Ethical Risks and Challenges of Computational Intelligence

Xin Yao

School of Data Science

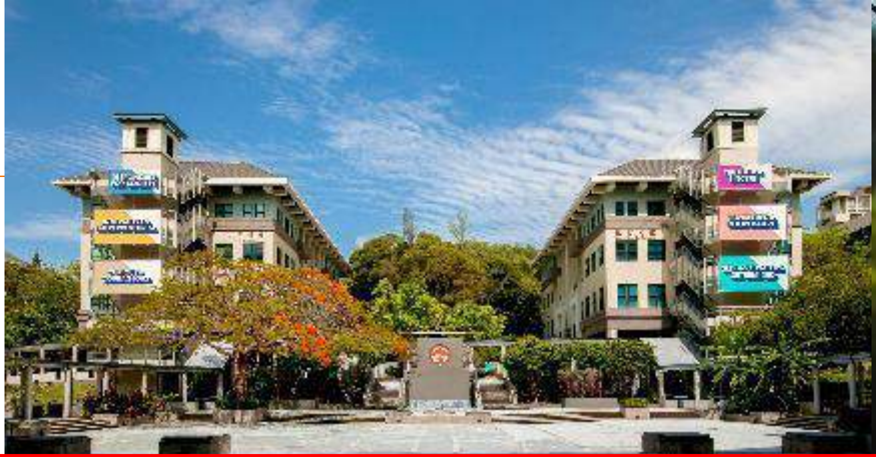
Lingnan University

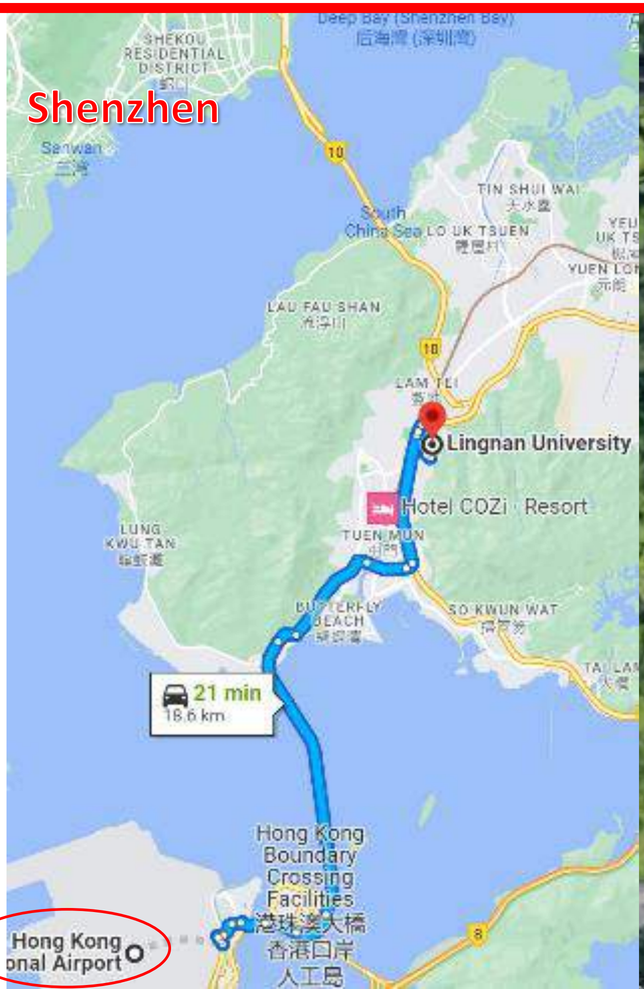
Hong Kong SAR, China

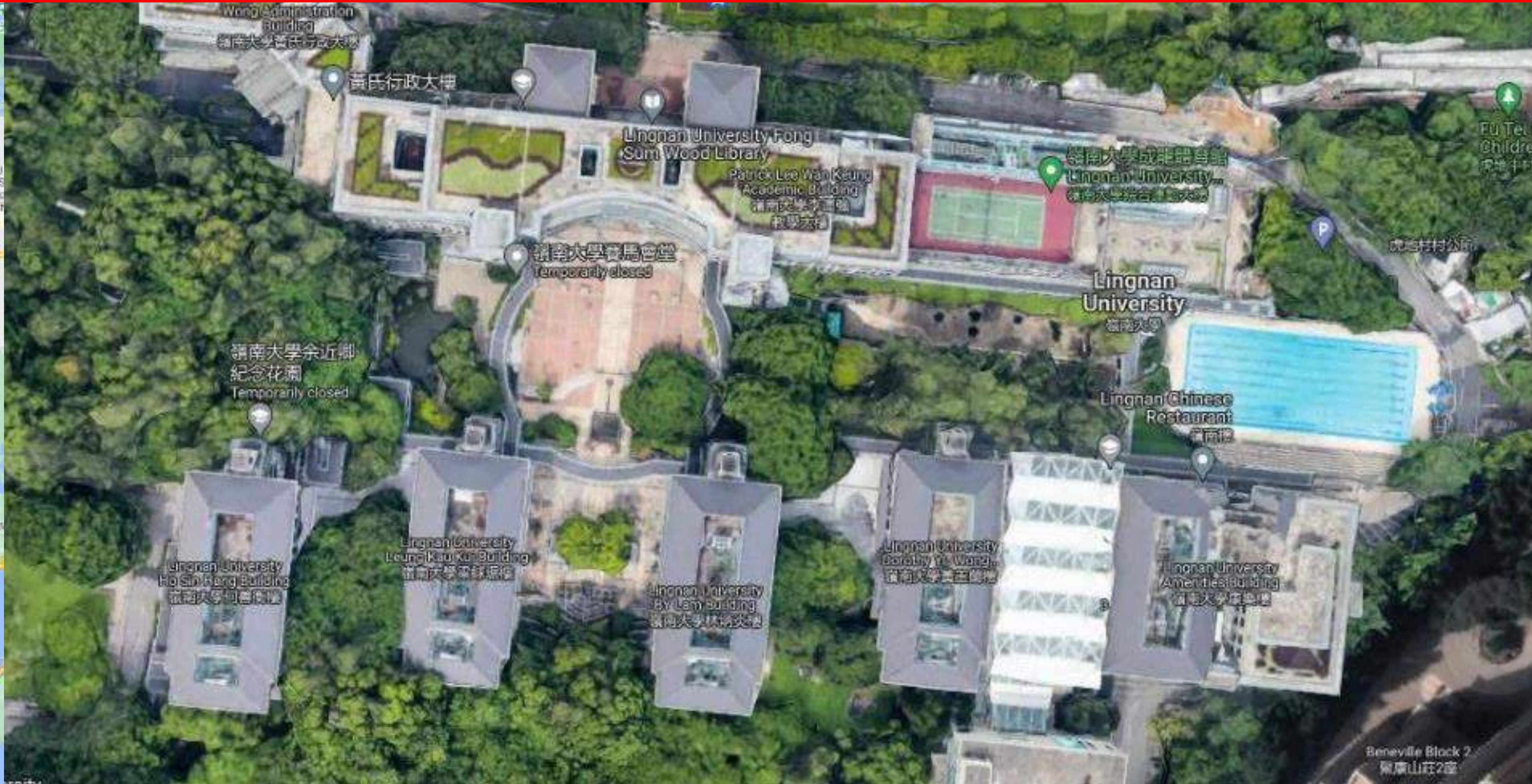
Location

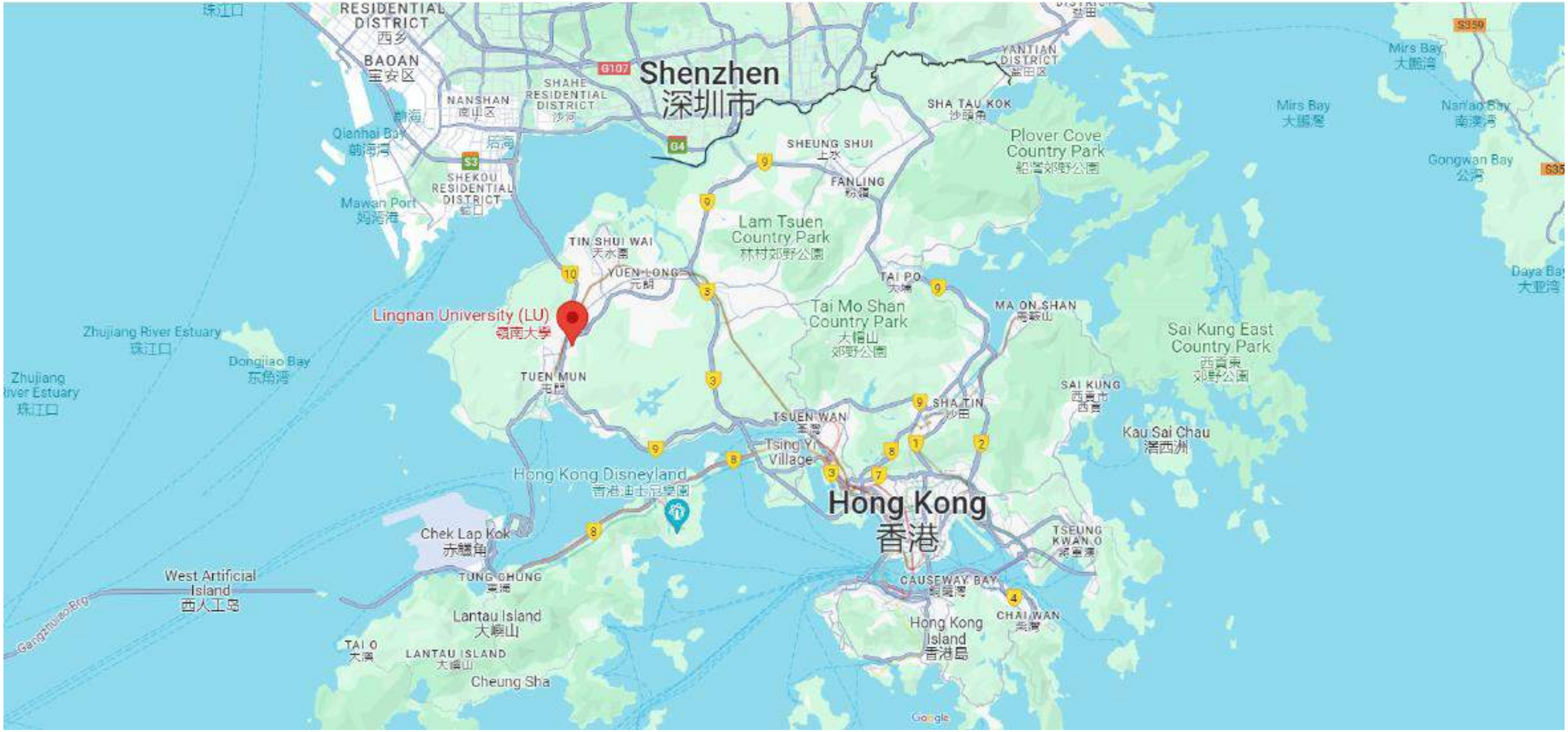
Lingnan University

香港，屯门

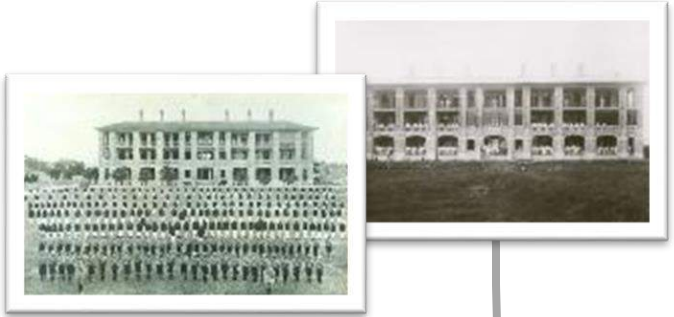





Wong Administration Building
嶺南大學黃氏行政大樓
黃氏行政大樓
Lingnan University Fong Sum Wood Library
Patrick Lee Wan Keung Academic Building
嶺南大學李運強教學大樓
嶺南大學成龍體育館
Lingnan University...
嶺南大學賽馬會堂
Temporarily closed
嶺南大學余近卿紀念花園
Temporarily closed
Lingnan University
嶺南大學
Fu Tei Children
P
Lingnan Chinese Restaurant
嶺南樓
Lingnan University Ho Sin Hang Building
Lingnan University Leung Kau Kui Building
Lingnan University B.Y. Lam Building
嶺南大學林炳炎樓
Lingnan University Dorothy Y. L. Wong...
Lingnan University Amenities Building
Beneville Block 2


Lingnan University (LU)
嶺南大學
Shenzhen
深圳市
Hong Kong
香港
BAOAN
宝安区
NANSHAN
南山区
SHAHE RESIDENTIAL DISTRICT
沙河
SHEKOU RESIDENTIAL DISTRICT
蛇口
Qianhai Bay
前海湾
Mawan Port
妈湾港
YANTIAN DISTRICT
盐田区
SHA TAU KOK
沙頭角
SHEUNG SHUI
上水
FANLING
粉嶺
Plover Cove Country Park
船灣郊野公園
Lam Tsuen Country Park
林村郊野公園
TIN SHUI WAI
天水圍
YUEN LONG
元朗
TAI PO
大埔
Tai Mo Shan Country Park
大帽山郊野公園
MA ON SHAN
馬鞍山
Sai Kung East Country Park
西貢東郊野公園
SAI KUNG
西貢市
Kau Sai Chau
滘西洲
TUEN MUN
屯門
TSUEN WAN
荃灣
SHA TIN
沙田
Tsing Yi Village
Hong Kong Disneyland
香港迪士尼樂園
Chek Lap Kok
赤鱲角
TUNG CHUNG
東涌
Lantau Island
大嶼山
LANTAU ISLAND
TAI O
大澳
Cheung Sha
West Artificial Island
西人工岛
Zhujiang River Estuary
珠江口
Dongjiao Bay
东角湾
Mirs Bay
大鵬灣
Nan'ao Bay
南澳湾
Gongwan Bay
Daya Bay
大亚湾
TSEUNG KWAN O
將軍澳
CAUSEWAY BAY
銅鑼灣
CHAI WAN
柴灣
Hong Kong Island
香港島
Google

Lingnan’s History（Ⅰ）



1888

Christian College in China established in Guangzhou

1903

New campus set up at Kangle village of Guangzhou. Renamed **Canton Christian College** in English and adopted Chinese name **Lingnan Xuetang**

1912

Renamed **Lingnan Xuexiao** in Chinese after 1911 Revolution

1918

Began offering university-level programmes, recognized by fifteen prestigious universities including Harvard, Yale, Columbia, and Stanford

1927

Renamed Lingnan Daxue in Chinese and Lingnan University in English. Dr Chung Wing-kwong and Dr Lee Ying-lam were elected as President and Vice-President.

