ISO/IEC 42001に基づく新たな適合性評価への期待

国立研究開発法人産業技術総合研究所
情報人間工学領域
杉村領一

ＡＩとは

強いAI
- 人間の「知能」と同等のものを作ることを目指す

弱いAI
- 道具として，「知的な処理」を行うことを目指す

双方とも
- 基本的に，ソフトウェアとデータで実現される機能.
- ソフトウェアはオープンソースで提供されることが多い
- データは，目的に応じて集められる，作られる

JIS X 22989における定義 (ISO/IEC 22989:2022 IDT)

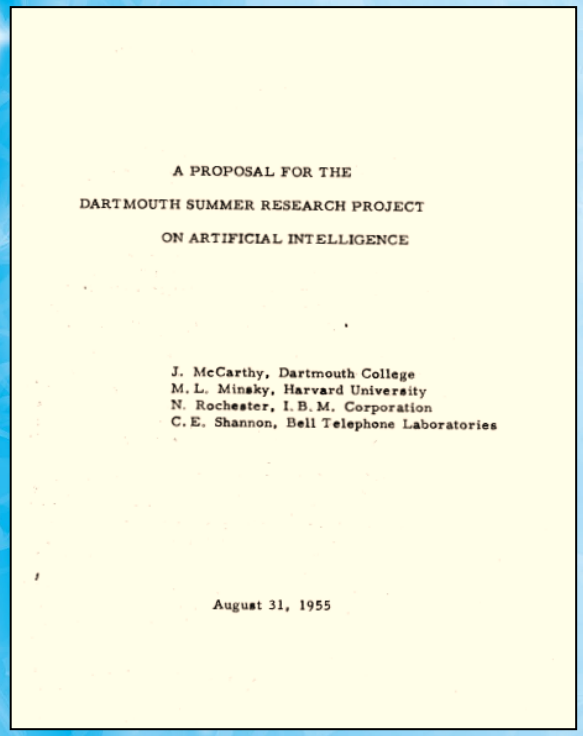
人工知能システム，AIシステム（artificial intelligence system，AI system）
人間が定義した所与の目標の集合に対して，コンテンツ，予測，推奨，意思決定などの出力を生成する工学的システム

注釈1 工学的システムは，人工知能（**3.1.3**）に関連する様々な技法及びアプローチを使用して，作業（**3.1.35**）の実施に使用可能であるデータ，知識（**3.1.21**），プロセスなどを表すモデル（**3.1.23**）を開発することが可能である。

注釈2 AIシステムは，様々な自動化（**3.1.7**）のレベルで動作するように設計されている。

AIの始まり(Inception of AI)

Dartmouth Summer Research Project on Artificial Intelligence in 1956



A PROPOSAL FOR THE

DARTMOUTH SUMMER RESEARCH PROJECT

ON ARTIFICIAL INTELLIGENCE

J. McCarthy, Dartmouth College
M. L. Minsky, Harvard University
N. Rochester, I. B. M. Corporation
C. E. Shannon, Bell Telephone Laboratories

August 31, 1955

Proposal document in 1955

We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

我々は、1956 年の夏、ニューハンプシャー州ハノーバーのダートマス大学で 10 人による 2 か月間の人工知能研究を行うことを提案します。この研究は、学習やその他の知能の特徴のあらゆる側面は原理的に非常に正確に記述できるため、機械でそれをシミュレートできるという仮説に基づいて進められます。機械に言語を使用させ、抽象化と概念を形成し、現在は人間にしかできない種類の問題を解決させ、機械自身を向上させる方法を見つける試みが行われます。厳選された科学者のグループが夏の間一緒に取り組めば、これらの問題の 1 つ以上で大きな進歩が達成できると考えています。

From https://en.wikipedia.org/wiki/Dartmouth_workshop

注目を浴びている機械学習とは？

第二世代AI

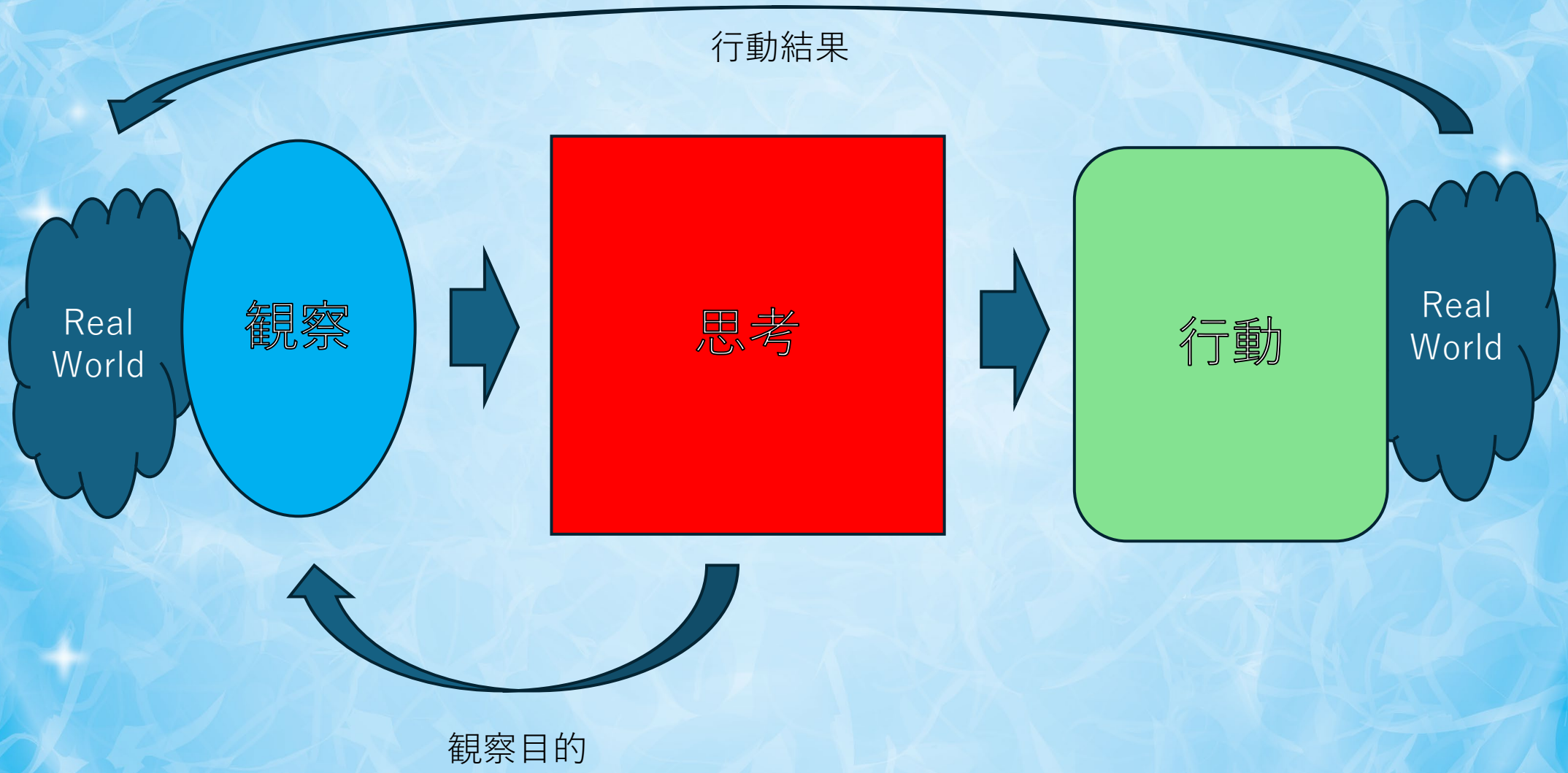
データ内の**パターン**を機械的に（人手で）見つけし、
予測や分析に用いる手法の総称

出典：データサイエンティスト養成読本　機械学習入門編　技術評論社 pp.20

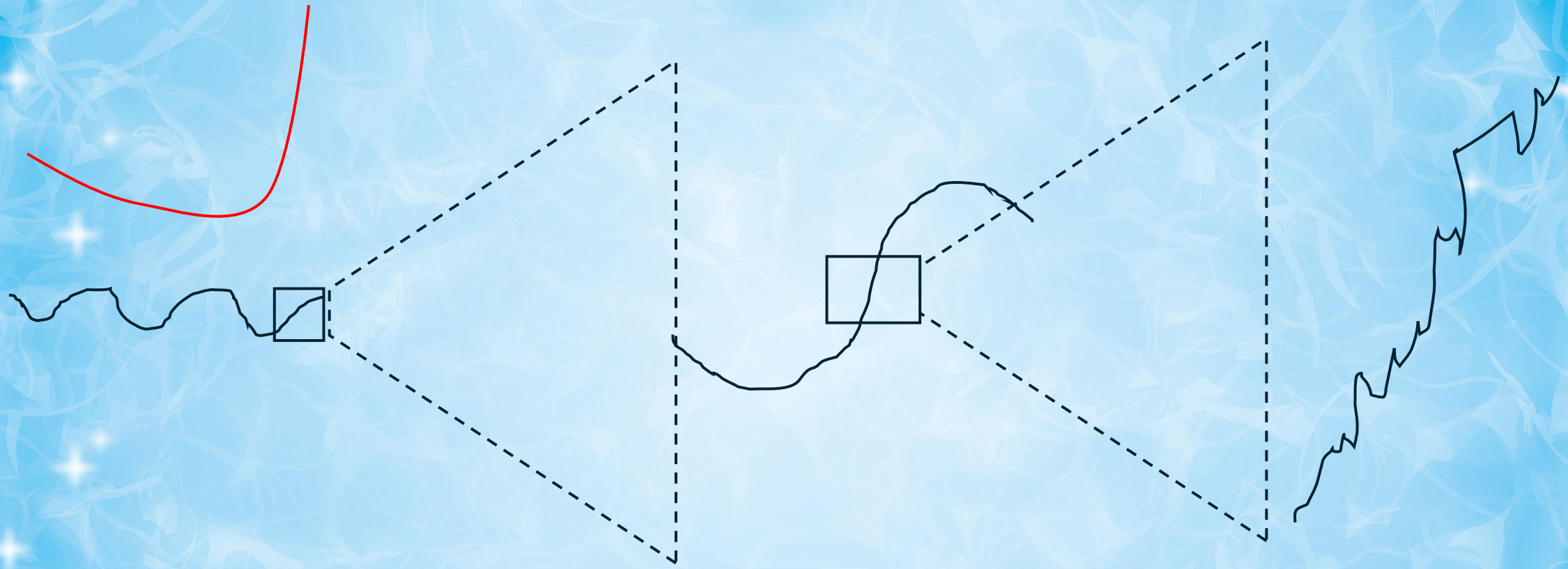
パターン

- 絵,柄,像 
- 家庭内幼児の転倒事故のパターン 
- 試験問題の出題パターン 
- 弁護士の弁論パターン 
- あなたの消費行動パターン 
- 日本人の行動パターン 
- 国際企業間の取引パターン 

AIの大まかな処理（例）
底流にあるのはデジタル化



デジタル化の壁：
データの取得方法は無限にある



取得　範囲　種類　時間帯　等

取得粒度

手っ取り早いのは、自分と似た状況にある先行成功例を真似し，貢献すること
オープンソースなど、ソフトウェアでは基本的な姿勢
でも日本では、一から考えることが貴ばれる？　尊敬される？

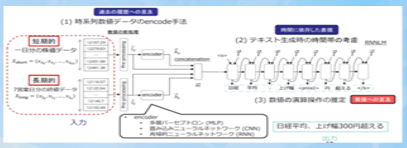
Recent Major Results (Function Modules)

Real-World

Natural Language Processing

Movie Caption Generation

Knowledge Extraction from Documents



Explanation of Time-Series Data

Observation Data Acquisition

Linked Living Lab

People Flow Measurement

VR Environment for Human-Robot Interaction

Information Sharing in Service Field

Detection and Tracking of Moving Objects/Humans

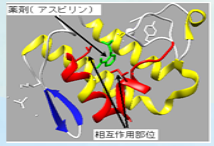
Measurement of Human Actions

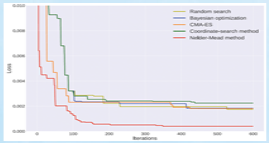
Modeling/Recognition/Prediction

Classification and Relation Modeling

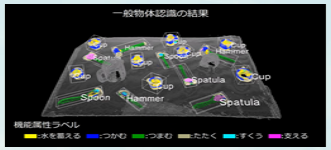
Pose Estimation from Multi-view Images

Prediction of Protein -Drug Interaction

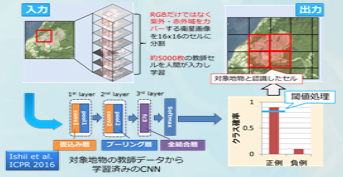




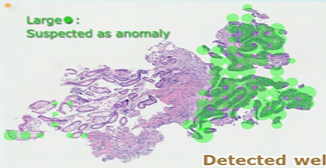
Hyper-Parameter Optimization of DNN



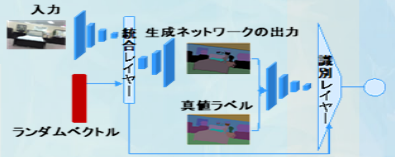
Recognition of Tool Function



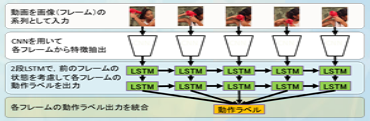
Recognition of



Anomaly Detection from Images/Signals



Object Recognition with GAN

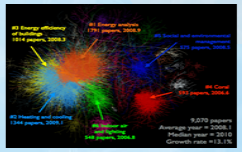


動画からの日常動作認識

Machine Learning

- 2値化重み{-1,+1}を{-1,0,+1}に拡張
- 接続関係の学習が可能になり、2値化DNNより精度が上がる

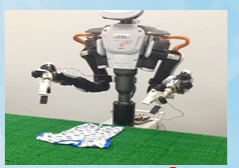
Trinary DNN for Low Power Devices



Visualization of Scientific Trends

Planning/Control

Planning Assembly Motion



Learning from Demonstration of Flexible Object Manipulation

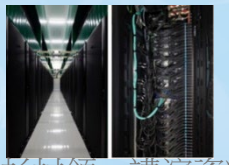
Manipulation of Tools with Function Recognition

