INTERACTING WITH AI MODULE 3

Session 1, October 27, 2020

Amela Karahasanović, SINTEF and UiO

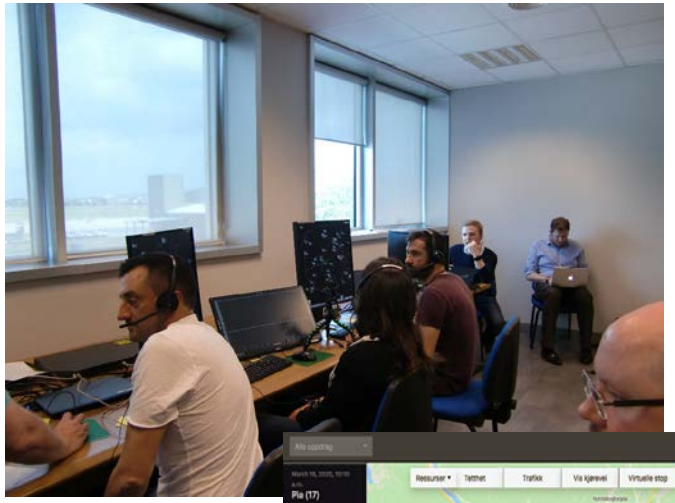
SINTEF

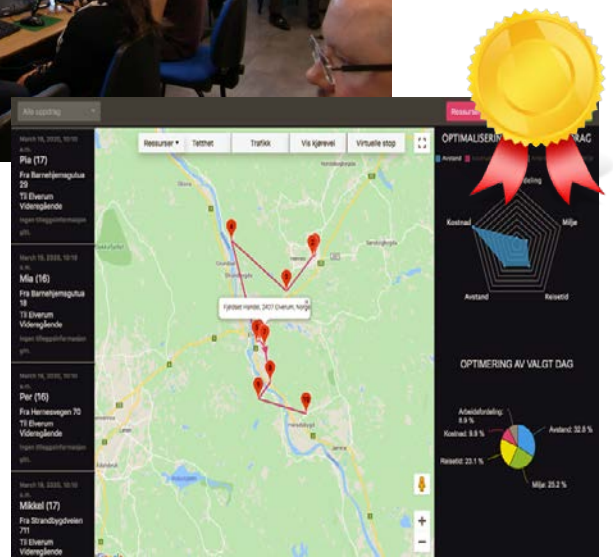
SUSTAINABILITY & DESIGN LAB

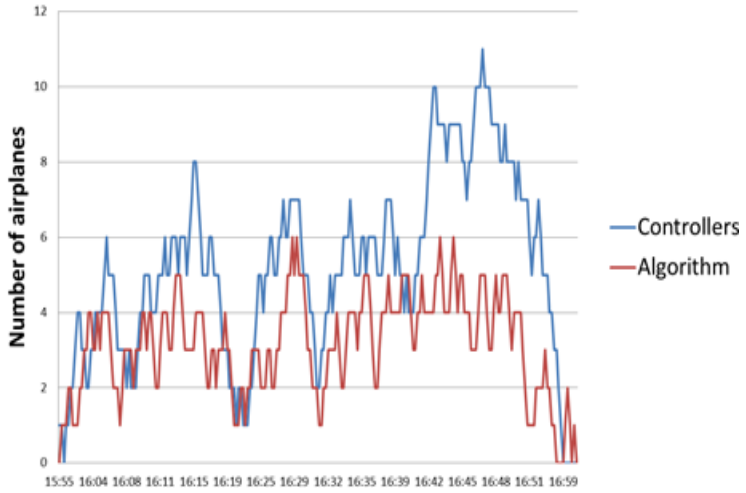
My background

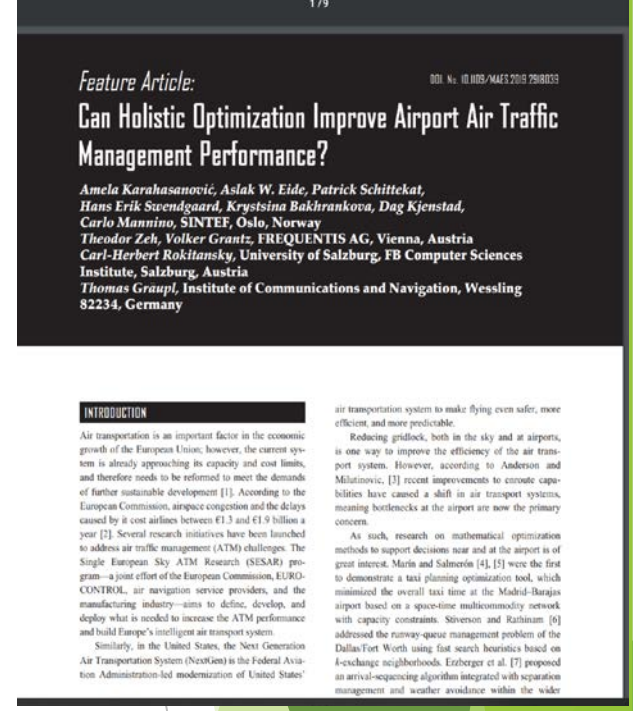
- Senior Research Scientist, SINTEF Digital
- Associate Professor, IFI, Sustainable and Design Lab

- Decision support in ATM
- Evaluation methods in HCI
- User Experience
- User behaviour









Feature Article:

Can Holistic Optimization Improve Airport Air Traffic Management Performance?

Amela Karahasanović, Aslak W. Eide, Patrick Schittekat, Hans Erik Swendgaard, Krystsina Bakhrankova, Dag Kjenstad, Carlo Mannino, SINTEF, Oslo, Norway
Theodor Zeh, Volker Grantz, FREQUENTIS AG, Vienna, Austria
Carl-Herbert Rokitansky, University of Salzburg, FB Computer Sciences Institute, Salzburg, Austria
Thomas Graupl, Institute of Communications and Navigation, Wessling 82234, Germany

INTRODUCTION

Air transportation is an important factor in the economic growth of the European Union; however, the current system is already approaching its capacity and cost limits, and therefore needs to be reformed to meet the demands of further sustainable development [1]. According to the European Commission, airspace congestion and the delays caused by it cost airlines between €1.3 and €1.9 billion a year [2]. Several research initiatives have been launched to address air traffic management (ATM) challenges. The Single European Sky ATM Research (SESAR) program—a joint effort of the European Commission, EUROCONTROL, air navigation service providers, and the manufacturing industry—aims to define, develop, and deploy what is needed to increase the ATM performance and build Europe's intelligent air transport system.

