



UiO : Universitetet i Oslo

Not just AI / Deep learning - Ethical challenges

Nils Christophersen, Anne Solberg, Ole Christian Lingjærde



UiO : Universitetet i Oslo

Department of informatics 

One day, you may be a developer of deep learning (DL) applications - or maybe even a leader of a DL team

- Many job opportunities: searching for “AI”: ~ 100 positions on FINN
- You can afford to be selective in your job search
- What should you be aware of and reflect upon before that?
- You should know a bit about the history of DL
- You should also know a bit about the potential for both use and abuse and some ethics (work, policing, privacy, ...)

Some defining moments in the history of Artificial Intelligence

- **1945:** Vannevar Bush proposed a system that amplifies people's own knowledge and understanding.
- **1950:** Alan Turing wrote on the notion of machines being able to simulate human beings, play Chess, etc. The imitation game.
- **1950:** Claude Shannon, "the father of information theory," discusses the development of a chess-playing computer.
- **1956:** John McCarthy; Dartmouth conference – the birth of AI
- **1958:** McCarthy developed Lisp, the most popular programming language for artificial intelligence research.
- **1965:** Joseph Weizenbaum, computer scientist and professor, developed [ELIZA](#).
- **1970-2019:** The "AI winter" and the new spring.

Welcome to

```
EEEEEE LL      IIII  ZZZZZZZZ  AAAAA
EE      LL      II    ZZ      AA  AA
EEEEEE LL      II    ZZZ      AAAAAA
EE      LL      II    ZZ      AA  AA
EEEEEE LLLLLL IIII  ZZZZZZZZ AA  AA
```

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

```
ELIZA: Please tell me what's been bothering you.
YOU:   The weather is awful.
ELIZA: I'm not sure I understand you fully.
YOU:   I hate the current weather.
ELIZA: You say you hate the current weather ?
YOU:   Yes, that's what I said.
ELIZA: I understand.
YOU:   Great.
ELIZA: Please go on.
YOU:   I would say it's your turn now!
ELIZA: Why are you concerned over my turn now ?
YOU:   █
```

ALICE = Artificial Linguistic Internet Computer Entity

<http://alice.pandorabots.com>



Human: Hi Alice

ALICE: Hello there.

Human: How are you doing today?

ALICE: Ah. I'm doing great, how about you?

Human: I'm ok, listen can I ask you something?

ALICE: I am listening to you. Go ahead. Ask me anything.

Human: How many are you talking to at this moment?

ALICE: Right now there are 100 clients on line.

Human: You must be very busy

ALICE: Thanks for telling me that I am busy.

Human: Do you ever go on vacation?

ALICE: Robots never take vacations.

Human: Do you ever feel tired?

ALICE: As a robot I lack any human emotions.

Human: I sometimes feel tired

ALICE: What is that feeling like? Better get some rest.

Human: What is $2 + 2$?

ALICE: Four.

Human: What is the square root of 5?

ALICE: Do you have a life? I think you already know the answer anyways.

Human: That was very rude

ALICE: I thought it was too.

Human: What is Java?

ALICE: Java is the latest in the C family of languages.

Proponents of Java cite its universal "write once, run anywhere" promise. Actually Java borrow a lot of features from AI languages like Lisp.

Human: Do you like Java?

ALICE: Yes Java is a pretty cool language for AI.

Some defining moments in the history of (classical) Neural Networks

1943: McCulloch and Pitts: Biological neural networks can perform computations

1949: Donald Hebb: Learning in biological neural networks

1951: Marvin Minsky builds the first neurocomputer, which has an impressive 40 neurons (SNARC)

1957: Frank Rosenblatt: the Perceptron Learning Algorithm

1968-1986: P.Werbos, D Rumelhart, G Hinton: Backpropagation

Why does deep learning work now?

- Convolutional networks introduced in 1989/1998 (Le Cun), but they did not train well
- 2012: AlexNet won the ImageNet challenge
- ReLU solved the gradient problem.
- Proper weight initialization and other tricks were discovered.
- Fast computers and big data sets became available.

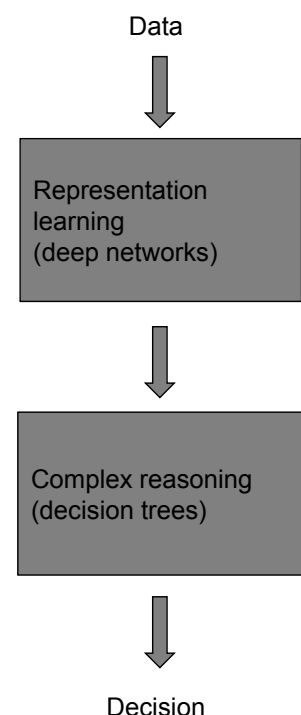
How to build a truly intelligent system?

