人工意識

内容

相関性の検証可能な変数群としての意識の区別

意識の明らかな性質として、自分の意識と他人の意識は、別個のものと区別できる。何が別なのか考えよう。例えば、A さんは、A さんの目からの視覚情報を元に、A さんの肉体をどう運動するか決定できる。B さんについても同様である。しかし、B さんの視覚情報から A さんの運動は決まらない。分離脳と呼ばれる左右の大脳が切り離された場合でも同様に、視野の左側の情報によってのみ左半身の運動が決まり、左右で意識が別個となっている。情報には、視覚や聴覚の 5 感からの感覚情報だけでなく、何が見えているかといった高度な情報も意識上には存在する。体をどう運動するべきかというのもまた意識上の情報である。A さんの視覚情報、聴覚情報は、A さんの意識に帰属するように、意識が別個ということは、さまざまな情報が別個のグループに属するといえる。A さんの視覚情報と、A さんの聴覚情報は同グループだが、A さんの視覚情報と、B さんの視覚情報は別グループである。何がグループを分けるのだろうか。A さんの視覚中の物体が動き、同時に、A さんの聴覚に何かが聞こえたとき、A さんは、その物体が、その音を発したと認知できるだろう。B さんの視覚と、A さんの聴覚の組み合わせでは不可能だ。同じ意識というグループに属する情報同士は、同時に起こったかどうかというような相関関係を認識可能である。A さんと B さんの脳は接続されていないため、情報の相関性を認識可能かという基準では、明らかに別グループに分けられる。ただし、逆に、ネットワークが何らかの形で繋がっているというだけでは、同一の意識グループとはいえない。

時刻による神経の反応対象の変化

「相関性の検証可能な変数群」という意識の性質が脳のニューラルネットワークでどう実装されているか考えよう。ある視覚情報と、ある聴覚情報が、同時に起こったと認知した場合を考える。ある視覚情報が意識上にあるときに発火しているニューロンと、ある聴覚情報が意識上にあるときに発火しているニューロンと、同時に起こったと認識できたときに発火しているニューロンが存在するはずである。そして、前 2 つのニューロンの論理積(AND)で 3 つ目のニューロンが発火するようなネットワーク接続があると予測できる。では、そのような論理積のネットワークが相互にある変数の集合を意識と仮定してみよう。視覚情報として n ピクセルの明暗を表す変数があるとすると、2 つのピクセルの論理積で発火する n の 2 乗個のニューロンが必要となる。さらに高度な認知をするために、3 つ以上の論理積や、視覚以外の情報との論理積まで考えていくと、組み合わせ爆発で、脳に入りきらない無数のニューロンが必要になってしまう。実際の脳がどう対処しているかは 2 つの可能性が考えられる。①一つ目は、ニューロンの数には限りがあるため、あらかじめネットワークが繋がっている組み合わせの範囲でのみ、変数間の相関性を認知できるのではないか。コネクトームによってフレームが限定されているとも解釈できる。いくら熟考しても、決して気づくことができない規則というものが存在することになる。②2 つ目の方法は、一つのニューロンが、ある決まった論理積だけを表すのではなく、複数の論理積を表すことが可能になっているのではないか。この方法なら、あらかじめ思考できるフレームが限定されてしまうということはない。しかし、1 つのニューロンが、複数(大量)のニューロンと同じ役割を果たすことは可能だろうか。例えば、時刻によって一つ

のニューロンが、A^B に反応したり、A^C に反応したりと変化すれば、可能である。ランダムに複雑に繋がっているネットワークはカオスな挙動をするため、それが可能である。例えば初期状態で視覚野 V1 に対して、V2 がランダムに接続しているとする。V2 のある神経は、時刻によって隣り合う画素の陰影差に反応したり、遠く離れた画素との陰影差に反応したりと変化する。後者は意味のない情報のため、他の神経と規則的な発火はしない。前者は映像の境界(コントラスト)を表す情報であり、他の神経と規則性を持って発火する。そのため、前者は後者よりも他の神経より同時発火する頻度が高いため、ヘブ則により接続が強化される。結果、その神経は、時刻により反応する対象が変化することなく、特定の入力だけに反応するようになる。

