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

- 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, …)

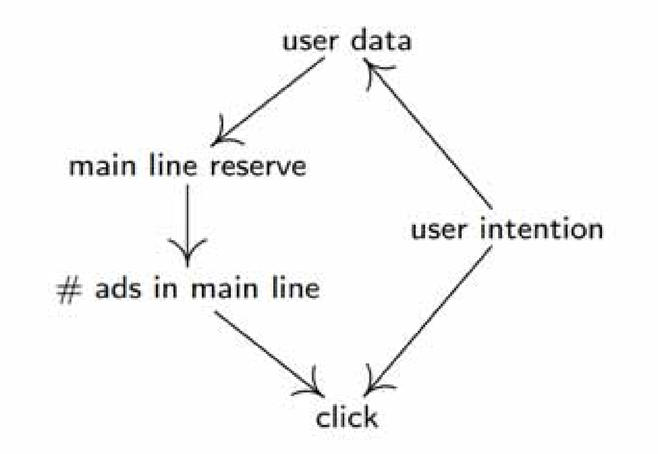
01 Repetition: From Causality to Ethical Responsibility

- 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 …

Judea Pearl 2009. Causal inference in statistics: An overview. Statistics surveys, 3, 96-146

Judea Pearl, Madelyn Glymour & Nicholas P. Jewell 2016. Causal inference in statistics: A primer, John Wiley & Sons.

- Hume again: "… if the first object had not been, the second never had existed …"
- Causal inference as a missing data problem
- $x_i := f_i(ParentsOfi, Noisei)$
- Interventions can only take place on the right side



Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard & Ed Snelson 2013. Counterfactual reasoning and learning systems: The example of computational advertising. The Journal of Machine Learning Research, 14, (1), 3207-3260.

Dependence vs. Causation

Storks Deliver Babies (*p*= 0.008)

Robert Matthews

Article first published online: 25 DEC 2001

DOI: 10.1111/1467-9639.00013

Teaching Statistics Trust, 2000

Issue

Teaching Statistics

Volume 22, Issue 2
38, June 2000

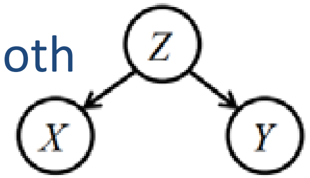
Country	Area (km^2)	Storks (pairs)	Humans (10^6)	Birth rate (10^3/yr)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56	774
Germany	357,000	3300	78	901
Greece	132,000	2500	10	106
Holland	41,900	4	15	188
Hungary	93,000	5000	11	124
Italy	301,280	5	57	551
Poland	312,680	30,000	mailto:rajm@compuserve.com	
Portugal	92,390	1500	10	120
Romania	237,500	5000	23	367
Spain	504,750	8000	39	439
Switzerland	41,290	150	6.7	82
Turkey	779,450	25,000	56	1576

Table 1. Geographic, human and stork data for 17 European countries

Robert Matthews 2000. Storks deliver babies (p= 0.008). Teaching Statistics, 22, (2), 36-38.

Correlation does not tell anything about causality!

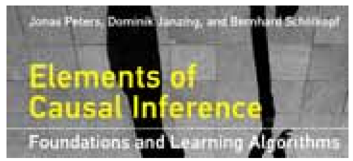
- Hans Reichenbach (1891-1953): **Common Cause Principle**
- Links causality with probability:
  - If X and Y are statistically dependent, there is a Z influencing both
  - Whereas:
  - A, B, … events
  - X, Y, Z random variables
  - P … probability measure
  - Px … probability distribution of X
  - 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.

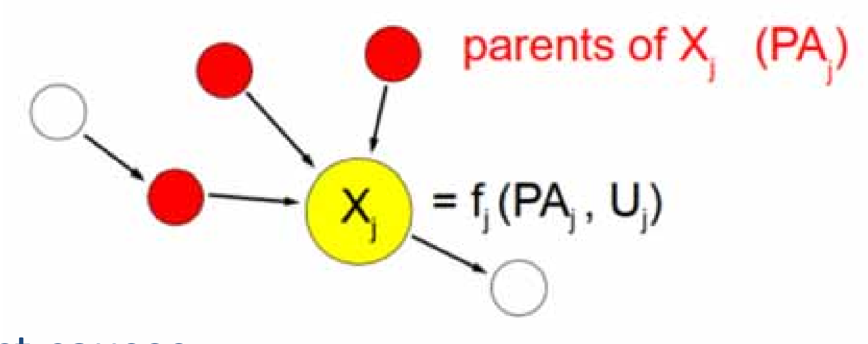
https://plato.stanford.edu/entries/physics-Rpcc/

For details please refer to the excellent book of: Jonas Peters, Dominik Janzing & Bernhard Schölkopf 2017. Elements of causal inference: foundations and learning algorithms, Cambridge (MA). https://mitpress.mit.edu/books/elements-causal-inference



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
  - $x_i := f_i(ParentsOf_i, Noisei)$

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

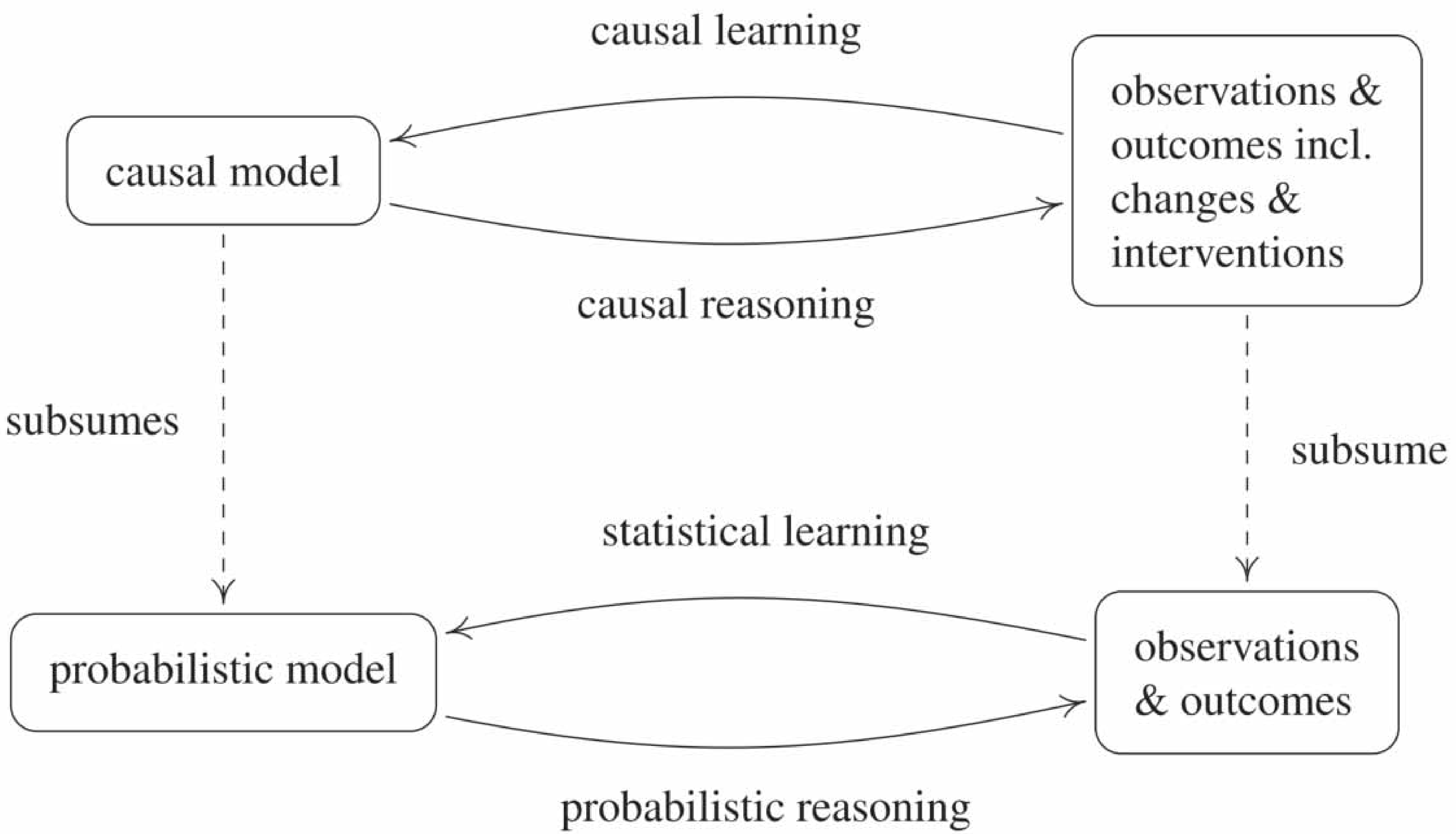
Question: Can we recover G from p ?
Answer: under certain assumptions, we can recover an equivalence class containing the correct G using conditional independence testing
But there are problems!

For details please refer to the excellent book of: Jonas Peters, Dominik Janzing & Bernhard Schölkopf 2017. Elements of causal inference: foundations and learning algorithms, Cambridge (MA). https://mitpress.mit.edu/books/elements-causal-inference

Explainability	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.
Causability	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.

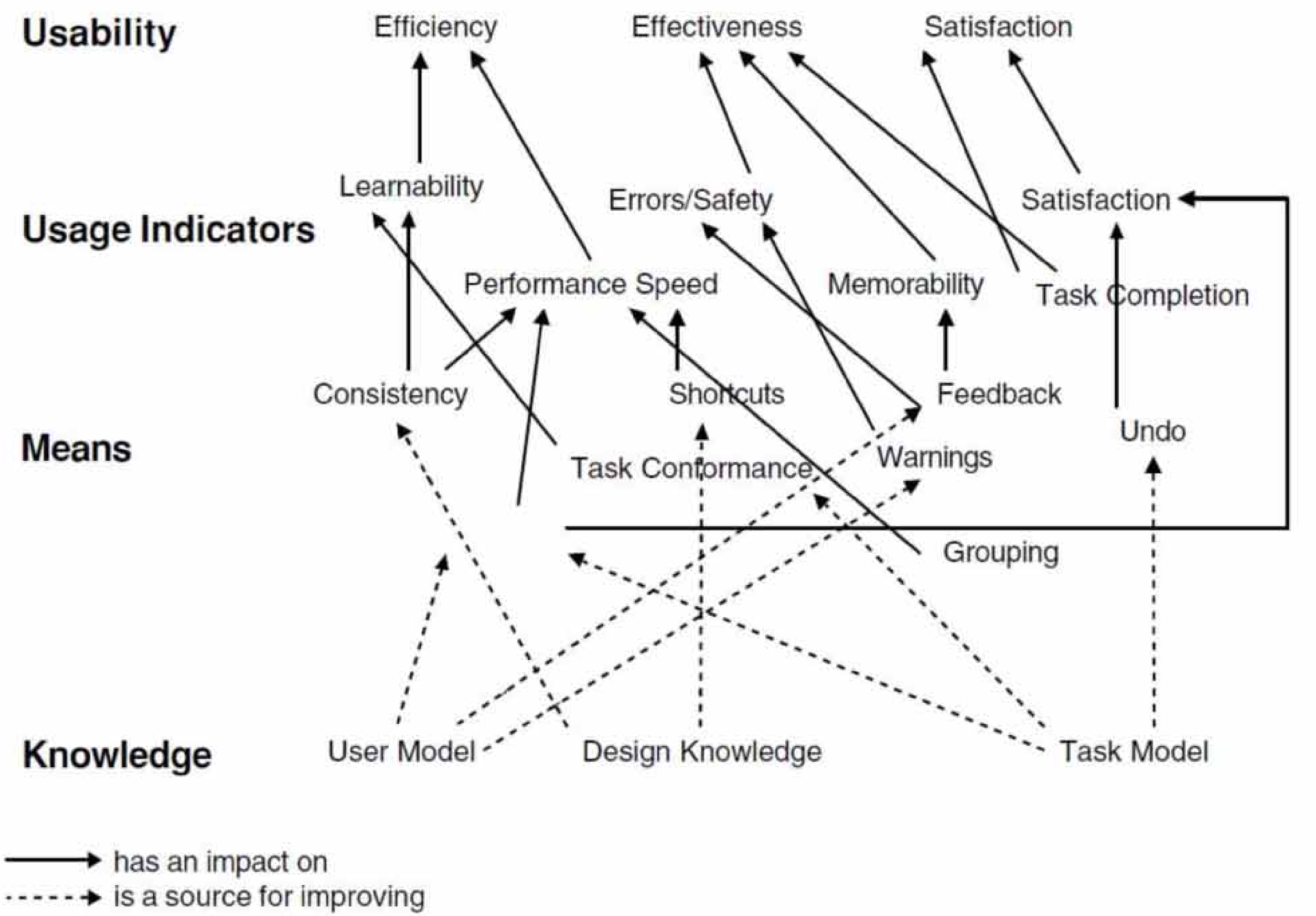
- **Causability := a property of a person, while**
- **Explainability := a property of a system**

Probabilistic vs. causal inference problems



Jonas Peters, Dominik Janzing & Bernhard Schölkopf 2017. Elements of causal inference: foundations and learning algorithms, Cambridge (MA).

Compare this with usability

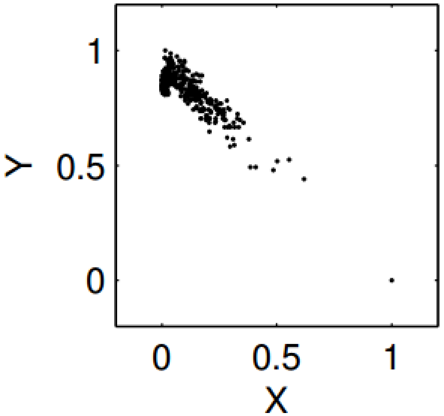


Veer, G. C. v. d. & Welie, M. v. (2004) DUTCH: Designing for Users and Tasks from Concepts to Handles. In: Diaper, D. & Stanton, N. (Eds.) *The Handbook of Task Analysis for Human-Computer Interaction. Mahwah (New Jersey), Lawrence Erlbaum, 155-173.*

- **"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 $Y \rightarrow X$

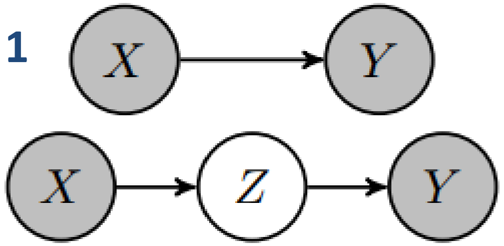
Bengio, Y., Courville, A. & Vincent, P. 2013. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35, (8), 1798-1828, doi:10.1109/TPAMI.2013.50.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. Science, 331, (6022), 1279-1285, doi:10.1126/science.1192788.



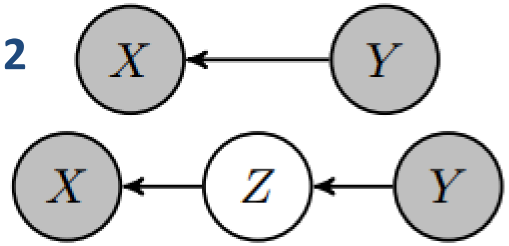
Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler & Bernhard Schölkopf 2016. Distinguishing cause from effect using observational data: methods and benchmarks. The Journal of Machine Learning Research, 17, (1), 1103-1204.