◀ Dr Chung Wing-kwong, the first President of Lingnan University

Lingnan's History（ II ）

1937

Dr Lee Ying-lam became the President

1938

Moved to Hong Kong after the fall of Guangzhou during anti-Japanese War and continued operation using the campus of The University of Hong Kong. School of Agriculture moved to Nam Tei of Tuen Mun.

1948

Chen Xu-jing became the President, CHEN Yinke, WANG Li, JIANG Lifu, Liang Fangzhong and other famous scholars joined Lingnan University

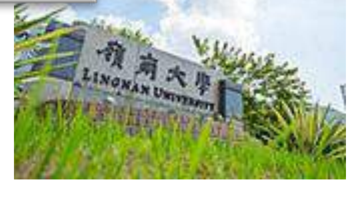
1952

Lingnan University was incorporated into other Guangzhou institutions, maintaining a unique reputation in South China's higher education.

1967

Lingnan College
Re-established at Stubbs Road of Hong Kong.

1999



Officially renamed Lingnan University, and became one of the UGC-funded public universities in Hong Kong.

Fine Print About This Part of the Tutorial

- **This tutorial is prepared for those who are unfamiliar with but interested in finding out the very basics of ethics.**

- **The tutorial is organized from more general concepts to specific topics.**

- **Further information can be obtained from the following papers:**
  - C. Huang Z. Zhang, B. Mao and X. Yao, "An Overview of Artificial Intelligence Ethics." *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 4, pp. 799-819, Aug. 2023 (doi: 10.1109/TAI.2022.3194503).
  - Q. Zhang, J. Liu, Z. Zhang, J. Wen, B. Mao and X. Yao, "Mitigating Unfairness via Evolutionary Multi-objective Ensemble Learning," *IEEE Transactions on Evolutionary Computation*, vol. 27, no. 4, pp. 848-862, Aug. 2023, doi: 10.1109/TEVC.2022.3209544.
  - Z. Wang, C. Huang, Y. Li and X. Yao, "Multi-objective Feature Attribution Explanation For Explainable Machine Learning," *ACM Transactions on Evolutionary Learning and Optimization,* Volume 4, Issue 1, Article No. 2, pp. 1–32. February 2024. https://doi.org/10.1145/3617380.
  - C. Huang, Z. Zhang, B. Mao and X. Yao, "Preventing Undesirable Behaviors of Neural Networks via Evolutionary Constrained Learning," Proc. of the *2022 International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy, 2022, pp. 1-7, doi: 10.1109/IJCNN55064.2022.9891926.

Outline

1. From Ethics to Artificial Intelligence (AI) Ethics

2. Fairer ML Through Multi-Objective Evolutionary Learning

3. Multi-objective Feature Attribution Explanation (MOFAE)

4. Regulating Machine Learning Behaviours Through Evolutionary Constrained Learning

5. Concluding Remarks

Outline

1. From Ethics to Artificial Intelligence (AI) Ethics

2. Fairer ML Through Multi-Objective Evolutionary Learning

3. Multi-objective Feature Attribution Explanation (MOFAE)

4. Regulating Machine Learning Behaviours Through Evolutionary Constrained Learning

5. Concluding Remarks

What is Ethics ?

A branch of philosophy.

- Ethics is a study of what are good and bad to pursue in life and what it is right and wrong to do in the conduct of life.
- It is concerned with systematizing, defending, and recommending concepts of right and wrong behaviors [1].
- Ethics seeks to resolve questions of human morality by defining concepts such as good and evil, right and wrong, virtue and vice, justice and crime [1].
- Ethics focuses on judging and determining which action would be good or moral in given circumstances [2].

[1] https://en.wikipedia.org/wiki/Ethics

[2] S. L. Anderson and M. Anderson, "AI and ethics," AI Ethics, vol. 1, no. 1, pp. 27–31, 2021.

[3] https://academyofideas.com/2013/08/introduction-to-ethics/#:~:text=As%20a%20philosophical%20discipline%20ethics,moral%20philosophers%20in%20Western%20Civilization.

What is Ethics ?

A branch of philosophy.

- **Ethics is a study of what are good and bad to pursue in life and what it is right and wrong to do in the conduct of life.**
- **It is concerned with systematizing, defending, and recommending concepts of right and wrong behaviors [1].**
- **Ethics seeks to resolve questions of human morality by defining concepts such as good and evil, right and wrong, virtue and vice, justice and crime [1].**
- **Ethics focuses on judging and determining which action would be good or moral in given circumstances [2].**

[1] https://en.wikipedia.org/wiki/Ethics

[2] S. L. Anderson and M. Anderson, “AI and ethics,” AI Ethics, vol. 1, no. 1, pp. 27–31, 2021.

[3] https://academyofideas.com/2013/08/introduction-to-ethics/#:~:text=As%20a%20philosophical%20discipline%20ethics,moral%20philosophers%20in%20Western%20Civilization.

What is Ethics ?

A branch of philosophy.

- Ethics is a study of what are good and bad to pursue in life and what it is right and wrong to do in the conduct of life.
- It is concerned with systematizing, defending, and recommending concepts of right and wrong behaviors [1].
- Ethics seeks to resolve questions of human morality by defining concepts such as good and evil, right and wrong, virtue and vice, justice and crime [1].
- Ethics focuses on judging and determining which action would be good or moral in given circumstances [2].

[1] https://en.wikipedia.org/wiki/Ethics
[2] S. L. Anderson and M. Anderson, "AI and ethics," AI Ethics, vol. 1, no. 1, pp. 27–31, 2021.
[3] https://academyofideas.com/2013/08/introduction-to-ethics/#:~:text=As%20a%20philosophical%20discipline%20ethics,moral%20philosophers%20in%20Western%20Civilization.

What is Ethics ?

A branch of philosophy.

- Ethics is a study of what are good and bad to pursue in life and what it is right and wrong to do in the conduct of life.
- It is concerned with systematizing, defending, and recommending concepts of right and wrong behaviors [1].
- Ethics seeks to resolve questions of human morality by defining concepts such as good and evil, right and wrong, virtue and vice, justice and crime [1].
- Ethics focuses on judging and determining which action would be good or moral in given circumstances [2].

[1] https://en.wikipedia.org/wiki/Ethics

[2] S. L. Anderson and M. Anderson, "AI and ethics," AI Ethics, vol. 1, no. 1, pp. 27–31, 2021.

[3] https://academyofideas.com/2013/08/introduction-to-ethics/#:~:text=As%20a%20philosophical%20discipline%20ethics,moral%20philosophers%20in%20Western%20Civilization.

What is Ethics ?

A branch of philosophy.

- Ethics is a study of what are good and bad to pursue in life and what it is right and wrong to do in the conduct of life.
- It is concerned with systematizing, defending, and recommending concepts of right and wrong behaviors [1].
- Ethics seeks to resolve questions of human morality by defining concepts such as good and evil, right and wrong, virtue and vice, justice and crime [1].
- Ethics focuses on judging and determining which action would be good or moral in given circumstances [2].

[1] https://en.wikipedia.org/wiki/Ethics

[2] S. L. Anderson and M. Anderson, "AI and ethics," AI Ethics, vol. 1, no. 1, pp. 27–31, 2021.

[3] https://academyofideas.com/2013/08/introduction-to-ethics/#:~:text=As%20a%20philosophical%20discipline%20ethics,moral%20philosophers%20in%20Western%20Civilization.

What is Ethics ?

A branch of philosophy.

- **Ethics is a study of what are good and bad to pursue in life and what it is right and wrong to do in the conduct of life.**
- **It is concerned with systematizing, defending, and recommending concepts of right and wrong behaviors [1].**
- **Ethics seeks to resolve questions of human morality by defining concepts such as good and evil, right and wrong, virtue and vice, justice and crime [1].**
- **Ethics focuses on judging and determining which action would be good or moral in given circumstances [2].**

[1] https://en.wikipedia.org/wiki/Ethics

[2] S. L. Anderson and M. Anderson, “AI and ethics,” AI Ethics, vol. 1, no. 1, pp. 27–31, 2021.

[3] https://academyofideas.com/2013/08/introduction-to-ethics/#:~:text=As%20a%20philosophical%20discipline%20ethics,moral%20philosophers%20in%20Western%20Civilization.

Three Major Branches of Study in Ethics

- **Meta-Ethics** investigates the nature, scope, and meaning of ethical principles or moral judgment.
- **Normative Ethics** seeks to arrive at moral standards and rules that regulate right and wrong behaviors.
- **Applied Ethics** studies the ethics in application fields, which consists of the analysis of specific, controversial moral issues, such as abortion, capital punishment, animal rights, environmental concerns, nuclear war etc.

Three Major Branches of Study in Ethics

- **Meta-Ethics** investigates the nature, scope, and meaning of ethical principles or moral judgment.
- **Normative Ethics** seeks to arrive at moral standards and rules that regulate right and wrong behaviors.
- **Applied Ethics** studies the ethics in application fields, which consists of the analysis of specific, controversial moral issues, such as abortion, capital punishment, animal rights, environmental concerns, nuclear war etc.

Three Major Branches of Study in Ethics

- **Meta-Ethics** investigates the nature, scope, and meaning of ethical principles or moral judgment.
- **Normative Ethics** seeks to arrive at moral standards and rules that regulate right and wrong behaviors.
- **Applied Ethics** studies the ethics in application fields, which consists of the analysis of specific, controversial moral issues, such as abortion, capital punishment, animal rights, environmental concerns, nuclear war etc.

Three Major Branches of Study in Ethics

- **Meta-Ethics** investigates the nature, scope, and meaning of ethical principles or moral judgment.
- **Normative Ethics** seeks to arrive at moral standards and rules that regulate right and wrong behaviors.
- **Applied Ethics** studies the ethics in application fields, which consists of the analysis of specific, controversial moral issues, such as abortion, capital punishment, animal rights, environmental concerns, nuclear war, etc.

Normative Ethics

- **Virtue Ethics.** Virtue ethics emphasizes the virtues or moral character and stresses the importance of cultivating good habits of character, such as benevolence.
- **Deontological Ethics.** Deontological theories, which are sometimes called duty theories, judge the morality of an action using certain moral rules that serve as foundational principles of obligation. There are three main schools of deontological theories, that is, agent-centered, patient-centered (also called victim-centered), and contractarian deontological theories.
- **Consequentialist Ethics.** Consequentialist ethics, as its name suggests, emphasizes the utilitarian outcomes of actions. Consequentialist theories can be divided into Ethical Egoism, Ethical Altruism, and Utilitarianism.

Table 1 Comparison of the three normative ethical theories [4]

Ethical Theory	Description	Deliberation Focus	Decision Criteria	Practical Reasoning
Virtue Ethics	An action is right if it is what a virtuous person would do in the situation.	Motives (Is action motivated by virtue?)	Virtues	Instantiation of virtues / human qualities
Deontological Ethics	An action is right if it is in accordance with a moral rule or principle.	Action (Is action compatible with some imperative?)	Duties/rules	Follow the rules
Consequentialist Ethics	**An action is right if it promotes the best consequences, i.e., maximizes happiness.**	**Consequences (What is outcome of action?)**	**Comparative well-being**	**Maximization of utility or happiness**

[4] Tolmeijer S, Kneer M, Sarasua C, et al. Implementations in machine ethics: A survey[J]. ACM Computing Surveys (CSUR), 2020, 53(6): 1-38.

