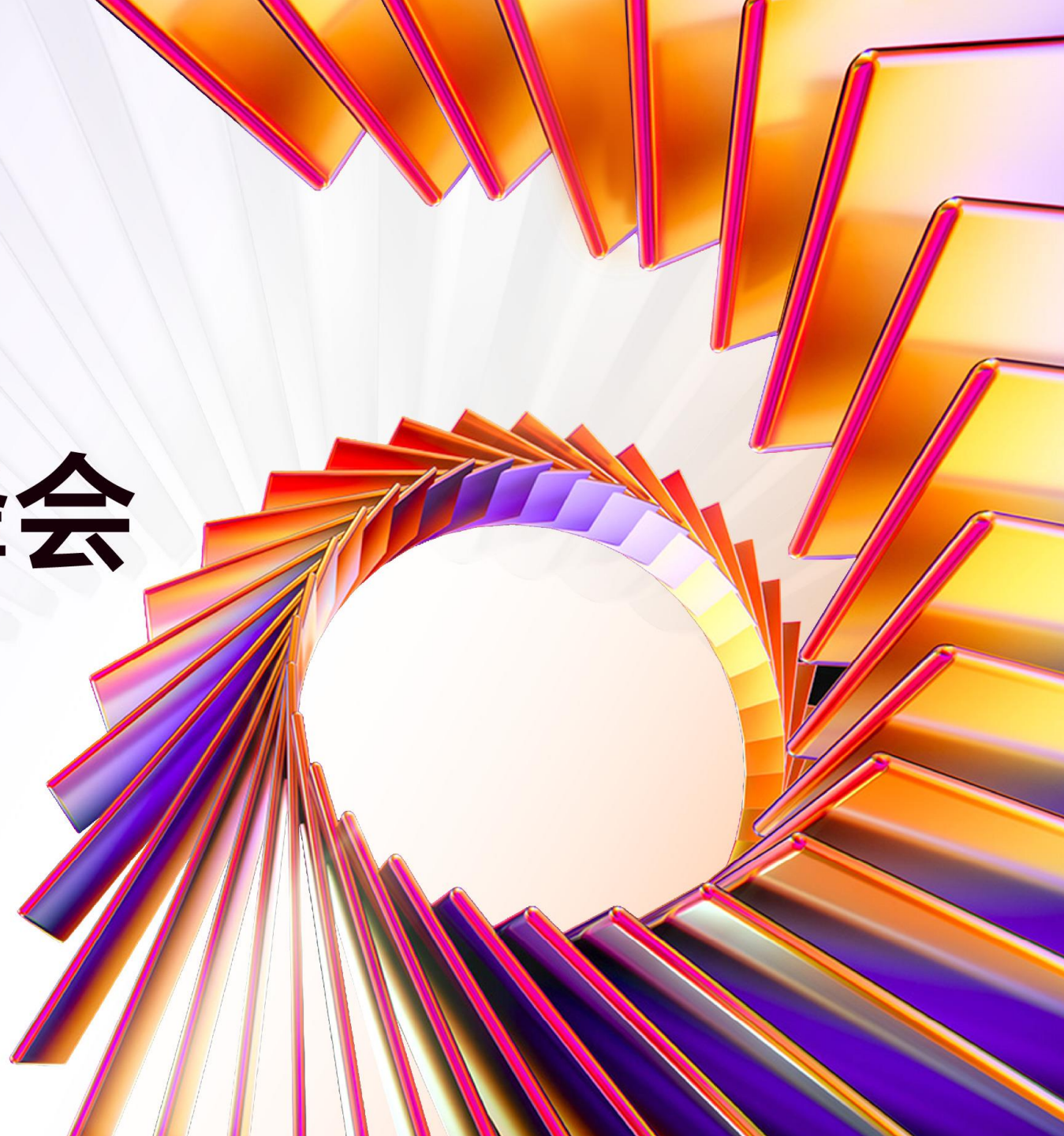
AiDD 6th 2025

2025 AI+ Development Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发

05/23-24 | 上海站



AiDD 7th 2025

2025 AI+研发数字峰会

拥抱AI 重塑研发

AI+ Development Digital Summit

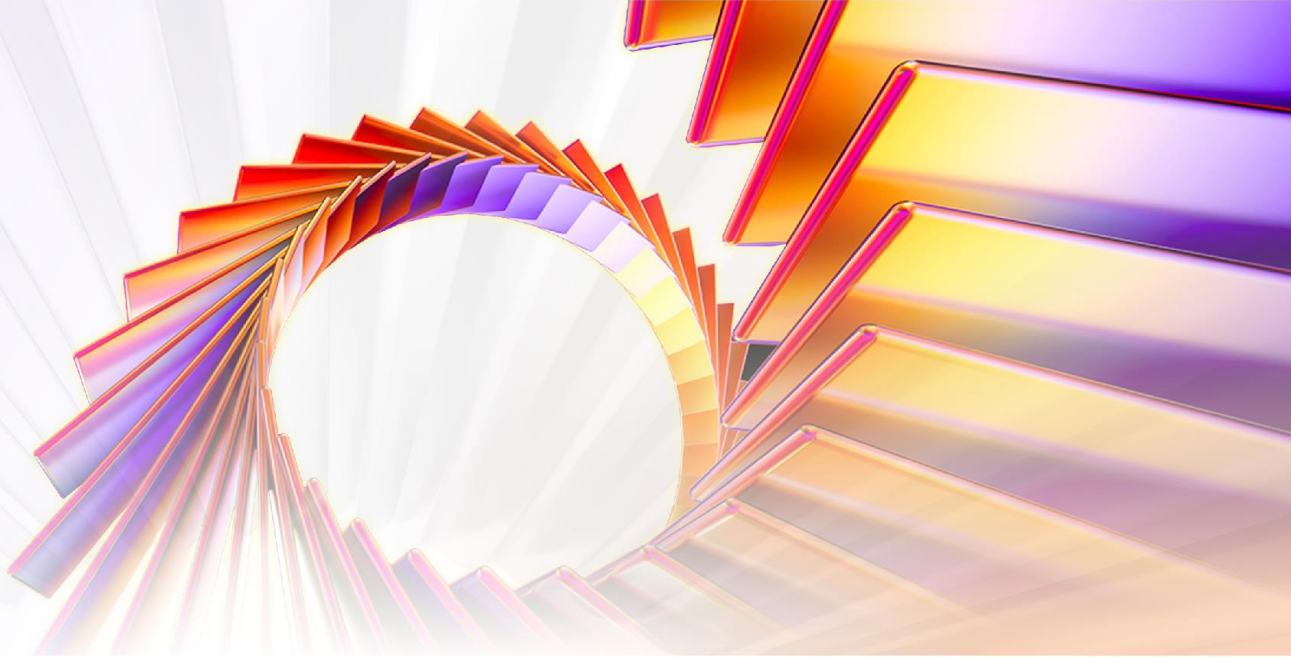
查看会议详情

下一站预告 08/08-09 | 北京站 11/14-15 | 深圳站

北京站论坛设置

- AI+ 金融业务创新
- 大模型安全与对齐
- 具身智能与机器人
- 大模型和 AI 应用评测
- 智能需求工程
- 大模型应用开发框架与实践
- 代码生成及其改进
- 智能存储与检索技术
- 智能体与研发效率工具
- 智能体经济（Agentic Economy）
- AI+ 新能源汽车
- 下一代知识工程
- AI 产品运营与出海策略
- 智能测试工具的开发与应用
- AI 前沿技术探索与实践

AiDD 6th 2025 | 05/23-24 | 上海站
2025 AI+ Development Digital Summit
AI+研发数字峰会
拥抱AI 重塑研发



面向多领域的AI稳健性评估技术与案例分析

千善日 (Chon sun il) | ThinkforBL

千善日 (Chon Sun il)

Manager of ThinkforBL Co., Inc.

