Tencent AI and ethics

Tencent Research Institute
April 2020

AI in ALL

Tencent AI lab (Shenzhen & Seattle)

70 world-class AI research scientists and 300 engineers. computer vision, speech recognition, machine learning, and natural language processing.

腾讯优图

Youtu Lab

Image understanding, face recognition, audio recognition, OCR recognition.

WeChat AI

Using Speech Recognition, Natural Language Processing technologies in Social scenarios

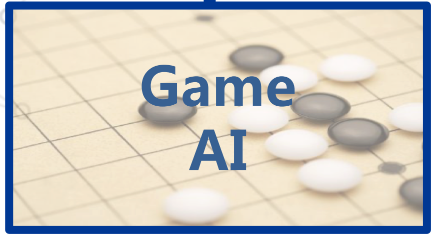
RoboticsX lab

embodied robots

Medical AI lab

medical image recognition, AI assisted diagnosis

AI Open Platform (Tencent Cloud, Tencent Opensource)



besides AI, Tencent has also established research labs in 5G, edge computing, quantum computing, IoT, and audiovisual technologies, to foster the next generation technologies

Currently AI is not capable of solving ”difficult, complex” problems

What can AI do?

Under restricted conditions , sensing and analytics

- Voice recognition
- Big data computing
- Machine translation
- …

Next Step

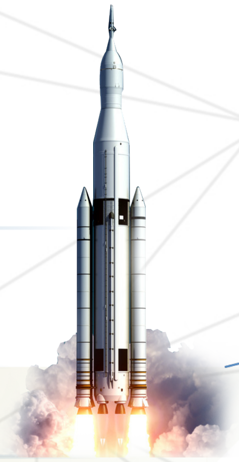
Solving complex problems in real environment with undefined conditions

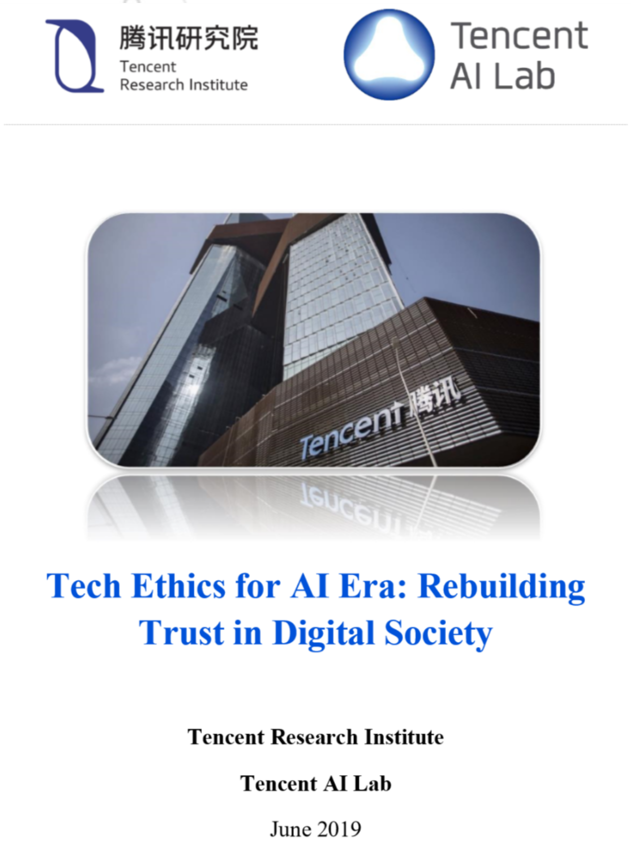
- Sensing emotion
- Smart Interface
- Reflection and Creation
- **multi-tasks and AGI**