Normative Ethics

- **Virtue Ethics.** Virtue ethics emphasizes the virtues or moral character and stresses the importance of cultivating good habits of character, such as benevolence.
- **Deontological Ethics.** Deontological theories, which are sometimes called duty theories, judge the morality of an action using certain moral rules that serve as foundational principles of obligation. There are three main schools of deontological theories, that is, agent-centered, patient-centered (also called victim-centered), and contractarian deontological theories.
- **Consequentialist Ethics.** Consequentialist ethics, as its name suggests, emphasizes the utilitarian outcomes of actions. Consequentialist theories can be divided into Ethical Egoism, Ethical Altruism, and Utilitarianism.

Table 1 Comparison of the three normative ethical theories [4]

Ethical Theory	Description	Deliberation Focus	Decision Criteria	Practical Reasoning
Virtue Ethics	**An action is right if it is what a virtuous person would do in the situation.**	**Motives (Is action motivated by virtue?)**	**Virtues**	**Instantiation of virtues / human qualities**
Deontological Ethics	An action is right if it is in accordance with a moral rule or principle.	Action (Is action compatible with some imperative?)	Duties/rules	Follow the rules
Consequentialist Ethics	**An action is right if it promotes the best consequences, i.e., maximizes happiness.**	**Consequences (What is outcome of action?)**	**Comparative well-being**	**Maximization of utility or happiness**

[4] Tolmeijer S, Kneer M, Sarasua C, et al. Implementations in machine ethics: A survey[J]. ACM Computing Surveys (CSUR), 2020, 53(6): 1-38.

Normative Ethics

- **Virtue Ethics.** Virtue ethics emphasizes the virtues or moral character and stresses the importance of cultivating good habits of character, such as benevolence.
- **Deontological Ethics.** Deontological theories, which are sometimes called duty theories, judge the morality of an action using certain moral rules that serve as foundational principles of obligation. There are three main schools of deontological theories, that is, agent-centered, patient-centered (also called victim-centered), and contractarian deontological theories.
- **Consequentialist Ethics.** Consequentialist ethics, as its name suggests, emphasizes the utilitarian outcomes of actions. Consequentialist theories can be divided into Ethical Egoism, Ethical Altruism, and Utilitarianism.

Table 1 Comparison of the three normative ethical theories [4]

Ethical Theory	Description	Deliberation Focus	Decision Criteria	Practical Reasoning
Virtue Ethics	An action is right if it is what a virtuous person would do in the situation.	Motives (Is action motivated by virtue?)	Virtues	Instantiation of virtues / human qualities
Deontological Ethics	**An action is right if it is in accordance with a moral rule or principle.**	**Action (Is action compatible with some imperative?)**	**Duties/rules**	**Follow the rules**
Consequentialist Ethics	**An action is right if it promotes the best consequences, i.e., maximizes happiness.**	**Consequences (What is outcome of action?)**	**Comparative well-being**	**Maximization of utility or happiness**

[4] Tolmeijer S, Kneer M, Sarasua C, et al. Implementations in machine ethics: A survey[J]. ACM Computing Surveys (CSUR), 2020, 53(6): 1-38.

Normative Ethics

- **Virtue Ethics.** Virtue ethics emphasizes the virtues or moral character and stresses the importance of cultivating good habits of character, such as benevolence.
- **Deontological Ethics.** Deontological theories, which are sometimes called duty theories, judge the morality of an action using certain moral rules that serve as foundational principles of obligation. There are three main schools of deontological theories, that is, agent-centered, patient-centered (also called victim-centered), and contractarian deontological theories.
- **Consequentialist Ethics.** Consequentialist ethics, as its name suggests, emphasizes the utilitarian outcomes of actions. Consequentialist theories can be divided into Ethical Egoism, Ethical Altruism, and Utilitarianism.

Table 1 Comparison of the three normative ethical theories [4]

Ethical Theory	Description	Deliberation Focus	Decision Criteria	Practical Reasoning
Virtue Ethics	An action is right if it is what a virtuous person would do in the situation.	Motives (Is action motivated by virtue?)	Virtues	Instantiation of virtues / human qualities
Deontological Ethics	An action is right if it is in accordance with a moral rule or principle.	Action (Is action compatible with some imperative?)	Duties/rules	Follow the rules
Consequentialist Ethics	**An action is right if it promotes the best consequences, i.e., maximizes happiness.**	**Consequences (What is outcome of action?)**	**Comparative well-being**	**Maximization of utility or happiness**

[4] Tolmeijer S, Kneer M, Sarasua C, et al. Implementations in machine ethics: A survey[J]. ACM Computing Surveys (CSUR), 2020, 53(6): 1-38.

Technology Ethics

- **Technology Ethics or Ethics of Technology (Technoethics)** is a sub-field of applied ethics addressing the ethical issues of technology. It is the application of ethical thinking to the practical concerns of technology.

- Its aim is to study and address ethical issues in technological activities, to identify the morally right courses of action when we develop and apply technology.

- Technology ethics seeks to ensure that technological activities have a positive impact on our society and to avoid unintended consequences.

Technology Ethics

- **Technology Ethics or Ethics of Technology (Technoethics)** is a sub-field of applied ethics addressing the ethical issues of technology. It is the application of ethical thinking to the practical concerns of technology.

- Its aim is to study and address ethical issues in technological activities, to identify the morally right courses of action when we develop and apply technology.

- Technology ethics seeks to ensure that technological activities have a positive impact on our society and to avoid unintended consequences.

Technology Ethics

- **Technology Ethics or Ethics of Technology (Technoethics)** is a sub-field of applied ethics addressing the ethical issues of technology. It is the application of ethical thinking to the practical concerns of technology.

- Its aim is to study and address ethical issues in technological activities, to identify the morally right courses of action when we develop and apply technology.

- Technology ethics seeks to ensure that technological activities have a positive impact on our society and to avoid unintended consequences.

Technology Ethics

- **Technology Ethics or Ethics of Technology (Technoethics)** is a sub-field of applied ethics addressing the ethical issues of technology. It is the application of ethical thinking to the practical concerns of technology.

- Its aim is to study and address ethical issues in technological activities, to identify the morally right courses of action when we develop and apply technology.

- Technology ethics seeks to ensure that technological activities have a positive impact on our society and to avoid unintended consequences.

Technology Ethics (TE): Approaches and Technologies

Three major approaches:

- **Responsibility-based TE：**As the impact or harm of technology on the future may be much greater than we imagine, the responsibility and obligation to protect environment and future human beings is the core of responsibility-based TE.
- **Utilitarian-based TE:** Utilitarian-based TE mainly concerns the consequences of technological development and attempts to avoid adverse consequences through the regulation and intervention in technological activities.
- **Value-based TE:** Value-based TE conducts ethical analysis of technological activities by focusing on the value or value conflict of technology.

- **Technologies**

Technology Ethics (TE): Approaches and Technologies

Three major approaches:

- **Responsibility-based TE：**As the impact or harm of technology on the future may be much greater than we imagine, the responsibility and obligation to protect environment and future human beings is the core of responsibility-based TE.
- **Utilitarian-based TE:** Utilitarian-based TE mainly concerns the consequences of technological development and attempts to avoid adverse consequences through the regulation and intervention in technological activities.
- **Value-based TE:** Value-based TE conducts ethical analysis of technological activities by focusing on the value or value conflict of technology.

- Technologies

Technology Ethics (TE): Approaches and Technologies

- Technologies

Three major approaches:

- **Responsibility-based TE：**As the impact or harm of technology on the future may be much greater than we imagine, the responsibility and obligation to protect environment and future human beings is the core of responsibility-based TE.

- **Utilitarian-based TE:** Utilitarian-based TE mainly concerns the consequences of technological development and attempts to avoid adverse consequences through the regulation and intervention in technological activities.

- **Value-based TE:** Value-based TE conducts ethical analysis of technological activities by focusing on the value or value conflict of technology.

Technology Ethics (TE): Approaches and Technologies

Three major approaches:

- **Responsibility-based TE：**As the impact or harm of technology on the future may be much greater than we imagine, the responsibility and obligation to protect environment and future human beings is the core of responsibility-based TE.
- **Utilitarian-based TE:** Utilitarian-based TE mainly concerns the consequences of technological development and attempts to avoid adverse consequences through the regulation and intervention in technological activities.
- **Value-based TE:** Value-based TE conducts ethical analysis of technological activities by focusing on the value or value conflict of technology.

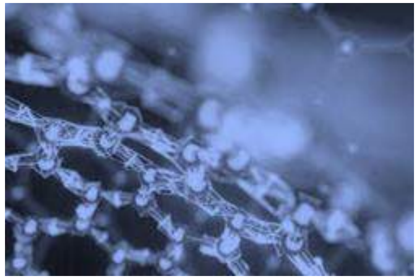
- Technologies

Technology Ethics (TE): Approaches and Technologies

Three major approaches:

- **Responsibility-based TE：** As the impact or harm of technology on the future may be much greater than we imagine, the responsibility and obligation to protect environment and future human beings is the core of responsibility-based TE.
- **Utilitarian-based TE:** Utilitarian-based TE mainly concerns the consequences of technological development and attempts to avoid adverse consequences through the regulation and intervention in technological activities.
- **Value-based TE:** Value-based TE conducts ethical analysis of technological activities by focusing on the value or value conflict of technology.

- **Technologies**


Ethics of Nanotechnology

Bioethics

Information Ethics

Nuclear Ethics

Ecological Ethics

……

Images are taken from Google Images

Technology-specific Ethical Issues

Fields	Main Ethical Issues
Ethics of Nanotechnology	• Safety Issues of Nanomaterials: Nanoparticles are "pervasive" and may enter human body and the environment during R&D, production, storage, transportation and thus causes harm to human health and ecology. • Ethical Issues in Nanomedicine: Is human enhancement through nanotechnology fair? • Issues arising from military applications of nanotechnologies
Bioethics	Bioethics is a broad and interdisciplinary field, which includes many issues: • Abortion; • Artificial life; • Eugenics; • Euthanasia; • Gene engineering; • Cloning; • Transsexual; • Suicide; ……
Nuclear Ethics	Nuclear ethics studies the ethical issues raised by the practice of nuclear technology, including nuclear warfare, nuclear deterrence, nuclear arms control, nuclear disarmament, and nuclear energy.
Ecological Ethics	There are many ethical decisions that human beings make with respect to the environment. For example: • Should humans continue to clear cut forests for the sake of human consumption? • Why should humans continue to propagate its species, and life itself? • Should humans continue to make gasoline-powered vehicles? • What environmental obligations do humans need to keep for future generations? • Is it right for humans to knowingly cause the extinction of a species for the convenience of humanity?
Information Ethics	**Information ethics focuses on the ethical issues arising in the following process:** **• Information production; • Information dissemination;** **• Information processing; • Information utilization.**

Artificial Intelligence (AI) Ethics

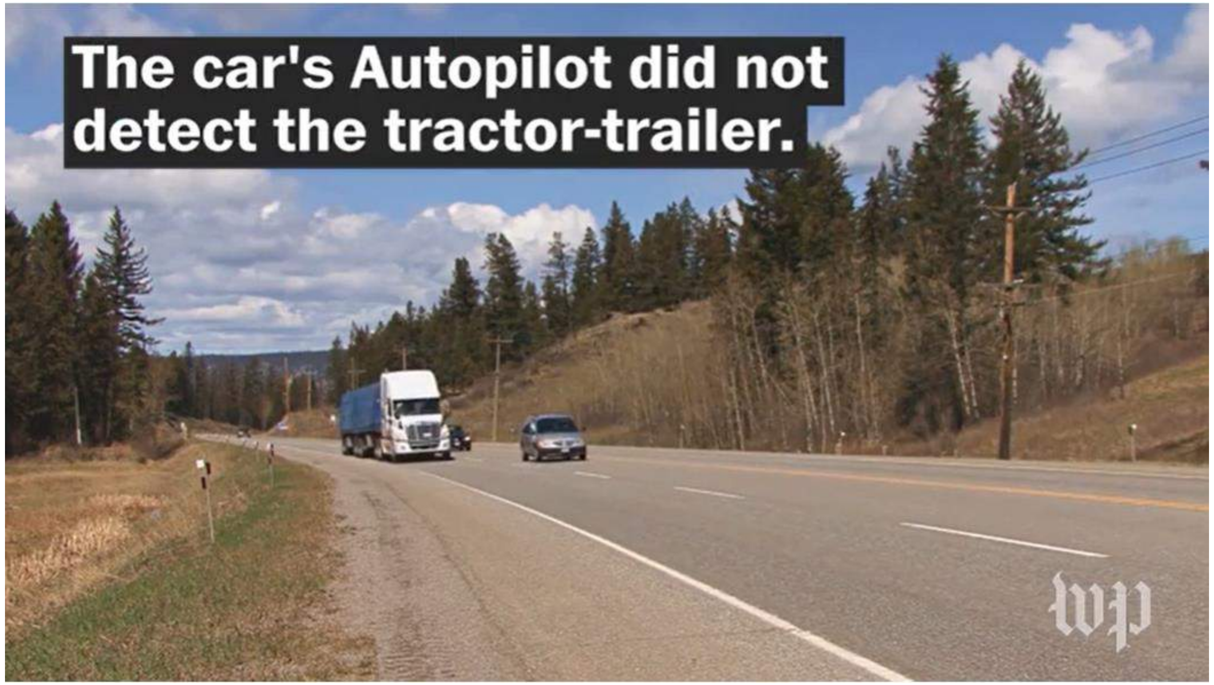
- AI Ethics (or Ethics of AI) is an emerging and interdisciplinary field concerned with addressing ethical issues of AI.

