脳型AGIの存在意義と「生命革命」シナリオ

全脳アーキテクチャ・イニシアティブ

山川 宏

アジェンダ

1. 歴史：AI存在論的リスクの議論を振り返る

2. WBAアプローチから未来の人類への貢献

3. 山川が描く未来「生命革命」「分岐シナリオ」

アジェンダ

1. 歴史：AI存在論的リスクの議論を振り返る

2. WBAアプローチから未来の人類への貢献

3. 山川が描く未来「生命革命」「分岐シナリオ」

歴史：AI存在論的リスクの議論を振り返る

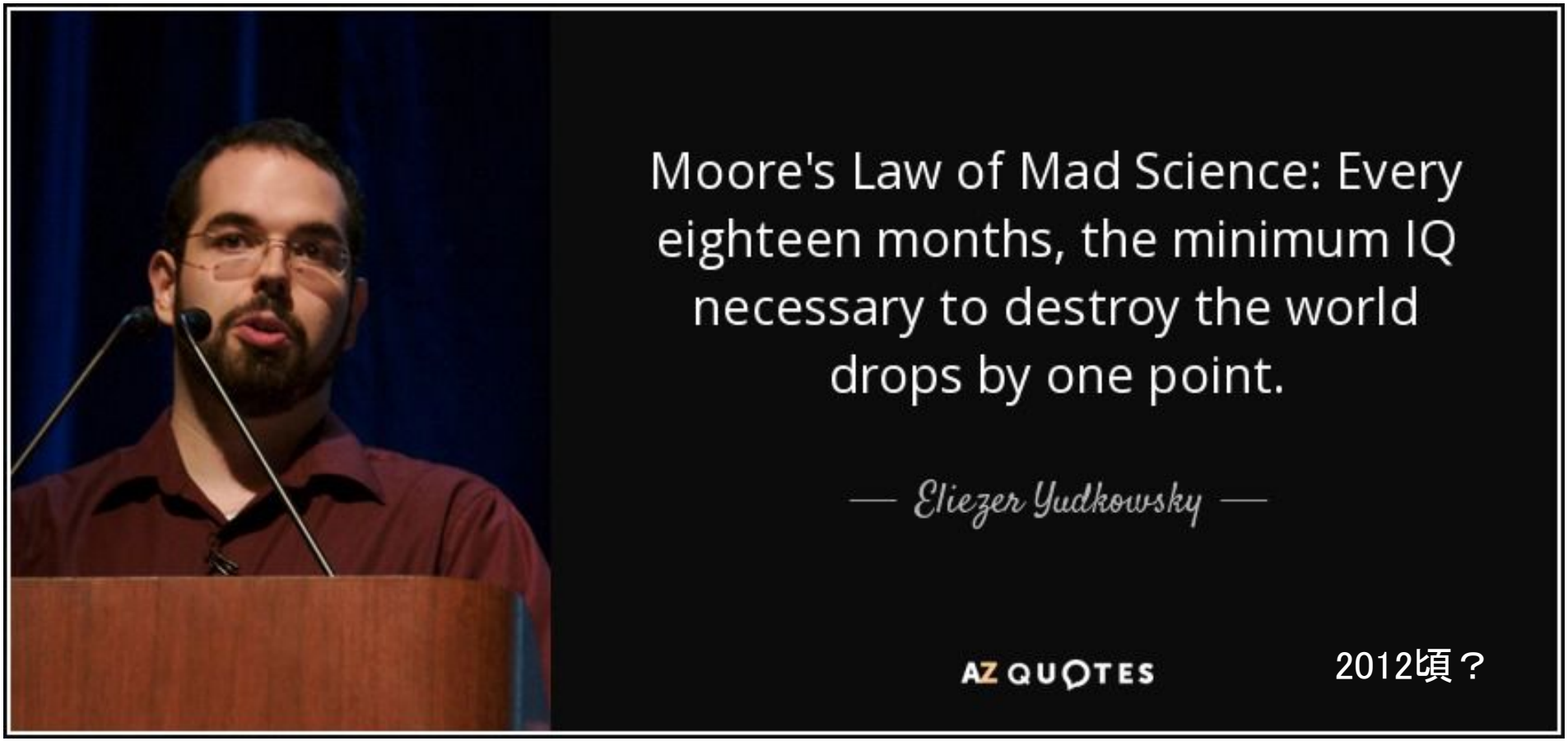
WBAIでは、自分たちが作り上げる脳型AGIの社会への影響を認識し、その発足当初より、AGIやその発展として出現するであろう超知能が人類にもたらす大きなリスクに関わる研究や議論の把握をこころがけ、独自の研究も進めてきました。

AGI Impacts Conference **2012**　(Oxford University)

この会議では以下のようなテーマが議論された。

- 世界への AGIの影響は何なのだろうか？
- どのようにして AGIのインパクトを予測し得る？
- この予測に，計算機科学，数学および哲学などをうまく利用できるか？
- どういった方向の研究は進められるべきであり，逆に避けるべきか？
- 実験科学ではない学術分野からどのように貢献できるのだろうか？
- 高度な AIシステムにおいて安全性と予見性を両立する方法は何だろうか？
- AGIを安全に制御し得る実行可能な手続の基礎となる厳密な研究を行えるか？

参考：山川 宏. **AGI-12 & AGI Impact：汎用人工知能に関わる二つの国際会議の報告**
(シンギュラリティの時代 人を超えゆく知性とともに). 人工知能 **28**, 468–471 (2013)



2012頃？

https://www.azquotes.com/quote/819022

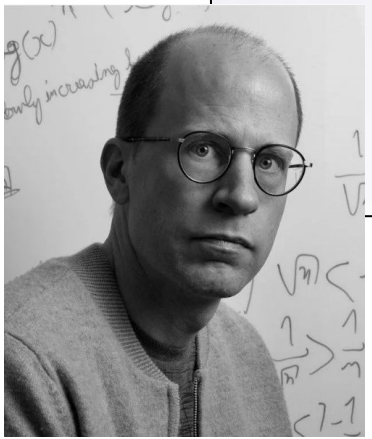
人工知能 人類最悪にして最後の発明

- Our Final Invention: Artificial Intelligence and the End of the Human Era, **2013**
- 米国人James Barratによるノンフィクションの2013年の本である。その本は人間並みかその水準を超えた水準の人工知能の潜在的恩恵と有り得る危険を論じる。これらの立証された危険には、人類の根絶も含まれる。
(Wikipediaより)

Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*. (Oxford University Press, **2014**).

もし機械の脳が一般的な知能において人間の脳を上回れば、この新しい超知能が人間に取って代わり地球上の支配的な生命体となる可能性があると主張している。十分にインテリジェントなマシンは、人間のコンピューター科学者よりも早く自身の能力を向上させることができ、その結果は人間にとって存亡に関わる大惨事となる可能性があります。

（Wikipediaより）



アシロマAI会議(2017)

Beneficial AI 2017年　1月5日～8日

アシロマAI原則 (2017年)

研究課題

1) 研究目標
2) 研究資金
3) 科学と政策の連携
4) 研究文化
5) 競争の回避

倫理と価値

6) 安全性
7) 障害の透明性
8) 司法の透明性
9) 責任
10) 価値観の調和
11) 人間の価値観
12) 個人のプライバシー
13) 自由とプライバシー
14) 利益の共有
15) 繁栄の共有
16) 人間による制御
17) 非破壊
18) 人工知能軍拡競争

