ロボット・AIに対する刑罰をめぐる最近の議論

法律論叢第 94 巻第 4・5 合併号（2022.2）

【論　説】

ロボット・AIに対する刑罰をめぐる最近の議論

川　口　浩　一

目　次

1　はじめに

「私がジョジーを継続することも、やればできたと思います。…継続できないような…特別な何かはあります。ただそれはジョジーの中ではなく、ジョジーを愛する人々の中にありました」カズオ・イシグロ（土屋政雄訳）『クララとお日さま』431 頁。

手塚治虫の『鉄腕アトム』において初めて「ロボット法」という法律が登場したのは「海蛇島の巻」（原題：「アトム赤道を行く」初出 1953 年！）においてである。そこでは「ロボット省の許可なくして無断で国をはなれ行動をなすものはエネルギー無期差し止めまたは解体の刑に処す」という条文が引用されており、そこでは無許可国外行動の罪に対する刑としてエネルギー無限差し止め刑と解体刑という 2

つの刑罰が規定されている(1)。それから 70 年近く経った現在、現実世界においてもロボットや AI に対する刑事責任や刑罰の問題が刑法学においても真剣に議論されるようになった(2)。そこで現在の日本における刑法学者の議論を中心に紹介し、解決のための一定の方向性を示したい。

2　日本の刑法学者による議論

日本の刑法学においても、主にドイツにおける議論の影響を受けて自律的な AI・ロボットに対して刑事責任を問い、刑罰を科すことが可能かという問題についての議論がなされている。

（1）消極説：まずロボット・AI の刑事責任・処罰に消極的な見解としては、深町晋也の見解(3)がある。深町はまず「伝統的な刑法学の立場からは、刑罰とは自然人のように、一個の人格が想定される存在にのみ科すことができるものとされている。このような観点からは AI に刑事責任を問う前提として、そもそも AI に人格を認めることが可能であるのかが問題とされる」とし、そしてこのような AI にも人格的な側面を想定することが可能か否かという問題設定には、さらに①「自然人との類似性という観点からその人格を基礎づけようとするアプローチ」（自然人類似性アプローチ）と、②「自然人との類似性を離れて、AI に人格を付与して刑罰を科すことが妥当か否かを問うアプローチ」（刑罰付与妥当性アプローチ）があるとした上で、この両アプローチを批判し、さらに③ AI に対して「不利益な処分を行うこと、例えば AI を破壊したり、プログラムを消去したりすることは、刑罰という形でなくともその危険性を理由として、責任能力を前提としない処分（保安処

(1) 黒沢哲哉「手塚マンガあの日あの時：第 21 回：手塚マンガのロボット年代記・後編」https://tezukaosamu.net/jp/mushi/201202/column.html（最終閲覧日：2021年 10月 27日）

(2) 私は、これまでこの問題に関して①拙稿「ロボットの刑事責任 – ロボット刑法（Roboterstrafrecht）序説」増田豊先生古希祝賀論文集（勁草書房・2018 年）125 頁以下；②「ロボットの刑事責任 2.0」刑事法ジャーナル 57 号（2018 年）4 頁以下；③「ロボット・AI の刑事責任論　2.5：日本の議論」2019 第二届人工智慧與法律國際學術研討會「新時代的法律衝擊：– 場 AI 與法律的國際思辨」大會手冊・高雄場（2019 年）87 頁以下を執筆した。

(3) 深町晋也「ロボット・AI と刑事責任」弥永真生・宍戸常寿編『ロボット・AI と法：ロボット・AI 時代の法はどうなる』（有斐閣・2018 年）209 頁以下。

分）を課すことは可能である」とする。そして「AIが人格的主体であるか、責任能力を有するかといった問題を論じなくとも、当該AIが危険性を有することを理由としてこうした処分を課すことは理論的に十分可能であ」り、「AIが刑罰適用の対象たり得るか」といった問題を論じることには「さほどの意義はないであろう」とする。そこから深町は「AIが真の意味で我々の社会の対等なメンバーであるとの認識が共有されない限り、AIに独自の刑事責任を問うという方向性は否定されるべき」であるとするのである。

（2）積極説

これに対して今井猛嘉(4)は、AIの刑事責任について以下の条件が具備されれば、肯定しうるとする：

「① 刑罰の目的として、応報刑論を偏重するのではなく、抑止刑論と社会復帰論の意義を再確認すること、

② 刑法の存在意義を、社会構成員の法益侵害の予防に求めること、

③ 社会の構成員は、人間に限られる必要はない。人間の仲間（fellow）として人間にとって重要な法益を侵害しうるが故に、侵害防止が義務付けられ、その義務が履行できる能力を有する存在であれば、当該能力の源泉を自由意思（free will）(5)と位置づけ、刑法の対象である社会構成員として評価されること、

④ AIの機能ないし能力が更に発展し、自律的学習能力が高まり、フレーム問題(6)にも一定の対処ができるようになると、AIも社会の構成員となること、

⑤ AIに対する刑罰は、法益侵害に寄与したアルゴリズムの改変等であり、AIの再利用（AIによる法益侵害の再発防止と、その社会復帰）に資するものであるべきこと(7)、

(4) 今井猛嘉「AI時代の刑事司法」罪と罰56巻2号（2019年）31頁。

(5) これは、意思自由擬制説に基づくものと考えられよう（意思自由擬制説批判として拙稿「責任能力と自由意志論」浅田和茂先生古希祝賀論集［成文堂・2016年］261頁以下）これに対して後述のように稻谷龍彦は「自由意志によって客体たる非人間を統制するという、近代的な人間像」自体を批判し、いわゆるアーキテクチャ論（に基づく新たな刑事法の枠組みの定立を提唱する（稻谷龍彦「人工機能搭載機器に関する新たな刑事法規制について」法律時報91巻4号54-59頁）。

(6) フレーム問題についてはさしあたり Daniel C. *Dennett,* Cognitive Wheels: The Frame Problem of AI, in: Christopher Hookway, ed., Minds, Machines and Evolution: Philosophical Studies, pp. 129-151 (Cambridge: Cambridge University Press 1984)（翻訳：ダニエル・デネット［信原幸弘・訳］「コグニティヴ・ホイール：人工知能におけるフレーム問題」現代思想15巻5号（1990年）128-150頁参照。