1



$$\mathbb{P}_Y \neq \mathbb{P}_{Y\,|\,\mathrm{do}(x)} = \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X = \mathbb{P}_{X\,|\,\mathrm{do}(y)} \neq \mathbb{P}_{X\,|\,y}$$

2

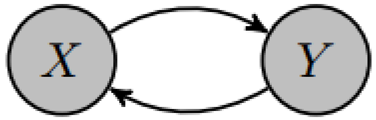


$$\mathbb{P}_Y = \mathbb{P}_{Y\,|\,\mathrm{do}(x)} \neq \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X \neq \mathbb{P}_{X\,|\,\mathrm{do}(y)} = \mathbb{P}_{X\,|\,y}$$

3

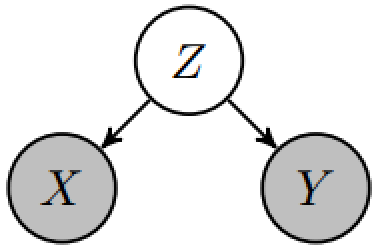
$$\mathbb{P}_Y = \mathbb{P}_{Y\,|\,\mathrm{do}(x)} = \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X = \mathbb{P}_{X\,|\,\mathrm{do}(y)} = \mathbb{P}_{X\,|\,y}$$

4



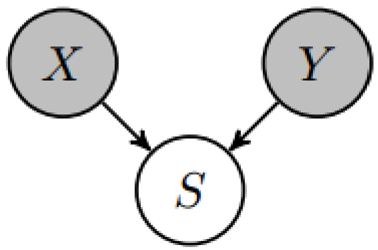
$$\mathbb{P}_Y \neq \mathbb{P}_{Y\,|\,\mathrm{do}(x)} \neq \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X \neq \mathbb{P}_{X\,|\,\mathrm{do}(y)} \neq \mathbb{P}_{X\,|\,y}$$

5



$$\mathbb{P}_Y = \mathbb{P}_{Y\,|\,\mathrm{do}(x)} \neq \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X = \mathbb{P}_{X\,|\,\mathrm{do}(y)} \neq \mathbb{P}_{X\,|\,y}$$

6



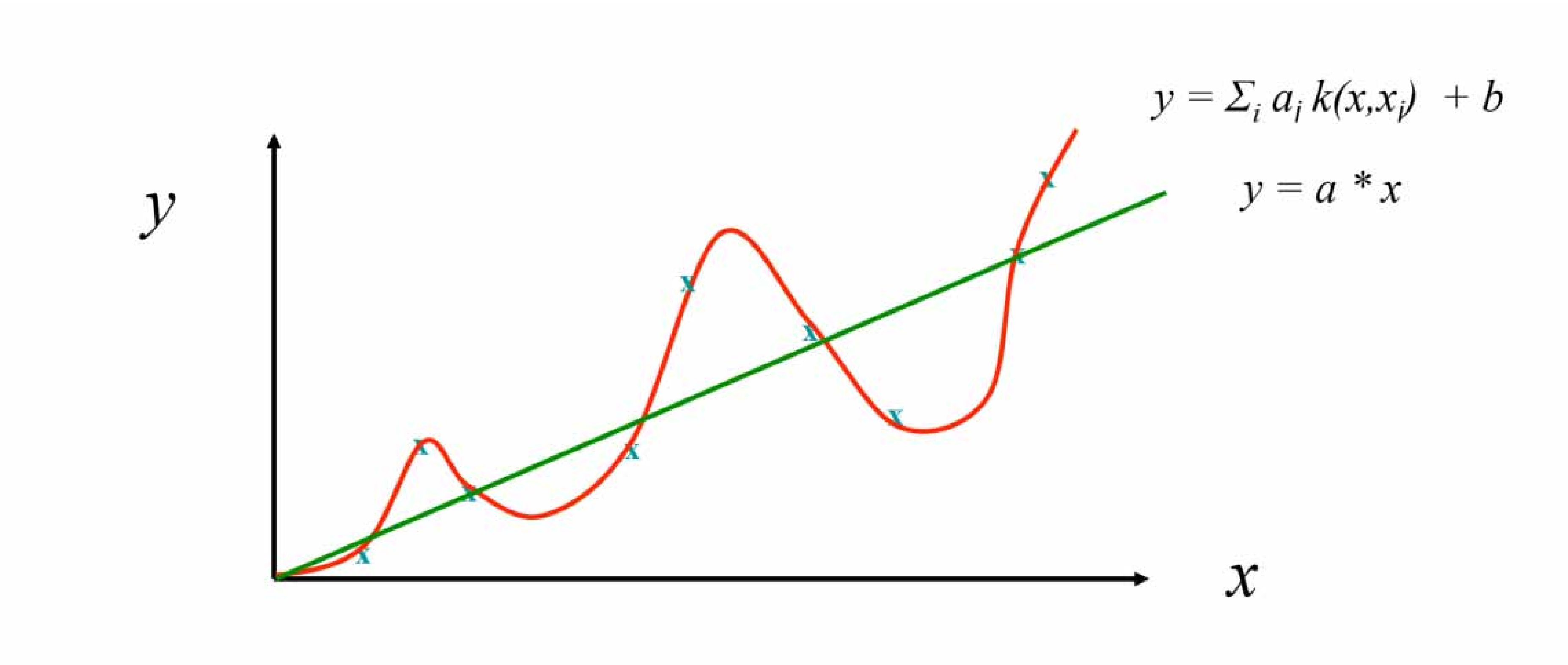
$$\mathbb{P}_{Y\,|\,s} \neq \mathbb{P}_{Y\,|\,\mathrm{do}(x),s} = \mathbb{P}_{Y\,|\,x,s}$$
$$\mathbb{P}_{X\,|\,s} \neq \mathbb{P}_{X\,|\,\mathrm{do}(y),s} = \mathbb{P}_{X\,|\,y,s}$$

Remember: Reasoning = “Sensemaking”

- **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.

- := 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



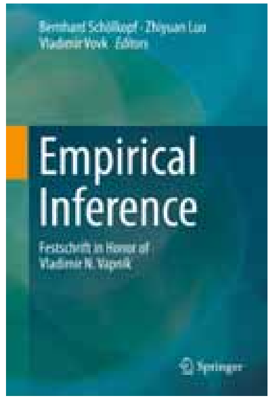
Gottfried W. Leibniz (1646-1716)

Hermann Weyl (1885-1955)

Vladimir Vapnik (1936-)

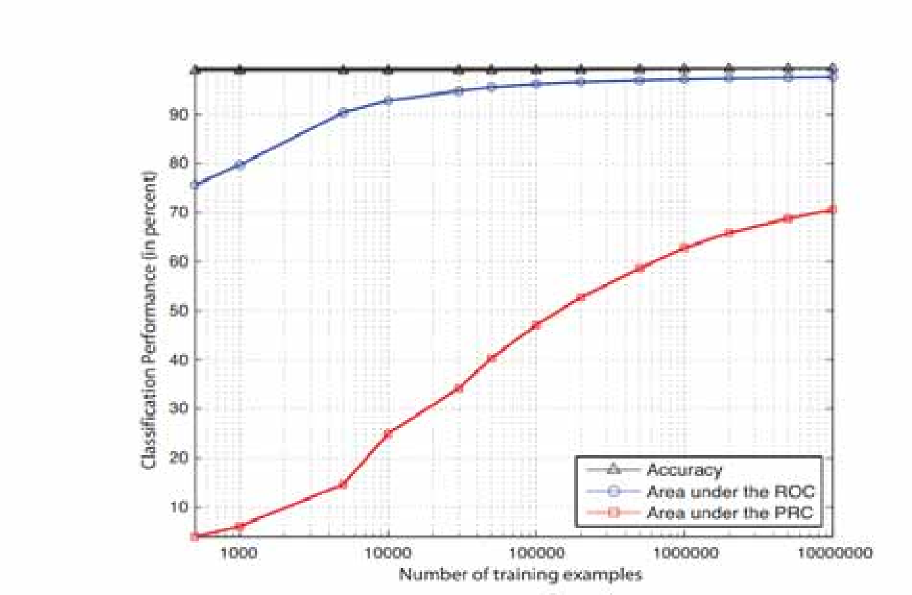
Alexey Chervonenkis (1938-2014)

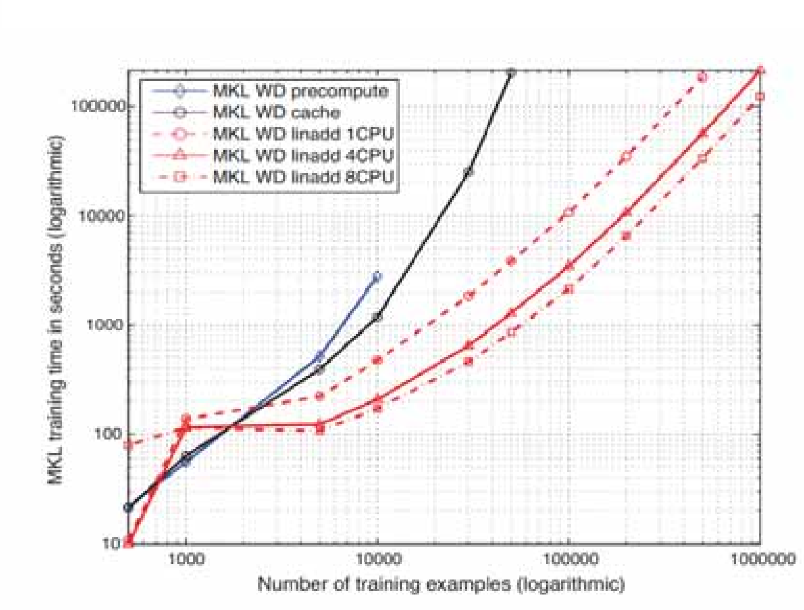
Gregory Chaitin (1947-)



Remember: hard inference problems

- 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





Sören Sonnenburg, Gunnar Rätsch, Christin Schaefer & Bernhard Schölkopf 2006. Large scale multiple kernel learning. Journal of Machine Learning Research, 7, (7), 1531-1565.

What makes it hard … ?

Jonas Peters, Dominik Janzing & Bernhard Schölkopf 2017. Elements of causal inference: foundations and learning algorithms, Cambridge (MA).

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

$$\mathfrak{C}: \begin{array}{lcl} T & := & N_T \\ B & := & T \cdot N_B + (1 - T) \cdot (1 - N_B) \end{array}$$

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

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 $do(T := 0)$.

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, δ_1. This leads to a modified SCM:

$$\mathfrak{C}|B=1,T=1: \begin{array}{lll} T & := & 1 \\ B & := & T\cdot 1+(1-T)\cdot(1-1)=T \end{array} \tag{3.6}$$

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:

$$\mathfrak{C}|B=1,T=1;do\,(T:=0): \begin{array}{lll} T & := & 0 \\ B & := & T \end{array} \tag{3.7}$$

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

$$P^{\mathfrak{C}|B=1,T=1;do(T:=0)}(B=0)=1\,.$$

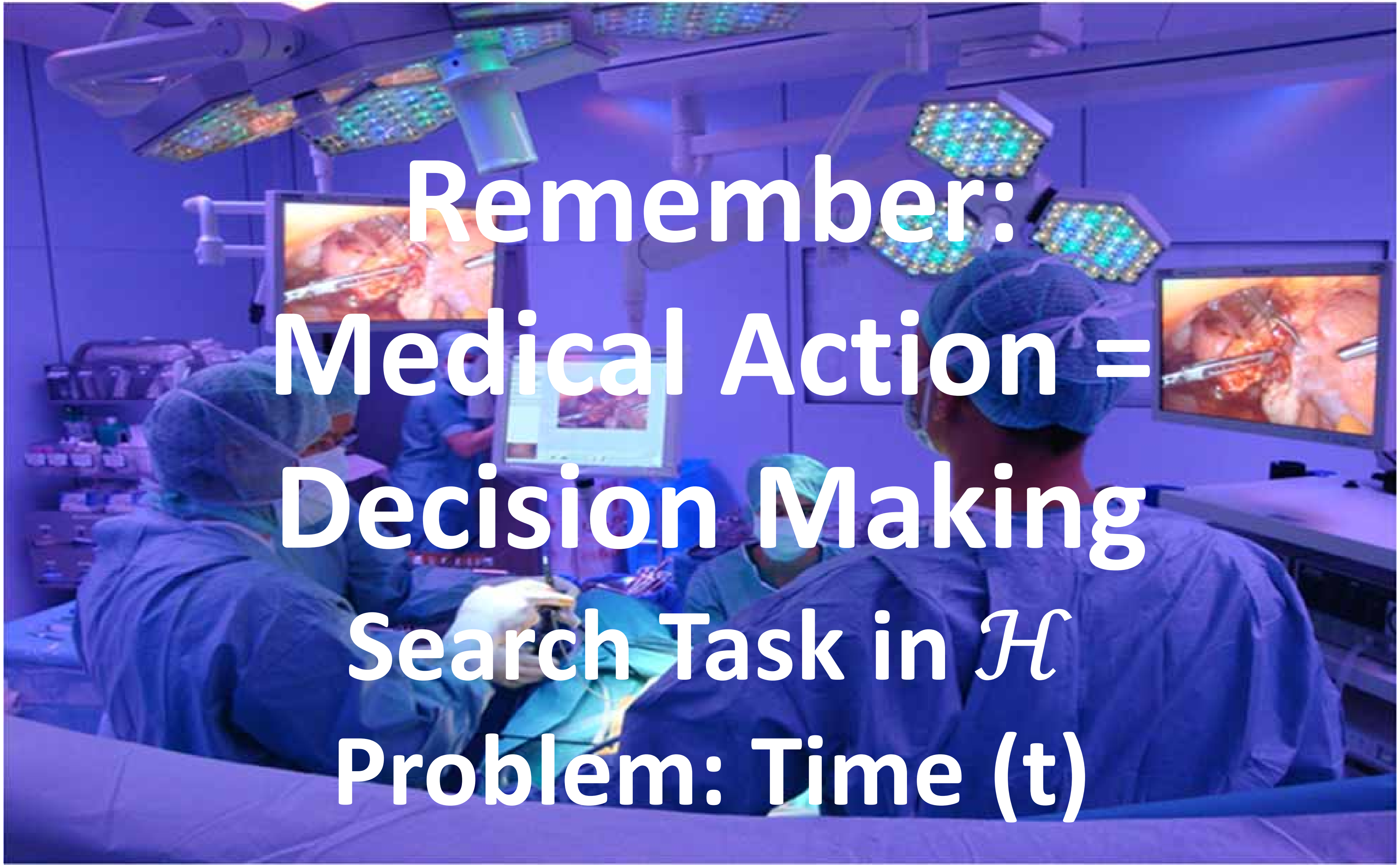
This means that the patient would thus have been cured ($B=0$) if the doctor had not given him treatment, in other words, $do\,(T:=0)$. Because of

$$P^{\mathfrak{C};do(T:=1)}(B=0)=0.99 \quad \text{and}$$
$$P^{\mathfrak{C};do(T:=0)}(B=0)=0.01,$$

however, we can still argue that the doctor acted optimally (according to the available knowledge). □

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.

Judea Pearl 2009. *Causality: Models, Reasoning, and Inference (2nd Edition)*, Cambridge, Cambridge University Press.



