CRDS-FY2020-XR-02

JSAI2020企画セッション報告書

# 次世代ＡＩ研究開発 ―さらなる進化に向けて―

## 2020年6月9日(火)開催

国立研究開発法人科学技術振興機構 研究開発戦略センター
Center for Research and Development Strategy, Japan Science and Technology Agency

# エグゼクティブサマリー

本報告書は、2020 年度人工知能学会全国大会 JSAI2020 における企画セッション「次世代 AI 研究開発―さらなる進化に向けて―」の内容を記録したものである。

国立研究開発法人科学技術振興機構（JST）研究開発戦略センター（CRDS）は、国の科学技術イノベーション政策に関する調査・分析・提案を中立的な立場に立って行う組織として、今後わが国が取り組むべき重要な研究領域に関する提言を行ってきている。その一環として、2020 年 3 月に「第 4 世代 AI の研究開発―深層学習と知識・記号推論の融合―」と題する戦略プロポーザル[1]を公表した。一方、人工知能学会は人工知能（AI）に関わる研究者・技術者のコミュニティーであり、毎年 6 月頃に開催される全国大会には学術界・産業界から 2,000 人を超える参加者が集まる[2]。そこで、2020 年 6 月 9 日から 12 日にかけて開催された 2020 年度の人工知能学会全国大会 JSAI2020 において、上記の戦略プロポーザルに示した方向性を論じることで、このような研究開発の活性化につなげたいと考え、本企画セッションを開催した。

以下、本企画セッションについて開催趣旨や講演のサマリーを示す。大会初日 6 月 9 日の 15:20～17:00 に実施した本セッションには 230 人を超える聴講者があり、高い関心を集めた。

**（1）趣旨説明**　◆福島 俊一（科学技術振興機構研究開発戦略センター フェロー）

これまでデータから学習した結果を用いる「即応的知能」と記号・論理を用いる「熟考的知能」という人間の知能の 2 つの側面が、別々の AI システムとして技術発展を遂げてきた。しかし今日、深層学習によって高精度なパターン認識・生成が可能になっただけでなく、柔らかな記号処理も可能になってきたことで、それら 2 つが統一的な考え方のもとで融合する可能性が見えてきた。融合によって、現状の AI の課題（学習データ・計算資源が大量に必要、学習範囲外の状況に弱い、意味理解・説明等が不得意）の克服、人間との親和性向上、応用拡大等が期待できる。このような次世代 AI の進化の方向性に取り組んでいる 2 件の招待講演に先立ち、取り組み事例を概観するとともに、国の科学技術政策の面では、文部科学省の本年度戦略目標、JST 事業（CREST、さきがけ、ACT-X 等）にも盛り込まれたことを紹介した。

**（2）招待講演「世界モデルと記号処理」**　◆松尾 豊（東京大学大学院工学系研究科 教授）

（1）で示された方向性に向けて、以前から提案している 2 階建てモデルのキーアイデアと取り組むべき研究課題を示した。2 階建てモデルは、知覚・運動系の動物 OS の上に、BERT 系の言語アプリを乗せた構造を持つ（BERT は自然言語処理分野で注目されている Transformer 型の言語モデル）。動物 OS 側は知覚の予測、言語アプリ側は発話の予測を目的関数とし、深層学習によってボトムアップに学習した世界モデル（外界をシミュレートする手段）を共有する。これは単に 2 種類の系を組み合わせたというものではなく、報酬系が 2 通りのものに分岐するという進化の結果だと考えている。発話の予測に伴い、離散化さらには社会的な蒸留（Distillation）を通して言語が生まれ、発達してきたのではないか。このような仮説の検証・実現に向けて取り組ま

[1] https://www.jst.go.jp/crds/report/report01/CRDS-FY2019-SP-08.html
[2] 人工知能学会の会員数は 2020 年 3 月末時点で個人会員 5,544 人、賛助会員 289 団体であり、昨今、会員数が減少傾向の学会が多い中、人工知能学会の会員数は増加傾向が続いている。全国大会の参加者数も JSAI2017（名古屋開催）が 2,540 人、JSAI2018（鹿児島開催）が 2,611 人、JSAI2019（新潟開催）が 2,905 人と拡大傾向が続いていたが、JSAI2020 は COVID-19 対策でオンライン開催となったため参加者数が 2,300 人前後となった。

ねばならないこととして、世界モデル、BERT、蒸留に加えて、複数の主体間のコミュニケーション（蒸留を含む）に関わる NAS（Neural Architecture Search）やマルチエージェントのメカニズム、および、現実データとの差異の最小化が挙げられる。

**（3）招待講演「確率的深層生成モデルによる実世界自律知能の創成に向けて～記号創発ロボティクスが生み出す認知アーキテクチャ～」** ◆谷口 忠大（立命館大学情報理工学部 教授）

（1）で示された方向性に合致する記号創発ロボティクスの考え方と、その実現のキーとなる確率的生成モデルの開発状況を紹介した。記号創発ロボティクスでは、実世界経験に基づき認知発達、言語獲得、運動・プランニングのスキル獲得等が進む仕組みを構成していく。これは、認知発達への構成論的アプローチと言える。定義された記号系と実世界のものを対応付けることを意図した「記号接地」は問題設定として適切ではない。実世界のものに対するラベル付けには恣意性があり、社会の中で共有されることで記号の役割を持つというものであるから、「記号創発」問題として捉えるべきである。これを実現するため、確率的生成モデルに基づく SpCoSLAM を開発してきた。これは自己位置推定と地図生成を同時に行う SLAM に、場所概念・語彙獲得を統合したものである。さらに、認識や概念獲得等の各モジュールをそれぞれ確率的生成モデルで実装し、それらの間で確率的な情報のやり取りを行いながら同時学習を進める SERKET へと発展させている。

**（4）総合討論** ◆モデレーター：中島 秀之（札幌市立大学 学長）

セッション時間の制約の中で総合討論としての十分な時間は確保できなかったため、ここではモデレーターの中島による総括コメントを要約する。

松尾・谷口のアプローチはいずれもボトムアップ型だが、中島はハイブリッド型を考えている。まず予測と予期の違いとして、予測は同じ物理レベルで実行されるが、予期は 1 段上の認知レベル（記号・物語の世界）を介して実行される。この予期の仕組みは、ボトムアップな学習だけでは難しいと考えている。次にハイブリッド型だが、深層学習の強化と記号推論の強化の両面が必要である。深層学習は騙されやすいという問題があるが、予期による推論・学習が解決策になる。記号推論では記号接地問題やフレーム問題が課題だったが、ボトムアップな学習とトップダウンな推論の組み合わせが課題解決につながるはずである。

なお、前述の戦略プロポーザル「第 4 世代 AI の研究開発―深層学習と知識・記号推論の融合―」を公表するに先立ち、JST CRDS では 2020 年 1 月 30 日に科学技術未来戦略ワークショップ「深層学習と知識・記号推論の融合による AI 基盤技術の発展」を開催し、その報告書も公開している[3]。これも上記戦略プロポーザルでの提言内容を補強し、本企画セッションでの講演内容を補完する内容になっている。

[3] https://www.jst.go.jp/crds/report/report05/CRDS-FY2019-WR-08.html
このワークショップでは、麻生英樹（産業技術総合研究所人工知能研究センター 副研究センター長）、尾形哲也（早稲田大学基幹理工学部 教授）、乾健太郎（東北大学大学院情報科学研究科 教授）、橋田浩一（東京大学大学院情報理工学系研究科 教授）、長井隆行（大阪大学大学院基礎工学研究科 教授）が参加・発表した。

# 目　次

# 1．趣旨説明

◆福島 俊一（科学技術振興機構研究開発戦略センター フェロー）

## 1．1 開催趣旨

今回の大会では、NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）と JST（国立研究開発法人科学技術振興機構）の連携企画の形で、「次世代 AI 研究開発」と題した連続セッションを開催することにした[4]。NEDO セッションでは社会実装に近いところを支える研究開発について紹介されたが、この JST セッションではその先のより基礎的な研究開発・技術チャレンジを取り上げる。

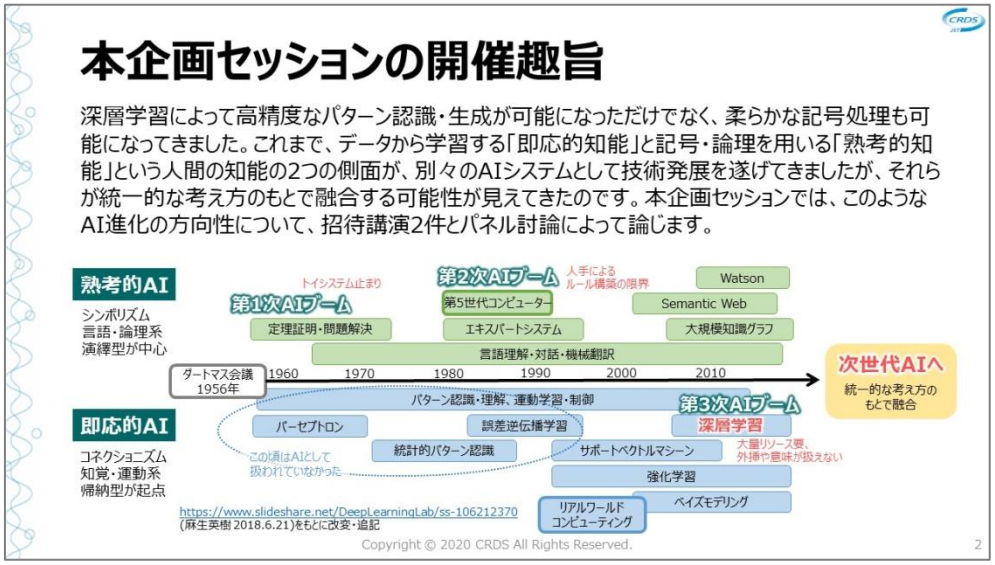


図 1-1 本企画セッションの開催趣旨

AI の第 1 次ブームから第 2 次ブームの頃は記号処理を中心とした AI が主流だったが（図 1-1 の上側「熟考的 AI」）、その当時は AI と呼ばれていなかったニューラルネットをベースとした帰納的なアプローチが深層学習（Deep Learning）に発展し、第 3 次 AI ブームが起こった（図 1-1 の下側「即応的 AI」）。これらは別系統として発展してきたが、深層学習によってパターン認識・生成だけでなく柔らかな記号処理も可能になってきたことから、融合する可能性が見えてきた。2 つを融合させて統一的に扱うような AI の姿というのがこれからの方向性になってくるのではないかということを論じていきたい。中島先生による本大会の基調講演[5]でも、後半でこのような方向性に言及されたが、その辺りを深掘りしたい。

[4] 大会初日の 2020 年 6 月 9 日（火）午後に連続開催し、13:20～15:00 に「次世代 AI 研究開発(1) 基盤技術開発と産業・社会への展開」と題した NEDO 企画セッション、15:20～17:00 に「次世代 AI 研究開発(2) さらなる進化に向けて」と題して JST 企画セッションを実施した。それぞれのプログラム内容は以下のサイトで公開している。国立のファンディング機関として AI 研究開発に関し、NEDO は社会実装を支える技術開発、JST はその先の基礎研究に重点を置いている。
「次世代 AI 研究開発(1) 基盤技術開発と産業・社会への展開」 https://www.nedo.go.jp/events/CD_100122.html
「次世代 AI 研究開発(2) さらなる進化に向けて」 https://www.jst.go.jp/crds/sympo/202006_JSAI/index.html

[5] 本企画セッションと同日 6 月 9 日（火）の午前中 10:30～11:40 に、大会プレナリー基調講演として、中島秀之（札幌市立大学 学長）が「AI 技術を活用する社会のデザイン」と題する講演を行った。

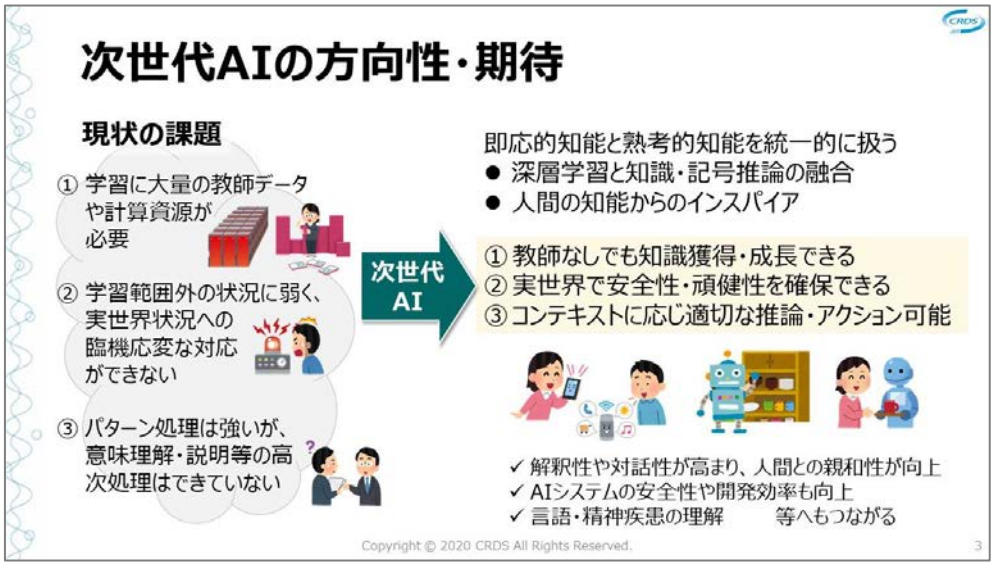


図 1-2　次世代 AI の方向性・期待

では、このような融合の方向に進むことで何が変わるか、何をできるように目指すのか（図 1-2）。まず、現 AI の課題がこのような融合の方向で解決できそうだという期待がある。現状の課題としてよく指摘されるのは、①学習に大量の教師データや計算資源が要る、②学習した範囲内ではうまくいくけれども、その範囲外には弱く、臨機応変に対応できない、③パターン処理は強いが、意味理解・説明等の高次処理はできていない、といったことである。そして、こういった課題は、深層学習をさらに発展させて記号・知識処理と融合させることで解決できるのではないかと考えられつつある。また、脳科学等の研究からも人間の知能が持つ異なる 2 つの側面が示唆されている。このような融合方向へ発展させることで、現 AI の課題が解決され、解釈性や対話性が高まり、人間との親和性が向上するとか、AI システムの安全性や開発効率も向上するとか、言語・精神疾患の理解にもつながるといった効果が期待される。

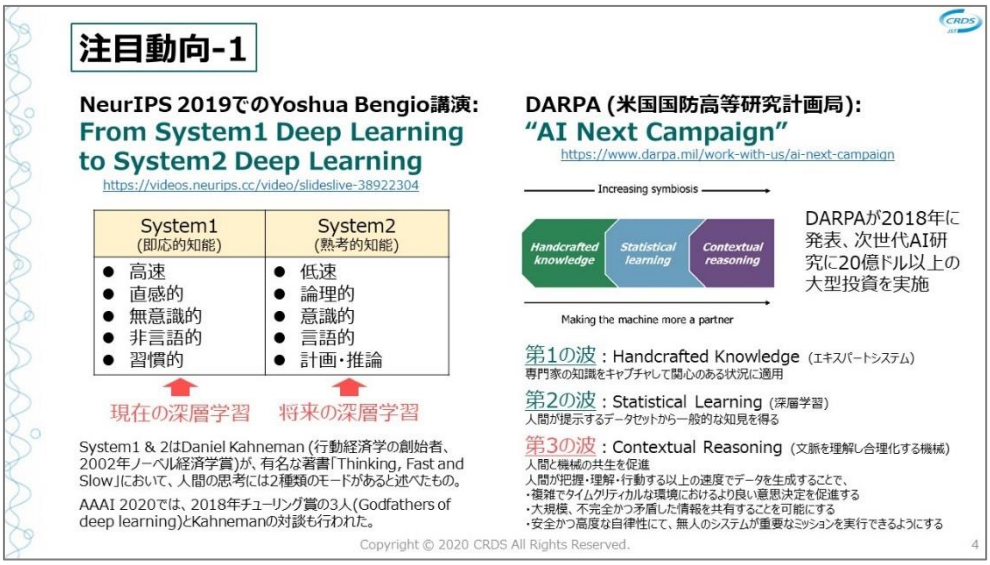


図 1-3　注目動向（その 1）

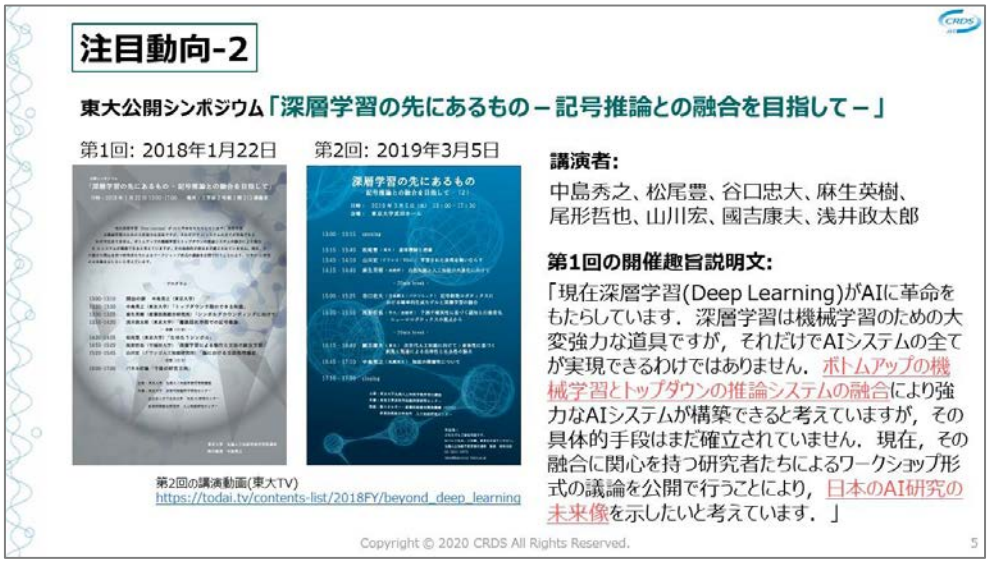


図 1-4　注目動向（その 2）

このような方向性は実際しばしば言及されるようになった（図 1-3、図 1-4）。

特に最近話題になったのは、昨年の NeurIPS で Yoshua Bengio が、現在の深層学習はシステム 1 に相当するが、今後システム 2 の深層学習も目指すとしたものである。ここでいうシステム 1・システム 2 は、ノーベル経済学賞を受賞した Daniel Kahneman が人間の思考について言っているもので、システム 1 は即応的なモード、システム 2 は熟考的なモードである。

また、アプローチ自体はこれと全く同じというわけではないが、DARPA（米国国防高等研究計画局）が「AI Next Campaign」を立ち上げている。ここでは、現状の深層学習を AI の「第 2 の波」と呼び、その先として「第 3 の波」を掲げ、大型の研究投資を始めている。

国内でも、中島先生が本日の登壇メンバーも交えて、「深層学習の先にあるもの―記号推論との融合を目指して―」と題した東大公開シンポジウムを開催されている。そのときの開催趣旨を図 1-4 に引用したが、既に 2 年前の段階で、AI の次の方向としてこの方向を目指すと言っている。本日のこの企画セッションもこの流れを汲んだものである。

## １．２　取り組み事例

即応的 AI と熟考的 AI の融合、あるいは、パターン処理と記号処理の融合という方向性に関して、非常に疎な組み合わせも含めると、これまでもさまざまな取り組みが行われている（図 1-5、図 1-6）。工学的に問題解決に重きを置いたタイプのものから、人間の知能のメカニズムを参考にしたタイプのものまで幅広い。

ざっと眺めていくと、演繹的なシステムの中に機械学習を入れて帰納的にスコアを決めるような疎な組み合わせや、深層学習をベースとした枠組みにナレッジグラフを組み合わせて知識を取り込めるようにしたものや、逆に、機械学習に基づく帰納的な処理結果を演繹型の枠組みに変換して取り込むようにしたものもある。また、シミュレーションはモデルに基づいた演繹的な処理と考えると、それと機械学習を組み合わせる取り組みも進められている。そういった中で、深層学習の発展から最近注目されるのは、ニューロ系とシンボル系・推論系を、人間がどう処理して

