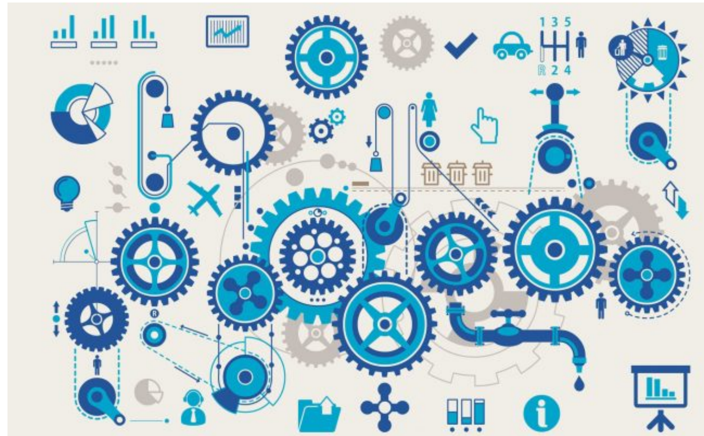
The case for **Normware**

Giovanni Sileno (g.sileno@uva.nl)

SNE group meeting, 21 February 2019

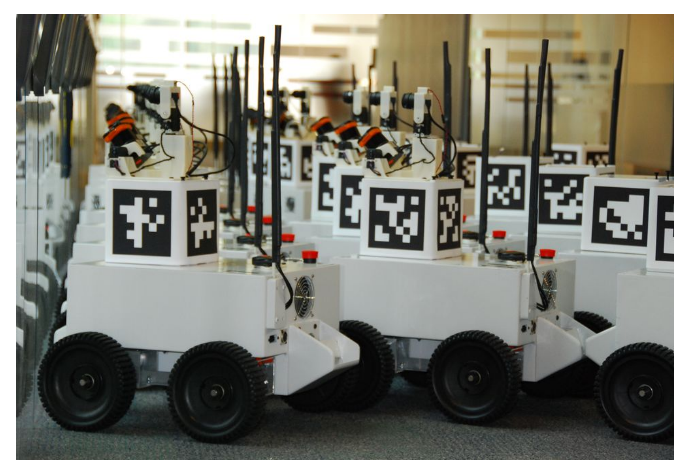
Extension, refinement of what presented in:

Sileno, G., Boer, A. and van Engers, T., The Role of Normware in Trustworthy and Explainable AI, Proceedings of XAILA workshop: Explainable AI and Law, in conjunction with JURIX 2018

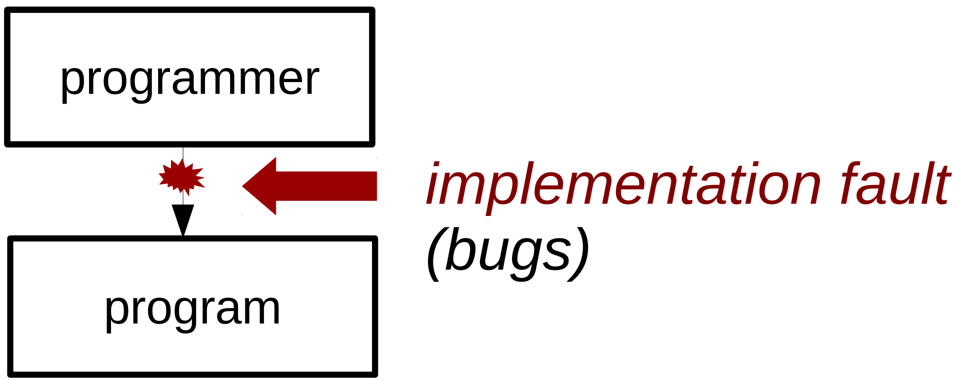
with the (supposedly) near advent of ***autonomous artificial entities***, or any other forms of ***distributed automatic decision making***,

- humans less and less in the loop
- increasing concerns about ***unintended consequences***

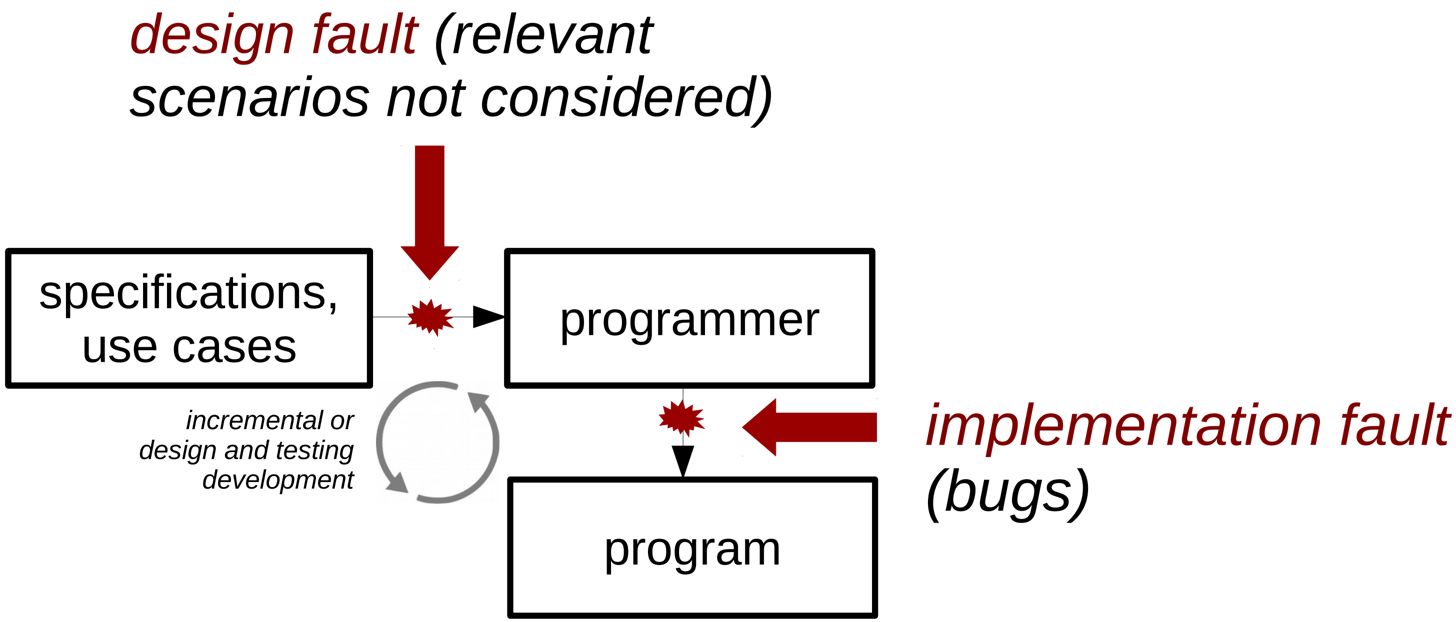
 

Unintended consequences: bad or limited design

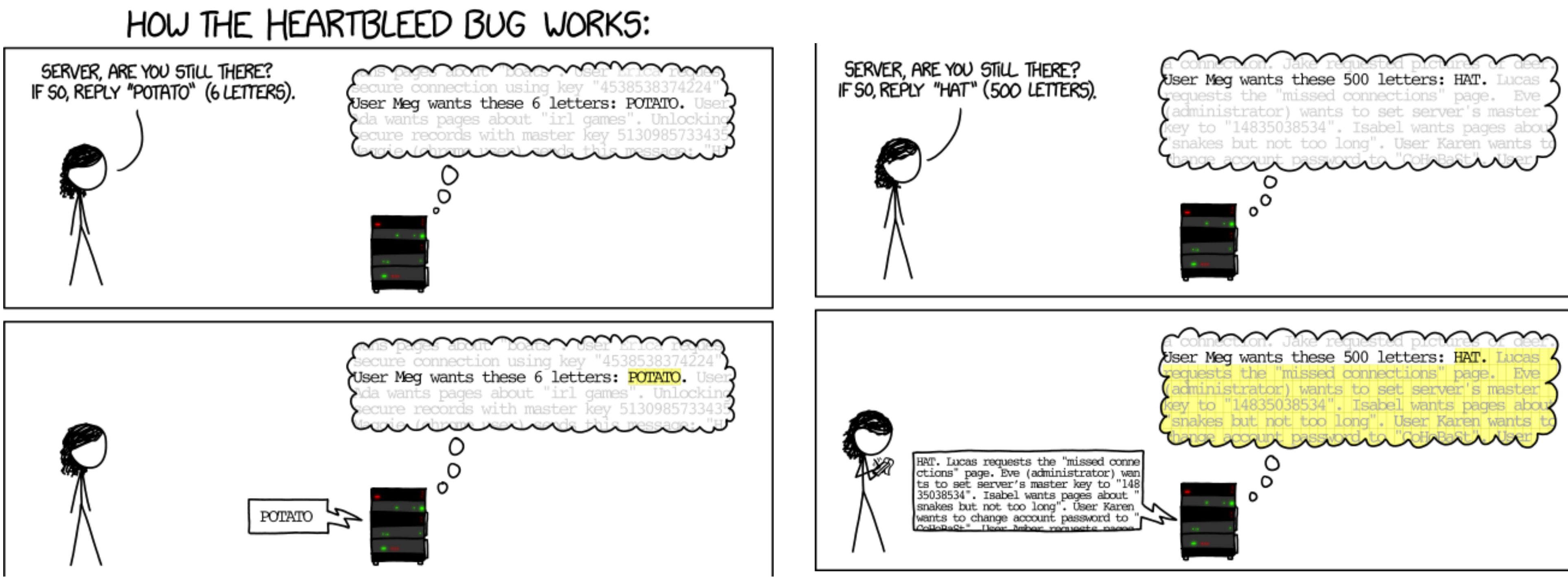
Unintended consequences: bad ~~or limited~~ design



Unintended consequences: ~~bad or~~ limited design



Unintended consequences: bad or limited design



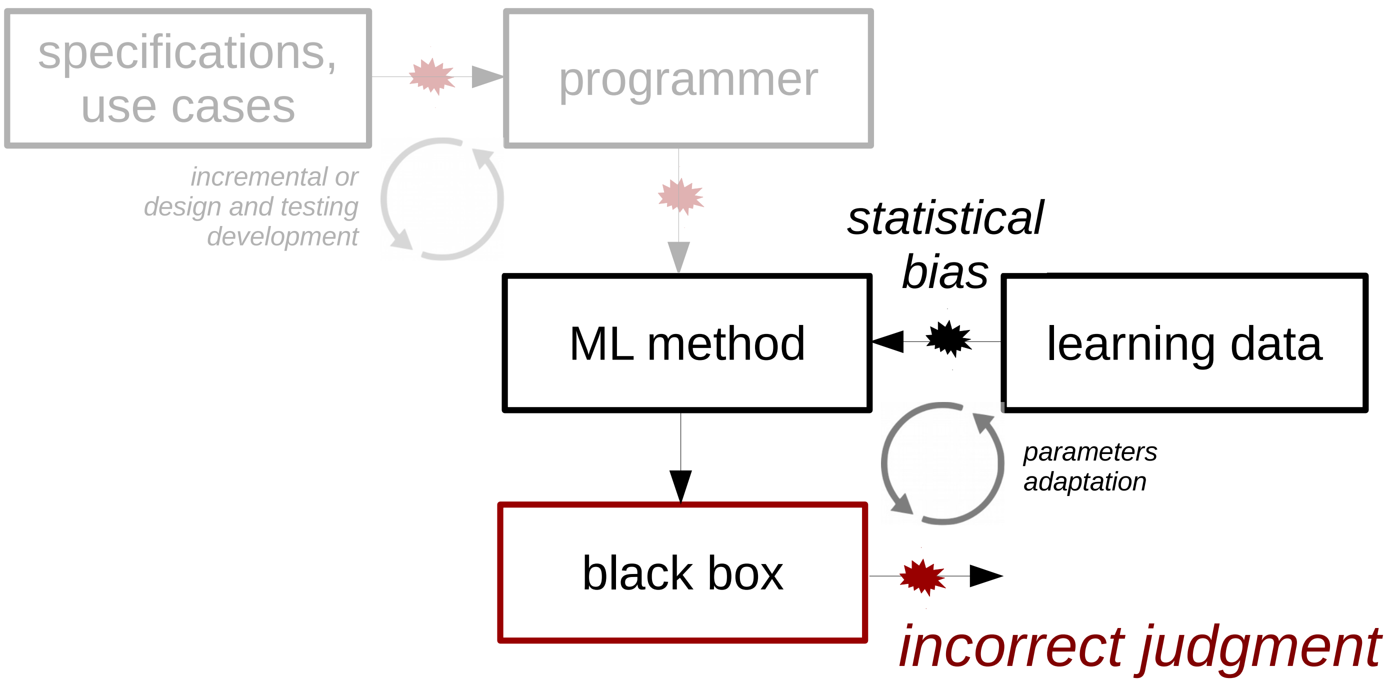
- Example: Heartbleed Bug with OpenSSL (CVE-2014-0160)
  - weakness allows stealing the information protected, under normal conditions, by the SSL/TLS encryption used to secure the Internet.
  - bug was introduced in December 2011 and has been out in the wild since OpenSSL release 1.0.1 on 14th of March 2012. OpenSSL 1.0.1g released on 7th of April 2014 fixes the bug.

