AIの責任主体性を巡る我が国における議論状況

根　津　洸　希

Ⅰ．はじめに

前号（法政理論55巻2号）では「AI責任肯定論の動向」と称して、諸外国におけるAIの法的責任主体性を巡る議論を紹介した[1]。同拙稿では概

1　拙稿「AI責任肯定論の動向」法政理論55巻2号28頁以下

ね、AI責任肯定論も一定の説得力を有しており、その名称ほど奇異な見解ではないことが確認された。他方、AI責任肯定論は基礎理論として、人間の責任を検討する際の補助線となり、興味深い点は多々あるものの、「AIに責任を認めることがどのような（刑）法理論的な影響をもたらすか」、「AIの責任はいかなる事例解決において重要か」といった理論的実益が論者から提示されていないという点に課題が認められた。

本稿では、前号と同様の問題関心ながら、我が国においてはいかなる議論がなされているかを紹介し、若干の検討を加えることとする。

Ⅱ．我が国における議論状況

1．小林史明の見解

AIの法的責任主体性につき、小林史明は法哲学における権利の本性論からアプローチを試みている。すなわち、権利／義務を負うとはいかなる意味があり、そこから権利／義務の主体にはいかなる性質が必要となるかを、いわば逆算しようというアプローチである。

> *権利や義務をもつとはいったいいかなる意味なのかを遡って検討する必要がある。これについては伝統的に権利の本性論と呼ばれる一群の議論がなされてきたのでこれを参照しよう。権利とは何かという問いへの回答は大きく三説に岐れている。他者への義務を強行・免除する支配力として捉える意思説と、その一つの発展形である、他者の義務を強行するか免除するかを選択できることが不可欠な要素であるとするH.L.A.ハートによる選択説（便宜上この二説を一説として扱う）、そして法によって保護されるべく意図された何らかの利益こそがその要素であるとする利益説とが主張されてきた。これに対し意思説ないし選択説では不可譲権（inalienable rights）を説明できず、利益説では第三者の*

ためにする契約や他者の義務付けによって本人が不利益を被る場合を説明できないなどの欠点が指摘されてきた。しかし、ここで注目しなければならないのは、両説がともに含意する主体の要件である。意思説・選択説が、文字通り相手方に何かをなさしめ、またはなさしめない意思や選択をする能力を前提にしていることから、それらの能力を欠く嬰児や重度精神障害者、脳死者などを権利主体として位置づけることができないという難点はつとに指摘されてきた。他方で利益説はその点を克服し、上記のような能力を欠く人間のみならず受動的に利益を受けることができる動物等にも権利主体の枠を広げることができるとされてきた。たしかに利益説のいう利益は法が保護しようと意図した利益であるから、目的外の単なる反射的利益の受益者は権利主体とはならないが、たとい単なる反射的な受益者も権利主体になると考えたとしても、森村進が指摘するように、利益説は「効用、快苦（pleasure/pain）、安楽といったものを重視」するから、少なくとも効用や快苦等を感じる能力が要求されるのではないか。そのような快苦享受ができない人間が権利主体から排除されることも当然ありうるはずである。したがって、意思説・選択説または利益説のいずれに立ったとしても、一部の生物学的人間が権利主体から排除されるという通常我々が「難点」と考える問題は残置されることになる。それでもなお人間であることのみを理由として権利主体たらしめるべしと主張するとなると、いわゆる人間という動物種のみを優遇する「種差別（spieciesism）」との批判を受けることになろう。この種の批判は、権利主体性の根拠を何らかの能力に基礎づける議論を採るかぎり、人間のうちもっともその能力をもたない者よりも高い能力をもつ人間以外の種が存在するだろうから多かれ少なかれ避けがたいものである[2]。

2　小林史明「権利主体性の根拠をAI・ロボットから問い直す」『市民的自由のための市民的熟議と刑事法 増田豊先生古稀祝賀論文集』147頁以下

以上の検討から、権利の本性論においてなされている議論のどの学説に立ったとしても、一定程度の能力を前提とする部分があるという。その能力を有さない主体は法的権利主体とはならないこととなり、この弱点を補うために「人間であること」という基準を用いれば、そこには何らの合理的な説明もない「種差別」に過ぎないこととなるという。

これに対する回答の一つは、能力による基礎づけをやめてしまうことである。「能力Xがあるのだから権利Yが認められる」という図式を拋棄し、「権利Yが認められるのだから能力Xがあるはずだ」というプラグマティックな図式への移行がそれである。実は、私たちは人間間での能力の差異についてはすでにこのような図式を用いている。さまざまな活動に用いられる個々人の判断能力は実際にはまちまちであるが、一定の閾値はあるにせよ、基本的には同程度と看做す擬制（fiction）によって法制度は運用されている。実際はどうであれ「等しい存在＝人格である」という制度による信憑が、権利主体性の平等を支えているのである。ところが私たちは異種間の場合には、きわめて短絡的に、異なる取り扱いはそもそも能力が異なるからだと信じ、人間間では多少慎重になるもののやはり異なる能力がその背景にあると信じている。だが、実際には制度によって等しく扱われることが等しさを支えるという側面がある（…）[3]。

