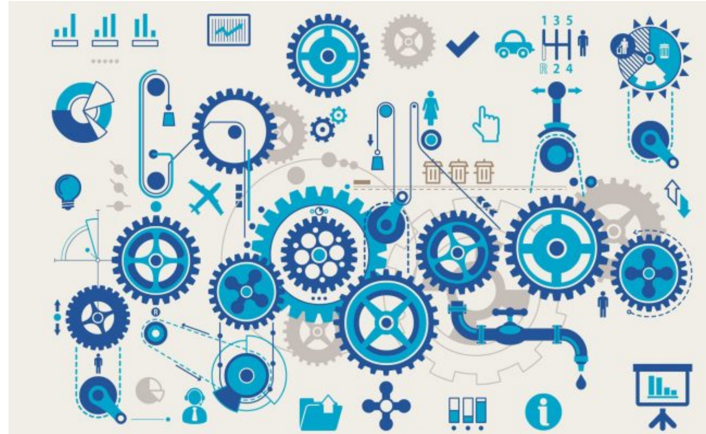
The Role of **Normware** in Trustworthy and Explainable AI

Giovanni Sileno (g.sileno@uva.nl),
Alexander Boer, Tom van Engers

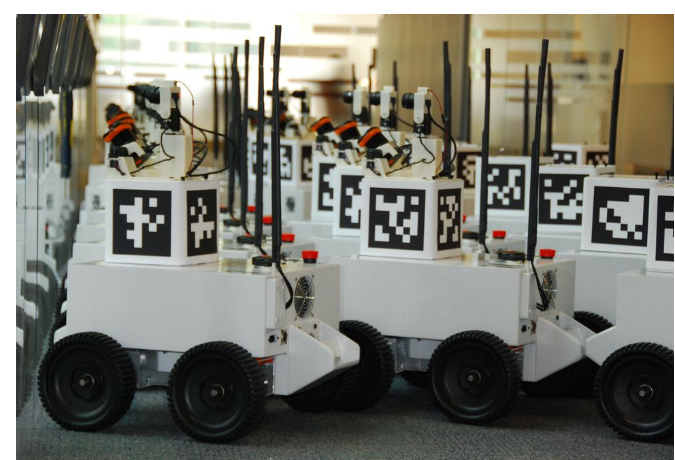
XAILA, eXplainable AI and Law workshop, JURIX 2018 @ Groningen

12 December 2018

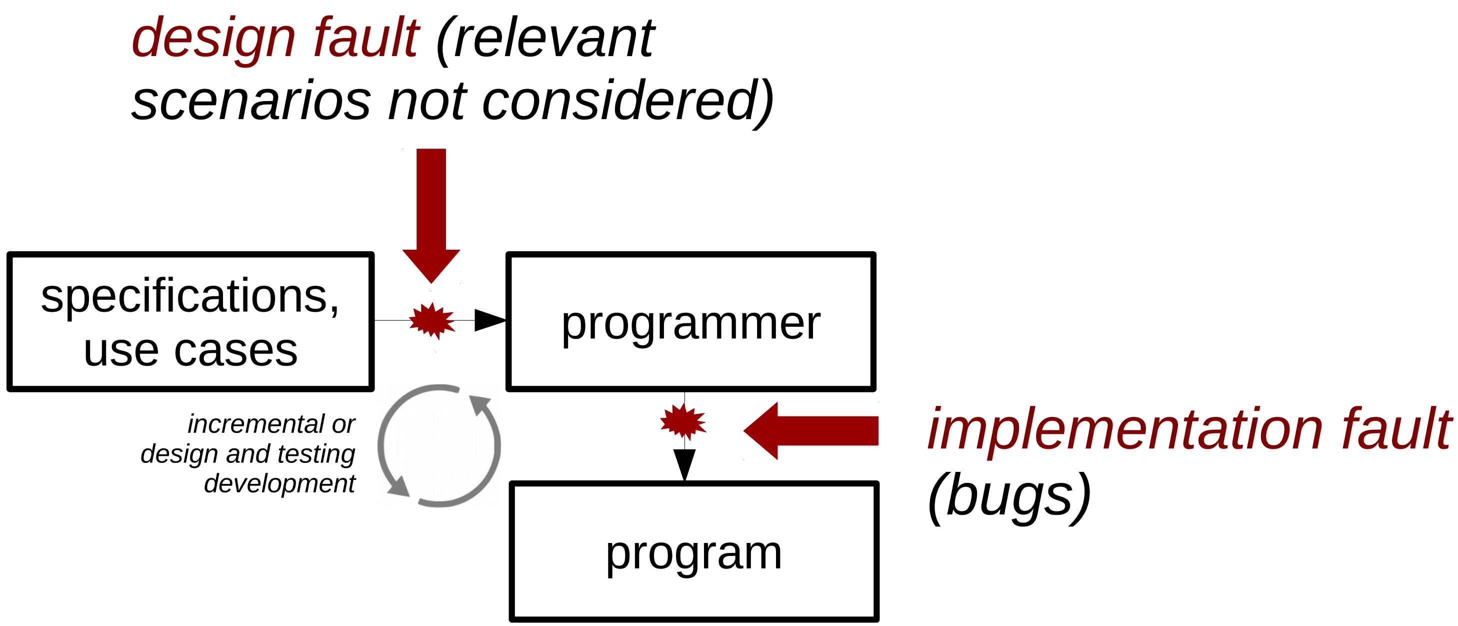

with the (supposedly) near advent of ***autonomous artificial entities***, or other forms of ***distributed automatic decision making***,

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

Unintended consequences: bad or limited design

Unintended consequences: bad or limited design



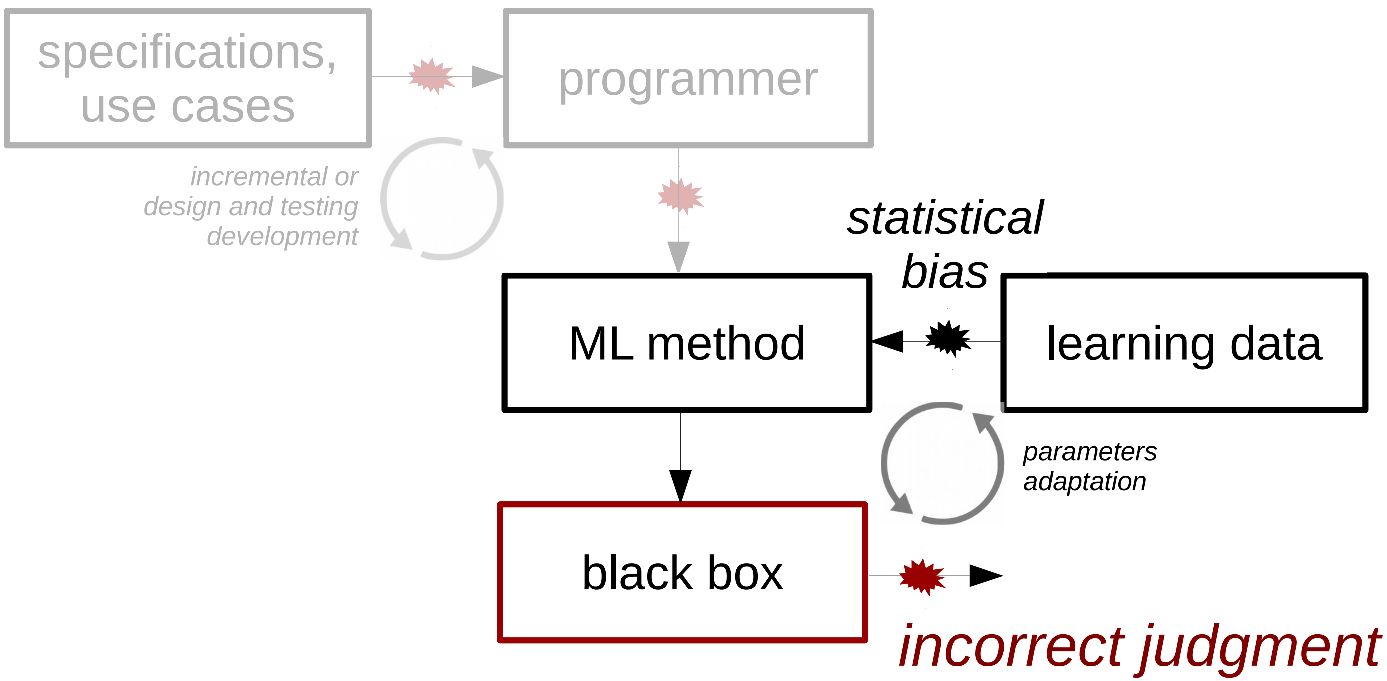
Unintended consequences: bad or limited design

