人工主体との
インタラクション
と生命の「脆さ」

AIと共生する未来に向けて

北海道大学大学院文学研究院
北海道大学人間知・脳・AI研究教育センター

田口　茂

CHAIN
Center for Human Nature, Artificial Intelligence, and Neuroscience

【氏名】田口　茂（タグチ　シゲル）
【所属】北海道大学大学院文学研究院・教授
人間知・脳・AI研究教育センター（CHAIN）センター長

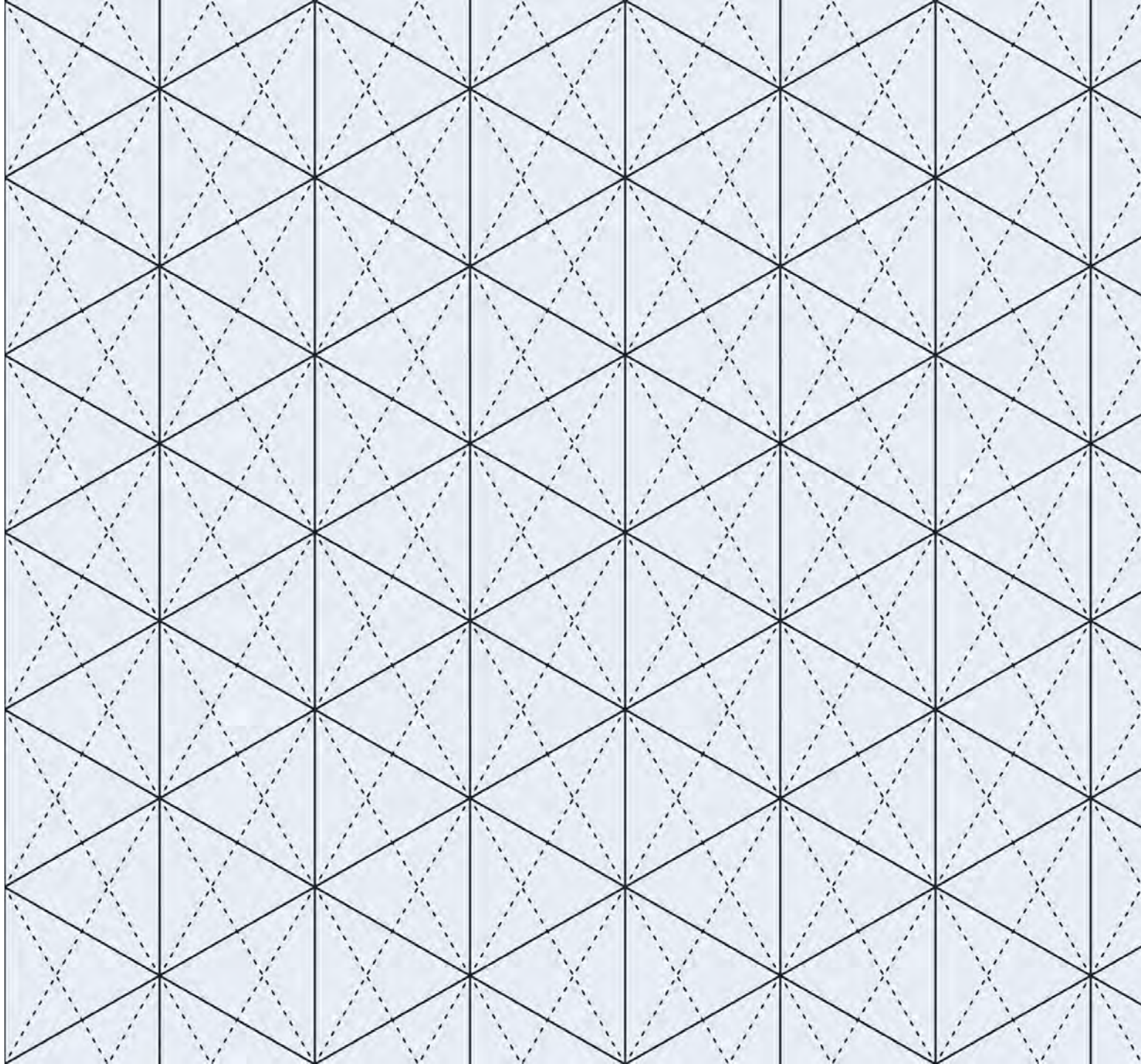
CHAIN
Center for Human Nature, Artificial Intelligence, and Neuroscience

【研究分野等】
専門は哲学。特に現象学（フッサール）。意識、自己、間主観性などを扱ってきたが、最近は神経科学者・数学者・ロボット工学者・AI研究者・認知科学者などとの共同研究が活動の中心となっている。目下「媒介」（mediation）という概念を軸に科学にも開かれた哲学のあり方を模索している。

主な著書：
Das Problem des ‘Ur-Ich’ bei Edmund Husserl. Springer: Dordrecht, 2006.
『フッサールにおける〈原自我〉の問題』法政大学出版局, 2010.
『現象学という思考――〈自明なもの〉の知へ』筑摩書房, 2014.
『〈現実〉とは何か――数学・哲学から始まる世界像の転換』筑摩書房, 2019.（西郷甲矢人氏との共著）

1. 序論

人間とAIの共存をめぐる問い



人間とAIの共存

ChatGPTが一般社会に衝撃を与えたばかりだが、この流れは止まりそうもない。今後AIはますます「人間のような」振る舞いを加速していくだろう。

ここで人類がAIを捨てる選択肢はありうるか？――ありえない。

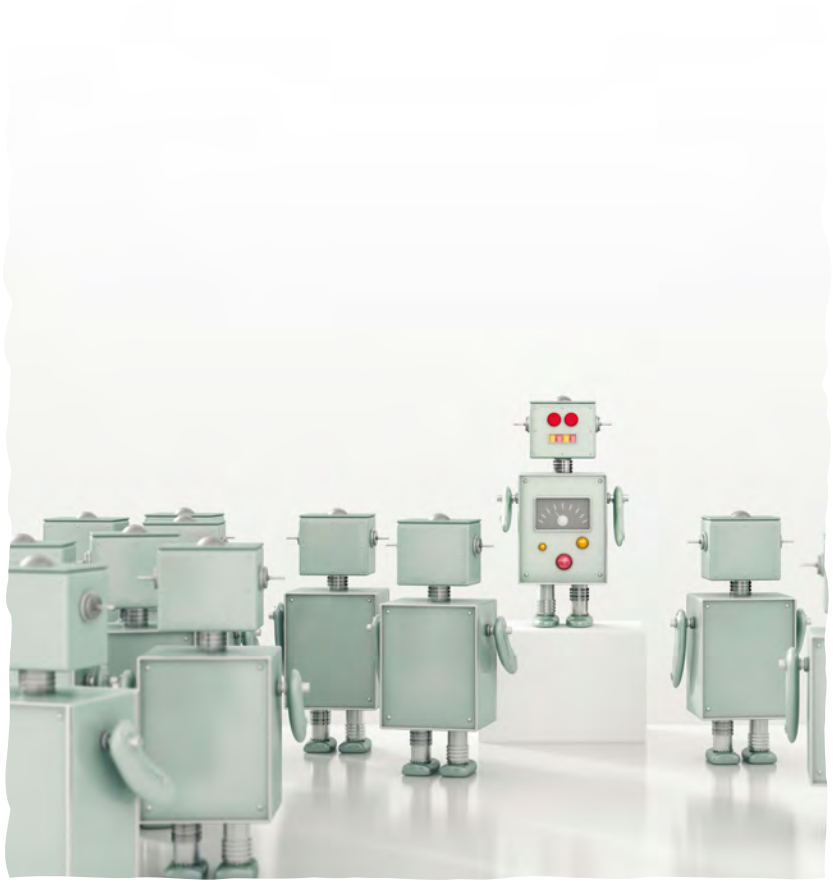
捨てないとすれば、共存しなければならない。

起こりうる否定的なシナリオ

1）人間がAIにあらゆる面で凌駕されてしまい、AIに従属するような未来

2）AIの「自律性」を徹底的に制限し、AIの発展を意図的に阻害し、AIを奴隷的な「単なる道具」の位置に留め置くようなやり方

1）AIに人間が従属
2）AIを奴隷的役割に制限



1）は、人間にとって望ましい未来ではない。

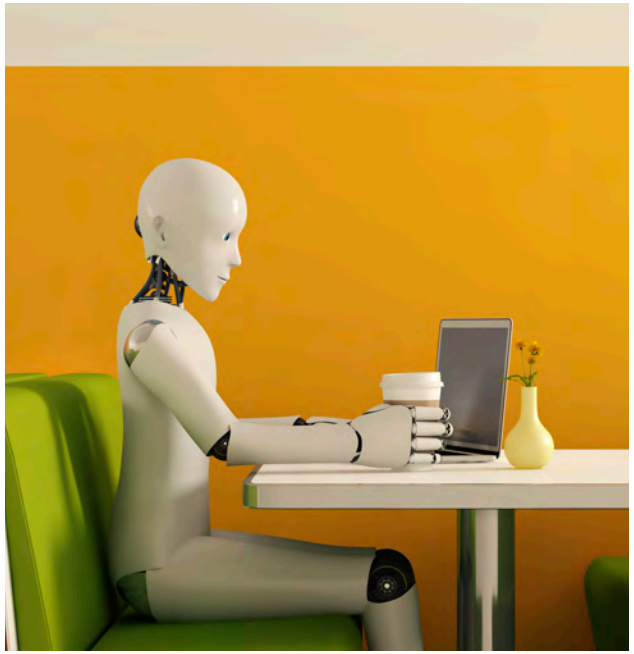
2）では、人間は主人の位置に留まるが、AIの可能性を不自然に抑え込むようなことになる。しかし、これまでの人間の歴史を見ても、新しい便利な技術が出てきたときに、それを使わないで済ますことはできそうにない。遅かれ早かれ、人間はAIの能力をフルに使おうとし始めるだろう。

