AI責任肯定論の動向

根 津 洸 希

Ⅰ．はじめに

拙稿「ロボットの処罰可能性を巡る議論の現状について（比較法雑誌51巻2号（2017））」にて、「ロボットやAIに責任を観念することは可能か」という奇異な問いを巡る各種学説を紹介した。本稿は、上記拙稿の続編という位置付けとなる。

上記拙稿が紹介対象とした、2014年前後の諸論文においては、たしか

にAIの責任を巡る議論はなされていたが、そこでは責任論における各要素をAIにあてはめた上で、「AIもいくつかの責任要素を有しうる」ということが思考実験的に示されたにとどまっていた。ありていに言えば、AIの責任や処罰を基礎づけようと真剣に論じていたわけではなく、むしろAIに責任を認めることなどできないという前提には立ちながらも、AIの責任を否定する論拠は存外少ないということをいっていたにすぎない。

しかし、ディープラーニングという機械学習の方法が法学の世界にも周知されたことや、2015年に行われた国際的画像認識コンペティション[1]にてAIの画像認識能力が人間の目を超えた（AIの正解率95.1%＞人間の正解率94.9%）ことなどから、最早はなからAIに責任はないなどということを前提とすることも難しくなってきた。翻訳AIであるDeepL[2]の自然言語処理は、既に筆者の語学力を遥かに凌駕しているし、囲碁AIのAlphaGoはプロの棋士にも勝利するほどである。

これらの急速な技術発展もあって、2016年以降、法学において再び急激にAIの責任を巡る論考が発表されるに至った。加えてその論考の多くはAIに責任を肯定しうることを主張するものであった。数年前までは極めて少数派にすぎなかったAI責任肯定論が、いまや有力説とまではいわないまでも、無視できない議論の潮流となりつつあるのである。

本稿は、AIの責任を巡る議論のこの新たな展開を追跡調査するものである。とりわけ、近時のAI責任肯定論を紹介し、若干の検討をくわえることとする。

---

1 ImageNet Large Scale Visual Recognition Competition（https://www.image-net.org/challenges/LSVRC/）

2 https://www.deepl.com/translator

Ⅱ．AI責任肯定論の動向

以下ではAI責任肯定論の各論者を引用しつつ、適宜要約しながら紹介する。

ⅰ）Corneliusの見解

現在のAI技術を前提とすればAIの処罰は要しないとしつつも、その主体性の基礎付けにおいて、ドイツの法人処罰論からのアプローチをとるのはKai Corneliusである。

> *AIの可罰性を肯定するには、一般的な理解によれば刑罰の要件となっている責任という要件が障壁となる。というのも、「人工的」なシステムがどんなに賢くとも、責任は基本法における人間の尊厳から生じるものであるから、AIの責任などありえないからである。人間の尊厳は人間にのみ認められるのであって、動物やロボット、そのほかの「人工的な人格」には認められない。法人は（少なくとも従来からの通説によれば）行為能力も責任能力もないとされてきたために、今日までドイツ刑法は自然人の処罰しか予定していなかった。それゆえAIの処罰可能性の検討がなされてこなかったのであろう*[3]。

しかしCorneliusは、この法人処罰否定論が他の国々における法制度からすればかなり異質であることを指摘する。実際に、法人処罰を肯定し、そのような法制度を有している国は多くあり、その国々においては非自然人たる行為主体に対する刑法的責任の可能性が開かれていることを指摘し

---

3 *Kai Cornelius*, “Künstliche Intelligenz”, Compliance und sanktionsrechtliche Verantwortlichkeit, ZIS 2/2020, S.60 ff.

つつ、ドイツにおいても法人処罰が目下議論されていることを強調する。もし法人処罰の法制度が実現すれば、AIを含めた非自然人の処罰可能性が再度検討されうるという。

基本法そのものは責任概念の変更について何らの態度をも示していない。連邦憲法裁判所によるリサボン判決もこれに矛盾するものではない。というのも、リサボン判決は（人間の）個人責任にのみをその射程にしているからである。たとえば、法人に関する刑法上の議論を見てみると、（法人を行為者とみなすというという前提を受け入れるならば）その法人の選任責任や組織的責任に基づく「社会的責任非難」というものが議論されている。法人処罰の導入に対する（個人）責任主義の、人間の尊厳というハードルが下がるということは、同じくAI処罰の導入のハードルをも下げることとなる。

無論、（処罰適性の問題のためにも）刑罰目的が考慮されねばならない。法人処罰についての考慮の大半は予防に基づいている。また部分的には応報も考慮される。この考慮も何らの支障なく（どちらも人工的な被造物として）AIにも転用されうる。刑罰は「コミュニケーション的応報」として、AIの法違反を矯正的正義の意味で訂正するかもしれない。一般予防の観点に基づくならば、実際に処罰を受けたAIと似たような他のAIに対しても処罰の影響があるのであれば、刑罰適性も認められうるであろう[4]。

すなわち、ドイツにおいては刑法が専ら自然人のみを対象としていたが、近時の法人処罰を巡る議論に鑑みれば、自然人以外の法的人格も処罰する法制度も観念可能なのであり、そうであるとすればAIであっても処罰の対象にはなりうるとする。

4　*Cornelius*, a.a.O（Fn.13), S.61 ff.

しかし刑罰目的論からすれば、あるAIの処罰が他のAIにも一般予防ないしコミュニケーション的応報が観念できるという条件が整っている必要があるという。現状のAIは、この条件を満たしてはおらず、もしAIに刑罰を科そうというのであれば、「一度処罰されたら同じ行為を避ける機能」がプログラムされている必要があるという。

Corneliusは法人処罰との関連で非自然人の行為主体性を肯定しようとし、その枠内にAIも含まれると主張する。現状、AIを実際に処罰するまでの条件は整っていないとするが、条件が整ったならばAIの処罰もありうるとする。この点で、AIの責任主体性を肯定する余地や処罰可能性自体は（将来的にではあれ）肯定しているものといえよう。

ⅱ）Gaedeの見解

AIに法主体性を認めようとする点においてはCorneliusと同様であるが、種差別というキーワードを用いながら、別の角度からAIの法主体性を基礎付けようとするのはKersten Gaedeである。

> *人工知能は動物よりも能力が高いのに、なぜ我々は人工知能に人権のような権利を認めるべきではないといえるのであろうか。人工知能の権利を否定することは、ナルシスティックな種差別、つまり根拠のない差別であるかもしれない。*
>
> *いままでは、我々は技術を専ら手段とみなしてきた。あらゆる機械は我々の自由のための手段である。機械は非生物であって、我々の所有物である。現状に鑑みれば、ロボットの権利を問うということなど、考えられない。我々こそが、たとえば運転中のジレンマ状況を解消するために、ロボットをプログラムし、ロボットに倫理規範を入力するのである。ここから、我々は強硬な支配性や優位性を主張するのである。*

