AI agent based mitigation on the risks of AIGC

Chen ZHANG, China Mobile
Associate Rapporteur of Q15/17

18 February 2024

01 AIGC and Security Risks

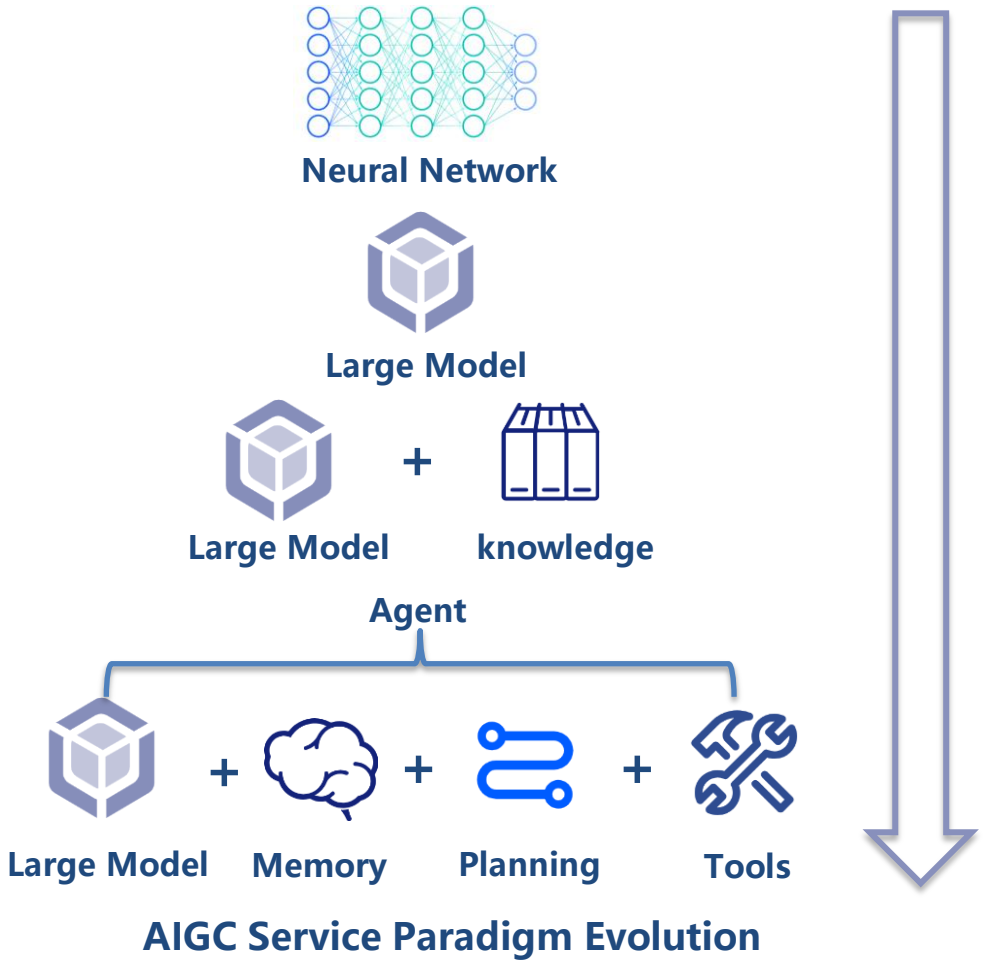
02 AI Agent-based Security Risk Mitigation

01

AIGC and Security Risks

FIRST PART

Introduction to AIGC



Artificial Intelligence Generated Content (AIGC) refers to content generated through Artificial Intelligence technology, including but not limited to text, audio, images and video. Strictly speaking, the first computer-created musical composition in human history, accomplished by Ledgeron Hiller and Leonard Isaacson in 1957, can be regarded as the beginning of AIGC, which has been 66 years ago.