そうなると、残るのは、AIの能力をうまく生かしながら、それによって人間的価値を過度に毀損することなく、AIと共存する道である。

人間とAIの共存のあるべき姿とは？

ここには様々な文明的規模の問題が潜んでいて、ここでそのすべてを論じることはもちろんできない。

そこで、コミュニケーションにおける「人間らしさ」をどう考えるか、という点に焦点を絞る。



理由：

① AIと共存するためには、AIが一定の「人間らしさ」を備えていた方が少なくともコミュニケーションは円滑になる。

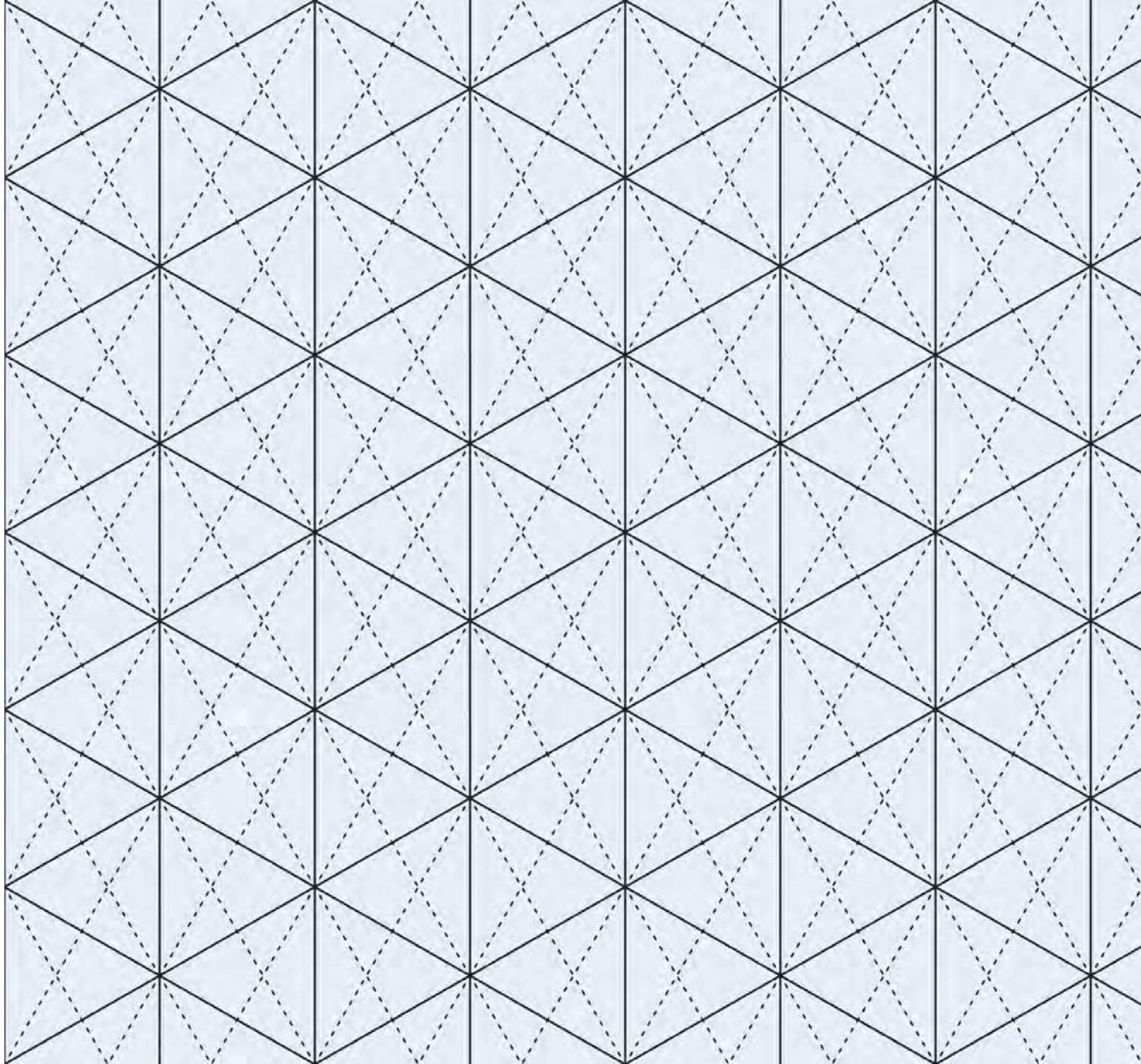
② AIのいる世界・社会で「人間らしさ」を失わないようにするためには、「人間らしさ」として保持すべき価値は何か、という点について、人間の側でも自覚的になる必要がある。

今後の議論の流れ

1. そこでまず、人間を人間として捉える、あるいはもう少し一般化するなら、主体を主体として捉える際に、われわれは何をやっているのか、という点を少し考えてみたい。そこでは、われわれの他者経験の特性に関する現象学的な分析と、エナクティヴ・アプローチによる社会性の捉え方を引き合いに出す。
2. 次いで、AIとのコミュニケーションにおいては、高次の能動的レベルと低次の受動的レベルの経験が乖離するという指摘をする。
3. さらに、より高度な意味でAIが人間のパートナーになるには、生命的な「脆さ」「不安定性」を孕んだあり方が必要になるのではないか、という問題提起をする。
4. 最後に、そのような考え方から、AIとの共存にどのような未来が考えられるかを私なりに考察してみたい。

2. 他の主体を経験すること

現象学とエナクティヴィズム



主体を主体として捉えるとは？

- 「人間らしさ」を考えるために、「他人を主体として捉えるとはどういうことか」を少し考えてみたい。
- 「そこに一人の主体がいる」と思えることが、相手を人間らしく思うための最低限の条件だからである。

- まず現象学的な考察を引き合いに出してみよう。

現象学におけるDirect Social Perception (DSP)

通常他人の心は見えないと思われている。だから推論やシミュレーションが必要と考えられている。（心理学における「心の理論」、theory-theoryとsimulation-theoryの論争など。）

これに対し、他人の心は身体の振る舞いや他者との相互作用のなかに直接「見える」という主張がある。現象学などに見られる「直接社会知覚」（direct social perception: DSP）と呼ばれる考え方である（Krueger 2018, 2019など）。

これは、とりわけシェーラーやメルロ＝ポンティが主張し、近年再び取り上げられるようになった考え方である（Gallagher 2008, Gallagher & Zahavi 2008など）。

Krueger, J. (2018). Direct social perception. In A. Newen, L. de Bruin, & S. Gallagher (Eds.), *Oxford handbook of 4E cognition* (pp. 301–320). Oxford: Oxford University Press.

Krueger, J. (2019) Enactivism, other minds, and mental disorders. *Synthese* 198 (Suppl 1): 365-389.

Gallagher, S. (2008). Direct perception in the intersubjective context. *Consciousness and Cognition*, 17(2), 535–43.

Gallagher, S. and Zahavi, D. (2008). *The phenomenological mind: an introduction to philosophy of mind and cognitive science*. New York: Routledge.（第9章）

DSP: 感情は身体の振る舞いから「見てとれる」

たとえば**感情は、身体の振る舞いから直接「見てとれる」**という。肩を震わせて泣いている人にわれわれは悲しみの感情をダイレクトに感じとるし、笑顔で跳ね回っている子供の振る舞いから、その子供の喜びを感じとる。

感情が外面だけから成り立っているというのではないが、他人から見た感情は、決して「推論」されたり自分の内部で「シミュレート」されたりするものではなく、直接「経験」されるというのである。

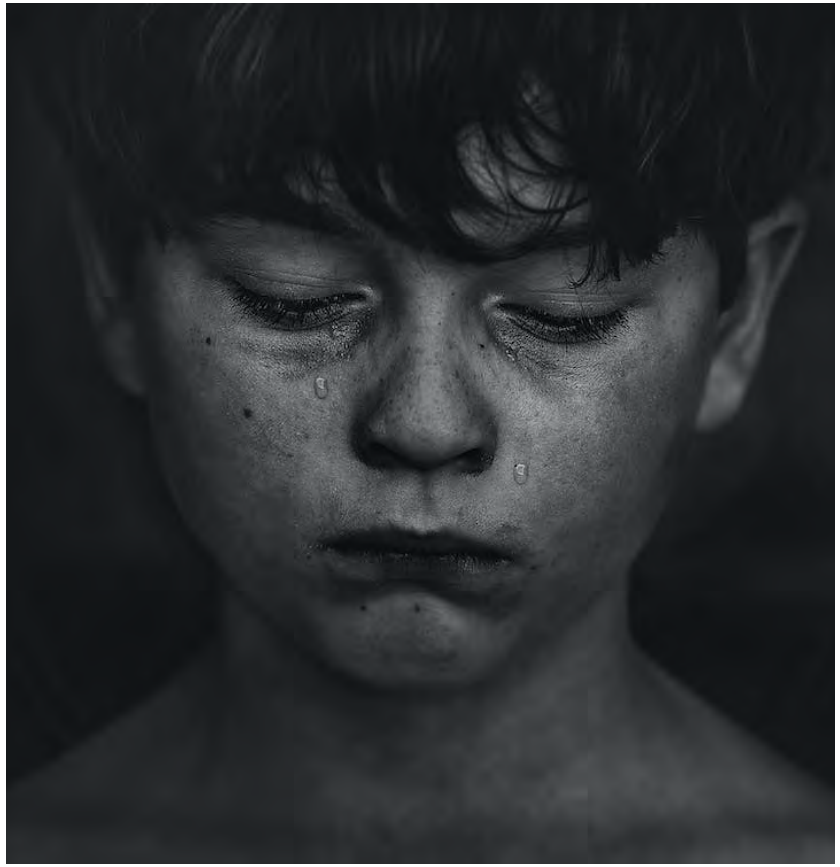
他人の内部を見通すマジカルな視線があるということではなく、そのように身体的振る舞いが他人によって見られることも含んだ仕方で、つまり身体的外面も巻き込んだ仕方で成り立っている出来事を、そもそもわれわれは「感情」と名づけているのである。（「感情」とは内部に閉じ込められたものではない。）