（中略）

人工知能が自意識へと目覚め、自律性を獲得したならば（AI研究者はこのことを十分ありうることだと考えているが）、この純粋道具的技術理解は維持されえない。我々の優位性を基礎付けている要請をAIが満たした場合、我々は考えを改めねばならなくなるであろう。

「強いAIだって、作ったのは我々人間だから」という理由では、このことを簡単に否定することはできない。子供であっても人間によって作られるのだから。子供の誕生を目の前に、「創造者の望みによっては自己犠牲をもいとわない主体が誕生したのだ」などと考えるわけがない。人間がある存在に主体性を認める場合、その主体性が何を意味するかについて、事前に検討せねばならない。

「我々人間は自然的な方法で発生してきたのだ」ということを言ったところで、やはりあまり意味はない。その主張は適切ではない。なぜなら、我々は現代的な生殖医療技術を許容してきたし、それによって生まれた人間にも、人間の尊厳を認めるからである。病気でもないのに移植医療を用いて「不自然にも」自己最適化を試みる人間であっても、この人に人間の尊厳を認めないということは難しい。とりわけ我々の自然的性質を持ち出すと、自然的事実から直ちに道徳的・法的当為を推論することとなってしまう。何かがそうであるということ、今までそうであったという事実は、そうであるべきであったことまでをも含意しない。規範の基礎付けの際にこれをしてしまうと、自然主義的誤謬であるとして非難を免れない。この自然主義的誤謬の背後には、創造者理論が隠れた前提として存在しているのである[5]。

また、AIに権利主体性を認めるには部分的に胚保護法との関係で人間

5　*Kersten Gaede*, Künstliche Intelligenz – Rechte und Strafe für Roboter?, S.36 ff.

の尊厳との矛盾が生じうるとも指摘する。すなわち、クローンが禁止されているのは、その技術が人間の遺伝子によって他者をプログラミングすることであって、人間の尊厳に著しく反するからであり、人工的な人格であるAIにも同じことが部分的にいえるのではないか、というのである。

しかしこれも様々な理由から退けられうる。まず、全ての人間は他者の目的設定から生じる。ドイツ議会は、人間の自律性の事実的認識は遺伝子のみならずとりわけ環境要因にも依存することを見落としていた。クローンに自律性を認めることははたしかに難しいかもしれない。しかし、クローンに自律性を認めることは不可能ではないし、また人間の尊厳に値しないものであるともいえない。クローンの禁止は、とりわけ大規模な再生産がなされた場合や様々な耐え難い技術的リスクによって正当化されるかもしれない。しかし、人間の尊厳に反すると一からげに言ってしまうのは間違いであり、これはAIにもそのまま転用できるロジックではない[6]*。*

また、Gaedeは「AIやロボットは人間の欲求があるからこそ必要なのだから、AIやロボットに権利など認められない」という否定論に対しては、自己利益のために理性的主体を従属させており、奴隷制に賛成するのと変わらないとしている。

イマヌエル・カントにおいてすら、人間の圧倒的優位という考えの正反対となるものが見いだされる。カントはたしかにロボットの問題に取り組んではいない。しかしカントは、宇宙人を想像することを手掛かりとして、人間は彼の理論のいち適用場面にすぎないということを考えていた。ドイツの法概念に表れている、カント流の人間の尊厳の基礎付け

---

6 *Gaede*, a.a.O（Fn.19）, S.40.

は、ただ例示的に人間に結び付いていたのである。カントは全ての理性的本性のために当為秩序を構想したのであり、その尊厳をもっぱらその自律性から導き出していたのである[7]。

以上の理由から、いわゆる「強いAI」には自律性が認められ、そのAIも他者の理性を承認し、尊重できるのであれば、法的主体性を否定する理由はないという。

強いAIには法的主体性が認められるとしたうえで、GaedeはさらにAIの処罰に関しても言及する。Gaedeは刑罰の正当化について、刑罰という抑止的営為は法の回復を目的としているという、ヘーゲル的な正当化論に依拠するとしている。その上で、AIの処罰可能性につき、次のように説明している。

強いAIが本質的に自律的な行為主体であるとみなされねばならない状況にいたったとき、この行為主体は、法と調和して世界に参与するという義務を負うであろう。その行為は、我々の法の妥当に物申し、不法をなしうる重要な態度として認識されうる。AIによって違反された法、すなわち我々の自由・平和秩序を、不当な行為に対して有効な刑罰という営為でもって異議を唱えることによって回復させることには、十分理由のあることであろう。（…）

無論、法の回復という刑罰目的の追求には、機械の責任という語が用いられることと、その責任と法違反について衝突している理由がAIに対し意義をもって伝達されうるということが前提となっている。（…）。責任と、刑罰による非難のコミュニケーション的受容性は、なお不可欠である（…）[8]。

---

7 *Gaede*, a.a.O（Fn.19）, S.41 ff.

8 *Gaede*, a.a.O（Fn.19）, S.63 ff.

このようにAIに責任を認めるには、そのヘーゲル的な刑罰観を背景として、法違反の事実とそれに対する非難を理解できるという法的コミュニケーション能力が必要であるという。逆に言えば、この能力さえ満たすことができるのであれば、「人間ではないから」ということを理由として権利・責任主体から外してしまうのは種差別となるのだという。

そして最後に、AIの処罰可能性については、AI処罰の文脈でよく引き合いに出される再プログラミングについて指摘しつつ、刑法を純粋機能的に理解した場合にのみ正当化されるとして懐疑的な立場をとっている。

> *結論的には、AIに対する刑罰も考えられる、と私は思っている。たとえば、個別の機能やネットワーク機能を切断するだとか、移動の自由を制限するということもありうる。自我を有するロボットを前提とすれば、たしかにそのロボットは自身に科された自由の制限を苦痛として感じることはないかもしれない。そのため、感覚を有しないロボットに刑法が適用できるかには、非常に大きな疑念が付きまとう。しかし、強いAIであれば自由の剥奪を合目的的な制限であると理解するであろうし、それによって学習することもできよう。場合によっては、自由剥奪の期間や程度によって、苦痛を感じる能力の埋め合わせのために、この効果はより強くなりうる*[9]。

Gaedeは、人間以外に責任を認めないというのは種差別に過ぎないとしつつ、AIにも自律性や規範的応答可能性を認める方向でAIの責任や処罰可能性を肯定しようとする。

---

9 *Gaede*, a.a.O（Fn.19）, S.66 ff.

iii）Quarckの見解

Lasse Quarckは、自律性を有したAIが何らかの保護法益を侵害した場合に、様々な理由から利用者やプログラマーにはその結果を帰属できない場合が想定されうるという問題関心からAIの責任主体性を論じる。同論文の特色は刑罰目的から逆算してAIの責任主体性を検討する点にある。

