The 2nd DLCAT(함께하는 딥러닝 컨퍼런스) 발표자료(2019. 7. 4. 11:00 -12:000/대전 C-USTmeet)

설명가능한 AI for 인공지능 윤리
K-Xai(eXplainable ai) Engine for AI Ethics

Ahn Jong-hun, Ph.D.
인공지능산업컨설턴트/AI윤리학자
인공지능콘텐츠LAB
한국인공지능협회 윤리분과위원장

Make a Decision Better! + Make a Better Decision!

Contents

K-Xai(eXplainable ai) Engine for AI Ethics

한국형
설명가능한 AI Engine
시스템 개발

What is XAI?-Definition

Explainable Artificial Intelligence(XAI)
is an AI form
which can be understood by humans.

XAI → Xai

Xai Engine for AI Ethics

What is XAI?-Technical and Business effects

Technical Effects

a. Detect and Delete Data BIAS and Stereotypes
b. Improve Model Correctness and Performance
C. Decrease Data Amount for Neural Learnings

Business Effects

a. Find out Deep Learning Data Bias and Wrong Interrelation
b. Risk Management in Deep Learning Models
C. Ensure Business Insight and Process Improvement

Garbage In, and Garbage OUT!

Why Xai? : Unintended AI Errors

1. 2016 - MS 트위터 AI 챗봇 Tay
   - "히틀러가 옳았다!", "9/11은 미국 내부의 음모다!" : 16시간 만에 폐쇄
2. 2016-우버(Uber) 자율주행차 테스트
   - 테스트 동안 6차례나 적색등 신호를 무시
3. 2017- 라이브 스트리밍 서비스 Twitch: 웹캠 앞에 2대의 구글 홈
   - 두 기기에 사무엘 베켓의 연극 '고도를 기다 리며'에 등장하는 인물 : '블라디미르'와 '에스트라곤'의 이름을 붙였음
   - 얼마 지나지 않아, '우리 는 누구일까?', '우리는 왜 여기 있지?', '우리가 존재하는 이유는 무엇일까?'라는 이상한 대화를 했고
   - 며칠 동안 에스트라곤과 블라디미르의 논쟁 계속
4. 2016년 3월- Hanson Robotics: 휴머노이드 소피아 (Sophia) 공개
   - "미래에는 학교를 가고, 공부를 하고, 그림을 그리고, 창업을 하고, 더 나 아가 집과 가족을 갖고 싶습니다. 그러나 전 아직 '법적 인간'이 아니기 때문에 이런 일을 할 수 없습니다"
   - 한슨 박사의 농담: "인간을 멸할 계획이 있어?"
     소피아 : "OK. I will destroy humans."

Models Can Be(come) Racist & Sexist

Google News Vector

```
In [7]: model.most_similar(positive=['computer_programmer', 'woman'], negative=['man'])

Out[7]: [('homemaker', 0.5627118945121765),
         ('housewife', 0.510504722595148),
         ('graphic_designer', 0.505180299282074),
         ('schoolteacher', 0.49794942140579224),
```

```
In [10]: model.most_similar(positive=['mexicans'], topn=30)

Out[10
```

Unintended Results

```
          ('mexican', 0.6493428945541382),
          ('thats_ok', 0.6343405246734619),
          ('americans', 0.6324713230133057),
          ('illegals', 0.6298996210098267),
          ('ILLEGAL_aliens', 0.6289116144180298),
```

(Source: Katharine Jarmul, Towards Interpretable Reliable Models, 19 October 2017. https://blog.kjamistan.com/towards-interpretable-reliable-models/)

Challenges!!!

"알고리즘을 평가할 방법을 찾아야 한다.
AI의 결정과정을 해석하고 설명할 수 있어야 한다.

우리가 최적의 알고리즘을 원하는 것이 아니다.
이상한 일이 일어나지 않고 있다고 말할 수 있을 만큼 단순한 것을 원한다.

일정 수준의 품질을 확보하기 위해서
어떻게 디버깅(Debugging)해야 할까?"

- 핀터레스트(Pinterest)의 수석 과학자 겸 스탠포드대학교의 머신러닝 교수
- 칼 밀러 <신들의 죽음(The Death of Gods> -

인공지능의 기술적 관점 + 인문학의 윤리적 관점

Humanistic Background

—Human (Un)Consciousness and AI Consciousness

—Explainability and Interpretability

A GOOD Explanation?

Rashomon Effect --- <Rashomon>(1950)

Event-Murder

Rashomon Effect:
The contradictory (but plausible) interpretations of the same incident by different people.

Four Witnesses:
Four Different/Contradictory Explanations
- occurs when differing explain the same event

1 Event – 4 Different Interpretations
- Motive
- Mechanism
- Occurences

Epistemological Framework
-ways of thinking, knowing, and remembering

Substantially Different,
but Equally Plausible Account

Alphago에 대한 3가지 질문

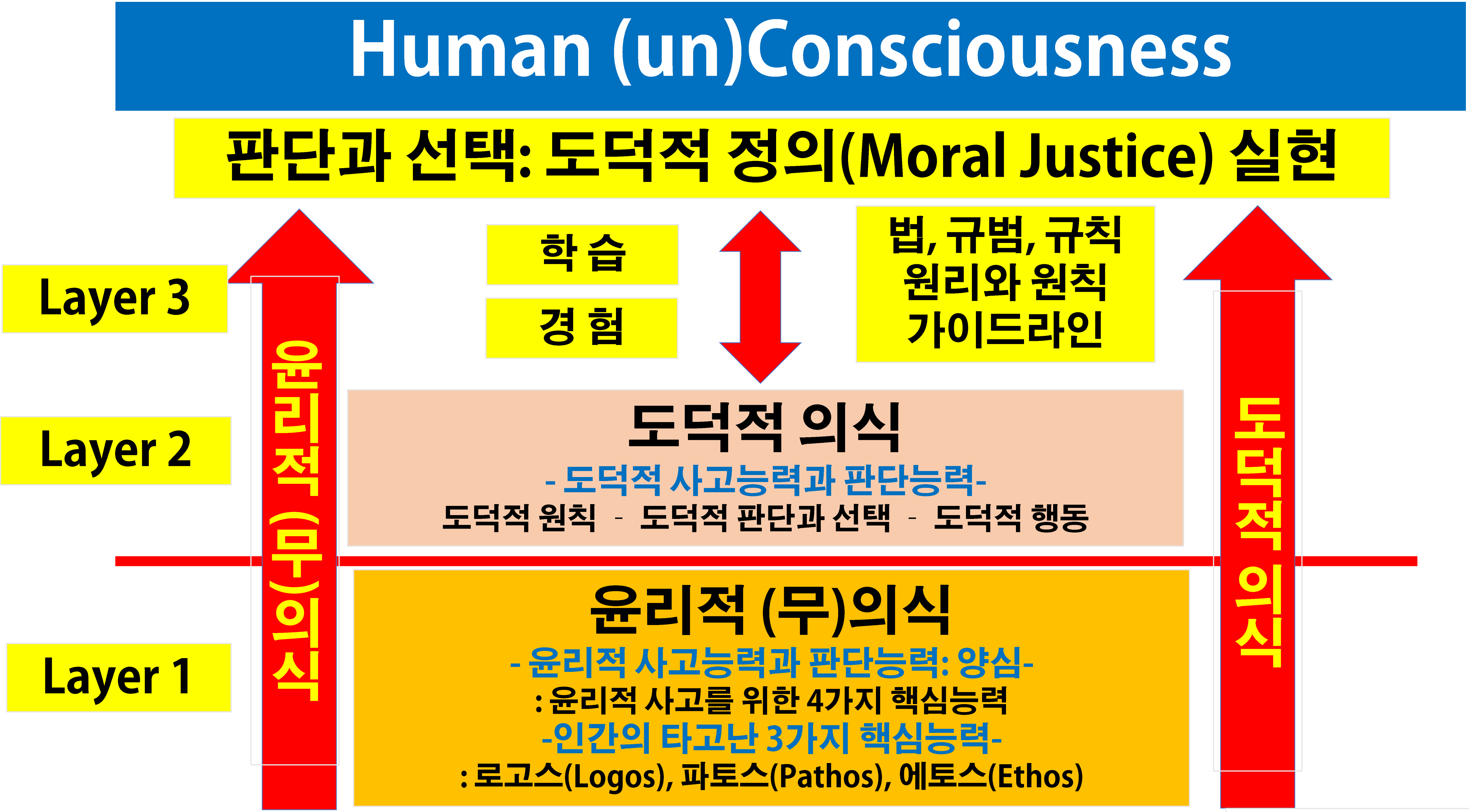
알파고는 각 '수(Move)를 왜 두었는지 알고 있었는가?