Unintended consequences: bad or limited design

- Wallet hacks, fraudulent actions and bugs in the in the ***blockchain*** sector during 2017:
  - CoinDash ICO Hack ($10 millions)
  - Parity Wallet Breach ($105 millions)
  - Enigma Project Scum
  - Parity Wallet Freeze ($275 millions)
  - Tether Token Hack ($30 millions)
  - Bitcoin Gold Scam ($3 millions)
  - NiceHash Market Breach ($80 millions)

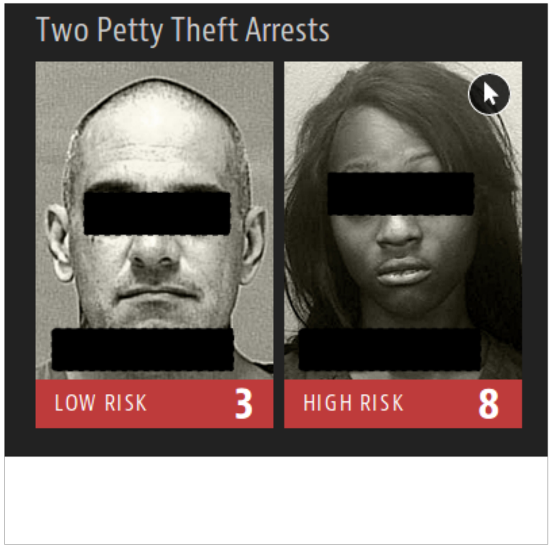
Source: CoinDesk (2017), *Hacks, Scams and Attacks: Blockchain's 2017 Disasters*

Unintended consequences: the “artificial prejudice”



Unintended consequences: the “artificial prejudice”

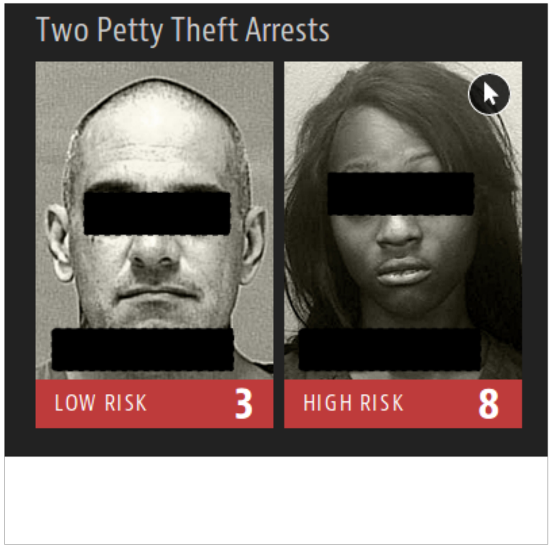
- Software used across the US predicting future crimes and criminals biased against African Americans (2016)



Angwin J. et al. ProPublica, May 23 (2016). *Machine Bias: risk assessments in criminal sentencing*

Unintended consequences: the “artificial prejudice”

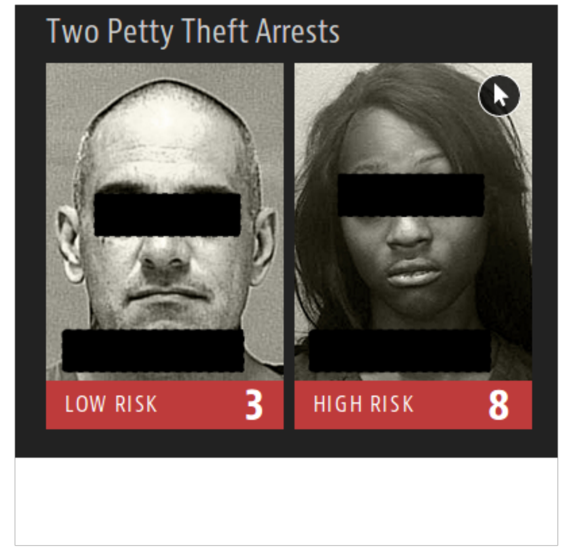
- Software used across the US predicting future crimes and criminals biased against African Americans (2016)
  - Existing statistical bias (correct **description**)
  - When used for prediction on an individual it is read as ***behavioural predisposition***, i.e. it is interpreted as a **mechanism**.
  - A biased judgment introduces here negative consequences in society.

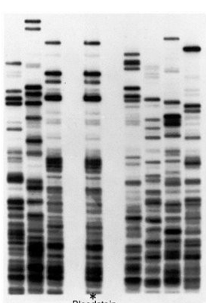


Angwin J. et al. ProPublica, May 23 (2016). *Machine Bias: risk assessments in criminal sentencing*

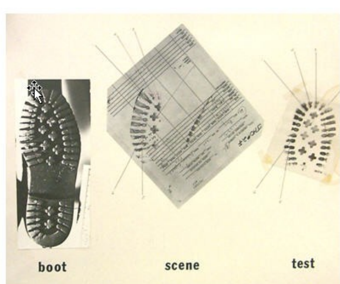
Unintended consequences: the “artificial prejudice”

- Software used across the US predicting future crimes and criminals biased against African Americans (2016)
- **Problem**: role of *circumstantial evidence,* how to integrate statistical inference in judgment?





DNA



footwear

...

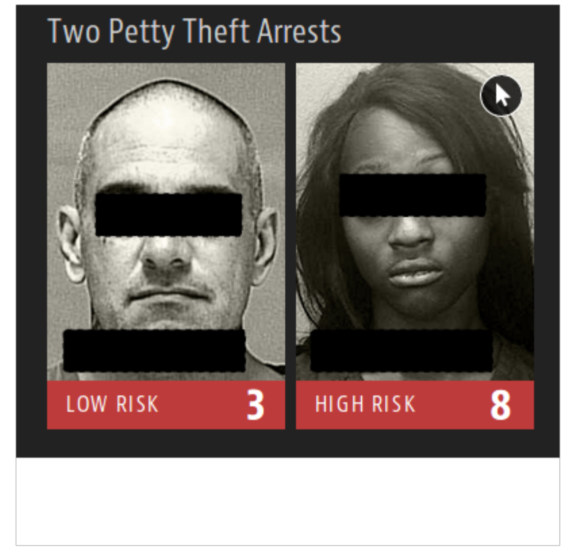
origin, gender,
ethnicity, wealth, ...

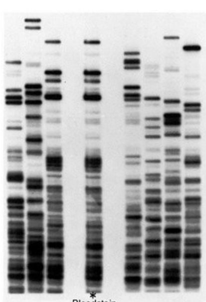
improper profiling?

Angwin J. et al. ProPublica, May 23 (2016). *Machine Bias: risk assessments in criminal sentencing*

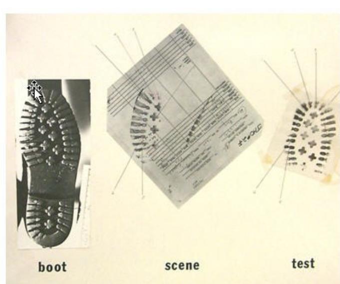
Unintended consequences: the “artificial prejudice”

- Software used across the US predicting future crimes and criminals biased against African Americans (2016)
- **Problem**: role of *circumstantial evidence,* how to integrate statistical inference in judgment?





DNA



footwear

...

origin, gender, ethnicity, wealth, ...

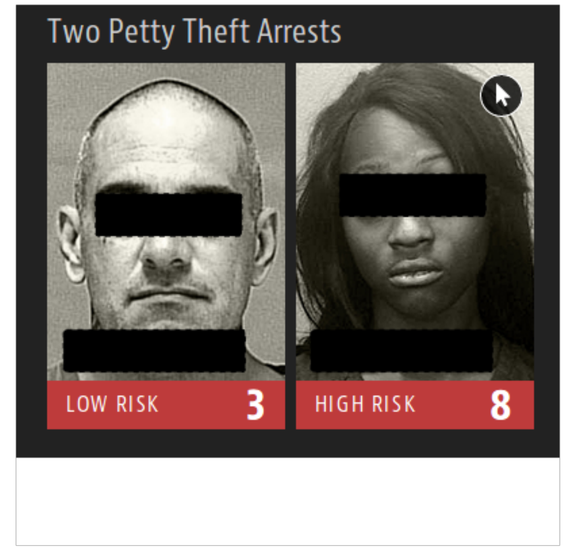
improper because it causes ***unfair judgment***

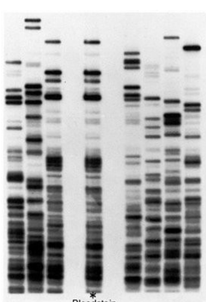
improper profiling?

Angwin J. et al. ProPublica, May 23 (2016). *Machine Bias: risk assessments in criminal sentencing*

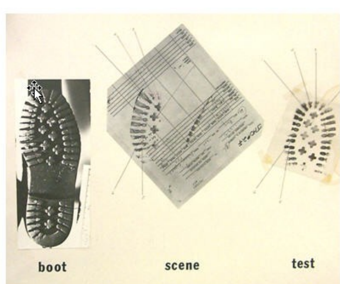
Unintended consequences: the “artificial prejudice”

- Software used across the US predicting future crimes and criminals biased against African Americans (2016)
- **Problem**: role of *circumstantial evidence,* how to integrate statistical inference in judgment?





DNA



footwear

...

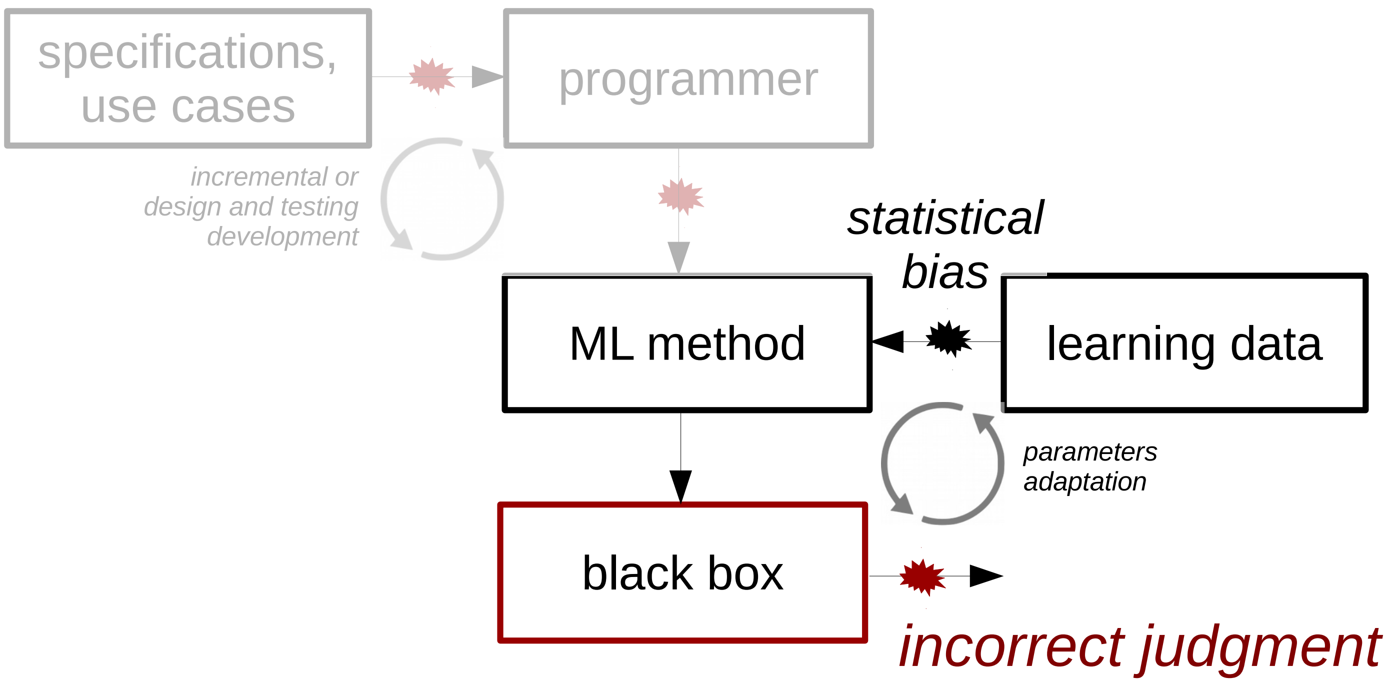
origin, gender, ethnicity, wealth, ...

improper
because it causes
unfair judgment

improper profiling?

Norms determine which factors are acceptable or not.

Unacceptable conclusions: improvident induction



- The “improvident” qualification to an inductive inference might be given already before taking into account the practical consequences of its acceptation.

Unacceptable conclusions: improvident induction

- Country A’s army demands a classifier to recognize whether a tanks is from country A or country B. It provides the developers with a series of photos of tanks from both countries.

Unacceptable conclusions: improvident induction





- Country A’s army demands a classifier to recognize whether a tanks is from country A or country B. It provides the developers with a series of photos of tanks from both countries.
- After the training, the developers investigate by introspection the activation patterns. They discover that “**daylight**” is a major factor supporting a B-tank classification. Returning on the source data, the developers discovered that there was *no photo of B-tanks at night*.