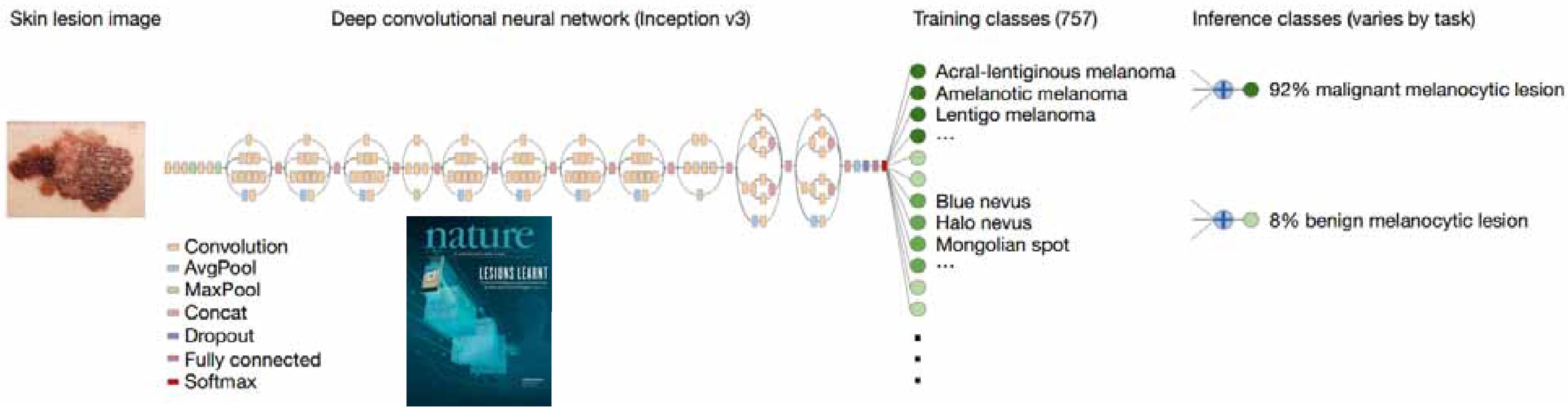
Remember: Medical Action = Decision Making

Search Task in $\mathcal{H}$

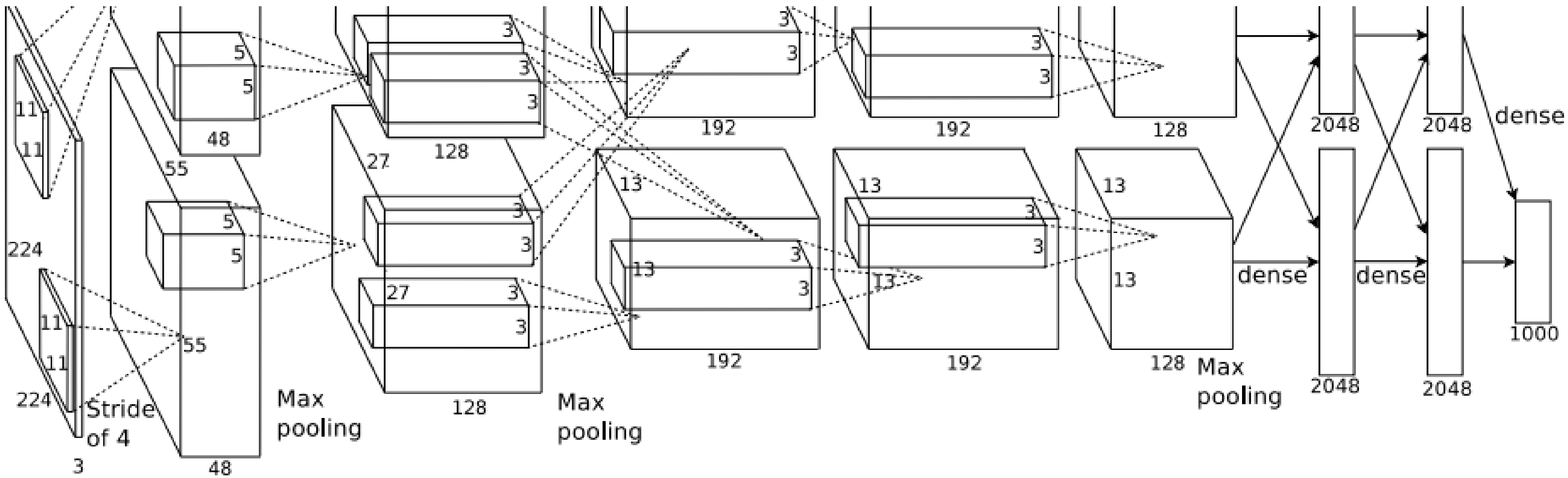
Problem: Time (t)

Why is explainability important for ethical responsible AI?

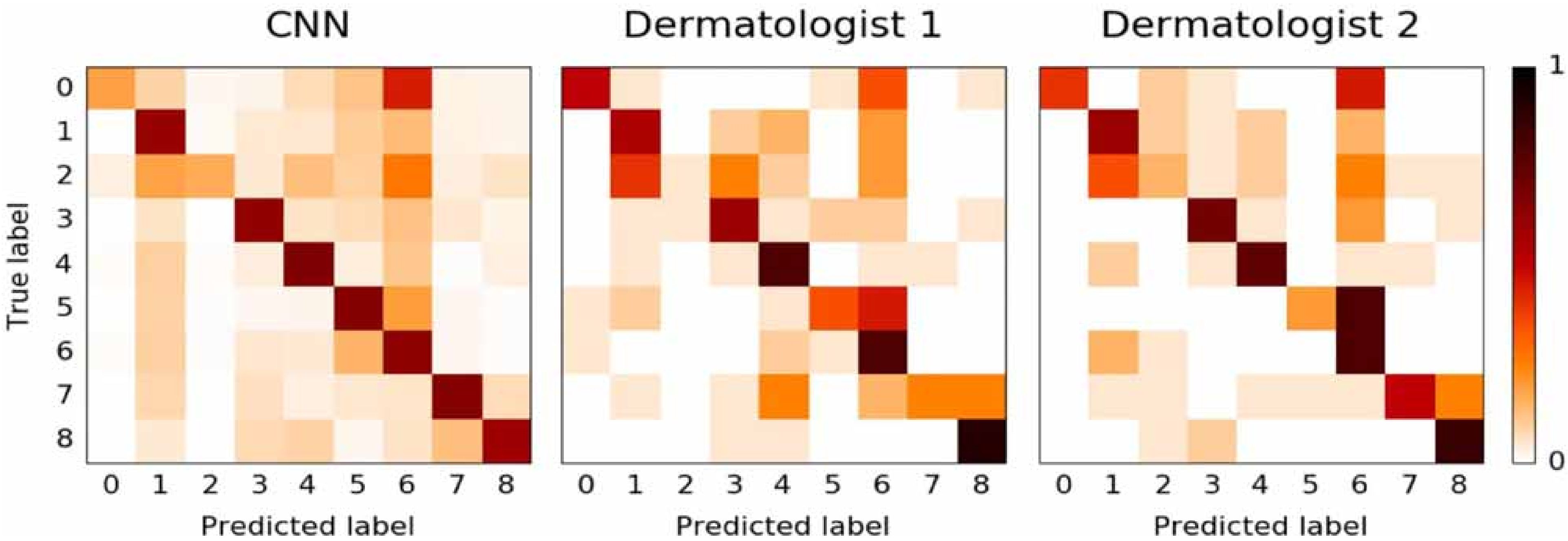
A very famous work as example …



Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. & Thrun, S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, (7639), 115-118, doi:10.1038/nature21056.

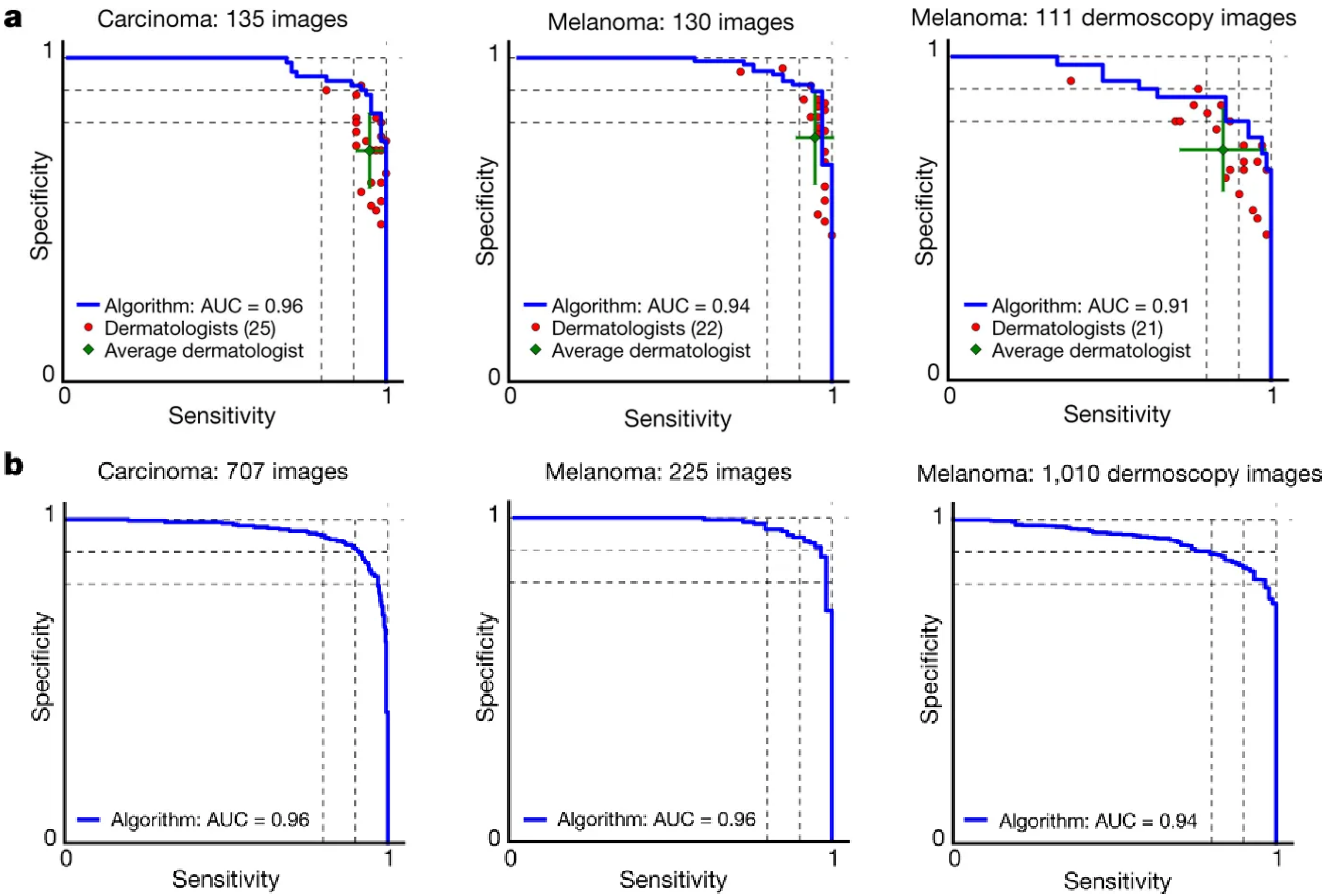


Do not be confused by the confusion matrix



Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, (7639), 115-118, doi:10.1038/nature21056.

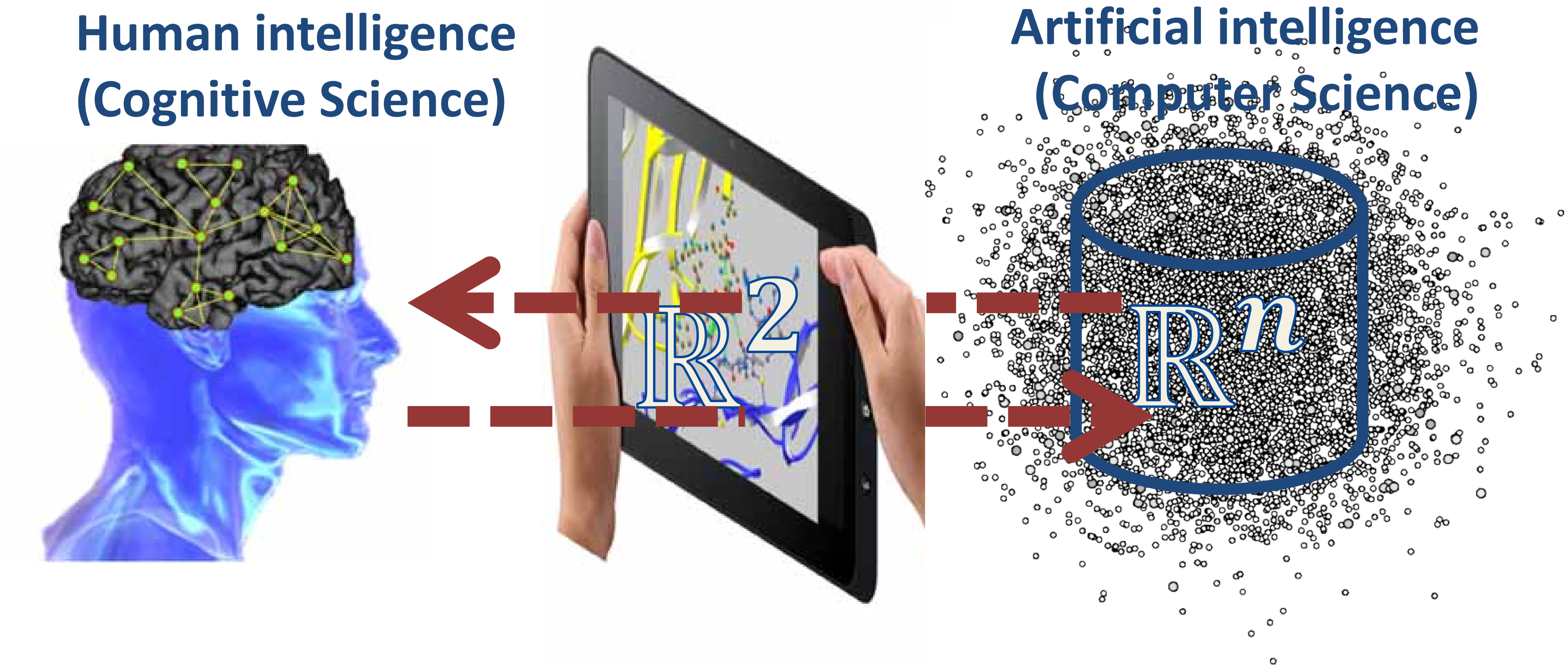
How would you interpret this results?



Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, (7639), 115-118, doi:10.1038/nature21056.

Why (!) are 8 % misclassifications?

- **Causability := a property of a person (Human)**
- **Explainability := a property of a system (Computer)**

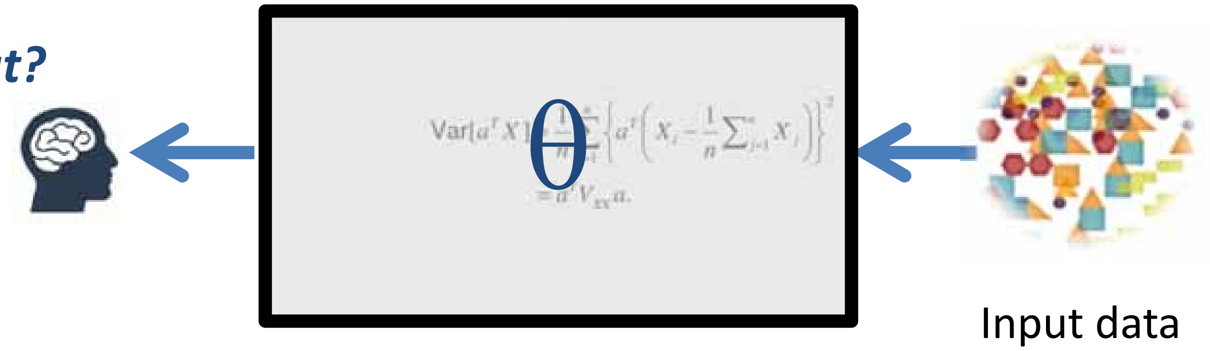


We need effective tools for Human-AI Interaction

Why did the algorithm do that?

Can I trust these results?

How can I correct an error?