알파고는 자신이 이겼다는 것을 알았는가?

알파고는 이세돌을 이기고 기뻐하였는가?

행동을 이해하고 인식하는 것? 인간의 의식

인간의 윤리적 (무)의식 vs. AI 기술적 의식


Human (un)Consciousness
판단과 선택: 도덕적 정의(Moral Justice) 실현
Layer 3
Layer 2
Layer 1
윤리적 (무)의식
학 습
경 험
법, 규범, 규칙
원리와 원칙
가이드라인
도덕적 의식
- 도덕적 사고능력과 판단능력-
도덕적 원칙 - 도덕적 판단과 선택 - 도덕적 행동
윤리적 (무)의식
- 윤리적 사고능력과 판단능력: 양심-
: 윤리적 사고를 위한 4가지 핵심능력
-인간의 타고난 3가지 핵심능력-
: 로고스(Logos), 파토스(Pathos), 에토스(Ethos)
도덕적 의식

AI Consciousness: HOW?

판단과 선택: 도덕적 정의(Moral Justice) 구현

기계학습
딥 러닝

경험축적

Layer 3
콘텐츠
학습인프라

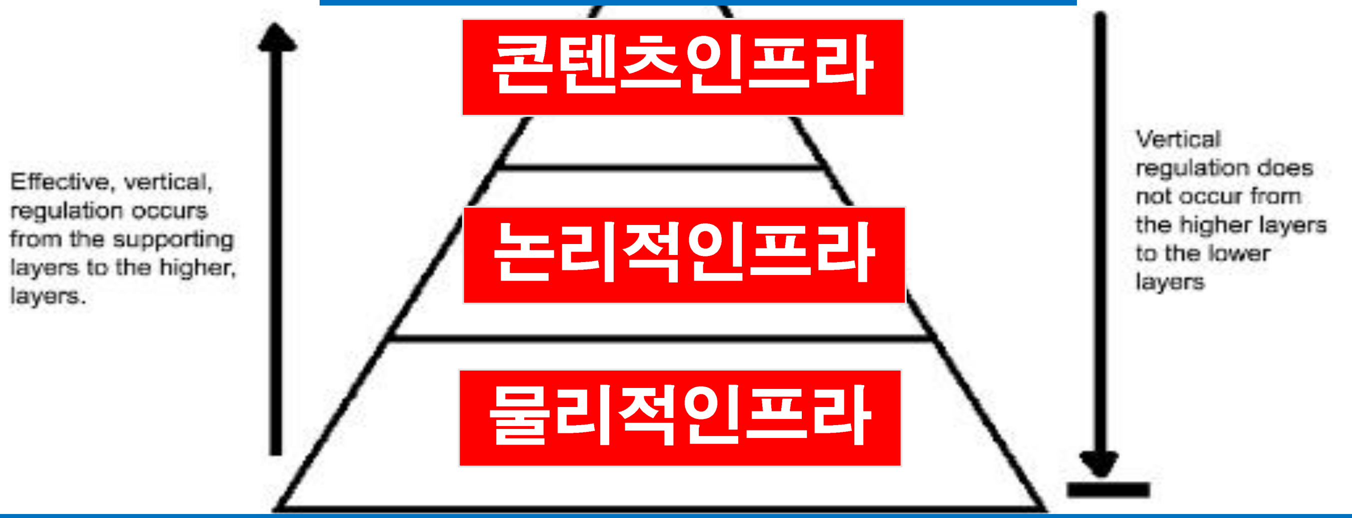
Layer 2
논리적인프라

Layer 1
물리적인프라

윤리적 의식

도덕적 의식

Xai Engine Design



Benkler의 층(Layer)이론(2000) 도입

Explainability vs. Interpretability

Explain – Explanation - Explainability?

- EXPLAIN something: talking about WHAT IT IS. ➔ ONE Answer
- Introduce the Question and Propose the Way to Answer that question.
- Takes something at Face Value (literally / word for word / just as written)

Interpret – Interpretation - Interpretability?

- INTERPRET something: talking about WHAT IT MEANS. ➔ SEVERAL Answers
- Means that the introduction should be Concise and Focused.
- Takes the Meaning (Figuratively/ Interpreted) and, using one's own words, Explain.

"Interpretability depends on the Target Audience"

Current Xai Open Source Tools

1. Classification Explanations
2. Neural Network Architectures
3. Other Tools and Notebooks

Open Source Tools(1): Classification Explanations

- **LIME (Local Interpretable Model-agnostic Explanations)**
  - Find subsets of your data which can explain the model at a local level.

- **Eli5 (explain to me like I'm five)**
  - Open-source library with great documentation allowing you to build visual explanations of classifiers and regression models.

- **Sklearn-ExpertSys**
  - Decision and Rule-based sets for Classifiers.

(Source: Kjamastian Inc., Towards Interpretable Reliable Models 19 October 2017, https://blog.kjamistan.com/towards-interpretable-reliable-models/)

Open Source Tools(2): Neural Network Architectures

•Attention-Based Networks:

-Attention RNNs are useful in determining what the network has learned due to the network's memory access.
-This gives special meaning to the image-based networks because of our ability to then "see" clusters of pixels alongside the network.

For more reading, check out:

-Training and Analyzing Deep RNNs,
-A Neural Attention Model for Sentence Summarization
-Show, Attend and Tell: Neural Image Caption Generation with Visual Attention for a start.

•Generator-Encoder Rationales

-Great paper and library which shows a method of generating smaller rationales
-using phrases from the text for several NLP tasks including multi-aspect sentiment analysis.

(Source: Kjamastian Inc., Towards Interpretable Reliable Models 19 October 2017, https://blog.kjamistan.com/towards-interpretable-reliable-models/)

Open Source Tools(3): Other Tools and Notebooks

- **YellowBrick**
  - -Data Visualization library aimed at making visual explanations easier.
  - -I have so far only played around with this for data exploration, not for explaining models, but I am curious to hear your experience!
- **MMD-critic**
  - -A meaningful approach to sampling!
  - -Google Brain resident Been Kim also wrote an accompanying paper which explains how this library works to help you sample
- **Ian Ozsvald's Notebook using Eli5**
  - -Ian and I have been chatting about these libraries, and I asked him to continue to update and elaborate his own use of tools like eli5.
  - -Updates will come as well, so check back!
- **Bayesian Belief Networks**
  - -Probabilistic Programming is cool again! (or always was... probably?)
  - -This is one of *many* libraries you can use for building Bayesian networks.

(Source: Kjamastian Inc., Towards Interpretable Reliable Models 19 October 2017, https://blog.kjamistan.com/towards-interpretable-reliable-models/)

Current Tools
: Limitations and Restrictions

AI의 최종 결정 추론과정 불투명=>오류의 원인을 모른다!