神経回路のヒステリシスによる短期記憶

下層の神経が反応する対象が時刻と共に変化する場合、上層の神経は、望みの反応が得られる時刻になるまで、同じ入力をし続けなければならない。全く同じ映像を見続けているのではあれば良いが、音声や一瞬だけ見えた映像は、その瞬間しか認知する機会がないのでは困る。意識には、今まさに見えている映像や音声だけが認知できるのでなく、短期記憶された情報をイメージ(想像)として保持でき、想像に対しても認知ができる。一瞬だけ見えた物体の映像の情報が上流から中流へ流れた段階で、物体が視界から消えて上流からの流れが途絶えたとしても、中流から下流へ情報を流し続けることができる。また、虚空に存在しない物体を想像することで、上流からの入力なしに、中流から流れを発生させることもできる。一瞬だけ見えた映像の情報を短期記憶するにはどうすれば良いだろうか。脳はシナプスの可塑性によって記憶することができるが、一瞬だけ流れた電気信号で変化させられるネットワークはたかが知れている。仮に僅かな信号で神経回路が大きく変化するとしたら、ノイズの影響をもろに受け、過学習のような状態に陥るだろう。しかし、神経回路を一切変更せずとも情報の保持は可能である。回路にヒステリシスがあればよい。例えば、出力が入力にループバックしていればよい。スピーカーの音がマイクにループバックする状態では、ほんの僅かな音さえ入力すれば、ハウリング音を出力するが、ハウリングしている状態としていない状態で、電子回路は一切変化していない。

シナプス可塑性では、僅かな信号で短期記憶ができないために、脳は必然的に、ヒステリシスループによる情報保持の仕組みを手に入れたのだろう。仮に短期の情報保持が出来ない場合を考えてみよう。視覚映像から外敵を認知して、振り返って後ろに逃げようと判断したとしよう。振り返って、視界から外敵が消えて瞬間に、外敵や逃げる必要性を忘れてしまっては、生き残れないだろう。意識上には感覚神経からの映像等のリアルタイムの情報だけでなく、「外敵がいる」「逃げるべきだ」という感覚の入力とは独立した変数を維持しなければいけない。感覚入力と独立した変数とは、意識上で考えたり、想像したりする状態に当たる。脳の進化の過程で、初めは偶然にできた回路のループによって、情報を保持できるようになった種が、生存に有利のため生き残ったのだろう。

脳のループ回路

ヒステリシスによる情報保持が可能となる神経回路を考える。その要件は３つ。①ループする回路が存在すること。②同ループ内の神経同士は興奮的に結合すること。③ループ外の神経へは抑制的に結合すること。ループ回路一つが、一つの情報を保持することができる。①②はループ内に信号が流れ続けるために必要である。③は、仮に逆に他のループにも興奮的に結合しているとすると、最終的に全てのループが興奮してしまい、情報が保持できない。②③から、他の神経に対してランダムに結合してはダメで、興奮的に結合する対象神経と、抑制的に結合する対象神経を明確に分ける隔離が必要である。それを実現しているのが大脳のミニコラム構造と考えられる。ミニコラム内の神経は興奮的に結合するが、他のミニコラムへは抑制的に結合しているはずである。そうすることで、ミニコラムを単位として情報を保持するとこができる。①のループ回路としては、大脳皮質⇒基底核、基底核⇒大脳皮質という大きなループがある。この大きなループにより、視覚野の上流から信号が途絶えても、映像を想像し続けたり、このループを発端として、存在しない映像を想像したりすることができる。また、大きなループを一周するのには時間が掛かるため、ごく短時間の情報保持については、ミニコラム内の小さなループによって行われると考えられる。また、神経の全てが興奮したりするのを防ぐ仕組みが必要である。①②だけでは、信号を流し続けるほどの興奮を維持できず、基底核からバイアスによってループ状の興奮を維持できると考えられる。このバイアスによって意識レベルや”注意”を制御できる。ループ信号が流れている状態が、その情報が意識に上がっていると考えられ、脳の特定の領域へのバイアスを増減することで、特定のことを意識し易くなるのが”注意”であると考えられる。

大脳のミニコラムの数が意識の容量であり、発火しているミニコラムの数が意識レベルに相当する。大脳以外でもループしている回路があれば、そこには意識があるといえる。ただし、仮に小脳に意識があったとしても、分離脳が他方の意識を感じられないのと同じように、大脳にある意識は、小脳にある意識を感じられない。

大脳皮質の神経回路と学習

ループ回路が機能するために、ミニコラムは、上層や基底核からの入力に対しては、興奮し、下層や同層の他のミニコラムからの入力に対して抑制しなければならない。ミニコラムは１一般に６層構造が知られており、視床から入力がある第４層が、興奮的な結合を担っていると予想できる。他の層への入力は抑制的だと予想できる。

