Seminar Explainable AI
Module 09

Ethical, Legal and Social Issues of xAI

Andreas Holzinger
Human-Centered AI Lab (Holzinger Group)
Institute for Medical Informatics/Statistics, Medical University Graz, Austria
and
Explainable AI-Lab, Alberta Machine Intelligence Institute, Edmonton, Canada
This is the version for printing and reading. The lecture version is didactically different.
Ethical Design principles at a glance:

- Explainability -> transparency, auditability, traceability
- Verifiability -> safety, security, reducing uncertainty
- Responsibility -> use, misuse, adverse social effects
- Fairness -> value align, human rights, shared benefit
- Privacy -> accessibility, human (data) protection
AGENDA

- 00 Intro: from Causality to ethical responsibility
- 01 Automatic – Automated - Autonomous
- 02 Legal accountability and Moral dilemmas
- 03 AI ethics: Algorithms and the prove of explanations
- 04 Responsible AI – examples from computational sociology (bias, fairness, ...)

a.holzinger@human-centered.ai 4
01 Repetition: From Causality to Ethical Responsibility
Causation – beware of counterfactuals

- David Hume (1711-1776): Causation is a matter of perception: observing fire > result feeling heat
- Karl Pearson (1857-1936): Forget Causation, you should be able to calculate correlation
- Judea Pearl (1936- ): Be careful with purely empirical observations, instead define causality based on known causal relationships, and beware of counterfactuals ...


Hume again: “... if the first object had not been, the second never had existed ...”

Causal inference as a missing data problem

$x_i := f_i(ParentsOf_i, Noise_i)$

Interventions can only take place on the right side

Remember: Correlation is NOT Causality

Dependence vs. Causation

Hans Reichenbach (1891-1953): **Common Cause Principle**

- **Links causality with probability:**
  - If X and Y are statistically dependent, there is a Z influencing both
  - Whereas:
    - A, B, ... events
    - X, Y, Z random variables
    - P ... probability measure
    - Px ... probability distribution of X
    - p ... probability density
    - p(X) .. Density of Px
    - p(x) probability density of Px evaluated at the point x

Hans Reichenbach 1956. The direction of time (Edited by Maria Reichenbach), Mineola, New York, Dover.

https://plato.stanford.edu/entries/physics-Rpcc/

Functional Causal Model

- $X_1, \ldots, X_n$ ... set of observables
- Draw a directed acyclic graph $G$ with nodes $X_1, \ldots, X_n$

- Parents = direct causes
- $x_i = f_i(\text{ParentsOf}_i, \text{Noise}_i)$

Remember: Noise means unexplained (exogenous) and denote it as $U_i$

Question: Can we recover $G$ from $p$?
Answer: under certain assumptions, we can recover an equivalence class containing the correct $G$ using conditional independence testing
But there are problems!

<table>
<thead>
<tr>
<th>Explainability</th>
<th>in a technical sense highlights decision-relevant parts of the used representations of the algorithms and active parts in the algorithmic model, that either contribute to the model accuracy on the training set, or to a specific prediction for one particular observation. It does not refer to an explicit human model.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Causability</td>
<td>as the extent to which an explanation of a statement to a human expert achieves a specified level of causal understanding with effectiveness, efficiency and satisfaction in a specified context of use.</td>
</tr>
</tbody>
</table>

- **Causability**: a property of a person, while
- **Explainability**: a property of a system
Probabilistic vs. causal inference problems

“How do humans generalize from few examples?”

▪ Learning relevant representations
▪ Disentangling the explanatory factors
▪ Finding the shared underlying explanatory factors, in particular between $P(x)$ and $P(Y|X)$, with a causal link between $Y \rightarrow X$


Deductive Reasoning = Hypothesis > Observations > Logical Conclusions
- DANGER: Hypothesis must be correct! DR defines whether the truth of a conclusion can be determined for that rule, based on the truth of premises: A=B, B=C, therefore A=C

Inductive reasoning = makes broad generalizations from specific observations
- DANGER: allows a conclusion to be false if the premises are true
- generate hypotheses and use DR for answering specific questions