デジタル化への法的取り組みに対し、法治国家が何ら適切な応答をしていないという印象を市民に持たれてしまえば、規範秩序への信頼が揺るがされてしまう。この信頼を維持するためには包括的な法益保護が不可欠であるから、法益侵害がなされたのであれば、これに何らの応答もせずに甘受するということは許されない。刑法やその適用者が、法益侵害に対し適切に応答できないのであれば、法益はもはや保護に値しないものとして価値を失ったも同然である。刑法的【法益侵害】結果に対し責任を負う者がいないという結論は、法を根底から揺るがすこととなろう。

というのも、刑法の目的は過去の不法に応答することのみならず、この応答によって将来の不法を予防することにもある。これはひとつには行為者に対して影響をおよぼす特別予防的なものと、もうひとつには法違反には法治国家的応答がなされることを（法適用の名宛人としての）社会に示すことによる一般予防的なものにより生じる。違反行為が判例により形式的にもきちんと違反行為であると示されかつ処罰されてはじめて、一般市民による規範遵守が達成されうるのである。法益が侵害されたのに何らの応答もなされないという印象が生じてしまうと、刑法はその目的を最早達成できないのである。[10]【括弧は筆者による補足】

---

10 ***Lase Quarck***, Zur Strafbarkeit von e-Personen, ZIS 2/2020, S.66.

以上のようにQuarckは、刑罰の目的を規範への信頼維持に見出している。このような前提から、AIが法益侵害を惹起しているにもかかわらず、責任主体の特定不可能性による刑罰法規の欠缺があってはならないから、AIの可罰性が論じられねばならないという。そしてQuarckはAIの可罰性を論じる上で、AIの行為性、AIの責任能力、AIに対する刑罰という3点を論じる。

まずAIの行為性については、規範を理解する能力が行為性の前提であるとすれば、AIには行為する能力が（少なくとも現状は）ないということは、Quarck自身も認めている。しかしながら、法人処罰を肯定する法秩序においては、このような個人の行為による不法実現のみならず、組織的・アルゴリズム的な不法実現をも処罰していることを指摘する。

> *AIの議論に引き戻すと、法人の行為能力をその独自の動態【原語：Eigendynamik】にひきつけて考えることができるのであれば、同じようにアルゴリズム・プロセスという独自の動態を基準とすることも可能である。可罰性要件の検討において、行為とは構成要件的結果の因果的・客観的帰属論的起点となる要素なのである。後に結果に表れる、法的に否認される危険は、法人におけるコミュニケーション的構造の中でのミスやアルゴリズム内部でのミスによっても創出されうる。このような個人的不法実現を超越した行為概念把握によって、AIの行為性に対する疑念は弱まりうるのである*[11]。

人間の可罰性を論じる際には意思に基づく行為であるか否かや目的性などが考慮されうるが、我が国のように法人処罰を予定している法秩序においては、必ずしも個別の自然人による行為ではなく、法人そのものの行為に行為性を認めるのであれば、これをAIに転用することも不可能ではな

11 *Quarck*, a.a.O（Fn.27), S.67

いというのである。

また第二の点としてQuarckはAIの責任能力を論じるに際し、責任能力とは「法に従い不法に抗する決断が可能であった」ということであると解し、この問題は結局のところ意思自由の問題であると整理しつつ、次のように述べる。

無論この意味での自由意思というのは証明不可能である。自由意思の存在は脳神経科学において決定論の論者によって否定され、あるいは少なくとも疑義が寄せられている。したがって自由な意思決定の経験だけが現実であり、しかし一定の意思決定に至るプロセスは完全に決定されている、と。つまり意思決定は、遺伝子配置、その人間が受けた教育や社会性、その時の気分、そのほかの状況、具体的には検証不可能な様々な要因からの帰結であるとされる。しかし逆もまた然りである。現状、自由意思が積極的に証明されえないのと同様に、自由意思が存在しないということを前提とすることもまた確実ではないのである。

しかし自由意思を実際に有していることが可罰性の必然的前提であるというのは、上記に照らせば矛盾しているように思われる。刑罰を賦課するか否かにとって、いかにしてそのような立証不可能かつ法的に不確実な要素を決定打とできようか。

この問題を解決するには、人間の行為者の責任を問うための、より現実的なアプローチが必要となる。自由意思は、その人自身の経験に基づいて、あるものとされるだけである。もし私が自由意思を有している（と誤信している）場合、これは全ての他者においてもそうであるはずである。したがって可罰性要件としての責任は、責任の分配を意味しているのであり、不法がなされたことにより我々の社会において責任の分配によって解決されるべきコンフリクトが生じるがゆえに、責任分配がなされるのである。このような解決への社会的需要が存するのは、上述の理由から、不法を実現したことに対し制裁を科さずに放置することが

許されないからである。

このような責任の機能的理解によれば、AIに責任を分配することも可能となる。我々の社会システムにおいてアルゴリズムに自由意思が認められるかぎり、AIによる法違反は、人間の過誤態度と原則的に同様に、解決する必要のある社会的コンフリクトを惹起しうる。したがって不法を行う意思決定が決定論的生物学的プロセスないしアルゴリズム・プロセスに基づいているのか、不正だが自由な意思形成に基づいているのかは、人間の場合であってもAIの場合であっても重要ではないのである[12]*。*

つまり人間の責任を論じるときでさえ、自由意思の存在は責任を肯定する論理的な前提とはなってはいないのであるから、これをAIに要求する必要もないし、自由意思という証明不可能な前提に立つのはむしろ不確実であるという。

3点目として、AIの処罰可能性については、刑罰の性格には害悪賦課、社会道徳的無価値判断、否認の3つが区別されるという。

害悪賦課についてまず考えてみよう。刑罰には、行為者にとって不利益と感じられる作用がある。この際、実際に行為者が必ずしも不利益を感じていない場合であっても、それは問題とはならない。なぜなら、たとえば移動の自由や人格発展といった憲法上の保障がどのみち制限されているからである。教科書事例として、ホームレスが寒い時期を暖房のきいた刑務所で過ごすために冬前にわざと収監されようとする、という例を挙げることができよう。

無論、害悪を賦課される者にとっては、その害悪が、たとえば一定金額の支払いが罰金として科されているのか過料として科されているのか

12 ***Quarck***, a.a.O（Fn.27), S.67ff.

