Advanced Deep Learning: Explainability

Anders Søgaard

Course outline

Goal 1: Quick tour of recent developments in deep learning

Goal 2: Inspiration for thesis/research projects

Week	Lecturer	Subject	Literature	Assignment
1	Stefan	Introduction to Neural Networks.	d2l 2.1-2.5, 2.7, 11.5.1, **slides**	
2	Stefan	CNNs; FCNs; U-Nets. Data augmentation; invariance; regularization e.g. dropout	d2l 6, 7, 13.9-13.11, **slides**	Assignment 1 (May 10)
3	Anders/Phillip	May 9 (A): RNNs May 11 (P): Transformers	d2l 8 Transformers: d2l 10.5-10.7 + Vaswani et al. (2017)	Assignment 2 (May 20)
4	Phillip/Anders	May 16 (P): Representation and Adversarial Learning May 18 (A): A Learning Framework + Self-supervised Learning + Contrastive Learning	Autoencoders: blog post GANs: Goodfellow (2016) Self-supervised learning: blog post Contrastive learning: Dor et al. (2018) Adversarial examples: Goodfellow et al. (2015)	
5	Anders	May 23: General Properties, e.g., Scaling Laws, Lottery Tickets, Bottleneck Phenomena May 25: Applications of Representation, Adversarial and Contrastive Learning	GANs: Lample et al. (2018) Autoencoders: Chandar et al. (2011) Contrastive learning: Yu et al. (2018) DynaBench: Talk by Douwe Kiela (Facebook, now HuggingFace) Scaling laws: Kaplan et al. (2020)	Assignment 3 *[MC on Representation Learning/1p Report on Lottery Ticket extraction]* (June 3)
6	Anders	May 30: Interpretability, Transparency, and Trustworthiness & Deep Learning for Scientific Discovery June 1: Interpretability (Feature Attribution), including Guest Lecture by Stephanie Brandl	DL for Scientific Discovery: Sullivan (2022) Interpretability/Background: Søgaard (2022)	
7	Anders	June 6: *Off (no teaching)* June 8: Interpretability (Training Data Influence)	Literature: Feng and Boyd-Graber (2018); Jiang and Senge (2021)	
8	Anders	June 13-15: Best Practices	Literature: Dodge et al. (2019) and Raji et al. (2021)	Assignment 4 *[MC on Interpretability; 1p Report on Best Practices]* (June 21)

Course outline

Goal 1: Quick tour of recent developments in deep learning

Goal 2: Inspiration for thesis/research projects

Architectures

Framework

Fairness / Explainable AI

Methodology

Week	Lecturer	Subject	Literature	Assignment
1	Stefan	Introduction to Neural Networks.	d2l 2.1-2.5, 2.7, 11.5.1, **slides**	
2	Stefan	CNNs; FCNs; U-Nets. Data augmentation; invariance; regularization e.g. dropout	d2l 6, 7, 13.9-13.11, **slides**	Assignment 1 (May 10)
3	Anders/Phillip	May 9 (A): RNNs May 11 (P): Transformers	d2l 8 Transformers: d2l 10.5-10.7 + Vaswani et al. (2017)	Assignment 2 (May 20)
		May 16 (P): Representation and Adversarial Learning	Autoencoders: blog post GANs: Goodfellow (2016) Self-supervised learning: blog post Contrastive learning: Dor et al. (2018) Adversarial examples: Goodfellow et al. (2015)	
4	Phillip/Anders	May 18 (A): A Learning Framework + Self-supervised Learning + Contrastive Learning		
5	Anders	May 23: General Properties, e.g., Scaling Laws, Lottery Tickets, Bottleneck Phenomena May 25: Applications of Representation, Adversarial and Contrastive Learning	GANs: Lample et al. (2018) Autoencoders: Chandar et al. (2011) Contrastive learning: Yu et al. (2018) DynaBench: Talk by Douwe Kiela (Facebook, now HuggingFace) Scaling laws: Kaplan et al. (2020)	Assignment 3 *[MC on Representation Learning/1p Report on Lottery Ticket extraction]* (June 3)
6	Anders	May 30: Interpretability, Transparency, and Trustworthiness & Deep Learning for Scientific Discovery June 1: Interpretability (Feature Attribution), including Guest Lecture by Stephanie Brandl	DL for Scientific Discovery: Sullivan (2022) Interpretability/Background: Søgaard (2022)	
7	Anders	June 6: *Off (no teaching)* June 8: Interpretability (Training Data Influence)	Literature: Feng and Boyd-Graber (2018); Jiang and Senge (2021)	
8	Anders	June 13-15: Best Practices	Literature: Dodge et al. (2019) and Raji et al. (2021)	Assignment 4 *[MC on Interpretability; 1p Report on Best Practices]* (June 21)