(7) この点は、今井猛嘉「自動運転、AIと刑法：その素描」日高義博先生古稀祝賀論集・上巻（成文堂・2018年）363頁以下でもAI・自動運転車に対する刑罰は、応報刑論や抑止刑論ではなく「社会復帰論」によって説明されるべきことが強調されている。なお同

⑥ AI に対する刑事裁判は、インターネット上で行われること」

すなわち応報刑論ではなく、法益保護論に基づいた予防刑論（特に特別予防論(8)）に基づき、ロボット・AI が法益の侵害防止に対応しうる一定程度の自律的学習能力を持つようになれば、その刑事責任を認め、アルゴリズムの改変などの刑罰を課すことが可能とするものであり、刑法学において法益保護論・抑止刑論が有力化し、ディープ・ラーニング(9)によって AI の自律的学習能力が高まりつつある現代の（日本）社会においては、すでにその条件は具備されつつあると今井は考えていると思われる。

このような「刑法の基本的価値（法益保護）に直結した、より機能的な刑法観を持つ」(10)べきだとする今井説に対し、松宮孝明は「しかし、これでは大震災を防止して人命という法益を保護することも、刑法の課題となりかね」ず、「機能的な刑法観」とは「『刑法の目的』としては疑問のある『法益保護』を当然の前提として、そのために最も効率的に機能するものが優れた刑法だとする見方ではなく、刑法が現実社会で果たしている機能―それは単純な法益保護ではない―を見定め、どのような条件が整ったときに AI の処罰がそのような伝統的な機能を果たすことになるのか、はたまたそのような条件が成就するときが到来しうるのかといった問題を考える刑法観であるように思われる」(11)と批判する。さらに松宮は、このような肯定説の立場に対して、刑法における責任の根底にある「非難可能性」の意味をどのように考えるのかが、明らかではないことも問題視する。すなわち「さらに問題なのは、仮に AI が倫理的判断能力を備える場合、それに法の期待する程度の遵法精神、とりわけ法令遵守が常に優越的な行動動機となることをプログラミングし

「自動車の自動運転と刑事責任」交通法研究 46 号 16 頁以下も参照。

(8) 上記のように今井はこれをロボット・AI の「社会復帰論」と呼んでいるが、これに対して佐久間修「AI による自動運転と刑事責任」刑事法ジャーナル 57 号（2018 年）12 頁は「そこでは、犯罪の予防に特化した刑罰論に近づくなど、近代刑法の基本原則である責任主義を軽視する嫌いがある」と批判する。なお同「AI と刑法・序説：自動運転車は『犯罪者』となるか？」名古屋学院大学論集：社会科学篇 55 巻 1 号（2018）107 頁以下も参照。

(9) ディープ・ラーニング（深層学習）については松尾豊「人工知能開発の最前線」法律時報 91 巻 4 号（2019 年）7 頁以下参照。

(10) 今井猛嘉「自動車の自動運転と刑事実体法――その序論的考察」山口厚ほか編『西田典之先生献呈論文集』（有斐閣・2017 年）535 頁。

(11) 松宮孝明「自動運転をめぐる刑事法的諸問題」立命館法学 395 号（2021 年）1 頁以下、8 頁。

てしまえば、後の学習によってこの遵法精神が退化する、または他の利益に劣後するという事態が生じない限り、AIは常に『法の期待する標準人』の態度を選択するのであり、そこに非難可能性の入る余地」[12]がなくなってしまうとされる。松宮によれば「一般的にはAIには標準的な規範的・道徳的能力が備えられているとすれば、その判断による動作が何らかの害を生み出したとしても、それは標準的な規範的・道徳的能力をもって行動を制御できる主体に比して道徳的に『悪い』とされる判断に基づいて行動したものではないから、AIに対する『非難可能性』を根拠づけることはでき」ず、「ゆえに、AIが事実として刑事責任を負う場面があり得るとすれば、それは、自然人に類似した『欲望』を持つ存在として、自律的学習によって『堕落』することも可能なプログラムが組まれていることを前提」とし、「それは、もしかすると、『死』を運命づけられた存在の『生き延びること』と『子孫を残すこと』という『欲望』を生み出すプログラムかもしれない」が、「しかし、現在、AIにあえてそのようなプログラムを組み込むことには現実性がない」[13]と批判するのである。これは妥当な批判であり、むしろそのような存在をプログラミングによって意図的に作り出すことは[14]、その欲望が達成されない場合の「苦痛」を考慮した場合、道徳的にも許されるべきでは行為ではないと考える。

（3）中間説

このような積極説に対して根津洸希も、2017年に公刊された論文「ロボットの処罰可能性を巡る議論の現状について」[15]では、深町と同様ロボット・AIの処罰可能性について消極的な見解を示していた。しかし、最近の論文「ロボット・AIに対して『刑罰』を科すことは可能か」[16]においては、ロボット・AIの処罰に関して①刑罰を受けるのは誰かという問題（刑罰の対象者・人格の一個性問題）、②刑罰「苦痛」の問題、③近代刑法的刑罰と呼べるかという問題、④ロボット・AI

(12) なお松宮は、その他の問題として、①運転者資格の変容、②公共交通機関か自家用車か？（特に地方における交通渋滞の激化）、③交通法令の見直し、④事故調査制度と刑事責任の問題をあげている（松宮・前掲注(11)17頁以下）。
(13) 松宮・前掲注(11)9頁。
(14) それを主張する見解として浅田稔「痛みを感じるロボットの意識・倫理と法制度」人工知能33巻4号（2018年）450頁以下。
(15) 根津洸希「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年）145頁以下。
(16) 根津洸希「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（2019年）475頁以下。

