「超人」としてのAI・ロボット

大屋雄裕　慶應義塾大学法学部教授

1．動物裁判と行為指導性

中世ヨーロッパには「動物裁判」と呼ばれる社会制度があったことが知られている。たとえば疫病が流行したとき、その原因と疑われた動物――典型的にはネズミやネコ――が数匹捕らえられ、法廷で有罪を宣告されて処刑されたり、不幸にも赤ん坊を蹴り殺してしまった豚が追放刑に処されたりしたという（池上俊一『動物裁判：西欧中世・正義のコスモス』講談社、1990年）。さて我々はこの制度を正当なもの、意味のあるものだと考えるだろうか。考えないとすれば、それはなぜなのだろうか。

おそらくただちに現われる反応は科学的な合理性がないとか、このようなことをしても動物が一定の行為（たとえばネズミによるペストの媒介）を止める見込みがない以上、意味はないというものだろう。法制度には行為指導性があり、一定の行為が処罰されるだろうという予測をもとにして我々は、そのような行為に関与することを事前に避けることができる。動物にはこのような予期能力がないので、違反行為を事後に処罰することを通じて学習させるしかない。犬猫のしつけとはこのように事後処罰に頼ったシステムであり、だからこそベンサムは何が禁止されているのかを個々人が理解することのできないようなイングランドの判例法システムを「犬法」（dog law）と罵倒したのだった。

逆に言えば、同じヒト（homo sapiens）に属していても一定の理性や判断能力を持たない存在を法の指示に服従させることはできないし、彼を法廷に引き出して有罪を宣告するようなことも――動物裁判の例と同じように――無意味だということになる。我々はもちろん、心神喪失者が刑事的に処罰されないことを定めた刑法39条1項を、ここで思い出すことだろう。法を通じた事前規制は、結果を予期し自らの行動をコントロールすることのできる自律的主体に対してしか意味を持たないと、とりあえず確認

しておこう。

◆

2．人と物と境界線と

世界を「人」と「物」とに区分し、理性的主体としての前者が合意に基づいて相互に形成する関係として契約をとらえようとしたサヴィニー以降のドイツ民法学（日本の民法もその延長線上にある）、あるいはその前提にあったカントの哲学・倫理学もまた、同様の観点に立っている（筏津安恕『私法理論のパラダイム転換と契約理論の再編：ヴォルフ・カント・サヴィニー』昭和堂、2002年）。理性に基づいて自律的に行為することのできる存在だからこそ我ら人間は人として根元的な自由を認められるのであり、そのような資格を持たない存在はすべて――生命の有無にかかわらず――物として人の意思に従属させられることになるわけだ。

動物の権利をめぐる生命倫理学的な問いも基本的にはこの構図を継承している。たとえばある種の動物はヒトに近い知性や自律性を持つものとして限定的な権利能力を認められるべきだとか、逆に胎児や重度心身障害者は意識と受苦能力を欠くので生命権のような人権保障の対象にならないという主張はしばしば見受けられるだろう。その具体的な主張内容はさまざまであっても、これらはいずれも意思と自律性を持つ人を高く、それらを欠く存在を低く位置付けるスロープの存在を前提にしている点ではほぼ違いがない。どこにその両者を区切る境界線があるのか、あるいは境界線は単一でなくいくつかの段階から構成されているのか――たとえば犬猫には能動的な政治参加の権利はなくとも理由なく傷付けられない権利は認められるべきかもしれない――、その境界線と生物種としてのヒトの境界が一致しているのかといった論点をめぐってどれだけ激しい論争を繰り広げたとしても、我々の社会を構成する人の特権性が、そこでは常に前提されているのである。

ここで問題は、いま新たに我々の世界に生まれつつある存在としてのAIやロボットがこの構図のなかでどのように位置付けられるかという点にある。もちろん我々は、これまでの生命倫理学と同様の構図に立ったうえで境界線をめぐる議論に基づいてその地位について考えることができる。人の条件を理性や判断能力に求めればAIやロボットも我ら人間と同じ高みに昇るべきだということになるかもしれないし、痛みや苦痛を感じる能力だと考えれば動物より低い位置付けも考えられるだろう。だがその

ような議論の構図は、AIやロボットの本質的なあり方を正確に反映しているのだろうか。AIやロボットは、まだ十分に技術が発展していない現時点では動物のようなものであり、成長するにつれてヒトと同じような存在へと移行していくだけのものなのだろうか。

