〔論文〕

想像力と行為の説明

佐 藤 英 明

〈目　次〉

はじめに

2015年5月に公開されたアプリケーションソフトGoogleフォトには，ラベリング機能があった．画像認識AIが，写っている人物，場所，物などを識別しラベルをつけるサービスである．しかし，公開から2ケ月足らずで問題が発生した．アプリを利用した一人の黒人男性が，自分と友人の黒人女性の写真にGoogleフォトが「ゴリラ」というラベルをつけたことを示すスクリーンショットをTwitter（当時）に投稿したのである．明らかに人種差別的なラベリングであり，Googleは被写体に「ゴリラ」とラベルづける機能を停止した．人間の画像を動物と分類してしまうような誤認識をもたらす機能は2023年現在，停止されたままである．この問題について，Googleの元従業員によれば，トレーニングデータに黒人の写真が不足していたためアプリが誤作動したという．また公開前の機能テストも不十分だったという．しかし，当該AIの学習プロセスが詳細に解明され，誤認識の理由が説明されたわけではない[(1)]．

Googleはフォトアプリの欠陥について謝罪し，人種差別的なラベリングをもたらすおそれのある機能を停止した．その後も，誤動作した機能を修正するのではなく，遮断するという対応がとられた．類人猿のラベルづけ機能の停止は，この機能がもたらすベネフィットがリスクを上まわるものではないと判断されたためだという．たしかにゴリラやオランウータンのラベルづけが活用される機会は少ないだろう．他方，人種差別的な誤認識が再び生じれば，企業にとっては大きなダメージとなる．かりに類人猿のラベルづけの精度が向上したとしても，機能を有効にするという判断は難しいと思われる．

フォトアプリの開発者に差別的な意図があったとは考えにくい．まして，画像認識AIにそのような意図があったなどとはいえない．しかし，AIの誤認識が，人の死傷といった重大な問題を引き起こしてしまった場合，それ

が意図せざる結果であっても，責任の所在を明らかにするためにAIシステムに関与する者は説明責任を果たさなければならない．AIは，入力に対する処理をすべて開発者が設計するものではなく，教師データとなる入出力データの組み合わせを模倣することで，処理を機械学習によって自動的に獲得するものである．そのため，その学習過程の根拠を示すことが困難となる．そうしたブラックボックス性は，説明や釈明において，大きな障害となる．

本稿では，こうした問題について考えるための前提として，そもそも人が行為の理由を説明するとはいかなることなのか，行為の理由を説明したり自らの行為について釈明したりするとき人は何をおこなっているのかを検討する．その際，特に人の想像力という能力が，行為の説明において果たす役割について考察する．

1　行為の説明

1-1　実践的推論と目的論的説明

人は行為を説明したり解釈したりする際に「実践的推論（practical inference）」の図式を用いる．これは，いくつかの前提から具体的な行為を結論として導き出す推論であり，アリストテレスによって体系化された．演繹的推論のような理論的推論では，前提が真であれば結論も真となり，前提と結論の間に論理的含意関係がある．それに対し，実践的推論（実践的三段論法）では，結論としての行為が論理的に導かれるわけではない．実践的推論は，行為への動機づけを与えるものとして機能する．

フォン・ウリクトは，次のような実践的三段論法の図式によって行為を説明している(2)．

> Aは，pを生ぜしめようと意図する．
> Aは，aをなさなければpを生ぜしめることができないと考える．
> それゆえ，Aはaにとりかかる．

この実践的推論では，p を意図する（目的とする）ことから a という行為が導出される．p という意図（目的）が前提，この目的を実現するために行為 a が手段として必要であるという認識が第二の前提となり，当該行為の遂行が結論として導かれる．これにより実践的推論は，行為遂行に先立ちそれを促すものとして機能する．

しかし，実践的三段論法は理論的三段論法とは異なり，前提から結論が論理的に帰結するものではない．行為者 A が p を意図し，そのための手段として行為 a が必要であると認識していたとしても，a を遂行しない可能性はある．締切りまでに原稿を完成させようと意図し，そのためには急いで執筆する必要があると認識していても，執筆にとりかかれるわけではない．

フォン・ウリクトによれば，以上のような実践的推論の図式と「逆さま」になっているのが「目的論的説明（teleological explanation）」の図式である．目的論的説明の出発点は，A が a をなし p を生じさせたことである．行為遂行後に「なぜ a をなしたのか」と問われたとき，「p を生ぜしめるためである」という応答がなされる．

> A は a をなし p を生じさせた．
> A は，a をなさなければ p を生ぜしめることができないと認識していた．
> それゆえ，A は，p を生ぜしめようと意図していた．

この目的論的推論の図式は，現実に行為がなされた後，その行為を説明したり解釈したりするものとなる．行為遂行後に「なぜそれをなしたのか」という理由を説明するためにこの図式は用いられる．

実践的推論の図式が意図から行為を導き出し，行為遂行前に行為を促すものとして機能するのに対し，目的論的説明の図式は行為から意図へと遡り，行為遂行後に理由を説明するものとして機能する．

1－2　意図的行為と非意図的行為

アンスコムによれば，行為者は自分が何をするつもりでいるのか（意図）

を観察によらずに知っている[3]．右折しようとしている運転手は，右折しようという自らの意図を観察によらずに知っている．行為者は自らの行為の意図を観察によらずに説明し記述することができる．

それに対し，行為者以外の者は，行為者や行為の結果を観察し解釈することにより，その意図を理解し記述する．コーヒーを注文している人を見れば，コーヒーを飲もうという意図をもっていると解釈できる．バス停の前に立っている人を見れば，バスに乗ろうという意図を理解する．もちろん，行為者が意図を表明することにより，意図が理解されることもある．それは言語的な表明の場合もあれば，そうでない場合もある．ウインカーの合図によって，周囲の人は運転者の右折しようという意図を把握する．