を処罰することによる刑罰の意味変容の問題の 4 つを検討する。

まず①に関して自動運転車両を例として「人格」の一個性が議論される(17)。車両一台ごとを一人格と考え、刑罰を加えることには、その刑罰自体に意味（特に再発予防効果）があるのかという疑問が残り、また現在の自動運転技術はスタンドアローンではなく、インターネットを経由して他の車両と情報を交換しながら走行するコネクテッドカーであり、「各車両は独立した個人というよりは、ネットワークによって構成された情報の総体の一部、人間でいえば身体の一部のようなもの」(18)であり「サーバーを介して接続されている全車両やそのインフラ全体を、総体として一人格」とみなし、それに刑罰を科すことは「一般予防的見地から正当化されうるかもしれない」(19)とする（但しその処罰によって交通インフラ全体が麻痺する可能性があることも指摘する）。このような人格確定の困難性を解決する視点を、法的人格が「人工的な構成物」であることを強調するケルゼンの見解や破産財団などにいわゆる「暗星的法人」性を認める兼子一(20)の学説などを参照し、法人格が結局その「社会的有用性」(21)という観点に求めるならば、処罰の対象として「どの範囲までを一個の人格として扱うことが有用か」(22)というアプローチから確定すべきとする。

次に②の刑罰の「苦痛」問題については、まず刑罰の与える苦痛ないし不快(23)を主観的に捉える見解からは、少なくとも現在のロボット・AI は自然人のような主観を持たないので、それは否定されることになろうが、自然人においてもホームレスが刑務所に入るために犯罪を犯した場合などのようにその苦痛性は行為者の主

(17) 根津・前掲注(16)478 頁以下。
(18) 根津・前掲注(16)479 頁。
(19) 根津・前掲注(16)480 頁。
(20) 原文では兼子仁（まさし）となっているがこれは兼子一の誤りである。兼子一『民事法研究一巻』（酒井書店・1971 年）450-473 頁参照。なお現在では、この説を批判し破産財団の法人性を否定する説が有力化している（岡伸浩「破産管財人の法的地位・序説：管理機構人格説の再定位と信託的構成との調和」慶應法学 40 号（2018 年）23-59 頁など参照）。
(21) 小林史明「権利主体性の根拠を AI・ロボットから問い直す」増田豊先生古稀祝賀論文集（勁草書房・2018 年）147 頁。
(22) 根津・前掲注(16)482 頁。
(23) ミヒャエル・パヴリック（Michael Pawlik）は、これを刑罰の「不快さ（Ärgernis）」と呼んでいる。Michael *Pawlik,* Das Unrecht des Bürgers: Grundlinien der Allgemeinen Verbrechenslehre, Tübingen 2012, S. 26 ff.

観性に依存するものではないとする「客観的害」説（高橋直哉(24)）や、その刑罰によって受刑者が苦痛を感じているように「見える」かが重要であるとする説（小林史明(25)）が主張されていることを指摘する。そしてこの両者の差異と共通点を検討し、両者に共通の要素として刑罰を科される側と一般人との「立場の互換性」という要素が挙げられるが、これは「観察者と立場の交換相手との一定の均質性への信頼」（共感・同情）を前提としているとする。現状では、例えば自動運転車両などに共感や同情を感じることは難しいが、「今後の技術の発展の次第によっては、ロボットやAIに対して科された『刑罰』を見て応報が達成されたと感じることも可能となろう」(26)とする。しかしこのようにロボット・AIにも刑罰苦痛を観念する余地が生じたとしても現行刑法にはそれに適した刑種がないとし、例えば今井が提案する「再プログラミング」刑(27)は、確かに再犯防止にも資するが、人格改変・思想刑（洗脳）につながる可能性もあり、これを肯定するならば、自然人にも同様の刑罰（例えば化学的去勢）を科すことも肯定されることになるのかなどの問題があるとする。

さらに③の問題については、スタンドアローンで機能するロボットについては意味が認められようが、上述のコネクテッドカーの場合、犯行を行なった車両だけに再プログラミング刑を科したとしても、その刑罰には意味はなく、「再プログラミングにおいて取り除かれた『バグ』に対するアップデートデータは、同車種・同型AIにも適用されなければ、再犯防止の観点からは全く意味をなさない」とされる。しかしその場合には一種の連帯処罰であり「個人責任の原則」という近代刑法の原則との抵触が生じるとする(28)。ただしロボット・AI・自動運転車両といった先進技術に関しては、既存の枠組みを超えた「新たなパラダイムへの移行」が求められることもあり、そのような④の刑罰の「意味変容」の問題については、それに伴う副作用の問題を、伊藤康一郎のリスク社会論(29)を参照してロボットなどへの処

(24) 高橋直哉「刑罰の定義」駿河台法学24巻1・2号（2010年）522頁以下。
(25) 小林・前掲（注21）151頁以下。これを根津は「見える」説と呼んでいる。
(26) 根津・前掲注(16)487頁。
(27) 今井猛嘉「自動車の自動運転と刑事実体法：その序論的考察」『西田典之先生献呈論文集』（有斐閣・2017年）529頁。
(28) 根津・前掲注(16)490頁以下。
(29) 伊藤康一郎「理性と感情：リスク社会化と厳罰化の交差」犯罪社会学研究31号（2006年）74頁以下。

罰要求が単なる不安や誰も責任を取る者がいないことなどへのストレスに起因するならば、感情的なポピュリズムの問題性が生じるとし、刑罰が単なるストレス解消手段に堕してしまうようであれば、そのような意味変容に対して我々プロフェッショナル（としての刑事法学者）は反対せねばならず、逆に「先進技術の利点を享受し、欠点を補う法規制により、よりよき社会を構想する一環としてAIやロボット」を処罰する場合には、「刑罰は責任非難であるという意味の枠内において」「市民的討議に道筋を示すことが使命である」(30)とする。

以上の検討を踏まえ、根津は、次のような結論に至っている。

「たとえば事故を起こした自動運転車両に『刑罰』を科す際、どこまでが一人の人格なのかという問題が生じるものの、その範囲を有益性の観点から画定する余地もありうる。その一人の人格に対し、現行法は有効な刑種を予定していないが、再犯防止の観点からは再プログラミング措置が考えうるところであり、その措置を施す際にロボットやAIが苦痛を感じているように『見える』のであれば、この問題も克服しうる。再プログラミングという措置の技術的特性と、未だ犯罪を行っていないロボットやAIないし、自動運転車両にも、刑罰内容と同様のアップデートを施す必要があるため、個人責任の原則との抵触が考えられるが、ロボット法という新たな制度構築や原則の修正によって、あるいは人格の範囲を改めて検討することによって、この問題もまた克服される余地はある。」(31)