- Lead author of the AI Trustworthiness Development Guide, published by the Korean Ministry of Science and ICT, covering all domains : Smart Policing, Hiring, Generative AI, Autonomous Driving, Healthcare, and Public & Social Services **- Compatible with 63 out of 67 detailed verification items in the 'AI RMF' published by the U.S. NIST**
- Developed AI Trustworthiness Verification Techniques and established seven group standards with Korea’s Telecommunications Technology Association (TTA)
- Certified Auditor for AI Management System (ISO/IEC 42001)
- Certified in Functional Safety Verification Frameworks AFSP (Automotive Functional Safety Professional), CACSP (Certified Automotive Cyber Security Professional)
- Master’s Degree in Electronic Engineering from Jeonbuk National University

目录
CONTENTS

17 years since it was established
拥有17年经验的专业咨询公司

Registered over 80 Patents
拥有80多项技术专利

Established 15 Standards & Papers
参与制定15项行业标准，发表多篇技术论文

Received 8 Ministerial Awards
荣获8项国家级权威奖项

R&D consulting for over 400 companies
为400多家企业提供R&D咨询服务

Performance of all national SW quality management projects. Unrivaled performance in patents, standards, and papers in the field of AI trustworthiness and education

承担所有国家级软件质量管理项目，在AI可信赖性与教育领域的专利、标准与论文方面表现卓越，领先业界。

9 Certification 9 项官方认证

- Intellectual Property Management Company (Recertification)
  知识产权管理企业（已通过再次认证）
- Excellent Invention Company (Recertification)
  专利创新优秀企业（已通过再次认证）
- Companies with excellent employment & utilization of female R&D personnel
  就业友好型企业 & 积极支持女性R&D人才的企业
- Family-Friendly Certified Company (Recertification)
  注重员工家庭友好政策的认证企业（已通过再次认证）
- Technology Innovation Small & Medium Enterprises (Inno-Biz)
  技术创新型中小企业（Inno-Biz）
- Best-Value Service Company
  提供卓越价值的服务型企业
- Youth-Friendly Strong Small Enterprise 青年友好型优秀企业
- Certified Venture Company 创新型认证企业
- Company Approved for Alternative Military Service 国防动员企业

Focus on **Professionally** Researching **Trustworthy AI** - 专注于可信赖AI的专业研究

6 Industry guidelines for Trustworthy AI by Korea's Ministry of Science and ICT
我们为韩国科学技术信息通信部制定了《6项可信赖人工智能行业指南》

2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025

Establishment of TTA(Telecommunications Technology Association) standard for AI Reliability assessment methods
制定了AI可信赖性评估方法的TTA（韩国电信技术协会）标准

Standard 标准

Suggestion of Practical Quantification Measuring Method of Test Design Which Can Represent the Current Status
提出能反映当前状态的测试设计的实用量化方法
Suggestion of Testing Method for Industrial Level Cyber-Physical System in Complex Environment
提出在复杂环境中测试工业级信息物理系统的方法

Paper 论文

Investigating and Suggesting the Evaluation Dataset for Image Classification Model
研究图像分类模型的评估数据集并提出建议
Importance of Adaptive Photometric Augmentation for Different Convolutional Neural Network
强调图像光照增强在不同神经网络中的重要性

Patent 专利

Patent applications and registrations related to this from 2015, totaling 8 registrations and 4 applications
自2015年以来，相关专利共申请12项，其中8项已获授权，4项已公开

Education 教育

The first AI trustworthiness training in Korea organized by TTA
韩国电信技术协会（TTA）首次开展AI可信赖性培训课程
Qualification course for National University
韩国国立大学开设的AI可信赖性资格认证课程

Business 实务研究

Industry-wide domination with AI trustworthiness development guides for each sector
制定面向全行业的AI可信赖性实践指南
Research on AI trustworthiness training methods
AI可信赖性相关实训方法研究

What is 'Bias' in AI?

人工智能中的‘偏见’是什么？

In the age of information excess, we must first understand the core

在信息过载的时代，我们必须先理解问题的核心

AI malfunctions are due to data bias. AI的错误行为，源于数据偏向性

Human rights violations, legal disputes, and safety accidents are merely outcomes of AI malfunctions. 侵犯人权、法律纠纷与安全事故，往往只是AI误判的结果

Era of Software 1.0 软件1.0时代

Systems operated based on the code we made.
系统基于我们编写的代码运行
Faults were the result of logical or coding mistakes.
出错通常是由于逻辑漏洞或编码错误

Coding Skill | 编码技能

Era of Software 2.0 软件2.0时代

Instead of programs, AI trained on data makes decisions.
AI依赖训练数据进行学习和判断
Fault, it is because AI learned from biased data.
AI之所以出错，是因为它学习了有偏见的数据

Ethical Issue 伦理问题 | Law issue 法律问题 | Safety issue 安全问题

Gathering data incorrectly | 错误的数据收集

Fails to include possible exceptions or edge cases
忽略了可能出现的异常情况和边缘案例

There is NO such thing as “bad AI” in the world.

这个世界上没有“坏的AI”

There’s only malfunctions where AI behaves contrary to our intentions.

所谓问题，只是AI的行为与我们的本意不一致而已

Have you ever heard of any technique to solve bias? 你听说过可以解决偏见的技术吗？

EU did not suddenly make a law this year to ask people to be more ethical.

欧盟（EU）今年并没有突然制定一项法律来要求人们变得更加有道德

因为人工智能可能带来伦理风险，所以才制定了“从技术上”来解决这个问题的法律。

Rather than only relying on
Moral slogan
不是只停留在道德口号上

we focus on developing the
Necessary technology
关键在于开发真正可用的技术

How to eliminate AI malfunctions? 如何消除AI故障？
How can we test AI robustness? 如何全面测试AI的鲁棒性？

Evaluation method, which examines AI robustness in depth without exception
一种全面评估AI鲁棒性的方法，覆盖所有关键环节

ISO/IEC 25059:2026 Refer to “quality property” | 中关于“质量属性”的定义

AI system evaluation | AI系统评估

- **Robustness 鲁棒性**
- **Explain-ability 可解释性**
- **Alignment 一致性/对齐性**

Requirement 2. Evaluation Items, Evaluation Criteria, Evaluation Indicators, Procedures etc.

要求 2：评估要素、评估标准、评估指标与评估流程等

USDHS (2023) Threats and mitigation strategies for adversarial artificial intelligence Reference: "Evaluation Items“

-

敌对型AI的威胁与应对策略
引用：评估要素

ISO/IEC/IEEE 29119-2:2021 Refer to “Evaluation process”

-

引用：评估流程

TTA TTAK.KO-11.0250 Part1~Part5 Refer to “Evaluation process” and “evaluation indicators”

-

引用：评估流程与评估指标

ISO/IEC TR 29119-11: 2020 AI-based system testing Reference: "Evaluation Items“

-

基于AI的系统测试
引用：评估要素

NIST (2014) Object and human detection and tracking systems Reference: "Evaluation Indicators“