Abductive reasoning = inference = to get the best explanation from an incomplete set of preconditions.
- Given a true conclusion and a rule, it attempts to select some possible premises that, if true also, may support the conclusion, though not uniquely.
- Example: "When it rains, the grass gets wet. The grass is wet. Therefore, it might have rained." This kind of reasoning can be used to develop a hypothesis, which in turn can be tested by additional reasoning or data.
Important Definition: Ground truth

- := information provided by direct observation (empirical evidence) in contrast to information provided by inference
  - Empirical evidence = information acquired by observation or by experimentation in order to verify the truth (fit to reality) or falsify (non-fit to reality).
  - Empirical inference = drawing conclusions from empirical data (observations, measurements)
  - Causal inference = drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect.
    - Causal inference is an example of causal reasoning.
Empirical Inference Example

\[ y = \sum_i a_i k(x, x_i) + b \]
\[ y = a \cdot x \]

Gottfried W. Leibniz (1646-1716)
Hermann Weyl (1885-1955)
Vladimir Vapnik (1936-)
Alexey Chervonenkis (1938-2014)
Gregory Chaitin (1947-)
High dimensionality (curse of dim., many factors contribute)
- Complexity (real-world is non-linear, non-stationary, non-IID *)
- Need of large top-quality data sets
- Little prior data (no mechanistic models of the data)
  - *) = Def.: a sequence or collection of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually independent

What makes it hard ... ?

Example 3.4 (Eye disease) There exists a rather effective treatment for an eye disease. For 99% of all patients, the treatment works and the patient gets cured \((B = 0)\); if untreated, these patients turn blind within a day \((B = 1)\). For the remaining 1%, the treatment has the opposite effect and they turn blind \((B = 1)\) within a day. If untreated, they regain normal vision \((B = 0)\).

Which category a patient belongs to is controlled by a rare condition \((N_B = 1)\) that is unknown to the doctor, whose decision whether to administer the treatment \((T = 1)\) is thus independent of \(N_B\). We write it as a noise variable \(N_T\).

Assume the underlying SCM

\[
\mathcal{C} : \begin{align*}
T & := N_T \\
B & := T \cdot N_B + (1 - T) \cdot (1 - N_B)
\end{align*}
\]

with Bernoulli distributed \(N_B \sim \text{Ber}(0.01)\); note that the corresponding causal graph is \(T \rightarrow B\).

Now imagine a specific patient with poor eyesight comes to the hospital and goes blind \((B = 1)\) after the doctor administers the treatment \((T = 1)\). We can now ask the counterfactual question “What would have happened had the doctor administered treatment \(T = 0\)?” Surprisingly, this can be answered. The observation \(B = T = 1\) implies with \((3.5)\) that for the given patient, we had \(N_B = 1\). This, in turn, lets us calculate the effect of \(\text{do}(T := 0)\).

To this end, we first condition on our observation to update the distribution over the noise variables. As we have seen, conditioned on \(B = T = 1\), the distribution for \(N_B\) and the one for \(N_T\) collapses to a point mass on 1, that is, \(\delta_1\). This leads to a modified SCM:
\[ \mathcal{C}|B = 1, T = 1: \begin{array}{c}
T := 1 \\
B := T \cdot 1 + (1 - T) \cdot (1 - 1) = T
\end{array} \tag{3.6} \]

Note that we only update the noise distributions; conditioning does not change the structure of the assignments themselves. The idea is that the physical mechanisms are unchanged (in our case, what leads to a cure and what leads to blindness), but we have gleaned knowledge about the previously unknown noise variables for the given patient.