These Works are supported by **NEDO** Project

Real-World

AI Bridging Cloud Infrastructure http://abci.ai

Computing Power Ranked #5@TOP500 for Everyone



ABCI
AI Bridging Cloud Infrastructure

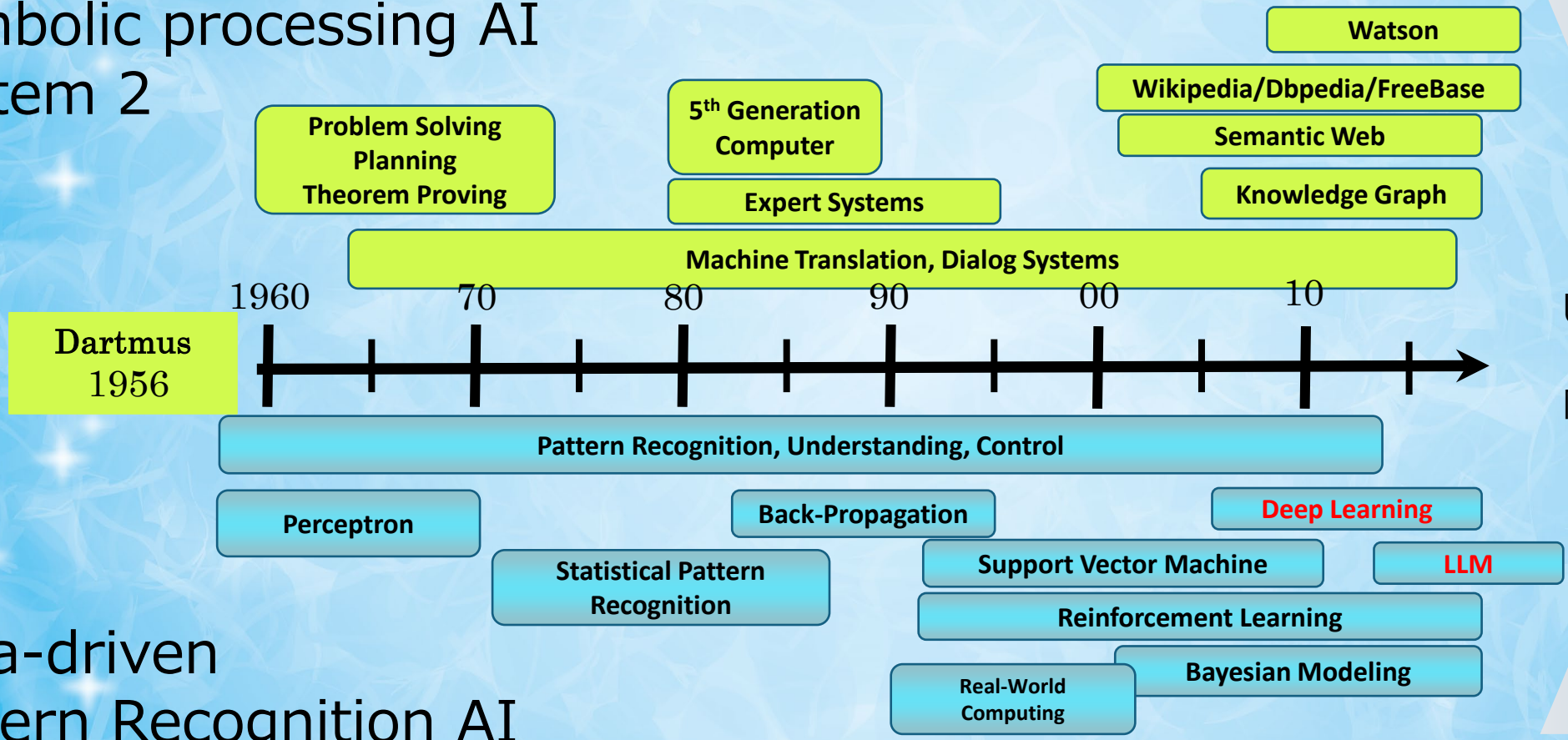
説明知/説明不可 × 既取得／未取得
AIは第4世代へ　（杉村の見解）

	我々自身が説明可	我々自身が説明不可
我々は既取得 (can be used)	*自身が意識的に獲得した知識，知見.* *国語，算数，理科，社会* *専門知識* 第一世代AI	獲得してはいるが，明示的に説明ができない知見. 自転車の乗り方，歩き方，言葉の使い方，等々 第三世代AI
我々は未取得 (Not available)	他人が持っていることは分かっているが，自信は持っていない知識 専門知識，地域情報， 第二世代AI	どこかにある，新しい知識 科学的発見！ イノベーションの元 👍 第四世代AI

出典：Roy Sugimura at GPAI November 21, 2022

System1 and System2 **in AI**

Knowledge-based
Symbolic processing AI
System 2



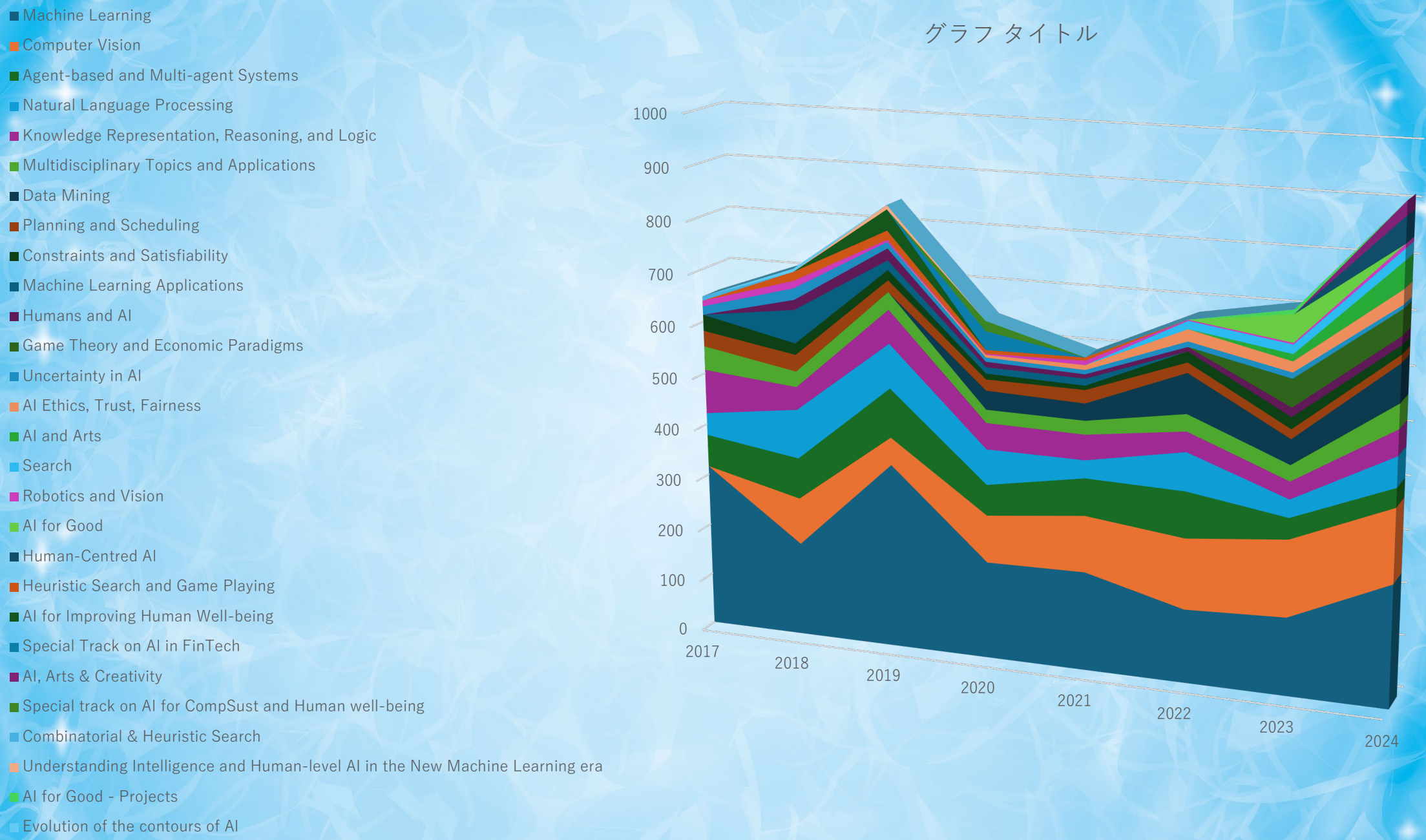
Unified Approach

Multi-modal (In/Out)

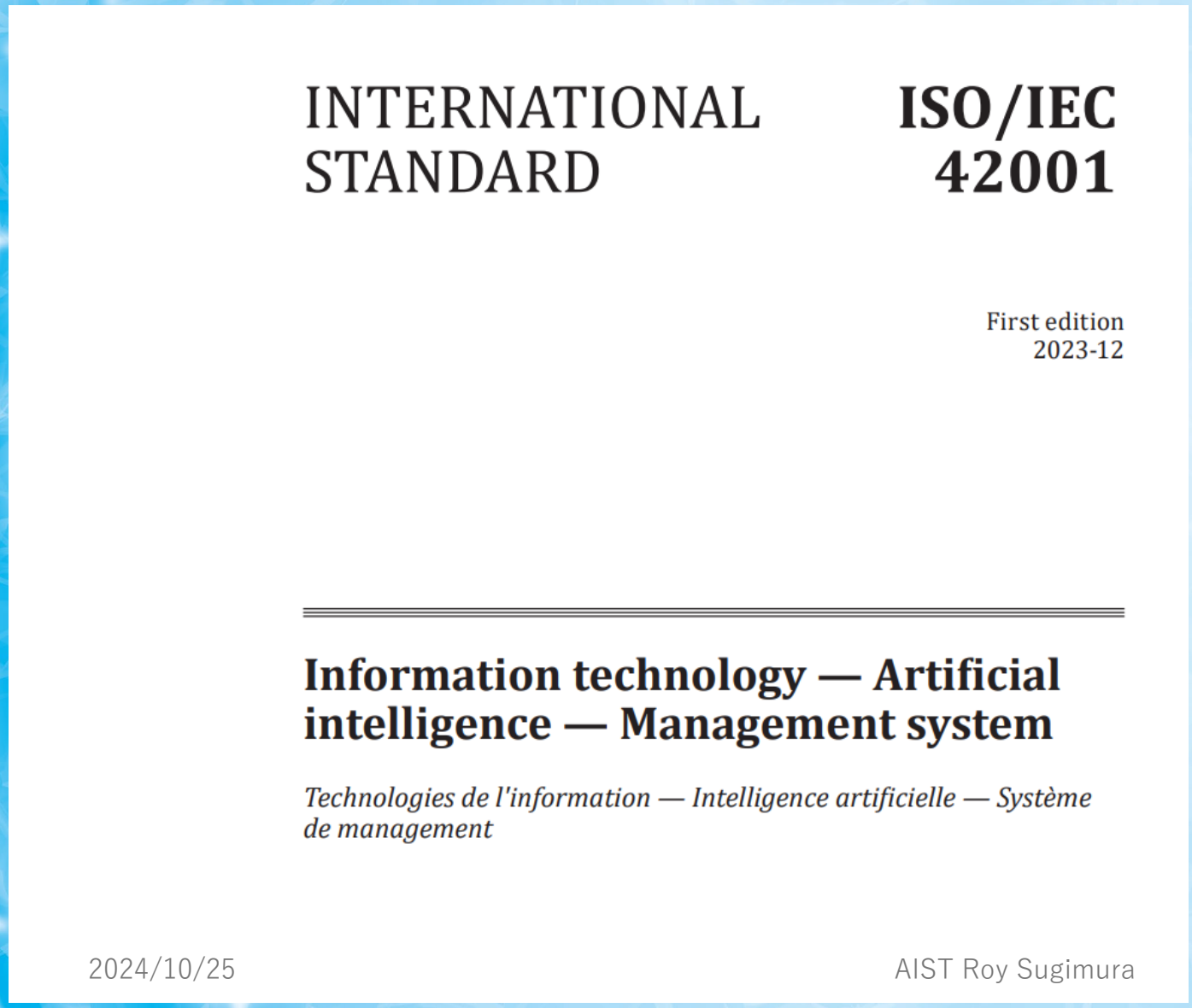
Data-driven
Pattern Recognition AI
System 1

IJCAI における発表傾向の変化

	2017	2018	2019	2020	2021	2022	2023	2024	総計
Machine Learning	316	181	356	187	190	141	149	234	1754
Computer Vision		90	53	91	108	135	146	140	763
Agent-based and Multi-agent Systems	62	79	94	59	71	88	40	36	529
Natural Language Processing	43	95	85	68	34	73	34	57	489
Knowledge Representation, Reasoning, and Logic	85	44	64	50	48	38	33	48	410
Multidisciplinary Topics and Applications	46	30	33	25	26	32	30	46	268
Data Mining				36	32	75	47	71	261
Planning and Scheduling	30	32	22	21	25	19	18	15	182
Constraints and Satisfiability	31	22	19	11	8	20	22	17	150
Machine Learning Applications		64	18	12	13				107
Humans and AI		19	22	10	8	8	17	16	100
Game Theory and Economic Paradigms							52	47	99
Uncertainty in AI	16	22	12	8	8	10	11	7	94
AI Ethics, Trust, Fairness				5	9	22	20	27	83
AI and Arts							13	49	62
Search						15	17	21	53
Robotics and Vision	11	14	4	2	8	2	3	8	52
AI for Good							51		51
Human-Centred AI								48	48
Heuristic Search and Game Playing		17	17	7	6				47
AI for Improving Human Well-being			39						39
Special Track on AI in FinTech				35					35
AI, Arts & Creativity								26	26
Special track on AI for CompSust and Human well-being				19					19
Combinatorial & Heuristic Search	8								8
Understanding Intelligence and Human-level AI in the New Machine Learning era			8						8
AI for Good - Projects							8		8
Evolution of the contours of AI		6							6


