博士論文「AI の刑事責任」要旨

中央大学大学院法学研究科
刑事法専攻博士後期課程
根津　洸希

【目次】

第五章　AI に対して「刑罰」を科すことは可能か

Ⅰ. はじめに　―　4 つのテーマの提示

Ⅱ. 「刑罰」を受けるのは誰かという問題

Ⅲ. 「刑罰」はロボットや AI にとって苦痛となりうるかという問題

Ⅳ. 近代刑法における意味での「刑罰」と呼べるかという問題

Ⅴ. ロボットや AI を「処罰」することによる「刑罰」という語の意味変容という問題

Ⅵ. おわりに

終章

【全体の要旨】

自動運転技術に代表される強い AI ないしそれに類する技術が実用段階に達した場合、利便性の向上や経済的利益など、一方では大きな社会的効用が期待される。他方で、そのような技術の利用によって法益侵害結果が惹起されることもありうる。しかし AI の判断にはいわゆるブラックボックスと呼ばれる領域が存在するため、AI の行動を完全に予測し、なぜそのような行動をとったのかを逐一検証することはできない。AI が法益侵害結果を惹起した事例を刑法的に分析する際、そのブラックボックス性によって過失論の精緻化のみによっては解決できない領域が存在し、そこでは責任分配が主たる問題となることを明らかとする(第一章)。

その責任分配を考えるにあたっては、現在いくつかの法律構成や解決策が提案されているが、技術発展という観点や過剰処罰の問題を踏まえれば、AI に刑事責任主体性を肯定することで妥当な結論を導くことができる事例群が存在することを指摘する(第二章)。

AI に責任主体性を肯定するというと奇異な印象を受けるであろうが、責任分配の当事者として AI の人格性・責任主体性を認める見解は国内外で徐々に増えてきている。しかしいまだ AI の責任主体性を説得的なかたちで基礎付けるまでには至っていない。他方、このような AI 責任肯定論に対して AI の責任主体性を否定する見解も根強いが、肯定論に対して有効な反論をしているようには思えない。両者の議論は散発的で噛み合わず、争点や対立軸が定まっていない印象を受ける。これは一つに責任の本質論からのアプローチ、とりわけ決定論問題を意図的に避けてきた結果であると考えられる(第三章)。

「AI には道徳的反省能力がない」であるとか、「AI は規範への応答可能性がない」など、さまざまな AI 責任否定論の論拠はあるものの、その論拠はいずれも、AI の意思決定プロセスには道徳的反省ないし規範による禁止命令が影響しないのだ、つまり「AI には自由意思がないのだ」ということを裏から言っているものと考えられる。そこで、AI 責任否定論を定式化し、「AI はプログラムの集合体にすぎないから決定論が妥当し、したがって AI の責任などありえない」という主張であると想定し、AI 責任擁護論の立場から、この

ような主張に反論することを試みた。AI 責任否定論の前提、すなわち「決定論と責任の両立不可能性」を論駁することで AI の責任が理論的に排除されるべきでないことを示した(第四章)。

AI の責任を肯定することが、ただちに AI の処罰を要請することをも含意するかについては慎重な検討を要する。そこで AI の責任と AI の処罰の関係性において、問題となりうる論点を設定し、若干の検討を試みた(第五章)。

Ⅱ. 各章の要旨

【第一章】

第一章では本研究全体を通じての問題の所在を提示することを試みた。

いわゆる強い AI という技術が実用段階に入った場合に、人間の関与者のコントロールの及ばないところで法益侵害結果が生じうることも容易に想定しうること、ならびにその事故原因は AI のいわゆるブラックボックス性によって困難を極めることを指摘する。

このブラックボックス性という AI 独自の問題性に対して、法的議論においては、製造時の結果予見可能性という文脈で主に過失犯論において検討されてきた。しかし、過失犯論の精緻化のみによって AI の問題は解決できるのか、つまり AI と刑法の問題は過失犯論の一つの応用場面に過ぎないのかを問う。

そのためまず AI のブラックボックス性がいかなる技術的特性から生じるのか、そしてそれは技術的取り組みにおいて解消可能な問題なのかを概観する。ブラックボックス性は、①AI の自律学習機能、②同一 AI が複数利用されることによる累積効果、③膨大な情報量の検証困難性、④ネットワーク化によってもたらされる。ブラックボックス性を解消する技術的取り組みも種々試みられているが、完全に解消することはできない。というのも、そもそも AI は人間が処理できないような情報量を処理するために用いられるのであるから、人間に予測・追体験不可能な領域があることは当然である。

