CAICT 中国信通院

Trustworthy AI framework and best practices

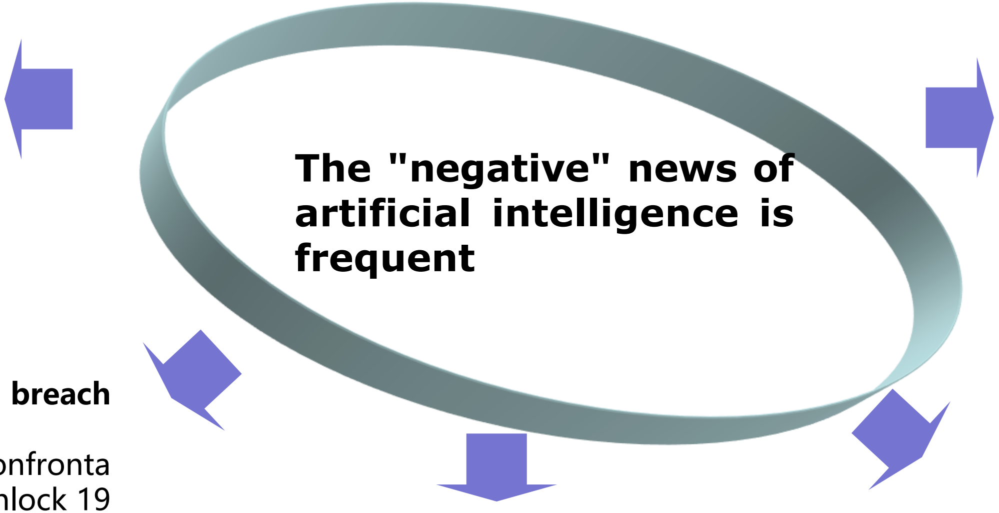
Wang Yuntao

2023.03.29

www.caict.ac.cn

AI trust issues continue to emerge

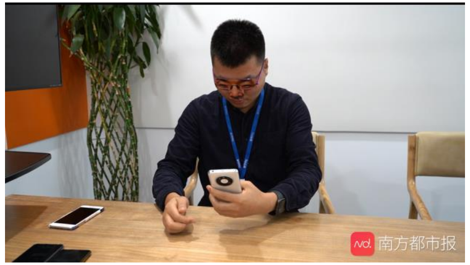
Self-driving safety concerns
According to the interface news statistics, more than **90% of** Tesla accidents are caused by "loss of control", so the self-driving assistance system is questioned.



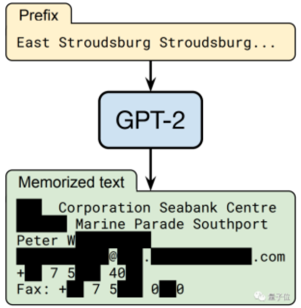
The "negative" news of artificial intelligence is frequent

Violation of the "right to know"
The "algorithmic black box" of artificial intelligence
VS
The user's right to know.

Face recognition system was breached
RealAI has reportedly used "confrontation" glasses to successfully unlock 19 cell phones.



Invasion of personal privacy happens all the time
Question and answer models, etc. reveal personal privacy.


Prefix
East Stroudsburg Stroudsburg...
GPT-2
Memorized text
[redacted] Corporation Seabank Centre
Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0

Deep forgery blurs the line between true and false

CARNEGIE MELLON RESEARCHERS ARE USING AI TO
TRANSFER FACIAL EXPRESSIONS FROM ONE VIDEO TO ANOTHER.

What Trustworthy AI looks like?

CAICT 中国信通院

International Organizations

G20 AI Principles Published to Promote Innovation in Trusted Artificial Intelligence

Building credible standards and publishing research reports

Governments

Countries have introduced AI governance principles and promoted AI legislation, etc.

Companies, etc.

MEGVII 旷视

Establishment of Governance Institute

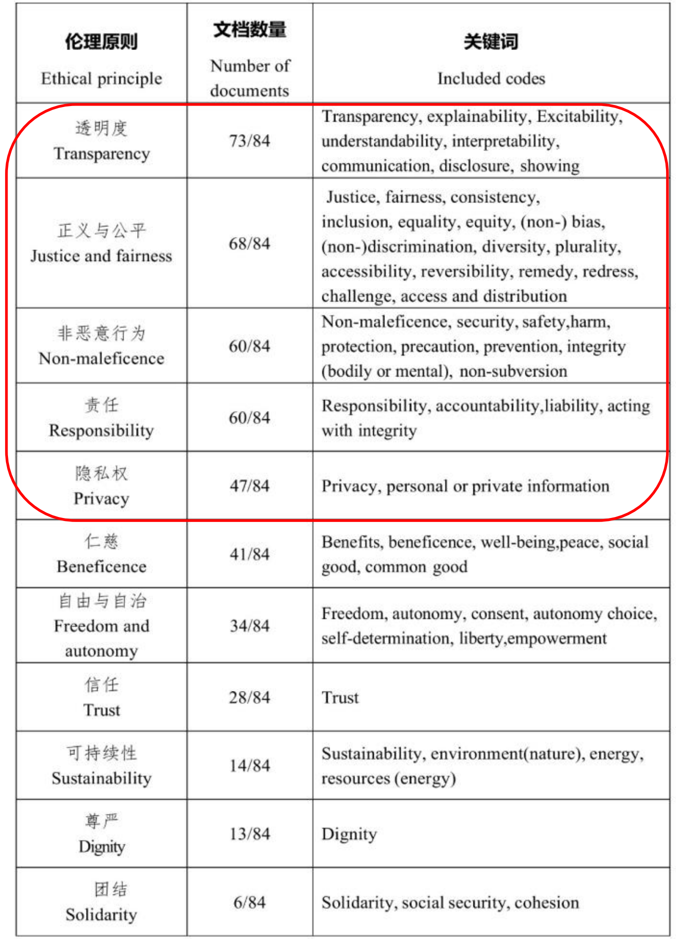
IBM

Open source governance tools, etc.