グラフ タイトル
1000
900
800
700
600
500
400
300
200
100
0
2017
2018
2019
2020
2021
2022
2023
2024
Machine Learning
Computer Vision
Agent-based and Multi-agent Systems
Natural Language Processing
Knowledge Representation, Reasoning, and Logic
Multidisciplinary Topics and Applications
Data Mining
Planning and Scheduling
Constraints and Satisfiability
Machine Learning Applications
Humans and AI
Game Theory and Economic Paradigms
Uncertainty in AI
AI Ethics, Trust, Fairness
AI and Arts
Search
Robotics and Vision
AI for Good
Human-Centred AI
Heuristic Search and Game Playing
AI for Improving Human Well-being
Special Track on AI in FinTech
AI, Arts & Creativity
Special track on AI for CompSust and Human well-being
Combinatorial & Heuristic Search
Understanding Intelligence and Human-level AI in the New Machine Learning era
AI for Good - Projects
Evolution of the contours of AI

本講義の趣旨



INTERNATIONAL STANDARD

ISO/IEC 42001

First edition
2023-12

Information technology — Artificial intelligence — Management system

Technologies de l'information — Intelligence artificielle — Système de management

出版された 左記規格について，
その作成経緯，関連する活動，そして
内容を理解する上で，特にAIに特有と
考えられる用語・概念，更には，管理策
を実装する際に，AIに特徴的な内容と
なっている内容について，あくまで，
AI研究者としての視点から，解説を
行います．

実務にて利用を検討される場合には，
是非，ISOの出版物を 42001, 229892 を
入手頂き，詳細をご確認されることを
お勧め致します．

なお，22989 は出版済で英文本文は無償
で提供されています．JIS X 22989 は
出版済です．また，JIS Q 42001 はJIS化
が現在推進されています．

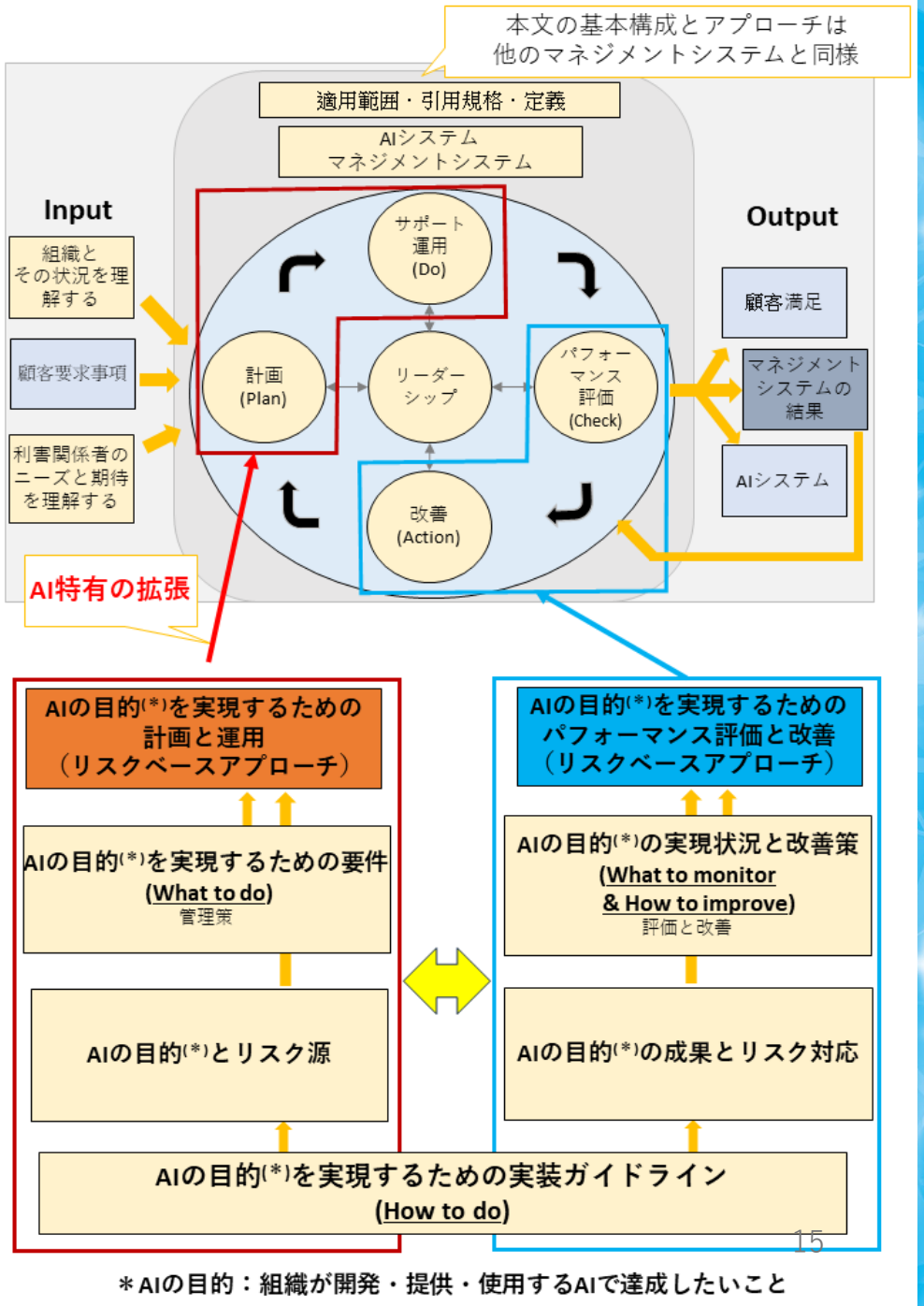
関連規格開発状況

	規格名称	エディタ
ISO/IEC 42006	Information technology — Artificial intelligence — Requirements for bodies providing audit and certification of artificial intelligence management systems	独
ISO/IEC 42005	Information technology — Artificial intelligence — AI system impact assessment	独
ISO/IEC 25336 (rnbred 42007)	Information technology — Artificial intelligence — High-level framework and guidance for the development of conformity assessment schemes for AI systems	独
ISO/IEC 42114	Information technology — Artificial intelligence — Guidelines for auditing of AI management system based on ISO/IEC 42001	日

ISO/IEC 42001:2023

- マネジメント標準としての共通部分は，ISO Directives Annex SL に Blue Text として規定．骨組みとしては，4 組織，5 リーダシップ，6 計画，7 サポート, 8 運用，9 パフォーマンス評価, 10 改善のループを回す体系
- AIに特有の部分について，追記等を行った．
- 実務的には，Annex A Reference control objectives and controls, B Implementation guidance for AI controls, C Potential AI-related organizational objectives and risk sources，D Use of the AI management system across domains or sectors なども，AI特有の追記部分として，参考となる
- Normative Reference としては，ISO/IEC 22989:2022 のみだが，22989:2022 にはAIの用語と概念が定義されている．

図出典：
https://www.meti.go.jp/press/2023/01/20240115001/20240115001.html



ISO/IEC 42001 Information Technology – Artificial Intelligence – management system 開発経緯

- 最初の提案：
  - NWIP ISO/IEC JTC 1/SC 42 N 649
  - Circulation date: 2020-05-28
  - Closing date for voting: 2020-08-21
- Proposed Project Leader (name and e-mail address)
  - Jim MacFie (jimacfie@microsoft.com)　⇒　後に，MacFie 氏退任に伴い, Marta Janczarski 氏へ．氏は後に，アイルランドへ移住
- Liaisons:
  - ISO/CASCO, ISO/TC 176, ISO/TC 262, ISO/IEC JTC 1/SC 27, ISO/IEC JTC 1/SC 40

- 提案活動時期と、COVIT 19 が重なり、殆どの議論はリモートで実施
- 当初のプロジェクトエディターは Canada の大ベテラン、のちに、当該エディタから新人Marta Janczarski 氏へ交代
- エディタは、その後、カナダからアイルランドへ移住
- 42001 は、アイルランドのエディタからの提案としてStandard Request への回答として提案

日本は、極めて基本的な所から詳細に至るまで継続的に貢献
To avoid confusion mentioned here I would recommend that identical core text and discipline-specific additions should be clearly distinguished from the starting of the drafting process as Japan requested in its comment to Q.6 of the ballot (WG 1 N 710) and the Convenor mentions in WG 1 N 727.

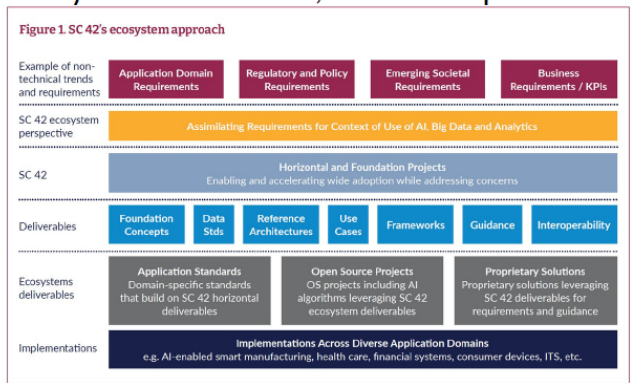
Comment to Q.6: (N 710)
Recommended to collaborate with AI System Lifecycle and AI Governance standardization activities. Request to distinguish between discipline-specific text and identical core text as instructed in Directives Part 1 Annex L.9.4 Clause 6. Request careful development considering the difference of two aspects of MSS, using AI and providing AI.

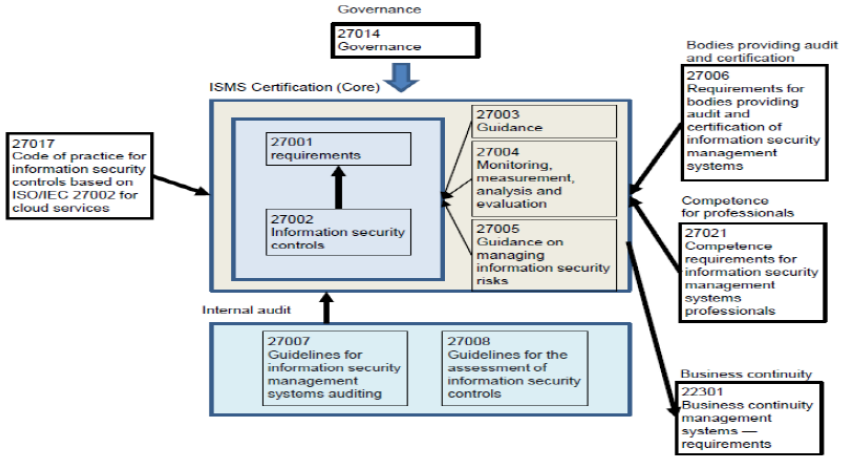
2700X を参照して、Constellation を提案

Competition and Collaboration among Standard Ecosystems?

A Series of Standards is achieving a competitive advantage

Proposal by Mr. Wael Diab, the Chair person of SC 42

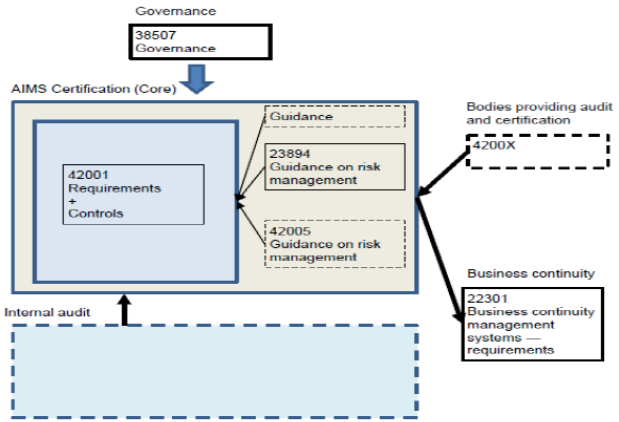




Set of Standards around ISO/IEC 27001

Series of Data related standards by SC42/WG2

Standard	Type	Title	Status
20546	IS	Big data - Overview and vocabulary	Published
20547		Big data reference architecture	
-1	TR	Part 1: Framework and application process	Published
-2	TR	Part 2: Use cases and devrived requirements	Published
-3	IS	Part 3: Reference architecture	Published
-5	TR	Part 5: Standards roadmap	Published
24668	IS	Process management framework for big data analytics	FDIS
5259		Data quality for analytics and machine learning	
-1	IS	Part 1: Overview, terminology, and examples	CD
-2	IS	Part 2: Data quality measures	WD
-3	IS	Part 3: Data quality management requirements and guidelines	CD
-4	IS	Data quality process framework	CD
-5	IS	Data quality governance framework	WD
8183	IS	Data life cycle framework	DIS



