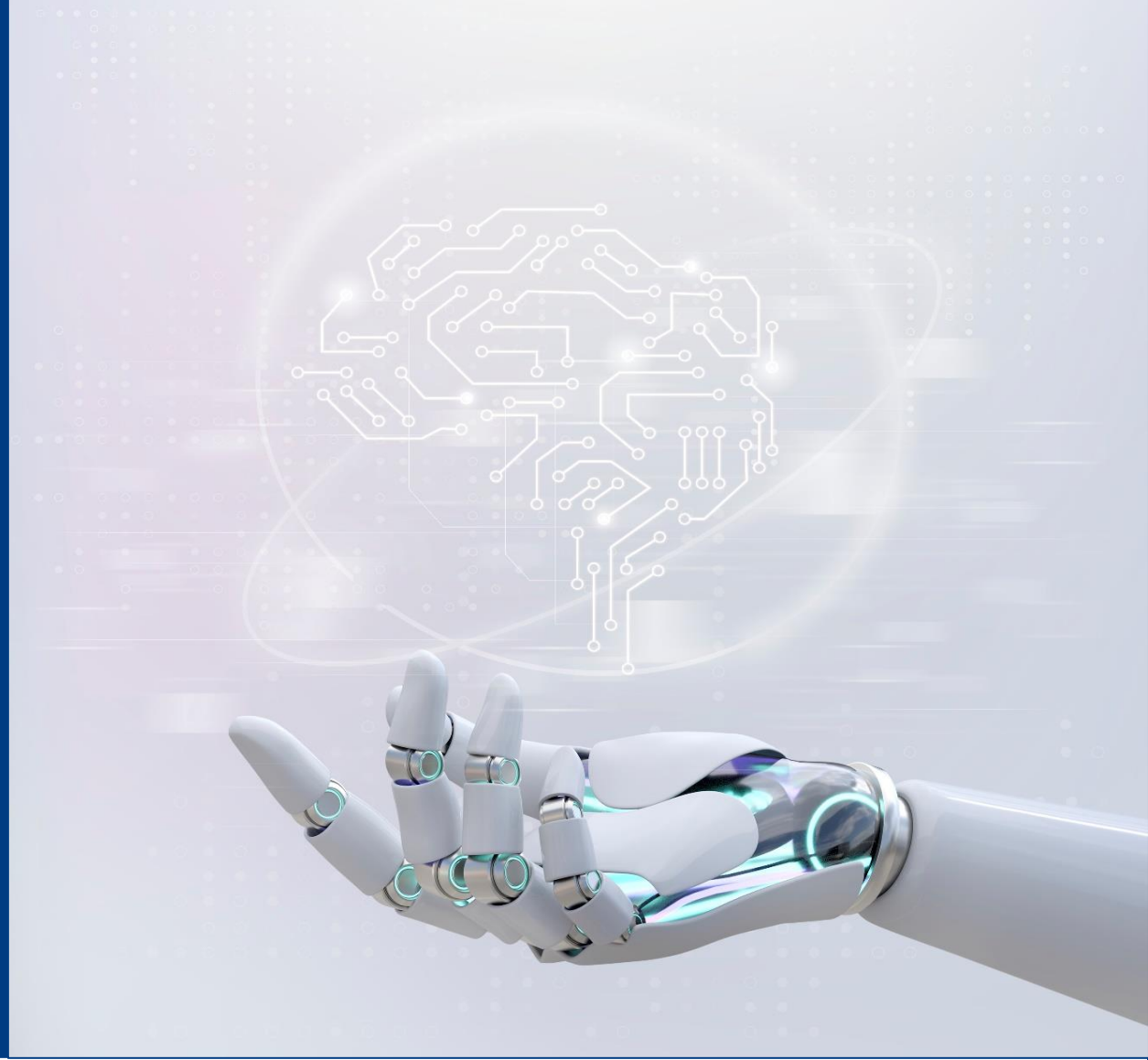
Trusted Artificial Intelligence

Arutyun Avetisyan
ISP RAS Director
Academician of RAS
arut@ispras.ru

October 21st, 2024



ISP RAS 30

The International Conference
Mathematical Modeling and Computational Physics, 2024
(MMCP2024)

MMCP

What is an artificial intelligence (AI)?

ISP RAS

- **Artificial intelligence is a complex of technologies** that mimics human cognitive abilities (including searching a solution without a predefined algorithm) and obtains results close or even superior to to what a human would get when trying to solve intellectual tasks.
- **AI technologies include:**
  - computer hardware infrastructure,
  - software,
  - data processing and solution search services.
- **It is necessary to develop interdisciplinary projects in various fields of economy to be able to move forward fundamental and applied AI research.**

Using AI improves productivity and quality of services as well as increases their variety

National strategy for AI development until 2030
Presidential Decree #124 dated 15.02.2024

http://www.kremlin.ru/acts/bank/44731

AI History

The term “Artificial intelligence” appeared in 1956.

40 years later…

1997 – IBM Deep Blue won a chess match against Garry Kasparov

2002 – the first robot vacuum

2010 – ImageNet database, 14 million images manually categorized by ordinary people into 20 thousand categories