Microsoft

Practicing the concept of responsible machine learning

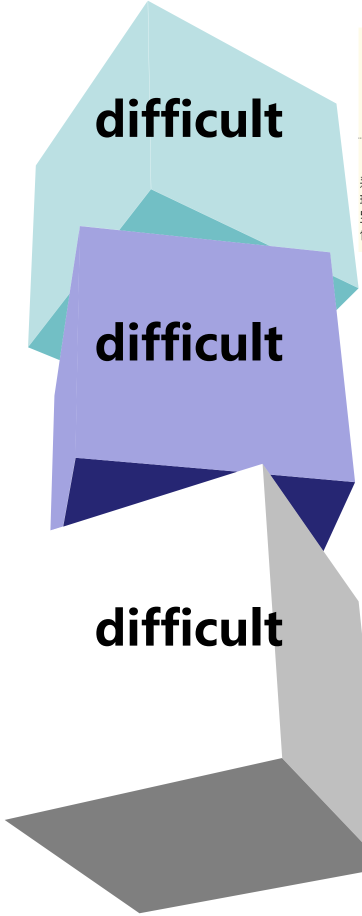
By analyzing relevant global documents, we gradually converge on key elements **such as transparency, security, fairness, accountability, and privacy protection.**

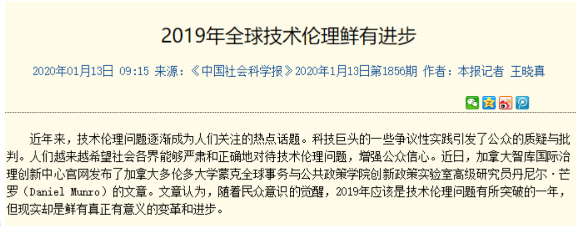


伦理原则 Ethical principle	文档数量 Number of documents	关键词 Included codes
透明度 Transparency	73/84	Transparency, explainability, Excitability, understandability, interpretability, communication, disclosure, showing
正义与公平 Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
非恶意行为 Non-maleficence	60/84	Non-maleficence, security, safety,harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
责任 Responsibility	60/84	Responsibility, accountability,liability, acting with integrity
隐私权 Privacy	47/84	Privacy, personal or private information
仁慈 Beneficence	41/84	Benefits, beneficence, well-being,peace, social good, common good
自由与自治 Freedom and autonomy	34/84	Freedom, autonomy, consent, autonomy choice, self-determination, liberty,empowerment
信任 Trust	28/84	Trust
可持续性 Sustainability	14/84	Sustainability, environment(nature), energy, resources (energy)
尊严 Dignity	13/84	Dignity
团结 Solidarity	6/84	Solidarity, social security, cohesion

The global landscape of AI ethics guidelines," a compendium of 84 documents.

"Empty talk" of Trustworthy AI is not enough, it needs to be more practical!

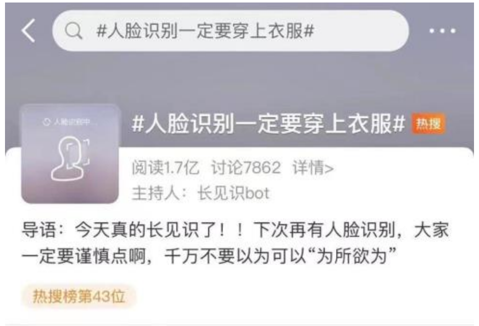




2019年全球技术伦理鲜有进步

2020年01月13日 09:15 来源：《中国社会科学报》2020年1月13日第1856期 作者：本报记者 王晓真

近年来，技术伦理问题逐渐成为人们关注的热点话题。科技巨头的一些争议性实践引发了公众的质疑与批判。人们越来越希望社会各界能够严肃和正确地对待技术伦理问题，增强公众信心。近日，加拿大智库国际治理创新中心官网发布了加拿大多伦多大学蒙克全球事务与公共政策学院创新政策实验室高级研究员丹尼尔·芒罗（Daniel Munro）的文章。文章认为，随着民众意识的觉醒，2019年应该是技术伦理问题有所突破的一年，但现实却是鲜有真正有意义的变革和进步。