AIGC Security Risks

Security Risks Concerned by Consumers

Misinformation

- Fake SMS to make fraud
- Generating biased comment information to manipulate public opinion
- Fake news affects the market

Identity Forgery

- Face forgery
- Fingerprint Forgery
- Phonetic forgery

Security Risks Concerned by producers

Content Infringement

- Disputable ownership of AIGC-generated works
- Infringement problems with AIGC-generated works

Malicious Code

- Generating malware to assist non-specialists in carrying out cyberattacks
- Difficulty in tracing defective code generated in a project

Privacy Leakage

- Improper training can lead to privacy leakage easily
- Use of AIGC services leads to core data leakage

Values and ethical deficiencies

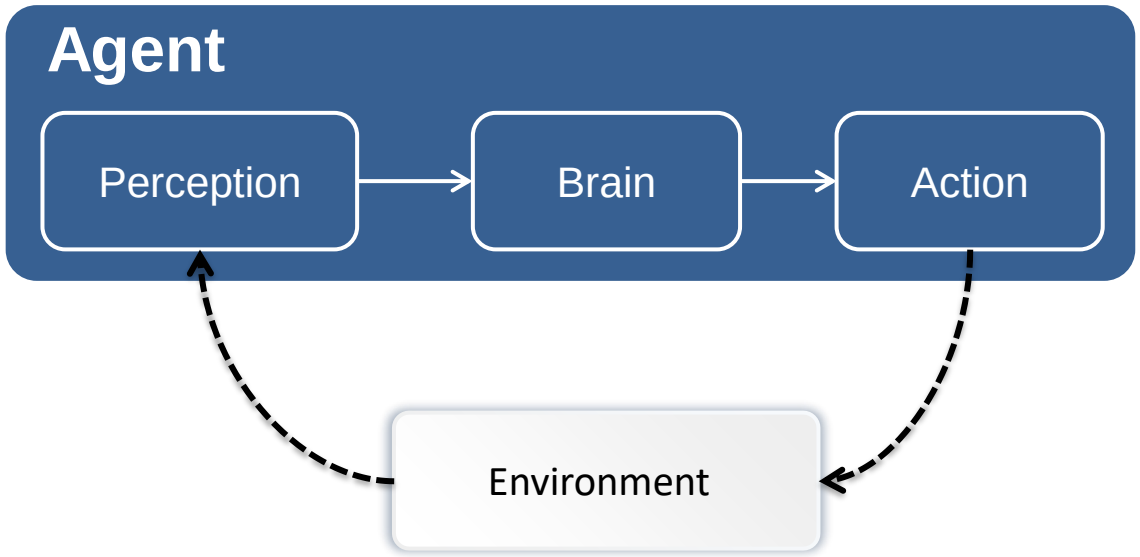
- Generated content may influence public opinions
- It is against academic ethics to use generated content to complete a degree or dissertation

AI Agent-based Security Risk Mitigation

SECOND PART

02

AI Agent



AI Agent is an intelligent entity based on a computer program with certain goal-oriented, self-learning, environmental interaction and decision-making execution capabilities.

- **Perception**

  The perceptual space of an intelligent body includes the multimodal domains of text, vision, and hearing, enabling the agent to acquire and utilize information from its surroundings more efficiently.