- 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*

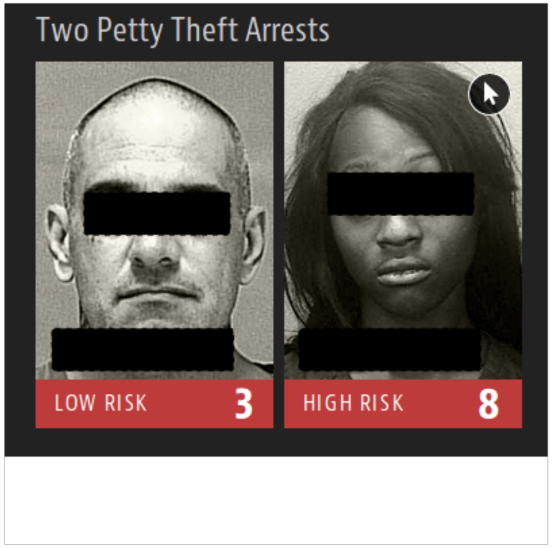
Unintended consequences: the “artificial prejudice”

Unintended consequences: the “artificial prejudice”



Unintended consequences: the "artificial prejudice"

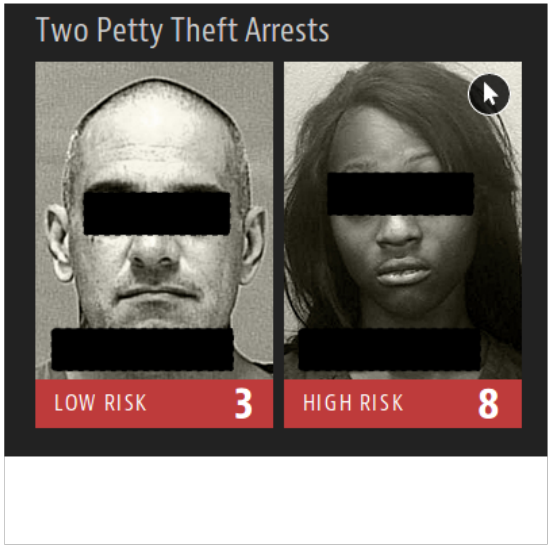
- Software used across the US predicting future crimes and criminals biased against African Americans (2016)



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

Unintended consequences: the “artificial prejudice”

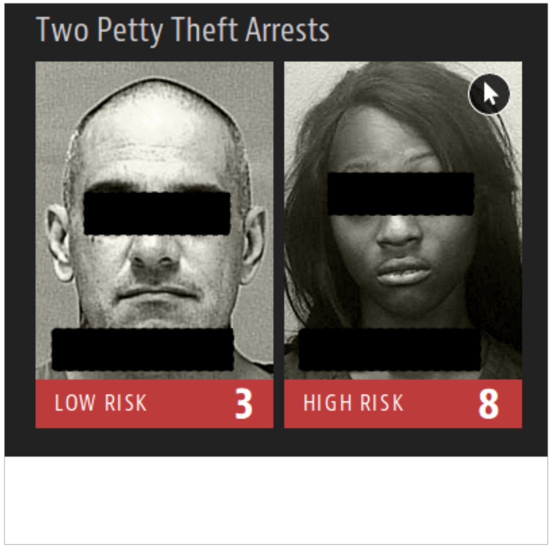
- Software used across the US predicting future crimes and criminals biased against African Americans (2016)
  - Existing statistical bias (correct **description**)
  - When used for prediction on an individual it is read as ***behavioural predisposition***, i.e. it is interpreted as a **mechanism**.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

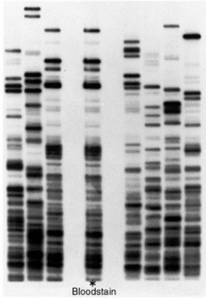


Angwin J. et al. ProPublica, May 23 (2016). *Machine Bias: risk assessments in criminal sentencing*

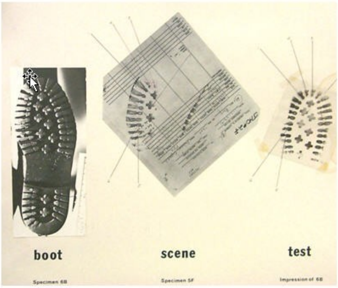
Unintended consequences: the “artificial prejudice”

- 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

...

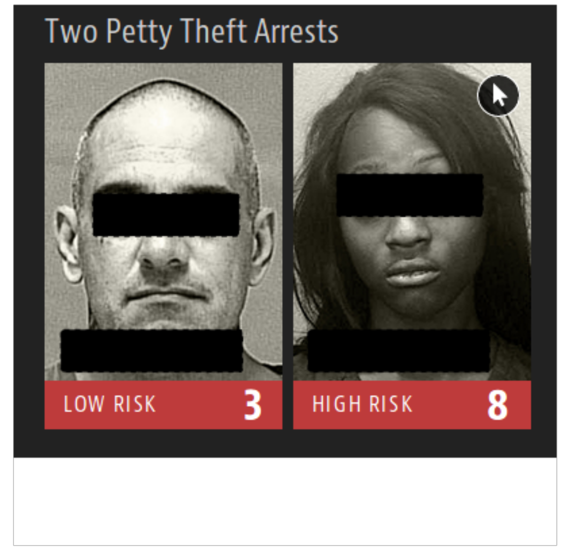
origin, gender, ethnicity, wealth, ...

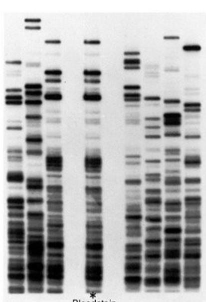
improper profiling?

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

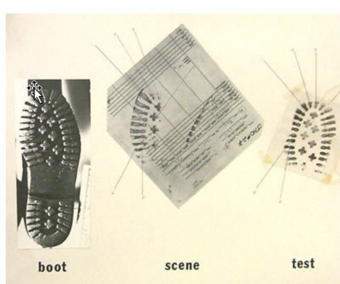
Unintended consequences: the "artificial prejudice"

- 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

...

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

origin, gender, ethnicity, wealth, ...

improper profiling?

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

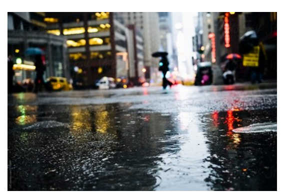
Unacceptable conclusions: improvident induction

- 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

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

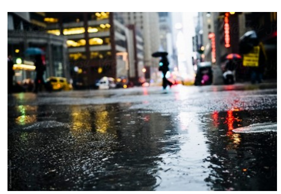
- Consider a diagnostic application predicting whether the patient has ***appendicitis***:
  - We would accept a conclusion based on the presence of fever, abdominal pain, or an increased number of white blood cells, but not if based e.g. on the length of the little toe or the fact that outside it is raining!



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.

- Consider a diagnostic application predicting whether the patient has ***appendicitis***:
  - We would accept a conclusion based on the presence of fever, abdominal pain, or an increased number of white blood cells, but not if based e.g. on the length of the little toe or the fact that outside it is raining!



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

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.