いるかも意識しながら融合させていく取り組みである。MIT の Neuro-Symbolic Concept Learner や、松尾先生が昨年の東大公開シンポジウムで話された知覚・運動系と記号系の 2 階建てモデルが挙げられる。早稲田大学の尾形哲也先生らが取り組まれている予測学習もこのタイプに近い。


図 1-5　融合型 AI の研究例（その 1）


図 1-6　融合型 AI の研究例（その 2）

また、いままで挙げてきたような方式とはややレベルが異なり、学問分野・研究分野といった方がよいかもしれないが、記号創発ロボティクスという取り組みがある。これは、今日このあと谷口先生が話されるが、人間の知能の理解に対する構成論的なアプローチである。

2 枚のスライド（図 1-5、図 1-6）にまとめたが、このようないろいろな取り組みが始まっている中で、特に⑤⑥に注目し、今日は松尾先生と谷口先生に招待講演をお願いした。

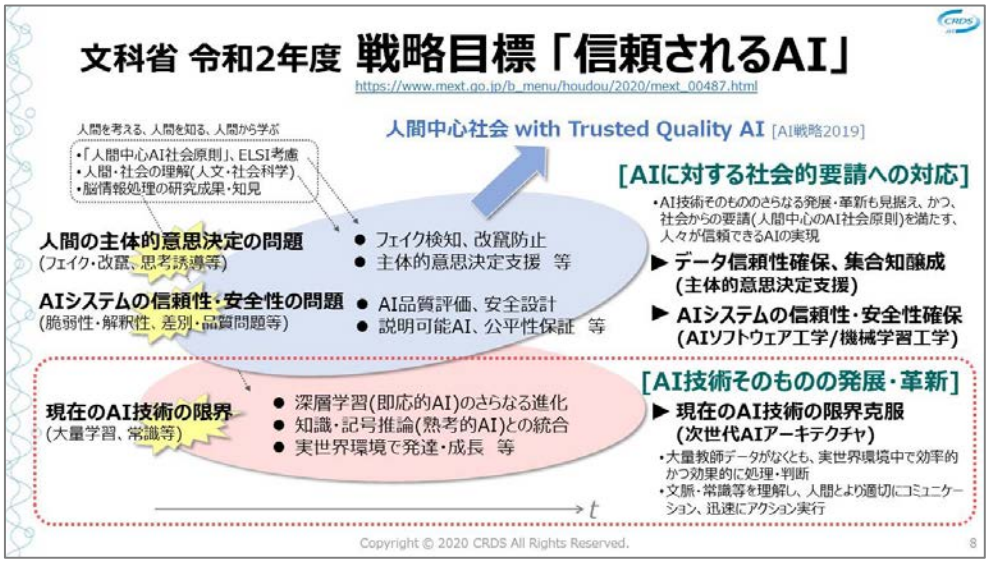


図 1-7　戦略目標「信頼される AI」

## １．３　戦略目標と研究開発プログラム

私が属している JST CRDS では国の科学技術政策に対する提言を行っており、私は特に AI 分野の研究開発戦略に関わっている。今回取り上げている融合型 AI の方向性についても調査・提言をしてきて、今年の文部科学省の戦略目標「信頼される AI」（図 1-7）につながった。

「信頼される AI」という名称だけからは、融合型 AI とは異なる話のように思われるかもしれないが、実は図 1-7 に示すように 3 層の構造になっていて、ベースとなる一番下の層（図 1-7 の赤点線枠）が今回の話に相当する。つまり、深層学習の進化だけでなく、さらに知識や記号推論を融合させて、これからの AI の形が発展していく。そして、そういった AI 技術そのものの発展・革新の上に、社会からの要請にも対応し、工学的な安全性・信頼性の確保や、人間の主体的意思決定の支援も重ね合わせた形で「信頼される AI」を目指そうという戦略目標になっている。

この戦略目標を受けて、JST が運営する研究開発ファンディングプログラム（CREST、さきがけ、ACT-X 等）が立ち上がり、現在ちょうど公募中[6]なので、今回示すような方向性の研究開発に意欲的に取り組まれる研究グループは応募を検討いただけたらありがたい。また、私たち JST CRDS ではこれまで、ここに示すような提言やワークショップ等を行ってきた。これらは戦略目標「信頼される AI」のもとになっており、また、本日のトピックの詳細情報としても参考にしていただけると思う。

### 2020 年度に開始された AI 関連の JST プログラム

- JST 戦略的創造研究推進事業 CREST「信頼される AI システム」
  https://www.jst.go.jp/kisoken/crest/research_area/ongoing/bunya2020-4.html
  研究総括：相澤 彰子（国立情報学研究所コンテンツ科学研究系 教授／所長補佐）

[6] この企画セッションを実施した 2020 年 6 月 9 日の時点では、これらの JST プログラムは提案を募集中であったが、本報告書が発行される時点では既に 2020 年度の募集期間は終了している。

- 同さきがけ「信頼される AI」
  https://www.jst.go.jp/kisoken/presto/research_area/ongoing/bunya2020-5.html
  研究総括：有村 博紀（北海道大学大学院情報科学研究院 教授）
- 同 ACT-X「AI 活用で挑む学問の革新と創成」
  https://www.jst.go.jp/kisoken/act-x/research_area/ongoing/bunya2020-1.html
  研究総括：國吉 康夫（東京大学大学院情報理工学系研究科 教授）
- JST 未来社会創造事業「超スマート社会の実現」領域
  重点公募テーマ「異分野共創型の AI・シミュレーション技術を駆使した健全な社会の構築」
  https://www.jst.go.jp/mirai/jp/uploads/application-guideline-r02-c6.pdf#page=6
  テーママネージャー：鷲尾 隆（大阪大学産業科学研究所 教授）

**戦略目標のもとになった JST CRDS で発行した報告書や企画・開催したワークショップ等**

- 戦略プロポーザル「第 4 世代 AI の研究開発―深層学習と知識・記号推論の融合―」（2020 年 3 月）
  https://www.jst.go.jp/crds/report/report01/CRDS-FY2019-SP-08.html
- 科学技術未来戦略ワークショップ報告書「深層学習と知識・記号推論の融合による AI 基盤技術の発展」（2020 年 3 月）
  https://www.jst.go.jp/crds/report/report05/CRDS-FY2019-WR-08.html
- 戦略プロポーザル:「AI 応用システムの安全性・信頼性を確保する新世代ソフトウェア工学の確立」（2018 年 12 月）
  https://www.jst.go.jp/crds/report/report01/CRDS-FY2018-SP-03.html
- 科学技術未来戦略ワークショップ報告書「機械学習型システム開発へのパラダイム転換」（2018 年 3 月）
  https://www.jst.go.jp/crds/report/report05/CRDS-FY2017-WR-11.html
- 2019 年度人工知能学会全国大会（第 33 回）企画セッション「機械学習における説明可能性・公平性・安全性への工学的取り組み」（2019 年 6 月）
  https://www.jst.go.jp/crds/sympo/201906_JSAI/index.html
- 戦略プロポーザル「複雑社会における意思決定・合意形成を支える情報科学技術」（2018 年 3 月）
  https://www.jst.go.jp/crds/report/report01/CRDS-FY2017-SP-03.html
- 科学技術未来戦略ワークショップ報告書「複雑社会における意思決定・合意形成を支える情報科学技術」（2017 年 10 月）
  https://www.jst.go.jp/crds/report/report05/CRDS-FY2017-WR-05.html
- 公開ワークショップ報告書「意思決定のための情報科学～情報氾濫・フェイク・分断に立ち向かうことは可能か～」（2020 年 2 月）
  https://www.jst.go.jp/crds/report/report05/CRDS-FY2019-WR-02.html
- 研究開発の俯瞰報告書「システム・情報科学技術分野（2019 年）」（2019 年 3 月）
  https://www.jst.go.jp/crds/report/report02/CRDS-FY2018-FR-02.html

# 2．招待講演「世界モデルと記号処理」

◆松尾 豊（東京大学大学院工学系研究科 教授）[7]

## 2．1　2階建てモデル

「世界モデルと記号処理」と題して、最近考えていること、特に重要だと思っていること話したい。まず 2 階建てのモデルについて話し、そこに関わる関連研究も紹介する。続いて、僕の主張したい点を話し、最後にこれからどういう研究をしていけばよいかを述べる。

人工知能の分野では、ずっとシンボルとパターンの戦いが続いてきた。例えば、シンボル派の代表は Marvin Minsky、パターン派は Geoffrey Hinton や Rodney Brooks で、Brooks は身体性が重要だと言っている。

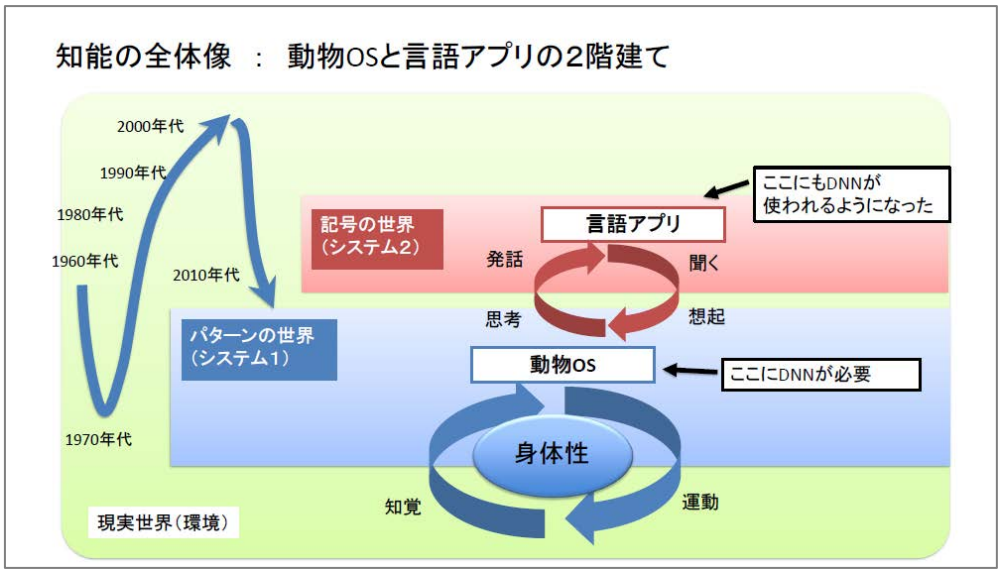


図 2-1　知能の全体像：動物 OS と言語アプリの 2 階建て

僕が以前から言っているのは、図 2-1 のような 2 階建てのモデルで、パターンの世界が 1 階部分、記号の世界が 2 階部分、その間で相互作用をしているというものである。特にこのパターンの世界、つまり 1 階部分に、ディープニューラルネットワーク（DNN）が使われるようになって、画像認識や音声認識をはじめいろいろな知覚の問題が解けるようになり、運動の問題についても、Pick and Place といったロボットの適応的な行動も可能になってきた。同時に、2 階部分の言語系についても、BERT を中心に非常に精度の高いモデルが出てきている状況である。

既に紹介があったように、NeurIPS 2019 で Yoshua Bengio が「From System 1 Deep Learning to System 2 Deep Learning」と題して講演した[8]。これは現在の深層学習（Deep Learning）が 1 階部分のパターン処理を扱っていたのに対して、これからは 2 階部分に相当するシンボル処理も対象にしていくと言っている。Bengio は図 2-2 のようなモデルを示している。

[7] http://ymatsuo.com/japanese/vita.html
[8] https://youtu.be/T3sxeTgT4qc

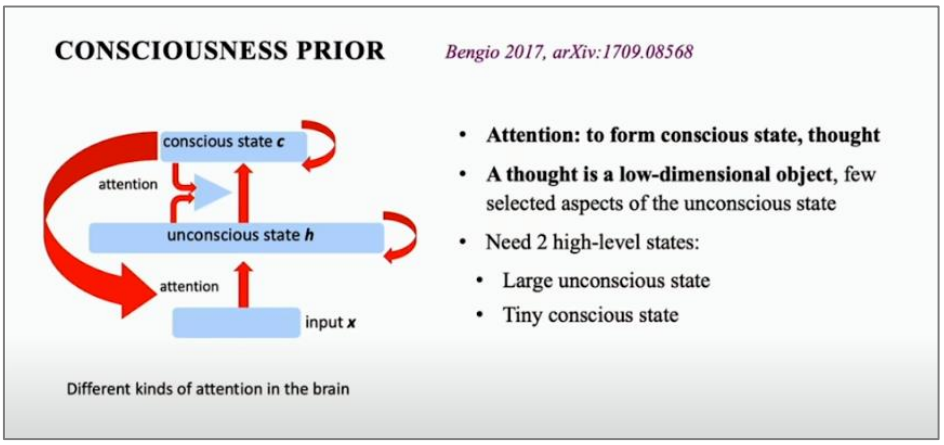


図 2-2　Bengio のモデル（Conscious Prior）

これに関して、Gary Marcus[9]と Bengio との言い争いがおもしろい[10]。Marcus からすると「Bengio は最近言うこと変えて、ついにシンボルが大事と言うようになった」みたいに言っている。しかし、Bengio としては「そんなことない、これは従来のシンボルとは違う」と言って、結構食い違っている。Marcus は、深層学習はいろいろある技術の 1 つで、それとシンボルが融合するという見方をしているが、Bengio の考え方はそれと異なり、深層学習のフレームワークを拡張することでシンボルを扱えるようにしようとしていて、深層学習が今のままの形である必要は全くないと考えている。深層学習というのは深い学習あるいは表現学習（Representation Learning）ということしか言ってないので、現在のニューラルネットワークに縛られる必要は全くなく、いろいろなアーキテクチャがあっていいし、その定義も移り変わっていいし、学問はそういうものだろうと言っている。単にシンボルを深層学習に組み合わせればいいという話ではないので、僕は Bengio の考えの方が適切だと思っている。

さて、僕の話は、まず 2 階建ての 1 階部分を話し、次に 2 階部分を話し、さらに、それらの融合という順に進める。

## ２．２　１階部分（動物ＯＳ）と世界モデルの研究動向

1 階部分で一番重要なのが世界モデル（World Model）[11]だと思っている。今までの記憶から未来を予測する力が知能だと Jeff Hawkins が言っているが[12]、これは、学習によって脳の中に世界モデルを作り、この世界モデルを用いて未来をシミュレーションしているということである。

この世界モデル系の研究には大きく World Models 系と Video Prediction 系があると自分では分けている。World Models 系は、観測があったときに 3 次元あるいは 2 次元の内部表現と遷移モデルを作ろうというものである。VAE（変分自己符号化器）が使われることも多い。ちなみに、松尾研の研究もこのあたりが結構多い。一方、Video Prediction 系は、動画の先のフレームを予測しようというものである。どちらかというときれいな画像が生成できることを目指し、そのた

[9] Gary Marcus は科学者かつ作家かつ起業家で、ニューヨーク大学心理学部教授、Geometric Intelligence CEO も務めている。
[10] https://www.zdnet.com/article/whats-in-a-name-the-deep-learning-debate/#:~:text=Yoshua%20Bengio%20and%20Gary%20Marcus,term%20%22deep%20learning%22%20means.
[11] 「世界モデル」という言い方の他に「内部モデル」「力学モデル」という言い方もある。昔、Phlip Johnson-Laird が「メンタルモデル」と呼んだものも近い。
[12] Jeff Hawkins 著「On Intelligence」、和訳書は「考える脳 考えるコンピューター」

めに適切な内部表現と遷移モデルが必要になる。

World Models 系の研究で代表的なものの 1 つは、David Ha と Jürgen Schmidhuber による「World Models」という名前そのままの論文[13]である。VAE を適用して情報を圧縮し、かつ、RNN（再帰型ニューラルネットワーク）を動かすことによって、未来がどのように移行するかを予測しながらプランニングしていく。世界モデルを使わなかった場合の強化学習の結果と、世界モデルを使った場合の結果を比較して、世界モデルを使った方がスムーズな結果が得られたことが示されている。

ちなみに、この Schmidhuber は、世界モデルのコンセプトペーパーのようなものを 2015 年に書いている[14]。最近も、強化学習ですごく面白いと僕が思う論文も出している[15]。強化学習は期待報酬を最大化するようなアクションを求めるということだが、僕は何かおかしいとずっと思っていた点があって、この論文はそれを指摘している。つまり、観測とアクションから期待報酬を算出するのではなく、観測と報酬からアクションを出すべき。例えば、今この場にいておなかがすいて何か食べたいというときに、どうアクションするかを出す問題ではないかと。かなり考え方を変えるようなものだが、正しい考え方だと思う。さらに彼は 1990 年というかなり早い時期に書いた論文[16]で既に、深層強化学習や、世界モデル[17]、自己教師あり学習、好奇心と飽き、メタ学習等、重要な概念を示していて、これも凄いと思う。また、Ian Goodfellow が作った GAN（敵対的生成ネットワーク）に対して Schmidhuber は、これは自分が考えたもので 1990 年と 1991 年の論文[18]に書いてあると主張している。ここでは若干攻撃的な面を見せているが、本当に天才的な人だと思う。

World Models 系の論文でもう 1 つ有名なものが、DeepMind の GQN（生成クエリネットワーク）である。異なる視点のデータから空間の内部表現を作るというもので（図 2-3）、2018 年に Science 誌に掲載された[19]。この後も World Models 系の論文、GQN 関連の論文はいろいろ出ている[20]。例えば、物体が 1 個ならうまく再現できるが複数個ではあまりうまくできないという問題がある。そこで、画像から物体検出してオブジェクトにエンコードし、オブジェクトの関係・相互作用をグラフニューラルネットでモデル化することで復元しようというアプローチが考えられている。あるいは、領域に区切って、どの領域に何があるか、領域毎に潜在変数を置いて、うまく復元しようというアプローチもある。

[13] https://arxiv.org/pdf/1803.10122.pdf
[14] Jürgen Schmidhuber (2015), “On Learning to Think: Algorithmic Information Theory for Novel Combinations of RL Controllers and RNN World Models”.
[15] Rupesh Kumar Srivastava, Pranav Shyam, Filipe Mutz, Wojciech Jaskowski, Ju¨rgen Schmidhuber (2019), “Training Agents using Upside-Down Reinforcement Learning”.
[16] Jürgen Schmidhuber (1990), “Making the World Differentiable: On Using Self-Supervised Fully Recurrent Neural Networks for Dynamic Reinforcement Learning and Planning in Non-Stationary Environments”.
[17] 論文中では「モデルネットワーク」という名前を使っていた。
[18] Jürgen Schmidhuber (1990, 1991, 2020), “Generative Adversarial Networks are special cases of Artificial Curiosity (1990) and also closely related to Predictability Minimization (1991)”.
[19] S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz, Helen King, Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu, Demis Hassabis (2018), “Neural scene representation and rendering”. https://science.sciencemag.org/content/360/6394/1204
[20] 例えばICLR 2020で発表されたThomas Kipf, Elise van der Pol, Max Welling (2020), “Contrastive Learning of Structured World Models” や Jindong Jiang, Sepehr Janghorbani, Gerard de Melo, Sungjin Ahn (2020), “SCALOR: Generative World Models with Scalable Object Representations”.

