Like Circles in the Water: Responsibility as a System-Level Function

XAILA Workshop on eXplainable and responsible AI and Law, joint with JURIX 2020
9 December 2020 @ Brno/Prague (virtual)

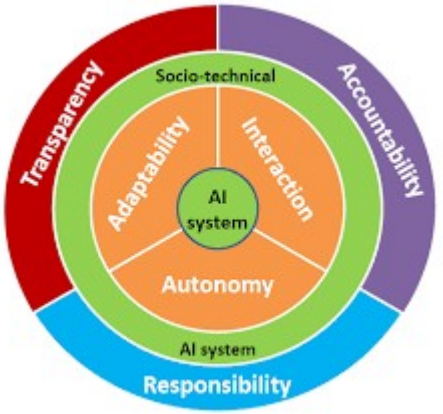
Giovanni Sileno[a] (g.sileno@uva.nl)
Alexander Boer[b], Geoff Gordon[a], Bernhard Rieder[a]

[a] Informatics Institute, University of Amsterdam, the Netherlands
[b] KPMG, Amsterdam, the Netherlands

Responsible, Ethical, Fair, Trustworthy, ... AI

higher-level contributions

ART principles



AI-HLEG guidelines and recommendations

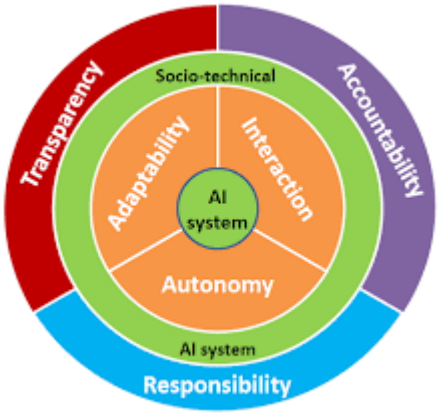
meaningful (human) control

Responsible, Ethical, Fair, Trustworthy, ... AI

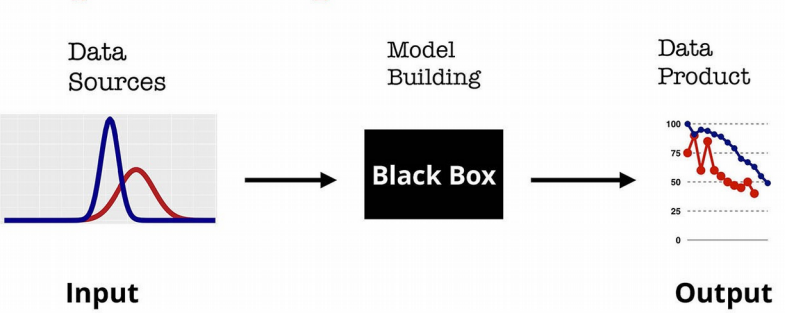
higher-level contributions

technical contributions

ART principles



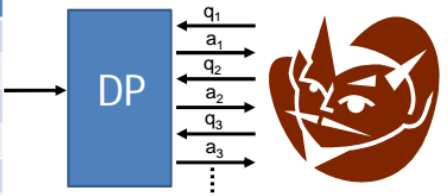
algorithmic fairness



AI-HLEG guidelines and recommendations

meaningful (human) control

Sex	Blood	⋯	HIV
F	B	⋯	Y
M	A	⋯	N
M	O	⋯	N
M	O	⋯	Y
F	A	⋯	N
M	B	⋯	Y



differential privacy

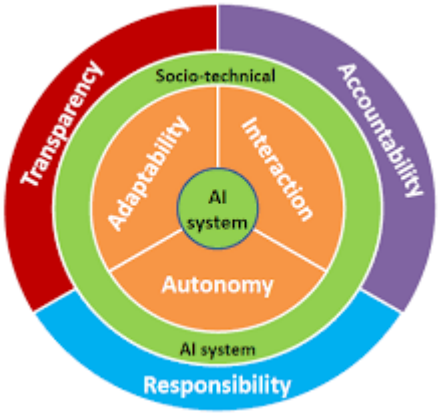
Responsible, Ethical, Fair, Trustworthy, ... AI

