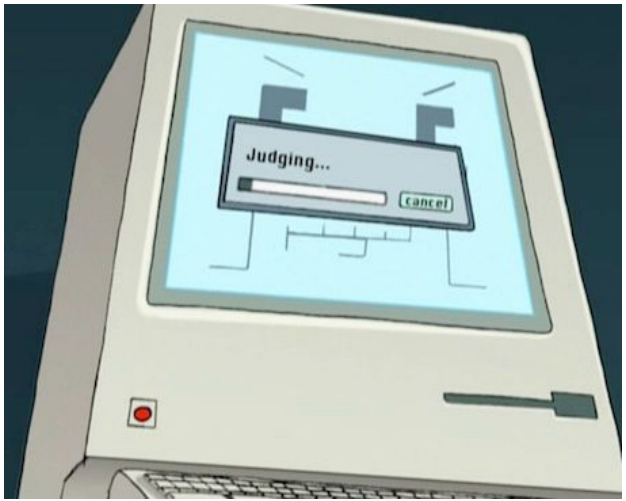
A Fairness-by-design framework for data-driven Automated Decision-making Systems

Claudio Lazo

Supervisors:
Dr. Christoph Lofi,
Agathe Balayn, MSc.

Web Information Systems - EEMCS Faculty

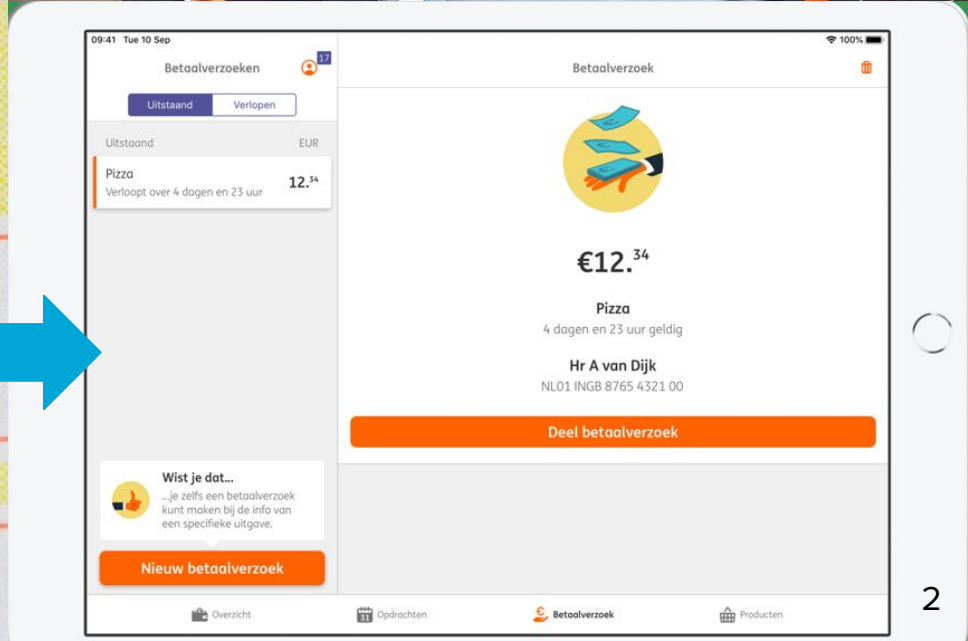


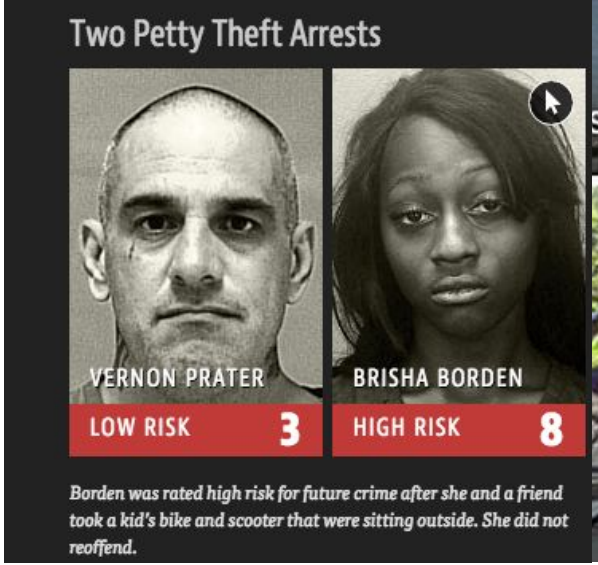
TU Delft


RECEIVING
TELLER


De lekkerste
ballen
van Utrecht
4 stuks 5.00
ING


Acceptgiro
over te schrijven
404
euro
van bankrekening (IBAN)
NL28INGB0009094394


09:41 Tue 10 Sep
100%
Betaalverzoeken
Uitstaand
Verlopen
Uitstaand
EUR
Pizza
Verloopt over 4 dagen en 23 uur
12.34
Wist je dat...
...je zelfs een betaalverzoek kunt maken bij de info van een specifieke uitgave.
Nieuw betaalverzoek
Betaalverzoek
€12.34
Pizza
4 dagen en 23 uur geldig
Hr A van Dijk
NL01 INGB 8765 4321 00
Deel betaalverzoek
Overzicht
Opdrachten
Betaalverzoek
Producten


Two Petty Theft Arrests
VERNON PRATER
LOW RISK 3
BRISHA BORDEN
HIGH RISK 8
Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.


Skyscrapers


Airplanes


Cars


Bikes


Gorillas


Graduation


CREDIT RISK

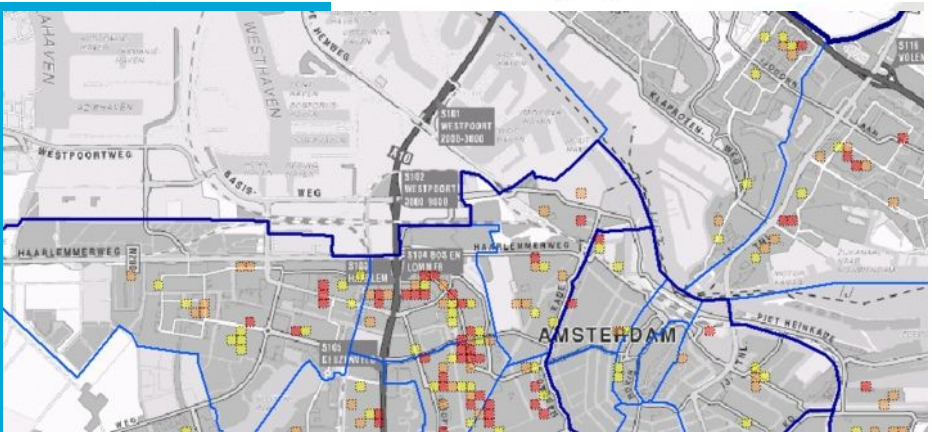
deVolkskrant

Voorpagina | Nieuws & Achtergrond | Columns & Opinie | Video | Weten

NIEUWS SYRI

SyRI, het fraudesysteem van de overheid, faalt: nog niet één fraudegeval opgespoord

Een omstreden systeem waarbij de overheid persoonsgegevens koppelt om fraude met uitkeringen, toeslagen en belastingen op te sporen, blijkt nauwelijks te werken. De kritiek op 'SyRi' neemt



Belastingdienst erkent: fraudesysteem deugde niet

AI ethical guidelines

Auditability

Human autonomy

Diversity

Fairness

Explainability

Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice & fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom & autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

Jobin et al, 2019

AI ethical guidelines

Auditability

Human autonomy

Diversity

Fairness

Explainability

?