Unacceptable conclusions: improvident induction

- Country A’s army demands a classifier to recognize whether a tanks is from country A or country B. It provides the developers with a series of photos of tanks from both countries.

- After the training, the developers investigate by introspection the activation patterns. They discover that “**daylight**” is a major factor supporting a B-tank classification. Returning on the source data, the developers discovered that there was *no photo of B-tanks at night*.

→ **statistical biases endanger ML predictive abilities**

Unacceptable conclusions: improvident induction

- Country A’s army demands a classifier to recognize whether a tanks is from country A or country B. It provides the developers with a series of photos of tanks from both countries.

- After the training, the developers investigate by introspection

*1. move the focus from **software engineering** to **data engineering***

photo of B-tanks at night.

statistical biases endanger ML predictive abilities

Unacceptable conclusions: improvident induction

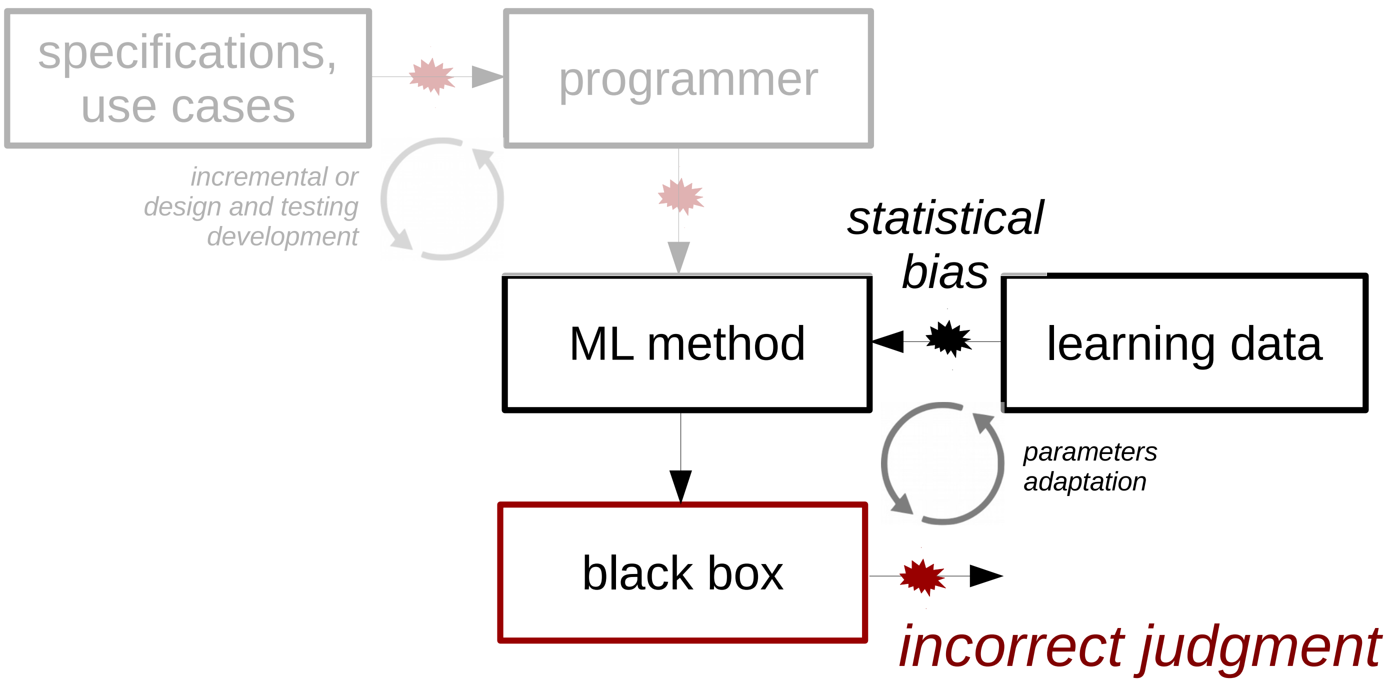
- Country A’s army demands a classifier to recognize whether a tanks is from country A or country B. It provides the developers with a series of photos of tanks from both countries.
- After the training, the developers investigate by introspection

*1. move the focus from **software engineering** to **data engineering***

photo of B-tanks at night.

*2. an **expert** would reject the conclusion when **no relevant mechanism** can be imagined linking factor with conclusion.*

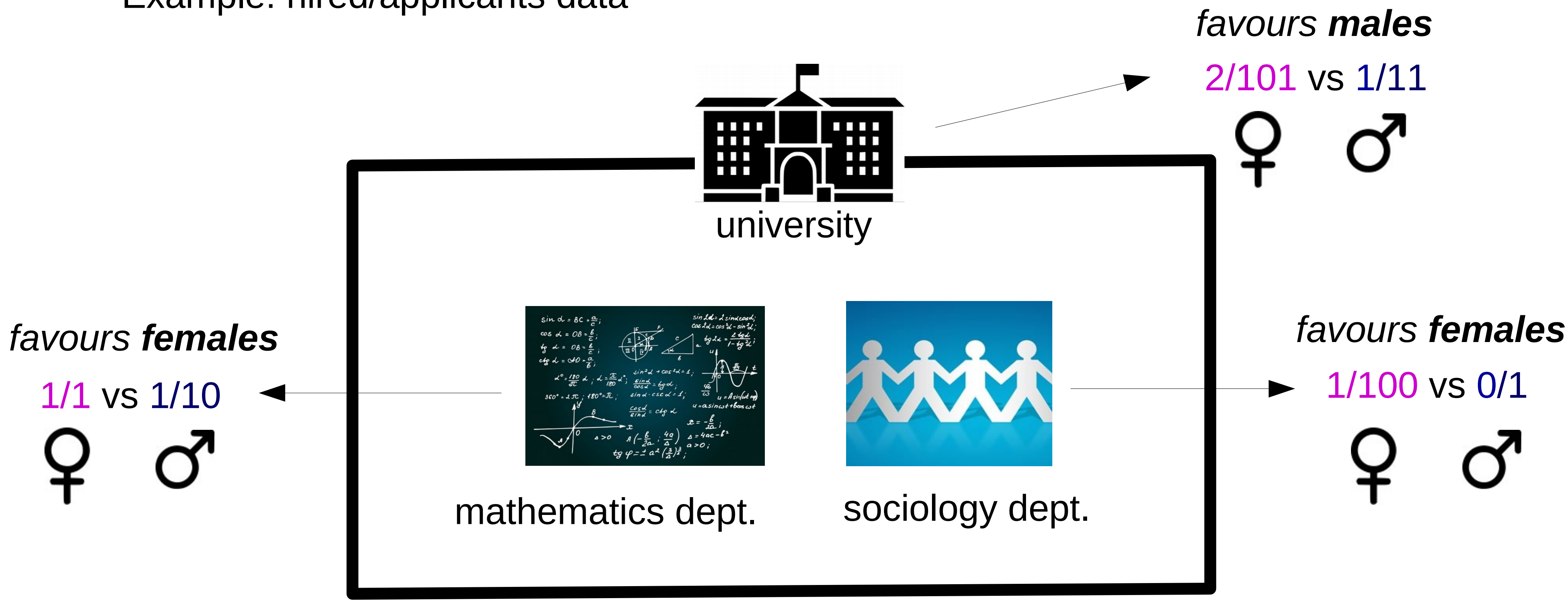
Unacceptable conclusions: improvident induction



- Problems may also arise for the statistical inference by itself, as shown e.g. by **Simpson's paradox**

Unacceptable conclusions: improvident induction

Example: hired/applicants data

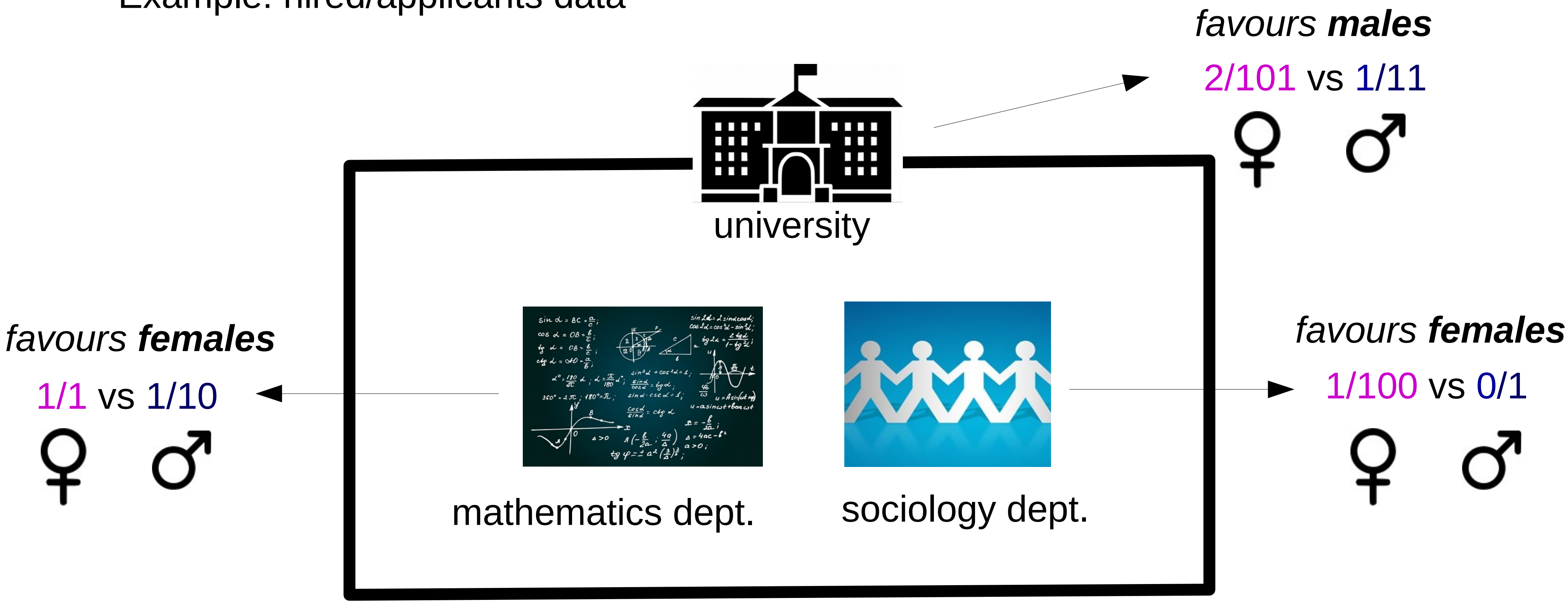


- Problems may also arise for the statistical inference by itself, as shown e.g. by **Simpson’s paradox**

Pearl, J. (2014). Understanding Simpson’s Paradox. The American Statistician, 68(1), 8–13.

Unacceptable conclusions: improvident induction

Example: hired/applicants data



- Problems may also arise for the statistical inference by itself, as shown e.g. by **Simpson’s paradox**

Only causal mechanisms enable to select an interpretation.

Explainable AI

- Explainable AI has basically two drivers:
  - *reject unacceptable conclusions*
  - *satisfy reasonable requirements of expertise*
- But what qualifies a conclusion as “unacceptable”? And what might be used to define an expertise to be “reasonable”?

Explainable AI

- Explainable AI has basically two drivers:
  - *reject unacceptable conclusions*
  - *satisfy reasonable requirements of expertise*
- But what qualifies a conclusion as “unacceptable”? And what might be used to define an expertise to be “reasonable”?
- claim: **normware**!

  i.e. *computational artifacts specifying shared expectations* (“norm” as in ***normality***)

Trustworthy AI

- **Trustworthiness** for artificial devices could be associated to the requirement of not falling into *paperclip maximizer* scenarios:
  - *of not taking "wrong" decisions, of performing "wrong" actions, wrong because having disastrous impact*
- How to (attempt to) satisfy this requirement?

Trustworthy AI

- **Trustworthiness** for artificial devices could be associated to the requirement of not falling into *paperclip maximizer* scenarios:
  - *of not taking “wrong” decisions, of performing “wrong” actions, wrong because having disastrous impact*
- How to (attempt to) satisfy this requirement?
- claim: **normware**!

  i.e. *computational artifacts specifying shared drivers* (“norm” as in ***normativity***)

A tentative taxonomy

hardware

physical device

when running →
physical mechanism

situated in
a physical environment

control structure

software

symbolic device

when running →
symbolic mechanism

relies on physical
mechanisms

control structure

?

normware

....

....

relies on symbolic
mechanisms

....

A tentative taxonomy

hardware

physical device

when running → physical mechanism

situated in a physical environment

software

symbolic device

when running → symbolic mechanism

relies on physical mechanisms

?

normware

………..