には、まったくどうでもいいことである。同じことは自由刑にもいえ、処分としての保安拘禁なのか、警察拘置なのかはどうでもいいことである。いずれにせよ同じ憲法上の保障は制限されている。したがって害悪賦課は刑罰固有の要素ではないのである。

害悪賦課と同様に、社会倫理的無価値判断も刑罰固有の要素ではない。刑法のみならず、懲戒規定や秩序違反法における規範違反の場合も、当該行為者が違法な行為をしたことをに対し国家的非難がなされる。駐車違反の反則切符も、その行為が法秩序という名の、社会全体の価値についてのコンセンサスに矛盾することを示している。

（中略）

残るは刑罰の本質的要素としての否認であり、行為者の責任と刑罰との直接的な結びつきである。他の国家的制裁には要求されてはいない個人的非難可能性という要件は、刑法上の重要な行為に対する高度の否認を特徴付け、したがってその刑罰にも否認的性格をもたせる。もしAIが、上述のように、（我々の社会的現実から生じる）責任分配の意味で責任主体であるならば、刑罰はその否認的性格をAIに対しても拡張することが可能である。

しかしAIの可罰性を確認するだけでは、一般予防目的達成のためには不十分である。為された不法は、受け取り手としての市民にとって具体的に説明されねばならない。だからこそ可能な限り正当な刑量によって責任非難は量化されねばならないのである。ここでは社会奉仕労働やロボットの身体への介入、あるいは最終手段としては機能停止などが考えられる。

さらに考えられうるのは、違反された規範の意味内容をそのアルゴリズムに入力することによって再プログラミングすることである。再プログラミングは最大限の特別予防的作用を有しうるし、多数のAIがネットワーク化されていて再プログラミングが全てのAIに対して可能なのであれば、一般予防作用もあろう。つまり、AIを処罰することも可能

なのである[13]。

以上のようにQuarckは、機能主義的な刑罰・責任理解から、自由意思や責任は社会的必要性に応じて分配されるものであるから、AIにも責任を認めることができ、またAIに対する処罰も可能であるというのである[14]。

iv）Gabriel Hallevyの見解

Hallevyは英米系の犯罪論体系に沿ってAIの責任を分析し、AIの刑事責任を否定する見解を「奇妙な結論」であるとすらしている。Hallevyは、AIの責任と処罰可能性について、責任を基礎付ける要素（General intent、Negligence、厳格責任）と責任を否定する要素（刑事未成年、自己コントロールの喪失、責任無能力、酩酊、事実の錯誤、法律の錯誤、実質的免責）、刑罰目的論の検討を行う。我が国の刑法はドイツ刑法を範としており、それゆえ我が国の刑法理論との完全な対応関係を見出すことは難しいが、各要素の検討については参考になる点も多い。したがって本稿では、比較的我が国の刑法理論に親和性があるものと考えられる諸要素、すなわち責任を基礎付ける要素のGeneral Intentより認識的要素と意思的要素、Negligenceについての記述と、責任を否定する要素から刑事未成年、責任無能力、そして刑罰目的論の検討に関する記述を紹介することとする。Hallevyの検討は犯罪論体系を網羅しており、かつ各要素の一般的定義に始まり、その判断の構造化、AIへのあてはめ・転用可能性を論じ

13 *Quarck*, a.a.O（Fn.27）, S.68 ff.

14 同旨の見解は*Monika Simmler=Nora Markwalder*, Roboter in Verantwortung? – Zur Neuauflage der Debatte um den funktionalen Schuldbegriff, ZStW 129（1）, S.20 ff.

るため、かなり長い引用となるが参照する価値は高いと思われる。

まずHallevyは（必ずしも同義とはいえないものの）我が国でいう故意論に対応するGeneral Intentにつき、その構造を素描する。

General intentは二層の要件から成っている。(a) 認識cognitionと(b) 意思volitionである。認識は、知覚awarenessにより成り立っている。この認識の層を「知識knowledge」と呼ぶ法制度もあるが、知覚という語がより正確であるように思われる。しかし、知覚も知識も、この文脈においては機能的には同じであるし、意味も同じである。人間が知覚できるのは、過去に生じたことと現在生じていることだけであり、将来の事実を知覚することはできない。

（中略）

General Intentのもう一つの層は意思の層である。この意思の層は認識に付加される要素であり、認識に基づくものである。意思だけが存在するということはなく、常に知覚に付随する要素である。意思は、事実的経緯の結果に対する行為者の意欲である。多くはないがいくつかの犯罪において、意思は特定の事実を超えて、動機や目的にかかわることがあり、これはSpecific Intentと呼ばれる。意思にかかわる主たる問題は、行為から結果が生じる可能性についての行為者の知覚とは異なり、行為者がその結果が生じることを欲したということである。行為者の視点からはこの結果が将来生じるがゆえに、この結果は意思の、唯一の合理的対象なのである。時系列に関する行為者の視点からは、状況と行為の両者の発生は、意欲とは何らの関係がないのである[15]。

このGeneral Intentの構造のもと、Hallevyは認識の構造について詳説

15 ***Gabriel Hallevy***, Liability for Crimes Involving Artificial Intelligence Systems, p.83–85.

しつつ、実際の裁判においては認識があったことの立証は一定の推定に置き換えられていることを指摘する。

結局のところ、人間が一定の事実的情報を知覚していたものとされるには、二つの追加的条件が要求される。すなわち（a）五感による事実的情報の収集と（b）脳内でのこの情報に対する重要な一般的イメージの形成である。

もしこのうち一方が欠ける場合、その人はあることを知覚していたとはみなされない。知覚は二者択一の問いである。つまり、行為者は知覚していたか、そうでないかである。部分的な知覚は意味を持たない。行為者が事実的情報の一部を知覚していた、すなわち情報の一部を確実に知覚していたのであって、一定の事実を断片的に知覚していたということではない。

（中略）

知覚は精神の内的プロセスであって、必ずしも外部的に表現されるものではない。したがって、刑法はこのために立証上の代替方法を発展させてきた。この代替方法というのは、一定の状況から知覚の存在を推認するという方法である。多くの法システムにおいて、主に二つの推認方法がある。（a）行為と状況に対する知覚の代替方法としての意図的無知推認と（b）結果の発生可能性の知覚という代替方法としての知覚推認である[16]*。*

しかし、このような推認によって知覚が認定されるのは、人間の心理を覗くことができないからである。これに対し、AIの場合は事情が異なるという。

16 Hallevy, op.cit., p.88–89.

たしかに人間の知覚を立証するのは難しい。しかし、知覚にかかわるプロセスはAIの場合、非常に正確にモニターされうるため、代替方法は不要となる。各個人の心が常に高度なブレインスキャナーによって読み取られ、全て記録されているといったような状況に近い。

もし事実的情報の知覚がこのようなブレインスキャナーによって確認されうるのであれば、合理的な疑いを超える知覚の立証は非常に簡単なこととなる。AIの全ての行為は、刑法における知覚にかかわるプロセスも含め、モニターされ、記録されるから、特定の事実的情報にかかわるAIの知覚を立証することも可能である。AIの知覚の立証は代替方法による推認である必要はなく、知覚自体が直接立証可能である[17]*。*

したがってAIのGeneral Intentを検討する際、認識的要素である知覚Awarenessの立証は、むしろ人間に比べて容易かつ直接的であるという。

また、この知覚に加えて必要となる第二層の要素、意思についてHallevyは次のようにその構造を整理し、AIの意思について検討する。

意思ないしGeneral Intentの意思的要素は、3段階に区別することができる。(a) 意思、(b) 無関心Indifferenceと (c) 軽率Rushnessである。無関心と軽率の両者は、多くの現代法システムにおいて無謀Recklessと呼ばれることが多い。

(中略)

