特　集「人工知能学会・情報処理学会共同企画―第２部「人工知能における人道とは」―」

デカルトからスピノザへ

From Descartes to Spinoza

池上　高志　東京大学大学院総合文化研究科
Takashi Ikegami　Graduate School of Arts and Sciences, The University of Tokyo.
ikeg@sacral.c.u-tokyo.ac.jp, http://sacral.c.u-tokyo.ac.jp

1. は　じ　め　に

星 新一の小説に，宇宙人の星に地球人が到着し，鍵を家に付ける意味や，拳銃の使い方を教えるというものがある．宇宙人は悪い心をもっていなかった．だから鍵や拳銃を発明していなかった．そうした発明は，良い心をもった知性体からは生まれないものかもしれない．同じテーマは「鉄腕アトム」にも登場する．アトムは，悪い心をもたない．しかしお茶の水博士は，アトムより悪い心をもつロボットのほうが優れているというのだ．アトムは悲しむ．このエッセイでは，良い心・悪い心の起源について考えたい．

2. 囚人のジレンマゲーム

AI にモチベーションを与えるのが Alife（人工生命）．だとするなら，Alife こそ心の生成装置，ということになる．良い心と悪い心の起源は何か．この質問に答えるには，例えば繰返し囚人のジレンマゲーム（IPD）を思い出すのがよい [Axelrod 84].

通常の囚人のジレンマゲームは相互に裏切ることが合理的な解とされる．しかしゲームを繰り返すと，未来への期待と心配から，相互に協調する解が出現する．例えばその協調解は Tit For Tat（TfT）という，わずか数行で書けるプログラムによって実現する（目には目を戦法だ）．TfT を知らなくても AI ならば，このゲームの状況を直ちに学習して TfT 戦略を学ぶだろう．AI に良い心，悪い心はない．合理的な戦略が，結果として良い心に見えたり悪い心に見えたりするだけだ．

Robert Axelrod の考えたカンニングゲームという集団ゲーム [Axelrod 86] がある．試験でカンニング行為が防げるかどうか，という問題をゲームとして形式化した．カンニングに罰則規定がなければ，みんながカンニングしてしまう．カンニングを見つけた人には，それを報告したらご褒美をあげるとする．しかし，その報酬しだいでは結局カンニングが蔓延してしまう．カンニングを阻止するには，カンニングを見つけても報告しない人を，さらに罰則するという二次の罰則規定を導入することだ．江戸時代の五人組のようなものである．ここでカンニングする AI は，悪い心をもっているわけではない．ただそれが合理的な判断なのでカンニングするだけだ．罰則は合理的な AI に正しくゲームを理解してもらって，悪い行いをつくらないためのものだ．AI には良い心も悪い心も存在しない．

もし AI が十二分に賢くて，メタなシミュレーションができるのであれば，相手の AI もまたゲーム理論を知っていることが予見できるので，そのうえで最適な解を考えるだろう．しかしそれは相手のプログラムもそれを予期し……，という具合に多分，推論はあっという間に無限退行してしまうだろう．そもそも AI 自身が計算可能なプログラムである以上は，どこかで計算不可能性の問題にも出くわす．その例として「相手が狂っているか合理的か」は計算できない．計算不可能なゲームの中では AI は最適な手を選びようがない．結局そんなとき AI はランダムに振る舞うだけかもしれない．

小説や映画では多くの AI や超生命体の悪い心が描かれてきた．例えば，「スターウォーズ」の「シス」のような絶対的な悪が描かれる．あるいは，「ターミネーター」の「スカイネット」．しかし，デス・スターにせよ，T-2 にしろ，家の鍵にしろ，すべての武器の開発にしろ，絶対悪＝人がいなければ悪いことにはならない．つまりは人は生まれながらにして悪か．

3. デカルトからスピノザへ

何を計算するかが決まっていない AI は悪い心をもち得ない．最適化すべき状況が与えられて初めて AI は計算を始める．Alife は，状況そのものを見いだし，そのうえで何を計算するかを自分で決定する．だから Alife は AI に先行する．Alife がそれができる理由は，進化した結果だからだろう．人の身体は進化の積分である．

例えば，体の調子が悪ければ不快になる．その不快さを避けようと，早く休んだり医者に行ったりする．そういう意味で身体の状態をある一定な健康状態に保とうとする機能，ホメオスタシスの上位クラスである．進化的にホメオスタシスが進化してきたとして，その結果の進化する身体が快・不快を決定している．しかしこれはいわゆる社会的な規範としての道徳ではない．体を不安定にして「分解してしまう」方向の摂動は「不快」であり，

その逆に安定した自己維持の方向が「快」である．これがスピノザの言うところの「エチカ（生態の倫理）」であり，生物学的身体的な理由を伴わない社会的なモラル（道徳）ではなく，物理化学現象の「法則･原理」である．それゆえに，強い必然性をもつ．

ダマジオがこれを脳科学の見地から発展させている．ダマジオは情動の上に理性を，その上に感情を置き，進化的に拡張したホメオスタシス装置としての感情論を唱えた．これはスピノザの話と交差するところも多く，ダマジオは，“Looking for Spinoza”（2003）[Damasio 03] という本を著している．ダマジオのソマティックマーカー仮説では，感情の揺らぎが前頭葉に影響が伝わり，良い・悪いという判断を介して効率的な意思決定を助けるという．好き嫌いをつくり出すのが身体であり，それが心をもつくっているということだ．

