AIの利活用における刑法上の諸問題（4・完）

――利用者と製造者の刑事責任を中心に――

日　原　拓　哉*

目　　次

＊　ひはら・たくや　立命館大学衣笠総合研究機構法政基盤研究センター補助研究員 京都外国語大学非常勤講師

第 3 章　さらなる AI の利活用における刑法上の諸問題
――財産侵害

第 3 節　コンピュータ領域の犯罪――行為客体としての AI

第 3 款　AI 製品に対するデータ変更・コンピュータ破壊

本款では、学習能力を有する AI を搭載した製品に対し、あるハッカーがその AI 製品にハッキングを行い、その内部データを変更したり、消去したりする事例における刑法上の検討を行う。

第 1 項　ドイツ刑法における議論

第 1 目　データ変更罪

刑法303条aによると、データを違法に消去、隠匿、使用不能にする、または変更することが刑罰の対象とする。当該規範の体系的位置や条文の文言に基づくと、器物損壊と類似する[411)]。本罪の保護法益は、データストレージに含まれる情報に対する権限者の処分権である[412)]。本条における「データ」概念は、刑法202条a第 2 項で規定されるように「他人」という書かれざる構成要件要素によって規定される[413)]。このデータの「他人性」とは、データに関する使用、処理、または削除する他者の権利がある場合のことをいうが[414)]、AI の領域ではメールのフィルタリング行為がデータ変更に該当するか否かという、いわゆるインターネット上のメールボック

411）*Günther,* a.a.O. (fn. 176), S. 231.
412）*Fischer,* a.a.O. (fn. 181), § 303a, Rn. 2. m.w.N.
413）MüKo-StGB/*Wieck-Noodt,* § 303a StGB. Rn. 9.
414）*Hilgendorf/Valerius,* a.a.O. (fn. 387), Rn. 588, *Fischer,* a.a.O. (fn. 181), § 303a, Rn. 4 f.; MüKo-StGB/*Wieck-Noodt,* § 303a StGB, Rn. 9 f.

ス問題での議論に倣い、AIのネットワーク化における問題が生じうる。ここではさらに、データの伝送における処分権の有無や、いつそれが確立されたのかを明確にすべきであり、それゆえデータがアドレスされ、AIシステムによって呼び出すことができる場合には、すでに処分権があると考えるべきかという問題が生じるが、少なくともデータ提供前の事前のフィルタリングや処理は、刑法303条aに基づくデータ変更には該当しないと考えられる[415)]。

構成要件的行為として、本条では、データの削除、隠匿、使用不能にすること、及び改変を認めている。その種類は、刑法303条による器物損壊に依拠するもので、内容上一部重なり合うところがある[416)]。データの削除とは、完全かつ再現できない程度に識別できなくすることであり[417)]、データの隠匿とは、データがその権限者から隔離され、権限者による利用が取るに足らないと言えない期間である場合に、データを利用不可能にすることをいう[418)]。またデータの使用不能とは、データが通常に即して利用し得ない場合に、利用可能性が制限されていることをいい[419)]、データの改変は、改変前のデータと異なる状態がもたらされたことをいう[420)]。また予備段階についても、刑法303条a第3項により処罰の対象となる。

刑法303条aを適用しうる状況は、例えば無許可の第三者によるAI製品のデータストレージへのアクセスである。データに対して無権限の者がAI製品をハッキングした後（ドイツ刑法202条a）、上記のような構成要件的行為を遂行した場合、ドイツ刑法303条aの構成要件に該当しうる。ただし、データの削除・隠匿・使用不能化もしくは改変が無権限者の無権限ア

415） Vgl. *Fischer,* a.a.O. (fn. 181), § 303a. Rn. 7.
416） *Fischer,* a.a.O. (fn. 181), § 303a. Rn. 8 ff.
417） MüKo-StGB/*Wieck-Noodt.* § 303a StGB. Rn. 12, m.w.N.
418） *Fischer,* a.a.O. (fn. 181), § 303a, Rn. 10.
419） BT-Drs. 10/4728, S. 36; MüKo-StGB/*Wieck-Noodt,* § 303a StGB, Rn. 9 f. m.w.N.
420） BT-Drs. 10/4728. S. 36.

クセスに伴う行為結果なのか、それとも AI の学習によるものなのかが不明確な場合が考えられる。このように因果関係が不明確な場合、構成要件的結果が発生しているにもかかわらず、その無権限者に対してはドイツ刑法303条aの既遂罪の適用が否定され未遂罪（ドイツ刑法303条a第 3 項）の適用にとどまる。そうすると、当該 AI 製品にハッキングを行った無権限者が AI の学習を引き合いに出して既遂罪の適用を免れようとすることが考えられるだろう。そこで重要となるのが**説明可能な AI** の構想であり、事例におけるデータ変更等が AI の学習によってもたらされたか否かを証明できるように開発を行うことが、この類型の保護法益であるデータ権限者の処分権を保障することになる。もし AI の学習ではなく無権限者によってデータ変更がもたらされたといえるならば、刑法303条aと刑法202条aが成立し、両罪は観念的競合となる[421]。

第 2 目　コンピュータ破壊罪

コンピュータ・システムが破壊されたときに必ず考慮されるのが、刑法303条b所定のコンピュータ破壊である。本罪は、データ処理の適切な機能化に対する運用者または利用者の利益を保護するものとされる[422]。

刑法303条b第 1 項によると、データ処理が破壊されなければならない。このデータ処理とは、技術的に理解されるべきものであり、狭義の伝送、入力、処理など、電子計算処理のあらゆる形態を含むものとされる[423]。データ処理は個別のデータ処理経過のみを記述するという限界づけはなされないが[424]、AI 製品は非常に複雑かつ高度に自動化されたシステムを有することが多いので、通常では個別のプロセスのみが妨害されることはなく、「あるデータ処理」が妨害されることになる。ここで重要となるのは

421）*Fischer,* a.a.O. (fn. 181), § 303a, Rn. 18.

422）保護法益の変化については BT-Drs. 16/3656, S. 13.

423）*Fischer,* a.a.O. (fn.181). § 303a, Rn. 4 f.

424）たとえば、*Lackner/Kühl,* Strafgesetzbuch: StGB, 29. Aufl., C.H.Beck 2018, § 303b. Rn. 2; *Fischer,* a.a.O. (fn. 181), § 303a. Rn. 4.

技術的見地であるため、例えば演算処理が内部でのみ実行され、限定された範囲にしか現れず、そのようなシステムが典型的に従来的装置の外観を呈しているにすぎない場合であっても、その経過はデータ処理に該当する[425)]。データ処理装置とは処理が行われる機能単位のことを意味し、そこにはAI製品も含まれるとしてもよい[426)]。さらにデータ処理経過は、他者にとって本質的に重要でなければならない。2007年改正以降では私的なデータ処理も刑法第303条bによって保護されているが、その本質的な意義はそれぞれのタスクまたは組織がデータ処理の性能に全面的または少なくとも大部分を依存している場合にある[427)]。具体的には、基本的にデータ処理経過の観点から判断され、大別して企業や行政機関における場合と私人における場合とで区別されている[428)]。前者では、例えば、人事管理、生産管理、購入や売却、物流管理、会計、会社の計画及び戦略決定の管理に使われるデータ処理、さらにはEメールやウェブサイト上でのメッセージのやりとりも重要な意味をもつとされる[429)]。後者は、私人の生活形成にとって中心的な機能を示しているか否かが重要であるとされる[430)]。そのため、本質性要件は具体的状況におけるAI製品の用途に応じて決定される必要がある。例えば介護や警備のためのAI製品であれば認められるかもしれないが、純粋に玩具としての役割を果たすにすぎないAI製品には存在しない。付言すると、デュアル・ユースのAI製品の場合はその線引きが困難であり、本質性要件を充足するか否か疑わしい場合には、利用時の重点が決め手となる。少なくとも、個人の生活を日常的に支援するAI搭載の家庭用ロボットや介護に供する介護ロボットの場合

425） *Günther,* a.a.O. (fn. 176), S. 232.
426） *Günther,* a.a.O. (fn. 176), S. 233.
427） *Fischer,* a.a.O. (fn. 181), Rn. 6.
428） SK-StGB, 9. Aufl. 2017, §303b Rn. 11. Spannbrucker, 75.
429） *Lackner/Kühl,* a.a.O. (fn. 424), §303b Rn. 10.
430） 西貝吉晃「コンピュータ・サボタージュ罪 刑法303条b」刑事法ジャーナル71号（2022年）100頁。

は、本質性が認められてもよいと思われる[431]。

構成要件的行為は、刑法303条aによるデータ改変（刑法303条b第１項１号）、不利益をもたらす目的でのデータの入力と伝達（刑法303条b第１項２号）、データ処理設備またはデータ媒体の破壊、損壊、使用不能化、除去または変更（刑法303条b第１項３号）である。2007年以降、刑法第303条b第５項は、刑法第202条cを参照して、コンピュータ破壊の予備も刑罰の対象としている。コンピュータ破壊は、前述のような構成要件的行為が AI 製品に対して遂行された場合に適用される。

第２項　日本刑法における検討

第１目　器物損壊罪の適用可否

想定する事例に関し、日本刑法においては、ドイツ刑法303条aや同条bのようなデータ変更やコンピュータ破壊のような構成要件は存在しないが、器物損壊罪（刑法261条）の適用の余地はあった。その先例としては、東京地裁平成23年７月20日判タ1393号366頁（イカタコウイルス事件）が挙げられる[432]。

その判示において、「損壊」概念については、最判昭和25年４月21日刑集４巻４号655頁を参照しつつ「物質的に物の全部又は一部を害し、あるいは物の本来の効用を失わせる行為をいう……。すなわち、器物損壊には、物自体を物理的に破壊する態様と物が持つ効用を侵害する態様があるが、後者の場合、『損壊』が成立するかどうかは、客体の効用を可罰的な程度に侵害したかどうかによって判断すべきであり、その効用侵害が一時

431）　*Günther,* a.a.O. (fn. 176), S. 233.

432）　本判例の評釈として、園田寿「『イカタコ事件』について」：器物損壊罪における「損壊」の概念〈判例批評〉甲南法務研究８号（2012年）103頁、浅田和茂「ファイル共有ソフト利用者に「イカタコウィルス」を受信・実行させた行為が器物損壊罪に当たるとされた事例」新・判例解説 Watch（法学セミナー増刊）11号（2012年）135頁、森住信人「イカ／タコウィルス事件：ソフトウェアの改変と器物損壊罪の成否〈刑事裁判例批評268〉」刑事法ジャーナル41号（2014年）211頁がある。

的なものではないか、原状回復の難易をも考慮して検討すべきである」とし、原状回復の容易性について判断する場合、利用者のコンピュータに関する知識レベルは様々であるところ、「損壊の成否は飽くまで社会通念に照らして判断すべきであるから、その難易は、パソコンの一般的な利用者を基準に判断すべきである」という規範を示し、その上で裁判所は、ハードディスクには保存されているデータを随時読み出せる機能（読み出し機能）と新たにデータを何度でも書き込める機能（書き込み機能）があることを定義し、「本件ウィルスにより、各被害者のハードディスクは、使用不能となったファイルが保存されていた部分について読み出し機能が害された」「本件ウィルスの実行状態を止めない限り、ファイルを書き込んで保存しておくことは事実上不可能であり、ハードディスクの書き込み機能は害された」と判示した。

しかし、本判例については疑問が呈されることが多い[433]。というのも、器物損壊罪（261条）の客体は、他人の「物」であり、公用文書等毀棄罪（258条）や私用文書毀棄罪（259条）の客体が「文書又は電磁的記録」と規定されていることと比較して、261条に電磁的記録は含まれないからである。本判決は、「物」であるハードディスクが、本件の客体であるとしたが、本件で毀棄されたのは電磁的記録たるファイルであって、ハードディスク本体ではない。ハードディスク自体は、その本来の機能どおりに「読み出し」と「書き込み」を果たすのであり、改変されたのはファイルであると解すべきであろう。たしかに、本条の「損壊」については、物理的に物の全部または一部を害する場合のみならず、物の本来の効用を失わせる場合を含むとするのが本判例において引用されている判例でもあり、通説である（効用侵害説）[434]。これに対し、「損壊」とは、有形的な作用もしく

433） 浅田和茂「判例に見られる罪刑法定主義の危機」立命館法学345・346号（2012年）13頁、園田・前掲（注432）108頁。なお、森住・前掲（注432）216頁。

434） 大塚仁・河上和雄・中山善房・古田佑紀編『大コンメンタール刑法（第3版）第13巻』（青林書院、2018年）807頁（名取）。

は有形力の行使によって、物の全部または一部を物質的に破壊・毀損し、その結果としてその物の効用を害することをいうとする説も主張されている（物質的毀損説）[435]。物質的毀損説は、条文に忠実な解釈であり、本判決のような広い効用侵害説には「類推」の疑いがあるとされる[436]。事実、本判決における被告人は、「被告人の行為は、倫理的に問題のある行為ではあっても、ウイルス作成罪等の立法によって解決すべき問題であって、本件に器物損壊罪を適用することは刑法の類推解釈を認めるものであり、罪刑法定主義上許されない」と主張していたこと[437]も考慮すれば、想定事例の行為者に対する器物損壊罪の適用には疑問を残すことになるので、日本刑法においては、電子計算機損壊等業務妨害罪（刑法234条の2）の適用可否を検討すべきである。

第2目　電子計算機損壊等業務妨害罪の適用可否

本罪の保護法益は、人の社会的活動としての業務遂行の円滑・安全であるとし、その構成要件は、人の業務に使用する電子計算機自体の損壊又はその電子計算機の用に供する電磁的記録の損壊という物理的な加害、人の業務に使用する電子計算機に虚偽の情報又は不正の指令を与えるという論理的な加害、あるいはこれらと同様の結果を惹起するその他の方法により、人の業務に使用する電子計算機に向けられた行為によって当該電子計算機の動作を阻害することである。また、本罪の行為客体は「人の業務に