Set of Standards around ISO/IEC 42001

ISO/IEC 17000:2020	Conformity assessment — Vocabulary and general principles
ISO/IEC 17007:2009	Conformity assessment — Guidance for drafting normative documents suitable for use for conformity assessment
ISO/IEC 17011:2017	Conformity assessment — Requirements for accreditation bodies accrediting conformity assessment bodies
ISO/IEC 17020:2012	Conformity assessment — Requirements for the operation of various types of bodies performing inspection
ISO/IEC 17021-1:2015	Conformity assessment — Requirements for bodies providing audit and certification of management systems — Part 1: Requirements
ISO/IEC 17021-3:2017	Conformity assessment — Requirements for bodies providing audit and certification of management systems — Part 3: Competence requirements for auditing and certification of quality management systems
ISO/IEC TS 17023:2013	Conformity assessment — Guidelines for determining the duration of management system certification audits
ISO/IEC 17024:2012	Conformity assessment — General requirements for bodies operating certification of persons
ISO/IEC 17025:2017	General requirements for the competence of testing and calibration laboratories
ISO/IEC TR 17026:2015	Conformity assessment — Example of a certification scheme for tangible products
ISO/IEC TS 17027:2014	Conformity assessment — Vocabulary related to competence of persons used for certification of persons
ISO/IEC TR 17028:2017	Conformity assessment — Guidelines and examples of a certification scheme for services
ISO/IEC 17029:2019	Conformity assessment — General principles and requirements for validation and verification bodies
ISO/IEC 17030:2021	Conformity assessment — General requirements for third-party marks of conformity
ISO/IEC TR 17032:2019	Conformity assessment — Guidelines and examples of a scheme for the certification of processes
ISO/TS 17033:2019	Ethical claims and supporting information — Principles and requirements
ISO 17034:2016	General requirements for the competence of reference material producers
ISO/IEC TS 17035:2024	Conformity assessment — Guidelines for validation and verification programmes
ISO/IEC 17040:2005	Conformity assessment — General requirements for peer assessment of conformity assessment bodies and accreditation bodies
ISO/IEC 17043:2023	Conformity assessment — General requirements for the competence of proficiency testing providers
ISO/IEC 17050-1:2004	Conformity assessment — Supplier's declaration of conformity — Part 1: General requirements
ISO/IEC 17050-2:2004	Conformity assessment — Supplier's declaration of conformity — Part 2: Supporting documentation

ISO/IEC 42001:2023 にある用語

42001:2023 にある用語

- 組織（organization），利害関係者（interested party），トップマネジメント（top management），マネジメントシステム（management system），方針（policy），目的　（objective），リスク（risk）　，プロセス（process），力量（competence），文書化した情報（documented information），パフォーマンス（performance），継続的改善（continual improvement），有効性（effectiveness），要求事項（requirement），適合（conformity），不適合（nonconformity），是正処置（corrective action），監査（audit），測定（measurement），監視（monitoring），管理策（control），経営陣（governing body），情報セキュリティ（information security），AIシステム インパクト アセスメント (AI system impact assessment），データ品質 (Data Quality), 適用宣言書（statement of applicability）
- 42001:2023 の Normative Reference は ISO/IEC 22989 のみ
  ISO/IEC 22989 において定義されている用語の理解が必要

ISO/IEC 22989:2022に定義された用語

日本語は著者メモ（必要な場合は JIS X 22989:2023をご利用下さい）

AIに関する用語

- **AIエージェント**（AI agent）【環境を感知及び環境に応答し，目標達成のための行動をとる自動化したエンティティー】,<SNIP>,
  ⇒　自動(Autonomy, autonomous) という概念だけを用いた技術文書も多数ある。私見ではあるが、自動だけではＡＩとは言えない。外界から認知された情報を処理するための変更可能かつ、学習可能なルールを持っていることが、ＡＩの特徴、ＡＩシステムのそれまでの情報システムとの差と考えるのが、私見ではあるが自然に思える。

- **人工知能，AI**（artificial intelligence，AI）

  ＜学問分野＞AIシステムのメカニズム及び適用の研究開発
  注釈1　研究開発は，コンピュータサイエンス，データサイエンス，自然科学，人文科学，数学など，幾つもの分野にわたって行うことが可能である。
  ⇒　学問分野以外の定義は、22989 の開発作業では断念している。作業過程において、学問分野でも非常に幅広く議論がされているように、多くの定義が提案され、いわゆるコンセンサスを取ることが難しかった。また、以降は私見であるが、ＡＩに対しては、その基礎となる、例えば知識 (knowledge) についての欧州関係者の抱える難しさ（Wisdom, Intelligence, Information, Data ハイアラーキーにおいて、intelligence は人間のみが扱えるという考え方など）が、ＡＩをデータを元に抽出した知識を用いるとする定義などを阻み、最終的な定義を困難にした。

- **人工知能システム，AIシステム**（artificial intelligence system，AI system）

  人間が定義した所与の目標の集合に対して，コンテンツ，予測，推奨，意思決定などの出力を生成する工学的システム

  注釈1 工学的システムは，人工知能に関連する様々な技法及びアプローチを使用して，作業の実施に使用可能であるデータ，知識，プロセスなどを表すモデルを開発することが可能である。

  注釈2 AIシステムは，様々な自動化 のレベルで動作するように設計されている。

- **自律性，自律，自律的**（autonomy，autonomous）【外部からの介入，制御，オーバーサイトを受けずに，意図された使用領域又は目標を変更できるシステムの特性】 <SNIP>

- **自動，自動化，自動的（な）**（automatic，automation，automated）【プロセス又はシステムによる，指定された条件下で，人間の介入なしに機能する】

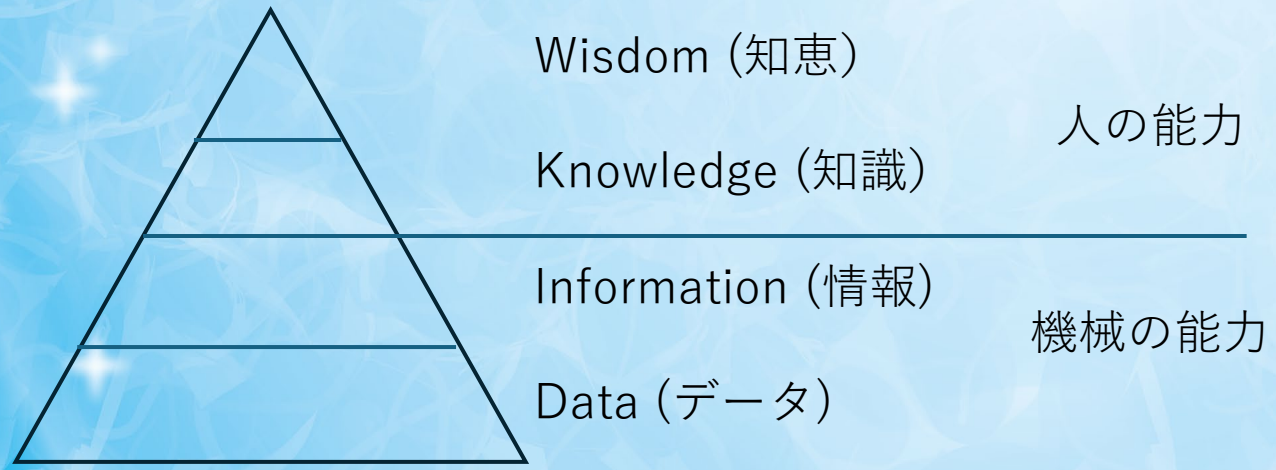
- **自律性，自律，自律的**（autonomy，autonomous）【外部からの介入，制御，オーバーサイトを受けずに，意図された使用領域又は目標を変更できるシステムの特性】
  ⇒　私見であるが、意図には、設計意図、システムの動作中に得られる意図、などが含まれ得る。これらは、例えば、Plan-Goal(sub-goal) という形で実装され得る。
  ⇒　変更できるという点は、正に、学習の寄与があり得る。この意味からは、当該定義は、学習の無い自動システムと学習のある自動システム双方を併せ持って定義しているともいえる
  ⇒　なお、当該定義以前の定義では、autonomy, autonomous を特徴付けるのは、人間の介入の有無となっている。
  （参考）autonomy
  ability to perform intended tasks based on current state and sensing, without human intervention　[SOURCE: ISO 8373:2012, 2.2]　ISO/TR 23482-1:2020(en), 3.1　**<SNIP>**
- **認知コンピューティング**（cognitive computing）【人間と機械との，より自然な相互作用を可能とするAIシステムのカテゴリー．**注釈1** 認知コンピューティングの作業は，機械学習（**3.3.5**），音声処理，自然言語処理，コンピュータビジョン及びヒューマンマシンインターフェースに関連する。】
  ⇒　Human Machine Teaming という規格提案は、当該分野の一つの取り組みとみなせる
- **継続的学習**（continuous learning，continual learning），**生涯学習**（lifelong learning）
  AIシステムライフサイクルの運用フェーズ中に継続的に行われるAIシステムの逐次訓練
  ⇒　フェーズは、22989 では、正確にはステージと記されている。
- **コネクション主義**（connectionism），**コネクショニストのパラダイム**（connectionist paradigm），**コネクショニストモデル**（connectionist model），**コネクショニストのアプローチ**（connectionist approach）
  - 多くの場合，単純な計算ユニットが相互に接続するネットワークを使用する認知モデリングの形態

- **データマイニング**（data mining）
  異なった観点及び次元から定量的なデータを分析し，それらを分類し，潜在的な関係及び影響を要約することによって**パターン**を抽出する計算処理
  ⇒パターン情報処理は、画像、自然言語などを対象とし、１９６０年代には多くの研究が開始されていた。私見ではあるが、現在のＬＬＭなどは、時系列に並んだ単語の組み合わせの膨大な、かつ数え上げることがもはやできないとも言える数のパターンを学習するものともみなし得るだろう。
- **宣言的知識**（declarative knowledge）
  事実，規則及び定理で表される知識
  注釈1通常，宣言的知識は，前もって手続的知識に変換しない限り，処理することは可能ではない。
  注釈2 宣言的知識は，信念，述語，フレームなどを含むことがある。
- **エキスパートシステム**（expert system）
  問題の解決策を推論する（**infer**）ために，特定の領域における人間の専門家から提供された知識を蓄積し，組み合わせ，カプセル化するAIシステム
- **汎用AI，汎用人工知能，AGI**（general AI，AGI）
  広範な作業（**3.1.35**）に，満足のいくレベルの性能で対応するAIシステム（**3.1.4**）のタイプ
  注釈1 専用AI（**3.1.24**）と比較。
  注釈2AGIはより強い意味で使用されることが多く，多種多様な作業を実行可能であるだけでなく，人間が実行可能である全ての作業も実行可能であるシステムを意味している。
- **遺伝的アルゴリズム，GA**（genetic algorithm，GA）
  最適化問題において個体（解）の集団を発生させ進化させることで自然とう（淘）汰をシミュレートするアルゴリズム
- **他律性，他律，他律的**（heteronomy，heteronomous）
  外部からの介入，制御，又はオーバーサイトの制約の下で動作しているシステムの特性

- **推論（inference）**
  既知の前提から結論を導き出す推論（**reasoning**）
  注釈1　AIでは，前提条件は，事実，規則，モデル，特徴又は生データのいずれかである。
  注釈2　“推論（**inference**）”という用語は，プロセス及びその結果の両方を指す。

- **モノのインターネット，IoT（internet of things，IoT），IoT機器**（IoT device），**IoTシステム**（IoT system）

- **知識（knowledge）**
  ＜人工知能＞オブジェクト，イベント，概念又は規則，それらの関係及び特性に関する抽象化された情報であり，目標指向の体系的な使用のために編成されたもの
  注釈1　AI領域の知識とは，他の幾つかの領域における用語の用法に反して，認知能力を意味しない。特に，知識は理解という認知行動を意味しない。
  注釈2　情報は，数値又は記号の形式で存在可能である。
  注釈3　情報は，状況に応じたデータであるため，解釈可能である。データは，抽象化又は測定によって世界から作成される。



追記情報 by 杉村
- 認知　は未定義述語
- AI の研究分野には、Knowledge Engineering があり、左記の階層議論と矛盾する.
- 自然言語処理の関連からは、情報には意味と、解釈が素材するが、広い意味で文脈が必要。ただ、客観的条件として検討されることは少ない

- **ライフサイクル（life cycle）**
  ISO/IEC 5338 - Information technology — Artificial intelligence — AI system life cycle processes（日本提案. エディターは富士通（株）鄭育昌氏）が出版済
- **モデル（model）**「システム，エンティティー，現象，プロセス又はデータの物理的，数学的又はその他の論理的な表現】
  ⇒ 学習結果をモデルと呼ぶことが一般的になっている
  ⇒ 第二次ＡＩ時代、モデルは、数学的な観点から議論されることが多かった。
- **専用AI（narrow AI）, 性能**（performance）, **プランニング**（planning）, **予測**（prediction）, **手続的知識**（procedural knowledge）, **ロボット**（robot）, **ロボティクス**（robotics）, **セマンティックコンピューティング（semantic computing）**, **ソフトコンピューティング（soft computing）**, **記号的AI（symbolic AI）**, **準記号的AI**（subsymbolic AI）, **作業（task）**
- **データアノテーション**（data annotation）【データに変更を加えずに一連の記述情報をデータに付加する処理　**注釈1**　記述情報は，メタデータ，ラベル（**3.2.10**），アンカーの形式をとる場合がある。】
  ⇒ 2024.10 SC42 Plenary における議論では、メタデータをフレームワークと呼ぶ傾向が見受けられる。
- **データ品質チェック**（data quality checking）, **データ拡張**（data augmentation）, **データサンプリング（data sampling）**,
- **データセット**（dataset）
  共有フォーマットによってデータを集めたもの
  例1　ハッシュタグ#rugby 及び#football に関連付けられた2020年6月以降のマイクロブログ投稿。
  例2　256×256画素の花のマクロ写真。
  注釈1 データ集合は，AIモデル（**3.1.23**）の妥当性確認又はテストに使用可能である。機械学習（**3.3.5**）においては，データセットは，機械学習アルゴリズム（**3.3.6**）の訓練に使用可能である。
  ⇒ set については、19世紀から２０世紀に多くの研究成果がある
  ⇒ 第二次ＡＩの波の時代では、共通フォーマット、という外苑的な議論だけでなく、データの持つ性質など、分析的な研究にも多くの努力が払われていた。

<snip>

- **真値**（ground truth）【ラベル付き入力データの，ある項目の対象変数の値．**注釈1**　真値という用語は，ラベル付き入力データが一貫して対象変数の実世界の値に対応することを意味するわけではない。】
  ⇒　当該説明の姿勢については，種々，議論があった．一般的には，現実世界の正しい情報を指す
- **補完**（imputation）【推定又はモデルによって得られたデータで，欠落データを置き換える手順】

