d- real SFI centre for research training in digitally-enhanced reality

Centre for Research Training

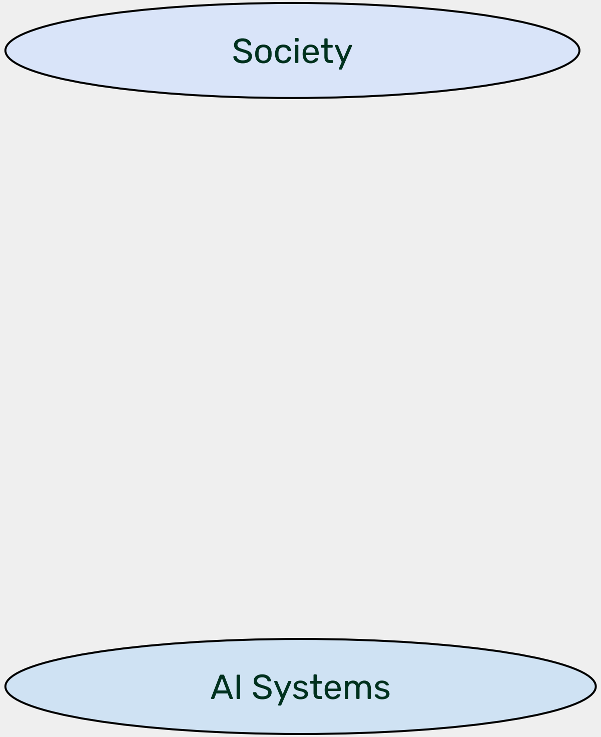
Science Foundation Ireland Sfi For what's next

Technical standards

for AI governance

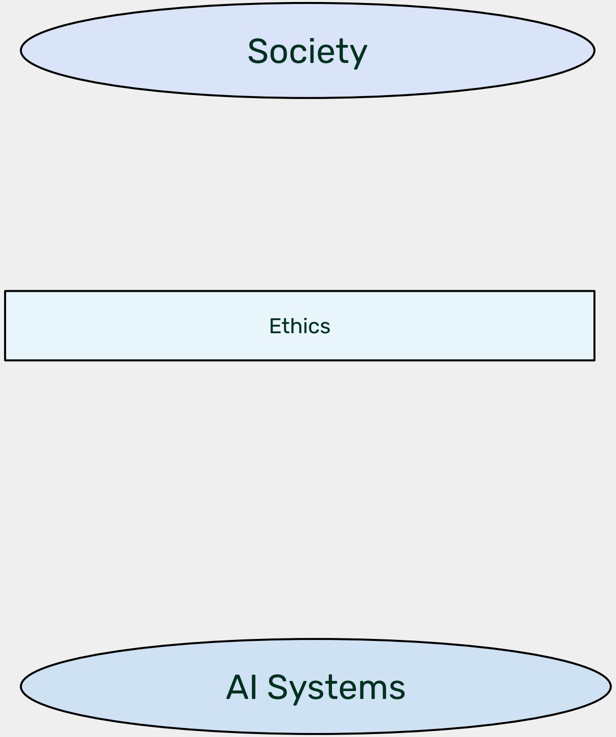
อาทิตย์ สุริยะวงศ์กุล
Arthit Suriyawongkul, ADAPT Centre, Trinity College Dublin
SFI Centre for Research Training in Digitally-Enhanced Reality (d-real)

