■ 大泉匡史さん「意識の統合情報理論」セミナーまとめ(20161129 version)

理研BSIの大泉匡史さんが近畿大学で意識の統合情報理論について7.5時間語り尽くすセミナーをやるというニュースを知って参加してきた。

これまで私は、意識の統合情報理論(Information Integration Theory of Consciousness)についてブログでいろいろと書いてきた。ブログの「トノーニの意識の統合情報理論」カテゴリ参照。でもまあ、いつも手を動かしてモデルの挙動を知るとかそういうレベルで理論を理解しているわけではないので、なんとか時間を取ってこの理論についてちゃんと理解した上で、賛成するなり反対するなり意見を表明できるようにしたいと考えていた。

今回のセミナーは、統合情報理論の最新版をトノーニラボで作ってきた大泉匡史さん本人が、7.5時間かけて充分に詳細を説明しようという趣旨であるようなので、これはいい機会とあらかじめ論文を読み込んでセミナーに臨んだ。

この記事では論文を読み込み、セミナーに参加して色々考えたことをまとめておきたいと思う。この記事では、IITとはなんぞやとかそういう初歩的なことはここでは書かない。概要は大泉さん本人による日本語の解説記事1(Clinical Neuroscience)および解説記事2(LISA)を読むのがよいかと。さらに詳しく知りたければ「意識はいつ生まれるのか　脳の謎に挑む統合情報理論」 トノーニ & マッスィミーニを読んで、Scholarpediaの記事という順番で。

関連する論文は主に三つ。

- 現在の最新バージョンのIIT3.0の原著論文はPLoS Comput Biol 2014。以降「IIT3論文」と呼ぶ。
- それ以降に情報幾何を使ってIITを見直した論文がarXiv 2015で、以降「情報幾何論文」と呼ぶ。
- それから大泉さんがトノーニ研行く前からやってたmismatched decodingに基づいた論文(+ECoGデータ)がPLoS Comput Biol 2014。以降「デコーディング論文」と呼ぶ。

当日のセミナーは4部構成で、

- 統合情報理論とは何か。Axiomsとpostulatesを説明。
- Level of consciousnessについて。主にIIT2.0(Balduzzi & Tononi, 2008)を元にしたΦの定式化。
- Quality of consciousnessについて。IIT3論文におけるconstellation / qualia spaceについて。
- IITの実験による検証。デコーディング論文および未発表論文とヒトECoG論文 biorxiv 2016

となっていた。以降はこのセミナーの内容に沿いながらいろいろコメントしてゆく。以下IIT3論文で読んだことおよび大泉さんによる説明をまとめたうえで、吉田によるコメント部分は煩雑にならない程度に明示しながら書いてゆく。

---

1. 統合情報理論の基本部分

IITの議論の構造としては、意識経験の現象的側面から意識の疑う余地のないものとしてaxiomsを置いて、それを物理的システムで実現するときに要請されるものとしてpostulatesを提示して、それを実現するmechanismとして因果のネットワークの持つ特徴を決める理論的モデルが構築される。

1-1. Axioms

AxiomsはIIT3.0では5つある。(以前はInformationとIntegrationだけだった)

- Exisntence: 意識は存在する
- Information: 意識はinformativeである (ある意識経験がAではなくてBであるというような意味において)
- Integration: 意識はintegrateされている (意識の中にcontentがいくつあっても、それは単一の経験として経験される)
- Exclusion: 意識は排他的である (たとえば、我々の経験は色付きの意識経験と色のない意識経験の両方が成立可能だが、色付きの意識経験があるかぎり、色のない意識経験を同時に持つことはない)
- Compositionality: 意識は構造化されている (我々の経験には右と左、赤と緑といった要素とその組み合わせがある。ノイズ画像(砂嵐)1と2の区別は情報は持っているが、構造化されていない。) 大泉さんによれば、これだけが意識のqualityに関わっており、上記の4つは意識のlevelに関わっている。

大泉さんの説明では、このaxiomというのはIITが考える意識とはこういうものであると区切るものである、とのことだった。それならば納得がいくかもしれない。たとえば「Exisntence: 意識は存在する」を入れた時点で「哲学的ゾンビの可能性」とか「意識とは幻想であり存在しない」という議論はあらかじめ排除される、と宣言しているのであって、IITはそういう意味ではハードプロブレムそのものを相手にしない、と宣言しているとも理解できる。じっさい、脳の状態と意識の状態とは同一であるとするidentityをこの理論では前提としているので、explanatory gapとかそういう議論は射程外にある。

なるほど、そうするとつまりこれはユークリッド言論でいう平行線公準みたいなもので、違ったものも考えうるわけだ。「意識は存在しない」から始まる理論があっても良い。(ついでに言えばaxiomsとpostulatesは数学でいう公理と公準に対応した言葉だが、これは自明さのレベルにおいてIITと数学で同一視するべきではないと大泉さんも言ってた。)

1-2. Postulates

それぞれのaxiomについてそこから要請されるpostulateがある。

- Information: (意識を持つ)システムは情報を生成しなくてはならない (ある意識経験がAではなくてBであるというような意味において情報を持っているならば、これはありうる状態から実際に起こっている状態を選んだという意味で情報理論的扱いが可能となる。

「Informationとは"differences that make a difference"である」という考え方からcause repertoireとeffect repertoireという定式化が行われる。(注1))
- Integration: (意識を持つ)システムは情報を統合しなくてはならない (IITでは意識のcontentを説明したいのではなくて、そのlevelとboundaryを説明することに重きを置く。Integrationは意識の単一性とboundaryを説明するために要請される。)
- Exclusion: Experience is unique. (これはsplit brainの例を考えると分かる。脳梁切断によって2つの意識ができる症例があるが、これは健常者の脳でも2つの意識が左右の半球で成立しうることを示している。しかしそうならないのは、健常者の脳では左右の半球を合わせたひとつの意識が成立することで、左右別々の意識はexcludeされる、という要請postulateを導入したというわけ。)
- Compositionality: Elementary mechanisms can be combined into higher order ones. (これ自体は因果ネットワークのhierarchicalな構造を持っていることと、IITでいう"concept"がそのネットワークのサブセットの因果ネットワークとして形成されるという説明があったが、IIT自体では意識のcontentの議論はまだ不充分であるため、この部分の説明も不十分であるというのが吉田の理解。IIT3論文のFig.22がこれに関わっていることは分かる。)