<snip>

- **個人識別可能情報，PII**（**personally identifiable information，PII**），**個人データ**（**personal data**）【a) 情報と，当該情報が関連する自然人とを結び付けるために用いることが可能であるか，又はb) 自然人と直接的又は間接的に結び付いている，若しくは結び付けることが可能である情報．**注釈1** 定義の“自然人”が，PIIの主体である。PIIの主体が識別可能であるかどうかを判断するには，PIIの集合と自然人との関連を確証付けるために，そのデータを保持するステークホルダー又は他の団体が合理的に利用することが可能である，全ての手段を考慮することが望ましい。**注釈2** この定義は，この規格で使用されるPII という用語を定義するためのものである。パブリッククラウドPII処理者は，クラウドサービスの顧客によって透明化されない限り，処理する情報が特定のカテゴリーに分類されるかどうかを明示的に知ることができる立場にはないのが一般的である。】

  ⇒ Ohio State University Prof. Hirsh の調査に依ると，企業がAI開発において，最も注意を払うべきデータの一つは，間違いなく個人情報との報告がある．
  https://moritzlaw.osu.edu/sites/default/files/2023-05/Final%20Report_10.15.21.pdf
  https://link.springer.com/book/10.1007/978-3-031-21491-2

- **本番データ（production data）**【AIシステム（**3.1.4**）の運用フェーズで取得したデータで，展開されたAIシステムが予測出力又は推論（**inference**）を算出するもの】
- **サンプル（sample）**【機械学習アルゴリズム（**3.3.6**）又はAIシステム（**3.1.4**）によってまとまった量で処理される分割不可能なデータ要素】
- **テストデータ（test data）**，**評価データ**（evaluation data）**，妥当性確認データ**（validation data）**，代替用語：開発データ（development data）**，**ベイジアンネットワーク（Bayesian network）**，**決定木**（decision tree）
- **ヒューマンマシンチーミング（human-machine teaming）**【人間と，機械の知的能力との相互作用の統合】
- **ハイパーパラメータ**（hyperparameter）【学習プロセスに影響を与える機械学習アルゴリズム（**3.3.6**）の特性．**注釈1** ハイパーパラメータは，訓練の前に選択され，モデルのパラメータの推定を支援するためのプロセスにおいて使用可能である．**注釈2** ハイパーパラメータの例としては，ネットワーク層の数，各層の幅，活性化関数のタイプ，最適化法，ニューラルネットワークのための学習率，サポートベクトルマシンにおけるカーネル関数の選択，決定木の葉の数又は深さ，K－平均クラスタリングのためのK，期待値最大化アルゴリズムの反復回数，混合ガウスモデルにおけるガウス分布の数などがある。】
- 機械学習，ML（machine learning，ML）【モデル（**3.1.23**）の挙動が，データ又は経験を反映するような計算技法を通してモデルパラメータ（**3.3.8**）を最適化する処理】
- **機械学習アルゴリズム（machine learning algorithm）**【機械学習モデルのパラメータを所与の基準に従って決定するアルゴリズム】**，機械学習モデル（machine learning model）**，**パラメータ（parameter）**，**モデルパラメータ（model parameter）**，**強化学習，RL（reinforcement learning，RL）**，**再訓練**（retraining），**半教師あり機械学習（semi-supervised machine learning）**，**教師あり機械学習**（supervised machine learning），**サポートベクトルマシン，SVM**（support vector machine，SVM），訓練済みモデル（trained model），訓練（training）**代替用語：モデル訓練**（model training），**訓練データ**（training data），教師なし機械学習（unsupervised machine learning）
- 活性化関数（activation function），**畳み込みニューラルネットワーク**，CNN（convolutional neural network，CNN），**代替用語：深層畳み込みニューラルネットワーク，DCNN**（deep convolutional neural network，DCNN），**畳み込み**（convolution），**深層学習**（deep learning）**代替用語：深層ニューラルネットワーク学習**（deep neural network learning），**勾配爆発**（exploding gradient），**フィードフォワードニューラルネットワーク，FFNN**（feed forward neural network，FFNN），**長短期記憶，LSTM**（long short-term memory，LSTM），長期及び短期の**ニューラルネットワーク，NN**（neural network，**NN**） **代替用語：ニューラルネット**（neural net），**人工ニューラルネットワーク**（artificial neural network），**ニューロン**（neuron），**リカレントニューラルネットワーク，リカレントNN，RNN**（recurrent neural network，RNN），

- **説明責任を負った**（accountable）【行動，意思決定及び性能について応えられる】 phonogram vs ideogram
- **説明責任**（accountability）【説明責任を負った状態 **注釈1** 説明責任は，割り当てられた責任範囲に関連する。責任は，規制若しくは規約に基づくか，又は委任の一部としての割当てに基づく場合がある。**注釈2** 説明責任には，特定の手段を通じて，及び特定の基準に従って，ある個人又はエンティティーが別の個人又はエンティティーに対して，何らかの説明責任を負うことが含まれる。】

AI事業者ガイドラインでの Accountability は アカウンタビリティ と表音文字で表現されており，説明は以下
注記）アカウンタビリティを説明可能性と定義することもあるが、本ガイドラインでは情報開示は透明性で対応することとし、アカウンタビリティとはAIに関する事実上・法律上の責任を負うこと及びその責任を負うための前提条件の整備に関する概念とする。

7) アカウンタビリティ	① トレーサビリティの向上 ②「共通の指針」の対応状況の説明 ③ 責任者の明示 ④ 関係者間の責任の分配 ⑤ ステークホルダーへの具体的な対応 ⑥ 文書化	i. AI 提供者への「共通の指針」の対応状況の説明 ii. 開発関連情報の文書化	i. AI 利用者への「共通の指針」の対応状況の説明 ii. サービス規約等の文書化	i. 関連するステークホルダーへの説明 ii. 提供された文書の活用及び規約の遵守

OED での Accountability は以下
The quality of being accountable; liability to account for and answer for one's conduct, performance of duties, etc. (in modern use often with regard to parliamentary, corporate, or financial liability to the public, shareholders, etc.); responsibility.
OEDでの Accountable は以下
Chiefly of persons (in later use also organizations, etc.): liable to be called to account or to answer for responsibilities and conduct; required or expected to justify one's actions, decisions, etc.; answerable, responsible.

<snip>

- **バイアス（bias）**【特定のオブジェクト，人々又はグループの対応において，他と比較したときの系統立った相違．**注釈1** 対応とは，認識，観察，表現，予測（**3.1.27**），意思決定など，あらゆる種類の行動である。】
  バイアスについては，ISO/IEC JTC1/SC42 WG3 N0013 にて杉村から貢献．１７２種類のバイアスの定義について共有を図った．
  ISO/IEC TR 24027 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making　が出版済
- Acquiescence bias, Actor observer bias, Ad hominem, Affect heuristic, Ambiguity effect, Anchoring, Anthropic bias, Apophenia, Argument from authority, Argument from Ignorance, Attentional bias, Attitude polarization, Attribute substitution, Attributional bias, Availability heuristic, Bandwagon effect, Base rate fallacy, Benefectance bias, Bias blind spot, Brand Loyalty, bystander effect, Catharsis, Choice-supportive bias, Clustering illusion, Cognitive bias, Cognitive Dissonance, Cognitive distortions, Confabulation, Confirmation bias, Congruence bias , Conjunction fallacy, Consistency bias, Contrast effect, Cryptomnesia,Cultural bias, Dunning-Kruger effect, Default standard unit bias, Déformation professionnelle , Disconfirmation bias, Dunbar‘s number, Dr Fox Effect, Egocentric bias, Emotional forecasting, Empathy gap, Endowment effect, Experimenter effect, Experimenter expectations, Exposure-suspicion bias, Extreme aversion, False consensus effect, False memory syndrome, Familiarity heuristic, Fluency (processing fluency), Focusing effect, Forer effect, Functional fixedness, Framing, Frequency illusion, Fundamental attribution error, Gambler’s fallacy, Generation effect(learning), Group-serving bias, Group think., Group attribution error, Halo effect, Herd instinct, Hindsight bias, Hostile media effect, Hyperbolic discounting, Illusion of asymmetric insight, Illusion of control, illlusory correlation, illusion of transparency, impact bias, Implicit cognition, Indoctrination, Inductive bias, Information bias, in-group bias, Irrational escalation, Introspection illusion, Isolation effect, Just-world phenomenon, Lake Wobegon effect, Learned Helplessness, List of cognitive biases, Loss aversion, Ludic fallacy, Magical thinking, Mere exposure effect, Mindset, Misinformation effect, Modesty bias, Negativity effect, Neglect of prior base rates effect, Neglect of probability, Normacy bias, Notational bias, Obsequiousness bias , Observer-expectancy effect, Oedipus effect, Omission bias, Optimism bias, Outcome bias, Out-group homogeneity bias, Overconfidence effect, Physical attractiveness stereotype, Picture superiority effect, Planning fallacy, Positive outcome bias, Positivity effect, Post-purchase rationalization, Primacy effect, Priming Effect,Projection bias, Pseudocertainty effect, Public goods game, Publication bias, Reactance, Recency effect, Regression fallacy, Reminiscence bump, Repetition bias, Representativeness heuristic, Response bias, Rosy retrospection, Selective Memory and selective reporting, Selective perception, Self-deception, Self-fulfilling prophecy, Self-efficacy, Self-serving bias, Selling out, Serial position effect, Spacing effect, Spotlight effect, Status quo bias, Strawman fallacy, Subadditivity effect, Subject-expectancy effect, Subjective validation, Supernormal Stimulus, Superstitions, Survivorship bias, System justification, Telescoping effect, Texas sharpshooter fallacy, The third-person effect, Trait ascription bias, True-believer syndrome, Ultimate attribution error, Ultimatum Game, Unacceptability bias, Unit bias, Univariate bias, Valence effect, Von Restorff effect, Wishful thinking, Worse-than-average effect, Zeigarnik effect, Zero-risk bias

- **制御**（control）【特定の目標を達成するための，プロセスに対する又はプロセス中における，目的をもった行動】
- **制御可能性**（controllability）**制御可能**（controllable）【人間又は他の外部のエージェントが，システムの機能に介入できるようにするAIシステムの特性】
- **説明可能性**（explainability）【AIシステムの結果に影響を与える重要な要因を，人間が理解可能な方法で表すAIシステムの特性．**注釈1**　とられた一連の行動が必ず最適であると実際に主張を試みることもなく，“なぜ”という疑問に答えることを目的としている。】
- **予測可能性**（predictability）【ステークホルダーが信頼できる，出力についての仮説を可能とするAIシステムの特性】

The fact that there's randomness here means that if we use the same prompt multiple times, we're likely to get different essays each time. And, in keeping with the idea of voodoo, there's a particular so-called "temperature" parameter that determines how often lower-ranked words will be used, and for essay generation, it turns out that a "temperature" of 0.8 seems best. (It's worth emphasizing that there's no "theory" being used here; it's just a matter of what's been found to work in practice.

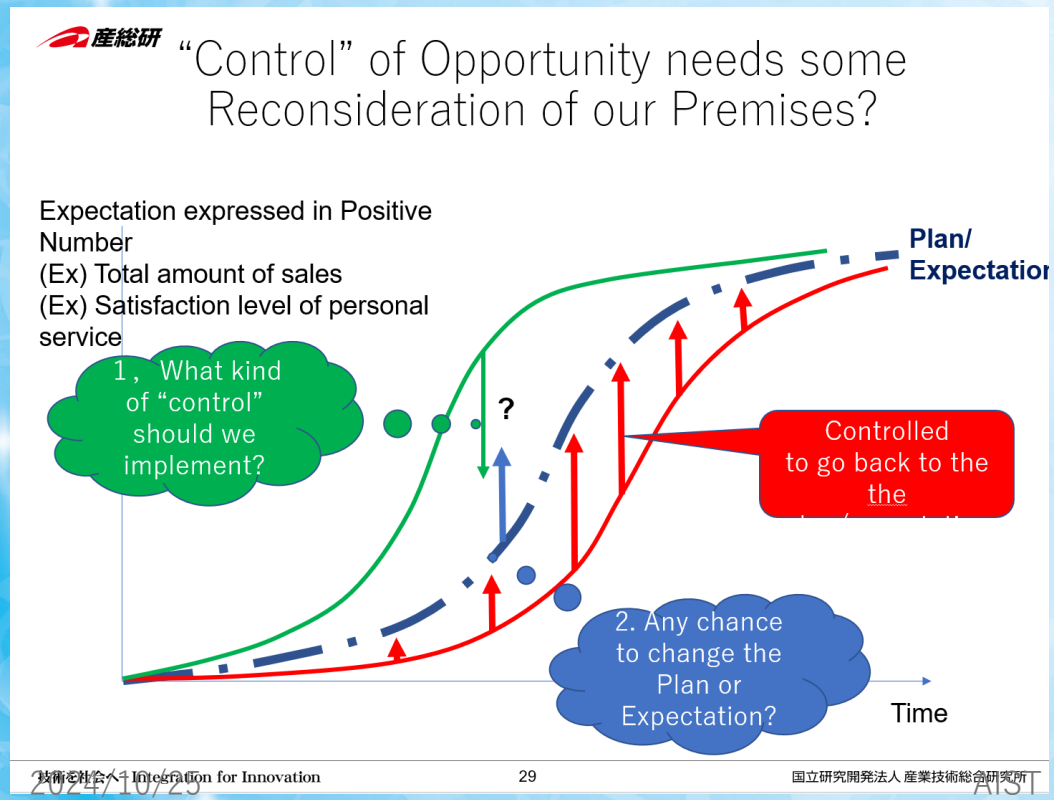
Wolfram, Stephen. What Is ChatGPT Doing ... and Why Does It Work? (English Edition) (pp.9-10). Wolfram Media, Inc. . Kindle 版.