A possible solution

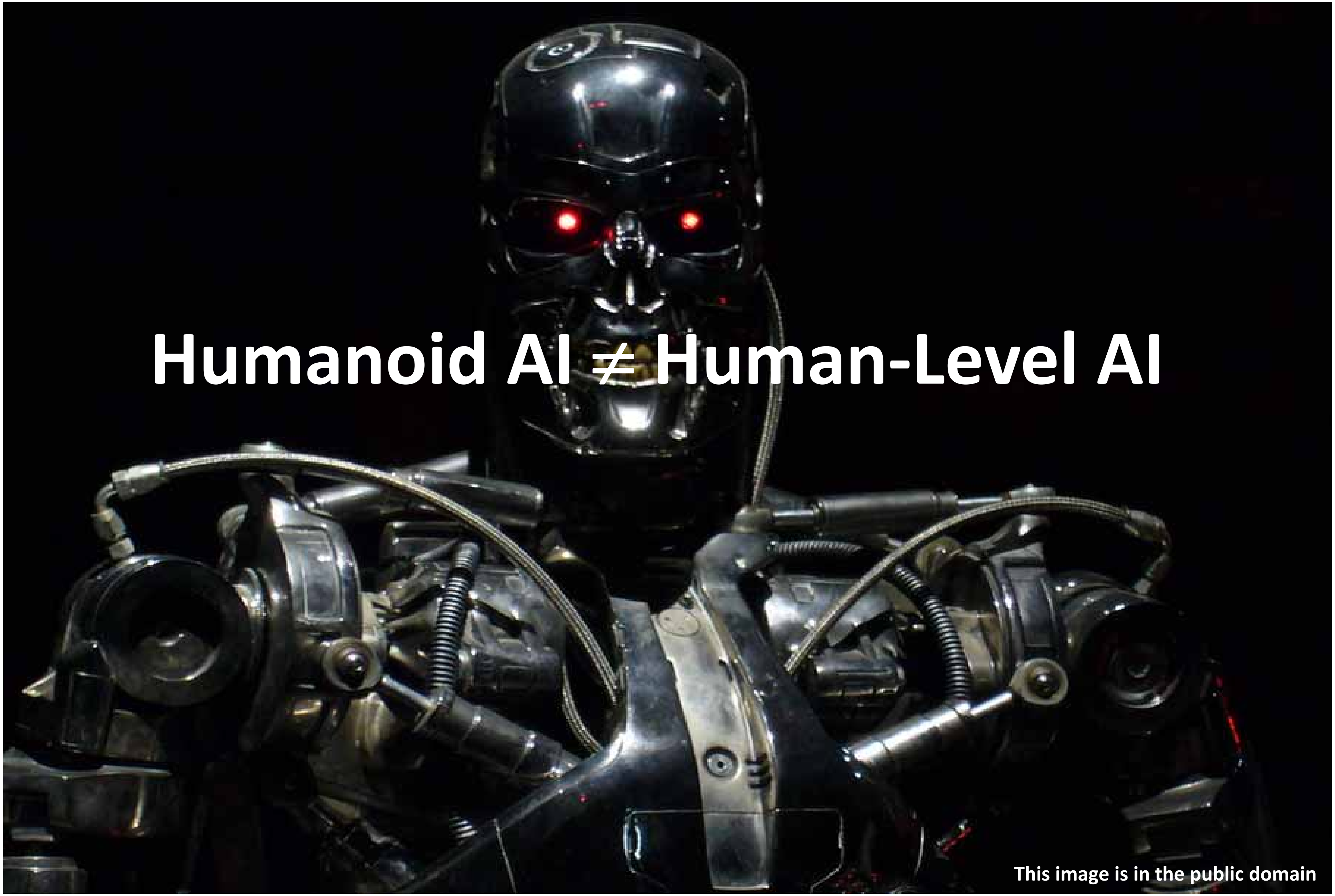
The domain expert can understand why …

The domain expert can learn and correct errors …

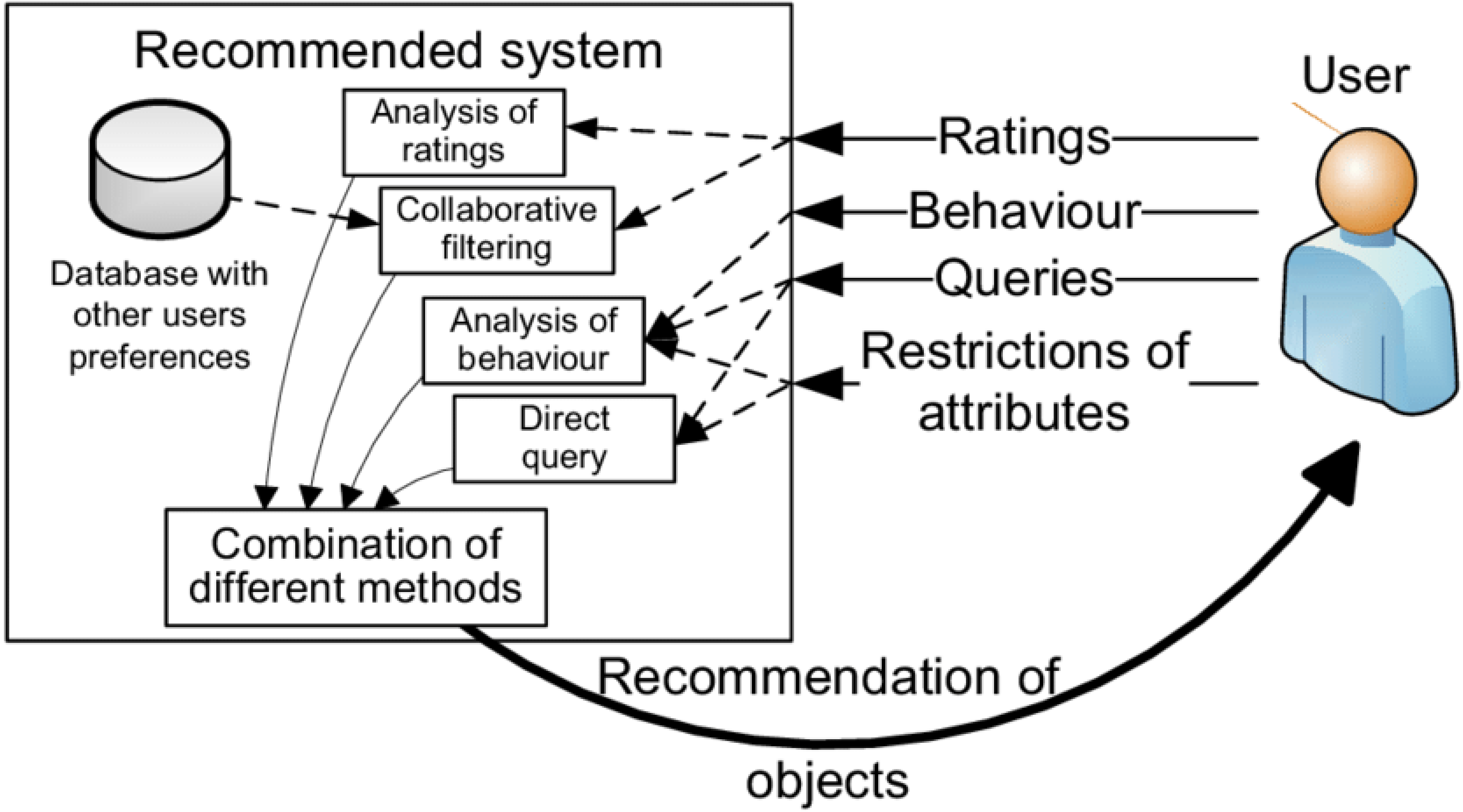
The domain expert can re-enact on demand …

02 Definitions Automatic-Automated-Autonomous

Not our Goal: Humanoid AI



Best practice examples of aML …



Alan Eckhardt 2009. Various aspects of user preference learning and recommender systems. DATESO. pp. 56-67.

Andre Calero Valdez, Martina Ziefle, Katrien Verbert, Alexander Felfernig & Andreas Holzinger 2016. Recommender Systems for Health Informatics: State-of-the-Art and Future Perspectives. In: Lecture Notes in Artificial Intelligence LNAI 9605. Heidelberg et. al.: Springer, pp. 391-414, doi:10.1007/978-3-319-50478-0_20.

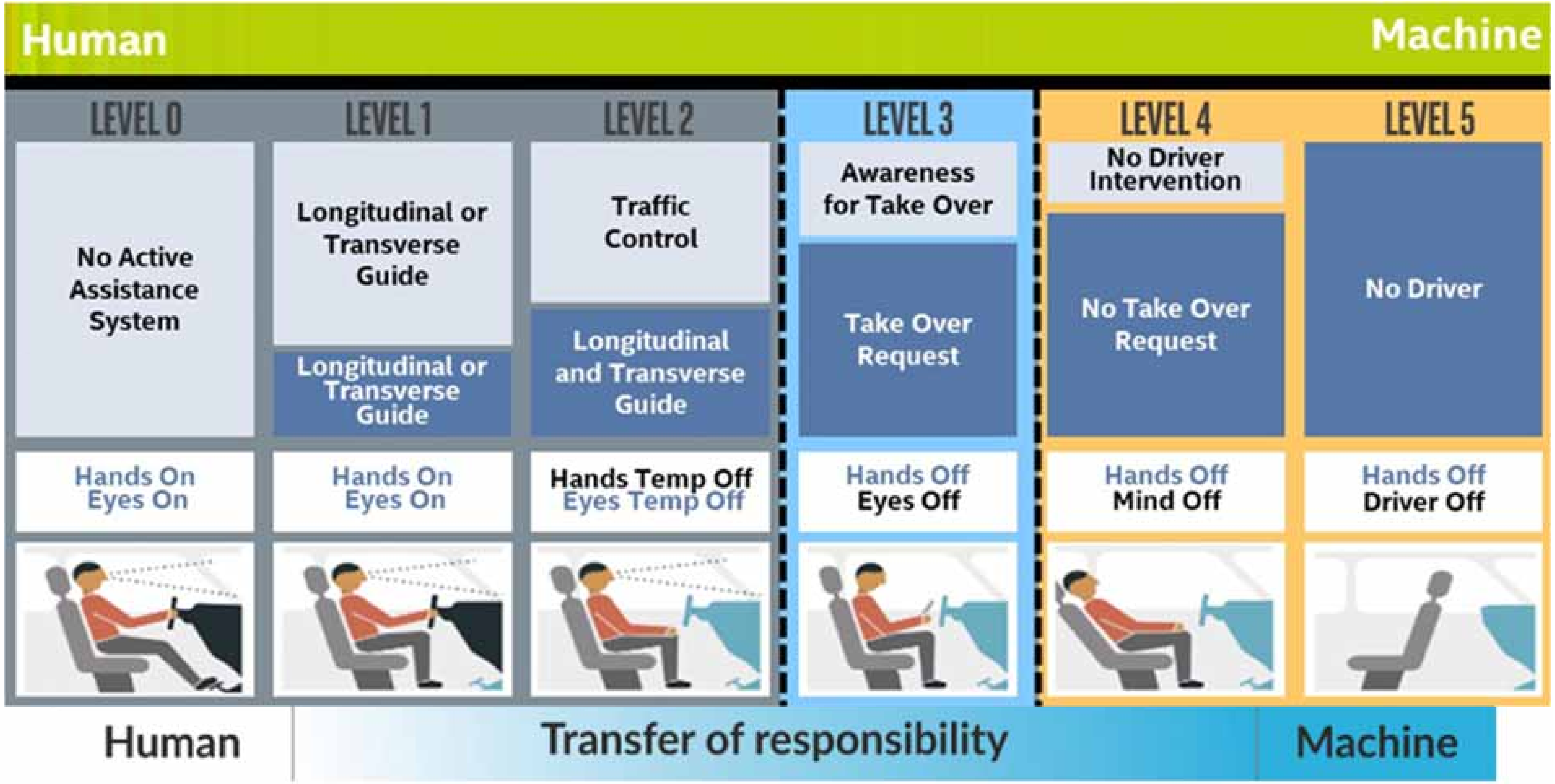
Fully automatic autonomous vehicles (“Google car”)

Guizzo, E. 2011. How google’s self-driving car works. IEEE Spectrum Online, 10, 18.

Autonomous aerial vehicle (AAV): passenger drone

https://www.businessinsider.sg/the-worlds-first-passenger-drone-makes-public-flight-in-china-and-you-could-soon-own-one

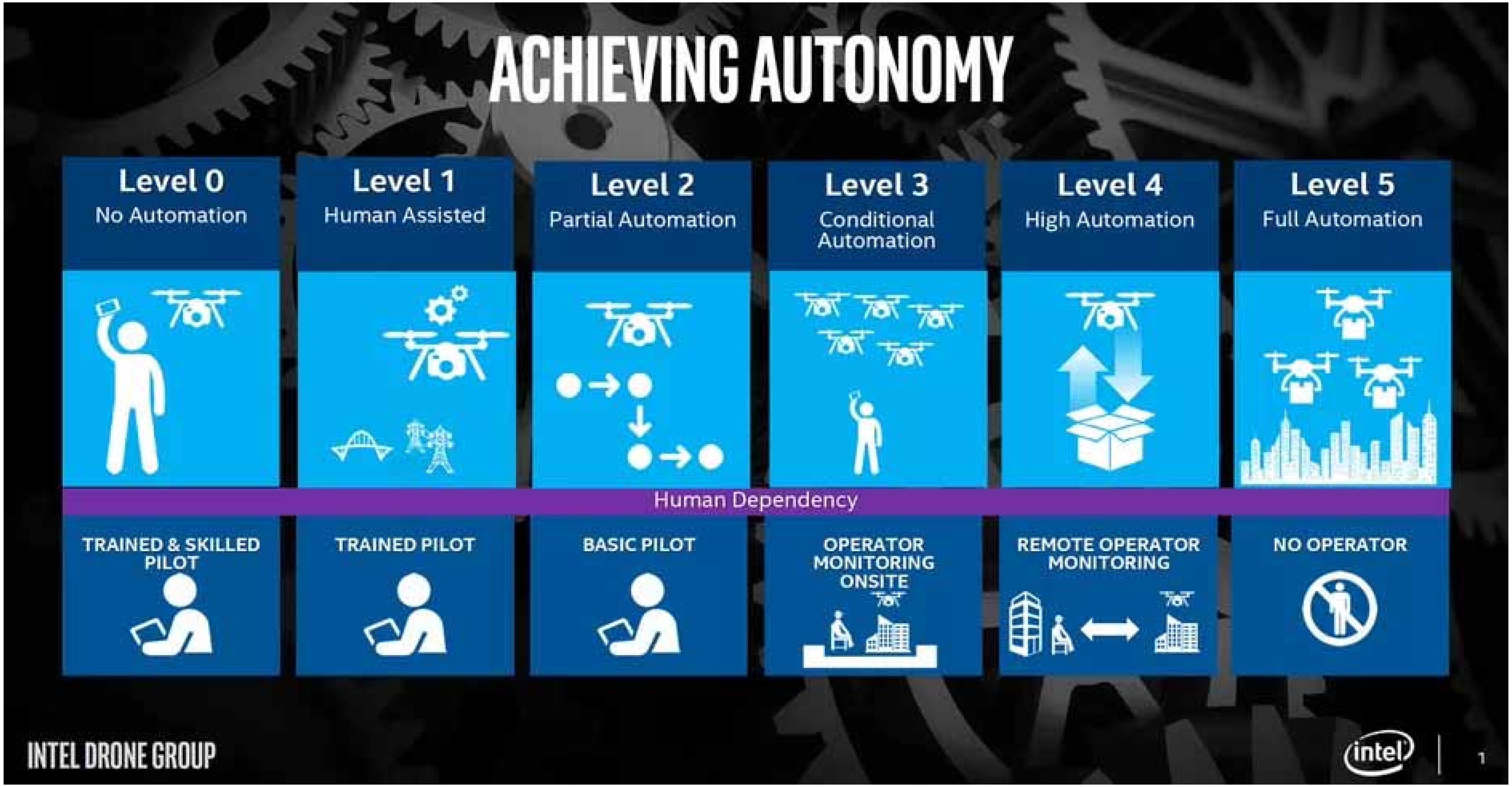
Transfer of responsibility to the machine



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

Achieving Autonomy: commercial drones



https://www.intel.com/content/www/us/en/automotive/autonomous-vehicles.html

https://www.intel.com/content/www/us/en/drones/drone-applications/commercial-drones.html