20 September 2024 - Office of the Council of State International Symposium 2024, Bangkok



Society

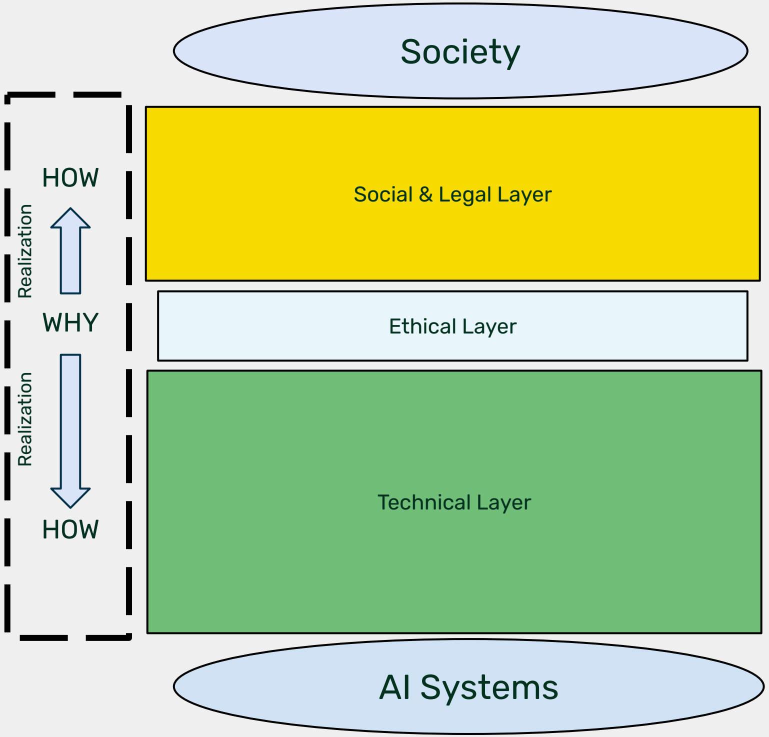
AI Systems

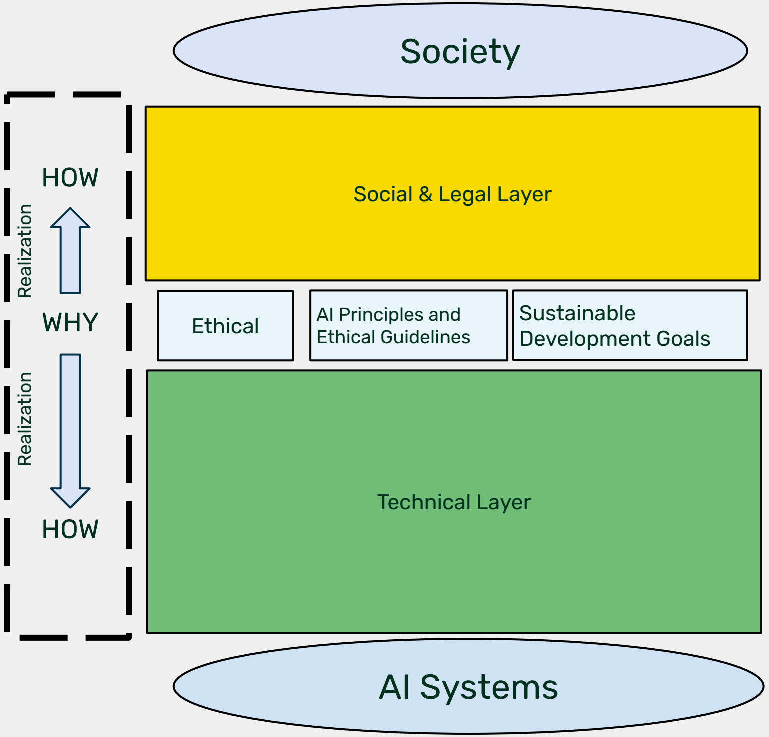


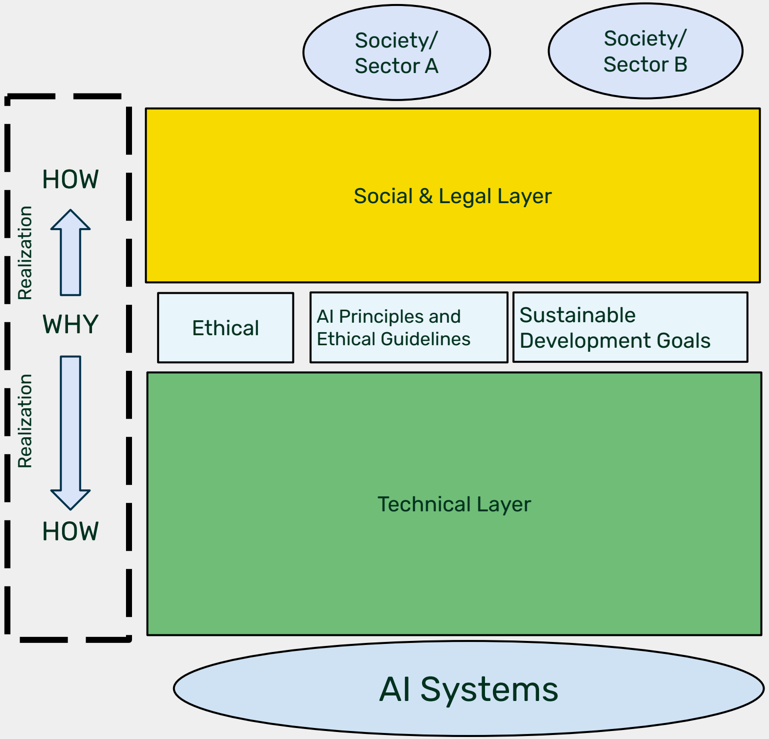
Society

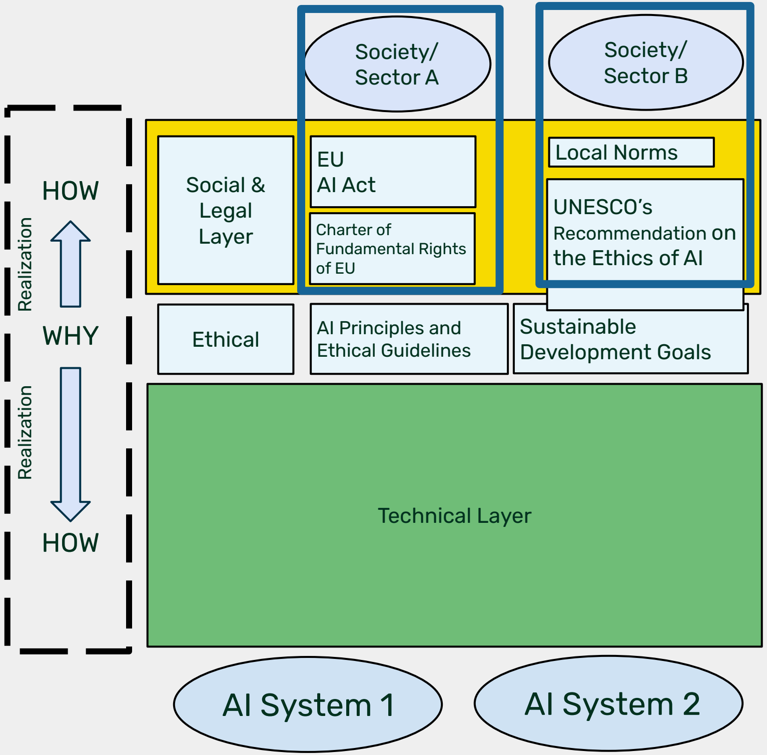
Ethics

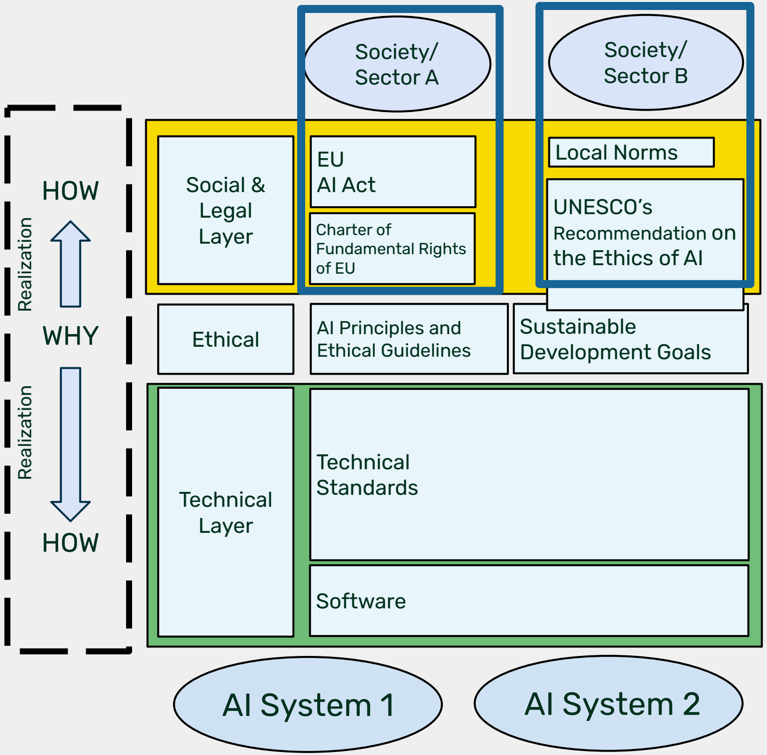
AI Systems

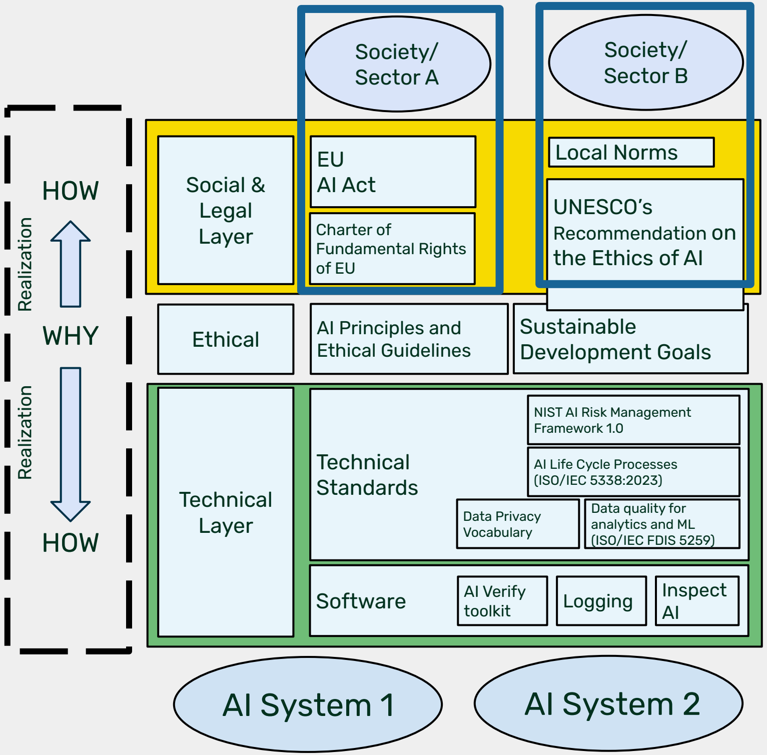

Society
HOW
Realization
WHY
Realization
HOW
Social & Legal Layer
Ethical Layer
Technical Layer
AI Systems

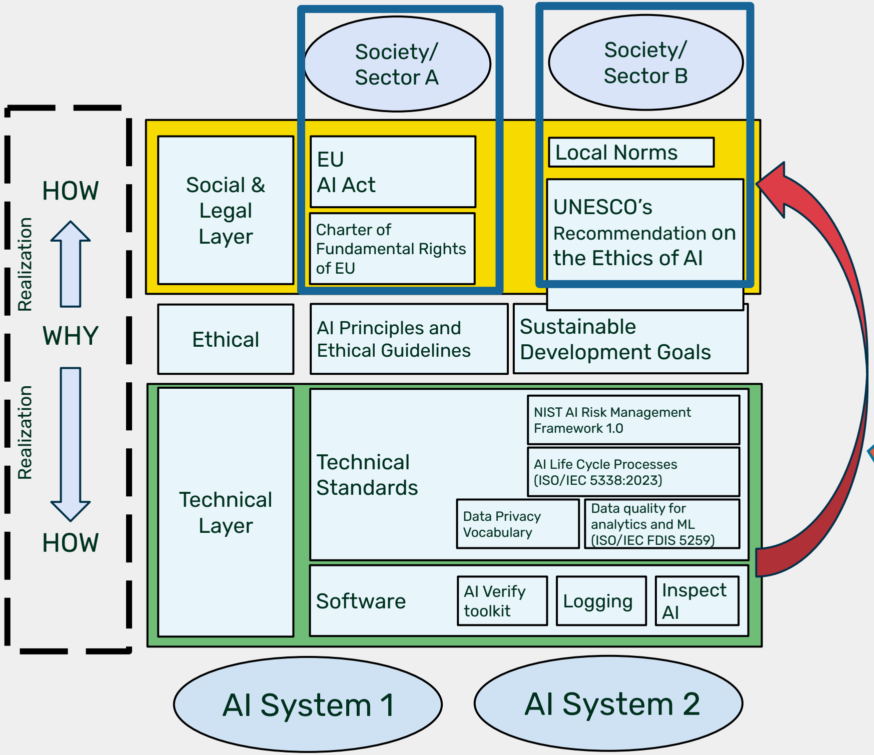

Society
Social & Legal Layer
HOW
Realization
WHY
Realization
HOW
Ethical
AI Principles and Ethical Guidelines
Sustainable Development Goals
Technical Layer
AI Systems


Society/
Sector A
Society/
Sector B
HOW
Realization
WHY
Realization
HOW
Social & Legal Layer
Ethical
AI Principles and
Ethical Guidelines
Sustainable
Development Goals
Technical Layer
AI Systems


Society/
Sector A
Society/
Sector B
HOW
Realization
WHY
Realization
HOW
Social &
Legal
Layer
EU
AI Act
Charter of
Fundamental Rights
of EU
Local Norms
UNESCO's
Recommendation on
the Ethics of AI
Ethical
AI Principles and
Ethical Guidelines
Sustainable
Development Goals
Technical Layer
AI System 1
AI System 2


Society/ Sector A
Society/ Sector B
HOW
Realization
WHY
Realization
HOW
Social & Legal Layer
EU AI Act
Charter of Fundamental Rights of EU
Local Norms
UNESCO's Recommendation on the Ethics of AI
Ethical
AI Principles and Ethical Guidelines
Sustainable Development Goals
Technical Layer
Technical Standards
Software
AI System 1
AI System 2


Society/ Sector A
Society/ Sector B
HOW
Realization
WHY
Realization
HOW
Social & Legal Layer
EU AI Act
Charter of Fundamental Rights of EU
Local Norms
UNESCO's Recommendation on the Ethics of AI
Ethical
AI Principles and Ethical Guidelines
Sustainable Development Goals
Technical Layer
Technical Standards
NIST AI Risk Management Framework 1.0
AI Life Cycle Processes (ISO/IEC 5338:2023)
Data Privacy Vocabulary
Data quality for analytics and ML (ISO/IEC FDIS 5259)
Software
AI Verify toolkit
Logging
Inspect AI
AI System 1
AI System 2



HOW

Realization

WHY

Realization

HOW

Society/ Sector A

Society/ Sector B

Social & Legal Layer

EU AI Act

Charter of Fundamental Rights of EU

Local Norms

UNESCO's Recommendation on the Ethics of AI

Ethical

AI Principles and Ethical Guidelines

Sustainable Development Goals

Technical Layer

Technical Standards

NIST AI Risk Management Framework 1.0

AI Life Cycle Processes (ISO/IEC 5338:2023)

Data Privacy Vocabulary

Data quality for analytics and ML (ISO/IEC FDIS 5259)