- Consider a diagnostic application predicting whether the patient has ***appendicitis***:
  - We would accept a conclusion based on the presence of fever, abdominal pain, or an increased number of white blood cells, but not if based e.g. on the length of the little toe or the fact that outside it is raining!

for that decision-making context

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

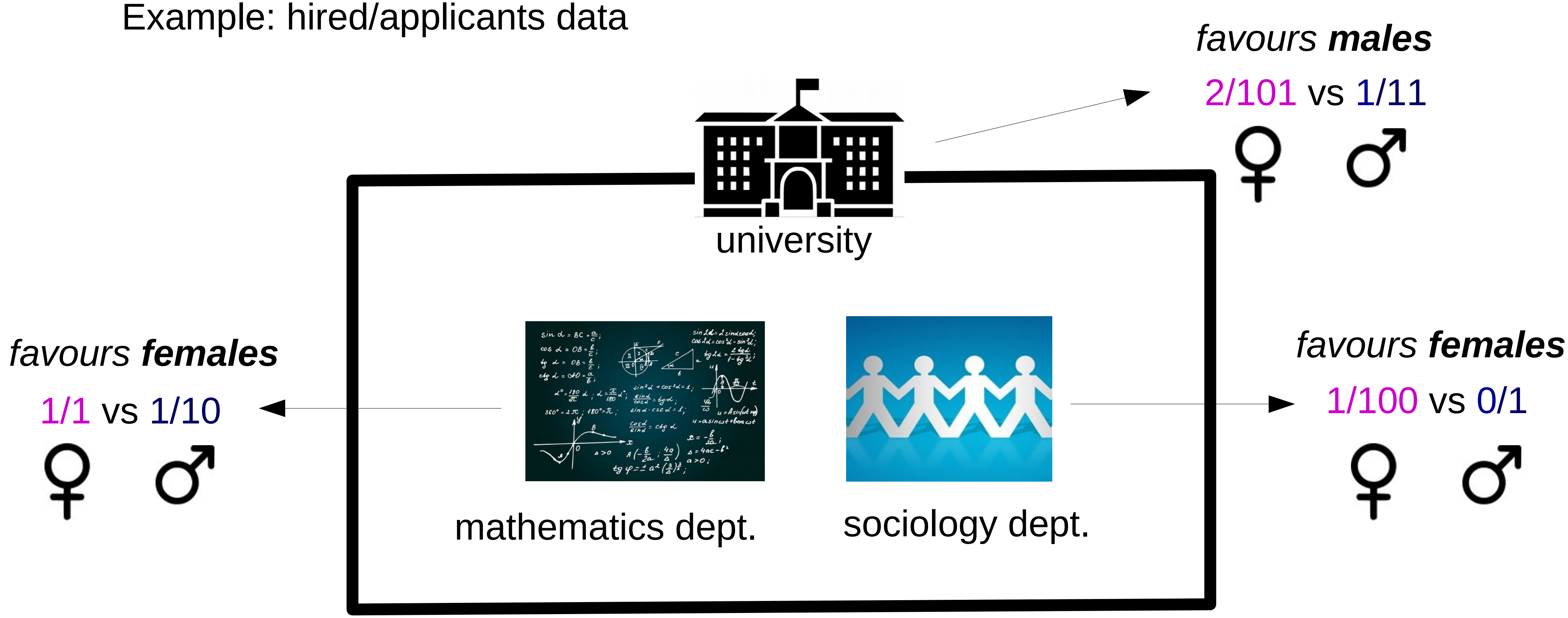
Unacceptable conclusions: improvident induction

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

Unacceptable conclusions: improvident induction

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

Example: hired/applicants data



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?
- claim: **normware**!

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

A tentative taxonomy

		?
hardware	software	normware
physical device	symbolic device	……….
when running → physical mechanism	when running → symbolic mechanism	……….
situated in a physical environment	relies on physical mechanisms	relies on symbolic mechanisms
control structure	**control structure**	……….

A tentative taxonomy

?

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

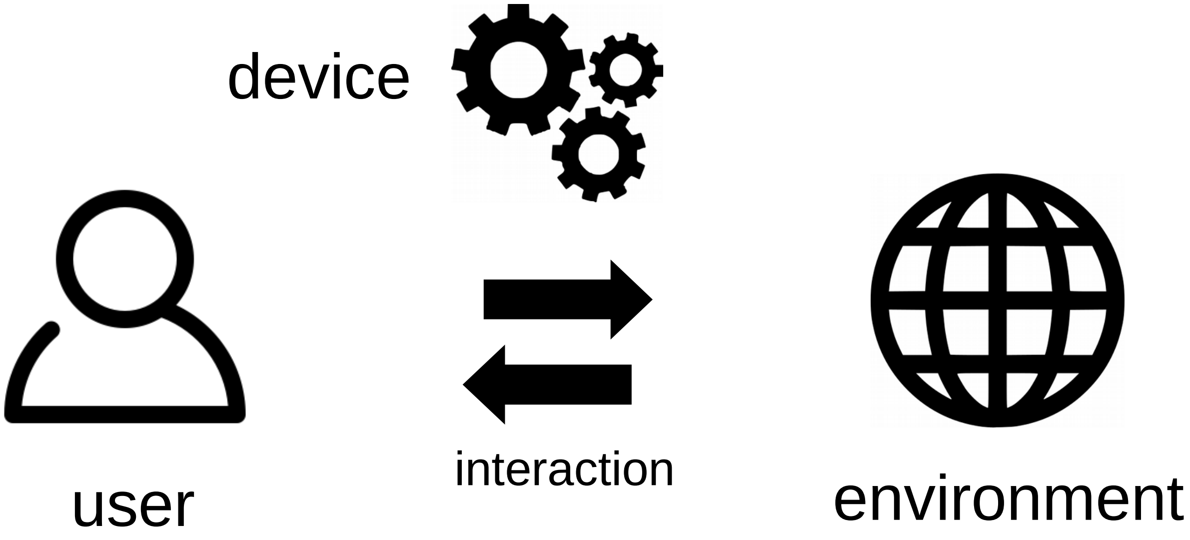
normative and epistemic pluralism?

interaction with sub-symbolic modules?

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

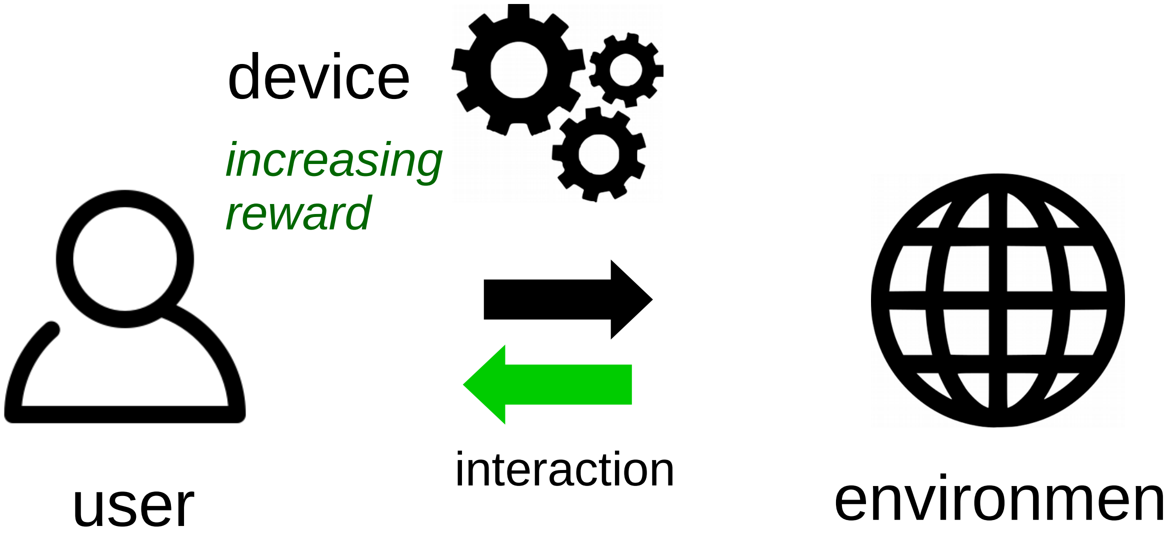
Impact *at large*

- 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:

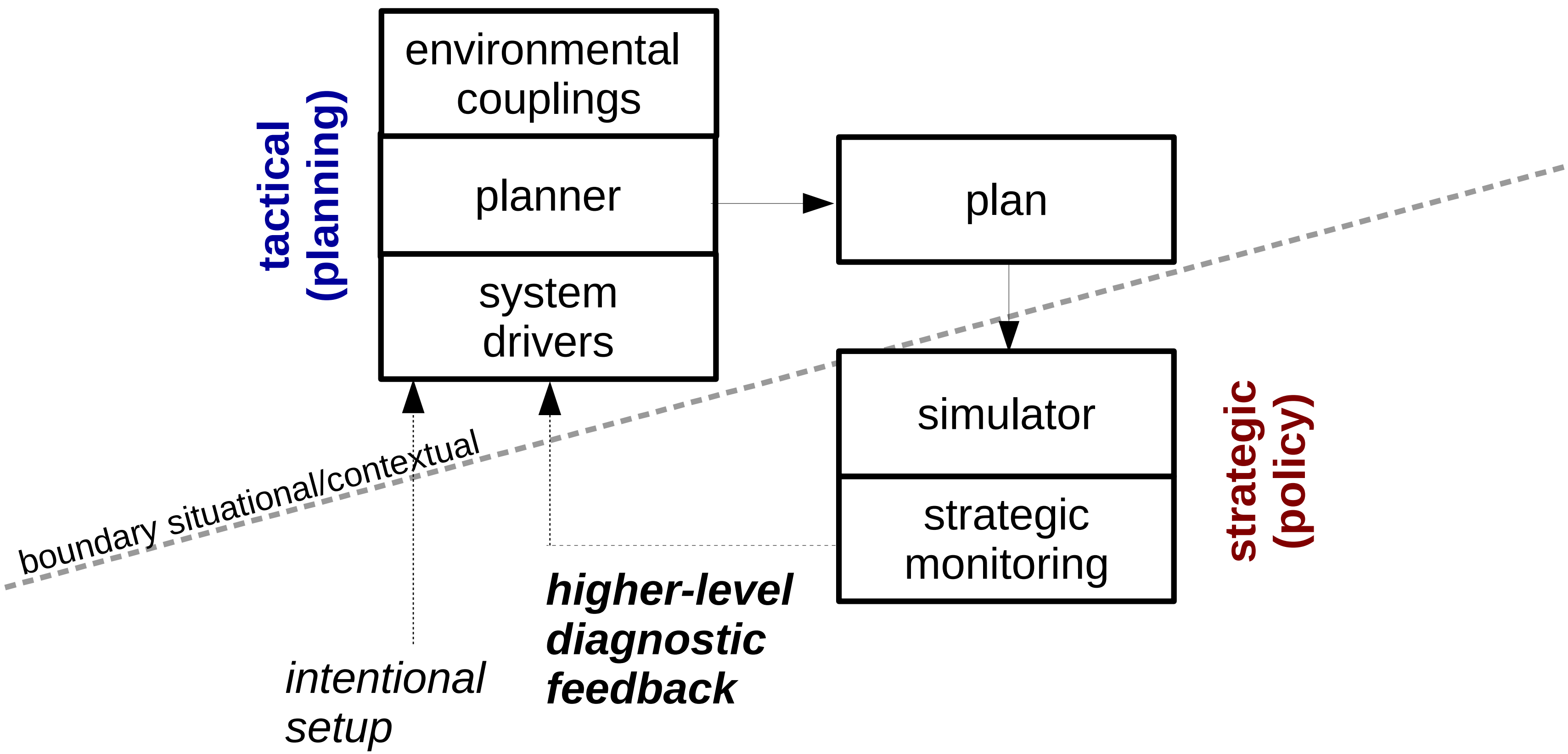
*“**fishing with bombs**”*

by whom? for whom?