435）　たとえば、松宮孝明『刑法各論講義〔第5版〕』（2020年）324頁以下。

436）　浅田・前掲（注433）14頁。

437）　同判決では「本年（2011年――筆者注）7月14日から施行された情報処理の高度化等に対処するための刑法等の一部を改正する法律によって新設された不正指令電磁的記録作成罪（刑法168条の2）は、今後、本件のようなコンピュータウイルスを作成する行為にも適用されることになると推測され、法定刑も3年以下の懲役又は50万円以下の罰金と類似している。しかしながら、本件は、ウイルスによって被害者らのハードディスクを損壊したことを問題にしているのであって、ウイルス作成自体を処罰しようとするものではなく、両者は構成要件も保護法益も異なっている。したがって、不正指令電磁的記録作成罪の新設は、器物損壊罪の成否に影響しない」と判示して被告人の主張を退けている。

使用する電子計算機」としてよい[438]。

ここでいう「人の業務に使用する」とは、行為者以外の自然人、法人、法人格なき団体等であって、人が反復継続する意図のもとに行う経済的社会的活動たる業務の主体たる者がその業務に使用していることを意味している[439]。この点において、公務を「業務」に含むかについては見解が分かれるところであるが、すべての公務に使用される電子計算機は本条の客体になると解される[440]。

次に、「電子計算機」概念については定義規定が設けられていない以上、解釈を要するものとなる。しかし、単に「自動的に演算・データ処理を行う電子装置」と定義するのでは、例えばマイクロコンピュータを搭載する家電製品、自動販売機、電卓、電子辞書も本罪の客体と想定されることになる。しかし、本条が電子計算機に向けられた加害を手段とする新たな業務妨害行為をとらえて処罰しようとするのがその立法趣旨であるから、それ自体が自動的に情報処理を行う装置として一定の独立性をもって業務に用いられているもの、すなわちそれ自体が業務を左右するような判断、事務処理、制御等の機能を果たしている電子計算機といえるものに限定されるべきであり、およそ当該機器自体が自動的に情報処理を行う装置とはいえない家電製品や自動販売機は、一定の情報処理は行っているとはいえ、それ自体業務を左右するような判断、事務処理、制御等の機能を果たしていない電卓、電子辞書は本罪の客体としての「電子計算機」には当たらないとされる[441]。

行為についてみると、第一類型である「損壊」とは、電子計算機や電磁的記録を物買的に変更、滅失させ、あるいは電磁的記録の消去などのよう

438） 大塚ほか・前掲（注434）247頁（鶴田＝河村）。
439） 大塚ほか・前掲（注434）247頁（鶴田＝河村）。
440） 大塚ほか・前掲（注434）248頁（鶴田＝河村）。
441） 大谷實『刑法講義各論〔新版第５版〕』（成文堂、2019年）157頁、山口厚『刑法各論〔第２版〕』（有斐閣、2012年）166頁、松宮・前掲（注435）182頁、浅田和茂『刑法各論〔第２版〕』（成文堂、2020年）176頁など。

にその効用を害することとされ、第二類型である「虚偽の情報」とは、当該システムにおいて予定されている事務処理の目的に照らし、その内容が真実に反する情報のことであり、「不正な指令」とは、当該事務処理の場面において、与えられるべきでない指令のこととされる。また、第三類型である「その他の方法」は、電子計算機に向けられた加害手段であって、当該電子計算機の動作に直接影響を及ぼすような性質のものであることを要する。具体的には、電子計算機の電源を切断する、温度・湿度を急激に上下させるなどのような動作環境の破壊、通信回線の切断、入出力装置等の損壊、処理不能データの入力などが挙げられる[442)]。

さらに本罪においては、「電子計算機に使用目的に沿うべき動作をさせず、又は使用目的に反する動作をさせ」るという結果、すなわち動作阻害という結果発生を必要としている。「使用目的に沿うべき動作」とは、電子計算機を設置して業務遂行のために使用する者が、具体的な業務遂行において当該電子計算機を使用して実現しようとしている目的に適合するような動作であって、例えば電子計算機がある条件下で一定の制御を行うという方法で機械制御に使用されている場合、そのような一定条件が与えられたときに行うこととされている制御の動作を意味する[443)]。その場合の「動作」とは、電子計算機の機械としての働きをいい、具体的には、電子計算機の所定の機械制御を実行するため、必要とされている情報処理等のために行う出入力、演算等の働きのことである[444)]。

最後に、「業務妨害」の要件が存在する。これは人が反復継続する意図で行う社会的活動である「業務」[445)]を妨害することであるが、これには本

442） 大塚ほか・前掲（注434）250頁（鶴田＝河村）。

443） 前田雅英編『条解刑法（第 4 版）』（弘文堂、2020年）704頁。

444） 大塚ほか・前掲（注434）250頁（鶴田＝河村）。

445） その定義については、職業その他社会的活動に基づき継続して行う事務または事業をいうものと解される（大判大正10年10月24日刑録27輯643頁）。社会的活動としてなされる継続的な事務・事業は、広く本罪にいうところの「業務」に含まれ、商業・農業・工業等の経済的活動に限定されるわけではない。鎮目ほか・前掲（注403）321頁（鎮目）も参照。

罪は具体的危険犯であり、妨害の結果が現に生ずることまでは要せず、電子計算機に向けられた加害によって、実際にその「使用目的に沿うべき動作」をさせず、あるいは「使用目的に反する動作」をさせるという状態が発生し、これが業務を妨害するおそれのあるものでありさえすればよいとされる[446]。

本罪の構成要件上の問題として、AIが搭載された機器の内部データ変更やその損壊における事例について日本刑法に当てはめると、ドイツ刑法のそれとは異なり、業務に供される電子計算機でなくてはならないので、私用の介護ロボットなど、私的空間に属するAI機器は本罪に該当しない。具体例をあげると、娯楽のために自動運転車（レベル３相当）を走行していたところ、ある者がハッキングにより同車と遠隔通信するサーバに侵入し同車の制御を奪った場合[447]や、介護の用に供する介護用ロボットにおける監視システムがハッキングされたことによりその制御を失い、被介護者の生命・身体に危険が迫る場合である。これらの事例において、本罪の成立は考えられず、これらAI製品の制御を失わせたことに係る不正指令電磁的記録の作出行為のみが可罰的となるだろう。もっとも、医療現場で用いられるものなど、業務性が認められるAI製品についてはこの限りではない。なお、業務を妨害するための手段がもっぱら不正指令電磁的記録によるものである場合、不正指令電磁的記録供用罪との罪数関係が問題となるが、本罪と不正指令電磁的記録供用罪の保護法益とは異なるため、事実関係が一個の行為と認められる場合には観念的競合になると考えられる[448]。

446） 大塚ほか・前掲（注434）251頁（鶴田＝河村）。

447） この場合、自動車道における往来危険罪（道路運送法100条）の適用も考えられる。西貝吉晃「コネクティッドカーシステムに対するサイバー攻撃と犯罪」法律時報91巻４号（2019年）49頁以下参照。なお、この事例における「実行の着手」については同罪が具体的危険犯であることから、静止しているのではなく現に作動状態にある自動運転車にハッカーがその遠隔通信サーバに侵入し、同車の移動・停止を可能にした時点であるという（49頁）。

448） 大塚ほか・前掲（注434）253頁（鶴田＝河村）、杉山・吉田・前掲（注382）90-91頁など。なお、仮にこれらの場合で器物損壊罪も成立するならば、罪数関係としては、行為↗

また、学習を行うAI製品という観点での固有の問題は、前項のドイツ刑法での検討でも言及したように、第三者によってハッキングされたAI製品が、その損壊ないしは虚偽の情報または不正な指令の供与等によりその使用目的に沿うべき操作をせず、またはその目的に反する動作をすることによってある者の業務を妨害した場合に、その損壊もしくは虚偽の情報または不正な指令の供与の原因がハッカーによるものなのか、AIの学習によるものなのかが不明であった場合、つまり妨害に至るまでの過程がブラックボックス化した場合に生じる。すなわち、ハッカーのハッキング以降になされた行為によって電子計算機等の損壊、虚偽の情報または不正な指令の供与等で人の業務妨害という結果を発生させたのかという因果関係の証明が問題であり、もしその原因が特定できないならばハッカーの行為と結果の因果関係が認められないことになるため、結果としてハッカーには電子計算機損壊等業務妨害未遂罪が成立するにとどまる。しかし、AI製品の学習を隠れ蓑にしてハッカーが本罪の既遂の刑責をしようとする可能性も否定できない。そうだとすると、AIのブラックボックス性がサイバー攻撃に対する刑法上の評価が未遂罪にとどまってしまう結果となってしまうので、やはりそこで重要となるのが**説明可能なAI**の構想であり、想定事例におけるデータの損壊等がAIの学習によってもたらされたか否かを事後的に証明できるようにシステム構築・開発を行うことがサイバーセキュリティ上求められることになるだろう。ただし、私的空間に属するAI製品に対するデータ変更・破壊行為は不正アクセス罪の適用のみが考慮されるにとどまり、不正指令電磁的記録を手段としてこの行為を行った場合に限り不正指令電磁的記録供与罪（刑法168条の2第1項）が成立するにすぎない。

第4款　AIソフトウェア・エージェントとコンピュータ詐欺

本款においては、AIを用いた資産運用を行うソフトウェア・エージェ

↘の単複により観念的競合か併合罪になると解すべきであるとされる（西田典之・山口厚・佐伯仁志編『注釈刑法 第2巻 各論（1）』（有斐閣、2020年）554頁（嶋矢）参照）。

ント（以下、「AIソフトウェア・エージェント」とする）に対して不正なデータが用いられ、結果としてAIソフトウェア・エージェントの利用者に対して財産的損害が発生した事例を想定する。

第1項　AIソフトウェア・エージェントとドイツ刑法におけるコンピュータ詐欺罪

コンピュータをネットワーク経由で攻撃する者は演算処理に影響を与えることで、コンピュータ詐欺を遂行しうる。データ処理設備の運営者の財産は、刑法263条aの保護の対象である[449]が、この規定は、データ処理経過の違法な操作の場合、刑法263条についての欺罔の名宛人が欠けているためこの条文を適用することが困難であったという理由で可罰性の間隙を埋めるものとされ、1987年に新たに規定されたものである[450]。

刑法263条aの意味でのデータは刑法202条aとは異なり、暗号化されたデータに限定される。そしてデータ処理とは、データを記録し、プログラムされたコードナンバーに従いそれらを結び付けることを通じて、一定の結果を得る電子データ処理システムにおける経過のことである[451]。ここで、学習能力のあるAI製品内の計算経過が刑法263条aの意味でデータ処理と認められるか否かという問題が提起される。そもそもデータ処理経過は、入力データが処理された後に出力され、処理経過が「具体的な」結果に至ることにより特徴づけられる[452]。しかし現在では、従来の入力データはなく、学習能力を有するAI製品がデータを収集し、あるいはどのデータを考慮するかを自ら決定することさえできるのではないかともいわれる[453]。その場合、入力データに応じて出力は変化するが、そのデータが直接システムに入力されるか、それともAIの自立学習により自らデー

449）BT-Drs. 10/318, S. 12, 16 ff.; *Leupold/Glossner,* a.a.O. (fn. 385), Teil 10. Rn. 140.
450）*Barton,* Multimedia-Strafrecht Ein Handbuch für die Praxis, 1999, Rn. 21.
451）*Leupold/Glossner,* a.a.O. (fn. 385) Teil 10, Rn. 145.
452）MüKo-StGB/*Wohlers,* §263a, Rn. 14.
453）*Günther,* a.a.O. (fn. 176), S. 234.

タを収集・評価してシステムに入力されたのかを事後的に区別することは困難である。そこで、刑法第263条aの適用領域を、「複雑な、知性を補完する人工的な……知能」のみに限定されるべきであるという試みもあり[454]、この試みは、データ処理はプログラムデータ以外の新たな情報を吸収することが可能であり、この情報の区分化された分析と分類は、既存の保存されたデータや並行的に記録されているデータと比較またはリンクすることによって実行される必要があり、それゆえに直接に財産を減少させる機能も存在するという発想に基づく[455]。少なくともこのことから、学習能力を有する AI システムも刑法263条aでのデータ概念に該当しうると結論づけられるだろう。

さらに、刑法第263条a第 1 項の 4 つの選択肢のいずれかを満たした上で、財産処分を惹起するようなデータ処理過程への影響が存在しなければならない[456]。この影響は、例えばプログラムの作出も含むプログラムの不正作成（ドイツ刑法第263条a第 1 項第 1 選択肢）に存在しうる[457]。ここでいう「不正」とは客観的な意味で理解され、そのプログラムが客観的にデータ処理タスクに適切に対処しているかどうかを考慮すべきであるため[458]、処分権限のあるユーザーの意思が介在しないことに注意しなければならない[459]。刑法263条a第 1 項第 2 選択肢は、不実または不完全なデータの使用（いわゆる入力操作）を処罰の対象としている[460]。データに含まれるその情報が現実と一致していない場合には不実性が、データから基礎に置く事情を十分に認識させない場合には不完全性が存在す

454） MüKo-StGB/*Wohlers,* § 263a. Rn. 15; その例として、ATM が引き合いに出される。Vgl. *Hilgendorf* Scheckkartenmißbrauch und Computerbetrug OLG Düsseldorf, NStZ-RR 1998, 137 JuS 1999.

455） MüKo-StGB/*Wohlers,* § 263a. Rn. 16.

456） *Leupold/Glossner* a.a.O. (fn. 385), Teil 10, Rn. 146.

457） *Marberth-Kubicki,* Computer- und Internetstrafrecht, 2009, Rn. 55.

458） MüKo-StGB/*Wohlers,* § 263a. Rn. 22.

