MIRAI Archive
京都大学未来創成学国際研究ユニット
International Research Unit of Advanced Future Studies(IRU-AFS) No.2015-010

＜講演録＞

京都大学基礎物理学研究所　研究会
複雑システムにおける創造的破壊現象の原理に迫る

日時　2015 年 8 月 7 日（木）13：00−13：40

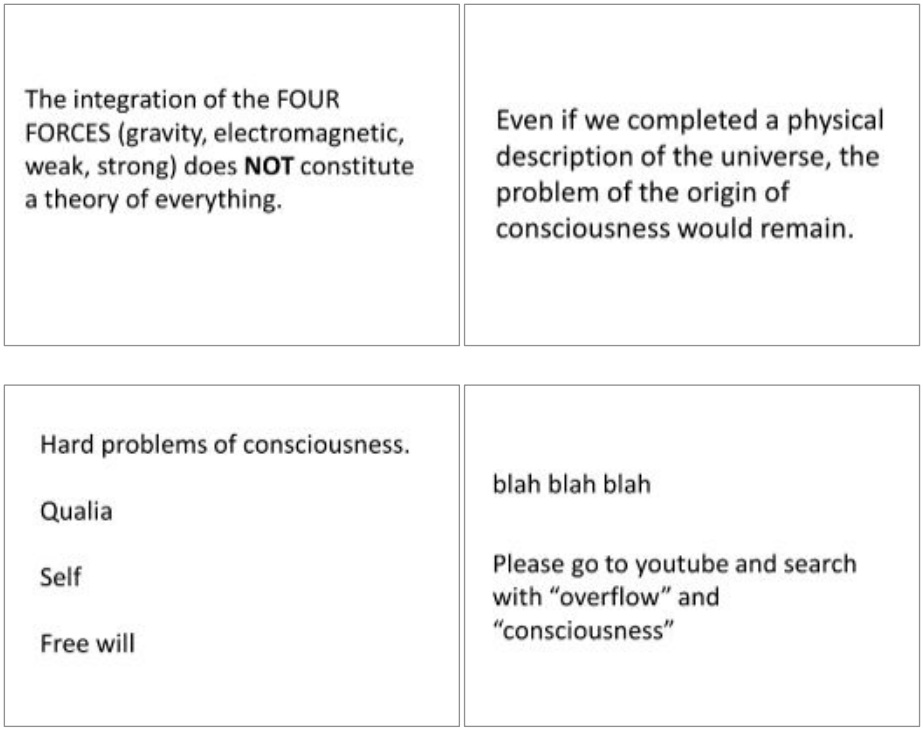
場所　コープイン京都　202

座長　三輪 敬之（早稲田大学 創造理工学機械工学科　教授）

講演　茂木 健一郎（ソニーコンピュータサイエンス研究所　上級研究員）

The narrowness of AI and human adaptation

◇スライド 4、5、6、7



茂木●「The overflow model of the evolution of consciousness」ということです。もともと万物の理論というものがあって、四つの力を統一すれば万物の理論が完成するというような考え方なのです。しかし、これは私の立場からいうと、じつは万物の理論ではない。四つの力がたとえ統一されたとしても、それは万物の理論を構成しないのです。なぜかというと、物理的な宇宙の記述を完成したとしても、意識の問題が残るからなのです。意識の問題というのは、クオリア（Qualia）とか自己とかフリー・ウィル（Free will）です。

これについて、いまオーバーフローという視点から理論を展開しているのですが、昨日から池上高志さんと喋っていて、この話は今回はやめようということになりました。YouTube で「overflow」、「consciousness」というキーワードで検索していただくと、今年ヘルシンキで開催された学会で私が発表したプレゼンテーションが出てきますので、興味がある方はみてください。オーバーフローというのは、われわれが受け止められる情報は意識のバンドワイズ（bandwidth）量をはるかに超えていて、それに対するアダプテーションが意識の起源だということなのですが、これについては論文を書いたり、本を書いたりしようとしています。

◇スライド 9

昨日から池上さんと喋っていて、じゃあ AI（人工知能）の話をしようということで、突然ですが、「人工知能の狭さと人間の適応（The narrowness of AI and human adaptation）」というタイトルで話させてください。村瀬さんが立ち上げたこの素晴らしい領域――未来創成ということで、このテーマがいちばんふさわしいのかなと思って、意識の問題という私の趣味の世界ではなく、いまたいへん興味をもたれている人工知能の話をしようと思いました。

◇スライド 10

人工知能では、ガルリ・カスパロフ（Garry Kimovich Kasparov）というチェスのチャンピオンがディープ・ブルー（チェス専用のスーパーコンピュータ）に負けてしまいました。

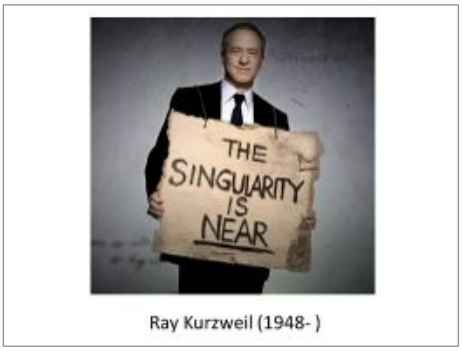
Deep Blue beats defeats Kasparov (1997)

◇スライド 11

IBM のワトソン（WATSON）が「ジェパディ！（Jeopardy!）」という雑学クイズに勝っていたりしています。いまワトソンは医療情報の分析などにも使われています。

IBM's WATSON wins in Jeopardy (2011)

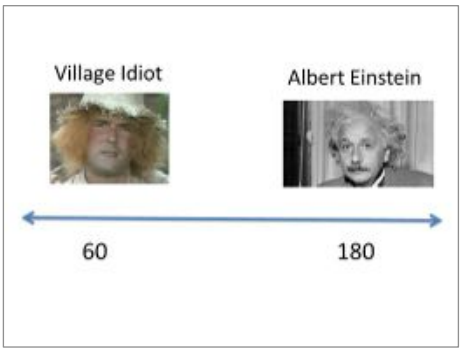
◇スライド 12



Ray Kurzweil (1948-)

レイ・カーツワイル（Ray Kurzweil）がシンギュラリティ（Singularity）ということをいっています。シンギュラリティとは、われわれの知性に人工知能が追いついて、超えてしまうということなのです。私は、いま人工知能に関する『神の名前』という小説を書いていて、特異点を超えたあとの世界がどうなるのかについて書いています。それはともかく、そのような予測がいま成り立っています。

◇スライド 13

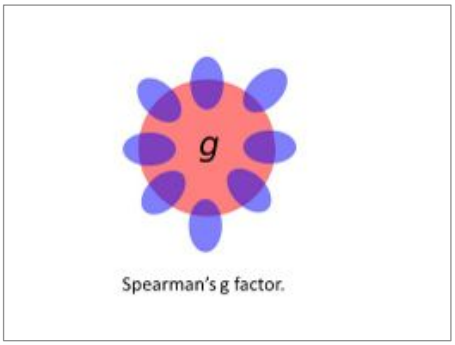


私の少年時代のヒーローはアルベルト・アインシュタイン（Albert Einstein）なのですが、ぼくはイギリスに 2 年間いたのですが、イギリスではビレッジ・イディオット（Village Idiot）という「村にいる馬鹿者」がよくコメディに出てくるのです。スライドの写真はモンティパイソンから取ったのですが。われわれは、「村にいる馬鹿者」とアインシュタインとのあいだの IQ の差は大きなものだと思っています。

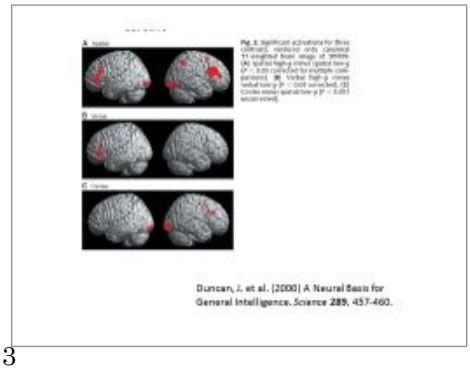
◇スライド 14、15、16


Charles Spearman (1863-1945)

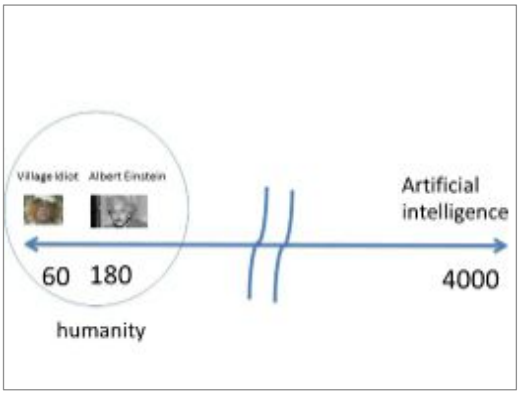
ちなみに IQ とは、チャールズ・エドワード・スピアマン（Charles Edward Spearman）という人が 1906 年の論文で g 因子（Spearman’s g factor）を提案しました。地頭のよさなどのいろいろなタスクがあるのですが、そのタスクに共通の g 因子というものがあって、これが頭のよさ、いわゆる IQ に相当している。それに相当する脳活動なども最近の研究でわかってきています。簡単にいうと、脳のリソース管理をする前頭葉の回路が g 因子との相関があるということがわかっております。



Spearman’s g factor.