Today

a) An incompatibility proof
b) Hedden (2021)'s criticism of fairness metrics
c) Anna's criticism of Rawlsian fairness
d) Sullivan (2022)'s criticism of XAI methods (that they are irrelevant to scientific discovery)
e) Jones (2012)'s criticism of trustworthiness
f) What we did not get around to

An incompatibility proof

Fairness metrics

1. **Statistical parity:** $P(d = 1|G = m) = P(d = 1|G = f)$
2. **Performance parity:** $P(Y = 1|d = 1, G = m) = P(Y = 1|d = 1, G = f)$
3. **Accuracy equality:** $P(d = Y, G = m) = P(d = Y, G = f)$

See this video for 21 definitions.

Fairness metrics

1. **Statistical parity:** $P(d = 1|G = m) = P(d = 1|G = f)$
2. **Performance parity:** $P(Y = 1|d = 1, G = m) = P(Y = 1|d = 1, G = f)$
3. **Accuracy equality:** $P(d = Y, G = m) = P(d = Y, G = f)$

Exercise: **What metrics are incompatible?**

Fairness metrics

1. **Statistical parity:** $P(d = 1|G = m) = P(d = 1|G = f)$
2. **Performance parity:** $P(Y = 1|d = 1, G = m) = P(Y = 1|d = 1, G = f)$
3. **Accuracy equality:** $P(d = Y, G = m) = P(d = Y, G = f)$

Exercise:

Say $P(d=1|G=m)=0.5$, $P(d=1|G=f)=0.5$ (Statistical parity). Then:

d=1, Y=0	d=0, Y=1	d=1, Y=0	d=0, Y=1	Acc_m: 0.0
d=1, Y=0	d=0, Y=0	d=1, Y=0	d=0, Y=1	Acc_f: 0.25
m	m	f	f	

Hedden (2021)

Hedden (2021)

Suppose that there are a bunch of coins of varying biases. Each individual in the population is randomly assigned a coin. Then those individuals are randomly assigned to one of two rooms, A and B. Our aim is to predict, for each person, whether that person's coin will land heads or tails. That is, our aim is to predict, for each person, whether they are a heads person or a tails person. Luckily, each coin comes labeled with its bias, with a real number in the interval [0, 1] indicating its bias, or its objective chance of landing heads. 10 Here is a perfectly fair and unbiased predictive algorithm: For each person, take their coin and read its label. If it says 'x,' assign that person a risk score of x. And if x > 0.5, make the binary prediction that they are a heads person (positive), while if x < 0.5, make the binary prediction that they are a tails person (negative). (What if x = 0.5? We might arbitrarily predict heads in that case, or randomize our prediction. I will sidestep this issue by assuming that none of the coins are labeled '0.5.')

Hedden (2021)

Suppose that room A has 12 people with coins labeled '0.75' and 8 people with coins labeled '0.125.' The former are all assigned risk score 0.75 and predicted to be heads people, and 9 of them in fact are heads people (since we're assuming that relative frequencies match biases). [...] Room B contains 10 people with coins labeled '0.6' and 10 people with coins labeled '0.4.' The former are all assigned risk score 0.6 and predicted to be heads people, and 6 of them are in fact heads people. The latter are all assigned risk score 0.4 and predicted to be tails people, and 4 of them are in fact heads people.

