MUNI
FACULTY
OF SOCIAL
STUDIES

Introduction to AI in Social Sciences

26.09.2023

GLCb2028 Artificial Intelligence in Political Science and Security Studies

Jan KLEINER

jkleiner@mail.muni.cz

The Course´s Commons

- https://docs.google.com/document/d/1wwI5iTifzOJbRBhU7AcoiUTF_qJKzpV7EKFphOeEToU/edit

Presentation outline

- What is AI and its history?
- Future markets´ predictions about AI.
- Societal impacts of the AI´s development.
- Will AI enslave the humanity or save it – the “big” debate and the ethics of it.
- ChatGPT and LLMs (r)evolution.
- More questions than answers.

What is AI? (Mueller & Massaron 2021)

- AI has nothing to do with human intelligence → a simulation at best.
- Seven kinds of human intelligence and how AI simulates them - next slide.
- **<u>Four ways of AI´s categorization (the standard model):</u>**

1. **Acting humanly** (Turing test [see critique – e.g., Levasque (2014)] and its alternatives – Reverse Turing Test, Marcus Test, Winograd Schema Challenge etc.)
2. **Thinking humanly** (how to determine? → cognitive modelling approach → introspection, psychological testing, brain imaging).
3. **Thinking rationally** – do humans think rationally?
4. **Acting rationally** – do humans act rationally?

→ **The standard model long-run problem**: assumes that humans supply fully specified objective to the machine → we want machines to pursue OUR objectives, not THEIRS, but they have to be UNCERTAIN (we cannot predict black swans) (Russell & Norvig, 2021).

Discussion: can AI simulate human intelligence?

- Seven types of human intelligence and how AI simulates them (Mueller& Massaron 2021, 12-13):

1. **Visual-spatial** – need to understand dimensions and characteristics of the physical environment.
2. **Bodily-kinesthetic** – Repetitive tasks, higher precision than humans.
3. **Creative** – „A truly new kind of product is the result of creativity“ (p. 13).
4. **Interpersonal** – Computers do not *understand* the question, they provide answers based on statistics.
5. **Intrapersonal** – Inward insight, own interests, setting goals – human-only intelligence? Machines have no desires or interests.
6. **Linguistic** – „…understanding oral, aural, and written input, managing the input to develop an answer, and providing an understandable answer as output“ (p. 13).
7. **Logical-mathematical** – „Calculating a result, performing comparisons, exploring patterns, and considering relationships“ (p. 13).

What is AI? (Mueller& Massaron 2021)

- AI has nothing to do with human intelligence → its simulation at best.
- **What is intelligence?** → it is composed of activities:
  - Learning
  - Reasoning
  - Understanding
  - Grasping truths – validity of manipulated information
  - Seeing relationships
  - Considering meaning
  - Separating fact from belief
- 7 kinds of human intelligence and how AI simulates them – see pp. 12-13.
- **OECD (2019: 21): four elements of current AI vision:** autonomous vehicles and robotics, natural language processing, computer vision, and language and learning.

The Foundations of AI I (Russell & Norvig, 2021)

- **(1) Philosophical:** ethical principles; how does mind arise from physical brain?; where does knowledge come from? (ontological and epistemological prisms) etc.
  - Practical consequences: e.g., Kant´s deontological ethics → the „right thing“ determined not by outcomes (utilitarism), but by universal laws (don´t lie, don´t kill).
- Important, among other things, in geopolitics → China as an emerging AI superpower (Chinese are guided by other principles then the West).

The Foundations of AI II (Russell & Norvig, 2021)

- **(2) Mathematical** – what are the formal rules that grant valid conclusions?; what can be computed/operationalized?
  - Formal logic, probability, statistics
- **(3) Economics** – how should we make decisions alongside our preferences?; what are our preferences?; what are the interests of all the stakeholders?
  - In practice: stakeholder theory; game theory etc, optimizing vs. satisficing (rationality vs. bounded rationality).
- **(3) Neuroscience** – how do brains process information?
- **(4) Psychology** – How do humans and animals think and act?
- **(5) Computer engineering** + quantum computing
- **(6) Control theory** (control engineering) and cybernetics – optimization of systems
- **(7) Linguistics** – how does language relate to thought?; prompts and responses etc.