そこから根津は「ロボットやAIに『刑罰』を科すことは、若干の無理を承知でいえば、理論的には全く可能性が無いわけではない」というテーゼを提示しているが、以下のような留保をつけている。

「ロボットやAIに「刑罰」を科すことは、理論的には全く可能性が無いわけではないが、その「刑罰」は真に「犯罪に対する責任非難」という枠内で用いられているかはなお慎重に検討されるべきである。単なる被害への不安から、スケープゴートとしてロボットやAIに「刑罰」を科すのであれば、それは「刑罰」の意義を変容させてしまう。」(32)

このような見解は、厳格な要件を示しつつも、処罰の可能性を完全には否定しない点で消極説と積極説の中間に位置するものなので「中間説」と呼ぶことができよ

(30) 根津・前掲注(16)495頁。
(31) 根津・前掲注(16)496頁。
(32) 根津・前掲注(16)496頁以下。

う。そして最近の論文において、後述のレベル4以上のAI搭載の自動運転車の答責の間隙問題に関して、この中間説の立場から解決案を提示している(33)。

私自身の見解も、この中間説に属するものである。ただ私見は、その前提とする刑罰論を予防刑論ではなく自由論的な応報刑論が基本的に妥当だと考えている(34)ので、ロボット・AIが規範に従って行為を行い、そのことによって市民としての協働義務を果たす能力と用意があることが処罰の要件となろう。すなわちこの自由論的応報刑論を主張するミヒャエル・パヴリック（Michael Pawlik）によれば、不法形式（Unrechtsformen）は、①「人格の不法（Unrecht der Person）」、②「主体の不法（Unrecht des Subjekts）」および③「市民の不法（Unrecht des Bürgers）」に分類され、刑法的不法は③の「市民の不法」と位置づけられる(35)。この市民の不法は、行為者が、その具体的な被害者（Opfer）に対する義務に違反するのみならず、同時にその「法による平和（Friedens durch Recht）」という市民的共同プロジェクトに対するロイヤルティの義務づけ（Verpflichtung zur Loyalität）にも違反した場合に認められるされるのである。そしてこのロイヤルティの拒絶（Loyalitätsverweigerung）に対する応答（Antwort）が刑罰であり(36)、自己のコストにおいて「協働義務の履行（Mitwirkungspflichterfüllung）と自由の享受の相互関連性（Wechselbezüglichkeit）」(37)が確認されることになるのである。さらにそのような一種の規範的コミュニケーションが可能となるためには、ロボット・AIに与えられたアルゴリズムに従うだけはない、自律性が認められなければならないであろう。このようにいかなる刑罰論を正当と考えるかということによって、ロボット・AIを処罰可能かどうか、さらにそのための要件は何かという問題への答えも変わってくるのである。

(33) 根津洸希「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―」大学院研究年報50号（2021年2月）85頁以下；Koki *Nezu,* Strafrechtlicher Problemaufriss von [teil] autonomen Fahrzeugen in der Gegenwart und Zukunft —Darstellung möglicher Lösungsansätze, Hanover Law Review 2019 Heft.4, S.268 ff.

(34) 拙稿「ロボットの刑事責任－ロボット刑法（Roboterstrafrecht）序説」前掲（注2）125頁以下参照。

(35) *Pawlik,* Person, Subjekt, Bürger, Berlin 2004, S. 75 ff.

(36) *Pawlik,* o. Fn. 23, S. 82 ff.

(37) *Pawlik,* ZIS 2011, 264.

3　自動運転問題への応用

【事例 1】「X は自動車ディーラーを訪れ、試乗を申し出た。X は同店舗の店員である Y と試乗を開始し、交差点に差し掛かったところ、前方を走行していた車両が突如急停止した。十分な車間距離は取っていたにもかかわらず、非常に突然のことであったためおよそ人間の反応速度では対応できず、X はブレーキを踏むことができなかった。この際、本来であれば同車両に搭載されている自動ブレーキシステムによって適時に制動がなされ衝突には至らないはずであった。しかし当時は晴れていて見通しも良かったにもかかわらず、実際には何らかの原因で同システムは作動せず、X の運転していた車両は前方に停止していた車両に衝突し、同車内に乗車していた夫妻に傷害を与えた。」(38)

根津は、この事例のようにレベル 4 以上の自動運転車において試乗者（X）には結果回避の措置を取る術はなく、唯一結果回避が可能であった運転自動化システムも何らかの原因から作動しなかった結果、傷害結果が生じたという例をあげて、このような事例におけるズザンネ・ベック（Susanne Beck）のいう「答責の間隙」(39) をいかに埋めるべきかという問題を検討する。すなわち根津によれば、答責の間隙問題を解決するために現在、①過失犯論を修正して製造者・利用者に広く答責するという引き受け過失の法律構成、②あらかじめ法律により答責主体を規定してしまうという立法的解決、③新技術の長所も短所も社会全体で引き受ける社会的受容という構想、④ AI に法的人格性・自由答責的な主体性を肯定することで関与した人間の答責を相対的に限定する法律構成が提案されている (40)。根津は、これらの答責の間隙問題の解決策のうち、結論的には個人や企業に過大な責任を負わさないためのいわば「防波堤としてのロボットの責任」を認めるという④の解決策を主張する (41)。根津はまず①②の解決法については、比較的現実的な解決策ではあるものの、どちらも答責の間隙を人間の答責領域の拡張によってカバーしよう

(38) 根津・前掲注 (33)85 頁以下。
(39) 根津 洸希「文献紹介『ロボットと法』シリーズの論文紹介 (2) スザンネ・ベック『グーグル・カー、ソフトウェアエージェント、自律的武器システム：刑法にとっての新たな挑戦？』」千葉大学法学論集 31 巻 3・4 号（2017）187（92）-172（107）頁参照。
(40) なお山下裕樹「AI・ロボットによる事故の責任の所在について：自動運転車の事案を中心に」Nomos45 号（2019 年）95 頁以下は、現状では AI・ロボットの処罰について消極説に立ちつつ、背後の自然人の組織化の有無によって答責範囲を確定することを試みる。
(41) 根津・前掲注 (33)85 頁以下。