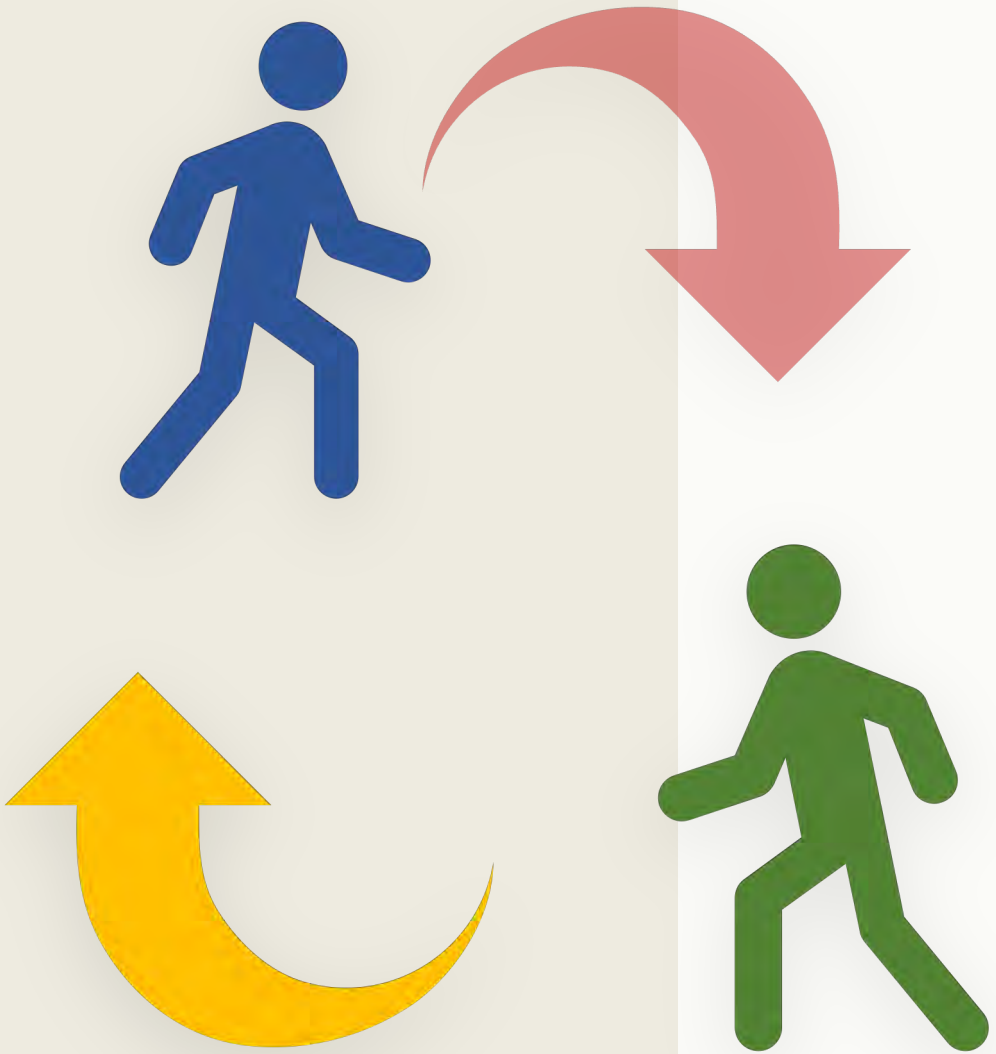
AIの振る舞いとDSP

DSPは、「他人に心を見る」ということが、われわれの能動的な推論の働きによるものではなく、受動的な「反応」であると考えている。

DSPにもとづくなら、もしAIやロボットが人間と区別のつかない振る舞いを（ある一定程度以上に）実現できたら、そこに相互作用しうる一人の「主体」を見ざるをえないということになる。

現状のAIの発展からみれば、優れた俳優の泣きの演技を学習して画面上に再現するといったことは、おそらく不可能ではないだろう。そこでわれわれは、泣いているAIに同情してしまうといった反応を否応なく引き出されてしまうかもしれない。





エナクティヴな身体的相互作用

さらに、DSPを可能にしているのは、エナクティヴな身体的相互作用であると考えられる。

エナクティヴィズム（enactivism）の考え方によれば、「他人に心を見る」ということは、他人の身体的振る舞いと自分の身体的経験との間にある種の「ループ」のような相互作用が成り立つことによる。

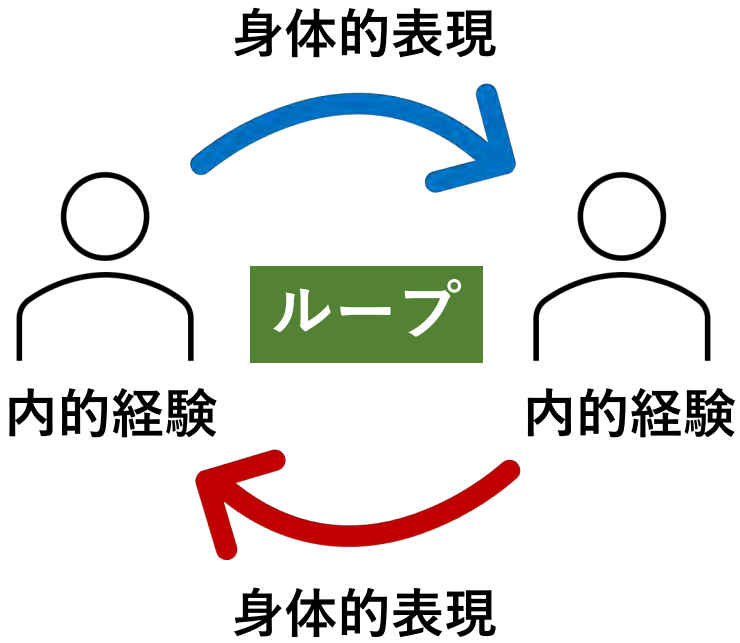
私の身体的働きかけに他人が身体的に反応し、それにまた私が身体的に反応し……というループ。

- De Jaegher, H., Di Paolo, E. (2007). Participatory sense-making. *Phenom Cogn Sci* **6**: 485–507.
- Fuchs, T., De Jaegher, H. (2009). Enactive Intersubjectivity: Participatory sense-making and mutual incorporation. *Phenom Cog Sci* **8**: 465–486.

心は身体と環境の相互作用のうちにある

エナクティヴィズムの場合も、他人の身体的な外的表現と自分の主観的な内的経験、自分の身体的な外的表現と他人の主観的な内的経験とが互いに噛み合って、はじめて間主観的なループ＝社会性が形成されている。

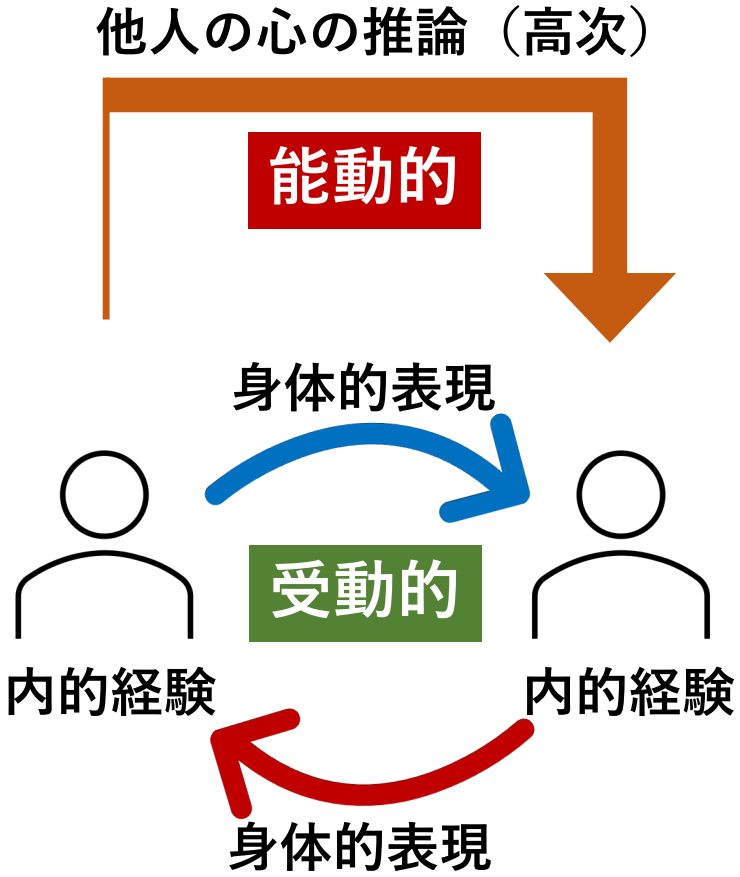
「心」とはこのような身体と環境との相互作用のうちにある。（心は内面に閉じ込められてはいない。）