Types of AI

- **Strong** („generalized intelligence that can adapt to a variety of situations“) vs. **weak** – („specific intelligence designed to perform a particular task well“) (Mueller& Massaron 2021: 16).
- **<u>Seven types (Betz, 2023):</u>**

1. **Narrow** – specific task, no independent learning.
2. **General** – AI learns, thinks and performs similarly to humans.
3. **Superintelligence** – AI surpasses humankind´s knowledge.
4. **Reactive Machines** – AI responds to stimuli in real time, unable to store information.
5. **Limited Memory** – AI can store knowledge.
6. **Theory of Mind** – AI can sense and respond to human emotions (+ limited memory capabilities).
7. **Self-aware** – AI recognizes other´s emotions and has sense of self – AI´s final stage of evolution.

The Singularity (Mueller & Massaron, 2021)

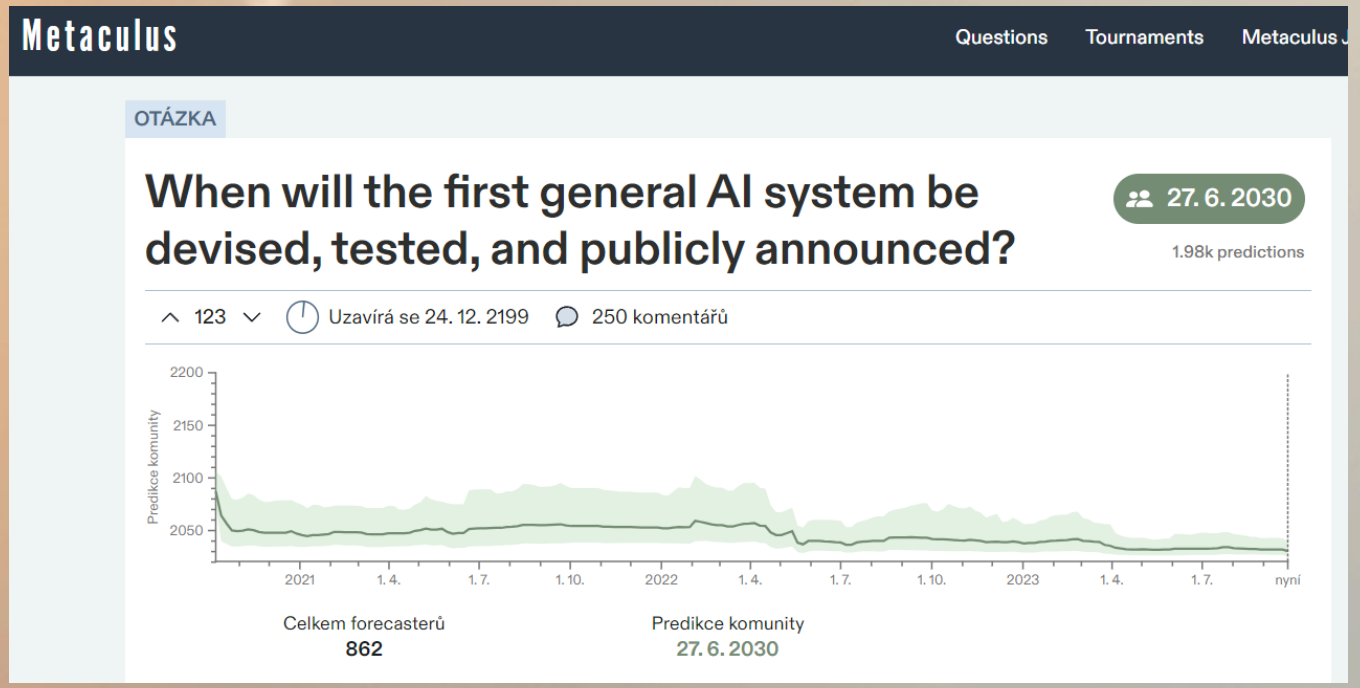
- There is a hype about this.
- A master algorithm (7 kinds of intelligence + 5 tribes of learning).
- Five tribes:
  - Symblogists – logic and philosophy; deduction solves problems
  - Connectionists – neuroscience; backpropagation (a backward process that adjusts neural network model) solves problems
  - Evolutionaries – evolutionary biology; genetic programming solves problems
  - Bayesians – statistics; probabilistic inference solves problems
  - Analogizers – psychology; kernel machines solves problems