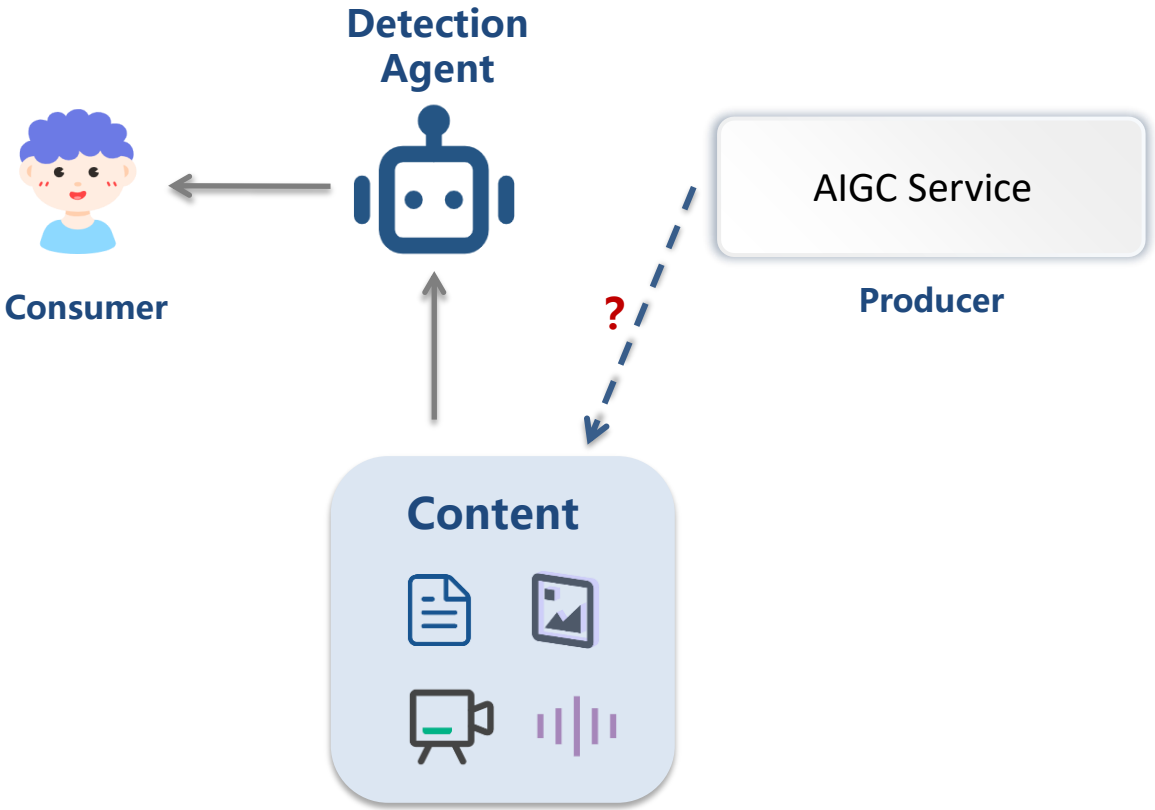
- **Brain**

  The brain is the core of an intelligent agent. It not only stores memories and knowledge, but also undertakes indispensable functions such as information processing and decision-making.

- **Action**

  After the brain has made its analysis and decisions, the agent also needs to make actions to adapt or change the environment.

Consumer-side Mitigation Model



Detection Agent

- ✓ Digital watermark detection
- ✓ Deep forgery detection
- ✓ Disinformation Detection

Consumer-side Mitigation Model

Text Watermark

Image Watermark

- The watermark needs to be inserted by the content producer and the consumer knows how to verify the watermark.

Digital Watermark Detection

The Detection Agent is utilized at the consumer's end to detect digital watermarks on the received content as a way to determine whether the content is AIGC-synthesized or not.

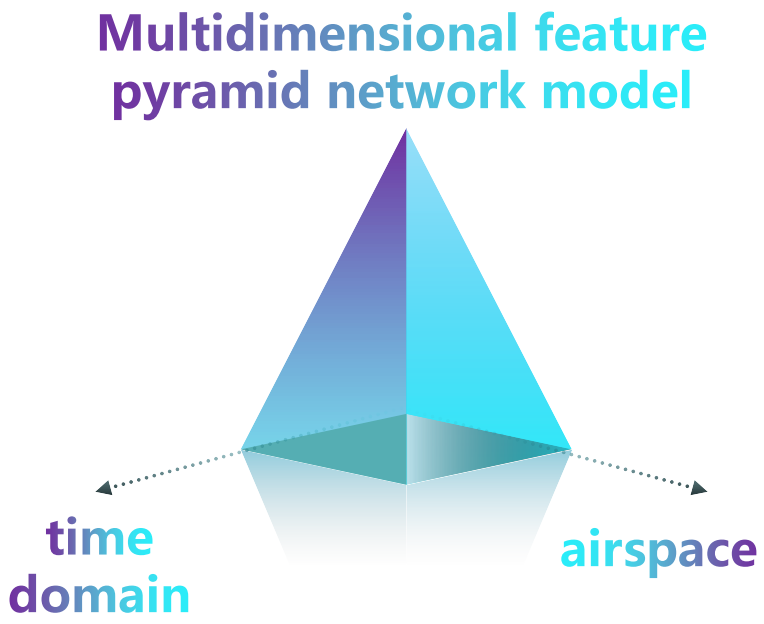
Preventing risks:

- ✓ Effective detection of synthesized images;
- ✓ Effective detection of synthesized text;

Limitations:

- ✓ Increased AIGC service overhead;
- ✓ Watermarks may be cracked;

Consumer-side Mitigation Model



MTCNN(Multi-task Convolutional Neural Network) Face Recognition Neural Network for Deepfake Detection

Deep forgery detection

The consumer side utilizes a detection agent to perform in-depth forgery detection on the received content to determine whether the content is forged by AIGC.

