腾讯研究院
Tencent
Research Institute

AI Principles and Ethics by Design

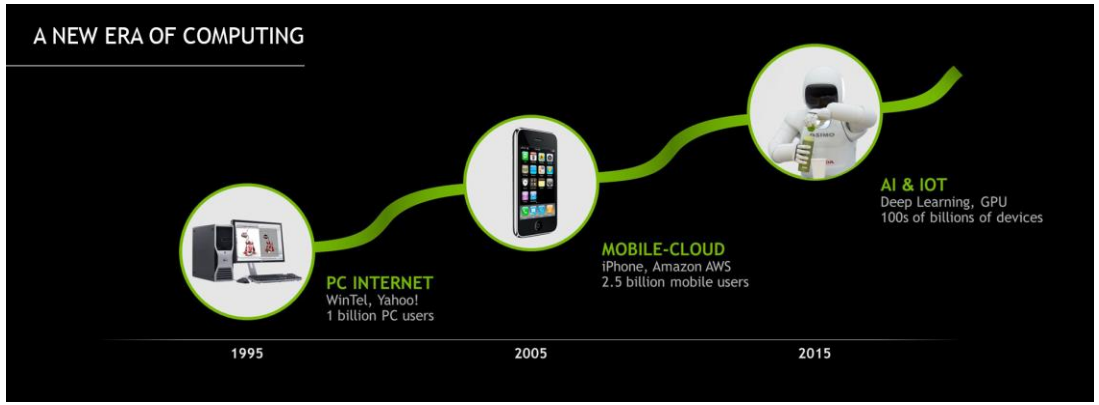
Jeff CAO

Tencent Research Institute

jeffcao@tencent.com

From internet era to AI era, data and algorithms are reshaping the world

- Ubiquitous connectivity
- Accelerating automation and autonomy
- Fusion of physical world and digital world, even human and technology



Data is “new oil”, AI is “new electricity”, what is “new pollution”?

AI brings various legal, ethical and societal issues (LESIs)

- Unintended behaviors
  - 2010 Flash Crash; editing wars between Wikipedia bots
- Lack of foresight
- Difficulty of oversight
- Distributed responsibility
- **Various risks**
  - Algorithmic bias; restraint of free choice; fake news and deepfake; amplification and perpetuation of bias and social stratification; abuse of data and algorithms; issues of surveillance, privacy and freedom; **safety and liability issues**; technological unemployment

