新進研究者 Research Note

毒パズルと脳走査装置
The Toxin Puzzle and The Brain Scanner

佐藤 広大

Abstract

The toxin puzzle is a famous thought experiment. In my view, this puzzle consists of several questions. In this research, I deal with the one question: is the toxin puzzle possible in principle? In this question, there are two questions: is the 'mind-reading' brain scanner which will correctly detect the presence or absence of the relevant intention possible in principle; if this scanner is not possible in principle, then is the toxin puzzle not worthy of consideration? I answer, to these two questions, that the brain scanner probably is not possible in principle, and that even if so, the toxin puzzle is still worthy of consideration.

(1) 研究テーマ

本研究のテーマは、G・カフカによって提出された「毒パズル（the toxin puzzle)」である（Kavka 1983)。毒パズルは次のような思考実験である。あなたが大富豪にある取引を持ちかけられるとする。すなわち、もし毒を明日の午後に飲むことを今夜の午前 0 時に意図できたら、それだけで大富豪があなたに 100 万ドルを明日の朝に支払うという取引である。意図できたかどうかは脳走査装置で確認されることになっている。その毒を飲むとまる一日ひどく苦しむことになるが、死んだり、後遺症が残ったりはしない。

あなたは 100 万ドルを獲得できるだろうか。100 万ドルを獲得することなど簡単だと最初は思われるかもしれない。なぜなら、毒を実際に飲まなくても、毒を飲むことを意図するだけで 100 万ドルが獲得できるという取引だからである。しかし、考えてみてほしい。「毒を実際に飲まなくてもいい」と思いながら、毒を飲むことを本気で意図することなどできるのだろうか。そのようなことは非常に困難に感じられる。「毒を実際に飲まなくてもいい」と思うことと、毒を飲むことを本気で意図することとが両立しないことに注目すれば、「毒を実際に飲まなければならない」と思う戦略が有望に思われてくるかもしれない。だが、「毒を実際に飲まなければならない」と思うことなどで

きるのだろうか。もし毒を飲むことを意図する前に、「毒を 100 万ドル獲得後に実際に飲む理由がない」と思っていたとしたら、「毒を実際に飲まなければならない」と思うことは難しいだろう。すると、毒を飲むことを意図することも難しいということになり、100 万ドルを獲得することも実は難しいということになる。毒パズルがパズルたるゆえんは、簡単だと最初は思われる100 万ドルの獲得が、よく考えてみると実は難しいという点にある。

以上の説明を読んでいて、腑に落ちていないところがあるとすれば、それは、毒パズルではそもそも何が問題になっているかが見通しにくいからかもしれない。筆者の見立てでは、毒パズルは実は複数の問題（たとえば、「100 万ドルを獲得できるか」という問題や、「毒を飲むことは合理的か」という問題）から構成されている。毒パズルが複数の問題から構成されているということを自覚しさえすれば、毒パズルは行為論における主要な問題、すなわち「意図と何か」という問題や「合理性とは何か」という問題それぞれについて考える機会を与えてくれるだけでなく、意図や合理性といった基本的な概念同士の関係について考える契機にもなるような豊かなパズルなのである。

本研究は、毒パズルを構成するそれら複数の問題のなかでも、毒パズルが原理的に可能かという問題の一側面を扱う。

(2) 研究の背景・先行研究

本節では、毒を飲む意図の有無を読み取る脳走査装置が原理的に可能かという問いと、そのような脳走査装置が原理的に不可能だとしたら毒パズルは論じるに値しないものになるのかという問いの二つを確認する。

第一の問いについて、毒パズルという文脈のなかでは、毒を飲む意図の有無を脳走査装置で読み取ることなど不可能だと主張する人々がいる。たとえば、D・ゴティエは「そのような機械は不可能だと思う」と述べている（Gauthier 1998, p.47）。

毒パズルの文脈を離れてみると、そもそも脳の状態と心的状態はどのような関係にあるのかという問題や、脳の状態から心的状態を読み取ることができるのかという問題は、心の哲学のなかではおなじみの問題である[i]。