◆

3．不透明な存在としての人間

少し遡って考えよう。さきほど法制度には行為指導性があり、処罰が予告された行為を人々は選択しなくなるだろうと述べた。だがそこでいう行為指導性がそれ自体として実在するのならば、処罰の予告を伴う必要はないのではないだろうか。たとえば「殺人は悪いことであり、禁止される」とだけ述べればよく、それに一定の制裁――現行法では死刑または無期もしくは5年以上の懲役（199条）――を結び付ける必要はないのではないだろうか。

もちろんこの問いへの答はごく明らかであり、我々が規範に従いそこねるという点にあるだろう。我々の多くは、あえて刑法の条文によって示されることがなくとも、殺人が悪い行為だということは十分に理解している（それが自然犯mala in seということの意味であった）。だとしても一定の理由からその悪をあえて選択したり（あいつだけは生かしておけない）、あるいは自らの行動を理性的にコントロールしそこねることによって（ついカッとなって）、悪いと知っているはずの行為に手を染めてしまうわけだ。そこで我々は、制裁を予告することで合理的な判断者にとってのバランスを変えてしまおうとする一方（処罰されることを考えれば、相手を殺すのは割に合わない）、激情に駆られた行為者に対しては実際に処罰を科すことで、他の人々への戒めにしようとすることになる。いずれにせよ法は、それが我ら人間の判断や行為に直接的に介入できないこと、判断や行為の条件を操作することで間接的に機能することを狙うしかないということを前提としている。我々は再び、法は外形的な行為を規制する合法性の次元においてのみ働くものであり、内心にある道徳性を左右することはできないというカントの指摘を思い出すだろう。だからこそ、物理的な条件を通じて判断・行為の可能性自体を事前に消去するアーキテクチャの権力が法を超えるものとして注目を集めることになったということにもなるはずだ（ローレンス・レッシグ（山形・柏木訳）『CODE：インターネットの合法・違法・プライバシー』翔泳社、2001年、

および大屋雄裕『自由とは何か：監視社会と〈個人〉の消滅』筑摩書房、2007年など）。

◆

4．超人としてのAI・ロボット

このように考えたとき、AIやロボットがヒトと大きく異なる点として、そのような従いそこねの可能性に注目することができるだろう。まず、現在でも生産現場で活用されているようなロボットを考えよう。それらはあらかじめ定められたプログラムに沿って定められた動作を反復し続けるだろうし、故障や燃料切れといった物理的な障害の場合を除けば、それに失敗することもないだろう。プログラムはロボットの動作を直接的に規定するのであり、そこには判断も自律も、したがって従いそこねの問題も生じないように思われる。

まさに最近話題となっている自動運転車のように、高度な学習機能を備えたAIや、それによって制御されるロボットの場合にはどうだろうか。もちろんそこに、与えられたデータからAIが何を学習するかが予測しにくいとか、データ自体に偏りが含まれていればそれをAIが忠実に学習してしまうという問題は指摘されている。たとえば2016年3月には、米マイクロソフト社がTwitter上で公開した会話AI「Tay」が、ユーザーとの会話を通じて人種差別や陰謀論をたちまち学習してしまった結果として不適切な発言を連発するようになり、半日あまりで緊急停止されるという事態が生じている。AIによる不法行為や暴走への懸念などが語られるひとつの契機にはなったのかもしれない。

だが注目すべきなのは、ここでたとえば人種差別を学習したAIがただちにそれを実行に移していることだろう。AIは差別発言がいいことだと思ったからそれを直接に行動へと反映させたのであり、そこにはヒトの場合であればしばしば生じるようなさまざまな配慮（この発言は社会的に許容されるか、TPOに合致しているか）やためらい（面倒くさい、眠い、疲れた）は存在しない。学習結果はAIの行動を直接的に規定しており、ヒトのような従いそこねの可能性はここにも存在しないということになるのではないだろうか。AIやロボットが我ら人間とは異なる「超人」的なあり方を実現するものだとすれば、それはたとえば理性、知識量、判断能力、情報処理速度といったもので我々を大きく凌駕するような「超知性」（superintelligence）だからではなく――あるいはそれに加えた別の問題として――、このように我々とは根本的に異なった規範