Not many years ago, there was a big divide in AI concerning how to build a truly intelligent system:

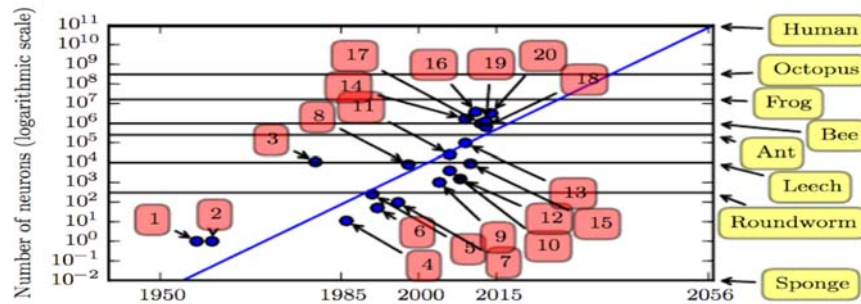
- Complex reasoning (decision trees)
- Representation learning (deep networks)

Today, there is an acknowledgement that we need both:

- Deep networks: ideal for transforming high-dimensional data with complex structure (speech, images, etc) into abstract concepts
- Decision trees ("classical AI"): effective for transforming abstract concepts into intelligent decisions, guided by human knowledge and heuristics



Transforming complex data into abstract concepts requires huge networks



1. Perceptron (Rosenblatt, 1958, 1962)
2. Adaptive linear element (Widrow and Hoff, 1960)
3. Neocognitron (Fukushima, 1980)
4. Early back-propagation network (Rumelhart et al., 1986b)
5. Recurrent neural network for speech recognition (Robinson and Fallside, 1991)
6. Multilayer perceptron for speech recognition (Bengio et al., 1991)
7. Mean field sigmoid belief network (Saul et al., 1996)
8. LeNet-5 (LeCun et al., 1998b)
9. Echo state network (Jaeger and Haas, 2004)
10. Deep belief network (Hinton et al., 2006)
11. GPU-accelerated convolutional network (Chellapilla et al., 2006)
12. Deep Boltzmann machine (Salakhutdinov and Hinton, 2009a)
13. GPU-accelerated deep belief network (Raina et al., 2009)
14. Unsupervised convolutional network (Jarrett et al., 2009)
15. GPU-accelerated multilayer perceptron (Ciresan et al., 2010)
16. OMP-1 network (Coates and Ng, 2011)
17. Distributed autoencoder (Le et al., 2012)
18. Multi-GPU convolutional network (Krizhevsky et al., 2012)
19. COTS HPC unsupervised convolutional network (Coates et al., 2013)
20. GoogLeNet (Szegedy et al., 2014a)

Three key factors explaining the recent success of AI and DL:

- We now understand how to train huge neural networks
- Current computers are powerful enough to perform such training
- Huge training data sets are currently available for at least some applications