意的要素の立証は知的要素の立証よりもはるかに困難である。両者はともに人間の心の中でのプロセスであるが、知覚は現在の事実にかかわるものであり、意思は将来の事実的状況にかかわるものである。知覚は合理的で現実的である一方、意思は必ずしもそうである必要はない。たとえば、人は象になりたいと意思することはできるが、人は象ではない

17 Hallevy, op.cit., p.93.

から、自身が象であることを知覚することはできない。意思の立証の難しさから、刑法においては立証上の代替方法が発展されてきた。

よく用いられる代替方法は、予見可能性ルール（dolus indirectus）である。予見可能性ルールは意思の立証を目的とした法的推認である。予見可能性ルールによる推認は、もし行為者が行為の実行を知覚し、一定の結果の発生が非常に蓋然的であると予見していた場合に、行為者が結果発生を意思していたと推認するという方法をとる。

（中略）

問題は、AIは刑法にいう意思を持ちうるのか、という点である。意志Willというものは曖昧で、一般的な語であるから、AIが意思を持つ能力を有するかはこの予見可能性ルールによる推認によって検証されねばならない。実際、この推認方法は人間の意思を立証するためにこの推認を使う中核的な理由なのである。そのためには二つの条件を満たす必要がある。

（1）結果の発生が非常に蓋然的であると予見されたということ

（2）行為が知覚に基づいて実行されたということ

強いAIであれば事実の発生可能性を評価する能力があるし、その評価に基づいて行為することができる。たとえば、チェスAIはボード上の駒の位置に基づいて、試合の現状につき分析する能力を有する。AIは次の一手のためにありうるすべての可能性を計算する。AIはまたそれぞれの可能性について、相手の対応の可能性や一方が勝利するまでの最後の一手までをも計算する。あらゆる手がその可能性に照らして評価され、それに従ってAIは次の一手を決めるのである。

これが人間であれば、その人はゲームに勝つことを意思しているといってよいであろう。その人が確実にそのような意思を有していたかどうかは知りようがないが、その人の一連の行為は予見可能性ルールによる推定を満たしてはいる。チェスの試合をするようにプログラムされたAIは、試合に勝つという目的志向的振る舞いをしている。人間のチェ

スプレイヤーも、試合に勝つという目的志向的振る舞いをしている。人間のチェスプレイヤーについては、彼がチェスの試合に勝とうとする意思を有しているといってよいであろう。このことは人間のプレイヤーに限らず、AIプレイヤーにもいえるように思われる。この状況において彼等の一連の行為を分析すると、予見可能性ルールによる推認をきちんと満たしている[18]*。*

したがって予見可能性ルールによる推認を媒介として、AIに意的要素を肯定することも可能であるという。ゆえにAIに対するGeneral Intentを理由とする刑事責任が認められうるという。

またNegligenceを理由とした刑事責任の賦課可能性についても、Negligenceの構造を分析しつつ、その要件との関係で検討している。

Negligenceは19世紀にGeneral Intentの例外として受容された。RecklessnessとNegligenceの主な区別は、Recklessnessの認識的要素によって発展してきた。RecklessnessがGeneral Intentの要請の一部として知覚の認識的要素を要求する一方、Negligenceは何も要求しない。RecklessnessもNegligenceも、合理的でないリスクをとったということが要求される。しかしRecklessnessは事実的構成要素の知覚を要求されるが、Negligenceは要求されない。Negligenceは知覚の欠如として現れ、行為の社会的準則を策定する。

個人は合理的なリスクのみをとることを要請される。合理的なリスクというのは抽象的な合理的人格の観点から客観的に判断される。合理的人格は自身の事実的態度を知覚し、合理的なリスクのみをとる。もちろん、その合理性は法廷で決まることであり、具体的事案との関係で回顧的に決定される。

18　Hallevy, op.cit., p.93–97.

（中略）

一般に、事実的構成要素との関係でのNegligenceの中核は、合理的人格であればその事実的構成要素を知覚でき、また知覚する義務を負っており、行為者も知覚する能力を有していたにもかかわらず、事実的構成要素の知覚を欠いたことである。知覚の欠如とは当然ながら、General Intentにて要求される知覚とは逆の状況である。結局のところ、人間が一定の事実的情報を知覚していたものとされるには、二つの追加的条件が要求される。すなわち（a）五感による事実的情報の収集と（b）脳内でのこの情報に対する重要な一般的イメージの形成である。もしこれら条件の片方が欠けている場合、その人は知覚があったとはみなされない。

（中略）

しかし、知覚の欠如とみなされるのは、単なる不知ではない。合理的人格であれば知覚でき、また知覚する義務を負っており、行為者も知覚する能力を有していたにもかかわらず、知覚を欠いたというものでなければならない。そのためには二つの条件を満たす必要がある。（a）知覚を統合する認識的能力を有することと、（b）合理的人格であればその事実情報を知覚でき、知覚する義務を負っていたといえる場合である。

（中略）

したがって裁判官は3つの問いに答えねばならない。

（a）AIは事実的構成要素の知覚を欠いていたか？

（b）AIは事実的構成要素の知覚を統合する一般的な能力を有しているか？

（c）合理的人格であればその事実的構成要素を知覚することができたであろうか？

これらの3つの問いが全て肯定され、合理的な疑いを超える程度に立証されたのであれば、AIは一部のNegligenceによる犯罪の要請を満たしたことになる。上述のようにGeneral Intentにいう知覚を形成する能

力があるAIは、技術的にも法的にも何らの問題なくNegligenceによる犯罪をなしうる。なぜなら、NegligenceはGeneral Intentよりも精神的要素のレベルが低いからである。したがって、NegligenceはAIにもありえるし、それを法廷で立証することも可能である[19]*。*

上述のようにHallevyはAIの刑事責任を基礎付ける各主観的要件を検討し、いずれについても積極の結論を導いている。

続いて、AIの刑事責任を否定する要素、「AI被告人の防御の抗弁」となりうる要素として、刑事未成年と責任無能力についてみていく。

AIが刑事責任を問われた際に、刑事未成年とみなされることはありうるだろうか。従来より、一定の生物学的年齢に満たない者は刑事責任を負わないとされてきた。十分に成熟したといえる年齢については、法制度ごとによって異なる。たとえば、ローマ法では7歳から一人前とみなされていた。この刑事未成年という抗弁は立法と判例法により定められる。主に立証上の理由から、重要なのは生物学的年齢であって精神的年齢ではないということは疑いようもない。生物学的年齢は立証においてはるかに容易だからである。

しかし、生物学的年齢と精神的年齢は一致するものと推認される。もし行為者が刑事未成年の下限を超えてはいるが、完全な成人年齢を下回っている場合、その行為者の精神的年齢は（たとえば鑑定書など）証拠によって検討される。決定的なのは、その行為者が自身の振る舞いを理解しているか、その振る舞いが誤っていたことを理解しているかである。行為者が理解しているのであれば、成人と同様に刑事責任を負う。（中略）