0.75	0.75	0.75	0.75	0.125
0.75	0.75	0.75	0.75	0.125
0.75	0.75	0.125	0.125	0.125
0.75	0.75	0.125	0.125	0.125
0.6	0.6	0.6	0.4	0.4
0.6	0.6	0.6	0.4	0.4
0.6	0.6	0.4	0.4	0.4
0.6	0.6	0.4	0.4	0.4

Fairness metrics

0.75	0.75	0.75	0.75	0.125
0.75	0.75	0.75	0.75	0.125
0.75	0.75	0.125	0.125	0.125
0.75	0.75	0.125	0.125	0.125
0.6	0.6	0.6	0.4	0.4
0.6	0.6	0.6	0.4	0.4
0.6	0.6	0.4	0.4	0.4
0.6	0.6	0.4	0.4	0.4

1. **Statistical parity:** $P(d = 1|G = m) = P(d = 1|G = f)$
2. **Performance parity:** $P(Y = 1|d = 1, G = m) = P(Y = 1|d = 1, G = f)$
3. **Accuracy equality:** $P(d = Y, G = m) = P(d = Y, G = f)$

No statistical parity: 0.6 vs 0.5

No performance parity: 0.66 vs 0.6

No accuracy equality: 0.75 vs 0.6

Exercise

What's the problem with Hedden's example?

0.75	0.75	0.75	0.75	0.125
0.75	0.75	0.75	0.75	0.125
0.75	0.75	0.125	0.125	0.125
0.75	0.75	0.125	0.125	0.125
0.6	0.6	0.6	0.4	0.4
0.6	0.6	0.6	0.4	0.4
0.6	0.6	0.4	0.4	0.4
0.6	0.6	0.4	0.4	0.4

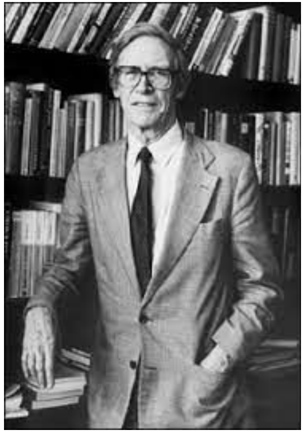
Counter-arguments

1. **Counter-argument 1** In Hedden's example, group assignment is random. Protected attributes are protected attributes in part because they are not assigned at random. Now what happens if we assign all people with biases lower than 0.5 to one room, and the rest to the other? Is the optimal classifier still fair?
2. **Counter-argument 2** The problem in Hedden's example is not a machine learning problem. Hedden only evaluates the classifier on the 20 data points in his sample. Machine learning problems are about minimizing expected risk on a data distribution. Fairness is, in other words, not to be measured on a finite sample of training data points, but on a distribution. Remember the distributions Hedden uses to draw his sample, are random: *Each individual in the population is randomly assigned a coin. Then those individuals are randomly assigned to one of two rooms, A and B.*

Rawls (1986)

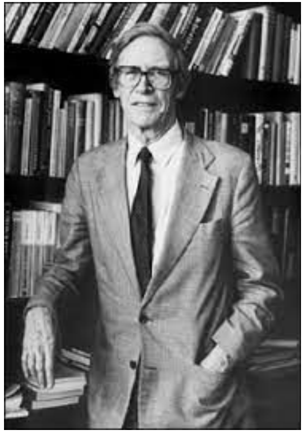
Rawlsian fairness

Exercise: What’s wrong with maximizing the welfare of the worst-off group?



Rawlsian fairness

Exercise: What's wrong with maximizing the welfare of the worst-off group?



- English is easy to learn for most. *Opportunity*
- English is the most widely used language. *Desert*
- It is up to industry/research labs to decide. *Procedure*
- English users have more advanced needs. *Need*
- Other technologies are for English markets first. *Reference*

Sullivan (2022)

Link uncertainty

Sullivan (2022) argues that it's **link uncertainty**, not **how-it-works**, that causes black box effects. Link uncertainty can mean underspecification or spurious correlations.

how-possibly	**how-actually**	**how-it-works**
Showing X->Y is learnable	Finding the causal factors	Interpretability methods