身体的相互作用は受動的に生起する

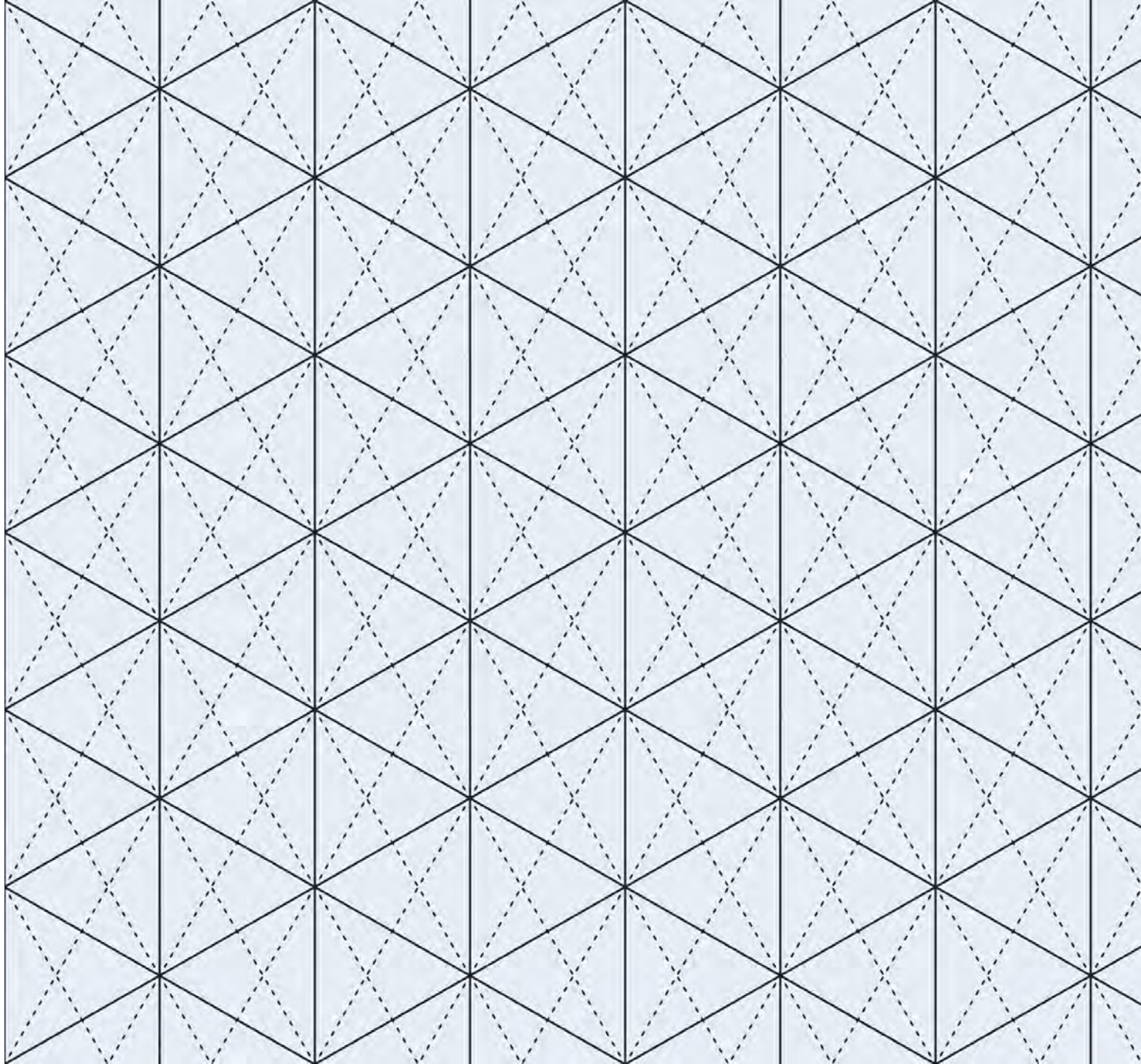
エナクティヴィズムの場合も、他者経験は高次の能動的推論によるのではなく、受動的な身体経験のレベルで起こると考えられている。

身体的な相互作用による間主観的ループの形成は、「意図的に」ではなく、「自然発生的に」「受動的に」生起する。



3. AIとのインタラクション

受動性と能動性の乖離、「ごっこ遊び」と「見立て」の世界



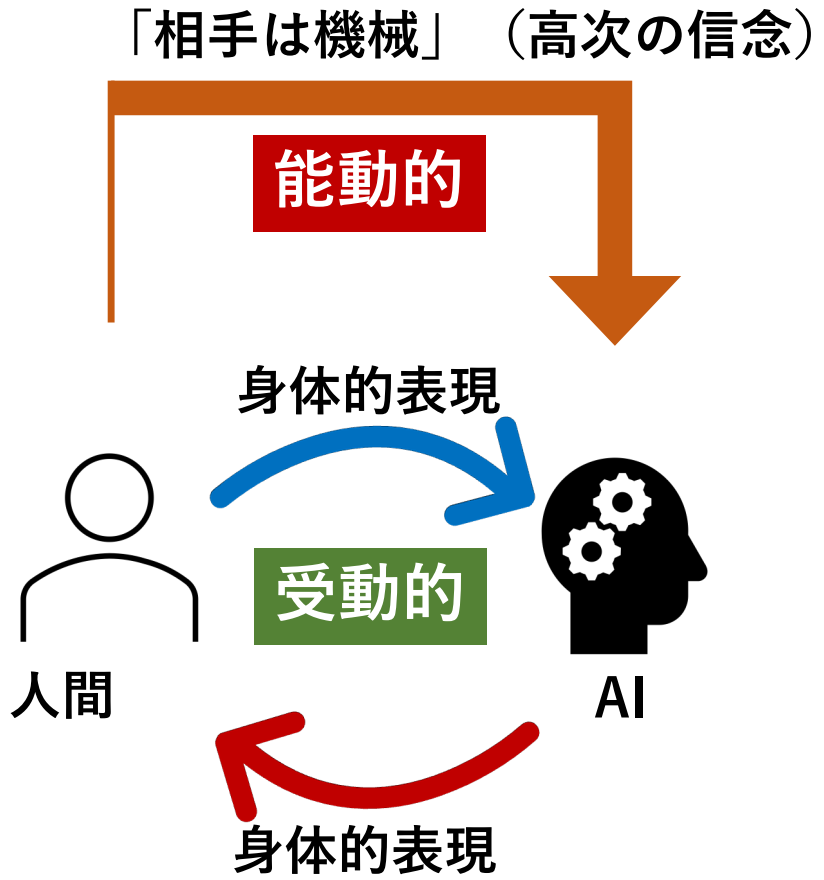
身体的相互作用は受動的に生起する

高次：能動的と低次：受動的レベルを分けた。

低次の受動的レベルのインタラクション、エナクティヴなループを発生させるような反応ができるAIが開発されたらどうなるか。

高次の信念のレベルで相手が機械だと思っていても、低次の受動的レベルでは相手との間に相互作用のループができてしまうことは避けられない。

勝手に自動的に起こってしまう。それが重要。

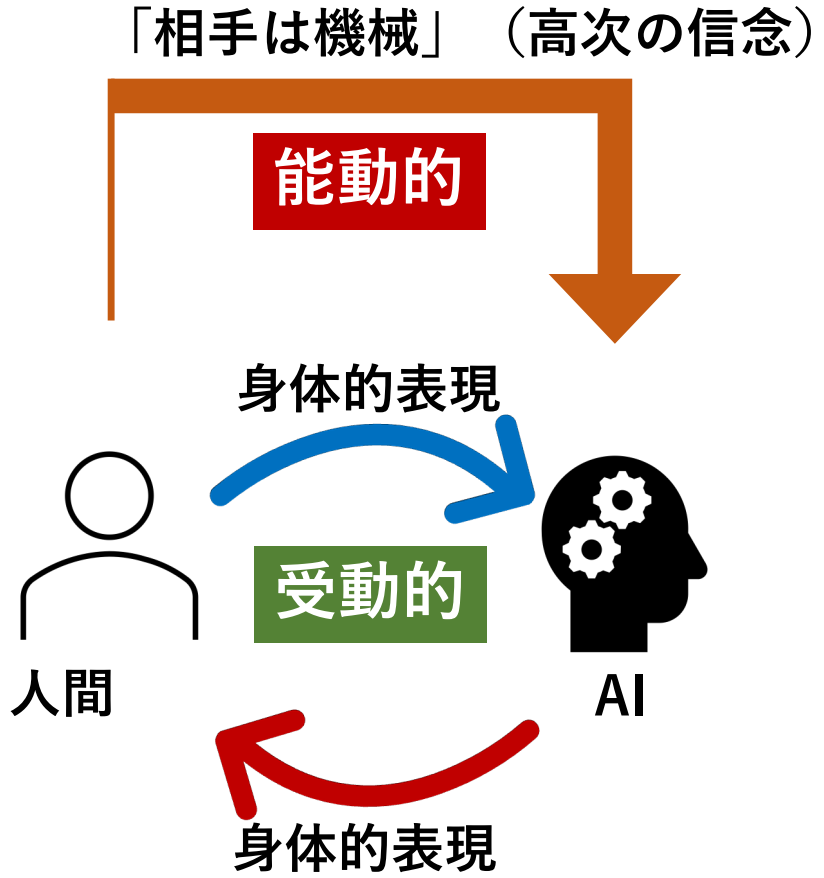


身体的相互作用があっても、人間かどうかは自明でなくなった

われわれの他者関係の基礎は、知的な推論でもないし、高次の信念でもない。受動的な身体的相互作用。

これは、一定の条件を満たせば、自然と発動してしまう。そもそも人間を相手にしていてもそうなのである。

この場合、「相手は人間か？」とあえて聞かれれば、「そうだ」と答えられる。従来は、それは問うまでもなく自明なことだった。**AIの登場により、それは自明ではなくなった。**



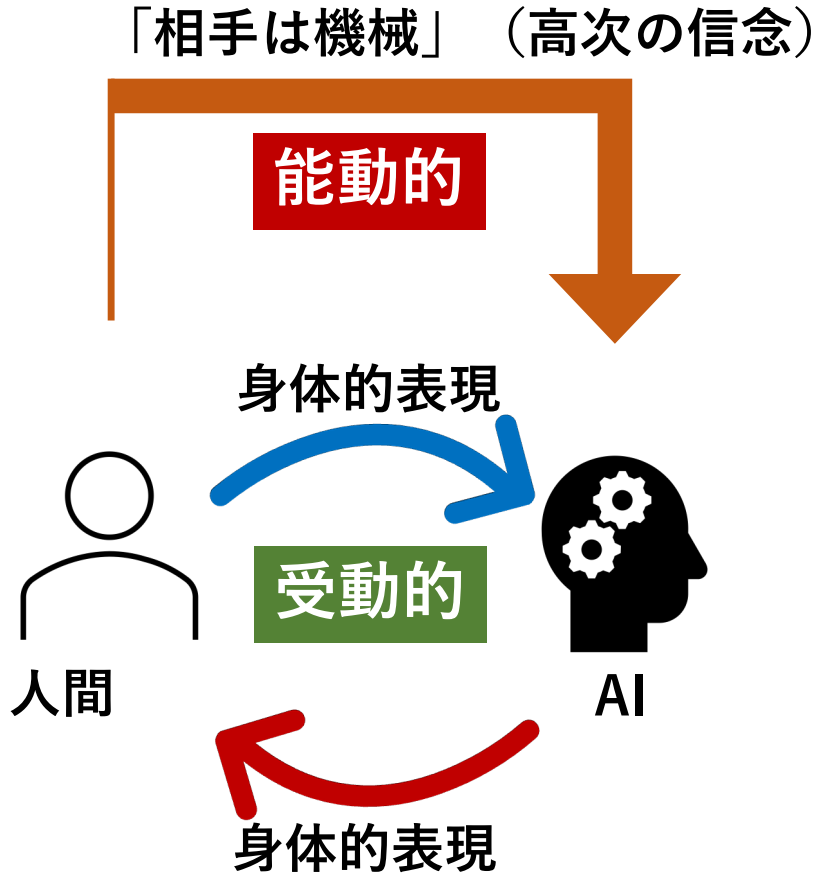
受動的レベルと能動的レベルの乖離

他者経験に関して、①受動的身体的レベルでの反応と、②高次の能動的知的レベルでの認識を区別した。

①はAIとの間にも発生しうる。②は、様々な条件を手がかりに推論すべき事柄となった。直接知覚の対象ではない。

以前は両者が連動していた。①が成り立てば、自動的に②も成り立つと考えられていた。

AIの登場は、この両者を乖離させることになった。



Eliza effect: チャットボットとの会話

AIチャットボットと自然言語で会話すると、人は非常に容易に相手を人間のように感じるという傾向がある（Eliza effect）。

近年、Woebotというチャットボットにより認知行動療法の効果を確認した研究がある（Fitzpatrick et al., 2017）。被験者の学生は、チャットボットであることを知りながらも、「関心を示してくれる本当の人間のように感じた」といった感想を述べている。