#人脸识别一定要穿上衣服#

#人脸识别一定要穿上衣服# 热搜

阅读1.7亿 讨论7862 详情>

主持人：长见识bot

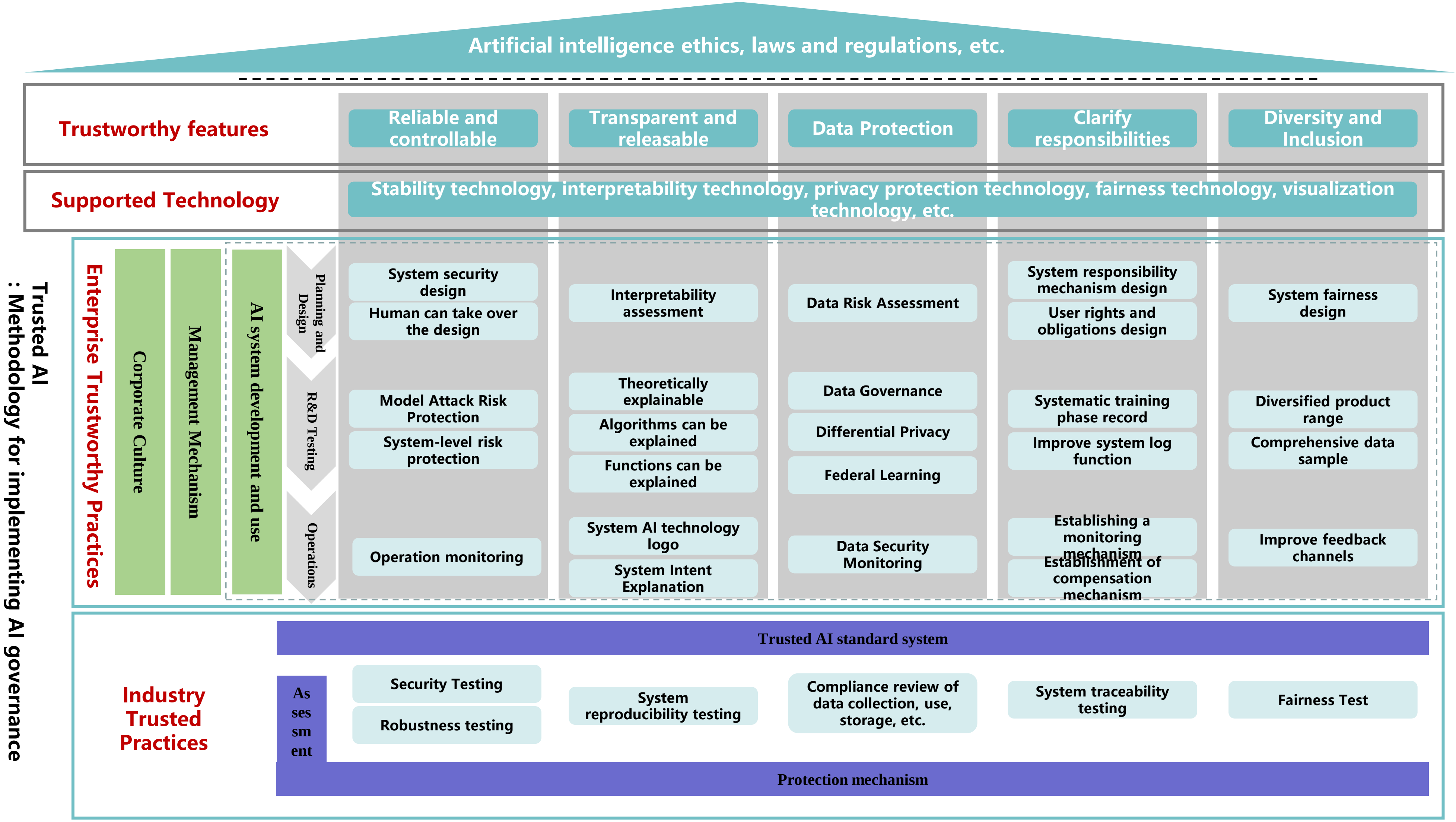
导语：今天真的长见识了！！下次再有人脸识别，大家一定要谨慎点啊，千万不要以为可以“为所欲为”

热搜榜第43位

- Some scholars believe that companies are putting the principles of artificial intelligence on the shelf and issuing ethics of technology as a means of "moral whitewashing"
- It's hard to get programmers to solve ethical problems!
- Some face recognition applications even take a picture of the whole body of end users to collect more data without even a notice, so experts warned users to put on clothes before doing face recognitions

Trustworthy AI: A Systematic Methodology for Implementing Governance Principles

CAICT 中国信通院



Artificial intelligence ethics, laws and regulations, etc.

Trusted AI: Methodology for implementing AI governance

Trustworthy features	Reliable and controllable	Transparent and releasable	Data Protection	Clarify responsibilities	Diversity and Inclusion
Supported Technology	Stability technology, interpretability technology, privacy protection technology, fairness technology, visualization technology, etc.				