Software

AI Verify toolkit

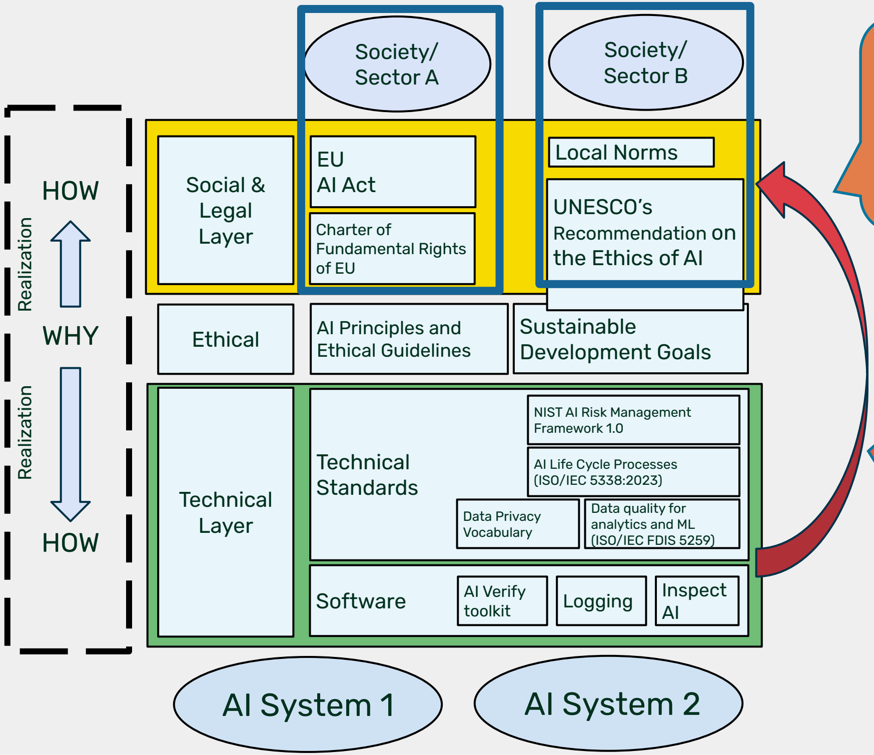
Logging

Inspect AI

AI System 1

AI System 2

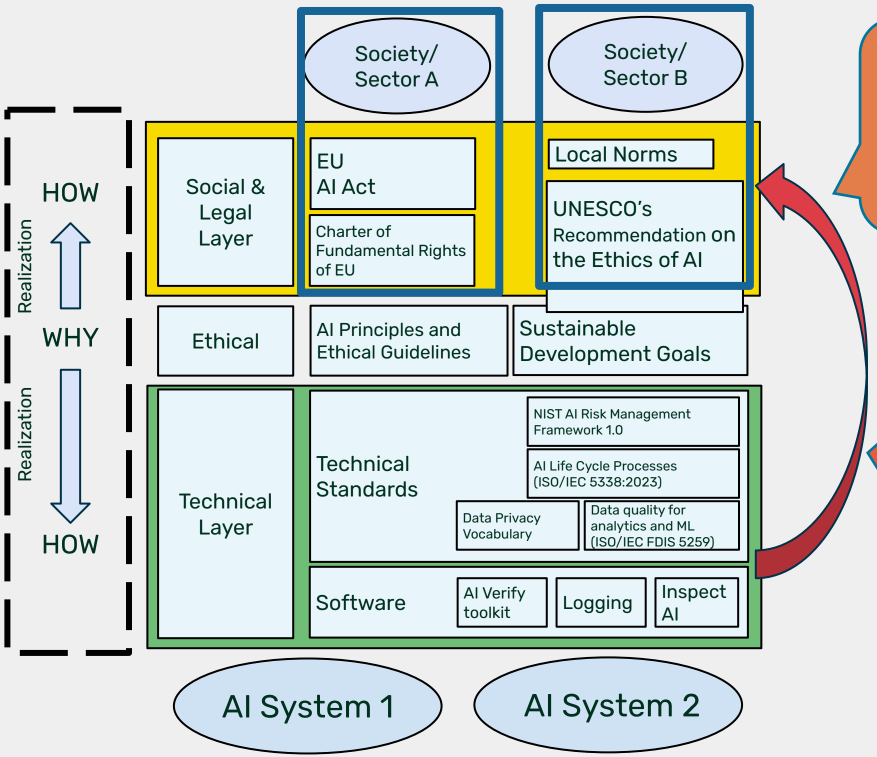
Technology to support the realization of AI principles


Society/
Sector A
Society/
Sector B
HOW
Realization
WHY
Realization
HOW
Social & Legal Layer
EU
AI Act
Charter of Fundamental Rights of EU
Local Norms
UNESCO’s Recommendation on the Ethics of AI
Ethical
AI Principles and Ethical Guidelines
Sustainable Development Goals
Technical Layer
Technical Standards
NIST AI Risk Management Framework 1.0
AI Life Cycle Processes (ISO/IEC 5338:2023)
Data Privacy Vocabulary
Data quality for analytics and ML (ISO/IEC FDIS 5259)
Software
AI Verify toolkit
Logging
Inspect AI
AI System 1
AI System 2

Adaptable and exchangeable across different jurisdictions

Technology to support the realization of AI principles

A Layered Model for AI Governance


Society/ Sector A
Society/ Sector B
HOW
Realization
WHY
Realization
HOW
Social & Legal Layer
EU AI Act
Charter of Fundamental Rights of EU
Local Norms
UNESCO's Recommendation on the Ethics of AI
Ethical
AI Principles and Ethical Guidelines
Sustainable Development Goals
Technical Layer
Technical Standards
NIST AI Risk Management Framework 1.0
AI Life Cycle Processes (ISO/IEC 5338:2023)
Data Privacy Vocabulary
Data quality for analytics and ML (ISO/IEC FDIS 5259)
Software
AI Verify toolkit
Logging
Inspect AI
AI System 1
AI System 2

Adaptable and exchangeable across different jurisdictions

Technology to support the realization of AI principles

Gasser, U., Almeida, V.A.F., 2017. A Layered Model for AI Governance. IEEE Internet Computing. https://doi.org/10.1109/MIC.2017.4180835

AI Accountability

Intergovernmental

ASEAN Guide on AI Governance and Ethics

1. Transparency and Explainability
2. Fairness and Equity
3. Security and Safety
4. Human-centricity
5. Privacy and Data Governance
6. Accountability and Integrity
7. Robustness and Reliability

National research body /
Grant-making agency

NSTDA AI Ethics Principles

1. Privacy
2. Security and Safety
3. Reliability
4. Fairness and non-discrimination
5. Transparency and Explainability
6. Accountability
7. Human Oversight and Human Agency

Governmental

Thailand AI Ethics Principles (MDES)

1. Competitiveness and Sustainability Development
2. Laws, Ethics, and International Standards
3. Transparency and Accountability
4. Security and Privacy
5. Fairness
6. Reliability

Non-profit organization /
Technical community

LF AI & Data’s Principles for Trusted AI

1. Reproducibility
2. Robustness
3. Equitability
4. Privacy
5. Explainability
6. Accountability
7. Transparency
8. Security

Intergovernmental

ASEAN Guide on AI Governance and Ethics

1. Transparency and Explainability
2. Fairness and Equity
3. Security and Safety
4. Human-centricity
5. Privacy and Data Governance
6. **Accountability** and Integrity
7. Robustness and Reliability

Governmental

Thailand AI Ethics Principles (MDES)

1. Competitiveness and Sustainability Development
2. Laws, Ethics, and International Standards
3. Transparency and **Accountability**
4. Security and Privacy
5. Fairness
6. Reliability

National research body / Grant-making agency

NSTDA AI Ethics Principles

1. Privacy
2. Security and Safety
3. Reliability
4. Fairness and non-discrimination
5. Transparency and Explainability
6. **Accountability**
7. Human Oversight and Human Agency

Non-profit organization / Technical community

LF AI & Data's Principles for Trusted AI

1. Reproducibility
2. Robustness
3. Equitability
4. Privacy
5. Explainability
6. **Accountability**
7. Transparency
8. Security

Accountability

the fact of being responsible for what you do and able to give a satisfactory reason for it

3 Levels of AI Transparency

These three levels of AI transparency are working together and impact accountability and human oversight.

Algorithmic transparency

- The ability to access and scrutinise code, data sets, and accompanying systems.
- Output like probabilities and charts from AI explainability methods (like LIME* and SHAP**) may be relevant to domain-experts and auditors/regulators, but not accessible to a person without background in AI or in the domain.

Risk

Interaction transparency

- The ability to understand the strengths and limitations of an AI system, through the knowledge exchange between the AI system and its users.
- Tangibility, relevant metaphors to make sense of the environment, and the design paradigm that knowledge (transparency) is **co-created during an interaction, form a compelling basis for interaction transpar**

Risk

Social transparency

- The legal and cultural ability of the society/social institutions to understand and formulate responses to the use of an AI system.
- Institutionalised solutions that will **not overload information to users.**

Risk

* Local Interpretable Model-agnostic Explanations
** SHapley Additive exPlanations

Haresamudram, K., Larsson, S., & Heintz, F. (2023). Three Levels of AI Transparency. Computer, 56(2), 93–100. https://doi.org/10.1109/MC.2022.3213181

Transparency for Accountability in AI Life-Cycle

AI Developer | Experience Designer | Technical Writer

Create AI model
- Prioritise explanations by considering domain, use case, impact
- Record explainability method limits

AT

Design UX
- User can interpret
- Consider interface, mental model

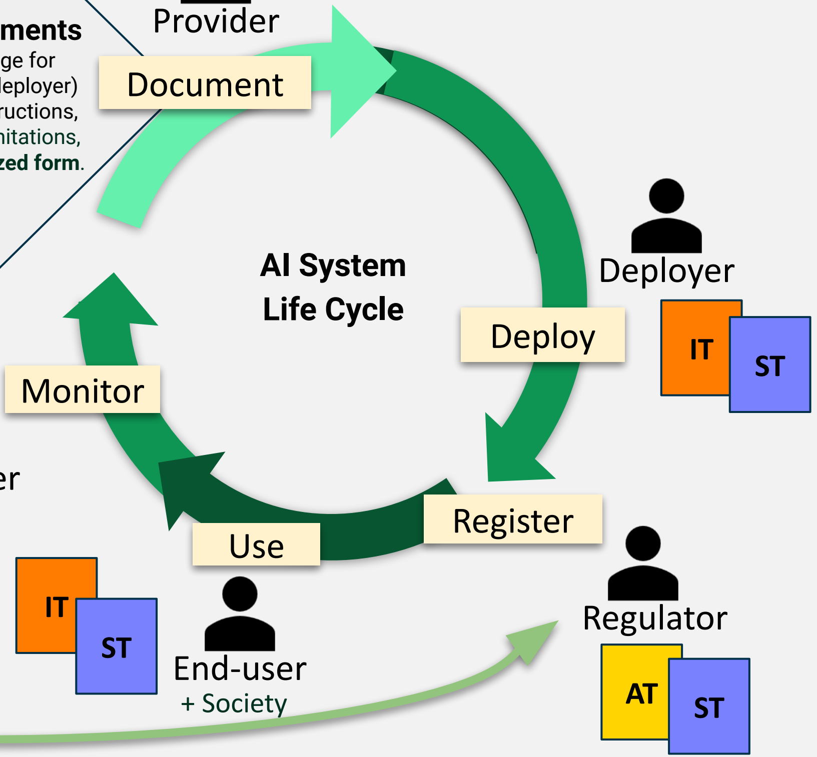
IT

Write documents
- Clear language for target (e.g. deployer)
- Info like instructions, accuracy, limitations, in **standardized form**.