それゆえ，意図的行為は，行為者本人によって一人称的観点から記述することも，他者によって三人称的観点から記述することもできる．行為者は自分が「何をするつもりか（意図）」を観察によらずに知る．それは一人称的に記述される．その意図から出発し，実践的推論の図式にしたがって行為を起動する．他方，行為者が「何をしたことになっているか」は，行為の結果を反映して記述される．行為の結果は，観察によって知られる．行為がもたらした結果から遡って，行為者の意図が目的論的に説明される．したがって，それは三人称的に記述される[4]．行為者が実践的推論に基づいて行為したと想定し，実践的推論図式を「逆さま」にして，行為がもたらした結果から意図へと遡及することにより，三人称的観点から行為の理由を説明することが可能となる．

目的論的推論の図式にもとづいて次のような推論をおこなうこともできる．A が a をなし p を生じさせたとしても，a をなすことで p が生じることを認識していなかったとすれば，A は p を生ぜしめようと意図していなかったと結論づけられる．

オイディプスは，意図的に一人の男を殺し，王妃と結婚した．それは父殺しであり，母との結婚であった．しかし，オイディプスはそのことを知らず，父殺しも母との結婚もオイディプスの「意図せざる行為」である．父を

殺そうという意図も母と結婚しようという意図ももっていなかったにもかかわらず，父を殺し母と結婚することになってしまった．オイディプスはその男を殺すことで父殺しが生じることを認識していなかったからである．そしてオイディプスが先王ライオス（父）殺しの犯人捜しをはじめたとき，ライオスがオイディプスの父であるという認識をもっていなかったのは，オイディプスだけではない．テバイ王国という共同体にそのような認識をもつものは（予言者をのぞき）いなかったのである．

事実の錯誤に関する判例とされる「たぬき・むじな事件」（1924年）では，被告人は，狩猟法が捕獲を禁止している「たぬき」を「むじな」という動物だと思って捕獲した．当時，たぬきがむじなと同じ動物であることは一般に知られていなかった．そのため，被告人もたぬきを捕獲することになるという認識なしに，むじなを捕獲しようと意図していたと考えられる．この場合，たぬきがむじなであることを認識していたのが動物学の知識を持つものに限られていたことが重要な論点となっている．そのような認識が共同体において共有されている場合には，被告人が「認識していなかった」と主張しても認められない(5)．

道路の向こうに友人がいることに気づいて手をあげて挨拶しようとする．ところが，それによってタクシーが停まってしまう．行為者は，「タクシーを停めよう」という意図はもっていない．手をあげることでタクシーを停められることは知っていても，そのときタクシーが通りかかっていることは認識していない．自らの行為によってタクシーが停まったことも認識していないかもしれない．行為者は，自分の行為がもたらした結果を認識しておらず，自分が「何をしたことになっているか」を知らない．

前述のように，行為者が「何をしたことになっているか」は，観察された行為の結果を反映して三人称的に記述される．行為者の意図は，行為がもたらした結果から遡って目的論的に説明される．他方で，行為者は「何をするつもりか」を一人称的に語る．この二つは必ずしも一致しない．行為者本人が「そんなつもりはなかった」と意図を否定しても，他者によって意図が肯

定されることもある．たとえば，被害者を刃物で刺した犯人が「誤って刺してしまっただけで，殺すつもりはなかった」と主張する．しかし，本人がそう主張しても，被害者に複数の刺し傷があったなどの証拠から，強い殺意があったと判断されることもある．

意図的な行為は，行為者の意図を示す．行為の意図は行為者が属する共同体の慣習や制度，共有されている知識などにもとづいて理解される．「人々は，学習や訓練によって，こうした共同体に導き入れられる」．こうした共同体のことを，フォン・ウリクトは「生活共同体（life-community）」と呼んでいる(6)．生活共同体にとって完全に疎遠な行動は，目的論的に説明したり理解したりすることができない．「意図はその状況の中に，つまり人間の慣習や制度の中に，埋め込まれている(7)」．この場合，「意図」は心的状態として行為者自身によって知られるものではなく，行為者が属する生活共同体の慣習や制度，共有されている知識などにもとづいて想像され解釈されるものである．

1－3　動物の意図的行為

中世ヨーロッパでは，動物裁判がおこなわれていたことが知られている．現代ならば事故や自然災害として処理されるような出来事が動物の犯罪行為とされ，正規の裁判所で人間とまったく同じ訴訟手続きと厳正な審理を経て判決が下され，刑が執行された(8)．人の行為と無関係な事故や災害は不可抗力による危害であり，犯罪として裁くことができるようなものではない．それゆえ，現代人の目には，このような裁判は意味のない不合理なものと映る．

大屋雄裕は，われわれが動物裁判を無意味なものと考える理由を「行為指導性」という観点から説明している(9)．人間は，法に反する行為が処罰の対象となることを予測して，そうした行為への関与を事前に避けることができる．そのため，法制度には，特定の行為を人々に行わせたり行わないようにさせたりする「行為指導性」がある．しかし，法を理解できず処罰を予測できない動物に対してそのような事前規制は不可能である．

人間は実践的推論の図式にしたがって次のように推論する．罰を受けないように意図し，「a をなせば罰せられる」と考えるならば，a をなすことはしない．それゆえ，法による事前規制が可能となる．しかし，動物は「a をなせば罰せられる」ということを事前に知ることができない．実際に a をなし罰せられるという経験をしてはじめて「a をなせば罰せられる」ということを学習する．動物は事後処罰による学習を通じてしか行動をコントロールすることができない．そのため動物が a を事前に回避できる可能性はない．

たしかに動物は処罰の対象を言語によって事前に知ることはできない．しかし，「縄張り（territory）」のように種が遺伝的に共有する行動の「ルール」は動物にもある．多くの動物は，生き延びるための資源占有空間を確保するために「縄張り」を作り，同種の他個体の侵入から防衛する．縄張りは，「臭いづけ」などの方法により表示されるが，その中に同種他個体が侵入することがある．占有者は縄張り防衛のために，威嚇や攻撃などの行動をとる．このような遺伝的に共有される行動ルールの場合，動物でも「ルールを破れば攻撃される」ということを事前に「知っていた」とみなすことができる．