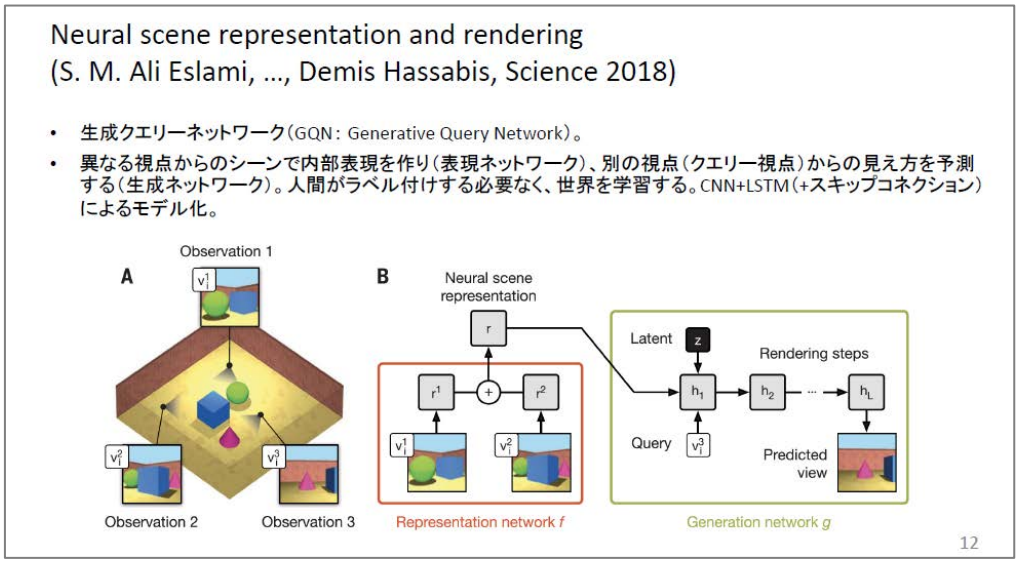


図 2-3　生成クエリネットワーク（GQN）

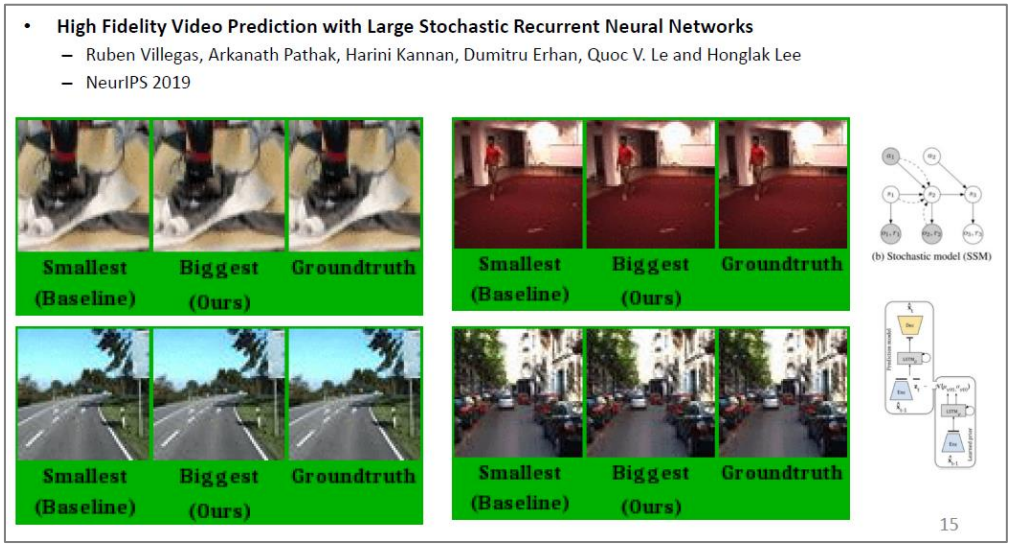


図 2-4　High Fidelity Video Prediction

一方、Video Prediction 系の研究もたくさん取り組まれている。特にコンピュータービジョンの研究分野で活発で、有名なものとして High Fidelity Video Prediction[21]が挙げられる（図 2-4）。SSM（状態空間モデル）、LSTM（Long Short-Term Memory）を使ったようなモデルをとても大きくして、パラメーターの数も非常に増やしたことで、いくつかのフレームが与えられると、その先までかなり精度よく予測できるようになったということである。

また、VideoFlow というノーマライジングフローを用いた技術[22]が提案されている。フロー系の技術を使って動画の予測をモデル化している。3 フレームだけ使って、残りの 100 フレームで何が起こるかを予測することができる。確定的ではなく確率的な分布を出せるので、複数のシナ

[21] NeurIPS 2019 で発表された Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V. Le and Honglak Lee (2019), “High Fidelity Video Prediction with Large Stochastic Recurrent Neural Networks”.
[22] ICLR 2020 で発表された Manoj Kumar, Mohammad Babaeziadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, Durk Kingma (2020), “VideoFlow: A flow-based generative model for video”.

リオがあったときに、どういう未来になるかを出せる技術になる。

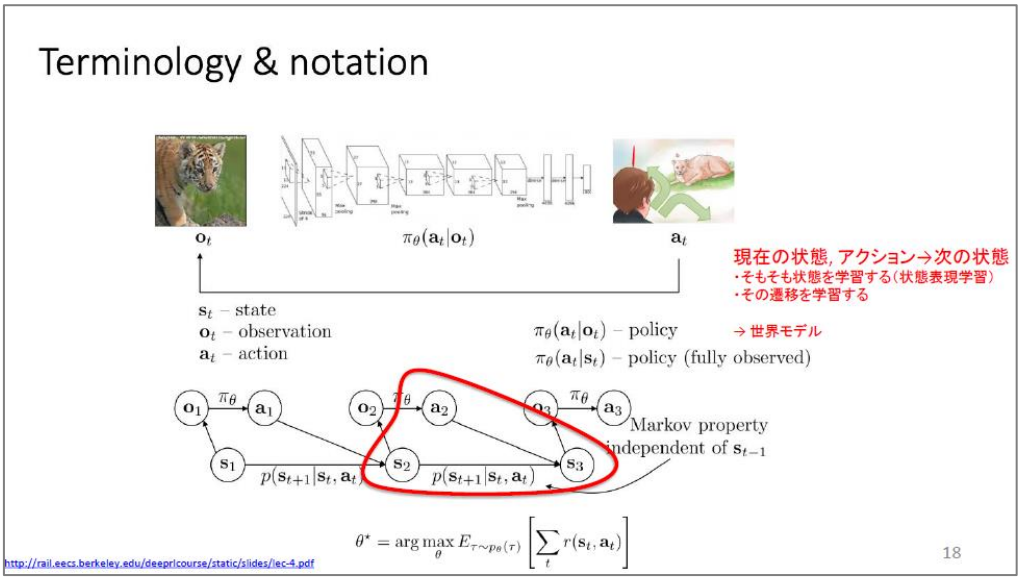


図 2-5　Terminology & Notation

このような世界モデルがどのように使われるかというと、強化学習、アクションの生成に使える。強化学習にはモデルフリー強化学習とモデルベース強化学習がある。図 2-5 の MDP（マルコフ決定過程）の図のように、ある状態からあるアクションを行うと次の状態に変わるという状態遷移が既知ならモデルベースと言う。例えばアルファ碁のような囲碁、そして将棋も既知なのでモデルベースの強化学習である。Deep Mind の Atari ゲームは、状態遷移が分からないのでモデルフリーである。最近も Atari57 の全てのゲームで人間を超えたという発表があったが[23]、モデルフリーでうまく学習できているのにはトリックがある。つまり、アクションが離散空間なので限定されているし、試行回数も極端に多くできるから実現できた。しかし、実環境で動くロボットを使う場合はそのような膨大な試行回数はできない。試行回数を減らすためにモデルを学習によって作ろうというのが、要するに世界モデルということになる。状態そのものを学習する状態表現学習と、その遷移の学習と、両方を学習しないといけないので結構難しかった。この世界モデルの学習ができるようになると、モデルベース強化学習ができるようになってくる。

これについてもいろいろな研究が進んでいて、Atari のゲームでフレームを予測するという 2015 年の研究[24]はよく知られている。これは、自分がどんな行動するとどんなことが起こるかをオートエンコーダーで予測し、それに基づいて動くというものである。また、NIPS 2017 で発表された Imagination-Augmented Agents（I2As）という研究（図 2-6）[25]がある。これも、どんなことをするとどんなことが起こるかを想像する。「倉庫番」等のゲームでうまくいくことを検証している。最近だと Dream to Control という研究（図 2-7）[26]がある。これは、アクションと状

[23] “Agent57: Outperforming the human Atari benchmark” (2020/3/31), https://deepmind.com/blog/article/Agent57-Outperforming-the-human-Atari-benchmark
[24] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Satinder Singh, Richard Lewis (2015), “Action-Conditional Video Prediction using Deep Networks in Atari Games”.
[25] Théophane Weber, Sébastien Racanière, David P. Reichert, Lars Buesing, Arthur Guez, Danilo Jimenez Rezende, Adria Puigdomènech Badia, Oriol Vinyals, Nicolas Heess, Yujia Li, Razvan Pascanu, Peter Battaglia, David Silver, Daan Wierstra (2017), “Imagination-Augmented Agents for Deep Reinforcement Learning”.
[26] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, Mohammad Norouzi (2020), “Dream to Control: Learning Behaviors by Latent Imagination”.

態を潜在空間で想像することによって、こういうことが起こりそうというのを想像して行動できるというもので、活発に研究されている。

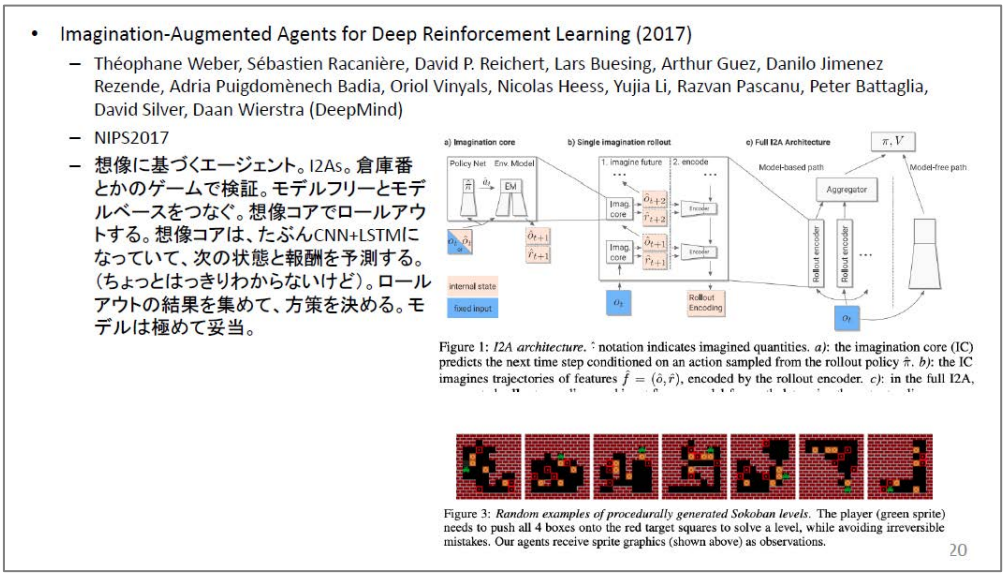


図 2-6　Imagination-Augmented Agents

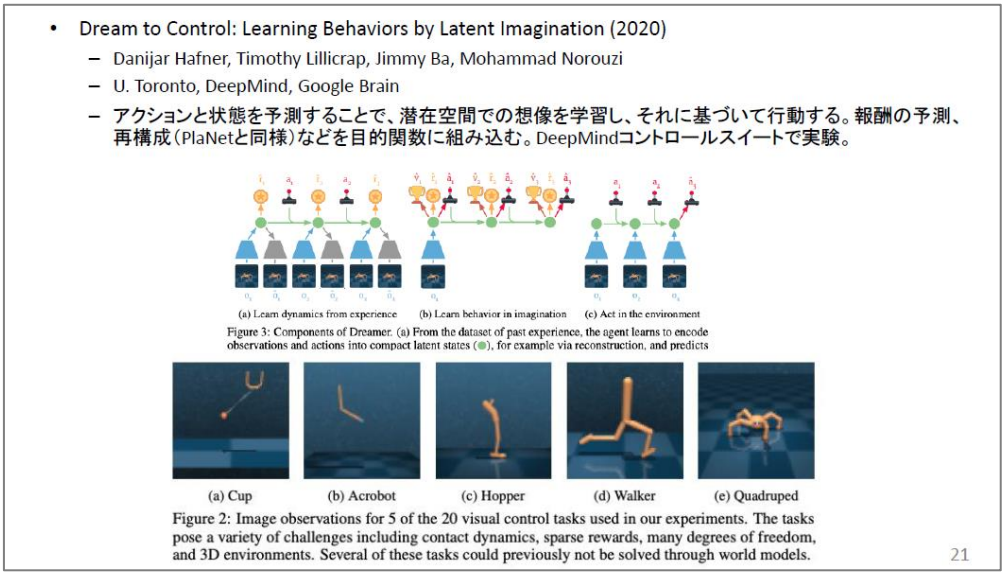


図 2-7　Dream to Control

世界モデルについて話してきたが、世界モデルは内部のシミュレーターと見ることもできる[27]。ただし、学習ベースである。というか、従来のシミュレーターという概念が限定的だったのではないか。従来のシミュレーターの概念は、基本原理や法則があって、それを回すというものだった。ところが、現実との誤差がある。そこで、現実との差分をとって、それを最小化する勾配を取ればよいので、微分可能なシミュレーターと考えることもできる[28]。また、アレン人工知能研

[27] シミュレータとの関係は、企画セッション当日は時間の関係でスキップされた話題だが、本報告書では簡単に触れておく。
[28] Yuanming Hu, Luke Anderson, Tzu-Mao Li, Qi Sun, Nathan Carr, Jonathan Ragan-Kelley, Frédo Durand (2019), "DiffTaichi: Differentiable Programming for Physical Simulation".

究所（Allen Institute for AI）の 3D シミュレーター環境 AI-THOR2[29]も参考になる。

## ２．３　２階部分（言語アプリ）の研究動向

次に言語アプリ系の研究では、やはりいま話題なのは BERT[30]で、Transformer を使って、さまざまなタスクで SOTA（State of the Art）を出している。この BERT が話題になったのが 2018 年で、その後、GPT-2 が出てきた。GPT-2 は高精度のテキストを簡単に自動生成できて、開発陣が「あまりにも危険過ぎる」と危惧して論文公開が延期された[31]。その後、2019 年 11 月になってそのコードが公開された。

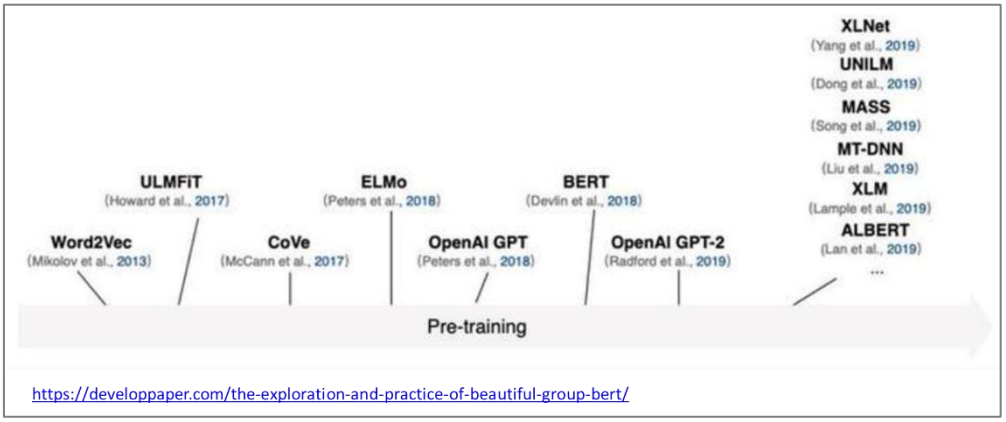


図 2-8　BERT 系の発展・派生

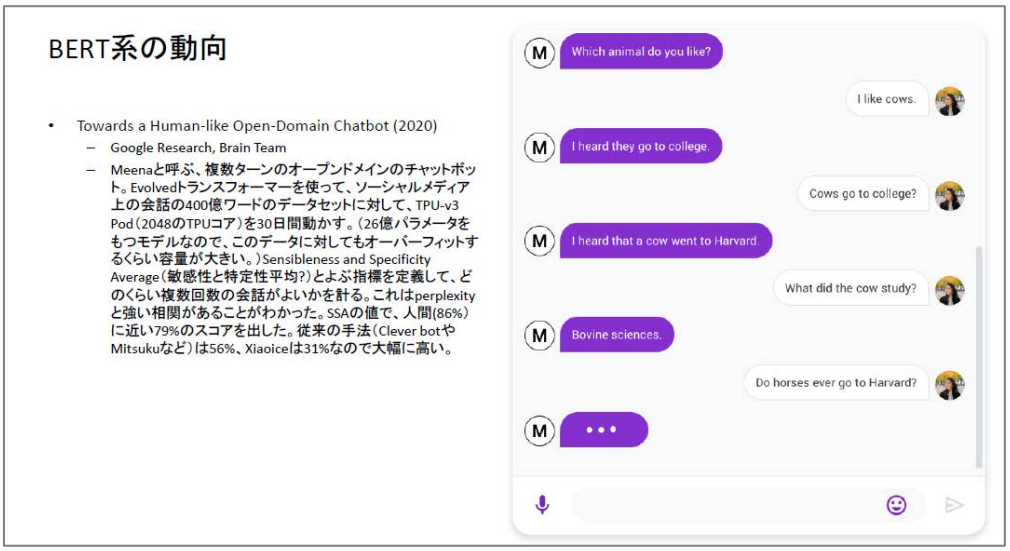


図 2-9　BERT 系の動向：Meena

この BERT 系の方式は、ALBERT とか、RoBERTa とか、XLNet とか、いろいろなものが次々

[29] https://ai2thor.allenai.org/
[30] https://arxiv.org/pdf/1810.04805.pdf
[31] https://gigazine.net/news/20191106-gpt-2-final-model-release/

に出ている（図 2-8）。また、Transformer を使った Google Research の対話システム Meena（図 2-9）は、従来手法や Xiaoice といったシステムと比べて精度が相当高く、人間並みのスコアを出している。それから、DeepL というドイツの会社が開発した機械翻訳システムも注目されていて、非常に精度が高い。たぶん BERT 系のモデルのとても巨大なものを使っているのではないかと思っているが、論文を読むのにこれ使うと全く問題なく読めるというぐらいレベルが上がっている。

## ２．４　１階部分と２階部分の統合に関する研究動向

1 階部分、2 階部分のそれぞれについて話してきたが、これらを統合するという研究もいろいろ取り組まれている。古いものだと 2015 年の Generating Images（図 2-10）[32]がある。画像からキャプションを作るというのとキャプションから画像を作るというのと両方向あるが、これはキャプションから画像が作れる。

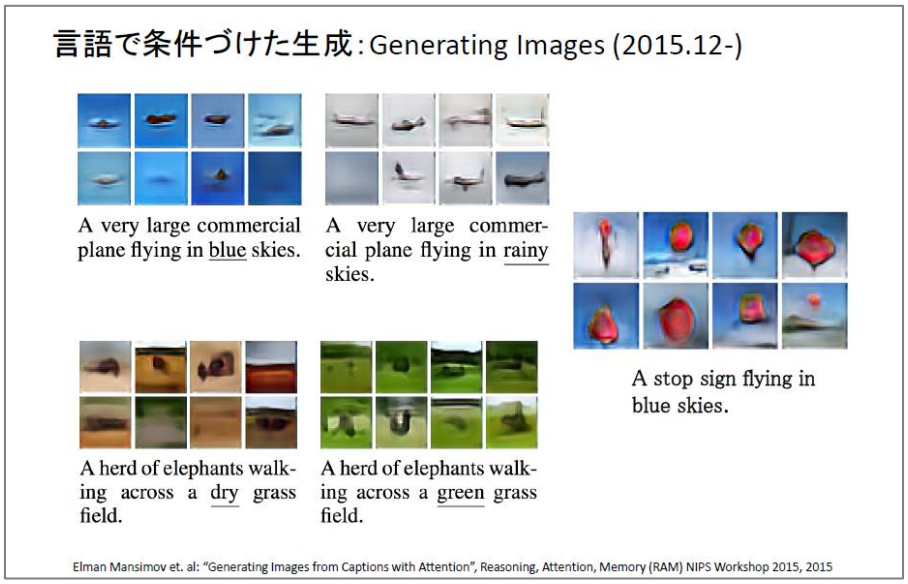


図 2-10　Generating Images

このあと、先ほどの BERT に画像を組み合わせた研究が ViLBERT（図 2-11）[33]、VL-BERT（図 2-12）[34]、VisualBERT（図 2-13）[35]などいろいろある。これらはみな似ていて、画像と文章をとにかく Transformer に突っ込むと、画像系のタスク、テキスト系のタスク、何でも精度が向上するという感じになっている。

[32] Elman Mansimov et. al (2015), "Generating Images from Captions with Attention", Reasoning, Attention, Memory (RAM) NIPS Workshop 2015.
[33] Jiasen Lu, Dhruv Batra, Devi Parikh, Stefan Lee (2019), "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks".
[34] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, Jifeng Dai (2020), "VL-BERT: Pre-training of Generic Visual-Linguistic Representations".
[35] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh & Kai-Wei Chang (2019), "VisualBERT: A Simple and Performant Baseline for Vision and Language".