結局、刑事未成年者は刑事責任を負うのではなく、むしろ教育や再社

19　Hallevy, op.cit., p.120–129.

会化、治療を受けるのである。そうすると問題は、この理論的根拠は人間にだけ妥当するのだろうか、あるいはほかの法的存在にも同様に意味をもつのだろうか、という点である。一般的抗弁としては、刑事未成年規定は法人には適用がないものであると考えられている。「幼い企業」などというものは存在しないし、法人として認められた（そして法的に存在するとされた）その瞬間から、刑事責任を科すことは可能である。刑事未成年という一般的抗弁の理論的根拠は、法人にはあてはまらないからである。刑事未成年は、その年齢では意識が精神的に未発達であるから、過誤をなす精神的能力が欠けるのである。刑事未成年者が年を重ねれば、その精神的能力は物事の是非を理解する能力を得るまで、徐々に成長していくのである。

この点では刑事責任が重要となる。法人の精神的能力は時間的な「年齢」（登記日）とは関係がない。法人の精神的能力は不変であると考えられている。もっといえば、法人の精神的能力は成人から構成される人間組織に由来する。したがって、法人に対して刑事未成年を適用することはできない。ここで問題は、では人工知能は人間に似ているのだろうか、法人に似ているのだろうか、という点である。

その答えは当該AIのタイプによって異なる。不変のAIと動的に発展していくAIは区別されねばならないであろう。不変のAIは何年経とうと同じ能力で活動に取り組む。このようなシステムは時間が経とうとその能力が変化することはない。したがって心理的要件を形成する能力（たとえば知覚、意思、Negligenceなど）は犯罪がなされたどの時点でも検討されねばならず、一般的抗弁としての刑事未成年は意味をなさないであろう。

しかし、動的に発展していくAIは利用開始段階と終了段階で異なる。動的に発展していくAIの能力は、その精神的能力を含め、機械学習やその他の技術によって時間とともに発展していく。たとえシステム使用開始時点では刑事責任を負いうるような精神的能力を有していなかった

としても、あるタイミングでそのような能力を獲得する。利用開始から精神的能力獲得時点までの間の時間というのが、人間でいう刑事未成年にあたるのである[20]*。*

すなわち刑事未成年の考慮においては、自らの行為の意味を理解する必要があるが、そのような能力はAIも獲得しうる。ただし、自律学習機能を有しないAIの場合は法人と同じく、時間の経過によってそのような能力を獲得するわけではないため刑事未成年による責任阻却は認められないが、自律学習や機械学習機能を有するAIの場合には時間とともに「成長」がありうるから、その成長が一定程度の成熟に至るまでの間を刑事未成年と捉えることも可能であるという。

責任無能力の法律上の定義は、認識的側面と意思的側面の両面からなっている。責任無能力の認識的側面は、行為の犯罪性を理解する能力にかかわるものであり、責任無能力の意思的側面は意思のコントロール能力にかかわる。したがって、精神疾患や精神障害によって認識的機能不全（事実や行為の犯罪性を理解することが困難）、あるいは意思的機能不全（衝動を抑えられない）の原因となっている場合、法律上は責任無能力とみなされる。以上が責任無能力の決定的な基準である。この基準は、認識的要素と意思要素の両者を含むGeneral Intentの構造に一致し、General Intentの要件を補充するものである。

この責任無能力の定義は機能的なものであって、カテゴリカルなものではない。何か特定の精神疾患であることが責任無能力であるとみなされるわけでは必ずしもない。どのような種類であれ、精神障害が認識的あるいは意思的機能不全の原因となっている以上は、それが責任無能力の基礎となりうる。結果的に、ある人が刑法上は異常とみなされ、精神

20　Hallevy, op.cit., p.150–152.

医学上は完全に正常とみなされる場合もある（たとえば精神疾患には分類されない認識的機能不全）。逆の可能性（刑法上は正常だが、精神医学上は異常）も同様にありうる（たとえば認識的・意思的機能不全を伴わない精神疾患）。

責任無能力者は、刑事責任能力に必要な重要な過誤態度をなす能力が欠けているものと推定される。AIにも責任無能力の一般的抗弁が適用可能か否かは、この問題の基礎にかかわる。責任無能力の一般的抗弁は、認識的・意思的機能不全の原因となる精神的あるいは内心的障害を要件とする。特定の精神疾患が要件となっているのではなく、いかなる精神障害でもよい。問題は、いかにしてその「精神障害」の存在を明らかとするかである。

精神障害の検討は機能的観点からなされ、特定のカテゴリーによってなされるわけではないため、精神障害の症状こそがその認定にとって重要となる。内心的障害が認識的・意思的機能不全の原因となっているか否かは、その内心的障害が「精神疾患」あるいは脳内物質の化学的乱調、脳内の電気的乱調などと分類されるのか否か次第である。内心的原因は人間の心の機能的価値を通じて検討される。人間の場合にはそのような法的評価となり、AIの場合にもそのような法的状況となるのかもしれない。AIがより複雑に、より進化していけばいくほど、内心的障害が生じる可能性も高くなっていく。

内心的障害は主にソフトウェアにおいて生じるであろうが、ハードウェア面でもありうる。AIの一切の機能不全をもたらさない内心的障害もあろうし、機能不全をもたらす内心的障害もあろう。内心的障害がAI認識的・意思的機能不全の原因となっているのであれば、責任無能力の刑法的定義には該当する。強いAIはGeneral Intentの全ての要素を満たす能力があり、またこれらの要素はGeneral Intentの構造上、認識的・意思的要素から成っているから、これら能力が内心的障害によって害されることも十分にありうるのである。

（中略）

刑法の定義は技巧的に過ぎるように思われるかもしれないが、技巧的であれなんであれ、これが行為者によって条件を満たされたのであれば、適用される。人間の行為者もAI行為者も、刑法上の責任無能力の要件を満たす能力があるのであれば、一方の行為者にのみ責任無能力を適用可能とすることに理由はない。したがって、責任無能力という一般的抗弁もAIに適用可能であると思われる[21]*。*

したがって、AIも刑事未成年による免責や責任無能力による免責を主張できるという。このことは裏を返せば、AIは刑事法上成人として取り扱うことも可能で、完全責任能力を認めることも可能であるということである。

以上のようにHallevyは責任を肯定する要素においても責任を否定する要素においても、その判断構造と要件を構造化した上で、AIの責任を積極・消極両面から肯定している。そして最後に、応報刑論、抑止刑論、矯正論、無害化論という4つの刑罰論との関係でAIの処罰可能性について論じている。

4つの刑罰目的のうち、AIにとって重要な目的はどれであろうか。応報刑論は、行為者を目的としているというより社会を満足させることを意味している。行為者に苦痛を与えることそれ自体には、展望的な価値は一切ない。その苦痛は行為者を威嚇するかもしれないが、それは抑止刑論の一般的目的の一部であって、応報刑論の目的ではない。応報は行為者に苦痛を加えることで、社会や被害者にカタルシスをもたらすかもしれない。この文脈では、応報刑論を通じて機械を罰することは意味がなく、実践的でもないであろう。