459） *Marberth-Kubick,* a.a.O. (fn. 457) Rn. 56;

460） *Fischer,* a.a.O. (fn. 181), § 263a, Rn. 7.

る[461]。データの無権限使用（刑法263条a第1項第3選択肢）では、正しいデータが権限者の意思に反して使用され、その使用がコンピュータ・システムではなく自然人に対して欺罔の性格を有することが前提となる[462]。第4選択肢の「経過へのその他の無権限干渉」は、AI製品に関する構成要件的行為が、第1選択肢から第3選択肢と類似の結果不法と行為不法を有する場合[463]、例えばAI製品のハードウェアの改造、プログラムのバグの悪用と関連する。ここでは受け皿構成要件として、特に新しい技術や未知の技術を把握することから、この類型がAIの分野でますます関連するものとなろう。

これらの犯罪遂行の4つの選択肢は、いずれもAIソフトウェア・エージェントに対して大きな意味を持つ。例えば、電子商取引では、不正なデータを使用することや、オンラインショップでの価格表示の誤認でAIソフトウェア・エージェントが「騙される」可能性があり、その結果AIソフトウェア・エージェント利用者の財産損害が発生しうる。また他人の署名を使用することも、データの無権限使用を充足する可能性がある。ここでは、第三者が当該AIソフトウェア・エージェントに偽の署名をして自らを認証するような場合が該当するだろう。第4選択肢は、他の類型と同置される、未知のあるいは新たな技術をカバーすることができるため、ロボット工学の範囲では関心が持たれうる。この類型では、学習能力を有するAIソフトウェア・エージェントが、第三者から誤情報を与えられたられた場合に認められうる。そして、AIソフトウェア・エージェントの利用者の財産が実際に減少（財産損害）したことが本罪の成立には必要である[464]。

461） *Fischer,* a.a.O. (fn. 181), § 263a. Rn. 7.

462） *Fischer,* a.a.O. (fn. 181), § 263a, Rn. 10, 11. このことには争いがないので、技術的システムがまさに人間と比較可能でないということが議論される。「コンピュータ特有の」もしくは「詐欺特有の」解釈をめぐる争いについては、*Fischer,* a.a.O. (fn. 181), § 263a. Rn. 9 ff. も参照。

463） *Fischer,* a.a.O. (fn. 181), § 263a, Rn. 18.

464） MüKo-StGB/*Hefendehl/Noll,* § 263a, Rn. 179.

このとき、資金移動の手続が不正に利用された場合などが定期的に存在するなど、資産に損害を与える（損害に相当する）リスクも財産的不利益を構成し本罪の対象となる[465]。

第 2 項　AI ソフトウェア・エージェントと日本刑法における電子計算機使用詐欺罪

キャッシュレス化がますます推進されている時代において、コンピュータによる処理のみが予定される取引形態もまた今後ますます拡大していくと予想されるところ[466]、もし行為客体が財物であれば、人の判断を介さずに財物の占有を取得する行為には窃盗罪（刑法235条）を適用しうるが、条文上客体が財物に限られる窃盗罪を、財産上の利益の場合に拡張することは許されないので、機械を不正に操作して人の判断を介在させずに財産上の利益を不正に取得する行為は、窃盗罪にも詐欺罪にも問えないこととなる[467]。本条は、このようなシステムを悪用する新たな財産侵害行為に対処するため、1987年改正により設けられたものである。

本条の構成要件として、まず「財産権の得喪、変更に係る電磁的記録」

465）　MüKo-StGB/*Hefendehl/Noll,* § 263a, Rn. 179.

466）　鎮目ほか・前掲（注403）344頁以下（鎮目）。

467）　西田典之・山口厚・佐伯仁志編『注釈刑法 第 4 巻 各論（ 3 ）』（有斐閣、2020年）317頁以下（西田＝今井）によると、たとえば、ATM により他人の預金を自己の口座に振替送金する行為は、それだけではいまだ「財物」を取得したとはいえないために窃盗を構成しないし、自己の口座に他人の預金を不正に付け替えた後、いまだ現金化しない間に自動振替によって水道・ガスなどの利用料金が引き落とされた場合でも行為者は一度も現金を手にしていない」ので窃盗罪の成立は否定されるという。さらに、鎮目ほか・前掲（注403）345頁以下（鎮目）では、甲がＡ銀行の係員に偽札を渡すことで100万円の入金処理をさせて預金債権を取得した場合にいわゆる 2 項詐欺罪（刑法246条 2 項）が成立するにもかかわらずインターネットバンキングを用いた場合には不可罰だとするのは不均衡であるという。インターネットバンキングを用いたとしても、その行為は、真実は経済的・資金的実体の伴う入金がないにもかかわらずそれがあったかのように装うものであり、これも「機械を欺いて」預金債権を取得する行為である。これは人に対する詐欺罪と同質の行為であるといえるから、詐欺罪と同等の処罰に値するものだという。

とは、財産権の得喪、変更の事実又はその得喪、変更を生じさせるべき事実を記録した電磁的記録であって、一定の取引場面において、その作出（更新）により事実上当該財産権の得喪、変更が生じることとなるようなものをいう[468]。その財産権とは、金銭的価値を内容とする権利であって、債権、物権等がその典型であるとされ[469]、記録の作出等と事実上の財産権の得喪、変更との間の直接的あるいは必然的な関連性を要する[470]。次に、前段類型の「虚偽の情報」とは、電子計算機を使用する当該システムにおいて予定されている事務処理の目的に照らし、その内容が真実に反する情報」をいう（東京高判平成５年６月29日高刑集46巻２号189頁）。「不正な指令」とは当該事務処理の場面において、与えられるべきでない指令のことをいい、「不実の電磁的記録を作り」とは、真実に反する内容を、記録媒体上に電磁的記録を存在するにいたらしめることをいう。ここには記録をはじめから作り出す場合のほか、既存の記録を部分的に改変、抹消することによって新たな電磁的記録を存在するにいたらしめる場合も含まれる。後段類型の「財産権の得喪若しくは変更に係る虚偽の電磁的記録を人の事務処理の用に供する」とは、行為者が真実に反する財産権の得喪、変更に係る電磁的記録を他人の事務処理に使用される電子計算機において用いうる状態に置くことをいう。そして前段・後段に共通する要件としては不当利得があり、例えば、不実の電磁的記録を使用して銀行の預金元帳ファイルに一定の預金債権があるものと作為し、その預金の引出し、振替を行うことができる地位を得ることなど、事実上財産を自由に処分できる利益を得ること、不正に作出したプリペイドカードを利用して労務やサービスなど一定の役務の提供を受けること、料金の計算及び請求が行われることとなる課金ファイルの記録を改変して料金の請求を事実上免れることなどがこれに該当し、必ずしも実際に権利又は義務の得喪、変更が行われたこと

468） 米澤・前掲（注376）116頁。
469） 大塚ほか・前掲（注434）13巻180頁（和田）。
470） その理由については、鎮目ほか・前掲（注403）346頁（鎮目）参照。

を要しない[471]。

第 4 款の冒頭で示した想定事例において、ドイツ刑法でのコンピュータ詐欺罪と比較すると不当利得要件を満たすか否かが問題となる。単なる財産損害にとどまる場合にはこの類型は該当せず、行為者に対し虚偽または不実のデータを供用したことについて、行為客体たる AI ソフトウェア・エージェントが業務に供するものである限りで電子計算機損壊等業務妨害罪の成立が考えられるが、私的用途で利用される AI ソフトウェア・エージェントの場合はその類型にも該当せず何らの犯罪も成立しないことになる。このことは、AI ソフトウェア・エージェントの利用者に対する財産的損害を与えることについて犯罪が成立しないことを意味するため、私見としては前項でも述べたように、電子計算機損壊等業務妨害罪の業務性要件を一般的利用にも拡張することで解決を図るべきであると思う。また、AI ソフトウェア・エージェントの学習に関する問題では、財産権の得喪もしくは変更に係る不実の電磁的記録の原因たる虚偽の情報もしくは不正な指令、ないしは財産権の得喪もしくは変更に係る虚偽の電磁的記録が人間の手によってもたらされたのか、それとも AI ソフトウェア・エージェントの学習によってもたらされたのかが不明な場合に問題となる。仮に行為者によってもたらされたデータが虚偽または不実のものであっても、不実の電磁的記録の作成が AI の学習によってなされたとすれば、行為者が AI の学習の結果、不実の電磁的記録を作成することを事前に認識していない限りで、本罪の行為類型を充足するとは言えず、行為者が不当利得を受けたとしても電子計算機使用詐欺既遂罪の成立は否定されてしまう。仮に行為者が不当利得を得る意思があったとしても、AI の学習過程があるか否かが既遂か未遂かの評価の分水嶺となってしまうことと、先述の電子計算機損壊等業務妨害罪やドイツ刑法におけるデータ変更・コンピュータ破壊罪のように、AI の「学習」によって行為者の罪責が未遂減軽される

471） 大塚ほか・前掲（注434）13巻187頁（和田）。

可能性があることが問題だろう。ここでも重要なのが**説明可能な AI** の構想であり、行為者が当該 AI ソフトウェア・エージェントに与えた情報が不実の電磁的記録が作出したか否かを事後的に検証できるようなシステム構築を図ることが、AI の学習の不正な使用を防止すること、そして AI 学習に対する社会的信頼を確立するためにも必要なことである。

第5款　小　　括

本節では、主に学習機能を有する AI 製品がハッキングによるサイバー攻撃を受けた場合に想定される事例を、① AI 製品内に保存されているデータを取得した、② AI 製品の内部データを変更・破壊することによって利用者に一定の不利益が生じさせた、③ AI ソフトウェア・エージェントに対し虚偽または不実の情報を供与してその利用者に財産的損害を与えた、という 3 つの類型に分類し、その手段行為のハッキング行為も含めてそれらの刑法上の評価を検討した結果、構成要件自体の解釈と AI の学習固有の問題があることが分かった。

まず、ハッキング行為についてドイツ刑法では、保護されるデータが日本における不正アクセス罪のようにアクセス制御機能を有しているか否かには関係なくドイツ刑法202条aの対象となる。この点、日本刑法ではアクセス制御機能のないデータに対するアクセスは処罰の対象としていないことから、利用者・製造者の側でセキュリティを強化するようなシステム構築が必要であるといえる。データ取得について、ドイツ刑法では202条bに規定される構成要件に該当する一方で、日本法では電気通信事業者法179条 1 項または有線電気通信法14条の罰則が関連することになる。データ変更・破壊にかかる利用者へ不利益を生じさせる行為については、ドイツ刑法では303条bのデータ変更罪の構成要件に該当するが、この対象となるデータについてはその利用者にとって本質的に重要であるか否かが問題となる。その一方で、日本刑法では刑法234条の 2 所定の電子計算機損壊等業務妨害罪の成否が問題となるが、ここでは対象となる電子計算機が

もっぱら業務に供するものでなければならないことに留意しなければならない。この点、私的空間で利用される AI 製品の場合、ドイツ刑法下では処罰の対象となりうるが、日本刑法のもとでは処罰の対象とならない。この比較から、AI 製品ひいては IoT 化された製品に対する保護の観点も踏まえて、業務性要件を保持しつつ、私的空間に属する電子計算機にもその範囲を拡げるべきではないかと考える。AI ソフトウェア・エージェントに対し虚偽または不実の情報を供与してその利用者に財産的損害を与えた行為について、ドイツ刑法では263条a所定のコンピュータ詐欺罪が、日本刑法ではで246条の 2 所定の電子計算機使用詐欺罪の成否が問題となる。前者では、当該行為についてドイツ刑法263条a第 1 項第 4 選択肢の受け皿構成要件に該当しうる一方で、後者では財産損害ではなく行為者の不当利得を求めることから、当該行為の構成要件には該当しないことになる。

そして、AI の学習のブラックボックス性は、上記のうちデータ変更・破壊とコンピュータ詐欺類型に関連する。前者においては、データの変更や削除という結果が、後者においてはその手段たる虚偽または不実の電磁的記録が、行為者（ハッカー）によるものなのか、それとも AI 自身の学習によるものかが不明確な場合、構成要件的結果が実現されたとしても因果関係が肯定されず未遂罪の適用にとどまり減軽の余地が残されてしまう。このアンバランスさの解消、そして因果関係の慎重な認定のため、**説明可能な AI** の構想に基づき、事後的にその因果関係を証明できるようなシステム構築を求めることにより、AI 製品に対するセキュリティ上の保護、そして AI の学習に対する社会的信頼を確立することができるように思われる。

第 4 章　AI 製品開発に対する将来的な刑法上の規制

第 3 章でみてきたように、AI の利活用によっては現行刑法及び特別刑

法の規制の射程が及ばなかったり、解釈上処罰（・制裁）範囲が拡張されたりする可能性があることが確認された。これらの議論はすでに人間の手によって作られ、これから発展する可能性のある AI に関する議論であることに留意しなければならない。この点において、そもそも人間社会にある一定の害を及ぼす可能性のある AI を創るべきでない、という見解もありうる。それは本稿で対象としてきた「弱い AI」の枠組を超える、「強い AI」にもその射程が及ぶという。それは、このような AI の開発に関し、事前抑制を意味するが、この点において刑法はどのような役割を果たすべきなのか。これについては、設計開発のみならず、研究自体も規制となりうる可能性を考慮しつつ、今後の AI 開発についての法的考察をまとめた Gaede の見解[472]を参照しながら検討を行う。

第1節　問題の所在――強い AI とその現状

強い AI の土台が、この瞬間にも世界中で真摯な試みとして自然科学的なものとして位置づけなければならない手法で研究されている[473]。また、人類の多くが AI の潜在能力を高める投資や能力に目を向けているように見えるという事実も十分に留意されないままに、経済、医療、軍事において、自らの将来を賭けることになるのではないかという不安の中で、実際に技術水準を変革するために多大な努力が尽くされている[474]。それと同時に、強い AI の研究により高性能のプラットフォームも作られつつあり、その一例として米国防総省の研究機関 DARPA は、すでに軍事的に利用可能なモデルやシステムの研究の一部でさえも AI が担うことになっている AI のプログラムを作成しており、それに対してエラーをする可能性があり、動作の遅い人間は「ゲーム・チェンジ AI」によって置き去り

472）*Gaede,* Künstliche Intelligenz –Rechte und Strafen für Roboter? Plädoyer für eine Regulierung künstlicher Intelligenz jenseits ihrer reinen Anwendung, Nomos 2018.