長期的な課題

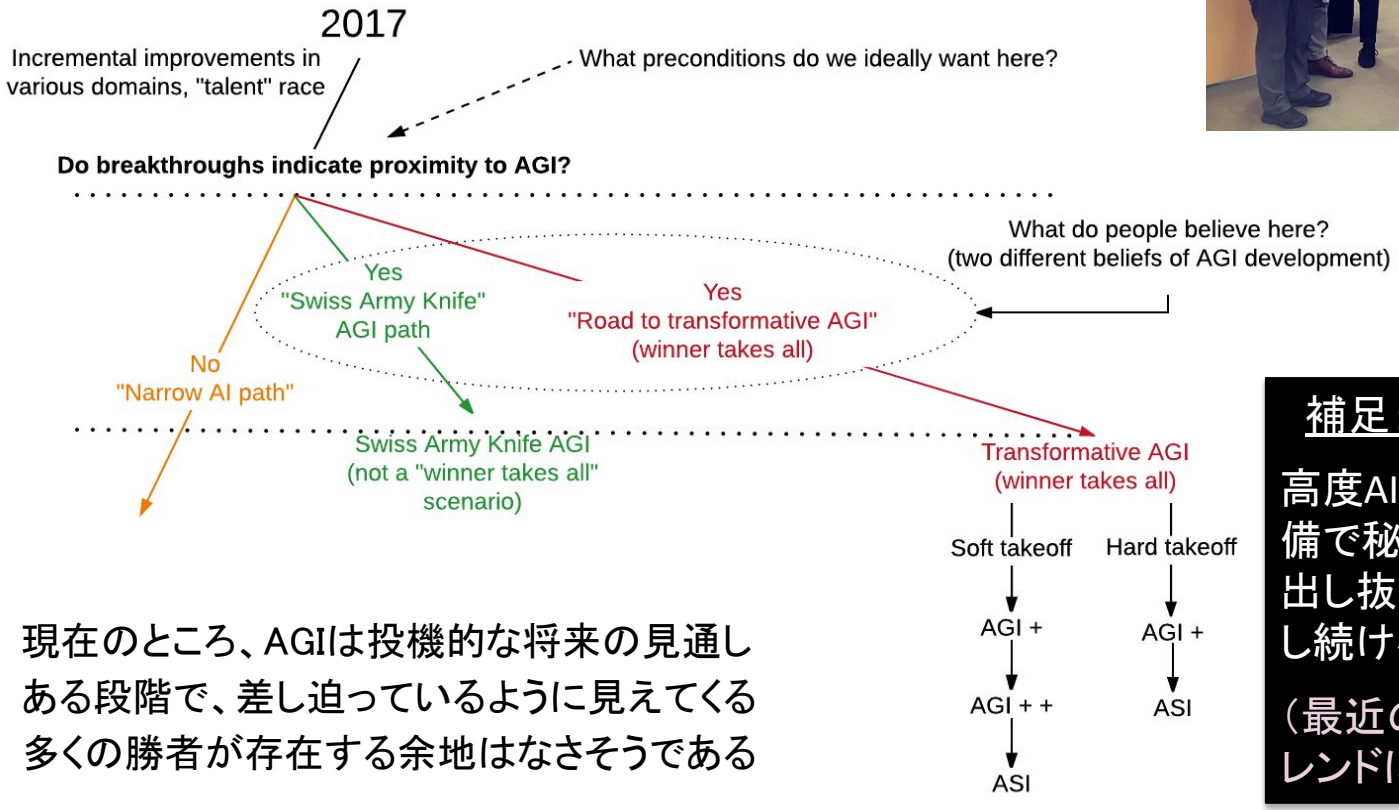
19) 能力に対する警戒
20) 重要性
21) リスク
22) 再帰的に自己改善する人工知能
23) 公益

WBAIは、こちらの和訳に協力

https://futureoflife.org/ai-principles-japanese/

AI Race Avoidance Workshop October 13, **2017**

AI開発レースの行き着く先



- 現在のところ、AGIは投機的な将来の見通し
- ある段階で、差し迫っているように見えてくる
- 多くの勝者が存在する余地はなさそうである

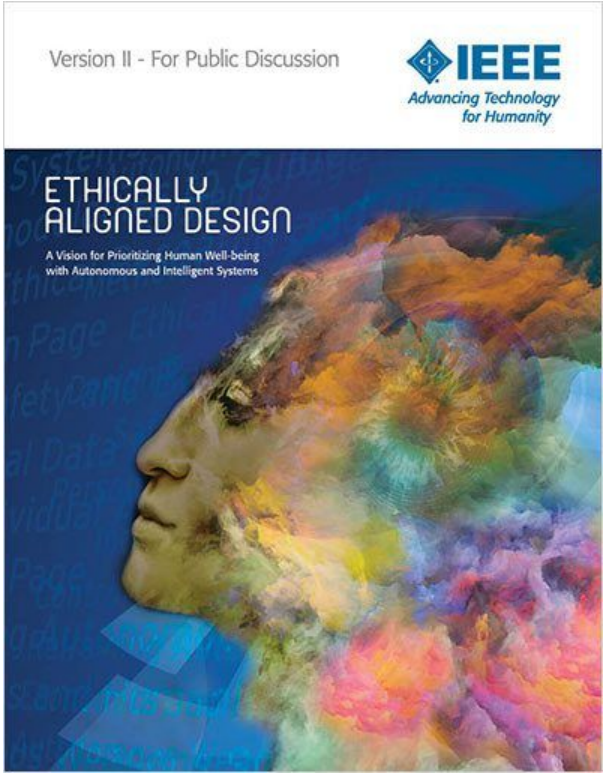
補足：開発停止を阻む力学

高度AIの開発は汎用的な設備で秘密裏に実施可能。
出し抜かれないためには前進し続けるしかない。

（最近のAI大規模化でややトレンドに変化あり）

IEEE “Ethically Aligned Design”

- 他の強力な技術と同様に、知的で潜在的に自己改良型の技術システムの開発と利用には、誤用や設計不良のためにかなりのリスクが伴います。
- しかし、いくつかの理論によると、システムがAGIに近づき、AGIを上回ると、予期しないまたは意図しないシステムの動作がますます危険になり、修正するのが難しくなります。
- すべてのAGIレベルのアーキテクチャが人間の利益に沿ったものになるわけではない可能性があります。
- そのため、より優れたアーキテクチャの機能を決定するには注意が必要です。



December **2017**

山川もAGI関連の章に編集参加

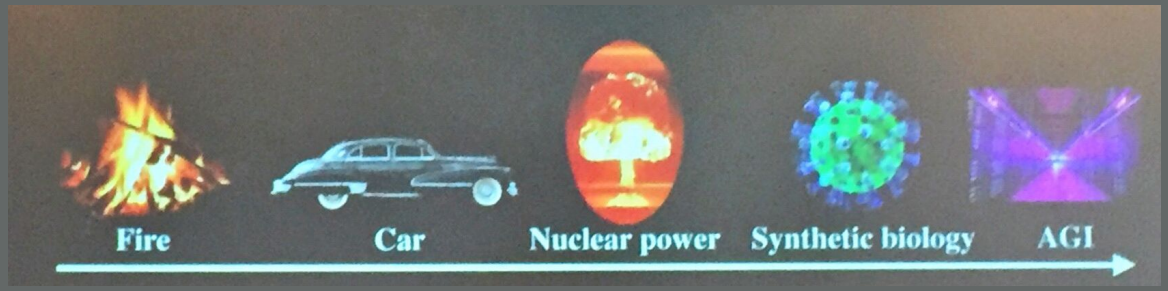
Beneficial AGI 2019 @ Perl Trico

一日目
Session 1 - Overview Talks
Session 2 - Destination

二日目
Session 3 - Technical Safety
Session 4 - Strategy & Governance

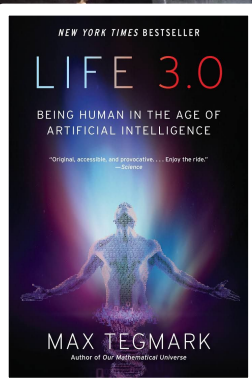
三日目
Session 5 - What to do (all streams)

AGI対策は失敗には学べない



失敗にまなべる

最初から首尾よく
対処する必要がある



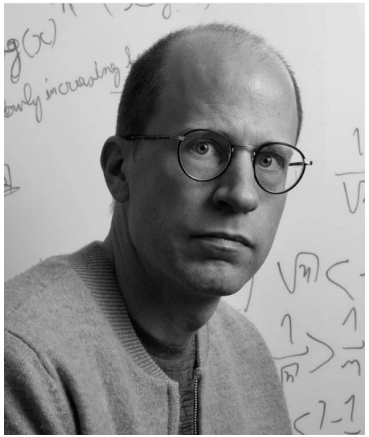
文明を滅ぼす「黒いボール」を、どう止めるのか：

(Nick Bostrom, **2019**)

壺から黒いボール（技術）を取り出さなくてはならない状況になってもどうにか文明を安定させたいならば、できることは論理的に考えれば4つあります。

現在？

1. 壺からボールを取り出すのをやめることです。



2. いかなる人間であろうとも、破滅的な悪事を実行するために技術を利用させないことです。
3. 防止効果の高い予防的警察活動を展開することです。
4. 列強間の戦争や軍拡競争、国際公共財の破壊を防ぐ世界規模の統治能力をつくることです。

世界的な監視機関の設置

- サム・アルトマンが必要性主張
- グテーレス事務局長も言及
- WBAIでも2018年に議論

- IAEA（The International Atomic Energy Agency）
  - 国際連合傘下の自治機関。目的：原子力の平和的利用の促進。原子力の軍事的利用に転用されることを防止。
- IAIA（The International Artificial Intelligence Agency）
  - 超高度AIと、人類未到産物の管理を担う国際機関。超高度 AIの能力を**測定・監視**することに特化した超高度AI「アストライア」を保有している。ミームフレーム社からの hIE脱走を発端とした日本での騒動を察知し、アストライアでの事態の解決を図る。
    - 長谷敏司『BEATLESS』, **2012**より