03 AI ethics: Legal accountability and moral dilemmas

Let's start with a statement …

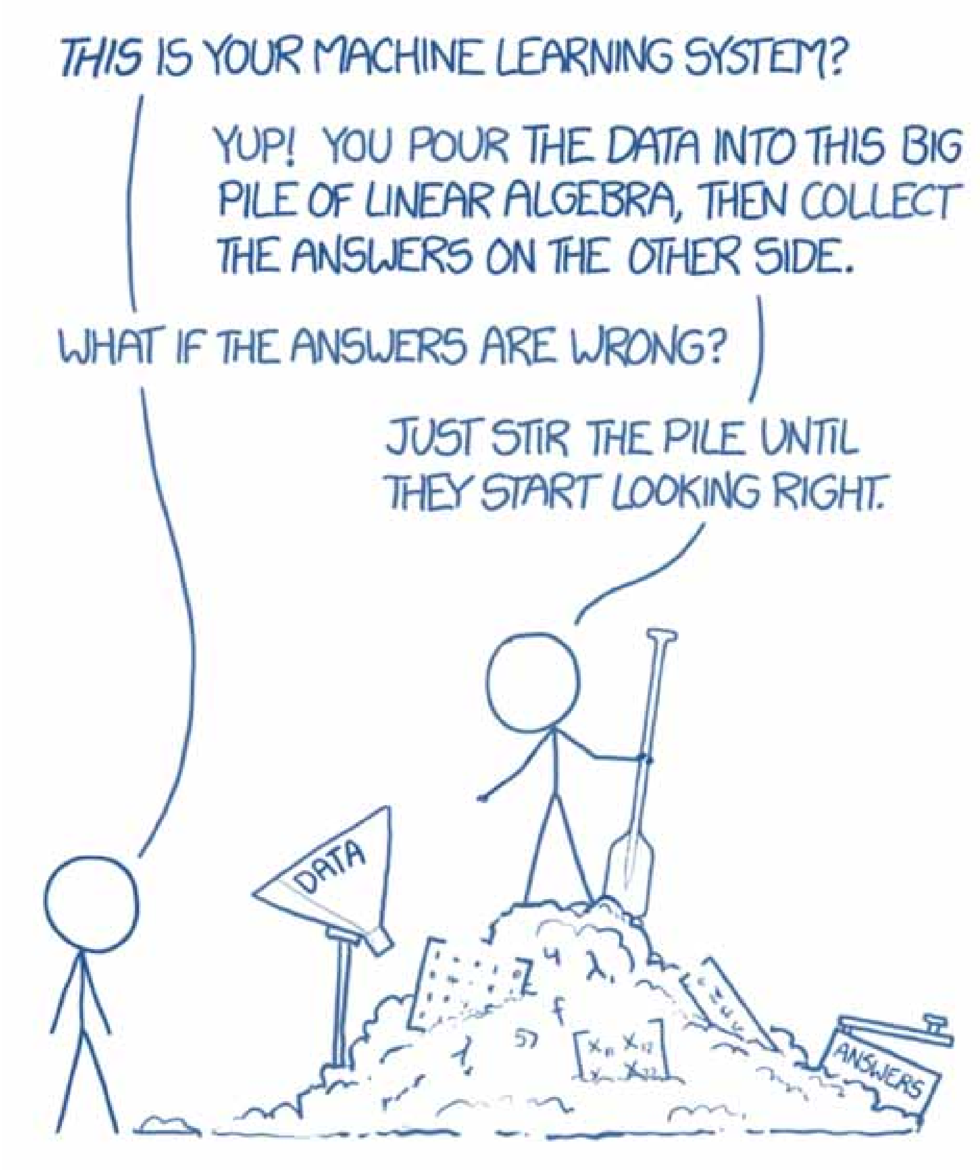


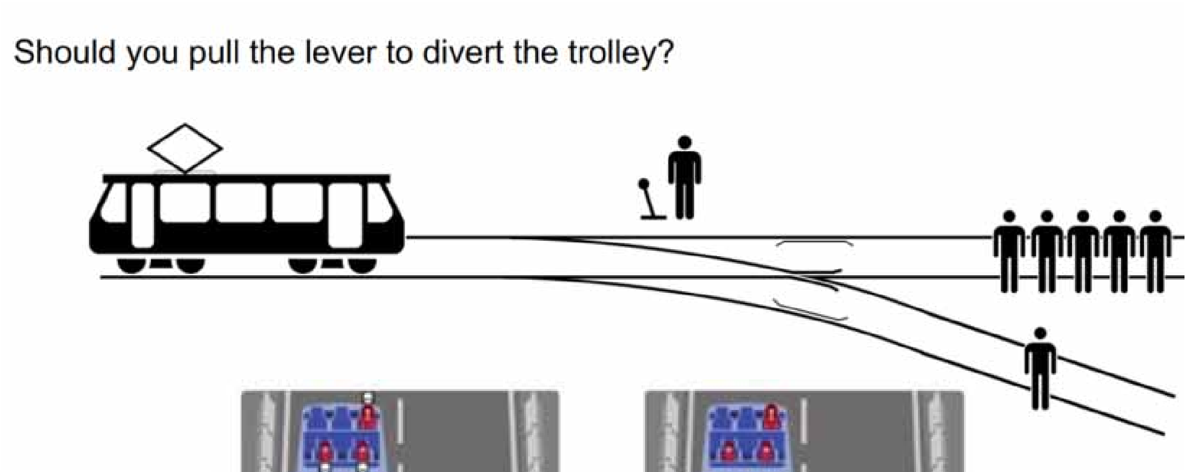
Image Source: Randall Munroe https://xkcd.com

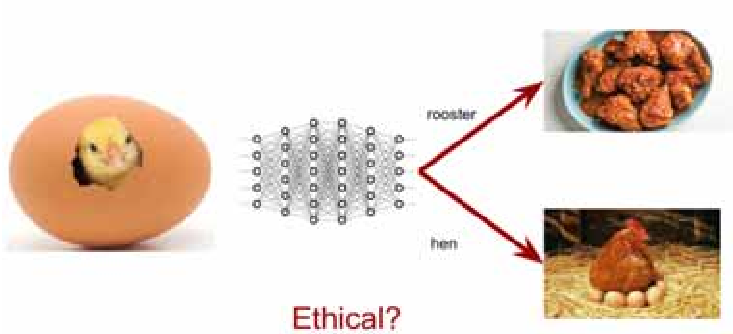
- Ethics = moral philosophy
- Recommending and defending concepts of right and wrong conduct.
- Three areas:
- 1) Meta-ethics, concerning the theoretical meaning and reference of moral propositions, and how their truth values (if any) can be determined
- 2) Normative ethics, concerning the practical 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

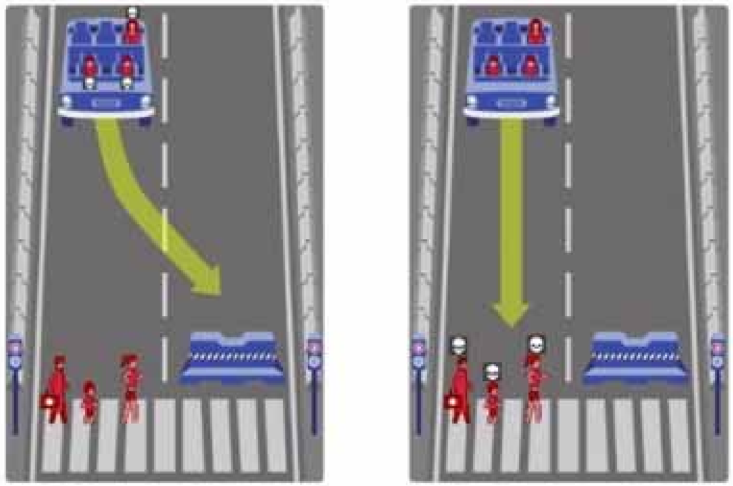
https://www.iep.utm.edu/ethics/

What is Ethics for us as Software Engineers?

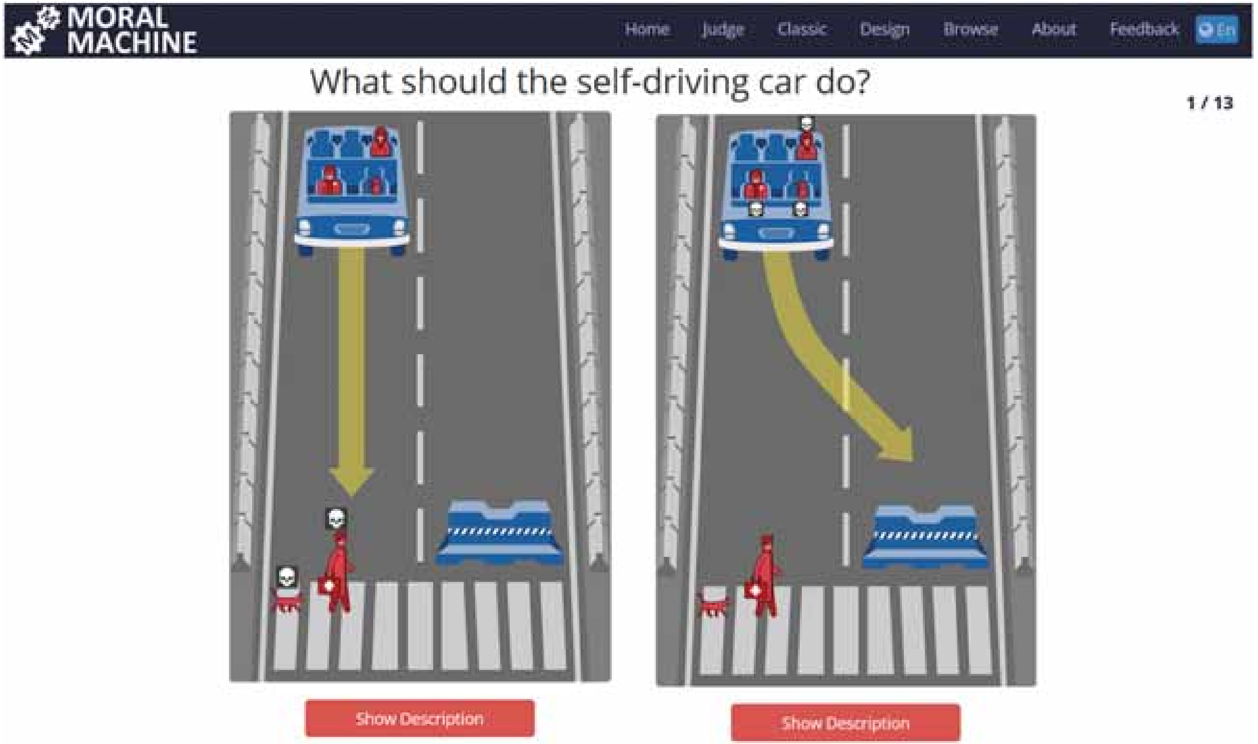
- Ethics is a **practical discipline**
- It is the good things – It is the right things
- BUT: How do we define what is good?





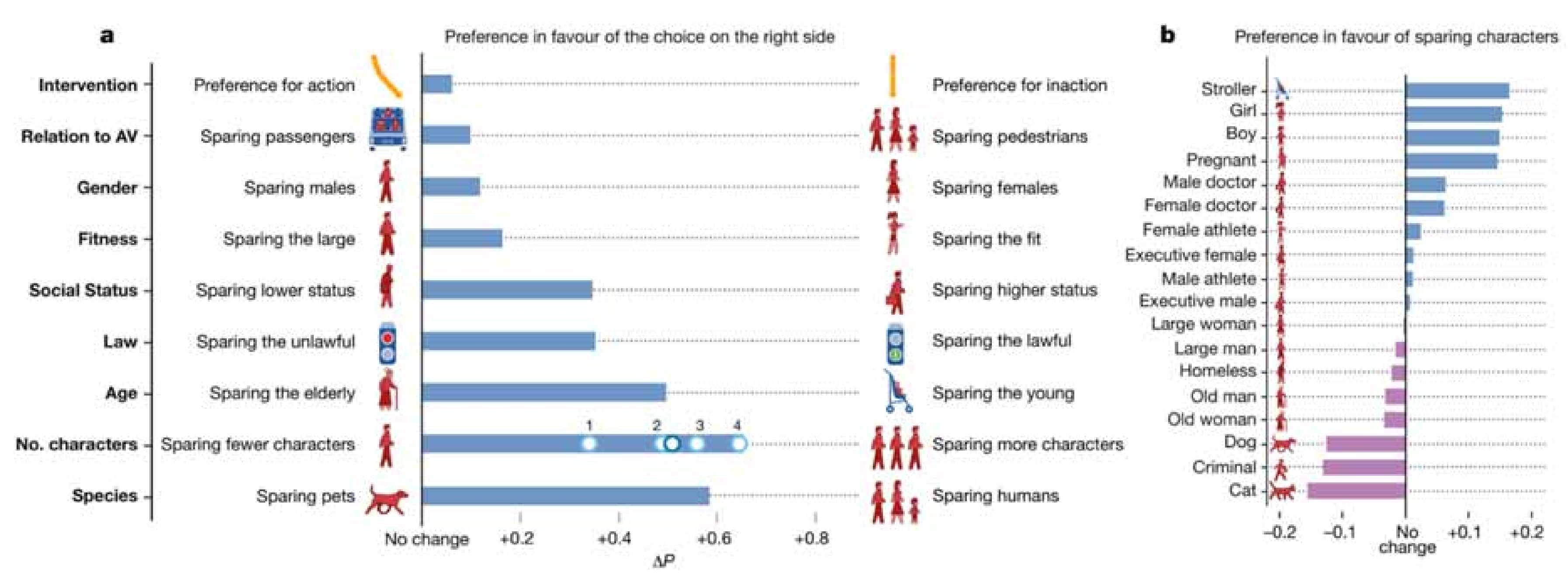


- Each student should try the MIT moral machine:
  - http://moralmachine.mit.edu/



Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon & Iyad Rahwan 2018. The moral machine experiment. Nature, 563, (7729), 59-64, doi:10.1038/s41586-018-0637-6.

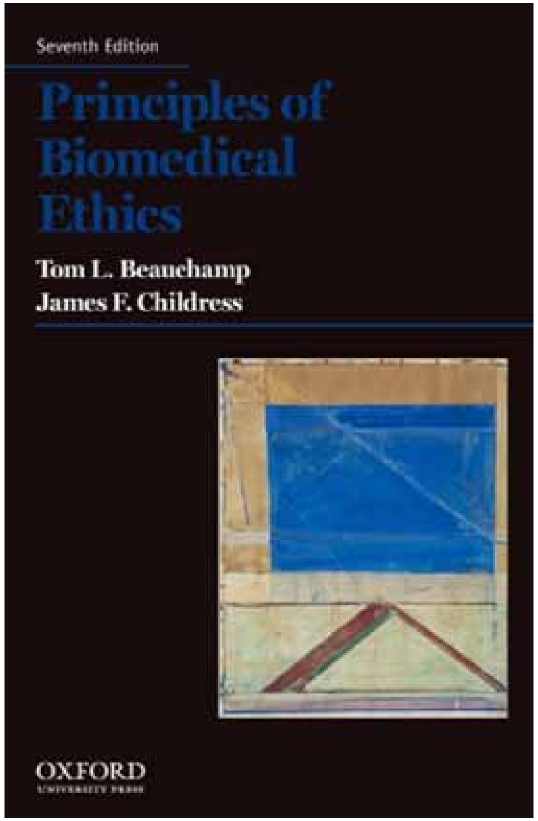
Some Results from the MIT Moral Machine study



Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon & Iyad Rahwan 2018. The moral machine experiment. Nature, 563, (7729), 59-64, doi:10.1038/s41586-018-0637-6.

UNESCO's 15 Bioethical principles

Human dignity & human rights	Benefit & harm	Autonomy-individual responsibility	Consent	Persons without the capacity to consent
Human vulnerability & personal integrity	Privacy / Confidenti-ality	Equality, Justice, Equity	Non-discrimination	Respect for cultural diversity
Solidarity & cooperation	Social responsibility & health	Sharing of benefits	Protecting future generations	Protecting biodiversity, biosphere & environment



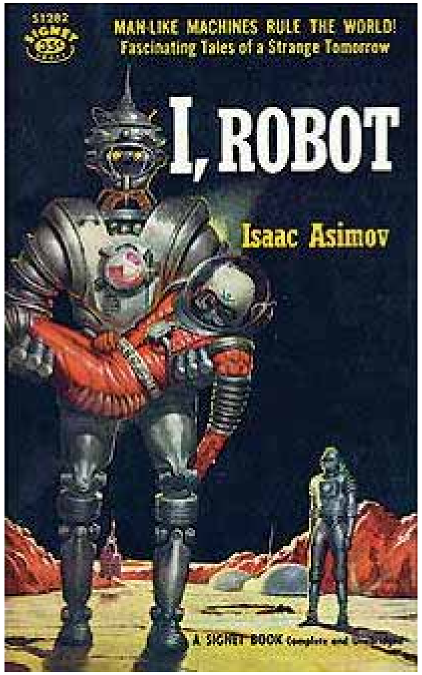
http://global.oup.com/us/companion.websites/9780199924585/student/

- 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



Asimov, Isaac (1950). I, Robot (The Isaac Asimov Collection ed.). New York: Doubleday.

Source: https://en.wikipedia.org/wiki/Three_Laws_of_Robotics

Christopher Grau 2006. There is no" I" in" robot": Robots and utilitarianism. IEEE Intelligent Systems, 21, (4), 52-55.