そのなかから、比較的新しい古田徹也の議論を取り上げてみよう。古田は「心の働きは脳内の物理的過程に付随する」という命題を否定している（古田 2013, p.130）ので、毒を飲む意図の有無を脳走査装置で読み取ることができるというカフカの考えも否定するだろう。

古田は、意図や信念の特徴をいくつか挙げた（*ibid*, pp.65-76）うえで、意図や信念が脳内の物理的過程に付随していながら、そうした特徴を持つと考

えると無理が生じると述べている（*ibid*, pp.76-79）。たとえば、古田は、長時間持続しうるという意図の特徴を取り上げたうえで、もし意図することが脳の特定の活動だとすると、意図が持続している間、その意図に対応する脳の特定の活動もずっと持続しているということになってしまい無理が生じると主張する。

他方で、近年、脳の信号を読み取って機械を操作することや、機械から脳に刺激を与えることを可能にする機器 Brain-Machine Interface（BMI）が、マインドリーディングの文脈で盛んに研究されている。意図が脳の特定の活動だということを前提にしている BMI は実用化もされていて、意図の有無を読み取るような脳走査装置の実現可能性が高まっているようにも見える。

BMI は、手術をして体の中に埋め込むのか（「侵襲的」）それとも手術せずに体の外から測定するのか（「非侵襲的」）という軸（「体に対する影響という軸」）と、機械から脳に情報が伝達されるのか（「入力型」）それとも逆に脳から機械に情報が伝達されるのか（「出力型」）という軸（「情報の伝達の向きという軸」）の二つの軸で分類される（長谷川 2008，pp.1066-1069；吉峰他 2016，pp.964-966）。

ここでは、毒パズルに関係しそうな実例として、侵襲的な出力型の BMI を利用した、ラットが脳波でロボットアームを動かすという実験を取り上げてみよう（Chapin et al. 1999）。まず、ラットが前足でレバーを押すとロボットアームが動き、水を 1 滴飲めるようにする。ラットの脳の運動野などにあらかじめ電極を刺しておき、ラットがレバーを押すたびにニューロン集団の活動を検出する。次に、ラットがレバーを押す直前に発生するニューロン集団の活動を検出するとロボットアームが動くようにし、レバーを押してもロボットアームは動かないようにする。つまり、ラットがレバーを押さなくても、レバーを押す直前にラットの脳で発生するニューロン集団の活動を検出するだけで動くように設定されたロボットアームという BMI を用意したことになる。すると、ラットはレバーを押さずにロボットアームを動かすようになる。すなわち、ラットは、実際にレバーを押さずに、レバーを押す直前に発生するニューロン集団の活動（つまり、レバーを押す「意図」）を発生させることができたように見える。

第二の問い、もし脳走査装置が原理的に不可能だとしたら、毒パズルは論じるに値しないものになるのかという問いについてはどうだろうか。

脳走査装置が原理的に不可能なら毒パズルは論じるに値しないと諦めてしまうのは早い。原理的に不可能なおそれがあるのは、毒パズル全体ではなく、毒を飲む意図の有無を脳走査装置で読み取るという部分だけである。その部

分さえ取り除いてしまえば、毒パズルを論じる価値が残るかもしれない。

たとえば、M・ブラットマンは、意図の有無を脳走査装置で読み取るという部分が取り除かれているのに毒パズルと似た、より日常的な次のようなケース（「恩返し事例」）を挙げている（Bratman 1998, pp.63f）。なお、以下の事例は、分かりやすさのために、オリジナルの事例の細部を変更したものである。飛行機のなかで、私が頭の上の棚にスーツケースを入れていた。そこにあなたが来て、あなたもその棚にスーツケースを入れた。飛行機が目的地に到着して、スーツケースを降ろすことになった。スーツケースを各自で降ろすよりも、協力して一緒に降ろす方がよいとする。ただし、私とあなたの座席の位置の関係から、まず私があなたのスーツケースを降ろすのに協力し、次に、あなたが私のスーツケースを降ろすのに協力することになる。私があなたのスーツケースを降ろすのに協力した時点で、自分のスーツケースを降ろすというあなたの目的は達成されてしまうということに私とあなたはお互いに気づいている。「あなたが私に恩返しするだろう」と、つまり、「あなたのスーツケースを降ろす際に私は協力してあげたのだから、私のスーツケースを降ろす際にもあなたは協力してくれるだろう」と私があなたのスーツケースを降ろす際に確信しているときにだけ、私はあなたのスーツケースを降ろすのに協力するだろう。