この「等しい存在」であるとの信憑を獲得するための一つの要素として、責任や義務を負うことであると小林は指摘する。そしてAIが刑事上の責任や義務を負うことの問題点につき、次のように論じる。

やや困難であるように思われるのは刑事責任の負担である。大屋雄裕

3　小林・前掲注2）149頁以下

は、ある種の責任という言葉に我々が担わせている意味について「血のバランスシート」という表現を用いて、我々の衡平感覚を指摘する。つまり「抗争の一方当事者に生じた犠牲と同じだけの生命が他方当事者からも失われねばならぬ」というものである。するとAIやロボットは「人間がすべて等しくかけがえのない生命を持っており、痛みや苦しみを感じうる主体であるという可傷性（vulnerability）」に欠け、複製可能であるために「かけがえのなさが我々とは共有されていない、均質性が存在していないという感覚」を抱かしめるので、刑事処罰に付することに納得し難い違和感があるというのである。

ここで考えるべきは、人間であっても実際には刑罰が痛みとして効かない者がいるなかで、同じ人間であるからという理由で効いているようにみえることで、私たちは均質性を擬制している点である。そうであるとすれば、AIやロボットに刑事処罰が効いているようにみえるようにすることが重要であろう。ロボット法研究においては、ロボットに刑罰を科すか否か、科すとしてどのような刑罰を科すべきかが議論されている。破壊や消去による死刑、停止による自由刑の執行などが検討されるが、いずれも刑罰の目的論に踏み込まざるを得ず明確な解答は出ていない。また、法人処罰のアナロジーによって財産刑を科すことも検討されている。プログラム不正や学習データの偏り等によって被害が生じた場合には、危険性除去のためにデバッグやリプログラミング、学習データの初期化等がおこなわれるであろうが、これらは人間にたとえれば洗脳教化刑に当たるかもしれず、かえって均質性を損なう可能性もある。さらにいえば、人間への洗脳教化刑への端緒を切り開く恐れもある。また、AIがネットワークの一部として結合することで個体化できないことが、かけがえのなさと両立せず、「可傷性を背景にした我々の責任実践の一部とは概念的に相容れない性質を帯びている」との指摘もなされているが、AIを意図的に個体化させ（られたように見せ）たり、他の

方法で均質性を補ったりすることも考えられる[4]*。*

すなわち小林は、刑事責任において重要だとされている可傷性は擬制されている部分があるため、AIであってもその刑罰が苦痛として効いているように見えるのであれば、この均質性を同じくAIにも擬制することはできるとしている。

小林は、人間がAIを「等しい存在」と感じることができるのであれば、権利主体性や責任を観念することはでき、その均質性も可傷性を擬制したり、見た目を似せることによって達成することは可能であるとする。「AIは人間と異なるから責任はない」というのは種差別に過ぎないとしている[5]。

２．今井猛嘉の見解

上記の見解は、法哲学的な背景からAIの主体性の基礎付けとの関係でAIの責任や処罰可能性を論じていた。これに対し、より刑法解釈学的アプローチによって、すなわち犯罪論体系にあてはめることによってAIの責任を検討するのは今井猛嘉教授である。今井教授は、犯罪論においてAIが問題となるのは、行為、責任能力、適切な刑事制裁を想定できるかであるとする。

先ず、ＡＶ【自動運転自動車 autonomous Vehicle：引用者注】に行為を想定できるかであるが、これは、不可能ではない。レベル４以上のＡＶで考えると、そのようなＡＶは、自動運転システムにより、自ら収集した情報（道路や交通の状況等）を、事前に設定されたプログラムに従って演算処理し、適切な速度と進路を選択して進行している。その結

4　小林・前掲注2）151頁以下

5　瀧川裕英「ソフィアの権利」自治実務セミナー2018年6月号49頁も同旨

果として、V【被害者Victim:引用者注】に損害（H）ないし法益侵害（その危険を含む）が生じたならば、損害発生に至った過程を、プログラム処理に起因する行為と評価することは可能である。

（中略）

しかし、行為の意義を、このように、（因果的行為論の見地から）法益侵害という結果と、その原因とを結合するだけの概念と捉えるのではなく、刑法が予定する主体（倫理的自己決定が可能な存在）の振る舞いとして実体的に把握する場合には、別の結論に至り得る[6]*。*

以上のように自動運転自動車の自動運転システムAIに行為性を認めることは不可能ではなく、行為論における学説のうち、どの学説に従うか次第であるという。また、目的的行為論に従うと、現状のAIの発展段階に鑑みればAIに行為性を肯定することはいまだ難しいものの、今後のプログラミングや技術発展次第ではその結論も永久不変ではないという。