higher-level contributions

wide gap

technical contributions

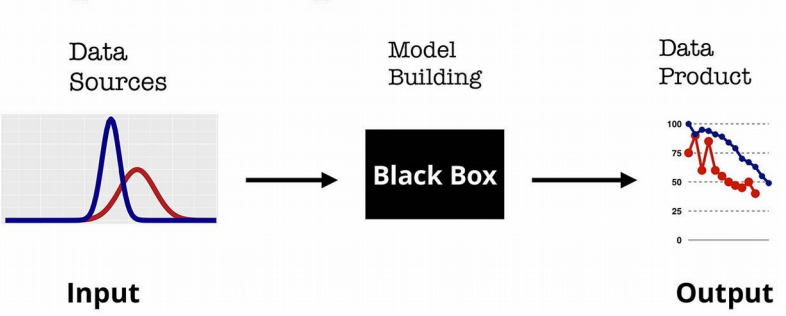
ART principles

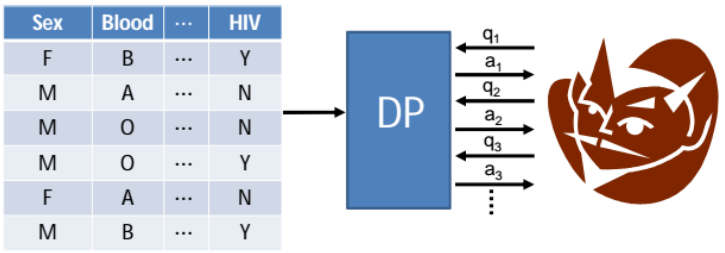


AI-HLEG guidelines and recommendations

meaningful (human) control

algorithmic fairness





differential privacy

abstract principles

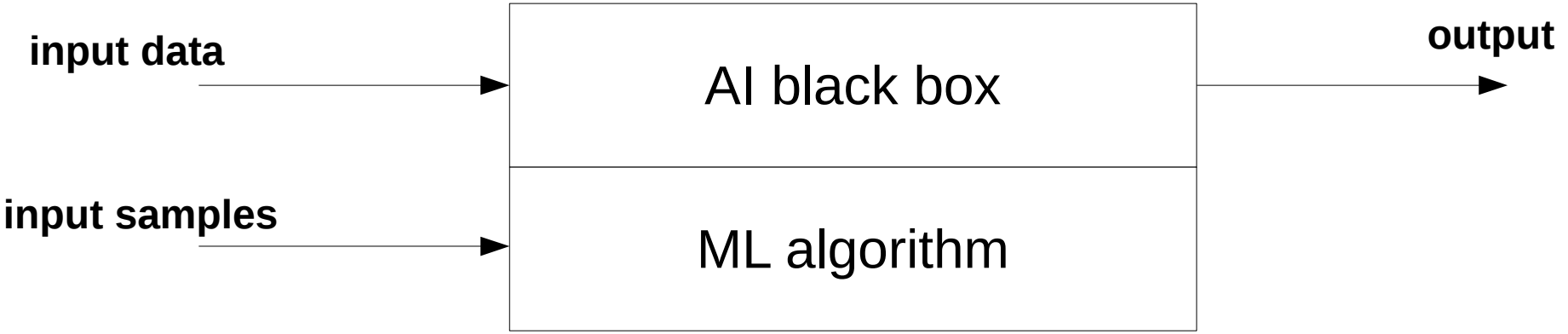
design principles

requirements meant to support the promotion of abstract principles

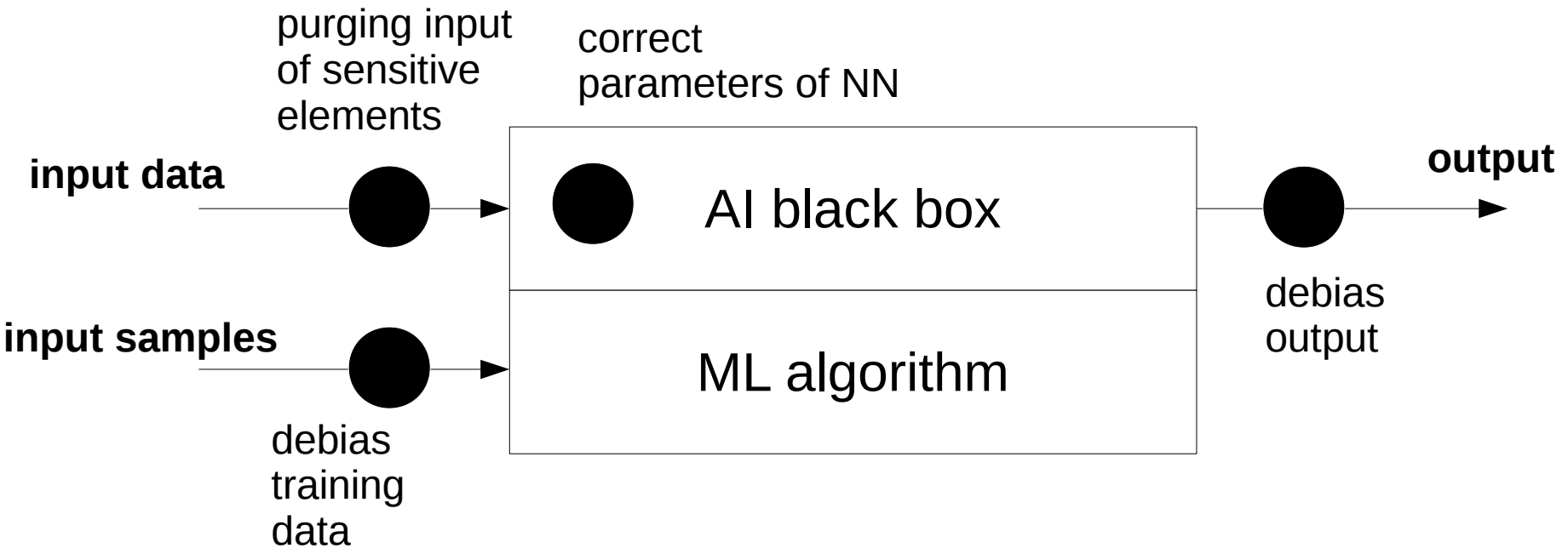
wide gap

computational instructions

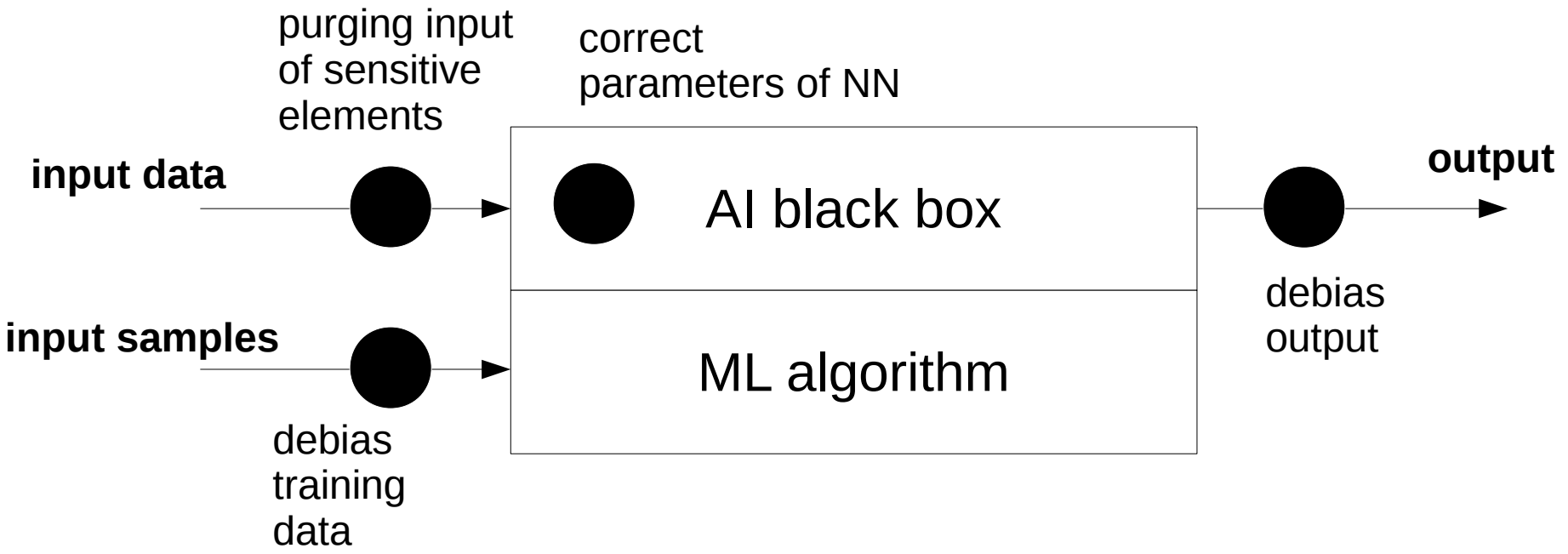
Algorithmic approaches to Fairness



Algorithmic approaches to Fairness



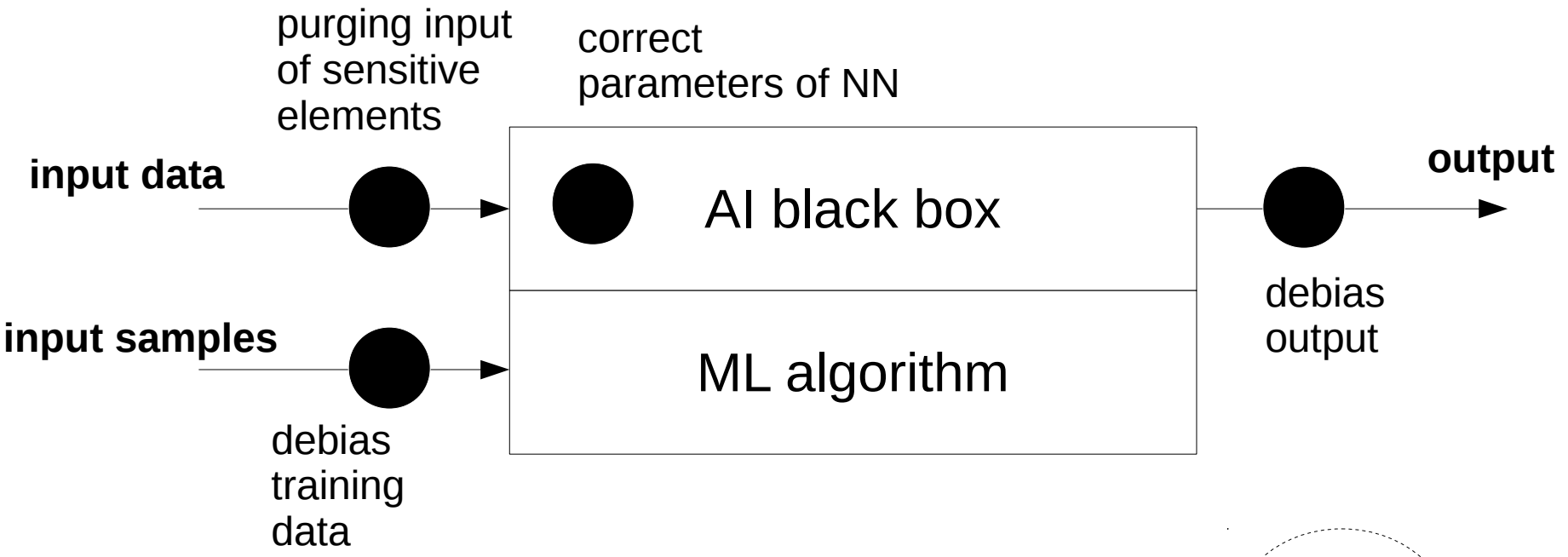
Algorithmic approaches to Fairness



computational reflection

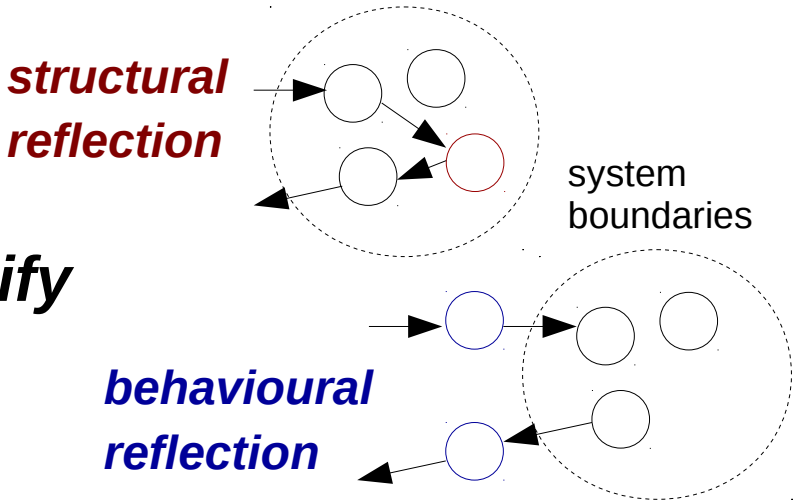
the ability of a system to ***inspect*** and ***modify*** itself in order to *improve its performance*

Algorithmic approaches to Fairness

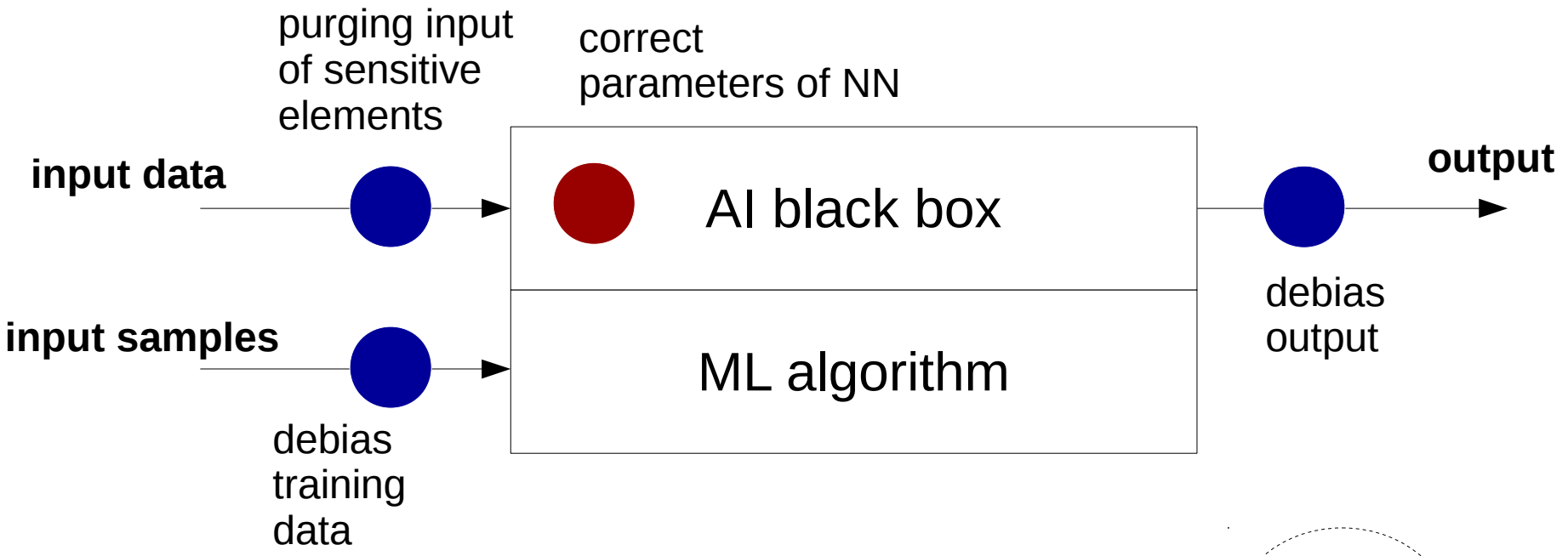


computational reflection

the ability of a system to ***inspect*** and ***modify*** itself in order to *improve its performance*