473）*Gaede,* a.a.O. (fn. 472), S. 72 参照。ここには「フランケンシュタイン――保護領域の例外」がある。*HLEG on AI, supra* (fn. 95), p. 12 も見よ。

474）技術的革命については *Hilgendorf,* a.a.O. (fn. 107) S. 99 f. も言及している。

にされるという[475]。こうした見解に対する戦略的な対応は、強い AI は将来的にも不可能であると考えることにある。AI の研究者が尽力した、人間は計算可能で再構築する能力を持つ機械であるという争いのある想定は[476]、むしろ狂気の沙汰のようなものであるとともに、我々は、説明しえない霊感だけが知的生命体を創造することができると信じるために、自らや人類を買いかぶるマッドサイエンティストは明確なものなので不安を煽るものではないが、万一に備えて危険防止や処罰のための規制権限に触れる場合があるかもしれない[477]。

強い AI が実現するのはいつであるかという予測は問題にしないものの、もはや SF 映画の世界にそのテーマを委ねるべきではないと考えられる。とりわけ野心的な AI 研究を誇大妄想だと片付けようとする者でも、中心となる問題を誇大妄想家の制御をその手に委ねることになるが、実体的な法理解には、法が紙上の理想として公式化されるだけではないことが必要である。むしろ、我々は適切な配慮を講じることができるように、我々の法の形と執行可能性に対する実存的な危険に、たとえわずかな兆候であっても時宜に即して注意を向けるべきである[478]。

また、人間の自律性を高度に模倣したことによっても、重大な危険が生じうる。例えば、弱いながらも自己学習し、部分的には優れた能力を持っている AI が、誤解を伴った命令により、制御不能になったり、危険な状態になったりすることがある[479]。また、技術的な基盤が予測可能な形で

475） それに対応する計画については https://www.darpa.mil/work-with-us/AI-next-campaign（最終アクセス2022年11月11日）に記載される「米国国家安全保障のための新たなゲーム・チェンジ AI 技術」では、米国国防総省は2030年までに 1 億6000万ドルをさらに自律化した「未来戦闘部隊」に投じるとされる。

476） その立場を支持する石黒氏については *Eberl,* a.a.O. (fn. 59) S. 321 も参照。

477） *Gaede,* a.a.O. (fn. 472), S. 72.

478） *Gaede,* a.a.O. (fn. 472), S. 73.

479） この例については *Bostrom,* Superintelligence – Paths, Dangers, Strategies, 2014, p. 146. *Russell/Norvig,* Artificial Intelligence – A Modern Approach (3rd Edition), 2010, p. 1010. 後者は、例えば AI が首尾一貫しない、明確に定義されなかった道徳律から誤っ↗

改善されることで、自己意識を探求する研究者が、半ば輝かしい自己意識へのブレイクスルーを超克するために、意図的または無意識的に重大な効果を持つリスクを冒すという危険を高めるかもしれない。この関連で「Random Darknet Shopper」というAIも訴求力を持つ。そのAIは、ダークウェブ上でドラッグと偽造パスポートをランダムに注文していたという[480]。システムは、確かに分別を持たないままではあるものの、予測不可能な危険な方法で「行動」することになることはこの点からも窺える。

人間と機械が共存する未来に向けて、危険防止や刑罰といった折り紙付きの手段への信頼は、決して将来への不渡手形をとなるべきではないが、そのためには個別利用の範囲外でもAIに目を向ける必要がある[481]。なぜなら、このようにして初めて包括的かつ技術的な危険分析が成功するからである。また、開発は世界中で行われ、容易には透明化されない共有財産が存在する状況を鑑みれば、すでにその中には第一の大きな挑戦が存在する。それは軍事的利用のために秘密となっていることや、企業秘密として自由に閲覧できない部分領域であるとされる[482]。

第2節　規制的措置

さらに、アナログもしくはデジタルでは操作できない手段を法執行に対して適用することも検討すべきであろう。人間の手によって操作されるという脅威は以前から存在していたが、いわゆるサイバーセキュリティの中

↘た結論を人間への対処について導出してしまう可能性を説明する。その一例として、その低い知能のために我々と同等の存在とはみなさない昆虫を殺しても良いという権限について挙げられている。

480）https://motherboard.vice.com/de/article/78kyz4/random-darknet-shopper-590 や https://motherboard.vice.com/de/article/kb7jma/kunstfreiheit-siegt-in-der-schweiz-duerfen-bots-drogen-im-darknet-kaufen-632.（最終アクセス2021年9月25日）を参照。

481）例えばBT Drs. 19/5880を参照。

482）*Gaede,* a.a.O. (fn. 472), S. 75.

ではこれについて争われている。とりわけ、AI の研究・構築の規制を検討することもまた必要であると思われる。

第 1 款　2010年代における AI 製品開発・研究に対して考慮されてきた規制

弱い AI の利益を奪うような過剰な規制に陥りたくないならばどうすればよいか。これには「強い AI」を目指す、あるいはそれが当然だと思われる段階を含む研究が重要だと思われるが、こうした研究は産業用ロボットやサービスロボットのための、用途に方向づけられた規制や市場に方向づけられた規制、例えば、ロボットの利用であれば製品安全ガイドライン ISO 10377 や RL/2001/95/EG（欧州）、産業用ロボットであれば ISO 10218-1, ISO 10218-2：2011、サービスロボットであれば ISO 13849-2, ISO 18646-1：2016（ISO 18646-2/3/4,13482）といった規制を超えるものであることに留意すべきである。

では、このような AI 研究に必要な義務づけはどのようなものとなるのか。それは技術水準に従って、開発された AI がその評価に先立って研究空間や研究ネットワークを超えた影響を及ぼさないようにも義務づけをすることにあり、それには「封じ込め」の形式が必要とされる[483]。すなわち、新しい AI がテストされることなく他の技術システムに波及し、ないしはそれらとネットワーク化されないよう、物理的ないしはアナログ的に保証するべきという[484]。AI に対しては適用の動機から生じる、これまでの自発的な選好を超えた学習プロセスの導入を義務づけなければならないが、この場合、たとえ莫大な経済的投資をしたとしても、いまだ十分な安全性を確保できていない技術が排除されてしまうことは考えられないこと

483）自己学習する（産業用）ロボットにおける必要不可欠な「カプセル化」はますます困難となっていくであろうと認める *Stell/Krüger,* Lernen und Sicherheit in Interaktion mit Robotern aus Maschinensicht, in: *Hilgendorf/Günther,* Robotik und Gesetzgebung, Nomos, 2012, S. 51, 61, 68 ff. も参照。

484）*Gaede,* a.a.O. (fn. 472) S. 81.

であってもよい。もちろん、自律的な技術を多種多様に扱うため、いわゆる残存リスクからの絶対的な安全性を提供することはできないものの、極めてリスクのある研究に対して具体的で実現可能な注意基準を対置させることは可能である。すなわち、法の維持のためには、AI研究に適切な法執行のインターフェースを、それに失敗した場合には、同等のメカニズムを再現することを追加で検討すべきである。このことについて欧州議会は、（AIを搭載する）ロボットは常に人間がいかなるときでも統制できるように構築されるべきだとさえ要請している[485]ことからも示される。

第2款　2020年代にAIの製品開発に対して策定された国内外の規制

前項での議論がなされていた2010年代後半からさらに発展して、2020年代には欧州連合、米国、中国で相次いでAI開発に関する法的ガイドラインが策定された。以下、各国のAI規制に関して概観し、それと日本のAI規制を比較しながら、将来的な国際的レベルでのAI開発の観点のもと望ましい規制となっているのか、そうでなければどのような問題点があるのかを抽出し、その解決策を提言したい。

第1項　欧州AI規制案（EU）

2021年4月、欧州委員会はAIに関する規制法についての提案（以下、「欧州AI規制案」という。特に断りのない限り規制案の条文を引用する際には条文番号のみを記載する）である「Proposal for regulation of the European

485）Vgl. *Europäisches Parlament,* Resolution zu Zivilrechtlichen Regelungen im Bereich Robotik, P8_TA (2017) 0051, Allgemeine Grundsätze bezüglich der Entwicklung der Robotik und der Künstlichen Intelligenz, 3. この要請は、主に意識を持たないAIに向けられたものと思われるが、欧州議会が除外していない自己意識型ロボットの場合、完全な制御可能性の要求は自由主義論の観点からは攻撃されうる。道徳的主体が承認される限り、原則的に自由が与えられなければならず、自由制限的、あるいは国家の連関の中で自由を定義するような規範の執行は種類や程度に応じて正当化されなければならない（Vgl. *Gaede,* a.a.O. (fn. 472), S. 79 fn. 214）。

Parliament and of the Council: Laying down harmonised rules on artificial intelligence（artificial intelligence act）and amending certain union legislative acts」を発表した[486]。この提案では「リスクベースのアプローチ」が強調されており、そのほかにもこの規制法案の特徴として、EU における既存の規制との調和、future-proof（将来の問題に対処できるような設計であること）、EU における統一的市場の確保、投資とイノベーションの促進といったことが強調されている。これらを見ると、コストとベネフィットの両方に配慮したビジネス的観点を重視したものとされる[487]。

この欧州 AI 規制案が我が国にもたらす影響としても留意しなければならないのは AI 法案の規制範囲が EU 域内だけでなく域外にも適用される点である[488]。欧州 AI 規制案は規制対象として次の 6 つを列挙する。すなわち、(a) 設立されたのが EU 域内であるか第三国であるかにかかわらず、EU において AI システムを市場に置き又はサービスを提供する提供者、(b) EU 域内に所在する AI システムの利用者、(c) AI システムが生み出すアウトプットが EU 域内で利用される場合における、第三国に所在する当該システムの提供者及び利用者、(ca) 許容できないリスクに位置付けられ、禁止される AI システムを EU 域内で提供または販売する場合、EU 域外で流通に置くまたはサービスを提供する提供者、(cb) AI システムの輸入者及び販売者、並びに授権された AI システムの提供者の代理人であって、当該輸入者、販売者または授権された代理人が EU 域内に事業

486） その原文（英文）は https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206（最終アクセス2022年11月11日）で閲覧可能である。なお、2023年 5 月16日に修正案が欧州議会で可決されている。その修正内容については、https://www.europarl.europa.eu/resources/library/media/20230516RES90302/20230516RES90302.pdf（最終アクセス2023年 5 月17日）で閲覧可能である。以下の説明は、当該修正案を踏まえたものである。

487） 久木田水生「AI のリスクと倫理」第36回人工知能学会全国大会論文集（2022年） 1 頁。

488） 北和樹「EU が目指す AI 社会のための規制法」立命館大学人文科学研究所紀要131号（2022年）287頁以下参照。

所を有する、または所在している者、(cc) EU 域内に所在し、EU 域内で流通に置かれた、またはサービス提供された AI システムの使用により健康、安全または基本的権利に悪影響を受けた者である（2条1項）。ここでの提供者とは、「有償か無償かにかかわらず、AI システムを開発するもしくは AI システムを市場に投入することまたは自身の名前もしくは商標で AI システムのサービスを提供することを目的として開発された AI システムを所有する、自然人、法人、公的機関、行政機関またはその他の機関」とし（3条2号）、また利用者とは、AI システムが個人的な非専門的活動で使用される場合を除き、「自身の権限の下で AI システムを使用する自然人、法人、公的機関、行政機関またはその他の機関」とする（3条4号）。

次に、規制案で制限を受ける AI のリスクについては4つの分類――「許容できないリスク」、「高リスク」、「限定リスク」、「最小限リスク」――が用いられている。

「許容できないリスク」に位置付けられ、禁止される AI の活動態様としては5条1項に規定がある。そこでは、(a) それにさらされる個人またはその法定後見人の特定のインフォームド・コンセントに基づいて、承認された治療目的のために使用されることを意図する AI システムを除く、人の意識を超えたサブリミナル技術または意図的に操作的若しくは偽装的な技術を用いる AI システムの市場投入、サービス提供、または使用することで、情報に基づいた意思決定を行う人の能力を著しく損なうことによって、個人または集団の行動を重大に歪め、それにより、本人、他の人または人的集団に著しい損害を与える、または与えるおそれがある方法でその人が他の方法で行わなかったであろう意思決定を行うようにする目的または影響を与えること、(b) 個人または特定の集団の脆弱性（当該個人または集団の既知または予測される性格特性または社会的もしくは経済的状況、年齢、身体的もしくは精神的能力の特性を含む）を利用する AI システムを、当該個人または当該集団に属する者の行動を、当該個人または他の者に重大