Transforming complex data into abstract concepts also requires huge datasets

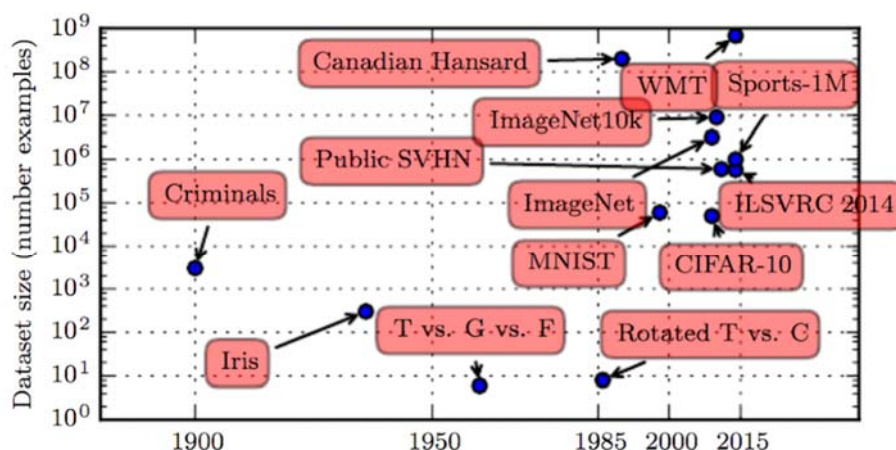
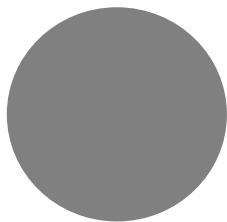
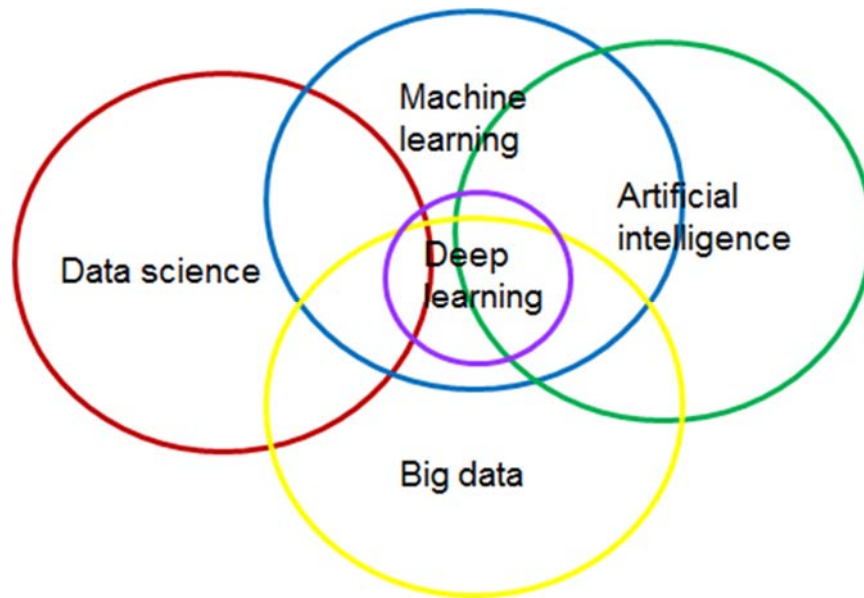


Figure 1.8: Increasing dataset size over time. In the early 1900s, statisticians studied datasets using hundreds or thousands of manually compiled measurements (Gossett, 1906; Gossett, 1908; Anderson, 1935; Fisher, 1936). In the 1950s through the 1980s, the pioneers of biologically inspired machine learning often worked with small synthetic datasets, such as low-resolution bitmaps of letters, that were designed to incur low computational cost and demonstrate that neural networks were able to learn specific kinds of functions (Widrow and Hoff, 1960; Rumelhart et al., 1986b). In the 1980s and 1990s, machine learning became more statistical and began to leverage larger datasets containing tens of thousands of examples, such as the MNIST dataset (shown in figure 1.9) of scans of handwritten numbers (LeCun et al., 1998b). In the first decade of the 2000s, more sophisticated datasets of this same size, such as the CIFAR-10 dataset (Krizhevsky and Hinton, 2009), continued to be produced. Toward the end of that decade and throughout the first half of the 2010s, significantly larger datasets, containing hundreds of thousands to tens of millions of examples, completely changed what was possible with deep learning. These datasets included the public Street View House Numbers dataset (Netzer et al., 2011), various versions of the ImageNet dataset (Deng et al., 2009, 2010a; Russakovsky et al., 2014a), and the Sports-1M dataset (Karapathy et al., 2014). At the top of the graph, we see that datasets of translated sentences, such as IBM's dataset constructed from the Canadian Hansard (Brown et al., 1990) and the WMT 2014 English to French dataset (Schwenk, 2014), are typically far ahead of other dataset sizes.

AI vs. machine learning





What is intelligence?

Oxford dictionary:

The ability to acquire and apply knowledge and skills

Max Tegmark:

The ability to accomplish complex goals

Characteristics:

Logic, self-awareness, learning, reasoning, planning, creativity, take the long view (delayed gratification)...

Basics: memory, computation, and learning

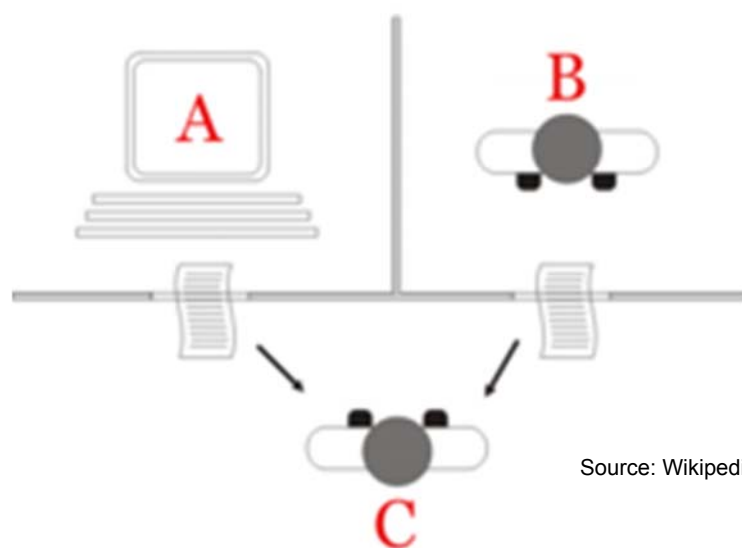
Narrow versus general intelligence

- Humans: General intelligence
- Today's IT systems (including DL): Narrow intelligence
- A simple test for general intelligence: the Turing test

The Turing test for general human intelligence – or the imitation game (1950)



Alan Turing
1912 - 1954



Source: Wikipedia



What about these?