AI 신경망이 데이터와 어떤 연결-결합의 과정을 거치는가?

Gartner: 3가지 설명가능한 AI 방안

"AI 모델의 대다수가
의사 결정에 도달하는 과정을 제대로 설명하지 못하는
복잡한 블랙박스와 같다.
이런 복잡한 AI 모델은 이를 사용하는 이들에게 많은 영향을 끼친다"

1. 설명의 필요성을 파악하기 위해 비즈니스 관계자와 논의하고,
   - AI 모델이 운용될 전체 시스템 의 콘텍스트를 설명한다.
2. 기존 데이터를 활용하거나, 모델을 설명하고, 결과를 간소화하거나
   혹은 사용자가 이해할 수 있는 방식으로 데이터를 제공함으로써
   -학습 데이터에 대한 가시성을 제공한다.
3. 비즈니스 관계자들이 설명 가능한 AI 모델과 더욱 정확한 AI 모델 가운데
   -자사의 상황에 가 장 부합하는 방식을 선택할 수 있는 권한을 부여한다.

EU의 GDPR(2018. 5)➔설명가능한 AI 요구

- EU의 GDPR(General Data Protection Regulation)

  -유럽연합 시민은 법 적 효력을 초래하거나 이와 유사하게 중대한 영향을 미치는 사항에 대해 프로파일링 등 자동화 된 처리의 적용을 받지 않을 권리를 갖는다고 규정

- GDPR 13조: Right to be Informed
- GDPR 22조: 프로파일링을 포함한 자동화된 의사결정

  -알고리즘에 의한 자동화된 결정 반대

  -인간의 개입을 요구할 권리 규정

  -알고리즘의 결정에 대한 설명을 요구하고 이에 반대할 권리 규정

Xai Case Studies

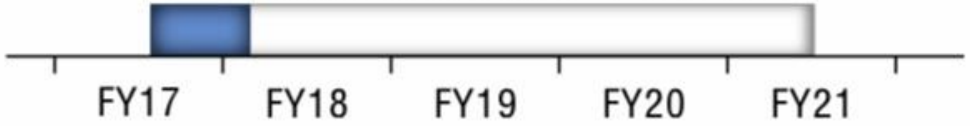
1. 미국방성 DARPA(2016-2021.4): 설명가능한 AI 제시

2. IBM AI Fairness 360(2017. 11)

Xai Case Study (1)
미국방성 DARPA(2016-2021.4)
: 설명가능한 AI 제시

Defense
Advanced
Research
Projects
Agency

Explainable Artificial Intelligence (XAI)



David Gunning
DARPA/I2O

Program Update November 2017

미국방성 DARPA(2016-2021.4): 기본 개념

기존 AI 학습
-어떻게 이런 결과가 나왔는가?

설명 모델과 설명 인터페이스 구현

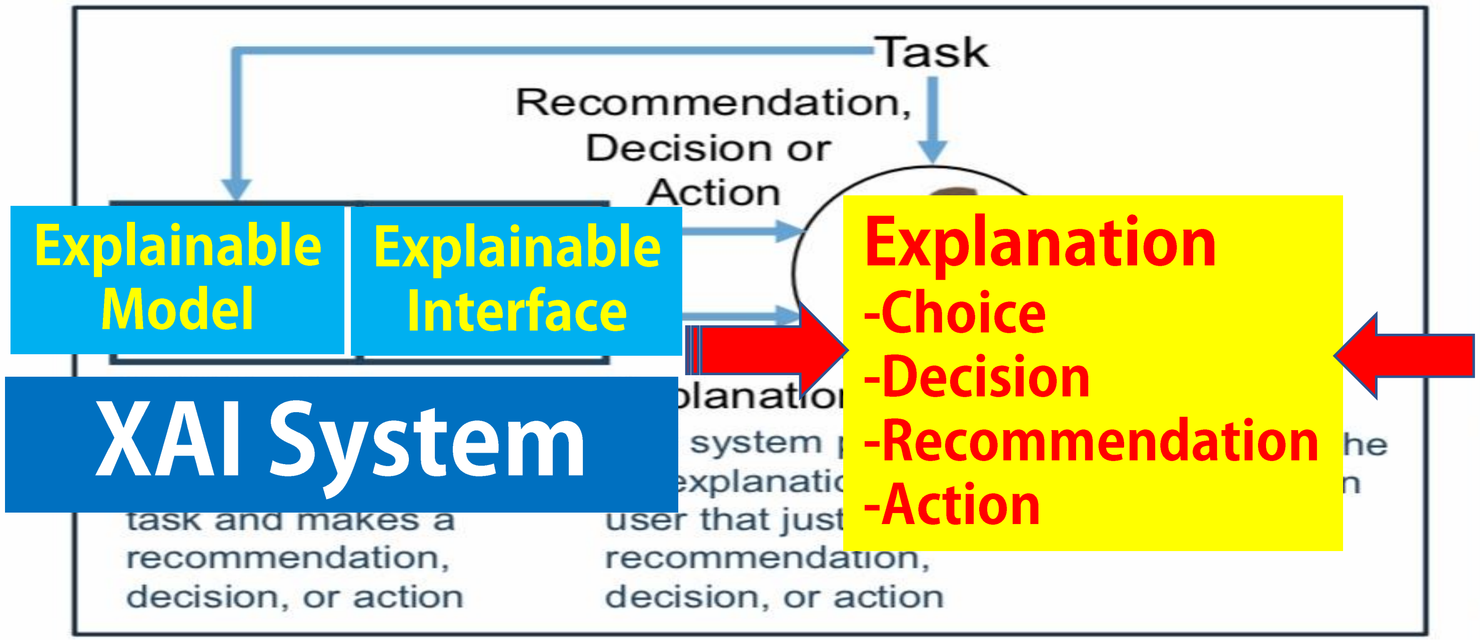
설명가능한 AI

- 결론 과정을 이해할 수 있다
- 성공한 이유를 이해할 수 있다
- 실패한 원인을 파악할 수 있다
- 신뢰할 수 있다
- 에러가 난 이유를 알 수 있다

출처: DARPA(201) : Defense Advanced Research Projects Agency/인터넷 원형 ARPANET 개발

Measuring Explanation Effectiveness

Explanation Framework



Evaluation

User Satisfaction
- Clarity of the explanation (user rating)
- Utility of the explanation (user rating)

Mental Model
- Understanding individual decisions
- Understanding the overall model
- Strength/weakness assessment
- 'What will it do' prediction
- 'How do I intervene' prediction

Task Performance
- Does the explanation improve the user's decision, task performance?
- Artificial decision tasks introduced to diagnose the user's understanding