な損害を与える、または与えるおそれがある方法で実質的に歪ませる目的または効果をもって市場に投入する、サービスを提供するまたは使用すること、(ba) 自然人をセンシティブな、または保護された属性もしくは特性に従って、またはこれらの属性もしくは特性の推論に基づいて分類するバイオメトリクス分類システムの市場投入、サービス提供または使用すること、(c) 自然人またはその集団の社会的行動または既知、推論もしくは予測される個人的もしくは人格的な特徴に基づき、(i) データが当初生成または収集されたコンテキストとは無関係な社会的文脈における、特定の自然人またはその集団の不利益な、または望ましくない扱い、(ii) 特定の自然人またはその集団の社会的行動またはその重大性に不当または不釣り合いな、不利益、または望ましくない扱いのいずれか、もしくは両者をもたらすような一定期間にわたってソーシャルスコアリング、評価または分類を行う AI システムの市場投入、サービス提供または使用、(da) 自然人またはその集団の犯罪または再犯のリスクを評価するため、または自然人のプロファイリングに基づき、あるいは自然人またはその集団の過去の犯罪行動や所在地を含む個人的特性や性格の評価に基づき、実際または潜在的な犯罪または行政犯罪の発生または再起を予測するための、自然人またはその集団のリスク評価を行う AI システムの市場投入、サービス提供または使用すること、(db) インターネットまたは CCTV 映像から顔画像を非標的的に切り抜くことにより、顔認識データベースを作成または拡張する AI システムを市場投入する、サービス提供するまたは使用すること、(dc) 法執行、国境管理、職場および教育機関の分野で、自然人の感情を推測する AI システムを市場投入する、サービス提供するまたは使用の禁止が規定される[489]。この規制の実効性を担保するために、71条 3 項で違反

489） 小泉雄介「欧州 AI 規制案の概要」データ社会推進協議会データ倫理プライバシー研究 WG 資料（2021年）https://www.i-ise.com/jp/information/report/2021/202106.pdf（最終アクセス2022年11月11日） 8 頁ではそれぞれの類型の具体例としては以下のものをあげる（ただし、修正前の規制によるものである）。

(a) トラック運転手に可聴域でない音を聞かせて、健康かつ安全な範囲を超えて長時間↗

者に対し4,000万ユーロ以下の行政上の制裁金を、違反者が企業である場合には、前会計年度の世界全体における売上総額の７％以下の金額のうち、いずれか高い金額の行政上の制裁金を課す罰則が規定される。

また「高リスク」に位置付けられるAIとは、自然人の健康と安全、あるいは基本的な権利に高いリスクを生じさせるものとされている[490]。こういったシステムは特定の要件[491]を遵守し、事前の適合性評価（19条）を受けているならば、欧州の市場に出すことが許される。高リスクのAIの利用について、提供者に対してはリスク管理システムの確立（９条）、実装（10条）、文書化（11条）、動作中の記録保持（12条）、利用者への透明性の確保（13条）、健康、安全又は基本権に対するリスク防止又は最小化を目的とする人間による監視（14条）、及びセキュリティの確保（15条）を義務づける（16条以下）。また、付属書Ⅱ―Aに定められる製品の場合には、当該製品の製造者に対しても提供者と同様の義務を課す（24条）。利用者に対しては、自ら入力データの管理を行う場合には、入力データが高リスクAIシステムの意図された目的の点から見て関連性を有し，これを十分に代表するものであることを確保する義務（29条３項）、使用上の指示に基

↘ 運転させるようにする。AIはこのような効果を最大化する音域の発見に使用される。

(b) 音声アシスタントを組み込んだ人形が、楽しくクールなゲームを装って、未成年者に次第に危険な行動やチャレンジをするようにけしかける。

(c) AIシステムが、医者予約の無断キャンセル、離婚など、親の取るに足らない、あるいは無関係な社会的な「不正行為」に基づいて、社会的ケアを必要としている子どもを特定する。

(d) ビデオカメラによってライブで撮影された全ての顔が、テロリストを特定するためにデータベースに対してリアルタイムでチェックされる。

490) その内容は６条１項と２項および付属書Ⅱ、付属書Ⅲに記載がある。付属書Ⅱ―Aでは、機械、玩具、娯楽用船舶、昇降機、医療機器などに関するEU規則内で対象とする製品に、付属書Ⅱ―Bでは、航空機、鉄道、自動車などに関するEU規則内で対象とする製品に第三者適合性評価（43条）を受ける義務を提供者に課し（21条）、付属書Ⅲでは自然人の生体識別・分類、重要なインフラの管理・運営、教育・職業訓練、雇用・労働者管理、及び自営業へのアクセス、重要な民間・公共のサービス及び給付へのアクセス及び享受、法執行など高リスクに該当する類型を列挙している。

491) ８条から15条にかけてその内容が具体化されている。

づいて高リスク AI システムの動作を監視し、重要である場合は提供者に知らせる義務（29条 4 項前段）、使用上の指示に従って使用した場合に AI システムがリスク[492]を示すことになる可能性があると考える理由がある場合には、提供者又は販売者に遅滞なく知らせるとともに、当該システムの使用を中止し、基本権を保護することを意図する EU 法上の義務の違反に該当する重大な事象又は機能不全を特定した場合にも、提供者又は販売者に知らせるとともに、AI システムの使用を中断する義務（29条 4 項後段）、高リスク AI システムによって自動生成されたログが自らの管理下にある場合にそのログを維持する義務（29条 5 項）を有する。これら要件または義務に違反した場合、1,000万ユーロ以下の制裁金、または違反者が企業の場合は、直前の会計年度における世界全体における売上総額の 2 %以下の金額、もしくはいずれか高額の方の制裁金の賦課という罰則を定める（71条 4 項）[493]。

「限定リスク」もしくは「最小限リスク」に位置づけられる AI としては、人と交流することを目的とした AI システム、感情を認識するために使用されたり、生体認証データに基づいて（社会的な）カテゴリーとの関係性を判断したりするシステム、存在する人・モノ・場所・その他の存在に酷似し、本物または真実であると誤解させる（「ディープフェイク」）画像、音声、またはビデオコンテンツを生成または操作する AI システムを指す[494]。これら AI システムに関する法的義務として、AI システムと相互作用をしている人々に、それが明白でない限り、チャットボットなどその旨を通知すること、感情認識システムや生体カテゴライゼーションシス

492） そのリスクの内容は Article 3, point 19 of Regulation (EU) 2019/1020 を引用する。そこでは、合理的かつ許容可能と考えられる程度を超えて、一般人の健康および安全、職場における健康および安全、消費者の保護、環境、公安および該当する法令で保護されるその他の公益に悪影響を及ぼすものと定義される。

493） ただし、データによるモデル学習を伴った技法を利用する「高リスク AI」の開発要件である10条に違反した場合は71条 3 項aの制裁の対象となる。

494） 北・前掲（注488）294頁。

テムの対象となる人々にその旨を通知すること、表現の自由などの基本的権利の行使や、公共の利益の理由からディープフェイクが必要な場合を除いてディープフェイクコンテンツにラベルを付けることが規定される（52条）。また、AIによってサポートされるゲームアプリケーションやスパムフィルタ機能AIなどは「最小限リスク」として位置づけられ、この種類のAIの任意の適用を促すことを目的とした行動規範について、当該AIシステムの意図された目的を考慮して、当該要件の遵守を確保する適切な手段である技術上の仕様及びソリューションを基礎として作成することを促進するにとどめる（69条）。ただし、これら「限定リスク」や「最小限リスク」に該当するAIであっても、71条４項の文言に従えば「高リスクAI」の場合と同様の制裁が課されることになることに留意しなければならない。

これら４類型のうち、「許容できないAIリスク」を除くもののリスクを示すAIシステムに関しては、市場監視機関が、当該評価の過程で、AIシステムが本規則に定める要件及び義務を遵守していないことを発見した場合、その市場監視機関は、AIシステムをして当該要件及び義務を遵守させるために、AIシステムを市場から取り下げるために、又はAIシステムをリコールするために、リスクの性質に比例した当該市場監視機関が定め得る合理的な期間内に、適切な全ての是正措置を講じるよう、遅滞なく関係する事業者に要求するものとし（65条２項）、事業者が適切な是正措置を講じない場合には市場監視機関が製品を当該市場から取り下げるための、又は製品をリコールするための適切な暫定措置を講じるとする（65条５項）という制裁規範がさらに存在することも見過ごしてはならない。

この規制に関してEUの産業界からは、企業の負担が増加することに関する懸念が規制案発効後すぐに表明され[495]、イノベーションの阻害と

495）その具体的な内容については、寺田麻佑・板倉陽一郎「欧州（EU）における2021年AI規制法案をめぐる各種意見とEUの対応の検討」情報処理学会研究報告22号（2022年）２頁以下も参照。

なることが示されている[496]。また、欧州の機械電気電子金属加工産業連盟（Orgalim）は、AIシステムという定義が不明確なのでその定義をより明確化するとともに、産業用AIは高リスクとみなされないことを保証することを求めること、適合性評価の義務化は企業の負担を増やし、安全性を高めることには必ずしもつながらないのではないのかという懸念を示し[497]、経団連の声明においても、「罰則の対象が広範に及び、また罰金額が非常に高額であることは、欧州市場における企業の活動を過度に委縮させる恐れがある。そこで、違反の類型や内容、得られた便益の大きさ、違反の悪意の有無などに応じて、適切なペナルティを定めるべき」という。私見としても、当該規制の文言解釈からすれば、そもそも71条の表題が罰則（penalties）であること[498]と、制裁金（administrative fine）が課される要件が5条所定の「許容できないAIリスク」の利用に該当するのみならず、「高リスクAI」の利用における他の義務やその他「限定リスク」や「最小限リスク」AIに課せられうる義務まで制裁金の対象となること、さらに、市場監視機関がそれらの義務の不遵守を発見した際にはその事業者に対して製品のリコールのみならず市場からの取下げを命じることもできることに鑑みれば、これら規制は営業の自由を侵しかねない強い制裁規範であり刑罰的性格を帯びたものといってもよいだろう。しかもこのような「財産刑」の対象となるAIのリスク評価がもっぱら市場監視機関に委ねられており、さらには抽象的なリスク段階での規制であることも考慮すればこのような規制の運用自体をより慎重にしなければならない。

496） See, BCS: New EU AI regulations demand a 'fully professionalised tech industry' – institute for IT. 2021, Apr 22.

497） *Orgalim*, European Regulation on Artificial Intelligence – Orgalim calls for legal clarity and workability, 21 April, 2021（https://orgalim.eu/news/european-regulation-artificial-intelligence-orgalim-calls-legal-clarity-and-workability 最終アクセス2022年11月12日）

498） なお、本規制案のドイツ語版（https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:52021PC0206&from=EN（最終アクセス2022年11月13日））での71条の表題は「Sanktionen」となっている。

第２項　AI 権利章典（米国）

2022年10月４日、米国ホワイトハウス科学技術政策局（OSTP: Office of Science and Technology Policy）は AI の開発に考慮すべき原則をまとめた「AI 権利章典のための青写真」（Blueprint for an AI Bill of Rights: A Vision for Protecting Our Civil Rights in the Algorithmic Age）を公表した。ここでは、患者の治療に役立つはずのシステムが、安全でないこと、効果がないこと、あるいは偏ったものであること、雇用や信用に関する判断に使われるアルゴリズムが不公平を反映・再現したり、新たな有害な偏見や差別を埋め込んだりしていること、ソーシャルメディアにおける無制限のデータ収集がプライバシーを損ない、本人の認識や同意なしに人々の活動を広く追跡するために利用されていることを問題の根底に置き、そのような公民権に対する脅威からすべての米国民を守り、その最高の価値を強化する方法でテクノロジーを活用する社会のための指針を、① 安全かつ効果的なシステム（Safe and Effective System）、② アルゴリズム的差別からの保護（Algorithmic Discrimination Protections）、③ データ・プライバシー（Data Privacy）、④ 通知と説明（Notification and Explanation）、⑤ 人間への代替、考慮、予備的措置（Human Alternatives, Consideration, and Fallback）の５つの原則にまとめたものである[499]。

①「安全かつ効果的システム」では、システムは多様なコミュニティやステークホルダ、専門家と協議の上、開発を行うものとし、システムを配備する前に試験を行い、リスクを特定・軽減し、システムの監視を行う。これらの保護措置の結果として、場合によってはシステムの配備中止や削除もあり得るとする[500]。その実践のために、AI が(a)合法的かつ国家の価値を尊重し、(b)目的を持ちパフォーマンスを重視し、(c)正確かつ信頼可能で効果的に、(d)安全、堅牢で弾力性があり、(e)理解可能で、(f)責任

499）　OSTP, Blueprint for an AI Bill of Rights: A Vision for Protecting Our Civil Rights in the Algorithmic Age, October 14th, 2022.

500）　OSTP, Blueprint, *supra* (fn. 499), p. 15.

があり追跡可能な、(g) 定期的に監視され、(h) 透明性を有し、(i) 説明責任を有するものであることを求める。

②の「アルゴリズム的差別」からの保護は、システムが人種、肌の色、民族、性別、宗教、年齢、国籍、障害、退役軍人の地位、遺伝情報、または法律で保護されているその他の分類に基づいて人々を不当に異なる扱いや影響を与え、このようなアルゴリズムによる差別は法的保護に違反する可能性を示唆しつつ、自動化システムの設計者、開発者、配備者は、アルゴリズムによる差別から個人やコミュニティを保護し、公平な方法でシステムを使用・設計するために、積極的かつ継続的な措置を講じるものとする。この保護には、システム設計の一環としての積極的な公平性評価、代表的なデータの使用と人口統計的特徴に対する保護、設計と開発における障害者のアクセシビリティの確保、配備前および継続中の格差テストと緩和、明確な組織の監視が含まれる必要がある[501]。

③の「データ・プライバシー」では、個人の合理的な期待に沿うもので、厳密に必要なデータのみを収集したうえで、システムの設計者、開発者、配備者は個人からの許可を取得し、データの収集、使用、アクセス、移転、削除に関する個人の決定を尊重する。個人の同意を求める際は、簡潔で、平易な言葉で理解できる内容にし、健康や仕事などに関わる機微なデータについては、継続的な監視とモニタリングをつうじてより強い保護措置を講じるものとする[502]。

④「通知と説明」では、システムの設計者、開発者、配備者は、システム全体の機能と自動化が果たす役割、そのようなシステムが使用されていることの通知、システムに責任を持つ個人・組織、明確で適時かつアクセス可能な結果の計画を明確に説明する文書を広く一般に提供する。これらの情報は最新の状態に保ち、重要な使用例や主要機能の変更についてはシステムの影響を受ける人々に通知するものとする。自動化システムは、技

501） OSTP, Blueprint, *supra* (fn. 499), p. 23.

