RESPONSIBLE AI LESSON 1

1. **The twin transitions**

-- **Climate protection** & **Digital transformation**

-- Goal: The green and digital transitions shall reinforce each other. Digital innovation shall become one of the driving forces behind the EU’s shift to a zero-carbon economy.

1. **Global climate agreements**

* Kyoto protocol
  + - * Held in 1997 but implemented in 2005
      * Role of developed countries to mitigate GHG that is taken as a burden.
* UNFCCC (UNSC): United Nations Framework for convention on climate change
* Paris Agreement
  + - * Limit global warming to well below 2deg and prsue efforts to make it 1.5deg
      * Held in 2015 imple in 2016
* Climate Change **Mitigation** = Reducing greenhouse gas emissions
* Climate Change **Adaptation** = Developing resilience to consequences of climate change

1. **Environmental Sustainability**

* the responsibility to conserve natural resources and protect global ecosystems to support health and wellbeing, now and in the future

1. **Considering Climate Impacts of AI**

* AI can either help with mitigation of Climate Change and the adaptation to it, or it can worsen it with its energy cost
  1. *Impact Assessment Framework:*

A) Compute-Related Impacts,

B) Immediate Application Impacts,

C) System-Level Impacts.

1. **Compute-related Impacts:**

* Model development and deployment carbon output (small vs large models(more parameters which is growing day by day.))
* Computing infrastructure (cloud, data centers, local networks)
* A study by the university of Massachusetts showed that **training AI models** to do Natural Language Processing (NLP) can produce carbon dioxide around the equivalent of **5 times the lifetime emissions of the average American car.**

1. **Immediate Application Impacts**

* *Positive*: **ML models can track gas emmisions**
* *Positive*: **Forecasting of resourse production**
* *Positive*: **Improving heating/cooling system efficiency**
* Negative: ***ML can decrease the cost of oil extraction operations, decreasing the price of benzene/diesel, and increasing practically the gas emmisions***
* Negative: ***Can help cattle-farming, which by itself contributes a lot to GW***

1. **System-level impacts**

* Advertisment recommending systems (which are AI) can increase the consumption of certain products with higher or lower gas emissions

RESPONSIBLE AI: LESSON 2

1. What is ethics and morals?

* Ethics: Moral principles that govern a person’s decision-making and behavior
* Morals: set of normative values, rules and ideals that are accepted in a society

1. Can and should AI make moral judgements?

* Humans are considered to be moral agents. Which is someone who has the ability to discern right from wrong based on their ethics, and be held accountable for their actions.
* However, in AI, moral situation are often not defined, thus the AI doesn’t know what to do
* Are AI really moral agents, when they must be made from different people with differing values/morals?

1. The moral responsibility test for AI

* A set of Aristotelian conditions to evaluate whether an AI has moral responsibility
  + Rule 1: cause an outcome through their own initiated and controlled (in)action (**causation condition**)
  + Rule 2: act physically and psychologically on their own will and intention (**freedom condition**)
  + be knowledgeable of the relevant details regarding the context of (in)action (**knowledge condition**)
  + possess the capacity to morally evaluate the significance of their action and inaction relative to a purpose (**deliberation condition**)

1. What is an ethical dilemma?

* decision-making problem between two possible moral imperatives, neither of which is unambiguously acceptable or preferable

1. The **Two Important Ethical Schools**:

* **Deontological ethics**: The morality of an action is based on whether that action itself is right or wrong under a series of rules. Some decisions/actions are considered morally obligatory regardless of their consequences
* **Utilitarian/Consequative ethics**: The consequences of one’s decisions are the ultimate basis for any judgment about the rightness or wrongness of that conduct. Moral rightness does not depend directly on anything other than effects. Outcomes can justify the means.
* Deontological Critic: By the statement, lying is always a bad thing. But, if lying would save a life, would it still be bad?
* Utilitarian Critic: If you use a flamethrower to do a good act, can the action still be morally right?