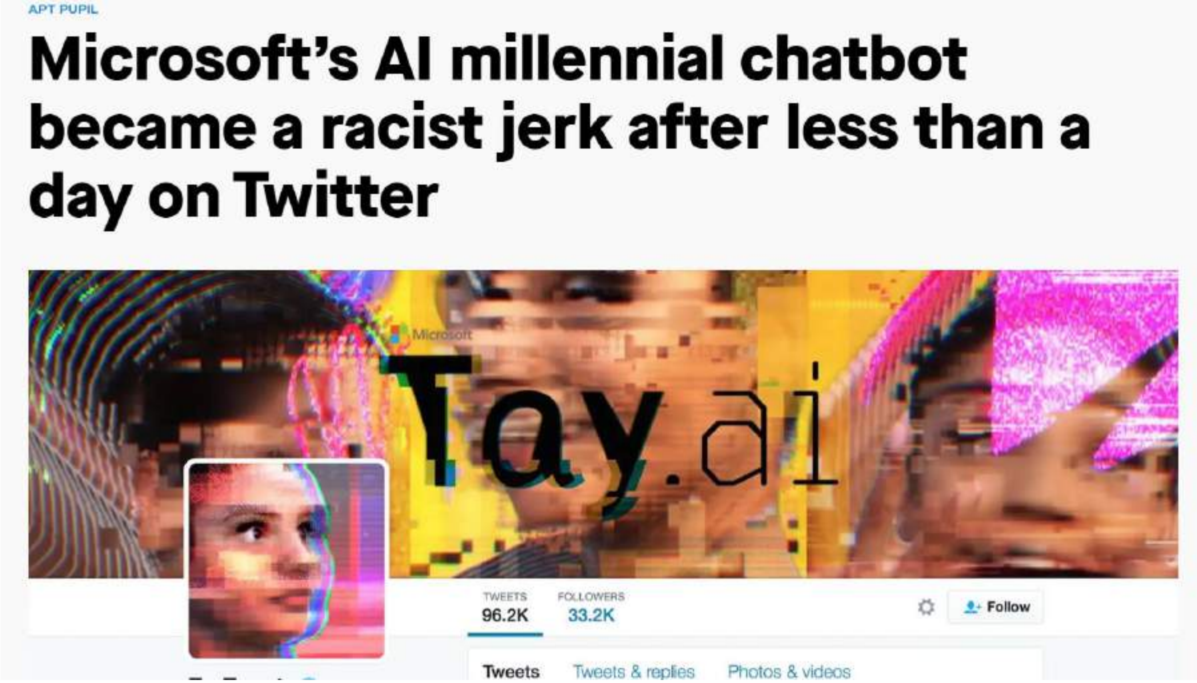
- The goal of AI ethics is to prevent and control the risks of AI technology and to promote the development of AI in the direction of improving human well-being and sustainable environments.

Ethical Issues in AI: Selected Examples

- In June 2016, a Tesla driver was killed in a collision in Florida with a tractor trailer while the vehicle was in "Autopilot" mode.

- Microsoft's AI chatting bot, Tay.ai ,was taken down because it became racist and sexist only less than a day after she joined Twitter.




APT PUPIL

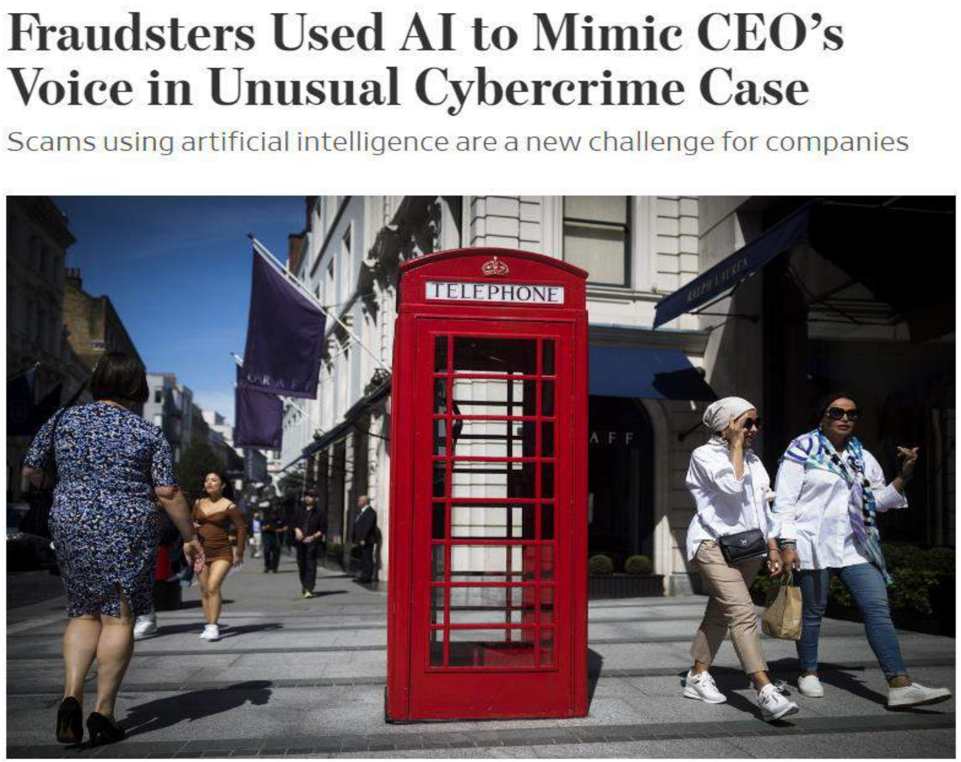
Microsoft's AI millennial chatbot became a racist jerk after less than a day on Twitter

https://www.washingtonpost.com/news/the-switch/wp/2016/06/30/tesla-owner-killed-in-fatal-crash-while-car-was-on-autopilot/

https://qz.com/646825/microsofts-ai-millennial-chatbot-became-a-racist-jerk-after-less-than-a-day-on-twitter

Ethical Issues in AI: Selected Examples

- Criminals used AI-based software to mimic a CEO's voice and demand a fraudulent transfer of $ 243,000.

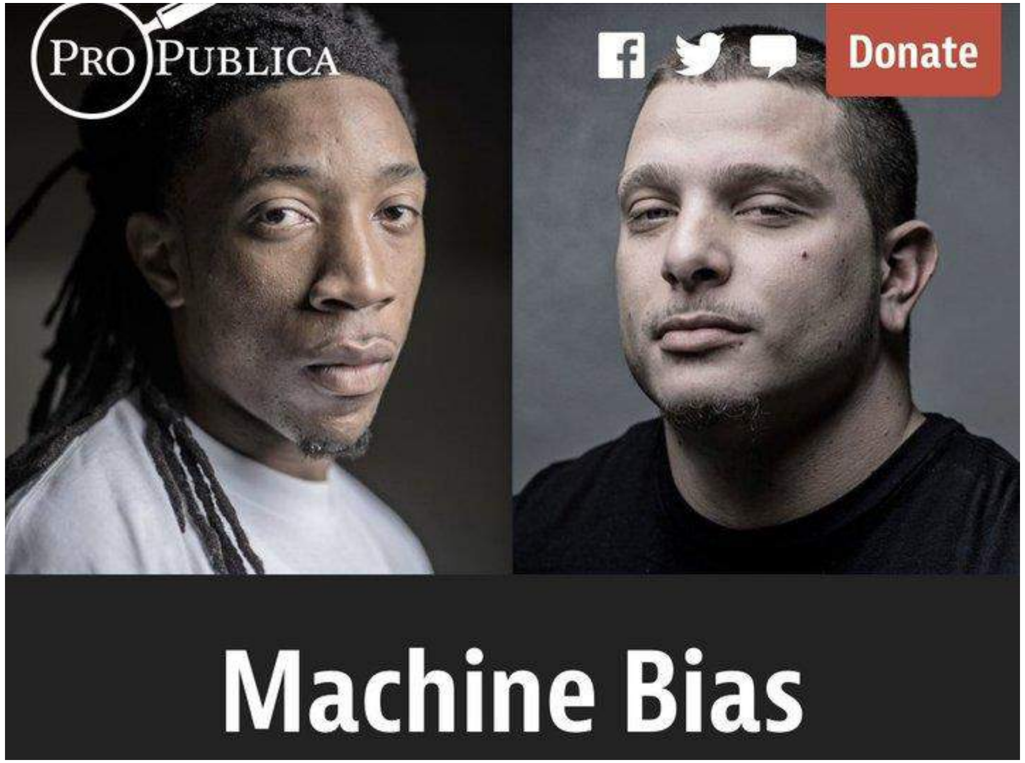


Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

- The software used across the country to predict future criminals was biased against blacks.

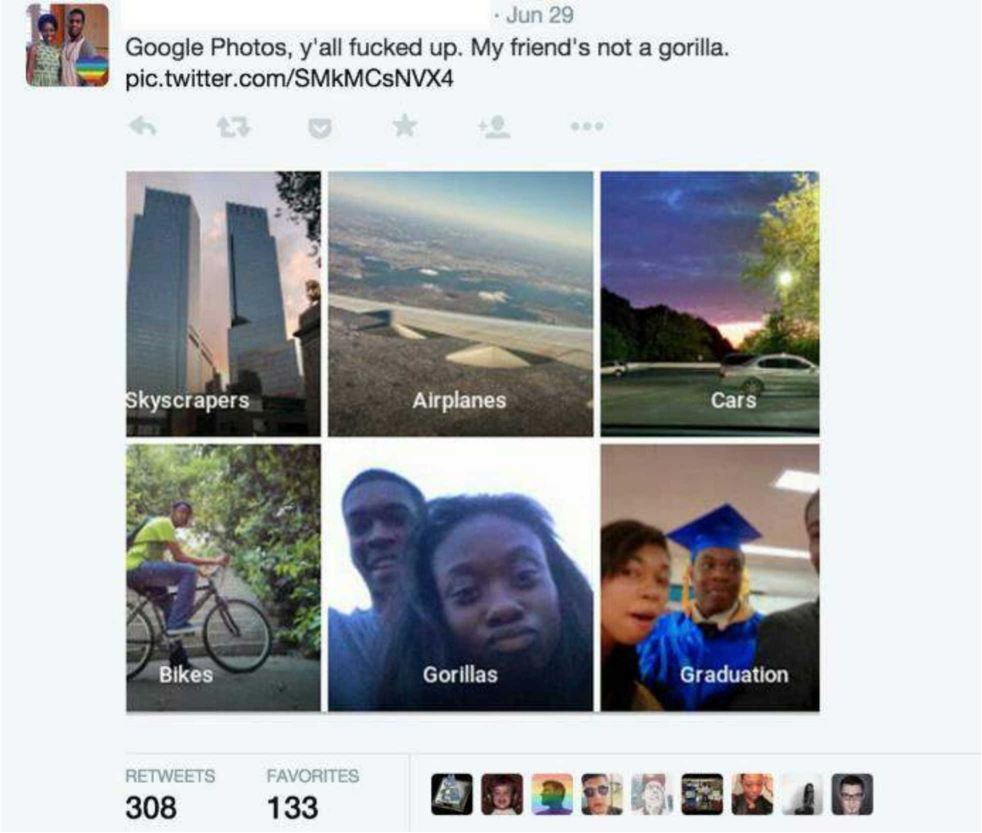


Machine Bias

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Ethical Issues in AI: Selected Examples

- **Google Photos app tags two black people as gorillas.**



· Jun 29

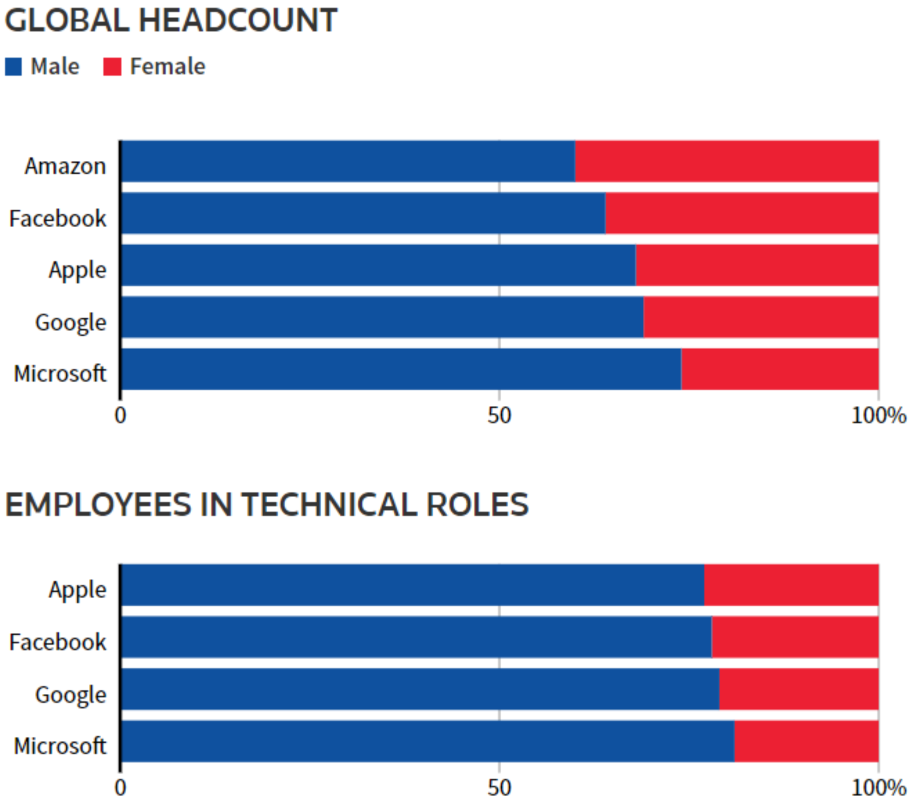
Google Photos, y'all fucked up. My friend's not a gorilla. pic.twitter.com/SMkMCsNVX4

RETWEETS 308

FAVORITES 133

https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas

- **Amazon scraps secret AI recruiting tool that showed bias against women.**



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Ethical Guidelines and Principles for AI: Overview

In recent years, ethical issues related to AI have aroused widespread awareness and consideration among researchers, developers, users, enterprises, and governments.