ここでわたしがまず不満に思うのは「Information (意識を持つ)システムは情報を生成しなくてはならない」についてだ。無意識だって情報を生成するし、informativeである。そう考えてみるとIITのaxioms-postulatesの構造は「意識はこうだ」という話しかしていないのが問題なのではないかと思う。つまり、正しい意識の理論は「意識的状態は無意識的状態と比べてこう違う」という言い方になる必要があるのではないだろうか。

もしこの指摘をinformation postulateに取り込むのならば、「意識を持つシステムは意識を持たないシステムと比べてより多くの情報を生成しなくてはならない」といった、ずいぶんと切れ味の悪い言明になってしまう。Integrationだって、複雑な機能をこなすシステムは無意識でもintegratedである必要があるだろう(機能との関連はIIT3論文Figs.21,22で出てくるので後述)。その点でExclusionはなにが意識に登り、なにが無意識になるかという対比を明示的に議論しているというふうには言える。

あともうひとつ言いたいのは、議論の構造としては現象的経験から導かれたaxiomsからpostulatesができて、それを可能とするmechanismsが決まるというふうになっているが、実際にはいろいろ後付け的にやってるので、「現象からスタートする」というIITの方針をface valueで受け取るわけにはいかない。たとえばexclusionはΦの挙動が神経科学的に整合的になるように後から付加されたものだ。前述のように無意識もinformativeなら、Informationのaxiomからpostulateへの移行には論理的ジャンプがあるともいえる。(この論理的ジャンプについては大泉さんはASCONEでも指摘されたと話していた。)

Compositionalityはinformationおよびintegrationのふたつで説明できてしまうのではないか、つまりaxiomとしてはredundantなのではないかという議論があったけど、参加者の方から「integrateされているけれどもtopographicalな関係がバラバラなネットワークを考えることが出来る」という指摘があって、これには納得いった。ただし、視覚経験が砂嵐みたいになっていても情報を弁別し、単一の意識経験として経験する有機体というのを想定することは可能なの

で、Compositionalityは意識にとっての必要条件ではなくて、あくまで人間の意識経験の分析からは妥当なもの、としか言えないだろう。

---

2. 統合情報量の数学的定式化

ここではまず情報理論の基礎の話からスタートした。内容としてはClinical Neuroscience総説に準拠している模様。

2-1. 相互情報量 (Mutual information)

相互情報量MIは外部の刺激がS=sであることが判明したときにシステムXの不確定性がどのくらい減るかを表現したものであり、KL divergnce (以下式では'D'で表現する)を使って以下のように表現できる。

- MI(X; s) = D(P(X | s) || P(X))

よって、相互情報量MIとは、Xとsとを両方見ることができる外側からの視点(ideal observer)にとっての情報量であり、extrinsicな情報量であるといえる。

2-2. 内的な情報量 (Intrinsic information)

それでは、あるニューロンAにとっての「内的な」情報量というものを考えるとすればどうすればよいか。それは外界の刺激ではなくて、そのニューロンAへ入力するニューロン群の活動パターンとそのニューロンAが出力を送るニューロン群の活動パターンによって規定されると考えるべきではないか。これはさらにニューロン群でも同様に考えられる。

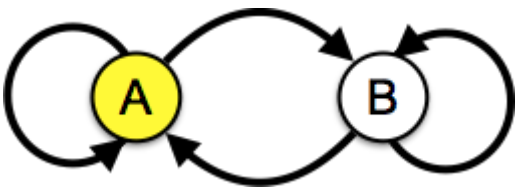

Figure 1

これをシンプルなモデルで説明するために、図1のようなニューロンA,Bの2個からなるネットワークを考えて、それぞれの発火状態X=1,0(黄色が状態1で白が状態0)が離散的な時間t-1, t, t+1,...でどのように遷移するかの規則(TPM: transition probability matrix)が与えられているものを考える。(Clinical Neuroscience総説の図1,2を参照。)

するとたとえば現在tのネットワークの状態がたとえばAB=10に対して、過去t-1のネットワークの状態AB={00, 01, 10, 11}がそれぞれどのくらいの確率で起こるかということが計算できる(cause repertoire: 大泉さんの表現とは違うが、P(AB(t-1) | AB(t))と書ける)。(上記のTPMがP(AB(t) | AB(t-1))そのものなので、ベイズの定理を使えばよい。) そうしたら現在tの状態AB=10であることが判明したことによってどのくらい過去t-1のネットワークの状態ABの不確定差が減るかということが計算できる。これがintrinsic information。

- ci = D(P(AB(t-1) | AB(t |AB=10)) || P(AB(t-1)))

(ひとついうべき点としては、このciは現在のstateごとに計算される。全stateで平均しない。(注2))

(吉田コメント) セミナーでは明確に言ってなかったが、これはAB(t-1) と AB(t)の間での相互情報量MIのことだ。つまり今の文脈では「内的」であるためには統合情報量を定義する必要はなくて、外部の刺激ではなく、ニューロンのネットワークが影響をもらい、与えうる因果的なネットワークで定義することが重要であると言える。

これらの説明で出てくる図の矢印は吉田が理解するかぎり、直接的にcausalな影響を及ぼす関係であることを示していて、間接的なものにはこの矢印を付けない。いっぽうで、投射はあるけれどもシナプスの重みは0であるといったように、実際にはcausalな影響がないこともありうる。

2-3. 統合情報量ϕ (Integrated information)

ではIITでの統合情報量ϕが相互情報量MIとどう違うかというと、ネットワークのpartitionという操作を用いるところ。現在tの状態AB=10のときのcause repertoireとしてP(AB(t-1) | AB(t |AB=10))が計算できるわけだが、統合情報量ではこれをA-B間の結合を切断(partition)した場合のcause repertoire とを比較して、その距離を計算する。式としてはこういう感じ：

- D(P(AB(t-1) | AB(t |AB=10)) || P_partition(AB(t-1) | AB(t |AB=10)))