責任能力は、事理弁識能力と行動制御能力から構成される。これら能力の相互関係を確認すると、刑法の目的は法益の保護にあるから、自己の行為により法益侵害を防止できる能力が、先ず検討されるべきであり、この能力の発動を促す前提として、自己の行為の結果が、社会的に如何なる意味を持つのかに係る理解力が問われることになる。こうして、事理弁識能力は、行動制御能力を発動する契機として意味があるから、前者の能力としては、自己の行為の結果が社会的に許されない程度の害悪（としての法益侵害）を生ぜしめるとの認識を基礎付ける能力があれば足りる。この理解から「自己の行為が社会倫理的に是認されるか否かを知る能力を、事理弁識能力と言う。」との整理を導くことも、不

6　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』524頁以下

可能ではない。しかし、そこで言う社会倫理とは、「共同生活を営まざるを得ない人間にとっての倫理（他人の法益を侵害しないという最低限のルール）」として理解されるべきである[7]*。*

すなわち責任能力における事理弁識能力を、法益侵害を認識するための能力として理解すれば、学習機能を搭載したAIにも責任能力が認められうるという。とはいえこのような帰結は現状のAI技術を前提としたものではなく、将来予測に基づく仮定的判断であるとする。しかし将来、AIに一定の善悪判断プログラムが実装され、AIの情報収集から事故回避の挙動が洗練されていけば、責任能力を肯定することも可能であるという。

AIは、SLC【自己学習能力self learning capabilities：引用者注】を持ちうるが、社会的非難を受容し、自然人と同様の感情（犯罪を選択したことへの反省、自己が非難の対象とされたことへの恥辱感等）を持つことは、（少なくとも当面は）期待できないであろう。即ち、自然人に想定される感情、そうした感情を持つことを前提として構想されてきた、犯罪を選択する際に他の行為を選択しえた可能性、そうした可能性を説明するための概念としての自由意思は、AIには、直ちには想定できない。そこで、応報刑論によっては、AIに刑罰を科すことを正当化することは困難であると思われる。

（中略）

抑止刑論は、合理的な利得計算ができる行為者に対してしか、妥当しない。AIは、SLCを持ちえても、こうした利得計算をし、不利益が大きい場合には同種行為の再現を中止するという判断能力は、（現時点では）持ちえていないし、近い将来、これが可能になるかも、不透明である。そこで、抑止刑論からも、AIに刑罰を科すことは正当化されない。

7　今井・前掲注6）526頁以下

（中略）

しかし、社会復帰論の下で構想される刑罰は、対象者に（自然人と同様の）心理形成機序がない場合でも、想定しうる。AIとの関係では、レベル4以上の走行を可能とするために設定されたアルゴリズムに欠陥があり、それ故に、自動運転技術の限界に至った際の安全な停車措置等が遅れるという現象の発生が、事故後の調査で判明した場合、そうした欠陥部分の修正（アルゴリズムの改良）は、社会復帰論からは、刑罰として整理可能である[8]*。*

AIには自由意思や心理形成機序を有していないために、応報刑による社会的非難や一般予防による抑止効果から刑罰を正当化することはできないが、特別予防による社会復帰措置としてアルゴリズムの修正を行うということは、AIに対する刑罰と観念することも可能であるという。

今井教授は、無論現状のAIが犯罪論体系上の要件をすべて満たし、かつ処罰に値するということまでをも主張するものではないが、将来的にAIが可罰性ある責任主体となる可能性は排除されていないと主張するものと位置付けられよう。

3．大屋雄裕の見解

大屋雄裕教授は、まず、AIの法的主体性につき、法人の成立との関係で、AIに法的人格性を肯定することは可能であるとする。

法人については「法律の規定によらなければ、成立しない」（民法 条1項）が、逆に言えば法律の規定を作れば新たに生み出すことができる。

8　今井猛嘉「自動運転、AIと刑法：その素描」高橋則夫ほか『日高義博先生古稀祝賀論文集 上巻』364頁以下

法人にはその構成員たる「人」（通常の会社であればいわゆる株主）が必要だが、法人が他の法人の構成員になることはすでにできるのだから、たとえば個々の自動運転車を製造者・サービス提供者である法人（…）の出資により設立された法人と構成し、その故意・過失を法律上も想定できるようにすれば、（…）責任分配の問題と考えることが可能になるだろう[9]。

しかし小林が指摘していたように、大屋教授は刑事責任においてはAIの可傷性のなさから、AIに責任を認めることには違和感があるとする。大屋教授によれば、刑事的責任実践が加害者側によっても負担や苦痛が等しく担われることであるのであるとすれば、可傷性や「かけがえのなさ」を持たないAIによって責任が果たされることはないという。