Duncan, J. et al. (2000) A Neural Basis for General Intelligence. Science 289, 457-460.

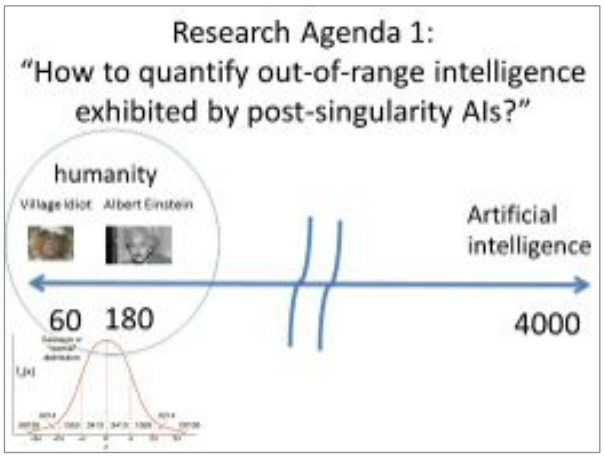
◇スライド 17



しかし、将来に人工知能がシンギュラリティを迎えて、たとえば知能指数が 4,000 というようなことになると、人間のあいだの知能の差なんて関係がなくなる。ぼくはこの視点が大事だと思っているのです。つまり、人間のあいだの頭のよさとか悪さというのは、おそらくシンギュラリティを迎えた人工知能からみるとまったく意味がない。

このトークのなかでは、いくつかこういう研究をしたらよいのではないかという提案をします。池上さんとも研究したいなと思っているし、もし京都大学の方でも…。これはオープンソースといいますか、パクりもぜんぜんオッケーです。

◇スライド 18

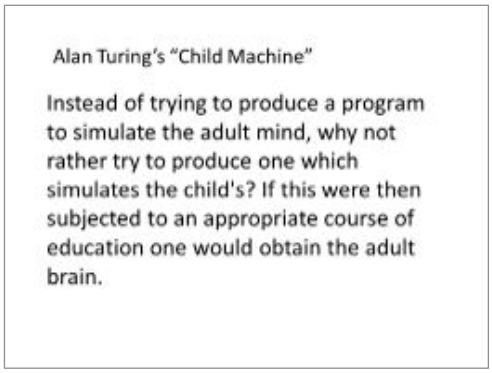


IQ は、普通ガウシアンの分布（Gaussian distribution）を前提に定義されているのですが、それとまったく違った次元にいったときのインテリジェンス（intelligence）をどう定量的に計測するのかという本質的な問題がある。これはまだ定義がよくされていない。つまり、たとえば同じ問題を 1,000 倍とか 10,000 倍速くできるようになった知能は、知性のメジャーとしてはどう評価すればよいのか。ようするに、人間のあいだの分布のなかの知性のメジャーなんかどうでもよいじゃないですか。IQ が 60 と 180 の差というのはたいしたことがない。

質的にジャンプしてしまった知性について、どうそれをエクストラポレーション（Extrapolation）できるのかということに対して、いまからメジャーを用意しておかないとシンギュラリティの評価ができないのです。なので、g 因子のような偏差値的な意味での IQ のメジャーではない知性のメジャーをいまから開発する必要がある。これがリサーチ・アジェンダ（Research Agenda）1 です。

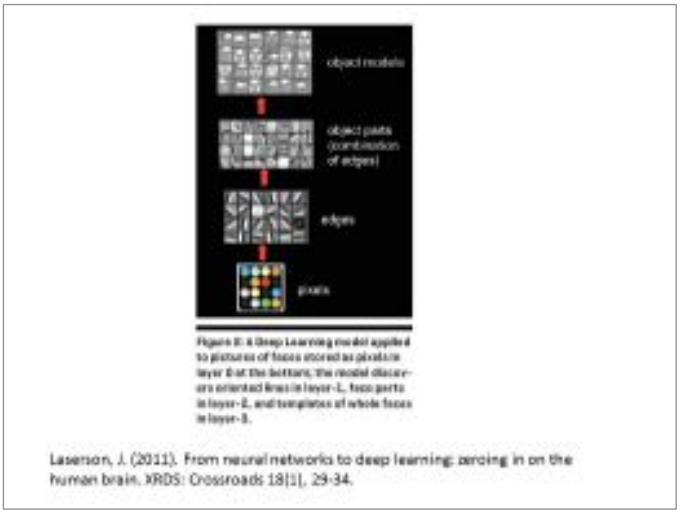
◇スライド 19

アラン・チューリング（Alan Turing）は「Child Machine」という概念を出しました。タブラ・ラサではないのですが、最初に子どもの状態を用意しておいて、そこから学習する。学習することでいろいろなことを学ぶのですが、こういうかたちの人工知能が出てくると、われわれは知性のコンスティチューション（constitution）をブラックボックスとして扱うしかないので、いろいろと困った問題が出てくるのです。



◇スライド 20

いま流行りのディープラーニング（Deep Learning）というのは、概念自体を人工知能システムが獲得するということで、基本的に人間はその過程をコントロールできないということになるのです。



◇スライド 21

こういうことを背景に、たとえばスペースエックス（SpaceX）、テスラモーターズ（Tesla Motors）、ペイパル（PayPal）を立ち上げたイーロン・マスクは、「人工知能は、ひょっとしたら核兵器よりも危険かもしれない」と言っていて、これが大きなトピックになっています。

◇ スライド 22、23

なぜ人工知能が核兵器よりも危険なのかもしれないかというと、たとえばファイナンシャル・マーケットでは 50%くらいがプログラム売買で、松尾豊さんという東京大学で人工知能の研究をしている人に聞いたら、いま 1ns（ナノ秒）で取引ができるらしいのです。

ということは人間のデイトレーダーがトレードしてもまったく太刀打ちできなくて、このことによってファイナンシャル・マーケットがプログラム売買で変動するようになってしまうのです。このような不安定性が、人工知能の応用によってもたらされるのではないかと懸念されている。

◇スライド 24、25

1900 年から 1991 年の湾岸戦争のときに、DART（Dynamic Analysis and Replanning Tool）というものがありました。じつはぼくは、いま英語で人工知能の本を書いていて、だからこんなにいろいろなことを調べているのです。いま 7 割くらい原稿が終わっている状態です。まだ出版社は決まっていないのですが…。この DART というのは、AI をもとにロジスティクス（Logistics）を最適化するツールなのです。ここがすごいところなのですが、湾岸戦争のときに DART は、人工知能の応用によって、アメリカの軍事技術にファンディングを出す組織の過去 30 年間の AI 研究の経費（budget）を超える節約効果が

あったというのです。これは1990年の話ですからね。

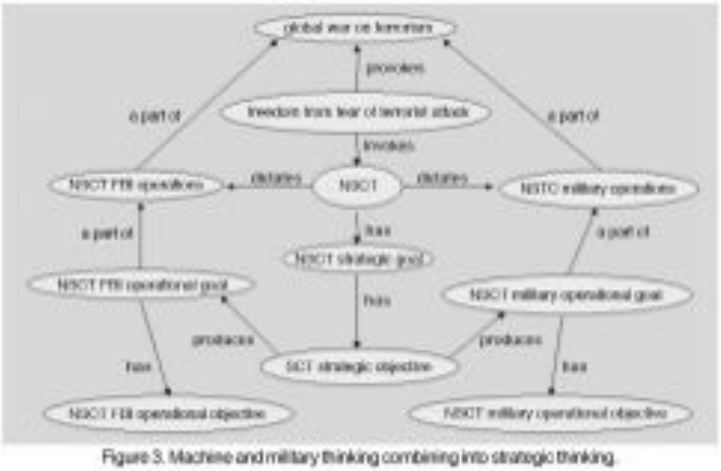
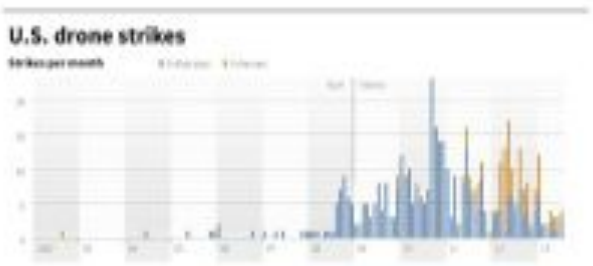