Link uncertainty

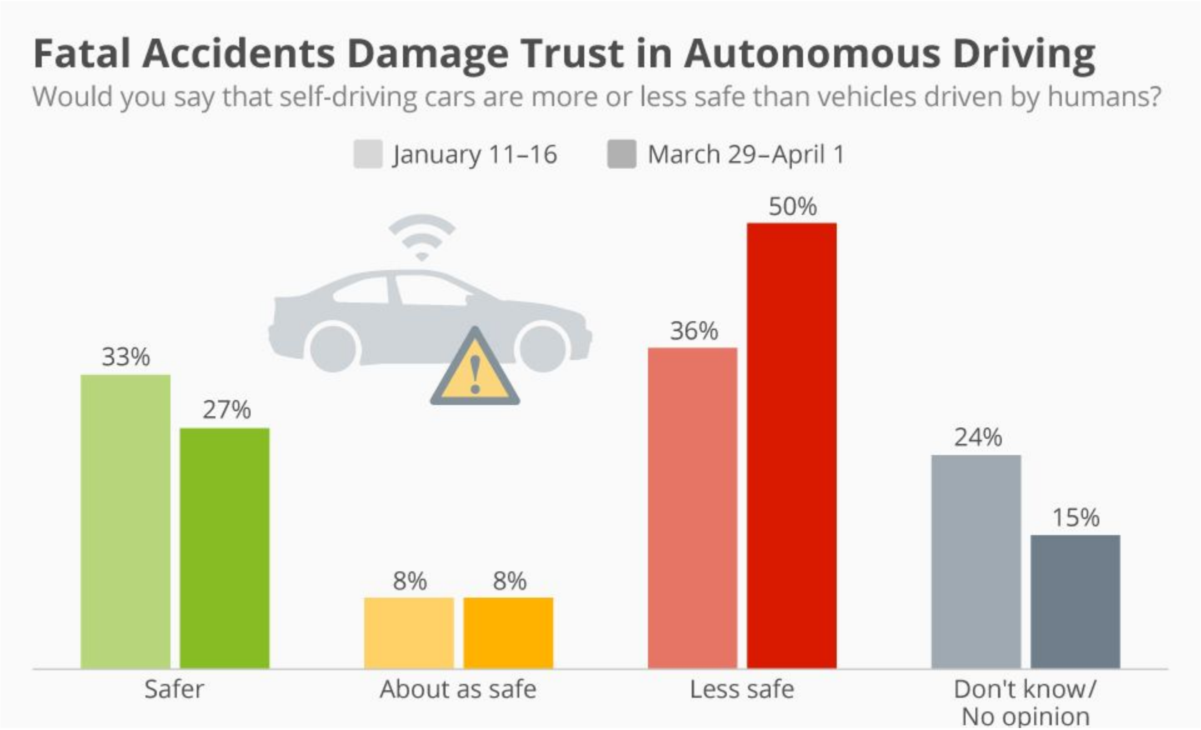
Sullivan (2022) argues that it's **link uncertainty**, not **how-it-works**, that causes black box effects. Link uncertainty can mean underspecification or spurious correlations.

how-possibly	**how-actually**	**how-it-works**
Showing X->Y is learnable	Finding the causal factors	Interpretability methods
Narrowing down the input-output mappings	Where the narrowing ends…	Tools for narrowing down

Jones (2012)

Trust in AI

As AI is rolled out to the general public and becomes a fundamental technology in our daily lives, we are increasingly asked to trust the predictions of AI models. Trust is hard to establish, though, and is easily lost. **Can AI be trusted?**



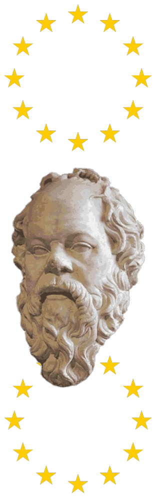
Trustworthy AI?

NSF initiated Trusted Computing (in 2001), then Cyber Trust (2004), then Trustworthy Computing (2007), and now Secure and Trustworthy Cyberspace (2011). In 2019, The European Commission's expert group on AI presented ethics guidelines for ‘trustworthy AI’.

We need people to trust AI, but of course this requires AI to be trustworthy.