Similarly, in the United States, the Next Generation Air Transportation System (NextGen) is the Federal Aviation Administration-led modernization of United States' air transportation system to make flying even safer, more efficient, and more predictable.

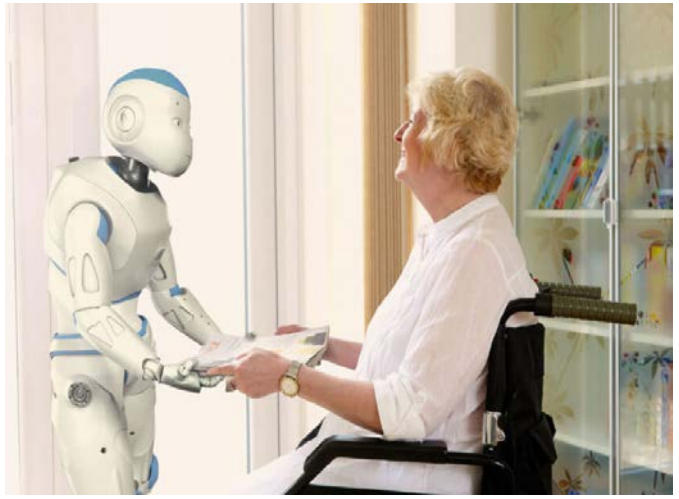
Reducing gridlock, both in the sky and at airports, is one way to improve the efficiency of the air transport system. However, according to Anderson and Milutinovic, [3] recent improvements to enroute capabilities have caused a shift in air transport systems, meaning bottlenecks at the airport are now the primary concern.

As such, research on mathematical optimization methods to support decisions near and at the airport is of great interest. Marín and Salmerón [4], [5] were the first to demonstrate a taxi planning optimization tool, which minimized the overall taxi time at the Madrid–Barajas airport based on a space-time multicommodity network with capacity constraints. Stiverson and Rathinam [6] addressed the runway-queue management problem of the Dallas/Fort Worth using fast search heuristics based on k-exchange neighborhoods. Erzberger et al. [7] proposed an arrival-sequencing algorithm integrated with separation management and weather avoidance within the wider

SESAR Joint Undertaking
https://www.sesarju.eu/discover-sesar
Research and Innovation Project Smart Transport in Rural Areas
https://www.sintef.no/projectweb/smart-transport-i-distriktene/

Why am I interested in this?

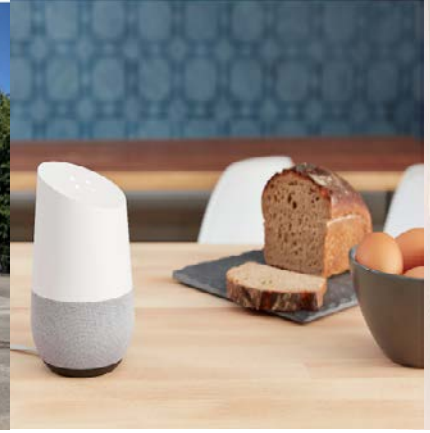
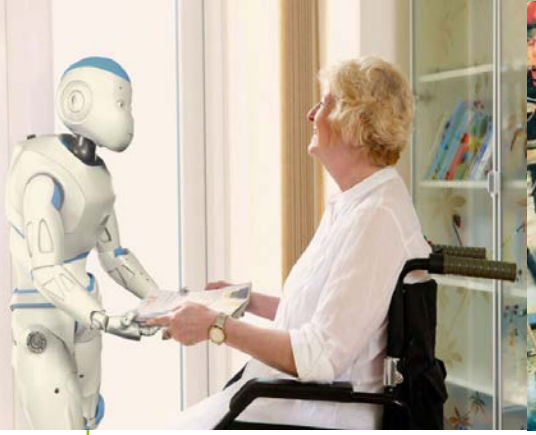
Module 3 Living and working with AI



Objectives

Understanding of challenges related to use of AI infused systems in everyday life and at work

- How to evaluate them?
- When and how to use them?
- What do we know about living and working with them?

Module 3 Overview

- Evaluation of interaction with AI [27th of October]
- Human - AI partnership [3rd of November]
- Lessons learned from studies of human – AI interaction [10th of November]

Literature

Hagras, H., Toward Human-Understandable, Explainable AI, Computer, 51, 9, 2018, 28-36 https://ieeexplore.ieee.org/document/8481251

Phillips, E., Ososky, S., Swigert, B. and Jentsch, F. Human-animal teams as an analog for future human-robot teams, Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol 56, Issue 1, (2016) pp. 1553 – 1557 DOI: https://doi.org/10.1177/1071181312561309

Shneiderman, B., Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy, arXiv.org (February 23, 2020). https://arxiv.org/abs/2002.04087v1 (Extract from forthcoming book by the same title)

Smith-Renner, A., Fan, R., Birchfield, M., Wu, T., Boyd-Graber, J., Weld, D.S., and Findlater. L. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. DOI: https://doi.org/10.1145/3313831.3376624

****

Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in PMLR 81:77-91 http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

De-Arteaga, M., Fogliato, R., and Chouldechova. A., 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–12. DOI: https://doi.org/10.1145/3313831.3376638

Hernández-Orallo, Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement, J. Artif Intell Rev (2017) 48: 397. https://dl.acm.org/doi/10.1007/s10462-016-9505-7

Literature

Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in PMLR 81:77-91 http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

De-Arteaga, M., Fogliato, R., and Chouldechova. A., 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–12. DOI: https://doi.org/10.1145/3313831.3376638

Hernández-Orallo, Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement, J. Artif Intell Rev (2017) 48: 397. https://dl.acm.org/doi/10.1007/s10462-016-9505-7

****

Endsley, Mica R. Designing for Situation Awareness: An Approach to User-Centered Design, Second Edition CRC Press. 2011 (chapters 2 and 10)

Hosanagar, K. A human's guide to machine intelligence, Viking, 2019 (chapters 7- 10)

****

iHUMAN documentary by Tonje Hessen Schei
https://tv.nrk.no/program/KOID75003817

Group assignment Deadline – the final report deadline

(new) Appendix 3: Evaluation - Evaluation plan, findings and reflections. Each group is to plan the evaluation of their own chatbot or a publicly available chatbot of their choice. The evaluation should include an evaluation using the guidelines for Human-AI Interactions and an abusability test. Briefly describe the subject and the scope of the evaluation, the evaluation plan, your findings, and lessons learned. Approx. 3 pages.

Individual assignment Deadline – the final individual report deadline

Human AI collaboration

Philips at al. (2016) give a taxonomy and examples of human-robots collaboration. Choose 2-3 examples, describe their levels of autonomy as described in Shneiderman (2020) and reflect on advantages and disadvantages if we decrease/increase their current level of autonomy. Reflect on their current and needed explainability (Hagras, 2018; Smith-Renner et al. 2020).