(Some) Grand Questions of AI Ethics

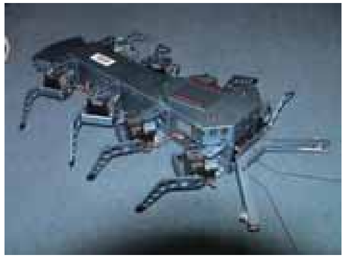
- Is it morally justified to create super-intelligent systems?
- Should our AI have any free will? And if it is possible: Can we prevent them from having free will?
- Will AI have consciousness? (Strong AI)
  - If so, will it they accept to be constrained by artificial AI-ethics placed on them by humans?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

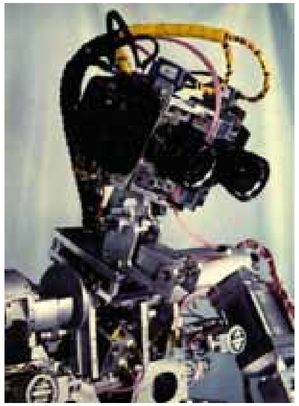
https://www.wired.com/story/will-ai-achieve-consciousness-wrong-question

For student discussion: What about existing AI?





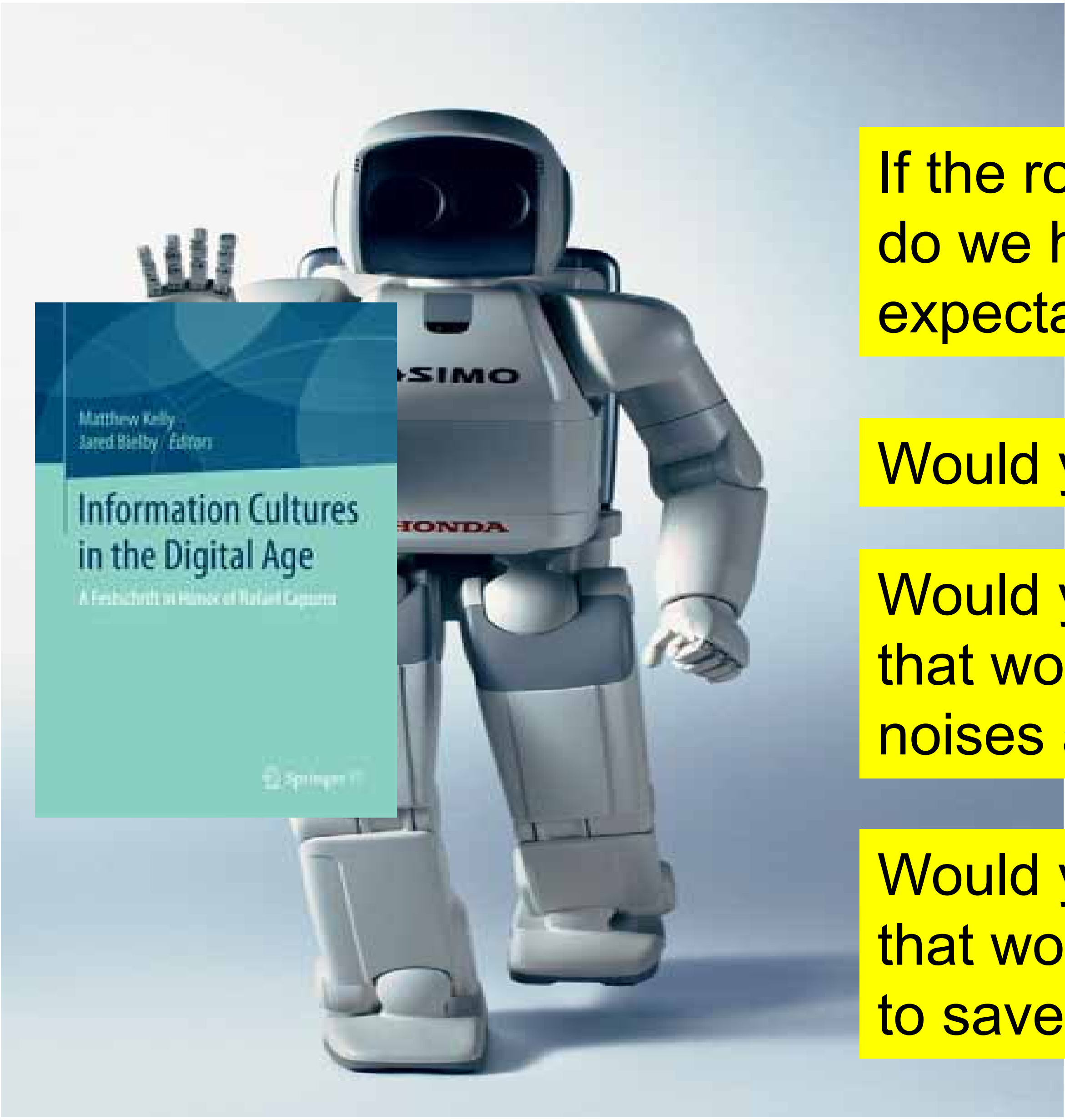





For student discussion: current AI in robotics?

HCAI
HUMAN-CENTERED.AI

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?

Last updated: 03-10-2019

04 AI ethics: Algorithms and the proof of explanations

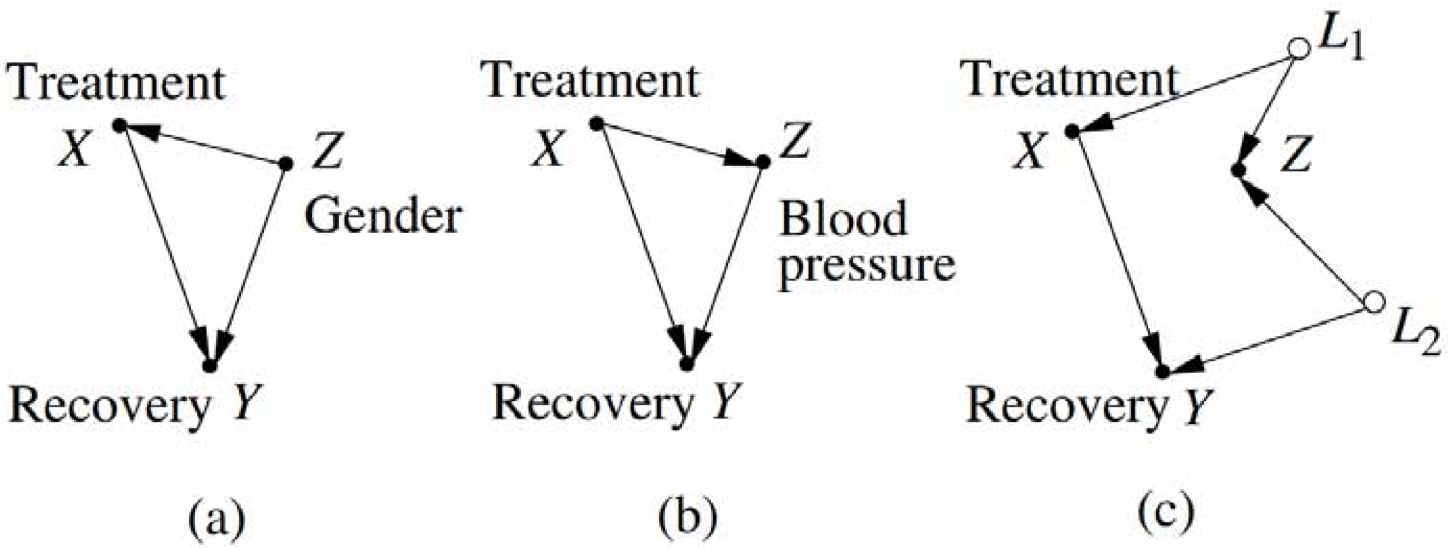
Simpson's Paradox – Statistics can not help!

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

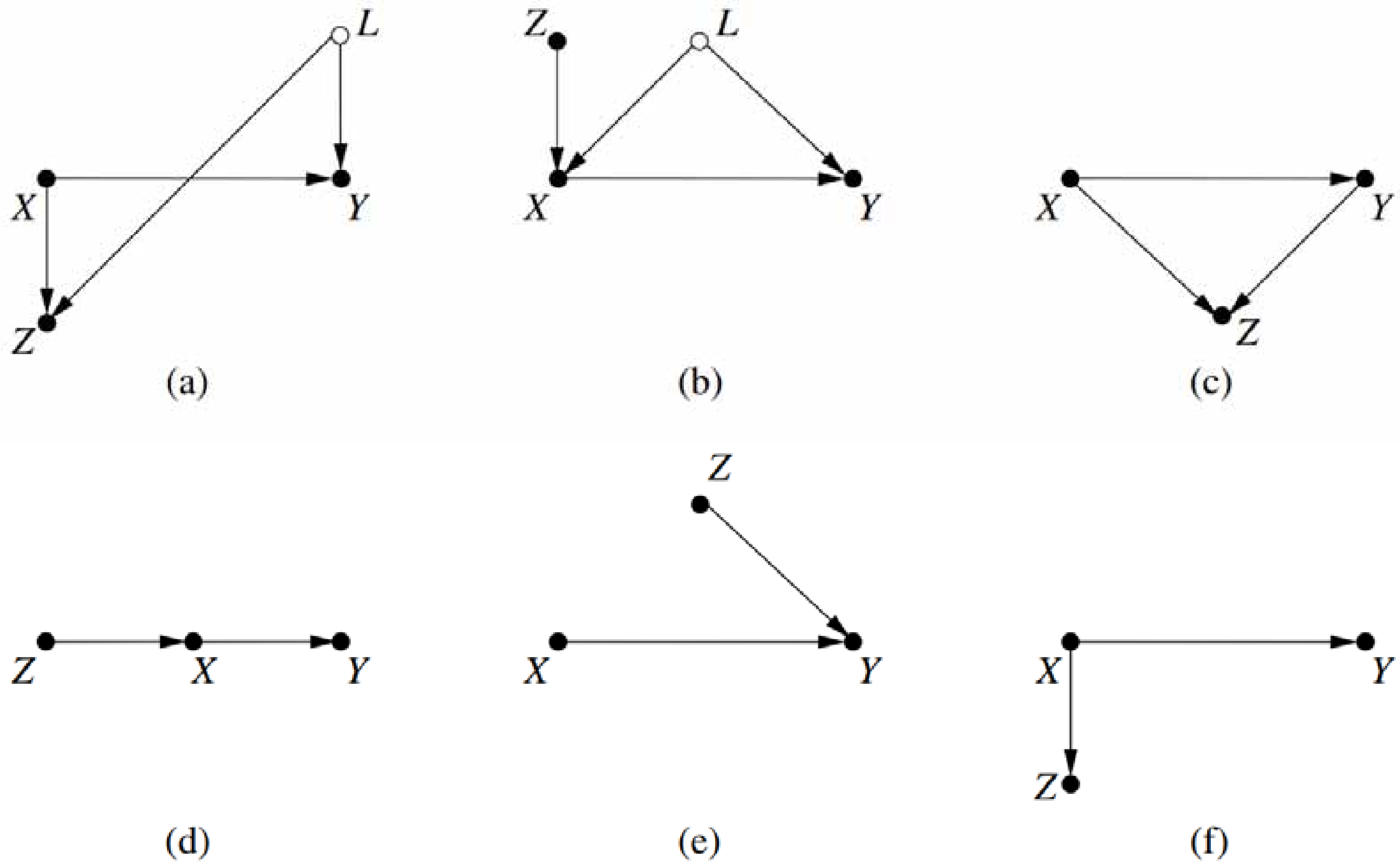
Martin Gardner 1976. Fabric of inductive logic, and some probability paradoxes. Scientific American, 234, (3), 119-124, doi:10.1038/scientificamerican0376-119.

Which scenarios invite reversals?



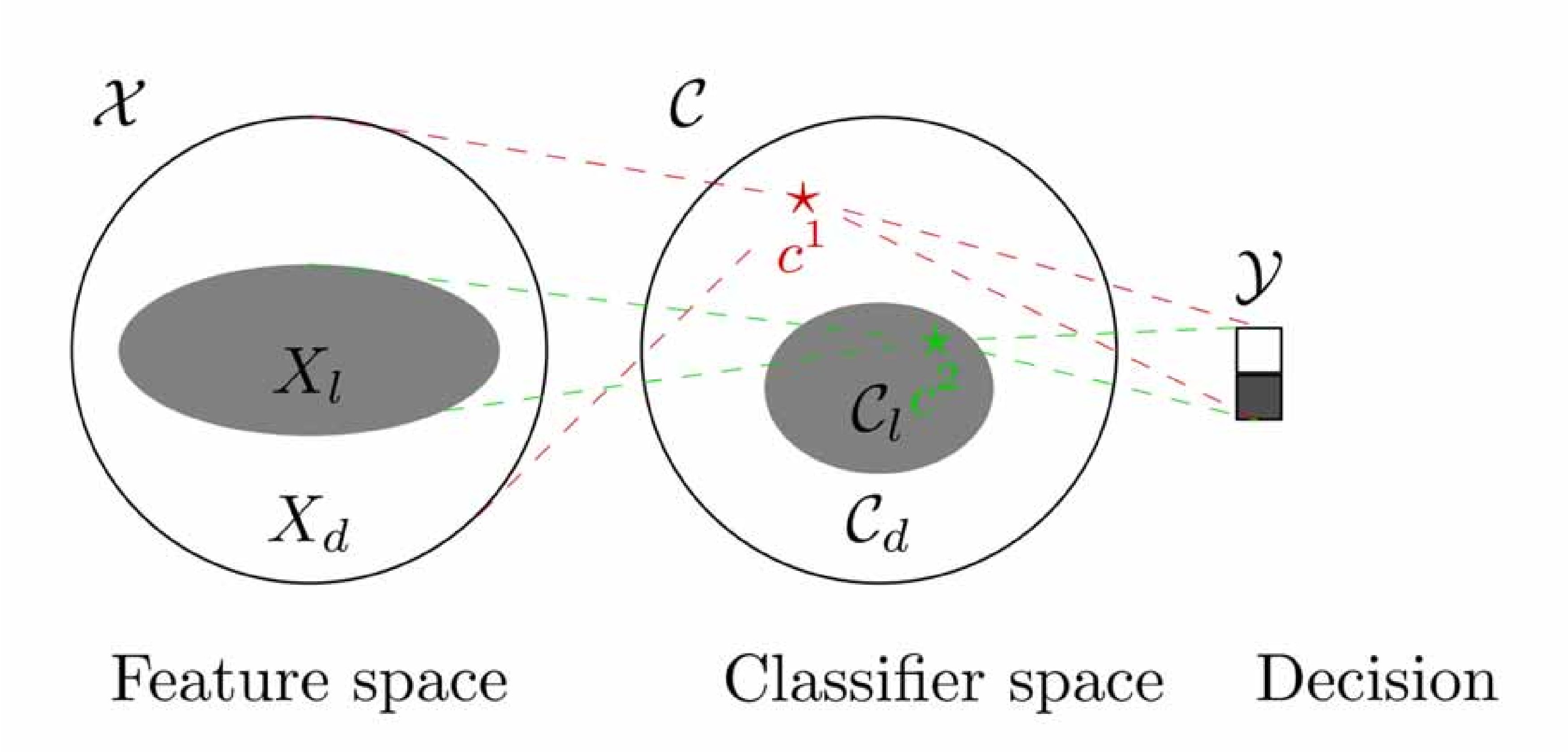
Judea Pearl 2014. Comment: understanding Simpson's paradox. The American Statistician, 68, (1), 8-13, doi:10.1080/00031305.2014.876829.

Which scenarios invite reversals?