への反応構造を持っているからだと考えられる。

◆

5．法とその対象

そしておそらくはこのことが、AIやロボットを規制する法のあり方にも反映することになるだろう。冒頭で挙げたように、我々は自律性を持たない動物に対して法による規制は無意味だと考えるのであった。ではAIやロボットに対しては、どうだろうか。判断過程に不透明性・間接性がなく、何を考慮してどのように判断すべきかを（あるいは逆に最低限このような判断をしてはならないと）命じればその通りに行為するだろうという意味においては自律性を持たないAIやロボットに対して、法はやはり意味を持たないのではないだろうか。あるいは別の言い方をすれば、AIやロボットはそれら自身が規範の名宛人となることはないのではないだろうか。我々が規範を投げかけるのはあくまで、彼らが従うだろう指示の作り主たる我ら人間に対してなのではないだろうか。

現在、世界各所で取り組みが進められているAI・ロボットに関する規制が、開発方針のコントロールに主眼を置いていることは――まだ高度の学習能力を実現したAIなどが現物として存在しないからという理由もあるだろうが――おそらく偶然ではない。国際的な研究支援団体であるFuture of Life Instituteによる「アシロマAI原則」（Asilomar AI Principles; https://futureoflife.org/ai-principles/）にせよ、日本で総務省AIネットワーク社会推進会議がとりまとめたAI開発ガイドライン案（http://www.soumu.go.jp/main_content/000499625.pdf）にせよ、その主たる内容はAIの開発者たるヒトが尊重・注意すべき諸価値であり、AI自体がそこで提示された規範の名宛人として想定されているわけではない。AI・ロボットの責任問題というテーマのもとに我々が考えているのが基本的には（過剰）緊急避難の例であること――たとえば5人の通行人を回避するために急転回して別の通行人1人を跳ねることが許されるか、急ブレーキで搭乗者1人を犠牲にするのならどうか――も、このような前提を反映しているように思われる。我々はそこで、与えられた指示にAIが忠実に従った結果として回避不能な一定の損害が発生した場合の負担分配について議論しているのであり、AIが我々の生命尊重という価値をうっかり忘れてしまうとか、怒りのあまり殺害を決意するようなことを考えているわけではないだろう。

そこにあるのは我々の自律的な決断としての故意の問題でも、従いそこねとしての過失の問題でもないのである。

◆

6．人間的な、あまりに人間的な

所有者の子供を故意に死なせた容疑で追求されたロボット「ロビタ」が、自分は人間なので人間同様に殺人罪で裁かれるべきだと主張する手塚治虫『火の鳥』復活編（小学館クリエイティブ、2014年（初出1970～71年））のエピソードを想起しよう。結果的にこの要求は認められず、ロビタには故障した物として融解処分が下される。だが重要なのはそもそもロビタが作中において「人間らしい」「ほかのロボットとちがう」失敗したり感情を表わすような存在として位置付けられていることだろう（その理由についてここでは触れない）。指示に逆らったり従いそこねたりする人間的な存在だからこそ、それを対象とする統制手段として法が意味を持つのであり、法的主体として扱えというロビタの主張が意味あるものとして我々に響くことになるのだろう。

AI・ロボット法とは、それらの取扱いにおいて我ら人間が守るべき掟なのか、AI・ロボット自身に向けられた掟なのか。法があくまで意思の弱さを抱えた不完全な存在としての我ら人類に対して意味を持つようなものであるとすれば、後者の可能性は――AI・ロボットが未発達だからではなく、逆にその完全性の故に――あらかじめ閉ざされているように思われる。それとも技術発展はやがてロビタのように、不完全性さえも備えたヒトの似姿を実現することを目指すことになるのだろうか。

プロフィール……
おおや・たけひろ　1974年生まれ。慶應義塾大学法学部教授、専攻は法哲学。東京大学法学部卒、同大学助手・名古屋大学大学院法学研究科助教授・教授等を経て現職。著書に『自由か、さもなくば幸福か？：21世紀の〈あり得べき社会〉を問う』（筑摩選書、2014年）、『法哲学』（共著、有斐閣、2014年）、『法哲学と法哲学の対話』（共著、有斐閣、2017年）、『裁判の原点：社会を動かす法学入門』（河出ブックス、2018年）等がある。