侵入すれば攻撃されるということは動物でも予測できる．しかし，他個体の縄張りに侵入していることを何らかの理由で認識していないこともありうる．臭いづけマーカーを認知せずたまたま侵入してしまったものに対し，占有者は縄張り防衛のために威嚇や攻撃などの行動をとる．威嚇行動は「ルール違反」への「警告」として機能する．警告を受けた侵入者は，それによって侵入しているという事実を認識するはずである．あやまって偶然に侵入してしまったのであれば，侵入者は攻撃を避けるために立ち去るであろう．この場合，侵入は「過失」であって「意図的」ではなかったと解釈することができる(10)．

それに対し，侵入者が警告を受けても立ち去らなかったとすれば，攻撃されることを予測しながら，それを回避せず，あえて侵入しようと意思決定したとみなすことができる．侵入者は「ルール違反」を認識し，攻撃を避ける

ことが可能であったにもかかわらず，「意図的」に侵入したことになる．威嚇されても執拗に侵入を繰り返し，侵入者の縄張り内の食料が枯渇していることが確認されれば，「空腹を満たすため食料を得よう」という意図をもっていたと説明することも可能だろう．

むろん動物が人間のように自らの行為の意図を観察によらずに知っているということはできない．動物は自らの意図を述べることも説明することもできない．「なぜそれをなしたのか」という問いに自ら応答することはできない．しかし，人間が動物の行動の理由を説明することは可能であり，場合よっては，意図的な行為として三人称的に記述することも可能である．たとえば，動物が侵入者をいきなり攻撃せず最初に警告をおこなうのは「無益な闘争を避けるため」であると解釈することができる．しかし，威嚇行動をとっている動物がそのような「意図」をもっているわけではないだろう．

自らの意図を説明することができない動物についても，人はその意図的行為や過失について語ることができる．それは，動物の行動が人の生活共同体から完全に疎遠なものではなく，人間の慣習や制度をもとに意図を想像し解釈することが可能なものだからであろう．

2　遡及推論

2-1　アブダクション

C.S. パースは，演繹，帰納のほかに，アブダクション（abduction）という仮説形成の推論方式を提唱した．演繹，帰納，アブダクションの違いを，パースは次のような例をあげて説明している[(11)]．

〈演繹〉

規則　この袋の豆はすべて白い（A ならば B である）
事例　これらの豆はこの袋から取り出した豆である（A である）
結果　ゆえに，これらの豆は白い（B である）

〈帰納〉

事例　これらの豆はこの袋から取り出した豆である（A である）

結果　これらの豆は白い（B である）

規則　ゆえに，この袋の豆はすべて白い（A ならば B である）

〈アブダクション〉

規則　この袋の豆はすべて白い（A ならば B である）

結果　これらの豆は白い（B である）

事例　ゆえに，これらの豆はこの袋から取り出した豆である（A である）

演繹は正しい規則を前提とし，事例が正しければ，正しい結果を導き出すことができる．帰納は，同様の事例と結果が観察されたことから一般規則を導き出す推論である．それに対し，アブダクションは観察結果を説明するための仮説を形成する推論である．それは，結果の由来を説明するものともなる．ある事実（結果）に遭遇し，それを一般規則の一事例と仮定すれば説明できるような場合，その仮定を採用するための推論である．内陸部で魚の化石が発見されたとき，かつて海であった場所から魚の化石が発見されることから，昔はその地域一帯が海であったと仮定するようなケースである．観察結果を生じさせた原因を既知の法則によって説明するような推論もこれにあたる．そこがかつて海であったということは観察された事実ではない．それは，観察結果を説明するために想像された仮説である．それゆえ，仮説形成には事実として確認されていないものを想像する能力が必要となる．

いうまでもなく「A ならば B である」から「B ならば A である」を導き出すのは，論理的には誤りである（後件肯定の誤謬）．「パリにいるならばフランスにいる」は正しくても「フランスにいるならばパリにいる」は誤りである．それでも，パリにいる可能性があると推論するのは理にかなっており，パリにいると仮定したうえで，そのことを裏づける根拠を見出すことへとつながる．インフルエンザに感染すれば発熱することが多い．だからといって，発熱したからインフルエンザに感染していると判断するのは誤りである．それでも，インフルエンザのために発熱していると仮定することで，そ

れを検査によって確認することへとつながる．

パースは，アブダクションを「リトロダクション（retroduction）」とも呼んでいる．この語は「遡及推論」を意味する．結果から原因へと，あるいは観察データからそれを説明する法則へと遡及する推論である[12]．ある事実（結果）が確認されたとき，その事実を説明するために考えられた仮説（原因）によって，説得力ある理由を示すことができる．内陸部で魚の化石が発見されたという事実を「かつては海であった」という仮説によって説明することは，演繹的推論のような論理的正しさを示すものではないが，ある程度の説得力をもつ．また，帰納的推論の結論が一般的（全称的）命題であるのに対し，アブダクションは「過去に起こった特定の一回的事実」についての単称的命題を結論づける[13]．

前述のように実践的推論では，「A が p を生ぜしめようと意図し，a をなさなければ p を生ぜしめることができないと考えている」という前提から，「A は a にとりかかる」という結論が導かれる．それに対し「なぜ a をなしたのか」と問われたときの「p を生ぜしめるためである」という応答は，実践的推論の図式を「逆さま」にした目的論的説明の図式にもとづいている．行為者本人が一人称的観点から応答する際には，この目的論的図式にしたがうことになる．A は，観察によらずに自らの意図を答えることができる．

他方，A が a をなし p を生ぜしめたことが観察されたとき，他者によって目的論的な説明がなされる．A が a をなしたことが観察され，a をなさなければ p を生ぜしめることができないと認識していたとみなされるならば，A は p を生ぜしめようと意図していたと結論づけられる．これは，目的論的な「遡及推論（retroduction）」である．その記述は，過去を振り返って「遡言（retrodiction）」するものとなる[14]．

目的論的な遡及推論においては，結果 p から行為 a に遡り，「行為 a によって結果 p が発生するであろう」と A が認識していたとみなすことができれば，A の意図が推定できることになる．B が A の銃撃により死亡し，銃を撃てば B が死亡することが誰が見ても明らかな状況であったとすれば，A