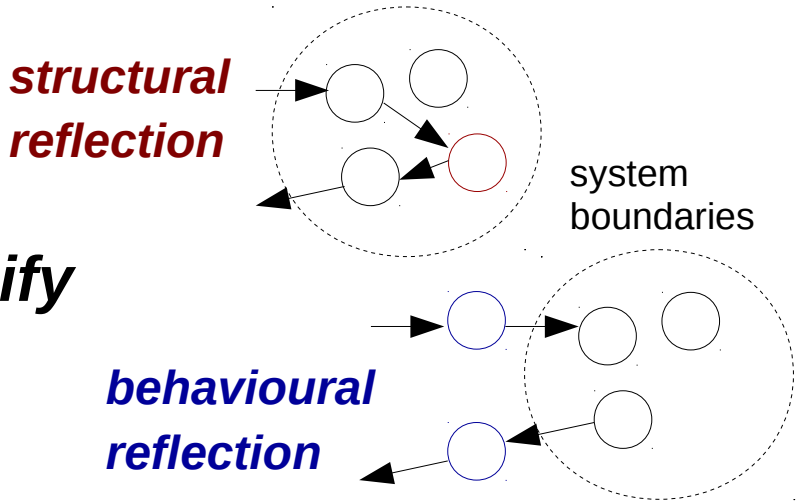


Algorithmic approaches to Fairness

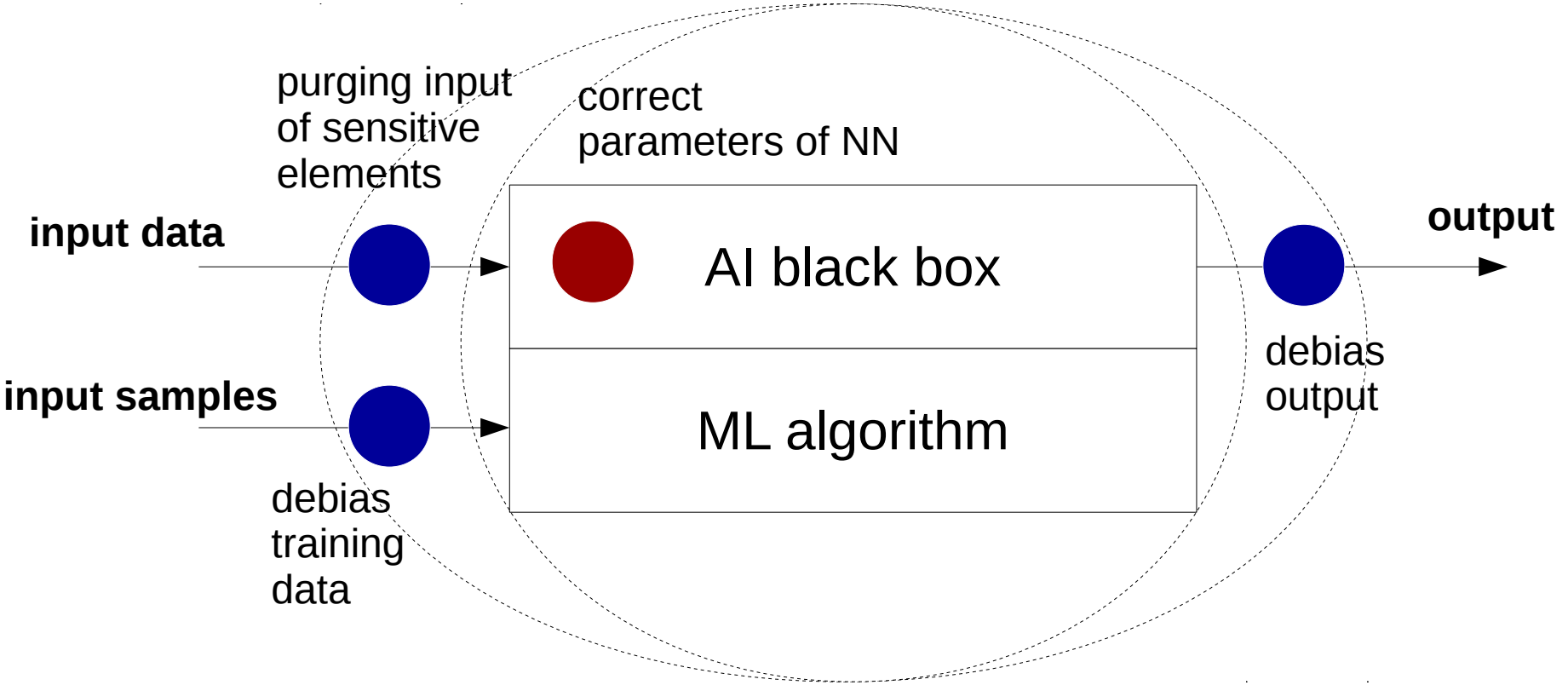


computational reflection

the ability of a system to ***inspect*** and ***modify*** itself in order to *improve its performance*

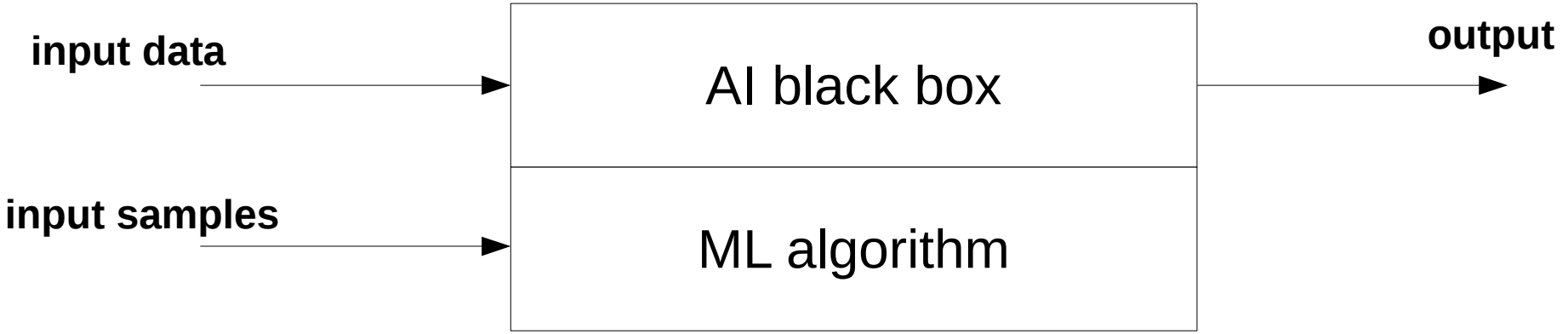


Algorithmic approaches to Fairness



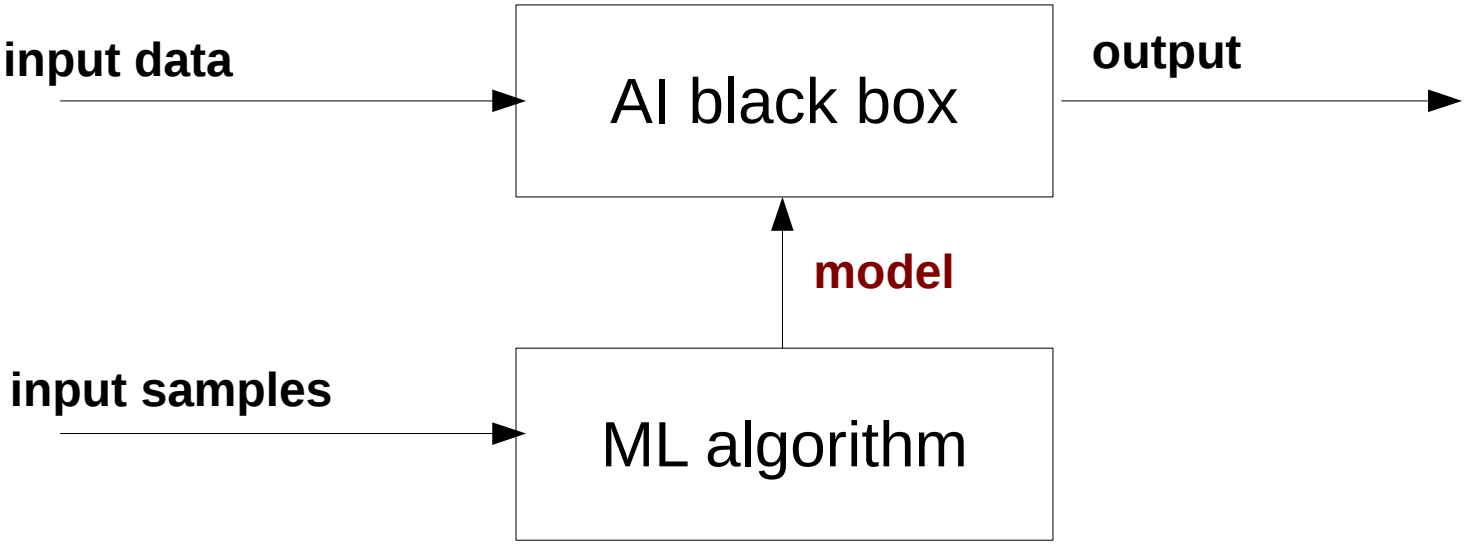
- All these methods focus on ***data***: input, output, relative to model.
- Behavioural reflection is only concerned by the very first layer beyond the system boundaries.