Ethical Guidelines and Principles for AI: Overview

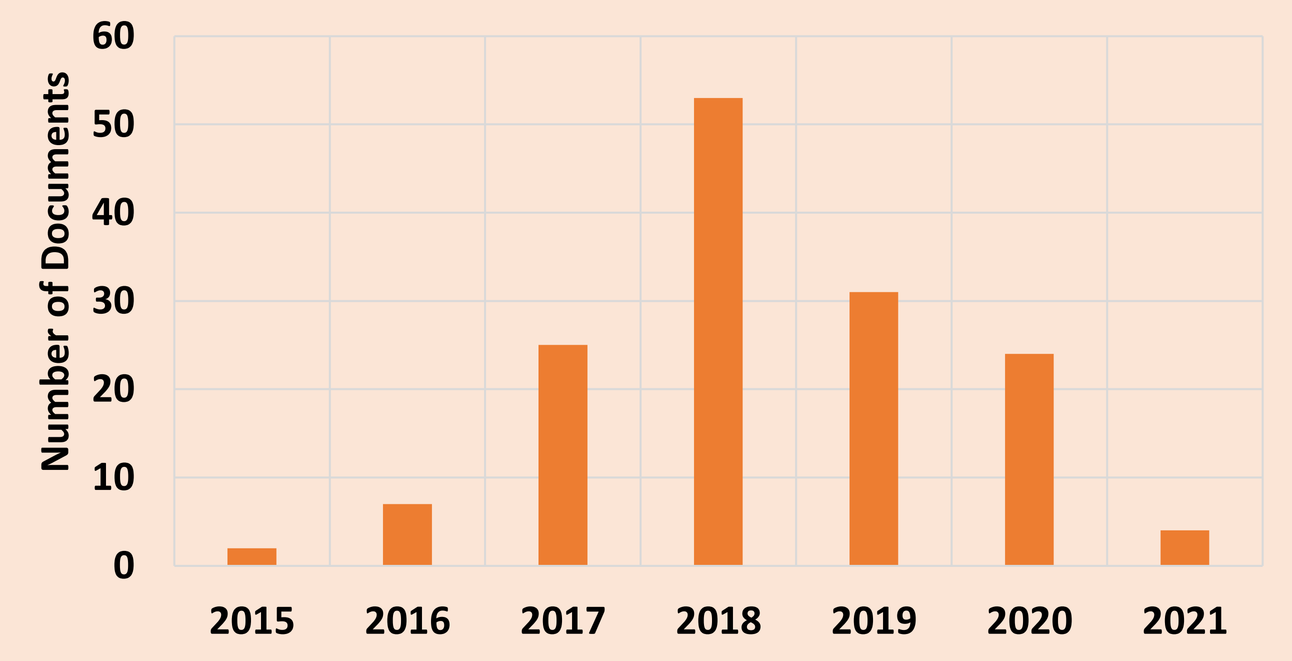
In recent years, ethical issues related to AI have aroused widespread awareness and consideration among researchers, developers, users, enterprises, and governments.

To guide their strategies in developing, adopting, and embracing AI technologies, many organizations, including governments, companies, academic associations, and other national/international organizations, have established ethical frameworks or guidelines for the planning, development, production and usage of AI technology.

Ethical Guidelines and Principles for AI: Overview

In recent years, ethical issues related to AI have aroused widespread awareness and consideration among researchers, developers, users, enterprises, and governments.

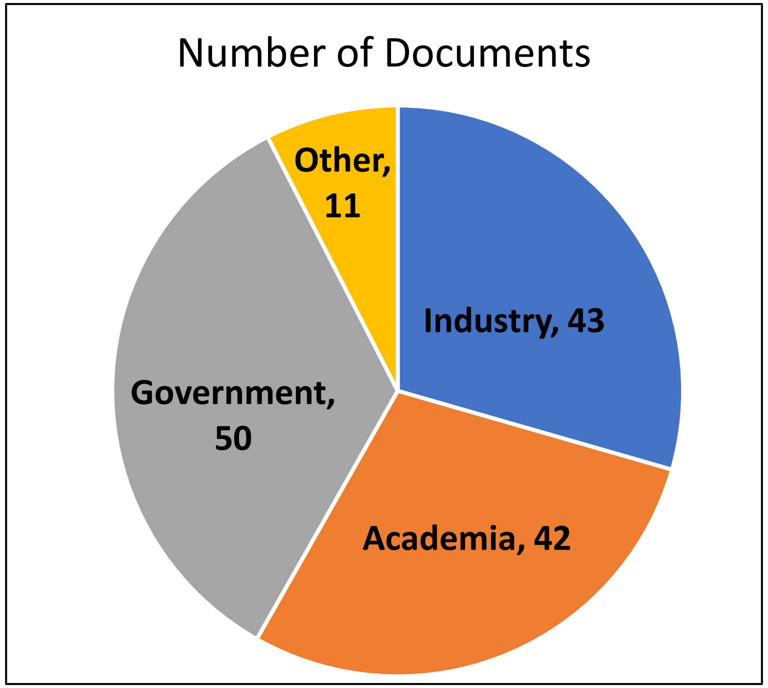
To guide their strategies in developing, adopting, and embracing AI technologies, many organizations, including governments, companies, academic associations, and other national/international organizations, have established ethical frameworks or guidelines for the planning, development, production and usage of AI technology.

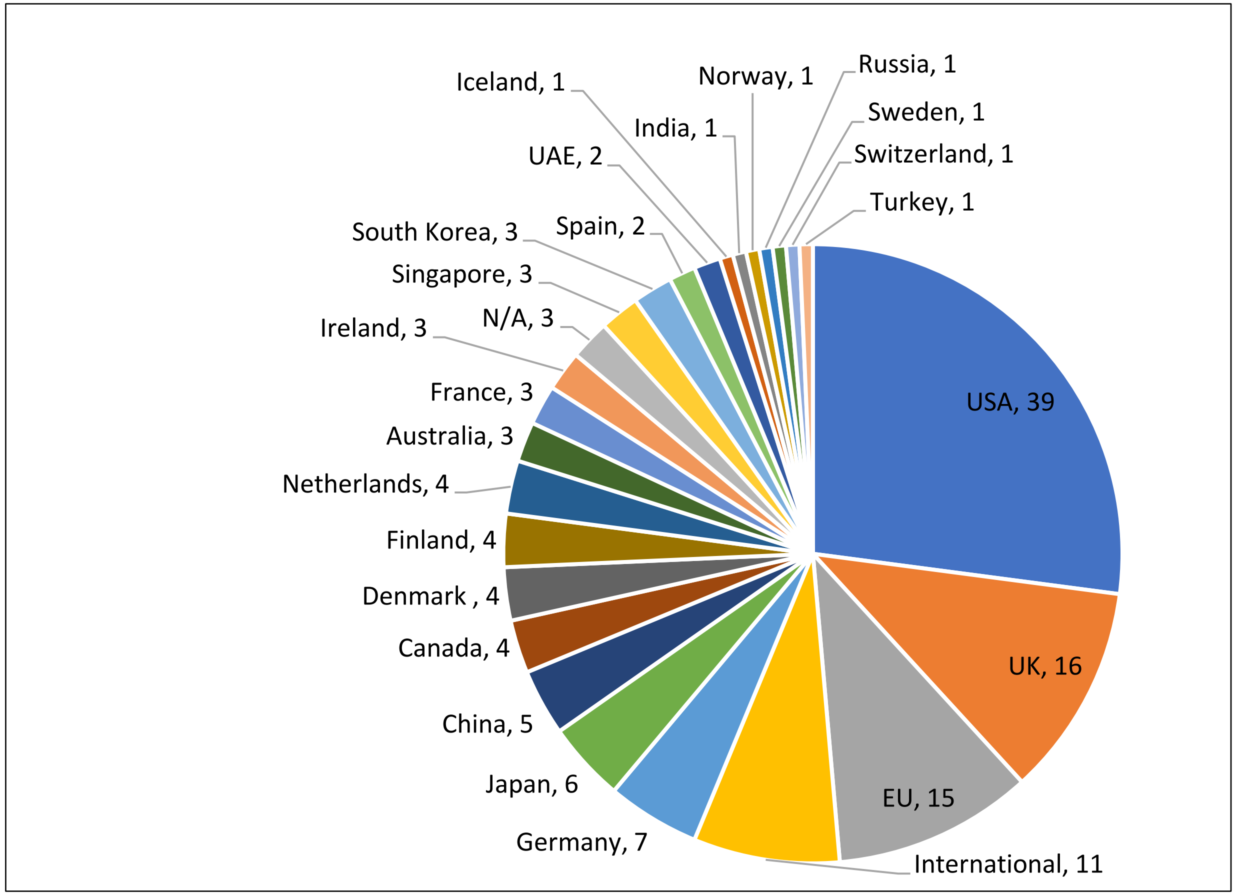


Year	2015	2016	2017	2018	2019	2020	2021
No. of Documents (146)	2	7	25	53	31	24	4

Ethical Guidelines and Principles for AI: Sources

Classification of AI ethical guidelines according to their sources.



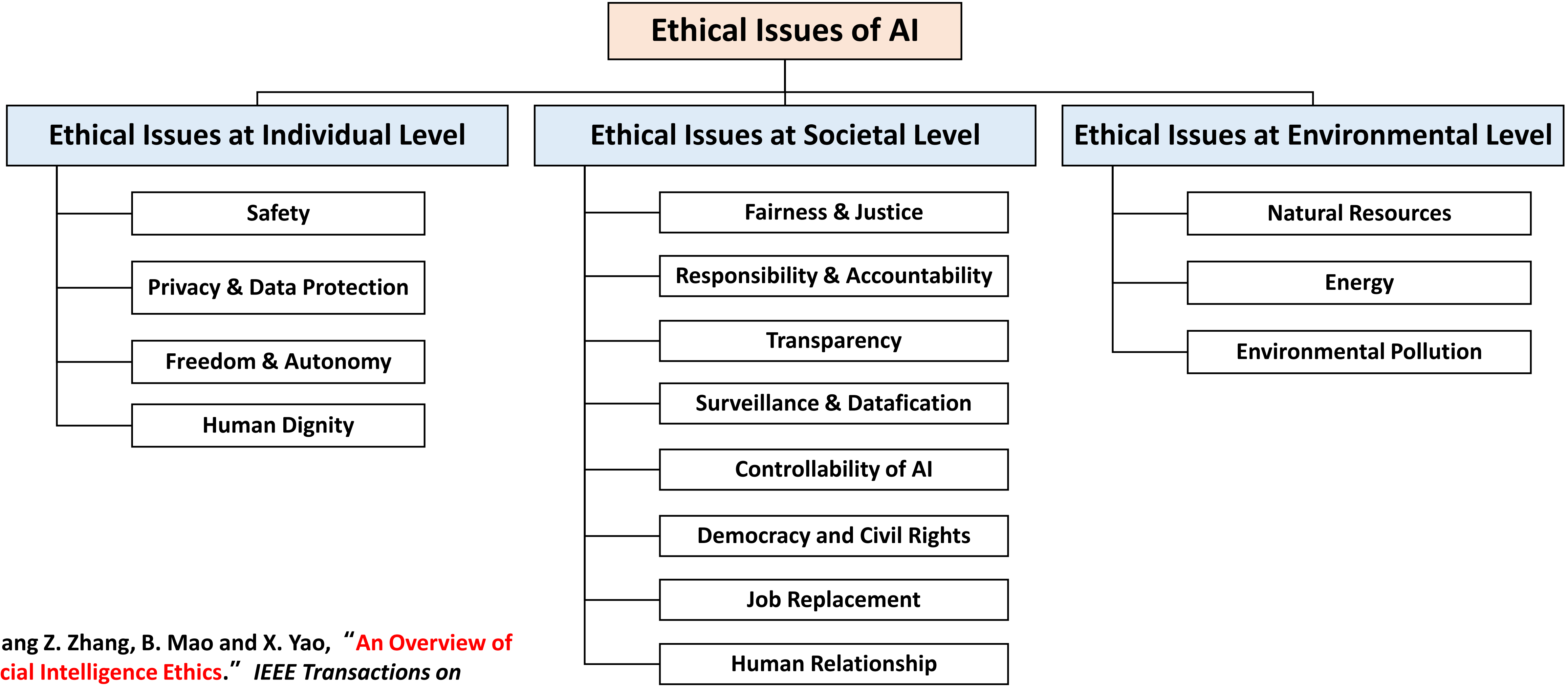


Ethical Principles Identified in Existing Guidelines

Ethical principle	Number of documents	Included codes/words
Transparency	107	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Fairness and Justice	107	Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-) discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Responsibility	100	Responsibility, accountability, liability, acting with integrity
Non-maleficence	81	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion, reliability, robustness
Privacy	73	Privacy, personal or private information, data protection,
Beneficence	41	Benefits, beneficence, well-being, peace, social good, common good, non-violence
Freedom and Autonomy	34	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment, human rights
Solidarity	20	Solidarity, social security, cohesion, inclusion, inclusiveness
Sustainability	19	Sustainability, environment, nature, energy, resources
Trust	14	Trust, trustworthiness, trustworthy
Dignity	13	Dignity

Ethical Issues of AI: Our Categorisation

Ethical issues of AI at individual, societal, and environmental levels.