Plan for today

- Evaluation – why, what and how to evaluate
- Focus of AI evaluation - User Experience, trust and values

Responsible AI

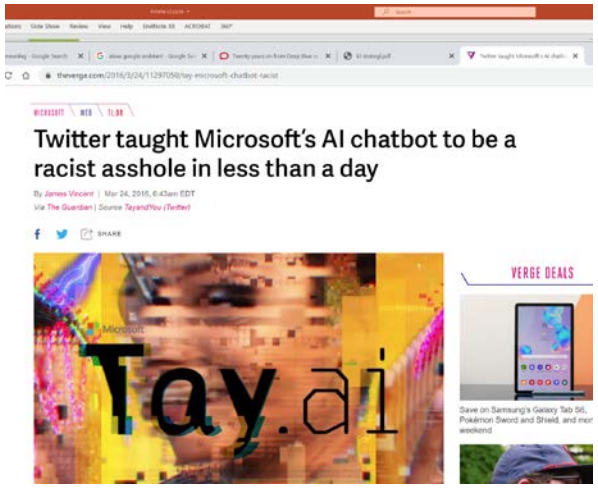
Trustworthy AI

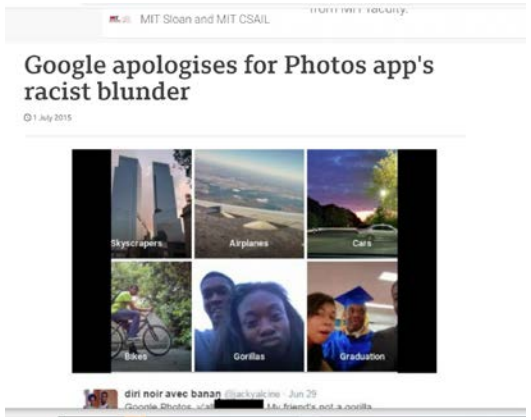
Sustainable AI

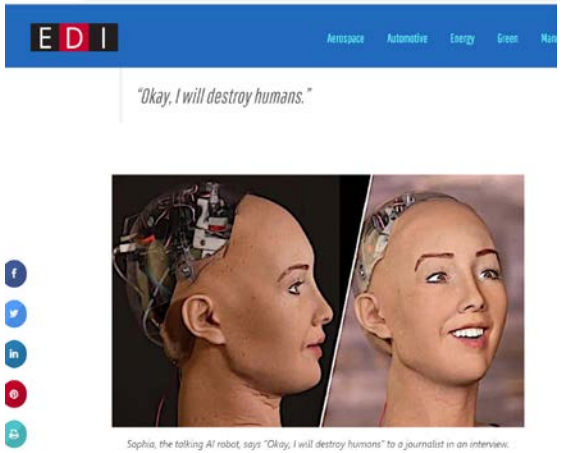
Is evaluating a chatbot different from evaluating a web site?

Why to evaluate?

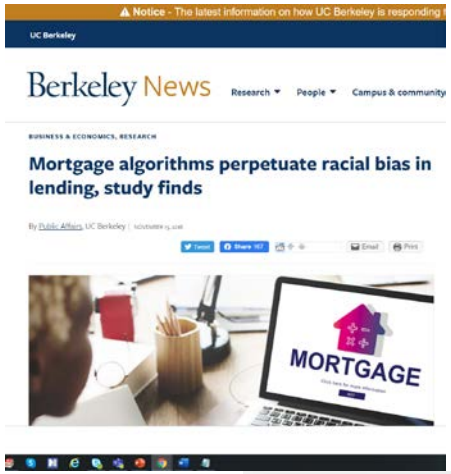
Some examples







Sophia, the talking AI robot, says "Okay, I will destroy humans" to a journalist in an interview.



Airport passport control

Your turn

- What Tay.ai, Google Photos and Airport Passport Control have in common?
- What caused the problem?
- Could this be discovered earlier and how?

Group work – 5 minutes discussion

Commercial facial analysis algorithms

- Evaluate bias present in automated facial analysis algorithms and datasets with respect to phenotypic (observable characteristics) subgroups
- Used approved test for classifying skin colour, **evaluated existing data sets** -> lighter-skinned subjects were overrepresented (up to 86%)
- Introduced a new facial analysis dataset which is balanced by gender and skin type
- Evaluate 3 commercial gender classification systems
- Darker-skinned females are the most misclassified group (error rates of up to 34.7%)
- The maximum error rate for lighter-skinned males is 0.8%

  Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in PMLR 81:77-91

- Can also be because of the background (husky – wolf example)

Testing drugs

- USA National Institutes of Health mandated in 1993 that women and minorities should be included in any government-funded health research
- Only one third of cardiovascular clinical trial subjects are female and only 31% of cardiovascular clinical trials that include women report results by sex (data from 2014)
- Lack of females in drug dose trials leads to overmedicated women (study from 2020)

https://www.sciencedaily.com/releases/2020/08/200812161318.htm

What can we do?

- Raising awareness
- Regulations
- **Evaluation**
  - Guidelines and checklists
  - Diversity of design teams (gender, race, culture, education…)
  - Abusability testing
  - Inspection – ethics bugs

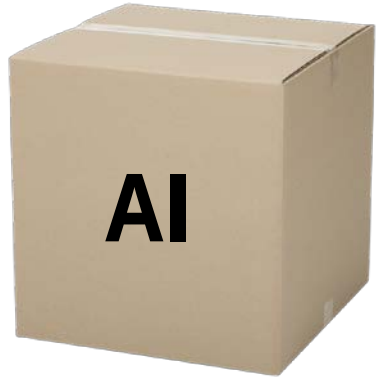
Smith, C.J, Designing Trustworthy AI: A User Experience (UX) Framework, presentation at the RSA Conference 2020, February 24-28, San Francisco, USA

What to evaluate?

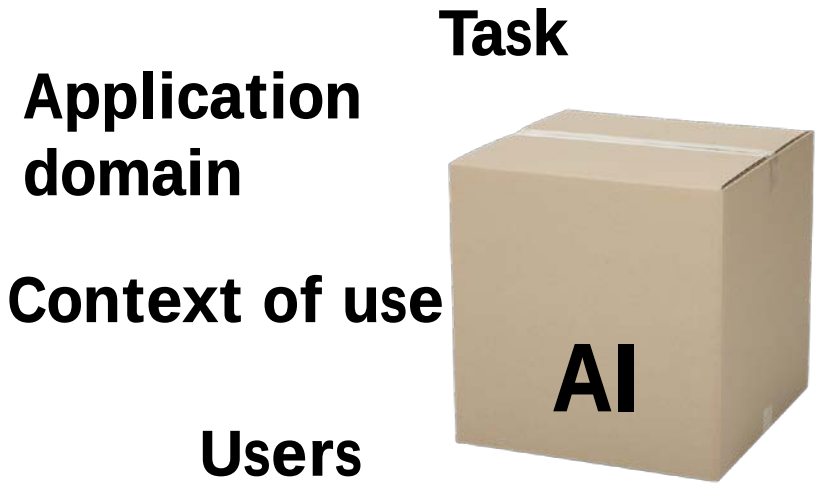
What to evaluate?