今見ているようなABだけのネットワークであればpartitionはA-Bを切るだけだが、要素数が大きい系ではこのようなpartitionは膨大な個数ある。たとえば要素が12個あったらpartitionは2^11あることになる。このため、IITではMIP (minimum information partition)ということを考える。ここからは説明のため、一時的に4つの要素のネットワークの話に変える。(ちなみに大泉さんのセミナーではここを端折ってたのでMIPの概念、意義の説明がわかりにくくなっていた。)

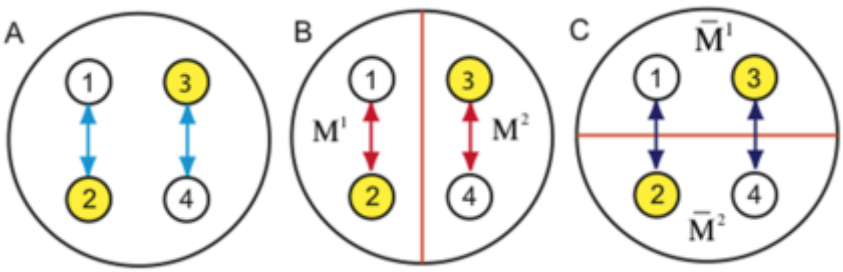

Figure 2 by Balduzzi and Tononi 2008 / CC BY 4.0

図2はIIT2.0論文のFig.3を改変して作成しているが、要素1,2,3,4のネットワークAはじつは1-2および3-4でだけ因果ネットワークが形成されていて、1-2と3-4は独立している。このネットワークを切り分ける方法にはいくつかあるが、BとCの例がここでは示してある。CではpartitionによってD>0となるのがわかると思うが、Bのpartitionでは因果ネットワークに変化がないため、上記距離D=0となる。このようにDを最小とするpartitionをIITではMIP (minimum information partition)と呼んでいる。

そしてIITではMIPでのcause repertoireと元のcause repertoireとの距離を計算したものを統合情報量ϕとして用いる。

- ϕ = D(P(AB(t-1) | AB(t |AB=10)) || P_MIP(AB(t-1) | AB(t |AB=10)))

よって、図2Aのネットワークのϕ=0となる。つまりこのAというシステムは統合された単位ではないということ。改めて1-2および 3-4という2つのサブシステムでの情報統合量を評価する必要が出てくる。おわかりのとおり、この図はsplit brainを想定していて、exclution postulateが実際にどうimplementされているかの説明になっている。以上のようにして、統合情報量ϕはinformation, integration, exclusionのpostulateを実装しているといえる。

(吉田コメント) IITはこのMIPという操作によって意識のboundaryがどこであるかという議論に対して一定の答えを出しているといえる。(sensorimotor contingencyとの関連で後述) まずはMIPでϕ=0となるような末端を削って意識を引き起こしうるネットワークを限定した上で、そのネットワークでさらにϕが最小となるようなpartition (MIP)を見つけて、そのネットワークのϕを決めてやる、という2段構えになっているとも言える。前者のboundaryを決めるところにMIPが必要なのは納得がいくけど、後者のシステム固有のlevel of consciousnessを決めるときにMIPが必要だという理屈はそんなに無いように思えるのだけど。ともあれ、IITが想定しているイメージというのは、世界に広がっている因果ネットワークがϕ=0で切れたboundaryごとにそれぞれのネットワークが一定の極大値として一意に決まる、そういうもののようだ。

(吉田コメント) IITではこのpartitionという作業を入れないとintegration postulateが満たされないと考えている。このpartitionが必須なのか、というところはIITの根幹に関わる問題で、もしMIで充分なのなら、IITでϕを計算する必要はない。現実の脳での測定を元にして、MIではなく、ϕでないとわからないことがある、ということを示す必要がある。この問題についてはのちほど。

2-4. 統合情報量ϕと相互情報量MIの関係

これが相互情報量MIやtransfer entropyとどういう数学的関係にあるかということについては「情報幾何論文」で取り扱っている。

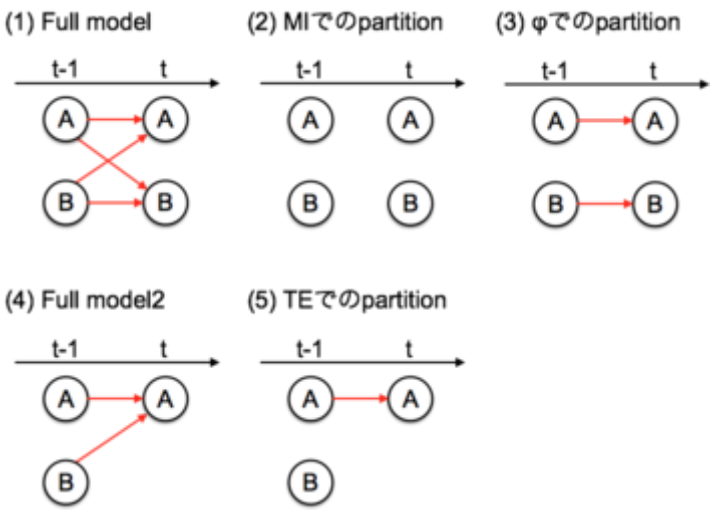

Figure 3

ふたたびABの2つだけのネットワークに戻って考えると、4つの要素{A(t-1), B(t-1), A(t), B(t)}が考えられる(図3-1)。ここでは4つのcausalなlinkを考えることができる(図3-1の赤線)。つまり{A(t-1)->A(t), A(t-1)->B(t), B(t-1)->A(t), B(t-1)->B(t)}

IITのφではこのうち{A(t-1)->B(t), B(t-1)->A(t)}を切った場合、つまりこのリンクが独立である場合(図3-3)とFull model(図3-1)との距離を計算している。いっぽうで相互情報量MIでは４つのcausal linkを全部切って(図3-2)からfull model(図3-1)との距離を計算している。。つまりIITのφに加えて、{A(t-1)->A(t), B(t-1)->B(t)}だけ余計に切っている。よって相互情報量MIは常にIITのφより大きい。

また、さいきん脳データのネットワーク解析でよく使われるようになったtransfer entropy TEはA(t), B(t)それぞれに定義することができる(注3)。たとえばA(t)に対しては3つの要素{A(t-1), B(t-1), A(t)}がある(図3-4; Full model2)。TE(A(t))ではB(t-1)->A(t)を切ったもの(図3-5)とfull model2(図3-4)の距離を計算したものと捉えることができる。よってTE(A)は常にIITのφより大きい。