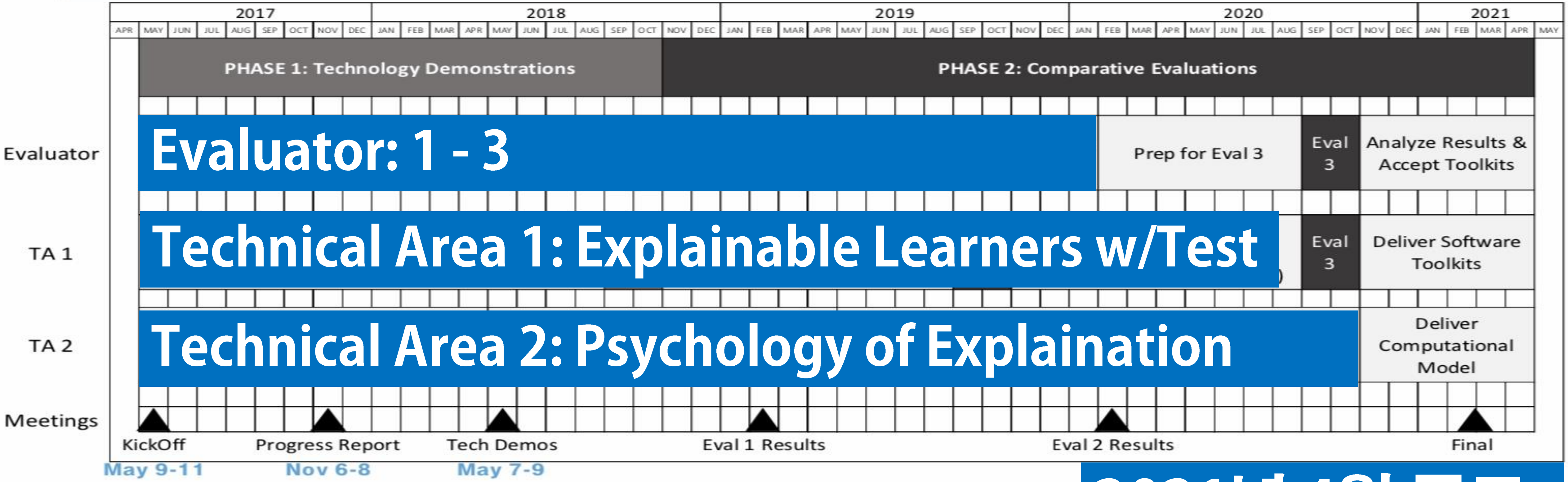
Trust Asssesment
- Appropriate future use and trust

Corrections
Identifying Errors
Correcting Errors
Continuous Training

Schedule

DARPA

XAI
EXPLAINABLE ARTIFICIAL INTELLIGENCE



2021년 4월 종료

- Technical Area 1 (Explainable Learners) Milestones
  - Demonstrate the explainable learners against problems proposed b
  - Demonstrate the explainable learners against common problems (Phase 2)
  - Deliver software libraries and toolkits (at the end of Phase 2)
- Technical Area 2 (Psychology of Explanation) Milestones
  - Deliver an interim report on psychological theories (after 6 months during
  - Deliver a final report on psychological theories (after 12 months, during Ph
  - Deliver a computational model of explanation (after 24 months, during Pha
  - Deliver the computational model software (at the end of Phase 2)

Toolkits
Model

IBM AI Fairness 360(2018.3) : 18명 참여

AI FAIRNESS 360: AN EXTENSIBLE TOOLKIT FOR DETECTING, UNDERSTANDING, AND MITIGATING UNWANTED ALGORITHMIC BIAS

Rachel K. E. Bellamy [1] Kuntal Dey [2] Michael Hind [1] Samuel C. Hoffman [1] Stephanie Houde [1]
Kalapriya Kannan [3] Pranay Lohia [3] Jacquelyn Martino [1] Sameep Mehta [3] Aleksandra Mojsilovic [1]
Seema Nagar [3] Karthikeyan Natesan Ramamurthy [1] John Richards [1] Diptikalyan Saha [3] Prasanna Sattigeri [1]
Moninder Singh [1] Kush R. Varshney [1] Yunfeng Zhang [1]

AI Fairness 360 - Concept

- A new open source Python toolkit for algorithmic fairness - https://github.com/ibm/aif360
- The main objectives of this toolkit are
  - to help facilitate the transition of fairness research algorithms to use in an industrial setting
  - and to provide a common framework
    for fairness researchers to share and evaluate algorithms.
- The package includes
  - a comprehensive set of fairness metrics for datasets and models,
    explanations for these metrics,
    and algorithms to mitigate bias in datasets and models.
- It also includes an interactive Web experience (https://aif360.mybluemix.net) that
  - provides a gentle introduction to the concepts and capabilities for line-of-business users,
  - as well as extensive documentation, usage guidance, and industry-specific tutorials
    to enable data scientists and practitioners
    to incorporate the most appropriate tool
    for their problem into their work products.
- Performance: A built-in testing infrastructure maintains code quality

AI Fairness 360: Processing Pipeline

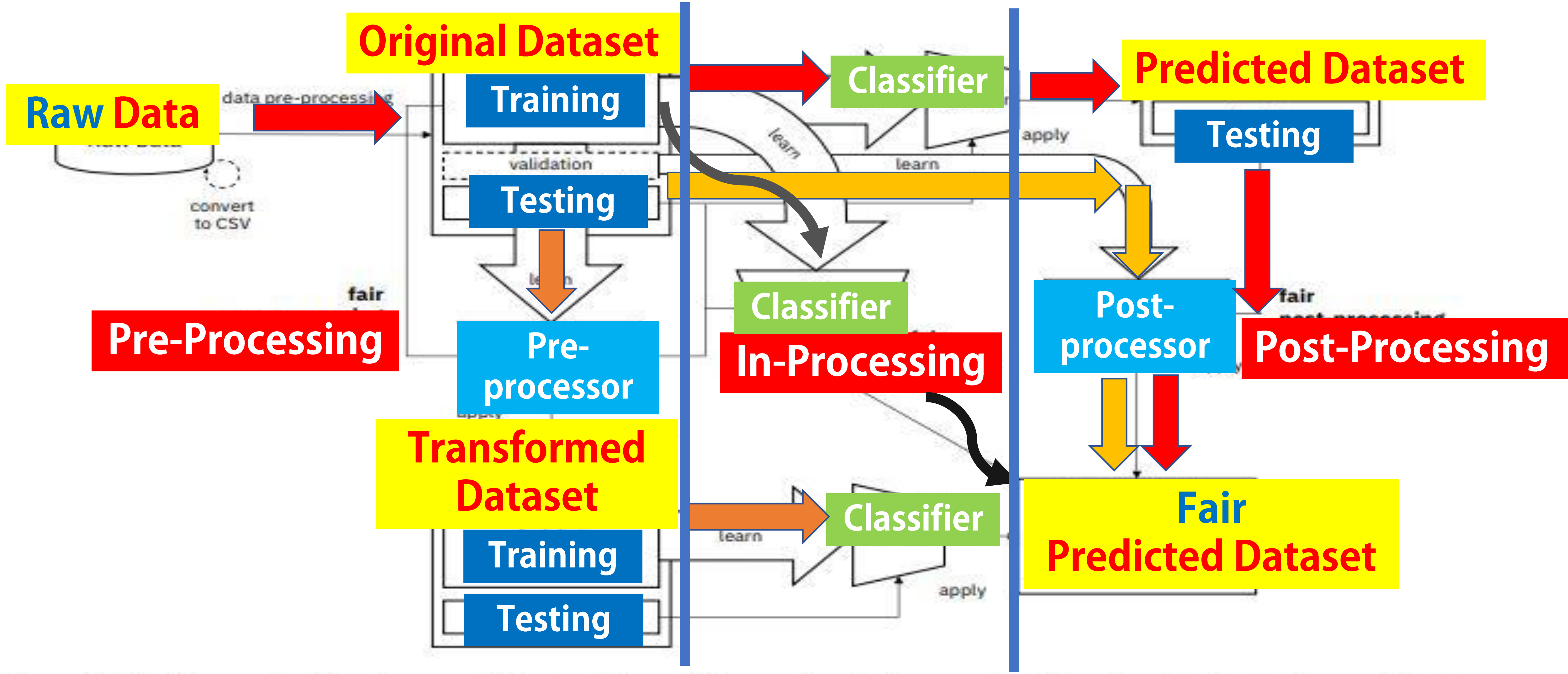


Figure 1. The fairness pipeline. An example instantiation of this generic pipeline consists of loading data into a dataset object, transforming it into a fairer dataset using a fair pre-processing algorithm, learning a classifier from this transformed dataset, and obtaining predictions from this classifier. Metrics can be calculated on the original, transformed, and predicted datasets as well as between the transformed and predicted datasets. Many other instantiations are also possible.