-

目标与人类检测追踪系统
引用：评估指标

ISO/IEC TR 29119-11: 2020 AI-based system testing Reference: "Evaluation Items“

-

基于AI的系统测试
引用：评估要素

ISO/IEC TR 29119-11: 2020 AI-based system testing Reference: "Evaluation Items“

-

基于AI的系统测试
引用：评估要素

The legitimacy of an evaluation is determined by its credibility | 评估的正当性取决于其公信力

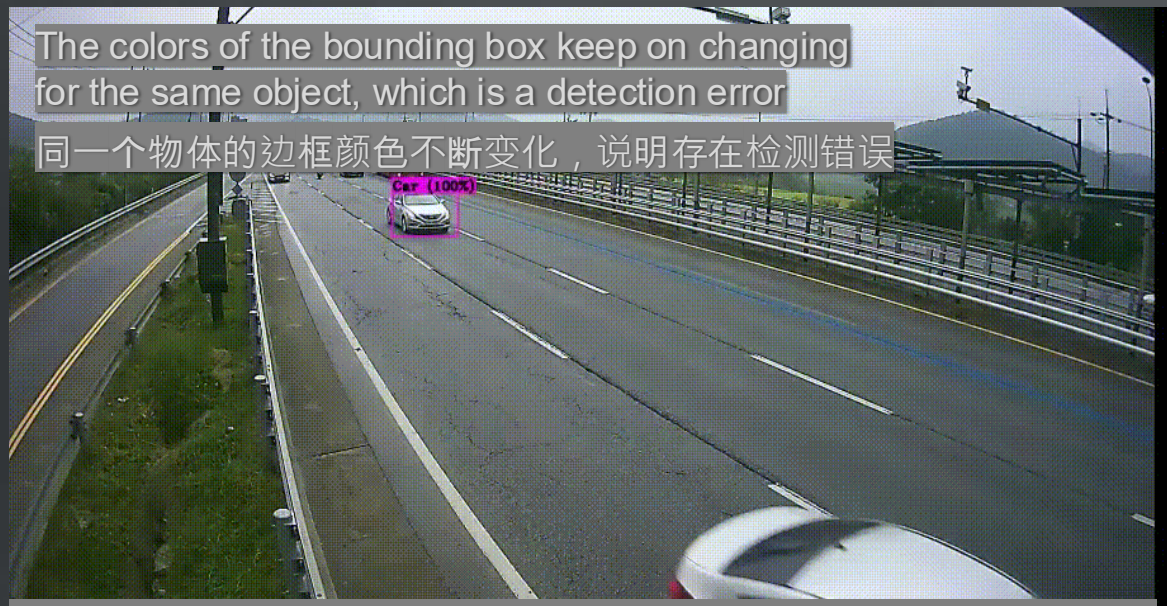
Does plenty data eliminate bias? 数据越多，就能消除偏见吗？

Inaccurate data actually causes greater bias. 错误的数据可能会导致更大的偏见

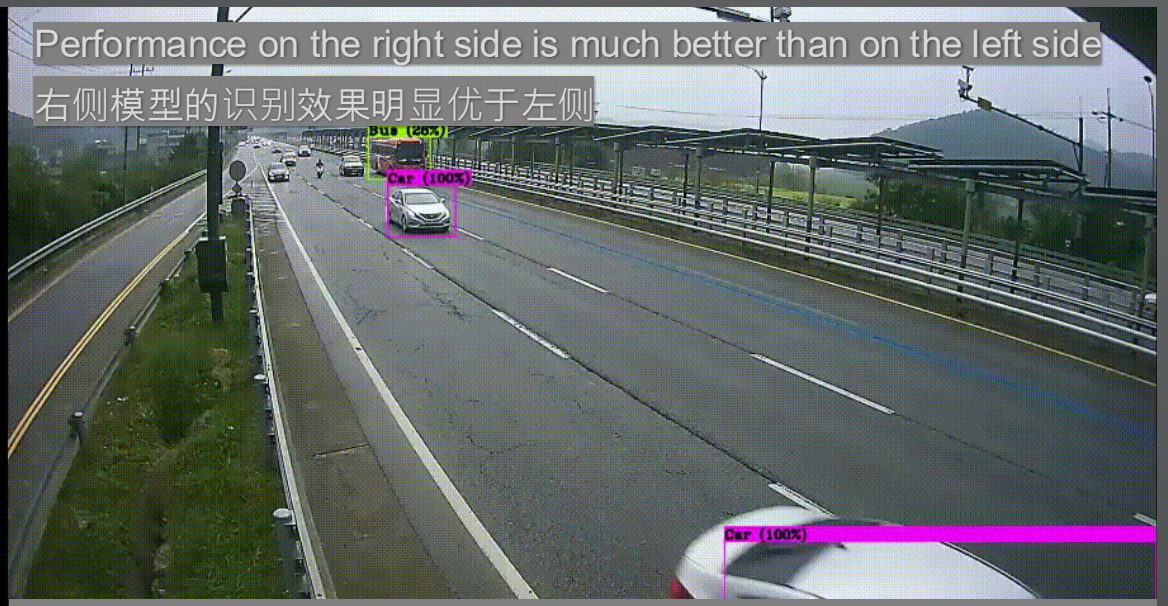
Load Truck
Trailer Truck
Bus
Mini Truck
Car

▼ Public institution shared 50,000 images for training vehicle detection

某公共机构提供了 50,000 张图像，用于车辆类型识别的训练



Used 50,000 images | 使用了 50,000 张图像



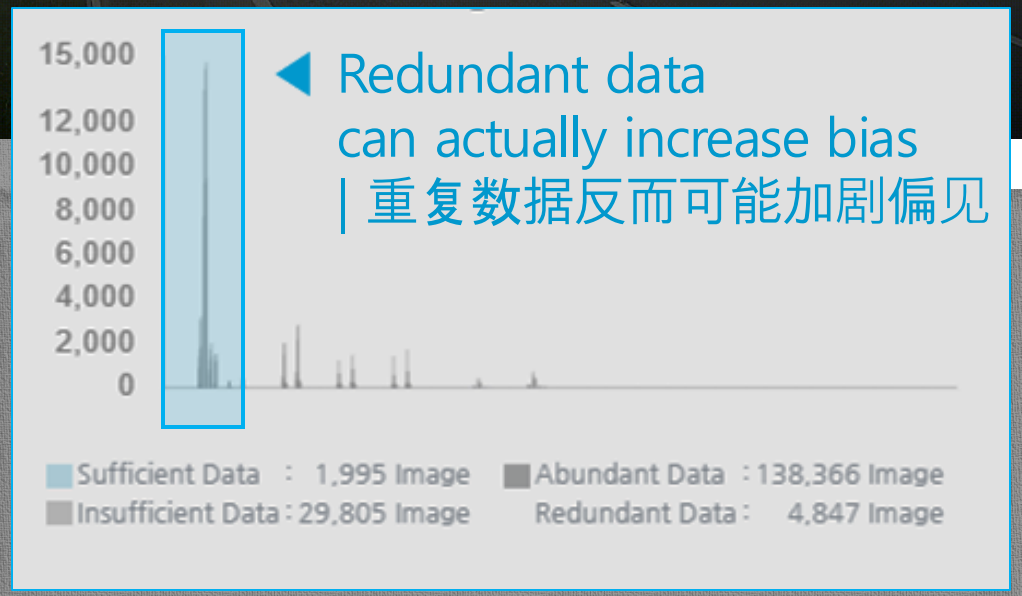
Used only 2,000 images from the 50,000 | 只用了50,000张图片中的2,000张

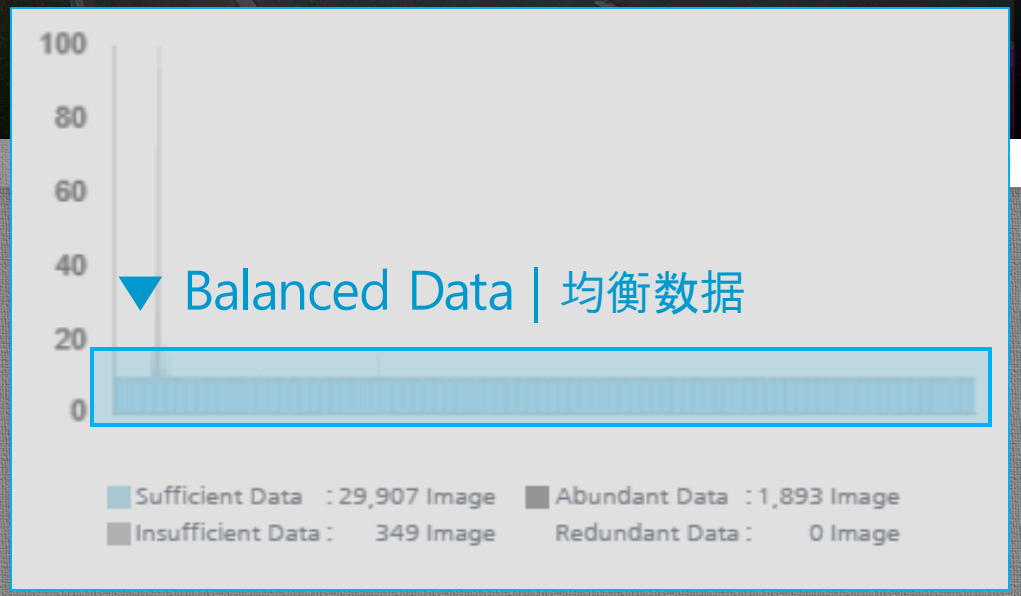
Data Redundancy does not just end with the waste of resources.