Looking at the bigger picture



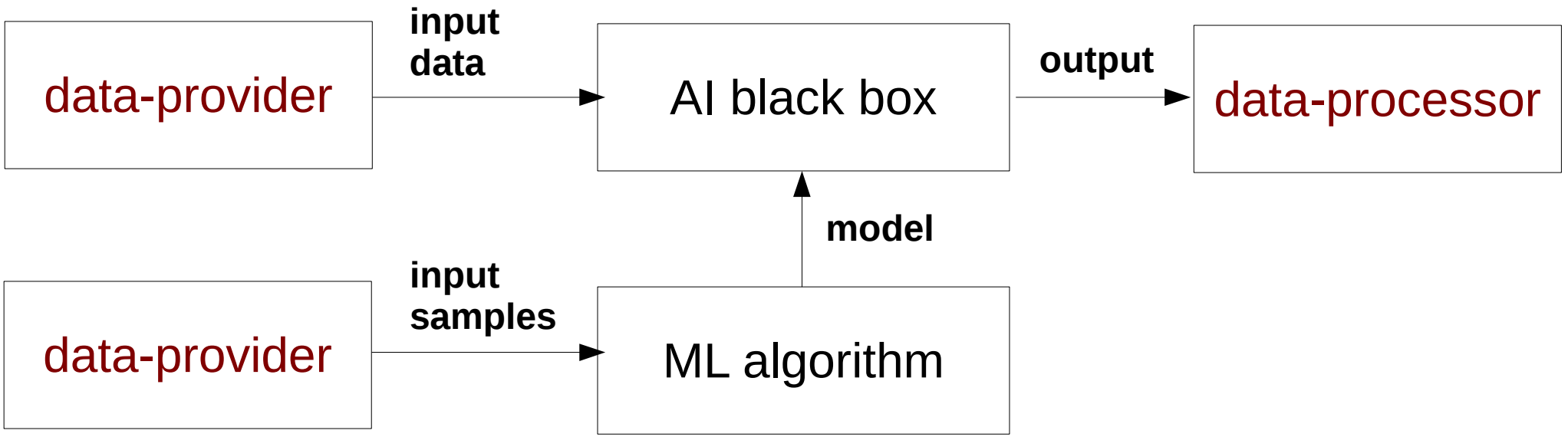
...starting from the previous core...

Looking at the bigger picture



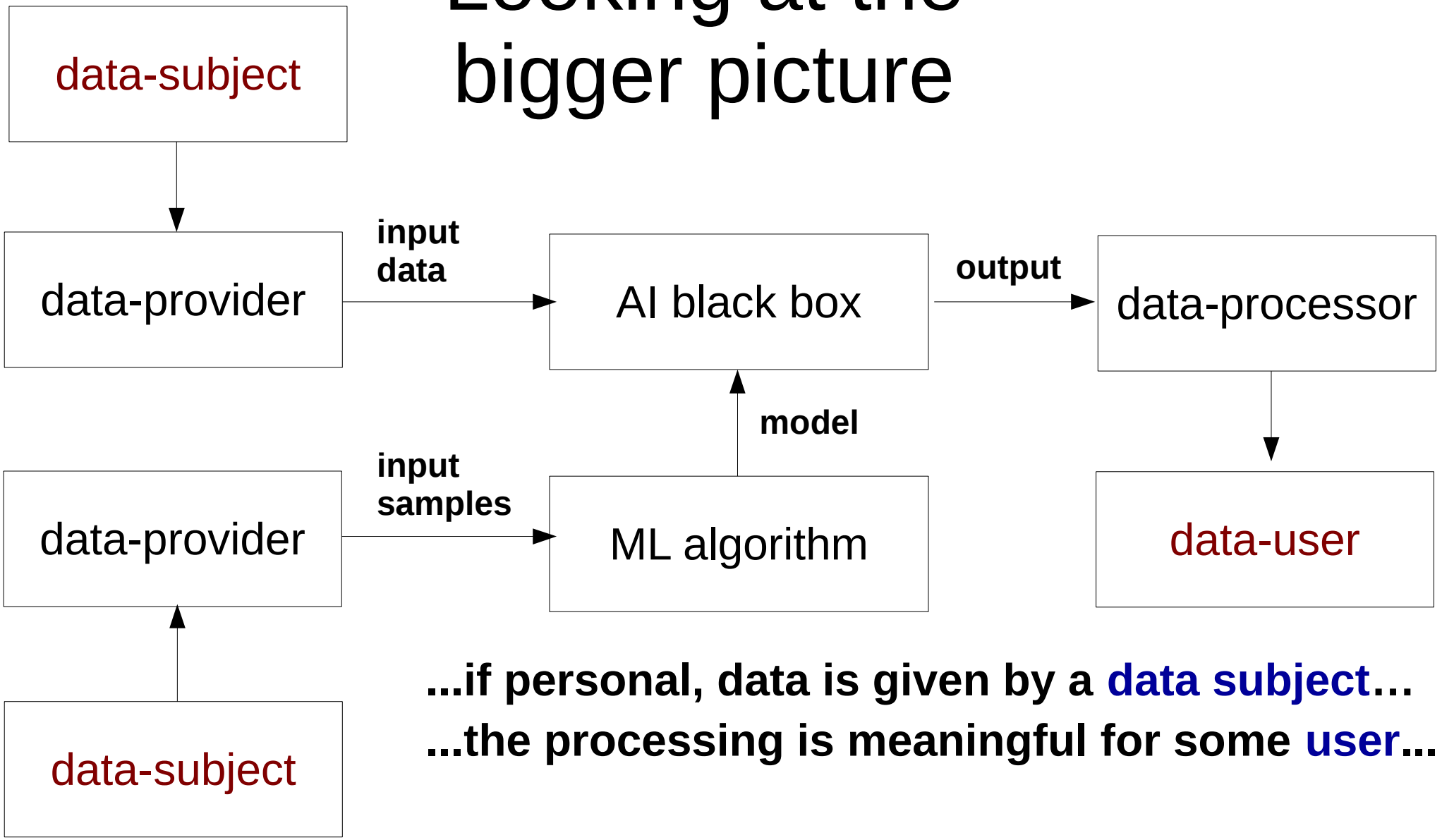
...we can decompose inference from training modules...

Looking at the bigger picture



...some other module is providing and taking data...