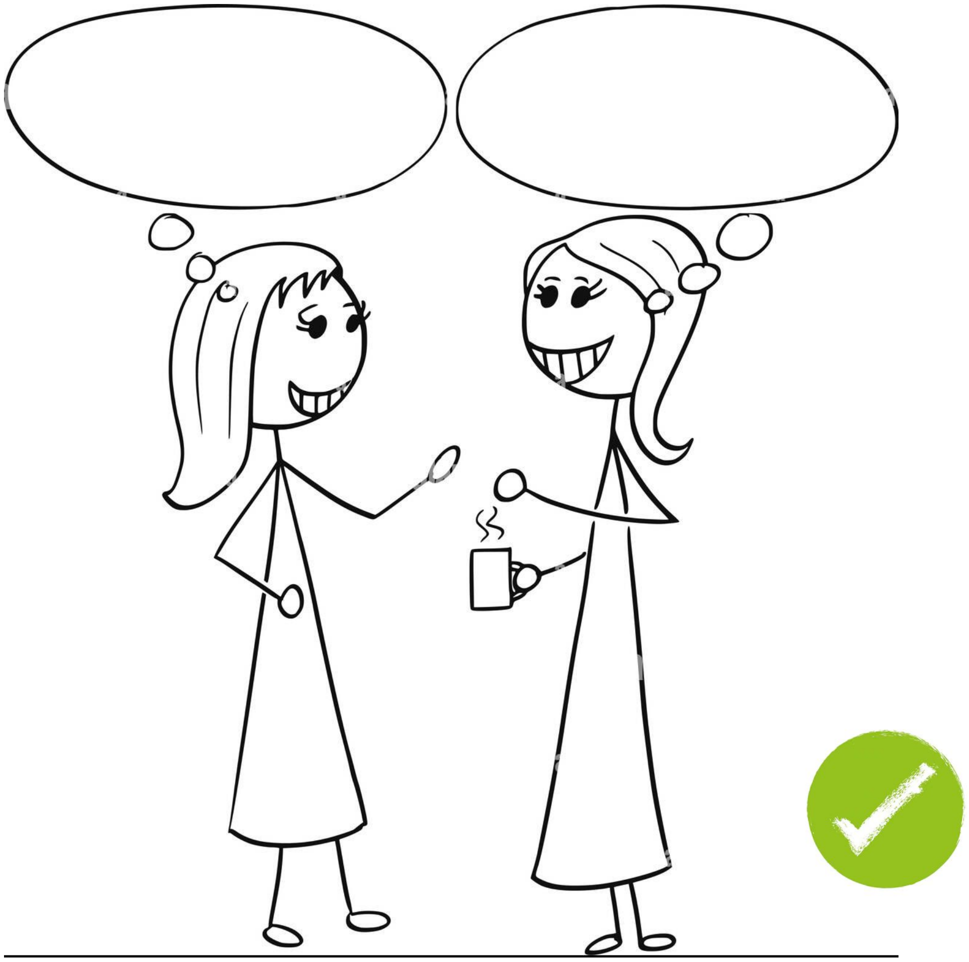
But AI cannot be trusted or distrusted! Only humans can.

Wait, what?



Karen Jones

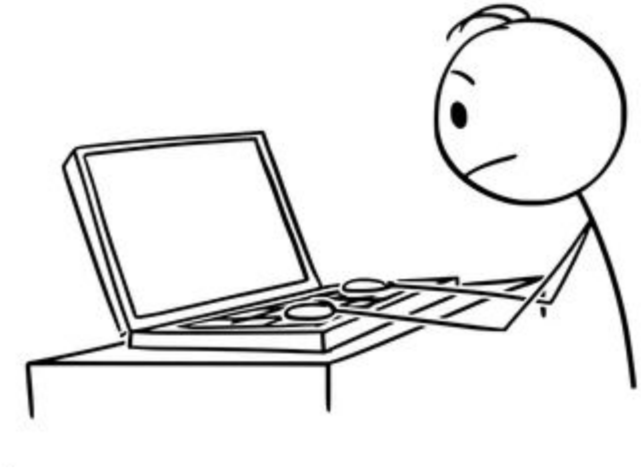
Karen Jones (2012) defines trustworthiness as the combination of competence and a 'direct responsiveness to the fact that the other is counting on you'.