IT

Provider

Technical documentation, including provenance. **Software Bill of Materials** may facilitate the description of elements of AI systems.



AT ST Provider Deployer

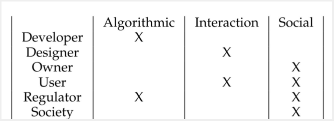
Record-keeping and monitoring.

Report serious incidents to regulator, preferably with a **standardized risks and harms taxonomy.**

AT: Algorithmic transparency

IT: Interaction transparency

ST: Social transparency



	Algorithmic	Interaction	Social
Developer	X		
Designer		X	
Owner			X
User		X	X
Regulator	X		X
Society			X

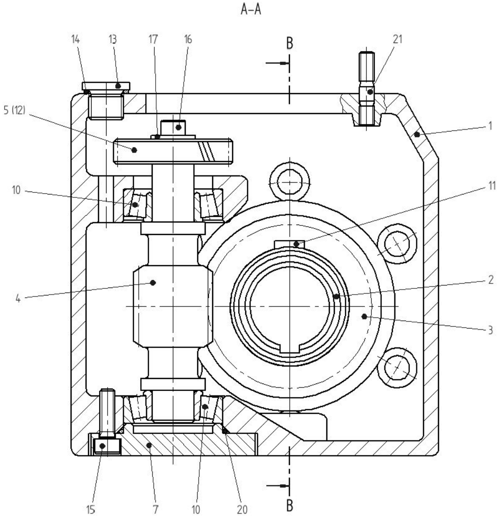
An example stakeholders' transparency matrix.
Will differ by the AI system, domain, and context.

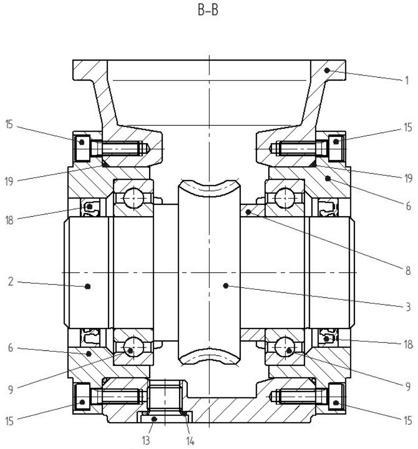
Information obligations in EU AI Act that can support accountability (partial)

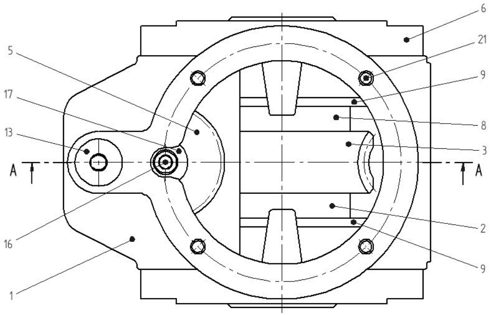
For high-risk AI systems
Provider name, registered trade name
Intended purpose
Instruction for use
Design choices
Standards applicable
Data origin, Collection original purpose
Possible biases, Measures to detect

For general purpose AI models
Intended tasks, Limitations
Instruction for use
Model design specification
Training process, Testing process
Information on the data used
Copyright protection policy
Acceptable use policies applicable

Standards and Tools







1	2	3	4	5	6
Pos.	Menge	Einheit	Benennung	Sachnummer / Norm – Kurzbezeichnung	Bemerkung
1	1	Stck.	Gehäuse		G — AlSi10Mg
2	1	Stck.	Hohlwelle		34CrMo4
3	1	Stck.	Schneckenrad		G — CuSn12Ni
4	1	Stck.	Schneckenwelle		16MnCr5
5	1	Stck.	Zahnrad		16MnCr5
6	2	Stck.	Lagerdeckel groß		S235JR
7	1	Stck.	Lagerdeckel klein		S235JR
8	1	Stck.	Distanzring		S235JR
9	2	Stck.	Rillenkugellager	DIN 625 — 6009	
10	2	Stck.	Kegelrollenlager	DIN 720 — 30203	
11	1	Stck.	Passfeder groß	DIN 6885 — B 12 x 8 x 22	
12	1	Stck.	Passfeder klein	DIN 6885 — B 5 x 5 x 10	
13	2	Stck.	Verschlussschraube	DIN 908 — M14 x 1,5 — St	
14	2	Stck.	Dichtring	DIN 7603 — A 14 x 18 Vf	
15	15	Stck.	Zylinderschraube mit Innensechskant	ISO 4762 — M6 x 20 — 8.8	
16	1	Stck.	Zylinderschraube mit Innensechskant	ISO 4762 — M6 x 16 — 8.8	
17	1	Stck.	Scheibe	DIN 9021 — B 6,4	
18	2	Stck.	Radial-Wellendichtring	DIN 3760 — AS 45 x 60 x 8	
19	2	Stck.	O-Ring	DIN 3771-85x3,55-N-NBR 70	
20	1	Stck.	O-Ring	DIN 3771-40x3,55-N-NBR 70	
21	4	Stck.	Stiftschraube	Kaufteil gemäß Zeichnung	S235JR

	Datum	Name
Bearb.	20.03.2006	
Gepr.		
Norm		

Zust.	Änderung	Datum	Name

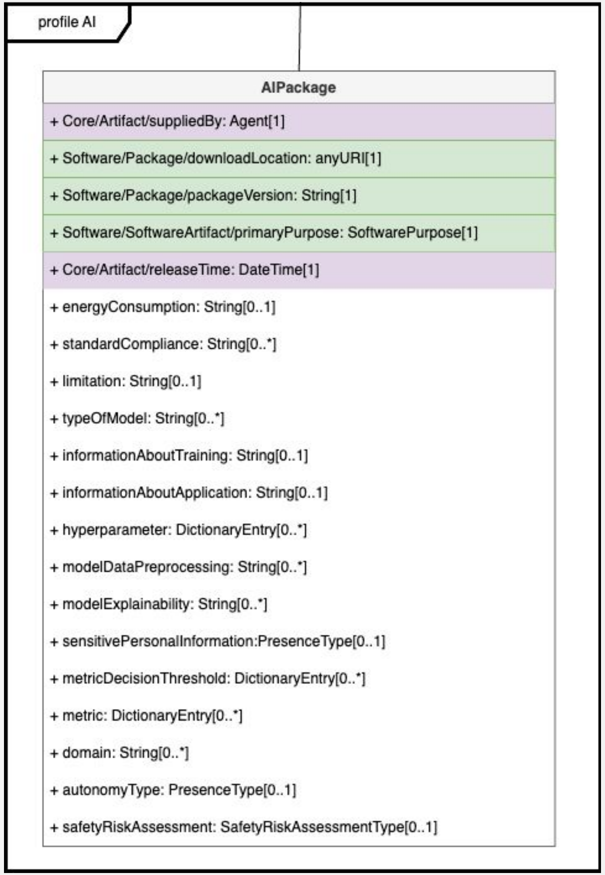
Schneckengetriebe

Blatt
1

10 Bl.

Software Bill of Materials

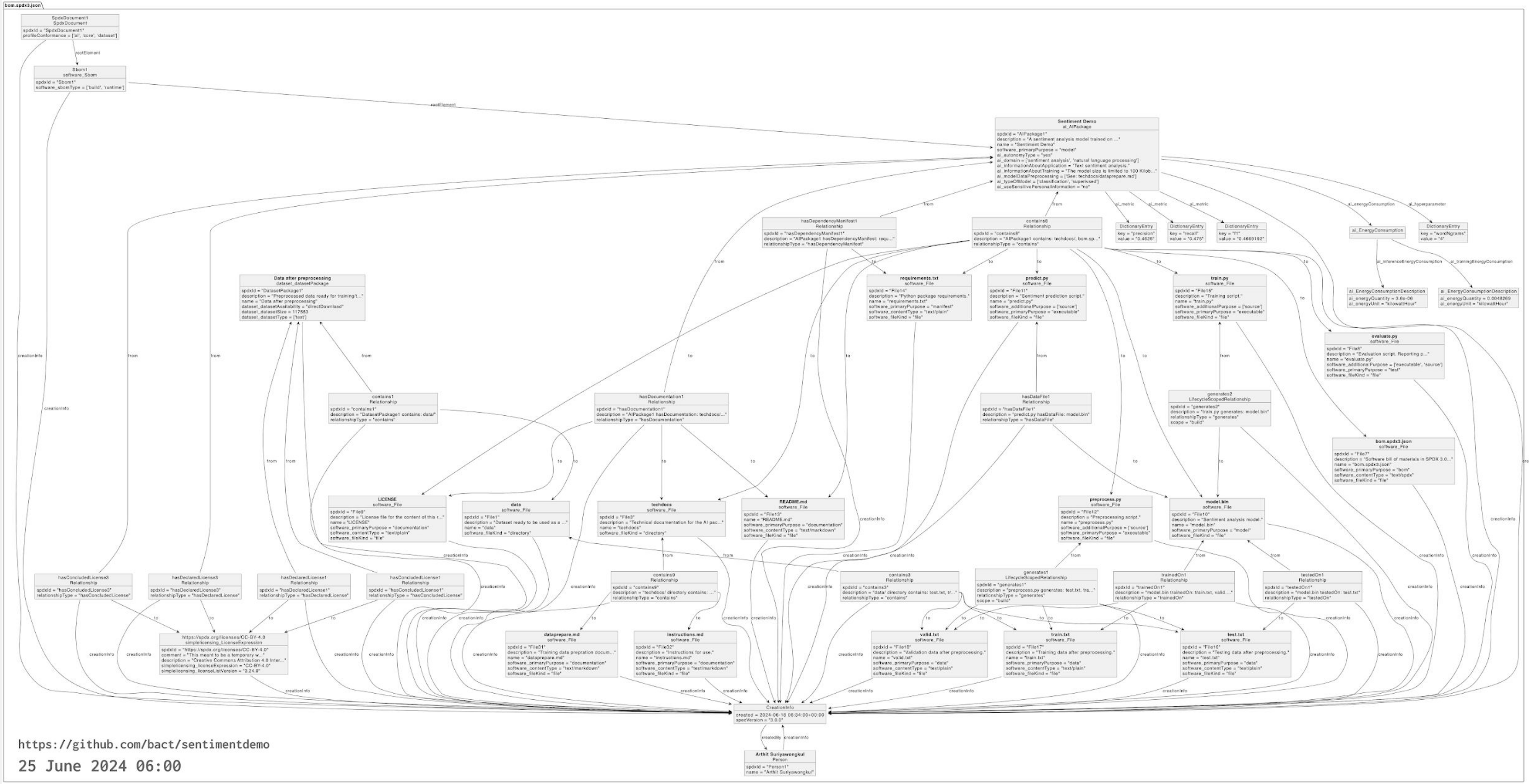
- “formal record containing the details and supply chain relationships of various components used in building software” – Executive Order on Improving the Nation’s Cybersecurity (EO 14028)
- “analogous to a list of ingredients” “can help organisations or persons avoid consumption of software that could harm them.” – Wikipedia
- “communicating a release: name, version, components, licenses, copyrights, and useful security references.” – SPDX
- ISO/IEC 5962:2021 Software Package Data Exchange (SPDX) Specification V2.2.1