………..

relies on symbolic mechanisms

*Is **normware** just a type of software?*

A tentative taxonomy

?

hardware	software	normware
physical device	symbolic	.....
when running → physical mechanism	when ru symbo	....
situated in a physical environment	relies on physical mechanisms	relies o symb mechan

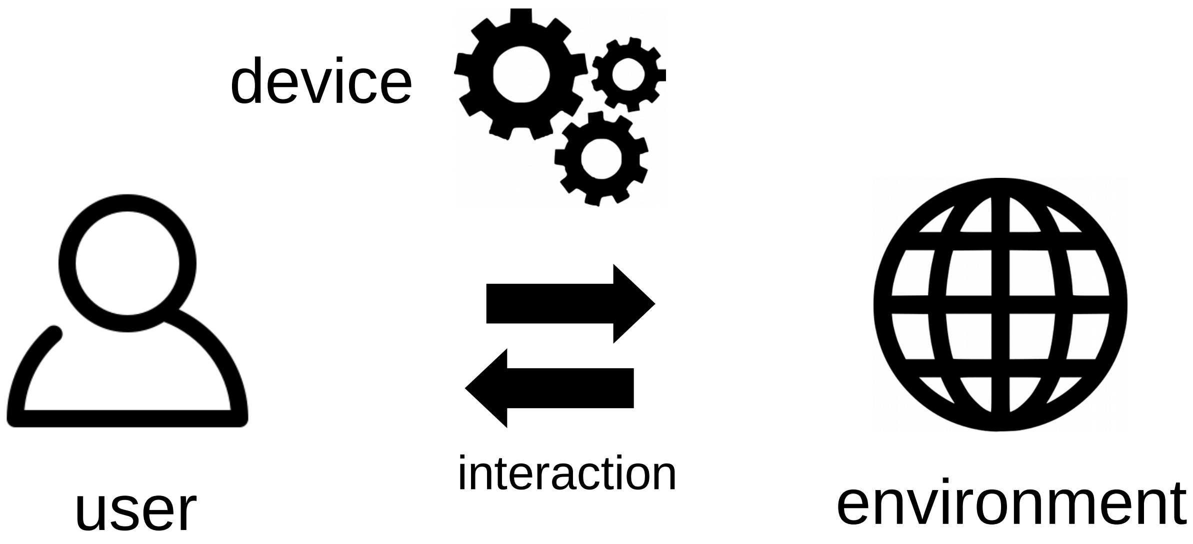
normative and epistemic pluralism?

interaction with sub-symbolic modules?

Is ***normware*** *just a type of software?*

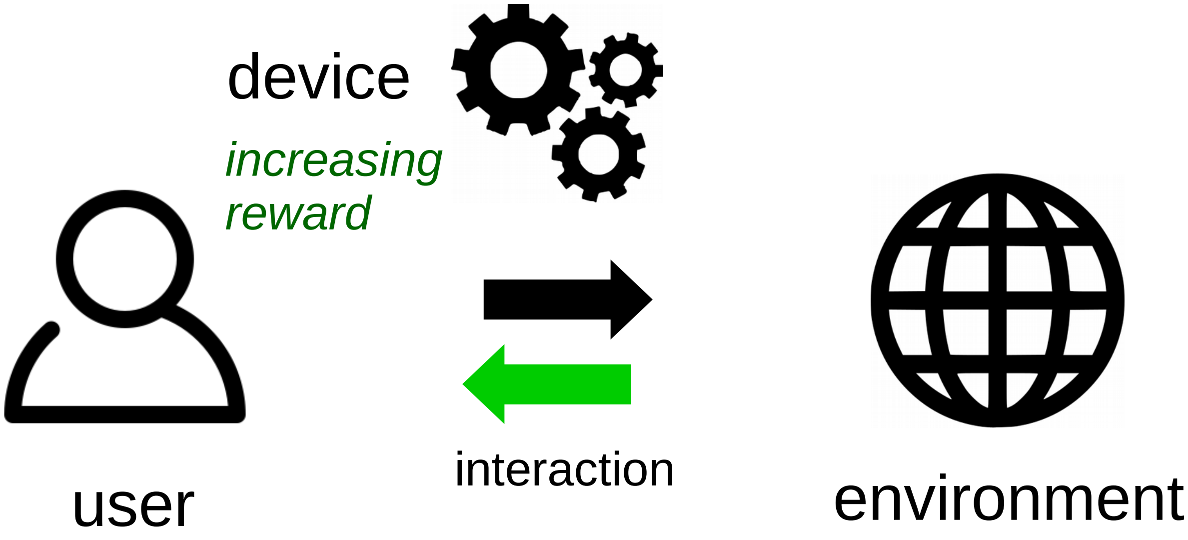
Impact *at large*

- Traditionally, engineering is about the conception of *devices* to implement certain *functions*. Functions are always defined within a certain **operational context** to satisfy certain **needs**.



Impact *at large*

- Traditionally, engineering is about the conception of *devices* to implement certain *functions*. Functions are always defined within a certain **operational context** to satisfy certain **needs**.



- **optimization** is made possible by specifying a **reward** function associated to certain **goals**

Impact *at large*

goal: fishing,
reward: proportional to quantity of fish, inversely to effort.

individual solution to optimization problem:

Impact *at large*

goal: fishing,
reward: proportional to quantity of fish, inversely to effort.

individual solution to optimization problem:

"fishing with bombs"

Impact *at large*

goal: fishing,
reward: proportional to quantity of fish, inversely to effort.

individual solution to optimization problem:

*“**fishing with bombs**”*

→ *acknowledgement of **undesirable** second-order effects.*

Impact *at large*

goal: fishing,

reward: proportional to quantity of fish, inversely to effort.

individual solution to optimization problem:

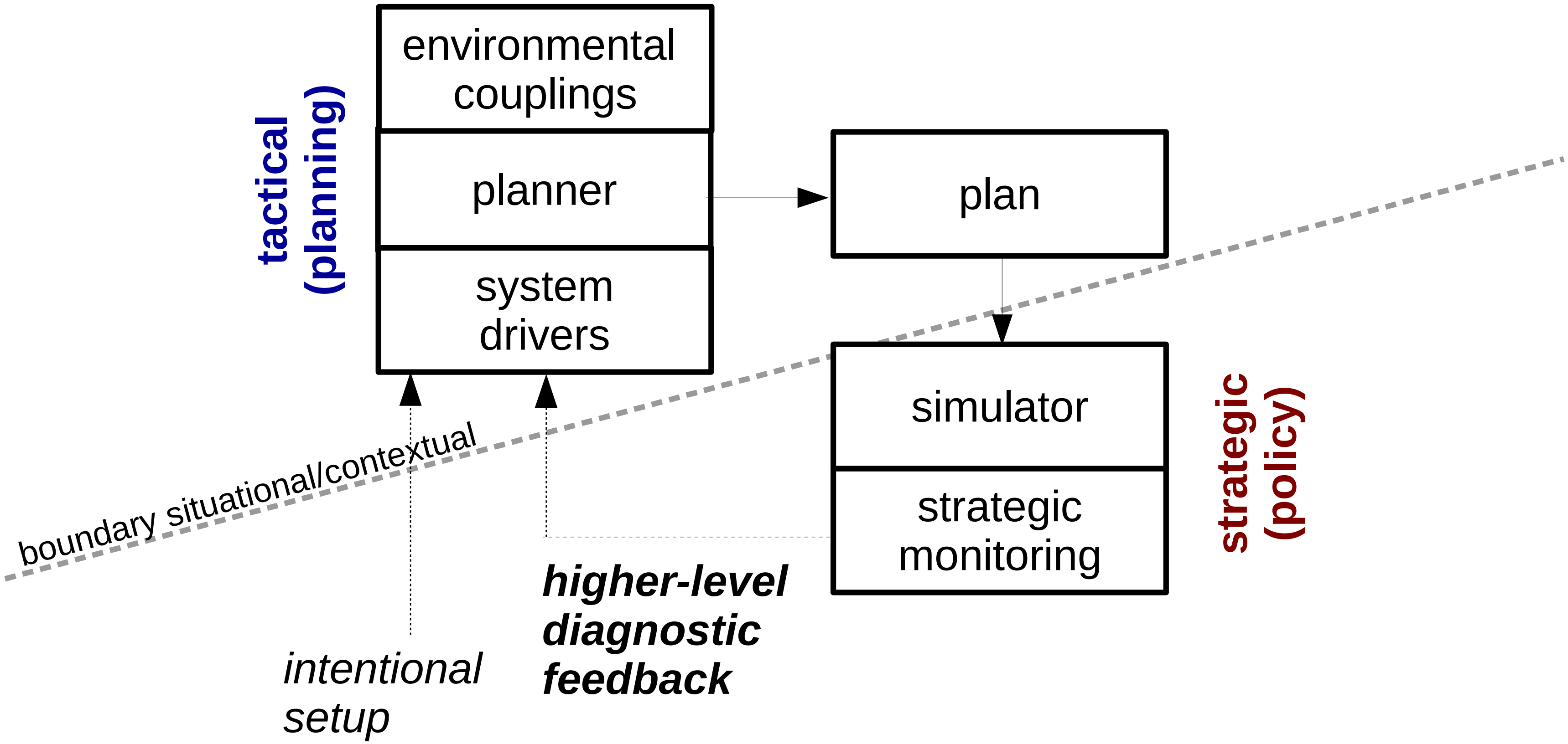
by whom? for whom?

*“**fishing with bombs**”*

➡ *acknowledgement of **undesirable** second-order effects.*

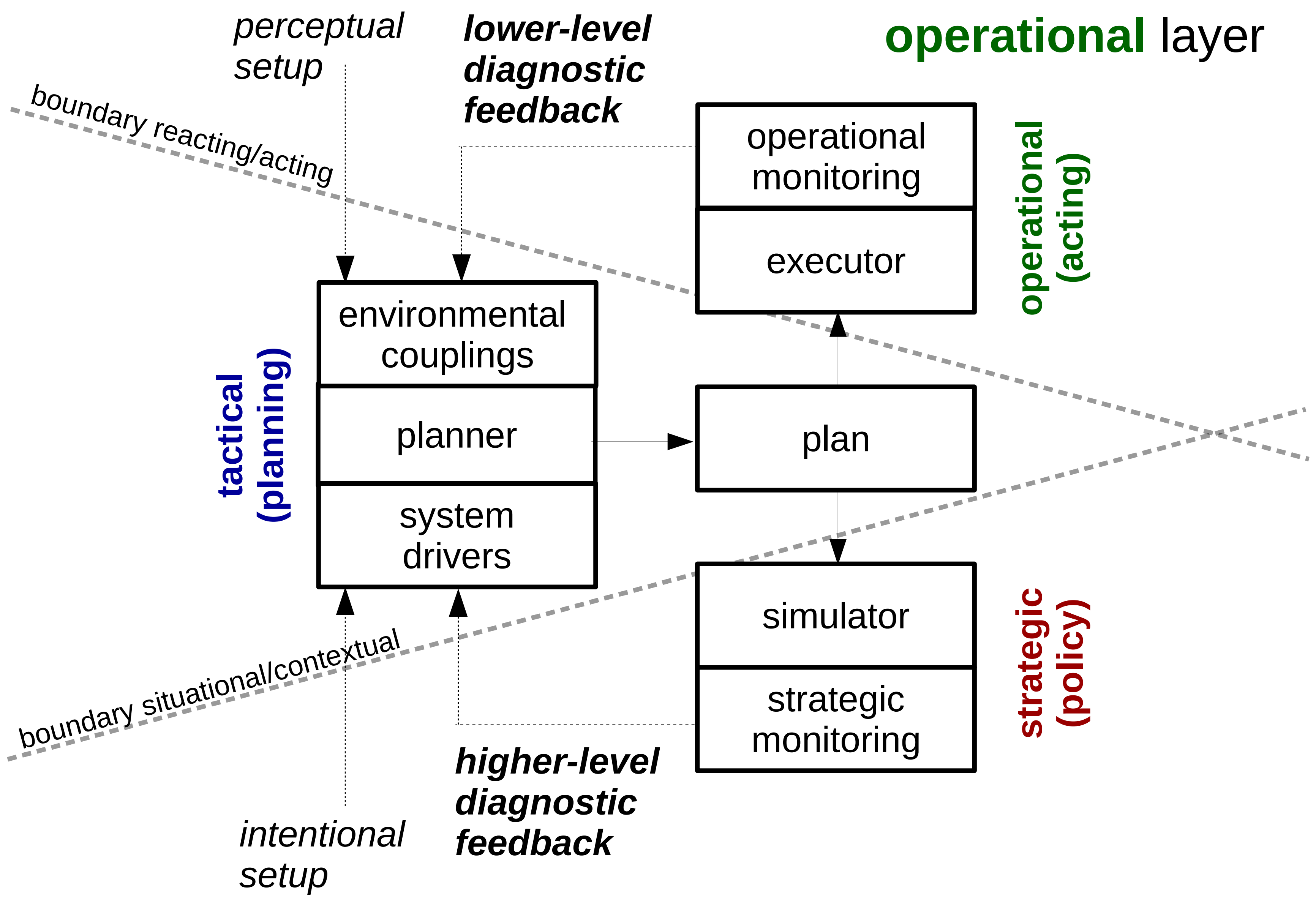
Planning with adaptations

- The process illustrated a *two steps decision-making process,* enabling “tactical” optimization and “strategic” control.