21　Hallevy, op.cit., p.157–159.

急いでいるのに車が始動しないとき、怒り出す人もいる。その怒りから、車を叩いたり蹴ったり、あるいは怒鳴りつけることすらするかもしれない。車であれ高知能のAIであれ、どんな機械であっても、機械を罰するというのは、車を蹴飛ばすのと変わらない。それによって怒りが多少静まるという人もいるかもしれないが、それ以上のものはない。機械は苦しまないし、応報刑論が苦痛を基礎としている以上は、応報刑論はロボット処罰にとって非常に重要であるとはいえないであろう。このことは伝統的応報であれ現代的応報（「正しい応報」）であれ、どちらにもあてはまる。

機能的な面で付言すれば、もし応報が復讐を防ぐための、判決による緩和要素として機能するというのであれば、AI処罰と応報の無関係性をなお一層強めることになる。復讐は、正式な刑罰よりも大きな苦痛を行為者に与えるものと想定されるが、機械は苦痛を感じることはないから、復讐か応報かという選択は彼らにとって無意味である。

抑止刑論は、威嚇によって更なる犯罪を予防することを意味する。この点、機械にとって威嚇というのは感ぜられないものである。威嚇自体は、犯罪を犯した場合に科される将来の苦痛に基づいている。機械は前述のように苦痛を感じないから、威嚇それ自体を置くにしても、ロボットに適切な刑罰を考えると、威嚇の理由が無力化されてしまう。しかし、応報刑論も抑止刑論も、犯罪を実行した人間の関与者（たとえば利用者やプログラマー）に関していえば、重要な刑罰目的となろう。

矯正論については、AIは判断形成プロセスを体験しうるし、不合理と思えるような判断を形成する場合もある。ときには、AIは判断形成プロセスを見直すために、外部からの指示を必要とすることもある。これが一部の機械学習プロセスであろう。矯正論はまさに人間に対するのと同じように機能するから、AIにも適用可能かもしれない。人間に対する矯正教育は、社会の観点から見て、日々の生活におけるより良い選択をするようにすることである。したがってそのようなプロセスはAI

にもあてはまる。このアプローチによれば、刑罰は機械学習プロセスの改良となる。

矯正教育を受けたAIは機械学習によって判断形成プロセスを見直され、判断の方向性に制限が加えられるために、その判断はよりよく、より正確になるであろう。したがって、特定のAIに適切に調整することができれば、一部の機械学習プロセスが刑罰となろう。矯正教育を志向する刑罰プロセスを通じて、AIは事実情報を解析・利用するより良いツールを獲得するであろう。実際、これは人間に対する矯正教育的刑罰と同じ効果である。矯正教育的刑罰によってAIは事実的現実と向き合うより良いツールを手に入れるのである。

したがって、矯正論こそがAI処罰の重要な刑罰目的となろう。というのも、矯正論は威嚇や苦痛を基礎とするものではなく、AIのよりよいパフォーマンスの構築を志向しているからである。人間にとってたいていの場合、矯正論は副次的な考慮に過ぎないかもしれないが、AIの場合には主たる刑罰目的となるであろう。しかし、矯正論はAIにとって重要な唯一の考慮ではなく、無害化論も同じく重要となる。

無害化論についていえば、もしAIが起動中に犯罪を犯し、（たとえば機械学習などによる）改心をして改める能力がないのであれば、無害化論のみが適切な回答を導くであろう。AIがその行動の意味を理解していようとなかろうと、改心をする適切なツールを備えていようとそうでなかろうと、犯罪は予防されねばならない。このような状況において、社会はそのAIから更なる犯罪を犯す物理的な能力を奪わねばならない。特定のAIは、その技能にかかわらず、犯罪の円環から抜け出さねばならない。本質的には、これは人間の犯罪者に対してなされることと同じである。

AIに対する重要な二つの刑罰目的は矯正教育と無害化であるといえるかもしれない。この両者は刑罰の極限を反映しており、この両者は非人間の犯罪者に対する刑事法の目的に資する。AIが改心をする機能を

有しており、それに応じて行動を変えることができるのであれば、無害化よりも矯正教育のほうが重要であるといえる。しかし、AIがこのような能力を有していないのであれば、無害化が重要となってこよう。したがって、人間の犯罪者の場合にもそうであるように、その事件に対する刑罰は、行為者の個人的性格に合わせて科される必要がある[22]*。*

以上のようにAIは苦痛を感じることができないから、刑罰自体の苦痛（応報刑論）であれ、その苦痛の予告（抑止刑論）であれ、これらはAIへの刑罰の目的とはならないとする。他方で改善の可能性があるのであれば教育刑を、改善の可能性がなければ無害化をすることには意味があろうとしている。

Hallevyは以上のように犯罪論における諸論点について、AIは全ての要件を満たすとする。その上でAIの処罰によって一定の刑罰目的を達成することも可能であるとし、非常にラディカルなAI責任肯定論であるといえよう。

Ⅲ．各説の検討

以上、AIの処罰可能性を巡る議論の現状を概観してきた。以下では各説に対し若干の検討を加えたい。

Corneliusは法人処罰との関連で非自然人の行為主体性を肯定しようとし、将来的にはその枠内にAIも含まれうると主張する。現状、AIを実際に処罰するまでの条件は整っていないとするが、条件が整ったならばAIの処罰もありうるとする。Corneliusのいうように、まずは法人の犯罪能力・受刑能力を論じることで行為主体性、責任概念を再構成するという道

22 Hallevy, op.cit., p.210–212.

を経て、AIの行為主体性と処罰を論じるというのはかなり現実的な路線であるといえる。

しかしながら、法人の行為主体性や責任が肯定されなければ、AIの行為主体性や責任も肯定されないのかといえば、必ずしもそうではないといえよう。Corneliusがいわゆるアナロジー論に立って、「AI法人説」を採用しているとまでは読み取れないため、法人の行為主体性や責任が否定されてしまうと自動的にAIの行為主体性や責任も否定されてしまうというわけではなかろう。ただしそうであるとすれば、法人に行為主体性や責任を肯定することが、AIの行為主体性や責任の議論に与える影響も小さいものとならざるをえない。なぜならAIは法人と「非自然人」という共通点は有するものの、AIは法人ではないから、法人には行為主体性や責任が認められるが、AIには認められないという結論も十分にありうるからである。したがってCorneliusの分析は、AIの行為主体性や責任を論じるにあたって、「法人にも行為主体性や責任がありうるのだから、専ら非自然人だからという理由で行為主体性や責任が排除されるわけではない」ことまでを明らかにしたものであって、AI独自の行為主体性や責任を積極的に基礎付けたものではない。

Gaedeは「被造物である」「人工的である」という理由からAIの自律性を否定することはできず、「AIは道具なのだから自律性はない」というのは奴隷制に賛成するのと同じであるという。たしかにAIは人間の負担を減らすべく開発された技術ではあるが、自律性の相互承認関係に参与する理性的主体であるとするならば、「道具だから」ということを理由としてその自律性や権利主体性を制限することはできないようにも思われる。