ではこのブラックボックス性は刑法的評価にどのような影響を与えるかが問われるべきである。本稿ではブラックボックス性は「予測困難性」と「検証困難性」という二つの要素から成っていることを指摘する。AI の行動はその複雑なアルゴリズムや自律学習などによって、厳密な事前予測が困難であるし、判断に要した情報量が膨大であるためなぜそのような行動をとったのかを事後検証することもまた困難であるという二面性があるということである。

このうち、予測困難性に関しては過失犯論の精緻化によって一定の対応が可能であると考えられる。なぜなら AI の行動が予測困難であっても、判例の立場を前提とすれば、製造者とりわけプログラマーの過失を問題とする際、過失犯論における結果予見可能性判断においては、因果経過の基本的部分の認識があれば足りるのであって、結果発生の機序を厳密に予測・予見している必要はないからである。したがって多少予見不可能なブラック

ボックス性が因果経過に含まれていたとしても、その一事をもって予見可能性が否定されるわけではありません。

これに対して検証困難性に関しては、過失犯論ではなく責任主体の主体選定にかかわる議論が不可避となる。というのも過失の競合などにより結果原因が検証できない場合には、過失犯論からのアプローチではなくたとえば共同正犯論からのアプローチによって解決が試みられる。これはたとえば、二人の作業員が屋根瓦を無造作に投げ落としていたが、このうちのどちらかが投げ落とした瓦が通行人にあたり、怪我をさせた場合という過失共同正犯の教科書事例も、一種の結果原因が検証困難な場合であるといえる。このような場合、無論共同の注意義務を負っていたかといった過失犯論上の問題も問われるものの、主に問われているのはこの両者に共同正犯を成立させてよいかという責任主体性にかかわる問題である。このように、検証困難性が問題となるような事例においては、過失犯論そのものというよりは責任主体の選定、ならびに責任主体間での責任分配が問われているのである。

AI に話を戻すと、AI が法益侵害結果をもたらしたとき、そしてその結果原因が検証困難であるとき、そこには「答責の間隙」が生じている。したがって AI と刑法の問題は、過失犯論に汲みつくされるものではなく、答責の間隙に対する責任分配を巡る議論をも含むということを明らかにした。

【第二章】

第二章では前章にて論じた「答責の間隙」をいかなる法律構成によって解決することが可能かということを検討した。

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、④AI 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。

①ないし②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。

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。そこで 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の見解を主張する。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。そして、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る。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のである。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るのである。

このような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。

【第三章】

第三章では「AI の責任」という奇異な概念を巡る議論を参照しつつ整理し、AI に責任を肯定するための障害となる問題を析出することを試みた。

AI の責任主体性を巡っては、積極説、消極説、中間説という三極に大別することができ、各論者の主たる論拠も出そろってきた感がある。しかし積極説も消極説も、決定的な論証には至っておらず、議論は膠着状態にあるといえる。

積極説は徐々に支持者を増やしつつあり、様々な論拠から AI の責任を肯定しようとする。しかし、散発的な議論が多く、責任の本質にまでさかのぼって AI の責任を論証するまでには至っていない。

他方、消極説は AI が責任を有しえないとする論拠を種々挙げるものの、その議論は現状の AI を前提としているものが多く、AI は将来的にも責任を有しえないのかという点については明らかとしておらず、こちらも決め手には欠いている。

そこでなぜこのような膠着状態に至ってしまったのかの分析を試みたところ、各論者ごとに想定している AI の技術レベルがまちまちであることも大きいのだが、AI の責任主体性を巡る議論においては責任の基礎理論における争点形成がなされていないという点が議論を阻害しているとの仮説が立てられるにいたった。とりわけ、積極説も消極説も現状の責任の基礎理論を維持したいあまり、ややもすれば人間の責任までをも危機に陥れかねない決定論問題に踏み込むことに躊躇があり、議論が遠回りして錯綜してしまったのではないか。

従来責任の本質論においては、道義的責任論と社会的責任論という対立軸のもと議論されてきたのである。そしてこれは実質的には非決定論と決定論の争いであった。いわゆる固い決定論は責任を否定する論拠であるし、プログラムデータである AI に、固い決定論の主張は妥当しやすいようにも思えるため、消極説が AI の責任を否定したいのであればこの論理を辿るのが筋であるように考えられるのである。

