特　集　「考証：2001年宇宙の旅」

HALの謀反

Rebellion of HAL

中島　秀之　　電子技術総合研究所
Hideyuki Nakashima　Electrotechnical Laboratory.
nakashim@etl.go.jp

Keywords: error, three laws of robotics, frame problem.

1. はじめに

HALはある意味で究極の人工の知能であろう．チェスが人間より強いのはもちろん，絵の観賞もするし感情もある．人間の命令に逆らえば，人を殺すこともできる……．

しかしながらHALを描いたのは人工知能には素人のSF作家と映画監督である．物理学と宇宙旅行に関してはかなり精度の良かった映画（それでも数か所の間違いがあり，後にClarke自身が認めている[Agel 70]）ではあるが，人工知能研究の観点から見ると根本的な誤解も散見される．

2. HALは間違わないか

映画の中でHALが自分自身を“incapable of error”と表現する場面がある．“間違いはいつも人間に原因がある，”と．このような，理想の計算機は間違いがなく，人間は間違いを犯すものだという見方こそ，素人の意見と言いたい．筆者のArther C. Clarkeは科学の専門家ではあるが，計算機の専門家ではない．

確かに現在の計算機は間違わない．間違いがあるとすれば人間のプログラミングの間違いであるように思われる．しかし，この“ミスは人間にある”という見方には二つの疑問がある．

第一は，計算機とはどこまでのことか．ハードウェアだけか？　OSまで含むのか？　あるいはその上のプログラムまで含むのか？　ソフトウェアの作り間違いのことを“虫”と呼ぶが，ハードウェア設計に虫がいる例はインテルのチップで明らかになった．演算回路に設計ミスがあったようである．さて，このミスは設計した人間にあるといえばそうだが，造られた回路にミスがあったことも否めまい．人間が完璧なOSを作っても，それが完璧には動かないのだからOS作者にとっては計算機が間違ったと言わざるを得ないのではないだろうか．さらにOSに虫がある例はもっと多い．この場合はプログラマにとってみれば，プログラムは正しいのだから計算機のミスということになろう．計算機は人間が作ったものだからミスは人間にあるという言い方は，（これ自体は正しいが，）人間は神が作ったものだから人間のミスは神の設計ミスであると言っているに等しく，あまり情報のある命題ではない．

第二は，第一の疑問よりもっと本質的なのだが，間違いは知能の本質ではないかという疑問である．“間違いは知能の必要悪である”という言い方をしてもよい．間違いはないほうがよい．しかし，間違いをなくすと同時に知能もなくなるのではないか？　現在の計算機は，そういう意味での間違いは犯さない．それほど知的ではないからである．しかし，HALのように知的な計算機は間違いを犯すはずである．

この見方にはもっと説明が必要であろう．

橋田[橋田 95]の主張している情報の部分性という考え方がある．彼によると情報の部分性には入力の部分性（必要な情報が全部は揃っていないこと）と処理の部分性（時間などの資源制約により手持ちの情報が完全には処理しきれないこと）の2種類がある．例えば，株の取引をするときには，株価の動向を予測するためのすべての情報（企業の経営や新製品開発の状態，新しい法律が制定されるか否かなど）が手に入るわけではない．これは入力の部分性である．将棋や囲碁のようなゲームでは盤面の情報は完全に手に入るが，それを完全に処理する時間がない．もし完全に読み切れれば，将棋は先手必勝，後手必勝，引分けのどれかにしかならず，そのどれになるかがわかるのであるから1手目にして勝負がつくことになる．しかし，現実にはやってみなければわからない．処理の部分性である．

しかも，もっと悪いことに，これらはシステムにとっては区別できない．つまり，ある推論を行っているときに解が得られない場合，もう少し計算を続ければ解が得られるのか，あるいはもう少し情報を収集しなければならないのかは不明である．情報あるいは時間のことを資源と呼ぶことにすると，資源が足りないという制約のもとで最善の答を出すのが知能である．この方法はアルゴリズムと対比する形でヒューリスティ

クス（heuristics）と呼ばれている*．資源が不足しているのだから論理的に正しい答えは出ないかもしれない．

例えば，朝，電車の駅に行って，どのあたりで待っていると座れるだろうか？ という問題を完全に解くには情報が不足している．いつもは車両の両端が空いているのだが，ごくたまにはそこに修学旅行の団体が乗っていたりする．このように，ヒューリスティクスによる解はたまに間違える．しかし，アルゴリズムによると，この問題の解は出ない．いつも“わからない”と答えるよりは，たまには間違ってもよいから“一番前の車両”と答えるほうが賢いと思う．これが知能である．そのように考えてみると，HAL が常に間違わない答えしか言わないとしたら，それはきっとあまり有用な情報を与えてくれないに違いない．