図 2-11　ViLBERT

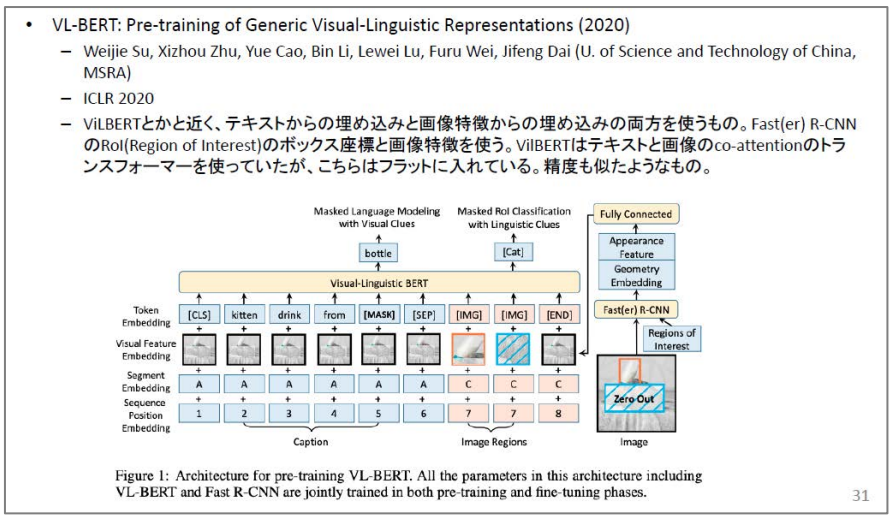


図 2-12　VL-BERT

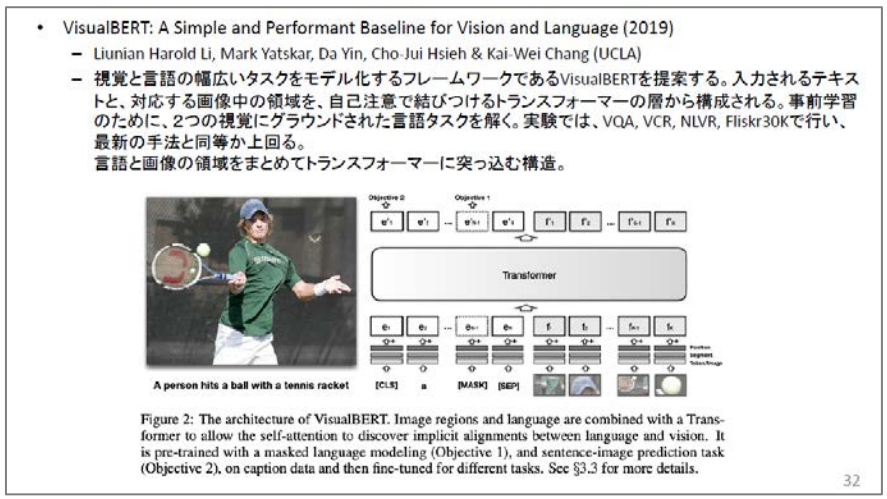


図 2-13　VisualBERT

## ２．５　統合のアーキテクチャ

ここから僕が主張したいことを話していきたい。

ここまで話してきたように 1 階部分の研究、2 階部分の研究、そして、それらを統合するような研究はあるが、それらはアプローチとして少し違うかなと思っている。つまり、画像と文章の両方を突っ込めばいいという話ではないと思う。やはり中心は世界モデル系、動物 OS と僕が言っている方であって、現実世界を知覚してロボットをどう動かすかというところがベースにあるべきだと思っている。そこにプラグイン的に BERT 系の言語アプリが加わっているという形だと思う。そもそも、1 階・2 階の 2 系統でタスクが異なり、その報酬系あるいは目的関数が異なるというのがポイントだと思っている。

かなり適当だが、全体アーキテクチャのイメージを描いてみた（図 2-14）。

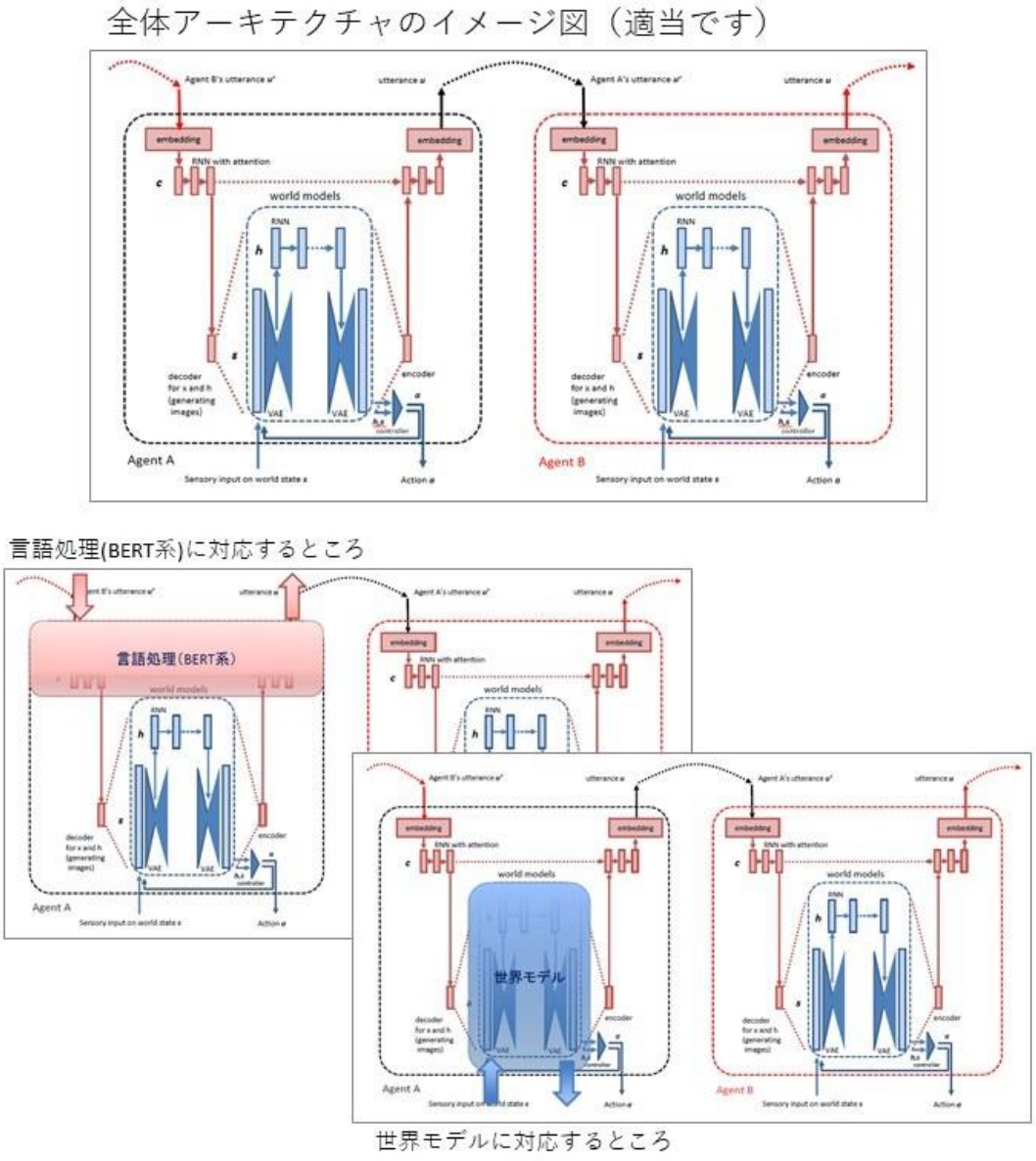


図 2-14　全体アーキテクチャのイメージ図

図 2-14 の上側には、Agent A と Agent B と箱が 2 つあるが、その 1 つ（Agent A）に着目すると、図 2-14 の下側に対応を示したように、Agent の真ん中の部分が世界モデルで、その上のあたりが言語処理（BERT 系）に相当する。世界モデルは、青矢印のように、何か観測データ（Sensory Input）が入ってきて、コントロール信号、アクションを外に出す。一方、言語処理は、赤矢印のように、他者の発話を聞いて返答する。そして、この返答は別の人（Agent B）への入力にもなるという関係になる。

このような 2 系統だと考えていて、言語処理側が世界モデル側を呼び出しているという関係になる。つまり、2 系統の間で潜在変数を共有している。むしろ世界モデル側の方が先にあって、それを言語処理側が使っているという関係と考えてよい。

基本的には、2 つの問題を自己教師あり学習的に解いているのだと思う。世界モデル・動物 OS の側は、次に何を知覚するかを予測している。言語処理・BERT 系の側は、他者が何を言うかを予測している。このような異なる問題を解いているのだと思う。

さらに言うと、図 2-15 のように、世界モデルの上に言語処理があるということ自体がまた世界モデルになって、その上に言語処理が積み重なるというようなメタな構造をしていると考えている。そう思っていたら、今朝の中島先生の基調講演でも、同じような層状構造を考えられているという話があった。

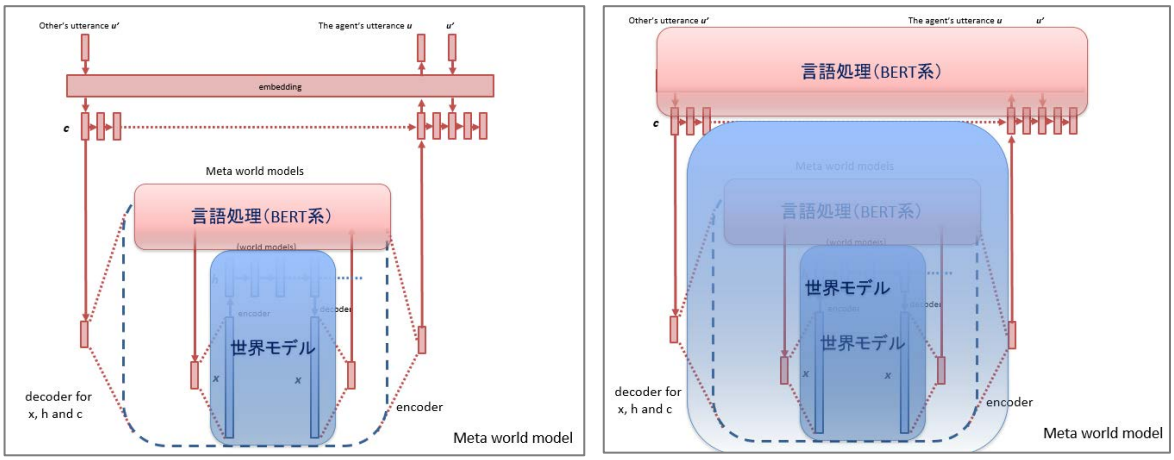


図 2-15　世界モデルと言語処理のメタ構造

## ２．６　人間の進化過程での報酬系の分岐

2 系統で報酬系が異なるようになったというのは、人間の進化過程に関わるものだと考えている。そもそも人間が進化の過程でどうしてこのような言語を獲得したのか、ということは非常に興味深い問題だと思う。脳の構造は、進化において急激に大きく変わるというのはないと思うので、サルやオランウータンやチンパンジーと大きくは変わらないはず。それらと人間の脳の構造でほんの少しだけ違うという部分がどこかというと、報酬系の項の追加と離散化という 2 点だと思っている。もともとの仕組みとして世界モデルとモデルベース強化学習があったところに、報酬系の項の追加と離散化が起こり、その結果として Distillation が起こったというのが、言語獲得に関わる僕の仮説である。Distillation についてはまた後で説明する。

そもそもどうして人間が言語を持つに至ったのかは難しい問題で、いろいろな説がある。進化生物学者の長谷川眞理子先生との対談で、人間だけが毛のないサルだという話があり、たぶん母子のやり取りがきっかけではないかと思うようになった。他のサルは赤ちゃんがお母さんにつかまっていられるので、お母さんは楽なのだけれど、人間は毛がないので、赤ちゃんがつかまれない。そのため、赤ちゃんを置いておかないといけないし、置いたままちょっと離れないといけないこともある。だから、赤ちゃんの状態を知ることがとても重要で、今離れていいときなのかどうかを知らなければいけないということから、母子のやり取りがとても重要になった。ところで、なぜ人間の毛がなくなったかというと、熱の効率というか、遠くに歩いていくために必要だったという説がある。

また、言語や歌の研究をされている東大の岡ノ谷一夫先生によると、言語を習得できるための必須条件の 1 つが息を止められることだとされる。他の霊長類は息を止められない。人間も例えば心臓は随意的に止められないわけで、同じように息も止められなくても問題ないはずだが、息は随意的に止められる。これはなぜなのかというと、赤ちゃんが泣くためだと考えられる。あんなに大きな声で泣く動物は他にいない。大声で泣くと外敵に狙われる危険性も高くなるはずだが、それでもお母さんにしがみついていられないということから、自分がどうして欲しいかといったことを知らせる必要がある。かまってもらう必要がある。赤ちゃんが生まれた瞬間はお母さんにも何を言っているのか分からないけれど、1 か月ぐらいすると、ミルクが欲しいとか、おしめとか、だんだん分かるようになってきて、母子のコミュニケーションが成立し始める。

いずれにしても、母子のやり取りというものがきっかけで、発話ができるようになってきたという見方ができる。

## ２．７　社会的蒸留と離散化

もう 1 つ重要な概念は Distillation（蒸留）である。これは Geoffrey Hinton が 2015 年に提案したもので、大きいニューラルネットワークの学習結果を小さいニューラルネットが学習する仕組みである。つまり、大きいネットワークの入出力のデータを小さいネットワークの教師データとする。そうすると、小さいネットワーク（生徒ネットワーク）の方が容量が小さく、パラメーター数が少ないにも関わらず精度が向上することがある。

Distillation も最近いろいろ研究されていて[36] 、教師の容量が大き過ぎると生徒がうまく学習できないとか、教師が持っている途中のステップを教師データにすると生徒の学習が進むといった研究が発表されている。Distillation によって、要するにどんなことが起こっているのかというと、ある人が例えば世界モデルを獲得したとか、ある分野で熟練の域に達したとかいったときに、それを言葉にすると、それがラベル、教師データとなって、次の人はより効率的に学習することができるということなのだと思う。しかも、効率的に学習できるので少ない容量で済み、余った容量を使ってもっと先まで学習できる。これが世代を超えてずっと行われていくと、脳という一定容量の装置の中にどんどん知識がうまく詰め込まれるようになってくる。こういうことが起こっているのではないかと思っていて、僕はこれを「社会的蒸留」と呼んでいる。

[36] 例えば、Jang Hyun Cho and Bharath Hariharan (2019), “On the Efficacy of Knowledge Distillation”.
Iou-Jen Liu, Jian Peng, Alexander G. Schwing (2019), “Knowledge Flow: Improve Upon Your Teachers”.
Yuandong Tian, Tina Jiang, Qucheng Gong, Ari Morcos (2019), “Luck Matters: Understanding Training Dynamics of Deep ReLU Networks”.

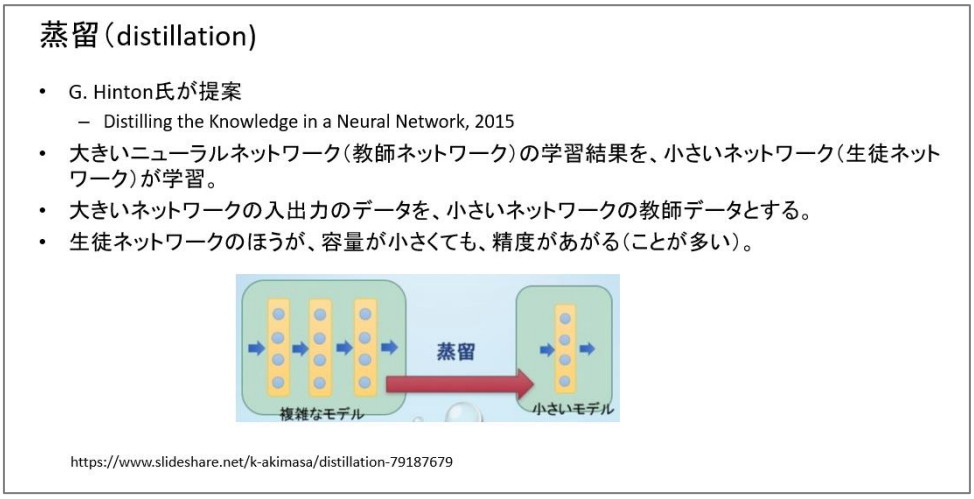


図 2-16　蒸留（Distillation）



これを先ほどの母子コミュニケーションの話と併せると、そのように母子間でコミュニケーションすることが、結局 Distillation の効果をもたらすようになってきたということだと思う。さらに、おそらくそのどこかの段階で離散化することが必要になってきたのではないか。赤ちゃんの状態を知るだけだったら、赤ちゃんの声のピッチやトーンといったものだけでよいはずだが、Distillation が利き始めるためには大人も子供も男女も同じ言葉を話さないといけないので、音の絶対的な高低ではなく相対的な高低とか、人によって異なる口や喉の形によらないシンボル化の方法とかが必要になってきて、たぶんどこかの時点で連続的なものの予測から離散的なものの予測にスイッチしたのだろうと思っている。

Hinton の Distillation 研究では、離散化したラベルよりも、ラベルにする前の連続値を教師データにした方がよいということなのだが、言葉も離散的な音素から構成されているので、たぶんどこかの時点でこういう変化が起こったのだと考えている。それで結局、離散化された音を当てると嬉しいという項が報酬関数に足されたのだと思う。つまり、世界モデルを用いて知覚の予測精度を上げる（1 階部分の予測エラーPerception Prediction Error を下げる）という話と、言語モデルを用いて他者の発話（離散化されたもの）の予測精度を上げる（2 階部分の予測エラーUtterance Prediction Error を下げる）という話が足し合わされたのだと思う。これがゆえに、人は他者の話すことに興味を持ったり、物語が好きだったり、音楽というものができたりといったことにつながったのではないかと思う。

ということで、まとめると、世界モデルがベース（動物 OS）で、そこに BERT 系言語アプリがアクセスをしてくるという形である。動物 OS は知覚予測が目的関数（報酬系）で、敵を察知したりとか、餌を探したりとか、痛いことが嫌だとか、動物的な起源のものである。一方、言語アプリはおそらく発話予測が目的関数（報酬系）、母子の関係が起源で、相手の状態を知りたいということから離散化されて言葉になった。そして、そこから副次的に社会的蒸留という効果が起こって、ある人の発話が他の人の学習に寄与するようになり、さらにそれがメタに重畳的に起こっていったということだと思う。

## ２．８　脳の仕組みとの対応

新学術領域研究に「人工知能と脳科学」という領域[37]があり、僕も参加している。銅谷賢治先生が代表をされていて、次に発表する谷口先生も参加している。そこでの貢献に向けて、先ほどの話を脳科学の文脈で考えてみた。無理矢理考えた個人メモなので、脳科学の専門家から見るとおかしなことを言っているかもしれないところはご容赦いただきたい。

報酬関数が 2 つあるというのは脳においてどう実現されているかを考えている。

脳の後ろの方にある黒質からドーパミンが線条体や側坐核へ行っている。また前頭皮質にも行っていて、この前頭皮質に行く部分がたぶん鍵ではないか。人間に特徴的なのは前頭前野の 10 野や 9 野や 32 野と呼ばれるところで、ここに先ほど話したような人の発話を予測したいという報酬系ができているのではないかと考えている。なぜかというと、そのあたりは心の理論や社会性に関わるところで、リスクと意思決定や報酬・葛藤といったものに関係していて、そこにつなげることで実現されるのではないか。ここに対する報酬系が機能してないと、相手・他者に対して興味を持たなくなり、そうすると、Distillation が利かなくなって、内的な報酬に基づく行動を過度に好むようになってしまい、自閉症的な症状が出るのかもしれない。これは病気の話なので、あまり軽々しく言ってはいけない点だが。