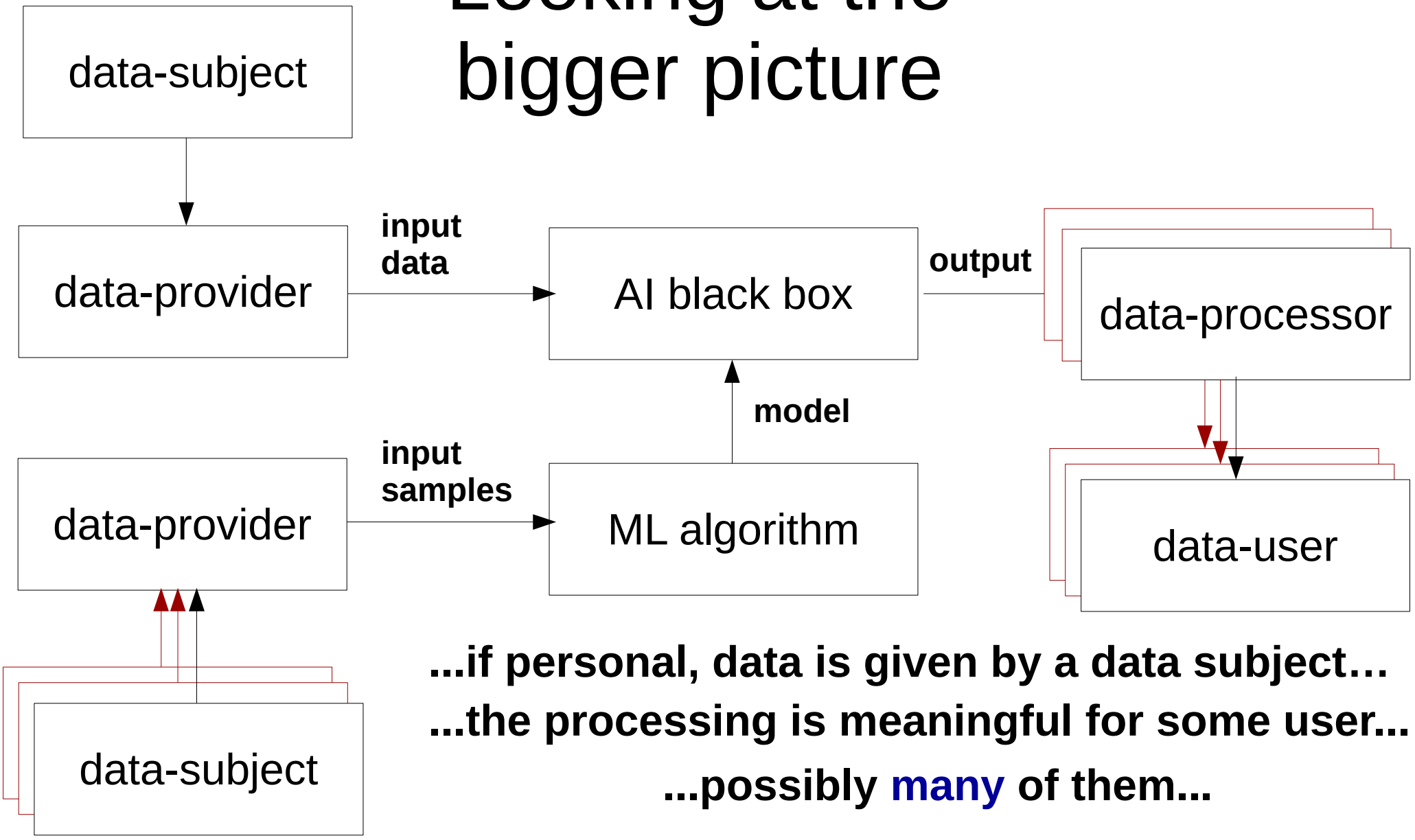
Looking at the bigger picture



...if personal, data is given by a data subject…

...the processing is meaningful for some user...

Looking at the bigger picture

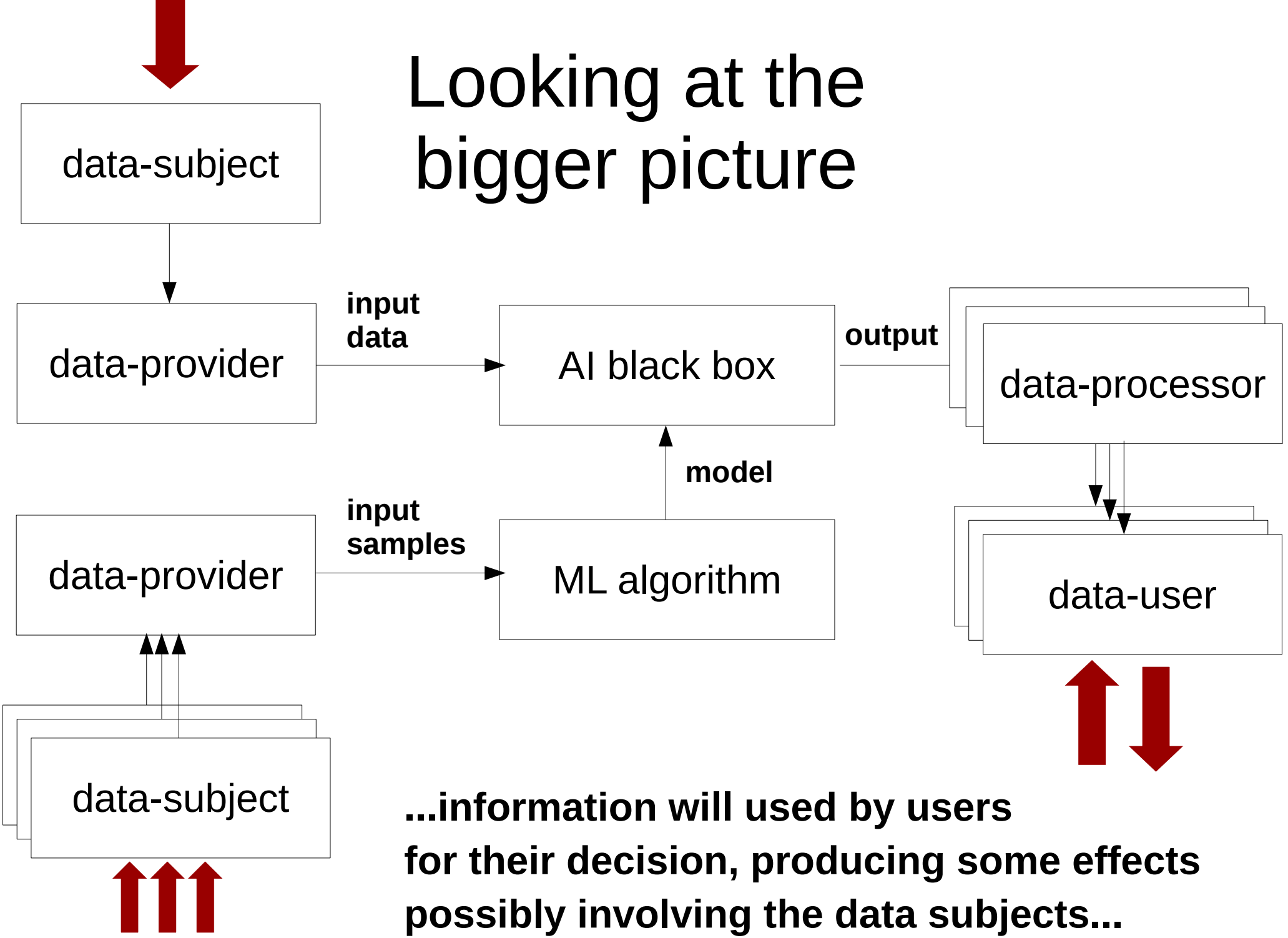


...if personal, data is given by a data subject…

...the processing is meaningful for some user...

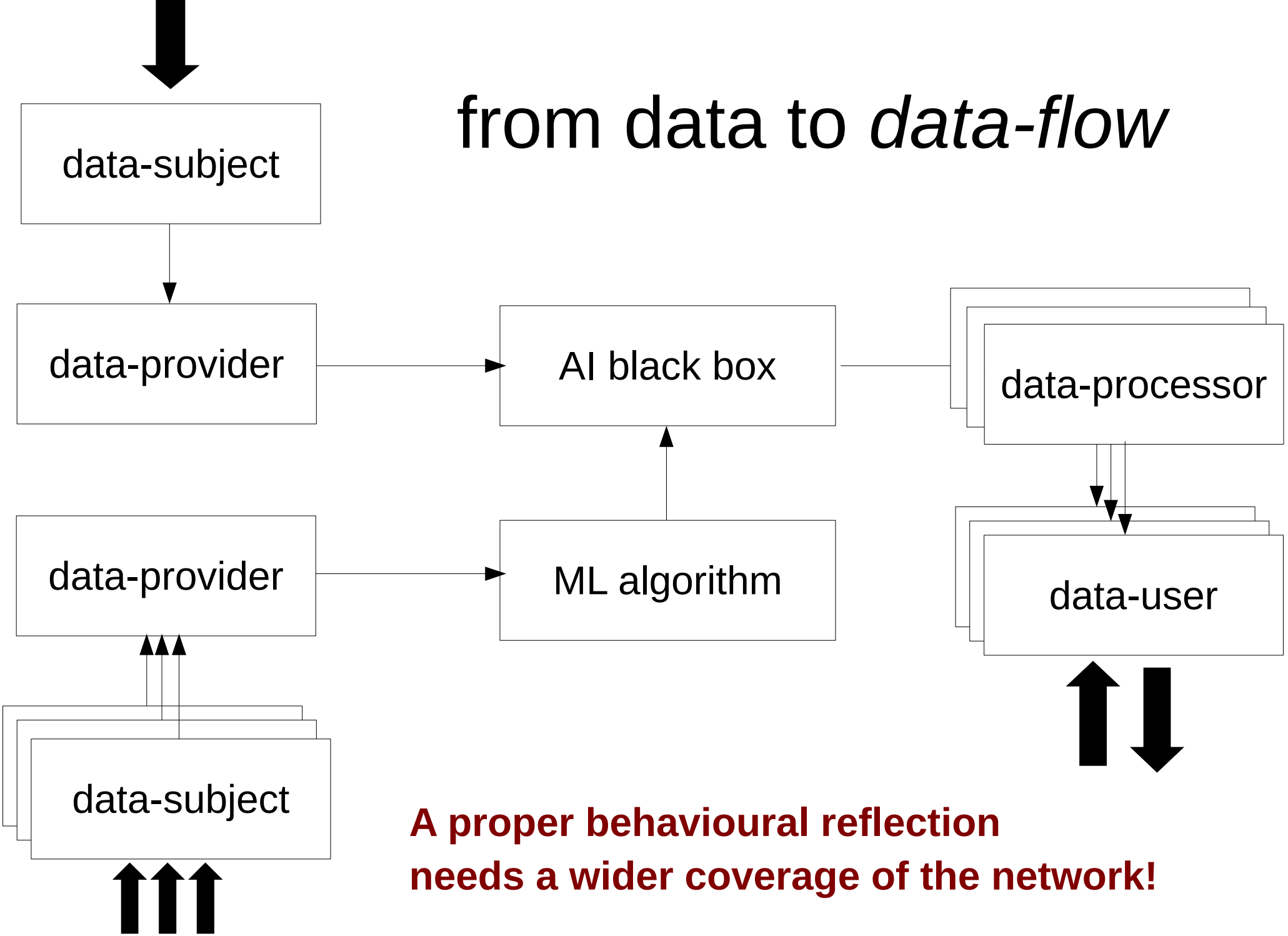
...possibly many of them...