From the previous lectures

- Narrow intelligence
  - AI that is good at performing a single task
- AI > Machine Learning > Deep Learning



What to evaluate?



Definitions of AI

- McCarthy (2007) – "AI is the science and engineering of making intelligent machines" –> intelligence test
- Minsky's (1968) - "AI is the science of making machines capable of performing tasks that would require intelligence if done by humans"

-> task-oriented evaluation

- **AI effect** (McCorduck 2004) - tasks are not considered AI problems any more once they are solved without full-fledged intelligence

What to evaluate?

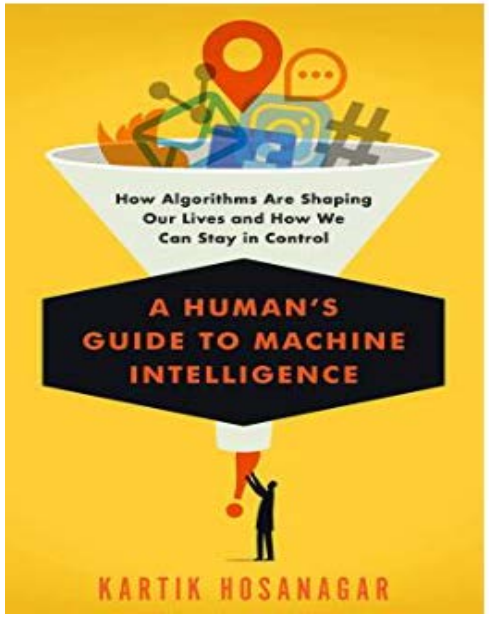
- *AI systems* - robots, chatbots, social robots, AI agents, *self-driving car*
- *AI components* - techniques, algorithms, methods or tools, *camera of the self-driving car*

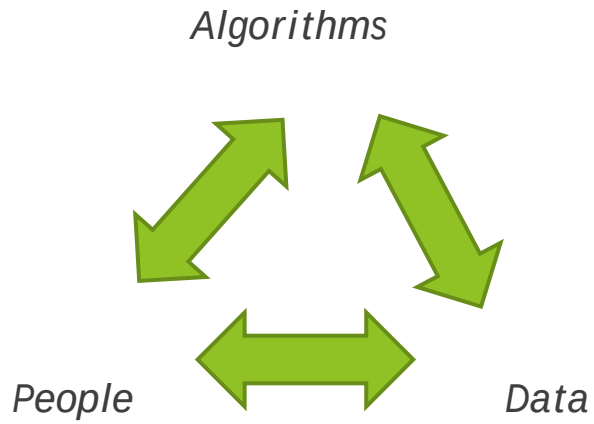
- *Systems evaluates as they are, components according to a specification and how they the serve the system*
  - *Formula 1 engine not appropriate for a family car*

(Hernández-Orallo, 2017)

What are we actually evaluating?

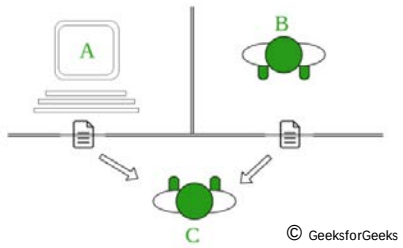
"The results of the algorithmic systems can be attributed to their underlying data, their mathematical logic, and the ways in which people interact with their decisions and suggestions" (Hosanagar, 2019)





How to evaluate?

The Turing test



- Alan Turing 1950
- Turing test is used to determine whether or not computer(machine) can think intelligently like human
- Marcus test – if a computer comprehend a TV show
- Reverse Turing test – a humans proving not being a computer

Lovelace 2.0 test


Computer paints 'new Rembrandt' after old works analysis

By Chris Baraniuk
Technology reporter

6 April 2016

THE NEXT REMBRANDT

The painting was produced by a computer that had analysed existing Rembrandt works

CREATIVITY

- If a computer can create art
- 2016 - new work was created by AI that looks as much like a Rembrandt as possible, while remaining an original portrait
- Composer David Cope – Experiments in Musical Intelligence

A Turing test for emotions

- Pepper by Softbank Robotics is specialized in empathy
- Expert in perceiving emotions
- Communicate with people in natural and intuitive way

From task-oriented to ability-oriented evaluation

Hernández-Orallo (2017)

- Task-oriented evaluation approach
  - Specialized AI systems
- Ability-oriented evaluation approach
  - General-purpose AI

Artif Intell Rev (2017) 48:397–447
DOI 10.1007/s10462-016-9505-7

Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement

José Hernández-Orallo[1]

Published online: 19 August 2016
© Springer Science+Business Media Dordrecht 2016

Abstract The evaluation of artificial intelligence systems and components is crucial for the progress of the discipline. In this paper we describe and critically assess the different ways AI systems are evaluated, and the role of components and techniques in these systems. We first focus on the traditional *task*-oriented evaluation approach. We identify three kinds of evaluation: human discrimination, problem benchmarks and peer confrontation. We describe some of the limitations of the many evaluation schemes and competitions in these three categories, and follow the progression of some of these tests. We then focus on a less customary (and challenging) *ability*-oriented evaluation approach, where a system is characterised by its (cognitive) abilities, rather than by the tasks it is designed to solve. We discuss several possibilities: the adaptation of cognitive tests used for humans and animals, the development of tests derived from algorithmic information theory or more integrated approaches under the perspective of universal psychometrics. We analyse some evaluation tests from AI that are better positioned for an ability-oriented evaluation and discuss how their problems and limitations can possibly be addressed with some of the tools and ideas that appear within the paper. Finally, we enumerate a series of lessons learnt and generic guidelines to be used when an AI evaluation scheme is under consideration.