Next, we calculate the effect of \( do(T = 0) \) for this patient:

\[ \mathcal{C}|B = 1, T = 1; do(T := 0): \begin{array}{c}
T := 0 \\
B := T
\end{array} \tag{3.7} \]

Clearly, the entailed distribution puts all mass on \((0, 0)\), and hence

\[ p_{\mathcal{C}|B=1,T=1;do(T:=0)}(B = 0) = 1. \]

This means that the patient would thus have been cured \((B = 0)\) if the doctor had not given him treatment, in other words, \( do(T := 0) \). Because of

\[ p_{\mathcal{C};do(T:=1)}(B = 0) = 0.99 \quad \text{and} \]
\[ p_{\mathcal{C};do(T:=0)}(B = 0) = 0.01, \]

however, we can still argue that the doctor acted optimally (according to the available knowledge). \( \square \)
Interestingly, Example 3.4 shows that we can use counterfactual statements to falsify the underlying causal model (see Section 6.8). Imagine that the rare condition $N_B$ can be tested, but the test results take longer than a day. In this case, it is possible that we observe a counterfactual statement that contradicts the measurement result for $N_B$. The same argument is given by Pearl [2009, p.220, point (2)]. Since the scientific content of counterfactuals has been debated extensively, it should be emphasized that the counterfactual statement here is falsifiable because the noise variable is not unobservable in principle but only at the moment when the decision of the doctor has to be made.

Remember: Medical Action = Decision Making
Search Task in \( H \) Problem: Time \( (t) \)
Why is explainability important for ethical responsible AI?
Do not be confused by the confusion matrix

How would you interpret this results?


a.holzinger@human-centered.ai
Why (!) are 8% misclassifications?
How can we map these two effectively?

- **Causability**: a property of a person (Human)
- **Explainability**: a property of a system (Computer)
We need effective tools for Human-AI Interaction

Why did the algorithm do that?
Can I trust these results?
How can I correct an error?

A possible solution

The domain expert can understand why ...
The domain expert can learn and correct errors ...
The domain expert can re-enact on demand ...

Input data

Explanation Interface

Explainable Model

The domain expert can understand why ...
The domain expert can learn and correct errors ...
The domain expert can re-enact on demand ...
02 Definitions
Automatic-Automated-Autonomous
Not our Goal: Humanoid AI

Humanoid AI ≠ Human-Level AI
Best practice examples of aML ...
Explanations in Recommender Systems


Fully automatic autonomous vehicles ("Google car")

Autonomous aerial vehicle (AAV): passenger drone


http://www.ehang.com/ehang184/
Transfer of responsibility to the machine

SAE International J3016_201806: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles [http://www.sae.org/standards/content/j3016_201806]

SAE = Society of Automotive Engineers

a.holzinger@human-centered.ai
Achieving Autonomy: commercial drones

[Diagram showing different levels of autonomy for drones, from Level 0 (No Automation) to Level 5 (Full Automation), with corresponding human dependency levels from Trained & Skilled Pilot to No Operator.]


03 AI ethics: Legal accountability and moral dilemmas
Let’s start with a statement ...

*This is your machine learning system?*

YUP! You pour the data into this big pile of linear algebra, then collect the answers on the other side.

What if the answers are wrong?

Just stir the pile until they start looking right.

Image Source: Randall Munroe https://xkcd.com
What is Ethics in general?

- Ethics = moral philosophy
- Recommending and defending concepts of right and wrong conduct.
- Three areas:
  - 1) Meta-ethics, concerning the theoretical meaning and reference of moral propositions, and how their truth values (if any) can be determined
  - 2) Normative ethics, concerning the practical means of determining a moral course of action
  - 3) Applied ethics, concerning what a person is obligated (or permitted) to do in a specific situation or a particular domain of action -> AI ethics

https://www.iep.utm.edu/ethics/
What is Ethics for us as Software Engineers?

- Ethics is a **practical discipline**
- It is the good things – It is the right things
- BUT: How do we define what is good?

**FROM KANT TO KIRK: ‘STAR TREK’S’ PHILOSOPHICAL ARGUMENTS**

Should you pull the lever to divert the trolley?