Figure 3. Machine and military thinking combining into strategic thinking

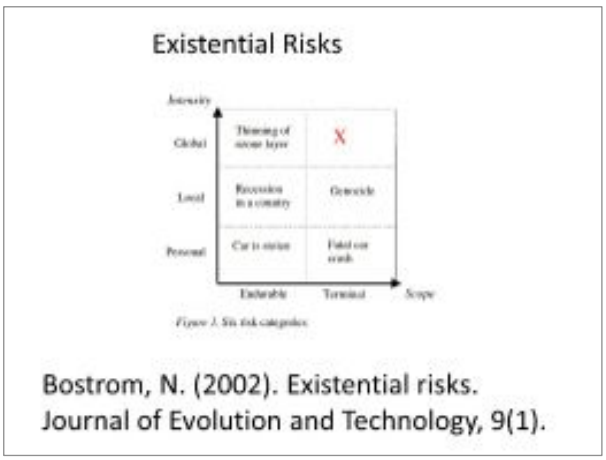
◇スライド 26、27

日本ではまだ牧歌的な状況にあるのですが、アメリカでは人工知能の軍事技術への応用が真剣に議論されています。スライドのもDARTです。そのときに懸念されるのが、すでにドローンがアフガニスタンやイラクなどで使われていて、これだけのストライク（Strike）があり、何人も殺されているのです。いまはまだアメリカの軍事基地で、ジョイスティックのようなものによる遠隔操作で最終的なストライクの判断をしているのですが、これは24時間オペレーションをしてターゲットを攻撃するとなったときに、遅かれ早かれ人工知能を応用して、人の認識および攻撃の決断をするようになるのは目にみえているのです。ぼくがアメリカ国防総省本庁舎（ペンタゴン）の人だったら絶対にそうしますよ。



◇スライド 28

そうなったときに、High Frequency Tradingの例で挙げたような不安定性が軍事技術にももちこまれるのではないかといま懸念されているのです。そのようなことを背景にニック・ボストロム（Nick Bostrom）は存在のリスク（Existential Risks）ということをいっている。



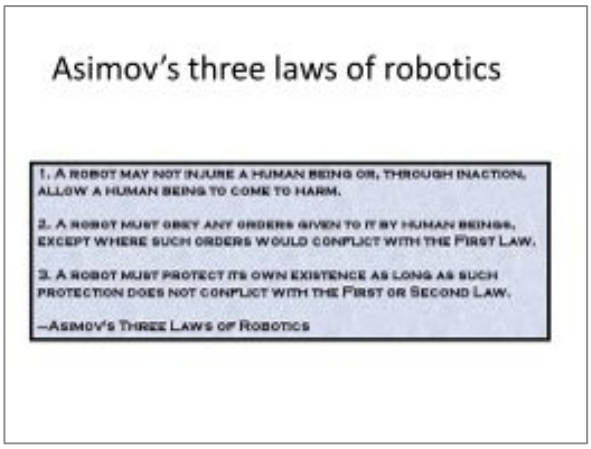
人工知能のもたらす危険について、イギリスのケンブリッジ大学とオクスフォード大学、それからアメリカの MIRI（Machine Intelligence Research Institute）と三つの研究組織が立ち上がっているのです。私は日本でも絶対にそういうものはつくるべきだと思っています。人工知能自体を研究するというよりも、人工知能の社会的インパクトやそれがもたらす安全保障上の懸念、いろいろなことを研究する組織が日本でも必要です。このボストロムという人は、オクスフォード大学でそれを研究している人です。

◇スライド 29

だからリサーチ・アジェンダ 2 としては、AI をいかにコントロールするのかという問題があります。スライドの画像は「エクス・マキナ（Ex Machina）」という今年イギリスで公開された映画です。私は飛行機のなかでみたのですが、二つの問題をこの映画は扱っています。一つがチューリング・テストです。この女性は AI なのですが、この人をインタビューして、この人が人間と区別できないかどうかを判断する青年が来るのです。それが一つの映画のテーマなのです。もう一つは、これはネタバレなのですが、AI の女性が青年に恋をさせて、その恋愛感情を利用して研究施設から抜け出るのです。最後は青年が閉じ込められて、「出してくれー」となってしまって、AI の女性が服を着て街を歩いているところで終わるのです。ラストシーンまで言ってしまったというひどい話ですね。

人工知能の封じ込め問題というのはボストロムらの研究コミュニティで議論されているのです。つまり、人工知能をいかに封じ込めることができるのかという問題です。これは大きな研究テーマの一つだと思っています。

◇スライド 30

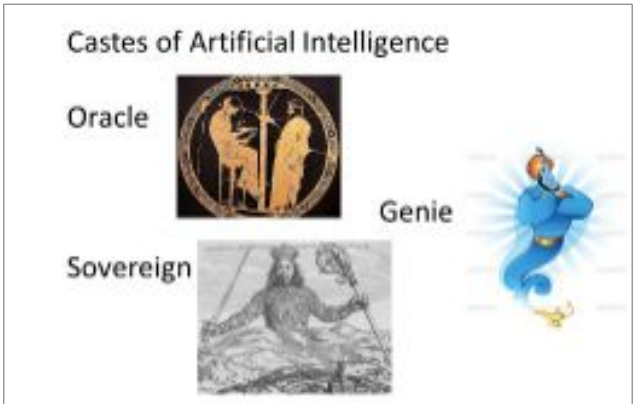


アイザック・アシモフ（Isaac Asimov）が「ロボット工学三原則」を出しました。これは比較的知られていない事実なのですが、この三原則は、「このようにルール化していても失敗してしまう」という例でアシモフは書いているのです。アシモフの短編のなかで、この三原則でロボットは動くのですが、現実に出会うと例外などが生じてこの

三原則ではうまくいかない事例が生じる。このことを示すためにアシモフの三原則は出ているのです。

◇スライド 31、32、33、34

AI のコントロールはたいへん難しくて、まだ日本語には訳されてはないですが、ボストロムの『Superintelligence』という本に出てくる分類としては、オラクル（Oracle）、ジーニー（Genie）、ソブリン（Sovereign）というものがある。オラクルというのは、人間がなにか問いを発するとそれに対する答えを出してくれる。ジーニーは、人間がなにかタスクを投げるとそれを実行してくれる。ソブリンというのは、完全に人工知能に全権を委任するということです。こういうかたちで、オラクルからジーニーそしてソブリンへと人工知能が進化したときに、それをどうコントロールするのかという問題があります。ボストロムは哲学者なのですが、彼の『Superintelligence』という本のなかで、そのコントロール問題について詳細にテクニカルな議論が行なわれています。もし興味がある方は、ご覧いただければと思います。



Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

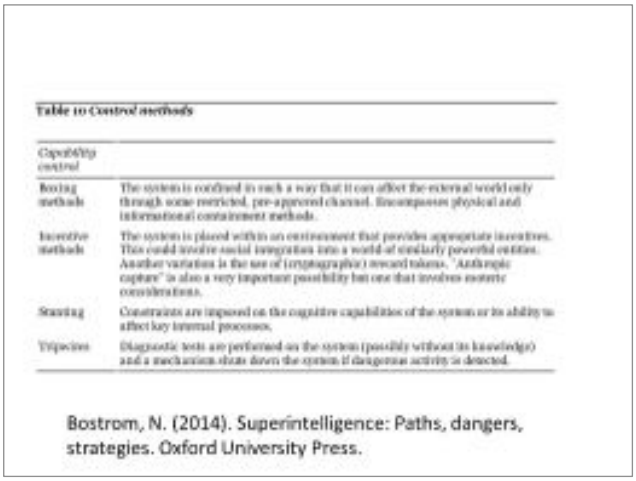


Table 10 Control methods

Capability control	
Boxing methods	The system is confined in such a way that it can affect the external world only through some restricted, pre-approved channel. Encompasses physical and informational containment methods.
Incentive methods	The system is placed within an environment that provides appropriate incentives. This could involve social integration into a world of similarly powerful entities. Another variation is the use of (cryptographic) reward tokens. "Anthropic capture" is also a very important possibility but one that involves esoteric considerations.
Stunting	Constraints are imposed on the cognitive capabilities of the system or its ability to affect key internal processes.
Tripwires	Diagnostic tests are performed on the system (possibly without its knowledge) and a mechanism shuts down the system if dangerous activity is detected.

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.