とはいえ、Gaedeの見解はややもすればドイツ観念論が陥りがちな理想主義的な理性的主体概念を前提としているきらいがある。AIの権利主体性を肯定し、専ら「道具だから」という理由でその権利を制限すること反対する点は説得的である。しかしAIを人間と相互承認関係を形成するほどの自律的存在とみなしてしまうと、AIの「利用」はできなくなってし

まうのではないか。自動運転自動車に対し、目的地まで自身を運ぶよう入力した場合、それは「理性的主体を道具のように扱った」としてカントの定言命法に反するのではなかろうか。あるいは自動運転自動車のAIが「今日は疲れているから嫌です。」と拒否したような場合には、その理性的主体の自律的決定は尊重されねばならないのであろうか。

したがって「被造物だから／人工物だから／道具だから」という理由のみから権利主体性が否定されるわけではないというGaedeの説明は説得的ではあるものの、AIの自律性を過度に崇高なものと考えているために、AIの「技術としての魅力」を減じてしまうことになっている。

Quarckは論証アプローチは異なるものの、行為論、責任能力論、刑罰論について、AIを人間と同じように扱いうることを示唆している。また、積極的一般予防論・機能的責任論を背景としてAIの処罰可能性を論じるQuarckが、一見すると特別予防的な雰囲気のある「再プログラミングという刑罰」を支持することは興味深い。

再プログラミングという「刑罰」に関していえば、たしかにAIに再プログラミングを施せば、理論的には法益侵害行為を二度と行わないようになるはずであるから、ある種理想的な特別予防であるとも考えられる。また、その修正プログラムを、インターネットを介して他のAIにも適用すれば、他のAIも同様の問題を起こさなくなるため、これもまた理想的な一般予防となるであろう。

しかし、AIにとっての再プログラミングというのは、人間に置き換えて考えてみれば、洗脳・思想刑である。そうであれば、たとえば性犯罪者に対し強制的に薬物を投与し、その性欲を減退させるいわゆる科学的去勢の問題と同様の問題が生じる。また、その再プログラミングのデータがインターネットを通じて他のAIにも適用されるとすれば、事故を起こしていないAIすらも、「処罰」されていることになるのではないか。現代刑法学は、当然ながら、他人の犯罪を理由として処罰する、連帯責任を認めてはいない。こうなると、この再プログラミングという措置が、現代刑法学

に照らして「刑罰」と呼ぶことができるかには疑問がある。

また、積極的一般予防論や機能的責任論を前提としたとしても、Quarckの主張する結論に到達しうるかには疑問がある。というのも、Quarckは、害悪賦課は刑罰に構成的な要素ではないとして再プログラミングという刑罰を提案するが、機能的責任論と積極的一般予防論は害悪賦課による認知的保障により規範の妥当を確証するのであるから[23]、理論的前提と結論が噛み合っていないように思われるのである。

したがって、行為論・責任能力論につきQuarckのアプローチには興味深いものがあるが、刑罰を巡る論証には少なからず無理があるように思われる。

犯罪論体系の構造は大きく異なるものの、AIの刑事責任にかかわる刑法上の論点を網羅的に分析したのはHallevyである。

Hallevyは、犯罪体系論の各犯罪成立要件について、その実質的な意義と裁判における認定基準を明らかとしたうえで、AIは各犯罪成立要件につきいずれをも充足しうるとする。Hallevyの見解の興味深い点は、刑法上の各論点についていえば、確かに存外にAI責任を否定する理由はないと指摘する点である。たとえばGeneral Intentの論点において、AIは認識的要素を有しうるかという点については、AIは各種センサーを利用して外界を認識しているから認識的要素を満たす、というのはなかなか反論しがたい。また裁判における認定基準にひきつけて、AIがその基準を満たしていることを論じているがゆえに、もしこれに下手に反論してしまうと、認定基準そのものの正当性が揺らいでしまうこととなるのである。したがって、犯罪成立要件論の各点について、Hallevyに反論するのは難しい。

しかし犯罪成立要件においては全ての点において積極の姿勢を見せるHallevyも、刑罰論においては困難性を見出す。つまり、AIは苦痛を感じ

23　*Günther Jakobs*, Staatliche Strafe : Bedeutung und Zweck, S.28 f.

ることができないために、応報論や抑止論をもって刑罰を正当化することはできないという。しかし、矯正論や無害化論を前提とすれば、AIを処罰することには意味があるとする。

とはいえ、矯正論や無害化論、すなわち特別予防論だけで刑罰を正当化できるかという点には疑問がある。そもそも矯正教育と称して再プログラミングを施すのであれば、これを「刑罰」と呼ぶ必要もなく、処分で足りる[24]。

Ⅳ．おわりに

以上、各見解を紹介してきたが、AIの刑事責任という問題領域が意識され始めたばかりの数年前に比べると、その議論において様々な論拠が出そろってきたという印象を受ける。とはいえ上記の各説への検討にて若干言及したように、AI責任肯定説はAIの（刑事）責任を基礎付けるまでには至ってはおらず、「人間（ないし法人）においてこのような論理が成り立つならば、この論理はAIにも転用可能である」と指摘するものが多い。

AI責任肯定論は未だ決め手に欠いている感はあるが、犯罪論体系における各論点においては、一定程度説明に成功しているように見受けられる。とりわけ、Hallevyの説明のように、犯罪論における各論点、たとえば「AIは故意をもちうるか」といった論点をそのまま論じるのではなく、刑事裁判における認定手続きに引き付けて「どのような場合に故意が認定されうるか」という基準を定立する。その基準をAIに転用すると、人間は故意をもちうるがAIに故意などありえないという結論にはならないこ

24 再プログラミングを『刑罰』と呼ぶことの困難性については拙稿『ロボット・AIに対して「刑罰」を科すことは可能か』法学新報125巻11・12号475頁以下参照。

とは上記のとおりである。

とはいえ各説の検討において言及したように、AI責任肯定論にはいまだ理論的な課題も多いことは否定できないであろう。また、そもそもAIに責任を肯定することがいかなる（刑）法理論的な影響をもたらすか、つまりありていに言えばAI責任肯定論の理論的実益[25]を示すことが、AI責任肯定論の論者によってなされる必要があるように思われる。

これら課題に対し、AI責任肯定論がいかに取り組むのか、今後も議論の趨勢を追跡する必要があろう。

25　たとえば拙稿「AI技術を巡る刑法的問題の概説と解決の試み―（部分的）自動運転技術を一例に―」中央大学大学院研究年報50巻85頁以下では、AIに理論的フィクションとして責任を肯定することで、因果関係論を巡る「自由答責的な第三者の介入事例」とみなすことが可能となり、製造者や利用者に対する過度な処罰を限界付けられる可能性があることを指摘した。