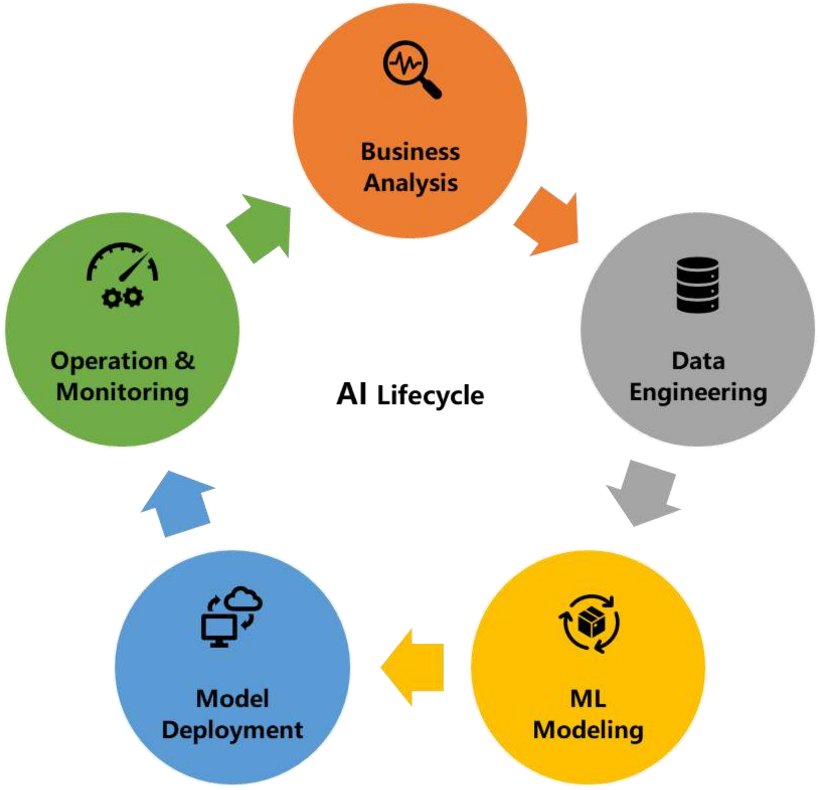
Ethical Issues of AI

- Ethical Issues at Individual Level
  - Safety
  - Privacy & Data Protection
  - Freedom & Autonomy
  - Human Dignity
- Ethical Issues at Societal Level
  - Fairness & Justice
  - Responsibility & Accountability
  - Transparency
  - Surveillance & Datafication
  - Controllability of AI
  - Democracy and Civil Rights
  - Job Replacement
  - Human Relationship
- Ethical Issues at Environmental Level
  - Natural Resources
  - Energy
  - Environmental Pollution

C. Huang Z. Zhang, B. Mao and X. Yao, “An Overview of Artificial Intelligence Ethics.” *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 4, pp. 799-819, Aug. 2023.

Ethical Issues of AI: AI System Life Cycle

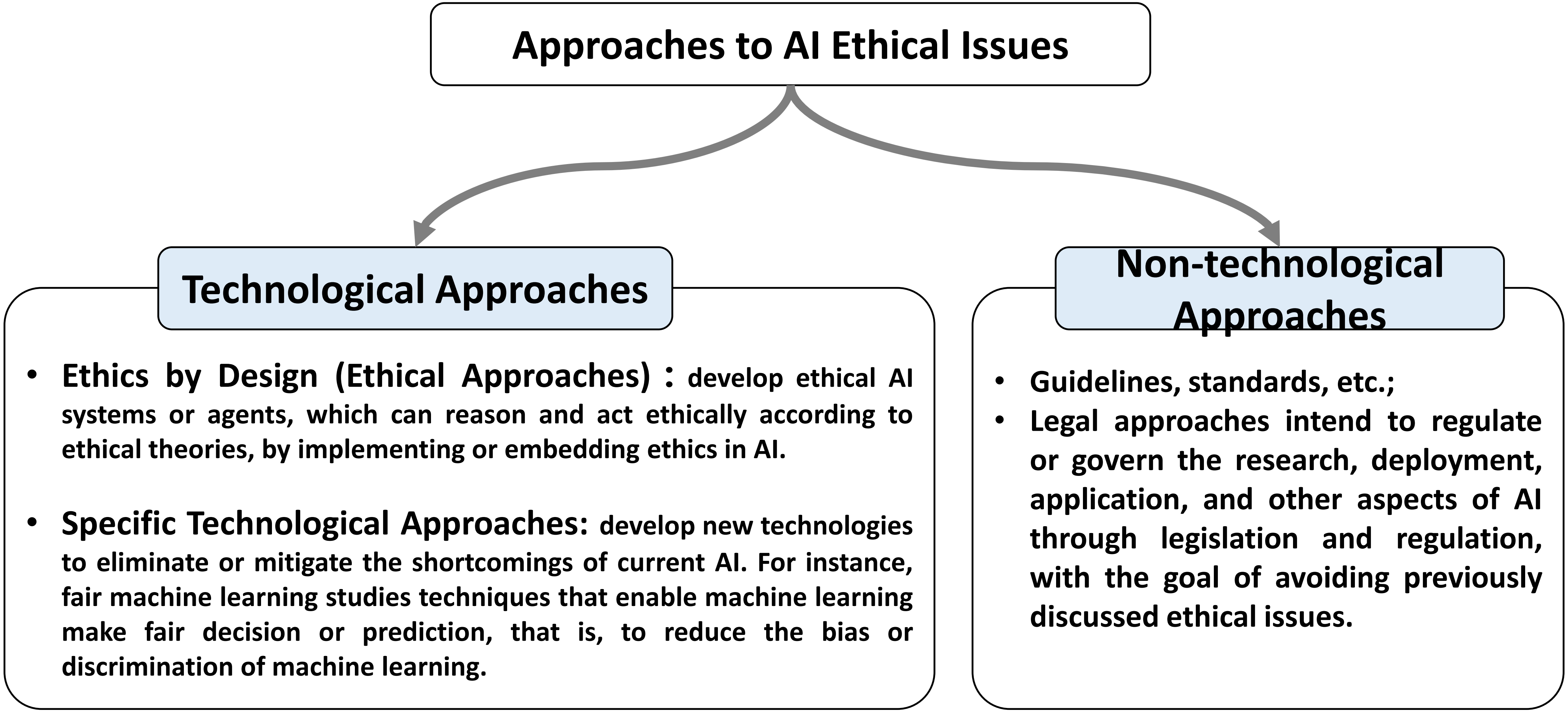
Key Ethical Issues Associated with the lifecycle of AI Systems, where we try to map ethical risks to all stages of the AI product life cycle.



Stage of AI Lifecycle	Ethical Considerations Exist Along the Stage
Business Analysis	Transparency, Fairness (Does the designed AI product includes any variables, features, processes that are unreasonable, morally objectionable, or unjustifiable?), Responsibility & Accountability, Democracy & Civil Rights, Sustainability
Data Engineering	Privacy (How to assure the data security and keep the private and sensitive information included in data set?), Transparency (How to make data collection procedures transparent to consumers?), Fairness (Are data properly representative, relevant, accurate, and generalizable ?), Democracy & Civil Rights (How will you enable end users to control use of their data?)
ML Modeling	Transparency (Does the decision or inference process of the model can be understood ?), Safety (Accuracy, Reliability, Security, and Robustness of the model), Fairness (Are the model outputs show disparate results on different groups of people ?)
Model Deployment	Privacy (Make sure that private information cannot be re-identified through the deployed model), Safety (How to ensure the safety of the deployed model , such malicious modification and attack ?)
Operation & Monitoring	Privacy (Privacy should be guaranteed during the operation & monitoring process), Fairness (Does the AI product has discriminatory or inequitable impacts on peoples they affect?), Democracy & Civil Rights (Do not infringe civil rights or the users)

Enough problems! How do we enhance AI ethics?

Approaches to Address Ethical Issues of AI



Approaches to AI Ethical Issues

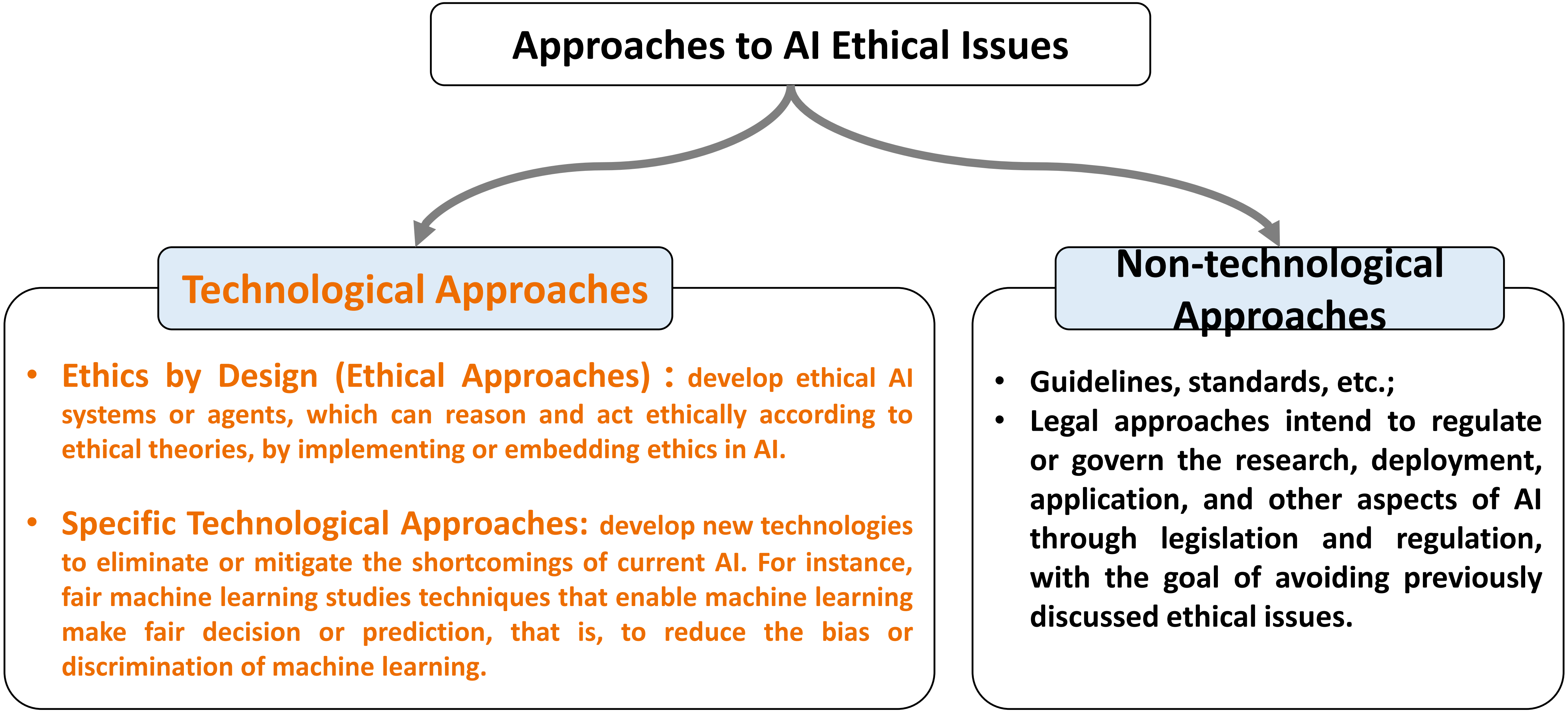
Technological Approaches

- **Ethics by Design (Ethical Approaches)**: develop ethical AI systems or agents, which can reason and act ethically according to ethical theories, by implementing or embedding ethics in AI.
- **Specific Technological Approaches:** develop new technologies to eliminate or mitigate the shortcomings of current AI. For instance, fair machine learning studies techniques that enable machine learning make fair decision or prediction, that is, to reduce the bias or discrimination of machine learning.

Non-technological Approaches

- Guidelines, standards, etc.;
- Legal approaches intend to regulate or govern the research, deployment, application, and other aspects of AI through legislation and regulation, with the goal of avoiding previously discussed ethical issues.