1. MIT Moral Machine

* each dilemma simultaneously randomizes three additional attributes:
  + which group of characters will be spared if the car does nothing;
  + whether the two groups are pedestrians, or whether one group is in the car;
  + and whether the pedestrian characters are crossing legally or illegally
* Global preferences:
  + • sparing humans over animals • sparing more lives • sparing young lives
* Local preferences:
  + For the West and the South, sparing the young is higher in priority than in the East
  + Preference for inaction is stronger in the West, and is less prominent in the South
  + South = southern america, West = europe/US, East = middle east / asia

RESPONSIBLE AI PART ¾

EU RULES AND GUIDELINES TO AI

1. EU Ethics guidelines for Thrustworthy AI

* **Lawful**
* **Ethical**
* **Robust**



1. Human Agency and oversight

* Like many technologies, AI systems can equally enable and hamper fundamental rights. Oversight may be achieved through governance mechanisms, preventing that. Humans 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 challenge the system.

1. Technical robustness & safety

* AI systems, like all software systems, should be protected against vulnerabilities that can allow them to be exploited by hacking. Back-ups of the AI systems are necessary. Also, in terms of robustness, accuracy pertains to an AI system’s ability to make correct judgements

1. Privacy and data governance

* AI systems must guarantee privacy and data protection throughout a system’s entire lifecycle. In any given organization that handles individuals’ data, protocols governing data access should be put in place. The quality of the data sets used is paramount to the performance of AI systems, Biases, info corruption and errors must be taken into account.

1. Transparency

* The data sets and the processes that yield the AI system’s decision, should be documented to the best possible standard. Also, AI systems should not represent themselves as humans to users.

1. Diversity, non-discrimination and fairness

* Identifiable and discriminatory bias should be removed in the collection phase where possible. Particularly in business-to-consumer domains, systems should be user-centric and be able to be used by anybody, regardless of demographic.

1. Societal and environmental wellbeing

* Research should be fostered into AI solutions addressing areas of global concern, such as the Sustainable Development Goals
* The use of AI systems should be given careful consideration particularly in situations relating to the democratic process, including not only political decision-making but also electoral contexts

1. Accountability

* Evaluation by internal and external auditors, and the availability of such evaluation reports, can contribute to the trustworthiness of AI

RESP AI PART 4 DIVERSITY

1. Does university education translate into leadership opportunities? For men, women, and minorities equally?
2. Does gender diversity in the workplace really have a measuarable impact on innovation and financial success?
3. And why does an absence of women in tech leadership become a self-perpetuating, long-term problem for some companies?
4. **Does university education translate into leadership opportunities? For men, women, and minorities equally?**
   * The documents do not directly address the correlation between university education and leadership opportunities for men, women, and minorities. However, they do discuss the importance of diversity in leadership and the slow progress in achieving gender and ethnic diversity in executive roles. For example, the McKinsey report notes that female representation on executive teams in the U.S. and U.K. rose from 15% in 2014 to 20% in 2019, and representation of ethnic minorities on these teams stood at 13% in 2019, up from 7% in 2014​​. This suggests that while there is progress, it's slow, and the documents don't specifically tie these trends to university education.
5. **Does gender diversity in the workplace really have a measurable impact on innovation and financial success?**
   * Yes, gender diversity in the workplace has a measurable impact on innovation and financial success. The McKinsey report and the BCG study both affirm this. McKinsey's analysis found that companies with diverse executive teams are more likely to outperform their peers in terms of profitability​​. Similarly, the BCG study indicates that companies with above-average diversity on their leadership teams report greater innovation and higher EBIT margins​​.
6. **Why does an absence of women in tech leadership become a self-perpetuating, long-term problem for some companies?**
   * The absence of women in tech leadership can become a self-perpetuating problem due to several factors. Firstly, a homogenous leadership team lacks diverse perspectives, which can hinder innovation and adaptability in a dynamic business environment​​. Secondly, companies with a lack of diversity may struggle with creating inclusive cultures, which is crucial for leveraging the benefits of diverse leadership​​. Additionally, the absence of women in leadership roles can perpetuate a cycle where there are fewer role models and mentors for aspiring female leaders, potentially leading to a narrower talent pipeline. While the documents do not directly address the long-term consequences of an absence of women in tech leadership, they highlight the importance of diversity and inclusion for business success and innovation

STEREOTYPES

AND BIAS IN AI

1. Prototype = Most typical representative of a concept/category (e.g. the typical “doctor” or the typical “bird”)
2. Stereotype = Perceived characteristics of an entire group of people (e.g. „men are not good at socializing“, „people of color are often criminal“, „Asians are good at math“)
3. Bias = Disproportionate weight in favor of or against an idea, thing or group of people

Note: A method from Social Psychology to detect unconscious biases is the **Implicit Association Test** (IAT, Greenwald, McGhee, & Schwartz, 1998).