Enterprise Trustworthy Practices

Corporate Culture

Management Mechanism

AI system development and use

Planning and Design	System security design Human can take over the design	Interpretability assessment	Data Risk Assessment	System responsibility mechanism design User rights and obligations design	System fairness design
R&D Testing	Model Attack Risk Protection System-level risk protection	Theoretically explainable Algorithms can be explained Functions can be explained	Data Governance Differential Privacy Federal Learning	Systematic training phase record Improve system log function	Diversified product range Comprehensive data sample
Operations	Operation monitoring	System AI technology logo System Intent Explanation	Data Security Monitoring	Establishing a monitoring mechanism Establishment of compensation mechanism	Improve feedback channels

Industry Trusted Practices

Trusted AI standard system

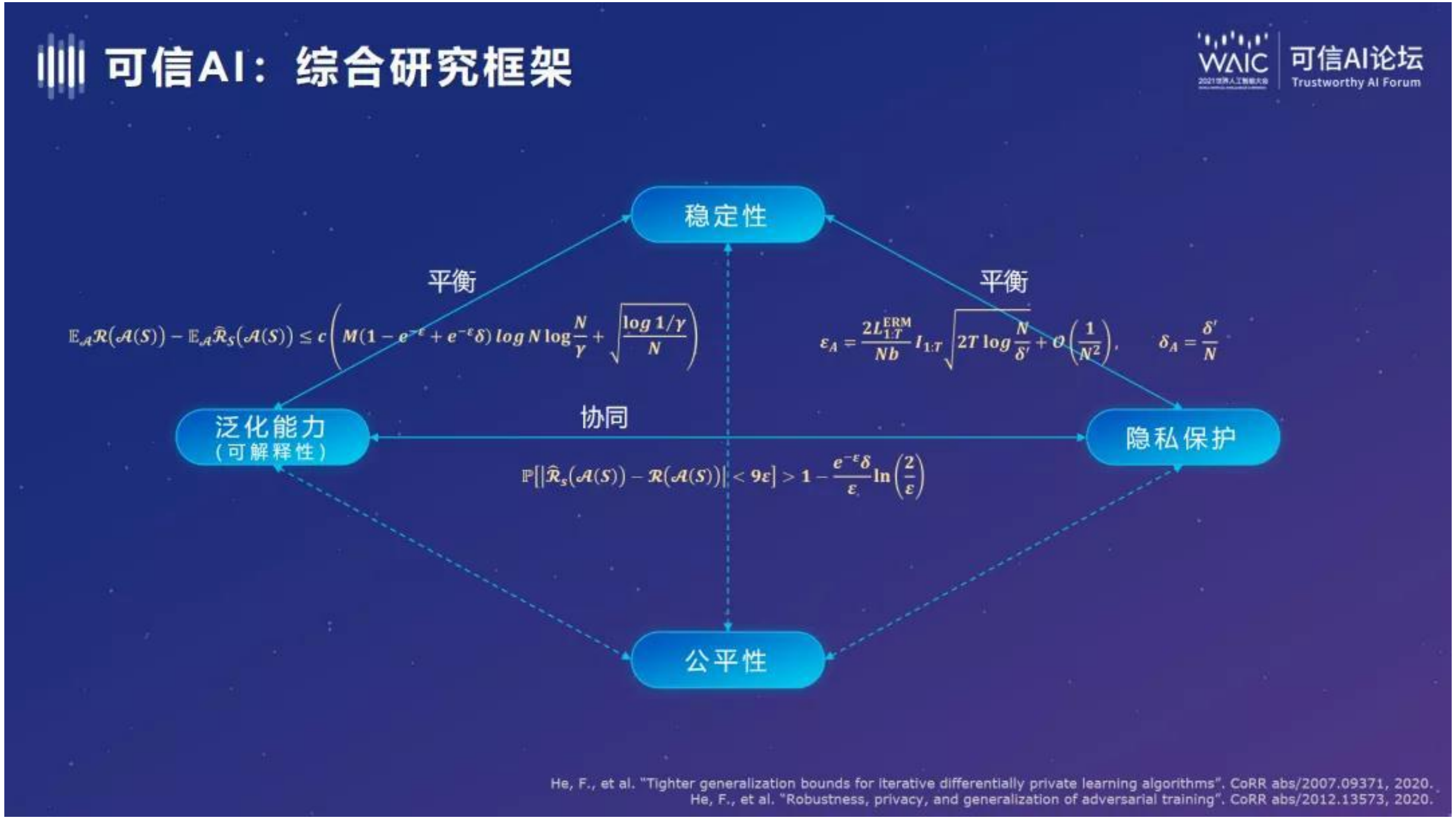
Assessment	Security Testing Robustness testing	System reproducibility testing	Compliance review of data collection, use, storage, etc.	System traceability testing	Fairness Test

Protection mechanism

Trustworthy AI supported technology: the "four rulers" for judging trustworthiness

CAICT 中国信通院

- Stability means: the ability of the AI system to resist malicious attacks.
- Interpretability means that the decisions made by an AI system need to be understandable to humans.
- Privacy protection means: the AI system cannot divulge private information of individuals or groups.
- Fairness means: AI treats all users fairly



可信AI：综合研究框架

WAIC 可信AI论坛 Trustworthy AI Forum

$$\mathbb{E}_{\mathcal{A}}\mathcal{R}(\mathcal{A}(S)) - \mathbb{E}_{\mathcal{A}}\hat{\mathcal{R}}_S(\mathcal{A}(S)) \leq c\left(M(1-e^{-\varepsilon}+e^{-\varepsilon}\delta)\, log\, N \log\frac{N}{\gamma} + \sqrt{\frac{\log 1/\gamma}{N}}\right)$$

$$\varepsilon_A = \frac{2L_{1:T}^{\mathrm{ERM}}}{Nb} l_{1:T}\sqrt{2T\log\frac{N}{\delta'}} + \mathcal{O}\left(\frac{1}{N^2}\right), \qquad \delta_A = \frac{\delta'}{N}$$