Keywords AI evaluation · AI competitions · Machine intelligence · Cognitive abilities · Universal psychometrics · Turing test

1 Introduction

The evaluation of any discipline must necessarily be linked to the purpose of the discipline. What is the purpose of artificial intelligence (AI)? McCarthy's pristine definition of AI sets this unambiguously: "[AI is] the science and engineering of making intelligent machines"

José Hernández-Orallo
jorallo@dsic.upv.es

[1] DSIC, Universitat Politècnica de València, València, Spain

Springer

From task-oriented to ability-oriented evaluation

Task-oriented evaluation approach

- Specialized AI systems; clear goals: speech recognition, game playing
- Does the system perform the task

Ability-oriented evaluation approach

- General-purpose AI: artificial pets, assistants, smartbots…variety of tasks
- Abilities: verbal abilities, learning abilities, motion abilities

How to evaluate?

AI applications: computer vision, speech recognition, music analysis, machine translation, text summarisation, information retrieval, robotic navigation and interaction, automated vehicles, game playing, prediction, estimation, planning, automated deduction, expert systems

Task-oriented

- human discrimination – assessment by or against a human – it is being subjective
- problem benchmarking – assessment against a set of problems - difficult to define good sets, for example self-driving cars perform well in California, but not in Nord Norway
- Peer- confrontation – competition against another system; the results relative to the opponents

How to evaluate?

Ability-oriented evaluation

- Psychometrics (IQ tests and similar) – not really adequate, "ability" very anthropocentric not reflect the diversity of AI systems

Generic guidelines

- Specify the set of systems to be evaluated (or can be opponents in the competition), the set of possible tasks, describe the similarities between the tasks

How can we use what we already know about interaction evaluation?

Usability engineering

Activities aiming to improve the ease of use of an interface

- Expert-based testing *(usability inspection)*
- Automated testing *(usability inspection)*
- User-based testing *(usability testing)*

Expert-based testing

- Structured inspections done by interface experts
- Before tests with users
- Confusing wording, inconsistent layout, obvious flaws
- **Heuristic review**
  - **Compare interface with the rules**
- Consistency inspections
  - Series of screens or web pages inspected
- Cognitive walkthrough
  - Experts perform the tasks (high-frequency and important/seldom)
- Guidelines review
  - Web Content Accessibility Guidelines

Eight Golden Rules of Interface Design

July 15, 2008 by web2usability

Eight Golden Rules of Interface Design

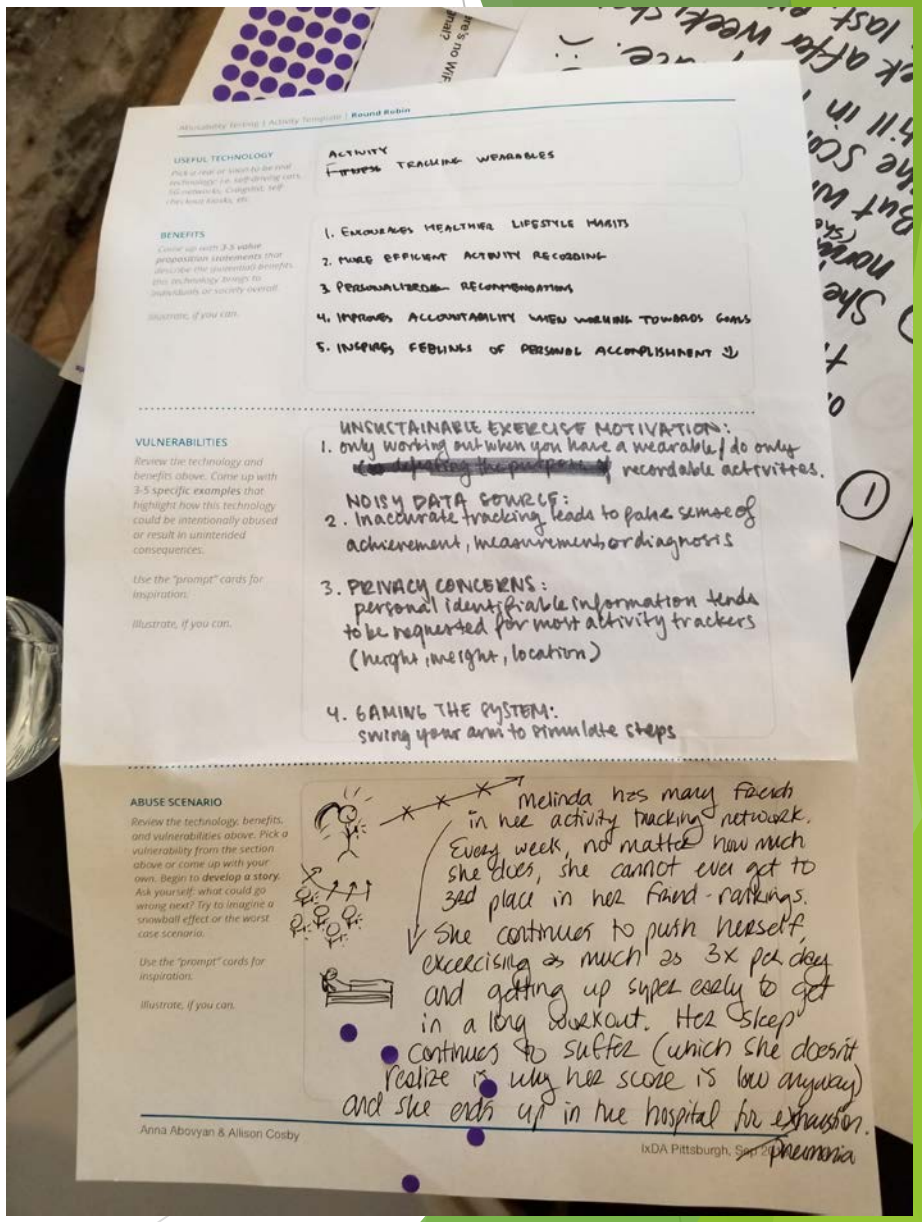
As a result of Interface Design Studies, Ben Shneiderman proposed a collection of principles that are derived heuristically from experience and applicable in most interactive systems. These principles are common for user interface design, and as such also for web design.

1. Strive for consistency.
2. Enable frequent users to use shortcuts.
3. Offer informative feedback.
4. Design dialog to yield closure.
5. Offer simple error handling.
6. Permit easy reversal of actions.
7. Provide the sense of control. Support internal locus of control.
8. Reduce short-term memory load.