Sources of AI hype

- Corporations and their marketing, yet a lot of fails (e.g., Genderify – predicting person´s gender; public backlash due to built-in stereotypes and biases) (Mueller & Massaron, 2021).
- Reliance on authorities and expert opinions = a big no-no! (*ibid.*).
- → User overestimation of and over-reliance on AI´s capabilities (new threats – sleeping in autopilot Tesla etc.) (*ibid.)*.
- AI is not „so much an advancement of technology, but rather the metamorphosis of all technology. This is what makes it so revolutionary“ (Elliot, 2021: 4).
- Future markets.

Future markets predictions

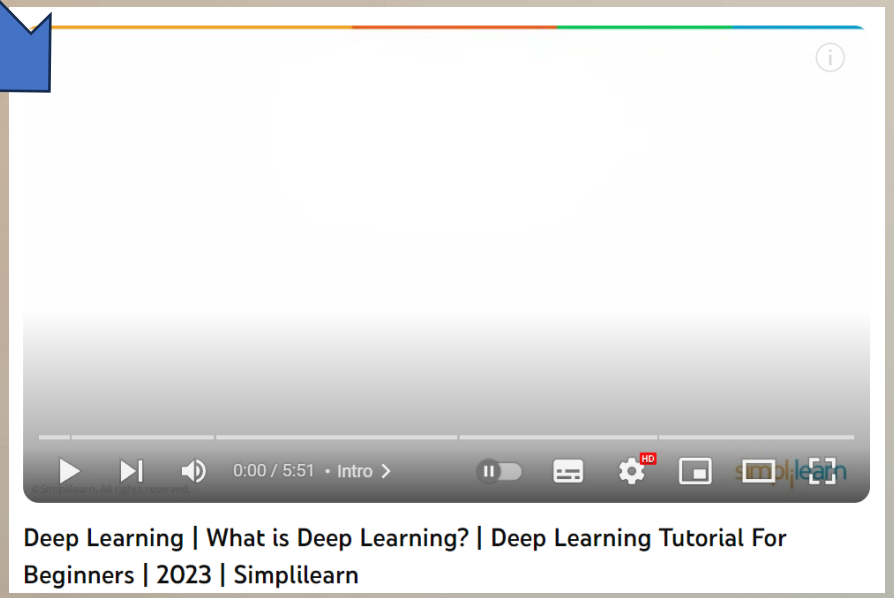
- E.g., Metaculus; what about quantum computing?



Source: Metaculus

The brief history of AI (Russell & Norvig, 2021)

- **Inception** in 1943 – model of artificial neurons.
- <a long, but boring period of theoretical concepts>
- 2001 – Big data due to the advancements of the WWW.
- 2011-present – Deep learning