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

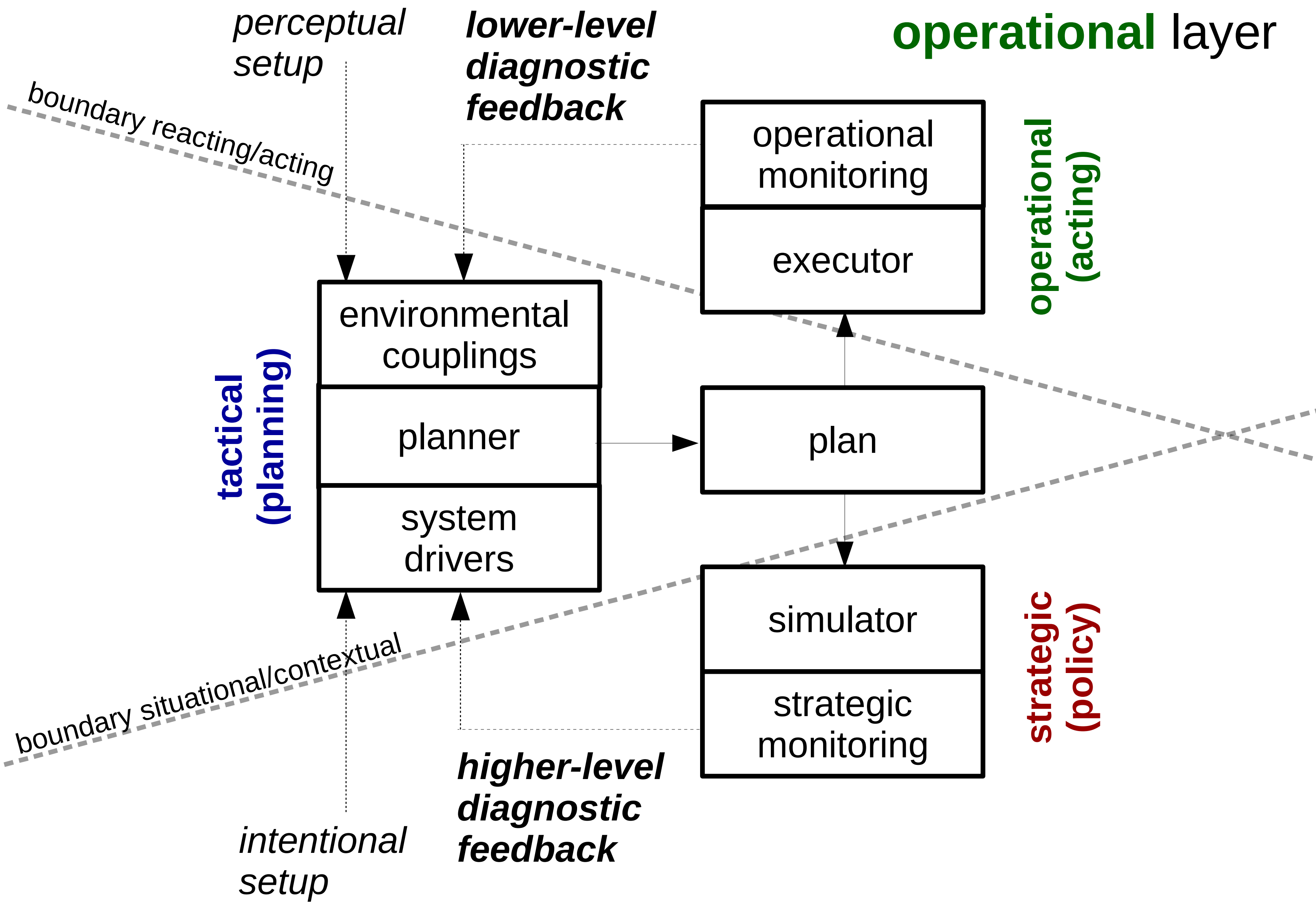
Planning with adaptations

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

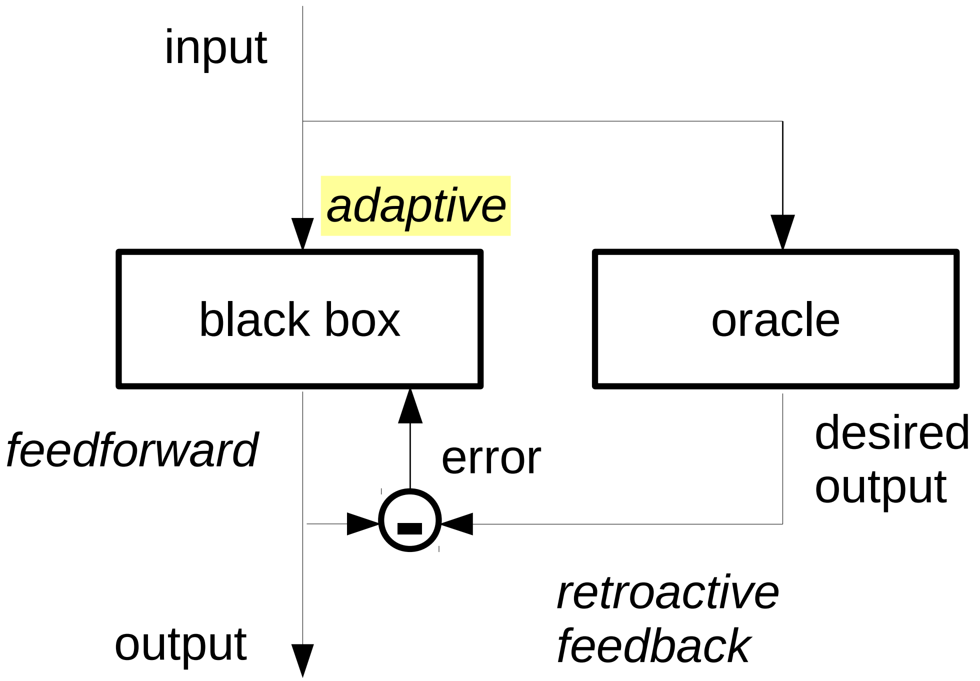


Planning with adaptations

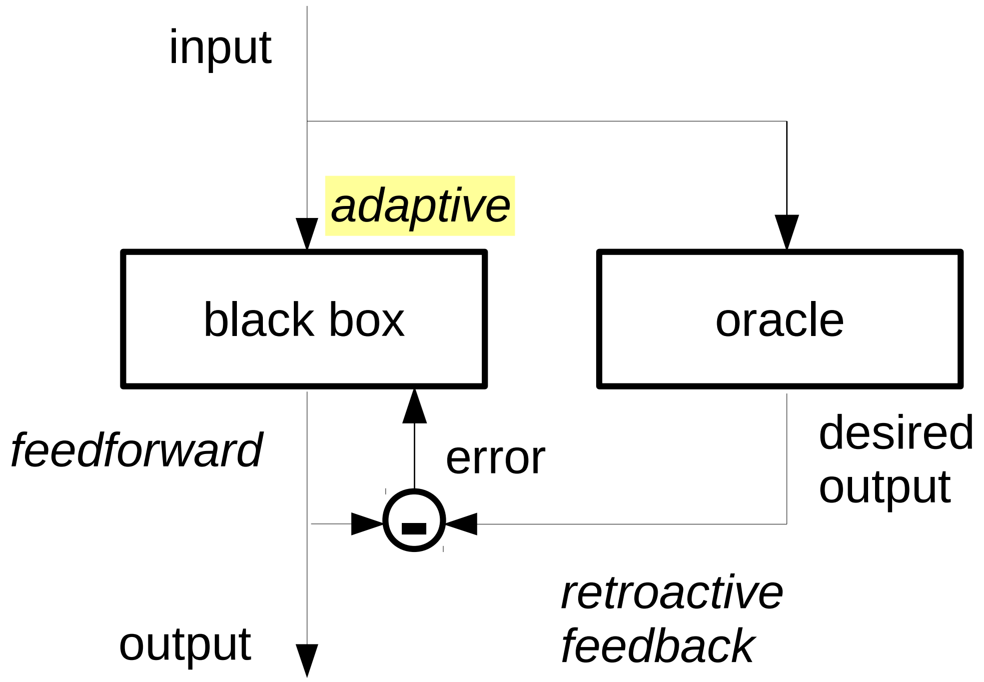
- We might add also the **operational** layer



Supervised Machine Learning

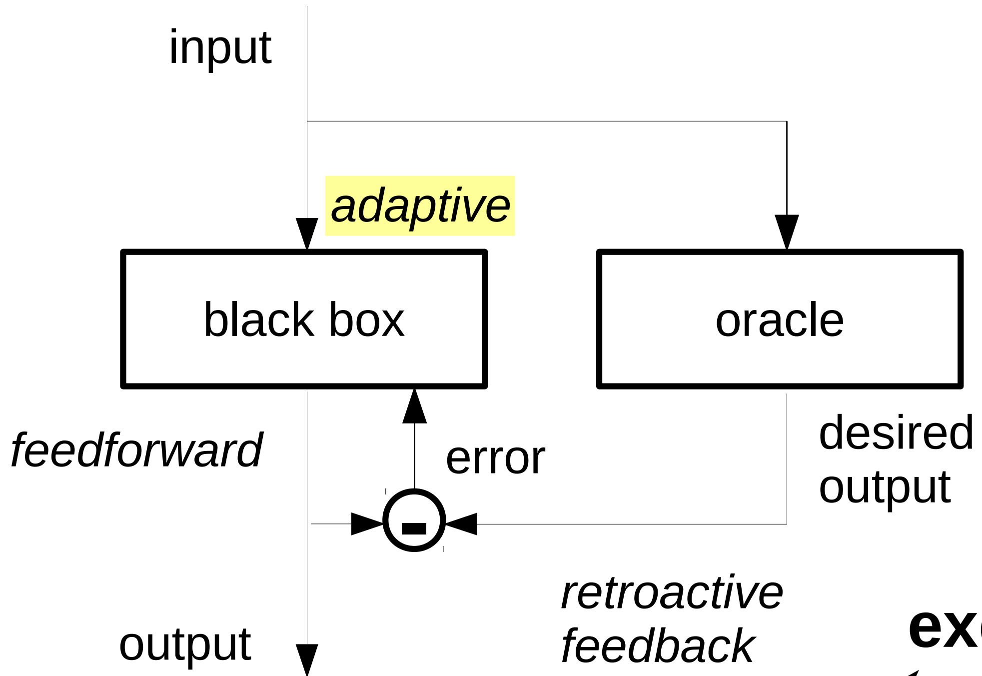


Supervised Machine Learning



- 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

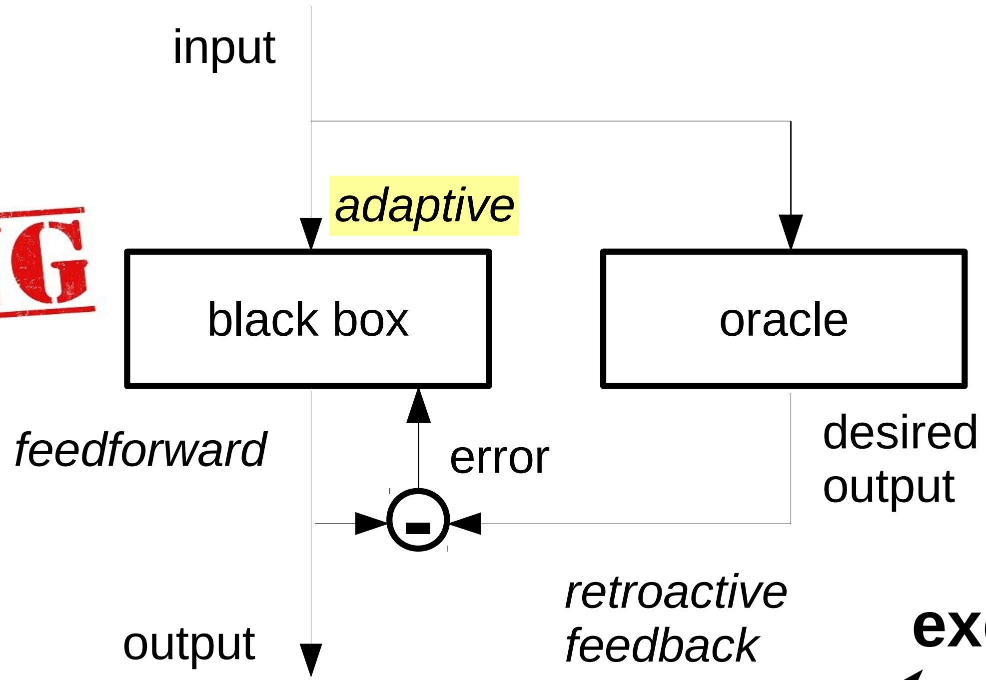
- In general, supervised machine learning involves:
  - a data-flow computational network
  - parameters distributed along the network
  - a ML method enabling adaptation of parameters against some feedback, e.g. output error in the training phase
  - an oracle making targets explicit **intentional setup**

plan

Supervised Machine Learning

higher-level diagnostic feedback?

MISSING



lower-level diagnostic feedback

planner

executor

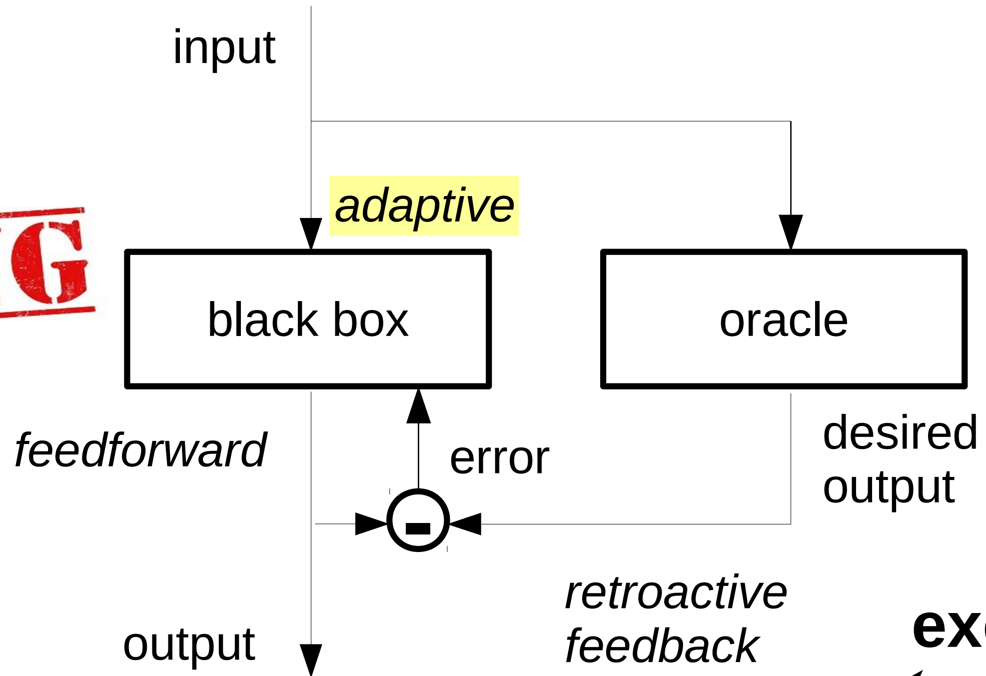
- In general, supervised machine learning involves:
  - a data-flow computational network
  - parameters distributed along the network
  - a ML method enabling adaptation of parameters against some feedback, e.g. output error in the training phase
  - an oracle making targets explicit

plan

intentional setup

Supervised Machine Learning

higher-level diagnostic feedback?