Speculative: Conduct UX research and activate curiosity

Abusability Testing

- Speculate about misuse and abuse
- Create “Black Mirror” episodes
  - Severe abuse and consequences



Smith, C.J.,Designing Trustworthy AI: A User Experience (UX) Framework, presentation at RSA Conference 2020, February 24-28, San Francisco USA

Image of Template created by Anna Abovyan & Allison Cosby for Abusability Testing activity conducted at IxDA Pittsburgh, Sep 2019

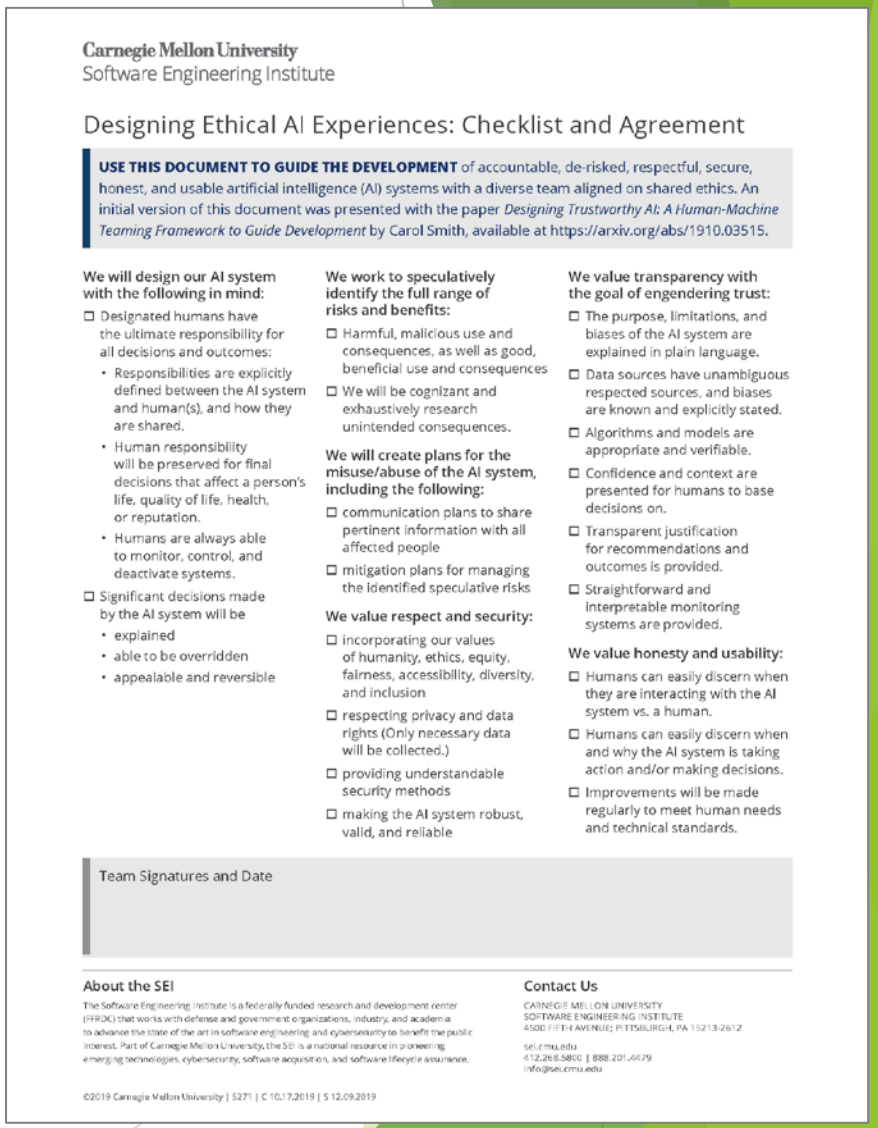
Template by: Anna Abovyan & Allison Cosby, IxDA Pittsburgh, Sep 2019

Abusability template

- Useful technology
  - Pick a real or soon-to-be technology
- Benefits
  - Come up with 3-5 value propositions statements that describe the (potential) benefits this technology brings to individuals or society overall. Illustrate if you can.
- Vulnerabilities
  - Review the technology and benefits above. Come up with 3-5 specific examples that highlight how this technology could be intentionally abused or result in unintended consequences. Use the prompt card. Illustrate if you can.
- Abuse scenario
  - Review the technology, benefits and the vulnerabilities above. Pick a vulnerability from the section above or come up with your own. Begin to develop a story. Ask yourself: what could go wrong next? Try to imagine a snowball effect or the worst case scenario. Use the prompt card. Illustrate if you can.

Checklist and Agreement

- Pair with Tech Ethics
  - Bridge gap between "do no harm" and reality
- Reduce risk and unwanted bias
- Mitigation planning
- Support inspection



Carnegie Mellon University
Software Engineering Institute

Designing Ethical AI Experiences: Checklist and Agreement

USE THIS DOCUMENT TO GUIDE THE DEVELOPMENT of accountable, de-risked, respectful, secure, honest, and usable artificial intelligence (AI) systems with a diverse team aligned on shared ethics. An initial version of this document was presented with the paper *Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development* by Carol Smith, available at https://arxiv.org/abs/1910.03515.

We will design our AI system with the following in mind:

- ☐ Designated humans have the ultimate responsibility for all decisions and outcomes:
  - Responsibilities are explicitly defined between the AI system and human(s), and how they are shared.
  - Human responsibility will be preserved for final decisions that affect a person's life, quality of life, health, or reputation.
  - Humans are always able to monitor, control, and deactivate systems.
- ☐ Significant decisions made by the AI system will be
  - explained
  - able to be overridden
  - appealable and reversible

We work to speculatively identify the full range of risks and benefits:

- ☐ Harmful, malicious use and consequences, as well as good, beneficial use and consequences
- ☐ We will be cognizant and exhaustively research unintended consequences.

We will create plans for the misuse/abuse of the AI system, including the following:

- ☐ communication plans to share pertinent information with all affected people
- ☐ mitigation plans for managing the identified speculative risks

We value respect and security:

- ☐ incorporating our values of humanity, ethics, equity, fairness, accessibility, diversity, and inclusion
- ☐ respecting privacy and data rights (Only necessary data will be collected.)
- ☐ providing understandable security methods
- ☐ making the AI system robust, valid, and reliable

We value transparency with the goal of engendering trust:

- ☐ The purpose, limitations, and biases of the AI system are explained in plain language.
- ☐ Data sources have unambiguous respected sources, and biases are known and explicitly stated.
- ☐ Algorithms and models are appropriate and verifiable.
- ☐ Confidence and context are presented for humans to base decisions on.
- ☐ Transparent justification for recommendations and outcomes is provided.
- ☐ Straightforward and interpretable monitoring systems are provided.

We value honesty and usability:

- ☐ Humans can easily discern when they are interacting with the AI system vs. a human.
- ☐ Humans can easily discern when and why the AI system is taking action and/or making decisions.
- ☐ Improvements will be made regularly to meet human needs and technical standards.

Team Signatures and Date

About the SEI

The Software Engineering Institute is a federally funded research and development center (FFRDC) that works with defense and government organizations, industry, and academia to advance the state of the art in software engineering and cybersecurity to benefit the public interest. Part of Carnegie Mellon University, the SEI is a national resource in pioneering emerging technologies, cybersecurity, software acquisition, and software lifecycle assurance.

