ADDRESSING AI ETHICS THROUGH CODIFICATION

ANDREY KULESHOV, CFA, MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY (A.KULESHOV@PHYSTECH.AI)

ANDREY IGNATIEV, EXPERT OF THE AIGO GROUP, OECD

ANNA ABRAMOVA, PHD, MOSCOW STATE INSTITUTE FOR INTERNATIONAL RELATIONS (ANNA.VL.ABRAMOVA@GMAIL.COM)

GRIGORY MARSHALKO, TECHNICAL COMMITTEE FOR STANDARDISATION "CRYPTOGRAPHY AND SECURITY MECHANISMS"

TECHNOLOGICAL INNOVATION AND ETHICS

- Most true innovations exist in a legal "gray zone", limited ability to prove that innovation "did nothing wrong"
- Society expects compliance with its values and moral principles in general, from any person, company or technology
- Ethics matters for any innovative technology; if it's to have a future, it has to be "ethical"
- AI systems do not have intrinsic ethical content, any ethical behavior has to be imported
- Ethics is largely abstract; technology is concrete. Codification of ethical behavior in creating and deploying AI system is proposed to bridge the gap

CURRENT UNDERSTANDING OF ETHICS BY INDUSTRY

- Aim to minimize risks created by innovation uncertainties
- System of "soft" law and self-regulation (Google FT article Feb 2020 calls for AI self-regulation)
- By the end 2019, 84 papers published on ethical principles for the industry
- Key themes: transparency, fairness and impartiality, non-harm, responsibility, privacy, well-being, freedom and autonomy, trust, dignity, sustainability and solidarity

ETHICS AND TRUSTWORTHINESS

- Trusted AI expect to meet a few criteria:
  - compliance with the law (compliance with all regulatory requirements);
  - compliance with ethics;
  - technical and functional reliability and security.
- Ethics fills the responsibility gap in situations where legal and technical regulations do not keep pace with the development of actual technologies and legislation
- Trustworthy AI must meet some kind of ethical standards

ETHICS AND BIAS

01

Bias, like ethics is based on human values and is hard to define for AI in technical terms

02

Ethical AI system may not show bias. However, the understanding of bias changes:

- With time, as societies change;
- Across geographies

03

Basic principles to prevent bias:

- Minimize cognitive biases;
- Look for bias-free and representative data
- Test for algorithmic biases

ETHICS IN AI, NOT ETHICS OF AI

- AI at the current state of development is neutral to ethical expectations of society, any ethical behaviour of AI systems is imported by human actors involved in design, production, deployment and application of a specific system using AI
- Ethics has no natural place in AI, ethics comes into AI practice through human actors
- Code of ethics for human actors makes sure that they all share a degree of responsibility for ethical behaviour of AI system

CODIFICATION 101 – ASIMOV LAWS OF ROBOTICS

- 1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- 2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
- 3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.
- Asimov laws of robotics assume that the robot can perform the role of a moral agent. At the current level of technology this provides little foundation for practical steps in securing ethical behaviour of AI systems

COMMON GENERAL ETHICAL CODES

- Informed consent:
  - prior, free, explicit and informed consent of the person
  - information must be adequate, provided in an understandable form
  - Applies to personal data in full
- Precautionary principle
  - ethical principle because it expresses the moral responsibility of decision-makers
  - agent implementing the action to assess the degree of danger (risk, threat), carefully weigh and prevent the worst-case scenarios, or those scenarios where there is no clear limit to risk
  - Practiacal benchmark for ethical behaviour
- Social responsibility: ISO 26000

PROFESSIONAL ETHICS

- The need for ethical rules is dictated by considerations of public trust in professional institutions (financial, medical, media, etc.) - a task that is substantially identical to the task of ensuring confidence in AI systems.
- Ethical issues emerge when specialist makes decisions that have significant consequences for other people in an environment of asymmetric and incomplete information.
  - decisions are made based on highly specialized knowledge;
  - decisions have serious consequences for people who do not have the knowledge to assess the quality of decisions;
  - there are likely cases when the validity of actions or decisions cannot be clearly assessed due to uncertain and compelling external circumstances, that is, the specialist makes decisions (i) based on incomplete information, and (ii) in a compromise between the speed of decision-making and the completeness of information
- Codes of professional ethics explain the principles and methods of specialists' work in ambiguous situations that have important, sometimes critical consequences and can carry significant risks
- Codes of professional ethics describe the behaviour of specialists, practicing in a particular ares where ethical issues arise. The agent is always a human, and the code is specific to the practical application.

CODE OF ETHICS IN IT

- Invariant set of moral attitudes **for IT specialists**, summarised as follows:
  - Respectful general attitude;
  - Personal/institutional qualities, such as conscientiousness, honesty and positive attitude, competence and efficiency;
  - Promotion of information privacy and data integrity;
  - Production and flow of information;
  - Attitude towards regulations.