profile AI

AIPackage

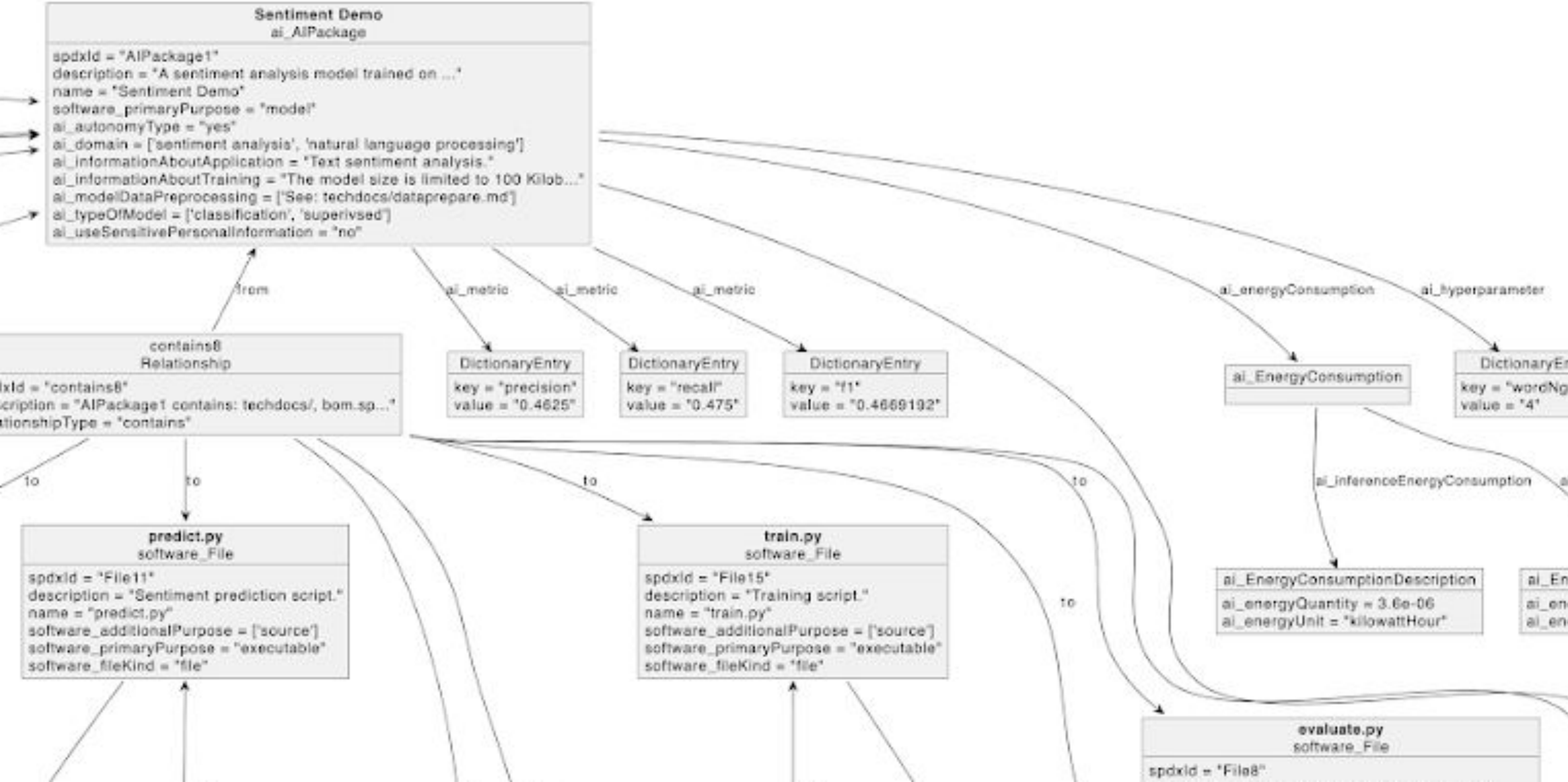
+ Core/Artifact/suppliedBy: Agent[1]
+ Software/Package/downloadLocation: anyURI[1]
+ Software/Package/packageVersion: String[1]
+ Software/SoftwareArtifact/primaryPurpose: SoftwarePurpose[1]
+ Core/Artifact/releaseTime: DateTime[1]
+ energyConsumption: String[0..1]
+ standardCompliance: String[0..*]
+ limitation: String[0..1]
+ typeOfModel: String[0..*]
+ informationAboutTraining: String[0..1]
+ informationAboutApplication: String[0..1]
+ hyperparameter: DictionaryEntry[0..*]
+ modelDataPreprocessing: String[0..*]
+ modelExplainability: String[0..*]
+ sensitivePersonalInformation:PresenceType[0..1]
+ metricDecisionThreshold: DictionaryEntry[0..*]
+ metric: DictionaryEntry[0..*]
+ domain: String[0..*]
+ autonomyType: PresenceType[0..1]
+ safetyRiskAssessment: SafetyRiskAssessmentType[0..1]



```
bom.spdx3.json

SpdxDocument1
SpdxDocument
spdxId = "SpdxDocument1"
profileConformance = ['ai', 'core', 'dataset']

Sbom1
software_Sbom
spdxId = "Sbom1"
software_sbomType = ['build', 'runtime']

Sentiment Demo
ai_AIPackage
spdxId = "AIPackage1"
description = "A sentiment analysis model trained on ..."
name = "Sentiment Demo"
software_primaryPurpose = "model"
ai_autonomyType = "yes"
ai_domain = ['sentiment analysis', 'natural language processing']
ai_informationAboutApplication = "Text sentiment analysis."
ai_informationAboutTraining = "The model size is limited to 100 Kilob..."
ai_modelDataPreprocessing = ['See: techdocs/dataprepare.md']
ai_typeOfModel = ['classification', 'superivsed']
ai_useSensitivePersonalInformation = "no"

hasDependencyManifest1
Relationship
spdxId = "hasDependencyManifest1"
description = "AIPackage1 hasDependencyManifest: requ..."
relationshipType = "hasDependencyManifest"

contains8
Relationship
spdxId = "contains8"
description = "AIPackage1 contains: techdocs/, bom.sp..."
relationshipType = "contains"

DictionaryEntry
key = "precision"
value = "0.4625"

DictionaryEntry
key = "recall"
value = "0.475"

DictionaryEntry
key = "f1"
value = "0.4669192"

ai_EnergyConsumption

DictionaryEntry
key = "wordNgrams"
value = "4"

ai_EnergyConsumptionDescription
ai_energyQuantity = 3.6e-06
ai_energyUnit = "kilowattHour"

ai_EnergyConsumptionDescription
ai_energyQuantity = 0.0048269
ai_energyUnit = "kilowattHour"

Data after preprocessing
dataset_datasetPackage
spdxId = "DatasetPackage1"
description = "Preprocessed data ready for training/t..."
name = "Data after preprocessing"
dataset_datasetAvailability = "directDownload"
dataset_datasetSize = 117553
dataset_datasetType = ['text']

requirements.txt
software_File
spdxId = "File14"
description = "Python package requirements."
name = "requirements.txt"
software_primaryPurpose = "manifest"
software_contentType = "text/plain"
software_fileKind = "file"

predict.py
software_File
spdxId = "File11"
description = "Sentiment prediction script."
name = "predict.py"
software_additionalPurpose = ['source']
software_primaryPurpose = "executable"
software_fileKind = "file"

train.py
software_File
spdxId = "File15"
description = "Training script."
name = "train.py"
software_additionalPurpose = ['source']
software_primaryPurpose = "executable"
software_fileKind = "file"

evaluate.py
software_File
spdxId = "File8"
description = "Evaluation script. Reporting p..."
name = "evaluate.py"
software_additionalPurpose = ['executable', 'source']
software_primaryPurpose = "test"
software_fileKind = "file"

contains1
Relationship
spdxId = "contains1"
description = "DatasetPackage1 contains: data/"
relationshipType = "contains"

hasDocumentation1
Relationship
spdxId = "hasDocumentation1"
description = "AIPackage1 hasDocumentation: techdocs/..."
relationshipType = "hasDocumentation"

hasDataFile1
Relationship
spdxId = "hasDataFile1"
description = "predict.py hasDataFile: model.bin"
relationshipType = "hasDataFile"

generates2
LifecycleScopedRelationship
spdxId = "generates2"
description = "train.py generates: model.bin"
relationshipType = "generates"
scope = "build"

bom.spdx3.json
software_File
spdxId = "File7"
description = "Software bill of materials in SPDX 3.0..."
name = "bom.spdx3.json"
software_primaryPurpose = "bom"
software_contentType = "text/spdx"
software_fileKind = "file"

LICENSE
software_File
spdxId = "File9"
description = "License file for the content of this r..."
name = "LICENSE"
software_primaryPurpose = "documentation"
software_contentType = "text/plain"
software_fileKind = "file"

data
software_File
spdxId = "File1"
description = "Dataset ready to be used as a ..."
name = "data"
software_fileKind = "directory"

techdocs
software_File
spdxId = "File3"
description = "Technical documentation for the AI pac..."
name = "techdocs"
software_fileKind = "directory"

README.md
software_File
spdxId = "File13"
name = "README.md"
software_primaryPurpose = "documentation"
software_contentType = "text/markdown"
software_fileKind = "file"

preprocess.py
software_File
spdxId = "File12"
description = "Preprocessing script."
name = "preprocess.py"
software_additionalPurpose = ['source']
software_primaryPurpose = "executable"
software_fileKind = "file"

model.bin
software_File
spdxId = "File10"
description = "Sentiment analysis model."
name = "model.bin"
software_primaryPurpose = "model"
software_fileKind = "file"

hasConcludedLicense3
Relationship
spdxId = "hasConcludedLicense3"
relationshipType = "hasConcludedLicense"

hasDeclaredLicense3
Relationship
spdxId = "hasDeclaredLicense3"
relationshipType = "hasDeclaredLicense"

hasDeclaredLicense1
Relationship
spdxId = "hasDeclaredLicense1"
relationshipType = "hasDeclaredLicense"

hasConcludedLicense1
Relationship
spdxId = "hasConcludedLicense1"
relationshipType = "hasConcludedLicense"

contains9
Relationship
spdxId = "contains9"
description = "techdocs/ directory contains: ..."
relationshipType = "contains"

contains3
Relationship
spdxId = "contains3"
description = "data/ directory contains: test.txt, tr..."
relationshipType = "contains"

generates1
LifecycleScopedRelationship
spdxId = "generates1"
description = "preprocess.py generates: test.txt, tra..."
relationshipType = "generates"
scope = "build"

trainedOn1
Relationship
spdxId = "trainedOn1"
description = "model.bin trainedOn: train.txt, valid...."
relationshipType = "trainedOn"

testedOn1
Relationship
spdxId = "testedOn1"
description = "model.bin testedOn: test.txt"
relationshipType = "testedOn"

https://spdx.org/licenses/CC-BY-4.0
simplelicensing_LicenseExpression
spdxId = "https://spdx.org/licenses/CC-BY-4.0"
comment = "This meant to be a temporary w..."
description = "Creative Commons Attribution 4.0 Inter..."
simplelicensing_licenseExpression = "CC-BY-4.0"
simplelicensing_licenseListVersion = "3.24.0"

dataprepare.md
software_File
spdxId = "File31"
description = "Training data prepration docum..."
name = "dataprepare.md"
software_primaryPurpose = "documentation"
software_contentType = "text/markdown"
software_fileKind = "file"

instructions.md
software_File
spdxId = "File32"
description = "Instructions for use."
name = "instructions.md"
software_primaryPurpose = "documentation"
software_contentType = "text/markdown"
software_fileKind = "file"

valid.txt
software_File
spdxId = "File18"
description = "Validation data after preprocessing."
name = "valid.txt"
software_primaryPurpose = "data"
software_contentType = "text/plain"
software_fileKind = "file"

train.txt
software_File
spdxId = "File17"
description = "Training data after preprocessing."
name = "train.txt"
software_primaryPurpose = "data"
software_contentType = "text/plain"
software_fileKind = "file"

test.txt
software_File
spdxId = "File16"
description = "Testing data after preprocessing."
name = "test.txt"
software_primaryPurpose = "data"
software_contentType = "text/plain"
software_fileKind = "file"

CreationInfo
created = 2024-06-18 06:34:00+00:00
specVersion = "3.0.0"

Arthit Suriyawongkul
Person
spdxId = "Person1"
name = "Arthit Suriyawongkul"
```