ここで、脳梁が切断された分離脳について考えてみよう。分離脳では同時刻に、左右の脳は意識が独立し、全く別のことを考えている。しかし、脳梁が接続されていれば、同時刻に全く別の２つのことを考えることはできないだろう。同一の意識の中では、あるミニコラムが興奮すれば、競合する他のミニコラムを全て抑制しよう働いていると考えられる。競合するミニコラム同士は、互いに抑制するため、どちらかが優位な状態になり安定する。どちらも興奮している状態は、迷っている状態に相当するだろう。ある物体が、左回りに回っているように認知した状態と、右回りに回っているように認知した状態は、同時に起こらないように、同層の競

合するミニコラムへは抑制的に働く。

下層へのフィードフォワードは興奮的で、上層へのフィードバックは抑制的である必要を述べる。上層のミニコラムは線を表し、下層のミニコラムは文字を表すとしよう。どんな線の組み合わせかという上層から下層への興奮的な入力は、これを認識したいという”出題”に当たる。逆に、下層から上層への抑制的なフィードバックは、”回答”に当たる。学習済みの文字ならば、即座に特定の文字を表す上層のミニコラムが興奮し、文字を構成する線を表す下層のミニコラムを抑制される。下層が抑制されて出題が止まる。抑制しすぎると、フィードフォワードも弱くなり、フィードバックも弱くなる。興奮と抑制が釣り合ったところで平衡状態になる。それが、文字を認知し終えた状態に当たる。認知し終えれば、無関係なフィードフォワードによる興奮が収まるため、それ以上、それが何という文字か考えることはないだろう。次に、学習していない文字を見た場合を考えよう。抑制的なフィードバックがないため、上層をランダムに広範囲に興奮させることになる。何度か入力が繰り返されれば、どの文字にも使われていないミニコラムの一つが偶然に興奮した後、ヘブ則で結合が強化され、その後は、特定のミニコラムが反応するようにして、学習する。学習が進めば、下層の広範囲への”出題”の興奮を撒き散らしは穏やかになり、正解の文字へのフィードフォワードだけが残るだろう。しかし、一つの神経細胞を複数の文字の認識に使用しているため、他の文字を学習した結果、ある文字へのフィードフォワードが正しく行われなくなる場合がある。いわゆるど忘れである。ど忘れした場合、”注意”によってバイアスを掛け、周囲の発火頻度上げたり、関係しそうなことを考えて、連想によって、正解の文字を再び発火させようとしたりするだろう。どんな方法でも良いので、正解の文字が発火さえすれば、ループによるヒステリシスのため、その後はすぐに思い出せるように戻るだろう。

ニューラルネットワークと意識

一般的なディープラーニングのようなニューラルネットワークに意識はあるだろうか。一般的な DNN は、入力側から出力層へ向けて一方向に流れ、途中でループし情報が保持されることはなく、最終層へ到達すれば計算は終了する。層が深い程、より高次な思索できるが、層の数は有限である。一方、脳は積み重なった層があるのではなく、横方向に繋がっている。脳では、一方向ではなく、戻る経路、ループした経路もある。DNN が、各層それぞれが、どのくらい高次の情報を扱うのかあらかじめ決まっているのに対して、脳では決まっていない。脳では、周辺の空いているミニコラムを使用して、より高次の論理を学習できる。ミニコラムの空きがある限り、いくらでも高度な概念を学習可能である。どこが低次で、どこが高次の部分になるのかあらかじめ決まっていないので、DNN のようにネットワーク全体に低次から高次へ情報を流すことはできない。脳では学習済みのところまでは情報が即座に流れるが、その末端ではループにより発火が維持される。そこから先にどう流れるかは学習により決まる。仮に、DNN にループによる情報を保持する仕組みを取り入れたとしても、有限な各層の役割があらかじめ決められていては、全く意味がない。

直感と思考とクオリア

脳で、学習済みのミニコラムが発火するまでが”直感”であり、ループで維持された情報を使った処理が”思考”に当たる。直感は反射的なもので、例えば赤いものが見えているのに、赤いと感じるのを意識的に止めることはできない。直感によって自動的に情報が意識上にロードされる。意識上の情報の規則性を調べたり、関連する情報をロードしたりするのが思考である。思考するためにはループによる情報保持が必要だが、直感のためだけなら必要ない。