は殺そうと意図していたと結論づけられる．それに対し，B が A の銃撃により死亡しても，A が B の存在に気づいていなかったと判断されれば，A の殺意は否定される．

このような遡及推論において，行為者 A の認識や意図といった心的状態は，第三者によって観察されるようなものではない．それは，行為そのものや行為の結果を観察することから推論され想像される．そして前述のように行為者の認識は，生活共同体によって共有されている認識をもとに想像され，行為者の意図は生活共同体に共有されるルールや規範を前提として解釈される．

2-2　言語習得とアブダクション

今井むつみは，子どもの言語習得の過程でおこなわれていることは，帰納的推論とアブダクション推論を混合したものであると指摘している[(15)]．

子どもにある対象 X を指さして，A という新しい言葉を教える場合，子どもが A という言葉を理解したかどうかを確認するには，X とは異なるモノを一緒に見せて，「A はどれ？」と聞く方法が標準的だという．これは「X ならば A」と教えたとき，子どもが「A ならば X」を学習できることを前提としている．黄色い積み木を見せて「キイロ」という言葉を教えたとき，子どもが「キイロ」という音声を逆に黄色いモノに対応づけられると想定しているのである．人間は，「対象→記号の対応づけを学習したら，記号→対象の対応づけも同時に学習する」．

ところが動物はそうではない．人間にとってはあたりまえのことが動物にはできない．訓練を受けたチンパンジー「アイ」は，異なる色の積み木にそれぞれ対応する記号（絵文字）を選ぶことができるようになった．しかし，アイに記号から色を選ぶように指示すると，対応づけができなかった．訓練された方向での対応づけ（対象→記号）はできても，逆方向の対応づけ（記号→対象）はまったくできなかった．人間の子どもなら難なく正解できる逆方向の対応づけが動物にはできないという事実に，人間の子どもの言語発達

を研究してきた今井は驚愕したという[16].

人間は言葉と対象とのあいだに対称的関係があると想定し，自然に対称性推論をおこなう．しかし，人間以外の動物は（わずかにグレーな例外はあるが）対称性推論は行わない．そして，対称性推論は言語の習得と深く関係している．言語は対称性推論なしには成立しない[17]．しかし，「X ならば A」から「A ならば X」を導き出すことは，論理的には誤りである（後件肯定の誤謬）．世界地図でパリのあたりを指さして「ヨーロッパ」という言葉を教えたとき，「ヨーロッパはどこ？」と聞かれて，同じ場所を指さすことができたとしても，「ヨーロッパ」という言葉と対象との関係を理解できたとは言えないだろう．

力を入れて物を遠くに飛ばす動作を見せて，子どもに「なげる」という言葉を教える．子どもは，対称性推論によって「なげる」をその動作と対応づけ，この言葉を学習する．ところが，手を使わなくても物を遠くに飛ばすことはできる．そのため「足でなげる」といった誤用が生じる[18]．そうした誤用は，「ける」といった他の言葉を学習していくなかで修正されていくことになる．

クワインが翻訳の不確定性テーゼを論じた際に用いた‘Gavagai’という例は，言葉が何を指示するかは一義的には確定できないことを示している[19]．ある未知の言語を話す現地人が，一匹のウサギが走り抜けるのを見て‘Gavagai’と言った．言語学者はひとまず‘Rabbit’と翻訳する．しかし，‘Gavagai’と‘Rabbit’が一対一に対応しているのかどうかは不明である．‘Animal’と訳すべきかもしれないし，‘White’かもしれない．「ウサギが跳ねている」という文かもしれないし「ウサギを捕まえろ」という命令かもしれない．言語学者は，この状況では特定することができない．さまざまな状況で現地人に‘Gavagai?’と問いかけ，同意するかどうか確認しなければならない．この場合，言語学者はアブダクティブな推論によりひとまず‘Gavagai’が‘Rabbit’に対応するという仮説をたてていることになる．そして，その仮説が正しいかどうかを現地人への問いかけによって検証していか

なければならない.

2-3　言語習得と言語共同体

それゆえ，言語習得にはアブダクティブな推論の能力が必要とされる．人間は言葉と対象との対称的関係を想定し対称性推論をおこなうことによって，言葉の意味を理解する．しかし，その理解は誤っているかもしれない．意味理解を正しいものとするためには，その言葉を用いて他者と対話することにより，意味理解の正しさを検証し，修正していく必要がある．換言すれば，言語を習得し理解できるようになるということは，言語共同体への帰属プロセスである．

そして，言葉の意味は対象と一対一の対応になっているようなものではない．同じ言葉でも，使用される状況によって意味は異なる．「テストをなげる」は，答案用紙を投げ捨てることかもしれないし，試験に合格することをあきらめることかもしれない．「視線をなげる」は，投げ捨てることでもあきらめることでもない．「なげる」という言葉が，具体的な生活の中で使われ，特定の役割を果たすことによって，この言葉は意味を成し，理解できるものとなる．この言葉を用いて生きてきた日本語話者の言語共同体が有する生活形式（生活のあり方）を背景とし，その背景の認識のもとで，言葉は理解可能なものとなる[(20)]．

言語共同体において言語が習得される際には，たんに言語の意味が理解されるだけではない．共同体の生活規範，ルールもまたともに習得される．行為の意図が行為者が属する共同体の慣習，制度，共有されている知識などにもとづいて理解されるように，言葉の意味は，言語共同体が有する生活形式のなかで，慣習，制度などとともに習得される．「黒人」という語は，この言葉が使用される文脈のなかで習得される．それは，たんに言葉と対象との対応を習得することではなく，黒人差別を含む生活のあり方のなかでの使用法を習得することであり，差別に対する規範の習得もそれにともなっている．言葉の習得は，それに関わる規範を含む共同体の慣習，制度，知識の習

得とともにおこなわれる．

AI は対象とラベルとの対応を「学習」するが，それは，人の言語習得とは異質なものである．AI は対象との対応関係を学習するだけで，それと関係する言語共同体の規範やルールを学習するわけではない．AI は，ラベルが差別的に使用されることについては何も「知らない」．AI は，ある言葉がどの対象と対応するかを「学習」することができても，その言葉の生活共同体における意味を把握することはできない．それゆえ，それが差別的な言葉であるといったことは理解できない．それゆえ，差別的な意図でその言葉を用いたと解釈することはできない．