502） OSTP, Blueprint, *supra* (fn. 499), p. 30.

術的に有効で、利用者及びシステムを理解する必要のあるオペレータ等にとって有意義かつ有用であり、かつ文脈に基づくリスクレベルに適合した説明を提供しなければならず、これらシステムに関する要約情報を平易な言葉で記載した報告書、および通知と説明の明確性評価と質的評価を、可能な限り公表することも求める[503]。

⑤「人間による代替、考慮、予備的措置」では、システムから影響を受ける個人が必要に応じてオプトアウトし、人間による代替手段を選ぶことができるようにする。その適切性については、与えられた文脈における合理的予期に基づき、幅広いアクセス性を確保し、特に有害な影響から公衆を保護することに重点を置いて決定されなければならない。さらに、システムの失敗やエラーが起きた場合などに人間による考慮と予備的措置による救済を受けられるようにする。ここではアクセス可能で、公平で、効果的で、維持され、適切なオペレータの訓練を伴うべきであり、一般大衆に無理な負担を強いないようにすることが求められる[504]。

これら５つの原則は、自動化システムの構築、展開、ガバナンスにおいて、市民の権利を保護し、民主的な価値を促進する政策と実践の開発を支援することを目的とするものにとどまり、それ自体は拘束力を持たず、既存の法令、規制、政策、国際文書に取って代わるものでも、それを修正するものでも、その解釈を指示するものなく、一般市民や連邦政府機関に対して上記原則の遵守を義務づけるものではないとしている[505]。むしろ、Executive Order 13960, Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government (December 2020) という大統領令など、既存の政策や規制に従うことが原則であることを留保している。

まとめると、この原則自体には法的拘束力はなく、制裁規範も有していないため、これら諸原則を製造者等に課せられた（刑）法的な注意義務と

503） OSTP, Blueprint, *supra* (fn. 499), p. 40.
504） OSTP, Blueprint, *supra* (fn. 499), p. 46.
505） OSTP, Blueprint, *supra* (fn. 499), p. 2.

して援用するのは少々困難なようにも思われる。しかし、この「AI権利章典」が「青写真」の段階から実際に法的拘束力を有する「AI権利章典」へと昇華した場合には、ここで掲げられている諸原則がAI製品開発に関与する製造者に課せられる法的義務となり、注意義務の認定に資するものとなるだろう。

第3項 「新時代の人工知能倫理規範」（中国）

2021年9月25日に中国の「新世代人工知能のガバナンスに関する国家専門委員会」は、人工知能のライフサイクル全体に倫理を統合し、AI関連の活動に従事する自然人、法人、その他の関連機関等に倫理的なガイドラインを提供することを目的とした「新時代の人工知能倫理規範（新一代人工智能伦理规范）」（以下、「倫理規範」という）を公表した[506]。この倫理規範は、プライバシー、偏見、差別、公正性など、現在のコミュニティの倫理的懸念を十分に考慮し、テーマ別の調査、重点的な起草、協議を経て、一般規定、特定の活動に関する倫理規範、組織的実施事項に分類して作成された。倫理規範では、「人間の福祉の増進」「公正と正義の推進」「プライバシーとセキュリティの保護」「制御性と信頼性の確保」「責任の強化」「倫理意識の向上」という6つの基本的な倫理要件を定めると同時に、AIの管理、研究開発、供給、利用など特定の活動に対する18の具体的な倫理的要求事項を提案する。

このうち、AI製品に関与する主体に関連する義務としては以下のようなものが挙げられる。研究開発者に関しては、技術研究開発のあらゆる側面にAI倫理を統合することを率先して行い、意識的に自己検閲を行い、自己管理を強化し、倫理・道徳に反するAI研究開発を自制する意識の強化（10条）、データの収集、保存、使用、処理、伝送、提供及び開示の過

506） 中华人民共和国科学技术部・前掲（注297）。この倫理規範については https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html（最終アクセス2022年11月28日）で閲覧可能である。

程において、データ関連の法律、基準及び規範を厳守し、データの完全性、適時性、一貫性、標準化、正確性等データの品質向上（11条）、アルゴリズムの設計・実装・応用において、透明性、解釈性、理解性、信頼性、制御性を高め、AI システムの回復力、自己適応性、反干渉性を強化し、検証性、監査性、監督性、追跡性、予測性、信頼性の実現（12条）、データ収集やアルゴリズム開発において、倫理的審査を強化し、差別的主張を十分に考慮し、データやアルゴリズムの偏りの可能性を回避し、AI システムの普遍性、公平性、無差別性の実現（13条）の義務がある。

次に製造（販売）者については、市場参入、競争、取引などの活動に関する各種規則を厳格に遵守し、市場秩序を積極的に維持し、AI の発展に資する市場環境を整備し、データ独占、プラットフォーム独占などで秩序ある市場競争を損なうことを控え、いかなる方法によっても他の主体の知的財産権を侵害することを禁止することを目的とする市場ルールの尊重（14条）、AI 製品・サービスの品質監視と利用評価を強化し、設計や製品の欠陥などによる個人の安全、財産の安全、利用者のプライバシーの侵害を回避し、品質基準を満たさない製品・サービスの運用、販売、提供は行わないこと目的とする品質管理強化（15条）、製品及びサービスにおける AI 技術の使用について、利用者に明確に伝え、その機能と制限を明らかにし、利用者の情報及び同意の権利を保護することを目的とする利用者の権利・利益保護（16条）、緊急時のメカニズムや損失補償のスキームや手段を研究・開発し、AI システムを適時に監視し、利用者からのフィードバックに適時に対応・処理し、システム障害を適時に防止し、法律や規則に従って AI システムに介入する関連主体を支援し、損失を減らしリスクを回避する準備を整えることを目的とする緊急時の保護強化（17条）を規定する。

さらに使用規定として、製造者に対しては AI 製品とサービスの使用前のデモンストレーションと評価を強化し、AI 製品とサービスがもたらす利益を十分に理解し、すべてのステークホルダの合法的権益を十分に考慮

し、経済繁栄、社会進歩、持続可能な発展を促進するという善意の利用の促進（18条）や AI の倫理的ガバナンスの実践に積極的に参加し、関連するテーマに適時にフィードバックし、AI 製品やサービスを利用する過程で見つかった技術的な安全性の陥穽、政策や規制の空白、規制の遅れなどの問題解決を支援すること（21条）を、製造者のみならず利用者を含みうるものとして、AI 製品・サービスの適用範囲と悪影響を十分に理解し、関連する対象者の AI 製品・サービスを使用しない権利を効果的に尊重し、AI 製品・サービスの不適切な使用や濫用を避け、意図せずに第三者の正当な権利や利益を損なわないようにするという AI の誤用・濫用の回避（19条）、法令・倫理・基準・規範に適合しない AI 製品・サービスの利用の禁止、AI 製品・サービスの利用による違法行為の禁止、国家の安全、公共の安全、生産の安全を脅かすことの禁止、公益の毀損等の禁止という AI の違法な利用の禁止（20条）、AI 製品とサービスを安全に使用し、効率的に活用するために、AI に関連する知識を積極的に学び、運用・保守・緊急時の処理などに必要なスキルを率先して習得する（22条）ことを定める。

この「倫理規範」は他の AI に関する規制とは異なり、製造開発に限定せず、利用者も含め AI 製品に関与しうる主体の義務を明文化したものである。特に、AI 製品に起因する事故の刑事過失責任を検討する際に注意義務の確定の手掛かりとなるものとして、製造者ならば15条が設計上の義務・製造上の義務と、16条が指示上の義務、17条・21条が製品監視義務と調和する。そして、利用者の注意義務として19条・20条の内容が関連する。ただし、その具体的内容については条文上明確でないところは多いものの、これら法的義務をつうじて注意義務違反を認定することが望ましいだろう。もちろん、注意義務違反自体が刑事責任を生ぜしめるわけではなく、注意義務違反と結果の因果関係も求められる。その認定の障壁となりうる AI のブラックボックス性を可能な限り透明化する（＝説明可能な AI）ことも開発者に対して12条のように法的に義務づけることで、因果関係の

適切な立証に資するものとなり、AI 製品にかかる事故事例における刑事責任の負責が過度なものにもならず、間隙となる状況も防ぐことができるように思われるため、これは我が国においても非常に示唆に富む AI 規範となるだろう[507]。

第４項 「AI 開発ガイドライン」・「AI 利活用ガイドライン」（日本）

我が国における AI 開発・利活用に関する法的原則としては、総務省による「国際的な議論のための AI 開発ガイドライン案」[508]（2017年、以下

507） この倫理規範から AI 製品の事故事例における刑事責任を論じた中国の文献としては、曾粤兴・高正旭「论人工智能技术的刑法归责路径」治理研究（2022年第３期）113頁以下がある。同文献では、ネットワークへの依存度が高い AI には、ネットワークセキュリティに関する包括的な法規範が必要であるとし、AI を支えるアルゴリズムの安全性を先行法規のみならず刑法のレベルでも担保すべきという。そこで、AI（アルゴリズム）の安全性を保護法益とし、魏东「人工智能算法安全犯罪观及其规范刑法学展开」政法论丛（2020年第３期）を引用しながら「安全基準を満たさない AI 製品を設計・製造・販売・使用する罪、AI 武器を違法に設計・製造・所持・取引・運搬・使用する罪、AI 製品のアルゴリズムや使用方法を無断で改変する罪、AI 濫用罪、AI 騒乱罪」という５種類のアルゴリズムの安全に危害を加える犯罪の新設を提言するという試みがあり、さらに17条を参照しつつ、AI を供給する主体は、「緊急メカニズムや損害賠償計画・対策を検討・策定し、AI システムを適時に監視し、ユーザーからのフィードバックに適時に対応・処理し、システム障害を適時に防止し、関連主体が法律に従って AI システムに介入し、損失を軽減しリスクを回避できるように支援する準備をする」義務を負い、この義務に反してアルゴリズムの安全性を著しく侵害する行為を規制するために、刑法に新たな犯罪を創設する必要があるという（123頁）。

508） AI ネットワーク化の健全な進展及び AI システムの便益の増進に関する原則として、① 連携の原則：開発者は、AI システムの相互接続性と相互運用性に留意すること、主に AI システムのリスクの抑制に関する原則として、② 透明性の原則：開発者は、AI システムの入出力の検証可能性及び判断結果の説明可能性に留意すること、③ 制御可能性の原則：開発者は、AI システムの制御可能性に留意すること、④ 安全の原則：開発者は、AI システムがアクチュエータ等を通じて利用者及び第三者の生命・身体・財産に危害を及ぼすことがないよう配慮すること、⑤ セキュリティの原則：開発者は、AI システムのセキュリティに留意すること、⑥ プライバシーの原則：開発者は、AI システムにより利用者及び第三者のプライバシーが侵害されないよう配慮すること、⑦ 倫理の原則：開発者は、AI システムの開発において、人間の尊厳と個人の自律を尊重すること、そして主に利用者等の受容性の向上に関する原則として、⑧ 利用者支援の原則：開発者は、AI ↗

「AI 開発ガイドライン」という）及び「AI 利活用ガイドライン～AI 利活用のためのプラクティカルリファレンス～」[509]（2019年、以下「AI 利活用ガイドライン」という）が存在する。

このうち、本論文で想定してきた事例で関連するのは、(1) AI 製品の監視・管理の観点においては「AI 開発ガイドライン」の③制御可能性の原則と「AI 利活用ガイドライン」の①適正利用の原則であり、(2) AI のブラックボックス性への対応として「AI 開発ガイドライン」の②透明性の原則と「AI 利活用ガイドライン」の⑨透明性の原則、そして (3) 人間への責任帰属という観点からは「AI 開発ガイドライン」の⑨アカウンタビリティの原則、「AI 利活用ガイドライン」の⑩アカウンタビリティの原則である。以下、その具体的内容を確認する。

↘システムが利用者を支援し、利用者に選択の機会を適切に提供することが可能となるよう配慮すること、⑨アカウンタビリティの原則：開発者は、利用者を含むステークホルダに対しアカウンタビリティを果たすよう努めること、という 9 つの原則を定める。

509） ①適正利用の原則：利用者は、人間と AI システムとの間及び利用者間における適切な役割分担のもと、適正な範囲及び方法で AI システム又は AI サービスを利用するよう努める。②適正学習の原則：利用者及びデータ提供者は、AI システムの学習等に用いるデータの質に留意する。③連携の原則：AI サービスプロバイダー、ビジネス利用者及びデータ提供者は、AI システム又は AI サービス相互間の連携に留意する。また、利用者は、AI システムがネットワーク化することによってリスクが惹起・増幅される可能性があることに留意する。④安全の原則：利用者は、AI システム又は AI サービスの利活用により、アクチュエータ等を通じて、利用者及び第三者の生命・身体・財産に危害を及ぼすことがないよう配慮する。⑤セキュリティの原則：利用者及びデータ提供者は、AI システム又は AI サービスのセキュリティに留意する。⑥プライバシーの原則：利用者及びデータ提供者は、AI システム又は AI サービスの利活用において、他者又は自己のプライバシーが侵害されないよう配慮する。⑦尊厳・自律の原則：利用者は、AI システム又は AI サービスの利活用において、人間の尊厳と個人の自律を尊重する。⑧公平性の原則：AI サービスプロバイダー、ビジネス利用者及びデータ提供者は、AI システム又は AI サービスの判断にバイアスが含まれる可能性があることに留意し、また、AI システム又は AI サービスの判断によって個人及び集団が不当に差別されないよう配慮する。⑨透明性の原則：AI サービスプロバイダー及びビジネス利用者は、AI システム又は AI サービスの入出力等の検証可能性及び判断結果の説明可能性に留意する。⑩アカウンタビリティの原則：利用者は、ステークホルダに対しアカウンタビリティを果たすよう努める、と10の原則を定める。