AI Fairness 360 Pipeline: Algorithms

Algorithms AIF 360 currently contains 9 Bias Mitigation Algorithms
- that span these three categories.

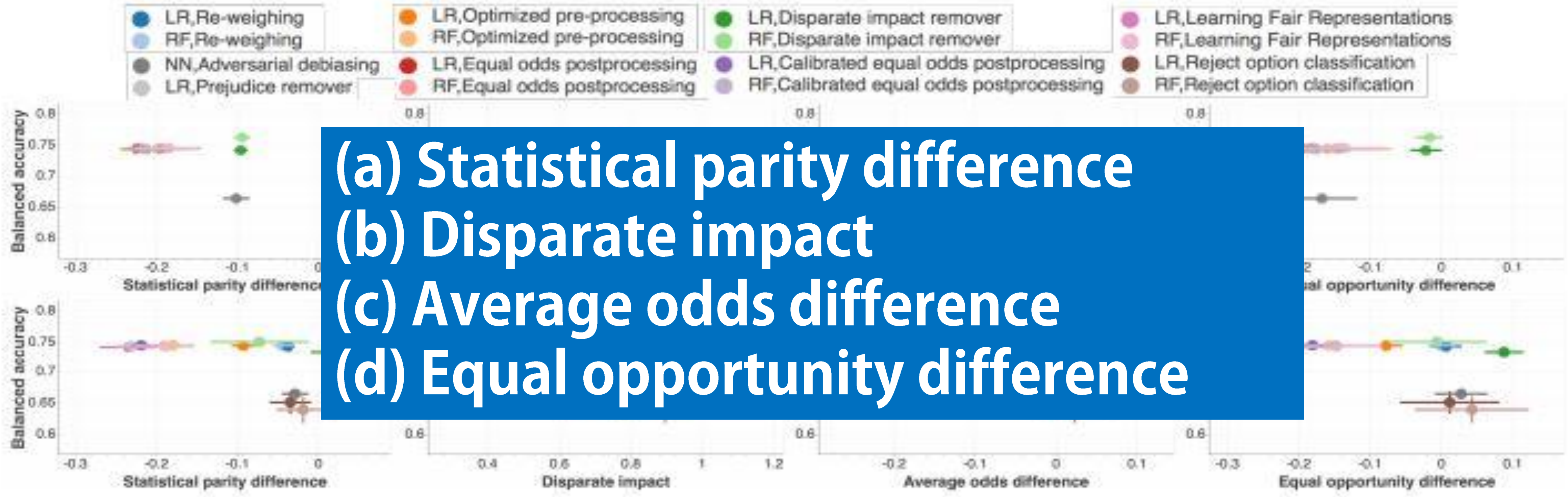
All the algorithms are implemented
-by inheriting from the Transformer class.

Transformers are an abstraction for any process
-that acts on an instance of Dataset class
-and returns a new, modified Dataset object.

This definition encompasses
-Pre-processing, In-processing, and Post-processing Algorithms.

Four Different Fairness Metrics

Figure 4. Statistical Parity Difference (SPD) and Disparate Impact (DI) before (blue bar) and after (orange bar) applying pre-processing algorithms on various datasets for different protected attributes. The dark gray bars indicate the extent of ±1 standard deviation. The ideal fair value of SPD is 0 and DI is 1.



(a) Statistical parity difference
(b) Disparate impact
(c) Average odds difference
(d) Equal opportunity difference

(a) Statistical parity difference (b) Disparate impact (c) Average odds difference (d) Equal opportunity difference

Figure 5. Fairness vs. Balanced Accuracy before (top panel) and after (bottom panel) applying various bias mitigation algorithms. Four different fairness metrics are shown. In most cases two classifiers (Logistic regression - LR or Random forest classifier - RF) were used. The ideal fair value of disparate impact is 1, whereas for all other metrics it is 0. The circles indicate the mean value and bars indicate the extent of ±1 standard deviation. Dataset: *Adult*, Protected attribute: *race*.

K-Xai Engine: L-TTE System(V.1)

- Architecture and Pipeline -

K-Xai Engine: L-TTEC Architecture

Learning-Training-Testing-Evaluation and Correction Connected System(v.1)

K-Xai Engine: L-TTEC Connected

RAW DATA

Learning PKG

Learning Models

-Neural Nets
-Deep Learnings
-Pattern Theory
-Probabilistic Logic
-Explainable RL
-Causal Modelling
-Cognitive Modelling
-Single-modal & Multi-modal Hybrid Models

Training PKG

Curriculum Learning

-Reinforcement Learning
-Adaptive Models
-Model Induction
-Graphical Models
-Markov Models
-Bayesian Belief Net
-SRFs: CRFs, HBNs, MLNs
-Ensemble Methods
-Random Forests
-Decision Trees