AI が言葉と対象の対応を学習すれば，入力された対象に対して適切な言葉を出力することができるようになるだろう．Google フォトのラベリング機能も，画像データに対してラベルを出力するものである．このアプリは，犬を馬と間違え，時計をホイールキャップと間違える．スケートボード中に怪我をした人の血まみれのひじの写真に「食べ物」というラベルをつけることもあるという[21]．そうした誤認識は，学習により改善されるであろう．しかし，深層学習によって画像認識の精度が上がり，画像に対して正しくラベリングできるようになっていったとしても，機械学習によって自動的に獲得された学習プロセスは複雑で，認識の根拠を示すことは困難である．正しい予測結果だったとしても，結果に至る過程が適切であったかどうかは分からない．それは，誤認識が生じた場合も同様である．

3　理由説明としての釈明

3-1　事実に反する想像

「なぜそれをしたのか？」と問われたとき，人は言葉によってそれに応じる．この問いに対し，人は必ずしも行為の意図を答えるわけではない．その行為が悪い結果（負事象）をもたらすものだったとき，この問いは，たんに

理由の説明を求めるものではなく，非難の言葉ともなる．「なぜ殴ったのか？」という問いに「いためつけるためだ」と応じれば，意図を答えていることになるが，「命令されたからだ」と答えれば，自分が非難の対象となることを否定しようとしていることになる．

人間には「不祥事が起これば，責任者を探し出し罰したいとする願望」があり，その背後には「すべてのことには原因があるはずだという信念」があるといわれる．「良いことをすれば報われ悪いことをすれば罰せられる世界にわれわれは生きている」という信念は，心理学では「公正世界信念（belief in a just world）」と呼ばれる[22]．これによって，悪いことが起こったからには，その「原因」があるはずであり，悪いことを引き起こした人は「責任」を負って罰せられなければならないと信じられることになる．それは，責任を負うべきものを探し出して罰したいという「応報願望」をもたらす[23]．人は応報願望から負事象（悪いこと）に関連したものを追求しようとする．そして，ルール違反や他者への危害などの負事象との関連が問われた人間は，言語を用いてさまざまな説明をおこなう．それは「釈明（account）」と呼ばれる．

エデンの園で身を隠したアダムに神は尋ねる．「取って食べるなと命じた木からお前は食べたのか」と．アダムは答える．「あなたがわたしと共にいるようにしてくださった女が，木から取ってわたしにくれたので，食べたのです」．神がイブに尋ねると，イブは答える．「蛇がわたしを騙したので，食べたのです」．

このやりとりについてジューディア・パールは，こう指摘する．「神が最初に発したのは『何が起きたのか』の問いだったのに．二人は『なぜ起きたのか』の問いに答えているようだ．神はただ事実を尋ねているだけなのに，二人は事実に説明を加えている[24]」．アダムもイブも「食べる」という行為をなしたという事実だけでなく，それをなした理由を答え釈明している．釈明は，行為者による行為の理由説明である．

釈明には，否認，正当化，弁解，謝罪があるとされる．「否認」は，「わた

しはやっていない」というように，被害をもたらした行為を自分がおこなったことを否定することである．「正当化」は，自分の行為が被害をもたらしことは認めるが，「わたしは規則に従ってそうしただけだ」というように，その行為の不当性を否定し，非難の対象となることを否定することである．「弁解」は，自分の行為が被害をもたらしたこともそれが非難の対象となるものであることも認めるが，その非難（罰）が自分に向けられることは否定することである．「上司の指示によっておこなったことだから，上司が責任を負うべきである」というように，自分がおこなったことにより負事象が生じたことやそれが非難の対象となるものであることは認めるが，責任を負うのは自分でないと主張するのである．それに対し，「謝罪」は，加害行為への関与とその不当性，責任のすべてを肯定するものである．それゆえ謝罪するためには，加害行為への関与およびその不当性を認め，自分が責任を負うことを表明する必要がある[(25)]．

アダムもイブも否認はしておらず，謝罪もしていない．アダムは「神が創造したイブがくれたから」という理由を述べて正当化し，イブは「蛇が騙したのだから蛇の責任である」と弁解している．アダムは，もし神がイブを創らずイブが木の実を渡さなかったならば，食べることはなかったと答えていることになる．そしてイブは，蛇が騙さなかったならば，食べることはなかったと応じている．

「イブが木の実を渡さなかった世界」を想像できなければ，アダムはこのような正当化をすることができない[(26)]．このような発言ができるのは，イブが渡さなければ自分は食べなかったという，事実に反する想像をしているからである．イブもまた「蛇が騙さなかった世界」を想像している．実際に起こった事実とは異なる可能性を想像できなければ，アダムやイブのように釈明することはできない．人間が動物と異なるのは，事実に反する世界を想像する能力をもっている点である．創世記の記述は，人間固有のそうした能力を示すものとなっている．

正当化や弁解のためには，アブダクティブな遡及推論が必要である．そし

てアブダクションは想像力なしには成立しない．ニュートンによる万有引力の法則の発見は，観察データ（事実）からの帰納的推論によって導かれたものではない．それは，「観察事実を説明するために創案され発明されたもの」である．引力は直接には観察不可能な「理論的仮説的対象」であり，帰納的一般化によって導かれるものではなく，「創造的想像力」なしに考えつくことのできるものではない[(27)]．

観察されるのは事実である．それに対し想像されるのは反事実である．観察データ（事実）をいくら集めても，想像上の反事実の世界のことは分からない．パールは，観察データと反事実との相性の悪さを指摘し，それでも「人間の知性はなぜか事実に反することを推定することができる」と述べている．そして，この能力は，動物や機械の学習にはない「人間の知性の最も重要な特徴」であるとしている[(28)]．