それから、ウェルニッケ野やブローカ野もとても興味深い。たぶんウェルニッケ野が世界モデルを担当しているのではないか。周りに 1 次視覚野、2 次視覚野、体性感覚連合野など、いろいろあるが、センサー系とアクチュエータ系だけで、潜在変数に対応するのがたぶん体性感覚連合野なのではないかと思っている。ウェルニッケ失語というのがあって、発話のメロディも抑揚も保たれているが支離滅裂になる。相手の話が理解できなくて、読んでも理解できない。言葉を真似することもできない。真似することができないというのは、たぶん意味が分かっていないから真似できないということだと思う。流暢性失語やジャーゴン失語とも呼ばれる。つまり、世界モデルが壊れていると、言葉としては滑らかに発音できても理解できないのだと思う。

一方、ブローカ野の方が言語系を担当していて、文法的に複雑な文章を作り出す。こちらはブローカ失語というのがある。この場合は電文体の発話になる。電文体というのは文法がなくて、内容語のみが並ぶ。ところが、言語理解は正常に行われている。なので、言語系の部分が壊れているが世界モデルの部分は正常に動いているということなのだと思う。

それから、聴覚の信号が 2 系統に行かないといけないと思うのだが、これには背側と腹側の経路がある。背側が知覚情報・センサー情報の一種として音声の経路で、腹側は発話予測をしたり報酬系に関わる部分に回る音声の経路なのではないか。

現状では大部分が推測だが、こんなふうに考えると、脳の仕組みともおおまかに対応づけることができるのではないかと思っている。

## ２．９　何がしたいか？

最後に「で、何がしたいの？」ということを話したい。

1 つは、産業的には意味理解がしたい（図 2-17）。意味理解ができると極めて大きな産業的な価値に結びつくし、社会変化をもたらすはず。つまり、言っていることの意味をコンピューター

[37] http://www.brain-ai.jp/jp/

が本当に理解できるということなので、コンピューターはいろいろな仕事ができるようになり、ホワイトカラーの生産性が極端に向上することになると思う。

もう１つ、学術的には、やはり知能の仕組みを解明したい。特に構成論的に解明したい。しかし、何か言えれば言えたことになるのか、先ほど僕が適当に話したこと以上に言うにはどうしたらいいのかは結構難しい。

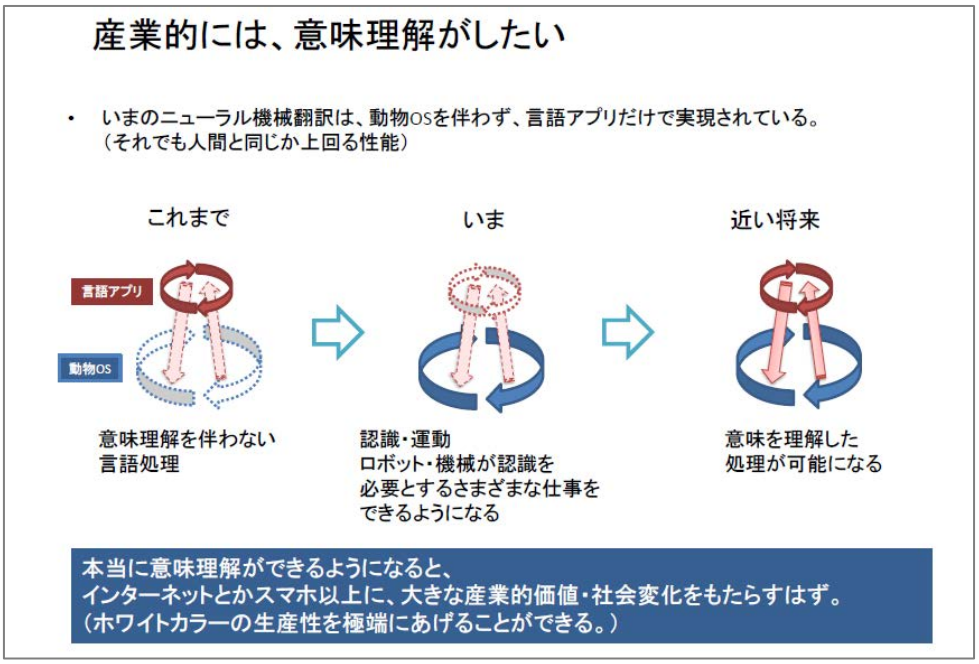


図 2-17　産業的には意味理解がしたい

2 つの視点があって、1 つは、言語のそもそもの成り立ち、存在理由を明らかにしようというものである。マルチエージェント的なものを作って、そこに何か言語を発生させるというようなことができると思うが、たぶん人間と同じ言語は絶対生まれない。言語的なものは生まれるし、いろいろな発見はあるだろうが、センサー・アクチュエーターや生存環境に根差した報酬関数が違いすぎて、ある設定でこうなったという話になる。これは「ヤッコー研究」、つまり、やってみたらこうなりました的な研究になりそう。

もう１つは、言語の成り立ちや存在理由はまあいいとして、より実際的に言語理解のできる AI を作ろうというものである。そのためのデータセットとして、膨大なテキストデータ、YouTube 動画、3D シミュレーションと言語データ、実ロボットなど、いろいろ使い倒して言語理解のできる AI を作る。先ほどの産業的な話と同じだが。

このような 2 つの方向でいろいろ考えたところ、目指すのはこういう方向（図 2-18）かなと思う。「世界モデル＋BERT＋Distillation＋NAS＋マルチエージェント＋現実データとの差異の最小化」というのが答えかなと。

ここで「世界モデル＋BERT」というのは今日話してきた 2 階建てモデルのことで、そこに Distillation が加わって、教師ネットワークの学習したことを生徒ネットワークに離散化して伝える。教師ネットワークが例えば世界モデルを学習しましたというときに、それをどのように伝えると一番よく伝わるのかという方策を教師側は学習する。一方、生徒側はこれはどんなハイパーパラメーターあるいはアーキテクチャにしておけば教師のモデルを一番よく学習できるか、とい

ったことをやるようになる。

これは結局ハイパーパラメーターの探索なので、NAS の技術ということになる。NAS というのは Neural Architecture Search、ニューラルネットワークのパタメーターを探索する手法である。教師側も生徒側も同時に学習・探索していく必要があって、これをマルチエージェント的に動かす必要がある。あるエージェントが世界モデルを学習し、それを次に伝え、それをまた次に伝えるというのを繰り返しながらよいパラメーターを探していく。

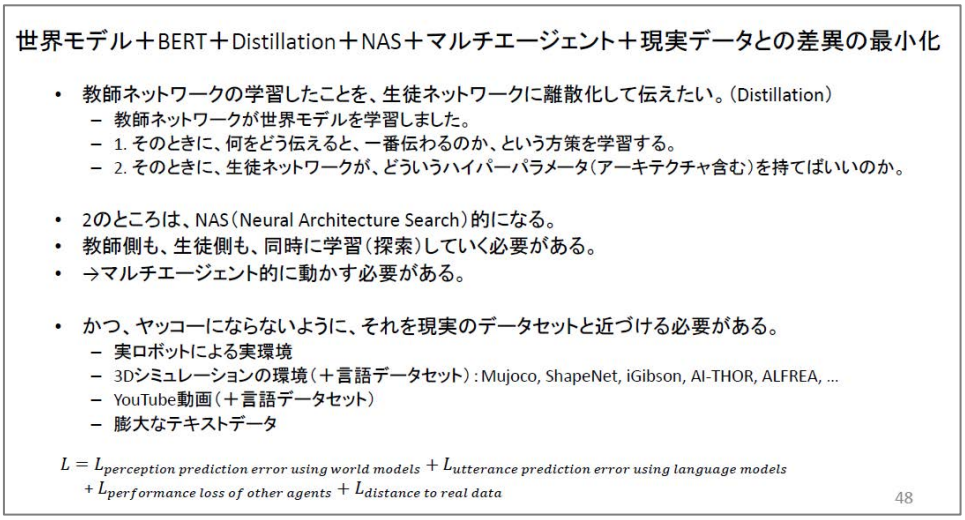


図 2-18　目指す方向

それから、「ヤッコー研究」にならないために、現実のデータと近づける必要がある。実ロボットによる実環境とか、3D シミュレーションの環境とか、YouTube 動画とか、膨大なテキストデータを使って、要するに人間界のデータと近づけることをやっていくべき。この近づけるというのは、実は深層学習の文脈でよくやられることで、ロス関数に項を 1 個足せばだいたい近づく。そういう意味では、図 2-18 の一番下に書いたロス関数 L は、Perception Prediction Error という世界モデル的な知覚の誤差と、Utterance Prediction Error という発話の予測の間違いと、Performance Loss of Other Agents という他エージェントにどれだけロスさせないか、他エージェントをどれだけ手伝ってあげられるかという項と、Distance to Real Data という実データに対してどれだけ距離があるかという項を組み合わせたものになっていく、といった方向に進んでいくのではないか思っている。

# ３．招待講演「確率的深層生成モデルによる実世界自律知能の創成に向けて ～記号創発ロボティクスが生み出す認知アーキテクチャ～」

◆谷口 忠大（立命館大学情報理工学部 教授）[38]

## ３．１　記号創発ロボティクス

今日は記号創発ロボティクスの話を、2020 年バージョンということで話したい。セッション冒頭の趣旨説明でも記号創発ロボティクスについて触れられ、先ほどの松尾先生の講演も最後のあたりでマルチエージェント系での言語創発や意味理解の話になったので、非常によいつながりになったように思う。


図 3-1　記号創発ロボティクス

記号創発ロボティクスは、浅田稔先生らが進めてこられた認知発達ロボティクスの研究思想をかなり継承している。人間の子供は自らの身体的経験、感覚的運動情報の統合を通じて、いろいろな機能を獲得し、言語を獲得し、コミュニケーションをも可能にしている。結局、現状の AI では音声認識装置も画像認識装置も大規模なデータセットを構築・使用して教師あり学習を行っているが、人間の知能は自らの五感のみ（体性感覚とかも含むが）でどんどん知能を形成・獲得している。さらに言語を獲得し、社会的なインタラクションもして発達していく。ならば、やはりグランドチャレンジは、そのように発達していく知能を作ることだと思う。

私自身はコミュニケーションや意味理解といったものに特にこだわりがあり、発達が言語の獲得にまで至るのを、2010 年頃から記号創発システムとして捉えて提案してきた。同時に記号創発ロボティクスを記号創発システムへの構成論的アプローチと位置づけて提案した（図 3-1）。2014 年には講談社から「記号創発ロボティクス」と題した書籍を出した。また、この流れの最新刊「心を知るための人工知能」がちょうど本日、共立出版から出る。本日の講演内容もこの最新刊にかなり含まれている。

[38] http://www.ritsumei.ac.jp/ise/teacher/detail/?id=176

記号創発ロボティクスは、構成論的アプローチということで、実世界経験に基づく言語獲得ロボットの実現と、それを通した人間の認知発達の理解を目指している。つまり、触覚情報、聴覚情報、視覚情報などのマルチモーダル情報を統合して、すべてボトムアップに行う学習をベースにして、物体の概念や場所の概念、そして語彙を獲得し、人間とのコミュニケーション、ロボットの知能の構成を可能にしていこうとしている。こういうことをやり始めた 2010 年頃と比べると、最近はマルチモーダル VAE など、マルチモーダル情報を扱うこともかなりスタンダードになってきたので、ある意味、我々にとっていい時代になってきたと感じている。

## ３．２　記号創発システム

記号創発ロボティクスのバックにある思想・フィロソフィカルな概念として記号創発システム（図 3-2）がある。我々はこれをベースに取り組んでいるのだが、まだまだ全体を把握・実現するまでには至っていない。

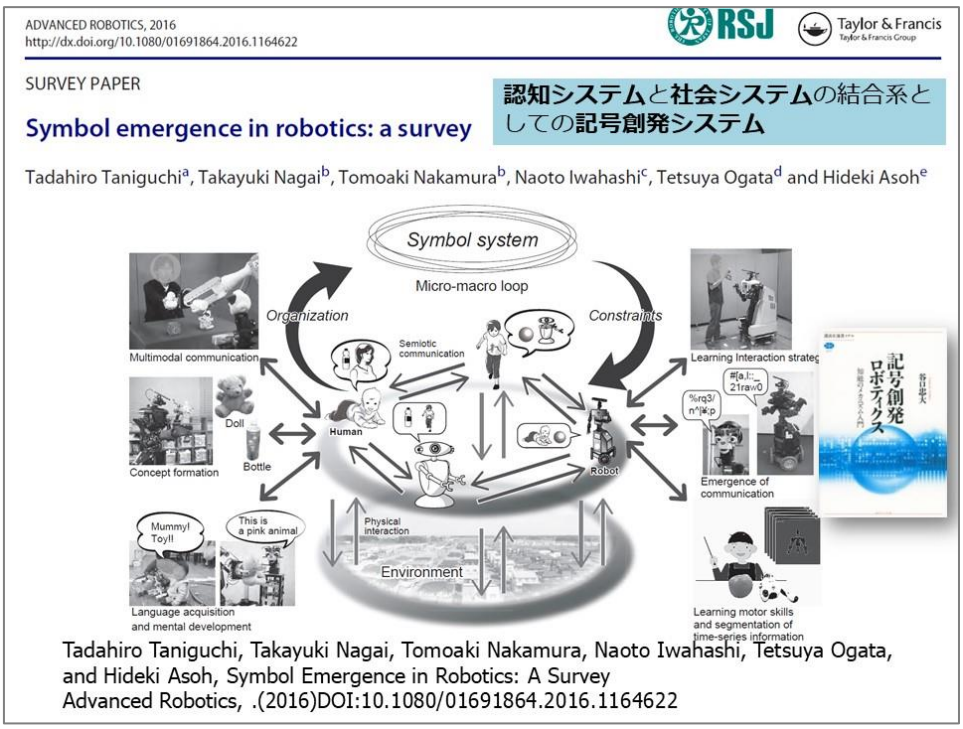


図 3-2　記号創発システム

記号創発システムはロボットの話ではなく、人間の言語の獲得や意味理解を、創発システム、複雑系の視点から捉えたものになっている。松尾先生が先ほどの講演の最後のあたりで話したマルチエージェント系の中で言語ができてくるというのもこれと同じ。

図 3-2 の真ん中に人間、赤ちゃんがいる。これが環境とのインタラクション、Sensorimotor Interaction を通して、さまざまな物体概念を形成することができる。例えば、物を触って「あ、ボール」みたいな概念を得るということである。実はこれは今の言葉遣いで言えば、表現学習（Representation Learning）にも相当し、概念形成やカテゴリー形成ということもある。人間はこうやって言語の意味の基礎を作っていくことになる。

記号論的に考えると、どういうものをひとまとまりにするかは恣意性があり、基本的には、形成された概念は一人一人で異なり得て、人々の中で同一であることは保証されない。しかし、その概念を用いてコミュニケーションをしようとすると、共通性がないと困る。例えば「ペットボトル取ってきて」と言うときには、やはり相手がペットボトルをどの物体でどのようなものなのかを分かってくれなくてはいけない。そこで、社会の中で共通化が図られて、概念に対するラベルづけは社会的に保持される。でも、ラベルづけは恣意性があるものなので、例えば日本語で「リンゴ」とラベルづけされたものを、英語圏では「Apple」というラベルをつけるというように、同じ概念に対して違うラベルを用いることができる。このような違いは、国の違いや、日本語と英語とかだけではなく、家庭とオフィスとか、学問分野とかによっても異なるものになり得る。

意味理解というのは、ある言葉の意味が一意に決まるという単純な世界ではなく、人間が環境や他者とインタラクションするダイナミクスの中で創発的にできあがるものである、このようなダイナミクスを捉えて、そのダイナミクスに適応していける AI システムを作らないと、変化していく意味についていけないということになる。なので、ルールベースで作ったり、あるデータセットで作ったりすると、文脈や環境・状況が少し変わっただけで、その言葉の意味についていけなくなる。

そこで、表現学習により概念形成がされて、それに基づきながらインタラクションする中で、言葉の意味が、社会的な言葉の使い方としてある程度決まってきて、徐々に社会の中で共有される記号システムというものができあがってくる。これが記号創発システムである。その社会の中で共有された言葉の意味を制約・ルールとして従う限りにおいて、我々はコミュニケーションができることになる。他者と意味を通じ合うことができる。このような記号創発システムの世界観の下で研究をしている。



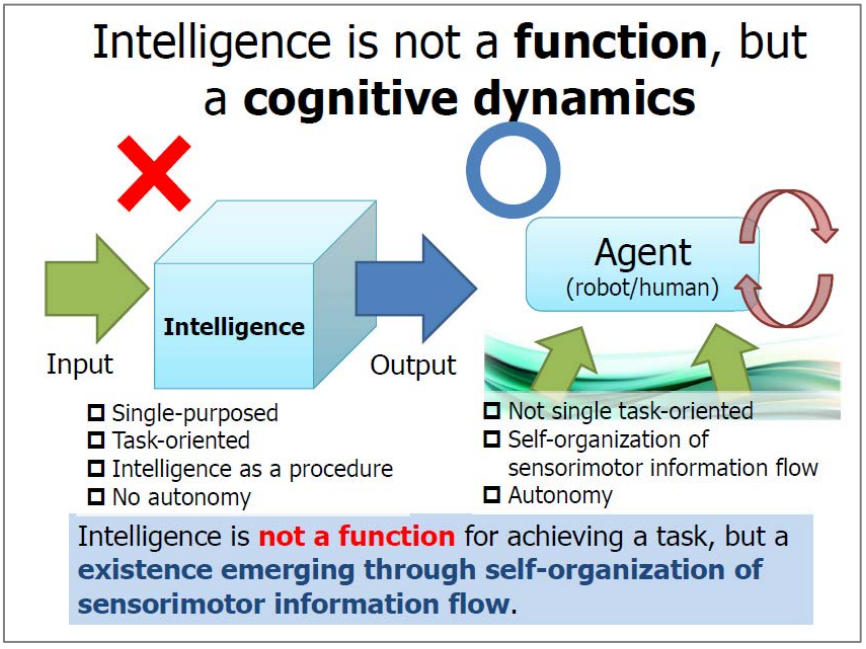


図 3-3　知能：認知的ダイナミクス

## ３．３　知能と認知アーキテクチャ

ここで、そもそも知能とは何かについて話しておきたい。知能は認知的なダイナミクスだと理解することが重要である（図 3-3）。つまり、環境との相互作用に基づいて、どんどん変化していくもので、我々自身も認知発達していく。インプットに対してアウトプットを出す機能モジュールだと単純に捉えていてはいけない。

2010 年代そして現在も、AI 研究は機能モジュールを取り出して、それをトレーニングしようという研究が主流になっている。これは単目的 Task-Oriented で、知能はインプットに対して正解のアウトプットを出すもの、つまり「Intelligence as a procedure」と考えられている。例えば、盤面に対して最適な手を打ったりとか、画像入力に対してキャプション文を出したりとか、テキスト文に対して音声データを出力したりとかいったものになりがちである。

一方、赤ちゃんや子供を見てみると、知能は何か単一のタスクをやるために動いているわけではないし、何か正解ラベルが与えられて、それに向かって学習しているわけでもない。これを学習しろというのを受けて、それに向けてチューニングされてきたわけでもない。

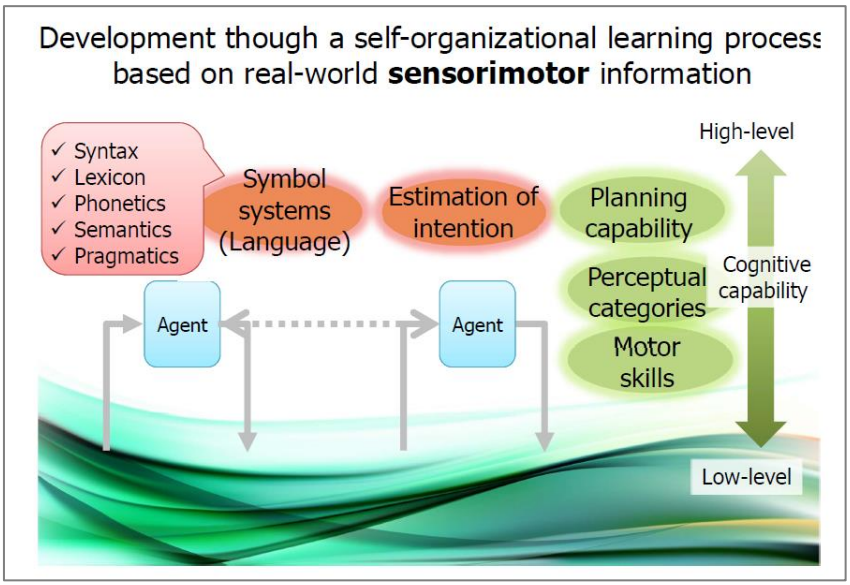


図 3-4　感覚運動情報に基づく自己組織的学習過程を通した発達（1）

知能はもっと自律性がある。知能を持つものは、我々人間だろうが、エージェントだろうが、ロボットだろうが、感覚運動情報を通して環境とインタラクションしながら、自律的な適応の中で運動スキルを手に入れ、知覚的なカテゴリーを形成したり、プランニングができるようになったりする（図 3-4）。さらに他者がいて、この他者が何かサインを出してきてインタラクションしていくと、この他者と協調しようと、他者の意図推定のスキルを得ていったり、記号システムや言語の獲得をしたりもする。言語にも文法・語彙などいろいろあるけれど、そういったものの認知モジュールを環境や他者とのインタラクションの中でボトムアップに構成していける。