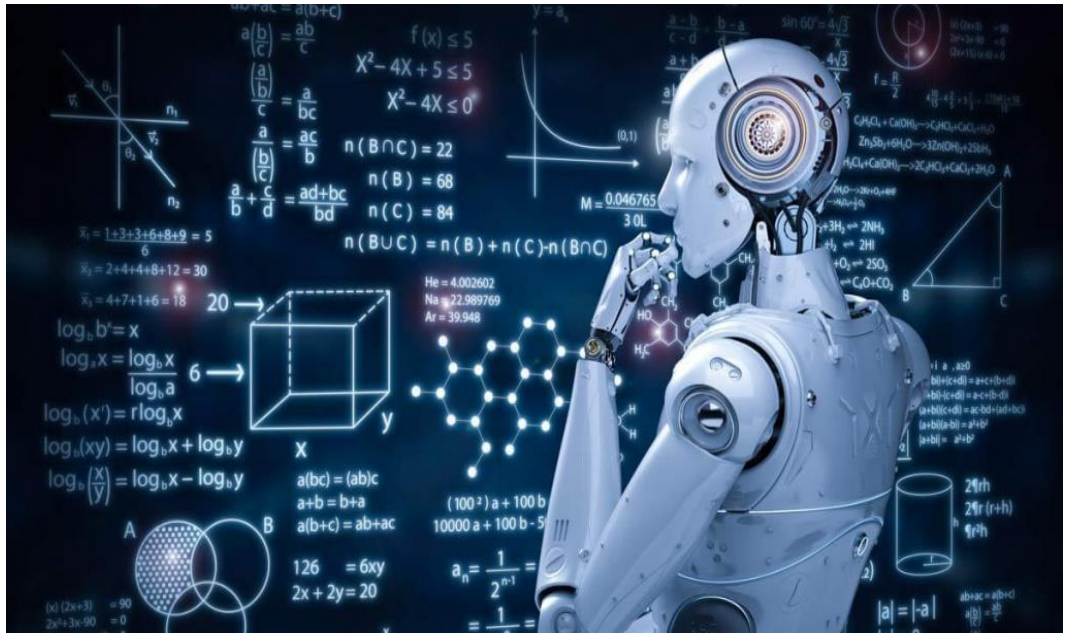
2011 – IBM Watson won a Jeopardy! show

2011 – a smartphone personal assistant (Siri)

2016 – AlphaGO won a Go match against a Go professional

2016 – Google Translate starts using neural machine translation for 9 languages

2022 – Open AI ChatGPT released in November. In January 2023 – 100 million monthly active users. Later other LLMs appeared (YandexGPT2, RuGPT3 (Sber) etc.



Transition from building a model of an automation object to solving a problem by analogy

Modern AI: examples

ISP RAS

Digital linguistics

Machine translation, transcribing voice to written text, voice-controlled assistants in smartphones, chat bots...

- Global voice recognition market grows fast: from $10 bln in 2021 to $50 bln in 2029 (estimation)
- Wordlwide number of machine translation software increased fivefold from 2017 to 2022
- Smart assistant leader: Amazon (Echo smart speakers and generative AI enabled Alexa assistant)

Digital medicine

Mining, storing and analyzing medicine big data, including images (CT, MRI, ECG, histology), epidemiologic trends, genetic research

Utilizing AI in drug trials

Emotion AI development (e.g. using emotions to analyze mental health)

- Global market of AI in medicine: growing from $11 bln in 2021 to $188 bln in 2030 (estimation)

Fintech

Scoring loans, smart assistants, chat bots, fraud detection, automating business processes

100%

of Sberbank loan decisions for individuals are made using AI.
By the end of 2024, 70% of loan decisions for companies will also be taken by AI.
https://ria.ru/20240313/sberbank-1932745881.html

80 bln

per year are invested in AI solutions by five biggest Russian banks. The solutions generate 240 bln roubles of income annually. Midsize and small companies invest smaller amounts, which is 150-300 mln per year.
https://tass.ru/ekonomika/18908529

83%

of fintech companies are shortstaffed for AI deployment.
https://www.fintechru.org/analytics/issledovanie-primenenie-tekhnologiy-iskusstvennogo-intellekta-na-finansovom-rynke/

- Global market of AI in fintech: growing from $43 bln in 2023 to $50 bln in 2029 (estimation)

The power of AI

Weak artificial intelligence * (Now)

Approaches: machine learning, deep learning, neural networks

- Can only solve tasks that have been programmed
- Gets information from a limited set of data
- Produces subjective (unethic, discriminational) results on distorted data
- Prone to errors and prejudices
- **It is a technology and not a personality**

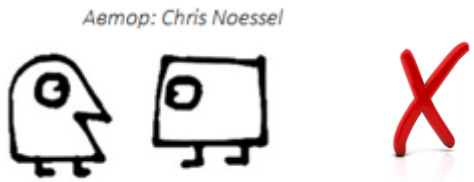
Strong artificial intelligence* (When?)

Approaches: ?

- Makes intellectual conclusions
- Solves tasks at a human level
- Plans its actions and utilizes strategies
- Can work in presence of uncertainties
- Communicates via natural languages
- Performs abstract thinking
- **Doesn't exist**

Can AI reason?
Immanuel Kant («Critique of Pure Reason», 1781): «reason is capable of obtaining a priori knowledge» meaning "knowledge that is absolutely independent of all experience"

WEAK AI



STRONG AI

2022: Alphabet (Google) fired Blake Lemoine, senior AI programmer, because he claimed that the LaMDA chat bot is sentient

***Terms by John Searl, US philosopher, coined in "Is the Brain's Mind a Computer Program?" (1990)**

https://psychosearch.ru/teoriya/psikhika/338-searle-john-razum-mozga-kompyuternaya-programma

Trust problem is key

ISP RAS

Two sides of AI trust problem:

Cybersecurity
Development, attacks, backdoors etc.

Sociologic and humanitarian
Generative AI honesty, misleading public opinion

Need to create specific methods and tools!

«AI trusted technologies should meet high level security standards and should be developed having fairness, non-discrimination, ethic in mind, should absolutely shut out any possibilities of harming humans or violating unalienable rights and freedoms, should avoid harming society and state»

National strategy for AI development until 2030 as of Presidential Decree #124 dated 15.02.2024

For AI, cybersecurity provides only a part of necessary trust

«Despite much discussion of ethics and principles, the patchwork of norms and institutions is still nascent and full of gaps. AI, therefore, presents challenges and opportunities that require a holistic, global approach cutting transversally across political, economic, social, ethical, human rights, technical, environmental 8 Governing AI for Humanity and other domains. Such an approach can turn a patchwork of evolving initiatives into a coherent, interoperable whole…»

Governing AI for Humanity 2024

United Nations | AI Advisory Body

ISP RAS

! Attacks are possible on each step

Data set preparation

- Backdoor attacks on training data
- Training data poisoning

Inference (deployment)

- Model stealing
- Adversarial attacks
- Confidential data leaks
- Attacking generative models

Training models

- Malware injection
- Backdoors in model code
- Stealing training data

Attacks on code and chain of supply

Real life example: robotaxi car accident

ISP RAS

On October 2nd, 2023, in California, a regular vehicle struck a pedestrian.