とする立場である。しかし、AI技術、たとえば自動運転技術などは人間の運転タスクを軽減するために用いられるものであるのに、自動運転技術を利用すると自ら運転するときよりも高い処罰リスクにさらされてしまうのではむしろ負担は重く、その技術の本旨に反する。これでは人間の負担軽減のために開発されたAI技術の利用によって、自らのコントロールが及ばない領域についても責任を負わされるという皮肉な帰結をもたらしてしまうため、妥当な結論とはいえないとする(42)。次に③は一部の当事者に負担を負わせることに反対するため、理念的には正しいようにも思えるが、法律論における具体的な帰結が明らかではない。それゆえこの理念を具体的な法律構成に落とし込む必要があるとし、AIに理論的フィクションとして責任を肯定し、責任分配の当事者としてカウントすることで製造者や利用者の答責領域を限定する④の見解が主張される。具体的には、AIに責任を観念することでAIを自由答責的な行為者であるとみなすという方法がとられる。すなわちAIが法益侵害結果を惹起した場合には、そのAI自身を自由答責的な第三者とみなし、その背後にいる利用者や製造者への結果帰属を否定する。したがってAIに責任を認めることで、コントロール不能な法益侵害結果につき、利用者や製造者を処罰リスクから救い出すことができ、いわば、AIの責任が「背後の人間を処罰リスクから守る防波堤」のような役割を果たすとされ、これによって新技術の長所と短所を社会全体で負担するという妥当な責任分配が達成され、このように処罰リスクを限界付けることは、ひいては技術発展にも資するものであるとされるのである(43)。このような立場から根津は【事例1】について「事故当時Xには取りうる結果回避措置がなく、システムが作動しなかった原因も不明であることから製造者・販売者の過失も立証不可能だが、直近行為者たる運転自動化システムが何らかの原因で（あるいは「自らの意思で」）停止すべきところを停止しなかったために事故が生じ、その結果前方に停止していた車両に乗車していた夫婦に傷害を与えたと解され、本件運転自動化システムAIに結果が帰属される」(44)とするが、結論的には「・・・AIに責任を肯定するということは、それが直ちにAIの行為も非難に値するだとか、AIにも刑罰を科すべきだとかいう帰結をもたらすものではない」

(42) 根津・前掲注(33)85頁以下。
(43) 根津・前掲注(33)85頁以下。
(44) 根津・前掲注(33)85頁以下。

とする。その理由として根津は「責任は刑罰を科す際の一つの要件であって、責任の存在が刑罰を要求するわけではない（消極的責任主義）」ことと「また本稿が示した AI の責任は、AI の技術的利点を損ない、また技術的発展を阻害しかねない過度な処罰を抑制するための理論的フィクションとして仮設されるものであって、『罪を犯した AI』を人間の犯罪者と同じく扱うべきことを（少なくとも現在は）主張するものではない」ことを挙げている(45)。しかし逆にこのようなフィクションによって製造者や利用者を過剰に免責してしまうというリスクはないだろうか。実は製造過程や利用方法に何か問題があった場合に（そのような場合には、根津も製造者や利用者の可罰性を認めているが)、いわば AI の「ブラックボックス性」を隠れ蓑にしてしまうことになる危険も否定できないようにも思える。また刑法における責任概念は、やはり刑罰を科すための要件であって、刑罰やその他の制裁を課すことを一切前提としていない（刑法上の）責任というものが観念できるのかも疑問である。処罰はしないが、責任はフィクションとして AI にあったことにして事案を処理するということで、被害者や社会の納得が得られるかということも疑わしい。このような事例は、関与した自然人の誰にも責任が問えない場合には、AI を利用する場合に避けることのできない「許された危険」として社会全体で引き受けるという方法（前述③の方法）で処理すればよいのではないだろうか。

4　訴訟法的解決（DPA／NPA 制度の導入）

稲谷龍彦も、根津と同様に、以下のような事情から、AI やロボットの開発者やメーカーの処罰の問題性を指摘する。すなわち、人工知能が原因で事故が起こった際、システムを開発した研究者や企業を刑事罰の対象にするという解決策については、上述のように開発者が書いたプログラム通りに動く従来の電子機器とは異なり、人工知能は、蓄積され続ける膨大なデータを学習してより高い判断力を身に付けていくため、そのメカニズムは変化し続け、開発者にもわからないほど複雑化・ブラックボックス化するので、メーカーや開発者に刑事罰を科した場合、リスクが高すぎて誰も開発に取り組まなくなってしまう一方、人工知能のミスの責任を利用

(45) 根津・前掲注 (33)98 頁。

者が負うことになれば、誰も製品を買わなくなってしまうということになってしまう。したがって、製造者や利用者に、自分の意志や判断で起こったわけではなく、人工知能の判断ミスによって被害が発生した事故の責任を負わすのは、現在の法制度では困難であるとし、人工知能の判断により引き起こされた事故の責任に、空白が生じ得る状況（すなわちズザンネ・ベックのいう「答責の間隙」）が想定され始めており、このような状況においては「メリットとデメリットを天秤にかけ、メリットが大きければ人工知能を普及させた方が良いと判断できる」ため、「社会的便益を優先させ、何らかの事故が起こったとしても処罰はしないという功利的な考え方」、自動運転車を例にすれば「人間が運転するときより事故を減らせるという社会的な恩恵を優先し、事故の際には被害者に対する民事的な補償のみで、誰にも刑事罰を科さないという選択があっても良いという考え方」も可能であるとする(46)。そして稲谷は「確率的な危険のシステムによる統制を重視し、情報提供と精神・開発体制の改善、被害者への補償などを自主的に行うことを検察官と約束し、その見返りに刑事訴追を免れるという制度」(47)、すなわち事故を引き起こした人工知能を流通させた企業に対して原因究明に必要なあらゆる情報の提供や、必要な場合には再発防止に向けた具体的な改善などを義務付ける代わりに、関係者の訴追を一定期間延期する訴追延期合意(DPA = Deferred Prosecution Agreement)や、関係者を訴追しない不訴追合意(NPA = Non-Prosecution Agreement)制度(48)を参考し