lower-level diagnostic feedback

planner

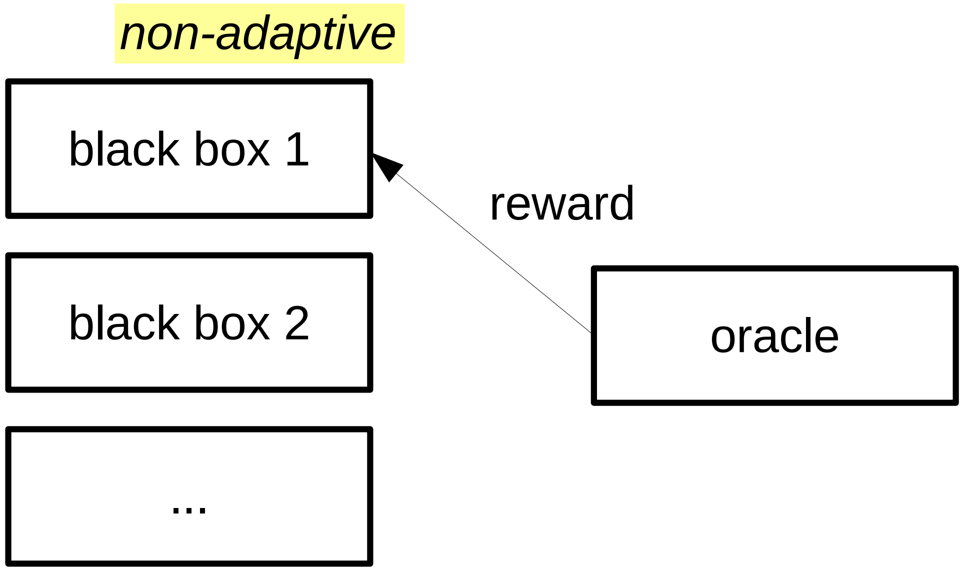
executor

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

plan

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

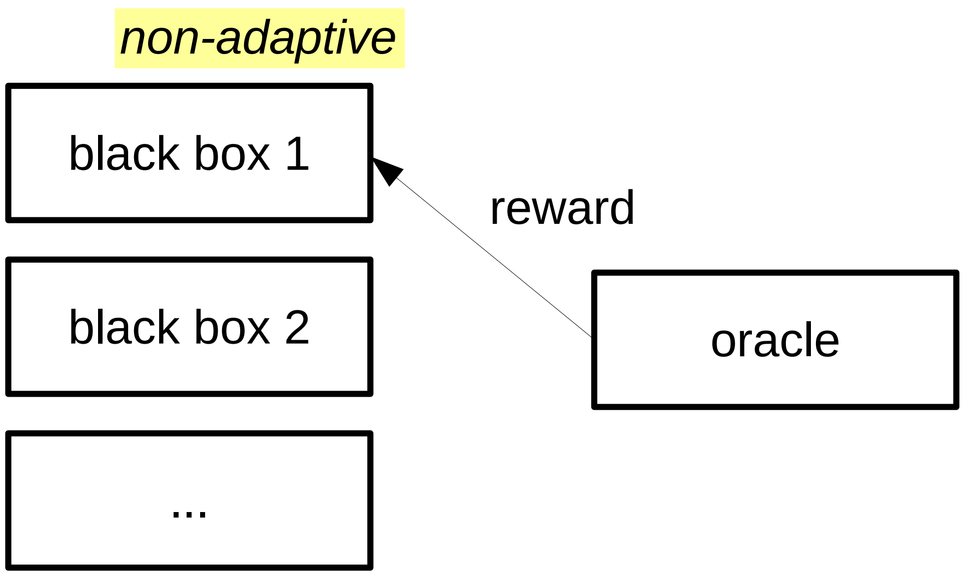
Evolutionary view



- 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?