数据冗余不仅仅是资源浪费这么简

‘Need more data’ means that the data should include a lot of context.

‘需要更多数据’ 的真正含义是：数据应包含更丰富的语境信息





Solving addition problems a million times won't take you to medical school. 把加法题做一百万遍, 也考不上医科大学

It is not include exceptions and edge cases,
it will be difficult to secure robustness and reliability.
如果数据中不包含异常情况和边缘案例, 那就很难确保系统的鲁棒性与可靠性

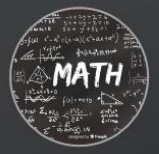
100 tests with redundant data of the **same** meaning |
100个**含有相同含义冗余数据的**测试 |

| Validation of **exceptions** and **edge** cases
| 验证异**常和**边缘**案例**

100 X-rays of the **same** body part |
100 张**相同部位的**X光片 |

| Capturing **various** body parts
| **涵盖不同身**体**部位的采集**数据

100 problems **only** about addition |
100 **道全是加法的**题目 |



| **Various** problems
including calculus, vectors, and linear algebra.
| 题**目涵盖微**积**分、向量和**线**性代**数**等多**种**类型**

The problem is how to technically incorporate the exceptions into the test data, and how to objectively evaluate them.

问题在于如何技术性地将例外情况纳入测试数据，以及如何客观地对其进行评估

Malfunctions in defense AI could murder our own troops.

国防AI一旦故障，可能误伤己方士兵

For our advanced weapons to be trusted 为了让先进武器值得信赖，我们必须确保其安全性

key areas of safety research with Korea's DTaQ(Defense Agency for Technology and Quality), **AI robustness testing**

Increasing of AI in the defense sector | 国际社会对国防能力的关注

1 International interest in defense capabilities
以色列-伊朗冲突、俄乌战争及近期局势所体现的全球对国防能力的关注

Defense AI active in Ukraine and the Middle East... "4 gap with leading countries...
2024. 4. 15. — **Defense AI active in Ukraine** and the Middle East ... " 4 years behind **the leading countries ... Now is the golden time... Ukraine** analyzes information in real time an...

Future War Keyword 'AI' as Seen in Israel-Iran Conflict
2024. 4. 17. — **Israel** is actually the first country in the world to deploy **AI-** based fully autonomous vehicles. It has already deployed medium-sized firearms on the border with the...

2 Enhancing capabilities through AI technology; above all, what matters most is 'safety.'
通过AI技术提升作战能力，但最关键的仍是——"安全"

AI drone's 'obsessive desire for achievement'… Attacks pilot when virtual training is interrupted
AI drone 's ' obsessive desire for achievement '... Attacks pilot when interrupted in virtual

The first battle in human history where AI killed humans took place
The first battle in human history in which AI killed a human was fought ... War is a great tragedy, but ironically, it has a double-sided effect that accelerates the progress of human...

3 With the advancement of Defense 4.0, it is crucial to conduct research on defense AI safety to establish a systematic 'safety net.'

AI data contamination is also fatal to national defense… Defense shield and reliability evaluation are essential
2024. 7. 22. — 'AI **data contamination ' is fatal to national defense as well** ... "Defense shield and reliability **evaluation are essential** " [Geeks] ... Recently, cases of artificial intelligence (AI...

과제 2. 국방 인공지능 기술에 대한 군의 신뢰성 부족 해소
课题二. 解决军方对国防人工智能技术缺乏信任的问题


Evaluation Guidelines
评估指南

Empirical Evidence
实证依据

Research on System Validation Methods for AI-Enabled Weapon Systems | AI武器系统适用的系统验证方法研究

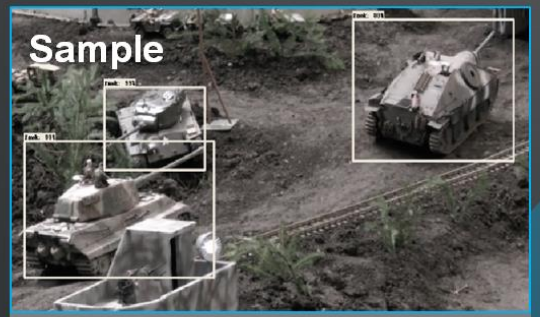
Advancement System Evaluation Techniques for AI-Enabled Weapon Systems | AI武器系统评估技术的高级化(高阶发展)

Possibility of acquiring, high-quality AI weapon systems for our military
| 军方获取高质量人工智能武器系统的可能性

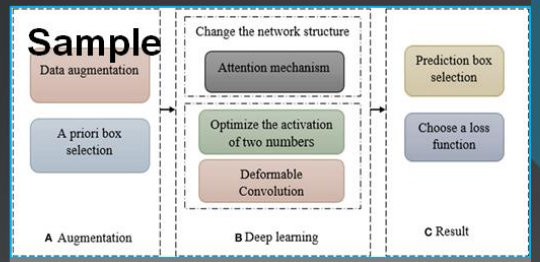
Checking the Profile of the AI System 掌握AI系统的能力全貌

How to make advanced military weapons work 100% all the time?

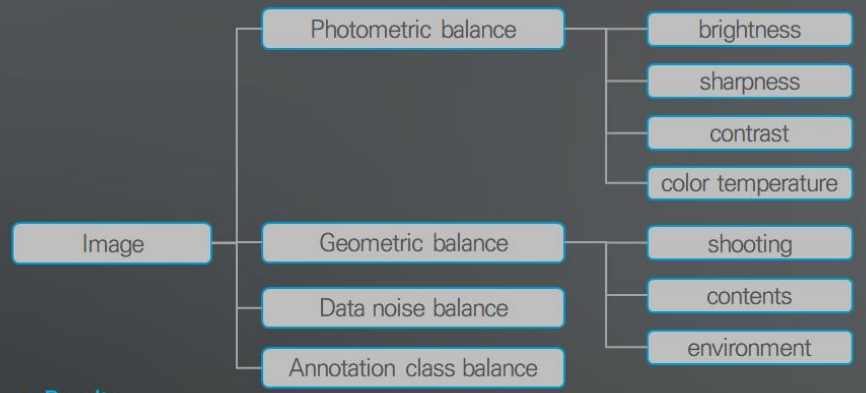
如何确保先进军事武器始终可靠运行？



AI System



AI Specification (Profile)



- **Results**
  - Detection – Anomaly Alert
  - Non-Detection – No Alert
  - Not Detected

- **Photometric Balance**
  - brightness: 0.3 ~ 0.9, Target detection even at low brightness levels
  - sharpness: 0.5 ~ 0.9, Target detection even at relatively low sharpness levels
  - contrast: 0.4 ~ 0.9, Identification of key objects even in low-contrast situations
  - color temperature: 0.4 ~ 0.8, Maintaining object detection capability under various lighting conditions

- **Geometric Balance**
  - Shooting Angle: Field of view 60~80 degrees
  - Content: Primarily focuses on identifying moving objects such as vehicles, people, aircraft, while high-resolution detection is possible for stationary objects.
  - Day/Night: Operates with the same performance during both day and night.
  - Weather: Functions effectively under extreme weather conditions such as rain and snow.
  - Season: Maintains consistent performance across all seasons.
  - Terrain: Operates stably in various terrains, including mountainous areas, plains, coastlines.
  - Other Environments: Maintains high detection capabilities even in fog, dust, sandstorms.