Trade-offs

	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓
3.3.2	Well calibration	[16]	81	✓
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–
5.2	No unresolved discrimination	[15]	14	–
5.3	No proxy discrimination	[15]	14	–
5.4	Fair inference	[19]	6	–

Table 1: Considered Definitions of Fairness

Bias mitigation

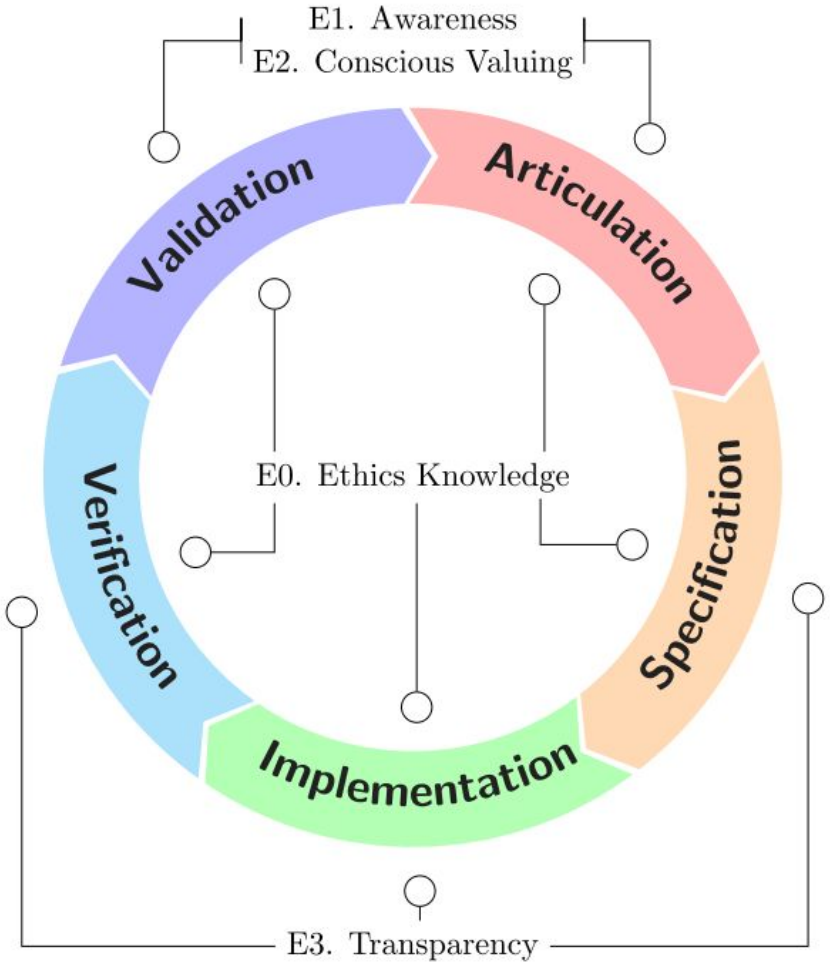
Fairness measures

Representation

Control