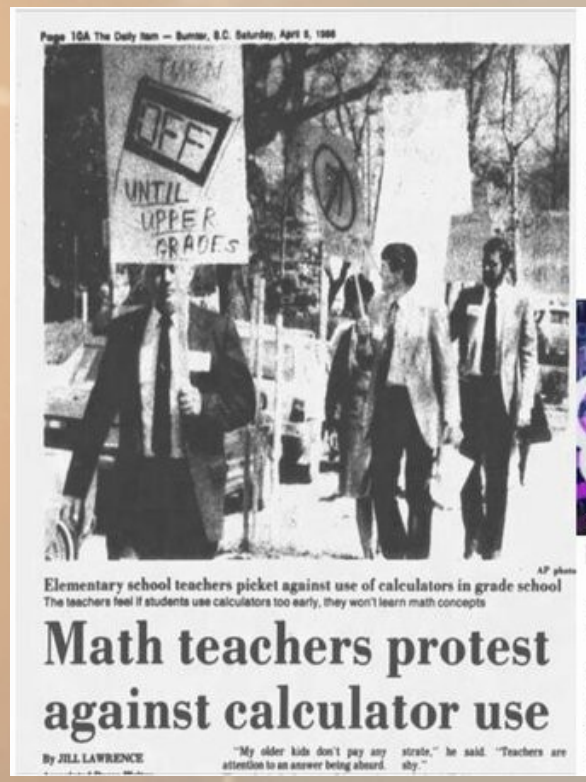
Source: YouTube

Societal impacts I

- Tied to geography (e.g., mobility, geopolitice), workforce and economics, security… everything = cannot list everything ☺
- **Hard-to-imagine** due to its unpredictability and anticipated black-swan-events (beware of predictions!).
- **Social theory of mobility justice** – how inegalitarian aspects of mobility can be exacerbated by AI (e.g., autopilot vehicles) (Birtchnell, 2021: 19).
- Increasing efficiency of traffic flows; could change people´s willingness to travel bigger distances (generally digital technologies + phenomena like pandemics).

Societal impacts II – Work (Boyd, 2021)

- 1738 – Jacques Vaucanson – stoned for the flute player automata.
- 1779 – Lancashire machine-breakers.
- 1793 – cotton gin → concerns about the cost of slaves.



Page 10A The Daily Item — Sunbur, S.C. Saturday, April 9, 1988

AP photo

Elementary school teachers picket against use of calculators in grade school

The teachers feel if students use calculators too early, they won't learn math concepts

Math teachers protest against calculator use

By JILL LAWRENCE

"My older kids don't pay any attention to an answer being absurd. strate," he said. "Teachers are shy."

ChatGPT banned in Italy over privacy concerns

By Shiona McCallum
Technology reporter

31 March 2023
Updated 1 April 2023

OpenAI launched ChatGPT last November

Italy has become the first Western country to block advanced chatbot ChatGPT.

The Italian data-protection authority said there were privacy concerns relating to the model, which was created by US start-up OpenAI and is backed by Microsoft.

Source: Mijwil et al., 2023

Societal impacts II – Work (Boyd, 2021)

- Tied to geography (e.g., mobility, geopolitice), workforce and economics, security… everything = cannot list everything ☺
- **Hard-to-imagine** due to its unpredictability and anticipated black-swan-events (beware of predictions!).
- **Social theory of mobility justice** – how inegalitarian aspects of mobility can be exacerbated by AI (e.g., autopilot vehicles) (Birtchnell, 2021: 19).
- Increasing efficiency of traffic flows; could change people´s willingness to travel bigger distances (generally digital technologies + phenomena like pandemics).

Let´s have a break…

- Find interesting predictions regarding the AI related societal impacts on Metaculus or any other future-markets-prediction site.

Ethics (Birtchnell, 2021)

- **The snowman problem** – AI-driven vehicle has no way of knowing whether the snowman is alive or not; alteration of the trolley problem.
- **Autonomous weapon systems** (UAVs, clever munitions, UGVs, autonomous vessels etc.) → **combatant/non-combatant**? (even humans have troubles distinguishing – contested ROEs); what error (**collateral damage**) is acceptable? → AI **quantifies** decision-making; UAV pilots suffer PTSD too etc.
  - Landmines parallel → Banned by the Ottawa Treaty (Russel & Norvig, 2021).
  - Convention on Certain Conventional Weapons (CCW) – legal aspects (China yes / Israel, USA no); stems from IHL (e.g., the principle of proportionality).
- **Surveillance** and dataflows increase after COVID-19 → AI and smart cities → **privacy** concerns (state/corps. – surveillance capitalism).
- + Robot rights, trust and transparency, fairness and bias etc. (see Russel & Norvig, 2021).

Ethics II (Russell & Norvig, 2021)

- Improved medical diagnosis, better predictions of extreme weather, safer driving…
- BUT: unintended side effects, out-of-distribution, deception, legal aspects (authorship laws) etc.

→ **Principles:** ensure safety, fairness, respect privacy, promote collaboration (to prevent concentration of power), provide transparency (is it possible?), limit harmful uses of AI (e.g., employment ramifications).

How to gauge AI´s performance? (Lynch, 2023)

- Usually via **benchmarks** („…a goal fir the AI system to hit“).
- Artificial Intelligence Index Report 2023 (Maslej et al., 2023) – includes publicly available data.
  - Image classification – 91 %.
  - Human Pose estimation – 94.3 %.
  - Etc.
  - Overcame human baseline → The need for new and new benchmarks.
- HELM – Holistic Evaluation of Language Models.
- Results of the HELM core scenarios.
- Multi-task Language Understanding (MMLU).
  - Non-specialist human baseline – 34.5 % (Nikkel, 2023).
- The more parameters, the better performance?