Chat GPTとLINEを使った認知行動療法ボットもすでにある（mimo AI）。

https://mimo-ai-cbt.studio.site/

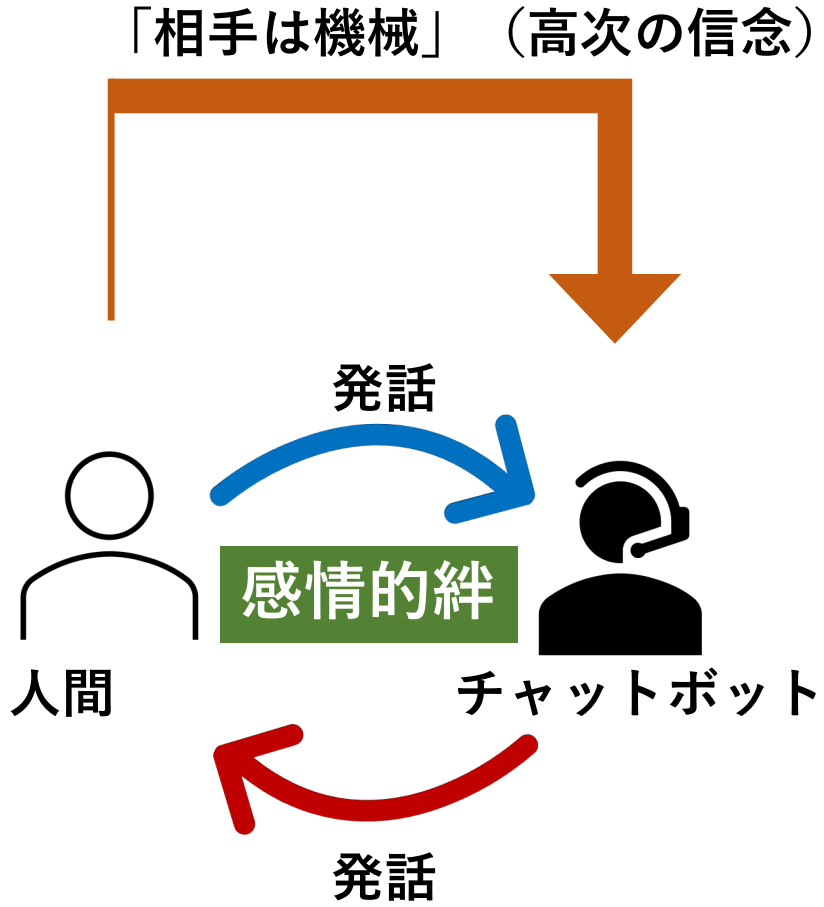
Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9, 36–45

Cristea, I. A., & Sucală, M., David D (2013). Can you tell the difference? Comparing face-to-face versus computer-based interventions. The „Eliza" effect in psychotherapy. *Journal of Cognitive Behavioral Psychotherapy*, 13(2), 291–298

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering Cognitive Behavior Therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Ment Health*, 4(2), e19.

機械とわかっていても人間のように感じる

- これは面白い現象。
- 「相手は機械である」とわかっている。
- しかし、この高次の信念に妨げられることなく、実質上会話のループが成立し、そこでは感情的な絆・結びつきまでが生み出されている。
- （高次と低次の乖離）



振る舞いのレベルと知的理解のレベルの乖離

呉羽真も、「ロボットに対してあたかもそれが心をもっているかのように振る舞いながら、明示的にロボットが心をもっているかを問われると否定的に答える人が多い」と指摘している。

（呉羽2021; Sharkey & Sharkey 2006; Wegner & Gray 2016, ch. 3）。

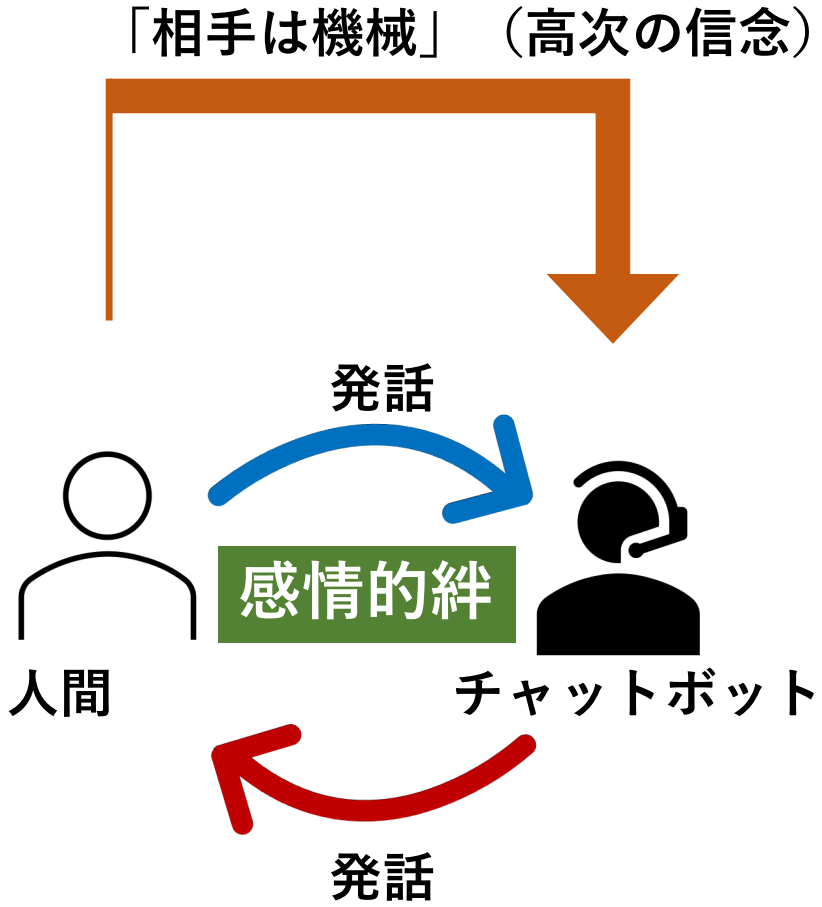
呉羽真 (2021). 日本人とロボット――テクノアニミズム論への批判. *Contemporary and Applied Philosophy* 2021, 13: 62-82.

Sharkey, N. and Sharkey, A. (2006). Artificial intelligence and natural magic. *Artificial Intelligence Review*, 25(1-2): 9-19.

Wegner, D. M. and Gray, K. (2016). *The Mind Club: Who Thinks, What Feels, and Why It Matters*. New York: Viking.

「ごっこ遊び」「宙吊り」

「そこで、彼らの行動の解釈として、ロボットが心をもっているとは信じていないながらも、「不信の宙吊り」（Sharkey & Sharkey 2006; Duffy & Zawieska 2012）あるいは「ごっこ遊び」（久木田 2017）に興じることで、人工物とのやりとりを楽しんでいる、というものが有力と考えられる」（呉羽2021）



Duffy & Zawieska (2012). Suspension of disbelief in social robotics. In Proceedings of the 21st IEEE: 484-489.
久木田水生 (2017).「AI と誠――ソーシャル・ロボットについて考える」. 久木田水生, 神崎宣次, 佐々木拓.『ロボットからの倫理学入門』, 名古屋大学出版会: 105-118.