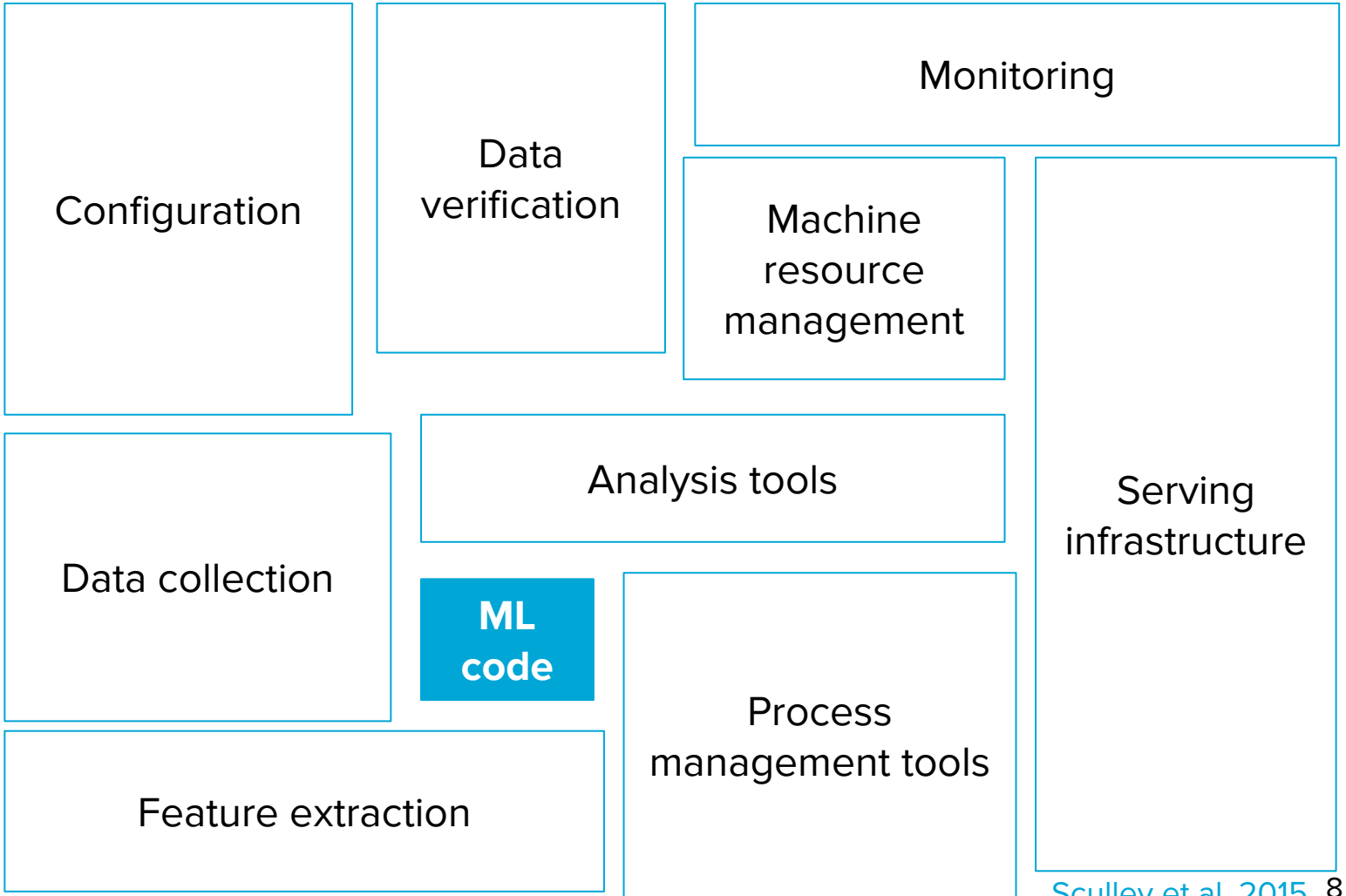
Justification

Due process

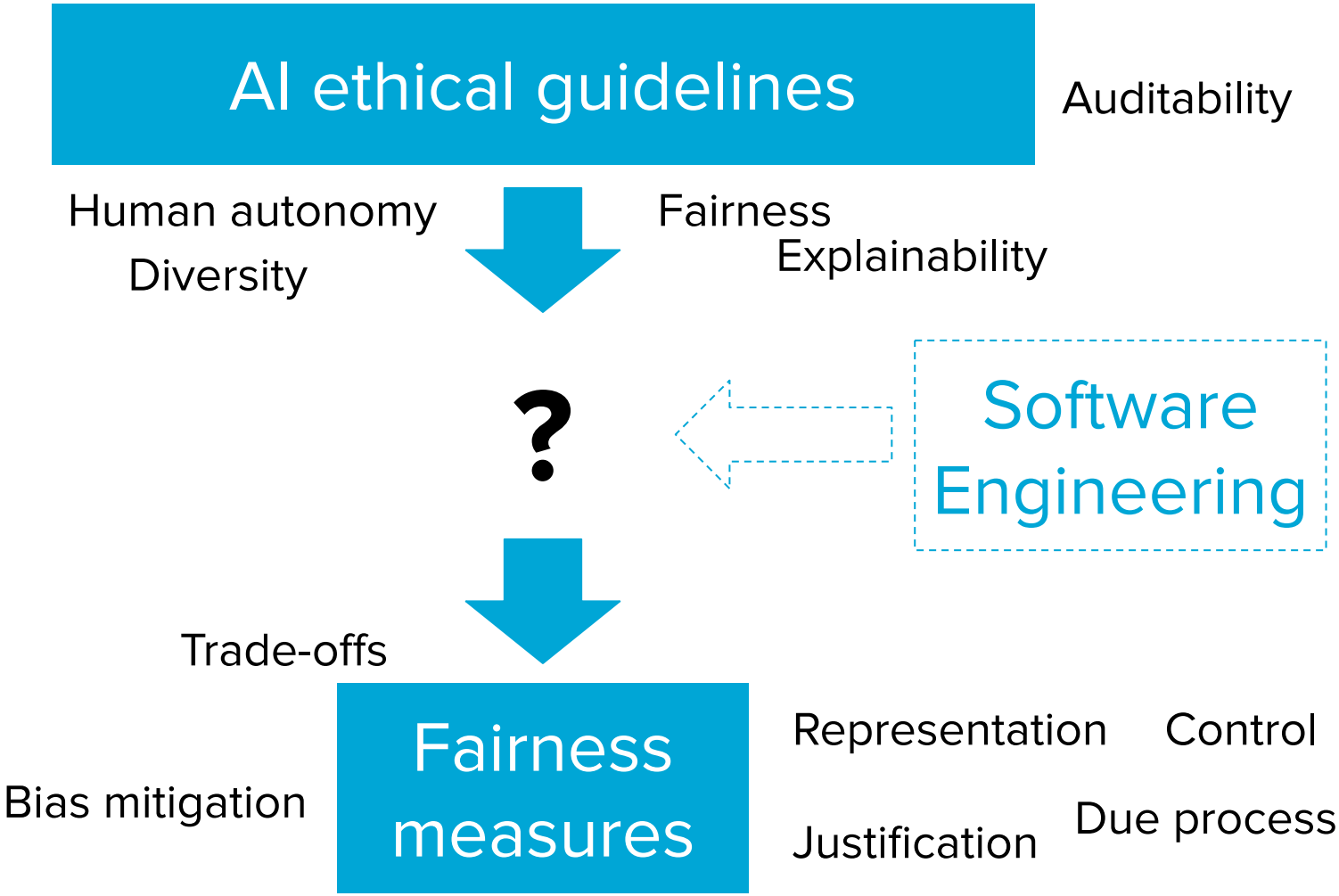


Aydemir & Dalpiaz, 2018

- Requirement Elicitation
- Requirement Analysis
- Requirement Specification
- Requirement Validation



Sculley et al, 2015


AI ethical guidelines
Auditability
Human autonomy
Diversity
Fairness
Explainability
?
Software Engineering
Trade-offs
Fairness measures
Bias mitigation
Representation