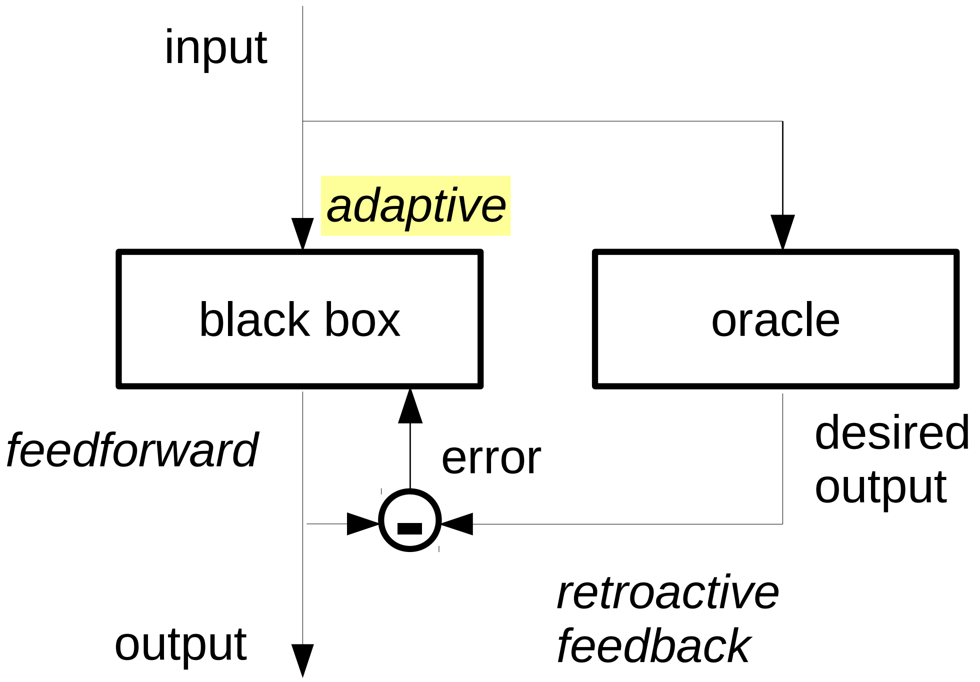


Planning with adaptations

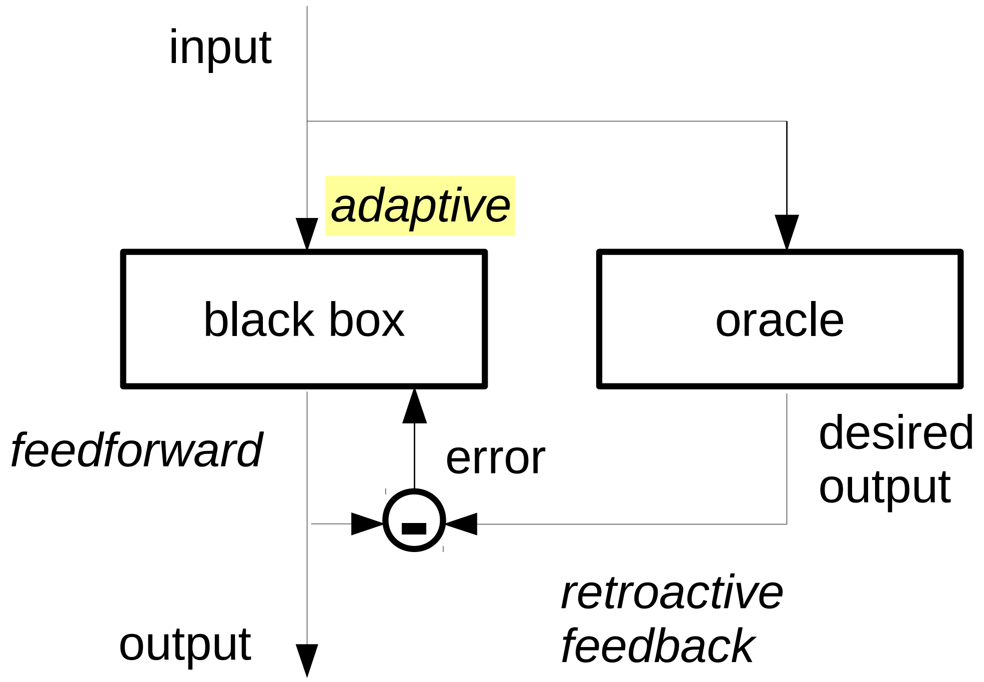
- We might add also the **operational** layer



Supervised Machine Learning

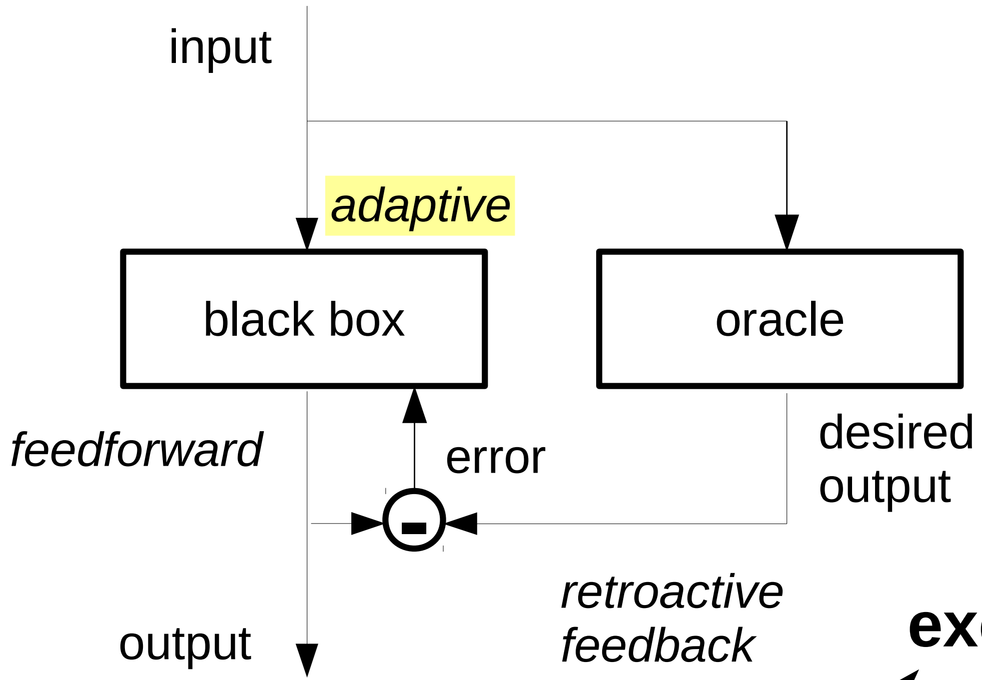


Supervised Machine Learning



- In general, supervised machine learning involves:
  - a data-flow computational network
  - parameters distributed along the network
  - a ML method enabling adaptation of parameters against some feedback, e.g. output error in the training phase
  - an oracle making targets explicit

Supervised Machine Learning



lower-level diagnostic feedback

planner

executor

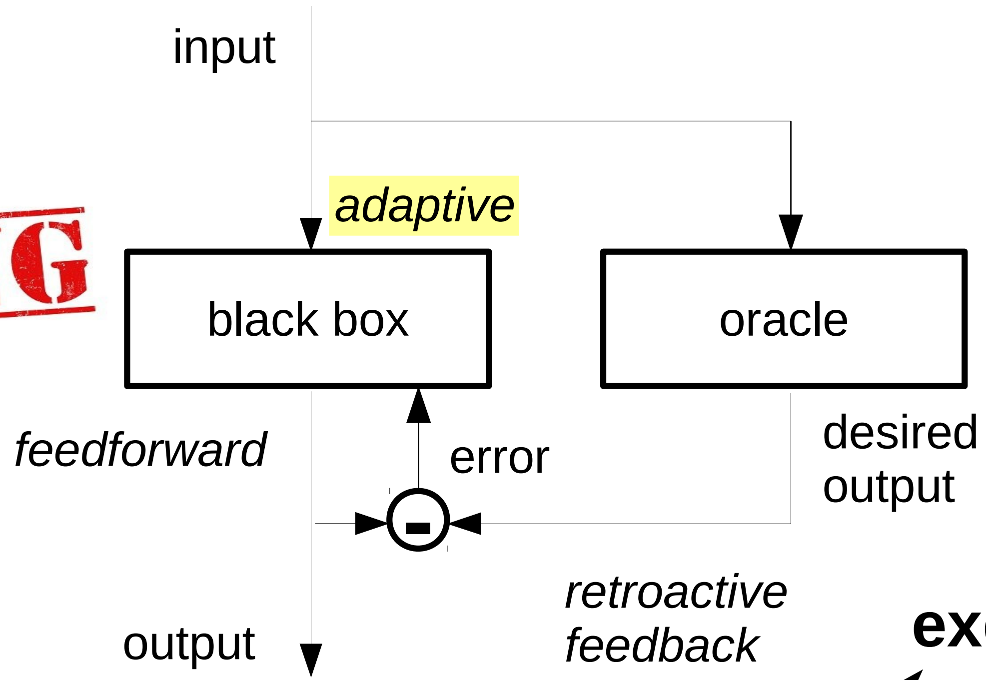
- In general, supervised machine learning involves:
  - a data-flow computational network
  - parameters distributed along the network
  - a ML method enabling adaptation of parameters against some feedback, e.g. output error in the training phase
  - an oracle making targets explicit

plan

intentional setup

Supervised Machine Learning

higher-level diagnostic feedback?



lower-level diagnostic feedback

planner

executor

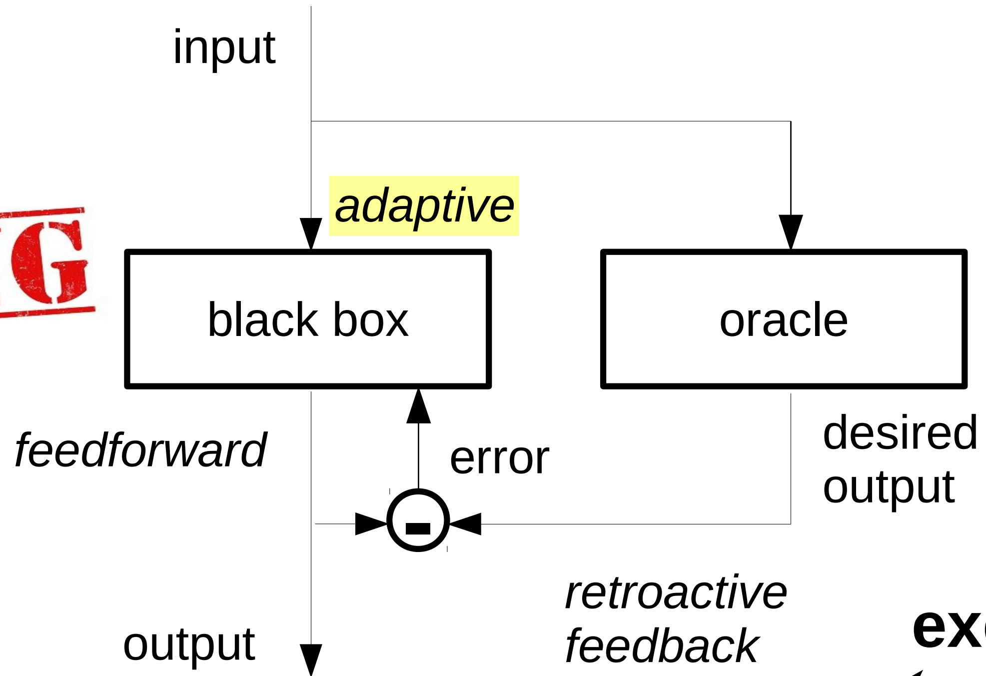
- In general, supervised machine learning involves:
  - a data-flow computational network
  - parameters distributed along the network
  - a ML method enabling adaptation of parameters against some feedback, e.g. output error in the training phase
  - an oracle making targets explicit

plan

intentional setup

Supervised Machine Learning

higher-level diagnostic feedback?



lower-level diagnostic feedback

planner

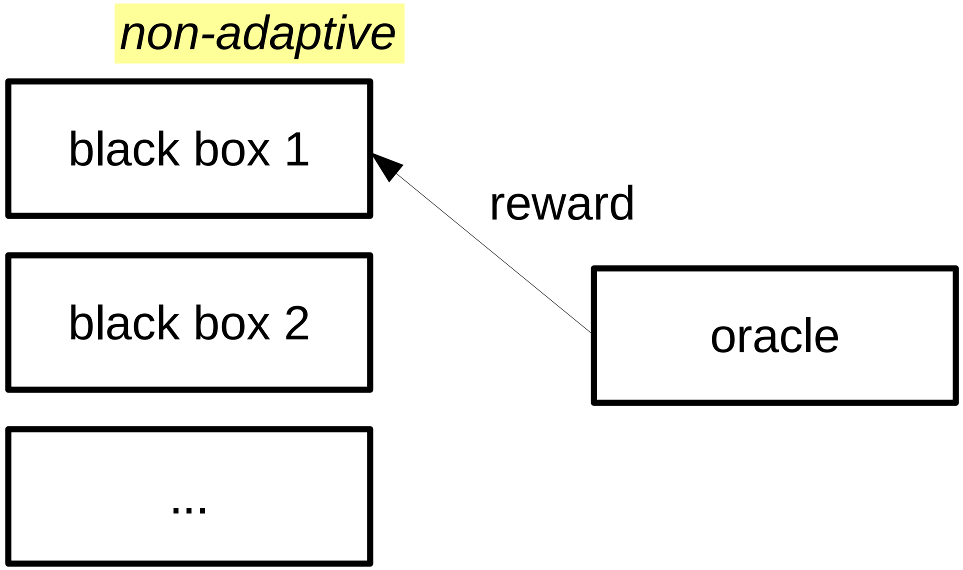
executor

plan

- In general, supervised machine learning involves:
  - a data-flow computational network
  - parameters distributed along the network
  - a ML method enabling adaptation of parameters