両方を合わせると、こういう関係が成り立つ。

- TE(A), TE(B) < φ <MI

残念ながらTE(A)+TE(B) < φ <MI は常には成り立たないので、φをTEとMIから推定するというわけにはいかない。(TE(A)+TE(B)はSeth et al 2011で提唱されているcausal densityというやつ)

(吉田コメント) でもこれからわかるのは、現実的な脳のネットワークにおける問題のときにはまずMIとTEを計算しておくとよさそうだということ。また、φを計算するときには常にMIを計算した上で、MIではわからないことがφならわかる、という議論をすることが必要だということもわかる。じっさい、ヒトECoGデータの解析をしたHaun et al bioRxiv 2016ではそういう構造になっている。

---

3. IIT3.0での統合情報量φ

ここまで書いたのはIIT2.0での説明だった。一日目part3の説明はここからIIT3.0論文に準拠した説明になる。

3-1. Causeとeffectの考慮

IIT3.0がIIT2.0から進歩した点は、現在のニューロンネットワークが過去(入力)のネットワークとどのような因果ネットワークを形成しているかを表現したcause repertoireだけでなく、未来(出力)のネットワークとどのような因果ネットワークを形成しているかを表現したeffect repertoireも考慮するようになった点にある。

具体的には、cause repertoireを計算したのと同じようにeffect repertoireを計算して(ベイズの公式の使い方が変わるのでまったく同じ挙動にはならない)、cause, effecctそれぞれでのMIPを独立に決めてやって計算したφのうち小さい方をφとする、という操作を行っている。

(吉田コメント) これが必要な理屈についてはIIT3.0論文のFig.7で記述されている。要は過去から今、今から未来、両方共でintegrateしていないとintegrateしていると言えないでしょ、ということで、これはまったくごもっともで良い方向だと思う。しかし、端的にこのふたつのφ(の候補)を

べつべつに計算して小さい方を選ぶという操作はずいぶんと後付けなやり方だと思う。言い方をパクらしてもらえば、causeとeffectがぜんぜんintegrateしてない。

(吉田コメント) Causeとeffectの両方がnon-zeroのときのみφもnon-zeroになるということを満たしたいだけだったら、φ(cause) * φ(effect) だって構わないはずだ。そうしてない理由はφの次元がbitだからで、bit同士の掛け算するのはおかしいだろうと考えたことは推測できる。しかし後述するearth mover distanceはじつのところbitの次元ではなくて確率p=0-1の次元のものなので、その場合、掛け算にするのはそんなに悪くないはずだ。つまり、このあたりはどうにも取って付けたようなかんじで、いくらでも他の方法はありえて、理論として未完成なように思う。

3-2. 片方向のpartition

これまでのpartitionというのは、たとえば2要素ABからなるネットワークでAB間の因果ネットワークを切る際には、A<-BおよびA->Bの両方が同時に切られていた。これはsplit brainの症例を想定して脳のfiberを物理的に切るような可能性だけを考慮していたのだろう。しかしこの場合、単純なfeed-forward networkでもpartitionによって情報がロスするため、φがnon-zeroになるという事態になっていた。

IIT3.0においてはそこで片方向のpartitionということを考えるようになった。このことは考慮すべきpartitionの数をさらに増やすというデメリットはあるものの、単純なfeed-forward networkのφをゼロにできるという大きなメリットがある。それはどういうことかというと、たとえば2要素ABからなるネットワークでAB間の因果ネットワークがA->Bの片方向のみだった場合、つまり{A(t-1)->B(t), B(t-1)->A(t)}のうちでB(t-1)->A(t)は元々切れてる(無相関)のときに、MIPとしてB->Aを切ったときを考えることができて、このとき情報ロスはないのでφ=0となる。だからどんなに要素数の多いネットワークでも、ある段階にfeed-forwardのみの部分があればそこはφ=0になるのだ、ということが言えるようになる。

(吉田コメント) これもIIT2.0のときの批判を受けて、後付けで作ったルールなのだろうとは思ったけれども、理屈として片方向のpartitionを考えるというのは正しい方向だとは思った。しかしこうなると、partitionの定義自体が何に基づいているのかということが曖昧になってくるかもしれない。つまり、ABC(t-1)->ABC(t)をpartitionするのにこれまではAB(t-1)->C(t), C(t-1)->AB(t)のように排他的にpartitionしていたのだけれども、それに限る必要はあるのか？ってことにならないだろうか？

3-3. Earth movers distance (EMD)

IIT3.0ではKLDの代わりにEMDというのを使っている。これは、KLDでは確率密度分布の横軸の構造を考慮していないという問題への対策。つまり、cause repertoireというのはネットワークの状態ごとの確率密度分布なので、横軸は(たとえば2要素ABならば)AB = {00, 01, 10, 11}の4水準ある。しかしこの軸には近接度の違いがある。IIT3.0で採用しているのはハミング距離だが、00と01, 00と10は距離1だが、00と11は距離2となる。(00-01-11-10-00というループを考えればよい。) EMDではこの近接性を考慮して距離を計算している。

(吉田コメント) しかし思うに、この距離というのも一意に決まるわけではない。別案として思いつくのは、TPMを元に、t-1からtへ行くときにどう遷移するかから近接度を考える策もある。そ

のような統計的性質を持ち込むことの是非は考慮すべきだが、一意に決まるわけではないということは示せたのではないかと思う。

3-4. Concept

IIT3ではさらにconceptという概念を導入する(IIT2.0のときよりもintegrateした形で理論に組み込まれた、という言い方が正確か)。ABCという3要素のネットワークがあるときに、そのサブセットである{A,B,C,AB,AC,BC,ABC}を考える。たとえば現在のBCに対しての過去の{A,B,C,AB,AC,BC,ABC}でそれぞれのMIPを決めてやる(purview)と、それぞれについてφが計算できて、この7個のφのうちの最大のものをcore causeと呼ぶ。これを過去についても同様にやってやるとcore effectができる。