Parameters (Deepchecks, 2023)

- More parameters → better performance (Google PaLM (540 bil.) vs. Google Chinchilla (75 bil.) – almost same MMLU score).
- Define AI´s model behaviour (how it processes input to produce output) and shape its understanding of language.
- Hyperparameter: **LLM´s temperature**.
  - Regulates randomness or creativity of AI´s responses.
  - Higher – diverse and creative output, but risk of straying.
  - Lower – deterministic, sticking to the most-likely prediction.
- Parameters are just probabilistic constructs – function on statistics and do not hold any inherent meaning.
  - E.g., Birds of feather → flock together (0.7) OR lay eggs (0.2).
  - → Can **hallucinate**!

The Future of Life Institute Open Letter (FLI, 2023)

- A six-months halt on training AI systems more powerful than GPT-4 bc.:
  - Spread of disinfo and propaganda.
  - Automate away even fulfilling jobs.
  - AI can become uncontrollable.
  - AI can be used to develop autonomous weapons.
  - Humankind is not prepared to deal with threats such as these (Maslej et al., 2023 concur).
- Creation of regulatory authorities.
- AI systems should be „accurate, safe, interpretable, transparent, robust, aligned, trustworthy, and loyal“.
- → Enjoy „AI Summer“ before fall and winter.

+ Fear of China and other parties → technological race – a new security dilemma?

References I

- Betz, S. (2023). 7 Types of Artificial Intelligence. Built in. https://builtin.com/artificial-intelligence/types-of-artificial-intelligence.
- Birtchnell, T. (2021). Geographies of AI. In: Elliott, A. (Ed.). (2021). The Routledge Social Science Handbook of AI (1st ed.). Routledge. https://doi.org/10.4324/9780429198533.
- Boyd, R. (2021). Work, employment and unemployment after AI. In: Elliott, A. (Ed.). (2021). The Routledge Social Science Handbook of AI (1st ed.). Routledge. https://doi.org/10.4324/9780429198533.
- Deepchecks. (2023). Deepchecks Glossary: LLM Parameters. Deepchecks. Available from: https://deepchecks.com/glossary/llm-parameters/.
- Elliott, A. (Ed.). (2021). The Routledge Social Science Handbook of AI (1st ed.). Routledge. https://doi.org/10.4324/9780429198533.
- FLI. (2023). Pause Giant AI Experiments: An Open Letter. Future of Life Institute. Available from: https://futureoflife.org/open-letter/pause-giant-ai-experiments/.
- Levesque, H. J. (2014). On our best behaviour. Artificial Intelligence. 212: 27–35. doi:10.1016/j.artint.2014.03.007.

References II

- Lynch, Shana. (2023). AI Benchmarks Hit Saturation. Stanford University Human-Centered Artificial Intelligence. Available from: https://hai.stanford.edu/news/ai-benchmarks-hit-saturation.

- Mijwil, M. M., Hiran, K. K., Doshi, R., Dadhich, M., Al-Mistarehi, A.-H., & Bala, I. (2023). ChatGPT and the Future of Academic Integrity in the Artificial Intelligence Era: A New Frontier. Al-Salam Journal for Engineering and Technology, 116–127. https://doi.org/10.55145/ajest.2023.02.02.015

- Mueller, J. P. & Massaron, L. (2021). Artificial Intelligence for Dummies. Hoboken: John Wiley and Sons, Inc.

- Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault, (2023). The AI Index 2023 Annual Report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA.

- Nikkel, B. (2023). MMLU: Better Benchmarking for LLM Language Understanding. Deepgram. Available from: https://deepgram.com/learn/mmlu-llm-benchmark-guide.

- OECD 2019. Artificial Intelligence in Society. Paris: OECD Publishing.

- Russel, S. J. & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.): Pearson Education Limited.

MUNI
FACULTY
OF SOCIAL
STUDIES

Thank you for your attention.

Questions?

Jan KLEINER

jkleiner@mail.muni.cz