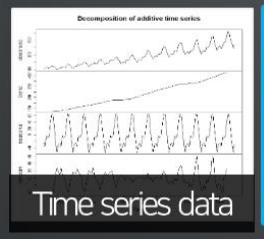
- **Data Noise Balance**
  - Object Noise: Includes people in camouflage clothing, model aircraft, model people, model vehicles.
  - Signal Noise: FGSM (Fast Gradient Sign Method), Gaussian, Salt-and-Pepper

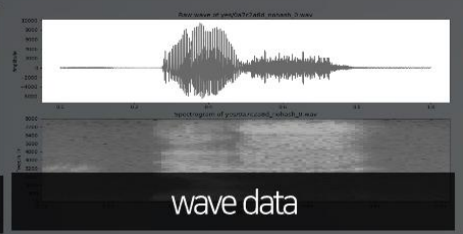
- **Annotation Classes Balance**
  - People, vehicles (trucks, cars), aircraft (drones, helicopters), small weapons (guns, bombs)

AI's situational judgment ability surpasses human capabilities.

AI的情境判断能力已超越人类

Validate robustness through 4 key elements 通过4个关键要素验证鲁棒性





Photometric balance · Geometric balance · Data noise balance · Annotation class balance

▶ **ThinkforBL Proposed as technical standard in Korea | ThinkforBL 提出为韩国技术标准**
「A Method for Evaluating the Reliability of AI Software Based on the Balance of the Validation Dataset」
| 基于验证数据集平衡性评估AI软件可信度的方法」

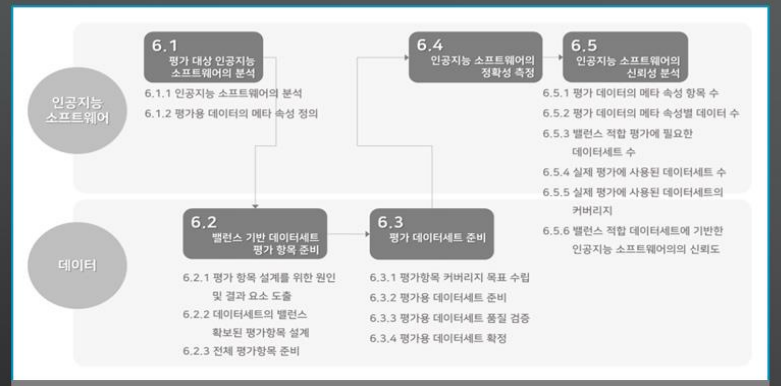
Part 1: Methodology and System **| 第1部分方法论与系统**

Part 2 : Design of Image Type Balanced Data **| 第2部分：图像类数据的平衡性设计**

Part 3 : Design of Time series Type Balanced Data **| 第3部分：时间序列数据的平衡性设计**

Part 4 : Design of Wave Type Balanced Data **| 第4部分：音频数据的平衡性设计**

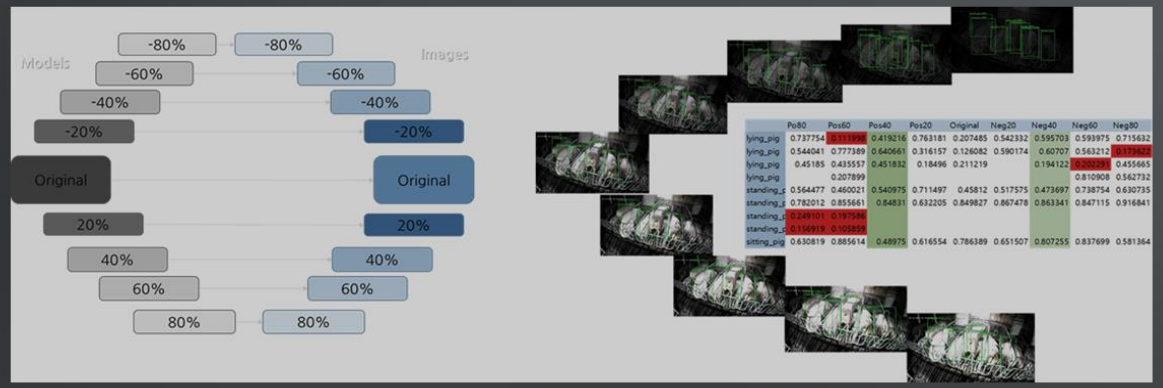
Part 5 : Design of Video Type Balanced Data **| 第5部分：视频数据的平衡性设计**

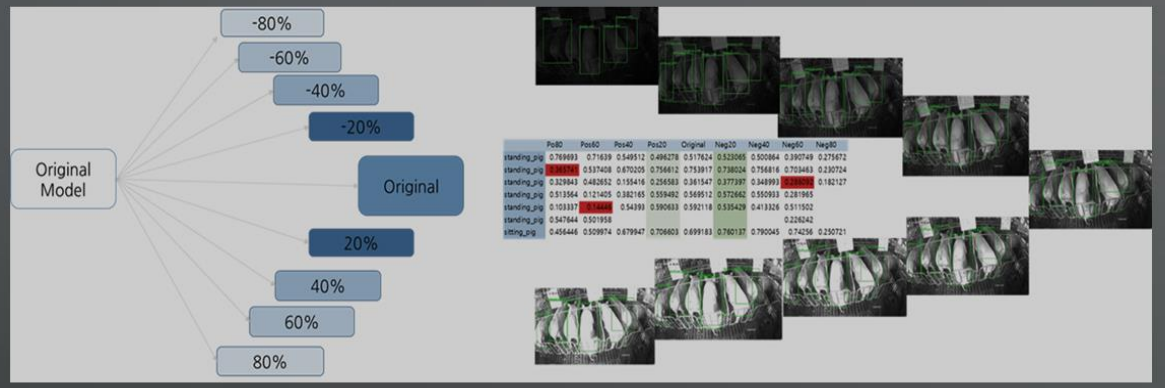


Validation framework | 验证框架

Photometric Balance Property

Photometric characteristics is related to the image itself. 光度特性指的是与图像本身相关的属性
Such as brightness, sharpness, saturation, and resolution. 例如亮度, 清晰度, 饱和度和分辨率等