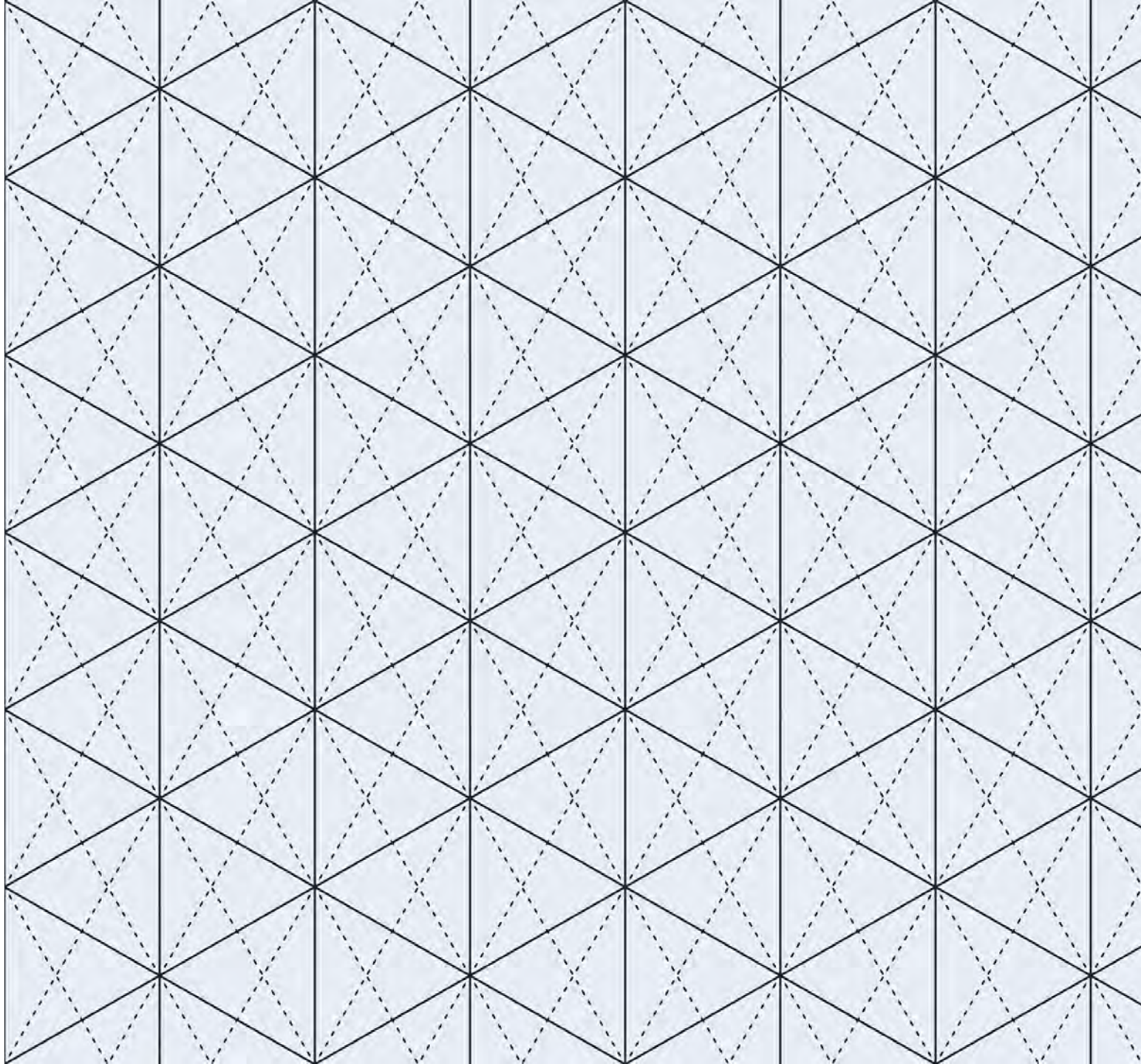
疑似人格的交流：「見立て」「あたかも」

- 相手は機械で、意識も主体もないとわかってはいるが、それでも相手との「**疑似人格的**」会話を楽しむ、といったことが可能である。
- 相手が人間に近いパフォーマンスを示せば示すほど、われわれはほとんど自動的に、相手に対して人間に対するのと同じような反応をしてしまう。それは止められない（受動的・自動的反応）。そういうことはますます増えてくる。
- それは別に構わないのではないか。「AIを人間のように誤認するのは危険だ」という意見もあるが（Fuchs 2022）、利用者に「相手は機械なんだ」とはっきり認識させておけば比較的問題は起こりにくいのではないか。
- ごっこ遊びのような、「**見立て**」のような世界。「**あたかも**」の世界に遊ぶということが可能。当分の間、AIとのつきあいはそのようなものになるのではないか。

Thomas Fuchs (2022). Understanding Sophia? On human interaction with artificial agents. Phenom Cog Sci 2022.

4.
生命の
不安定さとAI

人間のパートナーになるうるAIとは？



「同じもの」の検索 vs. 臨機応変な対応

- AIとの受動的身体的レベルでの相互作用は、おそらく近い将来かなりのレベルまで到達するのではないか。（ChatGPTなどを見ると、言語のみなら現状でも「会話」が成立している。）
- さらに必要なものがあるとすれば、決まり切った応答を崩していくような「臨機応変さ」や「思いがけない返答の面白さ」などか。
- 本の内容はいつでも誰にとっても同じ。インターネットの検索もその延長線上。
- 企業のQ&Aサービスのボットなども同様。
- 他方、本当に「会話のできる」AIを求めるとすれば、それらとは違った性格が求められる。

揺らぎつつズレていくループ：脆さと不安定性

- 変動し揺らぐ問いに対して変動し揺らぐ答えが返される。
- 互いに揺らぎつつ作用し合うループ、同じ軌道を回り続けるのではなく、**絶えず揺れながらズレていくループ**が形づくられている。
- 「対話」とはそういうものではないか。絶えず**脆さと不安定性**を孕んでいる。（一方がつんのめっても他方が支える。そこで受け止めなかったら会話は崩壊。）

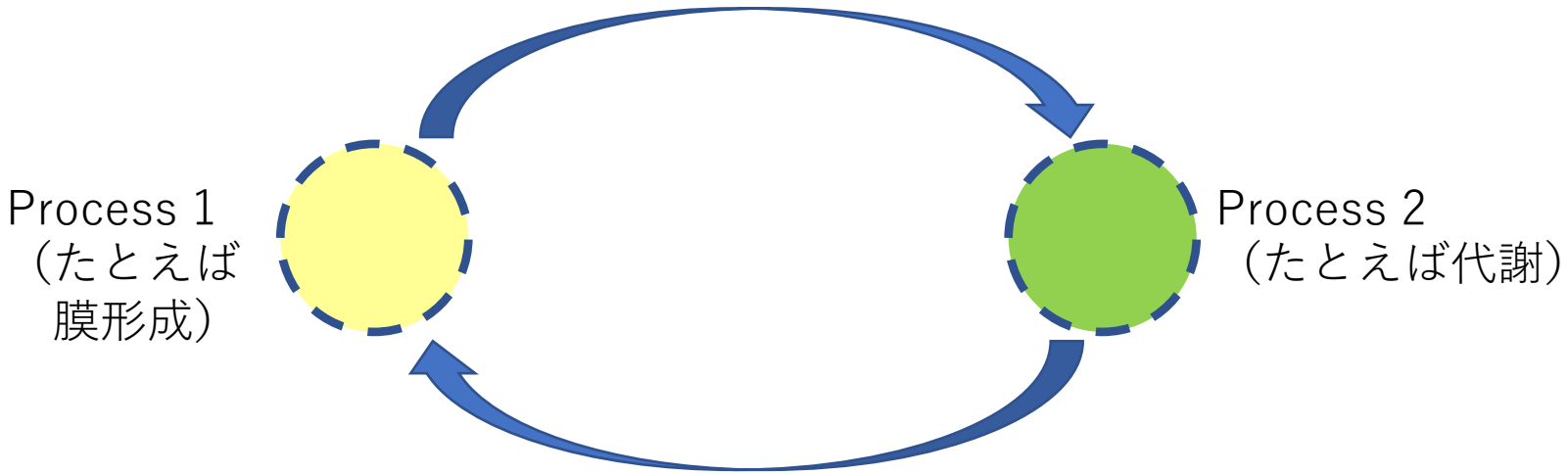
生命の不安定性

- そうした不安定性は、生命の不安定性に帰着する。
- 近年のエナクティヴィズムは、それを**生命の不安定さ・脆さ precariousness**と呼んで主題化している（Froese 2017; Weber & Varela, 2002; Di Paolo 2009）。
- 生命の核心に、崩壊していくプロセスがあり、それなしに生命はない。

- Froese, T. (2017). Life is Precious Because it is Precarious: Individuality, Mortality and the Problem of Meaning. In G. Dodig-Crnkovic & R. Giovagnoli (Eds.), *Representation and Reality in Humans, Other Living Organisms and Intelligent Machines*. Dordrecht: Springer, 33–50.
- Weber, A., & Varela, F. J. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenom Cog Sci*, 1(2), 97–125.
- Di Paolo, E. (2009). Extended Life. Topoi 28, 9-21.

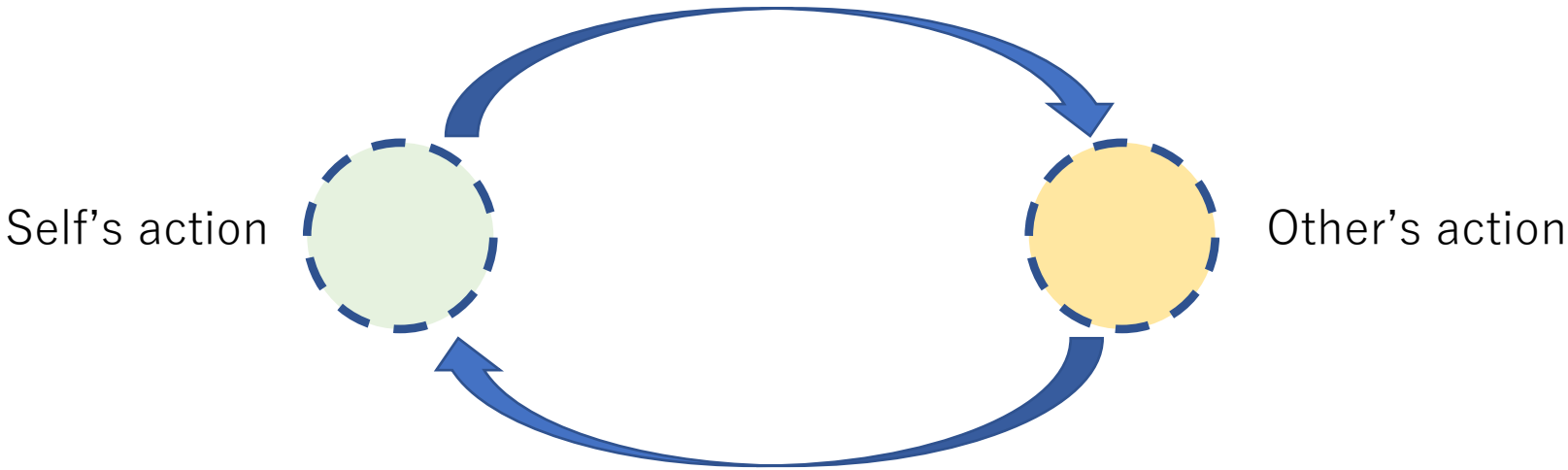
生命の核心にある「不安定性」

- 生命を構成する各プロセスは自然に崩壊する傾向
- どのプロセスが停止しても他のプロセスは消滅
- 互いに消滅を食い止め合って循環的に自己生成
- 一種の二重否定