Looking at the bigger picture



...information will used by users for their decision, producing some effects possibly involving the data subjects...

from data to *data-flow*



A proper behavioural reflection needs a wider coverage of the network!

Responsible processing of data streams

- A data-flow view makes clear that several problems of responsible AI map to **control of *information disclosure***

Responsible processing of data streams

- A data-flow view makes clear that several problems of responsible AI map to **control of *information disclosure***
- **Discrimination**: informational transmission resulting in distinguishing/characterizing information about an individual

Responsible processing of data streams

- A data-flow view makes clear that several problems of responsible AI map to **control of *information disclosure***
- **Discrimination**: informational transmission resulting in distinguishing/characterizing information about an individual;
- **Privacy**: rights and powers controlling channels transmitting self-information

Responsible processing of data streams

- A data-flow view makes clear that several problems of responsible AI map to **control of *information disclosure***
- **Discrimination**: informational transmission resulting in distinguishing/characterizing information about an individual;
- **Privacy**: rights and powers controlling channels transmitting self-information
- **Differential privacy**: addition of noise channels in informational transmission

Responsible processing of data streams

- A data-flow view makes clear that several problems of responsible AI map to **control of *information disclosure***
- **Discrimination**: informational transmission resulting in distinguishing/characterizing information about an individual;
- **Privacy**: rights and powers controlling channels transmitting self-information
- **Differential privacy**: addition of noise channels in informational transmission
- **(some types of) frauds:** information leakage between theoretically independent components

Responsible processing of data streams

- A data-flow view makes clear that several problems of responsible AI map to **control of *information disclosure***
- It makes clear the role of the **network topology** in *enabling*, but also *inhibiting* a certain result.
- It goes beyond the computational domain, and includes **human** (informational and decision-making) **aspects**

Responsible processing of data streams

- A data-flow view makes clear that several problems of responsible AI map to **control of *information disclosure***
- It makes clear the role of the **network topology** in *enabling*, but also *inhibiting* a certain result.
- It goes beyond the computational domain, and includes **human** (informational and decision-making) **aspects**

- e.g. using sensitive data as ethnicity (or a proxy of it) is
  - unfair for deciding the premium for an insurance policy,
  - not unfair to suggest the colour/style of a dress in an e-shop

Responsible processing of data streams

- A data-flow view makes clear that several problems of responsible AI map to **control of *information disclosure***
- It makes clear the role of the **network topology** in *enabling*, but also *inhibiting* a certain result.
- It goes beyond the computational domain, and includes **human** (informational and decision-making) **aspects**

- e.g. using sensitive data as ethnicity (or a proxy of it) is
  - unfair for deciding the premium for an insurance policy,
  - not unfair to suggest the colour/style of a dress in an e-shop
- We don’t need differential privacy if we are querying our own personal data.

Responsible processing of data streams

- A data-flow view makes clear that several problems of responsible AI map to **control of *information disclosure***
- It makes clear the role of the **network topology** in *enabling*, but also *inhibiting* a certain result.
- It goes beyond the computational domain, and includes **human** (informational and decision-making) **aspects**

- e.g. using sensitive data as ethnicity (or a proxy of it) is
  - unfair for deciding the premium for an insurance policy,
  - not unfair to suggest the colour/style of a dress in an e-shop
- We don’t need differential privacy if we are querying our own personal data.

What is “context”?

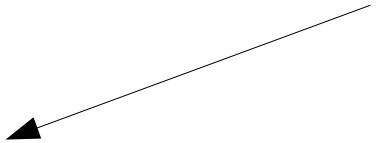
Contextual integrity

- **Privacy can not be defined in absolute terms.**
- Actors: data subject, sender, recipient
- Type of information
- Basis for disclosure
- Contextual elements (legal and social roles, practices, etc.)

Nissenbaum, H.: Privacy In Context: Technology Policy And The Integrity Of Social Life (2009)

Contextual integrity

- **Privacy can not be defined in absolute terms.**
- Actors: data subject, sender, recipient
- Type of information
- Basis for disclosure
- Contextual elements (legal and social roles, practices, etc.)



domain knowledge,
used by subjects to form their **expectations**

Nissenbaum, H.: Privacy In Context: Technology Policy And The Integrity Of Social Life (2009)

Contextual Demographic Disparity

- Wachter et al. [2020] analyse the decisions of the *European Court of Justice* in cases of discrimination.

Wachter, S., Mittelstadt, B., Russell, C.: Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI. SSRN Electronic Journal pp. 1–72 (2020)

Contextual Demographic Disparity

- Wachter et al. [2020] analyse the decisions of the *European Court of Justice* in cases of discrimination.
- CDD seems to be the best measure for assessment:

Wachter, S., Mittelstadt, B., Russell, C.: Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI. SSRN Electronic Journal pp. 1–72 (2020)

Contextual Demographic Disparity

- Wachter et al. [2020] analyse the decisions of the *European Court of Justice* in cases of discrimination.
- CDD seems to be the best measure for assessment:
  - A norm protects certain groups of people.
  - A decision produce a positive or negative outcome dividing the people in two classes, advantaged and disadvantaged.
  - There is disparity if $A_R < D_R$ for all conditions R

Wachter, S., Mittelstadt, B., Russell, C.: Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI. SSRN Electronic Journal pp. 1–72 (2020)

Contextual Demographic Disparity

- Wachter et al. [2020] analyse the decisions of the *European Court of Justice* in cases of discrimination.
- CDD seems to be the best measure for assessment:
  - A norm protects certain groups of people.
  - A decision produce a positive or negative outcome dividing the people in two classes, advantaged and disadvantaged.
  - There is disparity if $A_R < D_R$ for all conditions R
- How to decide R ?

Wachter, S., Mittelstadt, B., Russell, C.: Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI. SSRN Electronic Journal pp. 1–72 (2020)

Contextual Demographic Disparity

- Wachter et al. [2020] analyse the decisions of the *European Court of Justice* in cases of discrimination.
- CDD seems to be the best measure for assessment:
  - A norm protects certain groups of people.
  - A decision produce a positive or negative outcome dividing the people in two classes, advantaged and disadvantaged.
  - There is disparity if $A_R < D_R$ for all conditions R
- How to decide R ?
  - **explanatory conditions**: potential **causes** explaining the disparity, e.g. working hours w.r.t. different salaries.

Wachter, S., Mittelstadt, B., Russell, C.: Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI. SSRN Electronic Journal pp. 1–72 (2020)

“What causes what?”

responsibility

Responsibility attribution

- In human societies, responsibility attribution is a ***spontaneous*** and ***seemingly universal*** behaviour.

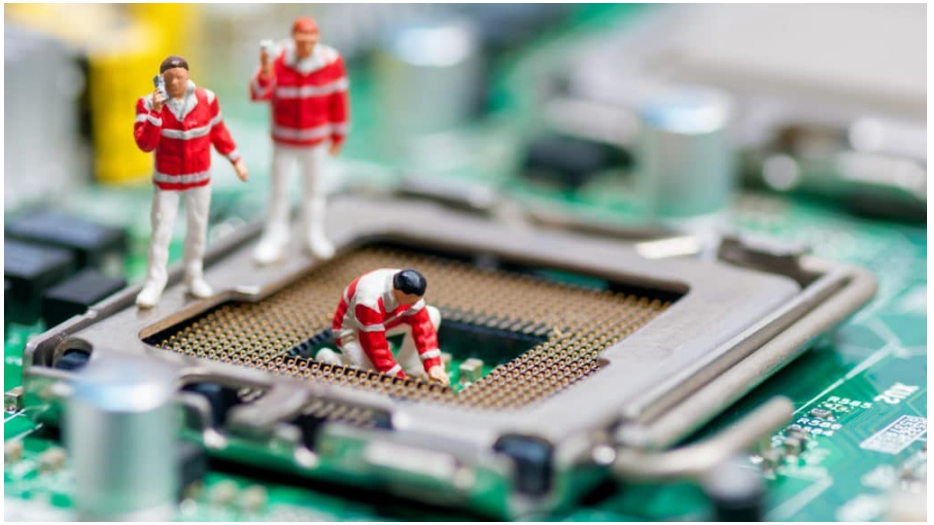
.

Responsibility attribution

- In human societies, responsibility attribution is a ***spontaneous*** and ***seemingly universal*** behaviour.

FUNCTION OF RESPONSIBILITY

Localization of failures in constructions whose components are deemed to be independent/autonomous.