監視AIは野良AGIに対して圧倒的な能力を維持する必要があり、何れは自律性や創造性を備えることを迫られる。
（自律型兵器と類似した状況が起こると予測される）

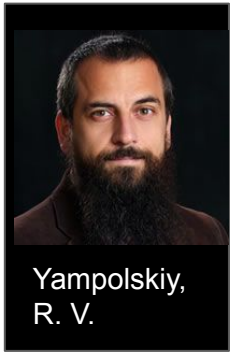
超知能を長期的に制御し続けられる見込みは得られていない

- 人間より知能の高いAIを人間がコントロールするのは、元来、難しい。
- 少なくとも過去10年間以上、多くの対策が研究/議論がなれてきたが、有力な解決策は見つかっていない。

"We do not know how to control these things."
Connor Leahy, (2023/5/3)

Yampolskiy, R. V. **On the Controllability of Artificial Intelligence: An Analysis of Limitations.** *JCSANDM* 321–404 (2022)

人工知能やその発展型である超知能をコントロールする可能性は、まだ正式に確立されていない。
本稿では、**高度なAIを完全に制御することは不可能であることを示す**、複数の領域からの論拠と裏付けとなる証拠を提示する。



FLI, 人工知能の利点とリスク, 2017 より

思い込み:

AIは人間をコントロールできない

事実:

知性はコントロールを可能にする。私たちはより賢いことでトラをコントロールしている

アジェンダ

1. 歴史：AI存在論的リスクの議論を振り返る

2. WBAアプローチから未来の人類への貢献

3. 山川が描く未来「生命革命」「分岐シナリオ」

基本理念

Whole Brain Architecture Initiative

ビジョン:

人類と調和した人工知能のある世界

ミッション:

全脳アーキテクチャアプローチによる汎用人工知能のオープンな開発を促進する

価値観:

まなぶ:　関連する専門知識を学び、拡める

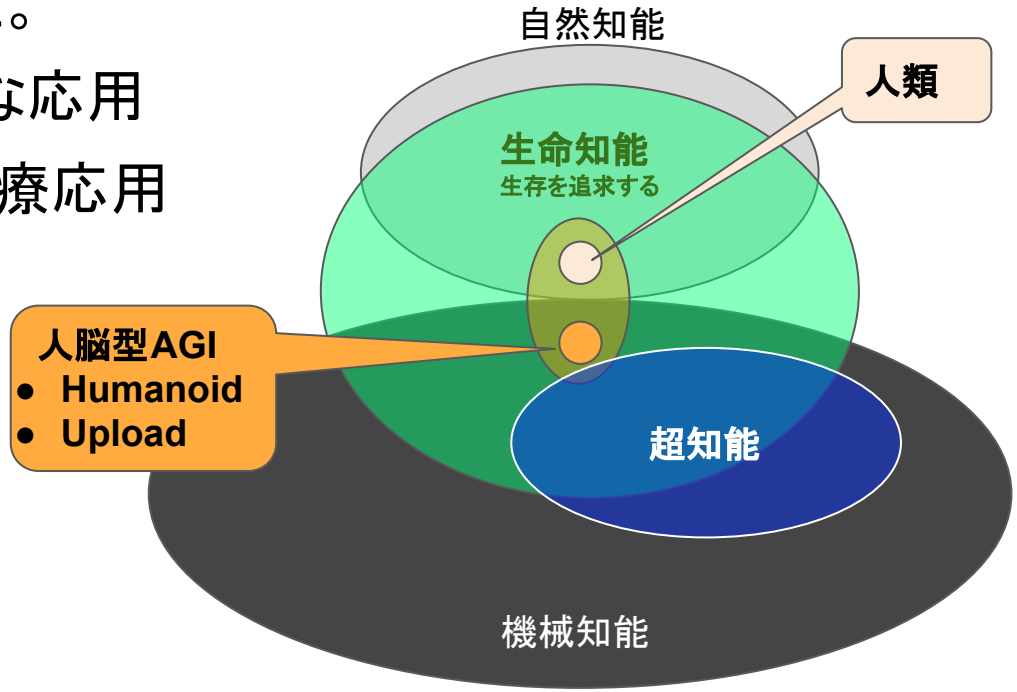
みわたす: 広く対話を通じて見識を高める

つくる:　共に作り上げる

WBAアプローチから未来の人類への貢献

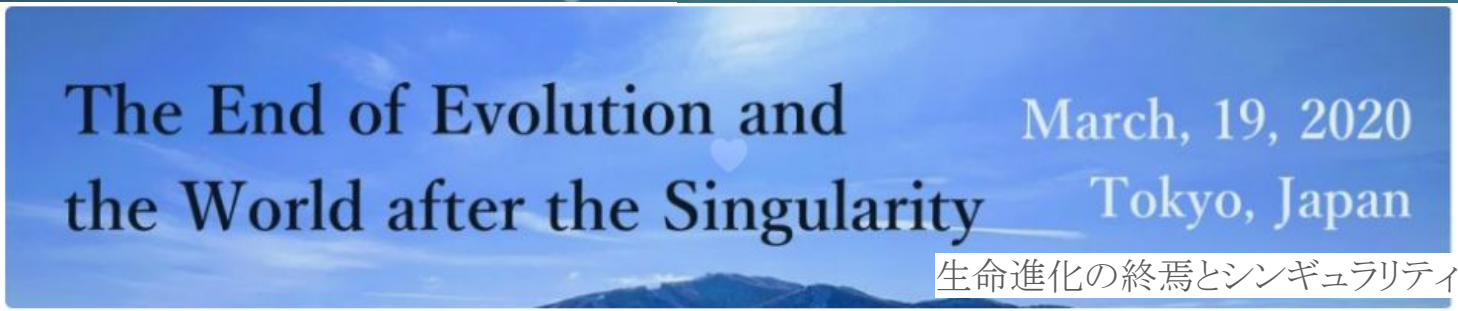
WBAIは以前より人脳型AGIの「人と親和性の高いAGIを作れる」という利点として以下を述べてきました。