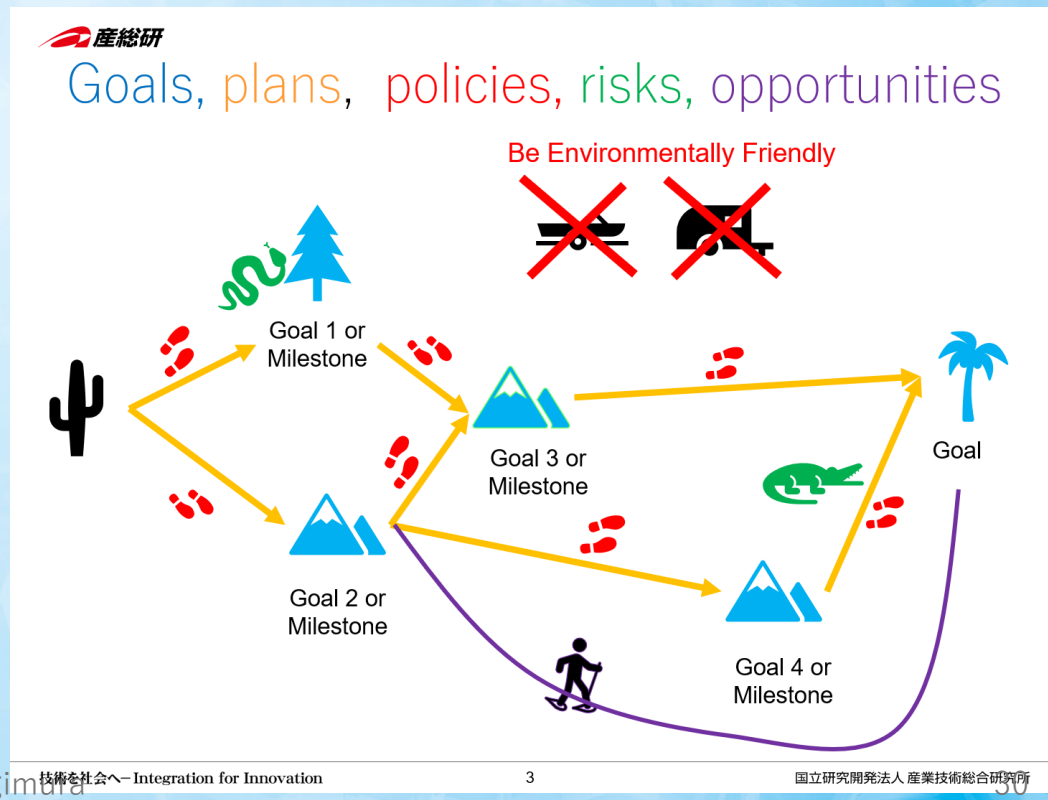
ここでランダム性があるということは、同じプロンプトを複数回使用すると、毎回異なるエッセイが得られる可能性が高いことを意味します。
　そして、ブードゥー教の考えに従って、下位の単語がどのくらいの頻度で使用されるかを決定する特定のいわゆる「温度」パラメーターがあり、エッセイの生成には、**0.8 の「温度」**が最適であることがわかりました。(ここで使用されているのは「理論」ではないことを強調しておきます。重要なのは、**実際に何が機能することが判明したかというだけです。**

<snip>

- **リスク（risk）【**目的に対する不確かさの影響. **注釈1** 影響とは，期待されていることからかい（乖）離することをいう。影響には好ましいもの，好ましくないもの又はその両方の場合がある。影響は，機会又は脅威を示したり，創り出したり，もたらしたりする可能性がある。**注釈2** 目的は，様々な側面及びカテゴリーをもつ場合があり，様々なレベルで適用可能である。**注釈3** 一般に，リスクは，リスク源，起こり得る事象及びそれらの結果並びに起こりやすさとして表される。**】**

⇒　リスクには，Opportunity と Hazard の考え方があるが，Britz-Scaling のように, ポジティブに例えば売り上げが予想以上に上がるような場合が，IT事業一般では見られ，特にAI利用により，ポジティブサイドへ予想以上に触れることも想定する必要があり，これらの点について，バランス良く検討することを提案し，多くの面で採用されています.





- **ステークホルダー**（stakeholder）
  - 意思決定又は活動に影響を与える，影響を受ける，又は影響を受けると感じる可能性のある，あらゆる個人，集団，又は組織

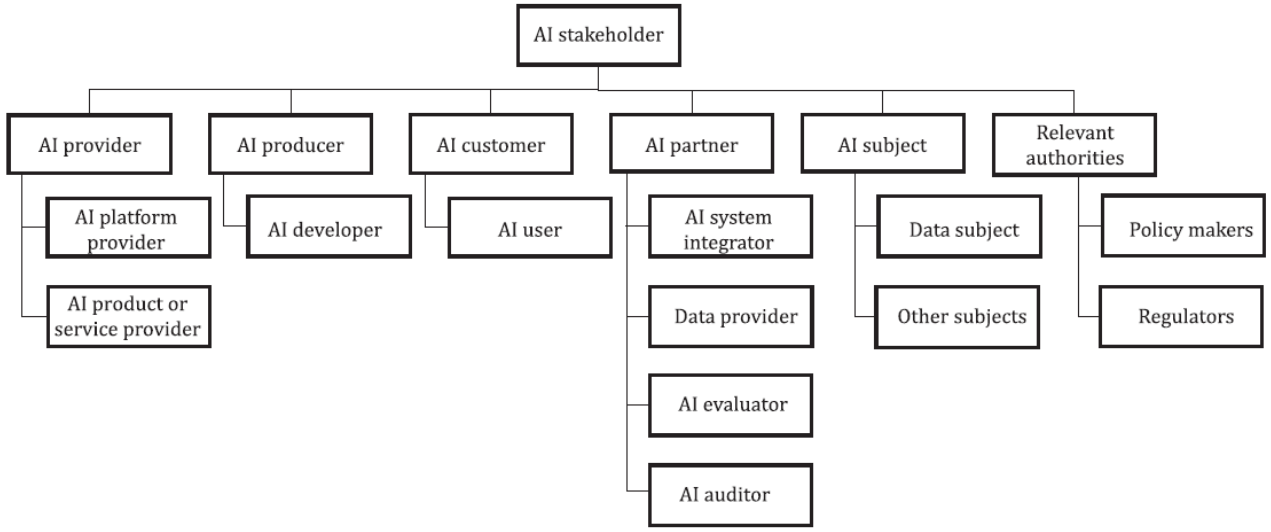


Figure 2 — AI stakeholder roles and their sub-roles

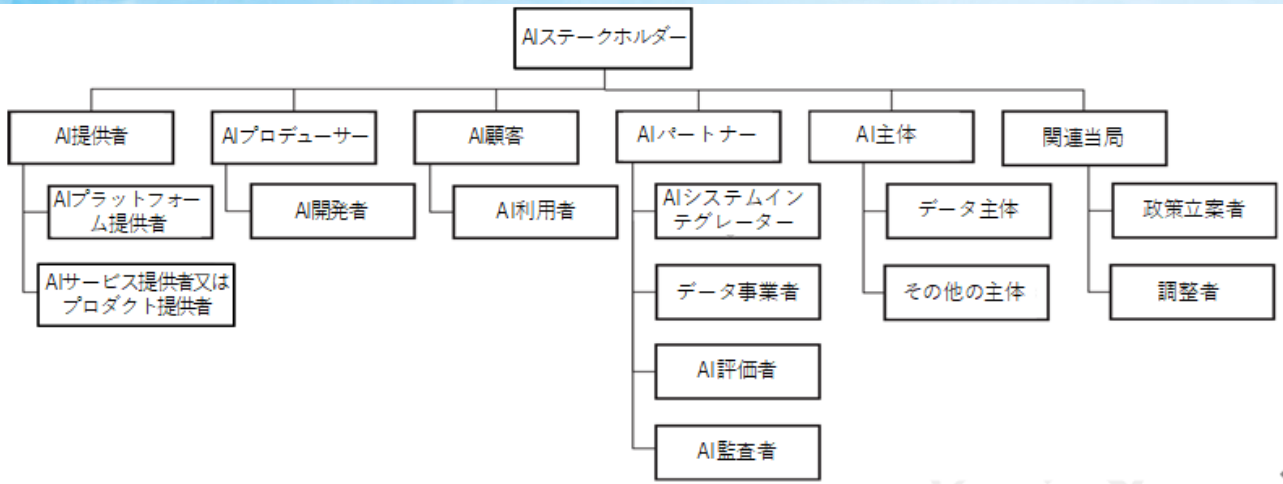


図2－AIステークホルダーの役割及びその副次的な役割

ISO/IEC 22989:2022 では，
Stakeholder role という整理で
AIシステムに関係する個人，集団，
組織を分類定義している

- **透明性**（transparency）【＜組織＞適切な活動及び意思決定が，包括的で，アクセス可能で，理解可能な方法で関連するステークホルダー（**3.5.13**）に伝達される組織の特性．**注釈1**　活動及び意思決定の不適切な伝達は，セキュリティ，プライバシー又は機密性の要求事項に違反する可能性がある。】
- **透明性**（transparency）【＜システム＞システムに関する適切な情報を関連ステークホルダー（**3.5.13**）に利用可能とするシステムの特性．**注釈1**　システムの透明性のための適切な情報には，特徴，性能，制限，コンポーネント，手順，対策，設計目標，設計の選択及び仮定，データソース及びラベリングの規約などの側面が含まれる場合がある。**注釈2**　システムの一部の側面を不適切に開示すると，セキュリティ，プライバシー又は機密性の要求事項に違反する可能性がある。】

⇒透明性については，AI事業者ガイドラインにて，詳細が定義されている．
また，注釈 pp.16 において，以下のように詳細が紹介されている．
「透明性については、諸外国でも様々な定義がある。例えば、NIST,Artificial Intelligence Risk Management Framework（January 2023）では、透明性（システムで何が起きたかについて答えられること）、説明可能性（システムでどのように決定がなされたかについて答えられること）及び解釈可能性（なぜその決定がされたかについてその意味又は文脈について答えられること）に分類されており、European Commission, ETHICS GUIDELINES FOR TRUSTWORTHY AI （April 2019）では、トレーサビリティ、説明可能性及びコミュニケーションが取り上げられている。また、国際標準(ISO/IEC JTC1/SC42)では、透明性（適切な情報が関係者に提供されること）と定義されている。本文書では、情報開示に関する事項を広く「透明性」とする。」

6） 透明性	①検証可能性の確保 ②関連するステークホルダーへの情報提供 ③合理的かつ誠実な対応 ④関連するステークホルダーへの説明可能性・解釈可能性の向上	i. 検証可能性の確保 ii. 関連するステークホルダーへの情報提供	i. システムアーキテクチャ等の文書化 ii. 関連するステークホルダーへの情報提供	i. 関連するステークホルダーへの情報提供

- **トラストワージネス**
  - 検証可能な方法でステークホルダーの期待を満たす能力
- **注釈1**　状況又は分野，並びに特定の製品又はサービス，使用されるデータ及び技術に応じて，様々な特性が適用され，ステークホルダーの期待が満たされることを確認するための検証が必要となる。
- **注釈2**　トラストワージネスの特性には，例えば，信頼性，可用性，回復力，セキュリティ，プライバシー，安全，説明責任，透明性，インテグリティ，真正性，品質，ユーザビリティがある。
- **注釈3**　トラストワージネスは，サービス，製品，技術，データ及び情報に適用可能な属性であり，ガバナンスにおいては，組織に対しても同様である。

EU AI-ACT 関連のCEN/CENELEC JTC21 における作業では，Trustworthiness Frameworkにおいて以下のSTD Rq をまとめる位置づけとなる可能性が高いと思われる

- Governance and quality of datasets used to build AI systems
- Record keeping through built-in logging capabilities in AI systems
- Transparency and information to the users of AI systems
- Human oversight of AI systems
- Accuracy specifications for AI systems
- Robustness specifications for AI systems
- Cybersecurity specifications for AI systems

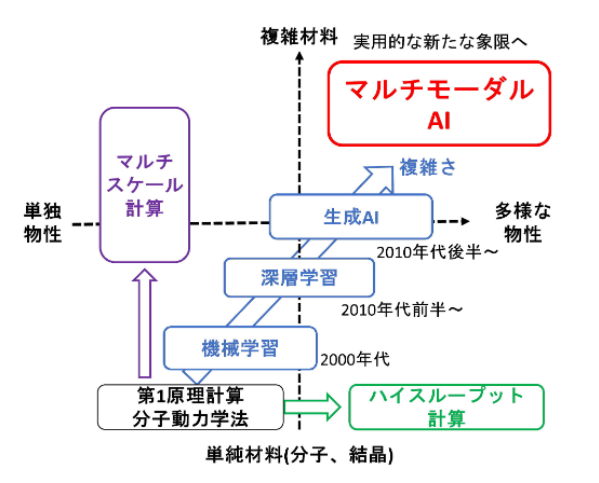
- **自然言語処理に関する用語**
- **自動要約**（automatic summarization）, **対話マネジメント**（dialogue management）, **感情認識**（emotion recognition）, **情報検索，IR**（information retrieval，**IR**）, **機械翻訳，MT**（machine translation，**MT**）, 固有表現抽出，NER（named entity recognition，NER）, **自然言語**（natural language）, **自然言語生成，NLG**（natural language generation，NLG）, **自然言語処理，NLP**（natural language processing）, **自然言語処理，NLP**（natural language processing）, **自然言語理解，NLU**（natural language understanding，NLU）, **代替用語：自然言語理解**（**natural language comprehension**）, **光学的文字認識，OCR**（optical character recognition，OCR）, **POSタグ付け，品詞タグ付け**（part-of-speech tagging）, **質問応答**（question answering）, **関係抽出**（relationship extraction，relation extraction）, **心情分析**（sentiment analysis）, **音声認識**（speech recognition）, **代替用語：音声テキスト変換，STT**（speech-to-text，STT）, **音声合成**（speech synthesis）, **代替用語**：テキスト音声合成，TTS（text-to-speech）

生成AIについては，種々の用語の追記がISO/IEC 22989 Amendment 2 Generative AI にて検討されている
以下は，あくまで例（今後，追加，削除が発生しますので，最終版とは乖離する可能性があります）

- Generative artificial intelligence system
- Generative model
- Probability distribution
- Foundation model
- Large language model
- Self-supervised machine learning
- Generative adversarial network
- Transformer
- Variational autoencoder
- Prompt
- Etc.