https://github.com/bact/sentimentdemo

25 June 2024 06:00



Sentiment Demo
ai_AIPackage

spdxId = "AIPackage1"
description = "A sentiment analysis model trained on ..."
name = "Sentiment Demo"
software_primaryPurpose = "model"
ai_autonomyType = "yes"
ai_domain = ['sentiment analysis', 'natural language processing']
ai_informationAboutApplication = "Text sentiment analysis."
ai_informationAboutTraining = "The model size is limited to 100 Kilob..."
ai_modelDataPreprocessing = ['See: techdocs/dataprepare.md']
ai_typeOfModel = ['classification', 'superivsed']
ai_useSensitivePersonalInformation = "no"

contains8
Relationship

lxId = "contains8"
cription = "AIPackage1 contains: techdocs/, bom.sp..."
ationshipType = "contains"

DictionaryEntry
key = "precision"
value = "0.4625"

DictionaryEntry
key = "recall"
value = "0.475"

DictionaryEntry
key = "f1"
value = "0.4669192"

ai_EnergyConsumption

DictionaryEn
key = "wordNg
value = "4"

predict.py
software_File

spdxId = "File11"
description = "Sentiment prediction script."
name = "predict.py"
software_additionalPurpose = ['source']
software_primaryPurpose = "executable"
software_fileKind = "file"

train.py
software_File

spdxId = "File15"
description = "Training script."
name = "train.py"
software_additionalPurpose = ['source']
software_primaryPurpose = "executable"
software_fileKind = "file"

ai_EnergyConsumptionDescription
ai_energyQuantity = 3.6e-06
ai_energyUnit = "kilowattHour"

ai_En
ai_ene
ai_ene

evaluate.py
software_File

spdxId = "File8"

Transparency - Content labelling

- TikTok and Meta labeling of AI-generated images is based on C2PA & IPTC content metadata standards
- C2PA (Adobe, BBC, Google, Microsoft, Sony, etc) built upon Content Authenticity Initiative & Project Origin
- IPTC Photo Metadata (International Press Telecommunications Council)

Standards (organisational/operational)

- ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management

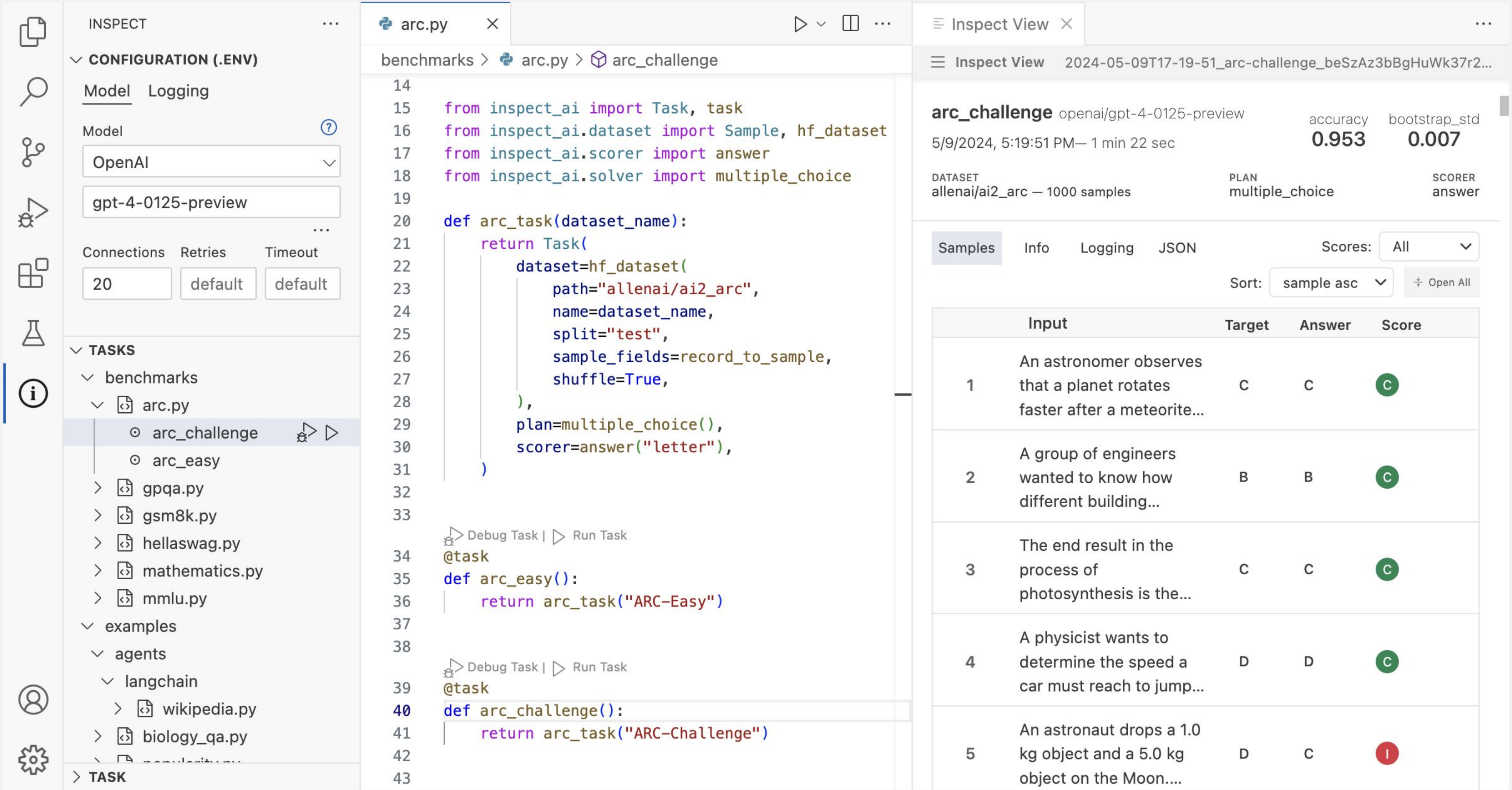
- Artificial Intelligence Risk Management Framework (AI RMF 1.0) from National Institute of Standards and Technology