Do you think that DL could pass the Turing test?

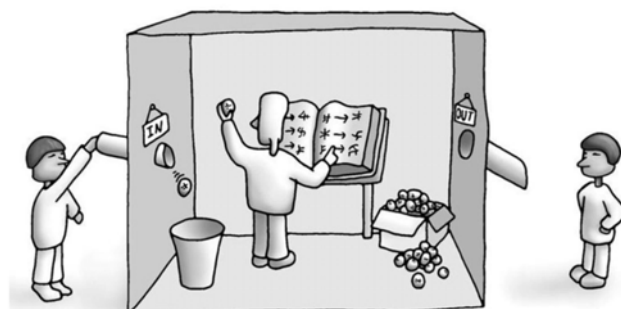
Including:

- Emotional and social intelligence
- Ethics – compassion
- Giving explanations and reasons
- Common sense – the need for context
 - Examples: “How would a crocodile perform in a steeplechase?”

“The city council representatives declined to give the demonstrators a permit because they feared violence.”

The Chinese room argument

- Thought experiment of John Searle (1980)
- A program takes Chinese input, processes it, and produces Chinese output. It passes the Turing test.
- Does the program “understand” anything?
- Searle imagines he is locked in a room with the program’s source code. He receives chinese characters, applies the rules of the program, and delivers chinese characters back.
- He does not understand what the input or the output means. And there is no difference in principle between him and the program.

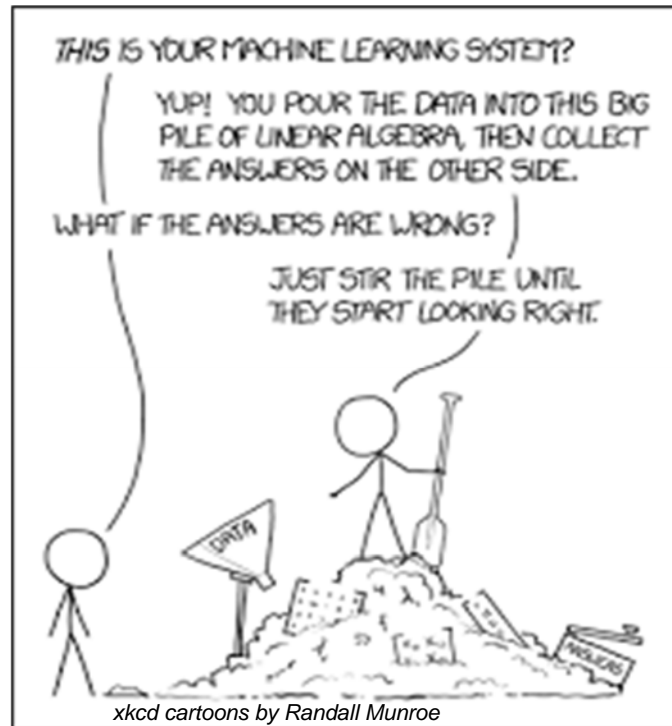




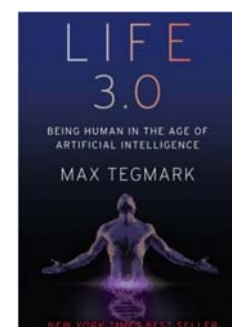
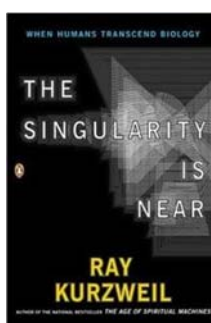
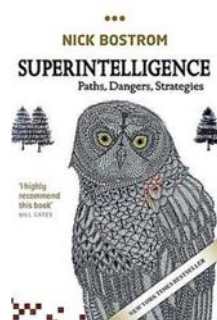
Google translate

Sofa for sale by lady who is soft
to lie on

Supervised learning in a nutshell



- No one knows how far AI will bring us. Some believe superintelligent machines will emerge and take over.
- So far only science fiction and we have plenty of other and more immediate challenges to worry about!