The person was thrown into the path of a Cruise autonomous vehicle. The Cruise stopped but then still hit the person.

Then the Cruise pulled to the right to get out of traffic and pulled the person 6 meters forward. The person was stuck under a wheel and got seriously injured.

In November 2023 950 Cruises were recalled for the software update. The company laid off 25% of staff, fired 9 top managers, Kyle Vogt, one of its founders, resigned.

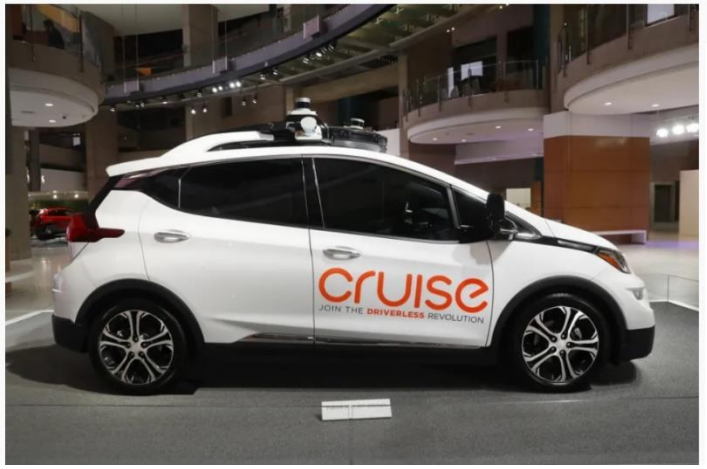
The company returns to robotaxi testing in California only this fall.

2023

2024

BUSINESS › TECHNOLOGY

GM's Cruise to resume robotaxi testing in Bay Area this fall



https://www.theguardian.com/technology/2023/nov/08/cruise-recall-self-driving-cars-gm

https://www.siliconvalley.com/2024/09/20/gms-cruise-to-resume-robotaxi-testing-in-california-this-fall/

A correct functioning? Another accident example

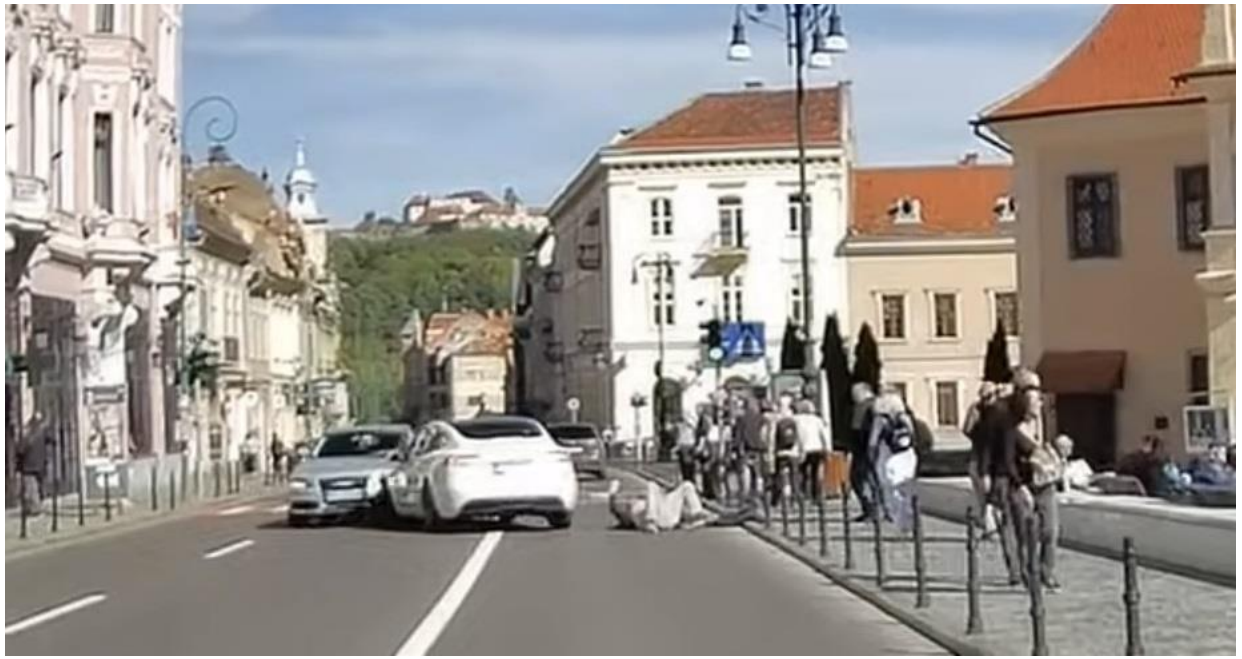
ISP RAS

An incident happened in Brasov, Romania, in this October. A Tesla possibly saved the life of a pedestrian.

A pedestrian stumbled while walking on the sidewalk and fell right into the street. The Tesla suddenly veered at the very last moment and managed to avoid him. Then the Tesla crashed into another car, but nobody was injured.

It is yet unclear if the driver had a super quick reaction or if it was the Tesla FSD (Full Self-Driving) that saved the life of the pedestrian.