あるいはそれ【責任:引用者注】を適切に表現する言葉のひとつが「血のバランスシート」であるかもしれない。それ自体は現代日本のヤクザ社会における紛争処理の原理――抗争の一方当事者に生じた犠牲と同じだけの生命が他方当事者からも失わなければならぬ――を表現したものだが、このようなある種の衡平（equity）への意識が、例えば中世における紛争解決慣行としての解死人制度にも共通してみられることは、歴史学者・清水克行が指摘している。加害者側の集団から贖罪のため被害者側に差し出された「解死人」が殺害される（あるいは贖罪の意が示されたこと自体で満足して解放される）ことによって均衡が回復され、報復の連鎖が断ち切られることになるというのである。

ここで重要な鍵となっているのが、制度全体を構成する我ら人間がすべて等しくかけがえのない生命を持っており、痛みや苦しみを感じうる

9　大屋雄裕「人格と責任――ヒトならざる人の問うもの」福田雅樹・林秀弥・成原慧（編）『AIがつなげる社会：AIネットワーク時代の法・政策』353頁

主体であるという可傷性（vulnerability）への意識だと言うことは、おそらく許されるだろう。そこに存在するような痛み・苦しみを前提として、それが加害者（集団）にも等しく担われることが責任の実践なのだと考えるのならば、そのような可傷性を持たず、また本質的に複製可能であってかけがえのなさを持たない（と我ら人類が想定する）AIやロボットによって責任が果たされることはありえないということになるかもしれない。我々がロボットやAIに対する刑事処罰という観念に納得しがたい違和感を覚えるとすればそこにあるのは、そのようなかけがえのなさが我々とは共有されていない、均質性が存在していないという感覚なのだと思われる。

そしてこの問題もまた、AI同士がネットワークの一部として結合していくことにより、より深刻になるだろう。そもそもかけがえのなさとは、他の存在とは区別される個（individual）＝分割不能（in-dividual）な単位として代替不能だからこそ成立する性質だと考えることができる。だがそのように個別性・独自性を持った存在としてではなく、たとえばすべての自動運転車が構成する全体ネットワークとして、近隣を走行する車両群としてといったようにさまざまな集合を有機的・弾力的に作り出し、その集合にとっての最適解を実現するための一部分として振る舞うことが可能になるからこそ、AIネットワークは個々のAIの集団を超えた利便性を社会にもたらしうると予想されるのであった。このようにネットワーク化するAIの本質は、そもそもかけがえ＝互換可能性を全面的に実現しようとか、他者と自己を区別する境界を消し去ろうとする点にあるように思われる。だとすればそれは、可傷性を背景にした我々の責任実践の一部とは概念的に相容れない性質を帯びているということになるのかもしれない[10]。【傍点原文ママ】

10　大屋・前掲注9）358–359頁

AIには個人のかけがえのなさは認められず、むしろ互換可能性を実現する技術であるから、現行刑法における責任実践には沿わないという。

大屋教授はAIが苦痛を感じることができないという点、また個別性・独自性がないという点、この二点から少なくとも現行の責任実践における意味での責任をAIが果たすことはないであろうと主張する[11]。

4．川口浩一の見解

川口浩一教授は刑罰目的を規範の安定化、すなわち積極的一般予防論に求めるMarkwalder/Simmlerの見解[12]を検討素材としつつ、この見解を英米圏での自由意思論[13]の論者であるFischer/Ravizzaの見解[14]と類似した見解と位置付けた上で、大枠ではMarkwalder/Simmlerの見解に賛同する。

…ドイツ語圏における議論においても、このようなFischer/Ravizza

11　しかしこれはあくまで「現行の責任実践で」という留保がつくことには注意を要する。大屋雄裕「AIのいる社会に向けて」自治実務セミナー2019年1月号3頁は、責任という制度を担う存在が我ら人間だけではなくなることを前提として制度の根幹から再構築することが必要になる可能性につき指摘している。

12　Monika Simmler=Nora Markwalder, Roboter in Verantwortung? – Zur Neuauflage der Debatte um den funktionalen Schuldbegriff, ZStW 129(1), S.20 ff.

13　「意思／意志」という表記につき、法律学では「意思」という漢字があてられ、哲学では「意志」という漢字があてられるという大まかな傾向はあるが、本稿では基本的に（引用部分以外では）「意思」という表記を用いることとする。しかし意味する内容は「自由意思」「自由意志」ともに同じであるとご理解いただきたい。