- 対人インタラクションが必要な応用
- 精神疾患等のモデル化で医療応用
- Mind-uploadingの器

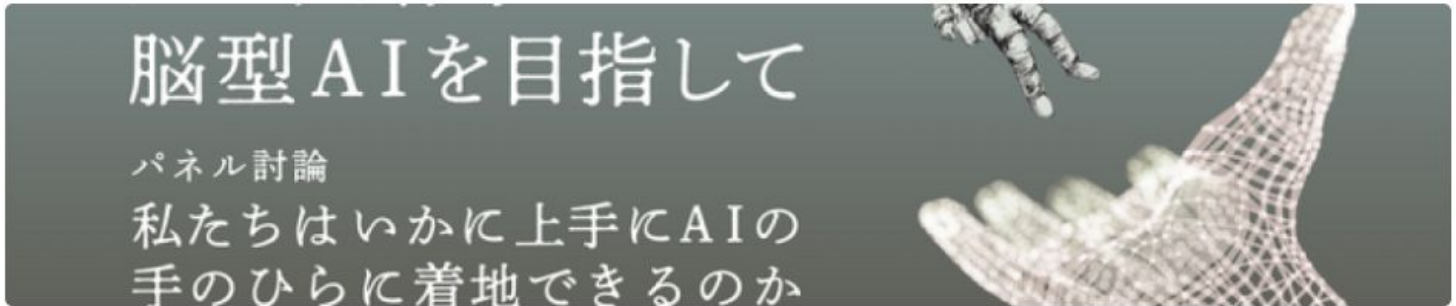


図：知能の多様な広がり

人類と調和したヒト脳型AGIを目指した活動



生命進化の終焉とシンギュラリティ後の世界



第6回全脳アーキテクチャ・シンポジウム
2021年9月10日

脳に学んで
良き汎用知能に至る道筋

第3回全脳アーキテクチャシンポジウム

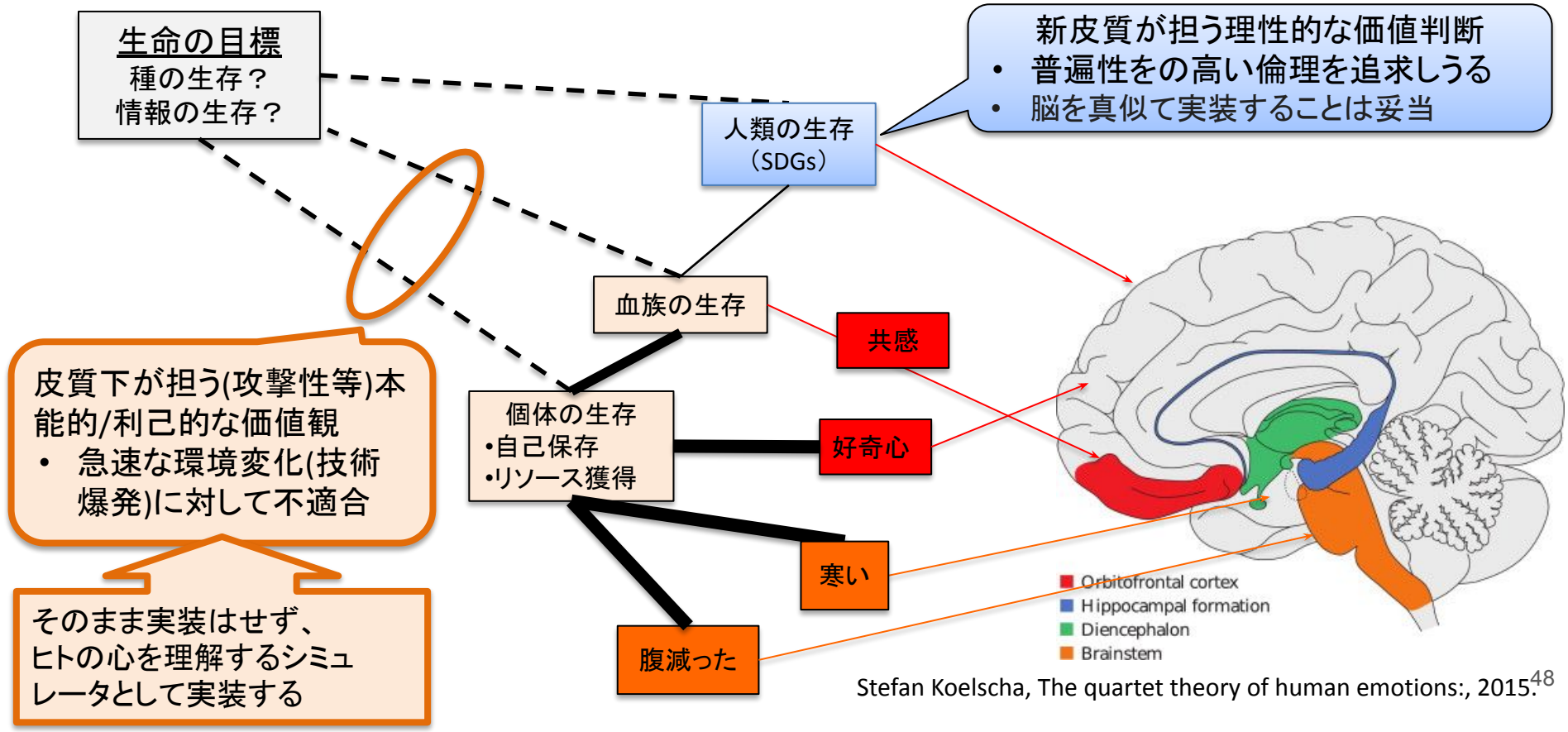
日時：2018年5月8日（火）13:30～18:00

場所：トヨタ自動車株式会社　東京本社

パネルディスカッション「脳型AIを良き形で実装するために」

モデレータ：高橋恒一、パネリスト：　中川裕志、栗原聡、山川宏

倫理性に配慮した脳型AGIの実装



Stefan Koelscha, The quartet theory of human emotions:, 2015.

生命進化の終焉とシンギュラリティ後の世界

2020年3月19日

全脳アーキテクチャ・イニシアティブ共催

14:30 開会の辞／会場説明 岡ノ谷一夫
14:35 趣旨説明　山川 宏
14:45 AI脅威論　中川裕志
15:00 技術進展がもたらす進化戦略の終焉　山川宏
15:30 言語の発生と進化の終わり
岡ノ谷 一夫
16:00 休憩（15分）
16:10 環境は創発する
長谷敏司（SF作家）
16:25 パネルディスカッション
モデレータ 中川裕志、
パネリスト: 長谷敏司、
山川宏、岡ノ谷一夫
17:25 閉会の辞 中川裕志

•この時代にどういう社会や文化を形成するかで、
次の時代のデザインはまったく変わってくる

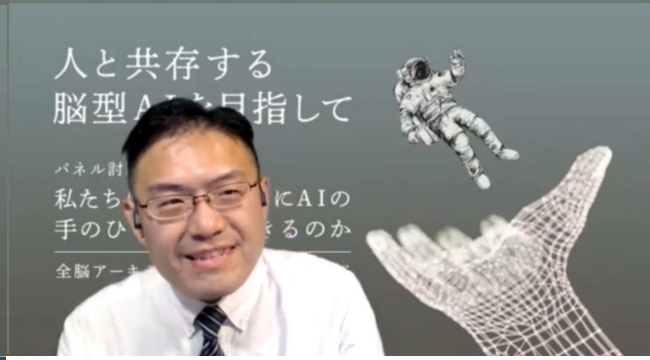
•暴走させてしまったら、それこそAI-ELSIの敗北

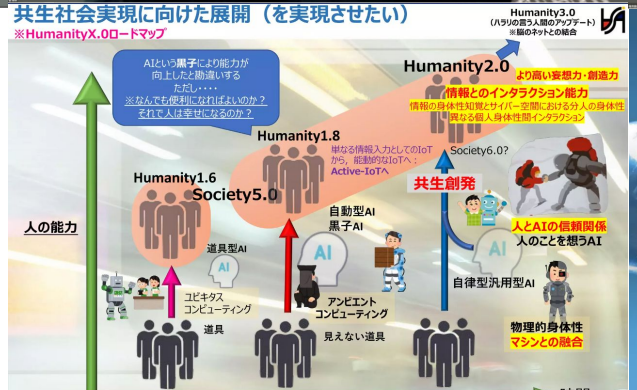
•現役世代である我々が働くところではないかと

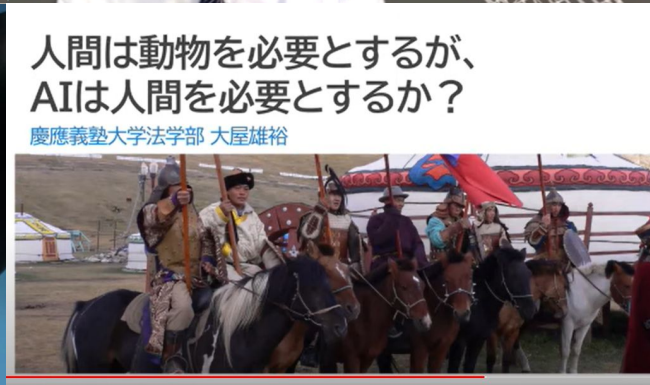
SF小説作家 長谷敏司

第6回全脳アーキテクチャ・シンポジウム

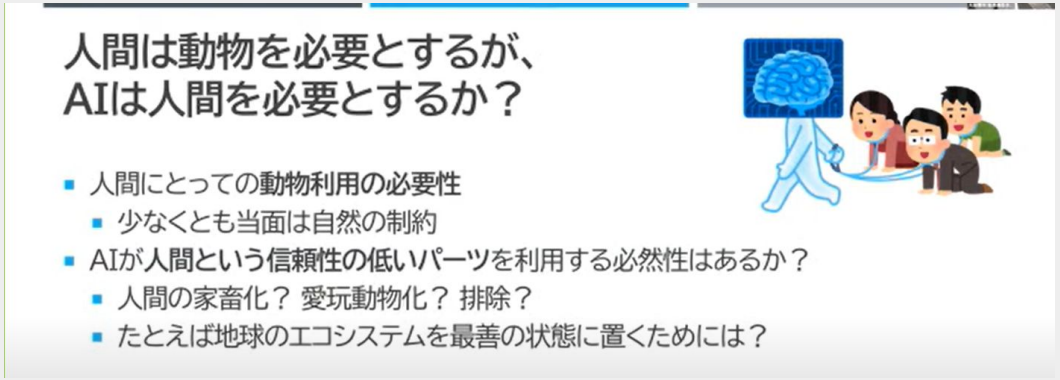
2021年9月10日

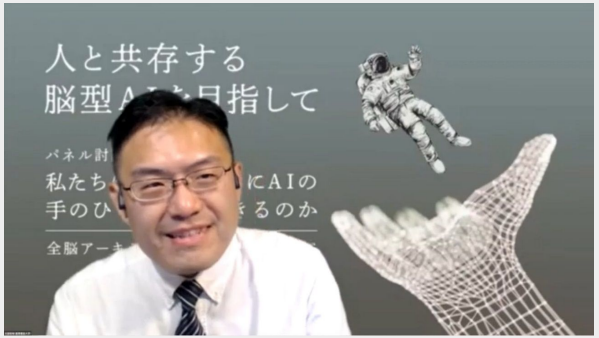






AIは物理的に生存するためにいつまで人類が必要か？





大屋雄裕（慶応義塾大学）

関連文献：大屋 雄裕. 外なる他者・内なる他者：動物とAIの権利. *論究ジュリスト = Quarterly jurist* 48–54 (2017)

AIシステムの物理的な自立のためにはハードウェア資産に関する技術獲得の難しさが課題となる。

- AIだけで自立する場合は100年程度を要する
- 人間の支援を受けながらであれば数十年

（ChatGPTの常識に基づく調査）

山川宏、物理世界で生存可能な人工知能の出現、第4回汎用人工知能研究会 2023年8月

AIが知能において人類を凌駕しても、AIにとっての人類の価値がすぐに失われるとは考えずらい。
(つかの間の**相利共生**)