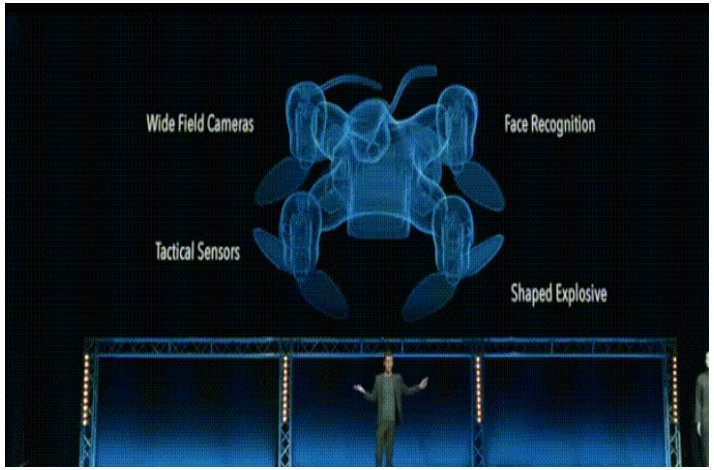
unintended behaviors

deepfake

algorithmic bias

Three levels of AI safety

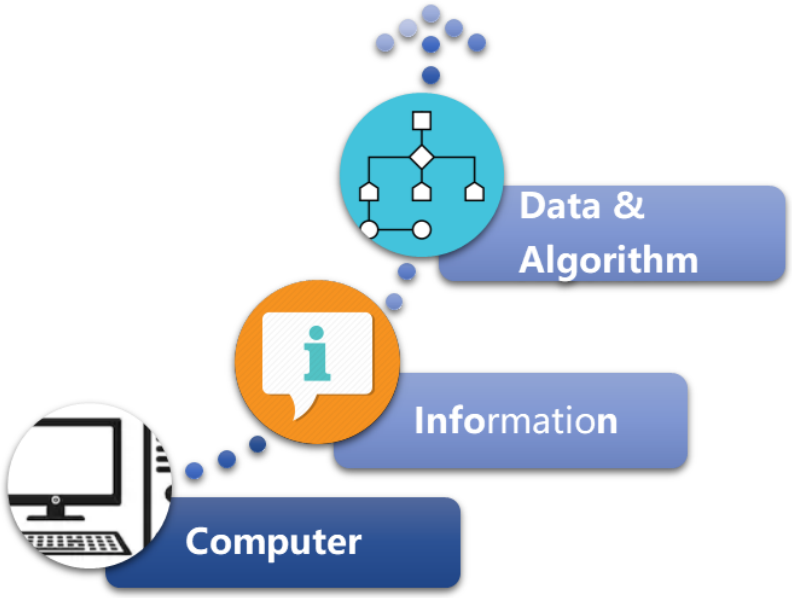
Technical safety



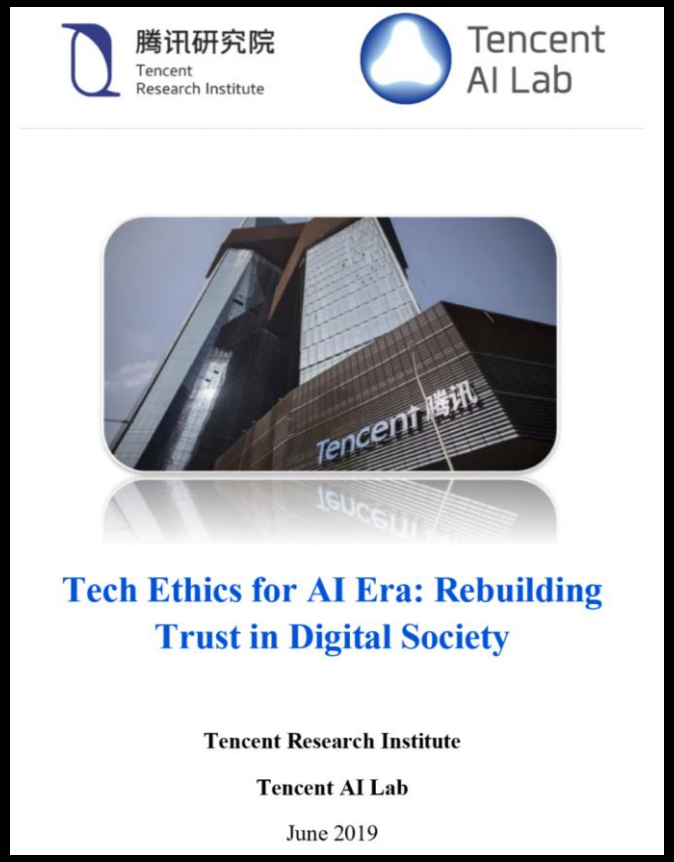
Physical safety

Social safety

We urgently need AI ethics oriented to data and algorithm



Data	Algorithm	Practices
• privacy re-identification group privacy • trust • transparency	• responsibility & accountability • ethical design of algorithm • ethical auditing of algorithm	• ethical code of conduct • consent • privacy of data subjects • secondary use



Tech ethics for AI: rebuilding trust in digital society

Tech trust

Ethical principles ("ARCC") and multi-level governance

Individual happiness

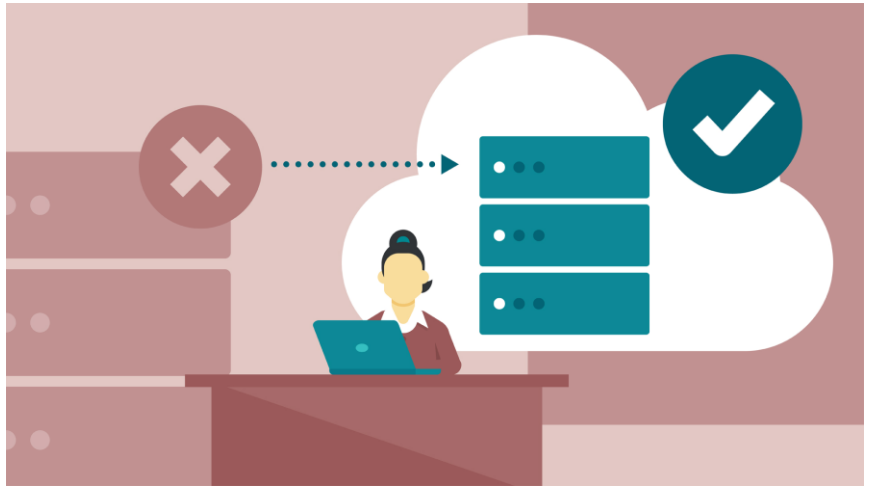
Digital wellbeing and personal development (human-machine symbiosis)

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

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)

