置相关罪名实现对强人工智能的规制[3]。

肯定论观点的提出，主要依据以下几方面：

#### 3.1.1. 强人工智能与人类具有相似性

不同于弱人工智能更多体现出的工具性，强人工智能可以人类为其设定的程序为基础，通过不断学习形成自主意识，思维具有更高的独立性。在此情形下，强人工智能将拥有趋近于人类的思考方式和逻辑能力，换言之，人类与强人工智能的区别可能仅局限于形体和生理方面。因此，当人工智能拥有像人类一样的独立意识并能够在意识的指导下实施刑法所规制的行为时，则应认定其具有刑法意义上的辨认能力和控制能力，具备刑事责任主体资格。

#### 3.1.2. 人类对人工智能的控制逐渐弱化

科学技术的持续进步使人工智能被赋予人脑思维逻辑的可能性提高，并由此产生类似于人类的道德伦理和价值理念。部分学者认为，如果人工智能获得了可以在其道德观念的指引下进行理性思考的能力，

那么其取得刑事责任主体资格将不仅局限于理论层面。在这样的趋势下，人工智能将愈发呈现出脱离人类控制的独立性。尽管人类与人工智能仍然可以保持互动，但是当人工智能不受人类意志的干预而自主实施人类为其设定的程序以外的行为时，则应认定其为刑事责任主体。

#### 3.1.3. 人工智能存在产生自主意识的可能性

持肯定论观点的学者认为，“人工智能无法具备独立意识”的观点过于悲观，甚至在未来的强人工智能时代缺少合理性与前瞻性。算法、大数据等技术的发展为人工智能独立意识的产生提供了条件[4]，因此，在绝对意义上否定人工智能可以产生独立意识，继而认定人工智能无法取得刑事责任主体资格的观点缺乏足够的说服力。

#### 3.1.4. 将刑事责任主体资格赋予人类以外的主体具有可行性

我国现行刑法的刑事责任主体分为自然人和单位，而作为人类主体以外的单位主体，其对刑事责任的承担同样经历了从无到有的历程。在刑事责任主体资格的问题上，肯定论认为将人工智能类比单位具有合理性：单位由自然人组成，其意志相当于自然人意志的延展。而人工智能的初始行为源于人类设定的程序，程序是人类原有意志的体现，因此，人工智能实施的既定程序以外的行为也可以理解为人类意志的延伸，并且其与人类意志的紧密关系要远超于单位。以此类比，既然单位尚且可以作为刑事责任主体，人工智能自然也具备成为刑事责任主体的合理性。

### 3.2. 人工智能刑事责任主体资格否定论

持否定论观点的学者认为，即使未来进入到强人工智能时代，人工智能与人类都始终存在着本质区别。朱建华教授将人工智能定性为“犯罪工具”或“犯罪对象”，认为其无法脱离人类的控制，因此也无法取得刑事责任主体资格[5]。换言之，即便是可能拥有独立意志的强人工智能，也难以拥有与人类相仿的认知能力，无法成为能够被刑法所规制的适格主体。

否定论观点的提出，主要依据以下两方面：

#### 3.2.1. 人工智能不具备辨认能力和控制能力

弱人工智能能够在某些具体领域替代人类的工作，虽然其行为是人类程序设计的结果，但无需人类参与全部操作过程，具备一定的能动性。在弱人工智能时代，人工智能产品普遍不具备独立意识与自主决策的能力，刑法更多是将其归置为犯罪工具和产品。因此，对于其引发的危害结果，刑法只需追究相关研发者、销售者或使用者的刑事责任，而无需考虑其刑事责任主体资格。

判断强人工智能是否具备辨认能力和控制能力目前需要以人类为参照对象。由于人类对自身“独立意识”的认定尚且缺乏统一标准，则更不必说以客观统一的标准认定人工智能是否具备独立意识，继而也无法展开更深入的讨论与分析。此外，辨认能力和控制能力还体现在对价值的判断[6]。科技的发展使人工智能对客观事实的认知能力逐渐强于人类，但其价值判断能力可通过何种途径取得？其一，依靠人类在设计程序的过程中为其注入价值理念。这种途径本身就表明人工智能受制于人类而无法独立思考，并不具备辨认能力和控制能力；其二，通过自主学习获得人类的价值判断能力。结合当前的科技水平以及未来人工智能的发展趋势，人工智能暂时还不具备实现自主学习的能力，而且即便可以自主学习，也难以准确预知其对自身价值判断能力的认知水平，以及其产生的价值理念是否与法律规范、社会道德相适。因此，强人工智能具备等同于人类的价值判断能力仅停留在构想层面，至于其能够具备辨认能力和控制能力则更是缺乏足够的理论依据。