Approaches to Address Ethical Issues of AI



Approaches to AI Ethical Issues

Technological Approaches

- **Ethics by Design (Ethical Approaches) :** develop ethical AI systems or agents, which can reason and act ethically according to ethical theories, by implementing or embedding ethics in AI.
- **Specific Technological Approaches:** develop new technologies to eliminate or mitigate the shortcomings of current AI. For instance, fair machine learning studies techniques that enable machine learning make fair decision or prediction, that is, to reduce the bias or discrimination of machine learning.

Non-technological Approaches

- Guidelines, standards, etc.;
- Legal approaches intend to regulate or govern the research, deployment, application, and other aspects of AI through legislation and regulation, with the goal of avoiding previously discussed ethical issues.

Ethics by Design: Three Paradigms

The existing methodologies or approaches for implementing ethics in AI can be divided into three main types [7]: top-down approaches, bottom-up approaches, and hybrid approaches.

Top-Down

Top-down approaches, which infer individual decisions from general rules. The aim is to implement a given ethical theory within a computational framework and apply it to a particular case.

Bottom-Up

Bottom-up approaches, which infer general rules from individual cases. The aim is to provide the agent with sufficient observations of what others have done in similar situations and the means to aggregate these observations into a decision about what is ethically acceptable.

Hybrid

Hybrid approaches, which combine elements from bottom-up and top-down approaches in support of a careful moral reflection which is considered essential for ethical decision-making.

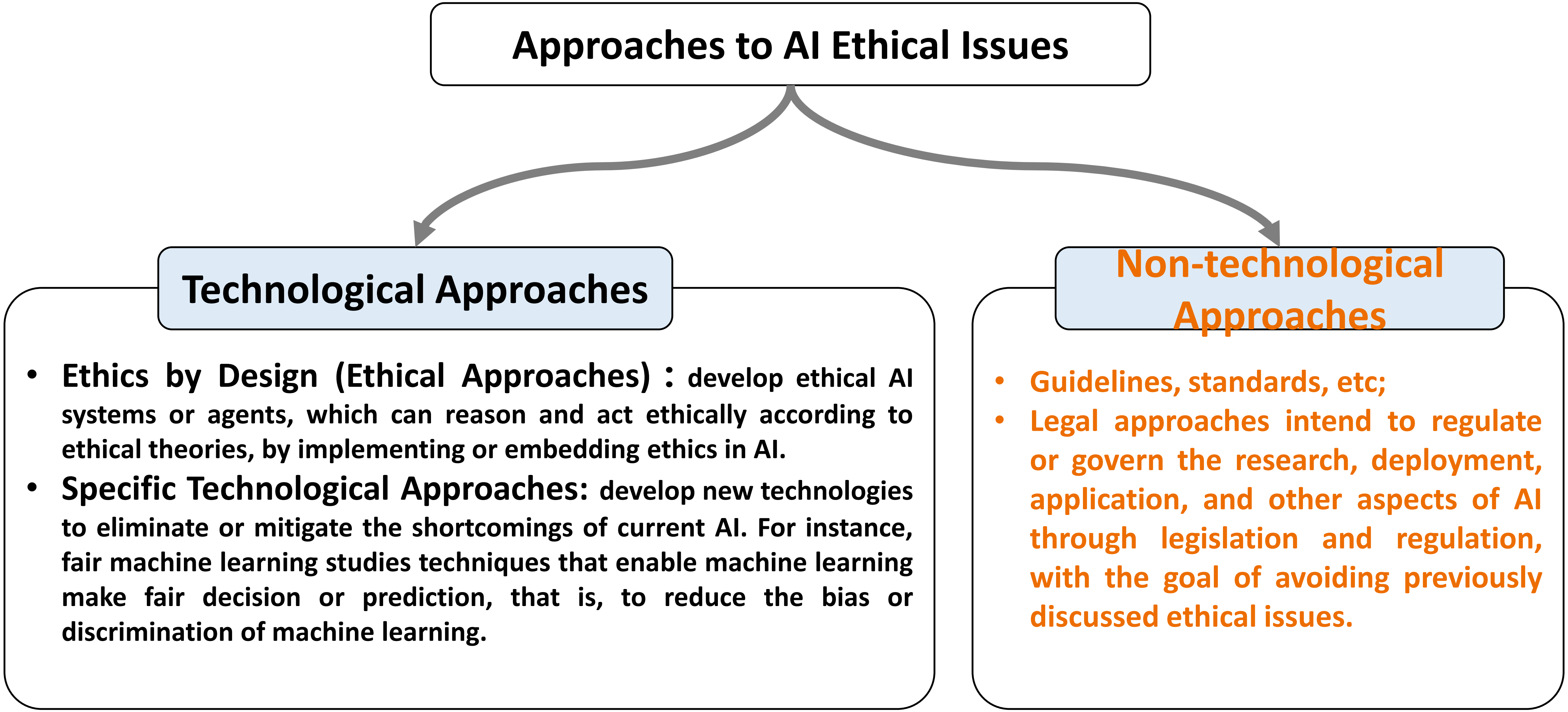
[7] C. Allen, I. Smit, and W. Wallach, "Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches," *Ethics Inf Technol*, vol. 7, no. 3, pp. 149–155, 2005

Specific Technological Approaches

The exiting work mainly focuses on a few major issues and principles. Other issues and principles are less studied.

Principle	Representative Research Topics or Direction
Transparency	Explainable AI (XAI) or Interpretable AI
Fairness & Justice	Fair AI
Non-maleficence	Safe AI Secure AI Robust AI
Responsibility & Accountability	Responsible AI
Privacy	Confidential Computing Differential Privacy Federated or Distributed Learning

Approaches to Address Ethical Issues of AI



Approaches to AI Ethical Issues

Technological Approaches

- **Ethics by Design (Ethical Approaches)：**develop ethical AI systems or agents, which can reason and act ethically according to ethical theories, by implementing or embedding ethics in AI.
- **Specific Technological Approaches:** develop new technologies to eliminate or mitigate the shortcomings of current AI. For instance, fair machine learning studies techniques that enable machine learning make fair decision or prediction, that is, to reduce the bias or discrimination of machine learning.

Non-technological Approaches

- Guidelines, standards, etc;
- Legal approaches intend to regulate or govern the research, deployment, application, and other aspects of AI through legislation and regulation, with the goal of avoiding previously discussed ethical issues.

Non-technological Approaches: Legislation & Regulation

At the regional, national and international levels, some relevant laws and regulations have been established by governments and organizations to govern the development and applications of AI.

2016	European Union	General Data Protection Regulation (GDPR)
2017	USA	Safely Ensuring Lives Future Deployment and Research in Vehicle Evolution Act
2018	Brazil	General Data Protection Law
2020	China	Personal Privacy Protection Act《个人信息保护法（草案）》 Data Security Act 《数据安全法（草案）》
2021	European Union	Artificial Intelligence (AI) Act
2021	Shenzhen, China	Shenzhen AI Industry Promotion Regulation《深圳经济特区人工智能产业促进条例（草案）》

Outline

1. From Ethics to Artificial Intelligence (AI) Ethics

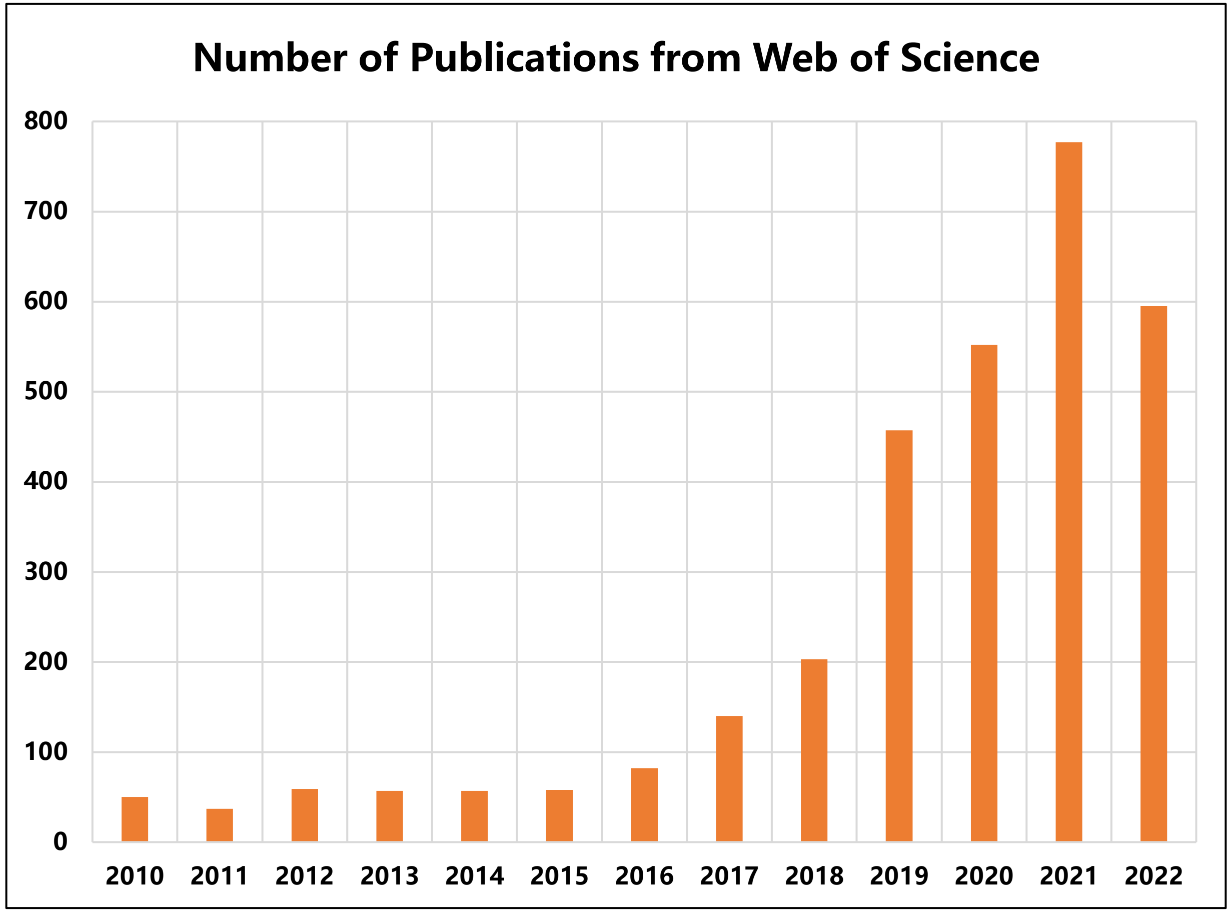
2. Fairer ML Through Multi-Objective Evolutionary Learning

3. Multi-objective Feature Attribution Explanation (MOFAE)

4. Regulating Machine Learning Behaviours Through Evolutionary Constrained Learning

5. Concluding Remarks

Fair Machine Learning Is a Hot Research Topic



Search on Web of Science by using keys of “fairness and bias in artificial intelligence” or “algorithmic bias” or “algorithmic fairness” or “fairness-aware machine learning” or “fairness in machine learning”. (Accessed on Nov. 3, 2022)

Many Metrics Have Been Proposed to Measure Fairness

There are over 20 fairness metrics proposed so far, which can be divided into 5 main categories [1]：

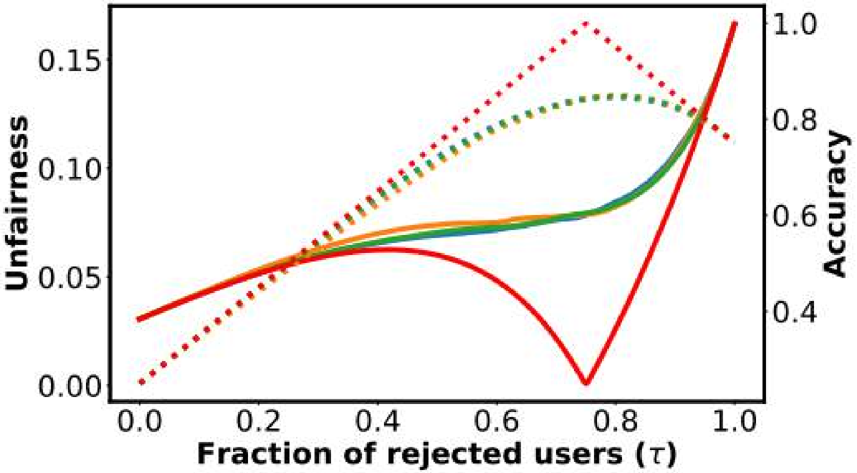
1. **Metrics based on predicted outcome**
2. **Metrics based on predicted and actual outcomes**
3. **Metrics based on predicted probabilities and actual outcome**
4. **Metrics based on similarity**
5. **Metrics based on causal reasoning**

[1] S. Verma and J. Rubin, “Fairness definitions explained,” in 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), IEEE, 2018, pp. 1–7.