A. Holzinger@human-centered.ai

Last updated: 03-10-2019
Each student should try the MIT moral machine:

http://moralmachine.mit.edu/

Some Results from the MIT Moral Machine study

Example: Biomedical Ethics is well-established

### UNESCO's 15 Bioethical Principles

<table>
<thead>
<tr>
<th>Human dignity &amp; human rights</th>
<th>Benefit &amp; harm</th>
<th>Autonomy-individual responsibility</th>
<th>Consent</th>
<th>Persons without the capacity to consent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human vulnerability &amp; personal integrity</td>
<td>Privacy / Confidentiality</td>
<td>Equality, Justice, Equity</td>
<td>Non-discrimination</td>
<td>Respect for cultural diversity</td>
</tr>
<tr>
<td>Solidarity &amp; cooperation</td>
<td>Social responsibility &amp; health</td>
<td>Sharing of benefits</td>
<td>Protecting future generations</td>
<td>Protecting biodiversity, biosphere &amp; environment</td>
</tr>
</tbody>
</table>

Independent review and approval by ethics board:

1) Informed consent
2) Risk-Benefit ratio and minimization of risk
3) Fair selection of study population (inclusion-, exclusion-criteria)
4) Scientific validity (‘scholarly review’)
5) Social value
6) Respect for participants and study communities
7) Confidentiality and privacy, data security
8) No Conflict of interest
Now, Why do we need AI Ethics? To ensure ...

Accountability ... we have to take responsibility for our developments, governments have to take responsibility for decisions and laws affecting all citizens

Trust ... confidence in the reliability, truth, ability (a trustee holds the property as its nominal owner for the good of beneficiaries

Transparency ... implies openness, communication, accountability, trust, ...

Understandability ... property of a system according to the principles of usability, we can say it is a kind of domain usability, and can be perceived as the relation and good fit between the “language of the human” and the “language of the machine”
First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm.

Second Law: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.


(Some) Grand Questions of AI Ethics

- Is it morally justified to create super-intelligent systems?
- Should our AI have any free will? And if it is possible: Can we prevent them from having free will?
- Will AI have consciousness? (Strong AI)
  - If so, will it they accept to be constrained by artificial AI-ethics placed on them by humans?
  - If AI develop their own ethics and morality, will we like what they do with us?

https://www.wired.com/story/will-ai-achieve-consciousness-wrong-question
For student discussion: What about existing AI?
For student discussion: current AI in robotics?

http://www.rob.cs.tu-bs.de/teaching/courses/seminar/Laufen_Mensch_vs_Roboter/

If the robot looks like a human, do we have different expectations?

Would you “kill” a robot car?

Would you “kill” a robot insect that would react by squeaky noises and escape in panic?

Would you “kill” a robot biped that would react by begging you to save his life?
04 AI ethics: Algorithms and the proof of explanations
Simpson’s Paradox – Statistics can not help!

https://www.youtube.com/watch?v=ebEkn-BiW5k

The Yule-Simpson effect describes the paradox that a trend which appears in several different groups of data disappears or reverses when these groups are combined, often in computational sociology or in medical science statistics. The paradox can only be resolved when causal relations are appropriately addressed in the statistical modeling.

Which scenarios invite reversals?

Which scenarios invite reversals?

Explainability of remote decisions

Explainability of remote decisions


\[ x = (x_l, \emptyset) \rightarrow y \]

Remote

\[ x = (x_l, x_d) \rightarrow y \]

Remote

\[ (x_l, x_d) \rightarrow y \]

PR\(C, (x_l, x_d), y \rightarrow C'\)

s.t. \(C'(x_l) = y\)

Explainability of remote decisions

- Explainability in a remote context is propagated as the society’s demand for transparency facing automated decisions
- It is unwise to blindly trust those explanations:
- Similar to humans, algorithms can easily hide the true motivations of a decision when “asked”.
- Consequently a huge future research direction is to develop secure schemes in which the involved parties can trust the exchanged information about decisions and their explainability, as enforced by new protocols!!

05 Responsible AI
Examples from Computational Sociology (Bias, fairness, ...)

Last updated: 03-10-2019
Watch the Obama Interview on how artificial intelligence will affect our jobs:

https://human-centered.ai/2016/10/14/obama-on-humans-in-the-loop
Explainability is an enabler for ensuring ethical responsible AI ...
A man and his son were involved in a terrible accident and are rushed to the intensive care unit.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?
Cognitive Bias

Please have a look at the list of cognitive biases: https://en.wikipedia.org/wiki/List_of_cognitive_biases

- Biases in Interpretation:
  - Confirmation bias (favour info confirming beliefs)
  - Overgeneralization (similar to overfitting, e.g. a cat says all dogs have four legs therefore I am a dog)
  - Automatization bias (humans favour suggestions from machines)
  - **Correction fallacy** (most people confuse correlation with causation !!)