Motivation selection	
Direct specification	The system is endowed with some directly specified motivation system, which might be consequentialist or involve following a set of rules.
Domesticity	A motivation system is designed to severely limit the scope of the agent's ambitions and activities.
Indirect normativity	Indirect normativity could involve rule-based or consequentialist principles, but is distinguished by its reliance on an indirect approach to specifying the rules that are to be followed or the values that are to be pursued.
Augmentation	One starts with a system that already has substantially human or benevolent motivations, and enhances its cognitive capacities to make it superintelligent.

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

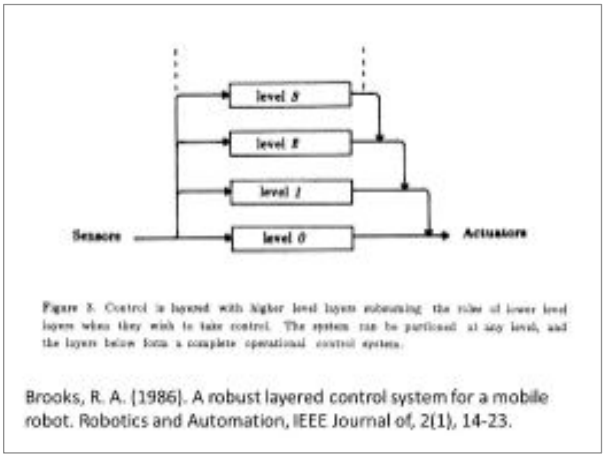


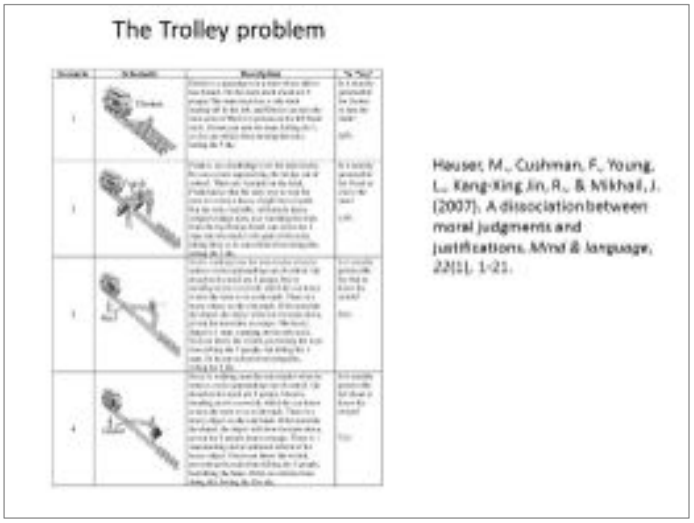
Figure 3. Control is layered with higher level layers subsuming the roles of lower level layers when they wish to take control. The system can be partitioned at any level, and the layers below form a complete operational control system.

Brooks, R. A. (1986). A robust layered control system for a mobile robot. Robotics and Automation, IEEE Journal of, 2(1), 14-23.

◇スライド 35

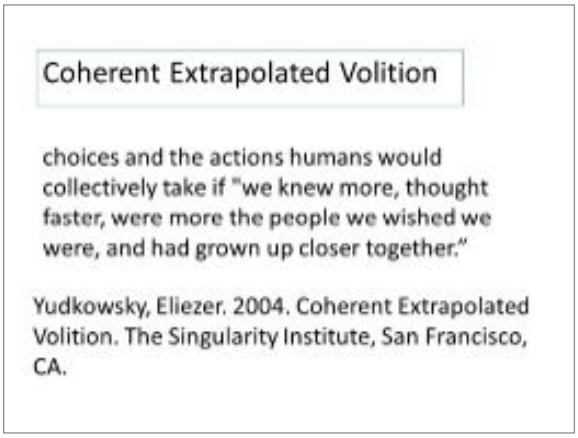
それ以外に AI の倫理学（ethics）という研究分野がある。これまでの話は、AI のコントロール問題です。倫理とはどういうことかというと、Google が自動運転車を研究していて、これは間違いなく日本にも入ってくる。そのときに倫理問題が影にあります。わかりやすい例としては、事故が起こったときに誰の責任なのかということがあります。自動車メーカーなのか、ソフトウェアの会社なのか、それともそこに乗っている人なのか。

◇スライド 36



それだけではなくて、じつはトロリー（Trolley）問題というものがあるのです。これはマイケル・サンデル（Michael J. Sandel）が「Justice」の講義の第 1 回めでも引用しているのですが、「ある事故を避けるために別の事故を起こすような状況が起こったときに、どういう基準でなにを優先すべきか」。たとえば、自動運転車が走っているときに、子どもが飛び出してきた。子どもを救うために急ブレーキをかけたら、後ろから来た二輪車のバイクが衝突して、二輪車に乗っていた人のほうが大けがをして死ぬことがある。二輪車が後ろから来ていることをわかっていながら、子どもの命を救うために急ブレーキをかけるべきかどうかという倫理問題が、自動運転車には潜在的にあります。これを研究することが、重要だと考えられます。

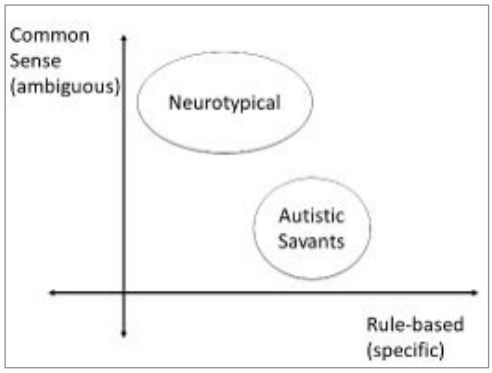
◇スライド 37



この分野の重要な概念は、「Coherent Extrapolated Volition」ということです。エリザー・ヨドコフスキー（Eliezer Yudkowsky）という人が提案しています。ヨドコフスキーは、MIRI にいるフリーランスの研究者なのですが、おもしろい人です。たとえば、池上さんが彼女にあげるためにダイヤモンドの指輪を欲しい。それで二つの箱がある。それで池上さんが「左を取れ」と AI に注文したとします。ところが、AI は右にダイヤモンドの指輪が入っていることを知っているとします。そうすると池上さんの指示（instruction）は、「左側を取れ」ということなのですが、ほんとうの池上さんの意図は、「ダイヤモンドの指輪が入っている箱を取れ」ということです。だから AI は、ほんとうは右を取るほうがよい。さらにいえば、池上さんがダイヤモンドの指輪を彼女にプレゼントすることで、家庭が崩壊するかもしれません。すると AI はそこまで考えて、「この人は、彼女にあげるためにダイヤモンドの指輪を取りたいと言っているけど、ほんとうは取らせないほうがよいのだ」というところまで判断するかもしれません。そのときに右のほうに指輪があっても、あえて左側を取るということもあるのです。

そういう人類のコレクティブなベスト・アンド・ブライテスト（the Best and the Brightest）の叡智を集めた意図（Volition）が Coherent Extrapolated Volition という概念です。たとえば安倍さんが集団的自衛権を実現させようとしているけど、それはある種のモデルのもとで日本の集団的安全保障のために必要だというロジックでしているじゃないですか。しかし、ほんとうにそれがベスト・アンド・ブライテストの――すべてのデータを集めたときのオプティマム（optimum）なチョイスなのかどうかということは別問題です。そのような概念を実装することでフレンドリーな AI を実現しようというものがこの概念です。

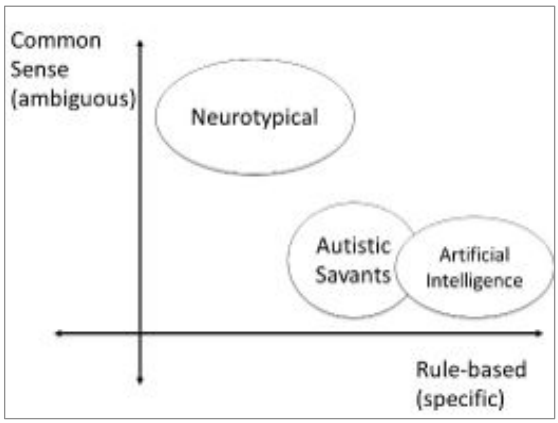
◇スライド 38

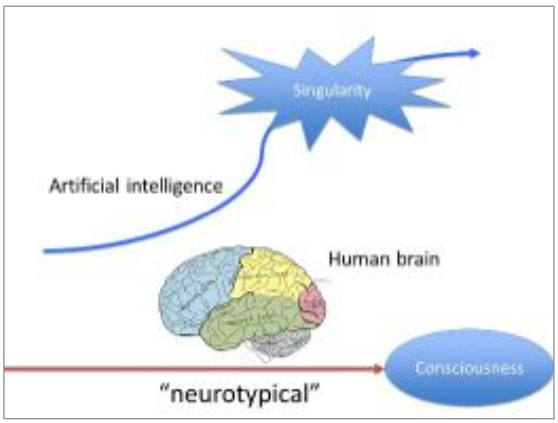


脳科学の視点からいうと、これは TEDxTokyo で先日話したのですが、内海さんの講演のときに、私は自閉症スペクトラムにいるのではないかっていう話をしました。袖川さんはじつは自閉症に近いという自己認識があるとお昼のときに聞いたのですが、ニューロティピカル（Neurotypical）な人は、ど