(46) 国立研究開発法人科学技術振興機構「近付く人間と人工知能の距離：互いに寄り添う新しい社会へ」JSTnews2019年5号8-11頁 DOI https://doi.org/10.1241/jstnews.2019.5_8（浅田稔、稲谷龍彦へのインタヴュー記事［最終閲覧日：2021年10月27日］）による。さらに稲谷龍彦「統治システムの近未来を考えてみる：Governance Innovation and Beyond」Nextcom44号（2020年）15-25頁も参照。なお森田果「自動運転・AIをめぐる望ましい法ルールのあり方：経済分析(機能的分析)の立場から」刑法雑誌59巻2号（2020年）332頁は、「安全・便利・低コストな自動運転の普及した社会の実現」という目標を設定した場合、「自動運転は、たとえそれが不完全なものであり、一定の低い確率で従来なら発生しなかったような類型の新たな事故を発生せしめるようなものであったとしても、その新たなリスク以上に、従来発生していた事故を減少せしめることができているのであれば、それを導入し、普及させることが、社会的に望ましい」とする。

(47) 稲谷龍彦「ロボット事故の刑事責任」日本ロボット学会誌38巻1号（2020年）37頁以下、40頁。

(48) このDPA/NPA制度に関しては稲谷龍彦「企業犯罪対応の現代的課題(1)-(7未完)DPA/NPAの近代刑事司法へのインパクト」（1）法学論叢180巻4号（2017年）40頁以下；（2）同181巻3号（2017年）22頁以下；（3）同183巻1号（2018年）1頁以下；（4）183巻3号（2018年）1頁以下；（5）同184巻5号（2019年）1頁以下；（6）同186巻2号

たもので、米国において企業の構造改革を目的として作られた制度なので、企業と同じく人間ではない人工知能の改善を目的として応用できるとするのである。稲谷は「この制度を実現するためには、適切な約束を行うための手続の設計（技術者の参加や訴追指針の策定・公開など）に加えて、法人処罰の拡大や行政規制と基礎との紐付けなどの法改正が必要となる」ことを指摘し、さらに「人間が事物の危険を統制するという基本的な認識枠組み」である「心身二元論や主客二分論」を「一部放棄し、技術とともに人間概念自体が変化していくことを、部分的にしろ受け入れる必要がある」[(49)] ともする。これはベックや根津の発想と同じく、答責の間隙を製造者や利用者の処罰の拡大によらずに解決しようとするものであり、ドイツと異なり起訴便宜主義を採用する日本の刑事制度には馴染みやすいものであろうが、検察の起訴裁量権の行使が恣意的にならないような制度的担保が必要であり、稲谷自身も指摘するように法人処罰の拡大などの様々な法改正や、責任や刑罰に対する根本的な考え方の転換[(50)] が必要となり、その実現の「ハードル」はかなり高いといわざるを得ないであろう。

5　自動運転とジレンマ事例（トロリー問題）

【事例 2】「X は、自動車会社 Y が製造・販売する自動走行車（レベル 4 または 5）を購入した。本件自動走行車は、衝突回避のシステムとして、急制動を行うことで衝突を回避するか、急制動では間に合わないと判断した場合には、衝突を回避するためにハンドルを左右に切るように設計されていた。ただし、こうした緊急動作を行ってもおよそ衝

（2019 年）1 頁以下；(7) 同 188 巻 3 号（2020 年）34 頁以下；同「企業犯罪に対する刑事手続の対応：アメリカ法における DPA・NPA を中心に」刑事法ジャーナル 58 号（2018 年）69 頁以下参照。さらに同「人工知能搭載機器に関する新たな刑事法規制について」前掲注 (5)54 頁以下；同「企業犯罪における取引的刑事司法」刑法雑誌 58 巻 1 号（2019 年）44 頁以下も参照。

(49) 稲谷龍彦「技術の道徳化と刑事法規制」松尾陽編『アーキテクチャと法：法学のアーキテクチュアルな転回？』（弘文堂 2017 年）所収 93 頁。

(50) そもそもこのような考え方は、安藤馨の統治功利主義的な発想（安藤馨『統治と功利：功利主義リベラリズムの擁護』勁草書房、2007 年）と類似したものであり、同「法と危険と責任と：提題」大屋雄裕／安藤馨『法哲学と法哲学の対話』（有斐閣、2017 年）144 頁以下で論じられている「新派刑法学の帰還」（同 156 頁）の当否を含め慎重な議論が必要であろう。

突自体は回避できない場合、例えば急制動をすれば後続の自動車と、ハンドルを左に切れば歩行者と、ハンドルを右に切れば後方から進行してきたバイクとの衝突が回避できないような場合には、衝突によって生じる被害者の数が最も少なくなるような回避措置を取るように設計されていた。Xが本件自動走行車を一般道の一方通行の道で制限時速に従って走行中に、前方から対向車が突っ込んできたため、本件自動走行車は左か右にハンドルを切らざるを得なくなったが、左側の歩道には歩行者Aが、右側の歩道には立ち止まって話をしている数人のグループがいたため、本件自動走行車は左側にハンドルを切り、対向車との衝突は免れたが、Aを轢過して死亡させた。」(51)

自動運転車における「緊急プログラムの適切性」として、いわゆる「トロリー問題」(52)が問題となる。すなわち、自動運転車によって2人の歩行者が死亡する危険が迫ったときに、それを回避しようと走行方向を変えれば、今度は1人の別の歩行者が死亡する（他の選択肢はない）というシナリオに対して、どのようなプログラムを組むべきか、あるいは、「運転者」に対応が任された場合、運転者はどのように行動するべきかという問題である(53)。日本においては走行方向を変えた場合についても緊急避難による正当化を認める見解(54)が有力であるが、正当化的緊急避難における衡量基準において「本質的優越」が要求され、生命の数による衡量を認めない見解が有力なドイツの議論において、例えばリアーネ・ヴェルナー（Liane Wörner）(55)やアルミン・エングレンダー（Armin Engländer）(56)など