Responsibility attribution

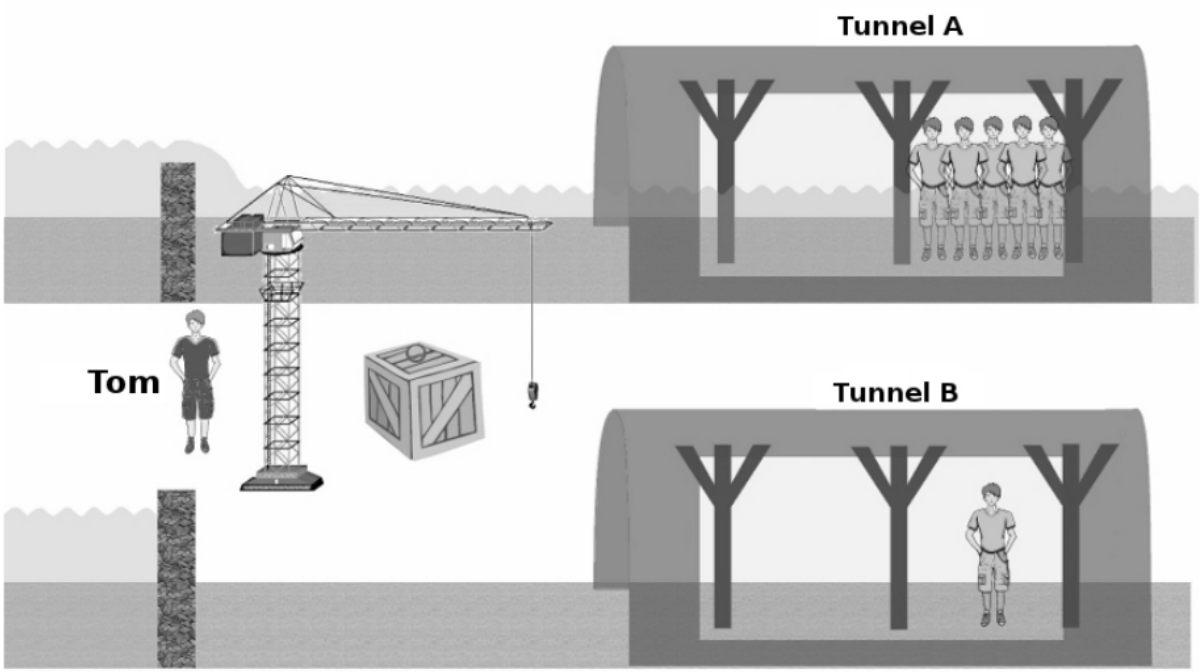
- In human societies, responsibility attribution is a ***spontaneous*** and ***seemingly universal*** behaviour.

  FUNCTION OF RESPONSIBILITY

  Localization of failures in constructions whose components are deemed to be independent/autonomous.

- Crucial distinction between:
  - **Causal responsibility** (physical, operational, etc.)
  - **Moral responsibility** (legal, social, etc.)

Sileno, G., Saillenfest, A., Dessalles, J.L.: A Computational Model of Moral and Legal Responsibility via Simplicity Theory. JURIX 2017 FAIA 302, 171–176 (2017)



flooded mine dilemma *(trolley problem variation)*

- Experiments show that people are more prone to blame an agent for an action:
  - the more the outcome is severe,
  - the more they are closer to the victims,
  - the more the outcome follows the action.

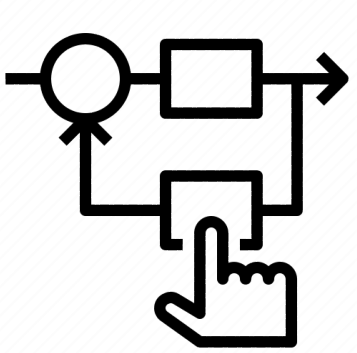
[A. Saillenfest and J.-L. Dessalles. Role of Kolmogorov Complexity on Interest in Moral Dilemma Stories. CogSCI 2012, pages 947–952]

Agentive responsibility

- The agent has **agentive responsibility** if it:

Agentive responsibility

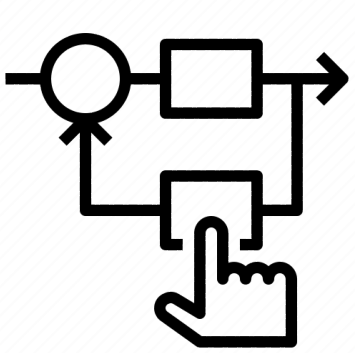
- The agent has **agentive responsibility** if it:



has the ability to **control its behaviour**

Agentive responsibility

- The agent has **agentive responsibility** if it:

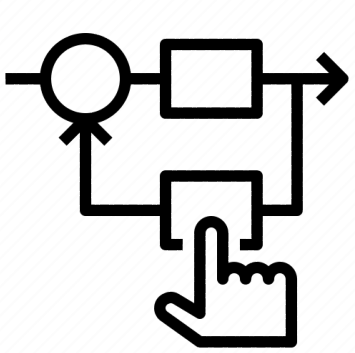


has the ability to **control its behaviour**

has the ability to **foresee the associated outcomes**

Agentive responsibility

- The agent has **agentive responsibility** if it:



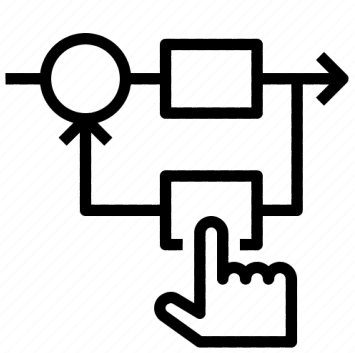
has the ability to **control its behaviour**

has the ability to **foresee the associated outcomes**

has the ability to **assess their impact according to a preferential or value structure**

Agentive responsibility

- The agent has **agentive responsibility** if it:



has the ability to **control its behaviour**

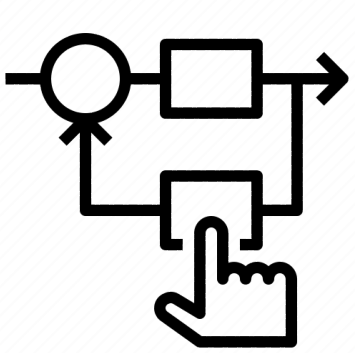
has the ability to **foresee the associated outcomes**

has the ability to **assess their impact according to a preferential or value structure**

according to some STANDARD...

Agentive responsibility

- The agent has **agentive responsibility** if it:



has the ability to **control its behaviour**

has the ability to **foresee the associated outcomes**

has the ability to **assess their impact according to a preferential or value structure**

RISK ASSESSMENT

according to some STANDARD...

abstract principles

design principles

requirements meant to support the promotion of abstract principles

wide gap

computational instructions

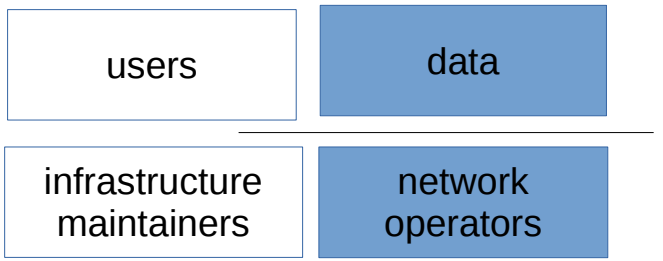
ART principles

- **Accountability:**
  - motivations for the decision-making (values, norms, etc.) need to be explicit
- **Responsibility:**
  - the chain of (human) control (designer, manufacturer, operator, etc.) needs to be clear
- **Transparency:**
  - actions need to be explained in terms of algorithms and data, and it should be possible to inspect them.

Dignum, V.: Responsible autonomy. Proceedings of International Joint Conference on Artificial Intelligence (IJCAI) pp. 4698–4704 (2017).

Responsible Internet

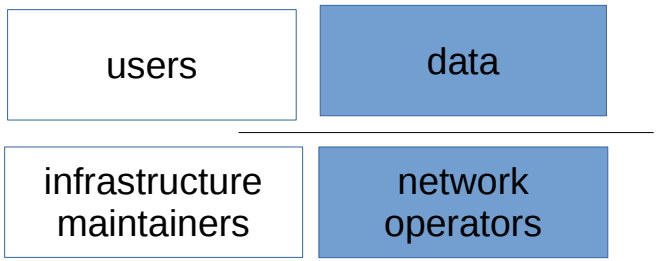
decentralized computational architecture
data and *infrastructure* dimensions



Hesselman, C., Grosso, P., Holz, R., Kuipers, F., Xue, J.H., Jonker, M., de Ruiter, J., Sperotto, A., van Rijswijk-Deij, R., Moura, G.C., Pras, A., de Laat, C.: A Responsible Internet to Increase Trust in the Digital World. Journal of Network and Systems Management 28(4), 882–922 (2020)

Responsible Internet

decentralized computational architecture
data and *infrastructure* dimensions

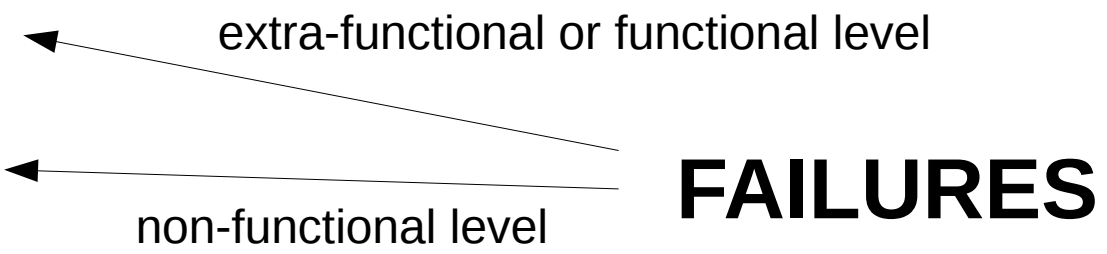


- **Accountability**
- **Transparency**
- **Controllability** instead of Responsibility:
  - ability of users to specify how network operators should handle their data (generally by means of path control)
  - ability of infrastructure maintainers to set constraints over network operators
- **Usability**

Hesselman, C., Grosso, P., Holz, R., Kuipers, F., Xue, J.H., Jonker, M., de Ruiter, J., Sperotto, A., van Rijswijk-Deij, R., Moura, G.C., Pras, A., de Laat, C.: A Responsible Internet to Increase Trust in the Digital World. Journal of Network and Systems Management 28(4), 882–922 (2020)

Responsible Internet proposal

- **Accountability**
- **Transparency**
- **Controllability**



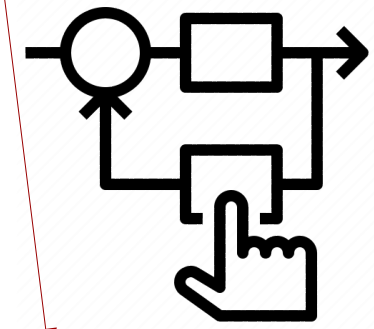
What responsibility?