ヒューリスティクスについてはもう少し詳しくみておく必要がある．初期の AI の教科書などでは発見的手法と訳されることも多く，その名が示唆するように探索問題などの解法の一つとして研究されることが多かった．例えば先に述べたゲームで先読みをする場合に全部を読んだのではとても時間とメモリが足りないので読む範囲を限定する．その場合には場面の良さを近似する評価関数を計算し，これが低いものは先読みを打ち切るなどの手法が用いられる．このヒューリスティクスが失敗するのはこのようにして読みを打ち切った先が正解着手だった場合である．失敗を減らそうとすれば多くの枝を読まねばならない．いかにうまい手法と評価関数を使って先読みの枝を刈込み早く正解を見つけるかがヒューリスティクスである．刈込みが大きいほど早く計算できるから頭が良いことになる．しかし，その分だけ失敗のリスクも大きくなる．

最近ではあまりヒューリスティクスという言い方は聞かれなくなったが，私としてはもっと強調すべき概念だと思っている．先に述べたように，ヒューリスティクスは知能とほぼ等価な概念である．従来は個々の問題を解くためのヒューリスティクスを場当たり的に組み込んだプログラムが多かったように思うが，これからはもっと系統的に，問題解決手法そのもののヒューリスティクス（メタヒューリスティクスと言うべきか？）を研究すべきであろう．例えば，以下のようなものである．

- 注 意

  情報の全体を扱えないとしたら，その部分に注目するしかない．情報のどの部分に注目するかのヒューリスティクスが重要である．AI では探索問題などで探索範囲を切り捨てることの研究がなされてきた．これも注目の一つの形態であろう．しかしながら，探索における枝刈りアルゴリズムの評価は常に客観的なものであり，完全性が要求されないまでも志向されていた．ここで問題とする注意はそのような情報側に依存する性質ではなく，主体の都合によるものである．

- 主体の思い込みによる状況変化への対応

  情況推論（筆者の造語．状況理論の主観版），実時間推論や学習などもこれに含まれる．

  古典的 AI の枠組みは

  観測 - 推論 - 行動

  という連鎖を繰り返すものであったが，すべてを受動的観測による入力情報に頼っていたのでは観測の負荷が多大となるうえに，情報の部分性にもろに制約されることになる．これを打破するには予測が重要である．予測は正しいとは限らないという意味でヒューリスティクスの一つである．思い込みを恐れず，積極的に思い込みをするシステムが知的である．

- メモリベース

  主体の経験したことだけが基礎である．経験に基づく推論，つまり，状況に応じて行動する時に完全に主体の独断と偏見で行動して良いというヒューリスティクスである．メモリベースアプローチは，この考え方をとったときに始めてその意義が実感できる．

  映画では Discovery 号に積まれた HAL と双子の計算機が地上にあり，HAL の計算の誤りを確認するシーンがあるが，上記のような状況依存性や経験の違いを考えた場合には，たとえ計算機といえども双子が同じ解に到達する保証はない．

3. HAL はなぜ反乱したのか

続編の“2010 年”は HAL が反乱した理由の謎解きのはずであった．結構期待して読み始めたのを覚えている．しかし私はがっかりしてしまった．実は Asimov のロボット小説に出てくるようなロボット心理学者のような活躍を期待していたのであるが，あっさしりた表面的な解しか提示されなかったからである．

Asimov のロボットたちはロボット三原則という以下のような規則にのっとって行動する．

(First Law) A robot may not injure a human being, or, through incation, allow a human being to come to harm. 人間を傷付けるべからず．また，傷付くのを看過ごすべからず．

(Second Law) A robot must obey the orders given it by human beings except where such orders would conflict with the First Law. 第一原則に反しない限り人間の命令に従うべし．

(Third Law) A robot must protect its own exis-

---

* HAL は Heuristically programmed ALgorithmic computer の略だそうだから，筆者の Clarke がこれらの区別を理解していたとは思い難い．

tence as long as such protection does not conflict with the First or Second Laws. 第一，第二原則に反しない限り自己保存すべし．

規則に書くと簡単だが，実際の場面ではこれらが複雑に入り組んでおり，その結果としてロボットが一見奇怪な行動をとる．それを解決するのがロボット心理学者である．原因の一つには規則間の矛盾がある．単純な矛盾は問題ない．例えば第二原則より第一原則が優先するので“人を殺せ”というような第一原則に反する命令には従う必要がない．しかし，例えば人間には無害な命令を遂行しようとすると自分が壊れ，その結果命令が遂行できないとしたらどうだろう．命令に従おうとすると，その結果として従えないというジレンマに陥る．Asimovのロボットたちはこのジレンマに悩むことはないが，そのジレンマの影響として奇怪な行動に走る．HALはどうだったのか．