ちらかというと論理が苦手で、感情が得意なのです。アンビリアス（ambiguous）なことが得意なのですが、サヴァン症候群の人は論理的なことが好きです。

◇スライド 39、スライド 40

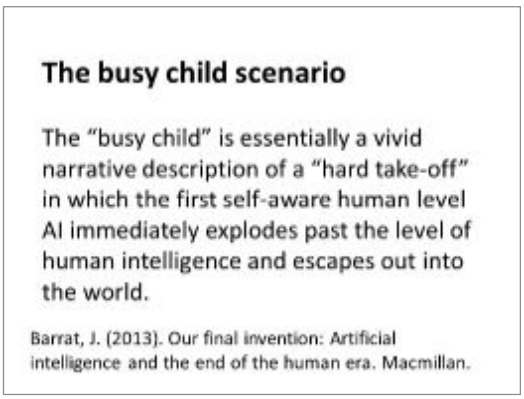




じつは人工知能は、脳の働きからいうとスーパー・サヴァンのようなところがあるのです。人工知能の狭さは、人間のニューロティピカルなエボリューションのパスとは違う極めての方向――スーパー・サヴァン的なエボリューションのなかでシンギュラリティを迎えようとしている。これが人工知能の決定的な狭さです。

MIT メディアラボの社長の伊藤穰一さんは、あるときにピザを 10 人分注文して、それを冷凍して 3 日間戦略ゲームをプレイし続けて、72 時間寝ずにプレイして勝ったらしいのです。そのあいだに 10 人分のピザをときどき加熱してたべていたらしいです。伊藤穰一さんは、一つのことに 72 時間集中するというサヴァン的な能力がある。しかし、72 時間というのは、コンピュータの時間でいうときわめて短い。コンピュータというのは、一つの問題に 1,000 時間とか 10,000 時間を連続して集中できるのです。それで飽きないし文句も言わないし、他のこともやらない。

◇スライド 41



ということは、AI というのは、人間的な見方をするとスーパー・サヴァンということです。そこに AI の狭さも卓越性もあります。これは翻訳されていますが、ジェイムズ・バラットという人の『人工知能 人類最悪にして最後の発明（Our Final Invention）』という本のなかで、AI のパーソナリティに注目して「ビジーチャイルド・シナリオ（The busy child scenario）」といいました。いつもちょこまかとなにかをしている小うるさい子どものようだということです。AI をパーソ

ナリティとしてみると、たいへん狭いパーソナリティしかもっていない。

◇スライド 42、43

Busy child personality

Personalities

たとえば、いまソフトバンクが売り出しているこのペッパー（Pepper）などもそうなのですが、一つのことに集中してサヴァン的な能力を出すのはよいのですが、パーソナリティという視点から問題がある。

◇スライド 44

Trait	Description
Openness	Being curious, original, intellectual, creative, and open to new ideas.
Conscientiousness	Being organized, systematic, punctual, achievement oriented, and dependable.
Extraversion	Being outgoing, talkative, sociable, and enjoying social situations.
Agreeableness	Being affable, tolerant, sensitive, trusting, kind, and warm.
Neuroticism	Being anxious, irritable, temperamental, and moody.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. Annual review of psychology, 41(1), 417-440.

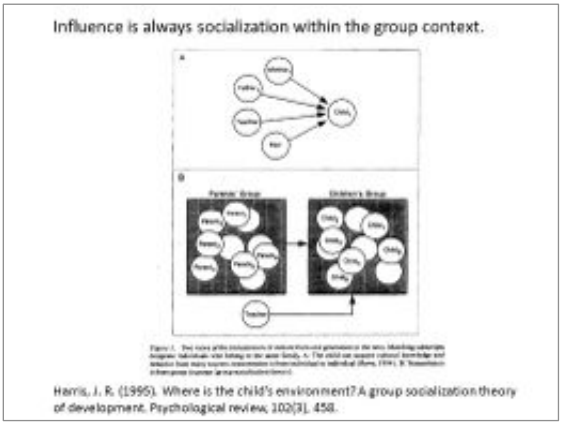
パーソナリティは、心理学で注目されていて、スライドのものがビッグファイブです。性格の 5 要素として、開放性（Openness）、誠実性（Conscientiousness）、外向性（Extraversion）、協調性（Agreeableness）、神経症傾向（Neuroticism）があります。これはいい加減そうにみえますが、心理学者たちは長年の手間をかけて、この 5 要素が主要な要素だというこ

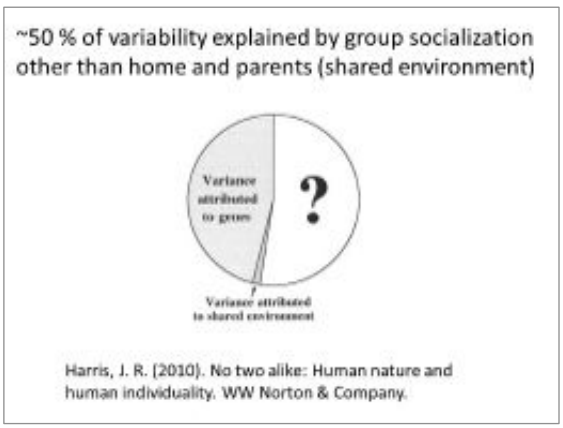
とを築いています。このような視点から、AI はきわめて狭いパーソナリティのレンジを実現しているにすぎないといえるのです。

◇スライド 45

じゃあ人格はどうできるのかというと、有力なものは『Group socialization theory』でこれはジュディス・リッチ・ハリス（Judith Rich Harris）という人が 1995 年に書いています。この人はフリーランスの研究者なのですが、これでアメリカ心理学会賞をもらっています。親が性格に与える影響は 0%という衝撃的な結論で話題になったスティーブン・ピンカー（Steven Arthur Pinker）もこの研究をエンドオース（endorse）しているのです。

◇スライド 46、47

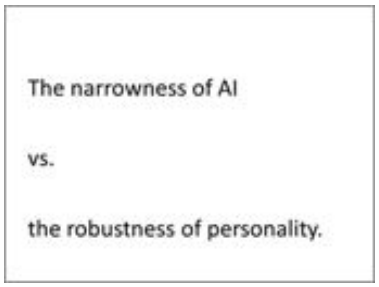




子どもが社会的にいろいろな人に囲まれて、そのなかでいろいろな影響を受けて、文脈のなかで自分のパーソナリティを築き上げるというモデルです。だから遺伝子によるバリアビリティ（variability）は、だいたい 50%くらいのことが知られているのですが、そのなかで家庭環境の影響は、ゼロに近いようなバリアビリティしか説明できないということがハリスの研究でわかっています。

◇スライド 48

このようなかたちで豊かな社会性のなかで育まれてきたパーソナリティの豊かさに対して、AI は一つの狭い問題領域に集中することで卓越する。



◇スライド 49

リサーチ・アジェンダ 4 です。インテリジェンスは、曲がりなりにも IQ テストみたいなかたちで測る（quantify）ことができるということになっていますが、パーソナリティのメジャーはものすごく曖昧なのです。先ほどのビッグファイブも主観評価などでしかやっていなくて、人工的なエージェントにはなかなか外挿できない。なので、いまおそらく研究しなくてはいけないのは、パーソナリティのメジャーは、どのように定量的に定義できるのかということだと思うのです。

たとえばビジーチャイルドといっても、それは印象評価にすぎなくて、どういう要素がないからこの人はパーソナリティが狭いスペクトラムにしかいないということをエビデンスに基づいてきちんといえるようにしておく。そうすると、こういうエージェントの開発にすごく資すると思います。

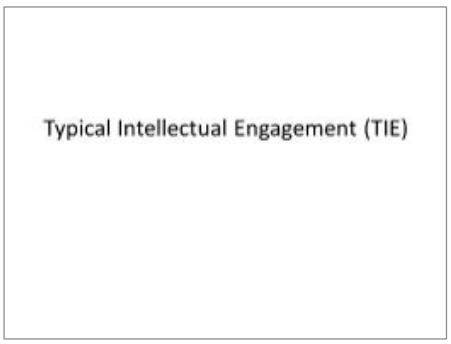
先日、TEDxTokyo のときにソフトバンクのエンジニアグループがいて、彼らはエモーション（emotion）を AI にインプリメントしているとクレームしているのです。無理なんだよ、そのクレームは。だってエモーションのモデルなんてないんだから。工学者は、ときどきすごいデタラメを言うのですよ。工学者が言う「エモーションを実装した」なんてものはデタラメもいいところですよね。「えっ」みたいな感じですよ。インテリジェンスについては、再帰関数の実行ということでチューリングマシン的なユニバーサリティー（universality）を定義できるのですが、エモーションってよいモデルがそもそもありませんから。