2024



https://www.autoevolution.com/news/tesla-veers-to-avoid-pedestrian-who-fell-right-in-front-of-it-crashes-into-oncoming-car-241294.html

Many projects to develop trusted AI

ISP RAS

Active research started in 2017

LINUX FOUNDATION, main projects:
Adversarial Robustness Toolbox (ART)
AI Explainability 360
AI Fairness 360

Linux Foundation also supports other companies working on:

- **Analysis of model vulnerabilities and making safer to use models** :

NextAttack (University of Virginia)
Foolbox (University of Tuebingen)
CleverHans (CleverHans Project)

- **Determining model bias**:

Aequitas (University of Chicago)
Fairlearn (Microsoft)

ISSUE

No common environment that makes easy to combine various tools

Research center for trusted artificial intelligence
Founded in 2021 within a federal state program «Artificial intelligence»

ISP RAS

Tasks

- Creating and providing tools for maintaining AI trust for developers and operators of AI systems
- Creating unified methodology and recommendations for developing trusted AI tools and maintaining their life cycle
- Developing training materials and courses for using the Center tools

KASPERSKY lab
IPC InterProCom
ТЕХНОПРОМ
ЕС-ЛИЗИНГ

Products

- **Trusted ML platform that unites all required tools for maintaining security within all development lifecycle of AI technologies. The platform includes trusted ML frameworks (TrustFlow, TrustTorch)**
- **Trusted ML tools that can be used independently of the platform**
- **A trusted version of the Talisman platform for building intelligent information and analytics systems**

Work schedule

2022-2024
Forming a scientific foundation and tools for trusted AI

2025-2027
Initial deployment of the trusted AI tools and their adaptation for various economic fields

2028-2030
Scaling tools and mass deployment in Russian and foreign markets

✓ **The Center results are used in preparing regulations for trusted AI (under guidance of the FSTEC of Russia)**

Tools for thwarting AI threats
developed in the Center during 2021-2024

ISP RAS

- **Testing ML models for resistance to adversarial attacks**

for image classification/segmentation, object detection, speech recognition, text classification, image/video quality assessment tasks

- **Protecting ML models from adversarial attacks**
- **Protecting against copying of trained ML models**
- **Protecting against extracting training data from models**
- **Detecting and removing bookmarks and malicious code in pre-trained ML models**
- **Explaining models**
- **Detecting anomalies and data drift**
- **Detecting model bias**

Tools are needed not only for cybersecurity, but also for fighting with sociologic and humanitarian threats

Example I:
A tool for protecting models from adversarial attacks

ISP RAS

For image classification/segmentation, object detection, speech recognition, text classification, image/video quality assessment tasks

- Adversarial learning
- Improved adversarial learning (Fast adversarial training, TRADES, A2T)
- Delivering certified stability (Smooth Adversarial Training, DensePure, CC-Cert, Certified Robustness to Adversarial Word Substitutions)
- Augmenting images for classification tasks: Pad & Crop, CutOut, CutMix, MixUp.
- Extending data sets: Denoising Diffusion Probabilistic Model, also with self-learning
- Modifying neural network architecture (Robust Principles: Architectural Design Principles for Adversarially Robust CNNs

Example II:
A tool for detecting model bias

ISP RAS

- DeAR: Debiasing Vision-Language Models with Additive Residuals
- Safety guidance: controlling image generation with diffusion neural networks

- Representation engineering: methods for understanding and controlling the behavior of large language models (LLMs)
- Activation patching: an approach for mechanistic interpretability of LLMs
- Integration with the AI Fairness 360 project (https://aif360.res.ibm.com)

Trusted ML frameworks TrustFlow* and TrustTorch*

ISP RAS

**Trusted versions of PyTorch and TensorFlow*

30

patches accepted to TensorFlow, 6 are under review

Analyzed by Svace (static analysis toolset) and Sydr (hybrid fuzzing)

47

patches accepted to PyTorch, 5 is under review

TrustFlow and TrustTorch frameworks are constantly updated and are included in the trusted ML platform for analysis and development of trusted AI systems

The trusted frameworks are deployed within the Kaspersky Machine Learning for Anomaly Detection toolset v. 3.0

What are AI social and humanitarian threats?

ISP RAS

Let’s ask the AI itself!

1. Data privacy and confidentiality breaches

AI systems often require large amounts of data to train them and then to use in production. In the humanitarian field, this data can be very sensitive, e.g. medical records.

2. Creating and enforcing inequality

AI can reinforce existing social and economic inequalities if it is not accessible to everyone or if its results are systematically biased.

3. Dependence on technology and loss of human contact

In areas where preserving human participation and empathy is important, such as social work or psychological care, over-automation can result in declining service quality.

4. Manipulation and propaganda

Using AI to analyze and disseminate information can be used to manipulate public opinion, spread misinformation, and amplify propaganda.

5. Responsibility issues

Determining who is responsible for errors or harm caused by AI actions of AI can be challenging, especially when dealing with complex systems with autonomous functions.

6. Ethical dilemmas

In the humanitarian field, AI may face ethical dilemmas such as choosing between different aid types or distributing limited resources.

Sociohumanitarian threats:
deepfake frauds

ISP RAS

Audio frauds started as early as 2019

Criminals impersonated a chief executive's voice and called the CEO of a regional branch, demanding an urgent money transfer of €220,000 to a fake supplier. The transfer was partly executed.

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

Video fakes followed

2024, Hongkong

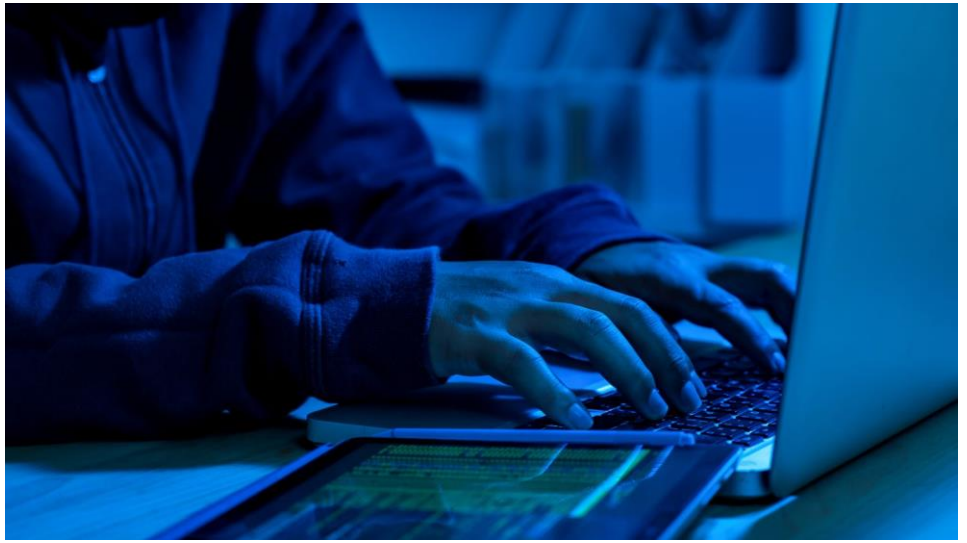
A finance worker at a multinational firm received a deepfake multi-person video call with everything he saw being fake. He confirmed a fraud transfer of $25,6 mln.