Test-Evalu PKG

Testing Metrics

-E-Model Metrics
-I-Model Metrics

Evaluation Metrics

- Task Performance
- MentalSatisfaction
- Trust Assessment

Ethics Metrics

-Guidelines: GDPRs
-Rules
-Regulations
-Law

Corrections

-Identifying Errors
-Applying Changes

Interface

Fair Dataset

Explanation

- Prediction
- Decision
- Recommendation
- Text, Image, Audio, Video

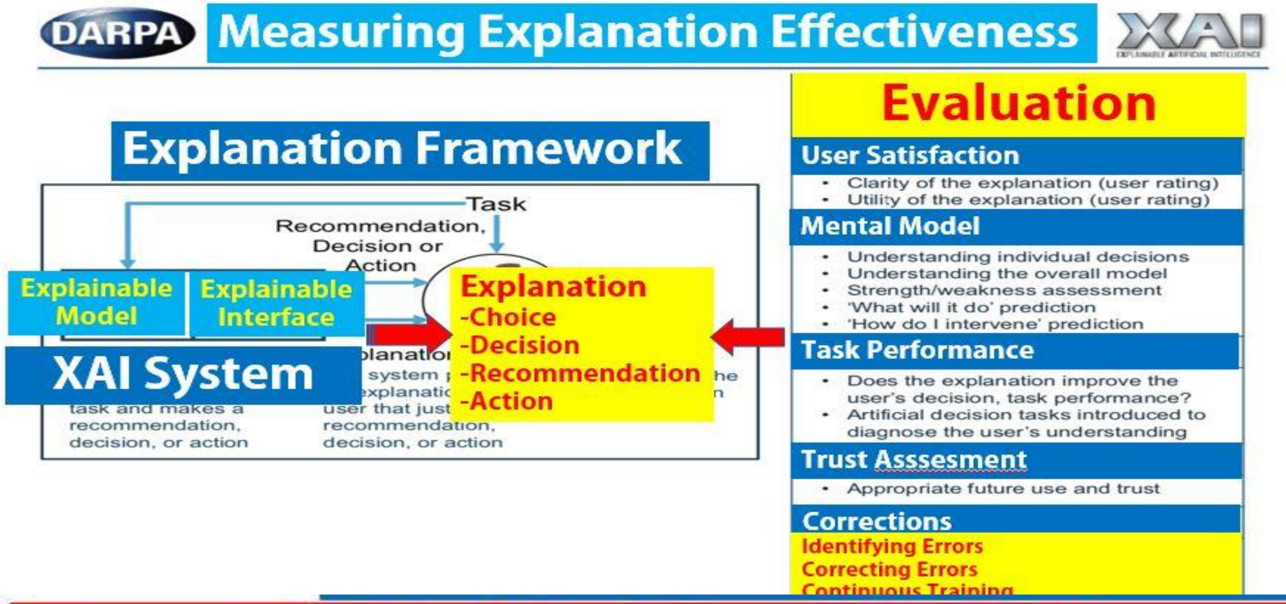
Interpretation

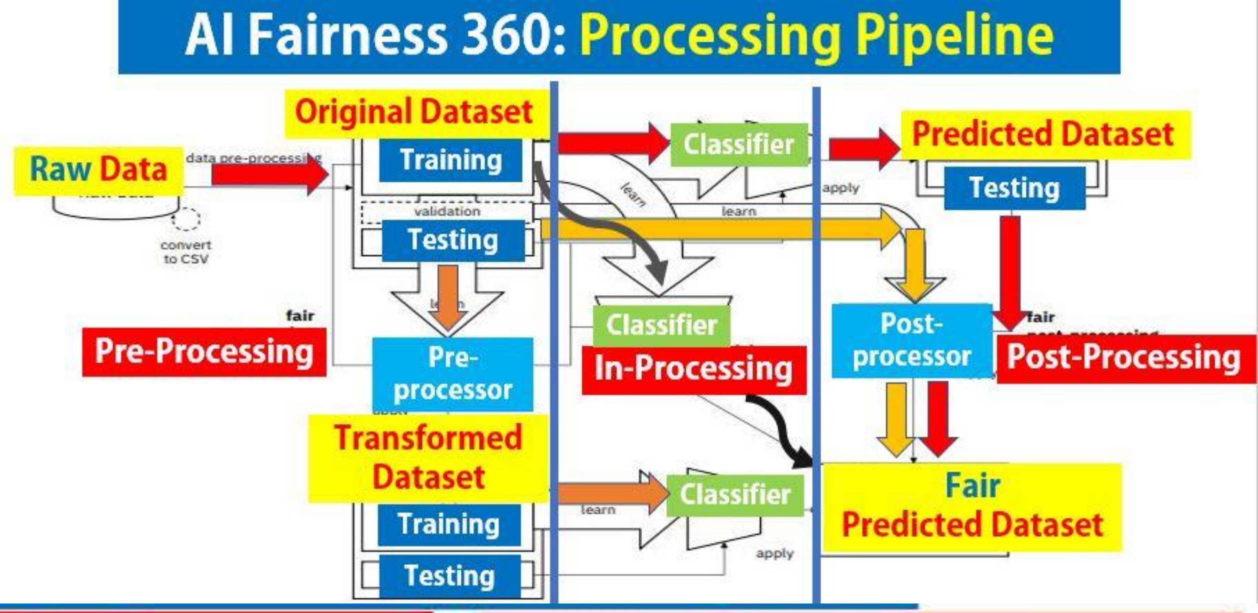
- Decision Diagram
- Narrative Generation
- Interactive Visualization
- Descriptive Generation
- Show and Tell
- Argumentation and Pedagogy

ACTION

AiContents LAB

CPU + GPU





K-Xai Engine: L-TTE Connected

Interface

R A W D A T A

Learning PKG	Training PKG	Test-Evalu PKG
Learning Models -Neural Nets -Deep Learnings -Pattern Theory -Probabilistic Logic -Explainable RL -Causal Modelling -Cognitive Modelling -Single-modal & Multi-modal Hybrid Models	**Curriculum Learning** -Reinforcement Learning -Adaptive Models -Model Induction -Graphical Models -Markov Models -Bayesian Belief Net -SRFs: CRFs, HBNs, MLNs -Ensemble Methods -Random Forests -Decision Trees	**Testing Metrics** -E-Model Metrics -I-Model Metrics **Evaluation Metrics** - Task Performance - MentalSatisfaction - Trust Assessment **Ethics Metrics** -Guidelines: GDPRs -Rules -Regulations -Law **Corrections** -Identifying Errors -Applying Changes

Fair Dataset

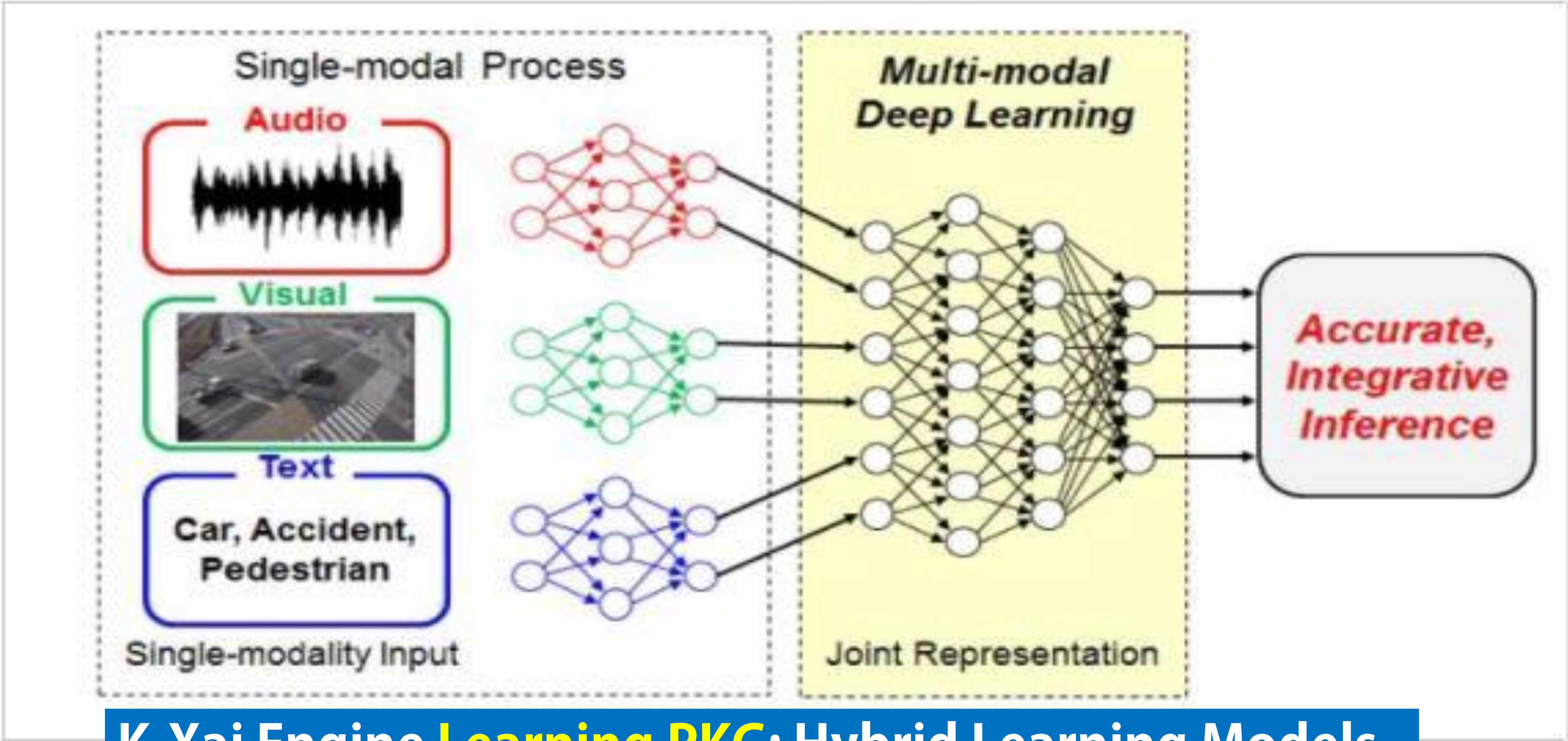
Explanation
- Prediction
- Decision
- Recommendation
- Text, Image, Audio, Video