毒パズルと恩返し事例の共通点を強調するブラットマンに対して、ゴティエは相違点を強調する（Gauthier, *op. cit*, pp.50f）。その相違点とは意図が果たす重要性の違いである。毒パズルのなかで、大富豪にとって重要なのは、あなたが毒を飲むことを意図するかどうかであって、毒を実際に飲むかどうかではない。一方、恩返し事例のなかで、私にとって重要なのは、あなたが私のスーツケースを降ろすのに実際に協力するかどうかであって、協力することを意図するかどうかではない。ゴティエが相違点をこのように強調している理由は、毒パズルと恩返し事例では、あなたが相手（毒パズルでは大富豪、恩返し事例では私）に利益を与えるのが相手よりも先なのか後なのかという違いがあるからである。毒パズルでは、あなたが先に大富豪に利益を与える。つまり、あなたが毒を飲むことを意図することが大富豪の利益となり、その見返りとして大富豪は 100 万ドルをあなたに支払う。一方、恩返し事例ではあなたは後に私に利益を与える。つまり、私があなたのスーツケースを降ろすのに協力することがあなたの利益となり、その見返りとして今度はあなたが私のスーツケースを降ろすのに協力する（*ibid*, p.53）。

(3) 筆者の主張

前節では、脳走査装置は原理的に可能かという問いと、もし脳走査装置が原理的に不可能だとしたら毒パズルは論じるに値しないものになるのかという問いの二つを確認した。一つ目の問いについては、脳走査装置は原理的に不可能だと主張する議論として古田の議論を取り上げた。一方で、脳走査装置の成立可能性を高めているように思われるものとして BMI を取り上げた。二つ目の問いについては、もし脳走査装置が原理的に不可能だったとしても、毒を飲む意図を脳走査装置で読み取るという部分を取り除き、毒パズルを恩返し事例のようなものとして扱えば、毒パズルは依然として論じるに値すると主張する議論としてブラットマンの議論を取り上げた。一方で、意図の果たす重要性という点で毒パズルは恩返し事例とは異なるので、毒パズルを恩返し事例のようなものとして扱うことはできないと主張する議論としてゴティエの議論を取り上げた。

これら二つの問いそれぞれにどのように答えるべきだろうか。

一つ目の問いに対して、筆者は、脳走査装置が原理的に可能な見込みは現時点では高まっていないと答えたいが、その根拠は古田とは異なるものである。脳走査装置が原理的に不可能だと考える古田の根拠は、意図が脳内の物理的過程に付随するにもかかわらず、概念的な特徴（たとえば、長時間持続しうるという特徴）を持っていると考えると、無理が生じるというものだった。しかし、この根拠は、長時間持続しうる意図が脳内の物理的過程（ニューロンの発火）にずっと付随していると考えてしまうと無理が生じるということを示しているだけであって、長時間持続しうる意図の形成がその時点の脳内の物理的過程に付随すると考えることを、つまりは、毒を飲む意図の有無を脳走査装置で読み取ることができると考えることを否定するものではない[ii]。したがって、古田の議論は、脳の物理的状態から心的状態を読み取る可能性を論駁するものではない。