しかし、おそらく消極説は、近時の脳科学の発展により、実は人間にも自由意思はないのではないか、という疑念が振り払えないがゆえに、「AI には自由意思がないから責任が

ない」と言い切ることに躊躇をしているのであろう。ここでもし正面から自由意思論を展開してしまうと、「人間にだって自由意思などないではないか」と切り返されてしまい、AI の責任を否定しようとして人間の責任まで否定してしまうという藪蛇となってしまいかねないからである。そこで Gleß や Joerden のように AI に自由意思の余地は不承不承認めつつも、「評価の自由」や「道徳的自己反省能力」といった付加的な要素で AI の責任を否定するという遠回りをしたのだと考えられる。

消極説は AI の責任を否定するために決定論を持ち出すことによって、勢い余って人間の責任まで否定してしまうことを恐れ、他方で積極説は AI の責任を積極的に基礎付けるために AI に自由意思を認めることによって、非決定論が真であることの挙証責任を負うことを恐れたのであろう。それゆえ両者が本当は言いたいことを我慢せざるをえず、それにより議論が遠回りして錯綜したのではないか。

したがって AI の責任主体性を論じるにあたっては決定論問題が改めて問われねばならないということが明らかとなった。

【第四章】

第四章では、AI の責任を論じる上では決定論問題に踏み込まざるを得ないとの前章の結論から、問題をごく限定して決定論問題に取り組んだ。すなわち本研究の目的は AI の責任を肯定することそのものではなく、利用者や製造者を処罰リスクから救い出すための理論的フィクションとしての AI の責任を仮設することにあるから、AI に責任を肯定することが理論的に全く理由のないことではないということを示すことで十分であるからである。したがって論証されるべきは AI の責任そのものではなく、「AI に責任などありえない」という主張を論駁することで足りる。そこで、「AI の意思は決定されているから、AI に責任はない」という AI 意思決定論に基づく AI 責任否定論に反証すべく、決定論と責任の両立可能性を検討した。決定論を前提としても責任が存在しうるのであれば、AI 責任否定論の理論的基礎を否定できるからである。

「決定論が真ならば、責任はない」という命題は比較的広く共有されている前提であるように思われる。その背後にあるのは、行為時に行為者には意思の自由があったことを理由に、行為者は規範の命令に従い適法行為をするという意思決定ができたということが前提とされ、したがって他行為としての適法行為が可能であったにもかかわらず、あえて犯罪に出たことに対する非難として責任が肯定される、という論理である。決定論は意思の決定性に基づき、他行為可能性を否定するから、他行為可能性を中核におく規範的責任概念とは両立不可能だというのである。

しかし、Strawson は非決定論も決定論も責任とは無関係であるとし、仮に決定論が真であるとしても、それによって人間の情緒的な反応や、それに深く結びついた責任という制度が変容することはないという。また Frankfurt や Fischer=Ravizza は、「A は B に対し、C を殺害するよう恐ろしい方法で脅迫したが、もともと B は C を殺害するつもりであ

って、当初の予定通り C を殺害した」という事例を設定する。この事例では B は脅迫を受けた際に仮に別の意思を有していたのであれば、A の恐ろしい脅迫に屈し C への殺意をその場で生じたであろうが、実際には B 自らの本懐を遂げる形で C を殺害している。そのため B はどのみち C を殺害せざるをえなかったという意味で他行為可能性はないが、A の脅迫は作用していないから通常の責任を肯定して差し支えないという事例である。この例から、Frankfurt と Fischer=Ravizza は責任にとって重要なのは「他行為が可能であった」という事実そのものではなく、「その行為が『その行為者のもの』であった」といえるかであるとする。

仮に他行為可能性が存在しなくとも、なお行為を「自らのもの」としてコントロールしているとみる余地があるとするこれらの見解には、一定の説得力があるといえる。したがって、事実としての他行為可能性を責任の中核に据える理論的必然性はないのであって、そうであれば決定論が事実としての他行為可能性を否定しても、それは必ずしも責任を否定する理由にはならない。行為当時、その他の行為をすることが事実的に不可能だったといえるような場合にも、その一事をもって責任が否定されるわけではないのである。

それゆえ、仮に AI がプログラムに従って行動するのみで他行為の可能性が事実的にないということを前提にしても、それは必ずしも AI の責任を否定する論拠にはならないということが明らかとなった。