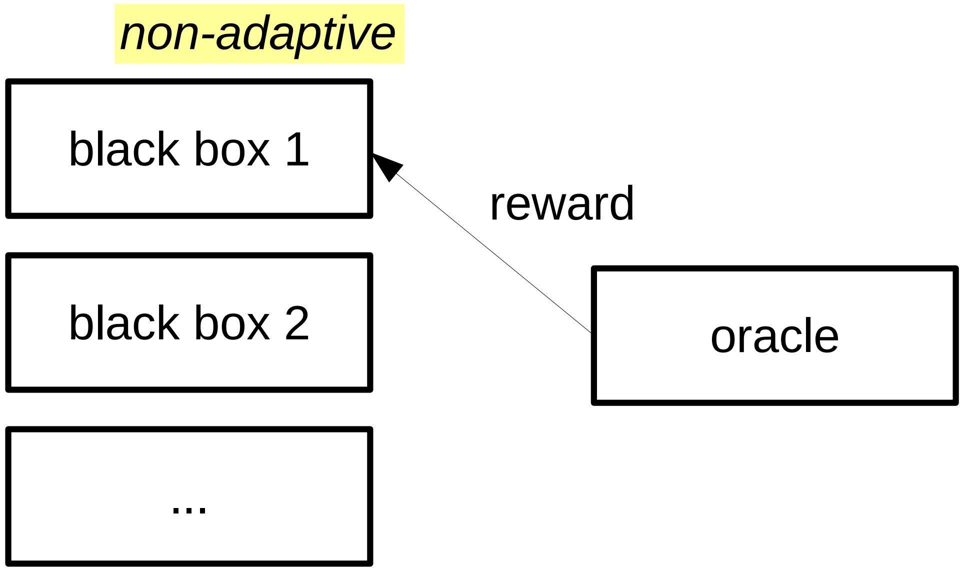
This seems the root of our problems with ML. ***Can we repair it?***

Evolutionary view



- In evolutionary terms, we could consider a multitude of different non-adaptive black-boxes, covering several configurations of parameters, competing for computational resources.
  - For each learning step, the oracle sets the means to select the best performing black-box(es), for which access to computational resources for future predictions will be granted as a ***reward***. [...]
- But who “pays” the oracle?

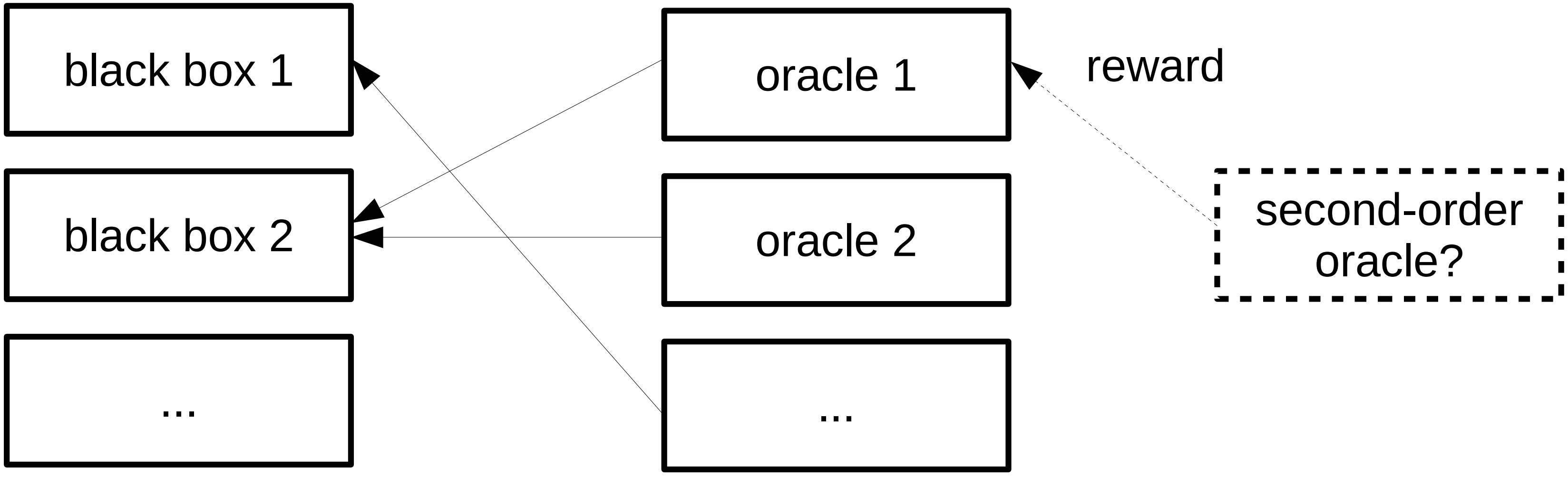
Evolutionary view



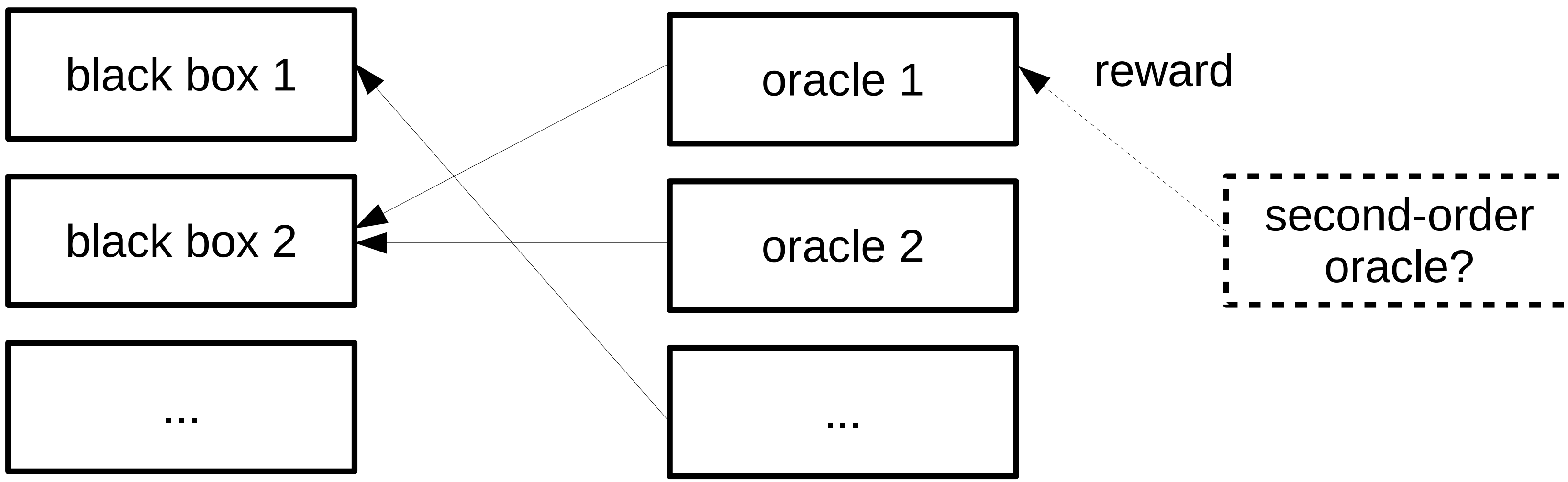
- In evolutionary terms, we could consider a multitude of different non-adaptive black-boxes, covering several configurations of parameters, competing for computational resources.
  - For each learning step, the oracle sets the means to select the best performing black-box(es), for which access to computational resources for future predictions will be granted as a ***reward***. [...]

- *The **higher-level diagnostic feedback** implies that also the system drivers should pass from a **selection mechanism**.*

Evolutionary view

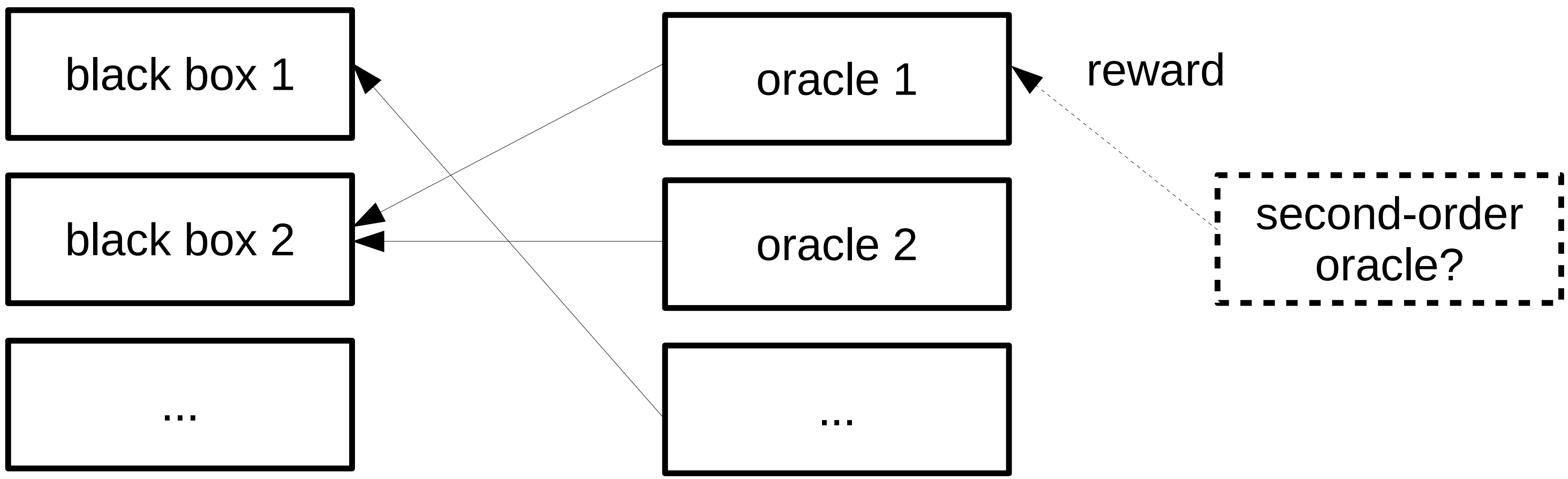


Evolutionary view



- Let’s use this architecture on a concrete example: IBM Watson (building upon a network of intelligent QA agents).
  - a question is given
  - the system has to guess
    - what the question demands (~ **oracles**)
    - what is the answer (~ **black-box**),
  - correct response is given by the jury (~ **second-order oracle**)