しかし、古田の議論が上手くいっていなかったとしても、そのような読み取りが不可能だということは依然としてありうるように思われる。なぜなら、筆者の考えでは、現在の BMI は、脳状態から心的状態を読み取る可能性を少しも高めてはいないからである。そのように考える根拠は二つある。第一に、筆者の知る限り、BMI 研究では、行為の直前の意図の読み取りが中心に研究されていて、毒を明日の午後に飲むことを目指す今夜の午前 0 時の意図といった、遠い未来の行為を目指す意図を読み取ることに関する研究がほとんど存在しておらず、遠い未来の行為を目指す意図が脳のどのような状態と関係しているかが明らかではないことである。第二に、行為の直前の意図の読み取りだけを考えてみても、BMI によって意図の内容まで読み取ることができ

ているようには見えないことである。先程挙げた J・K・チェーピンらの実験で、レバーを押すという内容を持ったラットの「意図」を読み取ることができているように見えたのは、実験環境が極めて限定されているからではないだろうか。この実験では、ラットは水を飲むためにレバーを押すように訓練されていて、チェーピンらもレバーを押すという意図を検出することに専念している。このように環境が限定されているからこそ、検出されたニューロンの発火に、レバーを押すという内容の意図を帰属させられるのではないだろうか。さらに、筆者には、毒を明日の午後に飲むという複雑な内容を持った意図を読み取るためにどのように実験環境を設定すればよいのか見当さえつかない。

二つ目の問いに対して、筆者は、ゴティエの議論に同意し、意図の果たす重要性という点で毒パズルは恩返し事例とは異なるので、毒パズルを恩返し事例のようなものとして扱うことはできないと答える。意図の有無に主眼を置いた毒パズルという特殊な事例のなかには、日常的で理解しやすい恩返し事例と置き換えてしまうと、失われてしまうものがある。

一つ目の問いと二つ目の問いに対する筆者の答えをまとめると次のようなものになる。脳走査装置が原理的に可能な見込みは現時点では高まっておらず、かつ、もし脳走査装置が原理的に不可能な場合、毒パズルを恩返し事例のようなものとして扱うと失われてしまうものがある。毒パズルが原理的に可能かという問いに対しては、毒パズルにおいて脳走査装置が欠かせないならば、毒パズルが原理的に可能な見込みも現時点では高まっていないと答えることになる。

(4) 今後の展望

まず、二つ目の問い（もし脳走査装置が原理的に不可能だとしたら毒パズルは論じるに値しないのかという問い）について今後の展望を述べていく[iii]。

二つ目の問いに対して、筆者は、もし脳走査装置が原理的に不可能だった場合、毒パズルを恩返し事例のようなものとして扱うと失われてしまうものがあると答えたが、だからといって、脳走査装置が原理的に不可能だった場合、毒パズルを救い出すことがまったくできないと主張しているわけではない。恩返し事例のようものとして扱っていては、毒パズルを救い出すことができないと主張しているだけである。それでは、脳走査装置を使わずに、かと言って、恩返し事例に同化させてしまうことなく毒パズルを救い出すためにはどのような方法があるだろうか。たとえば、毒を飲む意図の有無を読み取る脳走査装置が不可能だと考えていたゴティエは、意図の有無を読み取る

脳走査装置の部分を、あなたが本当に意図しているかを鋭く見抜く人に置き換えて考えている（Gauthier, *op. cit*, p.47）。毒パズルを成り立たせるためには、「見抜く人」が全知の神のようにあなたの意図を完璧に読み取る必要はなく、あなたが「見抜く人」をだますよりも本当に意図した方が楽そうだと（つまり、意図したふりをしてだますという戦略は見込みがないと）考える程度に正確に読み取るだけで十分である。このような「見抜く人」を導入することによって毒パズルを救い出すことができるかもしれない。このとき、「見抜く人」の代わりに全知の神を導入するとどうなるだろうか。哲学の議論では、全知の神のような存在が当然のように不可欠な前提とされていることがあるが、そもそも神のような存在を導入することはいかにして正当化されるのだろうか。こうしたより一般的な観点から、神という存在の導入について検討することになる。あるいは、「見抜く人」よりも現実的な嘘発見器のようなものを導入したほうがよいだろうか。毒パズルを恩返し事例と同化させてしまうことなく救い出す方法を今後このように考えていくことができる。