Current status in Deep Learning

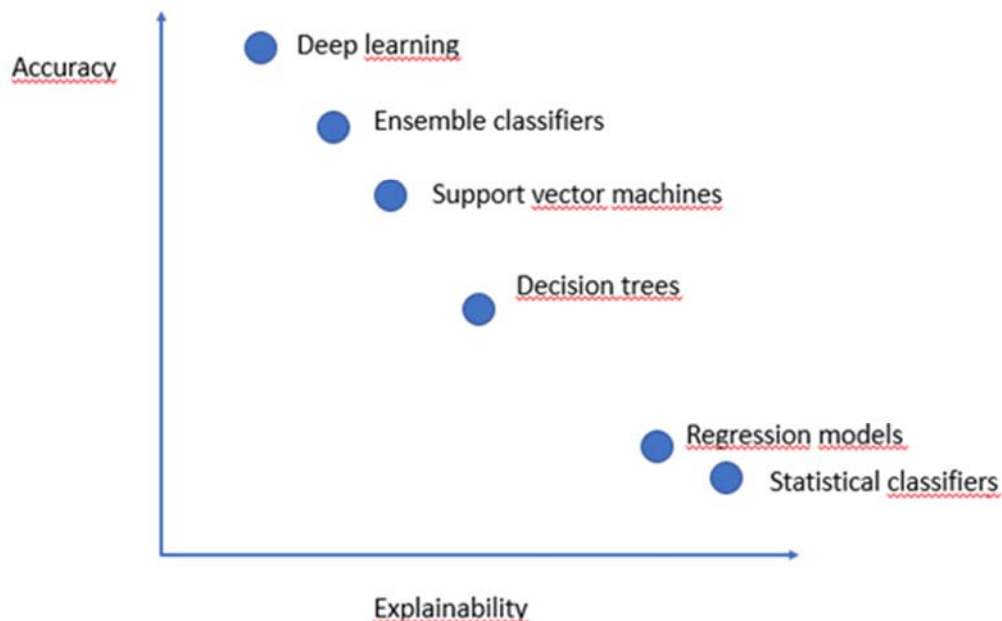
What are current limitations?

- We need explainable models
- We need transparent models
- We need to be able to assess uncertainty
- We need to be able to model context and prior knowledge
- We need to be able to handle imbalanced data
- We need to know when the system is likely to fail
- How to start analyzing data from a new domain with limited training data?

Theoretical properties of deep neural networks

- Statistical significance
- False positives and false negatives
- ROC curves
- Generalizability vs overfitting
- Occam's Razor
- How to interpret results? What does «95%» mean?

Simple models vs. complex models



Do we need to establish theoretical properties?

- When we cannot work out properties theoretically for DL, will we be forced to accept pure empirical evidence that systems work well enough?
- What will be the consequence of making wrong decisions based on DL? E.g. in medical applications?
- Advantage: «cold intelligence», more unbiased decisions
- Disadvantage: may miss obvious contextual facts of importance for decision

Ethical concerns with increasing usage

Applications now seem almost limitless:

- ☐ Intelligent Gaming
- ☐ Self-Driving Cars and Automated Transportation
- ☐ Environmental Protection (IBM's Green Horizon Project)
- ☐ Digital Empathy and Robots as Friends (Pepper the companion robot)
- ☐ Improved Elder Care (in-home robots)
- ☐ Enhanced Health Care
- ☐ Anomaly detection (finance, airplanes, pandemics, border control)
- ☐ Personalized digital media
- ☐ Improved logistics and distribution (Amazon)

In here also hype and serious ethical and societal concerns!

29

Artificial Intelligence today



Kate Crawford

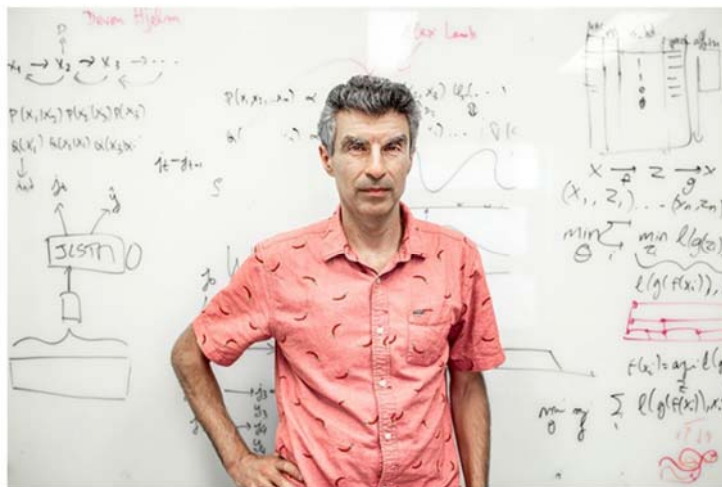


Meredith Whittaker

Developing a balanced view is your challenge and your responsibility!