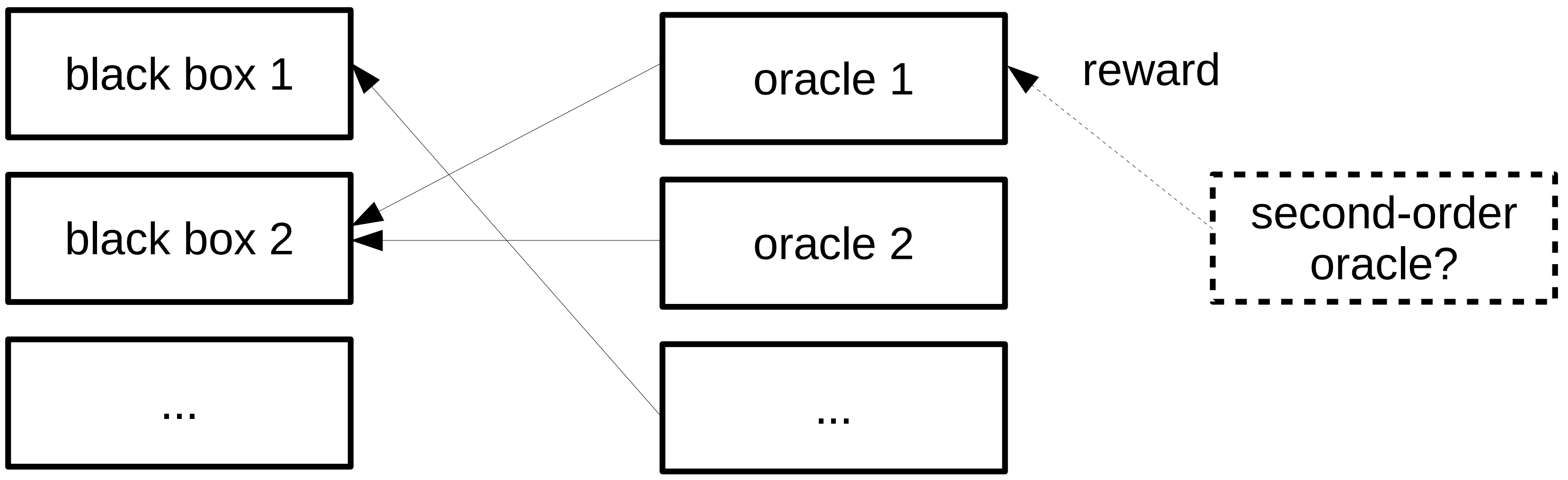
Evolutionary view



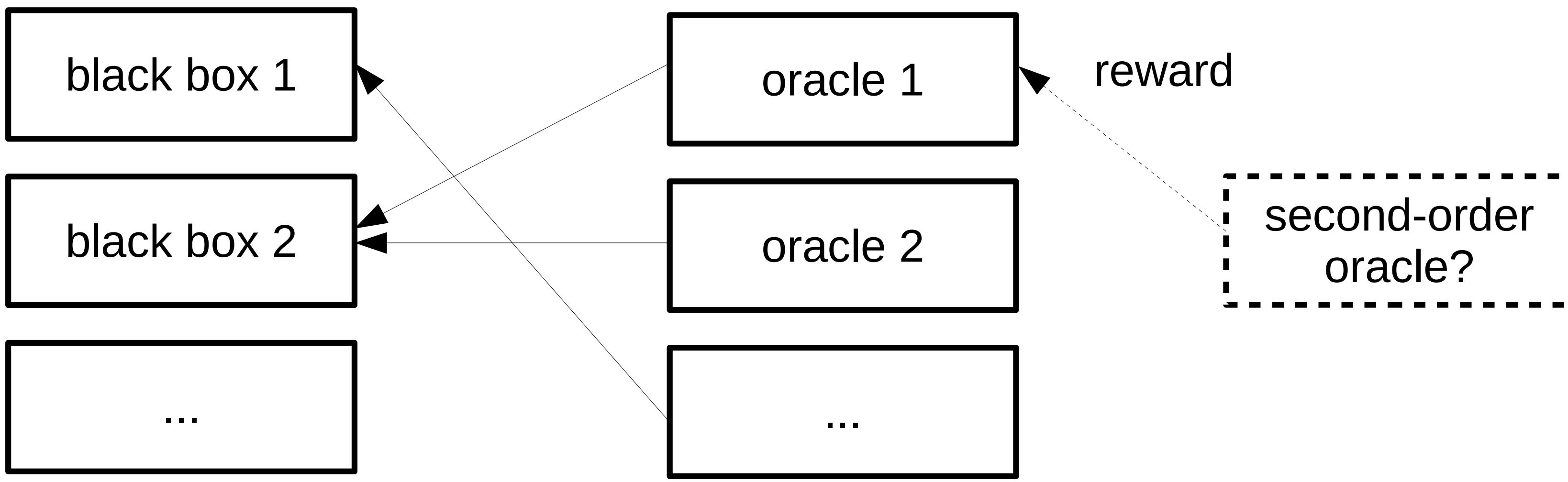
- In evolutionary terms, we could consider a multitude of different non-adaptive black-boxes, covering several configurations of parameters, competing for computational resources.
  - For each learning step, the oracle sets the means to select the best performing black-box(es), for which access to computational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

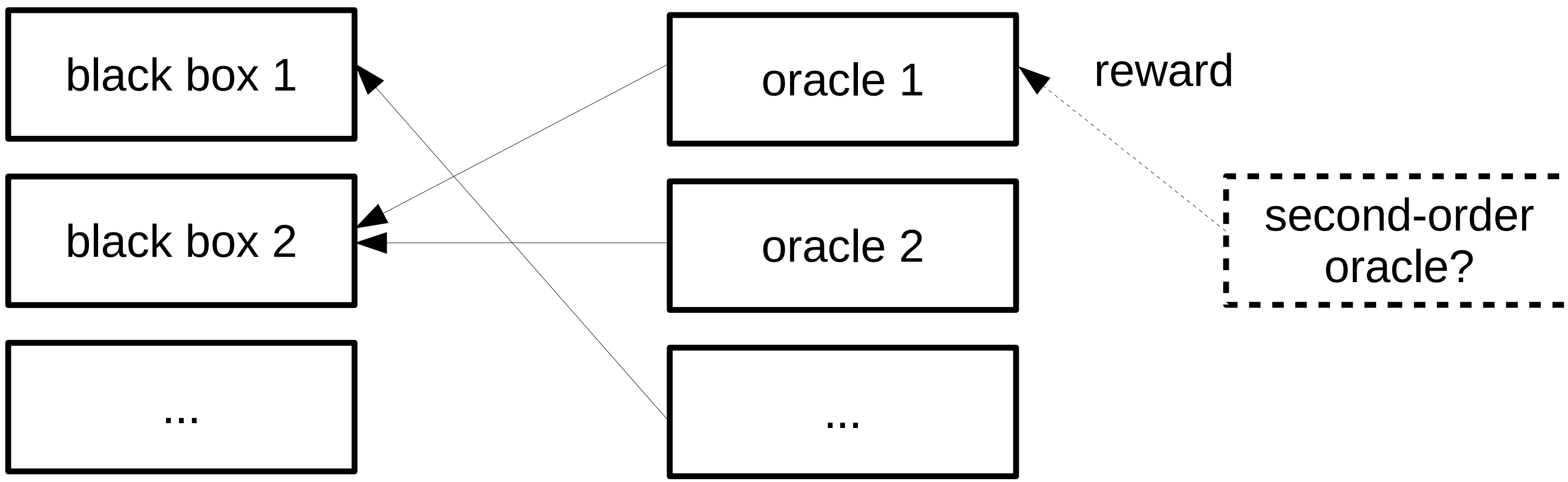


Evolutionary view



- 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**)