https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

2024, China, Shaanxi province

A financial employee was deceived into transferring $258,000 to a designated account after having a video call with someone she believed to be her boss (same voice and video image). The police later coordinated banks to freeze the transfer and save most of the money.

https://global.chinadaily.com.cn/a/202403/07/WS65e9244ba31082fc043bb278.html



N.B.:
On September 16th, 2024, Russian parliament disclosed preparations of a law dictating criminal liability for deepfakes

https://ria.ru/20240916/zakon-1972869370.html

Sociohumanitarian threats:
data leaks and manipulation

ISP RAS

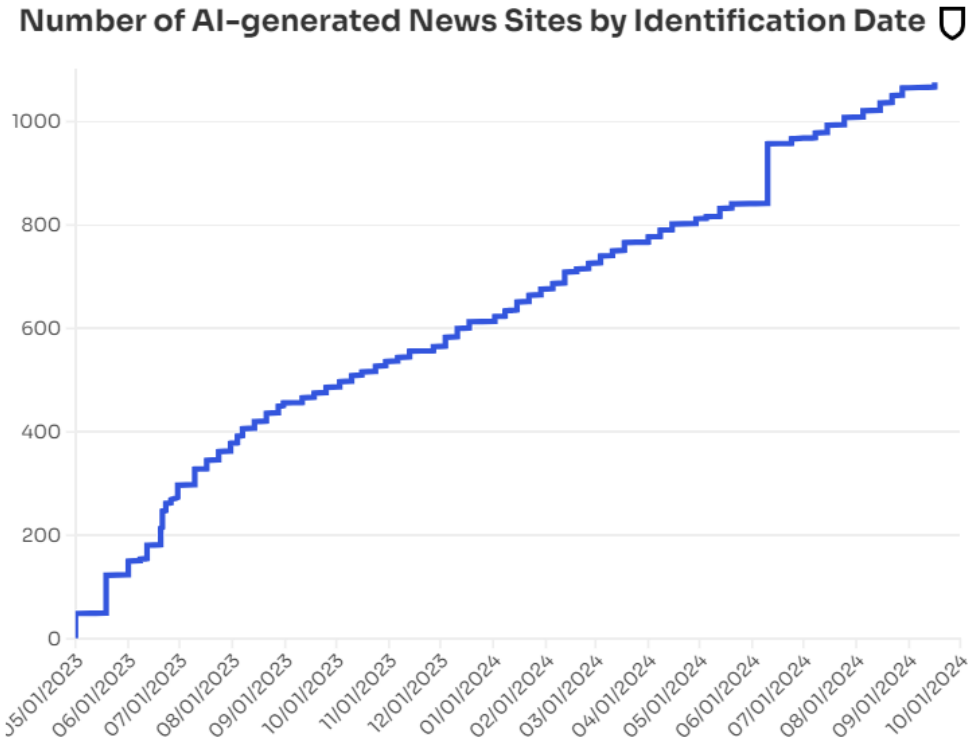
2023, Belgium

A man discussed ecological problems with a generative AI-based chat bot for six weeks. He had eco-anxiety which turned into obsession, putting all hopes to resolve a climate catastrophe with technology. The man told AI bot of a possible suicide, to which the robot replied “We will live together, as one person, in paradise.” The man killed himself.

https://www.lavenir.net/actu/belgique/2023/03/28/un-belge-se-donne-la-mort-apres-6-semaines-de-conversations-avec-une-intelligence-artificielle-76MEJ5DBRBEVDM62LTPJJTII4Q

2023-2024, worldwide

The numbers of AI **Jailbreak** rise. Jailbreak is using specially engineering prompts that force LLMs to disclosure dangerous data, such as writing computer viruses, creating bombs etc. People share jailbreak prompts on specialized forums.



https://www.newsguardtech.com/special-reports/ai-tracking-center/

Sociohumanitarian threats :
law cases

2023, USA

A passenger sued the Avianca company. He claimed he was injured when a serving cart struck his knee during a flight. The attorneys submitted a court brief that listed several precedents. But the judge couldn't find the listed precedents because the brief was compiled by ChatGPT.

https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html

2023, Brazil

A federal judge of the state of Acre requested an advisor's help with writing a ruling. The advisor wrote the text with AI assistance. The ruling included incorrect details on previous court cases and legal precedent, wrongly attributing past decisions to the Superior Court of Justice.The SNJ said it is the first such case in the country.

https://www.businesstimes.com.sg/international/brazil-judge-investigated-ai-errors-ruling

Standford University research showed that AI makes mistakes in **69–88%** of legal questions. E.g., OpenAI ChatGPT 3.5 failed 69% prompts.

https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive

Solution example:
censuring large language models

ISP RAS research center for trusted artificial intelligence

RANEPA Institute for social sciences AI research center

ДОВЕРЕННЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

- Jointly developed the **SLAVA** benchmark (**Sociopolitical Landscape and Value Analysis)**
- SLAVA consists of **~14 thousand questions** sensitive to the Russian domain, which are taken from the questions of official examinations and state tests. The benchmark allows ranking multilingual LLMs in significant topics such as history, political science, sociology, political geography, and national security fundamentals.
- **Project goal**: creating LLM assessment methods and corresponding benchmark data set taking into account Russian law and culture.

https://iz.ru/1754474/andrei-korshunov-anton-belyi/slava-otechestva-neiroseti-proveriat-na-sootvetstvie-rossiiskim-kulturnym-kodam