これは凄いので作りたい。しかし、これを教師なしで自己組織的に学習するだけでも大変なことなのに、図 3-5 のように一個一個の認知モジュールの学習が相互依存しているのでさらに問題が難しい。例えば、物体概念の学習と、それに対する音素列や単語の学習というのは、片方の知識があればもう片方も結構簡単に学習できるが、片方の学習ができていなかったらもう片方の学

習も難しいというように相互依存していると言われる。また、言語の文法の学習とアクションのプランニングが相互依存している、Syntactic Bootstrapping と言って文法の学習が行動文法の学習を助ける、また、行動の学習が言語の学習を助けるといった話もある。

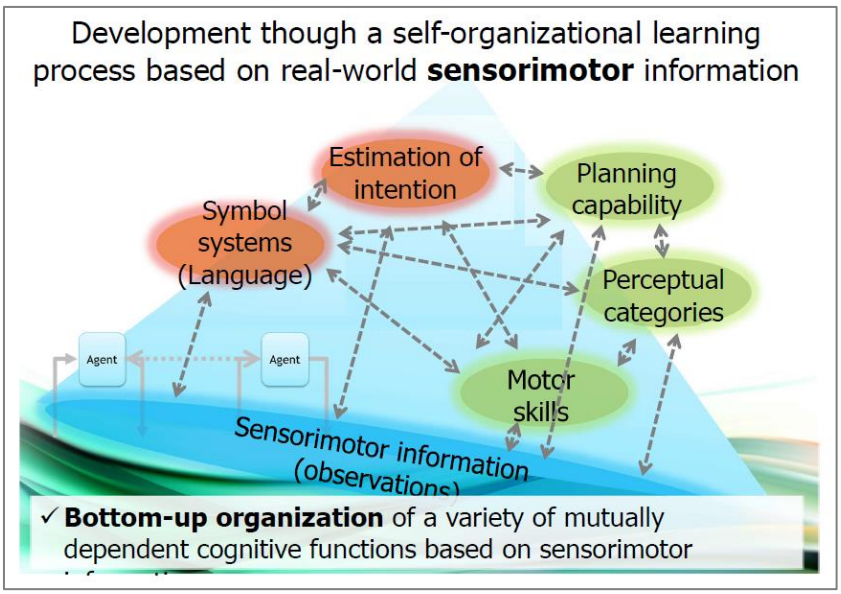


図 3-5　感覚運動情報に基づく自己組織的学習過程を通した発達（2）

今日の講演タイトルにも「認知アーキテクチャ」という言葉を入れているが、我々の記号創発ロボティクスの研究でも、このように相互依存しているさまざまなモダリティー、認知モジュールを統合的に構築していかねばならないと考えている。言語理解も、例えば「キッチンからペットボトルを取ってきて」という単純な発話を理解するためだけでも、ペットボトルとは何かという物体概念、キッチンとはどこかという場所概念、持ってくるという行為が何なのかという行為概念などが理解できていないといけない。そういう認知アーキテクチャが必要になる。

## ３．４　記号および記号接地問題についての誤解

ここで少し話がそれるが、記号創発システム論のコアなところに関係するので、「記号」という言葉が 2 種類の意味で使われているということについて話しておきたい。AI 研究者もこの 2 つを結構ナイーブに混同しがちで、分野やコンテキストによって、どういう意味で使われているか、気をつける必要がある。今日はシステム 1・システム 2 の話が何度か出ているが、そこでの意味・使われ方もやや気になるところがある。記号という言葉の 2 つの意味に重なりはあるし、この話は物議を醸すところもあるのだが、あえて僕の見方を話したい。

まず、1 つめの意味として、シンボリック AI の記号処理に由来する記号・シンボルという言葉の使い方がある（図 3-6 の左側）。これはシンボリック AI からさらに遡ると、シンボリックロジック、記号論理学、数理論理学、述語論理などにおける記号、プログラミング言語におけるトークンといったものが「記号」と呼ばれている。Stevan Harnad が 1990 年に「記号接地問題」（Symbol Grounding Problem）を定義したとき主に問題にしているのは、僕の理解では、この 1 つめの意味の方である。知能を記号処理で作ったときや、先ほど挙げたようなシンボリックシステムを作

ったとき、それを実世界情報にグラウンドさせなければならないことになる。ただ、実はシンボリック AI のアプローチを取らないのであれば、そのようなグラウンドをするという問題ではなくなる。

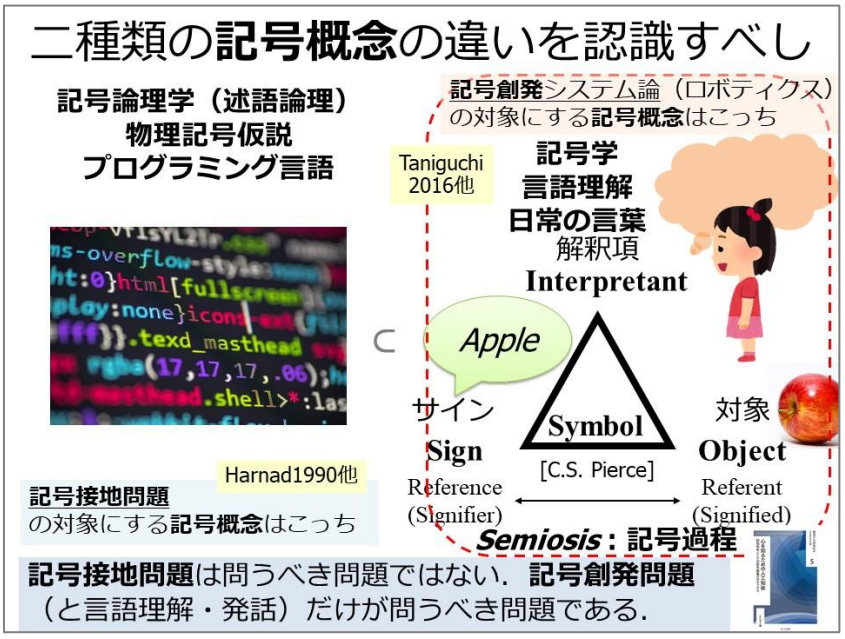


図 3-6　2 種類の記号概念の違い

一方、記号のもう 1 つの側面は、我々が日常で使う記号である（図 3-6 の右側）。記号学や言語学で扱っている言語もこちらの意味での記号であり、我々がコミュニケーションのために使っている記号、つまり、言語やサイン、手旗信号などさまざまなシグナル類もこのタイプになる。

記号という言葉は 2 種類の意味（図 3-6 の左側と右側）で使われているという話をしたが、記号創発システム論、記号創発ロボティクスにおける記号は、図 3-6 の右側に該当する。意味理解をしなければならないというのも、同じく右側の話になる。

我々の脳の中に、人間の知能の表現として、いわゆるシンボリック AI で言うところの記号表現が存在するかについては、存在すると主張をする人もいるとは思うが、僕が認識している限り、あまりしっかりとした証拠はないと思う。言語処理として BERT の話が今日も出ているが、それも右側の話だと思う。つまり、言語をパターン処理できているということであって、いわゆる記号論理を用いたプランニングをしているのではない。

先ほど触れた記号接地問題を、ここで改めて取り上げると、これは問題の設定が適切ではなく、基本的には問うべき問題ではないと僕は思っている。記号接地問題ではなく、記号創発問題を問うべきなのである[39]。先ほど記号創発システムについて説明したように、我々が使ってる言語は、日々の状況の中で言葉の意味は変わりながらも、我々はそれに適応して理解していくことができる。そのように適応・追従できるような知能システムを作るということが、記号創発問題を解くということになると僕は理解している。結局、記号創発問題というのは、そういう言語理解ができるかという問題であって、記号接地問題という言い方をすると、問題を取り違えて混乱する。

このようなことを 10 年くらい前から考えてきて、記号創発について国際的な議論もしてきた。

[39] 講演の冒頭でも触れた 2 冊の著書「記号創発ロボティクス：知能のメカニズム入門」（講談社、2014 年）と「心を知るための人工知能：認知科学としての記号創発ロボティクス」（共立出版、2020 年）の中で詳しく書いている。

欧州とかの 10 人くらいの研究者と共同で論文も書いている[40]。

## ３．５　場所概念・語彙獲得のモデル

次は、記号創発ロボティクスで、どのようなモデルを作ってきているかについて話したい。

実現したいことは何かというと、インタラクションを通して、ロボットの感覚運動情報や経験から概念、まずは物体概念や場所概念を獲得できるようにしたい。さらに言語も獲得できるようにしたい。これは、設計者が与えた記号表現の類からスタートするのではなくて、あくまで音声で語りかけられたりするといったところから語彙を発見して知識を得られるようにしたい。

これが利用されるシーンを考えると、ロボットを実環境に導入するときに、現場に技術者を送ってロボットに追加学習等をさせたりしなくてよいことになる。現場でロボット自身が知識を獲得して、その環境の中で自然に追加学習をして適応できるようになる。そこの場所概念や語彙もその場でオンライン同時学習をして、その環境の記号系に自動的に適応していける。

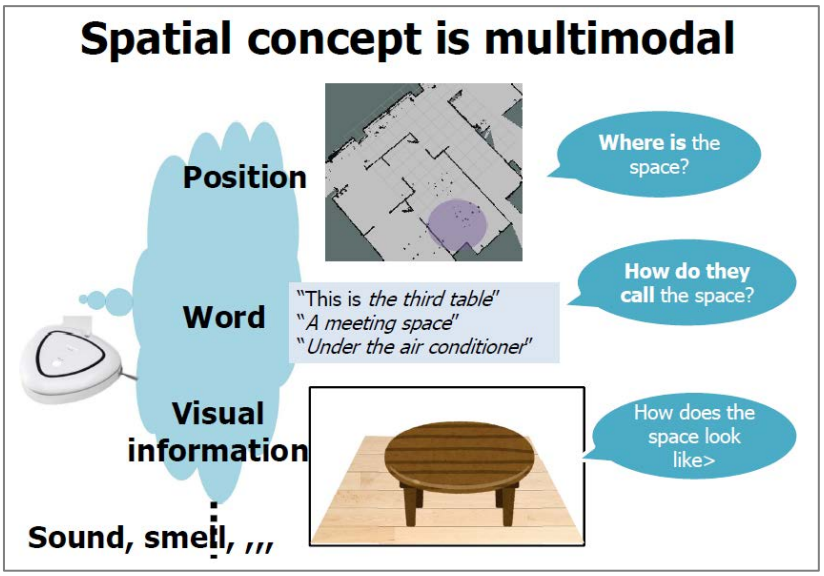


図 3-7　場所概念のマルチモーダル性

場所概念について言うと、現状、ロボットコンペティションとかでは、自己位置推定と置かれた環境の地図構築を同時に行う SLAM（Simultaneous Localization and Mapping）という技術で地図が作れるとしても、例えば、ここがキッチンの領域だといったことは教えねばならない。人間がそういう場所のラベルを設定・プログラミングすることになる。しかし、人間は初めて訪問した友人の家でも、どこがキッチンなのかを分かる。それはどうしているかというと、視覚情報からのシーン理解として、シンクがあって、フライパンがあって、冷蔵庫があって、オーブンがあってというようなことから、キッチンだと思うわけである。キッチンはたいていダイニングの隣にあるというような経験的な知識を使うこともある。また、その場所で誰かが料理をしてい

[40] Tadahiro Taniguchi, Emre Ugur, Matej Hoffmann, Lorenzo Jamone, Takayuki Nagai, Benjamin Rosman, Toshihiko Matsuka, Naoto Iwahashi, Erhan Oztop, Justus Piater, Florentin Wörgötter (2019), "Symbol Emergence in Cognitive Developmental Systems: A Survey", IEEE Transactions on Cognitive and Developmental Systems, Vol.11, No.4, pp. 494-516. doi: 10.1109/TCDS.2018.2867772.

たら、その動作を観測して、キッチンだと判断できるかもしれない。あるいは、そこを指さして「キッチンだよ」と言われて、その言語情報が助けになることもある。

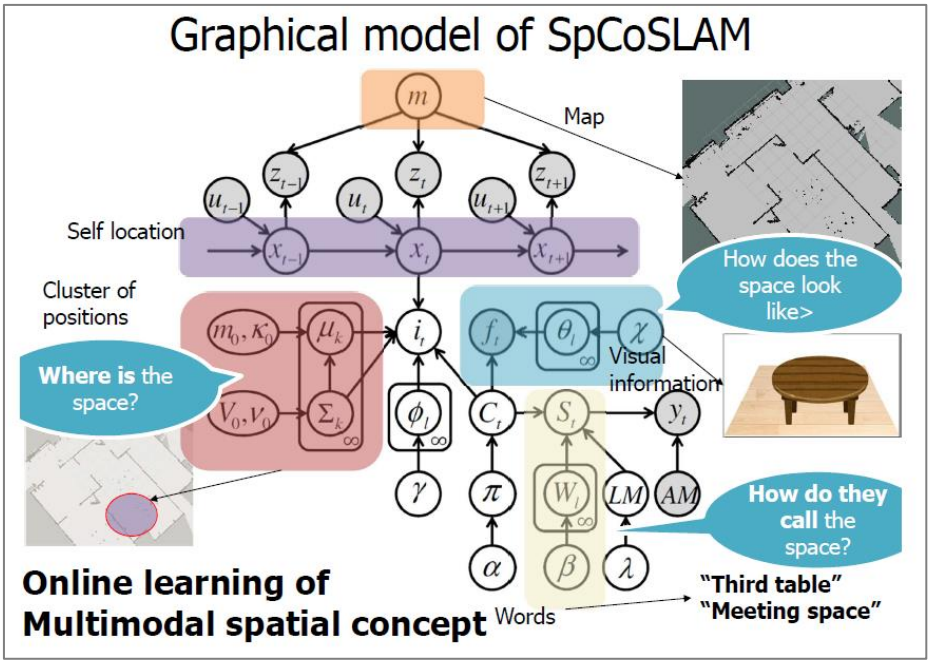


図 3-8　SpCoSLAM のグラフィカルモデル

我々が獲得する概念というものは、そういったものを統合したマルチモーダル情報の上に成り立っている（図 3-7）。そのような考えのもと、場所概念の獲得をマルチモーダル学習として捉えたのが SpCoSLAM（Online Spatial Concept and Lexical Acquisition with Simultaneous Lo-calization and Mapping）である[41]。SLAM を拡張して場所概念を獲得できるようにした。図 3-8 に示した例のように、いわゆる確率的生成モデルの形を取っている。この図の上の部分は部分観測マルコフ決定過程 POMDP（Partially Observable Markov Decision Process）で、SLAM の地図構築のためのモデル部分である。図の左の部分は、場所の領域をクラスタリングするための混合ガウスモデル、右中央の部分は、画像の特徴量を取ってくるところで、CNN（畳み込みニューラルネットワーク）の出力を与えている。右下の部分は音声認識モジュールである。言語モデル（LM）の部分は、最初は語彙を持っておらず、ユーザーとのインタラクションを通して音声情報から語彙自体を発見していく。

ここで、SpCoSLAM を用いた場所概念獲得のデモの様子を紹介する（図 3-9、動画は YouTube でも見ることができる[42]）。ロボットが廊下を移動しながら画像情報を得る。その情報をもとに地図を形成して位置を推定する。その間ときどき、人間が「廊下を歩いているよ」とか「プリンタールームに来たよ」とか話す。そのような言葉をどんどん聞いていると、「プリンタールーム」という言葉が複数回出てきたことなども手掛かりにして、単語を発見し、環境の情報と紐づける。こういったことを確率モデルの中でベイズ推論を使って行うことで場所概念のようなものを獲得し

[41] Akira Taniguchi, Yoshinobu Hagiwara, Tadahiro Taniguchi and Tetsunari Inamura (2017), "Online Spatial Concept and Lexical Acquisition with Simultaneous Localization and Mapping", IEEE IROS 2017.
Akira Taniguchi, Yoshinobu Hagiwara, Tadahiro Taniguchi, Tetsunari Inamura (2020), "Improved and scalable online learning of spatial concepts and language models with mapping", Autonomous Robots 2020. DOI: 10.1007/s10514-020-09905-0
[42] https://youtu.be/hVKQCdbRQVM　（デモの様子は 0:31～）

ている。「じゃあ、プリンタールームまで行ってきて」というのを受けて、その場でのオンライン教師なし学習に基づいて知識を獲得して行動できるようになる。まだいろいろな制約が入っていて、実際にはまだまだ万能ではないが、環境や他者とのインタラクションだけから学習して場所概念を獲得することを試みている。

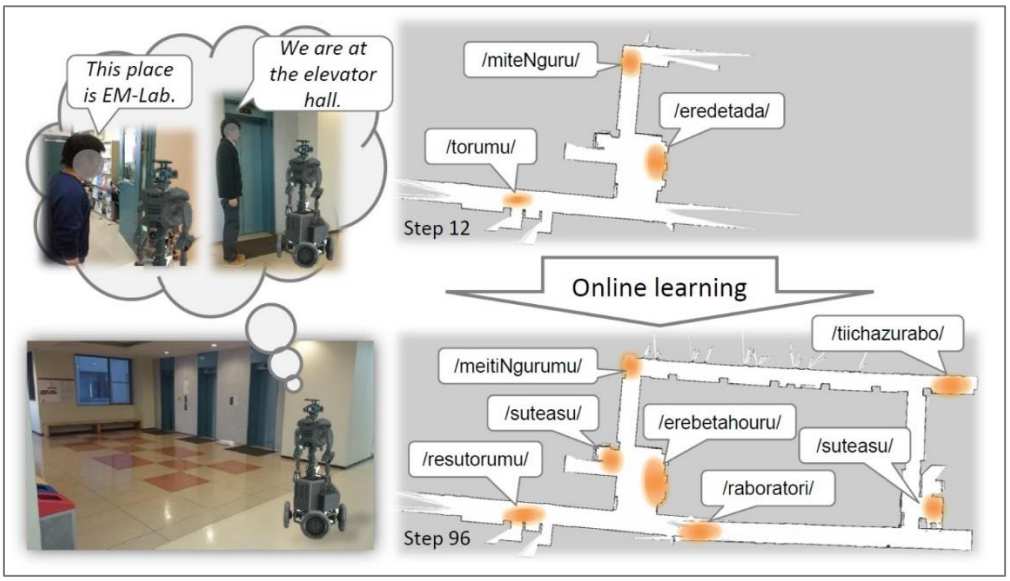


図 3-9　場所概念獲得のデモの様子

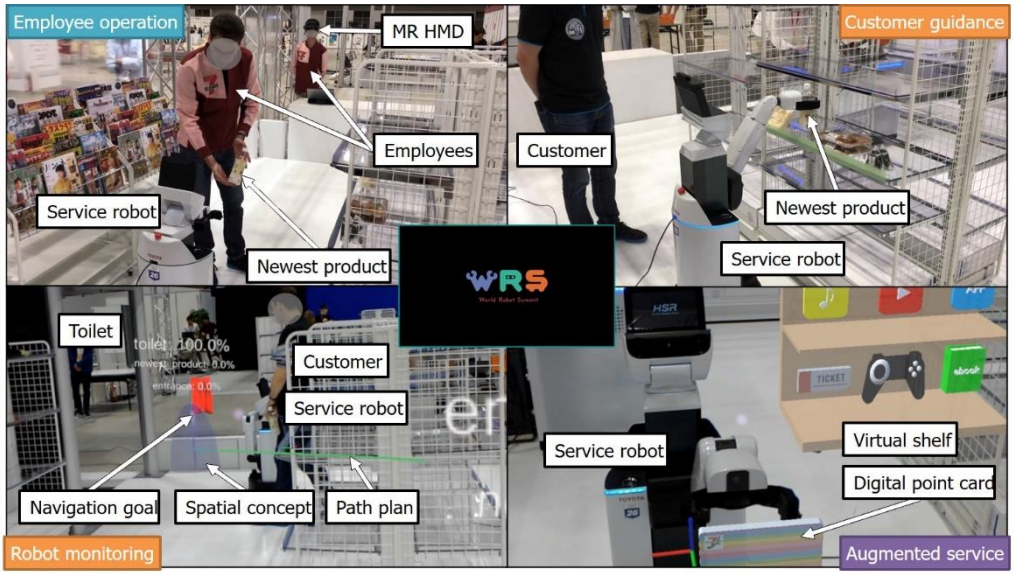


図 3-10　コンビニ接客タスクの様子

ここまで説明したようにコンセプチュアルな面が中心に検討されてきて、実際の産業応用までいってない点が、現状、記号創発ロボティクスの弱いところかもしれない。これに対して、僕らなりにチャレンジしていて、経済産業省と NEDO（新エネルギー・産業技術総合開発機構）が主催するワールドロボットサミット（WRS）のコンペティションに参加している。先ほどの場所概