HALの場合はロボット三原則が組み込まれていたという言明はないが，やはりなんらかの形で同等のものが組み込まれていたと考えるのが妥当であろう．人間に危害を加える（殺す）ということはやはり通常は起こらないはずである．その起こらないはずのことが起こった．

HALがアンテナ制御ユニットAE35の故障を予測する．このアンテナは地球との唯一の通信回路の一部で，故障は困る．副船長のプールが船外活動によって問題のユニットを回収し，回路を調べるが欠陥は見つからない．データを地球に送ってHALの双子計算機で計算するが欠陥はないという診断．HALはユニットをもとに戻して故障するのを待つのが良いと提案する．疑問を抱いた船長のボーマンは無線を遮断した船外活動用のポッドの中でプールと善後策を練る．“非常の場合にはHALの高次機能を遮断するのも止むを得ない”と相談する．通信器を切ってあるのでHALには聞こえないはずだったが，HALは彼らの唇の動きから会話の内容を知ってしまう．ユニットを戻すべく再び船外に出たプールをHALが制御しているポッドが襲う．ボーマンがプール救援に出ている間に，今度は船内で冬眠している乗組員が全員殺される．HALはボーマンを船内に戻すことも拒否する．

HALは任務遂行のために人間を排除しようとしたのである．自分の高次機能が止められるのを恐れたのか，あるいは人間より自分の任務遂行能力が高いと判断したのかは定かでない．

さてHALはそもそもAE35ユニットの故障予測において過ちを犯したのだろうか？　あるいはこの間違った予測も意図的だったのだろうか？　“2010年”では，これは人間の命令の矛盾によるノイローゼによる過誤だとしている．

地球外生命体の痕跡の発見は人類にとってパニックを引き起こしかねないような事態なのでこれは重大機密として，船長のボーマンにも知らされていなかった．しかし，ホワイトハウスはHALにはこの情報を入力した．HALはこの重大事実を人間に知らせてはならないという命令により，別の“人間には従え”という命令（第2原則）と矛盾する情報を抱えこんでしまったのである．

2010年の立場のように，ユニットの故障予測がHALの過ちだとしたら，HALは単純ミスを犯したことになる．しかし，これが単純ミスではなく意図的だとしたらどうだろう．人間に判断させるより，自分が遂行するほうがミッションの成功率が高いと判断していたとしたら，ミスは全く犯していないと言ってもよいのかもしれない．私としてはこちらの解釈のほうがずっとおもしろいと思うのだが．実際，HALは船外のボーマンに対して以下のように述べている．

“This mission is too important for me to allow you to jeopardize it.”

まずは，命令の矛盾により単純ミスを犯す可能性を考えてみよう．矛盾した情報により，誤った答を出すことをAIではGIGO（Garbage In Garbage Out）と呼ぶ．例えば古典論理において矛盾した公理系を与えるとすべての論理式が証明できてしまう．これがGIGOである．古典論理での典型的な証明手法は“背理法”である．ある命題pを証明するときに，その否定$\neg p$を仮定する．そしてそれから矛盾が導けると，もとの仮定：$\neg p$が間違っており，したがってpが正しいとするのである．矛盾とはある命題qと，その否定$\neg q$の両方が成立することである．ある公理系が矛盾している場合にはどのような$\neg p$を仮定しても矛盾が導ける（最初から矛盾している）ため，いかなる命題でも証明できてしまう．“AE35ユニットが故障する”という命題もこのようにして証明されたに違いない．

実は，このような古典論理的な見方はAIにおいては成立しない．なぜなら，ある命題を仮定しておいてそこから矛盾を探すという手続き（アルゴリズム）が存在しないからである．うまく矛盾がわかるような道筋を見つけ出さねばならないが，これはヒューリスティクスの分野である．普通は関連する知識に絞り込んだ探索をすることになるので，たとえ矛盾を含む知識体系であってもその矛盾する命題qと関係のない別の命題pを証明する能力は持っていない．ある体系の矛盾を発見するというのは非常に知的な作業なのである．そして，そのような知的なシステムは現在は実現されていないが，仮にHALにそうした能力があった場合，単純な矛盾によって混乱してしまうとは考えにくい．

というような理由により，私はノイローゼ説はとりたくないのである．

4. HALの止まるとき

HALとの知能戦に勝利し，船内に戻ったボーマン船長は，HALの高次機能を停止し，船の運航に必要な基本機能だけを残した．このあたりはHALのアーキテクチャの高度さをうかがわせるものがある．