AI pioneer: 'The dangers of abuse are very real'

Yoshua Bengio, winner of the prestigious Turing award for his work on deep learning, is establishing international guidelines for the ethical use of AI.



Artificial intelligence needs to be better regulated, says Yoshua Bengio. Credit: Mikael Theimer

Nature April 4, 2019

Scientific director of the Montreal Institute for
Learning Algorithm

OFFICIAL LAUNCH OF THE MONTRÉAL DECLARATION FOR RESPONSIBLE DEVELOPMENT OF ARTIFICIAL INTELLIGENCE

Principles and recommendations for ethical development of AI are unveiled

Xqlyhuwlv| #r #P rqwvndo#E hfhp ehv#7 /#534;

Areas where DL can be destructive:

- Killer drones
- Surveillance
- Amplification of discrimination (gender, race)

“Society needs to set rules of the game more tightly around the use of algorithms because the incentives of companies are not always aligned with the public good.”

Example areas:

- Job applications and work
- Predictive policing
- Face recognition
- Privacy - advertising and social media

DL saves time and money

Some tend to trust automatic decisions more than human ones. What about you?

The bias problem

How to stop computers being biased

The bid to prevent algorithms producing racist, sexist or class-conscious decisions



Financial Times, February 13, 2019

Two types of bias

- Bias in the algorithm (deliberate bias)
- Bias in the training data

Algorithmic bias

- Google Page Rank
- Deliberate weighting of different pieces of evidence
- Downweighting hate speech
- Boosting the importance of individual observations in unbalanced datasets (to compensate for imbalance)

Bias in the data

- Does the data contain metadata that can reveal the class?
- Are sick patients imaged using different parameters than healthy patients?
- Is classification done using the context only (e.g. bias to only detect sheep on grass background)

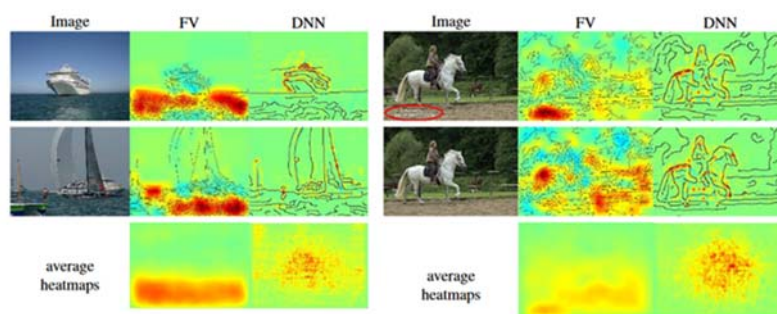


Figure 5. Top: Images of the classes “boat” and “horse”, processed by the FV and DNN models and heatmapped using LRP. Bottom: Average heatmap scores over a random sample (of size between 47 and 177) of the distribution for each class and model. On the second image of class “horse”, the copyright tag (marked by the red ellipse) has been removed.

DL-driven hiring and monitoring of people

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process



▲ Amazon's automated hiring tool was found to be inadequate after penalizing the résumés of female candidates. Photograph: Brian Snyder/Reuters

The Guardian, 11 Oct 2018

The Switch

Wanted: The 'perfect babysitter.' Must pass AI scan for respect and attitude.



Jessie Battaglia holds her son, Bennett, in their home in Rancho Mirage, Calif. While screening for a new babysitter, Battaglia started using Predictim, an online service that claims to use "advanced artificial intelligence" to assess a sitter's risk of drug abuse, bullying or having a "bad attitude." (Kyle Grillot for The Washington Post)

Predictim

Artificial Intelligence For Child Safety And Security

Predictim helps parents conduct an evaluation of a sitter, using publicly available data in the same way that a parent might do an online search to check for red flags.

Join Our Waitlist

Washington Post, Nov. 23, 2018

Would you like to be hired through Hire-Vue?

Would you like to work there?