CFA INSTITUTE CODE OF ETHICS

- Key characteristics of ethical behaviour **by a professional practicing in finance**:
  - (I) professionalism;
  - (II) integrity;
  - (III) obligations to clients;
  - (IV) obligations to employers;
  - (V) analysis, recommendations and actions;
  - (VI) conflict of interest;
  - (VII) informing about compliance with the code.

PRACTICAL ETHICS IN AI: BASIC APPROACH

1) Acting in the interests of society
2) Knowledge and respect of the law
3) Respect for citizens and society
4) Acting responsibly and with integrity
5) Act with professionalism and competence
6) Strengthen public trust in AI

ACTING IN THE INTEREST OF SOCIETY

- promote the achievement of the Sustainable Development Goals [28];
- preserve spiritual and cultural values, promote creative development of people and their cognitive abilities; promote the preservation of diversity, identity, socio-cultural traditions and foundations of various nations, ethnic groups and social groups;
- identify dangerous and sensitive areas where decision-making should not be delegated to AI systems;
- practice the precautionary principle; identify solutions and processes that may pose a threat to society or the individual and involve a neutral third party or authorized official bodies in an independent assessment of the balance between progress and security.

KNOWLEDGE AND RESPECT OF THE LAW

- keep up to date with the laws and regulations applicable to the field of professional activity in designing, implementing and using AI;
- act within the law and observe applicable regulatory requirements and technical standards at all stages of the life cycle of AI systems;
- avoid and prevent any forms of discrimination and inequality associated with the use of AI systems;
- respect existing regional and international agreements.

RESPECT FOR CITIZENS AND SOCIETY

- respect the autonomy of a person, honour and dignity, individual (unique) qualities; avoid any kind of explicit or hidden purposeful influence on social groups and individuals, exclude micro-targeting, manipulation, imposition of behavioural models;
- prevent the use of AI systems to restrict citizens ' rights and freedoms, promote the implementation of the provisions of the Bill of Human Rights [29];
- do not take actions that threaten the life and health of citizens, do not act to the detriment of the welfare of society; do not undertake or facilitate operations with unpredictable risks to the person, society, property or assets; observe prudence and caution;
- ensure the privacy and security of personal data; promote up-to-date and socially secure data management policies, including access and data sharing;
- use high-quality, representative, unbiased data obtained legally from reliable sources and in compliance with the principle of informed consent and other applicable regulatory requirements;
- test AIS for bias using representative test datasets which represent the expectations of the society;
- prevent the transfer of moral choice authority to the AI system;
- human oversight: provide for the possibility of vetoing (cancelling) any actions and decisions of AI systems by a human (user).

ACTING RESPONSIBLY AND WITH INTEGRITY

- approach professional activities responsibly and conscientiously, with due diligence and attention to the impact of AI systems on society at every stage of the AI lifecycle;
- document significant issues related to the development and use of AI and its impact on people, society and the government, maintain an audit trail of the system and clearly define the responsibility of actors for decision-making;
- separate facts from opinions and conclusions in informing third parties about the development and use of AI systems and in evaluating their functioning.

ACT WITH PROFESSIONALISM AND COMPETENCE

- follow the code of conduct of the professional community;
- apply sound / scientific methods at all stages of the life cycle of AI systems;
- use AI systems consistently with their intended purpose and in accordance with the intended application task;
- maintain up-to-date professional knowledge and certifications appropriate to professional practice of developing and implementing AISs.

STRENGTHEN PUBLIC TRUST IN AI

- promote public awareness and knowledge about the technologies of AI and the results of their application; promote public integrity;
- provide users with complete and reliable information about the use of AI and about the functionality of AIS and its intended areas of application;
- contribute to equitable regional and international cooperation in the field of technology transfer and data exchange;
- compete on an equal and fair basis; strengthen the credibility of products and services based on technologies AI;
- develop complete and comprehensive descriptions and instructions for the technical operation of various AI systems;
- improve the means of protecting AI systems from cyber-attacks, failures, unintentional accidents and malfunctions;
- promote the availability of modern AI technologies, products and services based on AI among the general population, prevent the digital divide and conflict between social groups due to difference in access to modern AI systems;
- inform the public about ethical issues and decisions in AI systems, and submit important decisions to public discussion.

WAY FORWARD FROM BASIC APPROACHES

- practical standards of ethical behaviour targeted at specific sectors, industries and applications;
- international agreements in the field of AIS and their use to advance the digital economy, e-Commerce, data and technology transfer;
- implementation of strategic alliances and corporate agreements involving the development of AI technologies for the good of the society;
- in general, for the development of practical methodology in the development of soft law tools and AI regulation.

THANK YOU!

A.KULESHOV@PHYSTECH.AI