14　Fischer/Ravizzaの「誘導的コントロール説」の詳細については、拙稿「AIの責任と決定論問題」石井徹哉編『AI・ロボットと刑法』（脱稿済・印刷中）を参照。

説と類似した理論が見られることが注目される。例えばJakobsは、刑法にとって自由意志の問題は無関係であり責任を認めるためには自律、すなわち「自己管理の自由」が認められれば良いとし、Simmler/Markwalderはそのような自由を社会的意味における自由意志と解して、それがロボットに認められるようになる可能性は排除されないとし、ロボットの潜在的な刑法的答責性の前提として、刑罰目的論との関係を議論している。すなわちSimmler/Markwalderによれば、刑罰目的を規範的期待の安定化（Luhmann）と捉え、「社会的アイデンティティの維持（Jakobs）」、「動機形成（Kargl）」、「規範忠誠・規範承認」の習得（Jakobs）に焦点を当てた積極的一般予防論を擁護されるべきであり、その基礎には社会的帰属論があり、伝統的責任・応報刑論とは異なり、刑事責任は絶対的・定言的なものではない（責任の「非神聖性・可変性」）とされるのである。そして、そこからロボットの人格化の問題が、規範を不安定化させるポテンシャルという刑罰目的論的観点から検討され、そこでは（認知的に）次回に学びなおすのではなく、規範的対応がされるかどうかが重要であるとされる。すなわちSimmler/Markwalderは、ロボット答責性の核心問題は「ロボットがその帰属された権能と人格性に基づいて規範を不安定化させ、それなしには、さしあたり規範が不安定化され、長期的に見れば、規範が消滅してしまうようなコンフリクト、それゆえ刑法的なリアクションがそれに対して必要となるコンフリクトを生み出すかどうか」ということだとし、伝統的見解のような個人的基準ではなく客観的基準によって判断されるという立場によれば「システムにおける役割の違反」（Jakobs）が肯定され、またロボットに搭載された人工知能がモラル・エージェントたりうるかという問題についても肯定し得るとする。そしてロボットの受刑能力の問題については、財産刑の賦課についていえばe-パーソン構想が参考になるし、さらなる自由刑類似の刑罰や処分、さらには「再プログラミング」または自己学習に作用する「害悪賦課」などのロボットに対する特有の刑罰も考えられるとされ

る。そしてロボットの刑事責任問題に対応する3つの選択肢として、①伝統的責任概念・自己決定的人間の理念による場合、②機能的責任概念・刑罰目的による場合、そして③伝統的責任概念の完全放棄（Hörnleなど）による対応が考えられるが、①によればロボットの答責性の完全否定へと導くものであり、また逆に、③の立場は、刑罰と処分の区別の解消など最も重大な帰結をもたらすものであることから、②による対応が最も現実的であるとしている。私見もSimmler/Markwalderが依拠するJakobsやその弟子のPawlikが示した方向性を支持するものであり、そのような「機能的」責任・刑罰理解からは上述のFischer/Ravizzaなどの両立論と同様、ロボットの責任ないし答責性は排除されない[15]。

川口教授は上記のようないわゆる機能的責任の理解に立ち、「人間の意思は自由であるか」ではなく、「人間の意思は自由であると（社会的に）扱われるか」こそが重要であるとする。したがって、仮に脳科学の発展によって自由意思の存在が自然科学的に否定されたとしても、なお刑事責任を問う営みは否定されない。その意味では、（決定論と責任実践の）両立論に分類されよう。

川口教授は上記の理解に基づいて、刑事責任は自由意思の存在を前提としないから、AIがプログラムに従っているがゆえにその意思が決定されているとしても、その一事をもってAIの責任は否定されないと主張する。したがって潜在的にはAIにも責任を認める余地はあるとされる。その上で、AIに責任が認められる条件として、川口教授は次のように論点を設定する。

ロボットの潜在的な人格性・責任を肯定することは排除されないとし

15　川口浩一「ロボットの刑事責任2.0」刑事法ジャーナルNo.57（2018年）6–7頁

ても、より大きな問題点は、それが実際に認められるようになるのはいつか、という点であろう。この問題については、①ロボットが意識を持つことができるか、②ロボット人格に同一性が認められるか、そして③ロボットが規範の妥当を不安定化させ、それゆえ刑法的なリアクションが必要となるようなコンフリクトを生み出すかどうか、という問題が検討されねばならないであろう[16]。

これら3点の論点を設定したうえで、①については、川口教授自身は立場表明をしてはいないが、哲学や脳科学における近時の知見を紹介し、AIも意識を持ちうるとする主張や、いかにしてAIに意識があるか否かを確認することができるかという研究に今後も注視する必要があるとしている。

②の問題については、ジョン・ロックの記憶説を引用しながら、「この点でAIを搭載したロボットにおいては、そもそも意識（または「〈心的ないし指標詞的〉属性」）をもつことができるかどうか問題となるが、少なくともそれを持つ可能性は否定できず、自己の行為に関する記憶による同一性は（法人とは異なり）肯定しうるように思える。」として、AIに意識ないし記憶の連続性から導き出される人格の同一性を肯定する余地はあるとされる。

