RESPONSIBLE AI

INFO381
MARIJA SLAVKOVIK
SKYPE: SLAVKOVIK

Responsible use of AI

- AI Guidelines: the landscape
- Enforcing responsibility
- Humans Supervising AI
- ~~AI for non-civilian use~~
- ~~Norwegian AI strategy discussion~~

Responsible use of AI

"...the street finds its own uses for things”

Burning Chrome by William Gibson.

All technology and science can be used as tool, used as weapon and abused. As a society and as scientists we must ensure to make technology as abuse proof as possible.

Recall accountability

- Accountability is a relationship between **an actor** and **a forum**, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences.

Recall accountability

- Accountability is a relationship between **an actor** and **a forum**, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences.

Responsible use of AI is concerned primarily with **legal enforcing of accountability.**

In general: ensuring that the forum has power over the actor.

How to guarantee responsible use?

Professional code of conduct

- Teach future developers right from wrong, the laws and guidelines
- Part III - THE HANDBOOK OF INFORMATION AND COMPUTER ETHICS

http://www.cems.uwe.ac.uk/~pchatter/2011/pepi/The_Handbook_of_Information_and_Computer_Ethics.pdf

Value sensitive design

- Concerned with how to make AI “products” abuse-proof.
- Chapter 4 - THE HANDBOOK OF INFORMATION AND COMPUTER ETHICS

  http://www.cems.uwe.ac.uk/~pchatter/2011/pepi/The_Handbook_of_Information_and_Computer_Ethics.pdf

- Definition: Value Sensitive Design is a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process.
- Instead of giving account to a forum, preempt what the forum wants and build it in.

AI guidelines

- Resource: https://inventory.algorithmwatch.org/
- Noble goal: help people who design AI systems know what to be careful about
- Nefarious goal: moral posturing and ethics washing
- Comparison work
- The Ethics of AI Ethics An Evaluation of Guidelines Dr. Thilo Hagendorff

  https://link.springer.com/article/10.1007/s11023-020-09517-8
- Artificial Intelligence: the global landscape of ethics guidelines Anna Jobin, Marcello Ienca, Effy Vayena

  https://arxiv.org/pdf/1906.11668.pdf

Standards

Standards

Standards

- International Standard Organisation https://www.iso.org/standards.html

Standards

- International Standard Organisation https://www.iso.org/standards.html
- “ISO standards are internationally agreed by experts. Think of them as a formula that describes the best way of doing something.”

Standards

- International Standard Organisation https://www.iso.org/standards.html
- “ISO standards are internationally agreed by experts. Think of them as a formula that describes the best way of doing something.”
- ISO26000 - Guidance on social responsibility

Standards

- International Standard Organisation https://www.iso.org/standards.html
- “ISO standards are internationally agreed by experts. Think of them as a formula that describes the best way of doing something.”
- ISO26000 - Guidance on social responsibility
- The ISO 26000 Scope states "This International Standard is not a management system standard. It is not intended or appropriate for certification purposes or regulatory or contractual use. Any offer to certify, or claims to be certified, to ISO 26000 would be a misrepresentation of the intent and purpose and a misuse of this International Standard. As this International Standard does not contain requirements, any such certification would not be a demonstration of conformity with this International Standard.”

Standards

- International Standard Organisation https://www.iso.org/standards.html
- “ISO standards are internationally agreed by experts. Think of them as a formula that describes the best way of doing something.”
- ISO26000 - Guidance on social responsibility
- The ISO 26000 Scope states "This International Standard is not a management system standard. It is not intended or appropriate for certification purposes or regulatory or contractual use. Any offer to certify, or claims to be certified, to ISO 26000 would be a misrepresentation of the intent and purpose and a misuse of this International Standard. As this International Standard does not contain requirements, any such certification would not be a demonstration of conformity with this International Standard.”
- ISO for AI is under construction https://www.iso.org/committee/6794475/x/catalogue/p/0/u/1/w/0/d/0

Certification and verification

- Certification is the formal attestation or confirmation of certain characteristics of an object, person, or organisation. This confirmation is often, but not always, provided by some form of external review, education, assessment, or audit. Accreditation is a specific organisation's process of certification.
- Verification and validation are independent procedures that are used together for checking that a product, service, or system meets requirements and specifications and that it fulfils its intended purpose.
- Benchmarking is the practice of comparing business processes and performance metrics to industry bests and best practices from other companies.