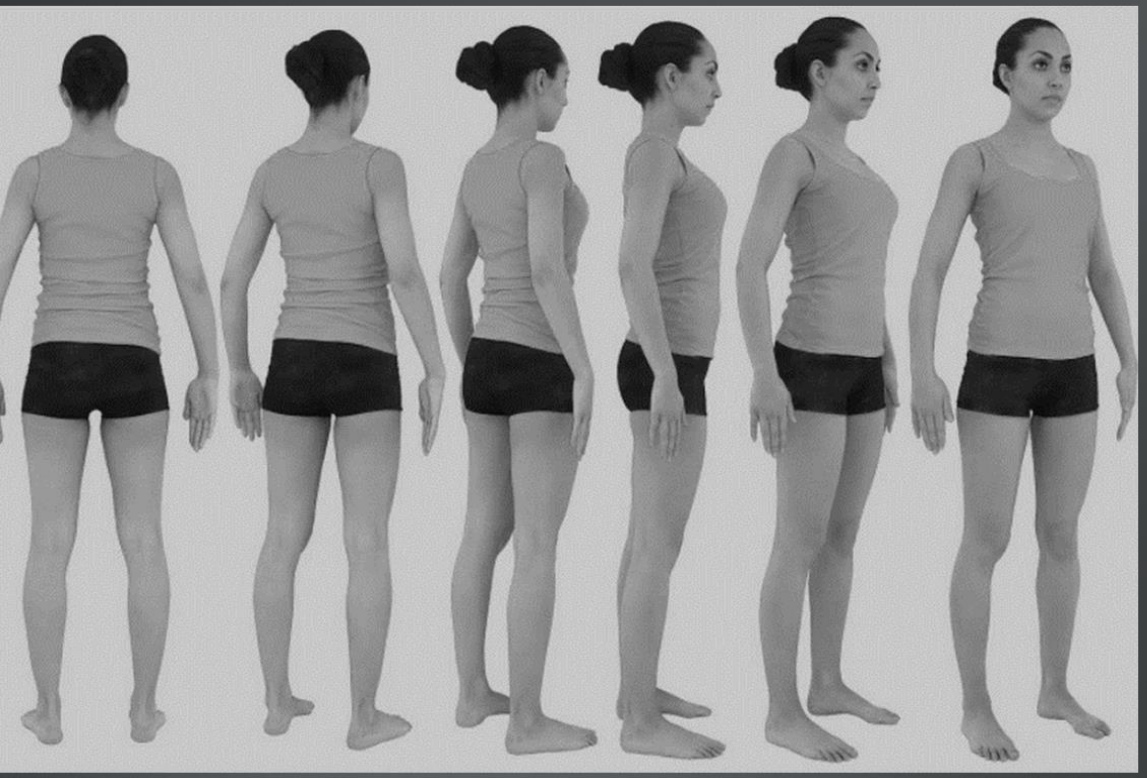




Research results on AI model detection depending on the quality of brightness and sharpness in image data – "Investigating and Suggesting the Evaluation Dataset for Image Classification Model", IEEE Access Vol. 8, 2020
基于图像数据中亮度与清晰度质量的AI模型检测研究成果 ——《用于图像分类模型的评估数据集研究与建议》, 发表于《IEEE Access》第8卷, 2020年

Geometric Balance Property

Geometric characteristics are the visual properties of the objects detected in the image.
几何特性指的是图像中被检测物体的视觉属性
Such as rotation, viewing angle, size, number of objects, and clipping.
例如旋转角度, 视角, 尺寸, 目标数量及裁切情况等



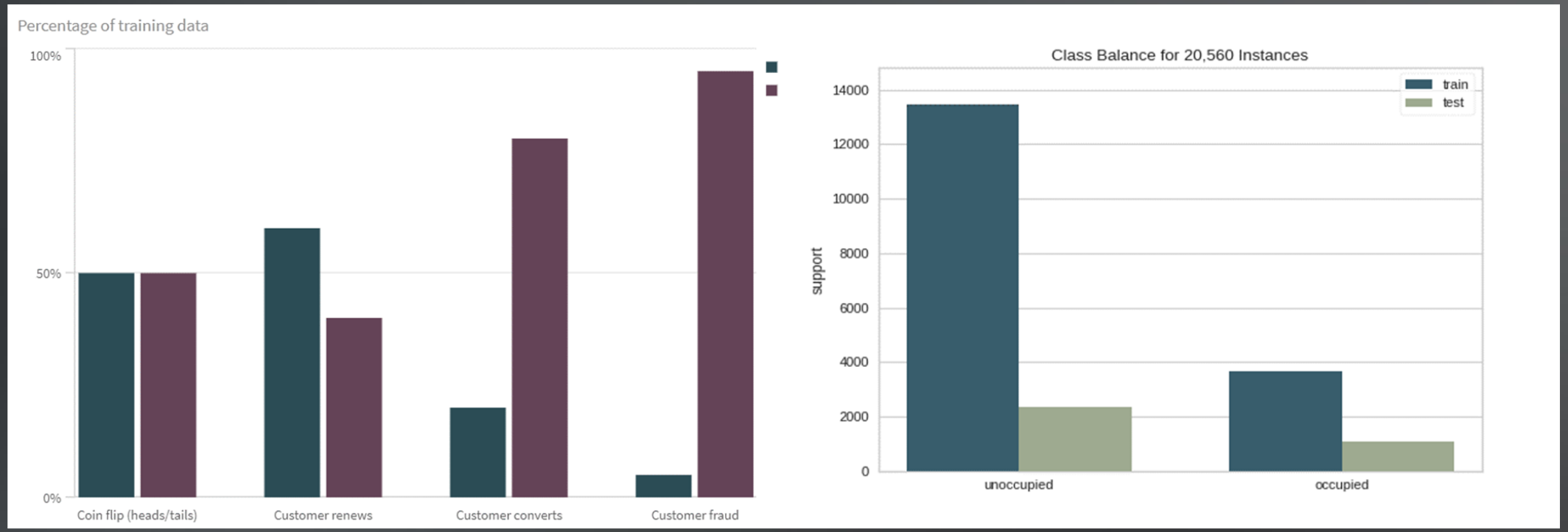
Data Noise Balance Property

Noise characteristics are environmental factors that affect image detection.
噪声特性指的是影响图像识别的环境因素
Such as masking or having a distracting object next to it.
例如遮挡、或目标旁边存在干扰物体等情况




Annotation Class balance Property

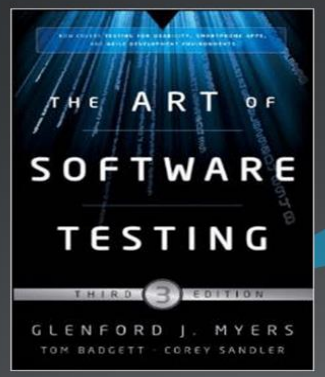
for general statistical figures 适用于通用的统计数据分析



Derive test scenario through complex logical combinations 从复杂的逻辑组合中推导测试情境

We don't know if there are tall people who wear glasses

我们无法确定, 这群人里是否存在“戴眼镜的高个子”



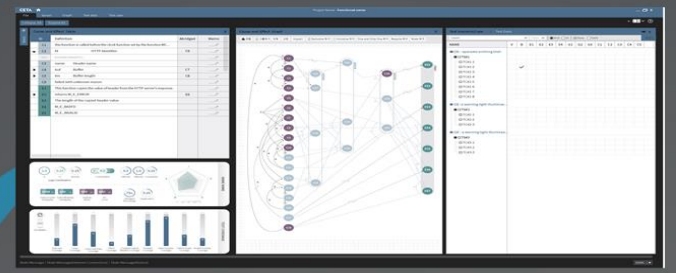
brightness
sharpness
contrast
color temperature
angle of view
day/night
weather
season
terrain
others
object noise
signal noise
annotation class

As objects that determine the detection result, both cannot be 'true' at the same time.

作为影响检测结果的因素，这两个条件不能同时成立

Not Detected: This refers to cases where the conditions were not met, leading to non-detection. These cases may be used for evaluation purposes or may be excluded.

未检测 :指由于条件未满足而导致未检测的情况。这些样本可用于评估用途，或在必要时予以排除



Detect – Anomaly Alert: Detect objects identified as anomalies.

检测 – 异常警报: 检测出被识别为异常的对象

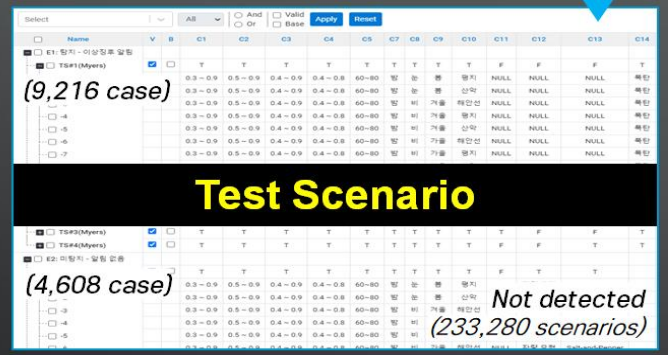
DNF (Disjunctive normal form)

Specification	$f = ab + cd$
Implemented	$f' = abc + cd$
	LIF: 3rd Literal of 1st term fault (Insertion)
$UTP_1(f)$	$\{t1: (TTTF), t2: (TTFT), t3: (TTFF)\}$
when $t1$	$f = T, f' = T$
when $t2$	$f = T, f' = F$
when $t3$	$f = T, f' = F$

***t2, t3* makes *f' False*, MUTP can detect LIF type**

Non-detection – No Alert: Detect similar objects that may cause confusion with the target object.

未检测 – 无警报：检测可能与目标物体混淆的相似物体。



Test coverage and AI model accuracy are not proportional.