Objectives

AI Backed

Industry Upgrade





腾讯研究院
Tencent
Research Institute

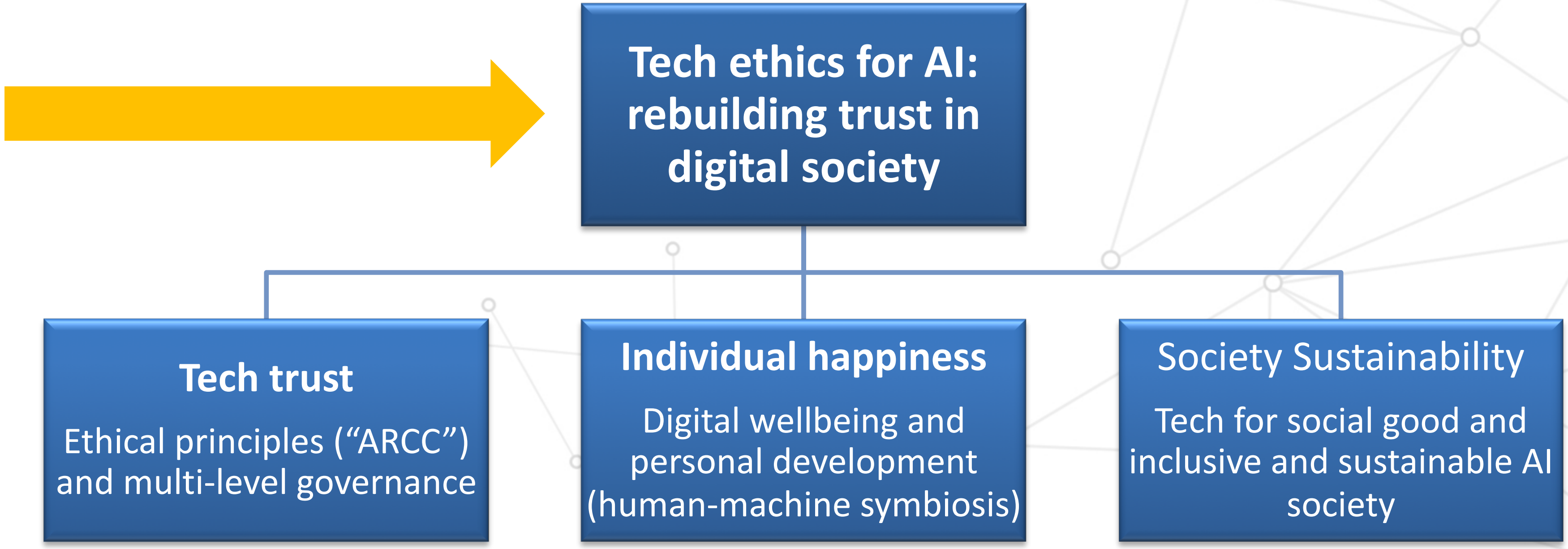
Tencent
AI Lab

Tech Ethics for AI Era: Rebuilding Trust in Digital Society

Tencent Research Institute

Tencent AI Lab

June 2019



Tech ethics for AI: rebuilding trust in digital society

Tech trust

Ethical principles (“ARCC”) and multi-level governance

Individual happiness

Digital wellbeing and personal development (human-machine symbiosis)