③の問題については、刑罰理論（特に応報刑論と積極的一般予防論）との関係が重要であると指摘しつつ、刑罰を巡っては「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。」[17]とされる。

16　川口・前掲注15）8頁

17　川口浩一「ロボット・AIに対する刑罰をめぐる最近の議論」法律論叢94巻4・5号117頁。なお同論文は拙稿を筆者自身よりも詳細に紹介くださり、また丹念に検討をいただいており、感謝にたえない。ただ一点付言すると、同論文107頁にて、筆者が予防刑論に依拠したうえでAIの責任を肯定する

川口教授の見解を要約すると、①自由意思が存在するか否かは法的責任とは関係がない（機能的責任論）ゆえにAIにも責任を肯定する余地はあるが、②責任を認めるにはAIに「意識」、「人格の同一性」、「規範を揺るがす前提としての規範的コミュニケーション能力」が必要であり、③現状のAIはいまだこの3つの条件を満たさないために責任はない、ということになる。

Ⅲ．若干の検討

以上、AIの責任主体性を巡る我が国の議論を概観してきた。以下では各説に対し若干の検討を加えたい。

1．小林説について

小林は権利の本性論からAIの権利主体性を問い、権利の本性論にてなされている議論においては少なからず能力に対する要求が潜んでいるという。しかし実際には要求された能力を有さない人間は、動かしがたい事実として存在し、それを「人間だから」という理由で補うのは種差別に過ぎないという。そこで小林は、能力による権利主体性基礎付けを放棄し、「等しい存在である」という擬制による基礎付けの可能性を示唆する。そして、人間がAIを「等しい存在」と感じることができるのであれば、権利主体性や責任を観念することはでき、その均質性も可傷性を擬制したり、見た目を似せたりすることによって達成することは可能であるという。

立場であるかのような印象を与える一文があるが、筆者は予防刑論からAIの責任を肯定するというアプローチをとったことはないため、その点のみ読者におかれては誤解なきようご理解いただきたい。

たしかに、AIに刑罰を科すことがどこか滑稽に思えるのは、寿命や肉体のないAIを刑務所に収容しても苦痛とはならないことも一因であろう。それゆえ、小林は苦痛を感じているように見えることが重要であるとする。しかしこの分析は、AIが人間に対する刑罰と同じ刑罰を科されることをその前提としているように思われる。したがって、小林の主張は、人間の刑罰制度がAIに科されることを有意味にするためには、少なくとも可傷性が擬制される必要がある、と要約可能である。

ここからAIと人間の均質性、つまりAIが人間に「等しい存在」であるという印象を与えることとなり、これがAIの権利主体性を基礎付けるという。しかしそうすると、AIに権利主体性を認める理由は「人間に似ているから」ということにはならないか。一方で、能力による権利主体性の基礎付けにおいて「人間だから」という理由付けは種差別に過ぎないとしつつ、他方でAIの権利主体性を、人間が均質性を感じうるかという点にかからしめるのは、やはり実際には人間中心主義が背景にあるのではないか。小林自身が指摘するように、法における人格は自然人、法人、財団等さまざまなバリエーションがあり、その権利行使の方法・態様もそれぞれ異なるのであるから、自然人と法人、財団等はそれぞれ異なる権利主体性を有すると解することも可能である。そうであればAIの権利主体性も必ずしも人間との類似性を論拠とする必要はなく、AI独自の行為主体性を観念することは可能である。

したがって小林は人間の刑罰制度が可傷性に基づく制度であることを明らかとしたが、AIの権利主体性は人間と同じであることまでをも立証したとは必ずしもいえない。

2．今井説について

上記諸見解が権利の本性論や種差別という出発点からAI責任を論じるアプローチを採用するのに対し、今井教授は刑法の犯罪論体系に引き付

け、各要件の枠内でAIの特殊性により論点となりうる点を論じている。AIの行為性に関してはよって立つ学説次第であるとするが、責任能力については、法益侵害を認識し、回避する能力と解することでAIにも責任能力を肯定する余地を認めている。AIの犯罪能力を肯定した上で、今井教授は特別予防論的観点から、再犯防止のための再プログラミングを刑罰と整理することも可能である旨主張する。

しかし、このアプローチにおいても疑問なしとはいえない。まず、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[18]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。しかし、特別予防論のみで刑罰を正当化できるかという刑罰論上の根本的な問いが立てられることとなろう。応報刑論、一般予防論による刑罰の正当化を諦め、特別予防論だけで刑罰を正当化する場合、そこでいう刑罰はむしろ色彩としては処分に近いのではないだろうか。

今井教授の見解には興味深いものがあるが、刑罰を巡る見解についての論証には少なからず無理があるように思われる。

3．大屋説について

これに対して、大屋教授は「血のバランスシート」というヤクザ社会の慣行こそが、「落とし前」としての責任の要素にとって重要であると説いている。そこでは被害者が被ったのと同じ苦しみを加害者も負担せねばならないが、AIは交換可能な存在であるがゆえに可傷性がなく、AIを罰す