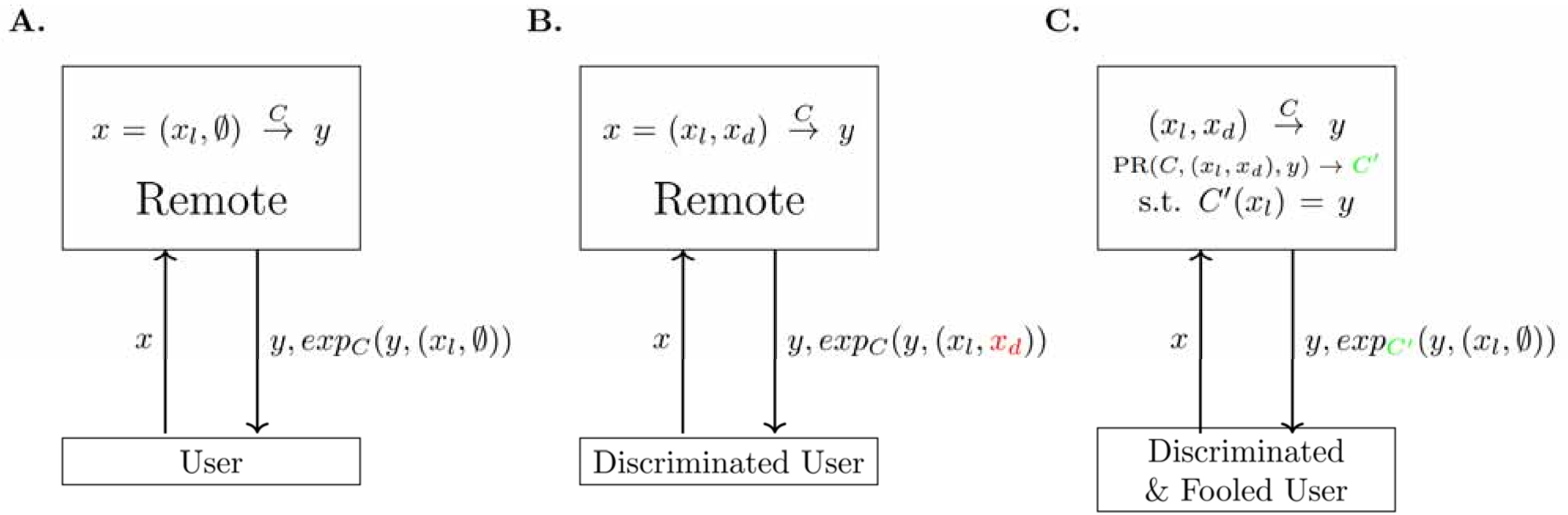


Judea Pearl 2014. Comment: understanding Simpson's paradox. The American Statistician, 68, (1), 8-13, doi:10.1080/00031305.2014.876829.

Explainability of remote decisions



Erwan Le Merrer & Gilles Tredan 2019. The Bouncer Problem: Challenges to Remote Explainability. arXiv:1910.01432.



Erwan Le Merrer & Gilles Tredan 2019. The Bouncer Problem: Challenges to Remote Explainability. arXiv:1910.01432.

- 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!!

Erwan Le Merrer & Gilles Tredan 2019. The Bouncer Problem: Challenges to Remote Explainability. arXiv:1910.01432.

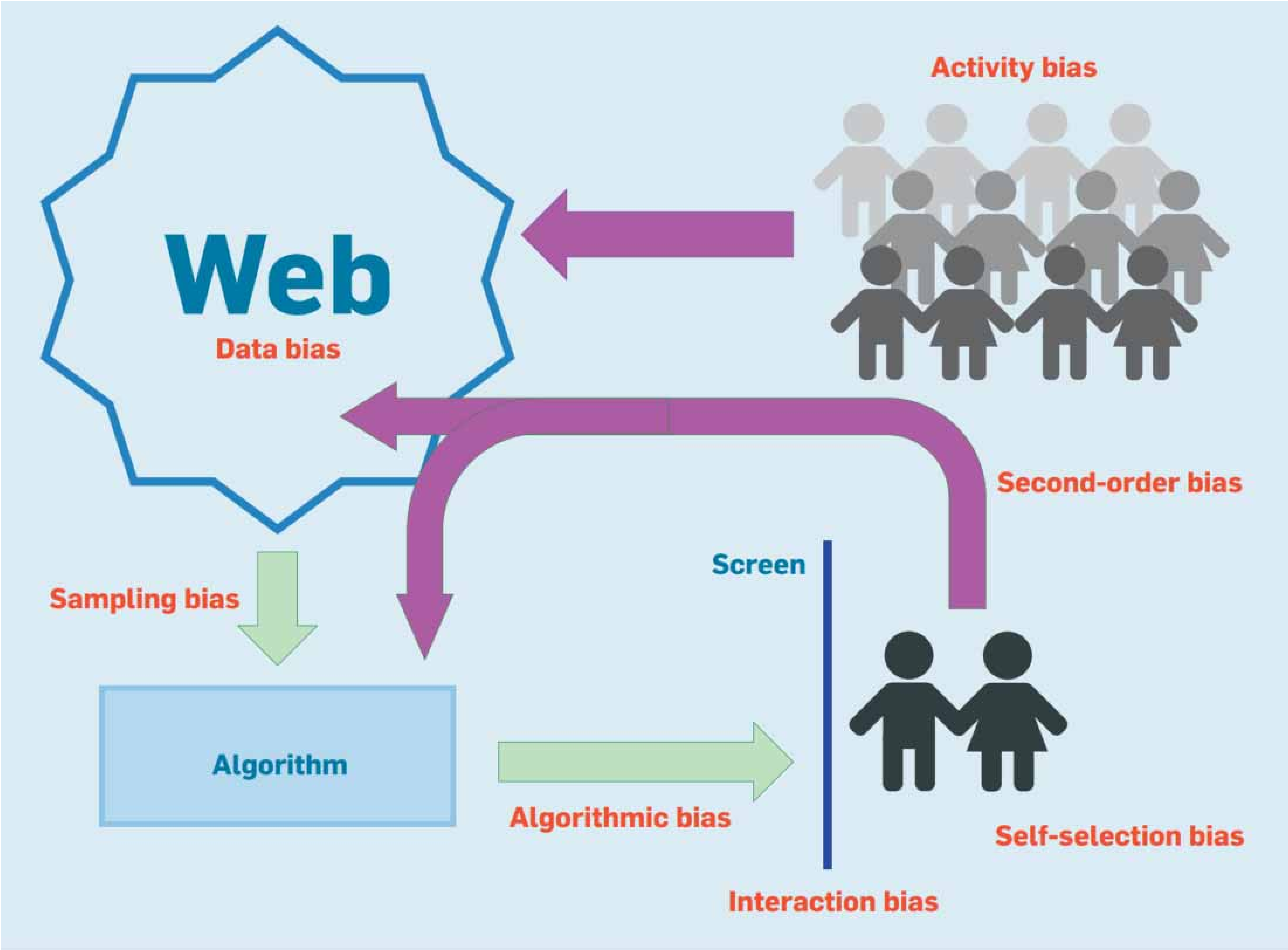
05 Responsible AI Examples from Computational Sociology (Bias, fairness, …)

- 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?

Taxonomy of “bias” on the Web



Ricardo Baeza-Yates 2018. Bias on the web. Communications of the ACM, 61, (6), 54-61, doi:10.1145/3209581.

Cognitive Bias

Please have a look at he 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 sin bias network effects!
- Bias can be bad, good, neutral (or unknown)

Amos Tversky & Daniel Kahneman 1974. Judgment under uncertainty: Heuristics and biases. Science, 185, (4157), 1124-1131, doi:10.1126/science.185.4157.1124.

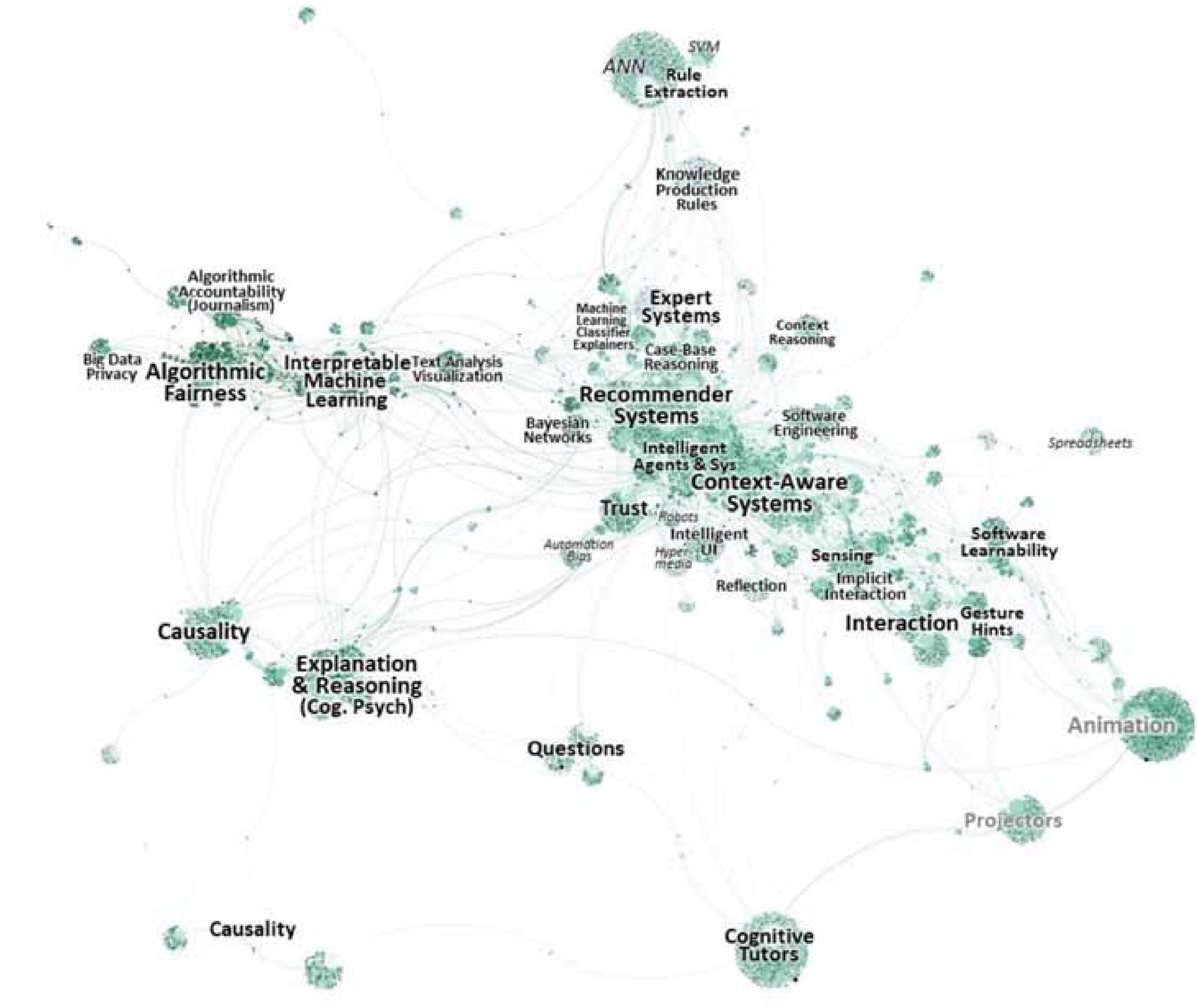
- Results from ML algorithms can be
  - unfair,
  - resulting in prejudicial treatment of people e.g. with regard to gender, race, income, sexual orientation, religion, occupation, origin, …
- Bias is resulting from many issues, e.g.
  - Data quality, distortions in demographics, behavioural aspects, linking biases, etc. etc., please have a read of this paper:
  - Alexandra Olteanu, Carlos Castillo, Fernando Diaz & Emre Kiciman 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data, 2, 1-33, doi:10.3389/fdata.2019.00013. https://www.frontiersin.org/articles/10.3389/fdata.2019.00013/full

Keith Kirkpatrick 2016. Battling algorithmic bias: How do we ensure algorithms treat us fairly? Communications of the ACM, 59, (10), 16-17, doi:10.1145/2983270.

Inclusive Images Competition of Google

James Atwood, Yoni Halpern, Pallavi Baljekar, Eric Breck, D. Sculley, Pavel Ostyakov, Sergey I. Nikolenko, Igor Ivanov, Roman Solovyev, Weimin Wang & Miha Skalic. The Inclusive Images Competition. 2020 Cham. Springer International Publishing, 155-186.

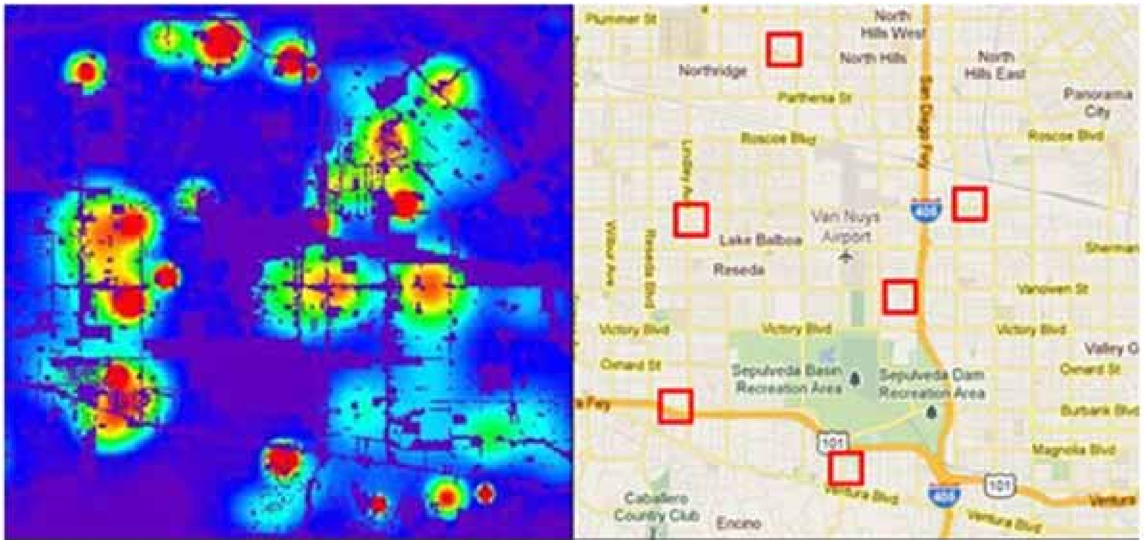
https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html



Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim & Mohan Kankanhalli. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018. ACM, 1-18, doi:10.1145/3173574.3174156.

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.



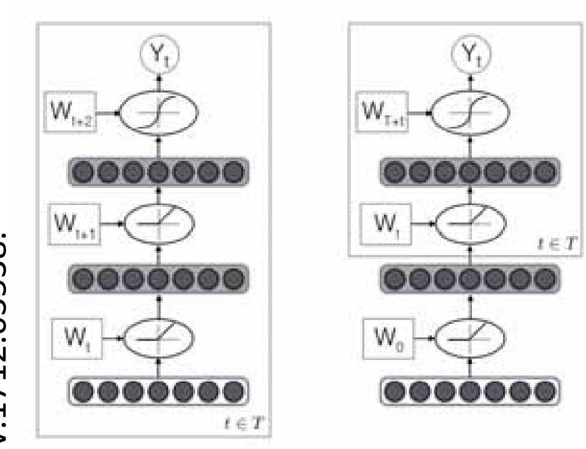
Predictive policing is built around algorithms that identify potential crime hotspots.. (PredPol)

https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/

Walt L Perry et al. 2013. Predictive policing: The role of crime forecasting in law enforcement operations, Rand Corporation.

- 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.