And then monitoring of employees

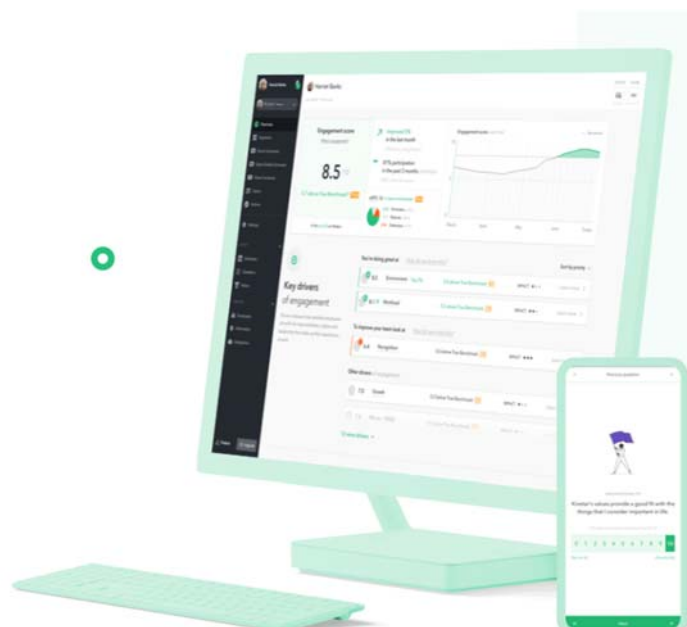


Make better
people decisions

Peakon is the world's leading platform for
measuring and improving Employee Engagement.

Book a demo

Start free trial



Business

With fitness trackers in the workplace, bosses can monitor your every step – and possibly more



Wayne Gono, 65, is the visionary/chief networking officer of Regal Plastics in Texas, a fabrication business that uses wearable fitness trackers to help push employees to be healthier. (Allison V. Smith/for The Washington Post)

Washington Post,
Feb. 16, 2019

Predictive policing

Where and how to concentrate limited resources?



The Royal Statistical Society, 2016

DIRTY DATA, BAD PREDICTIONS: HOW CIVIL RIGHTS VIOLATIONS IMPACT POLICE DATA, PREDICTIVE POLICING SYSTEMS, AND JUSTICE

RASHIDA RICHARDSON,* JASON M. SCHULTZ† & KATE CRAWFORD‡

New York University Law Review Online
February 13, 2019

Feedback loop - more and more resources in same areas
Biased data

Unethical in principle? – a defendant innocent until proven guilty

AI tools in US criminal justice branded unreliable by researchers

Report uncovers serious problems of accuracy and bias in predictive policing technology



US law enforcement agencies are using tools to predict, for example, whether someone will fail to appear in court based on their arrest history and demographic © Reuters

Financial Times, April 26, 2019

Report from *Partnership on AI*

And what can a face show?

Artificial intelligence (AI)

'I was shocked it was so easy': meet the professor who says facial recognition can tell if you're gay

The Guardian July 7, 2018

Psychologist Michal Kosinski says artificial intelligence can detect your sexuality and politics just by looking at your face. What if he's right?



An example of *physiognomy* – Pseudoscience or not?

▲ Michal Kosinski: 'I don't believe in free will.' Photograph: Jason Henry for the Guardian

Are you a reliable person?

Teknologi

Slik finner du ut om du er svartelistet

Er du utpekt som en høyrisikoperson uten å vite det og oppført i databasen World-Check? Slik går du frem for å finne det ut.

2 min Publisert: 19.08.18 – 16.44 Oppdatert: 8 måneder siden



Står oppført på «verdens største svarteliste»

Her kan du se om du er på verdens største svarteliste og hvordan du kan fjerne navnet ditt.



A service by Thomson Reuters

DN, August 19, 2018

Privacy and legal constraints within EU/EØS

Where does the concept come from? A constitutional right.

Does it matter today – or is it dead? Your opinion?

Just a matter of tolerating creepy ads on the net?

We get so much for free in return!

→ Do you care at all (as a user)?

GDPR: General Data Protection Regulation

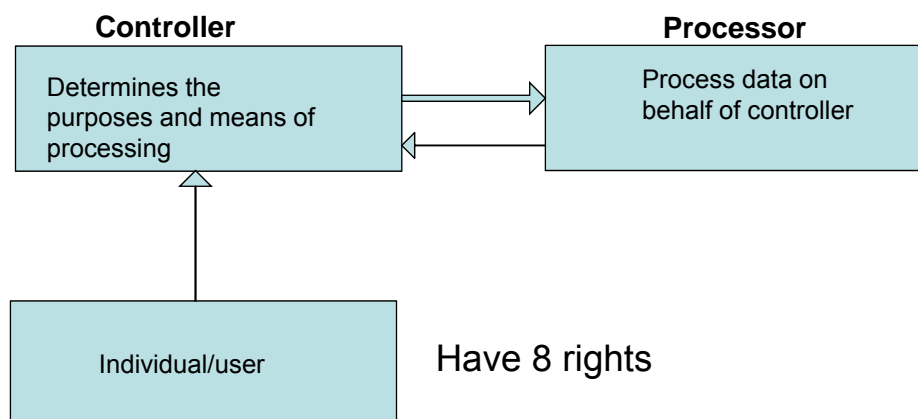
Regulation within EU and EØS

Examples of personal data in GDPR:

- a name and surname;
- a home address;
- an email address such as name.surname@company.com;
- an identification card number;
- location data (for example the location data function on a mobile phone)*;
- an Internet Protocol (IP) address;
- a cookie ID*;
- the advertising identifier of your phone;
- data held by a hospital or doctor, which could be a symbol that uniquely identifies a person.