Society Sustainability

Tech for social good and inclusive and sustainable AI society

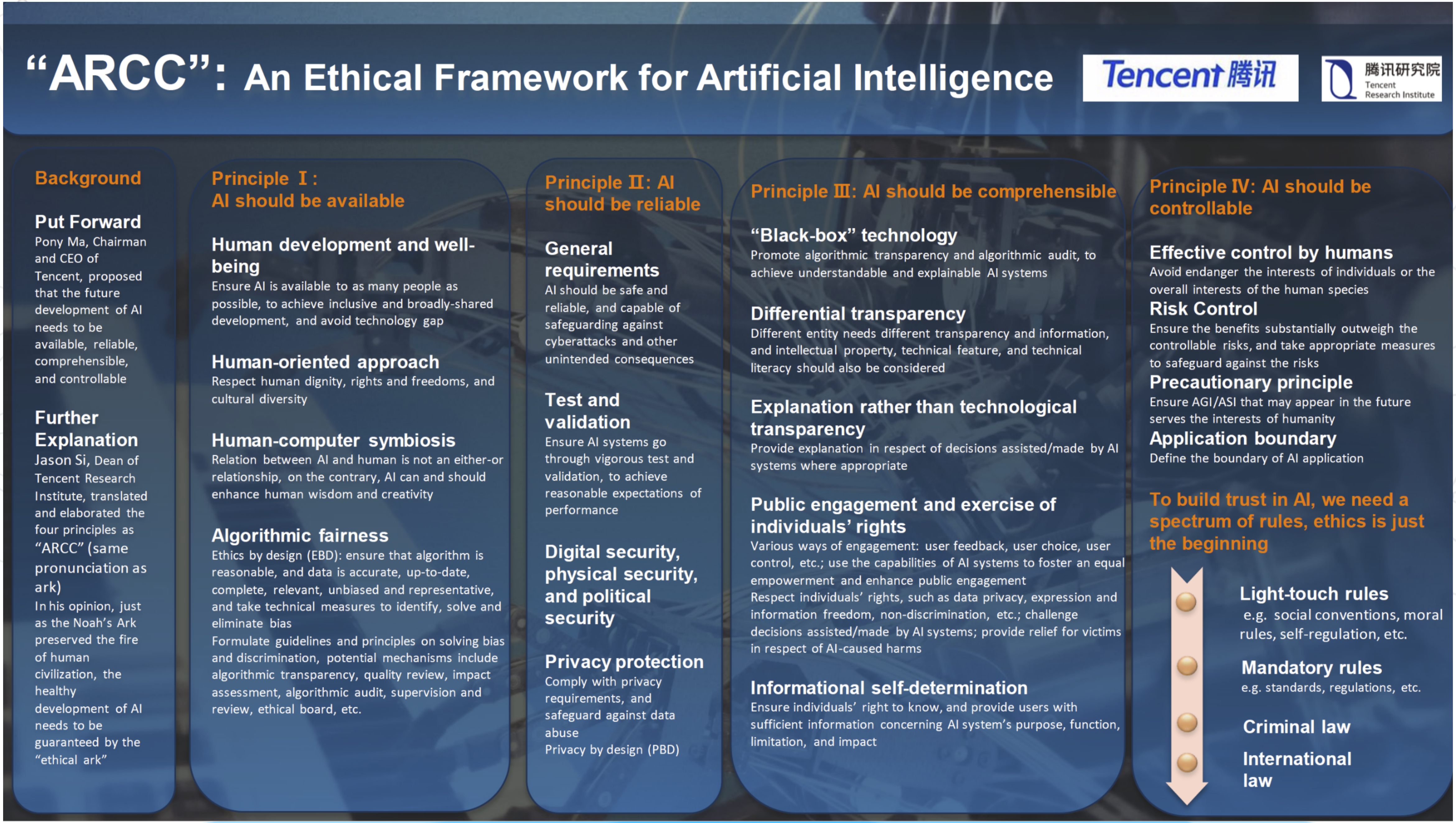
The “Ethical Ark” (ARCC) for the healthy and safe development of AI

available

reliable

comprehensible

controllable



"ARCC": An Ethical Framework for Artificial Intelligence

Tencent 腾讯

腾讯研究院 Tencent Research Institute

Background

Put Forward

Pony Ma, Chairman and CEO of Tencent, proposed that the future development of AI needs to be available, reliable, comprehensible, and controllable

Further Explanation

Jason Si, Dean of Tencent Research Institute, translated and elaborated the four principles as "ARCC" (same pronunciation as ark)

In his opinion, just as the Noah's Ark preserved the fire of human civilization, the healthy development of AI needs to be guaranteed by the "ethical ark"

Principle Ⅰ: AI should be available

Human development and well-being

Ensure AI is available to as many people as possible, to achieve inclusive and broadly-shared development, and avoid technology gap

Human-oriented approach

Respect human dignity, rights and freedoms, and cultural diversity

Human-computer symbiosis

Relation between AI and human is not an either-or relationship, on the contrary, AI can and should enhance human wisdom and creativity

Algorithmic fairness

Ethics by design (EBD): ensure that algorithm is reasonable, and data is accurate, up-to-date, complete, relevant, unbiased and representative, and take technical measures to identify, solve and eliminate bias

Formulate guidelines and principles on solving bias and discrimination, potential mechanisms include algorithmic transparency, quality review, impact assessment, algorithmic audit, supervision and review, ethical board, etc.