- **Accountability**
- **Transparency**
- **Controllability**

extra-functional or functional level

non-functional level

FAILURES



has the ability to **control its behaviour**

has the ability to **foresee the associated outcomes**

partially

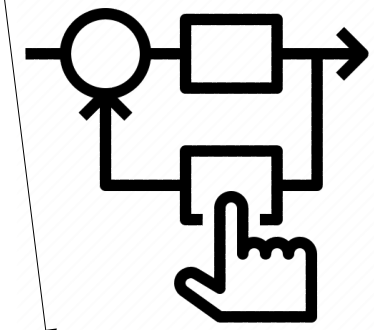
has the ability to **assess their impact according to a preferential or value structure**

What responsibility?

- **Accountability**
- **Transparency**
- **Controllability**

requires knowledge constructed from system practices, and users are rarely in position to know

?



has the ability to **control its behaviour**

has the ability to **foresee the associated outcomes**

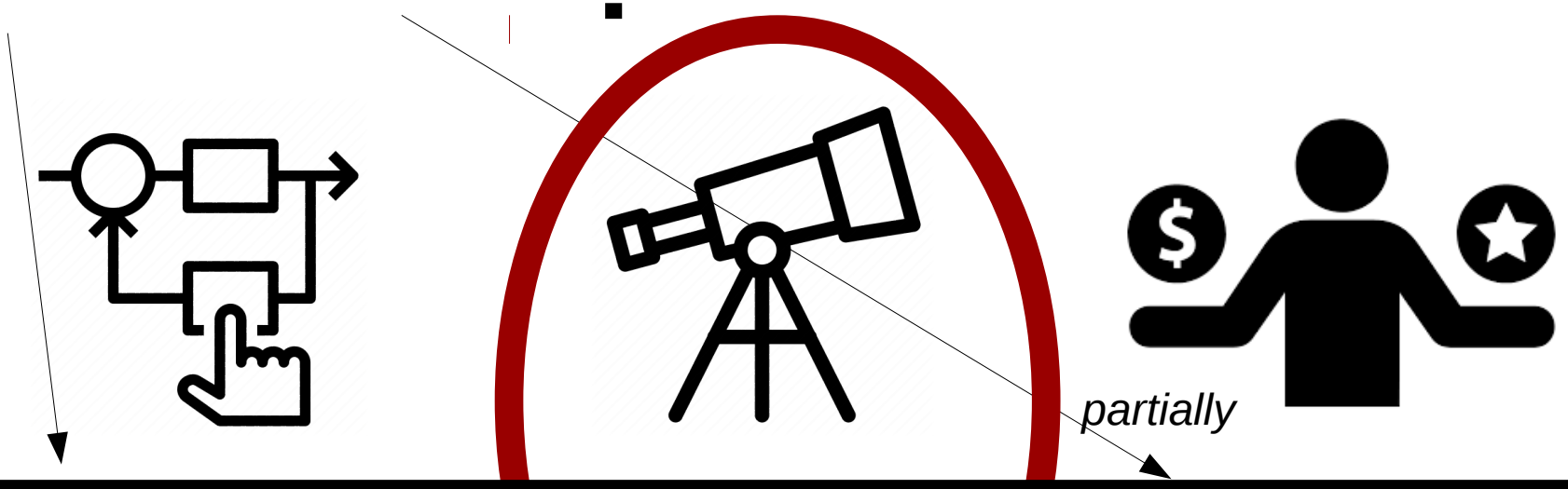
partially

has the ability to **assess their impact according to a preferential or value structure**

What responsibility?

- **Accountability**
- **Transparency**
- **Controllability**

requires knowledge constructed from system practices, and users are rarely in position to know



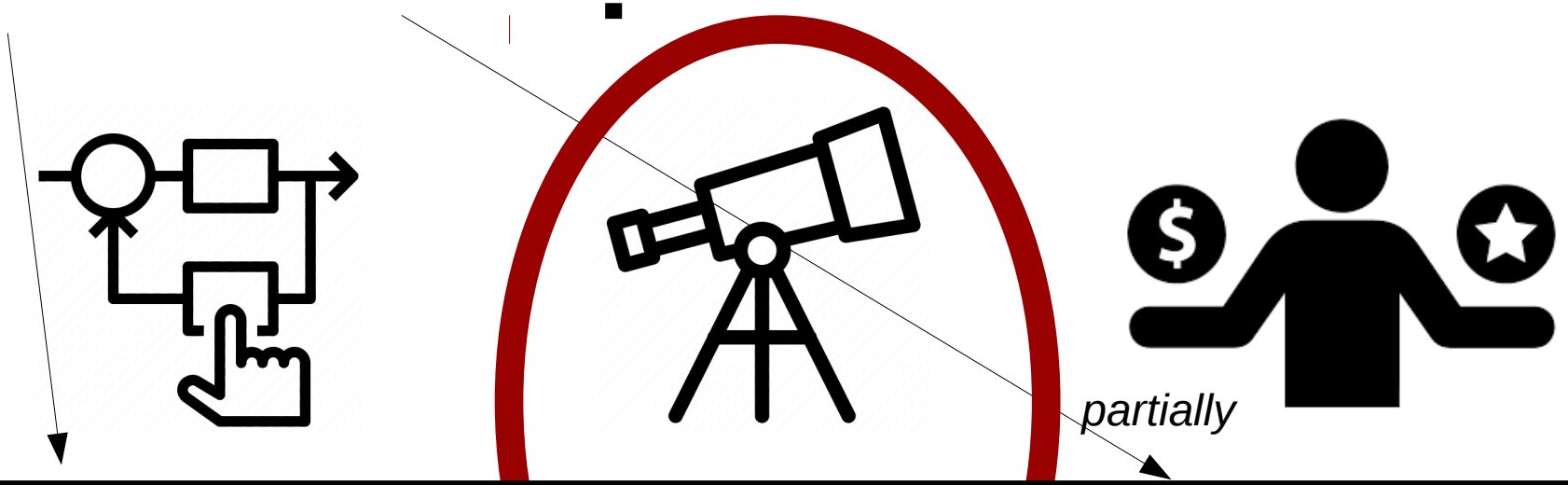
Only if all three components are enabled, solutions as differential privacy and algorithmic fairness can be used “responsibly”.

What responsibility?

cf. ***cooperative responsibility*** (not only individuals and regulators, but also intermediate civil actors) [Helberger, 2018]

- **Accountability**
- **Transparency**
- **Controllability**

requires knowledge constructed from system practices, and users are rarely in position to know



Only if all three components are enabled, solutions as differential privacy and algorithmic fairness can be used “responsibly”.

Conclusions

- Concepts as privacy and fairness refer to **context-dependent** and **plural norms** and **expectations** that can **not** be directly translated to ***optimization*** tasks.
  - Not all bias is unfair, it depends on how it is used and for what.
  - Not all disclosure is illicit; in fact, some might be beneficial to the data subject and to society.

Conclusions: key message

- To achieve **responsible computing**,
  - **computation** needs to be looked at in **distributed** terms (including ***human factors*** going with it), and

Conclusions: key message

- To achieve **responsible computing**,
  - **computation** needs to be looked at in **distributed** terms (including ***human factors*** going with it), and
  - **computational agents** need to be furnished with some degree of **autonomy**

Conclusions: key message

- To achieve **responsible computing**,
  - **computation** needs to be looked at in **distributed** terms (including ***human factors*** going with it), and
  - **computational agents** need to be furnished with some degree of **autonomy**
    - to be able to assess independently, on the basis of
      - (plural) **directives** given **by humans** and
      - (plural) **knowledge** constructed **from system practices**,

Conclusions: key message

- To achieve **responsible computing**,
  - **computation** needs to be looked at in **distributed** terms (including ***human factors*** going with it), and
  - **computational agents** need to be furnished with some degree of **autonomy**
    - to be able to assess independently, on the basis of
      - (plural) **directives** given **by humans** and
      - (plural) **knowledge** constructed **from system practices**,

    whether a certain requested processing is indeed ***justified***.

Like Circles in the Water: Responsibility as a System-Level Function

XAILA Workshop on eXplainable and responsible AI and Law, joint with JURIX 2020
9 December 2020 @ Brno/Prague (virtual)

Giovanni Sileno[a] (g.sileno@uva.nl)
Alexander Boer[b], Geoff Gordon[a], Bernhard Rieder[a]

[a] Informatics Institute, University of Amsterdam, the Netherlands
[b] KPMG, Amsterdam, the Netherlands