- **コンピュータビジョン**（computer vision）【画像又は映像を表すデータを取得，処理，解釈する機能ユニットの能力．**注釈1** コンピュータビジョンには，視覚シーンのデジタル画像を作成するためのセンサーの使用が含まれる。これには，赤外線画像のような可視光の波長を超える波長を捕捉したもののような画像を含めることが可能である。】
- **顔認識**（face recognition）【人の顔の保存画像と実際の顔の画像とを比較し，一致するものがあればそれを示し，また，何らかのデータがあればそれを示し，顔が属する人物を識別する自動的なパターン認識】
- **画像**（image）【<デジタル>視覚的に提示することを意図したグラフィカルなコンテンツ．**注釈1** これには，個々のピクセルで構成されるフォーマット（例えば，ペイントプログラム又は撮影手段によって作成されるフォーマット）及び数式で構成されるフォーマット（例えば，スケーラブルなベクトル描画として作成されるフォーマット）を含むが，これらに限定されない，任意の電子フォーマットで符号化されるグラフィックも含まれる。】
- **画像認識**（image recognition）【画像（**3.7.3**）の中でオブジェクト，パターン又は概念を分類する画像分類処理】

AI関連用語として，モダリティという言葉で，情報の種類（文字列・音声，画像・動画，３D情報，におい振動，等々）を指すことがあるが，昨今は複数のモダリティを同時に扱う研究が進んでいる．



出典：https://unit.aist.go.jp/ncdrc/ja/multimodal_ai/AI_top.html

認知科学的知識
Bias の定義と種類

Automation bias	the tendency to over-rely on automation	https://pmc.ncbi.nlm.nih.gov/articles/PMC3240751/
Bias	systematic difference in treatment of certain objects, people, or groups in comparison to others	ISO/IEC TR 24027
Human cognitive bias	Bias that occurs when humans are processing and interpreting information	ISO/IEC TR 24027

認知バイアスについては，それそのものを，合理的ではないとする定義がある．しかし，SC42における議論では，合理的かどうかは，バイアスがもたらす情報処理の結果についての評価が関わっており，定義自体，価値観（合理的かどうか）を持ち込むべきではないとの議論があった．

日本（その時は杉村からですが）からも，情報処理を非常に短時間で行うためのものであり，確かに，沈思黙考するものでは無いが，結果が合理的かどうかは，思考・行動の評価軸をどう定めるかにかかわるという見解を具申．結果的に，現在の定義は，合理性についての記述は入っていない．

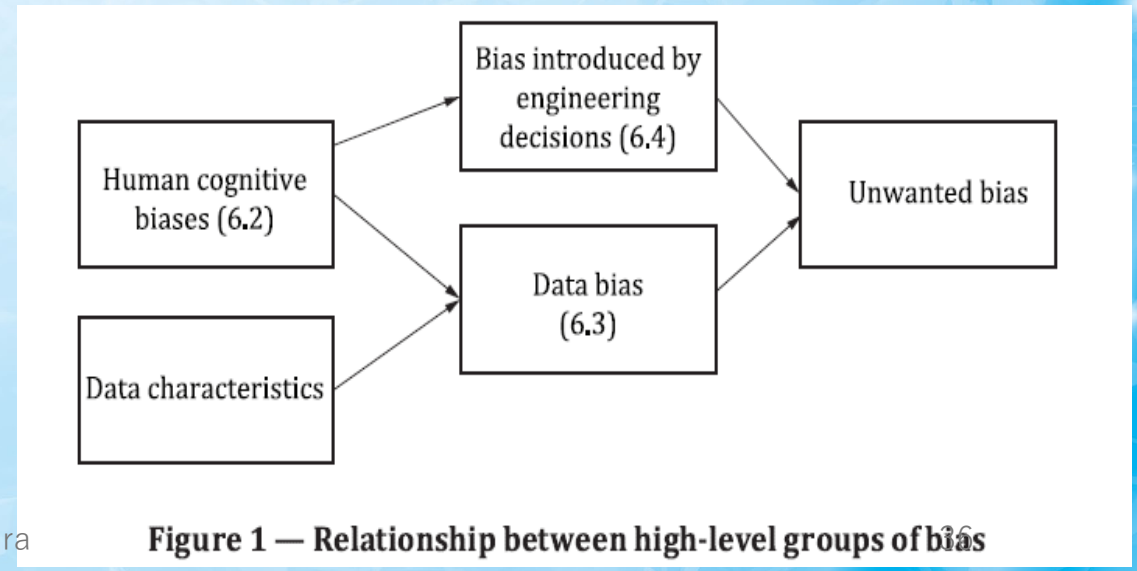
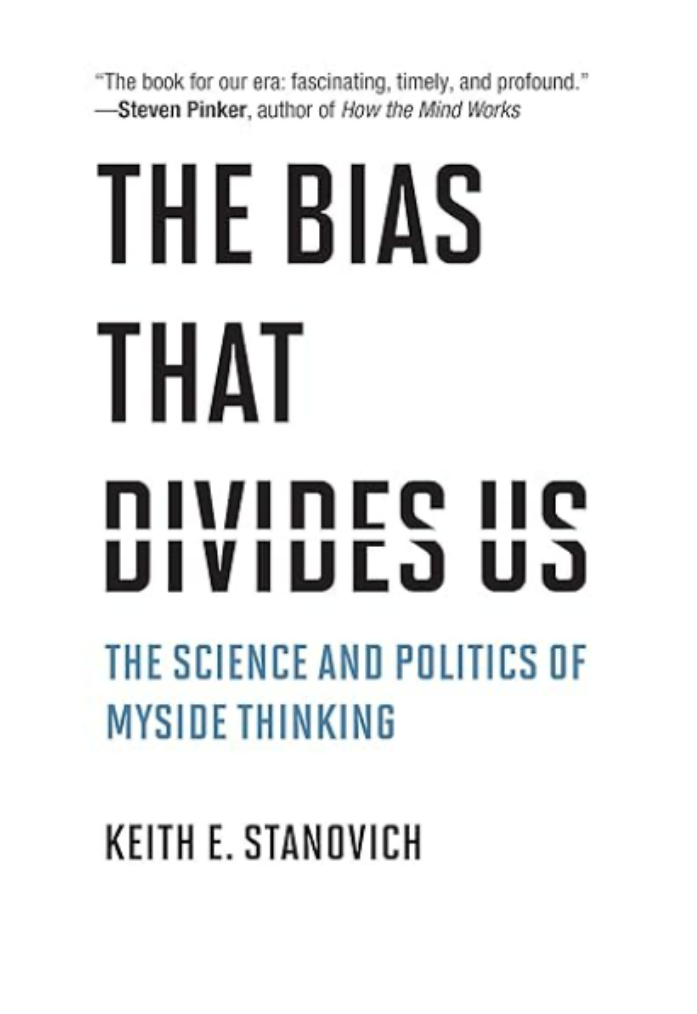


Figure 1 — Relationship between high-level groups of bias

Bias の研究は，まだ途上です



バイアスの盲点は、それ自体がマイサイド・バイアスの一種であるように見えるが、これは素朴なリアリズムの一種に由来するものである。他人の判断が自分の判断と異なるとき、私たちはその違いの原因を、証拠の正当な別解釈としてではなく、彼らの側にあるバイアスとみなす。

The bias blind spot itself appears to be a form of myside bias that may derive from a form of naive realism which causes us to believe that we perceive the world objectively (Keltner and Robinson 1996; Robinson et al. 1995; Skitka 2010). When other people's judgments differ from our own, we view the source of that difference as bias on their part rather than as their legitimate alternative interpretations of the evidence.

Stanovich, Keith E.. The Bias That Divides Us: The Science and Politics of Myside Thinking (English Edition) (p.95). MIT Press. Kindle 版.

管理策について，特徴的な内容（例）

Annex B において特徴的な内容（例）
組織の役割と責任

- risk management; ISO/IEC 23894:2023　Information technology — Artificial intelligence — Guidance on risk management 開発中
- AI system impact assessments; ISO/IEC 42005 Information technology — Artificial intelligence — AI system impact assessment
- asset and resource management;
- security;
- safety; ISO/IEC TR 5469:2024 Artificial intelligence — Functional safety and AI systems, ISO/IEC AWI TS 22440-1,2,3 Functional safety and AI systems (Requirements, guidance, examples of application)　開発中
- privacy;
- development;
- performance;
- human oversight; ISO/IEC AWI 42105 Information technology — Artificial intelligence — Guidance for human oversight of AI systems, ISO/IEC PWI 18966 Artificial intelligence — Oversight of AI systems 開発中
- supplier relationships;　一般に Lattice, 個人情報の越境など課題は多い
- demonstrate its ability to consistently fulfil legal requirements; 欧州では Machinery Regulation, Data Act, 多重
- data quality management (during the whole life cycle)：ISO/IEC 5259-1,2,3,4,5,6, ISO/IEC 8183:2023 Information technology — Artificial intelligence — Data life cycle framework開発中

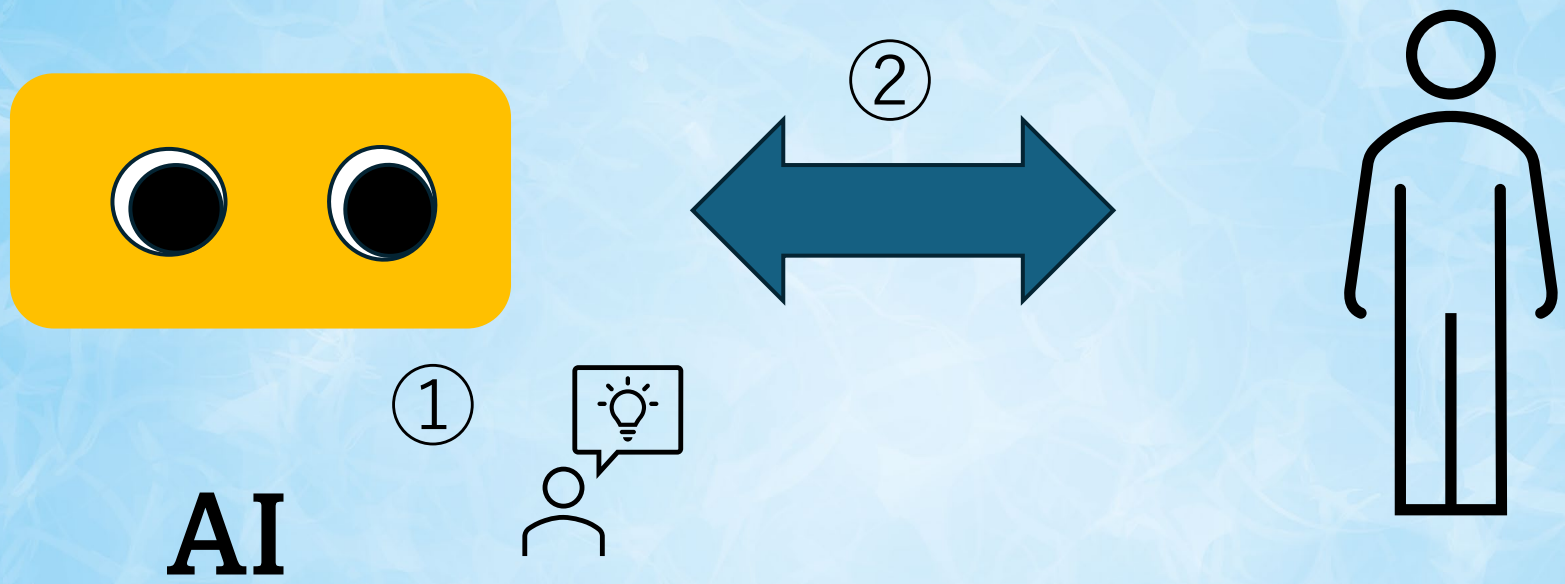
Annex B において特徴的な内容（例）
データ資源と管理策

- the provenance of the data; ISO/IEC AWI 24970 AI system logging開発中
- the date that the data were last updated or modified (e.g. date tag in metadata);
- for machine learning, the categories of data (e.g. training, validation, test and production data);
- categories of data (e.g. as defined in ISO/IEC 19944-1);
- process for labelling data;
- intended use of the data;
- quality of data (e.g. as described in the ISO/IEC 5259 series2));
- applicable data retention and disposal policies;
- known or potential bias issues in the data; ISO/IEC TS 12791 Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks開発中
- data preparation

Annex B において特徴的な内容（例）
AIシステムの個人，個人のグループへのインパクト

- fairness;
- accountability;　説明責任 or アカウンタビリティ
- transparency and explainability; DIS 6254Information technology — Artificial intelligence — Objectives and approaches for explainability and interpretability of ML models and AI systems開発中
- security and privacy;
- safety and health;
- financial consequences;
- accessibility; —
- human rights

２つの分かりやすさ



専門家への説明可能な出力が
求められている

人間にとって，解るとは？
理解できるとは？
技術の視点から定義が必要

Explainability （説明可能性訳語は杉村）

２つの意味

① 専門家が，AIが何故，特定の出力を出したのか説明できること
② 非専門家へ，AIの出力について分かりやすく説明できること

①が出来て，それを元に②の説明が可能.