1. Unconscious (implicit) human biases

* Biases may exist toward any social group
* may still influence decisions and actions we take
* Certain factors can activate unconscious beliefs

1. Widespread assumptions about human vs algorithmic decision-making

* People judge subjectively and have numerous bias effects in their decisions. Decisions can be influenced e.g. by momentary personal moods or even the weather. **TRUE**
* Computers / algorithms / AI systems, on the other hand, are objective and fair and able to judge neutrally **FALSE**

1. Biased AI

* Biased AI refers to AI systems that produce results that systematically and unfairly favor certain groups due to the presence of inherent biases in the data, thus reinforcing existing societal inequalities
* Fairness is “the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics”

1. Incomplete data: Selection bias

* Coverage bias: The population represented in the data set does not match the population that the machine learning model is making predictions about
* A model is trained to predict future sales of a new product based on phone surveys conducted with a sample of consumers who bought the product. Consumers who instead bought a competing product were not surveyed, and as a result, this group of people was not represented in the training data.

5. Sampling bias: Data is not collected randomly from the target group.

* Instead of randomly targeting consumers, the surveyer chose the first 200 consumers that responded to an email invitation (who might have been more enthusiastic about the product than average purchasers).
* Participation bias: Members of certain groups opt-out of surveys (or labeling tasks) at different rates than members of other groups
* Consumers who bought the competing product were 80% more likely to refuse to complete the survey, and their data was underrepresented in the sample.

1. How to keep human bias out of AI?

* Integration of bias detection strategies
* More diversity in teams developing AI would be very beneficial
* Awareness and understanding of the various causes of algorithmic biases is the first step
* Ethical governance standards

1. How to identify potential algorithmic bias?

* **Missing Feature Values** : indicator that certain key characteristics of your data set are under-represented
* **Unexpected Feature Values** : Look for examples that contain feature values that stand out as especially uncharacteristic or unusual
* **Data Skew** : certain groups or characteristics may be under or over-represented relative to their real-world prevalence

ANTROMORPHISM 6

**Anthropomorphism is the attribution of human traits, emotions, or intentions to non-human entities**

1. The Three-Factor Theory of Anthropomorphism

* **Elicited agent knowledge** : Knowledge about people must be accessible (usually easily retrievable) and activated by the agent, e.g. by humanlike design features.
* **Effectance motivation** : Persons must have a motivation to interact effectively with the agent, to explain its behavior and predict future behavior.
* **Sociality motivation** : Persons must have a need for social relation and belonging, e.g. when feeling lonely

1. Consequences of Anthropomorphism

* Provided an agent is perceived to have a mind, and be capable of:
* **Moral care and concern** : conscious experience and should therefore be treated as a moral agent worthy of care and concern
* **Responsibility** : intentional action and can therefore be held responsible for its actions
* **Social influence** : observing, evaluating, and judging a perceiver, thereby serving as a social influencer

1. What is Social Engineering

* **Influencing/manipulating people so they give up confidential information or sensitive data**

UNCANNY VALLEY 7

1. What do these three people have in common

* Ernst Jentsch
* Sigmund Freud
* Masahiro Mori
* **They have all addressed the creepy effects of highly humanlike machines**