Formal Verification

- Specify the system in a formal language
- Specify the property you are interested in in the same formal language
- Use model checking or theorem proving to see if property hods for system
- Formal verification of ethical choices in autonomous systems by Louise A. Dennis Michael Fisher Marija Slavkovik Matt Webster

https://www.sciencedirect.com/science/article/pii/S0921889015003000?via%3Dihub

Certification challenges

- Lack of definitions - what is AI?
- Certifying people vs certifying objects - which is AI?
- Verification of blackbox systems
- Verification of autonomous systems
- Benchmarking standards - who decides?

(The lack of) Laws

- Genral Data Protection Regulative

  https://eur-lex.europa.eu/eli/reg/2016/679/oj

- California Consumer Privacy Act of 2018

  http://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=CIV&division=3.&title=1.81.5.&part=4.&chapter=&article=

Human supervisory control

- Ethics Guidelines for Trustworthy AI (https://ec.europa.eu/futurium/en/ai-alliance-consultation/human-in-the-loop-/5068587.article guidelines/1) specify

  “Users should be able to make informed autonomous decisions regarding AI systems. They should be given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or chalenge the system.”

Human supervisory control

- Ethics Guidelines for Trustworthy AI ctd.

  “Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such as a human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach.”

Human supervisory control

- Ethics Guidelines for Trustworthy AI ctd.

  “Human in the Loop refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable.”

Human supervisory control

- Ethics Guidelines for Trustworthy AI ctd.

  “Human on the Loop refers to the capability for human intervention during the design cycle of the system and monitoring the system’s operation.”

Human supervisory control

- Ethics Guidelines for Trustworthy AI ctd.

  “Human in Control refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation.”

Society in the loop

Society-in-the-loop: programming the algorithmic social contract by Iyad Rahwan

https://link.springer.com/article/10.1007/s10676-017-9430-8

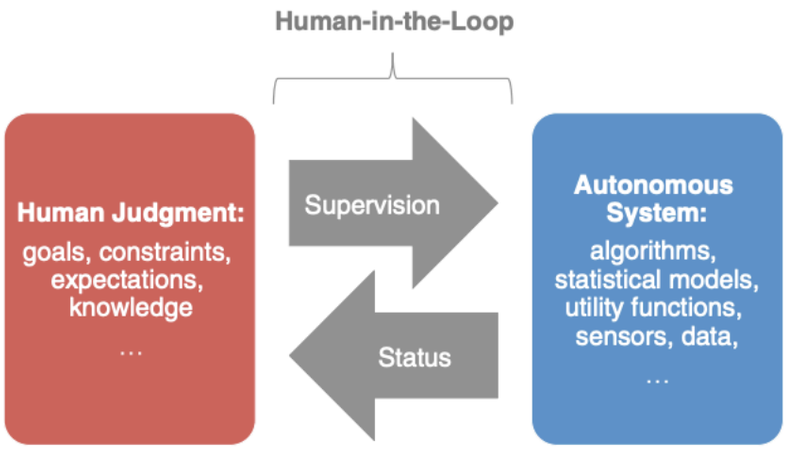

Fig. 1 In a HITL system, a human provides monitoring and supervisory functions at crucial junctions in the system's operation

Human supervisory control as a process by which "one or more human operators are intermittently programming and continually receiving information from a computer that itself closes an autonomous control loop through artificial effectors to the controlled process or task environment"

Reminders

- First deadline-3rd: 10.05.2020 (new date)
  (submission is mandatory, feedback will be given, work can still be done after this deadline, but with care!)

- Second deadline-3rd: 20.05.2020 (locked)
  (feedback is implemented by May 20)
  The submission is just the *project report*, all other material as link

- Approval of presentation paper: 04.05.2020
  (first come first served)

- Meetings can be scheduled as groups need them. but feel free to ping me on Discord any time. Scheduled meeting after May 10 submission.