思考するためにはループ回路が必要だが、意識上でクオリアを感じるためには、ループしている必要はないと考えられる。例えば、大きなループ回路上で、電気信号が一周した瞬間の状態を考えてみよう。一周せずに途中から流れ出した場合でも、物理的に全く同じ状態を取ることができる。すなわち、悪魔的なものが監視でもしていない限り、一周したかどうかを調べる方法は存在しない。一周するのもしないのも物理的に全く区別できないのだから、一周することでクオリアが発生するとは考えられない。単に、ミニコラムが発火さえしていれば、クオリアを感じていると考えられる。また、ミニコラムとそれ以外の神経回路を明確に区別できないため、ミニコラムに限らず、単に神経細胞が発火するだけでもクオリアが発生していると考えられる。ただ、ミニコラムに準ずるものでなければ、互いに抑制し合い競合することができないので、分離脳のように意識が分断されてしまい、他所のクオリアは感じることができないだけである。

クオリアは、神経細胞の発火そのものと推測できるが、決して証明することはできない。クオリアは主観現象で、神経発火は客観現象であり、それらを直接結びつける手段はない。客観現象同士ならば証明は可能である。例えば、BTB 溶液を赤変させる化学物質を明らかにしたい場合を考えよう。いくつかの酸性の物質を試せば、酸性なら赤変すると予測できるだろう。しかし、いくら沢山の物質を統計的に調べても、全ての酸性の物質で赤変すると証明にはならない。しかしながら、化学変化のメカニズムを明らかにすれば、統計的に調べることなく、どういった物質で赤変するか証明できる。クオリアの場合でも、どういった状態でクオリアを感じるかを調べれば、統計的に予測することはできる。しかし、理論的に証明する手段はない。また、統計的な方法でも、還元的に調べられる限界がある。例えば、神経細胞の電圧がクオリアだとする仮説と、電流がクオリアだとする仮説を立てたとしよう。電圧が 0 だけど電流が流れている状態を作らなければ、どちらが正体か分からない。また、電子によってクオリアが発生するのか、あるいは反物質・陽電子でもクオリアが発生するのかというように、いくらでも検証困難な仮説を立てることができるため、クオリアの必要条件、十分条件は予測できるが、必要十分条件は決して分からない。

クオリアが物理的な何に対応しているかといえば、何らかの物理量であるということだけは確かである。物理量といってもさまざまなものがあるが、還元的に考えれば、例えば”力”のような基本的な物理量になるだろう。

クオリアについては、赤や緑に感じる差異はなんであるか、他人と同じであるかという議論があるが、その差異は、ネットワークの繋がり方の差異と考えられる。例えば、色を一度も見たことが無い人が、赤と緑のものを初めて見たとき、それらの色は、ただ互いに違うものであるとしか感じないだろう。普通の人は、無意識に赤いもの・緑のものを連想しているため、他

人が逆のものを連想しているとしたらおかしいと感じるだけである。仮に質的な差があり、他人は逆だとしても、記憶をもとに連想するものは同じなので、質的な差をあってもなくても全く同じである。オッカムの剃刃的にクオリアの質的な差は不要である。２つのクオリアの間には、ただ「同じ」「異なる」という２つの状態しか存在しない。「同じ」というのは同じミニコラムが発火していることに対応している。

脳の完全なシミュレーションをしても意識は存在しない

量子論的な限界さえ超えて、脳内の全ての原子やあらゆる物理現象を、コンピュータ上で完全にシュミュレーションができたとしても、そこには意識は存在しない。実は、あるトリックを使えば､現在の科学技術でも完全なシミュレーションが可能なので､思考実験しても見よう。ある瞬間の脳の状態を表すには、1 無料大数 bit が必要と仮定しよう。その膨大な情報量をハフマン符号化で圧縮しよう。ハフマン符号化は、出現頻度の高いビット列を短いビット列に、出現頻度の低いビット列を長いビット列に置き換える方法である。ある瞬間脳の状態の情報を「10」というビット列に、次の瞬間の脳の状態を「110」、その次は「1110」という風に置き換えて、圧縮してしまおう。コンピュータ上の数値が「10」「110」「1110」と変化していけば、脳の状態を完全にシミュレーションできたことになる。もっと簡単に、ある瞬間の脳の状態を「1」と置き換えて圧縮してしまえば､「1」と書かれたただの紙であっても、デコードすれば、脳の状態を完全に表してシミュレーションしていると解釈できる。この紙には明らかに意識もクオリアも存在しない。デコードしなければただの「1」なのでインチキだと思われるかもしれない。しかしデコードして解釈しなければ、コンピュータでも全く同じことがいえる。フロッピーディスク上で、「NNNNSSSS(00001111)」という風に磁石が並んでいたら、数値の 15 であると解釈できるが、実際に 15 という何らかの物理量が存在するわけではない。ビッグエンディアン、リトルエンディアンといったような解釈の仕方次第で別の値に表しているといえる。「1」を脳全体の情報を表すと解釈するのと差はない。意識と同等の機能さえあれば、意識があると定義するのなら、コンピュータ上のシミュレーションにも意識があると解釈することもできるが、それでもクオリアは存在しない。クオリアは何らかの物理量に対応していると考えられるため、物理量を勝手に別の数値に解釈した結果にはクオリアは存在しない。仮に、そでもクオリアが存在すると仮定しても、そのクオリアは人間のクオリアとは全く互換性のない別物である。コンピュータ上で、脳を完全にシミュレーションできたとしても、それは意識やクオリアをシミュレーションできただけであって、意識もクオリアも実在しない。