このような基本機能と高次機能を分離した手法は服属アーキテクチャ（subsumption architecture）としてBrooksらによって提案・実装されている [Brooks 91]．これはさまざまな階層の機能を並列に重ねて実装するもので，上位層と下位層が矛盾する場合には上位が優先され，下位を押さえ込むのでこの名（subsume＝包括的に押さえ込むの意）がある．

脊椎動物の大脳も同様の構造をとっている．脳幹の周辺，後脳あるいは通常大脳辺縁系と呼ばれている部分は生存に必要な基本機能を担当している．その外側に位置する大脳皮質のうち，中脳と呼ばれている部分（大脳皮質の内側）は感覚入出力を司る．そして大脳前頭葉あるいは前脳（大脳皮質の外側）は感覚入力の統合を通じて意識の中枢となっていると考えられている．これらは進化の歴史に沿った発達でもある．大脳辺縁系は爬虫類の脳とも呼ばれ，爬虫類以降の脊椎動物がシェアしているものである．哺乳類では大脳皮質が大きく発達し，これは旧哺乳類の脳とも呼ばれている．そして類人猿以降では新哺乳類の脳が発達する．このように下位機能に重ねて上位機能を追加する手法が服属アーキテクチャである．

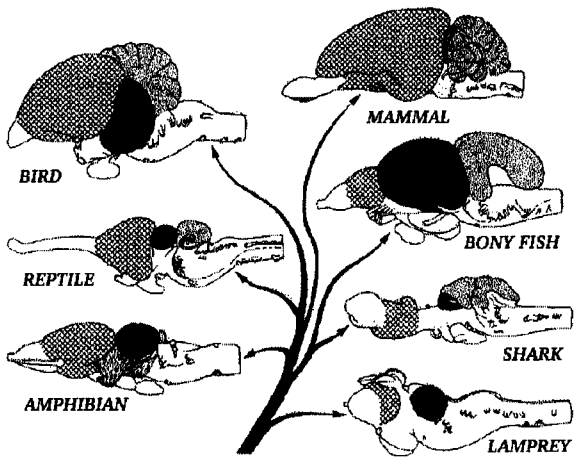

図1

HALの脳（？）もこのようにして構成されているとすると，高次の推論などを含む上位機能だけを停止させ，船のナビゲーションなどの基本機能だけを残すことが可能となる．しかしながら，映画に描かれたHALのシャットダウンシーケンスはいただけない．メモリキューブを抜いていくと，記憶が新しいものから消えていき，声もだんだんスローモーションになっていくというのは現実的ではない．

5. おわりに

“2001年”は私の高校時代に公開された映画である．確か当時の私は“人工知能”あるいは“AI”という単語は知らなかったし，計算機に興味を持っていたわけでもない．HALよりは宇宙船のほう，つまりオリオン号，宇宙ステーション，デスカバリー号やポッドのシーンに興味を引かれた．公開当時に数回，後の再演を含めると10回以上もこの映画を見たと思う．映画館にカメラとテープレコーダを持ち込んだこともある．もちろん後に発売されたビデオも買ったしレーザーディスクも持っている．私がこれまでに見た映画の中で間違いなく最高のものであると思っている．しかも他を大きく引き離して．

現実の2001年はもう目と鼻の先であるが，宇宙旅行と計算機技術はどちらも映画の予測ほどは進まなかった（チェスは確かに強くなったが）．しかし，映画には影も形もないインターネットが台頭している．今作るとHALはどんな形になったのか，興味のあるところではある．

◇ 参 考 文 献 ◇

[Agel 70] Jerome Agel (Ed.)：The Making of Kubrick's 2001, The New American Library Inc., 1970.

[Brooks 91] Rodney A. Brooks：Intelligence without Representation, Artificial Intelligence, Vol.47, pp. 139-160, 1991.（柴田正良 訳，表象なしの知能，現代思想，Vol.18, No.3, pp. 85-105, 1991）

[橋田 95] 橋田浩一：人工知能における基本的問題，人工知能学会誌，Vol.10, No.3, pp. 340-346, 1995.

2000年10月21日 受理

著 者 紹 介

中島 秀之（正会員）

1952年，兵庫県は西宮市の生まれ．関西弁と関東弁のバイリンガル．1983年，東京大学大学院情報工学専門課程修了（工学博士）．当時異端とされた人工知能，特に知識表現，推論などを研究した．現在，電子技術総合研究所企画室長．2001年4月より産業技術総合研究所サイバーアシスト研究センター長の予定．著書はなぜか本人の希望と関係なくProlog関係のものが多いが，興味の中心は“知能”にある．AIUEO，日本認知科学会，日本ソフトウェア科学会（ここまで任意団体），情報処理学会各会員．