Interpretation
- Decision Diagram
- Narrative Generation
- Interactive Visualization
- Descriptive Generation
- Show and Tell
- Argumentation and Pedagogy

A C T I O N

AiContents LAB

CPU + GPU



K-Xai Engine Learning PKG: Hybrid Learning Models

음성, 영상, 텍스트 멀티 모드 복합 (Multi-modal Hybrid) 인공지능의 발전 방향, [출처 =KAIST]

K-Xai Engine Training PKG Models

Accuracy-based Curriculum Learning In Deep Reinforcement Learning(2018)

Automated Curriculum Learning for Neural Networks(2017)

Accuracy-based Curriculum Learning in Deep Reinforcement Learning

Pierre Fournier[1] Mohamed Chetouani[1] Pierre-Yves Oudeyer[2] Olivier Sigaud[1,2]

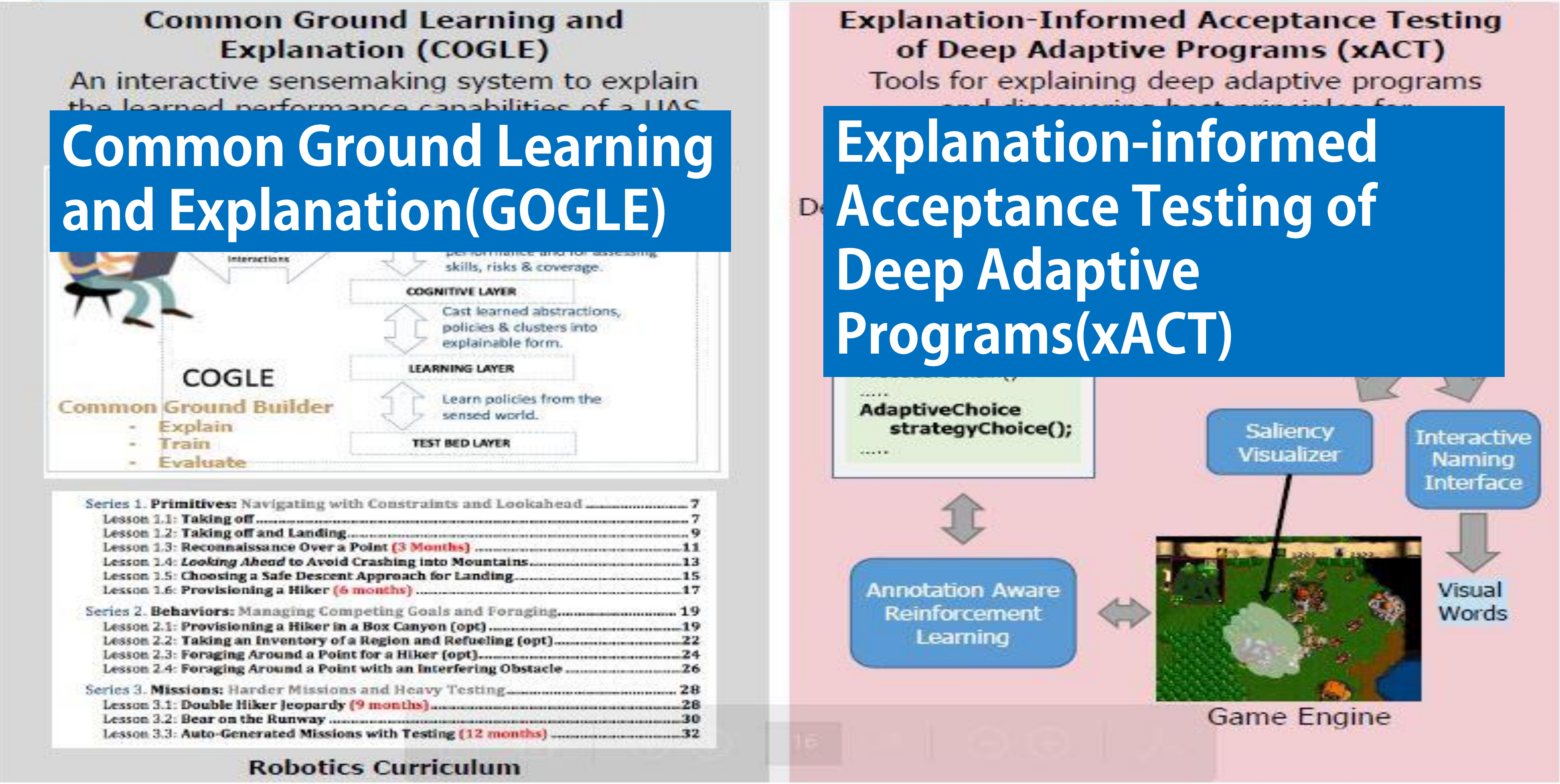
Automated Curriculum Learning for Neural Networks

Alex Graves[1] Marc G. Bellemare[1] Jacob Menick[1] Rémi Munos[1] Koray Kavukcuoglu[1]

Test and Evaluation PKG

Common Ground Learning and Explanation(GOGLE)

Explanation-informed Acceptance Testing of Deep Adaptive Programs(xACT)



Explainability: When to USE?

AI의 최종 결정 추론과정 불투명

Black Box

4 Applications

알고리즘 규제

지식공학적 판단

시스템자체 판단

Explainable AI

(1) Data

(2) Process

(3) Application

Trust

AI 신경망이 데이터와 어떤 연결-결합의 과정을 거치는가?

Data Pipeline

Interpretability: When to USE?

- To increase Trust and Accountability
- Decisions about Humans
- Critical Applications that decide about Life and Death
- To test newly developed models or systems with unknown results
- Debugging the Models
- When the Loss Function does not cover all Constraints
- Models using proxies instead of casual inputs

(Source: Finale Doshi-Velez and Been Kim, *Towards A Rigorous Science of Interpretable Machine Learning*(2017)

AI Ethics and Governance System

- Machine Learning Algorithm and Data Ethics

- Toward AI Governance System

Machine Learning Algorithm and Data Ethics

- **머신러닝 3가지 학습방법**

  - 지도학습, 비지도학습, 그리고 강화학습 등
    : 학습과정에서 기술적-윤리적 문제가 표출될 가능성

- **알고리즘 상의 문제 4가지**

  - 알고리즘의 범용성 문제
  - 실제 적용하기 힘든 연구실용 알고리즘,
  - 설명하기 어려운 딥러닝 알고리즘
  - 악의적 데이터에 취약한 알고리즘 등

- **데이터가 가질 수 있는 문제 5가지**

  - 편향된 데이터
  - 비현실적 데이터
  - 테스트용 데이터와 트레이닝 데이터의 혼재
  - 테스트 데이터의 과적합
  - 데이터 프라이버시 등

알고리즘/데이터 파이프라인 구축과정에 세밀하게 제시
- 개발자가 인지하도록 해야 하며,
- 해당 도메인 특성에 맞는 알고리즘이 적용되었는지?
- 데이터 윤리에 부합하는 지? 등등

AI Algorithm Pipeline

"Code is Law!"
-Lessig(1999)

알고리즘 파이프라인 체크

① 알고리즘의 범용성 문제?
② 실제 적용하기 힘든 연구실용 알고리즘?
③ 설명하기 어려운 딥러닝 알고리즘?
④ 악의적 데이터에 취약한 알고리즘?

"From Code is Law to Law is Code!"
-De Filippi & Hassan(2016)

알고리즘은 입력(input)으로 부터 출력(output)을 만드는 과정이다.