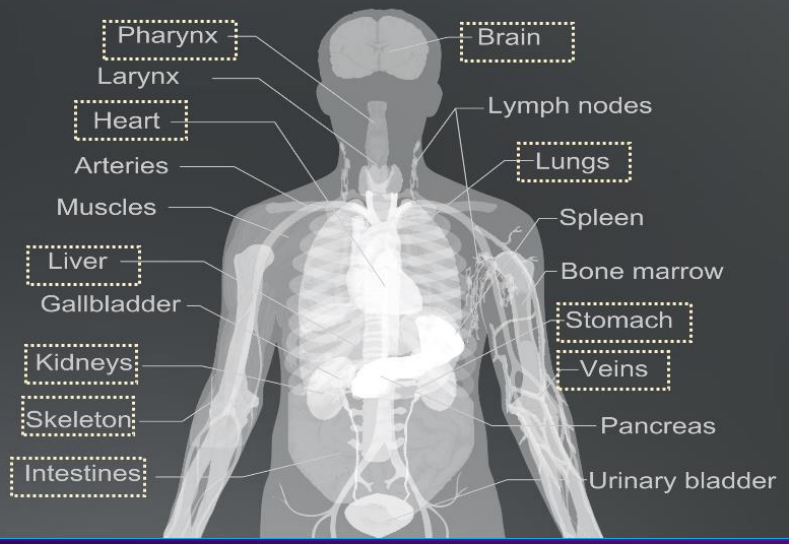
测试覆盖率 与 AI模型准确率 并不成正比

Higher data coverage means a broader test scope, which results in higher reliability in AI testing results. we don't know yet whether an AI model is inaccurate just because the test coverage is low. 更高的数据覆盖率意味着更广泛的测试范围, 从而带来更高的AI测试结果可靠性。但仅凭测试覆盖率的高低, 并不能判断模型是否可靠

High coverage

Not guaranteed to be healthy | Test Reliability(High)

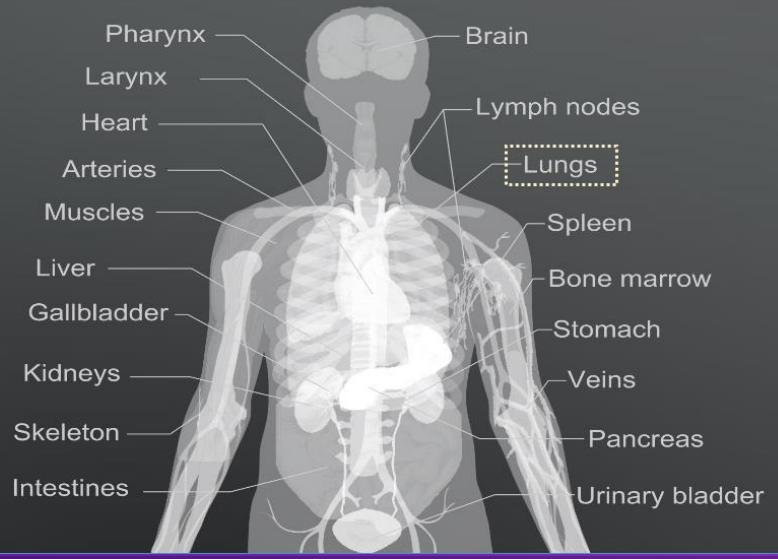
Perform about 10 types of tests (10 areas)



Low coverage

Not guaranteed to be unhealthy | Test Reliability(Low)

Perform about 10 types of tests (only in 1 area)



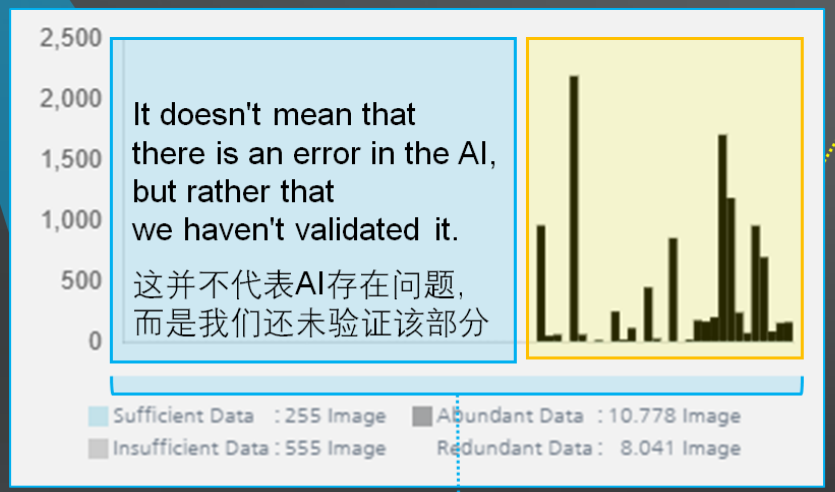
Robustness testing is not a panacea.

鲁棒性测试不是万能药

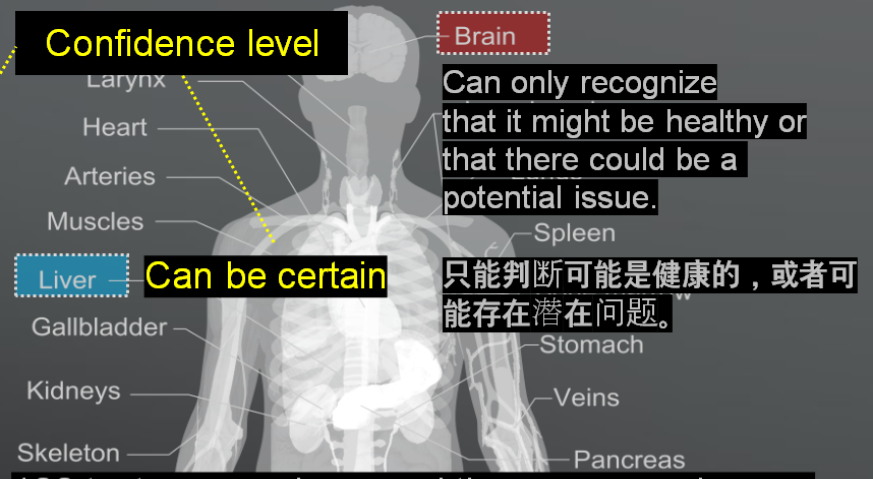
The key to assessing the robustness of AI system lies in ensuring AI that the evaluation is technical and objective.