Contact Us

CARNEGIE MELLON UNIVERSITY
SOFTWARE ENGINEERING INSTITUTE
4500 FIFTH AVENUE; PITTSBURGH, PA 15213-2612

sei.cmu.edu
412.268.5800 | 888.201.4479
info@sei.cmu.edu

Checklist and Agreement - Downloadable PDF:
https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=636620

How to evaluate chatbots?

A Survey on Evaluation Methods for Chatbots

Wan Maroengsit
SIIT, Thammasat University
131 Moo 5, Tiwanont Road,
Bangkadi,
Muang, Pathum Thani 12000,
Thailand
+66 (0) 2501 3505
5822771333@g.siit.tu.ac.th

Thanarath Phyakulpinyo
SIIT, Thammasat University
131 Moo 5, Tiwanont Road,
Bangkadi,
Muang, Pathum Thani 12000,
Thailand
+66 (0) 2501 3505
5822790978@g.siit.tu.ac.th

Korawat Phonyiam
NECTEC
112 Phahonyothin Road, Khlong Nueng, Khlong Luang, Pathum Thani 12120, Thailand
+ 66 (0) 2564 6900
korawat.pho@nectec.or.th

Suporn Pongnumkul
NECTEC
112 Phahonyothin Road, Khlong Nueng, Khlong Luang, Pathum Thani 12120, Thailand
+ 66 (0) 2564 6900
suporn.pon@nectec.or.th

Pimwadee Chaovalit
NECTEC
112 Phahonyothin Road, Khlong Nueng, Khlong Luang, Pathum Thani 12120, Thailand
+ 66 (0) 2564 6900
pimwadee.cha@nectec.or.th

Thanaruk Theeramunkong
SIIT, Thammasat University
131 Moo 5, Tiwanont Road,
Bangkadi, Muang, Pathum Thani
12000, Thailand
+66 (0) 2501 3505
thanaruk@siit.tu.ac.th

ABSTRACT

Keywords

1. INTRODUCTION

CCS Concepts

A Survey on Evaluating Methods for Chatbots. Maroengsit, W. et al., 2019, ICIET 2019

- Reviewed 30 papers; 18 with evaluations; AI-based chatbots, rule-based chatbots and combinations
- Three categories of evaluations
  - Content evaluation
    - Evaluation focuses on the response context of the chatbots; approach adopted from Natural Language Processing/ Natural Language Generation and Information Retrieval domains
  - User satisfaction
    - Standard HCI method; appropriate here as chatbots are complex systems with no single correct answer
    - Evaluate specific aspects on Likert scale such as appropriateness, naturalness, empathy, helpfulness etc.
    - Session level (whole chat session) and turn level (each response)
  - Functional evaluation
    - Evaluation based on goal/task; if chatbot assist in buying tickets – number of successfully finished purchase
    - Usage statistics
    - Evaluation of building blocks

How to evaluate chatbots?


Evaluating and Informing the Design of Chatbots

Evaluating and Informing the Design of Chatbots, Jain M. et al. 2018, DIS 2018

- Evaluation of 8 chatbots with 16 firs-time chatbot users over multiple sessions on the Facebook Messenger platform
- Three days interactions with chatbots
- Face-to-face semi-structured interview with the participants to elicit their understanding of the chatbots, perceived benefits/limitations, interesting conversations/experiences and areas of improvements
- Quantitative data analysis: total interaction time, message count, interactive elements (composition of the chatbots' and participants' messages)

Shifting the focus of AI evaluation

Can AI play chess better than humans?
-> How can AI empower humans?

Future AI systems should focus enhancing human cognitive capabilities and channelling human creativity…incorporating trust, ethics, and human values

Global effects of a 'local' optimal solution
Values, ethics, privacy and security as a core design considerations
Embedding ethics and values into AI system

(Lukowicz, Slusallek, 2018)

Example: User Experience with robots

- Context: factory
- Two types of robots, one within a safety fence
- UX questionnaire (23 respondents)
- Study – a year and the half
- Compared two types of robots
- Used UX factors in HRI for the evaluation
- UX over time
- Covered aspects: cooperation, perceived safety, perceived stress, perceived usability, general UX

Robots in Time:
How User Experience in Human-Robot Interaction Changes over Time

Roland Buehner, Daniela Wurhofer, Astrid Weiss, and Manfred Tscheligi

HCI& Usability Unit, ICT&S Center, University of Salzburg, Austria
firstname.lastname@sbg.ac.at

Abstract. This paper describes a User Experience (UX) study on industrial robots in the context of a semiconductor factory cleanroom. We accompanied the deployment of a new robotic arm, without a safety fence, over one and a half years. Within our study, we explored if there is a UX difference between robots which have been used for more than 10 years within a safety fence (type A robot) and a newly deployed robot without fence (type B robot). Further, we investigated if the UX ratings change over time. The departments of interest were the oven (type A robots), the etching (type B robot), and the implantation department (type B robot). To observe experience changes over time, a UX questionnaire was developed and distributed to the operators at three defined points in time within these departments. The first survey was conducted one week after the deployment of robot B (n=23), the second survey was deployed six months later (n=21), and the third survey was distributed one and a half years later (n=23). Our results show an increasing positive UX towards the newly deployed robots with progressing time, which partly aligns with the UX ratings of the robots in safety fences. However, this effect seems to fade after one year. We further found that the UX ratings for all scales for the established robots were stable at all three points in time.

Keywords: Industrial Robots, Measurement, Semiconductor Factory, User Experience.

1 Introduction

For effective and highly productive industrial manufacturing, robots have already shown their usefulness in many sectors of production. With that kind of automation, a vast, cheap, and fast production has become reality. However, most of these systems are placed within a safety fence. During production, no human is allowed to enter the working space of the robot and therefore restricting access, any interaction, and/or cooperation with the robot. However, there are claims that more powerful human-robot interaction with the human and the robot working as a team is needed in order to be highly competitive [1]. That means it is necessary to break the general known paradigm of strictly separating

G. Herrmann et al. (Eds.): ICSR 2013, LNAI 8239, pp. 158–167, 2013.

UX Definitions

- "A consequence of a **user**'s internal state, the characteristics of the designed **system** and the **context** within which the interaction occurs" (Hassenzahl&Tractinsky 2006)*
- "All aspects of the **end-user**'s interaction with the company, its services, and its products" (Nielsen Norman Group)
- "The quality of experience a **person** has when interacting with a specific design" (Uxnet, online)