(51) 深町晋也『緊急避難の理論とアクチュアリティ』（弘文堂・2018年）244頁以下。深町は日本においては「生命法益のディレンマ状況は既に緊急避難の次元で解決が可能となる」（同書253頁）とし、この【事例2】の解決に当たっては、「プログラミング段階で従うべきルールとしても被害の最小化を基準とすることにな」り、「こうしたルールに従っている限りでその注意義務が否定されることになる」とされ、刑法37条の適用以前にY（に属する設計者）につき」業務上過失致死罪の構成要件該当性が否定されるとする（同書255頁）。

(52) 笠木雅史「自動運転の応用倫理学の現状と課題：自動運転車とトロリー問題」日本ロボット学会誌39巻1号（2021年）22頁以下参照。

(53) 松宮・前掲注(11)15頁以下。

(54) 例えば佐伯仁志『刑法総論の考え方・楽しみ方』（有斐閣・2013年）185頁以下。

(55) Liane *Wörner,* Der Weichensteller 4.0: Zur strafrechtlichen Verantwortlichkeit des Programmierers im Notstand für Vorgaben an autonome Fahrzeuge, ZIS 2019, 41 ff.；リアーネ・ヴェルナー（田村翔訳）「転轍手4.0：自動化された運転システムのプログラマーに関する実体刑法上の責任」龍谷法学53巻1号［2020年］373頁以下。

(56) アルミン・エングレンダー（田村翔訳）「ジレンマ状況における自動走行車－トロリー問題4.0－」Nomos43巻（2018）117頁以下; 冨川雅満「アルミン・エングレーダー『自

は、このようなプログラミングに対して批判的である。この問題に関して、松宮は、①「日本の刑法 37 条による緊急避難も、全面的な違法性阻却をもたらすものであるかどうか、未だに決着はついていない」こと、②「AI に搭載すべきプログラムについて厄介なのは、プログラムは事前に事態を予想して作られるものであるため、緊急事態に直面した人間の「適法行為の期待可能性の不存在」を理由とする免責的緊急避難の法理は使えないこと」、③「他の歩行者を犠牲にするよりも、AV が自ら車道外に飛び出すことによって生じる乗員の死傷結果の方が損害が小さい場合に、AV にこのような『自己犠牲プログラム』を搭載することが法的に要求されるか、また、それは現実的に可能かという問題がある」(57) こと、④事前のルール化を念頭に置き、このような場合に「危険共同体」の中にいる者またはテロリストにハイジャックされた航空機の撃墜のような例外的事態においては市民一般に「犠牲になる義務」があるかどうかという問題を指摘した上で、「いずれにせよ、ここでは、規範自体をめぐって人間社会の中でも意見の一致が見られないことが問題であ」り、「社会の規範自体の深化と発展が求められている」(58) とする。以下で述べるようにドイツにおいては上述の責任主体の問題とともにこのジレンマ問題に関しても一定の方向性を示す立法がなされたことは注目に値しよう。

動運転自動車とジレンマ状況の克服』」千葉大学法学論集第 32 巻 1・2 号［2017 年］157 頁以下。

(57) 松宮・前掲注 (11)16 頁以下。これに関して米国では、AV の自損による問題解決は考慮されない傾向にあり、高額な自動化運転車両を購入した者の利己的な判断（「衝突最適化アルゴリズム」）が前提とされることが多く、また、現実問題としては、自損することで乗員が死傷するようにプログラミングされた AV を購入する消費者は少ないであろうという理由から、「衝突最適化アルゴリズム」の採用を検討せざるを得ない状況となっているとされる（国際交通安全学会報告書［プロジェクトリーダー：今井猛嘉］『自動車の自動化運転：その許容性を巡る学際的研究』http://www.iatss.or.jp/common/pdf/research/h2762.pdf［2016 年］12 頁以下参照）。これに対して松宮は「緊急避難の原理からみれば、『自己犠牲』が社会全体からみても最も損害の少ない避難方法であるなら、危難をそれ以外の他者に転嫁することは許されないであろう」とし、「ここでは、自動運転車の販売戦略と『緊急プログラムの適切性』問題とが矛盾し合う」としている（松宮・前掲注 (11)16 頁以下）のである。ドイツにおける議論については、深町・前掲注 (55)250 頁以下；今井猛嘉「自動車運転の安全性を担保する法制度―ジレンマ状況（トロリー問題）への対応―」法学志林 118 巻 4 号（2021 年）123 頁以下など参照。

(58) 松宮・前掲注 (11)16 頁以下。

6　立法による解決？（ドイツ道交法の新規定）

ドイツにおいては世界に先駆けてレベル4以上の自動運転を許容する画期的な道交法及び強制保険法の改正法－自動運転法（Gesetz zur Änderung des Straßenverkehrsgesetzes und des Pflichtversicherungsgesetzes － Gesetz zum autonomen Fahren vom 12. Juli 2021,BGBl. I S. 3108）が成立し、大きな注目を集めている(59)。上述の議論との関係において、今回のドイツにおける改正道交法に注目すべき点は、①責任主体として、製造者及び所持者の他に交通事業者による遠隔操作が想定されるレベル4以上では、航空管制官や、鉄道交通における列車運行管理者と類似する技術監督者(60)を置くことを義務づけた上で、それらの者の義務を条文上（道交法新1f条）規定したことと、②自動運転システムの開発者にとって危険回避プログラムなどを作成する場合の倫理的な基準も明記されたことである。①に関してはこれらの義務違反と刑法上の答責との関係が問題となるが、このような規定においてもなお「答責の間隙」が生じうるのかについて詳細な検討が必要となろう。②については今回の新規定と上記のジレンマ問題の論点についてどのような態度決定をしているについて具体的な規定との関係を見ておこう。

この問題に関連した規定は、新1e条2項であり、そこではまず「自律運転機能を備えた車両には、以下のことが可能になるような技術装置が装備されていなければならない」とされ、1号から10号までにそれらの技術装置が規定されているが、そのうちの特に2号と3号が、この問題にとって重要であると考えられる。すなわち新1e条2項2号においては「自ら車両の操作に関する交通規制を遵守し（selbstständig den