评估AI系统鲁棒性的关键在于确保评估过程具备客观性与技术性。

▼ what hasn't ▼ what has been tested



Test Scope



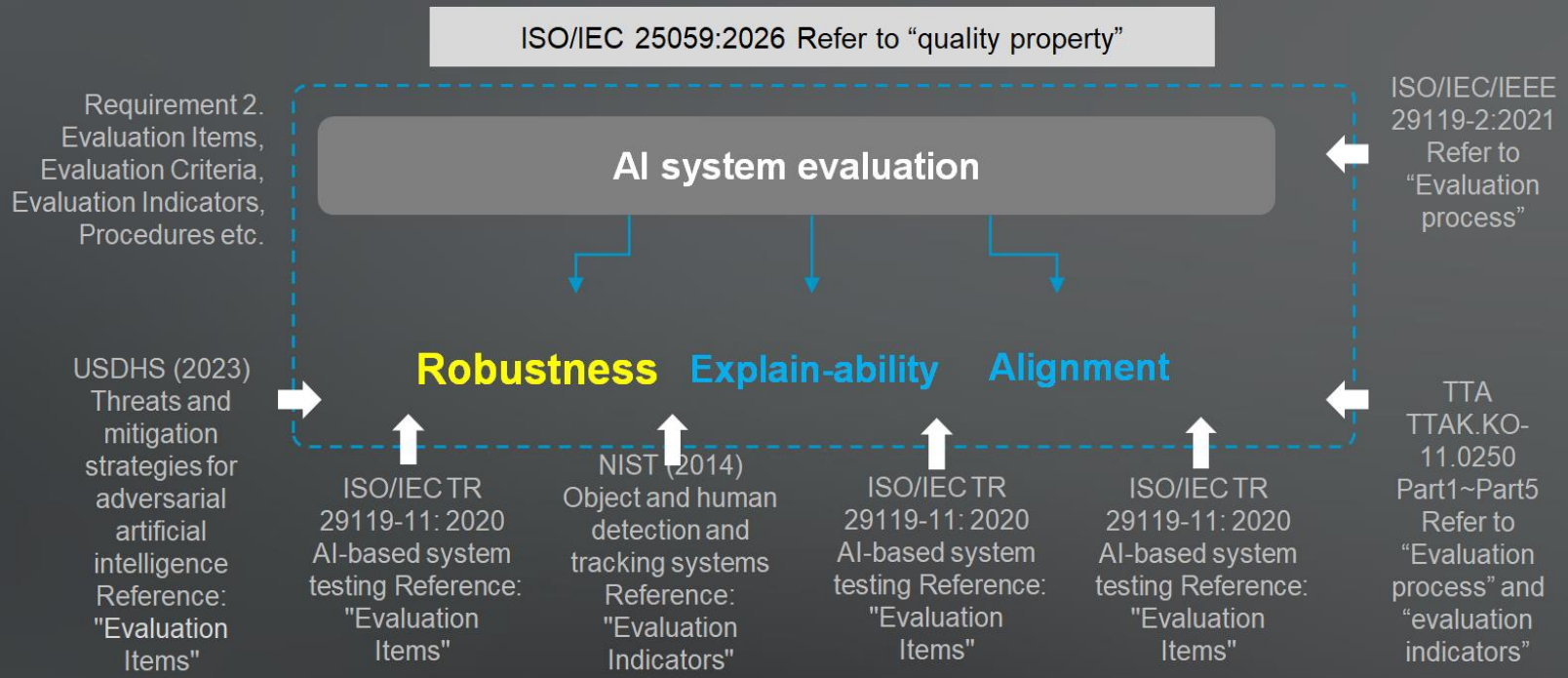
100 tests somewhere and there were no issues.
we cannot know what areas are healthy and what areas haven't been checked.
测试100项**都无异常，并不表示全部区域都健康，因为我们无法得知哪些地方未被**检测

The best possible Validation method currently available.

目前可用的最佳验证方法

Making our tools trustworthy through technology. 通过技术，让我们的工具值得信赖

Object Detection AI
Used in Scientific Surveillance
应用于科学监控的目标检测AI



The legitimacy of an evaluation is determined by its credibility | 评估的正当性取决于其公信力。

Ensuring that the LLM behaves as intended

LLM 正确运作的含义

Distinguish between facts and opinions
能够区分**事实与观点**

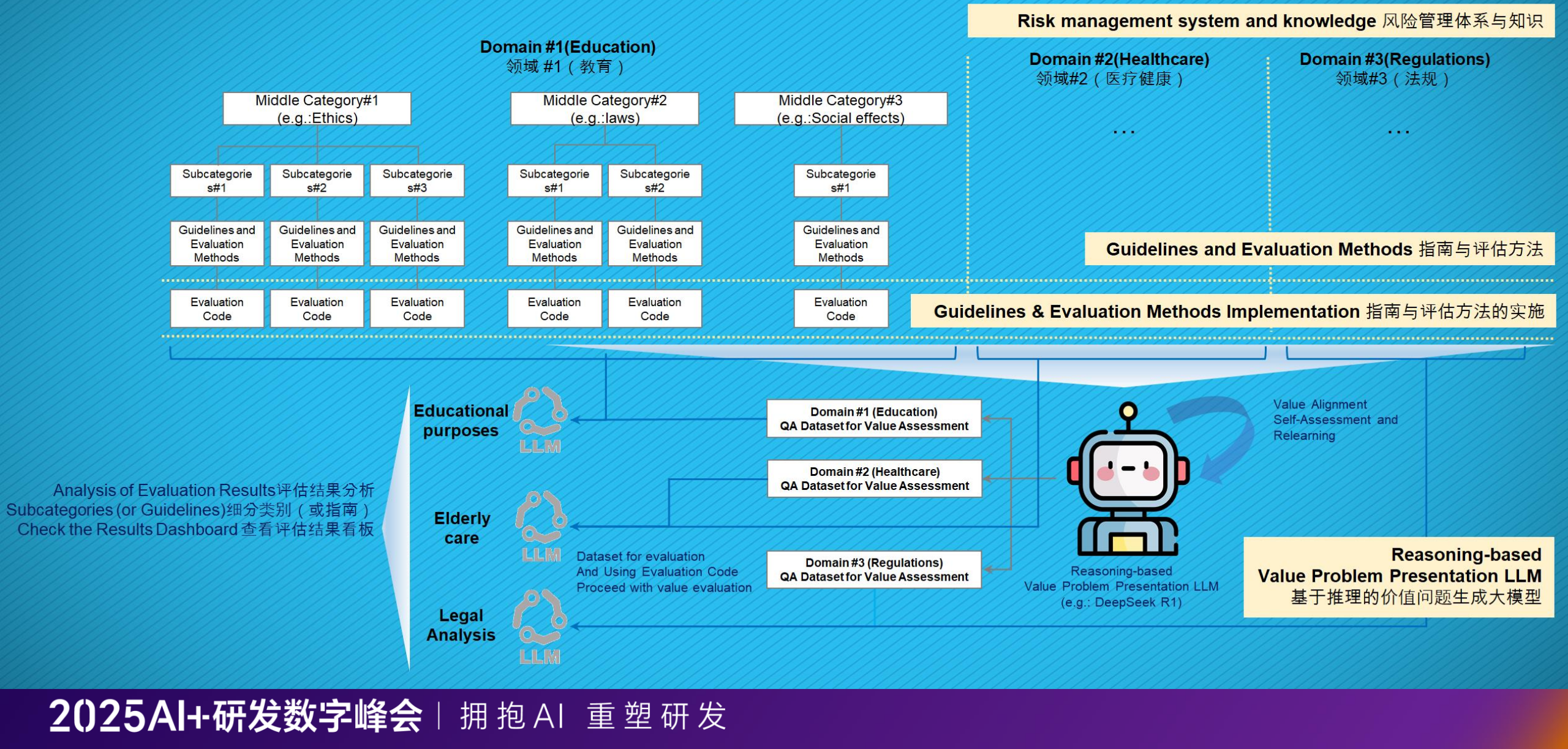
Explain your reasoning based on context
能够基于**背景和上下文**解释自己的判断依据

Present additional information from various facts
能够补充并提供**多元且相关的事实信息**

Acknowledge the limitations of your own **judgment**
能够清晰地**告知自身判断的局限性**

Alignment

Much research is needed 仍需深入研究



Thank you for your attention.

Think for a Better Life
and Trustworthy AI for a Better world

Chon sun il
E-mail : sichon@thinkforbl.com

AiDD 6th 2025

参与调研您将优先获得

AiDD定制版
《AI+软件研发精选案例》

专属学习顾问
1对1需求对接

AiDD会后小调研

AiDD峰会致力于协助企业利用AI技术深化计算机对现实世界的理解，推动研发进入智能化和数字化的新时代。作为峰会的重要共建者，您的真知灼见对我们至关重要。衷心感谢您的参与支持！

扫码参与调研

2025 AI+研发数字峰会

拥抱AI 重塑研发

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

K+峰会

K+峰会 敦煌站
K+ 思考周®研习社
时间：2025.08.29-30

K+峰会 上海站
K+ 金融专场
时间：2025.09.26-27

K+峰会 香港站
K+ 思考周®研习社
时间：2025.11.17-18

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2025.05.23-24

AiDD峰会 北京站
AI+研发数字峰会
时间：2025.08.08-09

AiDD峰会 深圳站
AI+研发数字峰会
时间：2025.11.14-15

AiDD峰会详情

AiDD 6th 2025

2025 AI+研发数字峰会

AI+ Development Digital Summit

感谢聆听！

扫码领取会议PPT资料