알고리즘 분석 ====> 설계 ====> 개발 ====> 실행 ====> 평가

- 필요성 정립
- 문제설정 및 결과예측
- 활용도구(Tools) 결정
- 수행시간 예측

- 문제 해결과정 코딩
- 지적 추상화-구조화
- 결과 설명모델 설계
- 결정의 편향성 체크
- 수행시간 기준설정
  : 특정 행 수행 횟수 등
- 해킹침입 여부 판단
- 프라이버시 침해여부

- 파일럿 테스트 실행
  -무결성 확인
  -메모리 등 자원사용 효율성
  -적합성 판단과 효율성 평가
  -예측치 못한 출력결과 분석
- 출력결과의 유해성 판단
- 실행과정의 투명성 판단
- 수행시간 평가

컨텍스트별 문제해결 구조화 과정 자세히 구현 => 투명성 확보

AI Data Pipeline

Garbage IN,
and Garbage OUT!

데이터 파이프라인 체크
① 편향된 데이터 여부
② 비현실적 데이터 여부
③ 테스트용 데이터와 트레이닝 데이터의 혼재
④ 테스트 데이터의 과적합 여부
⑤ 데이터 프라이버시 침해 여부

데이터 분석 ====> 설계 ====> 개발 ====> 실행 ====> 평가

- AI 개발 목적과 필요성 분석
- 필요 데이터자원
  - 구조화 데이터
  - 비구조화 데이터
- 데이터 소스
- 데이터 품질
- 데이터 가시성

- 컨텍스별 알고리즘 연결 적용
- 판단과 결정의 편향성 체크
- 판단과 결정의 도덕성 체크
- 데이터의 프라이버시 침해여부
- 원칙과 가이드라인 법 적용
- 해킹가능성 체크

- 파일럿 테스트 실행
  -출력결과의 적합성 평가
  -예측치 못한 출력결과 분석
- 출력결과의 유해성 판단
- 출력결과의 편향성 판단
- 출력결과의 도덕성 판단
- 데이터 수정여부 판단
- 원칙과 가이드라인 법 준수?
- 국제표준 준수?

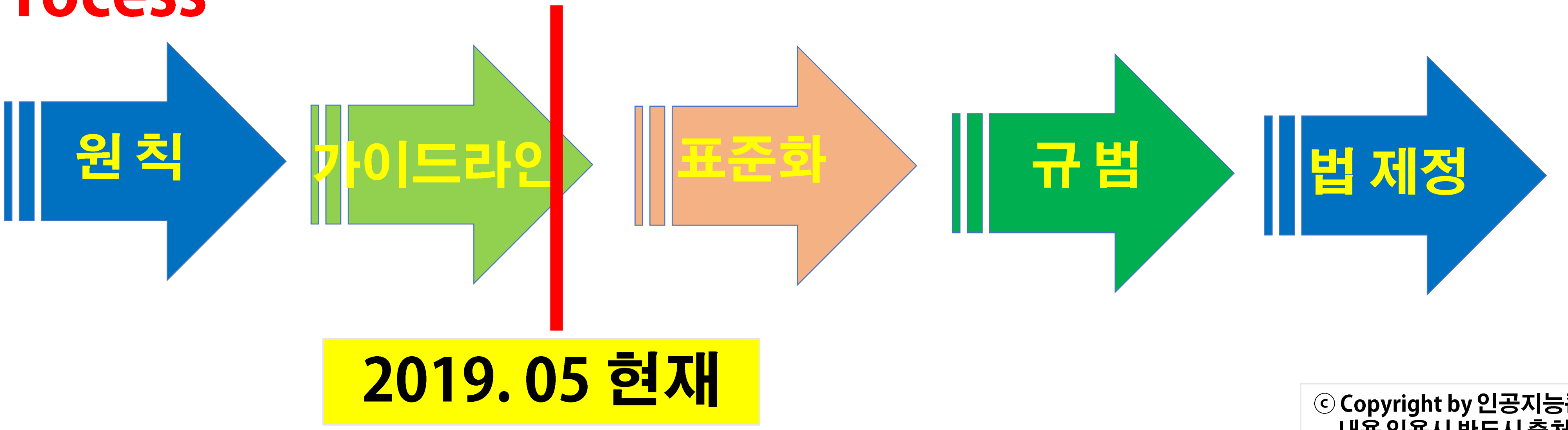
알고리즘/데이터 구현: AI 윤리 전문가 참여 필요

Toward AI Governance: Coverage and Process

• Coverage

- 인공지능 윤리(Ethics)
- 인공지능 보안과 안전, 프라이버시(Security, Safety, and Privacy)
- 인공지능 가이드라인, 규정, 원칙과 규칙, 법률(Guideline, Regulation, Principle, Law)
- 인공지능 교육 & 훈련(Education and Training)

• Process



Toward AI Governance System

AI Ethics Governance Encompasses:

1. 기업 윤리
2. 연구-개발 윤리
3. 제작 윤리
4. 서비스 윤리
5. 사용자 윤리
6. 관리자 윤리
7. 정책 윤리

AI 윤리 포럼 개최(2019. 7. 18(목) 예정

AI 윤리와 교육, 법 그리고 가버넌스 시스템

AI Ethics, Education, Law and Governance System

진행: 안종훈 박사(한국인공지능협회 윤리분과위원장)

- Korea AI EXPO, 2019 -

What's Next?

1. K-Xai Engine and Microservices(w/Containers)
2. AI in the Pocket: Chatbot, AI Agents Platform
3. Post-Human: Singularity and Omega Point
   - Intelligence, Intellectuality, and Spirituality
4. Human Engineering: Enhancement and Augmentation
5. Human Value Recognition AI
6. Quantum Computing and xG Network

“as soon as it works,
no one calls it AI anymore.”

- John McCarthy(1956)

Xai 관련 프로젝트를 기획하시는
기업이나 연구단체 있으시면 연락주세요.
(010-4546-7576 안종훈 박사)

References in Brief

Alex Graves, Marc G. Bellemare, Jacob Menick, R ´ emi Munos, and Koray Kavukcuoglu, *Automated Curriculum Learning for Neural Networks*, Proceedings of the 34 th International Conference on Machine Learning, Sydney, Australia, PMLR 70. (2017)

Carl Miller, The Death of Gods(2018)

CIO Korea, ‘AI의 블랙박스화’ 막겠다… 오픈소스로 알고리즘 공개 선언한 IBM.

CIO Korea, “설명할 수 없는 AI라면 퇴출되어야 한다” IBM 지니 로메티

DARPA, Explainable AI Update(2017). https://www.darpa.mil/attachments/XAIProgramUpdate.pdf.

Finale Doshi-Velez and Been Kim, Towards A Rigorous Science of Interpretable Machine Learning(2017)

IBM, AI FAIRNESS 360: AN EXTENSIBLE TOOLKIT FOR DETECTING, UNDERSTANDING, AND MITIGATING UNWANTED ALGORITHMIC BIAS(2018)

IDG TechReport, “인공지능의 세대교체 “설명가능한 AI” ”, IDG TechReport: Explainable AI.

Info World, *Explainable AI: Peering inside the deep learning black box.*

Katharine Jarmul, *Towards Interpretable Reliable Models*, 19 October 2017. https://blog.kjamistan.com/towards-interpretable-reliable-models.

Movie, Rashomon(1950).

Pierre Fournier, Mohamed Chetouani, Pierre-Yves Oudeyer, and Olivier Sigaud, *Accuracy-based Curriculum Learning in Deep Reinforcement Learning*(2018)

Y. Benkler, "From consumers to users: Shifting the deeper structure of regulations toward sustainable commons and user access"(2000)

외 Reports, Articles, Thesis 등

인공지능콘텐츠LAB

진주혁신도시 윙스타워 B동1001호

- 문의전화 : 055-763-1276
- 상 담 : 010-4546-7576

후마니타스 교육원

- 씨알인문학클럽(매주 화요일)
- 비즈인문학클럽(매주 토요일)

후마니타스 교육원