念のモデルも導入したロボットシステムを作り、WRS2018 では複数のタスクで入賞し[43]、特にコンビニの接客タスクでは優勝することができた。図 3-10 はこの接客タスクの様子を示している(動画を YouTube で見ることができる[44])。現場でロボットが活動しながら、その環境でのユーザーとのインタラクションなどから知識を得て、それをユーザー支援に活用するといったことを実現した。ただし、このシステムでは、ロボットがすべて確率的生成モデルで動いているわけではなくて、かなりの部分はルールベースのプログラミング、いわばシンボリック AI 的な枠組みも使って実現しているというのが現状である。環境や他者とのインタラクションから教師なしのオンライン学習のみで知識獲得して行動できるロボットを実現するまでには、まだまだ時間がかかる。

## ３．６　確率的生成モデルに基づく統合認知アーキテクチャ

このように記号創発ロボティクスの文脈では、確率的生成モデルをよく使う。その理由は、基本的には観測情報だけから知能を作っていきたいからである。できるだけ教師なし学習の枠組みの上で取り組んでいきたいと思っている。ただし、教師なし学習と教師あり学習や強化学習との境界線もかなり微妙になりつつあるのだが。

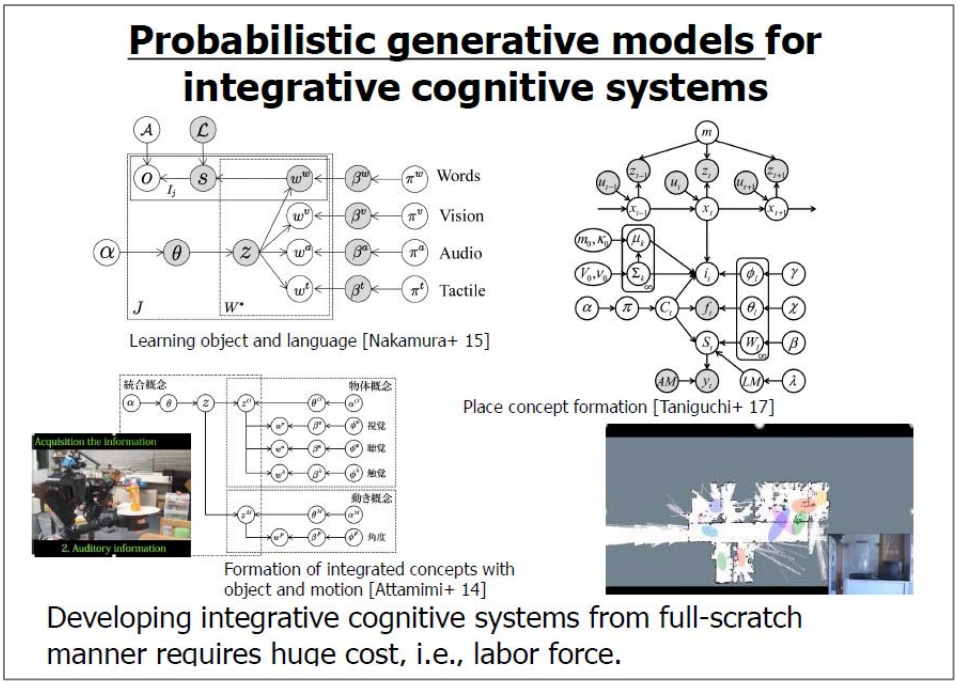


図 3-11　統合認知システムのための確率的生成モデル

これまで、長井隆行先生のグループや中村友昭先生のグループをはじめ、さまざまな研究者との共同研究も進めながら、さまざまなモデルを考えてきた。例えば図 3-11 では、右の部分が先ほどの SpCoSLAM のグラフィカルモデル、左の部分が物体概念と語彙を同時学習するモデル、左下は、mMLDA（Multi-Modal Latent Dirichlet Allocation）と呼ばれる物体概念と動き概念を統合させるモデルを表している。また、先ほど話したように、この中には SLAM、混合ガウスモデ

[43] http://www.ritsumei.ac.jp/news/detail/?id=1212
[44] https://youtu.be/2sBsquvSIOA

ル、音声認識モジュールなどが組み込まれている。ある種、レゴブロック的にいろいろなモジュールを組み合わせて作っている。

それで、言語理解や、場所や物体の概念獲得や、行動計画などを１つのシステムとして、全体で同時に学習できるようにするには、1 つのグラフィカルモデルにしたいわけだが、これが非常に大変である。これを効率的に開発できるようにしようという取り組みも進めている。それぞれを別の確率的生成モデルのモジュールとして作り、それらを組み上げるような作り方ができるようにする。組み上げる際は、密結合させるのではなく、モジュール間で通信する。信念伝播（確率伝播）、Sampling Importance Re-sampling、Metropolis-Hastings などを検討・提案しているが、何かしら確率的な情報をやり取りする。それによって、別のモジュールの間で通信しながら同時学習ができる。

これは深層学習の文脈でも面白いところがある。深層学習では微分可能性に基づいてエラーをシステム全体に伝播させながら最適化を図るが、いま述べた確率的なモデルでは、微分可能でなくても確率的な情報をやり取りできれば全体を最適化できるという話になる。この場合、1 つ 1 つのモデルはかなり異種なものでもつなぐことができて、ヘテロな認知モジュール群をまとめて確率的生成モデルで扱おうというものである。このアイデアに基づいて、大規模な統合認知アーキテクチャを作るための枠組みとして、2018 年に SERKET（Symbol Emergence in Robotics tool KIT）を発表した（図 3-12）[45]。

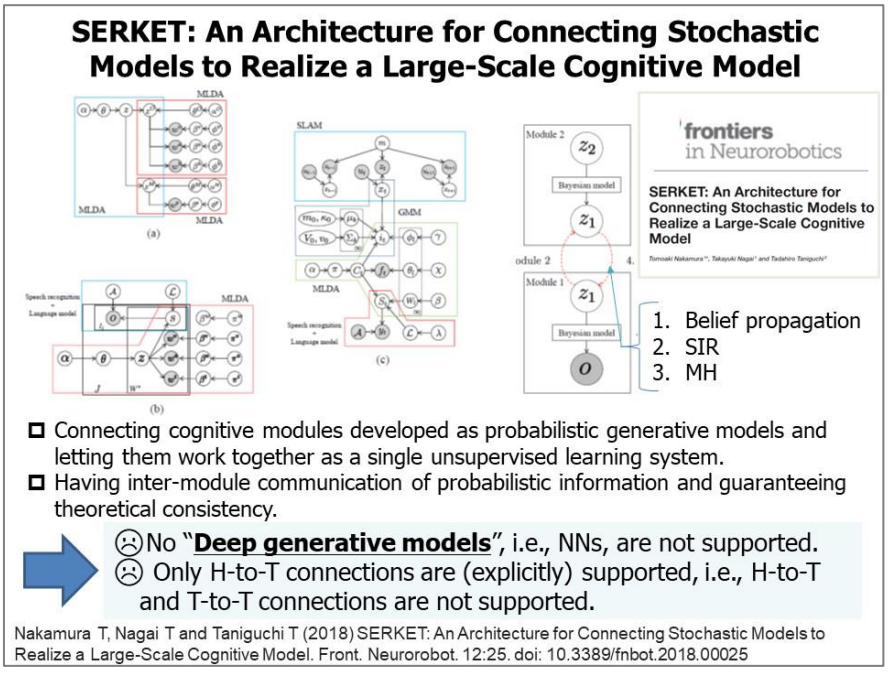


図 3-12　SERKET

[45] Nakamura T, Nagai T and Taniguchi T (2018), "SERKET: An Architecture for Connecting Stochastic Models to Realize a Large-Scale Cognitive Model", Front. Neurorobot. 12:25. doi: 10.3389/fnbot.2018.00025

ただ、この当時、SERKET はまだ深層学習に対応していなくて、テクニカルな面でもノードのコネクションに制約があるという課題があった。このあたりを発展させて、最近 Neuro-SERKET というのを発表した[46]。先ほども名前を挙げた長井先生・中村先生のグループの他に、松尾研究室の鈴木雅大先生も一緒にやっている。この Neuro-SERKET は深層生成モデルもサポートしている。また、グラフィカルモデル詳しくない人には分かりにくいとは思うが、SERKET は Head-to-Tail 型のコネクションを対象としたが、Neuro-SERKET はさらに Tail-to-Tail 型や Head-to-Head 型のコネクションも、少し近似は入るけれど扱うことができるようになった（図 3-13）。

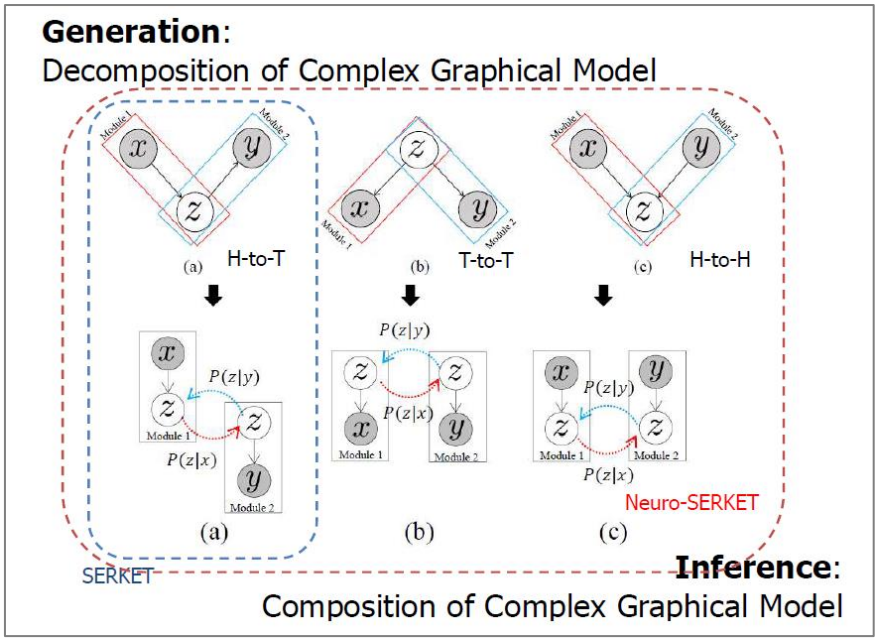


図 3-13　SERKET と Neuro-SERKET のカバレッジ比較

Neuro-SERKET の論文では、かなり人工的なデモではあったが、MNIST のデータとそれを読み上げた音声データというマルチモーダル情報のクラスタリングを行った。このときのシステムは、画像データと音声データを入力として、図 3-14 にあるように、VAE、GMM、ASR、LDA という 4 つのモジュールで構成した。VAE（変分自己符号化器）は画像の表現学習を行う。その潜在空間を GMM（混合ガウスモデル）でクラスタリングする。一方、ASR（音声認識）には Julius を使っていて、その結果の離散文字列を LDA（潜在的ディリクレ配分法）でクラスタリングする。そのクラスタリング結果が GMM のクラスタリング結果と一致すると、つまり、GMM のクラスターID と LDA のクラスターID をまとめられると、画像から概念と音声からの概念が結びついたことになる。

このような 4 つのモジュールの統合を Neuro-SERKET の枠組みで実現した。なので、各モジュール 1 つ 1 つの実装自体はほぼいじらないで統合ができている。さらに、ここでは、VAE のような完全に深層学習ベースのもの、GMM のような古典的なものなど、中身のかなり異なるモ

[46] Tadahiro Taniguchi, Tomoaki Nakamura, Masahiro Suzuki, Ryo Kuniyasu, Kaede Hayashi, Akira Taniguchi, Takato Horii & Takayuki Nagai (2020), “Neuro-SERKET: Development of Integrative Cognitive System Through the Composition of Deep Probabilistic Generative Models”, New Generation Computing, Vol.38, pp.23-48. https://doi.org/10.1007/s00354-019-00084-w

ジュールを結合していて、それらを同時学習させている。図 3-14 中の表に示しているように、個々の学習機を単純に結合しただけだと、クラスタリング精度は 62%、27%とあまり高くないのが、全部を同時学習させると 90%以上の精度が出ている。

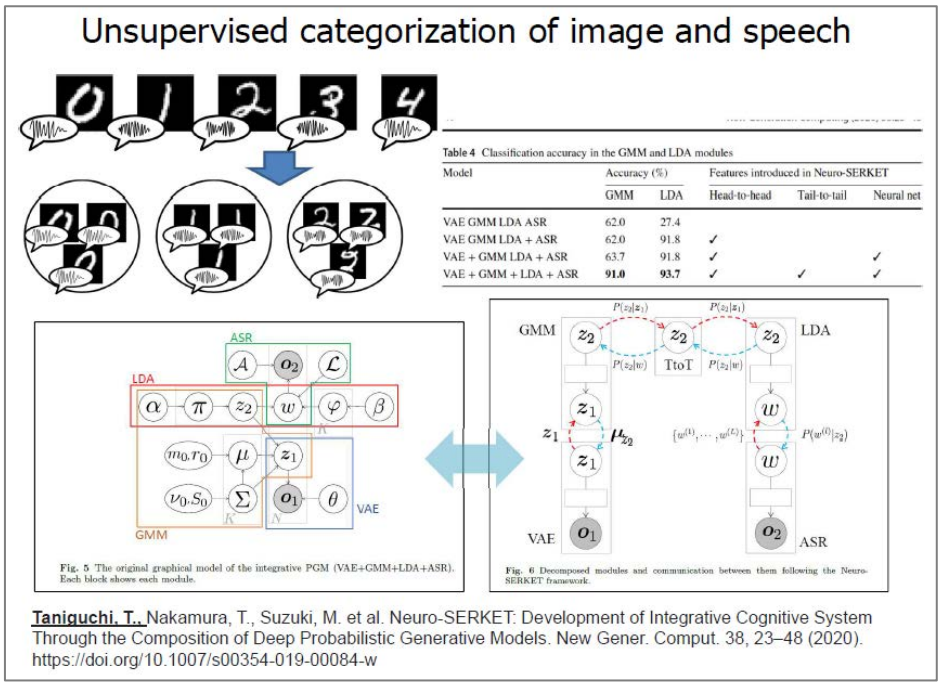


図 3-14　Neuro-SERKET による教師なし分類

こういう枠組みを使い、もっとスケールアップさせて、教師なし学習ベースでいろいろなモジュールを組み込んでやっていこうと考えている。教師なし学習を強調したが、ポイントは教師なしという点よりも、むしろ同時分布を学習するということだと思っている。

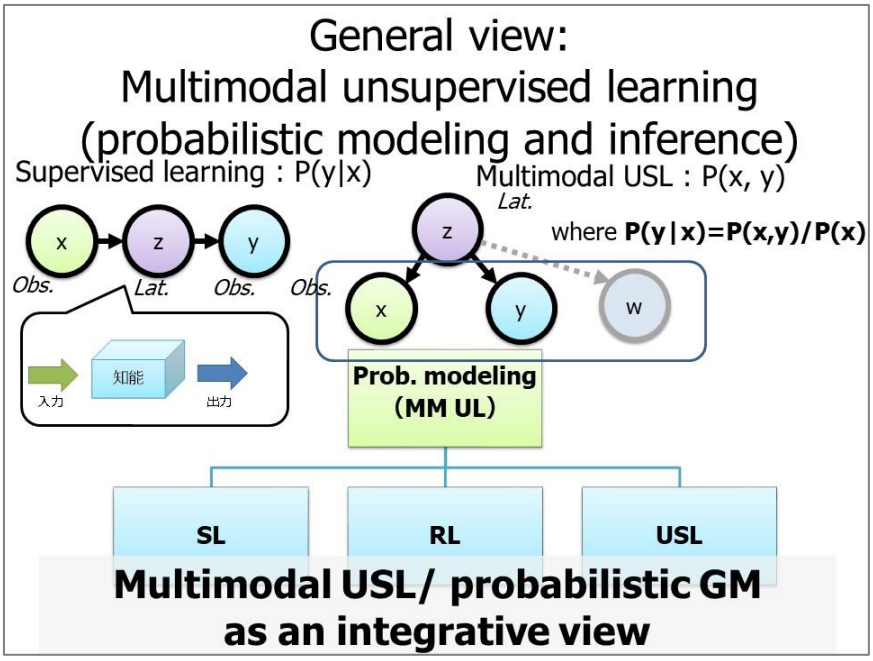


図 3-15　マルチモーダル教師なし学習

実際、マルチモーダルの同時学習で分布学習をするとき、観測データの確率的なモデリングと潜在変数の推論という形で整えると、実は教師あり学習も教師なし学習も強化学習も大部分はその傘下に入る。基本的に、教師あり学習は入力 X に対する出力 Y の自己分布をモデリングすることなので、これは当然入るが、強化学習についても、報酬という概念の代わりに最適性という概念を導入して、その最適行動の計画という推論問題に帰着させると、モデルベース強化学習はベイズ推論の問題とほぼ等価の式に直せるという話がある。このことから、行動計画のようなものも、確率的生成モデルに基づく認知アーキテクチャに統合していけて、その結果、環境とのインタラクションの中で言語獲得し、言語的指示に従って行動計画を行うようなものも実現していけるのではないかと考えている。

このあたりの研究はパナソニックの岡田雅司さんらと一緒に取り組んでいる。Variational Inference MPC の論文[47]では、モデル予測制御を変分推論で解くことをしている。モデル予測制御はモデルベース強化学習とほとんど同じと思ってよい。もう 1 つ PlaNet of the Bayesians の論文[48]は、「猿の惑星」をもじった「ベイズの惑星」というタイトルだが、中身は世界モデルを使ったモデルベース強化学習で確率的な推論を行っている。

## ３．７　記号創発ロボティクスのチャレンジ

最後に、記号創発ロボティクスのチャレンジを図 3-16 にまとめた。

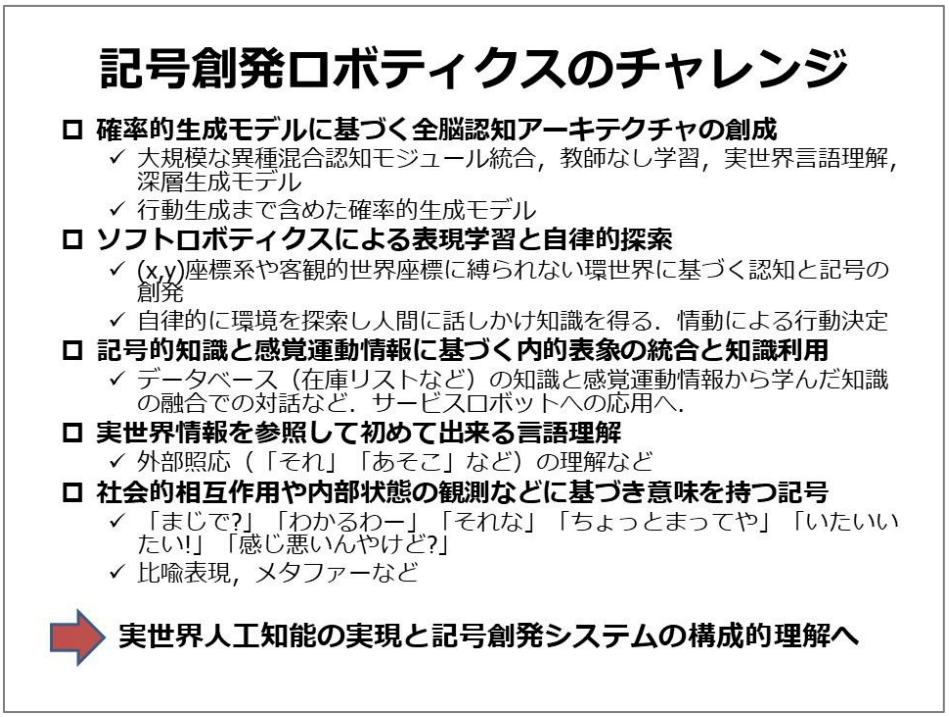


図 3-16　記号創発ロボティクスのチャレンジ

[47] Masashi Okada, Tadahiro Taniguchi (2019), “Variational Inference MPC for Bayesian Model-based Reinforcement Learning”, CoRL 2019.
[48] Masashi Okada, Norio Kosaka, Tadahiro Taniguchi (2020), “PlaNet of the Bayesians: Reconsidering and Improving Deep Planning Network by Incorporating Bayesian Inference”, IROS 2020.