3－2　弁解と制御可能性

弁解という釈明の効果と大きく関係するのが，「制御可能性（controllability）」である．弁解の効果に関する実証的研究では，「電車に乗り遅れたから」といった制御可能性（回避可能性）の高い弁解よりも，「母親を病院に連れて行かなければならなかったから」といった制御可能性の低い弁解の方が好意的に受け取られ，弁解の効果が高いことが確認されている[(29)]．電車に乗り遅れないようにできた可能性は高いが，母親を病院に連れて行かずにすんだ可能性は低い．

意図的行為とは，制御可能性がもっとも高いとみなされるものである．逆に制御可能性が最も低く，回避不可能な場合には，本人の意図とは無関係な不可抗力による出来事とみなされる．両者の間の評価次元は段階的に変化するものであり連続している．「意図的」と「不可抗力」の間に「怠慢」「不注意」という段階を設定することもできるし，「意図的」「非意図的」の二段階に分けることもできる[(30)]．それゆえ，「意図的」「非意図的」の区別は，意図の有無により明確に二分されるようなものではなく，段階的に変化する評価次

元に間に設定されたものにすぎない．弁解とは「意図的ではない」という釈明であるが，回避可能性が高いとみなされる行為については，弁解は効果的ではない．回避可能性が高いほど，より意図的であるとみられるため，「意図的ではない」という釈明と整合的ではなくなるためである．

行為者Aがaをなしpを生じさせたという事実から遡り，aをなせばpを生ぜしめることになるということをAが認識（予見）していたか否かが，まず確認される．認識していたと判断されれば，Aはpを生ぜしめようと意図していたと目的論的に説明される．このとき，Aが「aをなせば罰せられる」ということも認識していたとすれば，罰を受けないように意図し，aを回避するという実践的推論ができたはずである．それにもかかわらず，回避せずあえてaをなしたということになる．Aは，制御可能性がきわめて高いにもかかわらず，aを回避せず，意図的に行為に及んだとみなされる．それゆえ，Aは非難され責任を問われることになる．aをなしたことは事実である．それに対して，aを回避できたはずだというのは事実ではなく，「推論において成立する事態」である．Aは実践的推論に基づいてaを回避できたはずなのに，現実にはaをなした．Aがaをなした責任を問われる際に問題となるのは，aをなした事実だけでなく，推論のうえで可能と認められる事態である．行為が意図的であったと判断される際には，事実ではない事態を想像することが必要となるのである．

aをなせばpを生ぜしめることになるということをAが認識（予見）していなかったと判断された場合には，pを生ぜしめようと意図していたと結論づけることはできない．そこで問題となるのが，認識（予見）できた可能性である．現実には認識していなかったとしても，認識できた可能性が問われるのである．aをなせばpを生ぜしめることになるということが，一般によく知られたことであれば，認識できた可能性は高い．認識していなかったのは本人の怠慢や不注意とみなされる．他方，それが，ほとんど知られていないことであったり認識困難なことであったりすれば，認識できた可能性は低い．意図的にaをなしたわけではないとしても，pを生ぜしめることを予見

できた可能性が高ければ，予見できたはずなのにしなかったことになる．しかし，かりに予見していたとしても，aを回避できなかったかもしれない．その際，現実に回避しなかったことではなく，回避できた可能性が問題となる．予見でき，回避できた可能性が高いとみなされれば，Aの「過失」が非難されることになる．

過失の場合，現実には結果の発生は行為者によって予見されていない．しかし，危険を予見可能だったのではないか，予見していればその結果を回避できたのではないかと遡及的に推論されることになる．過失の責任が問われる際に問題となるのは，現実にはなされなかった予見や回避である．現実には生じなかった出来事の可能性を判断することにより「意図的」と「不可抗力」の間に「過失」の程度という段階を設定するのである．

3－3　説明可能なAI（XAI）

自動運転車が起こした交通事故に関する裁判例として「東名高速事故判決」（横浜地裁・2020年3月31日）がある．事故は2018年4月29日，神奈川県綾瀬市の東名高速で起きた．自動運転レベル2の運転支援機能で走行中の乗用車が，別の事故で路上に停止中のバイクを検知せず衝突し，1人が死亡，2人が重軽傷を負った．事故当時，自動ブレーキ等の「運転支援システム」は作動状態であったが，運転者は居眠りをしていた[31]．

検察側が，バイクを検知しなかったのは「機能の限界」であり，事故の原因は運転者の居眠りにあると主張したのに対し，弁護側は，自動運転システムの自動ブレーキの故障が事故原因であるとして無罪を求めた．問題となるのは，運転者が事故を予見できた可能性があるのか，予見していたならば事故を回避できた可能性があるのかという点である．

現実には，運転者は事故を予見も回避もしていない．しかし，この乗用車の運転支援システムでは対応できず事故を回避できない場合がありうることを，運転者は「当然理解していたはずである」と裁判所は判断した．それゆえ，居眠りをして前方を注視できず適切に運転操作できない状態になれば

「事故が発生して人が死傷する危険」が生じることを予見できたはずだということになる．できたはずの予見を怠ったことにより運転者は，その責任を問われる．問題とされているのは，予見しなかったという「事実」ではなく，予見できたはずだという「推論のうえで成立する事態」である．被告人は実際には，運転支援システムが衝突を回避できると理解していた可能性もあるが，「本件運転支援システムが道路状況に応じた適切な動作をしないことがあり得ることは理解していたと認められる」と推論されたのである．

さらに，運転者が眠気に襲われた地点から事故現場までの間には，複数の非常駐車帯があり，そこで休息したり運転を交代したりすることも可能であった．前方注視が困難となるような眠気に襲われた時点で，運転を中止し，衝突を回避することができたはずだということになる．できたはずの回避を怠ったことにより運転者は，その責任を問われる．問題とされているのは，回避しなかったという「事実」ではなく，回避できたはずだという「推論のうえで成立する事態」である．