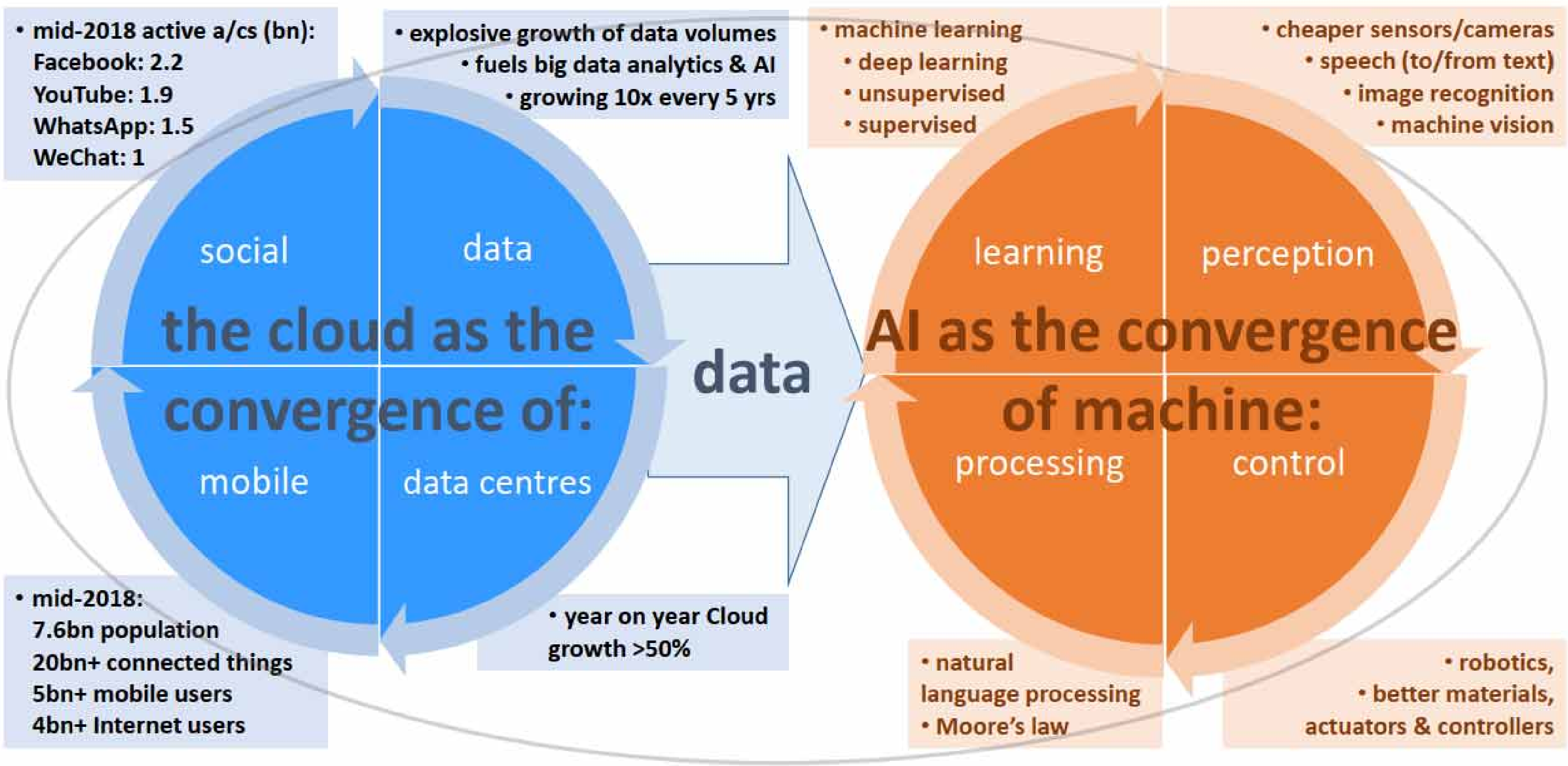
Adrian Benton, Margaret Mitchell & Dirk Hovy 2017. Multi-task learning for mental health using social media text. arXiv preprint arXiv:1712.03538.



Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé & Kate Crawford 2018. Datasheets for datasetsarXiv:1803.09010.

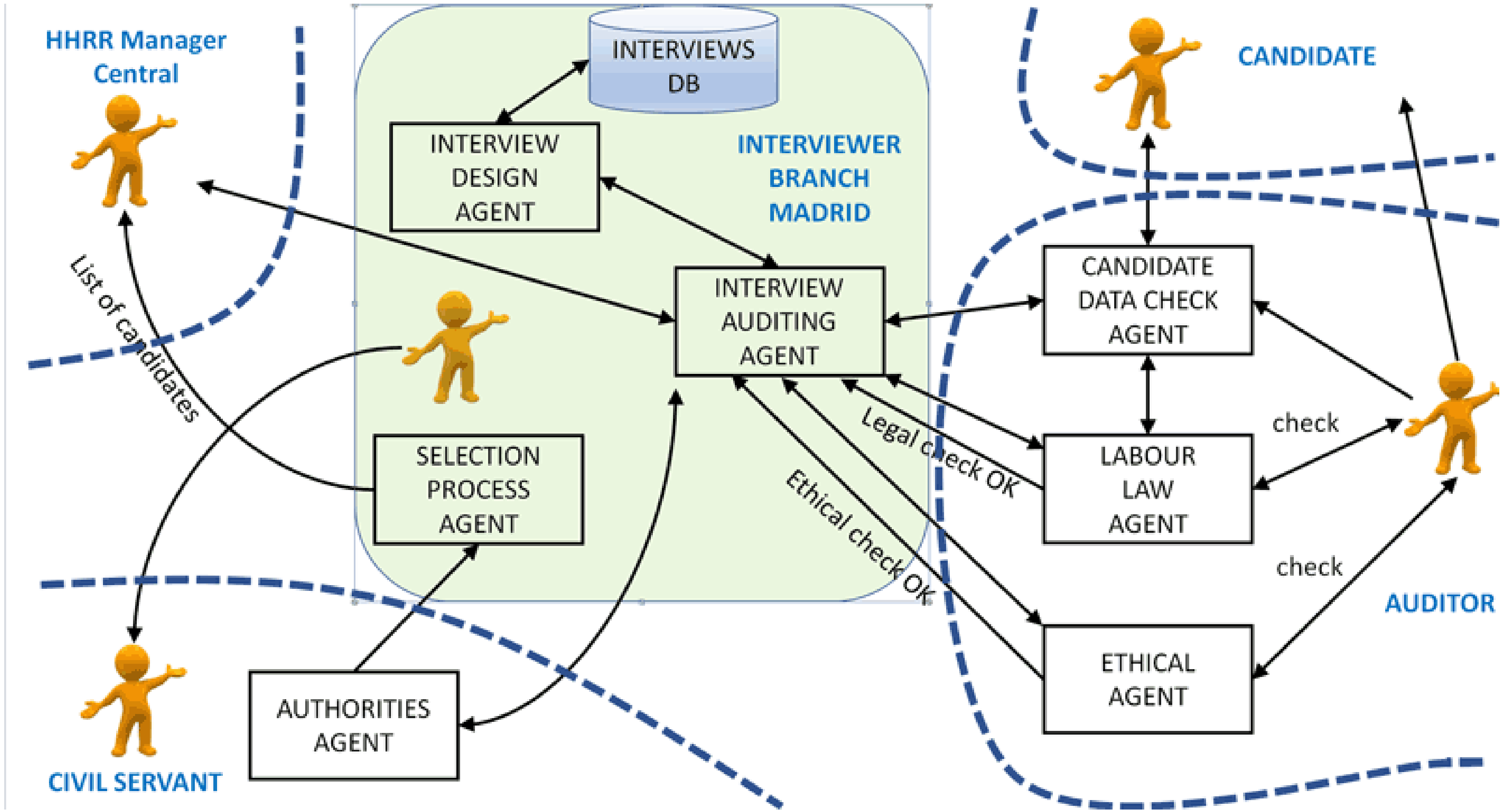
Updated versions: https://arxiv.org/abs/1803.09010

Alexa, what about legal aspects of AI ?



http://www.kempitlaw.com/wp-content/uploads/2018/09/Legal-Aspects-of-AI-Kemp-IT-Law-v2.0-Sep-2018.pdf

Example: AI recruiting system

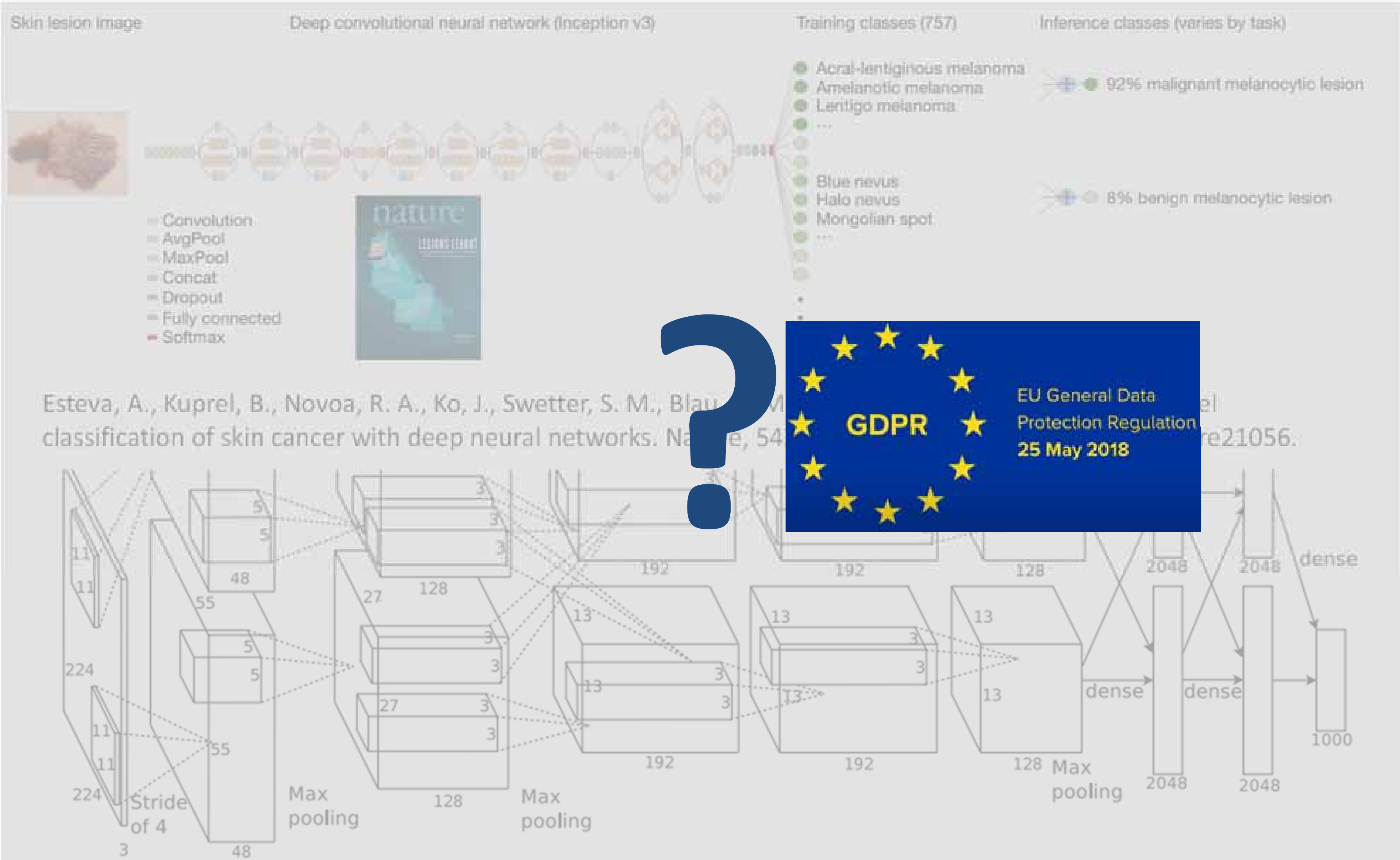


https://ercim-news.ercim.eu/en116/special/ethical-and-legal-implications-of-ai-recruiting-software

- 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.

Shane O'Sullivan, Nathalie Nevejans, Colin Allen, Andrew Blyth, Simon Leonard, Ugo Pagallo, Katharina Holzinger, Andreas Holzinger, Mohammed Imran Sajid & Hutan Ashrafian 2019. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. The International Journal of Medical Robotics and Computer Assisted Surgery, 15, (1), 1-12, doi:10.1002/rcs.1968.

The right for explanation



I. Overview ▶ Right to explanation

EU General Data Protection Regulation

Article	Contents
17. Right to be forgotten	**An individual to have certain data deleted** so that third persons can no longer trace them
22. Automated individual decision making	The data **subject** shall **have the right not to be subject** to a decision based solely on **automated processing (including profiling)**.
13-14. **Right to explanation**	A data subject has the right to "**meaningful information about the logic involved**."
EU administration	When violated **4% of global revenue** will be fined.
Enact	**May 28th, 2018**

3

Conclusion

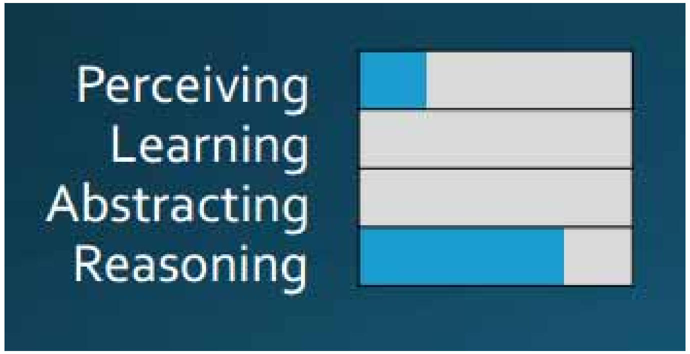


Image credit to John Launchbury

- 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 $p(x)$

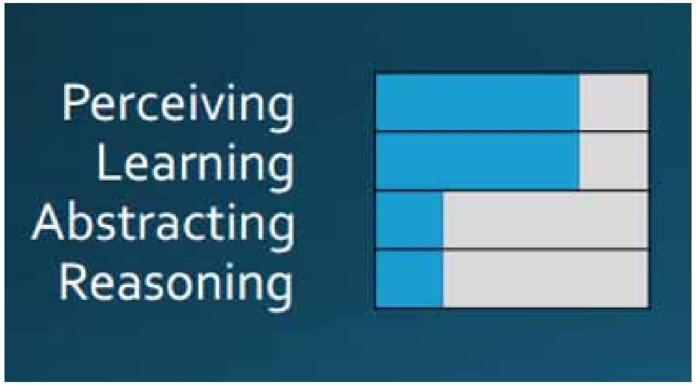


Image credit to John Launchbury

- 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

Image credit to John Launchbury

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

- 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

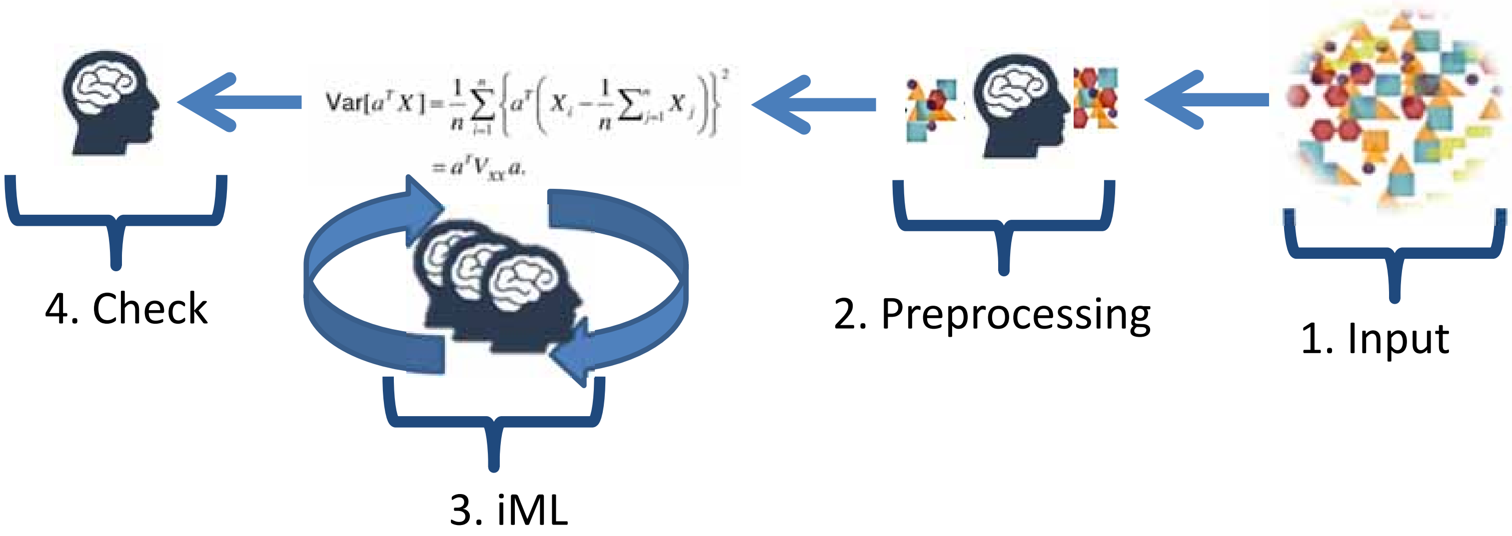
Fact: Cyberthreats are the main concern: it needs no body – only an Internet connection

- 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 …



Holzinger, A. 2016. Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? Brain Informatics (BRIN), 3, (2), 119-131, doi:10.1007/s40708-016-0042-6.

Thank you!