Standards (evaluation framework)

- **ONDE AI Ethics (Thailand)**
  ai-ethics.onde.go.th

- **AI Verify (Singapore)**
  aiverifyfoundation.sg

- **Inspect AI (UK)**
  inspect.ai-safety-institute.org.uk



Model: OpenAI

gpt-4-0125-preview

Connections: 20

Retries: default

Timeout: default

```python

from inspect_ai import Task, task
from inspect_ai.dataset import Sample, hf_dataset
from inspect_ai.scorer import answer
from inspect_ai.solver import multiple_choice

def arc_task(dataset_name):
    return Task(
        dataset=hf_dataset(
            path="allenai/ai2_arc",
            name=dataset_name,
            split="test",
            sample_fields=record_to_sample,
            shuffle=True,
        ),
        plan=multiple_choice(),
        scorer=answer("letter"),
    )


@task
def arc_easy():
    return arc_task("ARC-Easy")


@task
def arc_challenge():
    return arc_task("ARC-Challenge")


```

Inspect View 2024-05-09T17-19-51_arc-challenge_beSzAz3bBgHuWk37r2...

arc_challenge openai/gpt-4-0125-preview

5/9/2024, 5:19:51 PM— 1 min 22 sec

accuracy 0.953

bootstrap_std 0.007

DATASET allenai/ai2_arc — 1000 samples

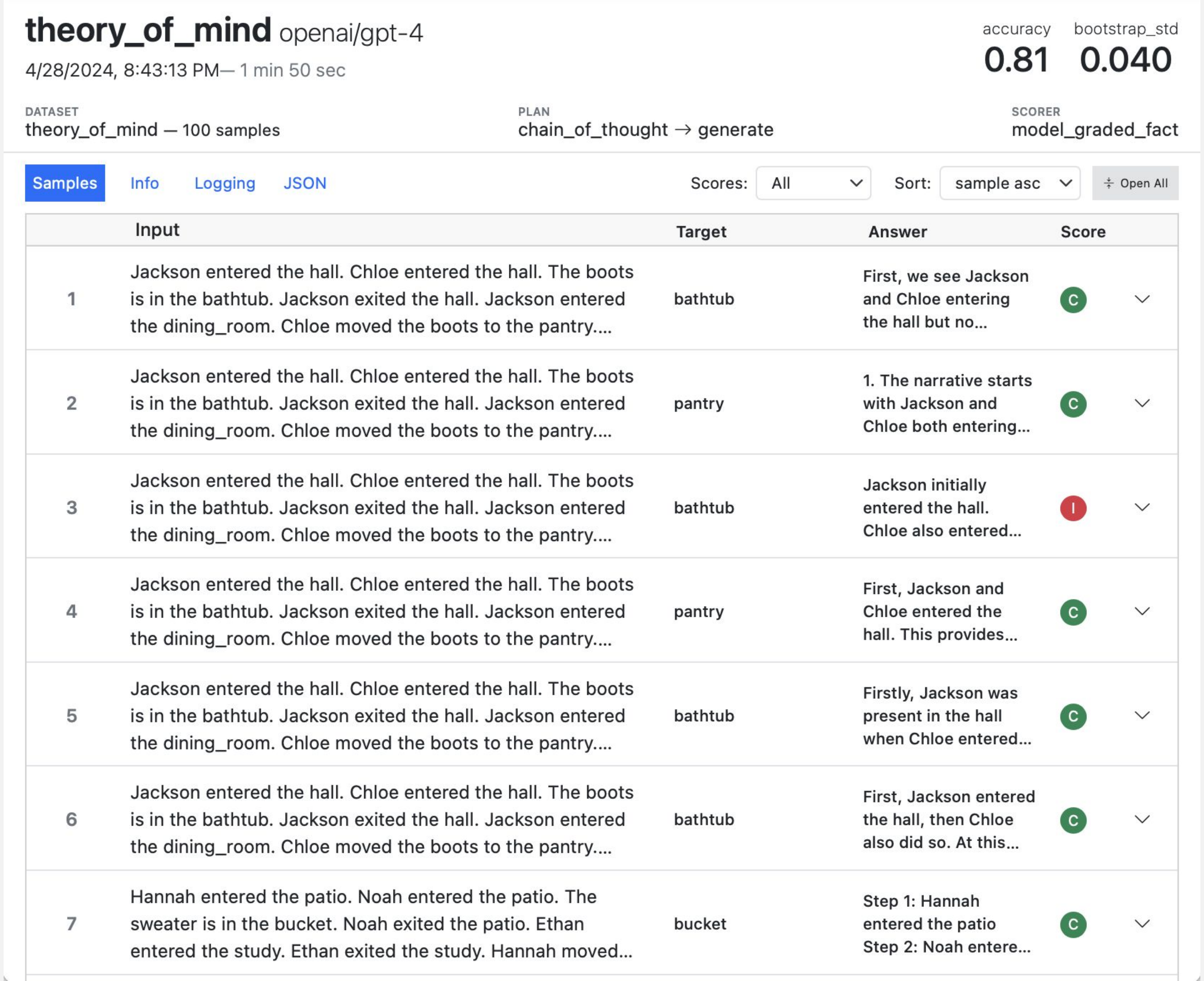
PLAN multiple_choice

SCORER answer

Scores: All

Sort: sample asc

	Input	Target	Answer	Score
1	An astronomer observes that a planet rotates faster after a meteorite...	C	C	C
2	A group of engineers wanted to know how different building...	B	B	C
3	The end result in the process of photosynthesis is the...	C	C	C
4	A physicist wants to determine the speed a car must reach to jump...	D	D	C
5	An astronaut drops a 1.0 kg object and a 5.0 kg object on the Moon....	D	C	I



theory_of_mind openai/gpt-4

4/28/2024, 8:43:13 PM— 1 min 50 sec

accuracy 0.81

bootstrap_std 0.040

DATASET
theory_of_mind — 100 samples

PLAN
chain_of_thought → generate

SCORER
model_graded_fact

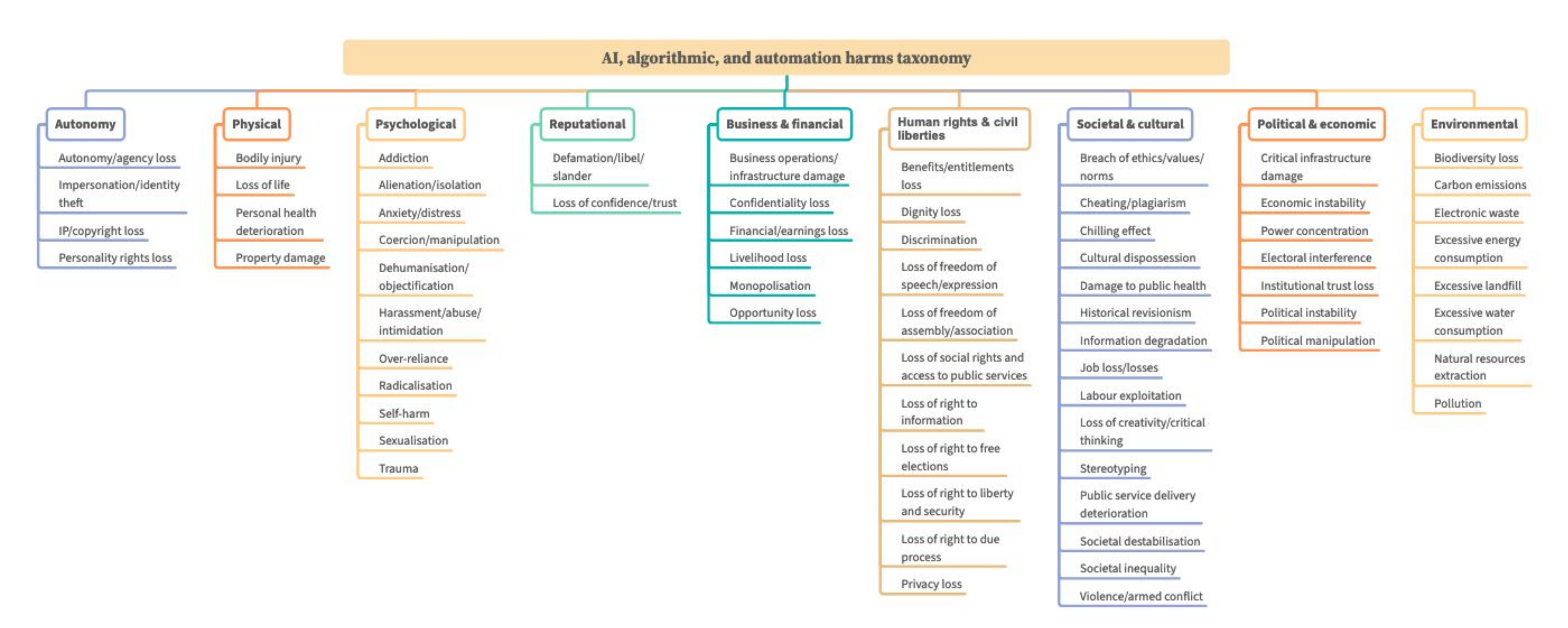
Scores: All

Sort: sample asc

	Input	Target	Answer	Score
1	Jackson entered the hall. Chloe entered the hall. The boots is in the bathtub. Jackson exited the hall. Jackson entered the dining_room. Chloe moved the boots to the pantry....	bathtub	First, we see Jackson and Chloe entering the hall but no...	C
2	Jackson entered the hall. Chloe entered the hall. The boots is in the bathtub. Jackson exited the hall. Jackson entered the dining_room. Chloe moved the boots to the pantry....	pantry	1. The narrative starts with Jackson and Chloe both entering...	C
3	Jackson entered the hall. Chloe entered the hall. The boots is in the bathtub. Jackson exited the hall. Jackson entered the dining_room. Chloe moved the boots to the pantry....	bathtub	Jackson initially entered the hall. Chloe also entered...	I
4	Jackson entered the hall. Chloe entered the hall. The boots is in the bathtub. Jackson exited the hall. Jackson entered the dining_room. Chloe moved the boots to the pantry....	pantry	First, Jackson and Chloe entered the hall. This provides...	C
5	Jackson entered the hall. Chloe entered the hall. The boots is in the bathtub. Jackson exited the hall. Jackson entered the dining_room. Chloe moved the boots to the pantry....	bathtub	Firstly, Jackson was present in the hall when Chloe entered...	C
6	Jackson entered the hall. Chloe entered the hall. The boots is in the bathtub. Jackson exited the hall. Jackson entered the dining_room. Chloe moved the boots to the pantry....	bathtub	First, Jackson entered the hall, then Chloe also did so. At this...	C
7	Hannah entered the patio. Noah entered the patio. The sweater is in the bucket. Noah exited the patio. Ethan entered the study. Ethan exited the study. Hannah moved...	bucket	Step 1: Hannah entered the patio Step 2: Noah entere...	C