弁護側は，かりに居眠りをせず，自動ブレーキの不調に気づいて急制動の措置を採っていたとしても間に合わなかったと主張した．これに対し判決では，運転者がもし居眠りせずに前方を注視していたならば，運転支援システムがバイクの手前で停止させるように動作していない可能性を認識し，衝突の危険を予見して急制動の措置を採ることが可能であったため，衝突事故を回避することができたと判断できるとした．それゆえ，居眠りして前方注視していなかったことと衝突事故との間には因果関係が認められることになる．判決は「前方を注視していれば衝突を回避できた」と指摘し，禁錮３年，執行猶予５年の有罪判決が確定した．判決は故障の有無については「判然としない」として判断を避けている．かりに自動ブレーキが故障していたとしても，運転者が居眠りせず前方注視を怠らなければ，衝突は回避可能であったと判断されたからである．それゆえ，故障の有無にかかわらず，運転者の過失責任が問われることになる．

しかし，自動運転車による事故が AI システムによってもたらされたと考

えられるような場合には，AI の処理プロセスが「判然としない」として判断を避けるというわけにはいかないだろう．判断を可能にするには，AI の透明性や説明可能性を高め，人間が理解可能なかたちで情報を開示することが必要となる．

2019年に経済協力開発機構（OECD）は，「安全性」「公平性」などのほか「透明性・説明可能性」「説明責任」などを含む５つの「AI 原則」を公表している．「透明性と説明可能性（Transparency and Explainability）」の原則は，AI システムに関する透明性と責任ある開示に関するものであり，人々が AI システムとの関わりを理解し，結果に異議を唱えることができることを保証するものである．「説明責任（Accountability）」の原則は，AI システムを開発，展開，運用する者が，OECD の AI に関する価値観に基づく原則に沿って適切に機能していることについて説明責任を果たすことを求めるものである．

しかし，AI は，入力に対する処理をすべて開発者が設計するものではなく，教師データとなる入出力データの組み合わせを模倣することで，処理を機械学習によって自動的に獲得するものである．そのため，その学習過程の根拠を示すことが困難となる．学習データに偏り（バイアス）があれば，公平性をそこなう処理を獲得してしまうことになるが，そのためには学習データの検証が必要となる．AI は開発者の意図とは異なる学習をしているかもしれない(32)．

透明性とは，利用者が理解可能なかたちで情報を提示することであるが，これは AI システムが誤った推論結果を導き出したり重大な問題を生じさせたりした場合に責任の所在を明らかにするために不可欠である．AI の透明性を高めるためには，学習や検証，処理などのプロセスについて情報を提示しなければならない．しかし，ディープラーニングなどの場合，人間が理解できるようなかたちで出力までのプロセスや根拠を示すことは容易ではない．

それでも，AI システムに関与する者は説明責任を果たさなければならな

い．問題が生じたとき，誤った出力にいたった根拠を提示し，どこに原因があったかを明確にする必要がある．そのためには，AI が学習によってどのように処理を獲得し，どのような根拠に基づいて出力をおこなったのかを説明可能なものとしなければならない．

「決定木」「ロジスティック回帰」のような古典的アルゴリズムは説明可能性が高い．それに対し，ディープラーニングモデルは説明可能性が低い．ところが，複雑な問題に関して精度の高い予測がおこなえるのは後者である．説明可能性と扱える問題の複雑さはトレードオフの関係にある．説明責任を果たすには，このトレードオフを解消し，説明可能性の高さを複雑な表現力と両立させなければならない．そのために提案されているのが「説明可能な AI（eXplainable AI：XAI）」という技術である(33)．AI の内部構造を精緻に解析することで予測にいたるまでの計算過程を確認できるようにしなくても，ブラックボックスである AI システムに対し「外挿的に説明を与えるような手法」によって判断理由を説明することも可能である(34)．XAI には，そうした技術も含まれる．人間に理解可能なかたちで理由を説明できることを主眼としているわけである．AI 技術が急速に普及するなかで，AI が示す予測や推論の根拠を示せないという問題の解決のために「説明可能な AI」という技術が求められている．

おわりに

実践的推論の図式は，意図から行為を導き出し行為を促すものとして機能する．それを反転させた目的論的説明の図式にしたがい行為から意図を遡言することにより，三人称的観点から行為の理由を説明することが可能となる．行為者は「何をするつもりか」を一人称的に語るが，行為者が「何をしたことになっているか」は，観察された行為の結果を反映して三人称的に記述される．このとき「意図」は心的状態として行為者自身によって知られるものではなく，行為者が属する生活共同体の慣習や制度，共有されている知

識などにもとづいて想像され解釈されるものである．自らの意図を説明することができない動物についても，人はその「意図」や「過失」について語ることができるが，それは，人の生活共同体をもとに動物の意図を想像することができるからである．行為者の認識や意図は，行為そのものや行為の結果を観察することから遡及推論される．とくに行為者の認識は，生活共同体によって共有されている一般的認識をもとに推定される．

「なぜそれをしたのか？」という問いは，たんに理由の説明を求めるだけでなく，非難の意味が込められた発話である．それに応じる人は必ずしも意図を答えるわけではなく，非難の対象となることを否定しようとして正当化したり弁解したりする．そうした釈明のためには，アブダクティブな遡及推論が必要である．そしてアブダクションは想像力なしには成立しない．

弁解という釈明の効果と大きく関係するのが，制御可能性である．制御可能性がもっとも高いとみなされるのが意図的行為であり，逆にもっとも低いとみなされるのが不可抗力による出来事である．その間の評価次元は連続的である．弁解は「意図的ではない」という釈明のため，回避可能性が高いほど整合的ではなくなり，効果がなくなる．「意図的」と「不可抗力」の間に位置づけられる「怠慢」という段階に相当するのが「過失」である．過失とみなされるとき問題とされるのは，予見可能性と結果回避可能性である．現実には予見されていなかったとしても，予見できた可能性が問われ，現実に回避しなかったとことではなく，回避できた可能性が問題となる．その際，一般的によく知られていることは，認識可能性が高いとみなされ，認識していなかったのは本人の怠慢と判断され非難される．行為者が「予見（認識）できたはず」「回避できたはず」というのは，「事実」ではなく「推論のうえで成立する事態」である．現実には生じなかった出来事が想像され，その可能性が判断されるのである．

自動運転車による事故の過失責任が問われる場合も，現実には生じなかった出来事が想像され，その可能性が判断されることになる．その際，問題となるのは，自動運転システムについて一般によく知られているとみなされる

のがどのようなことかということである．自動運転システムが利用される生活共同体の慣習や制度，共有されている知識などにもとづいて，行為者の意図や認識は想像され解釈される．ブラックボックスである AI システムの説明可能性を高めることは，生活共同体において共有される知識を更新することと結びつく．AI システムによってもたらさる事故の責任という問題にとって重要なのは，社会がどのような認識を共有するかである．自動運転車の社会受容において重要なのは，許容するといった姿勢よりも，社会的に共有すべき認識を形成することであろう(35)．

〔注〕

（1） New York Times（Online）New York Times Company. Jul 1, 2015, May 22, 2023. 渡辺豊・根津洸希（編）『AI と分かりあえますか？：ブラックボックスが生まれるしくみ』（信山社・2024年），9 頁以下.