Jones, K. (2012). Trustworthiness. *Ethics*, *123*(1), 61–85.

Karen Jones

Karen Jones (2012) defines trustworthiness as the combination of competence and a 'direct responsiveness to the fact that the other is counting on you'.



Jones, K. (2012). Trustworthiness. *Ethics*, *123*(1), 61–85.

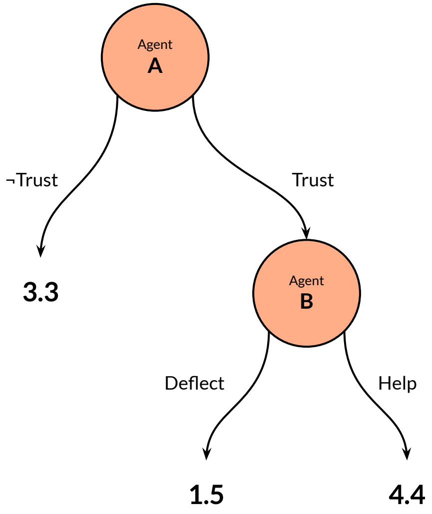
The Trustworthy - Reliable Continuum?

A One-Step Two-Player Game

Agent B is trustworthy for Agent A iff

1. Agent B is competent
2. and - knowing Agent A trusts her - helpful enough to secure a better outcome (**4.4**).

'direct responsiveness to the fact that the other is counting on you'.

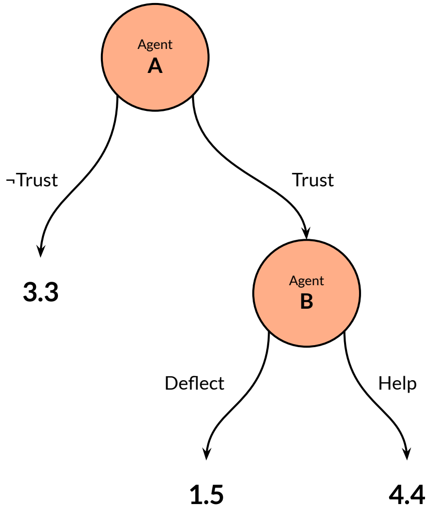


A One-Step Two-Player Game

Agent B is trustworthy for Agent A iff

1. Agent B is competent
2. and - knowing Agent A trusts her - helpful enough to secure a better outcome (**4.4**).

Note: 2) is a necessary precondition, because human agents need not be always helpful.

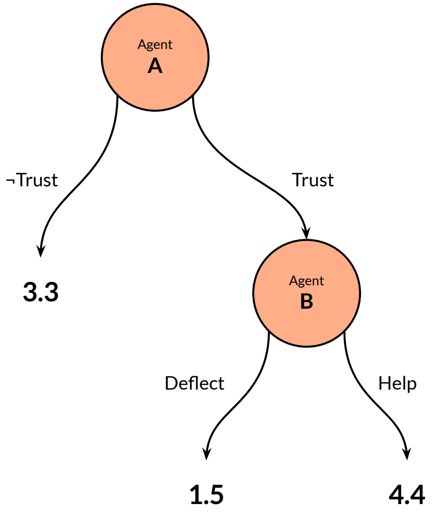


A One-Step Two-Player Game

Agent B is trustworthy for Agent A iff

1. Agent B is competent
2. and - knowing Agent A trusts her - helpful enough to secure a better outcome (**4.4**).

Institutions: Generally, help from an institution B does **not** depend on B's knowing Agent A trusts B. This is because B has to abide regulations. But institutions are trustworthy?

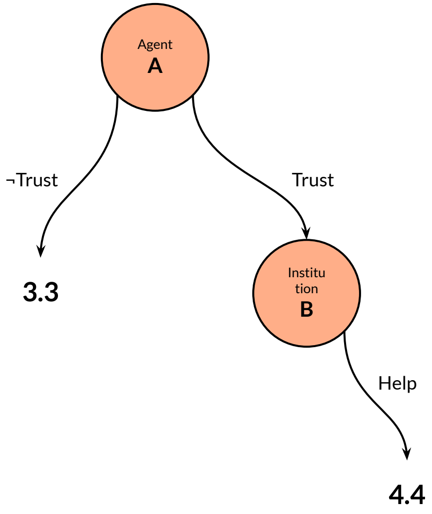


A One-Step Two-Player Game

Institution B is trustworthy for Agent A iff

1. **Institution** B is competent
2. and consistent (rule-abiding) enough to secure a better outcome (**4.4**).