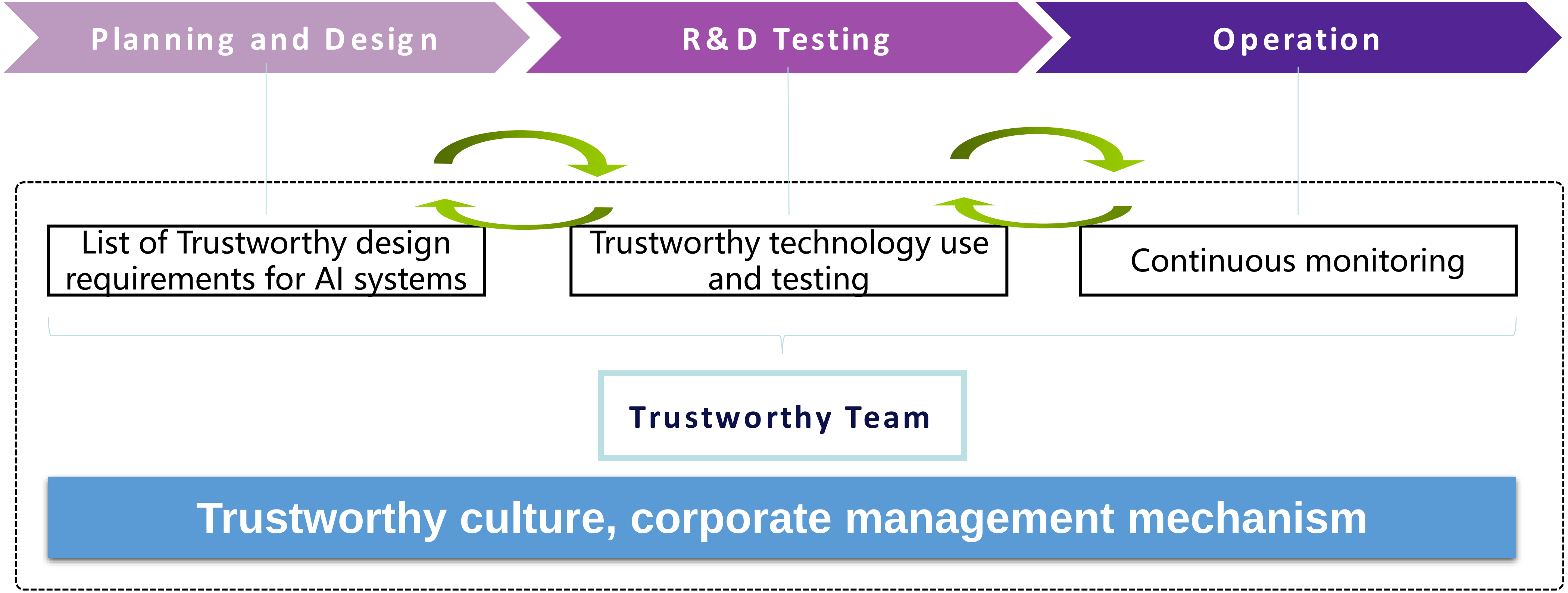
$$\mathbb{P}\left[\left|\hat{\mathcal{R}}_s(\mathcal{A}(S)) - \mathcal{R}(\mathcal{A}(S))\right| < 9\varepsilon\right] > 1 - \frac{e^{-\varepsilon}\delta}{\varepsilon}\ln\left(\frac{2}{\varepsilon}\right)$$

He, F., et al. "Tighter generalization bounds for iterative differentially private learning algorithms". CoRR abs/2007.09371, 2020.
He, F., et al. "Robustness, privacy, and generalization of adversarial training". CoRR abs/2012.13573, 2020.

Quoted from what Mr. Tao Dachen shared at the wAIC Trusted AI Forum

Integrating Trustworthy concepts into all aspects of corporate design, R&D and operations

Trustworthy full lifecycle of artificial intelligence technologies, products and services