Principle Ⅱ: AI should be reliable

General requirements

AI should be safe and reliable, and capable of safeguarding against cyberattacks and other unintended consequences

Test and validation

Ensure AI systems go through vigorous test and validation, to achieve reasonable expectations of performance

Digital security, physical security, and political security

Privacy protection

Comply with privacy requirements, and safeguard against data abuse

Privacy by design (PBD)

Principle Ⅲ: AI should be comprehensible

"Black-box" technology

Promote algorithmic transparency and algorithmic audit, to achieve understandable and explainable AI systems

Differential transparency

Different entity needs different transparency and information, and intellectual property, technical feature, and technical literacy should also be considered

Explanation rather than technological transparency

Provide explanation in respect of decisions assisted/made by AI systems where appropriate

Public engagement and exercise of individuals' rights

Various ways of engagement: user feedback, user choice, user control, etc.; use the capabilities of AI systems to foster an equal empowerment and enhance public engagement

Respect individuals' rights, such as data privacy, expression and information freedom, non-discrimination, etc.; challenge decisions assisted/made by AI systems; provide relief for victims in respect of AI-caused harms

Informational self-determination

Ensure individuals' right to know, and provide users with sufficient information concerning AI system's purpose, function, limitation, and impact

Principle Ⅳ: AI should be controllable

Effective control by humans

Avoid endanger the interests of individuals or the overall interests of the human species

Risk Control

Ensure the benefits substantially outweigh the controllable risks, and take appropriate measures to safeguard against the risks

Precautionary principle

Ensure AGI/ASI that may appear in the future serves the interests of humanity

Application boundary

Define the boundary of AI application

To build trust in AI, we need a spectrum of rules, ethics is just the beginning

Light-touch rules

e.g. social conventions, moral rules, self-regulation, etc.

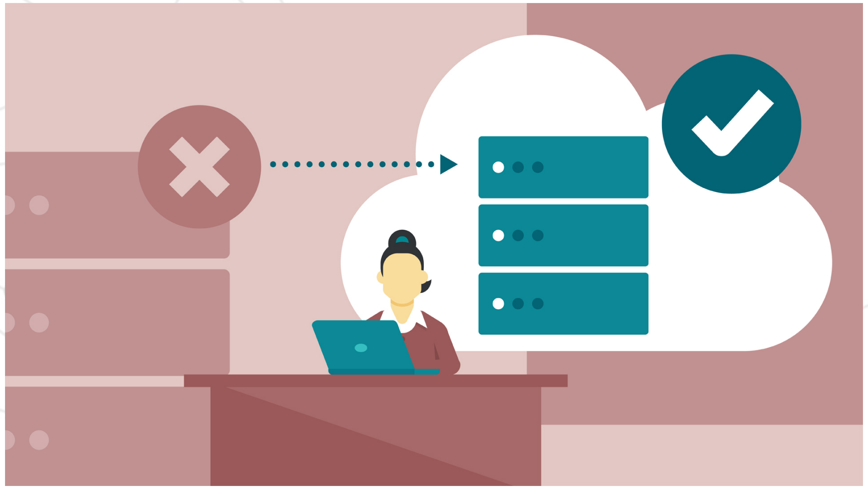
Mandatory rules

e.g. standards, regulations, etc.

Criminal law

International law

Principle Ⅰ: AI should be available



- **Human development and well-being**
  - Ensure AI is available to as many people as possible, to achieve inclusive and broadly-shared development and avoid technology gap
- **Human-oriented approach**
  - Respect human dignity, rights and freedoms, and cultural diversity
- **Human-machine symbiosis**
  - Relation between AI and human is not an either-or relationship, on the contrary, AI can and should enhance human wisdom and creativity
- **Algorithmic fairness**
  - Ensure that algorithm is reasonable and data is accurate, up-to-date, complete, relevant, unbiased and representative; take technical measures to identify, solve and eliminate bias
  - Formulate principles and guidelines on solving bias and discrimination; potential mechanisms include algorithmic transparency, quality review, impact assessment, algorithmic audit, supervision and review, ethical review, etc.