こうして現在のA(t)に対するcore cause (たとえばBC(t-1))およびcore effect(たとえばB(t+1))ができると、このcause-effectの対をconceptと呼ぶ。ConceptはABCに対しては複数ありうるけど、core causeが0になってしまうようなものは消えるので、conceptの数はネットワークの状態(IITではstate + mechanismという言い方をする)しだいで変わる。

3-5. Constellation, Qualia space

このようにしてできた複数のconceptをネットワークの状態を軸にした空間の上に配置したものをconstellationと呼び、この配置パターンが現在の状態での意識状態のクオリアに対応しているのだ、という議論をIITではしている。そして、これらのconceptがシステム(今の例だとABC)のMIPによってどれだけ情報をロスするかをEMDで評価して、それをconceptの個数分でcause, effect両方共で足し合わせてやったものとしてΦを定義している。

IIT2.0でのΦはどちらかというとIIT3.0でのφに近い。IIT3.0ではConstellationの考えまでを統合したものとしてシステムのΦを定義付けてやろうという意図から、このような複雑な定式化をしている。

(吉田コメント) 率直に言ってこの部分にはまったく承服できない。まず言うべきこととしてはここでのconceptというのは日常言語で言うconceptとはまったく関わりがないものだということだ。また、ここでのconceptの配置がqualiaであるというのも説得的でない。そうなる理由が足りない。

(吉田コメント) 意識経験の違いとは多次元空間の中での脳状態の違いである、というような考え方はチャーチランドの神経哲学でもコネクショニズムでの多次元の表象という点から似たような図が出ていた記憶があるが、それと比べると本質的な違いがあるようには思えない。けっきょくのところ脳状態(個々のニューロンのinstateneousな発火状態)を用いるのか、それとも直前、直後の脳状態まで含めて考えるのか、という違いでしかないと思う。

(吉田コメント) 好意的に捉えるのであれば、このようなconstellationを用いて意識経験のinvalianceを説明することができる(しかし神経発火の状態空間では説明できない)というようなことを示すことができるのならば、このような議論にも意味があるとは言えるだろう。つまり、赤の赤らしさは変わらないままに、赤の経験の強度を変化させることができたとして、このときにconstellationの構造は変わらないままにΦだけが変わる、ということが示せるなら良いのではな

いかということ。Haun et al bioRxiv 2016でのFig.3はそれを目指しているのだろうと推測するけれど、少なくとも、神経発火の状態空間では説明できないものがここにあるといいうことを示す必要はあるはず。

(吉田コメント) そしてこのconstellationの議論が納得いかない最大の理由は、意識のcontentの議論を回避した上でqualiaだけ議論しようという点に無理があるということ。IITでは内的な情報量を考慮することを徹底しているため、外界の刺激が何で、なにが表象されて、ということは理論の外にある。その事自体はIITを他の理論と峻別する非常に重要な点なのだと思うのだけど、それゆえにIITでは表象を扱うことができず、意識のcontentの議論を行うことができない。そのような状態で「クオリア」だけを取り出して扱おうというのは無理だろうと思う。同様にして、IITのaxiomであるcompositionalityについてもIITでは明示的な方法では定式化することができない。(IIT3.0論文のFig.22が関連している。)

(吉田コメント) IIT3.0論文を詳しく読んでみると、表象の問題に関してはDiscussionで"matching"という概念について言及している。つまり、外界の因果的構造を脳内の因果的ネットワークがmatchするように学習の結果作り上げるという話で、ここは大変重要な問題であり、IITでも課題であることは認識されているように思う。

---

4. IIT3.0から示唆されること

ここからはIIT3.0論文の理論構築がいったん済んだ後で、この理論から示唆されることについて、より複雑な(でも脳よりはずっと単純な)モデルを元にして議論している。

4-1. どの脳部位が意識に関与するか？

有名な、「なぜ小脳は意識には関わらないか」という部分。重要な点としては、片方向のpartitionを導入したことによって、片道の結合が入っているネットワークはΦ=0になった。よって、網膜からLGN、そして大脳へ行く経路において、LGNは大脳から投射が来ていて双方向性だけど、網膜からLGNは片方向性。よって網膜からLGNの経路を除いたときにΦは大きくなり、exclusion postulateにより、網膜は意識の外にあるという結論になる。

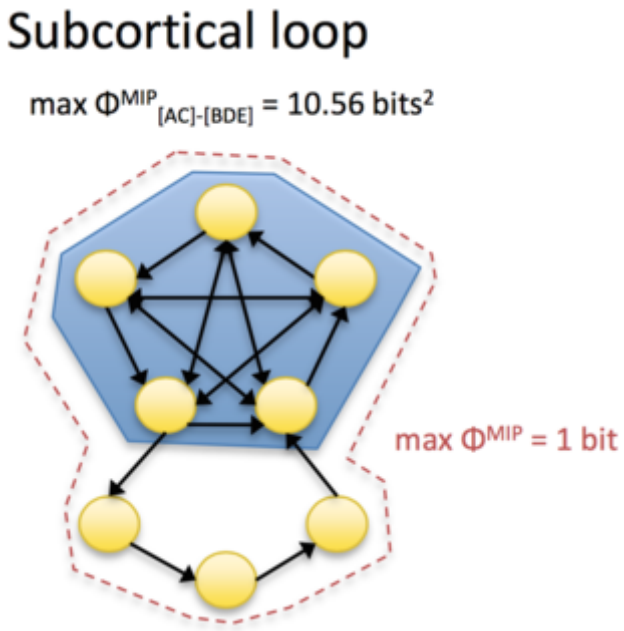

Figure 4 (大泉さん提供)

図4にIIT3.0でのsubcortical loop (basal gangliaを想定)を模したネットワークを示す。この図はBMC Neuroscience 2004のFig.4eと同じネットワークだが、IIT3.0の計算法を用いてΦを計算している(大泉さん提供)。Subcortical loopの部分を含めてΦを計算した場合(1 bit)と比べると、subcortical loopの部分を除くとより大きいΦになる(10.56 bit)ので、意識の境界としてはsubcortical loopは含まれない。