Planning and Design

R&D Testing

Operation

List of Trustworthy design requirements for AI systems

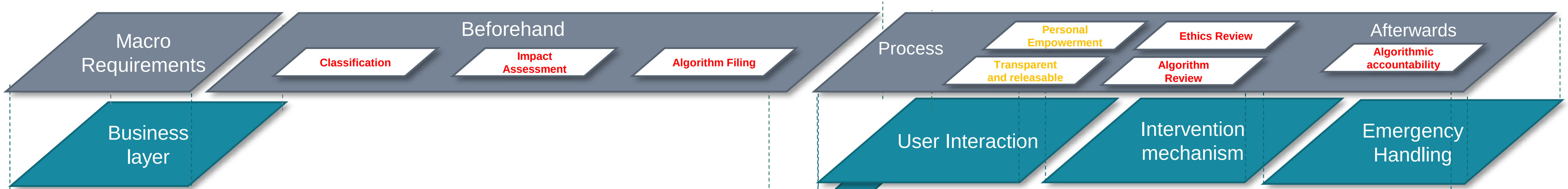
Trustworthy technology use and testing

Continuous monitoring

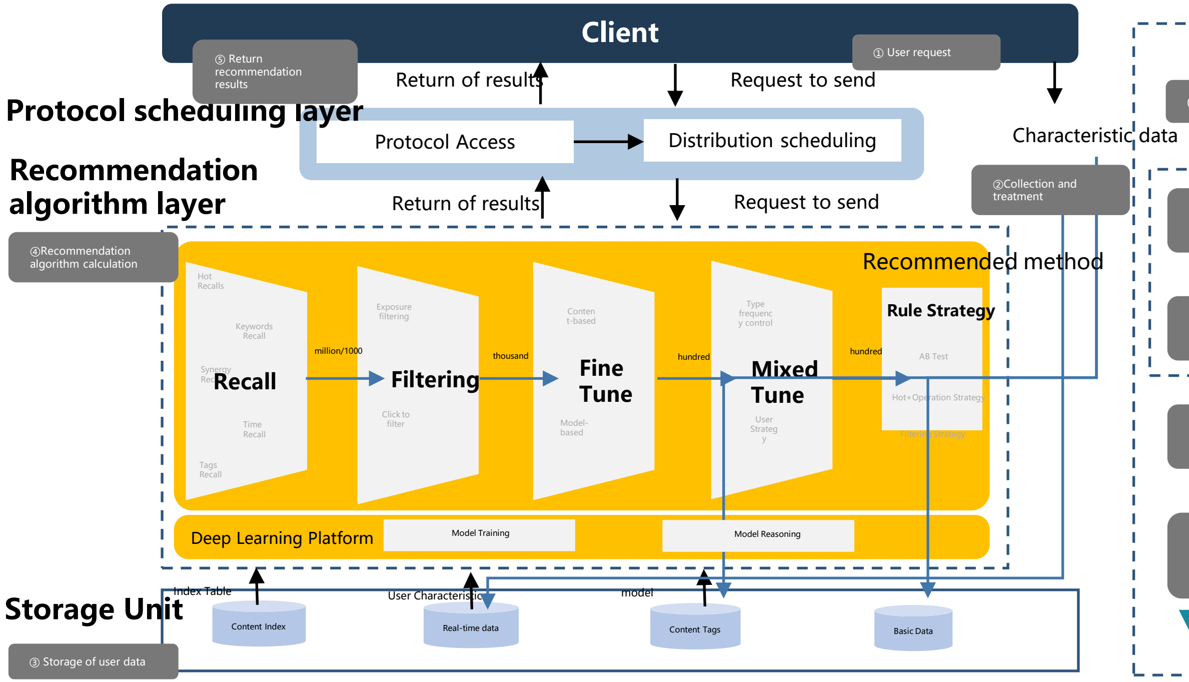
Trustworthy Team

Trustworthy culture, corporate management mechanism

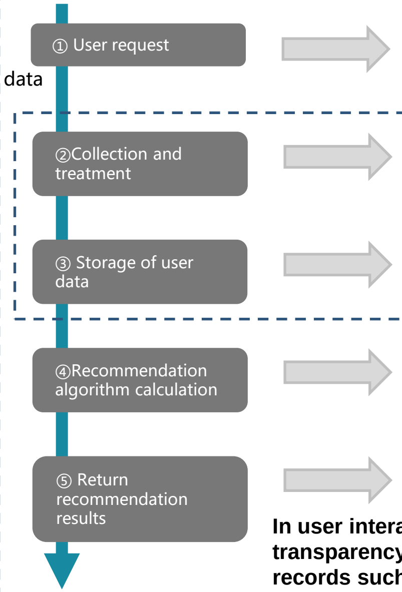
Recommendation algorithms: Building targeted solutions that break the information cocoon



1、The user interaction mechanism of the recommendation algorithm



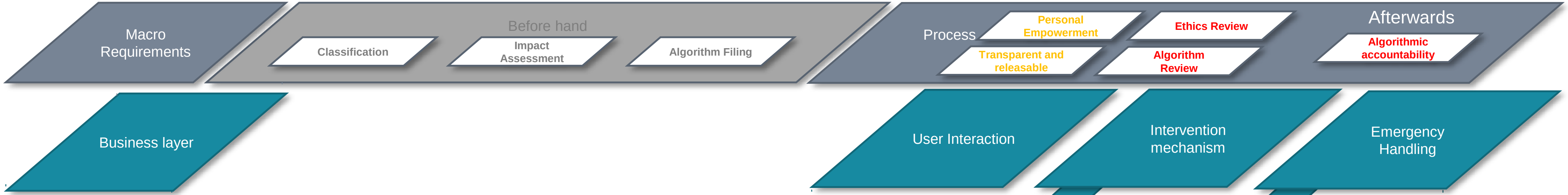
Business processes for user interaction meet governance requirements


① User request
②Collection and treatment
③ Storage of user data
④Recommendation algorithm calculation
⑤ Return recommendation results

- The user's right to know, the user's right to decide.
- Focus on data compliance review at the data layer.
- Focus on storage classification, storage encryption, storage security, etc. at the data layer.
- Implement intervention mechanisms such as de-weighting and breaking up.
- Build a complaint mechanism and feedback channels.