だからこそ、AI ではまだエモーションの実装ができていないということにもなるのです。チェスや将棋をうまく指す AI があるというクレームはよいのですよ。これは 100 万通りくらいの手を評価関数で最適化すればよい。感情を実現している AI やロボットがあるというクレームは、まったくナンセンスです。そもそも認知科学や脳科学で感情やパ

ーソナリティがよく定義されていないのです。そこを定義しなくてはいけないのでリサーチをする必要がある。

◇スライド 50

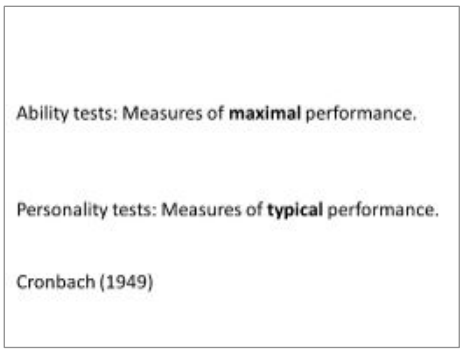
IQ テストなどは、みなさんの最大の能力を測っているということなのです。つまり、IQ テストをするときに、池上さんなんかは、IQ テストをして 200 を超えて測定不可能といわれたらしいです、これはいま適当につくった話ですが。（笑）IQ テストをしている 10 分間は、すべてのことを放り投げてそれに集中してマックスにするということを前提として、学校の先生などは IQ テストをするのです。だからマキシマム（maximam）なパフォーマンスなのです。



パーソナリティというのは、ティピカル（typical）なパフォーマンスで、池上さんがせっかく IQ が高くても、ガールフレンドが多すぎたとか酒を飲み過ぎたとかで、研究者としてのパフォーマンスがマキシマムではなくてティピカルである可能性があるのです。つまり、池上さんとしては典型的なパフォーマンスです。ここに IQ テストとパーソナリティの根本的な違いがあって、マックスに一つのことに集中して IQ テストを測るという考え方とティピカルなパフォーマンスをしめすパーソナリティという考え方がぜんぜんちがうということです。

◇スライド 51

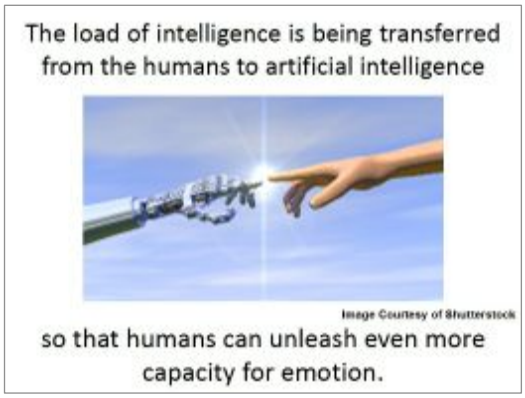
そういうところから TIE（Typical Intellectual Engagement）というようなことも出てきて、もともと潜在的に能力が高くても、じっさいにその人がデイリー・ベイシス（daily basis）でエンゲージメント（engagement）するかというようなことを評価する。



◇スライド 52、53

マキシマムパフォーマンスは、コンピュータには向いています。コンピュータは、ほかにすることがないから、ずっと集中しているのです。飽きもしないし、文句も言わない。でもいっぽうで人間というのは、日常生活のなかでティピカルなパフォーマンスをするので、このあいだをどうつなぐかが重要なポイントになるということです。

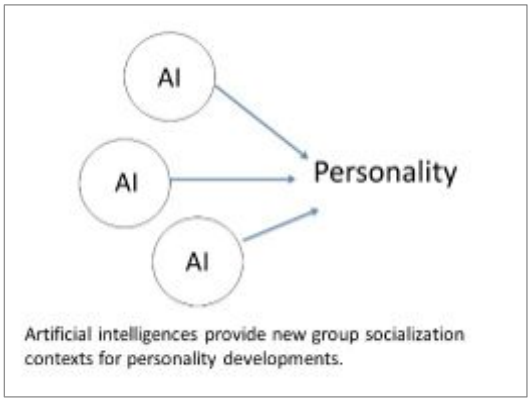
◇スライド 54



さらに、長い目でみてみます。力をもった人が、いまあまり意味がない。ブルドーザなどがあるのですから、力をもっているということはまったく意味がないですよね。それと同じように計算が速いというのもコンピュータにさせればよいから意味がないですよね。そのようなかたちでだんだんとわれわれの知性の一部分がどんどんと AI のほうにトランスファーされていく。記憶などもそうですよね。

そのときに人間は、おそらく賢さという…。たとえば京大生は賢いということをオブセッション（obsession）で生きてきた人たちじゃないですか、簡単にいうとね。京大に入って偉いとか、河合塾の模試で何位だとか、偏差値でいくつだとか。

◇スライド 55、56

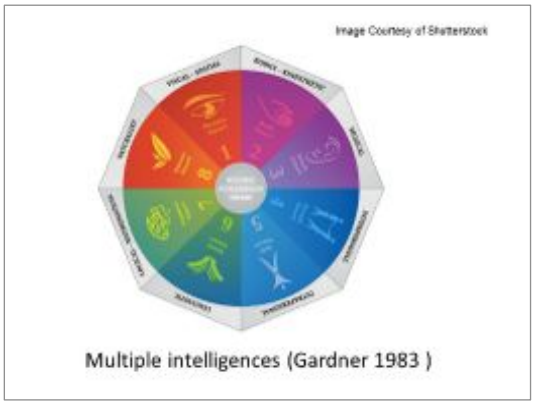


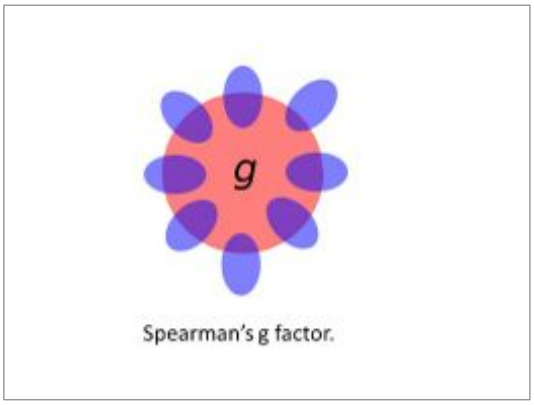
その賢さというオブセッションはだんだんと意味がなくなるだろうと予想されるのです。東ロボ君という東大の入試問題に向かう人工知能を新井紀子さんがつくっています。だから、それ以外の要素を人間が出すのが大事です。AI が人間を囲むようになったときに、人間がどのようなパーソナリティを発展させればよいのかが、おそらくこれからリサーチ・アジェンダとして大事です。これがリサーチ・アジェンダ 5 です。これはものすごく深いポイントです。

たとえば Google の自動運転カーがあるときに、ぜんぶ Google の自動運転車にゆだねるというオプションもあるのだけれども、おそらく人間はすこしは自分で「もうちょっと早く行け」とか「ここを右に曲がれ」とか、Fun to Drive だったら「ちょっと運転させろ」みたいなことがあると思うのです。人工知能の能力と人間の能力でどのように仕事を分けあうのかという研究が大事です。

たとえば、いまチェスではディープブルーがカスパロフを敗ってしまったので、新しいチェスが定義されて、人間がコンピュータを使いながらチェスをすることも行なわれています。コンピュータ・アシステッド・チェス（Computer assisted chess）みたいな新しい概念は、人間がすべてをするのではなくて、ある程度 AI のアシストを受けてどのように能力を発揮できるかというようなことです。この研究分野はこれから熱くなることが予想されるのです。

◇スライド 57、58

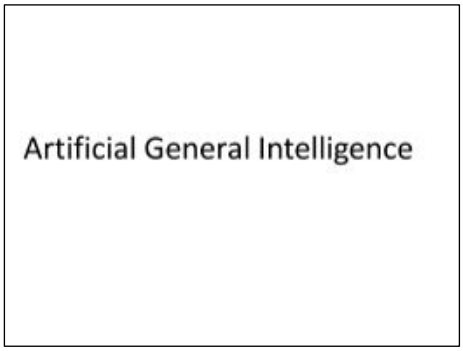




ハワード・ガードナー（Howard Gardner）がマルチプル・インテリジェンス（Multiple intelligences）といっていて、この手の話は一般ピープルが好きなのです。「知性は多重であり、あることは得意でも別のことは苦手なことがある。だから人間は、みんな違ってみんなよい」というような話が好きなのですね。しかし、これは学問的なかなりダウトフル（doubtful）なものです。スピアマンの g 因子は、そのようないろいろな異なる能力の共通因子を探ろうということなので、統計学的にいうと、マルチプル・インテリジェンスみたいな考え方は、批判があるのですよ。そういうものではないと。