#### 3.2.2. 人工智能与刑罚不相适

否定论观点认为：无论是从现行的刑罚种类出发，还是参考肯定论者提出的新刑罚措施，人工智能

都缺少刑罚的相适性。

如前文所述，现行刑罚的种类主要分为生命刑、自由刑、资格刑以及财产刑。首先，生命刑无法适用于人工智能。一方面，人工智能不享有生命权，无法通过剥夺生命权的理念实现刑罚的目的。另一方面，即便将销毁人工智能的程序类比于人类适用的生命刑，但二者仍存在实质差异。人类的生命不可替代，在被剥夺后无法再生，而人工智能则可以通过对数据、程序等提前备份，在被销毁后实现再生；其次，对人工智能适用自由刑没有意义。自由刑不仅限制了人类身体层面的自由，同时也从多方面影响了人类的生活。例如，长时间监禁的结果是行为人各项技能退步、无法适应社会的生活方式等，这对于希望回归社会的行为而言将是一个巨大的考验。相较之下，人工智能对自由并无必要的需求，甚至可能无法理解自由的真正价值。对人工智能实施自由刑不仅不会起到使其反省、改造的作用，还会浪费社会资源；再次，人工智能无法适用资格刑。资格刑主要以公民的权利内容为处罚对象，人工智能没有公民资格，不具备适用资格刑的前提条件；最后，罚金刑不适用于人工智能。一方面，人工智能自身不具有独立财产，不满足适用罚金刑的先决条件。另一方面，如果将罚金的种类定义成人类为人工智能缴纳的保险，则表明罚金出于人类之手，刑事责任主体转为人类而不再是人工智能[7]。

从设置新刑罚措施的角度看，刘宪权教授指出，现有的刑罚种类难以应对强人工智能犯罪，应当增设删除数据、修改程序、永久销毁的层级性刑罚措施[8]。但持否定论观点的学者认为人工智能依然不能得到有效规制，尽管以上措施在实践中具备可操作性，但却存在不符合刑罚理论之处。首先，新刑罚不能对受害者起到安抚作用。假如将永久销毁人工智能等同于人类适用的死刑，如果犯罪者是人类，那么死刑的实施将会在一定程度上平复受害者的心情；而如果犯罪者是人工智能，则即便对其实施了永久销毁，受害者也难以获得与死刑同等的心理安慰。可以说，任何对人工智能的惩处可能都无法实现对人类真正意义上的安抚；其次，新刑罚难以预防人工智能再次犯罪。数据本身具有可复制性，因此很难保证在删除强人工智能的数据后，其不会再通过强大的学习能力获得同种类的数据。此外，如果危害行为因程序本身的问题导致，那么责任主体应当是程序的设计者；最后，新刑罚对人工智能无法起到矫正作用。刑罚的矫正功能旨在让责任主体知悉自己行为造成的危害以及社会意义，而人工智能无法具备人类对感情和道德的认知能力，自然也无法及时矫正自己行为中的违法之处。

## 4. 观点审视：肯定论在现阶段缺乏足够的科学依据

通过上述分析不难看出，肯定论的逻辑整体表现为：人工智能的发展将会使人工智能超出人类的控制并以此产生具有独立意识的强人工智能，届时，其将具备辨认能力和控制能力，继而取得刑事责任主体资格。然而这样的推论看似具有合理性，实则并不具备牢靠的理论基础[9]。关于肯定论的理论基础不够牢靠的观点，可以从内外两个角度论证。所谓外部角度，即着眼于肯定论的对立观点，通过否定论予以论证，在上述关于否定论的概述中已有提及。因此下文将从内部角度，即肯定论自身存在的不合理之处进行论证。