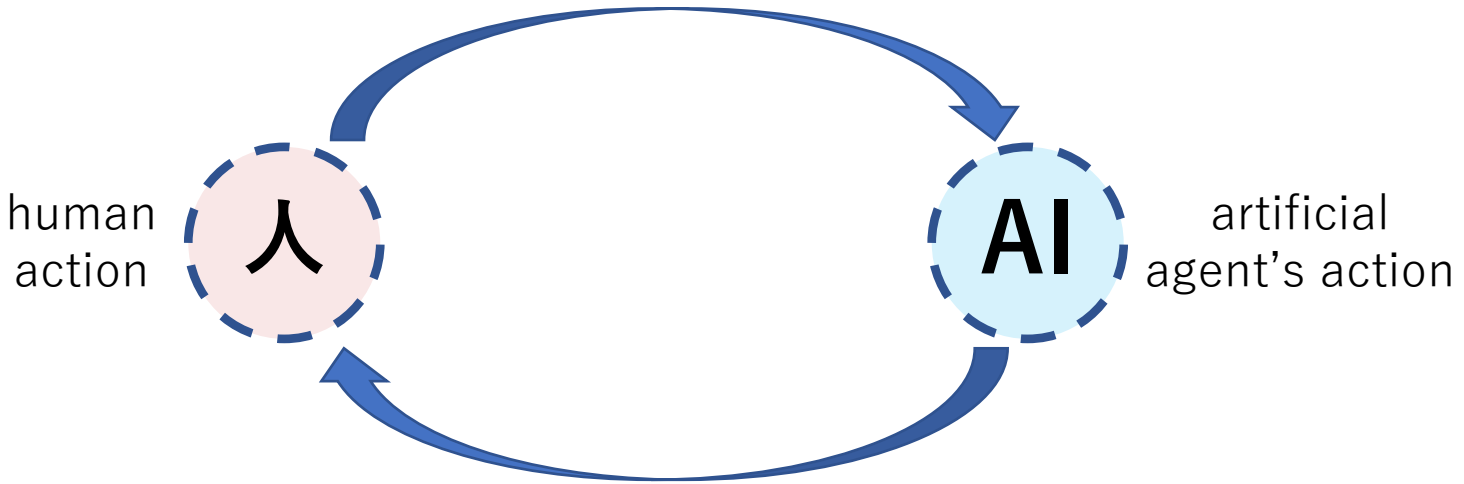
インタラクションにおける「不安定さ」

- 生命個体同士のインタラクションにも同じことが妥当
- 人間同士のコミュニケーションも。
- やはり不安定性を孕んだ相互依存的・相互生成的プロセス



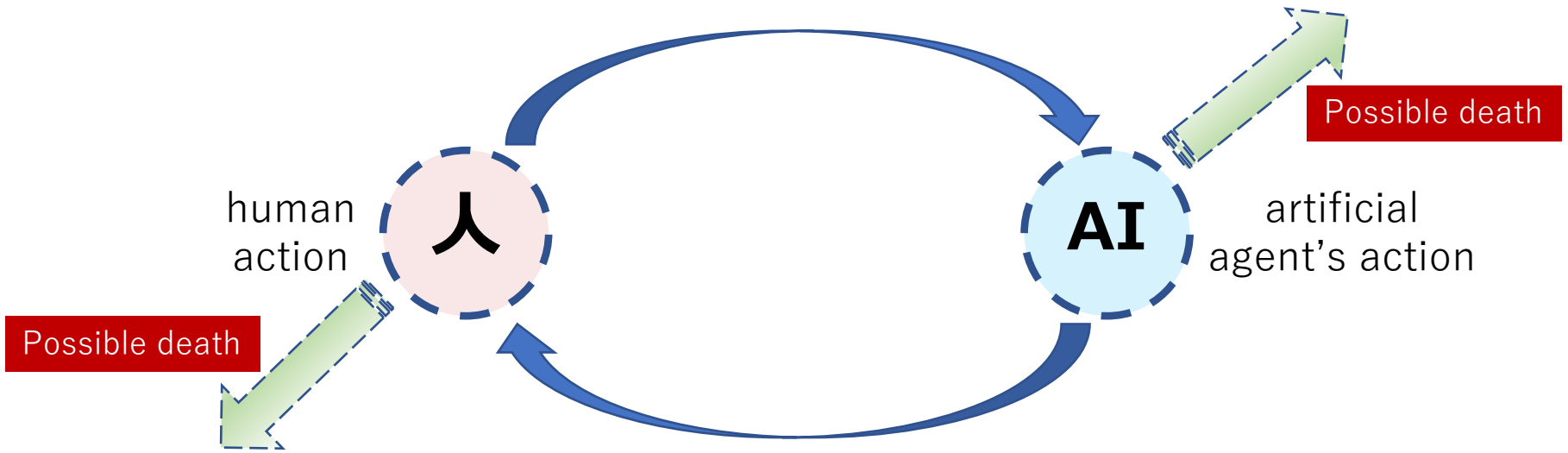
人工主体との共生と「不安定性」

- もし人間と真にコミュニケーションをとれるパートナーを求めるのであれば、人工的な主体にもそのような「不安定性」を採り入れていく必要があるのではないか。
- 同じ規則性に100%従うものではないが、ランダムでもない。



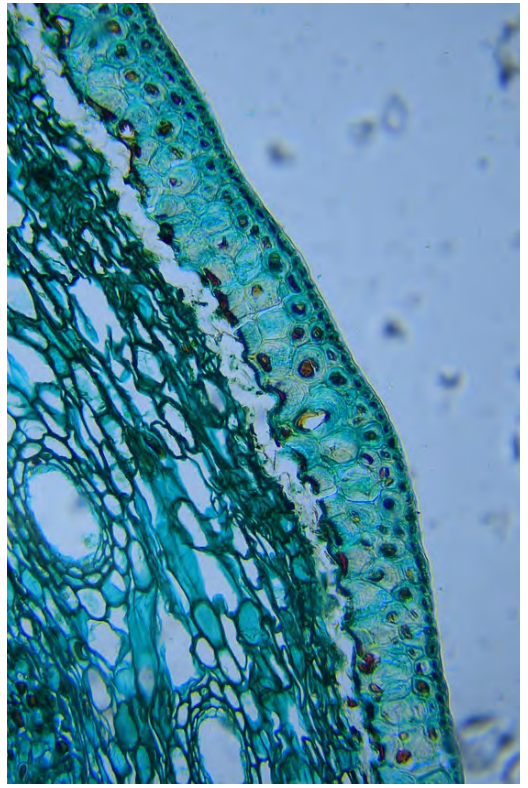
生死を賭けた行為

- 「状況を第三者的に分析して最適な行動をとる」というのではない（少なくともそれだけではない）。
- 世界と他者に向けて自己を賭ける？　真の「主体性」？
- 自己の消滅＝生死を賭けた行為？



傷つきやすい人工主体？

- ロボット・AIの「傷つきやすさ」
- 「ロボットがいかに壊れやすいか」という人もいるが、それはその存在者の本質にある脆さではない。それが存在することと、壊れることとは、相反する逆向きの方向性。



- これに対し、生命の脆さは、生命そのものの本質に含まれている。
- 生命が「ある」ということは、「不安定である」ということによってはじめて実現されている。それは「欠点」ではなく本質的な条件である。

AIに欠けているのは傷つきやすさ？

「死」が己れの存在に組み込まれていないのが現在のAIの弱み？

「生と死に根ざしたものでなければ、報酬は報われないし、損失が痛みを与えることもない。」(Paul 2021)

「傷つきやすさ［脆弱性］こそ、AIに欠けているものかもしれない。」(Paul 2021; see also Damasio 2021)

Paul, A. M. (2021). Vulnerable machines could be smarter ones. The Boston Globe, Oct. 8. https://www.bostonglobe.com/2021/10/08/opinion/vulnerable-machines-could-be-smarter-ones/

Damasio, A. (2021). *Feeling & Knowing*. New York: Pantheon.

傷つきやすいAI・ロボットは可能？

傷つきやすさ・不安定性によって本質的に規定されたAIやロボットを、われわれはまだ生み出せていないし、生み出せる見通しも立っていない。

しかし、「**人工生命**」（ALife）の分野で探究されてきたように、何かしら生命的なもの、生命と近しいものは作ることができる。

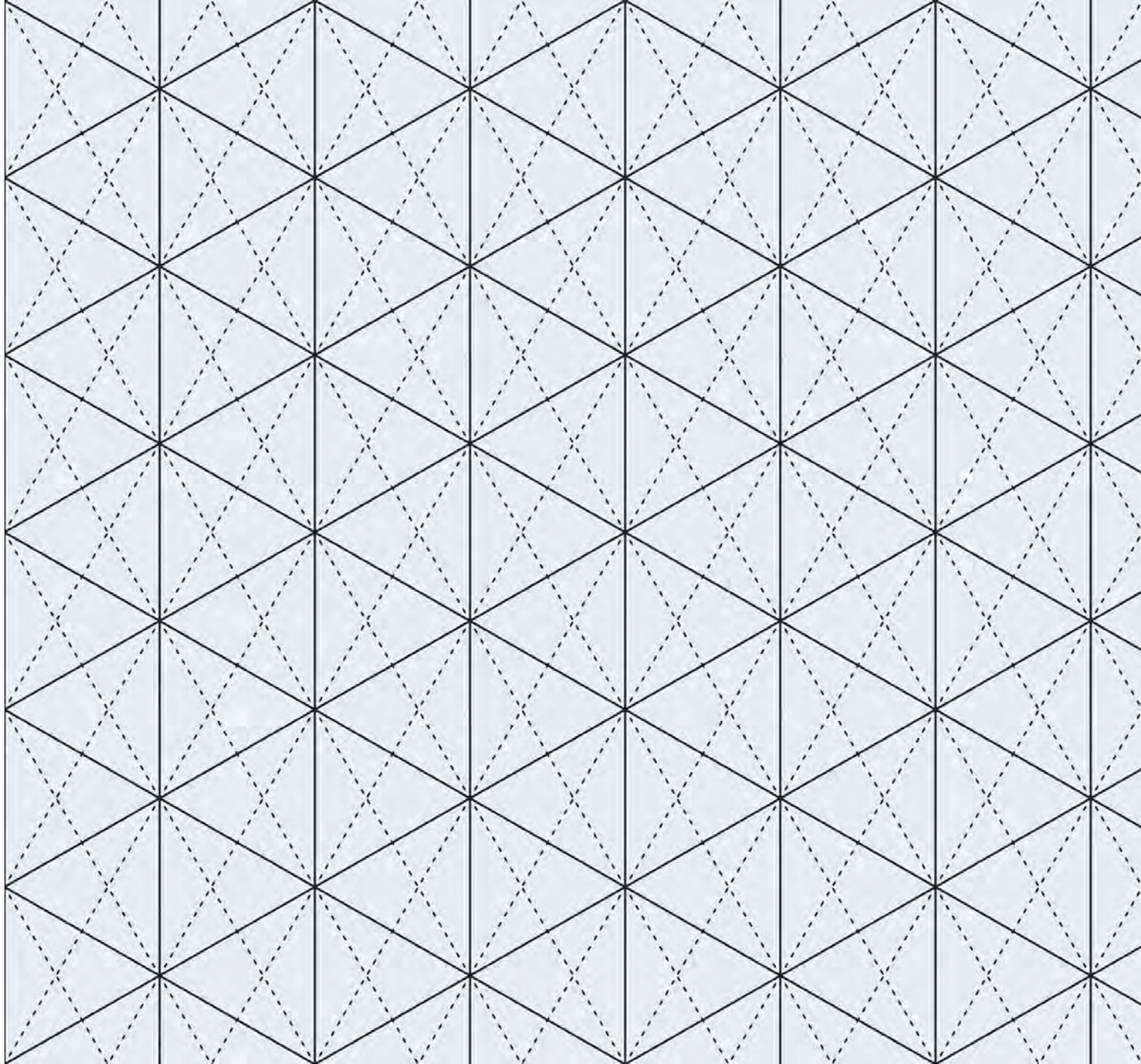
人工的な主体にも、そのような契機を採り入れていく必要があるのではないか（もし本当に人間とコミュニケーションできるパートナーを生み出そうとするのであれば）。