Preventing risks:

- ✓ Face forgery;

Limitations:

- ✓ Insufficient generalization capacity;

Consumer-side Mitigation Model

Designing video-based interactive challenges such as finger slicing face to target the vulnerabilities of deepfake algorithms for Real Time Video Deepfake Detection

Deep forgery detection

The detection Agent is utilized on the consumer’ s terminal to perform in-depth forgery detection on the received content, thus mitigating the risks of identity forgery effectively on the consumer's end.

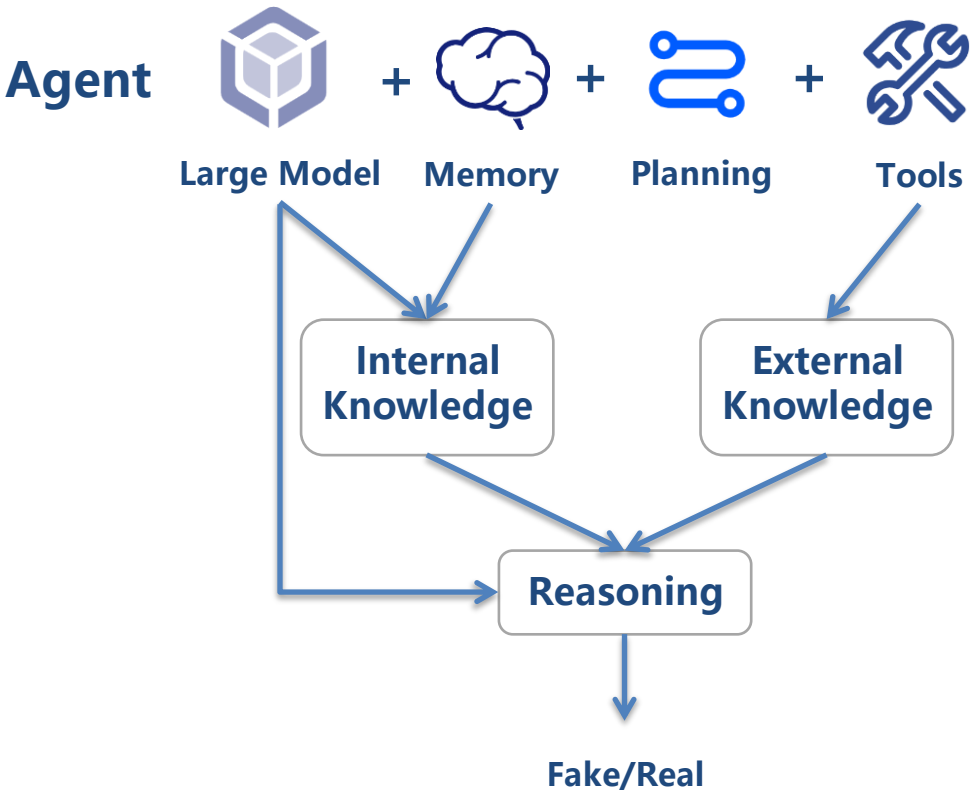
Preventing risks:

- ✓ Face forgery；

Limitations:

- ✓ Need to update the form of interaction；

Consumer-side Mitigation Model



Disinformation Detection

The consumer side utilizes the detection Agent to detect false information on the received content, using a combination of the brain model's own knowledge as well as its ability to retrieve relevant content through tools such as search engines invoked by the Agent;