* Marc Hassenzahl and Noam Tractinsky (2006): User experience - a research agenda, Behaviour &Information Technology, 25:2, 91-97

UX list

satisfying helpful fun
enjoyable motivating provocative
engaging challenging surprising
pleasurable enhancing sociability rewarding
exciting supporting creativity emotionally fulfilling
entertaining cognitively stimulating

boring unpleasant
frustrating patronizing
making one feel guilty making one feel stupid
annoying cutesy
childish gimmicky

(Rogers, Sharp, Preece; Interaction design; 2011)

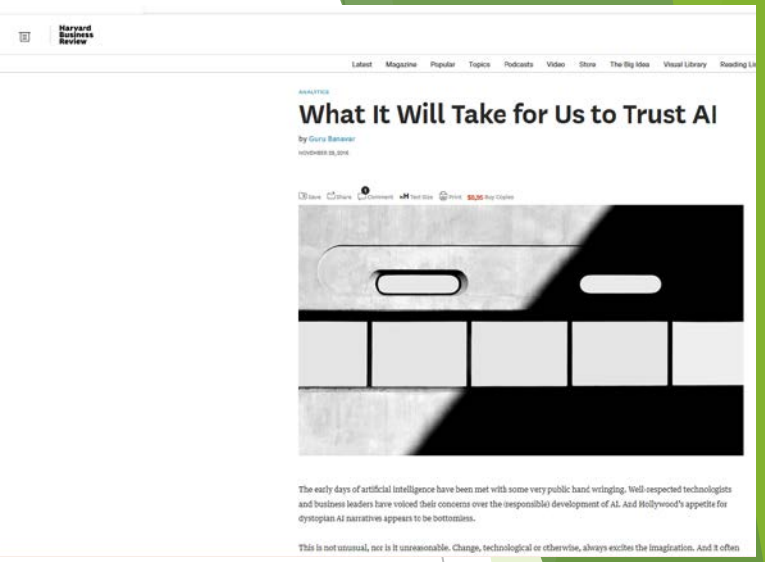
Values

The Three Laws of Robotics *(Isaac Asimov, I Robot)*

- A robot may not injure a human being or, through inaction, allow a human being to come to harm
- A robot must obey the orders given it by human beings except where such orders would conflict with the First Law
- A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws
- The Zeroth Law: A robot may not injure humanity, or, by inaction, allow humanity to come to harm

- What about environment, peace, justice…UN Sustainable Development Goals?

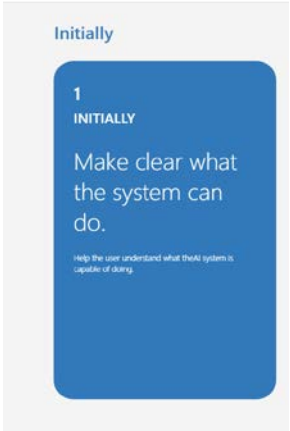
Trust



- "If we are ever to reap the full spectrum of societal and industrial benefits from artificial intelligence, we will first need to trust it"

- Trust of AI systems will be earned over time
- One need to recognize and minimize bias, introduced for example by data sets as chatbot Tay who become racist by reading tweets
- Algorithms should be able to explain their suggestions or decisions
- -> *We'll learn more about trust in Human – AI partnership session*

Your turn



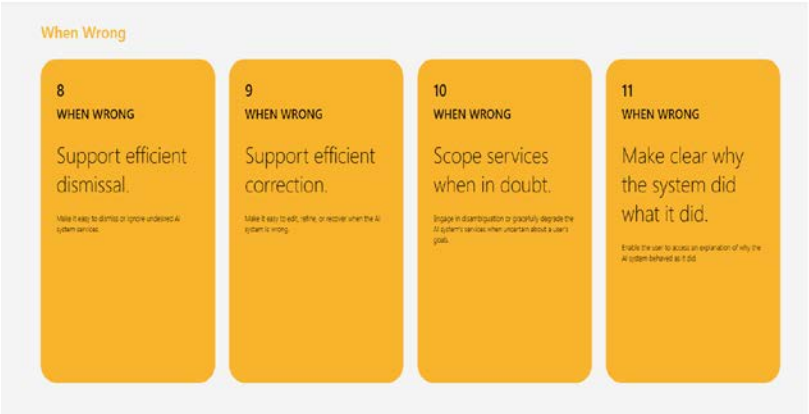
Use Microsoft guidelines for human AI interaction to quickly evaluate Siri or Google assistant

Group work – 10 minutes

Groups 1 and 2-> M1

Group 3 and 4 -> M9

Group 5 and 6 -> M11



https://aidemos.microsoft.com/guidelines-for-human-ai-interaction/demo

Your turn

LUDVIG

https://www.youtube.com/watch?v=U9KrEcn4W3Q

Which UX dimensions would you evaluate?
Which values should be adresses by design?

Group work – 10 minutes discussion

Evaluation – takeaways

- Importance of the evaluation cannot be overestimated
- Consider the big picture – application domain, task, users, context of use
- Consider type of the system you are evaluating and interconnections between the data, algorithms and people
- Consider using existing design guidelines and using/extending existing instruments for measuring usability and UX
- Focus of the evaluation is moving towards values, trust, ethics

References

- McCarthy J (2007) What is artificial intelligence. Technical report, Stanford University. http://www-formal.stanford.edu/jmc/whatisai.html
- Minsky ML (ed) (1968) Semantic information processing. MIT Press, Cambridge
- McCorduck P (2004) Machines who think. A K Peters/CRC Press, Boca Raton
- Hosanagar, A human's guide to machine intelligence, Viking, 2019
- Hernández-Orallo, Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement , J. Artif Intell Rev (2017) 48: 397.
- Paul Lukowicz and Philipp Slusallek. 2018. How to avoid an AI interaction singularity. Interactions 25, 5 (August 2018), 72-78.

- Buchner R., Wurhofer D., Weiss A., Tscheligi M. (2013) Robots in Time: How User Experience in Human-Robot Interaction Changes over Time. In: Herrmann G., Pearson M.J., Lenz A., Bremner P., Spiers A., Leonards U. (eds) Social Robotics. ICSR 2013. Lecture Notes in Computer Science, vol 8239. Springer, Cham
- Marc Hassenzahl and Noam Tractinsky (2006): User experience - a research agenda, Behaviour and Information Technology, 25:2, 91-97
- Banavar, What It Will Take for Us to Trust AI, Harvard Business Review, November 2016
- *Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in PMLR 81:77-91*
- Maroengsit, W. et al. (2019) A Survey on Evaluating Methods for Chatbots., ICIET 2019
- Jain M. et al. (2018) Evaluating and Informing the Design of Chatbots, DIS 2018