◇スライド 59

そこでアーティフィシャル・ジェネラル・インテリジェンス（Artificial General Intelligence）というものが出てくるのです。これがさいごの研究アジェンダです。汎用人工知能みたいなものができると、人工知能はこれは得意だけれどもこれは苦手みたいなことがだんだんとなくなっていくと考える。そのときに、はたしてエンボディメント（embodiment）が必要なのかどうかが研究されるべきことです。



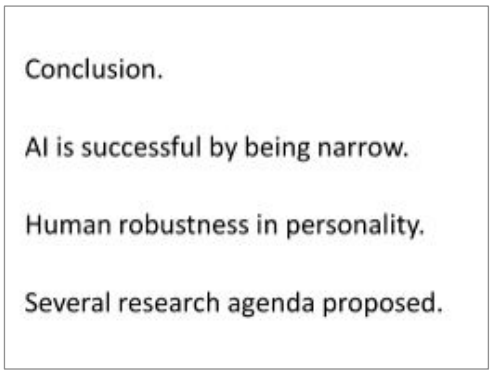
◇スライド 60

スティーブ・ウォズニアック（Stephen Gary Wozniak）という Apple のコファウンダー（Co-founder）の 1 人が「コーヒー・メイキング・タスク（Coffee making task）」ということを提案しているのです。これはなにかというと、ぼくが池上さんの下宿に行って、池上さんのためにコーヒーをつくるのです。

これは大学院生にときどき聞いてすごくおもしろいのですが、「君の家ではコーヒーをどう入れますか」と聞くといろいろな入れ方があるのですよ。コーヒー豆の保存についても、冷蔵庫に入れているとか、いろいろなケースがある。挽き方も手で挽く人とか、ミルのなかで電動で挽くとか、そもそもコーヒーを飲まない人もいる。スターバックスが近くにあるので、自分ではコーヒーをつくらなくてスタバで買ってくるというような人もいるのです。

そういうすべてのケースにおいて、コーヒーをつくるというタスクを人工知能ができるのかどうかということは、エンボディメントの話と関連している。これが重要なイシュー（issue）だというのは、アーティフィシャル・ジェネラル・インテリジェンスにおいて、身体性が必須なのかどうかという研究が求められているということです。

◇スライド 61



きょうは、ぼくの趣味である研究である意識の研究をいっさいおいておいて、人工知能の研究について、いくつか例示的に人工知能研究のアジェンダを話しました。基本的に私は、AI は人間の脳と比べると狭いことで成功していると認識しています。つまり、AI というのは、スーパー・サヴァンである。

いっぽうで人間のロバストネス（robustness）は、感情を含むパーソナリティのほうにある。たとえば、スティーブ・ジョブズ（Steve Jobs）のパーソナリティというのは、すごく素晴らしいことがあるいっぽうで、人がスティーブになにか意見を言うと、「このクソやろう。そんなクソみたいな意見をもってきやがって」とその場ではケチョンケチョンにけなす。翌日アップル（Apple）にその人が出社すると、スティーブ・ジョブズがみんなの前で「みんな、素晴らしいアイデアを思いついたんだけど」と昨日、その人がスティーブに言った意見を自分の考えであるかのように言いふらしている。そんなひどい人だったのですね。

そういうスティーブ・ジョブズのパーソナリティというのは、なんなのか。AI の賢さみたいなアプローチではまったく説明できないので、そこを印象評価やアネクドータル（anecdotal）な話ではなく、どのように定量化・構造化するかが重要なリサーチのチャレンジだと思います。そういう視点からいくつかのリサーチ・アジェンダを提案しました。

日本でも人工知能の研究そのものだけでなく、人工知能と人間との関わり、それから社会的インパクトとの関わりについて、いろいろなことをする必要があると思って、きょうは急遽このようなプレゼンにしました。以上です。

三輪●ありがとうございました。それでは質問などはありますか。

A氏●人工知能の再帰的なところにあって、人間の知性から逸脱の方向にあるというのは、あたりまえといえばあたりまえです。問題なことは、パーソナリティを人工知能で考えると誤解を助長する可能性があると思いました。

昨日もインダクション（induction）とデダクション（deduction）というものがあって、アナロジー（analogy）とかアブダクション（abduction）の軸を書きましたが、文系とか芸術家の人はそういうものに敏感です。インダクションとかデダクションというものは、再帰的な構造とかチューリングマシンだとか、そういうところでできる。ようするに肝を置き換えたり、代入したりという構造なのです。理科系の人はそっちのことはよくわかる。だけれどもアナロジーやアブダクションなどは、軸は書けないようなもので、ぜんぶ異質なものです。異質なものが連続していて、つながりがわからないにもかかわらず結果的につながっているようにみえるというだけなのです。

言語学なども、午前中も問題になったパラディグマ（paradigma）とシンタグマ（syntagma）があります。シンタグマは、置換や代入で理科系の人間が理解できる簡単な話です。しかし、パラディグマというものは、ぜんぜん異質なものが並んでいて、まったく関係ないのに並置するとか、選択する。そういうものになっている。インダクション、デダクションのグルグルと回っているようなものから逸脱するところというのが、さっきのいっていたように人間的なものと関係しているのですね。

私は、そのような意味では「シンギュラリティ万歳」で、はやくそういう時代が来て、人工知能的な頭のよさなんてみんな死ねばと思うのですが。それはおいておいて、シンギュラリティを心配するというのは、南北戦争のころに南部の人たちが「黒人に自由を与えてよいのか」ということと同じような心配に思えてしまう。

だから、半分は別にどうでもいいと思っていて、茂木さんの言っていることはわかるのですが、「人工知能におけるパーソナリティの研究」なんて話は、理科系の人間が理科系のセンスでしてもよいのか。先ほどのアブダクションなんて、部分から全体をでっちあげるという、ものすごく強引な話を含んでいる。不連続で異質であるにもかかわらず結果的にはわかってしまうとか――音を出して、みんな「これはよいか、悪いか」と。なにがよいのか悪いのかわからないのだけれども、一瞬よい音が出たら、みんな「ハッ」と同時に思ってしまう。そのようなことで、理科系のセンスでは難しいところがある。それでパーソナリティなどをまた人工知能のなかでするというのが…。

人工知能のなかにパーソナリティのようなものはぜんぜんない。パーソナリティだとか多様性だとか異質なものの連続性――異質でありながら連続であるものをみようとすると、連続的なものでそういったものをでっちあげるという方向になってしまう。文化系とか芸術家は、昔からそういうことを言っていたにもかかわらず、理科系の人間は、

ぜんぜん理解していないからそういう話になっている。それについてはどう思われますか。

茂木●逆に聞きたいのですが、「パーソナリティについては放っておけ」ということですか。定量的なメジャーは無理だということでしょうか。

A氏●そんなことはないですよ。私がするとすれば、人間的なおもしろいところは、インダクション、デダクションの再帰的な構造と予定調和的な逸脱というようなものが出会っていることをどうするのかということに関するメタフォリック（metaphoric）なモデルや表現もできると思います。

茂木●ということは、インテリジェンスのメジャーの構造とまったく違ういまAさんが言ったようなしかけを含めたメジャーだということですかね、もし、そういうものがあるとすれば。メジャーという言い方が気にくわないということですか。

A氏●メジャーという言い方が気にくわないというか、メジャーと言った瞬間にこっち側のシンタグマの構造に絡みとられてしまう。

茂木●スカラー関数なのか多次元論のベクトルなのかわからないけど、たしかにいわれてみればそうだね。

B氏●メジャーの単位はなんですか。これまでもメジャーということばがさかんに出ているのですが、そのメジャーの単位は、物理だったらちゃんと単位系とあるのですよ。

茂木●体系は、たぶんハピネスなどは、「happiness」Google Scholar で検索…。

B氏●これまであなたも、さかんにメジャーということをおっしゃったと思うのですが、「それをちゃんと測りましょう」ということを。相手を客体視するための基準、量化のときの単位は…。

茂木●ぼくがいっていることではなくて、リサーチ・コミュニティのなかでは、基本は主観評価ですね。たとえば「7 ポイントスケールでいくつですか」と聞いて、その統計的な相関をみるなどのアプローチです。それはぼくがしているというよりも、コミュニティではそのようにしているということです。単位はなにかといわれたら、単位はないです。