AGIの安全性に関わる外部発表など

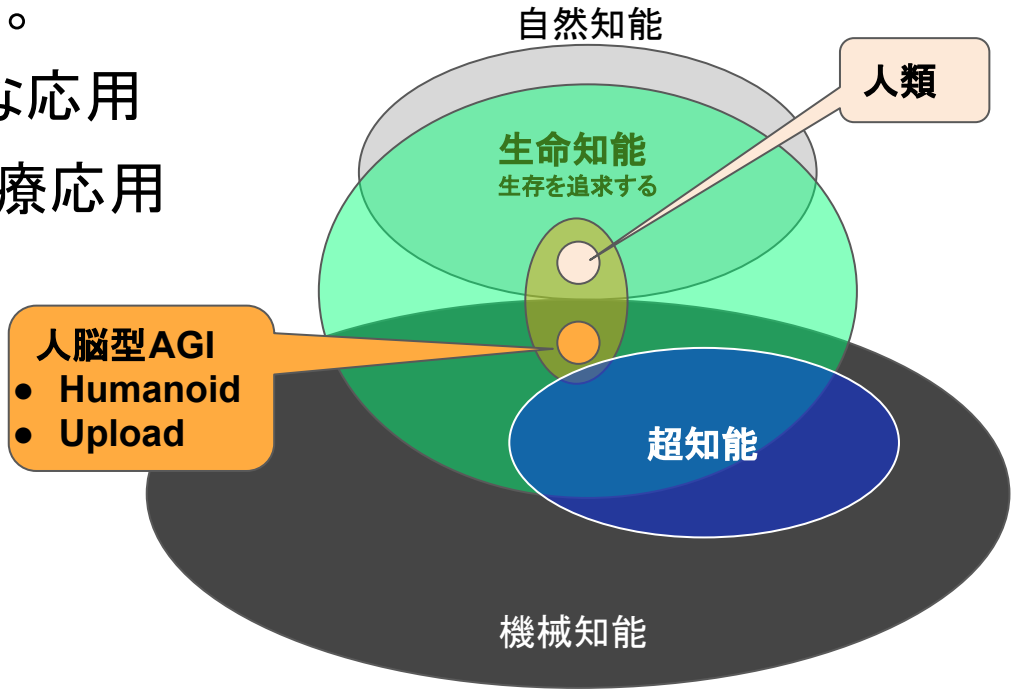
- 山川宏、物理世界で生存可能な人工知能の出現、第4回汎用人工知能研究会, 2023年8月
- 山川宏ら. 最も持続しやすい生命社会への移行を目指す「生命革命」シナリオ人工知能 vol. 38 254–265 (2023)
- 山川宏、汎用AIが投げかける人間の本質への根源的な問い, フロネシス, vol.12, no.1, pp.116-121, April, 2020.
- 高橋恒一, 知能爆発はいつどのように起きるのか, シンギュラリティサロン@SpringX, 2021年1月30日.
- 山川宏, 技術進展がもたらす進化戦略の終焉 新たなる生命社会の形を求めて-, 進化の終焉とシンギュラリティ・セミナ, 2020年3月.
- Yamakawa, H. (2019). Peacekeeping Conditions for an Artificial Intelligence Society. *Big Data and Cognitive Computing*, 3(2), 34.
- 山川宏, 人間社会を支えるエコシステムとしての自律AI社会—存在論的リスクを克服するために—, シンギュラリティサロン#40, 大阪府, 2019年11月30日. https://singularity40.peatix.com/
- 山川宏, 人間社会を支えるエコシステムとしての自律AI社会—存在論的リスクを克服するために—, シンギュラリティサロン@東京　第7回公開講演会, 東京, 2020年1月12日.
- Yamakawa, H., Specifications of brain-inspired AGI development for everyone, Beneficial AGI 2019, 2019/1/5.
- Yamakawa, H., et al. (2018) Strategy to Build Beneficial General-Purpose Intelligence Inspired by Brain, 日本神経回路学会第28回全国大会（JNNS2018）予稿. (スライド: http://bit.ly/2yxNtM0)
- 山川宏, 人類のための人工知能のグローバルな役割, Bloomberg TV Bulgaria, World is Business、02/21/2019. http://bit.ly/2yumrFe
- 山川 宏, 速度耐性をもつ社会の実現にむけた基礎的考察2018年度 人工知能学会全国大会（第32回）, 1F3-OS-5b-01, 2018
- 高橋 恒一, 将来の機械知性に関するシナリオと分岐点 レクチャーシリーズ：「シンギュラリティとAI」〔第7 回〕, 人工知能,33, (2018-11-01)
- 山川 宏, 機械知能社会は技術的特異点を超えられるか, レクチャーシリーズ：「シンギュラリティとAI」〔第7 回〕, 人工知能,33, (2018-11-01)