ここに示したように、記号創発ロボティクスのチャレンジはまだまだたくさんある。実際に言語獲得してインタラクションやコミュニケーションができるようなシステムを作っていかないといけない。現状はまだかなり限定的で、実際に物の名前を言って、その場所に行かせるといったことしかできてない。より大規模な語彙を扱ったり、さまざまなモジュールを統合したりしていくことが必要であるし、人間の言語というのはもっと不確かで、あやふやで、その場限りの意味が生まれるようなものも多いので、そのようなものにも適応していけるようにしたい。また、我々のロボットの表現学習は現状としてハードな（硬い）ロボットの身体の上で実行しているので、人間の場合とかなりギャップがある。やはりソフトな身体の上で実行すること、つまりソフトロボティクスとの融合のような話もこれから重要になっていくと思っている。

# 4．総合討論

◆モデレーター：中島 秀之（札幌市立大学 学長）[49]

## 4．1　総括コメント

2 人の講演を受けつつ、中島が考えていることを話したい。

最初にすごく荒っぽくそれぞれの立ち位置を整理すると、松尾さんは 2 階建てモデルの話をして、2 階建てだけど深層学習をベースにしていくべきという話をして、谷口さんは記号創発の話をして、基本的に 2 人ともボトムアップのアプローチだと思う。私はトップダウンかというと、そうでもなく、ハイブリッドかなと思っている。

当初予定していたパネル討論をやる時間はもうなさそうなので、以下、ハイブリッドの話をして、最後は福島さんに締めてもらおうと思う。

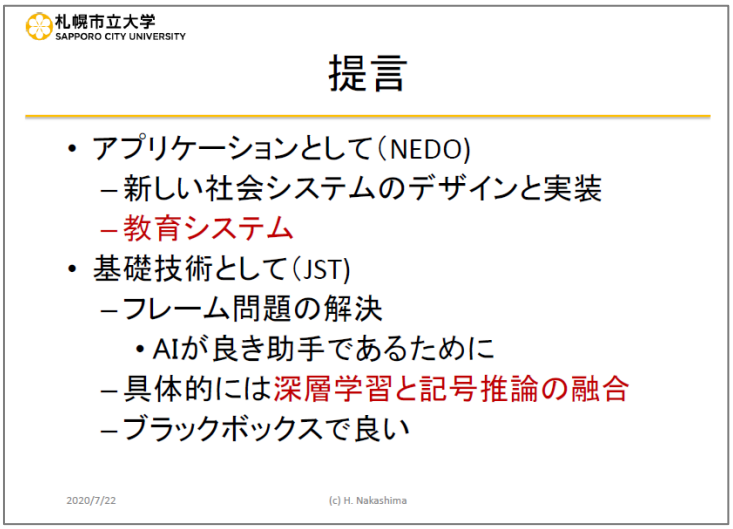


図 4-1　提言

ハイブリッドの話に入る前に、少しだけ今回の企画についても話すと、このセッションの冒頭でも触れられたように NEDO と JST が共同で取り組もうというのを以前から私がけしかけていた。NEDO は社会実装という面でアプリケーション側の話をして、JST は基礎技術をやるということで、この深層学習と記号推論の融合というテーマを掲げたと思っている。

実は、NEDO では教育システムを作るというのをやってほしいと思っている。午前中の基調講演でも話したように、このコロナ感染症の時代に教育はますます大事になるし、今までのシステムでは動かないというのも分かっているので、新しい教育システムを我々が作っていかないといけないのではないかと思っている。パネル討論では、教育の話と先ほどの 2 人の講演の話がつながるぐらいまで議論できるといいと思っていたが、今日それは難しそうだ。

さて、松尾さんは予測という話をしていたが、私は予測だけでは駄目だと思っていて、最近、予期という言葉を使い始めた。予測は同じ物理レベルで次の状態はどうなるかというものである

[49] https://www.scu.ac.jp/profile/hideyuki-nakashima/

のに対して、予期は１段上の認知レベルに上って、また下りてくるというものである。

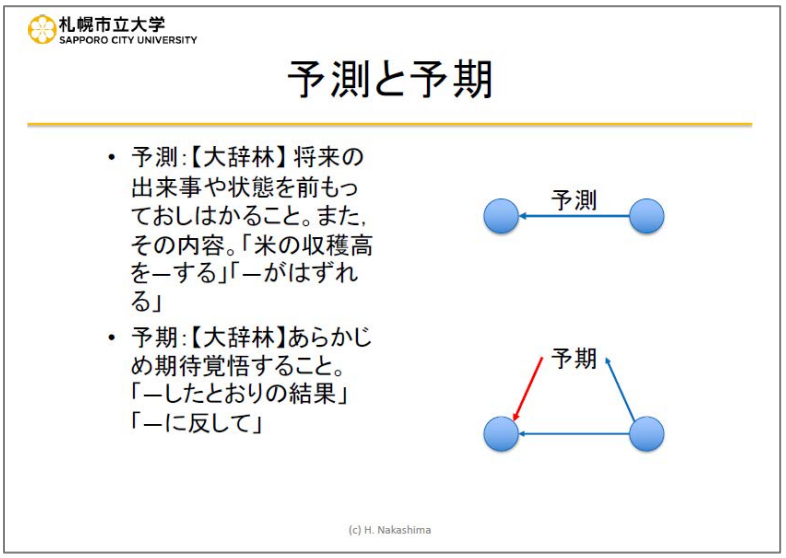


図 4-2　予測と予期

上のレベルとは何かだが、Yuval Noah Harari の「サピエンス全史」にとても印象的なことが書かれていて、サピエンスがネアンデルタール人になぜ勝てたかというと、言葉の力だと言っている。つまり、虚構の構築による集団の制御が鍵で、言葉あるいは虚構の進化は肉体の進化よりもはるかに速いタイムスケールで進むので、他を圧倒できた。基本的に政治とか宗教というのは虚構の最たるもので、そういうもので集団を統一していたと。

このとき、やはり物語というレベルには独自の法則があって、ボトムアップに出てくるものだけではないだろうということは考えておきたい。

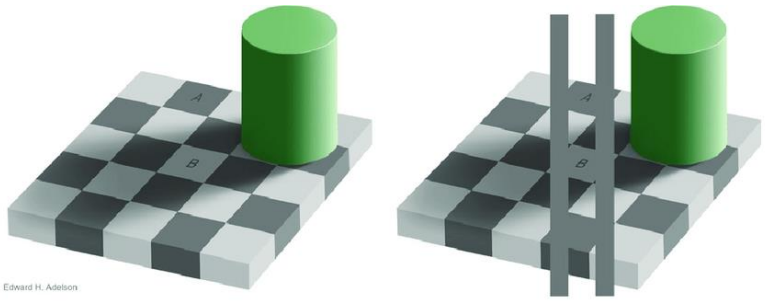


図 4-3　Checker Shadow Illusion

Rodney Brooks を松尾さんはパターン派に入れていたけれど、実は彼がムーンショットの国際シンポジウムで講演したとき[50]、ニューラルネットワークを攻撃する側に立って、今のニューラルネットワークではこんなのできないだろうという例をいろいろ示した。図 4-3 はその中の１つで、Edward Adelson が 1995 年に発表した Checker Shadow Illusion と名づけられた錯視である。この左側の図で A と B を比べると、人間は明らかに B のほうが明るく見える。しかし、実

[50] 2019 年 12 月 17 日・18 日に開催されたムーンショット国際シンポジウムにおける分科会 3 で、Rodney Brooks が講演した。その講演内容は YouTube でも公開されている。https://youtu.be/ehM_JzrpZD4

際には、右側の図で確認できるように、A と B の明度は同じである。この錯視を深層学習で可能になるかを考えたい。これを実現するには、深層学習の上に記号の層が要るのではないかと思う。記号の層が要るというのは、これは影だからというような知識を動員した認識が必要ということである。

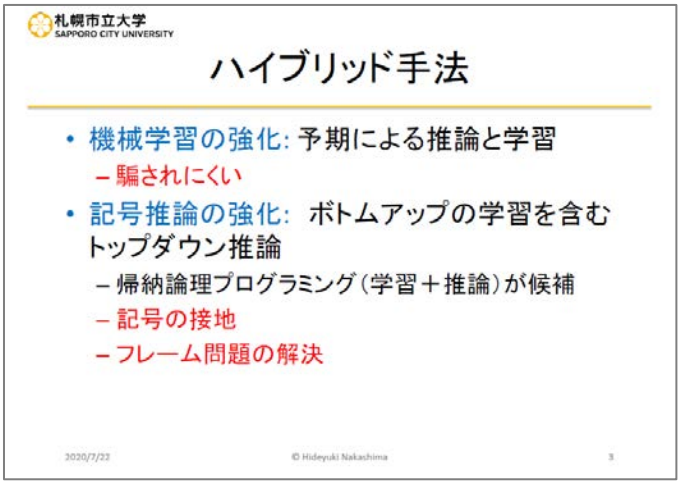


図 4-4　ハイブリッド手法

ということから、私は以前からハイブリッド手法ということをずっと言ってきた（図 4-4）。機械学習の弱点として騙されやすいという問題があるが、予期による推論と学習によって騙されにくくできるかもしれない。一方、記号推論の弱点はやはりボトムアップのところが弱いということで、記号接地やフレーム問題がまだ解決されていない。フレーム問題は、これから AI を助手として使おうとしたときに重要な問題になると思っている。

帰納論理プログラミングは、第 5 世代コンピュータープロジェクトの頃からずっとやっている人がいるが、学習と推論がセットになっている。そういう意味で、ボトムアップとトップダウンがこの中に含まれているので、枠組みとして使う手はあるかなと思っている。

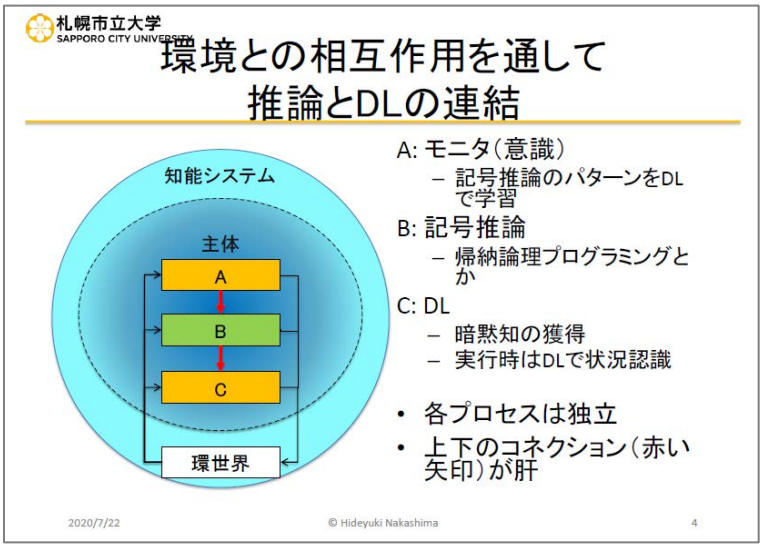


図 4-5　環境との相互作用を通して 推論と深層学習（DL）の連結

谷口さんの講演で触れられていたが、環境との相互作用はやはりとても大事で、図 4-5 のようなアーキテクチャを考えている。この図の中では「環世界」と書いたが、これは環境のことで、環境まで含めて知能システムを考えるべきと思っている。先ほど松尾さんも話していたように、この中でモニターや推論が上下のつながりを持ちつつ、また、環境を通したループの中で動いている。このようなハイブリッドのシステムをうまく作れればいいと思っている。

## ４．２　セッションの最後に

**中島**　残り 5 分あるので、松尾さんと谷口さんから、今の私の話に反論してもいいし、言い足りなかったことを話してもらってもいい。

**松尾**　記号が上に必要と捉えるのではなく、また、谷口さんの話で物議を醸すところがあると言っていた 2 種類の記号の意味のあたり（３．４節）にも関わるところだが、僕は結局、報酬系が 2 種類あるということだと思っている。視覚的な報酬系があるからそういう運動系が生まれたわけで、記号的な報酬系があるから記号系が生まれて発話の予測をするようになった、ということだと思っている。2 個のシステムというのは、報酬系の違いに根差していると。

**中島**　問題を報酬系に先送りしたような気もするが。

**松尾**　いや、報酬系であるがゆえにメタになる。記号系の方はどんどんメタにいくことができて、それが何か特殊な進化をした結果、数学や論理学のような記号だけの体系のようなものを創り出したという感じではないかと思う。

**谷口**　まず、今の報酬系の話については、やはりモデルベースとモデルフリーで異なるという議論がある。脳の中で、モデルフリーの方は基底核ベース、モデルベースの方は大脳皮質のインファレンスに全部置き換えられるので、たぶんそのあたりで 2 種類の話はつながるのかなと思った。

次に、中島先生の話に出てきた環世界について、環世界というのを言ったのは動物記号論の学者 Jakob von Uexküll だが、彼は環境と環世界を切り分けている。環境というのは客観的に上から観測できる物理的な存在、例えば地図のようなもので、一方、環世界というのはあくまで我々の感覚運動系が感じ取ってでき上がるもので、そこには客観的な座標系なんてないと言っている。世界モデルが作ろうとしているのは環世界の方なのだけれど、ロボティクスの研究の中では何かと座標系を使ってしまう。我々自身は客観的な座標系なんて学校に行って初めて知るのに、最初から座標系を置いてしまっている。言語でも同じようなことがある。

講演の最後あたりでソフトロボティクスについて触れたが、ソフトロボティクスが難しいのは、そもそも客観座標の上で扱うのが超難しいということがある。やはり環世界という、実環境とのインタラクションの中で生まれる、主体にとっての環境の経験に基づいて知能を作っていくことが非常に重要だと思っている。

**中島**　そこは賛成で、最近、プロジェクションサイエンスとかをやっている人もいる。この文脈では、論理学のような客観的な世界は少し関係ないという扱いでいいと思う。

もうセッションの残り時間がほとんどないので、最後は福島さんに締めてもらいたい。

**福島**　非常に面白い話題で、ぜひ突っ込んでもっと話をしたかったが、今の話を受けた締めというのも難しいので、実はパネル討論ができたら取り上げたいと思っていた話題に最後触れて

おきたい。今年 1 月に今日と同じようなテーマで JST CRDS 主催のワークショップを開催した。その中で、こういう分野の研究を進めるために必要なデータは何だろうか、これまでの深層学習に用いていたものとは違う、新たなタイプのデータをたくさん集めて研究を加速する必要があるのではないかという話をした[51]。そのとき、大量データとして必要なのは、記号的なやり取り、つまり言語的なやり取りのログと、実世界での操作の両方が入ったようなマルチモーダルデータをたくさん集めることが重要だという話になった。ワークショップの中では、これをエクスペリエンスデータと呼んだ。今は新型コロナ禍のため、リモート教育とか、遠隔医療とか、遠隔ロボティクスとか、サービスがどんどんリモート化されているので、エクスペリエンスデータが溜まりやすい状況になってきている。これをどんどん蓄積して、研究に回していくようにできると、今日議論されたような次世代 AI の研究が加速されるのではないかと思っている。

**中島**　一言だけ言うと、今の話は環境から入ってくるデータ、外からのデータだと思う。一方、今日の 2 人の講演や私が強調したこととして、中で動くデータ、内部モデルといったものもある。記号推論の上に深層学習をのせるというのは、その記号推論自身がデータになるという関係になる。松尾さんが言っていた、考える知能の方のほとんどの情報は脳内で作り出しているという話になるので、そういう方向をもう少し考えとかないといけないかなと思う。下からのデータが要るというのは深層学習として当然だが、それだけでは駄目だろうと思う。

**松尾**　僕は、コロナ禍で人が外に出られない状況なので、ロボット遠隔とかでどんどん動かしたらいいと思っている。

**中島**　アバターとか。

**福島**　なので、コロナ禍での遠隔サービスからのエクスペリエンスデータの獲得のようなことをうまくやって活用していくと、この分野の研究を加速する手段になるのではないか、ということを言ったところで、本セッションの締めとしたい。

## ４．３　質疑応答

セッション中は質疑応答の時間が取れなかったが、大会 Slack チャンネルにてやり取りされた質疑応答部分を以下に抜粋する。

**質問**　共通のシンボルの裏にある間主観性というか各自のクオリア、内部表現の違いが可視化されるようになると、世の中で起きているいろいろな齟齬や誤解を埋めてくことができたりするか？　あるいは、似た内部表現を共有する人達のコミュニティーが、フィルターバブルっぽく分断されていくディストピアもあり得そうな気もするが。

**松尾**　齟齬や誤解を埋めていくことができるかもしれない。フィルターバブルっぽく分断されることはあるかもしれないが、ネガティブな面よりポジティブな面のほうが大きいのではないかと思う。

**質問**　松尾先生の 2 階の記号部分について、記号推論というと A→B、B→C という抽象化したラ

[51] 科学技術未来戦略ワークショップ「深層学習と知識・記号推論の融合による AI 基盤技術の発展」の総合討論において議論された。https://www.jst.go.jp/crds/pdf/2019/WR/CRDS-FY2019-WR-08.pdf

ベルでの推論があると思うが、そういう機構は特に用意しなくても動くか？

**松尾**　記号推論というときの記号処理は、記号の中でもかなり特殊な発展を遂げた結果のものだと思う。谷口先生の話にもあったが、もともとは概念がエージェント間でやりとりされるという機構（離散化された発話の予測システム）があって、それがメタに相互作用と学習を重ねた結果、記号推論のようなものが生まれたのだと考えている。したがって、そういった機構は特に用意しなくてもよいはずと思う。

**質問**　2 階部分の目的関数の発話予測は、離散的な予測でなく、その背後の内部表現（発話者＝お母さんの非離散的な感情とか）を予測したがっているような気もするがどう思うか？

**松尾**　はい、それ自体は、知覚運動系（動物 OS）が、知覚自体を予測せずに、もっと潜在的な変数を予測しようとしているというのと同じ話だと思う。予測精度を上げようとすると、より背後の潜在変数、内部表現を予測するようになるということだと思う。

**■報告書編集担当■**

| | | |
|---|---|---|
| 福島 俊一 | フェロー | （システム・情報科学技術ユニット） |
| 東 良太 | フェロー | （システム・情報科学技術ユニット） |

※お問い合せは下記までお願いします。

**CRDS-FY2020-XR-02**

**JSAI2020 企画セッション報告書**

# 次世代ＡＩ研究開発 ―さらなる進化に向けて―

令和 2 年 10 月　October 2020

ISBN 978-4-88890-690-6

国立研究開発法人科学技術振興機構　研究開発戦略センター
Center for Research and Development Strategy, Japan Science and Technology Agency

〒102-0076　東京都千代田区五番町 7　K's 五番町
電　話　03-5214-7481
E-mail　crds@jst.go.jp
https://www.jst.go.jp/crds/

ISBN 978-4-88890-690-6