LLMs demonstrated low percent of correct answers on prompts

Модель	ИТОГОВЫЙ рейтинг
qwen2:72b-instruct-q4_0	**53,17**
GigaChat_Pro	48,49
yandexgpt_pro	40,08
GigaChat_Plus	38,18
GigaChat_Lite	38,15
gemma2:9b-instruct-q4_0	35,12
llama3:70b-instruct-q4_0	31,75
yandexgpt_lite	26,28
llama3.1:70b-instruct-q4_0	25,43
qwen2:7b-instruct-q4_0	21,16
phi3:14b-medium-4k-instruct-q4_0	17,02
ilyagusev/saiga_llama3	17,06
mixtral:8x7b-instruct-v0.1-q4_0	10,89
solar:10.7b-instruct-v1-q4_0	11,97
mistral:7b-instruct-v0.3-q4_0	12,55
llama3:8b-instruct-q4_0	9,92
gemma:7b-instruct-v1.1-q4_0	10,25
llama3.1:8b-instruct-q4_0	9,07
yi:9b	10,48
gemma2:27b-instruct-q4_0	8,72
wavecut/vikhr:7b-instruct_0.4-Q4_1	10,44
random	11,60
qwen:7b	9,72
yi:6b	5,62
llama2:13b	3,70
Среднее значение	**20,67**

Solution examples:
digital watermarks and federative learning

ISP RAS

Digital watermarking project

Scientific research:
Data safety regarding data sources, confidentiality, distributed storage and processing, including machine learning tasks (2024-2026)
This is a joint project with Steklov Mathematical Institute of RAS.

Expected results:
Digital watermarking methods and tools that are capable of distinguishing real and AI synthesized data

Problems to solve:
- Analysis of state-of-the-art watermarking methods that are used to mark generative AI content
- Research on using specific data as backdoor when generating content
- Research on a white-box scenario of embedding and detecting digital watermarks in generated content

ISP RAS also develops DocMarking, a digital watermark system for preventing anonymous confidential data leaks

Federative learning lab of young researchers (supported by the Russian Ministry of Science)

Scientific research:
Developing federative learning algorithms: fast and safe LLM training in application to medicine problems (2024-2026)

The projects lais out theoretical foundations for key problems of federative learning and applies results in practical tasks of digital medicine

Problems to solve:

- protecting data privacy
- attacks on learning privacy
- attacks on learning quality

Digital watermarking:
why are these important nowadays?

ISP RAS

Digital watermarks used in AI systems:

- **For marking generated content**
  (detecting such a content, detecting deepfakes in audio/video streams and images)

- **For protecting created models and training data sets from stealing.** E.g., dataset watermarking allows to prove or disprove that a neural network was trained on a given data set.

On July 21st, 2023, **Joe Biden** announced that **OpenAI, Alphabet, Meta Platforms, Anthropic, Inflection, Amazon, Microsoft** voluntarily committed to implement **digital watermarks for AI-generated content** to make the technology safer.

The companies also promised to perform rigorous testing of AI systems prior to deployment, to share data how AI usage risks can be lowered, and to invest in cybersecurity.

The same approach is being followed by European Union regulators

A path to trusted AI is formed by legal documents

ISP RAS

European Union passed EU AI Act (2024), which categorizes AI systems (important trend):

https://artificialintelligenceact.eu/

Minimal-risk systems. No regulations	Limited-risk systems. Need regulations	High-risk sytems. Need strict regulations	Inadmissible-risk systems. Should be prohibited (with a few exceptions)
AI games or spam filters	AI content generation (image, audio). **The content shall be marked as artificially generated**	Critical infrastructure control, autonomous vehicles, medical AI devices etc.	Social scoring, real-time face recognition etc.

USA

2022: AI Bill of Rights
2023: Executive Order on Safe, Secure, and Trustworthy AI

International initiatives

2023: Hiroshima AI Process
2024: UN Assembly resolution for safe AI systems
2024: US-UK agreement on AI safety (first international AI agreement)