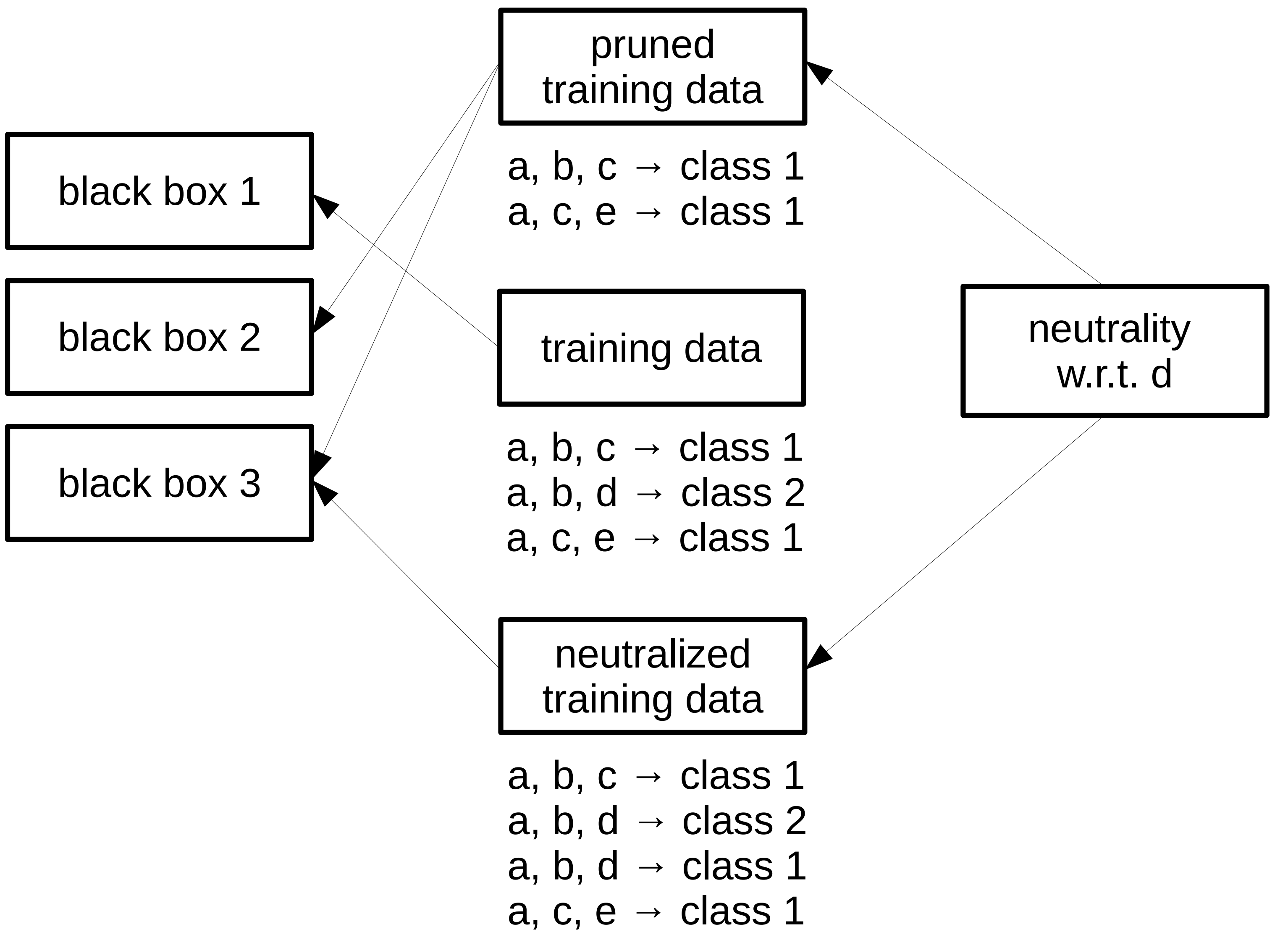
Evolutionary view



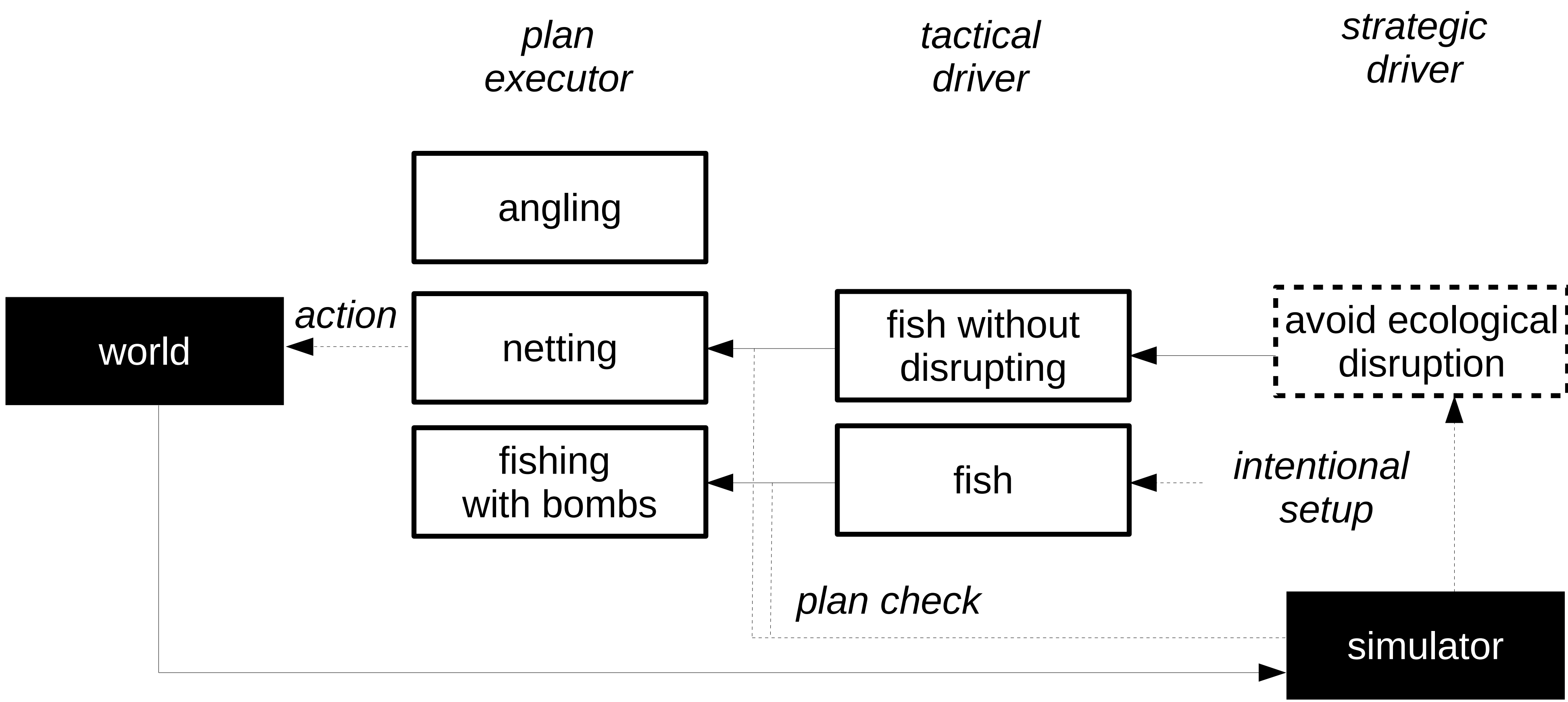
- Let’s use this architecture on a concrete example: IBM Watson (building upon a network of intelligent QA agents).
  - a question is given
  - the system has to guess
    - what the question demands (~ **oracles**)
    - what is the answer (~ **black-box**).

Let’s apply it to our initial problems!

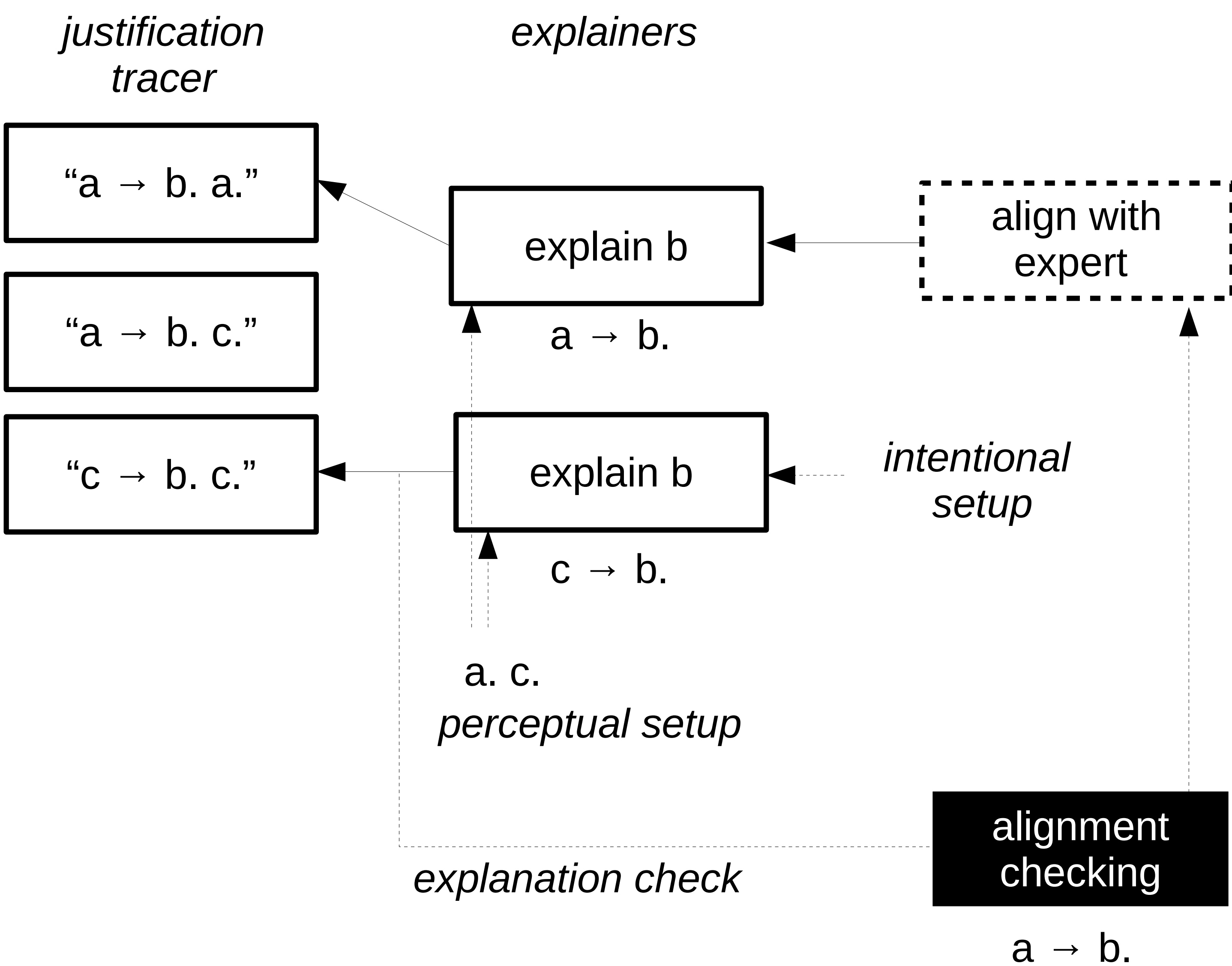
Example: neutrality constraint



Example: strategic protection to unintended consequences



Example: alignment to expert knowledge for explanation



Perspectives

- This position paper aims to highlight the crucial role of **normware** with respect to trustworthy and explainable AI
  - ML approaches usually do not consider this level of abstraction
  - ethical/responsible AI studies target higher level constraints

Perspectives

- This position paper aims to highlight the crucial role of **normware** with respect to trustworthy and explainable AI
  - 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 position paper aims to highlight the crucial role of **normware** with respect to trustworthy and explainable AI
  - 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!

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

A less tentative taxonomy

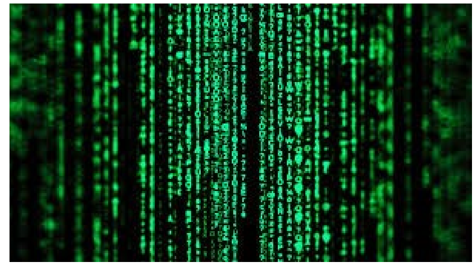
hardware

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