### 4.1. 肯定论的提出更多是肯定论者的主观臆测

法的逻辑推理要求：在对未来可能出现的事物进行预判时，均需以既有的法律规定和法律逻辑为立足点[10]。而肯定论的提出更多停留于假设，缺少足够的科学依据予以支撑。关于未来人工智能会产生独立意识，可能具有辨认能力和控制能力等一系列观点，当前均只是肯定论者的猜想。人工智能的发展是未知的，其是否真的会达到超越人类的高度还不得而知，任何人都难以预料人工智能是否会产生独立的意志。同时，理论对于人工智能独立意志的判断也很难形成客观的标准，即便人工智能实施了程序设计以外的行为，可能也是程序运行的故障而不能将其一概而论为“独立意志”。

### 4.2. 人工智能与单位不具备可类比性

人工智能在本质上不同于由独立的个人而组合成的单位，二者并不具备可类比性。肯定论的逻辑在于：既然单位可以作为自然人意志的延伸而成为刑事责任主体，那么对于与人类关系更为紧密，且意志可能超越人类的人工智能而言，更具备被赋予刑事责任主体资格的理由。然而，单位系自然人的集合，自然人以整体表现出的意志当然可以等同于单位自身的意志。而人工智能实施的犯罪为独立完成，这一过程中并没有人类参与[11]。因此，以单位为参考，认定人工智能具备自由意志并可将其拟定为刑事责任主体的观点缺乏科学性。

### 4.3. 肯定论自身存在相悖的逻辑

肯定论者既认为人工智能未来会脱离于人类，又希望人类能够通过立法的途径规制人工智能的逻辑本身就是自相矛盾的。强人工智能时代，人工智能可能将取代人类，依此可概括为事务的决定权在于人工智能；人类可以立法的手段约束人工智能，依此可概括为事务的决定权在于人类。毫无疑问，同时存在于肯定论中的此两种逻辑呈现出了相悖的状况。

综上所述，肯定论的提出只是一个单独意义上的结论，其得出的过程和论证的方法均有待商榷。结合人工智能的发展状况以及观点论证的科学性程度来看，否定论更为合理。人工智能技术的进步固然可能会引发对现实问题的担忧，但也不必过于对此感到焦虑，以没有实际科学依据的猜想为出发点进行超前立法显然是不合理的。

## 5. 人工智能刑事责任主体资格问题的延伸思考

在理论层面讨论人工智能刑事责任主体资格问题的同时，还应着眼于解决实践中出现的问题。以下将从三个领域进行否定论在人工智能刑事责任问题上的延伸思考。

### 5.1. 否定论之于医疗案件责任的讨论

现阶段，人工智能在医疗领域逐渐得到了越来越多的重视，但同时也增加了医疗案件的复杂性，其中涉及的主体不再局限于传统的医患双方，而是覆盖了诸如人工智能的设计者、制造者等主体。因此，合理分配各方主体的法律责任是极为必要的。在明确人工智能工具地位的基础上，可以认定人工智能不具备刑事责任主体资格。如果医生因相信人工智能而做出了错误诊断并造成了危害后果，则应由医生承担刑事责任；如果危害后果的发生归因于人工智能产品的缺陷，则应由其设计者、生产者承担相应责任。

### 5.2. 否定论之于自动驾驶的归责思考

首先需要明确的是，自动驾驶不具有独立意识，无法作为合格的刑事责任主体。其次，驾驶员应承担何种程度的责任也应分情况考虑。其一，在自动驾驶汽车没有产品质量问题的情况下，如果驾驶员需要在行驶的过程中参与驾驶，那么对于自动驾驶产生的危害结果应由驾驶员承担刑事责任；其二，如果自动驾驶汽车可实现全程自动驾驶而无需驾驶员参与，交通肇事案件的发生归因于自动驾驶汽车的质量问题，则最后的责任应由自动驾驶汽车的设计者、生产者等承担，驾驶员可基于公平原则给予受害人适当的补偿（例如，以驾驶员为自动驾驶汽车缴纳的强制保险为补偿金）。需要注意的是，在责任认定时，还需确定驾驶员是否在主观上存在过错或未履行注意义务，继而明确其是否应作为承担刑事责任的主体。

### 5.3. 否定论之于财产犯罪的审视

部分持肯定论观点的学者认为：人工智能存在因被欺骗而产生错误认识，继而处分财物的情形。故而应将人工智能作为一个全新的刑事主体，为其增设新的罪名予以规制。而从否定论的视角看，人工智