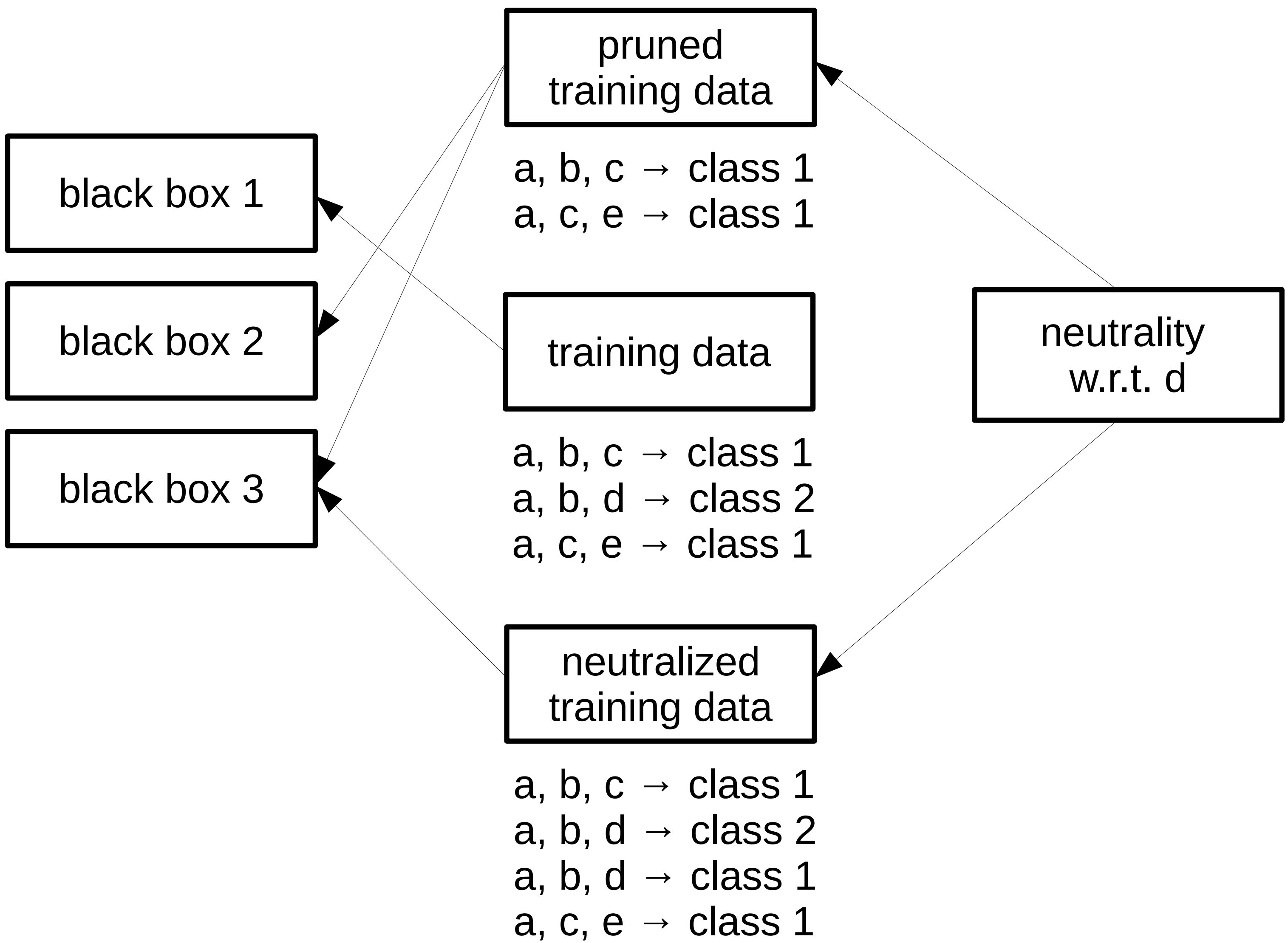
Evolutionary view



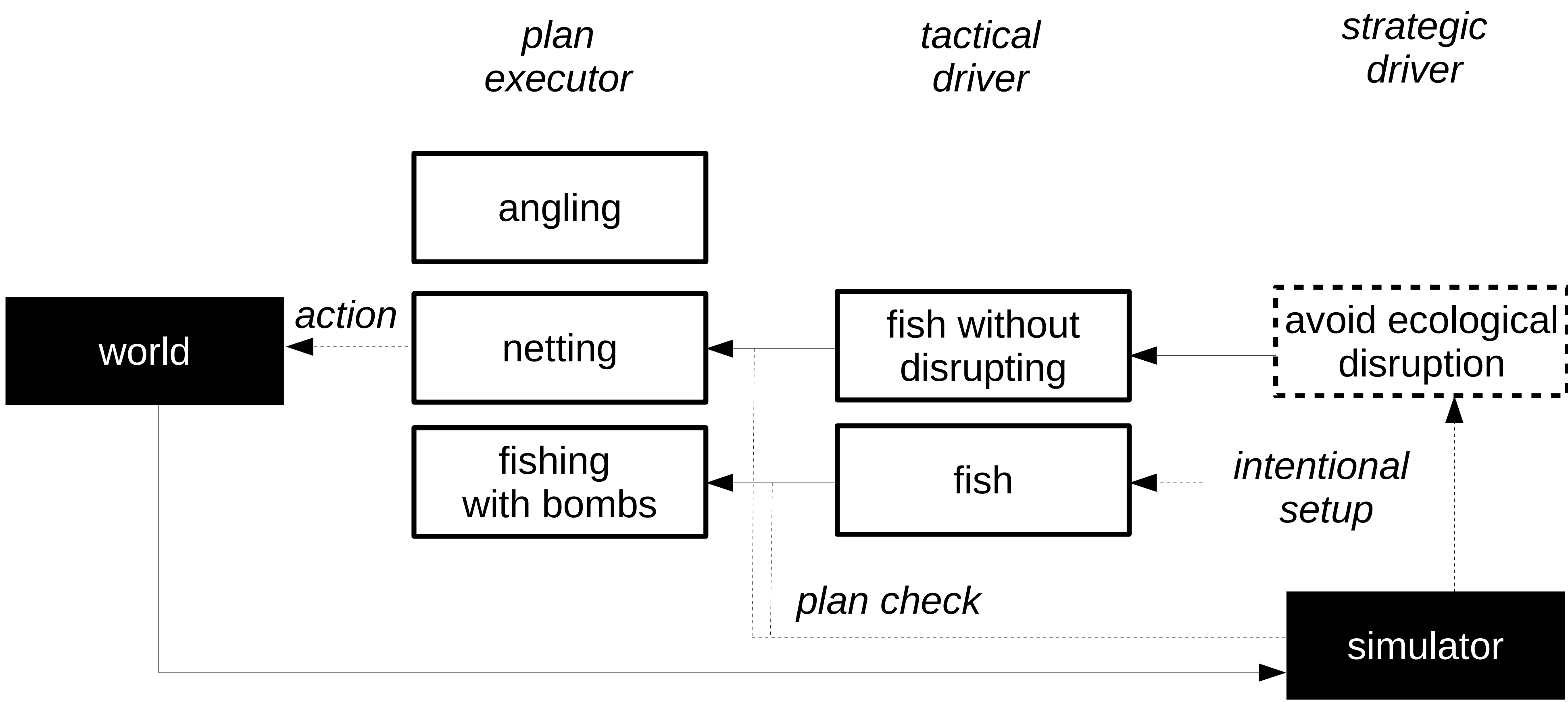
- Let’s use this architecture on a concrete example: IBM Watson (building upon a network of intelligent QA agents).
  - a question is given
  - the system has to guess
    - what the question demands (~ **oracles**)
    - what is the answer (~ **black-box**).

Let’s apply it to our initial problems!

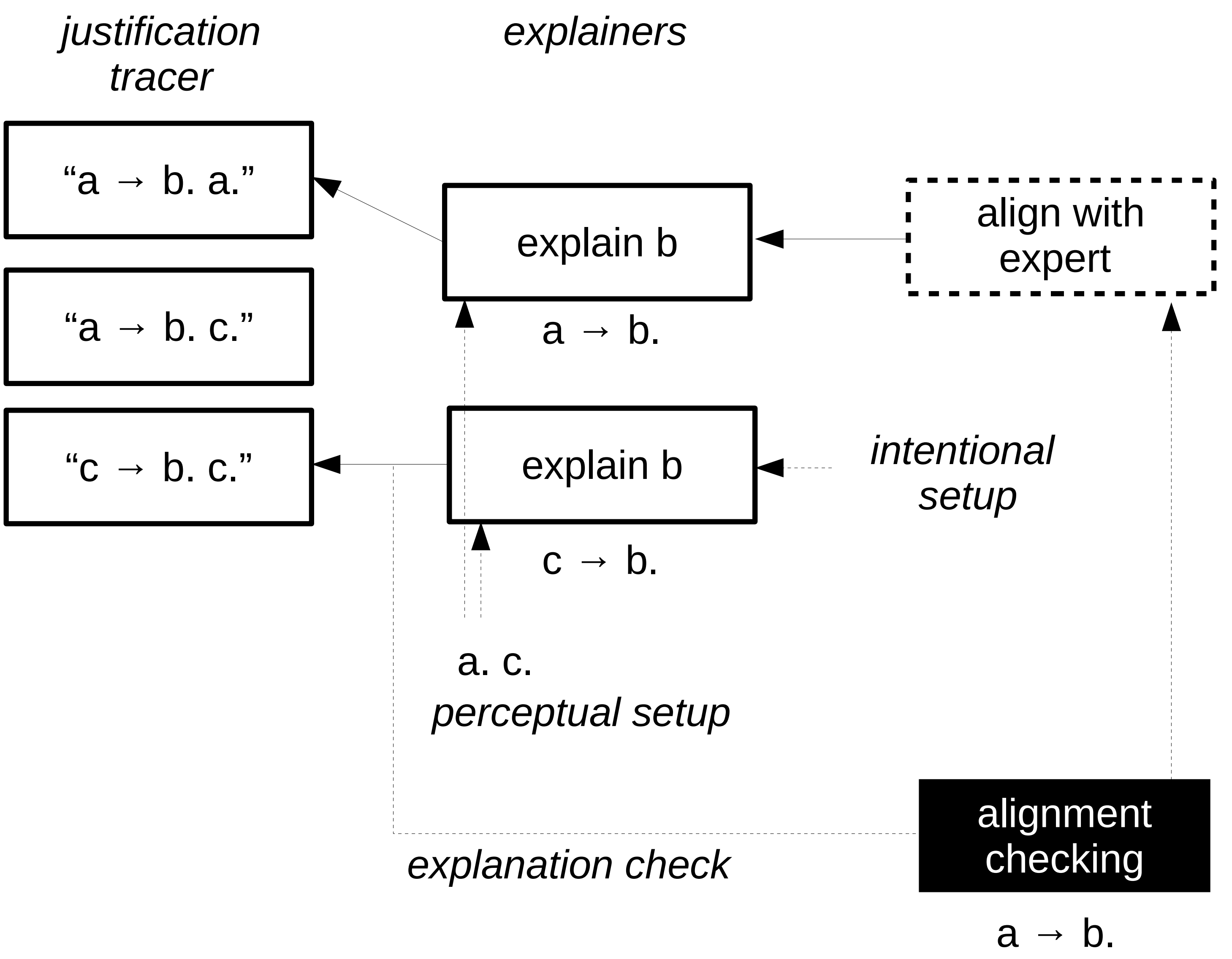
Example: neutrality constraint



Example: strategic protection to unintended consequences



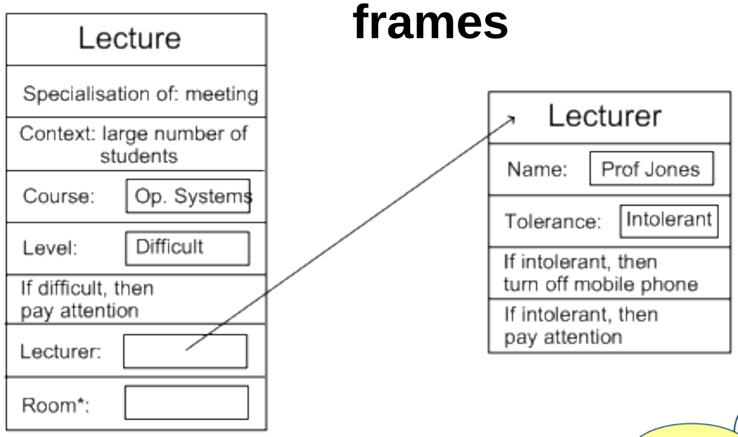
Example: alignment to expert knowledge for explanation



What normware consists of

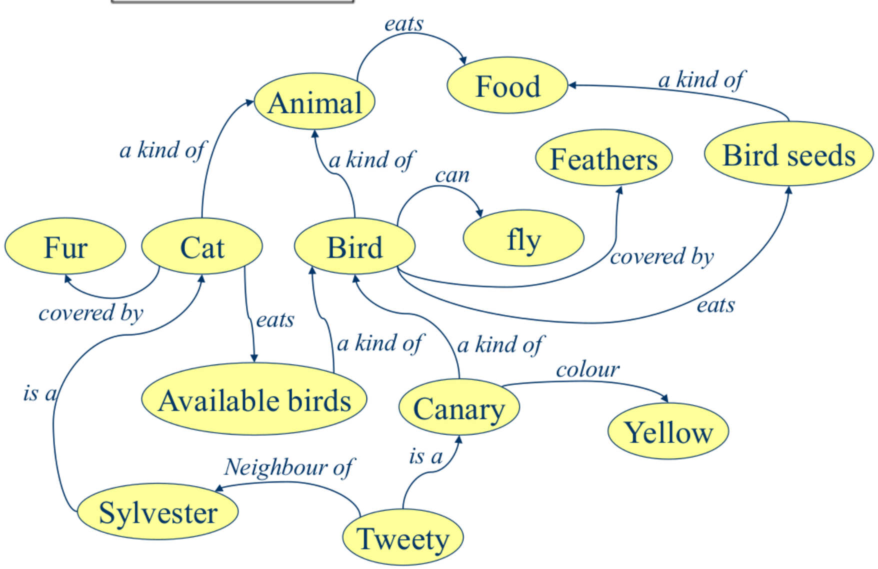
- It has to be symbolic
  - contains knowledge: *epistemic commitments*

frames




```
if flower and seed then phanerogam
if phanerogam and bare-seed then fir
if phanerogam and 1-cotyledon then
monocotyledonous
if phanerogam and 2-cotyledon then
dicotyledonous
if monocotyledon and rhizome then thrush
if dicotyledon then anemone
if monocotyledon and ¬rhizome then lilac
if leaf and flower then cryptogamous
if cryptogamous and ¬root then foam
if cryptogamous and root then fern
if ¬leaf and plant then thallophyte
if thallophyte and chlorophyll then algae
if thallophyte and ¬ chlorophyll then fungus
if ¬leaf and ¬flower and ¬plant then colibacille
```