Principle Ⅱ: AI should be reliable

- **General requirements**
  - AI should be safe, reliable and capable of safeguarding against cyberattacks and other unintended consequences
- **Test and validation**
  - Ensure AI systems go through vigorous test and validation, to achieve reasonable expectations of performance
- **Safety and security: digital, physical, and social**
  - **Privacy and data protection:** (1) comply with privacy requirements; (2) safeguard against data abuse; (3) privacy by design (PBD)

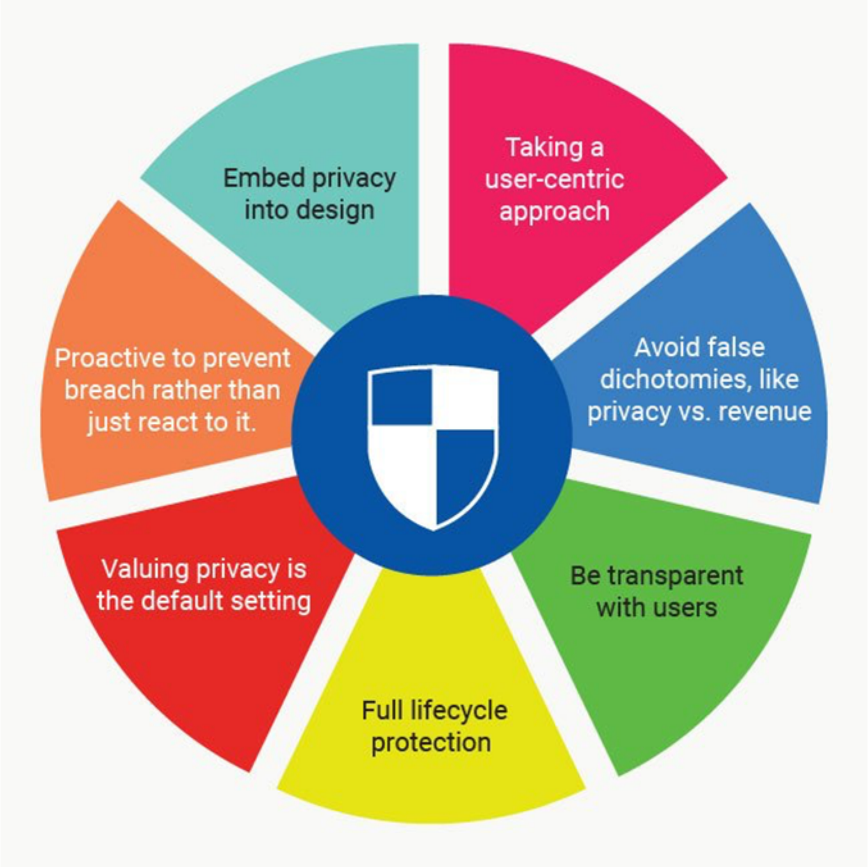
Principle Ⅲ: AI should be comprehensible

- **“Black-box” technology**
  - Committed to solve the “black-box” problem of AI, to achieve understandable and explainable AI models
- **Differential and reasonable algorithmic transparency**
  - Different entity needs different level of transparency information, and intellectual property, technical feature, technical literacy, **data privacy and safety of AI applications** should also be take into consideration
  - Provide explanation in respect of actions and decisions assisted/made by AI systems where appropriate rather than the complete detailed algorithm or the compete set of steps taken
- **Public engagement and exercise of individual’s rights**
  - Various ways of engagement: user feedback, user choice, user control, etc.; make use of the capabilities of AI systems to foster equal empowerment and enhance public engagement
  - Respect individual’s rights, such as data privacy, expression and information freedom, non-discrimination, etc.; challenge actions and decisions assisted/made by AI systems; provide relief and remedy for victims in respect of AI-caused harms
- **Informational self-determination**
  - Ensure individual’s right to know; provide users with sufficient information concerning the purpose, function, limitation, and impact of AI systems

Principle IV: AI should be controllable

- **Effective control by humans**
  - Avoid endanger the interests of individuals or the overall interests of humanity
  - Human takes responsibility for AI
- **Risk Control**
  - Ensure the benefits substantially outweigh the controllable risks, and take appropriate measures to safeguard against the risks
- **Application boundary**
  - Define the boundary of AI application
- **Precautionary measures**
  - Ensure AGI/ASI that may appear in the future serves the interests of humanity