- Note: data driven AI learns from human data – which may result in bias network effects!

- Bias can be bad, good, neutral (or unknown)

Algorithmic bias

- Results from ML algorithms can be
  - unfair,
  - resulting in prejudicial treatment of people e.g. with regard to gender, race, income, sexual orientation, religion, occupation, origin, ... 

- Bias is resulting from many issues, e.g.
  - Data quality, distortions in demographics, behavioural aspects, linking biases, etc. etc., please have a read of this paper:

Inclusive Images Competition of Google


a.holzinger@human-centered.ai
Recommender Systems/Context Aware Systems need explanation


a.holzinger@human-centered.ai

67

Last updated: 03-10-2019
Example: Predicting Policing

Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?
The software is supposed to make policing more fair and accountable. But critics say it still has a way to go.


What you should always remember:

- **Data matters most:**
  - Understand your data – have a look at the raw data, do not shuffle the data (look for skewness, etc.)
  - Combine inputs from multiple sources
  - Use technics for bias mitigation, e.g.


Updated versions: https://arxiv.org/abs/1803.09010
Alexa, what about legal aspects of AI?
Legal aspects of AI

Example: AI recruiting system

What about autonomous robotic surgery from legal aspects (such as civil law, international law, tort law, liability, medical malpractice, privacy and product/device legislation)?

Responsibility can be classified into the following: (1) Accountability; (2) Liability; and (3) Culpability.

Culpability is unthinkable in the current state of technology.

Similar problems as with autonomously driven vehicles.

Currently unsolved, much further research needed.
## I. Overview ➤ Right to explanation

### EU General Data Protection Regulation

<table>
<thead>
<tr>
<th>Article</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>17. Right to be forgotten</td>
<td>An individual to have certain data deleted so that third persons can no longer trace them</td>
</tr>
<tr>
<td>22. Automated individual decision making</td>
<td>The data subject shall have the right not to be subject to a decision based solely on automated processing (including profiling).</td>
</tr>
<tr>
<td>13-14. Right to explanation</td>
<td>A data subject has the right to “meaningful information about the logic involved.”</td>
</tr>
<tr>
<td>EU administration</td>
<td>When violated 4% of global revenue will be fined.</td>
</tr>
<tr>
<td>Enact</td>
<td>May 28th, 2018</td>
</tr>
</tbody>
</table>
Conclusion
Engineers create a set of logical rules to represent knowledge (Rule based Expert Systems)

- Advantage: works well in narrowly defined problems of well-defined domains (narrow reasoning)
- Disadvantage: No adaptive learning behaviour and poor handling of $p(x)$

Image credit to John Launchbury
Engineers create learning models for specific tasks and train them with “big data” (e.g. Deep Learning)

- **Advantage:** works well for standard classification tasks and has prediction capabilities
- **Disadvantage:** No contextual capabilities and minimal reasoning abilities
The third wave of AI (?): Adaptive Context Understanding

- A contextual model can perceive, learn and understand and abstract and reason
- Advantage: can use transfer learning for adaptation on unknown unknowns
- Disadvantage: Superintelligence ...

Image credit to John Launchbury
Three (selected) dangers and myths about AI generally

- Myth 1a: Superintelligence by 2100 is inevitable!
- Myth 1b: Superintelligence by 2100 is impossible!
- Fact: We simply don’t know it!
- Myth 2: Robots are our main concern
  Fact: Cyberthreats are the main concern: it needs no body – only an Internet connection
- Myth 3: AI can never control us humans
  Fact: Intelligence is an enabler for control: We control tigers by being smarter ...

https://futureoflife.org/ai-principles
Interactive Machine Learning: Human is seen as an agent involved in the actual learning phase, step-by-step influencing measures such as distance, cost functions ...

Thank you!