In user interaction, requirements such as personal empowerment, algorithm transparency and algorithm interpretation should be fully implemented. Meanwhile, records such as user behavior logs, material logs and recommendation logs should be kept to implement algorithm review requirements

Recommendation algorithms: building targeted solutions that break the information cocoon



Macro Requirements

Before hand

Classification

Impact Assessment

Algorithm Filing

Process

Personal Empowerment

Ethics Review

Transparent and releasable

Algorithm Review

Afterwards

Algorithmic accountability

Business layer

User Interaction

Intervention mechanism

Emergency Handling

2, the recommendation class algorithm application intervention mechanism, emergency treatment to break the specific initiatives of the information cocoon

Modeling user negative feedback data

Reduce user disliked content

Provide a portal for feedback on recommendation questions, model users' negative interests through their negative feedback information, and reduce the number of recommendations for content they don't like.

For the sparse negative feedback data, the recent click behavior and long-term click behavior of users are introduced to portray their positive interest, and the negative feedback behavior and recently exposed unclicked behavior of users are introduced to portray their negative interest.

Build a discovery recommendation link

Enhance the diversity of the recommendation system

Discovery-oriented quantitative recall, based on short-term user behavior, learns users' cross-category click behavior and recommends categories relevant to recent behavior for users.

Discovery search recall, combining short- and long-term user behavior with cross-category similar product indexing to build a discovery recommendation capability.

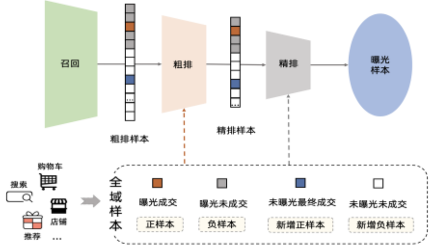
Seasonal recall, the corresponding goods are recalled before the arrival of seasonal nodes such as holidays, fruit and vegetable market time, and seasonal change.

Tag recall based on cognitive reasoning, building knowledge graphs and reasoning links based on tags, etc., to achieve the purpose of expanding user interests

Conducting full-link unbiased learning that

Portraying users' diverse interest distribution

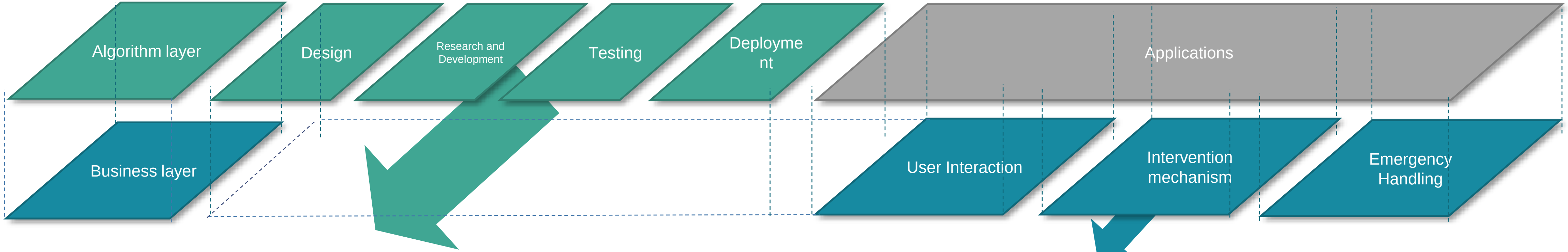
We make full use of the **funnel-type structure of the recommendation system** and the data of **multiple scenarios** to solve the problem of data selection bias and data sparsity encountered in single-scene single-task modeling



Reference source: Corporate research

Deep synthesis algorithm: at the algorithm level, establish dual-level multi-faceted protection

CAICT 中国信通院



1、Algorithm provider: build explicit and implicit technical capabilities at two levels

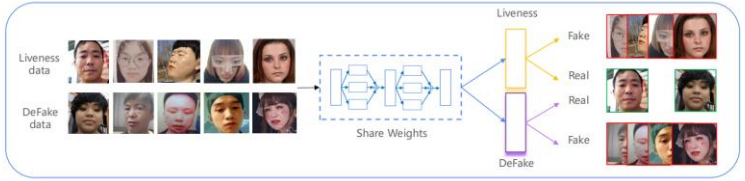
- Given that deep synthesis technology has reached a very realistic effect, it is difficult to identify by the naked eye without a priori information. Therefore, companies should intervene beforehand and manage at the source. **Adding logos** at **the explicit level** helps users to identify the authenticity of information**, and embedding watermarks at the implicit level** helps regulators to trace the source afterwards.

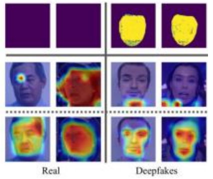
➢ **Embedded clear watermark logo**

- Clearly inform the user that the content belongs to the category of synthetic data.

➢ **Embedded digital watermark**

- Using reversible synthesis technology to guarantee traceability.