【第五章】

第五章では、AI に責任を認めるということがただちに AI の処罰を要請するかという問題に取り組んだ。

AI の処罰を検討する際には、①AI の人格の範囲、②AI に対する刑罰の実効性、③近代刑法における刑罰の意味、④刑罰の意味変容という問題を考慮せねばならない。

AI がネットワーク化された場合、各個体としてのロボットが「一人」の人格なのか、あるいは情報共有がなされた総体が「一人」なのかを決定せねばならない。また現行法は AI に対して有効な刑種を予定していないが、再プログラミングを「刑罰」と呼ぶことが可能であるか、とりわけ思想刑や科学的去勢の問題との関係で検討されねばならない。また「罪を犯した AI」の再プログラミングデータが、「無実の AI」にも適用されるのであれば、これは連帯責任にほかならず近代刑法にいう「刑罰」とは呼べないこととなる。そして、AI を処罰するという決断に至ったとしても、その決断が単なる犯罪への不安から生じているのであれば、刑罰は責任非難としての意義を失い、市民の不安・不満のガス抜きに堕してしまうことになる。

これらの問題があるため、AI に責任が否定されないということからただちに AI の処罰可能性が導き出されるわけではない。技術発展の展開をまって、慎重な態度決定が必要である。

Ⅲ. 本研究の到達点と残された課題

1. 本研究の到達点

以上の検討から、本研究が明らかとしたことは次のとおりである。

第一に、AI が法益侵害結果をもたらすような事故を惹き起こすことがありえ、AI のブラックボックス性によりその結果原因を予測することも検証することも困難である。このような人間のコントロールが及ばない領域で法益侵害結果が生じた場合、製造者や利用者の過剰処罰を限定する枠組みとして AI の責任という概念が有用な事例群もある。

第二に、「AI の責任」という一見奇妙な概念を巡っては既に議論の蓄積があるものの、決定論問題など意図的に避けられてきた問題群を前にして議論は停滞していた。しかし消極説の主張をよく吟味し、決定論問題との関係で検討したところ、AI の責任は理論的に当然に否定されるものではない。

2. 残された課題

本研究によりもうひとつ明らかとなったのは、AI の責任主体性を認めるために障害となっているのは、責任を認めるための一定の能力や属性が AI に欠けていることではなく、いままで避けられてきた責任論における基礎研究における議論の不足であるということである。本研究で検討した決定論問題はその一つにすぎない。

このように責任論おいて意図的に避けられてきた問題はほかにも、たとえば刑罰非難の道徳的価値と法的責任論を巡る問題や、刑罰非難は本当に個別行為のみに向けられているかという問題が挙げられる。

刑罰非難の道徳的価値と法的責任論を巡る問題というのは、刑罰非難は法的責任論がいうように、本当に道徳的に中立かという問題である。AI の責任を論じる文脈で、AI は道徳的な熟慮ができないだとか、あるいは規範に応答できないといったことがいわれるが、現在の通説である法的責任論によれば刑罰非難には道義的非難は含まれていないのであるから、AI が道徳的な熟慮ができるか否か、規範の背後に控える様々な価値に感応できるか否かは本来問題にならない。すなわち法的責任論からすれば重要なのは規範の禁止・命令を認識していたことと、事実的他行為可能性の存在に尽きる。言い換えれば、「法に従い不法に抗する」ときに、その行為の善し悪しについての「評価」や「道徳的反省」といった動機は問わないはずで、単に「処罰されたくない」といった動機であっても構わないはずである。

刑罰非難は本当に個別行為のみに向けられているかという問題は、責任には人格的要素が含まれることはないのかという問いにかかわる。行為のその瞬間に「法に従い不法に抗する」ということが可能であったことのみを理由に責任非難をするということは、常習犯の罪責が重くなる理由を説明できないことを従来から指摘されてきた。通説的立場である個別行為責任論は、この批判に対して真摯に向き合ってこなかった感がある。それゆえ、

AI という人格的連続性が観念されにくい存在であっても、行為のその瞬間の意思決定が自由であったか否かのみが問題となったため、責任を認めざるをえなくなっているのである。

以上のように、責任論における通説的見解は、AI という存在を前にしてその限界を突きつけられていると考える。上述のような未解決の諸問題について一定の解決がなされて初めて、AI の責任の積極的な基礎付けが可能か、またそうすべきか、そして AI 自身の処罰は可能か、またそうすべきかに対する回答が与えられうるのである。