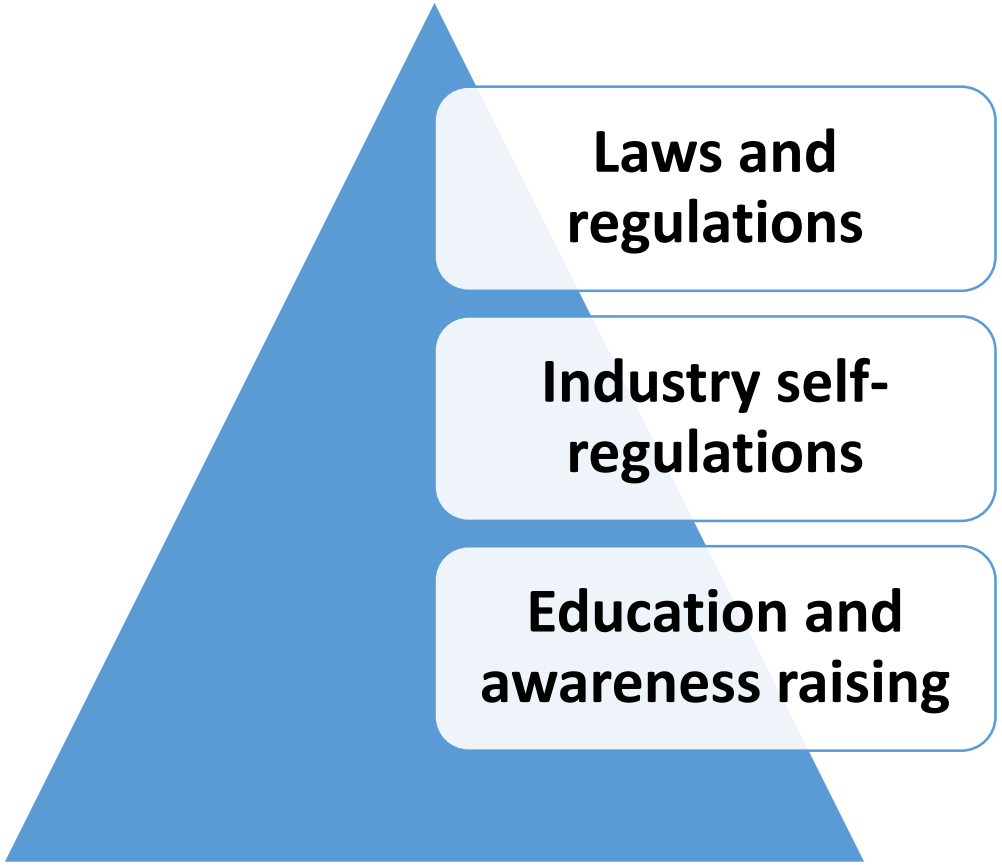
Principle Ⅲ: AI should be comprehensible

- **“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

To build trust in AI, we need multi-level governance

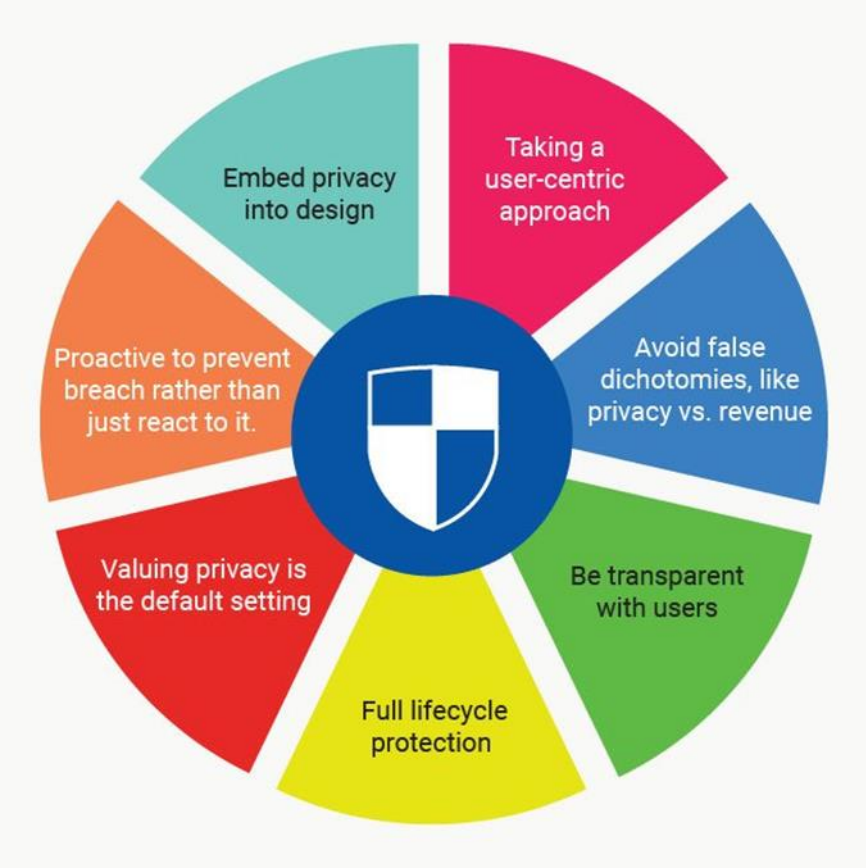


- Mandatory rules, criminal law, international law
- National committee on tech ethics