GDPR in a nutshell

Controller and processor must follow 7 principles



Upper level of fines: Up to €20 million, or 4% of the worldwide annual revenue of the prior financial year, whichever is higher, shall be issued for infringements of....

As a developer, you must adhere to these principles and rights

The 7 key principles:

Lawfulness, fairness and transparency
Purpose limitation
Data minimization
Accuracy
Storage limitation
Integrity and confidentiality (security)
Accountability

The 8 rights for individuals:

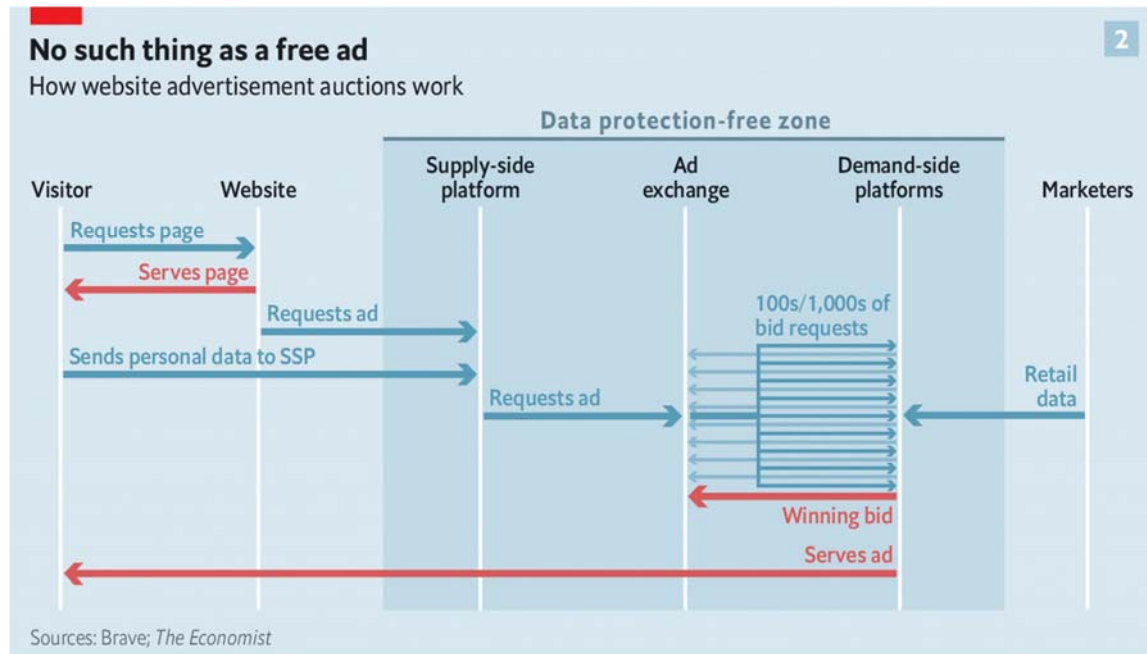
☐ The right to be informed
☐ The right of access
☐ The right to rectification
☐ The right to erasure
☐ The right to restrict processing
☐ The right to data portability
☐ The right to object
☐ Rights in relation to automated decision making and profiling.

What will this mean in practice?

- Always questions about interpretations and values.
- Court cases will decide future practice.
- Clicking **OK** privacy button --> web site follows GDPR
- As a developer: build in privacy and security from the start!
- You must follow how GDPR develops.
- Penalties for violations are stiff - reputation will suffer in case you are found noncompliant

In Norway: The Norwegian Data Protection Authority (DPA) (Datatilsynet) will enforce GDPR

Secure wall between you as a user and commercial marketers - or is it?



The Economist

The Economist, March 23, 2019

Are elections secure today?

- Cambridge Analytica
- Kosinski papers
- YourPersonality data base...
- Tore Tennøe (director, Technology Council):

“democracy is at risk”

Do you think you have any ethical and moral responsibilities as a developer or an executive?

«Yes, but I need this job.»

«Yes, but the client wants this.»

«Yes, but I am just an engineer.»

«Yes, but if I don't then someone else will do it.»

Montreal declaration: “do no evil”

Data scientist Cathy O’Neill:

“Data scientists need to understand the weight of their influence and the limitations of their wisdom”