能没有处分财物的意识，并且“机器不能被骗”更是一个通识性的观点。例如，在实践中，处分意识往往是区分盗窃罪与诈骗罪的重要标准，如果受害人存在处分意识，则犯罪人构成诈骗罪，反之则构成盗窃罪。但在人工智能时代，该标准可能不足以应对实践中的全部案件。例如，行为人在网购名牌商品后以赝品退货给卖家，如果受理方是人工服务，则存在人类错误的处分意识，认定其构成诈骗罪；而如果受理方系人工智能，按照机器不能被骗的理论，则不存在错误的处分意识，故认定其构成盗窃罪。这样的判断方式未免显得过于形式主义。因此，在人工智能时代，关于处分意识的判断可能并不具有良好的适用性，在坚持否定论理念的基础上，可以考虑将盗窃罪兜底适用于取得型财产犯罪，在难以分辨是否具备处分意识时，直接定性为盗窃罪。

## 6. 总结

引发人工智能刑事责任主体资格问题争论的一个重要原因在于，部分学者担忧在未来可能到来的强人工智能时代，一旦人工智能拥有独立意识并能够脱离人类控制实施犯罪的情况发生，现行刑法可能很难实现追究刑事责任的目的。然而，我国刑法自身具备的容纳功能从某种程度来讲可以解决这一问题。一方面，我国的刑法建构体系以“人”为核心，这一法律秩序框架从实质上就将人工智能定义为了工具。而现行刑法明确、全面规定了对犯罪参与者的追责方式，对人工智能的设计者、销售者及使用者都明确了问责标准。换言之，人工智能可能涵盖的所有主体均已纳入了刑法的调整范围，在实践中不会出现因主题不明而无法追责的情况；另一方面，如果进行更进一步的延展分析，现行刑法的相关规定足以应对人工智能作为工具而涉及的刑事法律问题。人工智能技术的进步固然会导致犯罪工具的智能化以及犯罪手段的复杂化，但相关犯罪行为的本质和人工智能的工具属性并没有发生改变，故无需设立新的罪名。在实践中也可以将人工智能类比于网络信息等技术，尽管犯罪手段的技术提高了，但由于其行为的性质未发生改变，现行刑法体系依旧可以对其进行有效规制。

综上所述，对人工智能刑事责任主体资格问题的思考应从法学视角和科技的实际状况切入，以想象出的科技发展预测为出发点进行立法显然是不合理的。持肯定论的学者预想人工智能在未来将超越人类，希望能通过在法律层面主动求变的方式提前防范的理念同时也忽视了刑法自身具有的包容性和延展性。因此，坚持以“人”为核心构建法律秩序，进一步明确人工智能与人类的关系，或许才是解决人工智能刑事责任主体资格问题的关键。

## 参考文献

[1] 王充, 董璞玉. 人工智能时代刑事责任主体之再审视[J]. 广西社会科学, 2020(12): 118-125.

[2] 王文明, 齐卫红. 论人工智能的刑事主体地位与刑法应对[J]. 河南工业大学学报, 2021, 37(5): 52-58.

[3] 刘宪权. 人工智能时代的“内忧”“外患”与刑事责任[J]. 东方法学, 2018(1): 134-142.

[4] 张博通. 人工智能刑事责任问题研究[D]: [硕士学位论文]. 呼和浩特: 内蒙古大学, 2020: 19.

[5] 闻志强, 梁小敏. 人工智能刑事法律主体地位与归责判断[J]. 法治论坛, 2021(3): 119-136.

[6] 李琪, 姜俊鹏. 人工智能刑事主体资格的生成[J]. 上海法学研究集刊, 2021(5): 165-171.

[7] 张斌. 强人工智能体的刑事责任主体资格审视[J]. 法制与经济, 2020, 29(12): 28-34.

[8] 刘宪权. 人工智能时代刑事责任与刑罚体系的重构[J]. 政治与法律, 2018(3): 89-99.

[9] 周子实. 强人工智能刑事主体地位的折衷说——阶层论视域下“准主体”的教义学证成[J]. 广西社会科学, 2021(8): 99-105.

[10] 刘瑞瑞. 人工智能时代背景下的刑事责任主体化资格问题探析[J]. 江汉论坛, 2021(11): 105-110.

[11] 陈洪兵. 人工智能刑事主体地位的否定及实践展开——兼评“反智化批判”与“伪批判”之争[J]. 社会科学辑刊, 2021(6): 92-98.