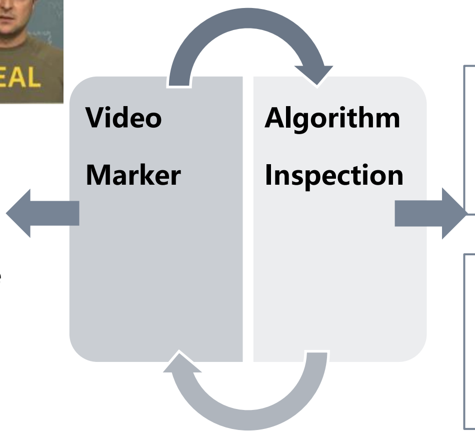


2, algorithm application side: implementation of algorithm mark to protect the user's right to know

- As the application side of the algorithm, the relevant platform should **clearly mark** the synthetic video to protect the user's right to know. In addition, potential synthetic videos should be further identified through technologies such as **digital watermarking, confrontation samples, and multimodal recognition.**

➢ **Significantly mark synthetic information**

- Protecting the user's right to know



➢ **Information Compliance**

- Securing content

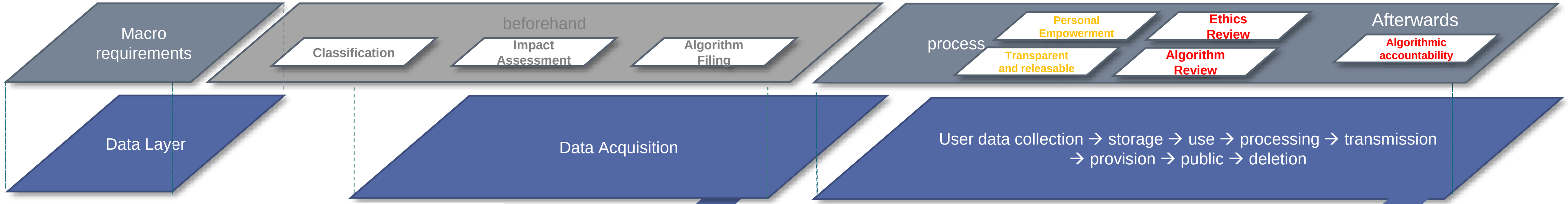
➢ **Synthetic Data Management**

- Image retrieval capability
- Digital watermark traceability

➢ **Forgery recognition capability**

- Single-mode recognition capability
- Multimodal recognition capability

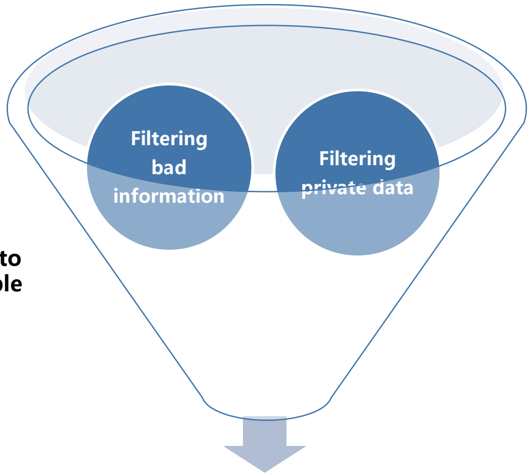
Deep synthesis algorithm: at the data level, forming continuous supervision



1、Training for data before going online

- Before the application goes live, the material data is screened in two ways, **filtering undesirable information on the one hand and private data on the other**.

➢ Filtering of undesirable information content and filtering of training data to **prevent the ability to generate undesirable information**.



➢ Filter private data data and **filter** data **with biometric features such as faces and pedestrians.**

Securing material data

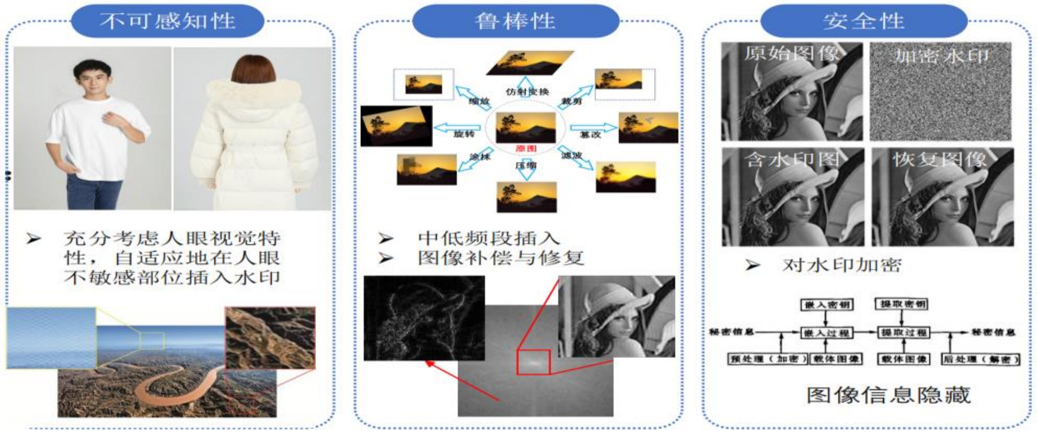
2, real-time monitoring of content security after the launch

- After the application goes live, the data is continuously tracked and **inspected in real time to ensure content security**.

➢ Information Protection

- Platform data cannot be tampered with: digital watermarking technology
- Platform data can be traced: counter-sample technology

➢ **Digital Watermarking Technology**



➢ 充分考虑人眼视觉特性，自适应地在人眼不敏感部位插入水印

➢ 中低频段插入

➢ 图像补偿与修复

➢ 对水印加密

图像信息隐藏

CAICT 中国信通院

Thanks!

Wang Yuntao, 18611547086,

wangyuntao@caict.ac.cn