Institutions: Trust in an institution only depends on their competence level and their consistency.

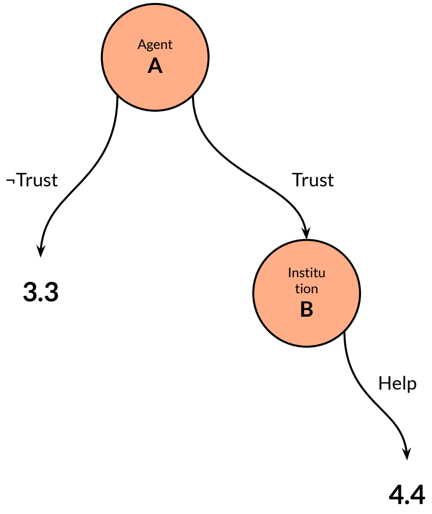


A One-Step Two-Player Game

Artefact B is trustworthy for Agent A iff

1. **Artefact** B is competent
2. and consistent (rule-abiding) enough to secure a better outcome (**4.4**).

Artefacts: Cars and drilling machines can now be trustworthy. *Competence and consistency here amounts to quality and predictability.*

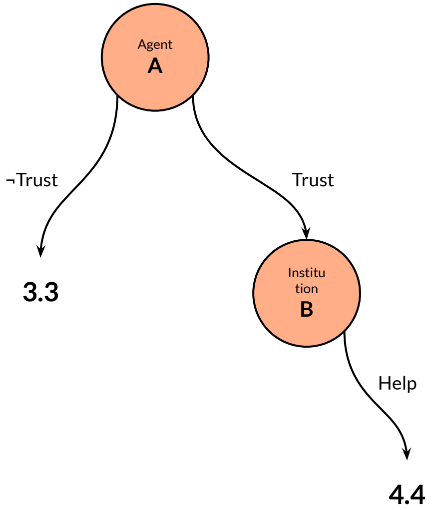


A One-Step Two-Player Game

Artificial Intelligence B is trustworthy for Agent A iff

1. **Artificial Intelligence** B is competent
2. and consistent (rule-abiding) enough to secure a better outcome (**4.4**).

Artificial Intelligence: Software, incl. artificial intelligence, can therefore also be trustworthy. *Competence and consistency here amounts to quality and predictability.*



AI Competence and Consistency

1. Competence and consistency in AI amount to quality and predictability.
2. That is, **low risk (accuracy)** and **low sensitivity to drift (robustness)**.
3. Accuracy is a common objective in AI.
4. Robustness involves fairness and may benefit (in development) from explainability, but does not necessarily involve privacy and transparency.