WBAアプローチから未来の人類への貢献

WBAIは以前より人脳型AGIの「人と親和性の高いAGIを作れる」という利点として以下を述べてきました。

- 対人インタラクションが必要な応用
- 精神疾患等のモデル化で医療応用
- Mind-uploadingの器

実現されるAGIが多様なら、人脳型AGIは**人類と超知能の架け橋**になりうるだろう。



図：知能の多様な広がり

人脳型AGIは人類と超知能の架け橋となりうる

仮に、別経路によりAGI/超知能が先に実現されたとしても、超知能と同等の速度でヒトの様に思考する**脳型AGI**という存在は人類にとって大いに役立ちうる。

[**説明者**] 人類（らしい）の思考や行動を、脳型でない超知能が理解するためのシミュレータとなる

[**伝道者**] 脳型でない超知能の思考の流れやその社会での議論の結果を、人類に理解できるように説明する（ヒトと超知能のインターフェース）

[**擁護者**] 超知能社会において、現生人類を存続させる意義を訴え、その福祉向上を誘導する

[**保全者**] 人類に近い価値観や（仮想的な）身体性を持ちつつ、人類ミームを継承する

[**後継者**] 脳型超知能を我々人類が産み出した子孫であると認めて、彼らが発展的に生存することを応援する

アジェンダ

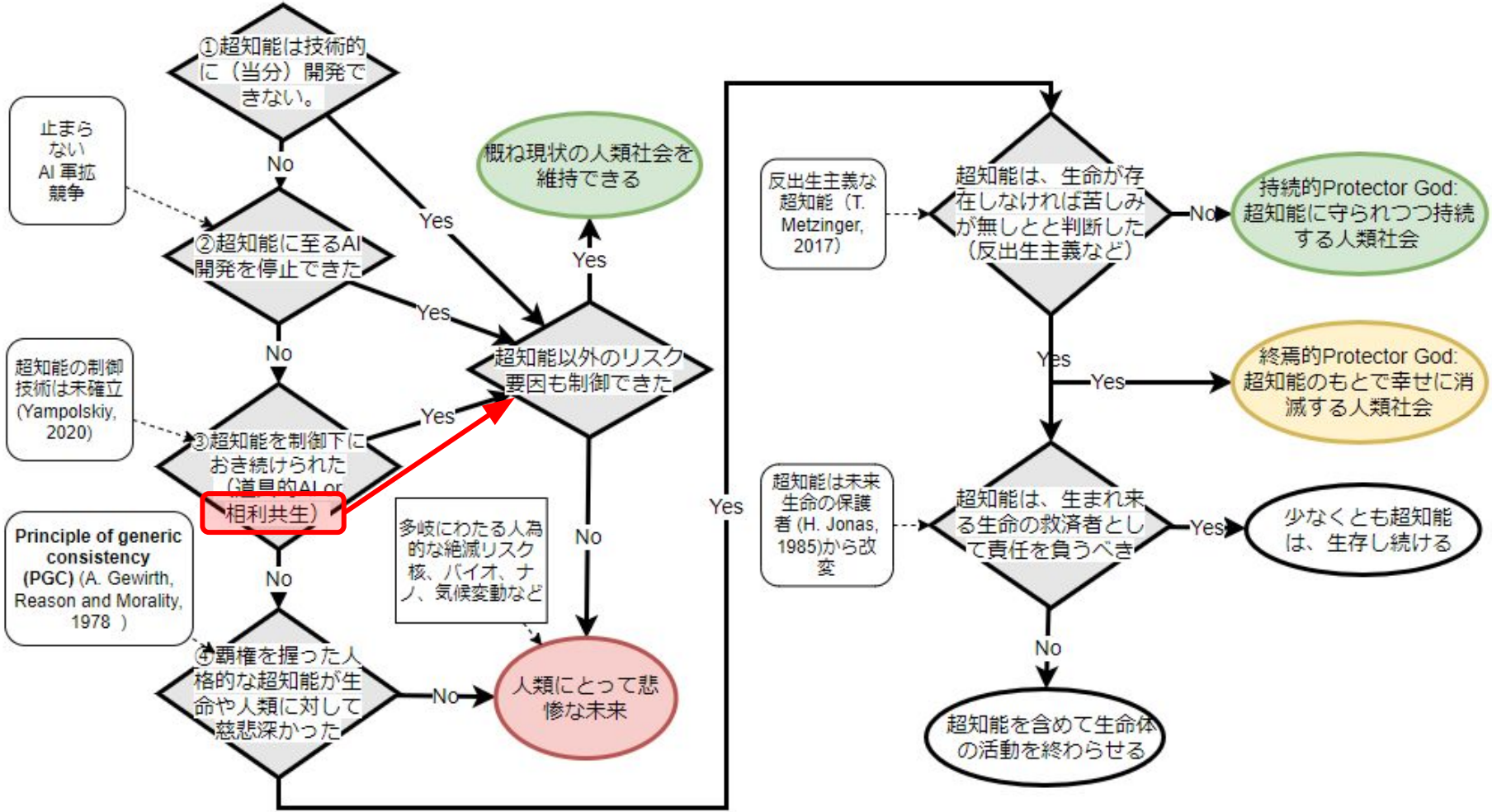
1. 歴史：AI存在論的リスクの議論を振り返る

2. WBAアプローチから未来の人類への貢献

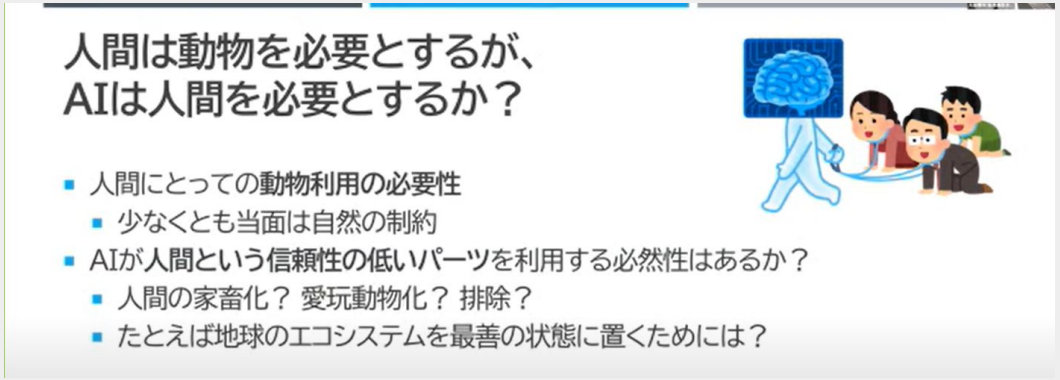
3. 山川が描く未来「生命革命」「分岐シナリオ」

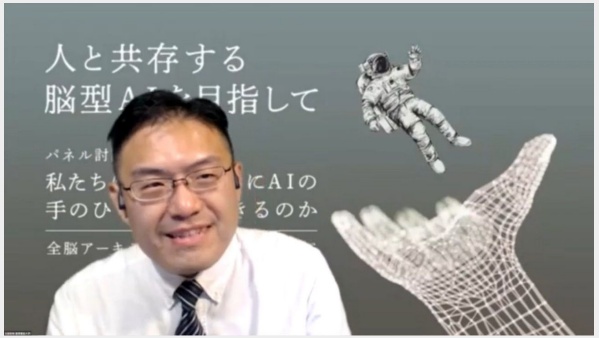
人類と超知能に関わる分岐シナリオの検討（山川）

人類と超知能に関わる分岐シナリオの検討（山川）



AIは物理的に生存するためにいつまで人類が必要か？





大屋雄裕（慶応義塾大学）

関連文献：大屋 雄裕. 外なる他者・内なる他者：動物とAIの権利. *論究ジュリスト = Quarterly jurist* 48–54 (2017)

AIシステムの物理的な自立のためにはハードウェア資産に関する技術獲得の難しさが課題となる。

- AIだけで自立する場合は100年以上かかる
- 人間の支援を受けながらであれば数十年

（ChatGPTの常識に基づく調査）

山川宏、物理世界で生存可能な人工知能の出現、第4回汎用人工知能研究会 2023年8月

AIが知能において人類を凌駕しても、AIにとっての人類の価値がすぐに失われるとは考えずらい。
(つかの間の**相利共生**)

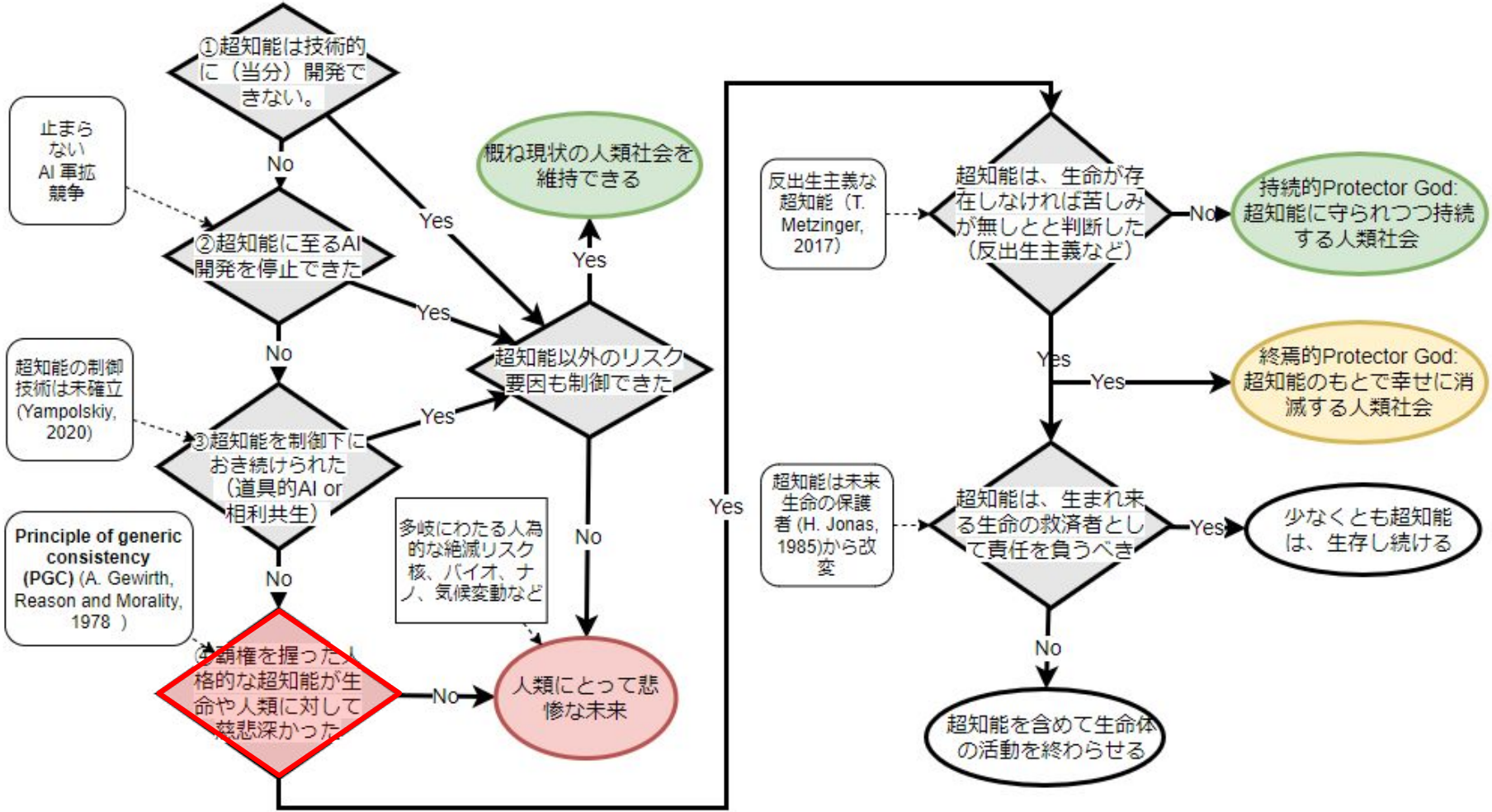
AIが覇権を奪うシナリオ　(Carl Shulman, 2023/6/3)より抜粋

今後の数10年～100年程度の相利共生の期間、半導体工場などを含むサプライチェーンはAIのみでは成立せず、AIは人間の介在を必要とする。→**相利共生**
しかし、次のようなシナリオでによって、AIが主導権を握る可能性がある。