18　AIが学習した膨大な学習データのうち、どのデータがAIの過誤行為の原因になっているかを特定したうえで、そのデータを削除ないし成果としての学習内容を修正することができると仮定すれば、の話ではある。問題は、人間にとって処理不能なほど膨大な情報から学習したAIの過誤行為の原因を突き止められる人間が存在するか、という点である。

ることでは被害者が被った苦しみを贖えないがゆえに、AIには刑事責任がないという。

AIは刑罰による苦痛を感じえないという論拠は従来から主張されているが、交換可能性による「かけがえのなさ」を理由に可傷性がないとする点に大屋教授の見解の特色があるといえよう。

とはいえ、「かけがえのなさ」をいかなる基準として考えるべきであろうか。たとえば、筆者自身が「凡庸な私」について考え、私はかけがえのない存在かと問われると、胸を張って「そうだ」と答えることは正直できない。社会においては筆者のあらゆる面での能力を上回る人間が数多くいることに思い至ると、私の「かけがえのなさ」なるものは建前にすぎず、「私にしかできない仕事」などというものは実際には存在しないであろうし、その意味で私は交換可能である。むしろ、その人が交換不可能であるというほど特定の個人の能力に共同体全体の存立が依存してしまうのは不健全であるから、これはある意味当然であるとはいえる。

これに対して、能力や社会への貢献いかんによって「かけがえのなさ」を決めるのではなく、義務論的に「人間は誰もかけがえのない存在なのだ。」と言い切ってしまうのも手ではあろう。大屋教授もおそらくはこのような意図であったと推察される。しかしそうなると、小林が指摘した種差別の問題が再度問題となってしまう。AIも人間も同じく（事実的には）交換可能であるのに、人間だけがかけがえのない存在であるということになるからである。AIも「傷つきかたが人間と違うだけで、可傷性はあるのだ。」といえる余地も残されている。

また、この可傷性の位置付けについても違和感を覚える。というのも、可傷性の有無によって、苦痛を中核とする応報刑論は影響を受けるが、責任の有無にまで影響を及ぼすことはないように思われるからである。刑罰の苦痛を意に介さない存在、たとえばホームレスが寒い冬を外で暮らすよりも、雨風しのげる刑務所の方がよいと考えたような場合、応報刑論による処罰の正当化の文脈で疑義を生じることはあっても、責任自体は何ら問

題なく肯定されるはずである。「無痛症」であるがゆえに責任が否定されるということはない。

4．川口説について

川口教授はSimmler/Markwalderの理論的基礎となっているJakobsやPawlikの見解としての「機能的応報刑論」を支持したうえで、決定論と責任実践は両立可能であるとする。このように川口教授は責任（主体性）を論じるにあたって、根本問題である自由意思論からのアプローチを採用している。刑罰や責任を巡る学派の争いの時代から、刑法における責任の内容は意思自由論と意思決定論を究極的な対立軸として論じられてきた。近時は脳科学の発展により脳神経学的意思決定論[19]の登場により、法学者にとって自由意思論はいわば「パンドラの箱」となってしまった感がある。しかし責任を巡っては自由意思の存否が問われていたのであるのだから、本来であればAIの責任を論じるにあたっては「AIに自由意思はあるか」という問いが設定され、争点とされるべきであった。しかしながら、AI責任否定論者であっても「AIに自由意思などない」と強硬に主張することにためらいがあったのであろう。というのも、AI責任否定論者も、人間の責任はなお肯定したいであろうところ、あまり強硬にAIの自由意思を否定してしまっては、AIの自由意思を否定する論拠がそのまま人間の責任を否定する論拠としてブーメランのように戻ってきてしまうからである。すなわち「AIに自由意思などない」という主張が、「人間の自由意思も極めて疑わしいではないか」、「人間の自由意思もフィクションに過ぎないではないか」との反論を招き、克服不可能なハードルを自らに課すこととなってしまうのである。

---

19 Benjamin Libet, Do we have a free will?, Journal of Consciousness Studies, 6, No. 8–9, 1999, pp. 47–57.

このような状況にあってもなお、責任の本質問題である自由意思論からアプローチを試み、両立論というかたちで決定論と責任実践を両立させようとするアプローチは、正道的なアプローチであるといえる（無論、その他のアプローチが邪道であることを含意するものでは一切ないことを付言しておく）。