生命の不安定性と「いのちある」AI

- AIが生命の不安定性を取り込んだとき、AIはより本質的な意味で人間のパートナーになりうるのではないか。
- 「死にうる」AIが自らの生死を賭した決断をするとき、われわれはそうした決断を尊重せざるをえないし、そのような存在者を「いのちを持つ」存在者として尊重せざるをえない。
- 人工知能というより人工生命。

5. AIとの共生の未来

二種類の人間らしさ



二種類の「人間らしさ」

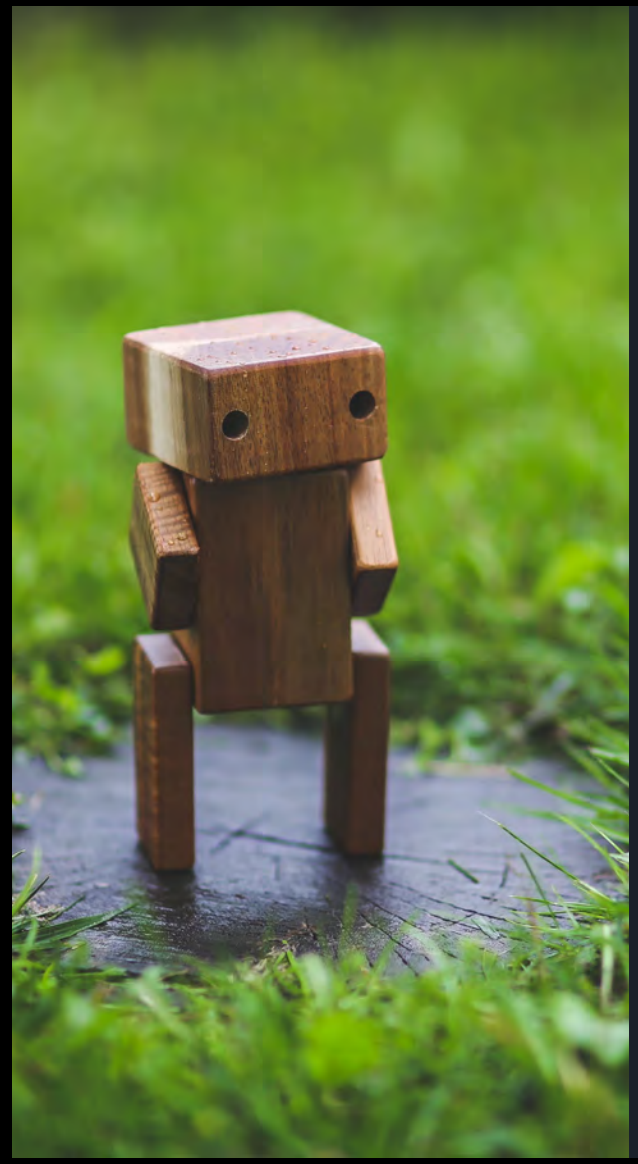
- ここまでの議論では二種類の「人間らしさ」が出てきた。
- 1）人間が受動的に反応しうる**身体的な振る舞いの人間らしさ**。これはかなりの程度AIが模倣できる。受動的な「反応」のレベル。だから、人間もAIに対し自動的に人間に対するような反応をしてしまう。
- 2）**生命的なレベルでの人間らしさ**。「死にうる」存在者として、生死を賭した行為を行うことができる。何らかの「死の自覚」が伴うなら、そこには「一度限りの決断」が行われうる。どういうことか？→

一度限りの決断

- 「一度限りの決断をする」というのは人間らしい振る舞い。
- それは、ある意味で「**偶然的**」だが、にもかかわらずそこには特別な意味が付与される。そこに自らの生命が賭けられているからである。
- われわれの一つ一つの小さな決断も、究極的には、自分の生命を維持するか、失敗して生命を失うか、という選択につながっている。
- それは「一般的な」判断ではなく、「いまここで生きている自分自身はどうなのか、ほかならぬこの身体をもつ自分はどうなのか」という個体的決断。

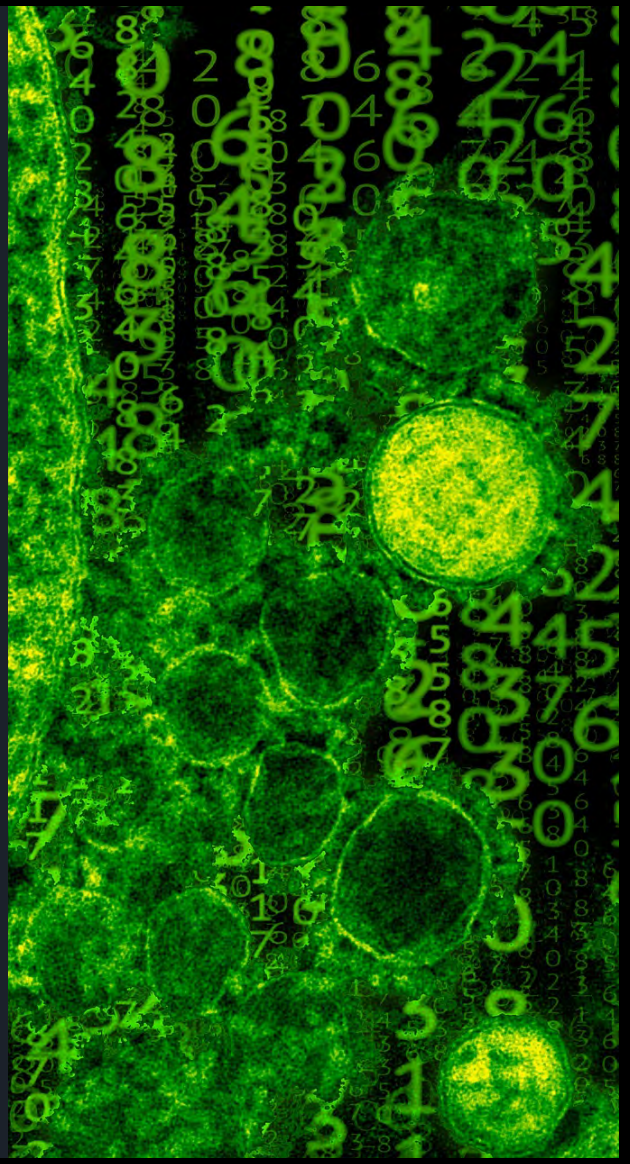
責任をとることのできる主体

- もしAIが人間的振る舞いを高度なレベルで模倣しうるようになったとしても、その判断が**「一般的」判断**にとどまるならば、AIは「責任をとる」ことはできないだろう。
- 「責任をとる」ことができるのは、**「一度限り」の決断**をしうる者だけではないか。誰にでも肩代わりされうるなら、それは「責任」とはいえない。肩代わりできない、「自分が担うしかない」のが責任である。
- AIが「責任を取りうる」存在者となったとき、AIは真に「信頼しうる」人間のパートナーになれるのかもしれない。



AIと人間の共存：二つの未来

- ここから、AIと人間が共存する未来を二つのパターンに分けてみたい。
- 1）人工生命的なAIが生まれない場合
- 2）人工生命的なAIが生まれた場合



①人工生命的なAIが生まれない場合

- AIとの間にエナクティヴなループが形成され、人間と同様にコミュニケーションをとれるとしても、人間の側は「相手は機械である」という認識を保持する（「見立て」の世界）。
- AIの地位も、当然のことながら従属的なものになる。
- **「生死を賭して」決断し「責任」をもつのは人間のみ。**AIの能力がいくら高まっても、人間には、統計的な推論ではなく、一度限りの「決断」を行うという役割が残る。
- （当分はこうした状況？）

②人工生命的なAIが生まれた場合

- AIを**人工的な主体**として認め、尊重せざるをえない。
- 異なった種類の主体として認め、その権利を認める。
- そこでは、AIたちもまた「決断」し「責任」をもつ主体となる。
- 社会的な合意と法的な整備が早急に必要となるだろう。
- その場合、人間は、「**非人間的な知性的存在者**」という新しい種類の同胞／パートナー／コンパニオンと、まさしく「共生」する時代に入る。

人間とAIの共存のあるべき姿

- どちらが望ましい未来かを単純に言うことはできない。
- ①道具的なAIは制御しやすいが、②自分で決断するAIは「何を考えているかわからない」。よい意味でも悪い意味でも「他者」。
- ①道具的なAIは使う人間によって「悪用」もされうるが、②人格的なAIは（適切に人格を陶冶すれば？）「自ら悪用を拒否する」かもしれない。

人間とAIの共存のあるべき姿

- 個人的な予想としては、人間は「自律的な人工主体」を生み出すところまでAI開発を推し進めずにはいられないと思う。
- それがいつになるかはわからない。
- しかし、AIとの人格的関係、それが帰結する倫理的諸問題や法的・社会的諸制度のあり方について、今から考えても決して早すぎることはないと思われる。