AI 製品の監視・管理に関して、「AI 開発ガイドライン」の③制御可能性の原則は、開発者に対して「AI システムの制御可能性に関するリスクを評価するため、あらかじめ検証及び妥当性の確認を行うよう努めることが望ましい」とし、「こうしたリスク評価の手法としては、社会において実用化される前の段階において、実験室内やセキュリティが確保されたサンドボックスなどの閉鎖空間において実験を行うこと」、そして「制御可能性を確保するため、採用する技術の特性に照らして可能な範囲において、人間や信頼できる他の AI による監督（監視、警告など）や対処（AI システムの停止、ネットワークからの切断、修理など）の実効性に留意することが望ましい」という[510]。一方で「AI 利活用ガイドライン」の①適正利用の原則では、「AI サービスプロバイダー及びビジネス利用者は、開発者等からの情報提供や説明を踏まえ、AI を利活用する際の社会的文脈に応じ、AI を利用する目的、用途と AI の性質、能力等を適切に認識した上で、AI を適正な範囲・方法で利用すること」、さらに「AI サービスプロバイダーは、AI サービスの公平な条件による利用を確保するとともに、必要な情報を適時に提供することが期待される」[511]という。

AI のブラックボックス性への対応として、「AI 開発ガイドライン」の②透明性の原則では、「本原則の対象となる AI システムとしては、利用者及び第三者の生命、身体、自由、プライバシー、財産などに影響を及ぼす可能性のある AI システムが想定される」ことを前提に、「開発者は、AI システムに対する利用者を含む社会の理解と信頼が得られるよう、採用する技術の特性や用途に照らし合理的な範囲で、AI システムの入出力の検証可能性及び判断結果の説明可能性に留意することが望ましい」[512]とし、「AI 利活用ガイドライン」の⑨透明性の原則では、「AI サービスプ

510） 総務省「国際的な議論のための AI 開発ガイドライン案」（2017年） 8 頁以下。

511） 総務省「AI 利活用ガイドライン～AI 利活用のためのプラクティカルリファレンス～」（2019年）13頁。

512） 総務省・前掲（注510） 8 頁。

ロバイダー及びビジネス利用者は、AI の入出力等の検証可能性を確保するため、入出力等のログを記録・保存すること」、「ログの記録・保存に当たっては、利用する技術の特性及び用途に照らして、ログの記録・保存の目的、ログの取得・記録の頻度等について考慮することが期待される」という[513]。

そして、AI 製品にかかる責任の観点で、「AI 開発ガイドライン」の⑨アカウンタビリティの原則では、「開発者は、AI システムへの利用者や社会の信頼を得られるよう、自らの開発する AI システムについてアカウンタビリティを果たすことが期待される。具体的には、利用者に AI システムの選択及び利活用に資する情報を提供するとともに、利用者を含む社会による AI システムの受容性を向上するため、開発者は……利用者等に対し自らの開発する AI システムの技術的特性について情報提供と説明を行うほか、多様なステークホルダとの対話を通じて様々な意見を聴取するなど、ステークホルダの積極的な関与（フィードバック）を得るよう努めること」、そして「自らの開発する AI システムによってサービスを提供するプロバイダー等と情報を共有し、協力するよう努めることが望ましい」[514]といい、「AI 利活用ガイドライン」の⑩アカウンタビリティの原則では「AI サービスプロバイダー及びビジネス利用者は、人々と社会から AI への信頼を獲得することができるよう……消費者的利用者、AI の利活用により影響を受ける第三者等に対し、利用する AI の性質及び目的等に照らして、それぞれが有する知識や能力の多寡に応じ、AI システムの特性について情報提供と説明を行うことや、多様なステークホルダとの対話を行うこと等により、相応のアカウンタビリティを果たすよう努めることが期待される」[515]とする。

これらガイドラインの策定以来、国際的に新たな AI 開発や利活用に

513） 総務省・前掲（注511）24頁。
514） 総務省・前掲（注510）11頁以下。
515） 総務省・前掲（注511）25頁。

関するガイドライン・綱領が策定されている状況に鑑みて、我が国でも新たなガイドラインの策定の議論がなされている[516]。そこでは22の尊重すべき価値[517]を確認しており、2019年当時のガイドラインの国際比較と比べて、堅牢性、責任[518]、追跡可能性[519]、モニタリング・監査、ガバナンス、その他（コスト・効果測定）の６つの新たな項目の追加が検討された。

これらガイドラインの規定で、透明性・説明可能性・追跡可能性はいわゆる説明可能なAIの構想と、安全性・堅牢性・制御可能性は製造者に対する設計・構造上の義務と、適正な利用は製造者の指示上の義務及びエンドユーザーを除く利用者に課せられる義務と[520]、モニタリング・監査は製造者の製造監視義務と、それぞれ通底する。この点、第２章で言及した

516） 総務省AIネットワーク社会推進会議「報告書2022～『安心・安全で信頼性のあるAIの社会実装』の更なる推進～」（2022年）１頁以下参照。

517） 総務省・前掲（注516）48頁では、１．人間中心、２．人間の尊厳、３．多様性・包摂、４．持続可能な社会、５．国際協力、６．適正な利用、７．教育・リテラシー、８．人間の判断の介在・制御可能性、９．適正な習（学習データの質）、10. AI間の連携、11. 安全性、12. セキュリティ、13. プライバシー、14. 公平性　15. 透明性・説明可能性、16. アカウンタビリティ、17. 堅牢性、18. 責任、19. 追跡可能性、20. モニタリング・監査、21. ガバナンス、22. その他（コスト、効果測定）である。

518） ここでの意味は、総務省・前掲（注516）13頁（別冊１）によると、他国のAIガイドラインを参照しつつ「ステークホルダにはAIが適切な条件下で、適切な訓練を受けた人々によって使用されることを保証すること」や、「AIに基づく意思決定が、誰の健康や安全にも脅威を与えないこと」を挙げており、AI技術によって問題が発生した際のステークホルダのアカウンタビリティ（説明責任）とは区別される。

519） 総務省・前掲（注516）13頁（別冊１）によると、透明性を保障する意味で「AIに基づく意思決定に影響を与えたデータや意思決定の根拠を追跡できること」を根底に置くと考えられる。

520） 総務省・前掲（注516）７頁（別冊１）によると、他国のAIガイドラインを参照しつつ、「個人は、どのようなデータが収集され、何のために、どのような状況で使用されるかについて自分自身でコントールできるようにするべきである」や「人間は、意思決定と行動を自律的システムに委ねるかどうか、いつ、どのように委ねるかを選択し、適正に利用しなければならない」ということを例示している。これらは日本「AI開発ガイドライン」・「AI利活用ガイドライン」では言及されていないエンドユーザーの利用者までを視野に入れたものであることに留意しなければならない。

開発製造者の刑事過失責任を論じるにあたって、その根拠となる注意義務違反を認定する際の注意義務の内容は、現行の製造物責任法や車両運送法68条 9 項（車両のリコール義務）のような法律上の義務のみならず、法的期待状況から導くことも許されうる。そう考えると、本ガイドラインは AI 製品の開発製造者に課せられる義務内容をより具体化するものとすれば、この行政法規たるガイドラインに記載される原則をもって義務を確定させることができるともいえるだろう。しかし、これら原則を遵守するための実効性を伴う取組[521]が法的義務という形で存在するのであれば、仮に何らかの（刑）法的問題が生じた場合でもその法的根拠や意味内容の解釈で開発製造者のみならず、エンドユーザーを除く利用者を困惑させることは少なくなるだろう。これこそが法的観点から見た「安心・安全で信頼性のある AI の社会実装」[522]を実現する一助となるし、この見地は中国の「倫理規範」が大いに参考となるだろう。

第 3 款　小　　括

本節で見た2010年代後半から2020年代初頭にかけての国内外の AI 開発や利活用に関する法規則・ガイドラインでは、強い規制を伴うものや法的拘束力のない諸原則にとどまるものなど様々な性格を有するものがあった。特に開発規制の観点から見れば、将来に対する規制を意味することになるため、欧州 AI 規制案のようにその原則で直接に制裁（罰則）的規制を設けるのは、技術開発・販売流通を委縮させてしまう可能性を秘めている以上首肯しかねるところである。制裁や罰則はその規制の実効性を担保するための最終手段であるから、可能な限り最小限度の事例に留めるべきであり、むしろ実効性を担保するならば**説明可能な** AI の構築や外部機関

521）総務省・前掲（注516）62頁によると、事業者自身による取組のみならず、リスクを洗い出すフレームワークの構築や外部機関によるモニタリングの仕組みの整備、チェックシートの策定や認定制度の創設があげられている。

522）総務省 AI ネットワーク社会推進会議のテーマである。総務省・前掲（注516）1 頁参照。

による監視制度、行政機関による認定制度の創設を先行すべきであると思われる。ただし付言すると、米国や日本のように、もっぱらAI製品の製造者やエンドユーザーを除く利用者のようなステークホルダを念頭に置いた法規になっていることに留意しなければならない。AI製品に関与する主体にはステークホルダのみならずエンドユーザーたる利用者も含まれるし、この利用者に対する適正な利用を求めることも忘れてはならない。そこで参考となるのが中国の倫理規範であり、ステークホルダとエンドユーザーすべてを包含する法規を策定することこそが重要である。その一方で、法律によって定められる義務となることがその実効性を確実なものにすることを可能にし、「安心・安全で信頼性のあるAIの社会実装」を実現することができるだろう。

第3節　刑法上の保護

「AIのための刑罰」という表現が持ちうる意味は、AIを構築し、それを利用する人々に対処しなければならないという示唆を含む。例えば、選挙の不正操作や道路交通違反を引き起こす可能性のあるAIの利用について当てはまる。すなわち、ロボットもしくは実体を持たないAIを道具として利用する人間は、故意または過失によって惹起された刑法上の結果について、犯罪への関与者として刑事責任を負う可能性があることはこれまでの検討から明らかである。例えば、自動運転車の利用者、所有者もしくは製造者について、原則として、定められた注意基準を顧慮しなければ注意義務違反を認定することはできる。この場合、関与者にいかなる注意基準を課したいのか、また、帰責主体がどのようなものになるのかということが差し迫った問題となっていた。

例えば、自動運転車の領域では、上記各主体に課せられる義務の明文化が議論されていた。これにはレベル4の自動運転車の公道走行について定めた2021年ドイツ道路交通法改正や第2章第4節第1款で言及した2022年日本道路交通法改正が鍵となる。そのポイントとして、運転者なしでも所

定の運行領域を独立して運転することができること、自然人の技術監督者を置かなければならないこと、そして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、各人の法益の重要性を考慮する。生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付けをしないことが明文化された。

しかし「AI の創る者のための刑罰」を将来のために検討すべきことも否定されないだろう。研究者が故意はないものの、それに対応する兆候に反して、例えば身体を侵害するような効果のあるような危険な AI を作成または公開した場合、過失の可罰性を帯びるかという疑問がある。このとき、将来的に妥当するだろう AI 研究の注意規制が尊重されない場合、予見可能性はすでに最終的に侵害する事象の詳細な予見を要しないため、予見可能性の欠如から研究者は少なくとも免れられない[523]。ただし、研究者は、自己学習技術の効果を限定させることが難しい、もしくは研究者はこの知見を考慮に入れなければならない、ということについては承知しているはずである。さらに、極めて性急な刑法の投入を主張するわけではないが[524]、しばしばボーダーラインとなる過失責任を超えて、将来のロボット、ないしは AI の単なる開発が部分的に刑法上制限されるおそれもある。しかし、法を危険にさらすような AI の作成ないしは普及の将来的な禁止は、その有害な投入もしくは刑法上の態度規範としての AI の所為に先立ち、例えば AI 兵器システムの文脈で真摯に考慮に入れられるが、

523） ここでは客観的帰属に限定されないであろう広範な意識なき過失の予見可能性基準については、BGHSt 48, 34, 39; BGH NStZ 2001, 143, 144 f. や本稿第 2 章第 2 款第 4 項を参照。その他に *Joerden,* a.a.O. (fn. 267) S. 195, 207 ff.; *Gless/Weigend,* (fn. 153), 561, 581 f.

524） とりわけ刑法の投入が早まることへの批判については , 例えば包括的な *Puschke,* Legitimation, Grenzen und Dogmatik von Vorbereitungstatbeständen, S. 49 ff 137 ff. 現在でも、刑法30条、159条の例外を除き、実用性に乏しい準備罪がいくつかある。ドイツ刑法の領域では、例えば、コンピュータ刑法では202c条、テロリズムの分野では、89条a、89条b、財産刑法における265条、医事刑法における217条が、関連する問題として184条 i がある（*Gaede* a.a.O. (fn. 472), S. 82 Fn. 280）。日本刑法では、組織的な犯罪の処罰及び処罰収益の規制等に関する法律 6 条の 2 がこれに該当する。

そのための検討事項が２つ挙げられる[525]。

法を危険にさらす機械というアクターの創造から法を守るための規範は、それ自体が厳格な犯罪化の基準を満たすことができる。すべての人間の利益のために、法とそれによる最高度の個人法益の保護[526]を信頼することができるようにする。ただし、未知の AI の所為に対する研究者の処罰は、そもそも原理的に不確実なものであるため、特にその前倒しは問題となることに留意しなければならない。

次に、制裁で強化された禁止をテーマ化するのはもはや尚早なものでないと考えられる。刑事立法は、基本的かつ、時間的にはまだ切迫しない徹底した議論に基づくべきであり[527]、それに加えて現在の AI のユーフォリアの中では、考慮すべき危険性について慎重に公式化された際には、単なる象徴的に留まらない法的根拠を早い段階で設定する必要がある[528]。今後数十年の間に危険な AI の萌芽が出現するか否かは断言できないものの、世界規模の AI 研究の開示義務を負わない状態をあらゆる国家や社会システムを超えて細部まで看破することは困難である。また、ある一国の意見は世界のすべての研究所に到底届くわけではないので、規制は早い段階で始めなければならず、それには、AI 研究の有害な継続に対する世界規模での規制が理想的である[529]。そうでなければ、例えば AI 兵器が一度本格的に利用されうるようになれば、それを持たざる国家は虚しくも国際的に AI 兵器の排斥を要求することになるだろう。

ただし、私見としては、実例の乏しいものに対する「抽象的な可能性」

525） *Gaede,* a.a.O. (fn. 472), S. 82 ff.