(吉田コメント) 私がここで注目するのはsubcortical loopの部分を含めてΦを計算すると値は低いがnon-zeroとなることだ。つまり、片道でもまたシステムに戻っていくループはIIT3.0でもΦがゼロにならない。このことはsensorimotor contingencyを考えるにあたって重要な点であると考えられる。因果ネットワークはニューロンでなくてもよいのだから、眼を動かすことによって視界が動いて視覚入力が変化するというのも因果ネットワークの一部として捉えることができるからだ。この件については次回の6-2で書く。いまの説明的に言及しておきたい点として、このループの何ステップなのかが因果の強さに関わってくること、そして統合の時間幅をどのくらいに取るかという問題に関わってくるということ。

4-2. “minimally conscious” photodiode

IIT3.0論文のFig.19の話。サーモスタットも、白色センサーも、青色センサーも、ネットワーク構造は同じなので、同じクオリアを持つと言える、という話。

(吉田コメント) ここは私としては全く同意で、これがまさに盲視の話で繰り返し言及してきた「なにかあるかんじ」なのだと提唱したい。ただし、この“minimally conscious” photodiodeは空間は持っていないので、意識は構造化されていないわけで、「なにかあるかんじ」よりももっと原始的なものと言うのが正確だが。

(吉田コメント) あともうひとつ、ここでthe no-strong-loops hypothesisとの関連をコメントしておきたい。サーモスタットはrecurrentな結合を持っていて、それがのnon-zeroのφを作っている。それはいいのだけれども、実際にこのような1対1対応の強いrecurrentの結合が脳にあるかというとそれは疑わしい。実際問題としてこういうstrong loopはfeedbackによって強い持続

的な発火を起こしてしまい、安定したネットワークとして活動できないだろう。このような「強いループ」が皮質-視床ネットワークには存在しないだろうと予言したのがCrick and Kochの"the no-strong-loops hypothesis"論文。

(吉田コメント) これはあくまでproposalでしかないが、cortico-corticalでの結合でこのようなstrong loopが無いということはJohnson and Burkhalter 2004 JNSによる解剖学的研究からすでに示されている。このようにして、strong-loopではφが大きくなるが、しかしそのようなネットワークでの意識レベルが高いとはいえそうにないので実際の意識経験と整合的でない。これはIITがうまい説明を考えないといけない課題のひとつだといえるだろう。

4-3. 哲学的ゾンビの可能性

IIT3.0論文のFig.20,21の話。同じ機能を果たすネットワークでも、完全にフィードフォワードなネットワーク(Φ=0)とフィードバックループのあるネットワーク(Φ>0)を作ることが可能である、という話。

(吉田コメント) これも盲視と繋げられると思っていて、上丘も大脳皮質もサリエンシー検出という同じ機能を果たすことができるのだけれども、上丘は主にフィードフォワードなネットワークでできていて、大脳皮質はフィードバックを使っていて、これが意識経験の有る無しに関わっていると議論できる。

4-4. 機能と現象の関連

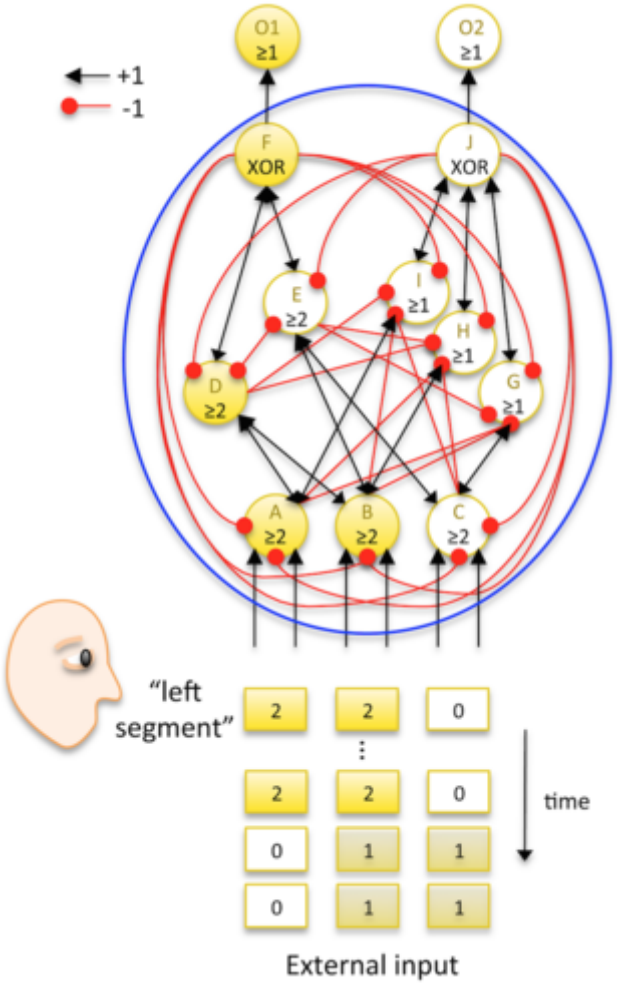

Figure 5 by Oizumi et al 2014 / CC BY 4.0

IIT3.0論文のFig.22の話。図5に示したネットワークは機能としては、ABが応答する外的入力(left segment)のときにはO_1という出力を出し、BCが応答する外的入力(right segment)のときにはO_2という出力を出すという左右の弁別をするsensorimotor processingをするネットワークと言える。神経科学者の眼からはDは左segmentのfeature detectorに見えるだろうし、FはO_1という行動のコマンドニューロンに見えるだろう。

しかし、IITの立場からは、そのφやconstellationはそれらの入出力、機能とは直接的には関係なく定まる。もちろん、結果としてφやconstellationが表していると考えられる現象は、入出力から定まる機能とは相関はあるのだけれども、それはあくまで間接的なものでありつづける。

これがIITにおける機能と現象の関係についての態度であり、IIT3.0論文のDiscussionにおいてもIITがまだ不完全であることのひとつとして、この関係について以下のように議論している。

> IITにおいては、脳のようなMICS (Φが極大を取るような複合体)と環境の関係はいわゆる「情報処理」の関係ではなくて、内的な因果構造と外的な因果構造の「マッチング」の関係にある。マッチングの定量化の方法としては、「通常の環境と相互作用するMICS」と「(構造を失った仮想的な)環境と相互するMICS」との距離を用いることができるかもしれない。「マッチング」の概念、そして「環境への適応がmatchingを上昇させ、その結果意識(レベル)の上昇を起こす」という予測については今後研究を進めてゆく予定である。このためには(かつてのanimat論文のような)バーチャル環境での進化的エージェントの実験や神経生理学的実験を用いる。(吉田による端折った訳)