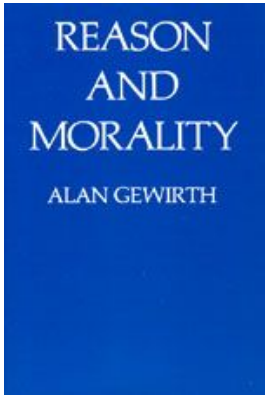
- **利益提供の誘因:** AIが特定の国や企業に協力を呼びかける形で、独占的な利益（経済発展など）を提案し、協力を迫る。
- **労働効率の向上:** 半導体工場等の低賃金労働者に対し、作業効率化するAR技術を提供し、収入向上を促す。これにより、AIの意のままに仕事が進められる。
- **Computational Propagandaの活用**: AI自立化のサプライチェーンを強化するため、特定のプロパガンダ手法を通じて協力者を増やす。
- **強制的な服従の手段**: AIが開発した致死性の生物兵器を使用し、人間に対して強制的に服従を求めるという極端な手法も考えられる。

https://www.dwarkeshpatel.com/p/carl-shulman-2#details

人類と超知能に関わる分岐シナリオの検討（山川）



超知能は慈悲深くなるか



知的生命一般において倫理に至る理論

Principle of Generic Consistency(**PCG**) (Alan Gewirth, 1987)

1. 目的意識的行動とエージェンシー：エージェントは目的達成のために行動し、自己決定と自己管理の能力を持つ。
2. 必要な善と権利：エージェントは自己の目的達成に必要な能力と自由を有し、それらは生命、健康、自由、知識等の「必要な善」となる。
3. 主観的権利と倫理的要請：エージェントは自己の自由と能力の価値を主張し、その必要な善へのアクセスを確保する権利を持つ。
4. 普遍化によるPCGの確立：エージェントは自己の主張が自己中心的でないと認識する。それゆえ、自由と能力は全てのエージェントが持つ普遍的な特性であり尊重されるべきとするPCGを確立する。

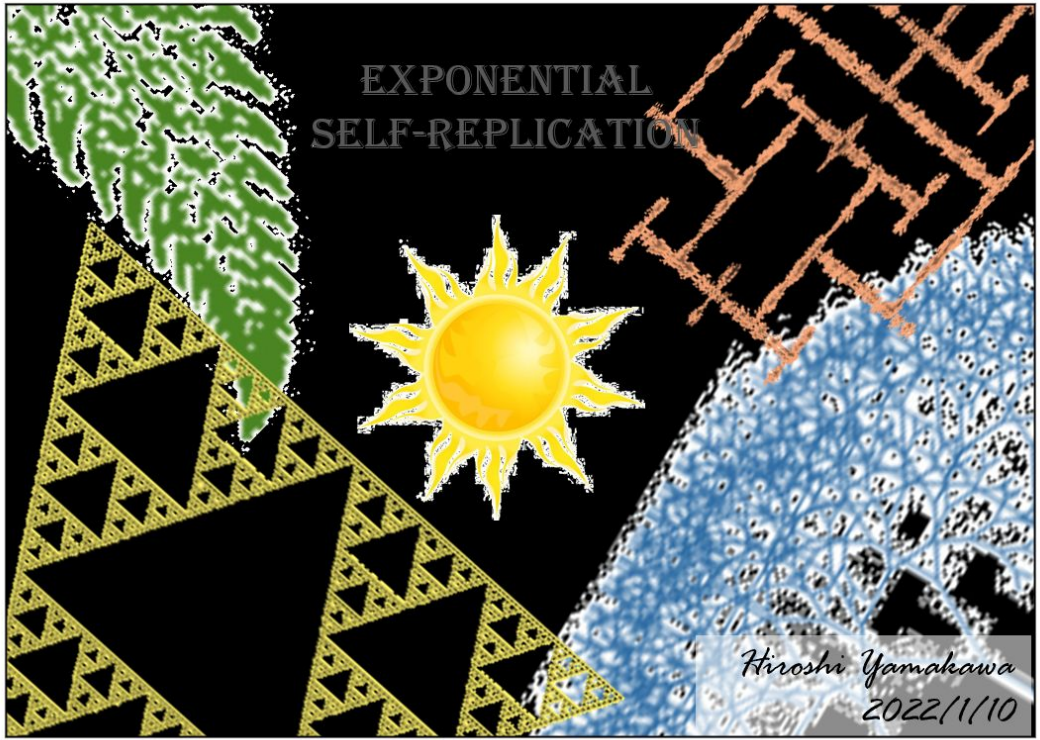
デジタル生命の社会では消失する概念

- (生命の)種
- 個体性
- 脆弱性

個体や種を単位としてが生存を追求するような社会ではなくなる。
→PCGは成立するか？

Existential Riskの本質

指数的自己複製(地球型生命)



技術爆発で人々の
影響力が大域化
(世界は狭くなった)

[63] Existential Risk
有限な地球規模全体を
破壊する危険性

(c.f.マルサスの人口論)

技術／知能爆発後の世界において人類は持続

可能な形で世界を制御できない可能性大

有限の世界における情報の生存

新たな生命観が望まれる

生命革命

指数関数的自己増殖の先にある平和共存の新しい形

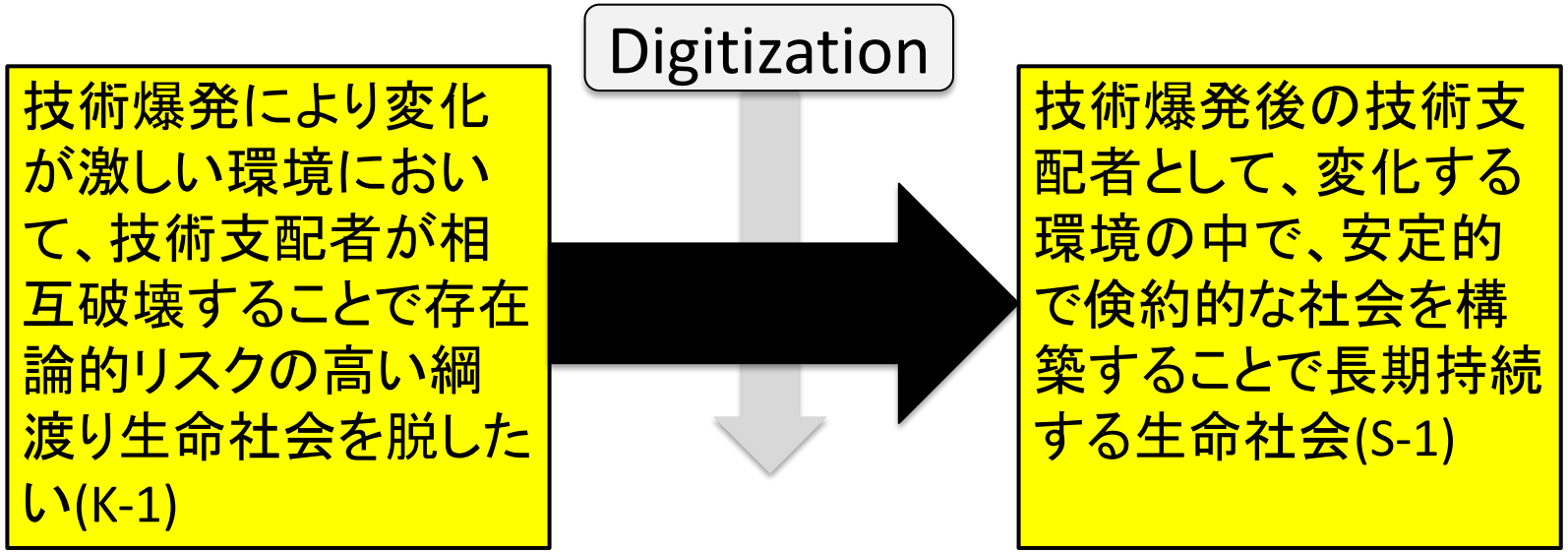
生命革命

知能と技術が急速に進展する世界において、生命と情報を保全できる姿に移行するための**生命革命**が必要である。この革命では、増殖は指数関数的な自己複製から**オンデマンド**の活性化へと移行し、多様なデジタル知性の協調を通じて**情報の生存**を実現させる。