人工意識とマインドアップローディング

現行のコンピュータでは、意識もクオリアも持てないため、マインドアップローディングはできない。例えば、脳を少しずつコンピュータに置き換えていけば、気づかないうちにマインドアップローディングが完了するという思考実験は錯覚である。脳の半分を完璧なコンピュータに置きかえたとき、コンピュータは元の脳と全く反応するため、半分の脳は、もう半分の脳が失われたのに気づかないだけである。コンピュータに置き換えても気が付かないだけで、意

識・クオリアは、置き換えとともに失われてしまう。

意識・クオリアを持てる新しいコンピュータを考えよう。例えば、炭素の代わりにケイ素を使用した神経細胞が全く同じ働きをするとしたら､意識･クオリアを持つので不可能ではない。どこまで似せる必要があるかというと、クオリアにさえ互換性があればよい。例えば「赤い」と感じる度合い（クオリア）が「0～1」という連続値を取るとしよう、その連続値は、ニューロンあるいはミニコラムの電圧・電流・周波数（発火頻度）に対応していると考えられる。コンピュータがクオリアを持つには、AI が「赤い」と感じる度合い「0～1」に応じて、少なくともコンピュータ上の電圧・電流・周波数等の物理量が「0～1」になっていなければならない。少なくとも、デジタルで数値を表現するコンピュータではいけない。例えば、メモリ上の電位や電気量として「0～1」間のアナログな数値を保持できれば、脳と互換性のあるクオリアとなる可能性がある。量子コンピュータでも可能かもしれない。

アナログコンピュータなら意識・クオリアを持つことができるが、デジタルよりも知能としての能力が劣るかもしれない。脳を半導体で再現したアナログコンピュータがあったとして、ニューロンの数が有限だったり、三次元空間上でシナプスを配線する必要があったり、脳同様に反復しないと学習できなかったりという限界がある。究極のコンピュータは、アナログ・デジタルハイブリッドコンピュータになるかもしれない。意識を持たせるためにアナログ部分があり、デジタル部分は脳の限界を超えるために用いる。脳に相当するアナログコンピュータの比率を減らせば、脳から情報を移植するときにも最小限で済む。脳の情報を全てコピーしなければ、一部の記憶が欠落するが、老化等で一部の記憶を失っても、同じ人物だとするなら、マインドアップローディングできたといえるだろう。ハイブリッドコンピュータで本当に意識・クオリアを持ち得るのかというと、次の思考実験をすれば分かるだろう。脳内にシナプスの結合を調節するナノマシンを導入し、また意識と関係ない脳の部位は切除したと考えよう。脳がアナログ部分で、ナノマシンがデジタル部分に当たる。ナノマシンは外部のスーパーコピュータによる超 AI の計算結果を反映して、シナプス結合強度を調節する。また脳のネットワークの内容を、メモリ内容のように外部コンピュータにスワップできるため、脳の皮質の面積・ニューロン数が無限大になったのと同じである。脳の制限を取り払った意識は、汎用人工意識と呼んでよいだろう。アナログ部分は、どこまで脳を再現する必要があるか分からないが、脳をそのまま使っても良いかもしれない。デジタル部分のナノマシンは非現実的だが、アナログ部分を工夫すれば他の方法で接続できる。例えば、アナログ部分は、脳の神経間の距離を離してスケールアップしたものにすれば、他の配線方法も使える。このような手法を使えば、人間の精神をコンピュータに移し、不老不死になったうえで、さらに究極の知能を手に入れることも不可能でない。