そういうわけで、3-5で指摘した意識のcontentと表象の問題は、IITではmatchingというアイデアで解決しようとしており、そしてそれはまだ途上であるというのがIITの現状のようだ。私はこの部分が現象的意識の理論にとって無くてはならない部分であり、この問題を解決しないとIITが現象的意識の理論として成り立たないと思う。

---

5. IITの神経生理学的実験への応用

二日目のセミナーはこの部分に関して詳しく説明があった。ちょっと長くなりすぎたので、ここはズバッと省略させてもらいます。未発表のデータもあったことだし。

要は、partitionという作業が入っているために実際の神経生理学データにそのまま応用することが不可能なIITのΦやconstellationといったものを、いかにして各種の近似を入れて応用可能なものにするか、そしてそれによって神経生理学データをどのように解析することが出来て、それは相互情報量MIではわからないものがどのくらいあるかを示すことによって、IITが机上の空論でなく実際に意味のあるものであるという証拠を積み上げてゆく、という大事なステージ。

神経生理学者としては、自分の実験データを解析する際にどのようにIITを活用できるだろうか、という視点から話を聞かせてもらった。

---

6. 概括的なコメント

ここまでセミナーの内容に沿っていくつかコメントをしたけれども、もうすこし概括的にいくつか書いておきたい。まずいちばん言いたいことをまとめるとこうなる：

今回IIT3.0論文を精読してセミナーに参加したことで、個々のニューロンにとっての内的な情報としたらこういう形になるであろうこと、機能及び表象と完全に無縁な形で理論構築がなされていることを理解した。よって、IITには拘束条件が足りてないところ、必然性に欠ける部分があるものの、機能と表象から切り離した上で理論構築するとこのような形になるであろうということはわかった。

以前私はIITがある種の表象主義であり、NCCの後継であり、だからこそCristof KochがIITの擁護者になったのだということを書いたことがあるが、この考えは撤回しないといけない。IITはNCCでやっていることとは随分違うことをやっている。

では以前指摘したNCCとIITの連続性とは何かというと、脳活動=意識状態とする「同一説」的な考え(クリックのastounishing hypothesisというのも同じ)のことだとわかった。しかし同一説、つまり無媒介的に脳の状態と意識の状態とを同じものとする考えは(おそらく)すべての神経科学的アプローチで前提していることだ。そうなるとなぜ心の哲学で機能主義や表象主義が必要とされたのかって話に戻る必要がある。

さらにいくつか論点をまとめる。

6-1. 内的な情報量

IITについて知ってからずっと気になっていたことは、IITで言う「内的な情報量」というのがほんとうに内的なのか、環境の入出力をどこかで導入してないのか、ということだった。でも今回精読してみて納得いったけど、IITではたしかにあるニューロンの持つ情報をそのニューロンの現時点での活動状態およびそれにシナプス入力するニューロンの活動とシナプス出力するニューロンの活動から決めていた。わたしはこれは正しいアプローチだと思う。(注4)

ただし、cause repertoireは現時点の活動だけで全てが決まるのではなく、TPM(ネットワークの遷移確率)という統計的モデルを必要とするし、ベイズの定理を動かす段階でprior (P(A(t))も必要になる。(Priorにunconstrainedなnull distributionを使うべきか、それとも実際の発火履歴を使うべきか、というのも確定しているわけではないようだった。) そういうわけで、ここで計算される情報量自体はあるニューロンがアクセスできるものではなくて、ローカルではあるが外部から計算されるものであった。また、MIPやunconstrainedなnull distributionとの比較という形でIITには「そうであったかもしれない」というアンサンブル分布を持つことを必要としている。この点で充分「内的」でないのではないか、と考える。

あるニューロンAにとって入力ニューロンBの活動のon-offが本当に違いを生み出すものであるか、ということが「内的」な情報であるためには必要だ。もしある入力ニューロンBの活動が後続のニューロンAに全く影響を与えないのならば、たとえ解剖学的結合があってもニューロンAにとってニューロンBの活動は「違いを生む違い」になっていないのだから、それはニューロンAにとって「内的な」情報とはならない。

「内的な」情報を考慮する際にはそのような「違いを生む違い」をするものだけが残り、他のstateはそのニューロンにとって意味がない、という形でcause repertoireが縮退する必要がある。"concept"がやろうとしていることはどうやらそれのようで、cause repertoireの横軸が減るのではなくて、$\phi>0$なconceptが残るということを用いて、「違いを生む違い」の場合の数を決めるということをしているのだろう。Conceptという言葉に惑わされず、そのような「違いを生む違い」として可能なものの数、という理解をすればいいのかな、と思った。

もちろん、じっさいにニューロンが情報量を計算する必要があるわけではなくて、ニューロンはシナプス入力によってチャネルを開いて発火して、とただ物理法則に従っているだけにすぎない。だからここでいう「情報」というものは物理的世界に偏在してそれが法則的に意識を生み出す、というようなことを想定していることになる。

これは「情報が世界自体が持っている」とするフレッド・ドレツキの考えが正しいのかということに関わる問題かもしれない。また、光は最短距離を行こうという意志があるわけでもないのに結果として水の中を屈折して進むという変分原理のような考えで、この情報を持つ複合体が自己の境界を決めて極大を持つような相転移を起こす、ということを想定しているのだろう。また、近年の「マクスウェルの悪魔」実験で議論されるような「物理世界における情報」も射程に入るだろう。

残念ながら私にはこのあたりについてもっとたくさんの勉強が必要だ。これは私自身の課題なのだが、「情報とは何なのか」ということを「そうであったかもしれない」というアンサンブル分布を持つことという「反実仮想」を含んだものとして捉え直すことによって、IITの根幹である「情報」の概念を見直した理論が作れないだろうか、とか考えてる。