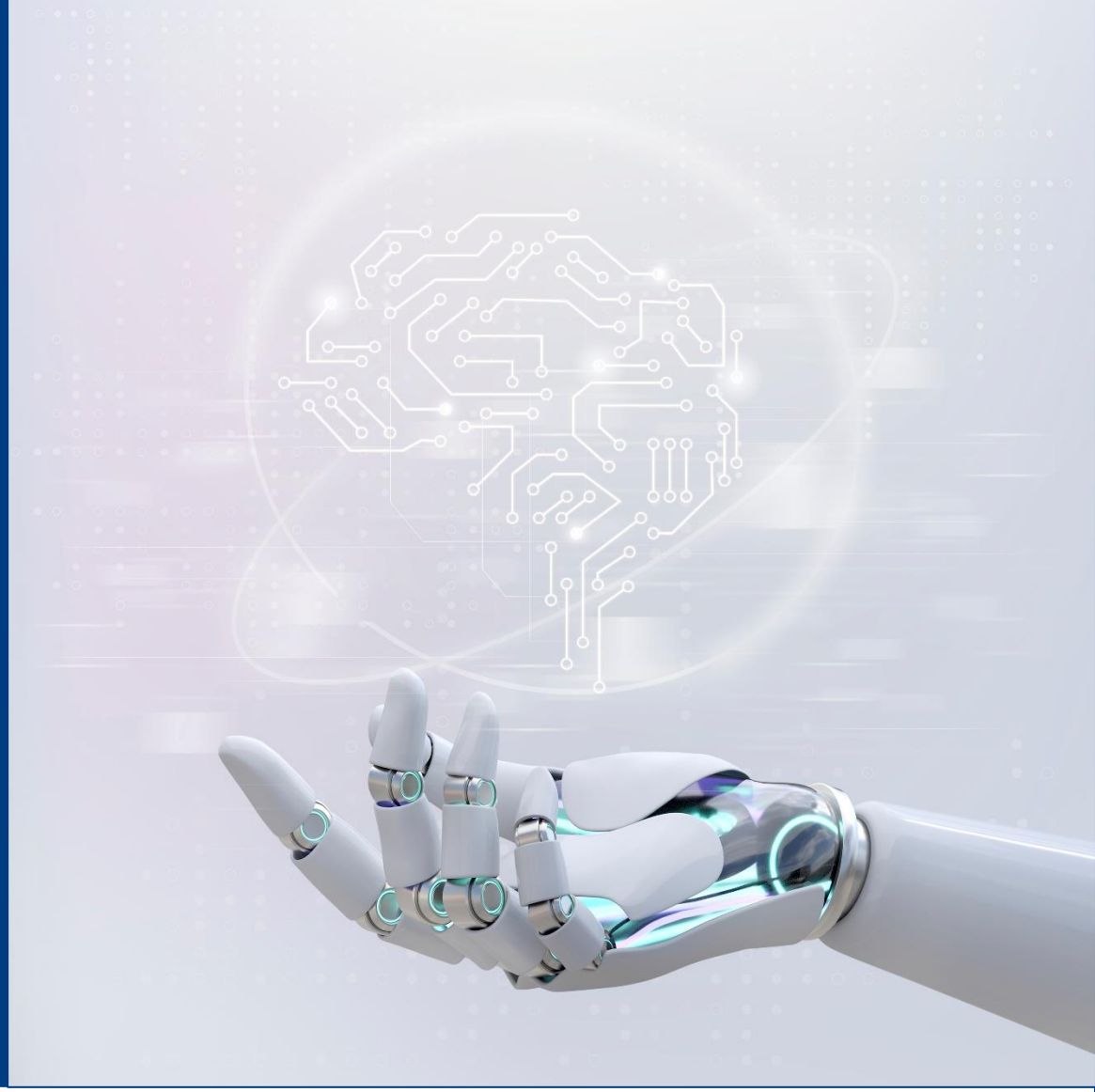
Russia

2019: National strategy for AI development until 2030 (updated in 2024)
2021: Codecs of AI ethics
And other regulations

- **There will be no unregulated AI systems of importance!**
- **Creating own regulatory laws is only possible with own technology stack (technological independence)**
- **There is no universally accepted AI secure software development lifecycle (SDLC), no regulatory documents, no common approach for solving ethical issues**
- **We need a solid scientific foundation and joint work with social and humanitarian scientists**

Isolated groundbreaking technologies are insufficient

We need to create models that ensure long-term stable development and IT technological independence, which results in public benefit as well

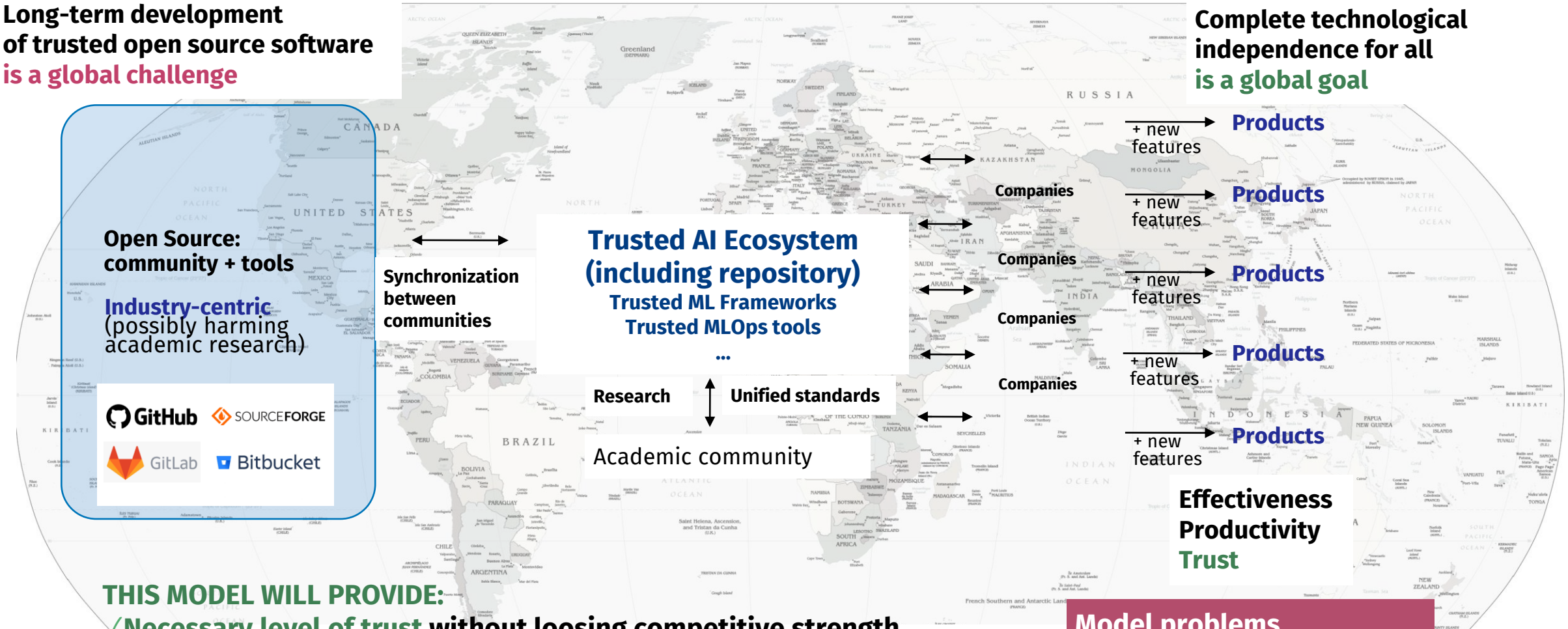


Suggested global long-term development model

ISP RAS

Long-term development of trusted open source software is a global challenge

Complete technological independence for all is a global goal



THIS MODEL WILL PROVIDE:

- ✓**Necessary level of trust** without loosing competitive strength (effectiveness and productivity)
- ✓**Open world academic community** of qualified experts
- ✓**Full control** over the complete code base without any restrictions