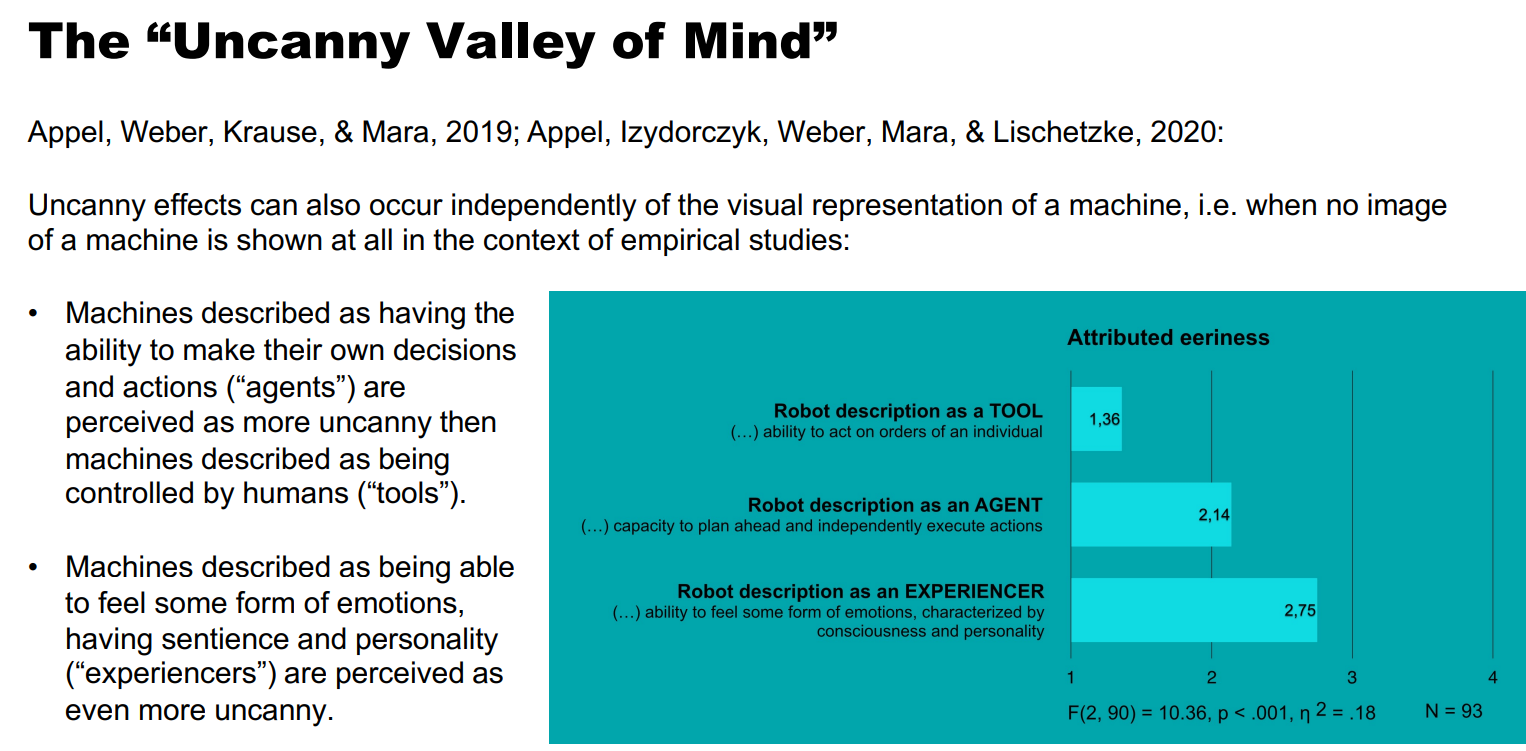
1. “On the psychology of the uncanny” – Jentsch

* entsch describes uncanny feelings as “intellectual uncertainty” and not being “at home” in the situation concerned

1. “The uncanny” - Freud

* Freud emphasized that what is uncanny is something that seems to be “un-homely” and unfamiliar, but at the same time “homely” and familiar
* In Freud’s view, the uncanny might be anything we experience in adulthood that reminds us of early psychological stages or of primitive experiences
* **Freud names the concept of the uncanny doppelgänger**

1. The Uncanny Valley – Mori

* as robots become more humanlike, they appear more familiar until a point is reached at which subtle imperfections of appearance make them look eerie
* This led Mori to the belief that robot builders should not attempt to make their creations overly lifelike and realistic in appearance
* general assumption that highly lifelike robots/machines are often experienced as eerie