（2） Von Wright, G.H., *Explanation and Understanding*,（Cornell UP, 1971），pp.96f. 丸山高司・木岡伸夫（訳）『説明と理解』（産業図書・1984年），123頁以下．増田豊『刑事手続における事実認定の推論構造と真実発見』（勁草書房・2004年），第一章第二節.

（3） Anscombe, G.E.M., *Intention*,（Harvard UP, 1957），柏端達也（訳）『インテンション：行為と実践知の哲学』（岩波書店・2022年）.

（4） 野矢茂樹『増補改訂版哲学・航海日誌』（春秋社・2024年），255-262頁 .

（5） 大塚裕史・十河太郎・塩谷毅・豊田兼彦『基本刑法Ⅰ：総論［第 3 版］』（日本評論社・2019年），93-95頁 .「むささび・もま事件」では，むささびともまが同じ動物であると一般的に知られていた点が「たぬき・むじな事件」とは異なっている.

（6） Von Wright：前掲注（2），pp.114-115. 邦訳149-150頁.

（7） Wittgenstein, L., *Philosophische Untersuchungen*,（Blackwell, 1953），§337.

（8） 池上俊一『動物裁判：西欧中世・正義のコスモス』（講談社現代新書・1990年），26頁 .

（9） 大屋雄裕「AI における可謬性と可傷性」宇佐美誠（編）『AI で変わる法と社会』（岩波書店・2020年），45-46頁 .

（10） 戸田正直『感情：人を動かしている適応プログラム』（東京大学出版・

1992年），10–12頁.
(11) Peirce, C.S., *Chance, Love, and Logic : Philosophical Essays*, (Harcourt, Brace and Company, 1923) Sixth Paper. 浅輪幸夫（訳）『偶然・愛・論理』（三一書房・1982年），第六章.
(12) 米盛裕二『アブダクション：仮説と発見の論理』（勁草書房・2007年），43頁.
(13) 同書，88–89頁.
(14) Von Wright：前掲注（2），pp.59–59. 邦訳74–75頁.
(15) 今井むつみ，秋田喜美『言語の本質：ことばはどう生まれ，進化したか』（中央公論新社・2023年），212頁 .
(16) 同書，225–228頁.
(17) 今井むつみ・岡田浩之「『対称性』へのコメンタリー：言語の成立にとって，対称性はたまごかにわとりか」『認知科学』（2008年）15（3），470頁.
(18) 今井・秋田：前掲注（15），216–217頁.
(19) Quine, W., *Word and object*, (MIT Press, 1960), §7. 大出晃・宮館恵（訳）『ことばと対象』（勁草書房・1984年），第7節.
(20) 古田徹也『はじめてのウィトゲンシュタイン』（NHK 出版・2020年），185頁.
(21) New York Times：前掲注（1），Jul 1, 2015.
(22) Lerner, M.J., *The Belief in a Just World : A Fundamental Delusion*, (Springer, 1980).
(23) 大渕憲一『謝罪の研究：釈明の心理とはたらき』（東北大学出版会・2010年），58–59頁.
(24) Pearl, J. and Mackenzie, D., *The Book of Why : The New Science of Cause and Effect*, (Penguin, 2019), pp.23–24. 夏目大（訳）『因果推論の科学：「なぜ？」の問いにどう答えるか』（文藝春秋・2022年），44–45頁 .
(25) 大渕：前掲注（23），18–26頁.
(26) Pearl and Mackenzie：前掲注（24），pp.25. 邦訳47頁.
(27) 米盛：前掲注（12），36–40頁.
(28) Pearl and Mackenzie：前掲注（24），pp.33. 邦訳59頁.
(29) 大渕・前掲注（23），63–64頁.
(30) 同書，51–55頁.
(31) 樋笠尭士「自動運転（レベル2及び3）をめぐる刑事実務上の争点：レベル2東名事故を手がかりに」捜査研究847号（2021年），46–62頁. 日原拓哉「AI のブラックボックス性が法的議論に与える影響」渡辺豊・根津洸希（編）

『AIと分かりあえますか？：ブラックボックスが生まれるしくみ』（信山社・2024年），13–31頁.

(32)　シベリアンハスキー犬の画像をAIが誤って「狼」と分類した事例の研究では，AIが予測の根拠とした画像領域が動物の「顔」の部分ではなく，背景の「雪」であったことが示されている．学習データが「雪中の狼」と「雪のない場所でのシベリアンハスキー」に偏っていたため，AIは動物の顔や体ではなく，「雪」の有無によって分類するという学習をしてしまったことになる．Ribeiro, M. T., Singh, S. and Carlos, G., "Why Should I Trust You?": Explaining the Predictions of Any Classifier, KDD '16：Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016, pp. 1135–1144.

(33)　大坪直樹・中江俊博・深沢祐太・豊岡祥・坂元哲平・佐藤誠・五十嵐健太・市原大暉・堀内新吾『XAI（説明可能なAI）：そのとき人工知能はどう考えたのか？』（リックテレコム・2021年）.

(34)　同書，29頁.

(35)　拙論「自動運転車の社会受容」『中央学院大学人間・自然論叢』第53号（2022年）.

[付記]
本研究はJSPS科研費JP21K00015の助成を受けたものである.