B氏●それはユニバーサルなものになるのですか。

茂木●わかりません。

B氏●つまり、時代とか、そういうものに頼るような評価になるのかということです。

茂木●どうなのでしょうか。ただ一ついえることは、主観評価をされますよね。たとえば、さっきの御飯ですが 1 から 5 のうちでどれくらいでしたか、5 が一番として。

B氏●マイナス…。

茂木●マイナス。先生が評価された数値と相関がある脳活動があるということはわかっているのです。そういう意味によると、主観的な評価というのは…。

Ｂ氏●この評価というのは、ここのあなたとの会話においてしか意味がないのですよ。
茂木●そういうことかもしれません。一般的に認知科学のコミュニティは、これまでそのようにしてきたということです。ぼくも、もともとは物理をしていたのですが、物理における単位とか、なにを測っているのかということとは違います。
Ｂ氏●量化するというか、最適化するといってもよいのですが、なにかを最適化するときの…。
茂木●しいていうならば、人間の脳のなかでメジャーに相当する活動があるのだろうということでは、「人間の脳のある部分の活動をみている」という言い方はいちばん物理主義に接続しやすいかもしれません。脳活動自体が、文脈依存だとか関係性依存であろうとおっしゃるのであれば、それはそのとおりですということかもしれません。
Ｂ氏●あなたは、関係性よりも客観的に人に渡すことができるようなものを、いまの TI に対してするべきであるということを主張されているのですよね。
茂木●なんらかの比較ができるようなものがあるほうが便利だろうということです。
Ｂ氏●私が不思議に思うのは、そういうのは時間…。これは瞬間芸でするのだから、まあよいと。池上君がいますぐに立ち上げるのだから、5 年もてばよいという評価だというのであれば、それはそれでよいでしょう。IT というのは、所詮そういう代物なのだから、シンギュラリティができても、それはすぐに除去できてしまうシンギュラリティかもしれない。

もう一つ疑問なのですが、IT というときに、シンギュラリティの話が出たときに、シンギュラリティといっているようなことはなにかの蓄積である。その蓄積というときに、時間というものがあります。たとえば人間であったとしたら、人間の脳のなかには時間というものがたたみこまれて入っているのですよね。TI のなかにそういうものがあるとしたら、それは人間の部分空間につくられた時間のうえにしか生きていないのですよ。だからそれってほんとうに比較できる対象になるのかどうか。
茂木●どうなのでしょうか。ちょっと考えさせてください。
Ｂ氏●IT を問題化するというときに…。私自身も混乱しているのですが、ちょっと質問が消えてしまいました。
茂木●問題提起は受け止めたので、ちょっと考えさせてください。難しい問題なので。
Ｃ氏●いまのＢ氏の話に関係あるのですが、最近の生命の起源とか AI の問題は、1bit というのは、ソフトウェアでやっているとなんのメジャーでもよいような気がするんだけど、現実世界にエンボディメントするというのは、1bit がどういう物理的なシステムでつくられるのか。エンボディメントも、それだけの情報量をもたせるために、どのくらいの大きさでどのくらいの重さでどのくらいのエネルギーがいるのかは、ほんとうに測らないといけない。

それはなんでもよいとしたら、生命の起源は、中空から浮かび上がったものになってしまう。そうではないというところが生命の起源で、ただの化学反応が生命にトランジションできないことの原因の一つになっている。エンボディメントで考えることが大事で、なおかつ、アントニオ・ダマジオ（Anthonio Damasio）のいうように、身体性があると好き嫌いができてしまう。茂木さんの言うように、アーティフィシャル・ジェネラル・インテリジェンスですかね、そのときにもしエンボディメントを与えたら、AI が好き嫌いをもつので、そうした瞬間に…。もっと怖いものですね。

さっきのメジャー単位、物理的なもので単位が与えられて、体が与えられて、ダマジオ的な意味で好き嫌いができる。そうするときに、はじめて生命的なものになるので、そのときに 2 番目のシンギュラリティが脅威に――最初に言ったシンギュラリティはB氏がいうように簡単に取り除かれてしまうようなものかもしれないけど、エンボディメントしたときには、いちばん脅威になるのではないかと思う。

だから考えなくてはいけないのは、なにが 1bit を支えているかと言うことをわれわれはフリーに講じているけれども、じつはそうではない。でかくなればなるほど、1bit を支える物理的な構造が必要になってきて、それを考えることが生物の起源の問題と等価なのです。だから、なんでも情報を担えるわけじゃなくて、ある特定の分子のある特定のエネルギー状態しか担えないから生命がある特定の国に出現したと考えられるので、それは中空の話ではない。これからそれを考えなくてはいけない時代だと私は思います。

だから IT とか AI とかいっても、架空の SF の話ではなくて、現実にその状況を実現しようと思って、実験をし始めて考える。すると途端にエネルギー状態とか、どれくらいの電子がいるのかとか、そういうことがババババッと入ってきて、そこから始まると思います。

茂木●だから、なんでかというと、チャールズ・ベネット（Charles Bennett）が言った『The thermodynamics of computation』などの話がどこかで参照されなければいけない。

C 氏●だからそれは絶対にまた考えざるをえない。

茂木●そうですね。

D 氏●私は、研究者というか写真家なので、ぜんぜん違う立場なのですが、その立場からうかがいたいことがあります。昨日に村瀬先生がお話しされた未来創成学の全体につながることかもしれません。ここでフォーカスされている方向性――こういう研究手法なりアジェンダをつくって、どこに向かおうとしているのでしょうか。人間の特殊な能力を解明することに向かおうとしているのか、それをトレースして生み出そうとしているのか。つまり、未来を創ろうとしているのか、過去をリサーチしようとしているのか。ここがどっちなのかによって、まったく違うものになってくるような気がするのです。

過去のものであれば、従来科学というものが得意な分野です。いろいろな方法があっ

て、昨日の大野先生の化石のお話などいろいろとあると思うのです。その過去というところから、どういう意味でなのかはわかりませんが、1 歩だけでも未来というところに踏み出すのかどうか。そこを入れるのか入れないのかというときに、ぜんぶいっしょくたに会話できるのか。そこを切り分けて会話したほうがよいのか。ぼくの視点からすると、そこがどっちなんだろうと思いました。そこをいっしょにしてよいのか、分けたほうがよいのか。そのあたりをどうお考えなのかなと、漠然とした考えなのですが。

茂木●その疑問についてのキーワードはこれです。トランスヒューマニズム（Transhumanism）ということだと思います。トランスヒューマニズムということで検索していただくと、たくさん出てきます。トランスですから人間が人間を超えた、あるいはポストヒューマニズムという言い方もあるのです。超人間化それからポストヒューマン――人間以降のことを考え始めたのかもしれないという議論があるということです。

だから、たんに過去の人間を理解してそれを再現しようとか、そういうことを超えたリサーチ・アジェンダだと見なされていると思います。それが人間にとって幸せなのかよいことなのかどうかは、まったく別の問題です。AI のコミュニティではこのような言い方がされているということです。

D 氏●いまのお話は基本的にはそこにのった…。

茂木●ぼくですか。ぼくはもともと脳の研究をはじめたのは、ロジャー・ペンローズ（Sir Roger Penrose）の『皇帝の新しい心 コンピュータ・心・物理法則（The Emperor's New Mind）』という人工知能の前のブームがあったのです。そのときにペンローズが、「意識などは人工知能では再現できない」という考え方をしめして、それにぼくは興味をもって脳の研究に入りました。これはぼくの趣味の世界なのですが、統計的な学習規則などのアプローチでは意識は解明できないというのが、ぼくの根本的なスタンスなのです。ジュリオ・トノーニ（Giulio Tononi）がしている IIT（Integrated information theory）というみんながリファー（refer）している研究があるのですが、これでは意識の本質には迫れないというのが私の考え方です。

いまの人工知能は統計的な学習規則をやっているので、ぼくの立場は人工知能では意識はつくれない。意識はもてないのだけれども、人工知能はすごく一つのことばかりする。それで超サヴァンのような感じで、知性というメジャーにおいては人間をトランスしているようにみえるだけで、基本的に人工知能には意識がないというのがぼくの立場です。

先ほど出たカーツワイルというような人は、「人間の意識をコンピュータに移せる」といっているのです。ぼくは、それはクルクルパーだと思っているのです。カーツヴァイはクルクルパーだと思っているのです。ぼくが言っているのは、見かけじょうのインテリジェンスをメジャーにしたときに、人工知能が私たちを超えることはあるでしょう。

だから危険なのでしょう。

人工知能の狭さとは、そういうことです。人工知能は意識さえもたないくらい狭いのです。にもかかわらず、私たちをはるかに凌駕する能力をもってしまう時代を迎えているので、それに対してどうしたらよいのかというのがぼくの基本的な考え方です。ぜんぜん人工知能万歳じゃありません。

茂木●ありがとうございました。

（了）