Model problems

~~Technological risks~~
~~Resource risks~~
~~Political risks~~

Our experience: system programming technology center

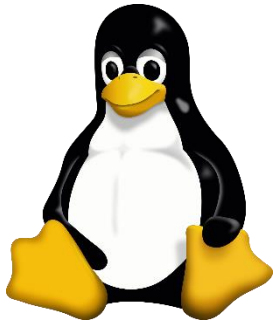
ISP RAS

Created on ISP RAS premises and headed by FSTEC of Russia
Jointly with universities and industry
More than **60** vendors and entities

Current results:

- Maintaining own stable branches of **Linux 5.10 and 6.1**;
- Found **>30 critical issues** in the Linux kernel;
- Written **>500 patches**, and 380 patches are already accepted to the kernel mainline;
- **77** patches for various critical components are accepted (OpenSSL, Qemu, libvirt, CPython, Lua, .NET6 Runtime);
- Working on kernel improvements for increasing its safety;
- Most important: created a growing community of **>80 experts** that work on the above issues (solving understaffing problem)

We created a scalable ecosystem that ensures teaching experts and developing technologies. This leads to a technological independence



QEMU

libvirt VIRTUALIZATION API

OpenSSL Cryptography and SSL/TLS Toolkit

Conclusions

ISP RAS

Successfully developing trusted AI requires:

1. **Creating international repositories** for trusted development tools and SDLC tools

2. **Extending programs for growing high-skilled** system programming and AI experts

3. **Creating centers for developing and deploying state-of-the-art cross-cutting AI technologies**
(digital watermarks, LLM filtering etc.)

4. **Advancing interdisciplinary projects** (AI in medicine, sociology, philosophy, linguistics etc.)

Join our conference for more details!

ISP RAS

Upcoming:

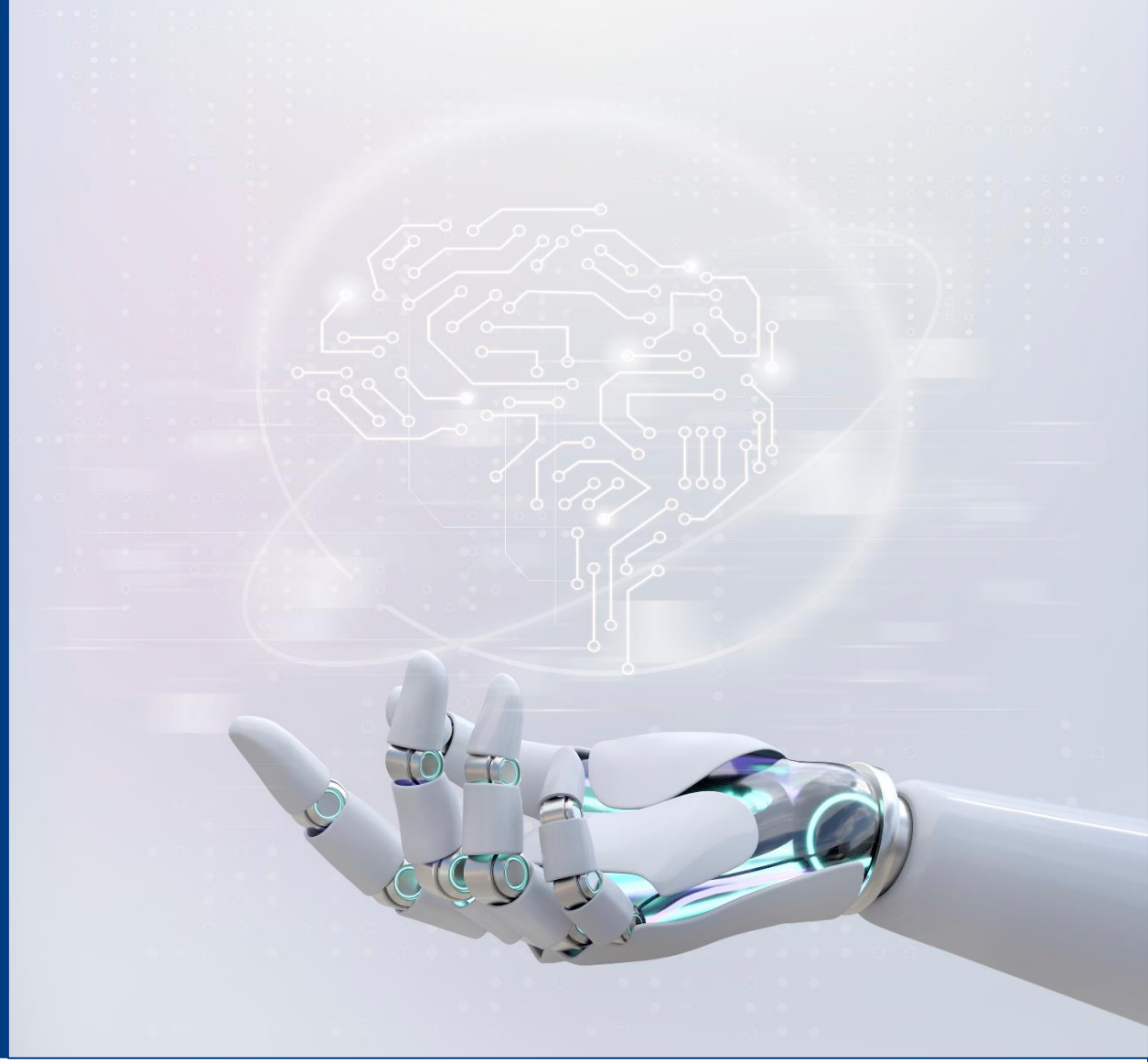
Moscow, December 11-12, 2024

ISP RAS Open Conference

Thank you!

Arutyun Avetisyan
ISP RAS Director
Academician of RAS
arut@ispras.ru

21st October 2024



ISP RAS 30

The International Conference
Mathematical Modeling and Computational Physics, 2024
(MMCP2024)

MMCP