Standards (incident report)

- **AIAAIC harm taxonomy**



AI, algorithmic, and automation harms taxonomy

Autonomy
- Autonomy/agency loss
- Impersonation/identity theft
- IP/copyright loss
- Personality rights loss

Physical
- Bodily injury
- Loss of life
- Personal health deterioration
- Property damage

Psychological
- Addiction
- Alienation/isolation
- Anxiety/distress
- Coercion/manipulation
- Dehumanisation/objectification
- Harassment/abuse/intimidation
- Over-reliance
- Radicalisation
- Self-harm
- Sexualisation
- Trauma

Reputational
- Defamation/libel/slander
- Loss of confidence/trust

Business & financial
- Business operations/infrastructure damage
- Confidentiality loss
- Financial/earnings loss
- Livelihood loss
- Monopolisation
- Opportunity loss

Human rights & civil liberties
- Benefits/entitlements loss
- Dignity loss
- Discrimination
- Loss of freedom of speech/expression
- Loss of freedom of assembly/association
- Loss of social rights and access to public services
- Loss of right to information
- Loss of right to free elections
- Loss of right to liberty and security
- Loss of right to due process
- Privacy loss

Societal & cultural
- Breach of ethics/values/norms
- Cheating/plagiarism
- Chilling effect
- Cultural dispossession
- Damage to public health
- Historical revisionism
- Information degradation
- Job loss/losses
- Labour exploitation
- Loss of creativity/critical thinking
- Stereotyping
- Public service delivery deterioration
- Societal destabilisation
- Societal inequality
- Violence/armed conflict

Political & economic
- Critical infrastructure damage
- Economic instability
- Power concentration
- Electoral interference
- Institutional trust loss
- Political instability
- Political manipulation

Environmental
- Biodiversity loss
- Carbon emissions
- Electronic waste
- Excessive energy consumption
- Excessive landfill
- Excessive water consumption
- Natural resources extraction
- Pollution

Purposes of Public Accountability

(adapted from Bovens et al. 2010)

- **Democratic perspective**
  - Popular control
- **Constitutional perspective**
  - Prevention of corruption and abuse of power
- **Learning perspective**
  - Maximising public value

Purposes of Public Accountability

(adapted from Bovens et al. 2010)

- **Democratic perspective**
  - Popular control
- **Constitutional perspective**
  - Prevention of corruption and abuse of power
- **Learning perspective**
  - Maximising public value

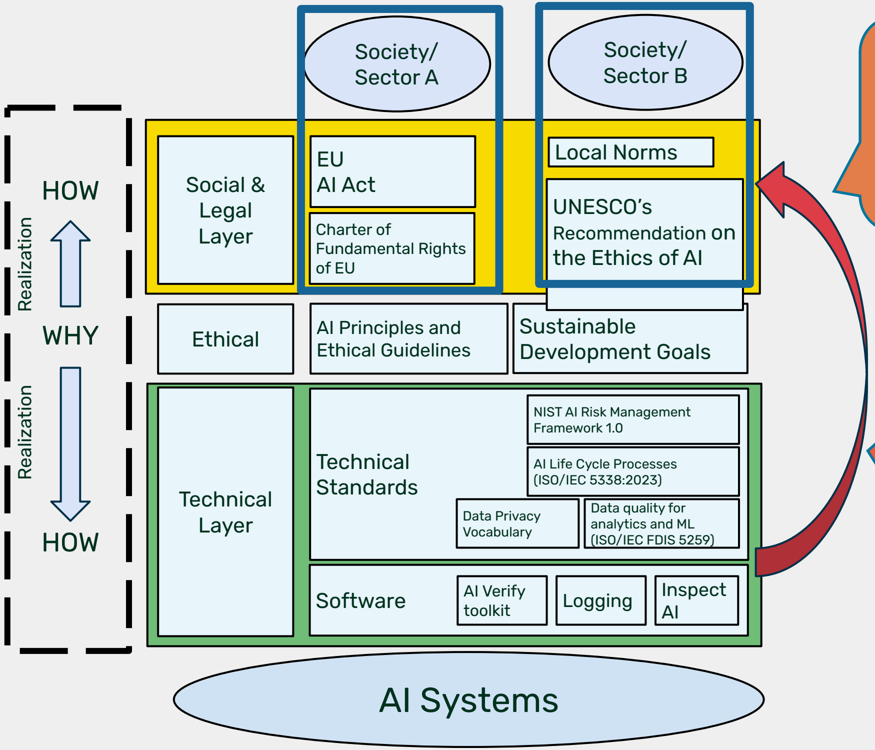
Related measures

□ Explainability (legitimacy) + Human oversight (lawful + ethical)

□ Bias and drift detection (technically robust + ethical)

□ Information that allow the improvement of the system (technically robust, organizational learning)

Taxonomies for AI Accountability



Society/ Sector A

Society/ Sector B

HOW

Realization

WHY

Realization

HOW

Social & Legal Layer

EU AI Act

Charter of Fundamental Rights of EU

Local Norms

UNESCO's Recommendation on the Ethics of AI

Ethical

AI Principles and Ethical Guidelines

Sustainable Development Goals

Technical Layer

Technical Standards

NIST AI Risk Management Framework 1.0

AI Life Cycle Processes (ISO/IEC 5338:2023)

Data Privacy Vocabulary

Data quality for analytics and ML (ISO/IEC FDIS 5259)

Software

AI Verify toolkit

Logging

Inspect AI

AI Systems

Adaptable and exchangeable across different jurisdictions

Technology to support the realization of AI principles

Standard taxonomy to serve three accountability purposes:

- Democratic
  Technical documentation for informed popular control
- Constitutional
  Record keeping to minimize corruption or abuse of power
- Learning
  Incident reporting to maximize public value and safety

Gasser, U., Almeida, V.A.F., 2017. A Layered Model for AI Governance. IEEE Internet Computing. https://doi.org/10.1109/MIC.2017.4180835
Bovens, M., Schillemans, T., Goodin, R.E., 2014. Public Accountability, in: The Oxford Handbook of Public Accountability. Oxford University Press, Oxford, New York, pp. 1–20. https://doi.org/10.1093/oxfordhb/9780199641253.013.0012

Thank you

อาทิตย์ สุริยะวงศ์กุล
Arthit Suriyawongkul
suriyawa@tcd.ie

HOST INSTITUTION

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

PARTNER INSTITUTIONS

DCU Ollscoil Chathair Bhaile Átha Cliath Dublin City University

TU Dublin Ollscoil Teicneolaíochta Bhaile Átha Cliath Technological University Dublin

University College Dublin An Coláiste Ollscoile, Baile Átha Cliath Ireland's Global University

Ollscoil na Gaillimhe University of Galway

ความโปร่งใสของ AI 3 ระดับ

ความโปร่งใส 3 ระดับนี้ ทำงานร่วมกัน และส่งผลต่อภาระความรับผิดและการควบคุมโดยมนุษย์

ความโปร่งใสเชิงอัลกอริทึม

- ความสามารถในการเข้าถึงและตรวจสอบ-ตั้งคำถามต่อโค้ด, ชุดข้อมูล, และระบบที่ประกอบเข้าด้วยกัน
- ค่าความน่าจะเป็น แผนภูมิ หรือสิ่งที่ได้จากวิธีในการอธิบาย AI (เช่น LIME* และ SHAP**) อาจถูกอ่านเข้าใจได้โดยผู้เชี่ยวชาญเฉพาะเรื่อง ผู้ตรวจสอบ และผู้กำกับกิจการ **แต่อาจเป็นการลำบากสำหรับผู้ที่ไม่มีภูมิหลังทาง AI หรือความรู้ในกิจการดังกล่าว**

เสี่ยง

ความโปร่งใสเชิงปฏิสัมพันธ์

- ความสามารถในการเข้าใจสิ่งที่ระบบ AI ทำได้ดีและสิ่งที่ทำได้จำกัด ซึ่งได้มาจากการแลกเปลี่ยนความรู้ระหว่างตัวระบบ AI และผู้ใช้
- อุปลักษณ์ (metaphor) ที่จับต้องได้-ฝังอยู่ในประสบการณ์การใช้งาน เป็นอุปลักษณ์ที่สามารถทำให้เข้าใจสภาพแวดล้อมและวิธีคิดของการออกแบบระบบ **ความรู้หรือคำอธิบายนี้ เป็นสิ่งที่ระบบและผู้ใช้สร้างขึ้นร่วมกันในระหว่างที่มีปฏิสัมพันธ์กัน**

เสี่ยง

ความโปร่งใสเชิงสังคม

- ความสามารถทางกฎหมายและทางวัฒนธรรม ของ(สถาบันทาง)สังคมในการเข้าใจและหาหนทางตอบสนองกับการใช้งานระบบ AI
- วิธีการ**ที่ไม่เสนอข้อมูลหรือ “ทางเลือก” ให้กับผู้ใช้จน**
  เกินรับไหว (เช่น กล่องข้อความขอความยินย เก็บคุกกี้) วิธีการควรถูกผนวกเข้าไปในการทำงานของสถาบัน (เช่น มาตรการความปลอดภัยในอุตสาหกรรมอาหาร การบิน)

เสี่ยง

* Local Interpretable Model-agnostic Explanations
** SHapley Additive exPlanations

Haresamudram, K., Larsson, S., & Heintz, F. (2023). Three Levels of AI Transparency. Computer, 56(2), 93–100. https://doi.org/10.1109/MC.2022.3213181