However, Fair Machine Learning Is Hard

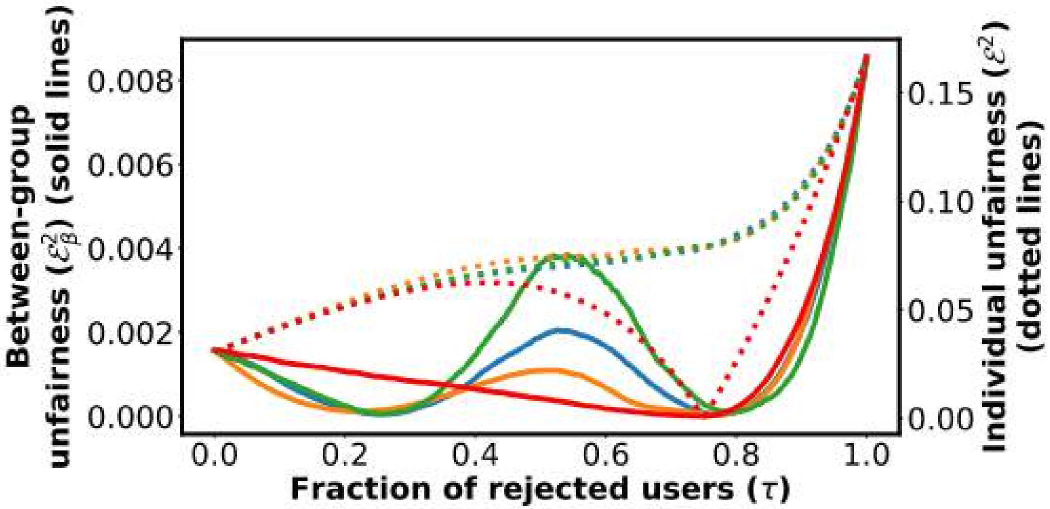
There are two inherent conflicts*:* (1) between model accuracy and fairness, and (2) among different fairness metrics.

(a) Conflict between accuracy and fairness



dotted line: unfairness
solid line: accuracy

(b) Conflict among multiple fairness metrics



dotted line: individual unfairness
solid line: group unfairness

Image source: Figure 4 from [17] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, and M. B. Zafarl, “A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices,” in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2239–2248

How Can We Find the Best Trade-off?

- We would end up with an accurate unfair model or an inaccurate fair model. Neither is desirable in practice.
- To make things worse, what’s fair according to one fairness metric would be unfair according to a different fairness metric.
- Key research question: How can we make machine learning fairer according to different fairness metrics without sacrificing accuracy too much?
    - Putting everything into a lose function with many terms balanced by hyper-parameters?

Evolutionary Computation Comes to Rescue: Multi-objective Fairer Machine Learning

Key Idea: Treating model accuracy and different fairness metrics as separate objectives in multi-objective learning [1].

Two studies were carried out：

Study 1

1. **Simultaneously** optimize several fairness measures without sacrificing accuracy significantly
2. Provide a group of **diverse** models

Study 2

3. Improve all fairness measures including those **not used** in model training
4. Generate an **ensemble** model combined from base models to balance accuracy and multiple fairness measures

[1] Q. Zhang, J. Liu, Z. Zhang, J. Wen, B. Mao and X. Yao, "Mitigating Unfairness via Evolutionary Multi-objective Ensemble Learning," in *IEEE Transactions on Evolutionary Computation*, vol. 27, no. 4, pp. 848-862, Aug. 2023, doi: 10.1109/TEVC.2022.3209544.

Does the Idea Work?

- Yes, it worked very well in comparison with the state-of-the-art when evaluated according to accuracy and multiple fairness metrics.
- The learned model even performed well according to metrics that were not used during training.
- We even provide a toolbox for practitioners to check and improve fairness of their machine learning models.

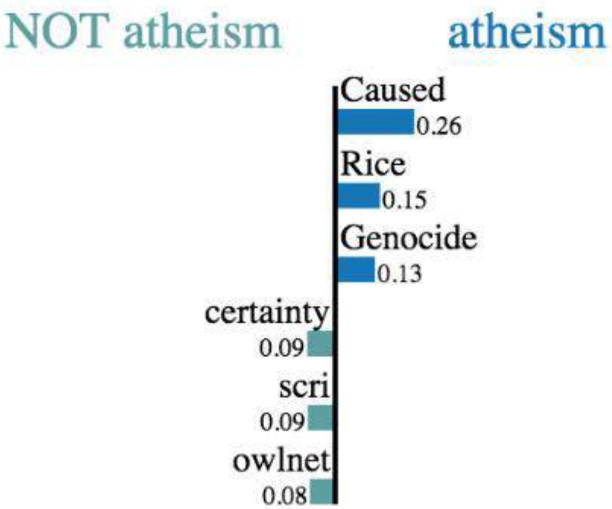
- Q. Zhang, J. Liu, Z. Zhang, J. Wen, B. Mao and X. Yao, "Mitigating Unfairness via Evolutionary Multi-objective Ensemble Learning," in *IEEE Transactions on Evolutionary Computation*, vol. 27, no. 4, pp. 848-862, Aug. 2023, doi: 10.1109/TEVC.2022.3209544.
- B. Yuan, S. Gui, Q. Zhang, Z. Wang, J. Wen, B. Mao, J. Liu and X. Yao, "FairerML: An Extensible Platform for Analysing, Visualising, and Mitigating Biases in Machine Learning," in *IEEE Computational Intelligence Magazine*, vol. 19, no. 2, pp. 129-141, May 2024, doi: 10.1109/MCI.2024.3364430.

Outline

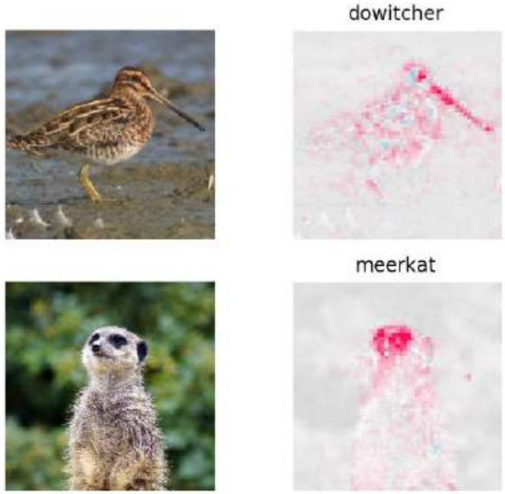
1. From Ethics to Artificial Intelligence (AI) Ethics
2. Fairer ML Through Multi-Objective Evolutionary Learning
3. Multi-objective Feature Attribution Explanation (MOFAE)
4. Regulating Machine Learning Behaviours Through Evolutionary Constrained Learning
5. Concluding Remarks

What Is Feature Attribution Explanation in XAI？

Local **feature attribution explanation (FAE) describes how much each input feature contributes to the output of a model *for a given data point*.**



FAE explains a **tabular** data point [1]



FAE explains an **image** data point [2]

[1] Ribeiro M T , Singh S , Guestrin C . "Why Should I Trust You?": Explaining the Predictions of Any Classifier[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. 2016.
[2]Lundberg S M, Lee S I. A unified approach to interpreting model predictions[C]//Proceedings of the 31st international conference on neural information processing systems. 2017: 4768-4777.

How to Evaluate Different Explanation Methods?

Various evaluation metrics have been proposed to assess the interpretation quality of FAE methods, such as:

- **Faithfulness [1]:** $u_F(f,g;\boldsymbol{x}) = \underset{S\in\binom{[d]}{|S|}}{corr}\left(\sum_{i\in S} g(f,\boldsymbol{x})_i, f(\boldsymbol{x}) - f(\boldsymbol{x}_{[\boldsymbol{x}_s=\bar{\boldsymbol{x}}_s]})\right)$

- **Sensitivity [1]:** $u_A(f,g,\boldsymbol{x}) = \frac{1}{|N_r|}\sum_{\boldsymbol{z}\in N_r}\frac{D\left(g(f,\boldsymbol{x}),g(f,\boldsymbol{z})\right)}{\rho(\boldsymbol{x},\boldsymbol{z})}$

- **Complexity [1]:** $u_C(f,g;\boldsymbol{x}) = -\sum_{i=1}^{d}\frac{|g(f,\boldsymbol{x})_i|}{\sum_{j\in[d]}|g(f,\boldsymbol{x})_j|}\ln\left(\frac{|g(f,\boldsymbol{x})_i|}{\sum_{j\in[d]}|g(f,\boldsymbol{x})_j|}\right)$

However, few existing FAE methods consider multiple metrics at the same time.

[1] Bhatt U，Weller A，Moura J . Evaluating and Aggregating Feature-based Model Explanations[C]// Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence IJCAI-PRICAI-20. 2020.

Consequences of Considering Only a Single Metric

- **We might end up with a very truthful explanation model, but too complex to understand. Or, we might end up with a simple model, but not very truthful to the original one.**

- **Neither is what we want. We need a reasonable trade-off between the two.**

- **Different stakeholders actually need different trade-off.**

- **Key research question: How can we satisfy such different requirements from different stakeholders?**

Evolutionary Computation Comes to Rescue: Multi-Objective Feature Attribution Explanation (MOFAE)

Key Idea: Treating each evaluation metric as a separate objective so that we can learn a set of diverse explanation models for different stakeholders.

Z. Wang, C. Huang, Y. Li and X. Yao, “Multi-objective Feature Attribution Explanation For Explainable Machine Learning,” *ACM Transactions on Evolutionary Learning and Optimization,* Volume 4, Issue 1, Article No. 2, pp. 1–32. February 2024. https://doi.org/10.1145/3617380.

Does the Idea Work?

Yes, it worked well. We have answered the following three research questions in our study.

Q1. Do faithfulness, sensitivity, and complexity conflict with each other?

Q2. Can MOFAE simultaneously optimize these conflicting metrics and be competitive against existing state-of-the-art FAE methods?

Q3. Can our method find a set of explainable models (i.e., explanations) with different trade-offs among the objectives (i.e., metrics)?

Z. Wang, C. Huang, Y. Li and X. Yao, “Multi-objective Feature Attribution Explanation For Explainable Machine Learning,” ACM Transactions on Evolutionary Learning and Optimization, Volume 4, Issue 1, Article No. 2, pp. 1–32. February 2024. https://doi.org/10.1145/3617380.

Outline

1. From Ethics to Artificial Intelligence (AI) Ethics
2. Fairer ML Through Multi-Objective Evolutionary Learning
3. Multi-objective Feature Attribution Explanation (MOFAE)
4. Regulating Machine Learning Behaviours Through Evolutionary Constrained Learning
5. Concluding Remarks

Machine Learning with Constraints

- **How do we ensure that a learned model meets given requirements, e.g., safety requirements of an autonomous vehicle?**
- **Problem formulation: Finding a model trained on $\boldsymbol{D}$ that has the minimal loss on $\boldsymbol{D'}$ and can satisfy the behavioral constraint $\boldsymbol{g_i(\cdot)}$ with a high probability $\boldsymbol{1-\delta_i}$.**
- **Key question: How to learn such a model?**

Evolutionary Computation Comes to Rescue: Evolutionary Constrained Learning (ECL) [2]

- **We can use the Stochastic Ranking Evolutionary Strategy (SRES) [1] to learn the model.**
  - **Our evolutionary constrained learning framework uses SRES as the training algorithm.**
- **Advantages of Evolutionary Constrained Learning (ECL) using SRES:**
  - **Differentiability of constraints and loss function is not required.**
  - **Penalty coefficient is not required.**

[1] T. P. Runarsson and X. Yao, "Stochastic ranking for constrained evolutionary optimization," *IEEE Transactions on Evolutionary Computation*, vol. 4, no. 3, pp. 284–294, 2000.

[2] C. Huang, Z. Zhang, B. Mao and X. Yao, "Preventing Undesirable Behaviors of Neural Networks via Evolutionary Constrained Learning," *Proc. of the 2022 International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy, 2022, pp. 1-7, doi: 10.1109/IJCNN55064.2022.9891926.

Does the Idea Work?

Yes. Considering *both* behavioral regulation and model accuracy, our proposed ECL approach can achieve better outcomes than others.

C. Huang, Z. Zhang, B. Mao and X. Yao, "Preventing Undesirable Behaviors of Neural Networks via Evolutionary Constrained Learning," *Proc. of the 2022 International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy, 2022, pp. 1-7, doi: 10.1109/IJCNN55064.2022.9891926.

Outline

1. From Ethics to Artificial Intelligence (AI) Ethics

2. Fairer ML Through Multi-Objective Evolutionary Learning

3. Multi-objective Feature Attribution Explanation (MOFAE)

4. Regulating Machine Learning Behaviours Through Evolutionary Constrained Learning

5. Concluding Remarks

Conclusion Remarks

- **AI ethics is a very important interdisciplinary research area.**
- **Many AI ethical issues are inherently multi-objective. Multi-objective evolutionary learning provides a natural approach to address them.**
- **There are numerous other AI ethical issues and solution approaches that have not been covered by this tutorial.**

Thank you for your attention!

Discussion Questions

- **Fairness**
  - How to deal with potential conflicts among different fairness metrics?
  - How should we make tradeoff? On what basis?
- **Explainability (WWWW)**
  - Who will explain what to whom for what purposes?
  - How to evaluate explainability? By whom? Why?
- **Safety**
  - How to ensure safety for data-driven AI?
  - Is it possible to guarantee safety for a purely data-driven approach?
  - Who is to provide such a guarantee?
  - What does safety mean exactly? Are there different dimensions of safety?