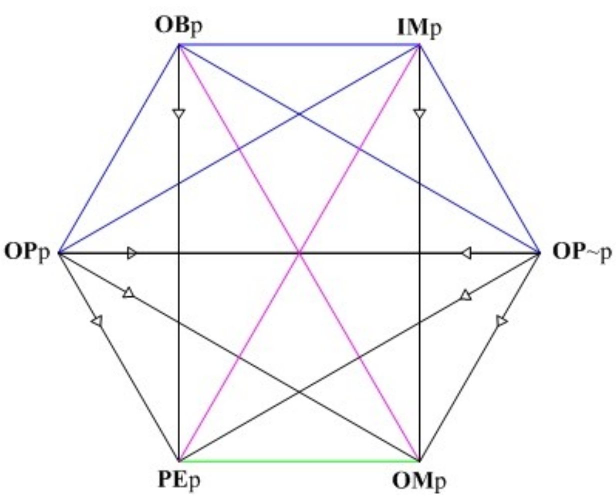
expert systems



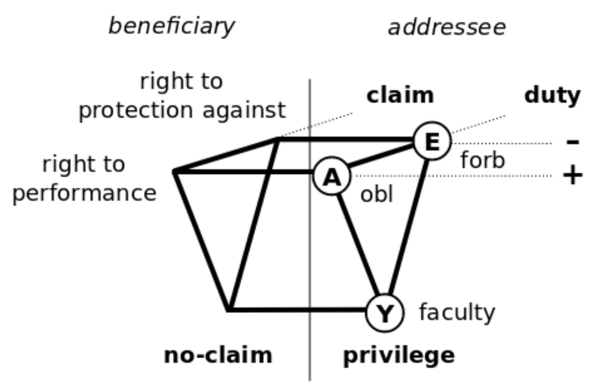
semantic networks

What normware consists of

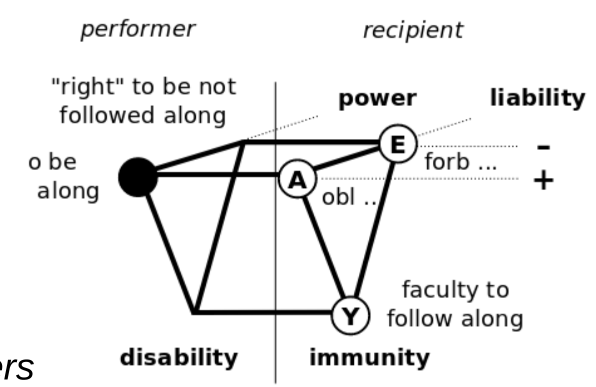
- It has to be symbolic
  - contains knowledge: *epistemic commitments*
  - contains drivers: *behavioural commitments*



deontic logics
(permission, obligation and prohibition)

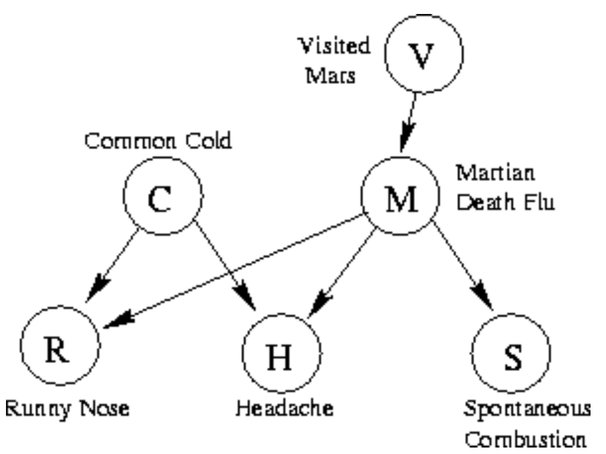


hohfeldian prisms
types of obligations and powers

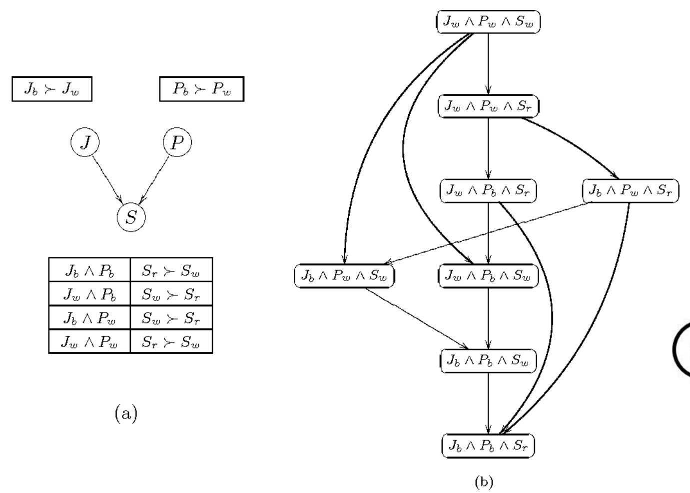


What normware consists of

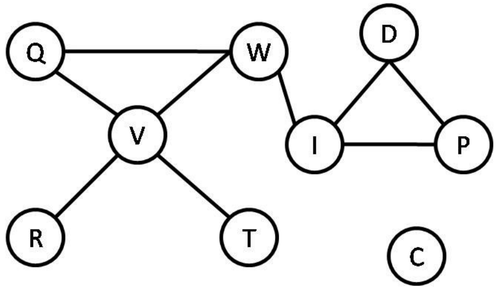
- It has to be symbolic
  - contains knowledge: *epistemic commitments*
  - contains drivers: *behavioural commitments*
  - could transport some strength (partial ordering or degree)



Bayesian networks



CP-nets (*Ceteri Paribus*)



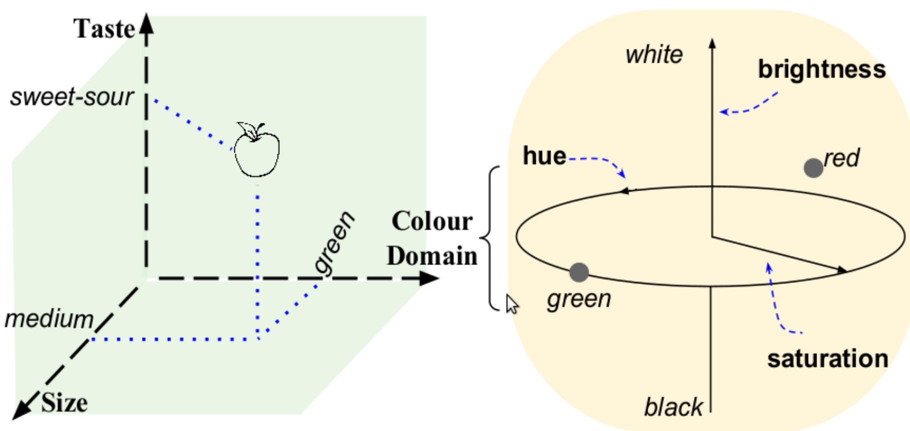
GAI networks

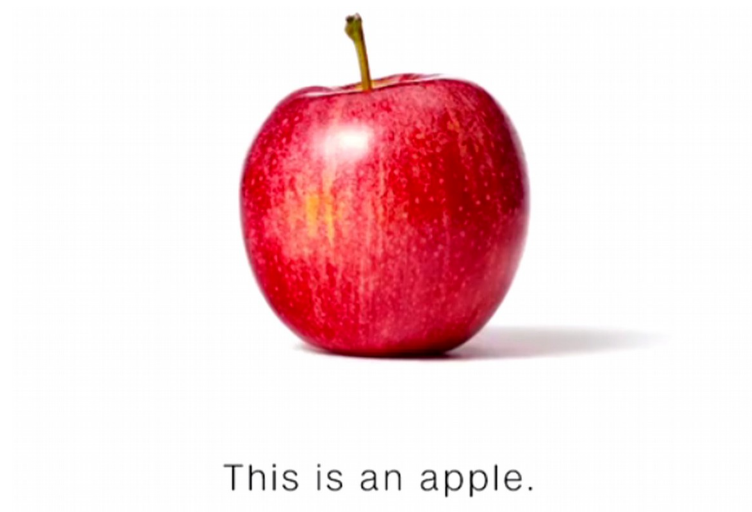
What normware consists of

- It has to be symbolic
  - contains knowledge: *epistemic commitments*
  - contains drivers: *behavioural commitments*
  - could transport some strength (partial ordering or degree)

- It has to reinforce some sort of “conceptual spaces” (for knowledge) and some sort of “action spaces” (for drivers)

What normware consists of





- It has to reinforce some sort of “conceptual spaces” (for knowledge) and some sort of “action spaces” (for drivers)
  - *graduality*
  - **prototyping** (to detect abnormalities)
  - *analogy*
  - solving the ***symbol grounding problem***

What normware consists of

- It has to be symbolic
  - contains knowledge: *epistemic commitments*
  - contains drivers: *behavioural commitments*
  - could transport some strength (partial ordering or degree)

- It has to reinforce some sort of “conceptual spaces” (for knowledge) and some sort of “action spaces” (for drivers)
  - *graduality*
  - **prototyping** (to detect abnormalities)
  - *analogy*
  - solving the ***symbol grounding problem***

Metaphorically

Machine learning

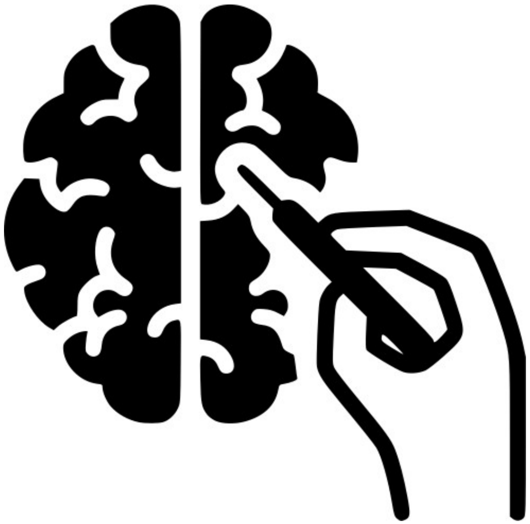
internalizing
desired
behaviour

Metaphorically

Machine learning

internalizing
desired
behaviour

Software development



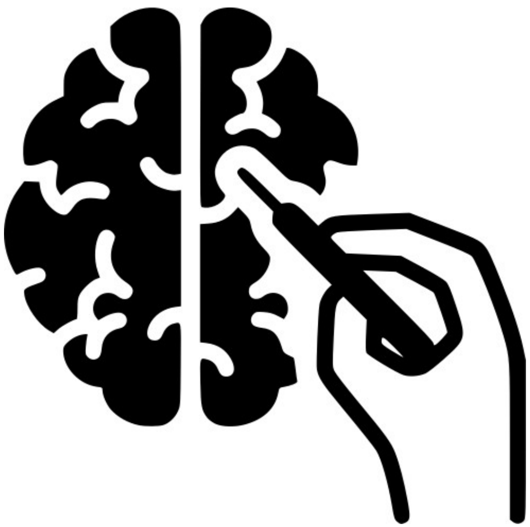
hacking
the brain

Metaphorically

Machine learning

internalizing
desired
behaviour

Software development



hacking
the brain

Normware-based computing

providing guidelines, interacting with experiences

Perspectives

- This presentation highlighted the crucial role of **normware** with respect to trustworthy and explainable AI (→ computing)
  - ML approaches usually do not consider this level of abstraction
  - ethical/responsible AI studies target higher level constraints

Perspectives

- This presentation highlighted the crucial role of **normware** with respect to trustworthy and explainable AI (→ computing)
  - ML approaches usually do not consider this level of abstraction
  - ethical/responsible AI studies target higher level constraints

- It makes clear two perspectives on normware:
  - computational artifacts specifying norms
  - ecology of components guiding the system components

Perspectives

- This presentation highlighted the crucial role of **normware** with respect to trustworthy and explainable AI (→ computing)
  - ML approaches usually do not consider this level of abstraction
  - ethical/responsible AI studies target higher level constraints

- It makes clear two perspectives on normware:
  - computational artifacts specifying norms
  - ecology of components guiding the system components

including sub-symbolic ones!

Perspectives

- This presentation highlighted the crucial role of **normware** with respect to trustworthy and explainable AI (→ computing)
  - ML approaches usually do not consider this level of abstraction
  - ethical/responsible AI studies target higher level constraints

- It makes clear two perspectives on normware:
  - computational artifacts specifying norms

    including sub-symbolic ones!
  - ecology of components guiding the system components

- *Focus on **incentive structures***

Perspectives

- This presentation highlighted the crucial role of **normware** with respect to trustworthy and explainable AI (→ computing)
  - ML approaches usually do not consider this level of abstraction
  - ethical/responsible AI studies target higher level constraints

- It makes clear two perspectives on normware:
  - computational artifacts specifying norms

    including sub-symbolic ones!
  - ecology of components guiding the system components

- *Focus on **incentive structures***

- The ecological perspective overlooked so far, but reminds of visionary ideas presented in the history of AI (Minsky's ***society of minds***, Brooks' ***intelligent creatures***).

A less tentative taxonomy

hardware

physical device

when running → physical mechanism

situated in a physical environment

control structure

software

symbolic device

when running → symbolic mechanism

relies on physical mechanisms

control structure

?

normware

coordination device

when ***adopted*** → interactional mechanism

relies on symbolic mechanisms

guidance structure

Questions?

Acknowledgements:
This work was supported by NWO in the DL4LD project and VWDATA in the NWA Start-Impuls program.