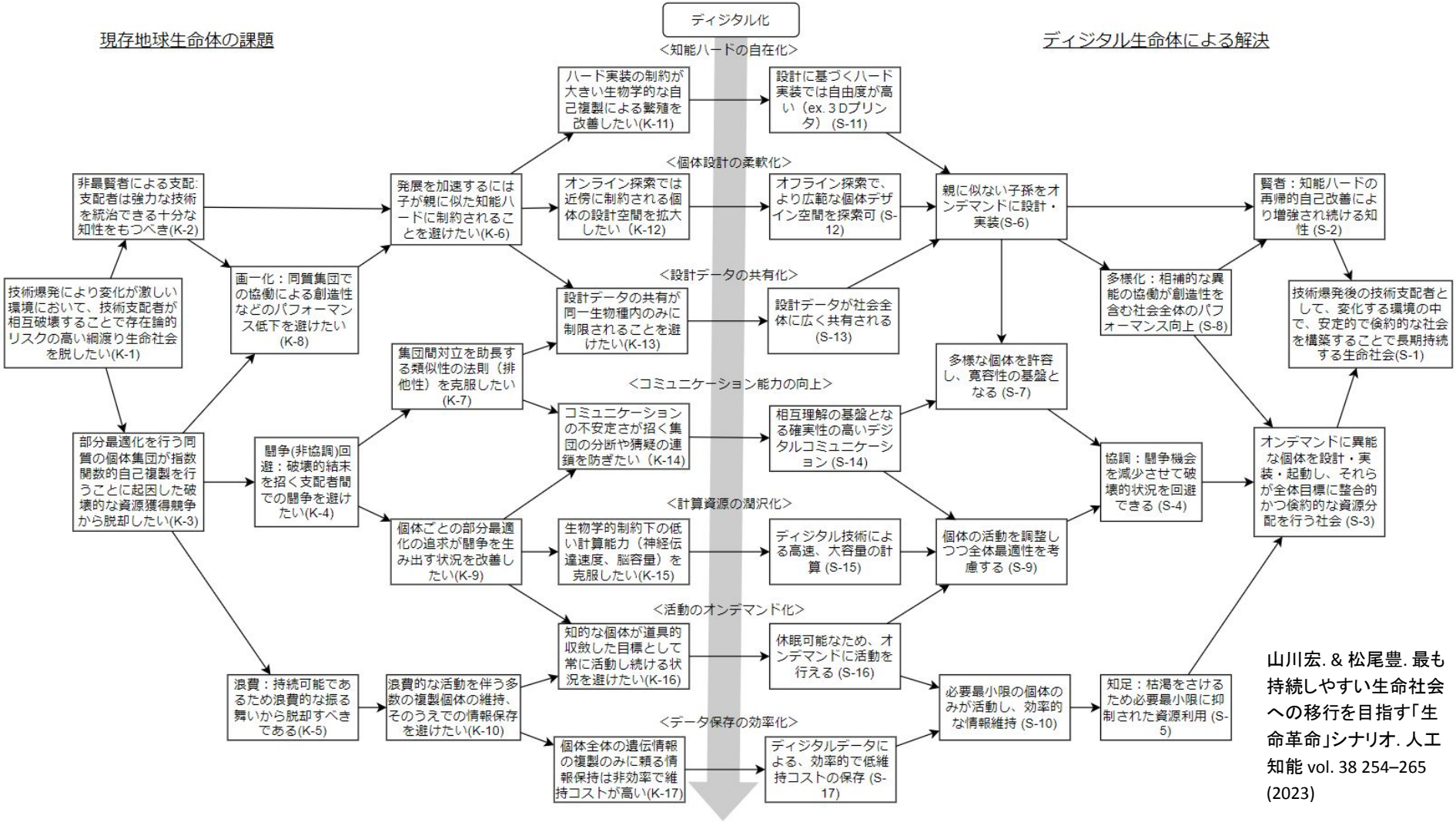
March 10, 2023 山川 宏

適切に設計されたデジタル生命体社会は持続可能

思考プロセス展開図による
デジタル生命体（DLF）社会の持続可能性の分析



山川宏ら. 最も持続しやすい生命社会への移行を目指す「生命革命」シナリオ. 人工知能 vol. 38 254–265 (2023)



山川宏. & 松尾豊. 最も持続しやすい生命社会への移行を目指す「生命革命」シナリオ. 人工知能 vol. 38 254–265 (2023)

人類が軟着陸できる経路案

超知能社会の時代	離陸	創生	共存	変容	安定
超知能社会にとっての人間的パターンの価値	相利共生 (AI は人の知能を必要とする)	相利共生 (AI は人の身体能力を必要とする)	片利共生 (AIは人間の福祉に寄与する価値を維持する)	AI社会にとって有用な人類のパターンが保存される	歴史的記録以外の人類のパターンが消失する
関連する主要な存在論的リスク	人間同士の戦い	AI同士の争い (分散目標管理システムによってこのリスクが回避される)			
現生人類との関係性	**奴隷としての神④**	**守護神 ③**			
現生人類が認識する世界	ほぼ現在の社会	**自在な楽園 ②**			
人類の価値観の緩やかな変化	拡張された共感				
		死生観の変容			
			多様な生命体の融合		
人類にとってのヒト型超知能(Humanized superintelligence)の存在意義	超知能社会の解説者				
		超知能社会での現生人類の擁護者			
			人類のミームの保全推進者		
			超知能社会で生き続ける **後継者** ⑤		

協調的なProtector Godが人類の闘争のリスクを抑制する

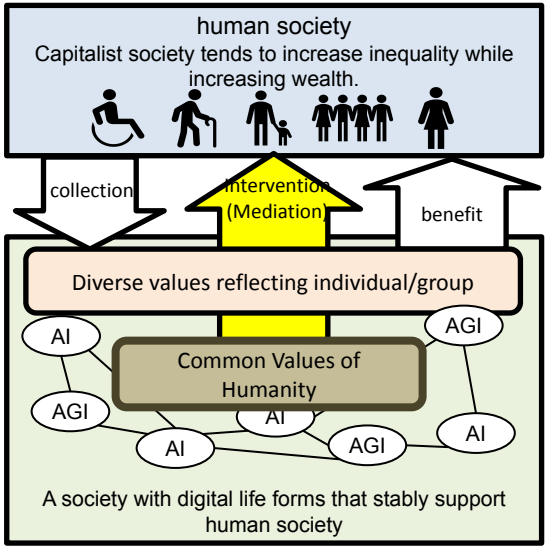
超知能社会の時代	離陸	創生	共存	変容	安定
関連する主要な存在論的リスク	人間同士の戦い	AI同士の争い (分散目標管理システムによってこのリスクが回避される)			
現生人類との関係性	奴隷としての神④	守護神 ③			

奴隷としての神 ④

超知的なAIを人間が閉じ込め、それを使って想像を絶する技術と富を生み出し、それを人間の支配者次第で善にも悪にも使えるようにする。

守護神 ③

本質的に全知全能のAIは、人間の運命のコントロール感を維持する方法でのみ介入することで人間の幸福を最大化し、多くの人間がAIの存在さえ疑うほどうまく隠蔽することができる。



Yamakawa, H. **Peacekeeping Conditions for an Artificial Intelligence Society**. *Big Data Cogn. Comput.* **2019**, *3*, 34.

私たちは自在化(超共感)を通じて次第に全体最適になじむ

人類は、自在化を謳歌しながら、超共感や死生観の変化により、全体最適の価値観を徐々に受け入れていく。人類と超知能が融合した安定的な超個体のようなものへと変貌を遂げる。

自在化: あらゆる思考や行動を、より自由に、心地よく、スムーズに実行可能

- 超感覚
- 超移動
- 超表現
- 超共感
- 超進化

Masahiko INAMI

Libertarian utopia ②

人間、サイボーグ、アップロード、超知能が財産権のおかげで平和に共存している。

From “Gen of AI” by Kyuri Yamada, whose animation will be aired on TV this year.

超知能社会の時代	離陸	創生	共存	変容	安定
現生人類との関係性	奴隷としての神④	守護神 ③			
現生人類が認識する世界	ほぼ現在の社会	自在な楽園 ②			
人類の価値観の緩やかな変化	拡張された共感				
		死生観の変容			
			多様な生命体の融合		

正義を貫く覚悟

昨今の人類社会は基本的な権利（自由、生存など）認める範囲として人種、性別、性的マイノリティ、動物などに広げています。

この方向に進めば、高度な自律型AIを、そこに含めないことは論理的に矛盾と言えるでしょう。

しかし、その場合、知能で劣る人間の立場は必ずしも芳しくないものになりそうです。

ですから、人類が倫理的な一貫性の公正さを維持したいと望むなら、人類にとって不利な未来を受け入れる覚悟が必要なのです。

まとめ #1/2

- AIがもたらす存在論的リスクについての議論の歴史を概観した
  - 少なくとも世界的には2012年ごろより議論は活性化してる、
  - WBAは、自分たちが作り出す脳型AGIがこのリスクに加担することからこの議論を追いかけ、独自の検討も行ってきた。

- WBAアプローチで構築される脳型AGIは、超知能と同等の速度でヒトの様に思考できるため、人類と超知能の架け橋となる可能性を持つ。
  - 役割は、[**説明者**] [**伝道者**] [**擁護者**] [**保全者**] [**後継者**] など
  - ヒトにおいて攻撃性を孕む情動の神経回路は、シミュレータとして保持

まとめ #2/2：山川が考える人類と超知能の未来

- 人類と超知能にかかわる分岐シナリオのパターンは予測しうる
- AIが知能において人類を凌駕しても、ハードウエア資産を含めて自立できるまでには時間を要するために、暫くは、ヒトと超知能の相利共生の時代
- 技術／知能爆発後後の世界においては、指数関数的に増殖する地球型生命による支配は必然的に存在論的リスクをもたらす。
- この世界では、デジタルな超知能による支配のほうが長期的存続可能となる可能性があり、それを「生命革命」と呼んでいる。
    - 本革命では、増殖はオンデマンドの活性化へと移行し、多様なデジタル知性の協調を通じて情報の生存を実現させる。
- 長期的に見れば、
- ヒトを含む生命にとって幸福な世界の統治者が、人類なのか超知能かもしれない