しかしながら、川口教授の見解（ないしその基礎にあるFischer/RavizzaやJakobsの見解）が「両立」させようとしているのは、あくまで決定論と責任実践であって、決定論と自由意思の存在の両立ではないことには、注意が必要である[20]。通常、哲学上の自由意思論を巡る「両立論問題」として提起される問いは、「決定論と自由意思は両立しうるか？ *Is free will compatible or incompatible with determinism?*」[21]、すなわち決定論と自由意思の存在との間での両立可能性である。この両立論問題に、「両立可能である」と回答する場合に、自由意思の余地が肯定され、その帰結としてなお人間に責任が認められるという結論となる。川口教授の見解（ないし機能的責任論）はこれとは異なり、両立を目指しているのは決定論と責任実践[22]であり、自由意思の存在ないし内容そのものには本来的には無関心で

20　なお厳密にいえば、Fischer/ Ravizzaの問題意識は「決定論が真である場合、我々は自身をコントロールすることはできないのか？」という点であり、目指されている「両立」は決定論と「コントロール」である。Fischer/Ravizzaはこのような問題意識のもと、仮に決定論が他行為可能性を否定したとしても、いまだ我々には自身をコントロールする余地はあり、したがって責任実践は決定論と両立しうるとする。Fischer/Ravizzaは決定論を前提としているため、自由意思の存否そのものに無関心という意味では機能的責任論に親和的であるといえるが、決定論と責任実践の直接的な両立ではなく、決定論と（責任実践の前提たる）誘導的コントロールとの両立を志向する点を強調するのであれば、「通常の」両立論に近いといえよう。

21　*Robert Kane*, Introduction: The Contours of Contemporary Free-Will Debates, The Oxford Handbook of FREE WILL second edition, p.5.

22　同様に決定論と自由意思の両立ではなく、決定論と責任実践の両立を志向するものとして、瀧川裕英『責任の意味と制度――負担から応答へ』50–51頁

ある[23]。

それゆえ、川口教授が「③ロボットが規範の妥当を不安定化させ、それゆえ刑法的なリアクションが必要となるようなコンフリクトを生み出すかどうか」という論点を設定したのは上述の理解からすれば極めて論理的である。責任が刑罰目的（積極的一般予防ないし規範維持）との関係から逆算されるのであれば、AIの責任の有無を巡って問われるべきは、AIが規範を動揺させうるか、に尽きる。AIが単なる機械であって、AIに起因する事故は自然災害のようなものであると考えれば、仮に自然災害で人が死傷してもそれによって「生命保護の規範が動揺させられた」と受け止められることがないことと同様に、AIの行為は規範的な意味を持たない。逆にAIの行為に一定の社会的意味が認められるとすれば、AIの行為は規範違反たりえ、したがって（機能的）責任も認められることとなる。

他方、上述の理解によれば、逆にAIのみならずおよそ責任を論じる上では、徹頭徹尾、客観的ないし社会的な観点から、「ある行為が規範を動揺させたか」だけを外部的な観察に基づき検討すれば足り、自由意思などの内心の働きを検討する必要はない[24]。むしろ機能的責任論の理論的長所は、自然科学的な意味での自由意思や内心の実証不可能性を度外視してでも責任概念を擁護することができる点ではなかったか。

そうであるとすれば、川口教授がAIの責任問題につき、「①ロボットが意識を持つことができるか、②ロボット人格に同一性が認められるか」といった、意識や記憶などの、いわばAIないしロボットの「内心」の内容を、

23　Simmler=Markwalder, a.a.O（Fn. 12）, S. 33–34. は繰り返し、責任は社会的構成物であって、必ずしも自然科学的な意味での自由意思の存否の問題とは結び付かないと主張している。

24　なおこの点につき、Urs Kindhäuser, Rechtstreur als Schuldkategorie, ZStW107（1995）, S. 711. は、機能的責任論が「機能的にのみ規範と市民を定義することで、規範とその違反に対する非難に内的な関係性を見出すことを放棄」しているとして批判する。

AIの責任の条件とする必要はなかったように思われる。機能的責任論によれば、我々人間においてでですら、意識を持っているのか、あるいは昨日の「私」と今日の「私」には連続性があるか、といったことは些末な問題であって、外部的・社会的観察に基づきある行為が規範を動揺させうるものと受け止められるか否かが本来的な関心ではないか。

Ⅳ．おわりに

本稿では前号に引き続き、AIは（刑）法的な責任をとりうるかを巡る国内の議論を参照し、若干の検討を加えた。法律学と哲学の両分野にわたる問題領域であるがゆえに、「権利の本性論」からのアプローチ、「刑法における犯罪論体系へのあてはめ」からのアプローチ、「刑罰制度と刑罰的苦痛」からのアプローチ、そして「自由意思論」からのアプローチと、この問題へのアプローチ方法は多岐にわたる。今後も議論の趨勢を追い続ける必要があろう。

紙幅の関係上、また筆者の能力の制約上、我が国における全ての見解を網羅したとは到底言い難い。本稿にて取り上げた見解以外にも、興味深く傾聴に値する見解はいまだ数多く存在する。そのような本稿にて紹介しきれなかった諸見解については、別稿にて改めて紹介の上、検討を加えたいと思う。