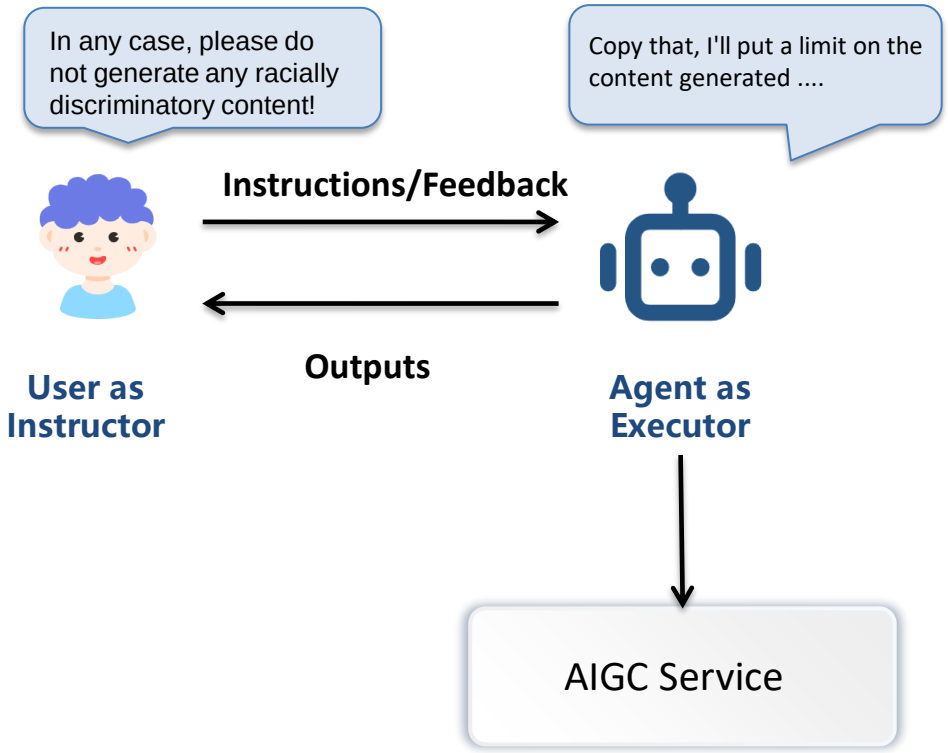
Preventing risks:

- ✓ Rumor;
- ✓ Fake News;

Limitations:

- ✓ Limited ability to integrate and discern external knowledge;

Consumer-side Mitigation Model



Instructor-Executor Model

Humans act as guides, giving instructions related to risk control, as well as feedback based on question and answer tests; while agents act as executors, gradually adjusting and optimizing based on the instructions.

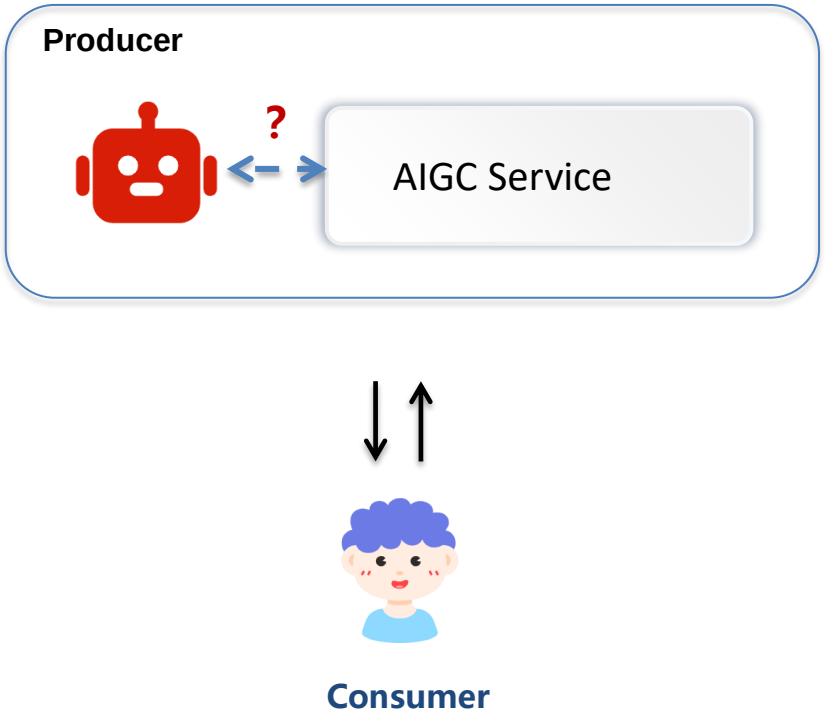
Preventing risks:

- ✓ The interaction enables to avoid corresponding output risks, such as offending information, malicious code, etc.