- Industry standards, best practice, technical guidance, ethical rules, self-discipline pact, etc

- Ethical education and training for AI researchers and practitioners
- Raising algorithmic literacy of watchdogs and the public

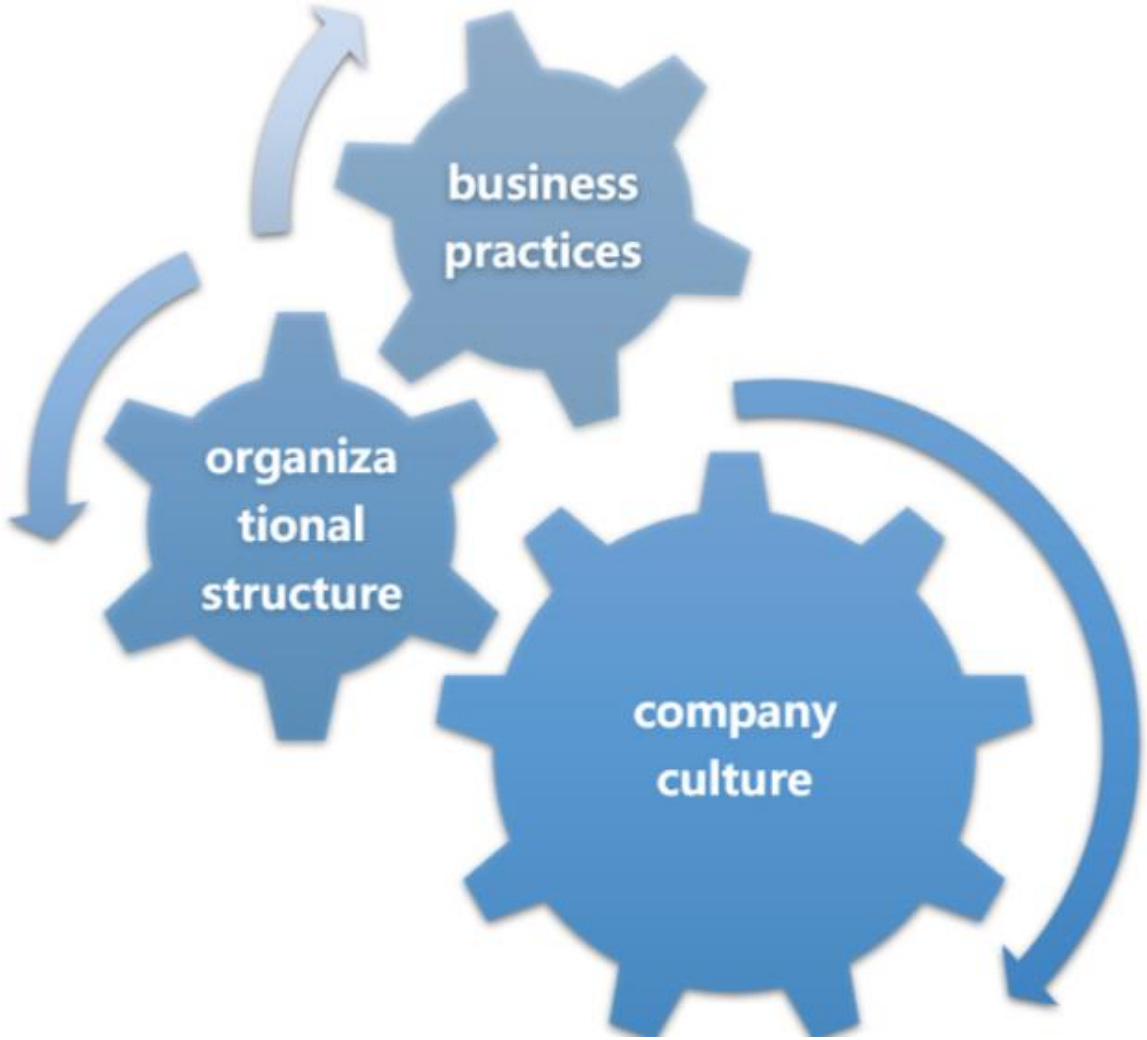
Follow the Ethics by Design approach to achieve “value-aligned” AI



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



Tencent's Pony Ma Declares Company's New Mission And Vision Statement — Tech For Social Good

INDUSTRY David Lee May 6, 2019

Thanks for listening