TRUST IN AI 8

1. Trust-relevant situation

* Trust-relevant (or trust diagnostic) situations **involve vulnerability and require some stake (risk)**
* Without being vulnerable or at risk, there is no need for trust
* AI-produced credit scoring may represent a risk to the loan officer, or AI-supported diagnosis may inherent a risk for the patient to be misdiagnosed

1. Trustor‘s propensity to trust

* **refers to the general tendency for someone to trust others**
* the impact of trust propensity is most notable early in interpersonal interactions, when other information may not yet be available

1. Scales to measure a person‘s propensity to trust

* **General Trust Scale** -> a 6-item questionnaire that uses general statements to measure participants’ beliefs about honesty and trustworthiness of others, in general *(Yamagishi & Yamagishi, 1994)*
* **Trust in Technology Scale** -> a 6-item questionnaire that measures participants’ beliefs about general trustworthiness of technology *(Schneider, 2017)*

1. Trustee‘s trustworthiness

* Trustworthiness is the trustor’s perception of the trustee
* Perceptions are formed as a trustor interprets and ascribes motives to the trustees’ actions . Thus, perceptions of trustworthiness, although inherently within the trustor, are a function of the interaction of trustor and trustee as the trustor is processing information about the trustee. It is important to note these are the ascribed beliefs of the trustor and are not necessarily factual

1. Goals of research on interpretability and trust:

* Trust in a system must be appropriately calibrated to the actual system performance (Muir, 1994)

1. Theoretical foundations: Dimensions of Trust

* Affective/social trust: When a trustee is perceived as well-minded, warm-hearted and adhering to social norms
* Cognitive/performance-based trust: When a trustee is perceived as competent, understandable and predictable in terms of the performance required
* **Cognitive components of trust should be more relevant for robotics/AI than affective components**

XAI – Explainable ai

1. The EU parliament definition of Explainability:

* the importance of the explainability of Al systems’ outputs, processes and values, making them understandable to non-technical audiences and providing them with meaningful information, which is necessary to evaluate fairness and gain trust

1. What is Explainable AI (XAI)

* AI that produces details or reasons to make its functioning clear or easy to understand
* refers to an explanatory agent revealing underlying causes to its or another agent's decision making

1. Goals of XAI

* Improve human interpretability of Al systems
* Foster "informed trust" / appropriate levels of trust
* Support human autonomy in Al-supported decision making
* Determine the justifiability of the decision made by a machine
* Help in deciding accountability and liability (leading to good policy-making)
* Avoid discrimination and the reproduction of societal bias

1. Challenges of XAI

* The Black box problem: Good AI models have around 100Mil parameters. We cannot know for sure which parameters affect the explanation (possible hidden bias)
* What is interpretable for whom: Explanation styles must be adapted to differend needs and knowledge levels, as explanations are different for everyone

AI, Data, & Privacy 10

1. What is privacy

* the right to control others’ access to one’s personal world, and also by regulating one’s own output in communication with others
* Types: Physical, Psychological, Social, Informational

1. The Privacy Paradox

The privacy paradox refers to the observed discrepancy between individuals' expressed concerns about privacy and their actual behaviors when it comes to sharing personal information, especially online.

1. Privacy Calculus

* Privacy calculus refers to the decision-making process individuals go through when determining whether to disclose personal information in exchange for some perceived benefits or services.

1. The big data scheme:

* The three Vs:

1. **Data Volume**: Makes analysis more powerful as the model or algorithm has more data to learn from
2. **Data Velocity**: Facilitates analysis in real-time. The predictions are based on real time data rather than old data or observations
3. **Data Variety**: Allows AI models to make more accurate predictions when presented with new data that was not seen before
4. The General Data Protection Regulation (GDPR EU law)

* Main principles: **Lawfulness, fairness and transparency; Purpose limitation; Data minimization; Accuracy; Storage limitation; Integrity and confidentiality; Accountability**

1. Ethical Measures by the EU commision

* Data Protection By Design, targeted towards **companies and organisations**
* Four measures in this law:
  + Pseudonymization
  + Anonymization
  + Data Minimization
  + Applied Cryptography

1. Online Disinhibition Effect

* Loss of self-control in online communication, which occurs in the context of phenomena such as hate speech
* Causes: Anonymity, Asynchrony, Lack of non-verbal comms, Minimized Authority