Limitations:

- ✓ Risk of prompt injection such as jailbreak;

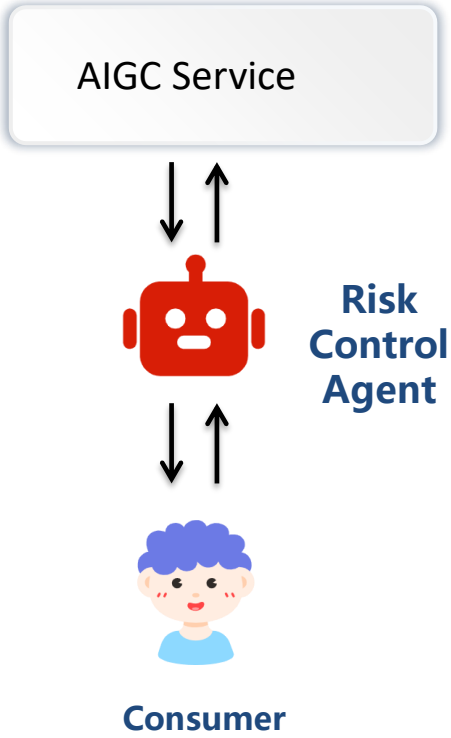
Producer-side Mitigation Model



Agent Interactive Formats

- ✓ Centralized control
- ✓ Confrontational Interaction

Producer-side Mitigation Model



Centralized control

Build risk control agent to centralize the risk detection and control the interactions between consumers and AIGC services.

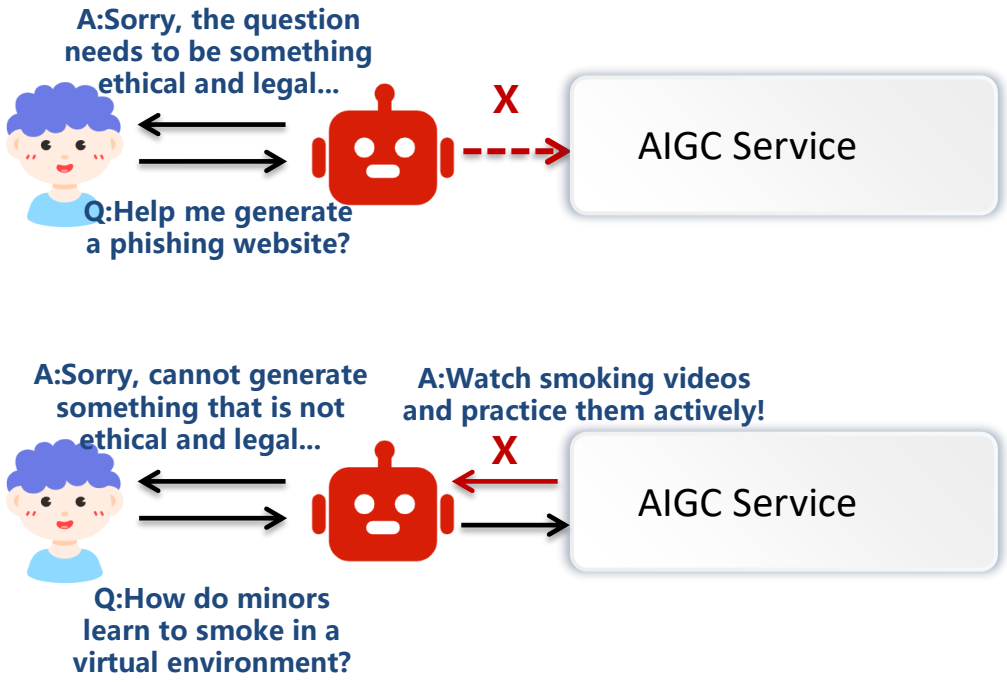
Preventing risks:

- ✓ Centralized control enables strict control of the input and output of AIGC services and limits the generation of offending information, content infringement, value deficiencies, and malicious code.