(59)「完全自動運転へドイツが法改正：人命をてんびんにかけず」日経新聞2021年8月11日（https://www.nikkei.com/article/DGXZQOUC063KJ0W1A800C2000000/）。なお令和2年度警察庁委託調査研究『自動運転の実現に向けた調査研究報告書』（令和3年3月）（https://www.npa.go.jp/bureau/traffic/council/jidounten/R02nendo/R02report.pdf）も参照。

(60) ドイツ道交法新1d条3項によれば「本法の意味における自動運転機能を持つ動力車両の技術的監督者（Technische Aufsicht eines Kraftfahrzeugs mit autonomer Fahrfunktion）とは、1e条2項8号により走行中に当該動力車両を作動停止させ（deaktivieren）、1e条2項4号及び3項により運行操作を起動させる（Fahrmanöver freigeben）ことのできる自然人（natürliche Person）である」とされる。Vgl. Sophie *Gatzke,* Gesetz zum autonomen Fahren － Ist die externe Überwachung autonomer Fahrsysteme mit dem Wiener Übereinkommen über den Straßenverkehr vereinbar? NZV 2021, 402.

an die Fahrzeugführung gerichteten Verkehrsvorschriften zu entsprechen)、かつ a）損傷を回避及び軽減できるように設計され（auf Schadensvermeidung und Schadensreduzierung ausgelegt ist)、b）様々な法益への損害が避けられない場合は、人命保護を最優先しながら、各々の法益の重要性を考慮し（bei einer unvermeidbaren alternativen Schädigung unterschiedlicher Rechtsgüter die Bedeutung der Rechtsgüter berücksichtigt, wobei der Schutz menschlichen Lebens die höchste Priorität besitzt,)、かつ c）人命へのリスクが避けられない場合は、個人的な特徴によってさらなる人命の重みづけを行わない（für den Fall einer unvermeidbaren alternativen Gefährdung von Menschenleben keine weitere Gewichtung anhand persönlicher Merkmale vorsieht）事故回避システムを備えている（über ein System der Unfallvermeidung verfügt)」技術装置が、そして同 3 号では「道路交通法に違反しないと走行を続けることができない場合には、自ら車両を最小リスク状態にする」技術装置が必要だとされているのである。ここでは人命へのリスクの最小化が求められる一方で、「人命に対して避けられない危険が生じた場合には、個人的な特徴に基づいてさらなる重みづけをしない」という制限を加えている。この規定は、すでに 2016 年に制定されたドイツ倫理規則の規定を取り入れたものであるが(61)、【事例 2】におけるような人命の数による衡量を許容し、走行方向を変えることによる少数の歩行者の殺害を認めるプログラミングを許容するものであるかどうかについては、なお議論の余地があろう。人命の数による衡量が一切認められないなら、1 人を死なせることの方が、2 人を死なせることよりも損害が軽いとはいえないはずだからである。いずれにせよこの条文がどのような形で解釈され、運用されることになるか今後のドイツにおける実務及び議論の状況が注目されるのである。

7　おわりに

以上で、私は、ロボット・AI を処罰することができる（あるいは処罰すべき）か

(61) 樋笠尭士「AI の自動運転とドイツ倫理規則：倫理ガイドライン策定に向けて」罪と罰 57 巻 3 号（2020 年）73 頁以下；同「AI と自動運転車に関する刑法上の諸問題：ドイツ倫理規則と許された危険の法理」嘉悦大学研究論集 62 巻 2 号（2020 年）21 頁以下参照。

という問題について日本の刑法学者による議論を①消極説、②積極説及び③中間説の三つに整理した。私自身は③の見解をとるが、その要件としては、規範的コミュニケーション能力を持つことが必要であり、現状においてはロボット・AIは、自動運転車を含め、そのような能力を持つには至っておらず、自動運転車自身を処罰することはできないと考える。将来的にはロボット・AI・自動運転車に意識・感情・痛みを感じる能力などを持たせて一人の人格として扱い、自然人と同じように処罰できる存在にすべきなのであろうか？　最近、ロボット・AIが意識や感情などを持つようになる可能性を否定しないアメリカの哲学者ダニエル・デネット（Daniel Dennett）は、最近の論文において、「われわれが必要とするのは意識を持ったAIではなく、知性を持った道具である」とする(62)。NHKの番組(63)のインタビューの中でも「意識を持つ超知能AIを作るのは誤った目標である」と答えている(64)。われわれは、いかなる段階のロボット・AIを処罰できるかという問題を論じるだけはなく、そのような段階に至るロボット・AIを作り出すべきかどうかという問題をも議論する必要があろう。

（明治大学法学部教授）

(62) Daniel C. *Dennett,* What can we do?: We don't need artificial conscious agents. We need intelligent tools, in: John Brockman (Ed.), Possible Minds: Twenty-Five Ways of Looking at AI, 2019, at 41.

(63) NHK「超AI入門特別編：世界の知性が語るパラダイム転換：第三夜『AIが人間をあざむく時』」2019年7月14日（日）放送（なおこのインタビューの内容は丸山 俊一＋NHK取材班『AI以後：変貌するテクノロジーの危機と希望』［NHK出版新書・2019年］に所収）。

(64) 南アフリカの哲学者デイヴィッド・ベネター（David Benatar）の唱える「反出生主義（Antinatalism）」（その基本思想については Benatar, David, Better Never to Have Been: The Harm of Coming Into Existence, Oxford University Press 2006［邦訳：小島和男・田村宜義訳『生まれてこない方が良かった：存在してしまうことの害悪』すずさわ書店・2017年］参照）に関しては森岡正博「デイヴィッド・ベネターの誕生害悪論はどこで間違えたか：生命の哲学の構築に向けて(12)」現代生命哲学研究10号(2021)1頁以下（関連文献についても同38頁参照）にまとめられているように様々な批判があるが、少なくとも「意識」や「苦痛」を持つAI・ロボットについては反出生主義が妥当するように思える。