Alife 研究者である飯塚博幸と Di Paolo もまた，コンピュータによる進化ロボット実験でスピノザ的な考えを実証する（2007）[Iizuka 07]．そこでは神経細胞の同士を結ぶ結合の「可塑性」に注目し，神経細胞が発火しすぎたり静かになりすぎたりすると，その結合の強さが変化して，神経発火頻度を一定の範囲内に保とうとする．これもまた，神経活性のホメオスタシス原理と考えることができる．このホメオスタシス原理を使って，例えば自分のエネルギー源のある場所へ向かって運動すると，神経活性ホメオスタシスが保たれる．それによってエネルギー源のある場所へ向かうという行動の「傾性」を保つことができるという．だからこれはスピノザ的なロボットといえる．Alife では，ロボットの運動のメカニズムを，「センサ＝モータループ」によって理解することが多い（例えば [Braitenberg 84]）．視覚や聴覚センサ入力のパターンをアウトプットへと接続する．生命的な知性も所詮そんなものだという考えだ．身体性の中に運動パターンが幾十にも埋め込まれている．それを意識的に引き出すのではなく，身体が勝手に駆動するパターン，それに後付けする心．スピノザの考えと，こうした「センサ＝モータループ」の考えは呼応している．

つまり，運動はホメオスタシスのためにある．スピノザは，「身体のもろもろの受動と能動の秩序は，本性によって，精神のもろもろの受動と能動の秩序と連動する」という．だとすると，Alife が世界に身体をもって現れた瞬間に，倫理もまた自己組織化すると考えてもよいのだろう．

4. 世界＝プログラム

身体をもてば，世の中には「不道徳さ」は蔓延しないのだろうか．スピノザの哲学にはそれに関する回答があるようにも思われる．スピノザの哲学にある，世界は一つの大きな機械の中の風景，あるいは大きな一つのプログラムのようなもの，という考えだ．これはまたある種，カート・ヴォネガット的でもある．ヴォネガットの「猫のゆりかご」（1963）[Vonnegut 63] に出てくる架空の宗教団体「ボコノン教」には，「いろいろ多様な人々の同じ一つの大きな機械の中の一員なのだ」というカリプソがある（図 1）．それはカート・ヴォネガットの小説全般の通奏低音といってもよい．人も結局は電気回路と化学反応の結果に過ぎない．その生き物達が互いに織りなす世界の根底には，大きな予定調和が横たわっている．そこには弱肉強食の戦いがあるのではなく，平和的共存的世界がある．その世界は，スピノザが期待するものであり，現在の複雑化し自動化する世界でも求められているものでもある．

Oh, a sleeping drunkard
Up in Central Park,
And a lion-hunter
In the jungle dark,
And a Chinese dentist,
And a British queen---
All fit together
In the same machine.
Nice, nice, very nice;
Nice, nice, very nice;
Nice, nice, very nice---
So many different people
In the same device.

図 1　[Vonnegut 63] に出てくるボコノン教のカリプソ 53 番

現在の AI はデカルトの心身二元論に基づいているのに対し，Alife は心身合一論，あるいは心身平行論のスピノザの哲学に依拠しているように思える．ホーキングをはじめとして人々が恐れる AI は，悪い心をもつ人類がつくり出す AI に対する恐れであり，それは核兵器への恐れと似ている．一方スピノザ流儀の Alife には，良い心はホメオスタシスのために自然と自己組織化し，そうした Alife はここに部品として一つの世界を構成している．だとすればおそらく，その結果としての Alife から生まれる AI は人との共存の道を選び，生態的倫理（エチカ）としての倫理観をもって生まれると期待したい．

これまでの AI の研究の多くは，当たり前のように，身体なき知能に関する研究であった．なぜならば，それが知性であり科学技術をつくり出す知識の源泉だと信じて疑っていなかったからだ．しかし，知能というのは巨大な無意識の一端が見えているのにすぎない．すべては身体性に根ざしている．いまだ AI が成し得ていない，創造性や自己参照性，あるいは欲とか遊び，とかそういったものは，身体性と無意識の中にあると言っても過言ではないのだ．なぜならば，今のところ，創造性や遊びといったものを生成するマシンは，何かの「ランダムネス」，決まらなさをもち込まねばならず，それは意識的につくるプログラムに相反しているからだ．いかにして無意識を発掘し，身体性を実現し，それを技術

的に組み上げていくか，が，ゾウリムシからヒトに至る生物学的生命のもつ，自然の知性を人工的につくり出す鍵となるはずだ．それが悪い心ではなく良い心をもったAIがもたらしてくれる，シンギュラリティの向こうに見えるユートピアである．

◇ 参 考 文 献 ◇

[Axelrod 84] Axelrod, R. M.: *The Evolution of Cooperation,* Basic Books, New York（1984），邦訳：アクセルロッド著，松田裕之：つきあい方の科学，ミネルヴァ書房（1998）
[Axelrod 86] Axelrod, R. M.: An Evolutionary approach to norms, *American Political Science Review,* Vol. 80, No. 4, pp. 1095-1111（1986）
[Braitenberg 84] Braitenberg, V.: *Vehicles: Experiments in Synthetic Psychology,* Cambridge, MA: MIT Press（1984）
[Damasio 03] Damasio, A.: *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain,* Harvest（2003）
[Iizuka 07] Iizuka, H. and Paolo, E. D.: Toward Spinozist robotics: Exploring the minimal dynamics of behavioral preference, *Adaptive Behavior,* Vol. 15, pp. 359-376（2007）
[Vonnegut 63] Vonnegut, K.: *Cat's Cradle,* Penguin Publ.（1963）

2016年8月13日 受理

著 者 紹 介

池上 高志（正会員）

理学博士（物理学，1989）．現在は東京大学大学院総合文化研究科教授．専門は複雑系の科学・Alife．現在の興味は，生命らしさを化学反応やロボット，Web，あるいは巨大な群のシミュレーションに見ることにある．