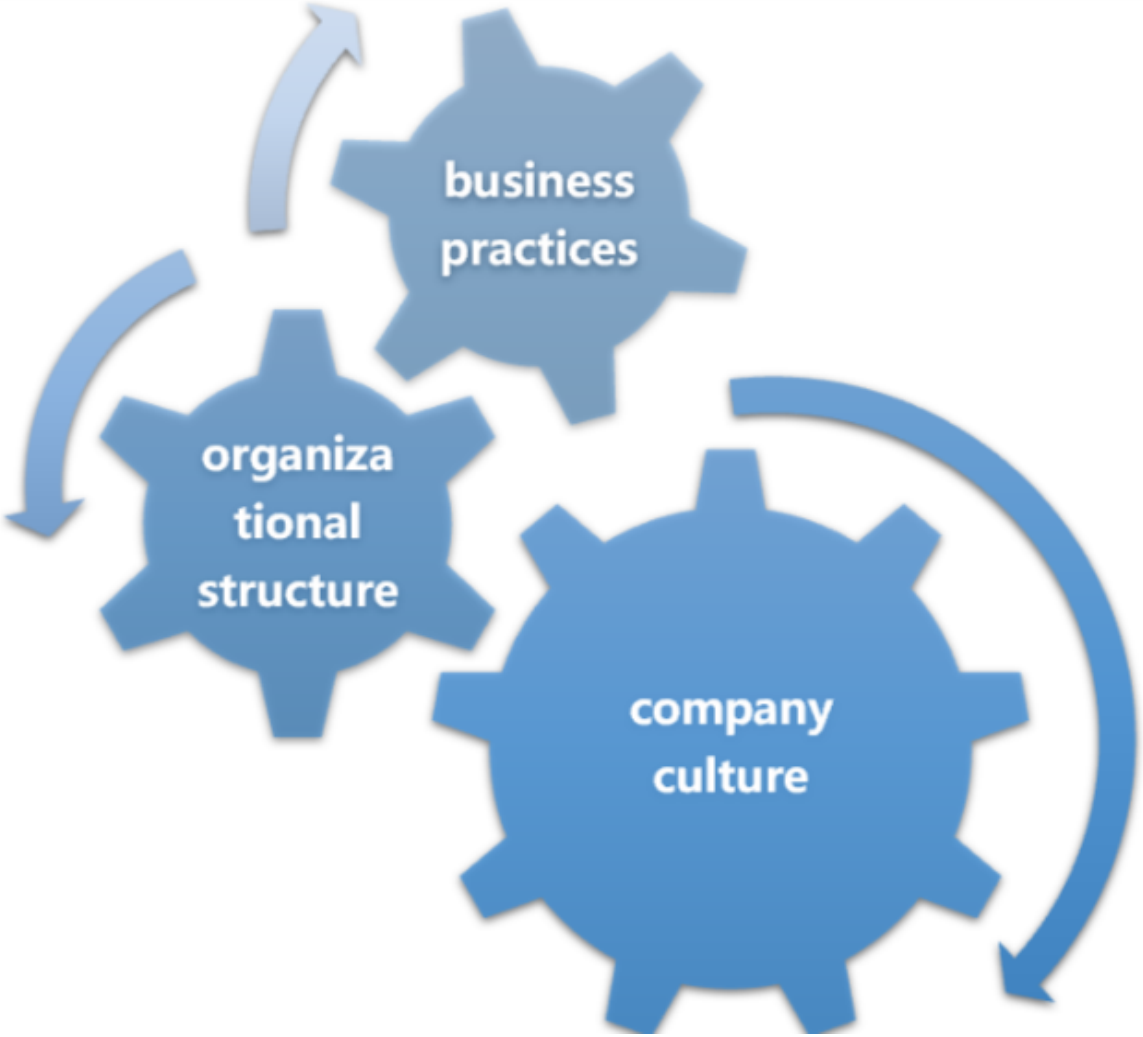
Follow the Ethics by Design approach to achieve "value-aligned" AI


Embed privacy into design
Taking a user-centric approach
Avoid false dichotomies, like privacy vs. revenue
Be transparent with users
Full lifecycle protection
Valuing privacy is the default setting
Proactive to prevent breach rather than just react to it.

From privacy by design to ethics by design

- Digital wellbeing and personal rights
- Algorithmic fairness
- Informational self-determination
- Value preserving AI methods, such as federated learning

Bring AI ethics to life



business practices

organizational structure

company culture

thanks!