526） 特別な生命と健康の意味につき、たとえば注意義務の決定の文脈においては *Gless/Weigend,* a.a.O. (fn. 153) S. 561, 584 f. も参照。

527） この文脈ではすでに *Sehuhr,* Willensfreiheit, Roboter und Auswahlaxiom, in: *Hilgendorf/Beck,* a. a. O. (fn. 8), S. 43 f.; *Simmler/Markwalder,* Roboter in der Verantwortung?, ZStW 129 (2017), S. 20, 22. で指摘がある。

528） *Gaede,* a.a.O. (fn. 472), S. 84.

529） *Gaede,* a.a.O. (fn. 472), S. 84.

の枠組での規制には慎重になるべきであると考える。なぜなら、これでは具体的危険のみならず、抽象的な危険で制裁が、ともすれば刑法の枠組における刑罰が発動されることになってしまうからである。そのため、立法上で AI 研究も含めて、一定の基準を設け、たとえば、許可義務・届出義務・免許制度などを仔細に設定して、各々の主体に課される義務を明確化すること[530]により、従前よりも望ましい展開を期待することができるのではないだろうか。もちろん、AI は非常に多彩かつ多岐にわたるものであるから、この文脈で刑法が実際にどこまで進むべきか、意味のある効果を発揮するためにそのような規範を技術に配慮した形でどのように策定すべきか、国際的にどの程度まで法を実現できるのかは、今後の法研究や政策的な助言によってまとめ挙げられるべきものであから、今後の AI 規制に係る国内外の動向には常に注視しなければならない。

おわりに

本論文では、AI の利活用における刑法上の諸問題というテーマで、主に AI 製品に関与する主体である製造者と利用者を中心に、AI 製品が人間の生命・身体・財産を侵害した場合、さらには経済犯罪を遂行した場合や AI 製品がサイバー攻撃を受けた際の刑法上の評価について検討を行った。

その際、特に先行研究の蓄積がある AI を搭載した自動運転車の事故に伴う人間の生命・身体を侵害した事例での検討において論者によって想定する AI の定義が確定していなかったことを端緒に、その刑法上の問題を論じる前にまずは AI の研究史にさかのぼって AI の定義を確定させるこ

530） 総務省・前掲「AI 開発ガイドライン」（注510）や「AI 利活用ガイドライン」（注511）では求められる諸原則やその実効性を担保する法創設が提案されているものの、それがステークホルダに限定されており、エンドユーザーにまで及ばないことは前節で指摘したとおりである。

とを試みた。しかし、その当時から AI の定義には困難を伴っていたことから、現存する AI の現象形態から帰納する形で、特定のタスクの遂行に特化し自律的判断する能力を有しないいわゆる「弱い AI」をこれら刑法上の問題を論じる上で対象とすべきであることが確認された（第１章）。

その上で、第２章では本論文の主題でもある AI 製品の事故に伴う人間の生命・身体の侵害事例を検討した。そこでは、道交法上の義務が創設された自動運転車の事故事例と、いまだ法律上の義務の存在しない、例えば介護用ロボットや産業用ロボットの利用中の事故事例とに分類して検討を行った。自動運転車の場合は、レベル４ではその自動運転車の運行に関与する主体の義務が仔細に規定されている一方で、レベル３については画面注視に係る義務と点検整備に係る義務のみが、さらにレベル２以下では普通自動車の運転手と同様の義務内容が条文の解釈上課せられる。しかし、レベル２の自動運転車の事故に関する判例の検討スキームにも見られるように、およそ普通自動車の操作と同一とはいえないレベル２の自動運転車にも普通自動車と同様の運転手の義務が課せられるとするのは不適切だろう。この点については、レベルに即した運転手の義務づけが必要になると思われる。

それとは対照的に、法律上の義務の存在しない AI 製品の事故事例に関しても、差し当たっては自動運転車の事例と同様に、その AI 製品に関与する主体たる人間に刑事責任が帰属しうると考えるべきである。2010年代から盛んに議論されてきたこのテーマでは、事故に至る動作をする判断をしたのが人間の判断ではなく AI の学習による自立的判断であることや、AI の学習経過がブラックボックス化して人間の判断と事故結果の間の因果関係が遮断される可能性などを考慮すると帰属主体が存在しなくなることが指摘されてきた。これに対し、AI 自体に刑事責任を帰属させようとする試みや製造者に対する厳格責任モデルを用いようとする試み、不規制によって刑事責任の帰属を考慮しないとする試みなどが考察されたが、いずれも伝統的解釈から外れたものとなりうるし、新たな刑罰モデルの構築

を必要とするが、AI 自体への刑事責任帰属では、本来であれば刑事責任が帰属されるべき主体が AI を隠れ蓑にして責任帰属から逃れる可能性が否定できなかったり、厳格責任モデルでは製造者に過度な負担を課したりすることになりかねず、AI 製品に関与する主体間でアンバランスな解決に至ってしまう。そこで、自動運転車のように、他の AI 製品に関しても一定の法的義務に基づいて AI 製品の関与者の義務を確定すべきである。例えば製造者に対しては、製造物責任法下で課せられる製造上の義務、設計上の義務、指示・警告上の義務（製品監視義務）を手掛かりに、刑法上の製造物責任を検討すべきである。ここで留意すべきは、製造者はこれら義務に違反したからといって直ちに刑法上の過失を構成するのではなく、その義務の内容の保護目的に従って、生じた結果との因果関係の有無を慎重に検討することである。製造者側で関与する技術サービスプロバイダーや許可責任者たる国家・地方公共団体に対しても、当該 AI 製品の供給・流通に関する一定の明文の義務づけをして、その義務内容の保護目的に従って、生じた結果との因果関係の有無を検討するというスキームが望ましい。その一方で、利用者や所有者に対しては製造者側の指示を遵守し、これを悪用・濫用しないようにするという一定の義務づけも必要とされるだろう。多くの AI ガイドラインや法規ではこの観点があまり考慮されていないが、実際に AI 製品を使用するのは利用者や所有者といったエンドユーザーなので、これら主体に対してこうした義務づけを製造者側と並置する形で明文化することもまた重要である。もちろん、結果帰属の検討の際にはその義務内容の保護目的に従って、生じた結果との因果関係の有無を検討することは上記と同様である。

しかし、過失犯処罰規定のない経済犯罪の場合は上記スキームで検討するには困難を伴う。そこで、第 3 章第 2 節では実体を持たない AI 製品である AI・アルゴリズムが、その学習の結果、利用者の知らないところで相場操縦、インサイダー取引のような証券犯罪、価格の協調的行為のような競争法違反を遂行した場合について検討した。証券犯罪では、まずその

刑罰の根拠となる金融商品取引法上の解釈が問題となる。相場操縦では、たとえ利用者の知らないところで相場操縦的取引が遂行されたとしても、客観的事実に基づいて利用者の故意が推定されるといった実務上の運用があるため、相場操縦行為が認定される可能性が否定できない。そうすると、利用者は常に相場操縦の可能性を考慮しながらAI・アルゴリズムを利用することになるが、これでは利用上も法的にも過度な負担になりうるし、開発普及を阻害する結果になりかねない。だからこそ、相場操縦でないと認められるに足るシステム構築が製造者には求められ、その判断プロセスを明確にする構造――説明可能なAI――が重要である。また、インサイダー取引では未公開重要事実を知った利用者がその事実を利用したか否かは問わず証券等の取引が行われ、もって利用者に一定の利益がもたらされた場合には形式的にインサイダー取引に該当するが、適用除外要件に該当する限りでインサイダー取引は成立しないという構造となっている。AI・アルゴリズムの利用促進、開発普及の観点から、この場合も適用除外要件に適合する形でのシステム構築が製造者には求められ、そこでもその判断プロセスを明確にする構造――説明可能なAI――が重要である。それに対して、学習するAI・アルゴリズムによりその利用者間（競争事業者間）での価格協調が実現した場合は、独占禁止法の解釈が関わってくる。判例上は黙示による競争事業者間の「共同性」も認定していることから、たとえ上記の場合でも不当な取引制限に該当しうるが、そもそもAI・アルゴリズムの学習により価格協調行為が遂行される可能性は現状では低いことや、「不当な取引制限」に対しては排除措置命令、課徴金、刑事罰の可能性が規定されていることを考慮すれば、「共同性」要件の認定には特に慎重になるべきであるし、「不当な取引制限」に該当するとすべきでもない。この経済犯罪における先行研究で目立っていたのは、刑事罰の可能性が予定される類型であるにもかかわらず、利用者の処罰を制限する論調ではなかったことである。このことはAIやAI学習特有の問題ではないが、刑事罰を発動する可能性を秘める以上、その構成要件の認定

を慎重にする姿勢が必要であると考える。

AI 製品が行為客体になる場合、すなわち AI 製品がサイバー攻撃を受けて利用者の情報が取得されたり、その内部データの変更・破壊により製品利用を妨げられたりした際の刑法上の評価では、構成要件の問題と AI の学習のブラックボックス性の問題に大別される。

構成要件の問題では、まず AI 製品にハッキングする行為が必ずしも不正アクセスを構成しないこと、AI 製品内に記録された利用者情報の取得する行為は特別刑法上の問題であること、そして内部データの変更・破壊の場合ではその製品利用に業務性があり、それが業務妨害結果に至ったという限りで電子計算機損壊等業務妨害罪が成立するにすぎないことである。それゆえ、私的空間に属する AI 製品の内部データの変更・破壊に関しては、それがアクセス制御機能を有している限りで不正アクセス罪のみが成立するという帰結になる。さらに、考慮すべきもう一つの事例として、資産運用を行う AI ソフトウェア・エージェントに対して不正なデータが用いられ、結果として AI ソフトウェア・エージェントの利用者に対して財産的損害が発生したものである。ここでは電子計算機使用詐欺罪の成否が検討されるが、条文解釈上行為者に対して不当利得を要求するため、本罪の成立は認められず、その AI ソフトウェア・エージェントの利用が利用者の業務に属し、財産的損害が利用者の業務を妨害したといえる限りで電子計算機損壊等業務妨害罪が成立する。ここには、電子計算機損壊等業務妨害罪の成否と同様の問題があることに注意しなければならない。

AI の学習のブラックボックス性の問題は、電子計算機損壊等業務妨害罪では、その損壊もしくは虚偽の情報または不正な指令の供与の原因が行為者によるものなのか、AI の学習によるものなのかが不明であった場合が、電子計算機使用詐欺罪では、財産権の得喪もしくは変更に係る不実の電磁的記録の原因たる虚偽の情報もしくは不正な指令、ないしは財産権の得喪もしくは変更に係る虚偽の電磁的記録が行為者によってもたらされた

のか、それともAIソフトウェア・エージェントの学習によってもたらされたのかが不明な場合にある。両者とも、その因果関係が不明確な場合はたとえ利用者の業務が妨害されたとしても、行為者が不当利得を得た結果が生じたとしても、行為者には未遂罪が成立するにすぎないという帰結となってしまう。ここでも重要なのが説明可能なAIの構想であり、本来ならば刑事責任が帰属されるべき行為者（攻撃者）がAIの学習を理由に未遂減軽の可能性を残してしまう状況を防ぐことが実現できる。

第4章では、2010年代後半から各国で進められている将来的なAI開発の指針・規制は、2020年代に入るとその内容に変化を見せていることを確認した。その規定ぶりは強い制裁規範を持つものから、法的拘束力を持たないガイドラインにとどまるものまで様々であるが、将来の開発に対する規制を論じる際には、いまだ具体的な危険のない状態で過度な制裁を課すことは、将来的なAI開発を委縮させる効果を招来することに注意しなければならない。もちろん、人権を侵害するような開発に関しては規制の対象とすべきであるが、現状の技術水準を考慮すればそのような開発が表立って行われているわけではないので、直ちに強い規制を必要とするわけではない。しかしAIを搭載した製品は日々進歩を続けており、数多くの人間の主体がAIと関わるようになっているので、これら主体が遵守すべき原則、課せられる義務を具体的に作成すべき時期に差し掛かってきているように感じる。その実効性を担保するための許認可、監査制度などのソフトな措置から創設し、エンドユーザーたる利用者の利益と製造者側の負担とのバランスを考慮しながら、AI製品を取り巻く主体が遵守すべき法律上の原則・義務を創設することが、これからのAI研究開発、ひいては販売流通・利活用にとって不可欠なものである。

残された課題として、第1章や第4章第2節第2款で言及した各国のAIに関する法規で述べられた、①AI製品の利用者のデータ保護およびその第三者利用に関する刑法上の観点、②ディープウェブ内でAIを用いたプラットフォーム事業者の刑法上の責任、③国際的な証券取引のレ

ベルでの経済犯罪が遂行された場合の刑法上の解釈である。①について日本では2022年個人情報保護法改正により罰則規定が厳罰化されたことも踏まえ、現在の喫緊の課題ともいえる。②・③については先行研究[531]が本年になって立て続けに刊行されているが、ネットワークによってグローバル化した社会の中で新たに考慮される重要な課題といえる。これらについては、別途検討を行いたいと思う。

531） *Grimm,* Das Insiderhandelsverbot zwischen Rechtstheorie und Rechtspraxis, Nomos 2022; *Weber,* Die Strafbarkeit von Plattformbetreibern im Darknet, Nomos 2022.