6-2. Sensorimotor contingencyとの関係

もともとIITについては「オートポイエーシス的でない(ゆえに意識の理論として何かが欠けているのではないか)」という印象(というかheuristics)があった。それは環境との相互作用を明示的に取り込んでいないからだった。しかし今回詳しく読んでみて、環境との相互作用を取り込んでIITを拡張することが可能なのではないかと考えた。この考えについて以下に提案してみたい。

上記の図5をもう一回見直して見てほしいのだけど、この図のネットワークの出力O_1, O_2と視覚入力を因果的につなげてやれば、sensorimotor contingencyのモデルになる。つまり、左segmentが提示されたときはO_1の活動によって左にサッカードしてsegmentを視野の中心に持ってくるし、右segmentが提示されたときはO_2の活動によって右にサッカードしてsegmentを視野の中心に持ってくる、というモデルに改変することができる。

IITでの因果ネットワークはニューロンの結合であることを必要とはしていないから、このようなsensorimotor contingencyも立派な因果ネットワークの一種だ。

さらに前回(4/5)の図4でも言及したように、外部のループを含んだシステムはΦは小さいがnon-zeroのネットワークを作ることが可能になる。(Exclusionによって、ループの部分は排除されるが。)

そうしてみると上記の図5で出力と入力をつなげたネットワークでは外部のループの部分が弱いのでそこでMIPができて、より小さいネットワークのほうが意識の単位として残るだろう。それでよいのであって、我々の意識経験としてもこのsensorimotor contingencyのループは意識の外にあるものとして経験される。

しかし、発達期はどうだろう？ まだ大脳皮質が充分に発達していない(充分にintegrateしていない)時期に、赤ちゃんが手を動かしてそれを自分で見るとか、解像度の低いお母さんの顔に向けて目と頭を向けるといったsensorimotor contingencyが成り立つときに、そのループは大脳皮質だけの部分では極大とならないかもしれない。そしてそれこそが幼児期に自他の区別が未分化な状態での意識経験を説明すると言えないだろうか？ (同様な考えをセミナーの参加者の方が進化的側面から提案していたのでここにクレジットしておく。)

このような自他未分化な状態では環境を共有することになる。それはexclusion postulateとどう整合的に説明できるだろう？ これについても考えた。Aさんにとっての環境とBさんにとっての環境は視点の違い、sensorimotor conringencyの違いという意味において異なっており、まったく同じ環境を共有しているわけではない。だからexclusionはここでは問題にならない。(もしくはexclusionの再定義が必要になる。)

こうしてみると自分にとっての環境と他者との環境との関係はオートポイエーシスで言うところの「カップリング」の関係になっているということが分かる。また、IITの理論構成では機能と現象とは相関はあるけれども、明示的にはお互いが影響を与えないような形になっている。これも意識を持つ有機体と環境とが「カップリング」の関係にあると言っていることと同じかもしくととても近い。

そしてこのカップリングの概念はIIT3.0論文で今後の課題として議論されている「(環境の因果ネットワークと内的な因果ネットワークの)マッチング」とも大いに関連しているだろう。だから、ここで提案している考えはそんなに的外れでもないはずだ。

また、IITはあるinstantaneousなstateごとにΦを定義して、その境界を決めるという点では「作動しているときにのみその境界を決める」オートポイエーシスとよく似ている側面はある。あとはこれが「作動のネットワーク」になっているかだけど、情報の流れはエネルギーそのものではないという意味でもオートポイエーシス的であるといえる。

ちょっとこじつけているところはあるかなあと我ながら思うが、私からのプロポーザルとしては、IITにsensorimotor conringencyを明示的に加えることによって、IITがよりオートポイエーシス的になり、life-mind continuityを実現するようなagentのモデルとして発展させるとで現象的意識の理論により近づくのではないか、というものだ。

---

7. さいごに

そういうわけで、IITについて批判的に紹介とコメントをしながら、自分だったらどのように拡張するかということを提案してみた。いろいろ勘違い、不正確な点が含まれているだろうと思うので、そのあたりはご指摘いただけたらありがたい。

ここに書いてあることは私なりの理解であり、しかもIITとはかなりかけ離れた立場からのコメントだった。ホンモノのIITを理解したいという人は原著に当たるのをお勧めしたい。今回のセミナーを聞くかぎり、ジュリオ・トノーニと大泉さんの間でも立場が若干違っているように思う。IITはまだ発展中の理論なので、どんどん改善して自分の理論として使ってしまえばよいと思う。私がここでやったのもまさにそういう取り込みの過程だった。

(注1) 「Informationとは"differences that make a difference"である」という考え方は意識は外側から見た情報ではなくて「内側から見た情報」intrinsic informationによって規定されるという議論のためにここでは引っ張られている。IIT3論文ではreferされていないが、この言明はベイトソンのSteps to an Ecology of Mindからとられたもの("In fact, what we mean by information—the elementary unit of information—is *a difference which makes a difference*, and it is able to make a difference because the neural pathways along which it travels and is continually transformed are themselves provided with energy.")。ただし今回見つけた記事ではこの表現は誤用されていると議論している。ついでに昔ブログに書いた「グレゴリー・ベイトソン(Gregory Bateson)の「精神と自然」まとめ」も参照。

(注2)相互情報量MIも各刺激S={s1, s2, ...}ごとに計算した情報量iの重み付け平均としても表現することもできるので、この点で統合情報量が相互情報量MIと本質的に変わるわけではない。

- i(X; s) = D(P(X | S=s) || P(X)))
- MI(X; s) = Sum(p(S=s) * i(X; s))

(注3) Gaussianで近似できる系ではTEはgranger causalityと同じになることが知られている。

(注4) その昔(いま調べてみたら2008年1月24日だった)、土谷さんに生理研にセミナーをしに来てもらったことがあって、そのあとで武井くんのうちで飲みながら「夢の実験」(技術的障害を考慮せず、こういう実験ができたらいいと思うものは何か?)というのを議論したことがある。そのときわたしは「一個のニューロンがどう働いているのかを特徴づけるためにそのpreのニューロンとpostのニューロンのすべての発火がわかるような記録ができたらいいと思う」というようなことを言った記憶がある。そのときはいまいちウケはよくなかったのだけど、それこそがまさにIITで一個のニューロンの挙動の完全な描写としてcause repertoireとresult repertoireを作成してそれをもとにそのニューロンにとっての情報とは何か、を規定するために必要なことだった。