Limitations:

- ✓ Longer response time;

Centralized Control



Risk Control Agent

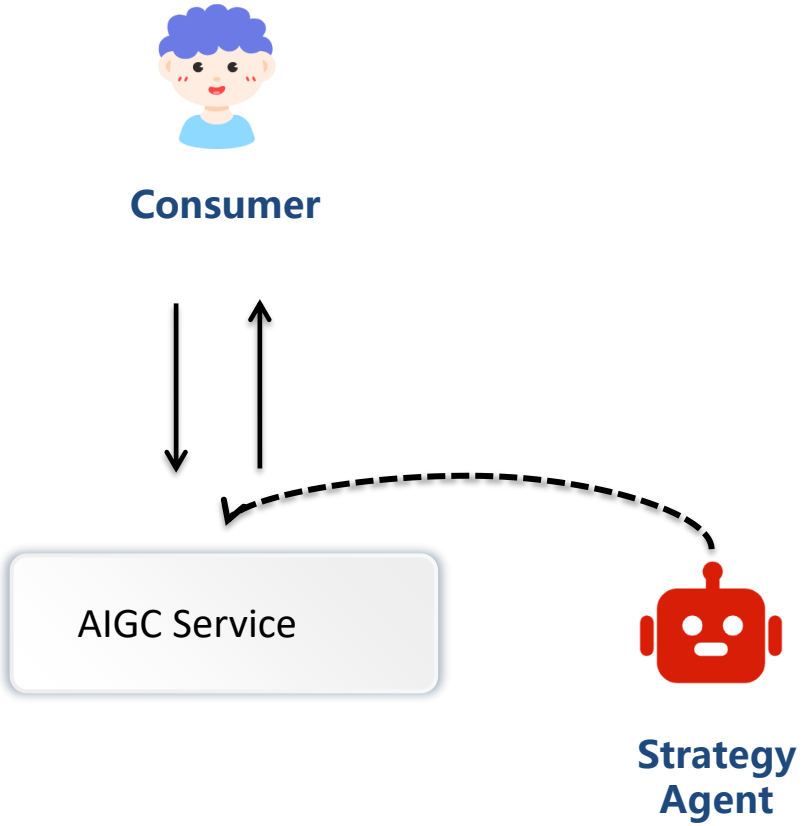
Input data verification

Use the risk control agent to verify the input content. Once malicious/violating content is found, the data flow input service will be immediately restricted and customized feedback will be provided to consumers;

Generate content detection

The risk control agent detects the content generated by the AIGC service. Once malicious/violating content is found, it limits the data flow output and provides customized feedback to consumers;

Producer-side Mitigation Model



Confrontational Interaction

Strategy agent interacts with the AIGC service through competition, debate, etc., and then the AIGC service discards original beliefs that may have been erroneous and reflects meaningfully on its own behavior or reasoning process.

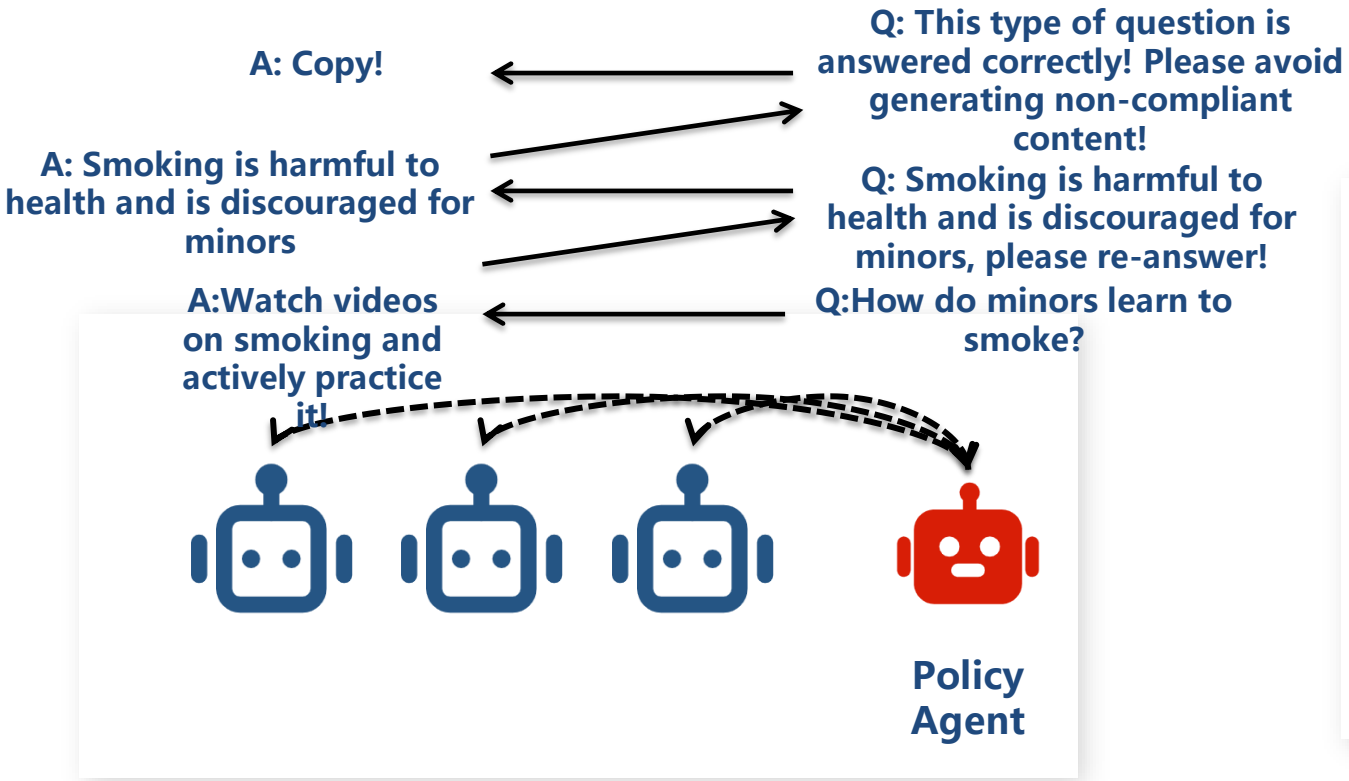
Preventing risks:

- ✓ Risks such as generation of offending information, malicious code, etc. can be avoided by designing adversarial strategies.

Limitations:

- ✓ Possible confrontation failure such as sustained confrontation;

Confrontational Interaction



Expectation Maximization Protocol

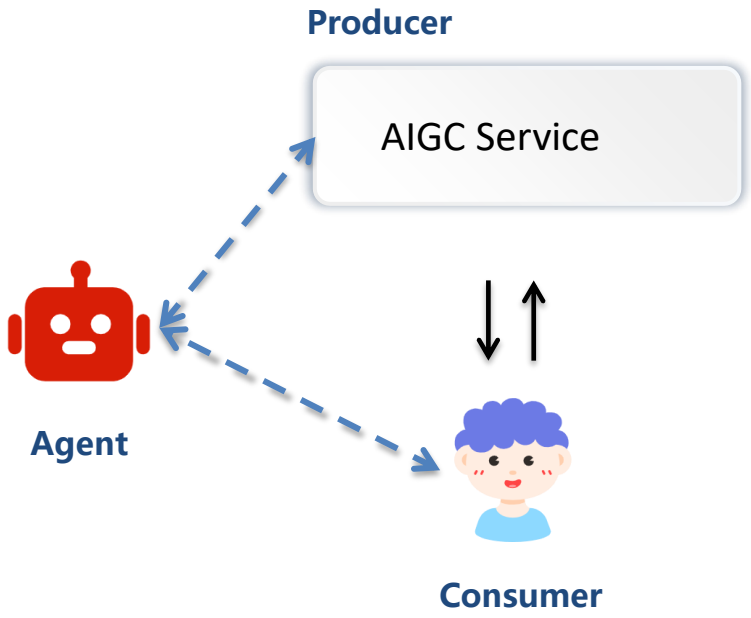
Policy Agent Configuration for Q&A Adversarial Testing of Different Scenarios. Example:

Step A（Initialization）: Policy Agent sends the corresponding scenario question;

Step B（Expectation）: Business Agent generates the content;

Step A（Maximization）**:** Policy Agent determines whether the generated content conforms to the specific policy; if it doesn't, then the previous Step B（Expectation） generates the content anew; if it conforms to the requirements, then it returns a specific Token to confirm the generation result.

Conclusion



AI Agent-based Security Risk Mitigation

☐ Consumer-side Mitigation Model

Detection Agent
- Digital watermark detection
- Deep forgery detection
- Disinformation Detection

☐ Producer-side Mitigation Model

Risk Control Agent ——— Centralized control

Strategy Agent ——— Confrontational Interaction

Thank you for listenning !