それでは、本研究全体の今後の展望はどのようなものになるだろうか。本稿の冒頭で、筆者は、毒パズルが実は複数の問題から構成されていると述べた。毒パズルが原理的に成立可能かという問題についての本研究の考察が、毒パズルを構成する他の問題とどのような関係にあるのかということも考えていく必要がある。そのように考えていくことによって、意図や合理性に関する概念的な制約と毒パズルの成立可能性がどのような関係にあるのかということも明らかにしていきたい。そのためにも、毒パズルに対する変更を最小限にとどめ、できるかぎりオリジナルに近い形のまま保っておくことによって、そのパズルとしての豊かさを守っていく必要がある。

---

[i] 脳の物理的な状態と心的状態が同一性を満たしていたとしても、その同一性がトークン同一性の場合、脳の状態から心的状態を読み取ることができるとはかぎらない（「非還元的な物理主義」）。また、心的状態がなんらかの物理的状態に付随していたとしても、その物理的状態が脳を含んだ周りの環境全体であることも可能であり、その場合、脳の状態だけから心的状態を読み取ることはできない（「外在主義」）。

[ii] さらに、古田のように、意図の持続が脳の物理的過程に付随していると考える必要はない。たとえば、意図の持続は脳の物理的な状態（意図の形成によって変化した脳の静的な状態）に付随していると考えることもできるだろう。

[iii] 一つ目の問い（脳走査装置は原理的に可能かという問い）に対して、筆者は、脳走査装置が原理的に可能な見込みは現時点では高まっていないと答えたが、その一つ目の根拠は、現時点の BMI 研究では、遠い未来の行為を目

指す意図を読み取ることに関する研究がほとんど存在していないことだった。しかし、遠い未来の行為を目指す意図を読み取る研究のための足掛かりが全く存在しないというわけではない。遠い未来の行為を目指す意図は「展望的記憶」と深く関わっていると考えられる。「展望的記憶」とは、「意図した行為をタイミングよく自発的に想起すること（梅田 2010，p.2)」である。たとえば、毒を明日の午後に飲むことを今夜の午前 0 時に意図した場合、そのように意図したことを明日の午後に思い出せなければならない。未来の行為を目指す意図と展望的記憶の関係についての研究に注目することによって、未来の行為を目指す意図の読み取りについて研究するための糸口を見つけ出すことができるかもしれない。

（慶應義塾大学）

(5) 参考文献

Bratman, M, 1998, “Toxin, Temptation, and Stability of Intention.”, in Coleman & Morris. eds, 59-83.

Chapin, J. K. et al, 1999, “Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex.”, *Nature Neuroscience* 2, 664-679.

Coleman, J. L. & Morris, C. W. eds, 1998, *Rational Commitment and Social Justice: Essays for Gregory Kavka*, Cambridge University Press.

Gauthier, D, 1998, “Rethinking the Toxin Puzzle.”, in Coleman & Morris. eds, 47-58.

Kavka, G, 1983, “The Toxin Puzzle.”, *Analysis* 43, 33-36.

梅田聡，2010,「し忘れはなぜ起こるのか：認知神経心理学から見た展望的記憶研究」,『認知リハビリテーション』, 1-10.

櫻井芳雄，2013,『脳と機械をつないでみたら―BMI から見えてきた』, 岩波書店.

戸田山和久他編，2003,『心の科学と哲学―コネクショニズムの可能性』, 昭和堂.

信原幸弘編，2004,『シリーズ心の哲学 I ―人間篇』, 勁草書房.

長谷川良平，2008,「ブレイン-マシン インタフェースの現状と将来」,『電子情報通信学会誌』91，電子情報通信学会，1066-1075.

古田徹也, 2013,『それは私がしたことなのか―行為の哲学入門』, 新曜社.

吉峰俊樹他, 2016,「ブレイン・マシン・インターフェイス（BMI）が切り開く新しいニューロテクノロジー」,『脳神経外科ジャーナル』25, 964-972.