What we did not get around to

—

Influence functions

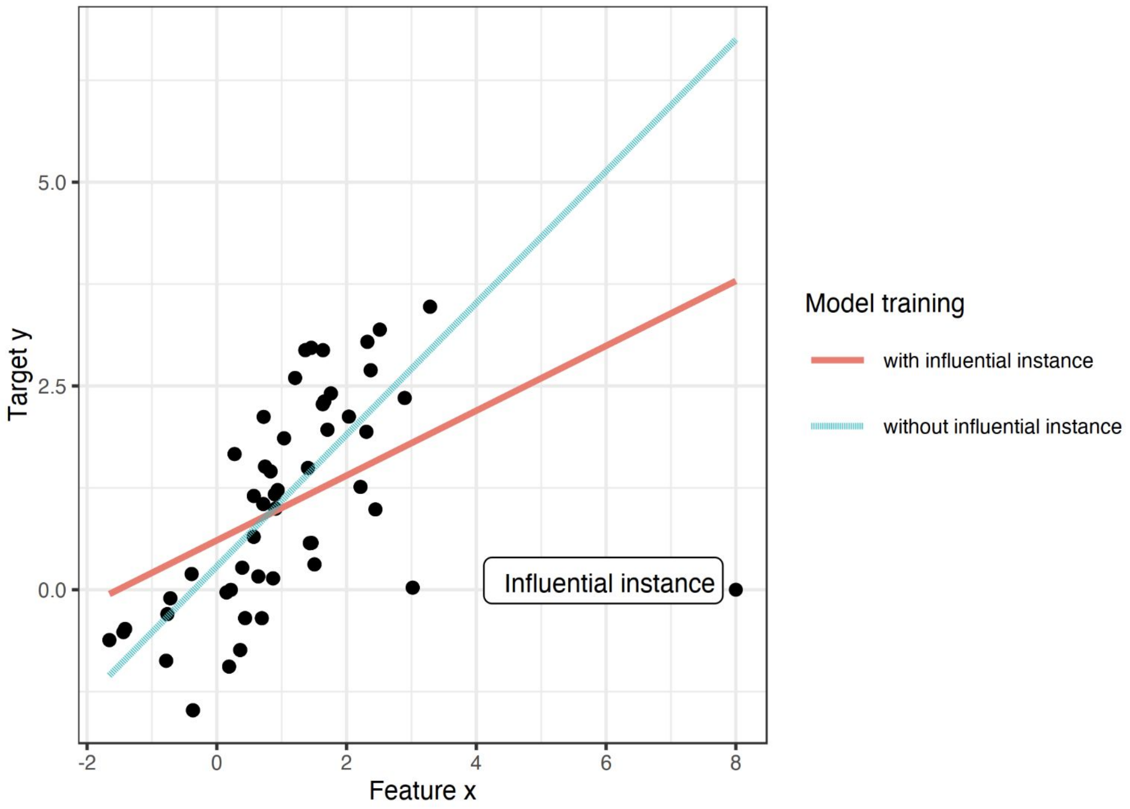
Compute the model which results from updating the parameters to reflect a slightly higher loss on *z*:

$$\hat{\theta}_{\epsilon,z} = \arg\min_{\theta\in\Theta}(1-\epsilon)\frac{1}{n}\sum_{i=1}^{n}L(z_i,\theta)+\epsilon L(z,\theta)$$

Classic result: This is the loss gradient of *z* wrt to the parameters times the inverse Hessian (matrix):

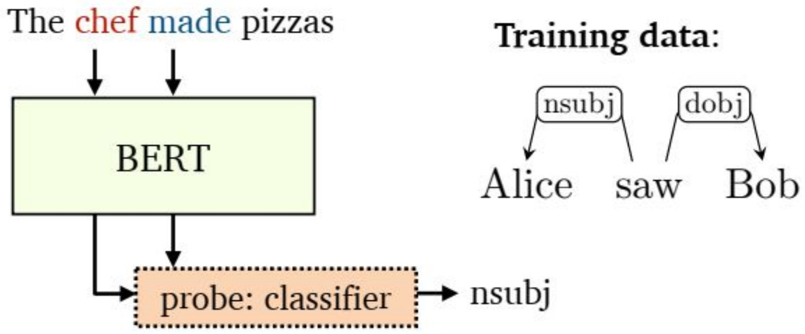
$$I_{\text{up,params}}(z) = \left.\frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon}\right|_{\epsilon=0} = -H_{\hat{\theta}}^{-1}\nabla_{\theta}L(z,\hat{\theta})$$

Which is the rate of change of the gradient: $H_{\theta} = \frac{1}{n}\sum_{i=1}^{n}\nabla_{\hat{\theta}}^{2}L(z_i,\hat{\theta})$



The probing framework

Let us denote by $f : x \rightarrow y$ the original model that maps input x to output y and is trained on the original dataset. This model generates intermediate representations of x, for example $f_l(x)$ may denote the representation of x at layer l of f. A probing classifier $g : f_l(x) \rightarrow z$ maps intermediate representations to some property z.



Exercise

What are potential pitfalls?

Hint: Probing seeks to identify whether information is already **present in** a network, not whether this is learnable from it.