２０１２年以来，AIのR&Dで課題となっているのは，①.
ただし，標準化やAI関連ポリシー，ガイドラインなどでは，
①，②，双方課題として認識されることが多い
②を技術的に達成できるかが課題

大阪大学　原聡教授

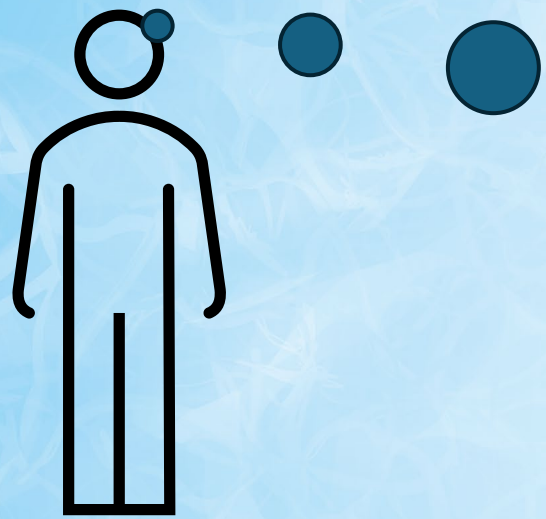
①

- **研究成果のポイント**
- AI が下した判断の根拠を明示する手法として、AI が判断を下すに際して参照した過去の類似事例を人間に提示するアプローチが注目されている。
- 類似事例を見つけるための技術的な方法は複数の提案があるが、人間への説明としてどの方法が適切かはわかっていない。
- 予測根拠として最低限満たすべき要件を定式化したところ、非常に人気の高い手法も含め、既存の手法の多くは要件を満たさないことを指摘した。
- 本研究を足がかりに、人間社会に自然に溶け込むAIの研究開発が加速することが期待される。

②

出典：https://resou.osaka-u.ac.jp/ja/research/2021/20210427_1

解る→人を技術的に把握することが必要



一般ユーザ

Resource Situation
インターネット上の行動から推定可能
良く使う単語，文章，良く触れている
物，場所，交流のある友達，社会，
等々

人間にとって，分かるとは？
理解できるとは？
技術の視点から定義が必要

Efficiency：聞き手が分かっていることを新しい情報として伝えない．
Effectiveness: 聞き手にとって馴染みのある言葉や絵，比喩等を用いて，「響く」ように伝える
Attractiveness: 聞き手が好む言葉や絵，比喩等を用いて「好ましく」伝える

上記定義は，杉村自身によるものです．

一般人にとって「わかる」とは何か？

- 「人工知能と人間」長尾真，岩波新書，2017
  - 第四章 理解することへの挑戦，
  - ２ わかるということ

説明においてもっと大切なことは次のようなことである．説明する側がいくら詳しく説明を行ったとしても，それがすなおに聞き手の頭に入り，納得するという状態にならなければ，それは本当の意味での説明にはなっていない．<snip> 相手にわからせるためには，相手に対してその人の知らない新しい情報を送らねばならない．そのためには相手は何を知っていて何を知らないか，本当に知りたがっていることは何かを知らねばならない．

  - ３知識

人工知能の立場は，知識は外界から獲得されるものであり，そのために認識という行動があること，また認識は知識に支えられて行われるものであるという相互依存の関係にあるものである．<snip>われわれが知識と総称しているものはいくつかの性質のちがった知識に区別するのが便利である．まず**事実知識と推論知識**に分けられる．後者は因果律といってもよいもので，<snip> 事実知識はまた常識的知識と専門的知識に分けられる．

日常の一般法則：（実は，部分情報に基づき，暗黙の前提あり）

- 雨が降れば傘をさす：一般法則A
- 風が強ければ傘はさせない：一般法則B
- 風雨の強い時，傘は？　（赤字：隠れた暗黙の前提）
  - 雨が降り，かつ，風が弱い場合には傘をさす．
  - 雨が降り，かつ，風が強い場合には，傘はささない

- 法則：Pattern も，普遍性を支える**暗黙の前提**を持つ．主なものは，環境要因．それらが成り立つ文脈
  - 固体対象
  - 性質
  - 変化
  - 因果法則

AIシステムの利用環境についても，暗黙の前提を，明示的に記載する方向で国際標準化では議論が進んでいます　（Operational Design Domain）

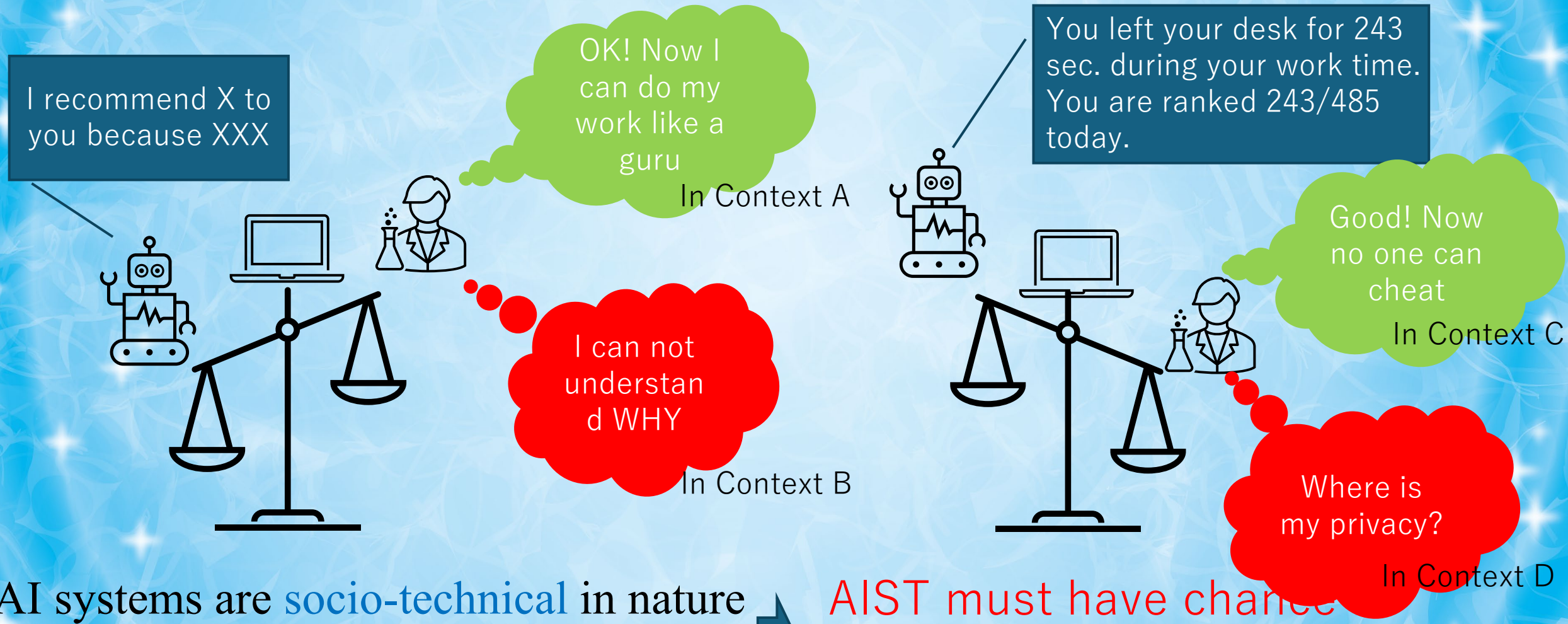
もう一つの課題

- 人，人間のチーム，人間社会へ，AIをどう組み込めばよいのか？
  - Human Machine Teaming（省３１プロジェクト），Partnership 等
- 例えば，ミッションクリティカルな状況，日常等で，どのような組み合わせが，今，可能で，将来は，どのような形が好ましい（良い結果を出しやすい，人間をスポイルしない，効率は高い，等）のか？
- AIと人間のベストミックスを，科学的・技術的に設計するには，

👍**人間についても，技術的に語る必要があるはず**だが？！

  - 新しいことでは無い！：工場のラインの効率を上げるのに，作業者の工程を詳しく分析して，作業計画を作成している．
- 作業→知的作業　へ範囲が拡大される「だけ」と思うと，簡単かも知れないが．

Socio-Technical Nature of AI
- Toward understanding ourselves embedded in contexts
-



AI systems are socio-technical in nature (from NIST AI-RMF Vr2. pp.1)

AIST must have chance to lead Standards, and R&D

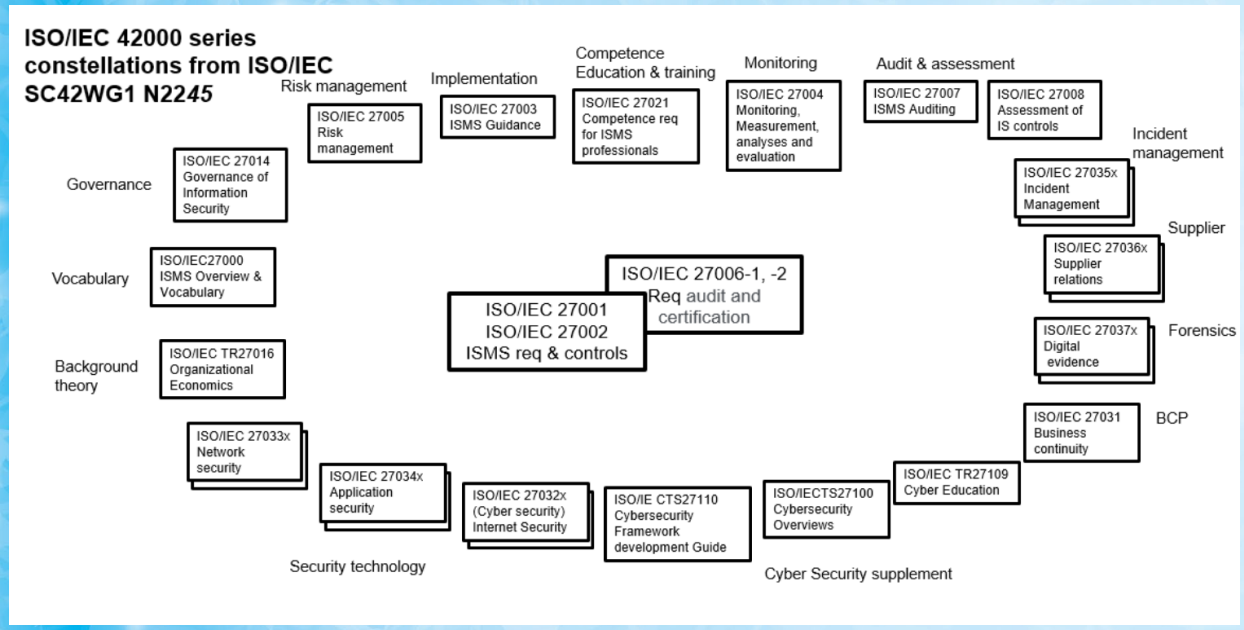
Annex D
AIマネジメントシステムのドメイン横断利用

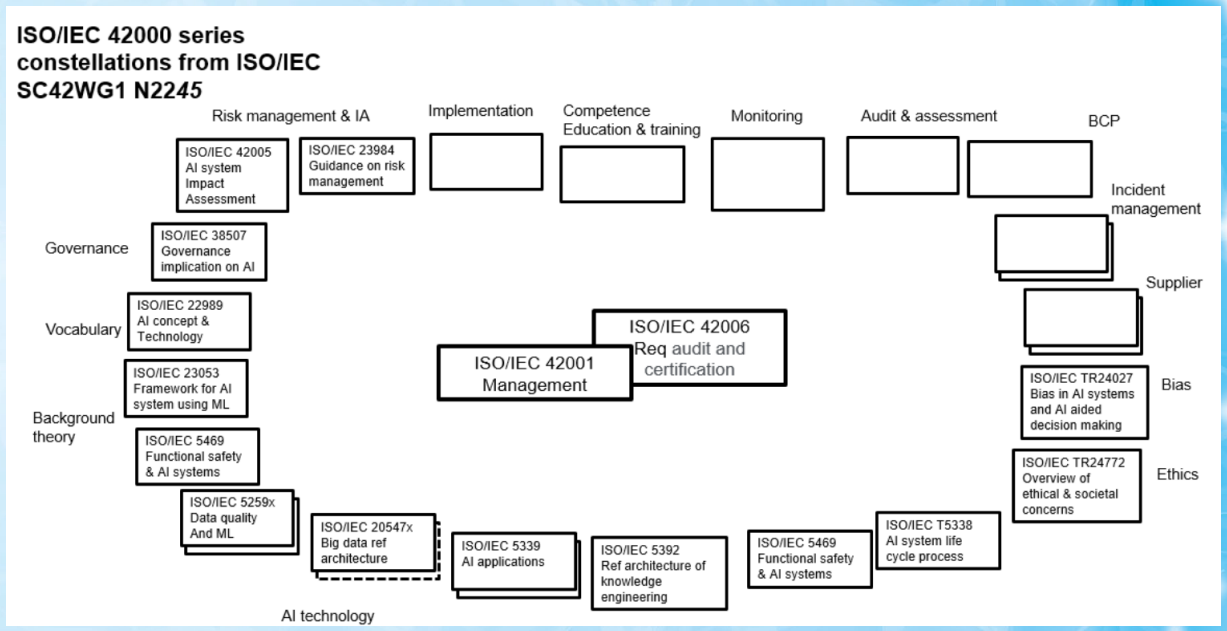
- Health; ISO/IEC JTC1/SC42 JWG3 にてドメイン利用検討中
- Defence;
- Transport; AI ACTではtransport はexemption だが，周辺環境はAI ACT に関連するとの議論あり
- Finance; SC42と ISO/TC 68 の間でCat A リエゾン
- Employment;
- Energy;

今わかっている範囲で、次に何が起こるか？

- Convergence of Understanding of Issues
  - NIST AI Symposium (2024/9/23,24) の論点
  - OECD の動き：GPAI+OECD
- マルチモーダルＡＩ
- 連合学習
- 分散協調AI
- 人間についての，技術的観点からの知見の蓄積

42001に続く標準群
ご期待下さい





先行例　2700Xを参考

ご静聴ありがとうございました