Control
Justification
Due process

Research goal

To help practitioners make fair-by-design software systems,

by guiding them through the Requirements Engineering phase,

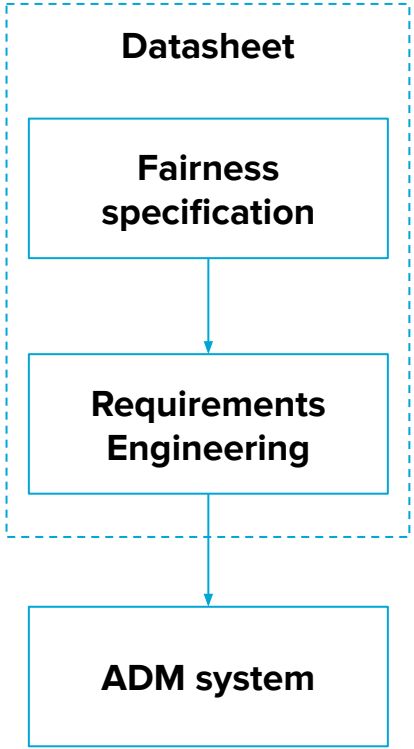
- Eliciting context-specific fairness requirements,
- From an integrative fairness specification,
- Implementable in a data-driven ADM system;

And making design decisions explicit and auditable.

TU Delft

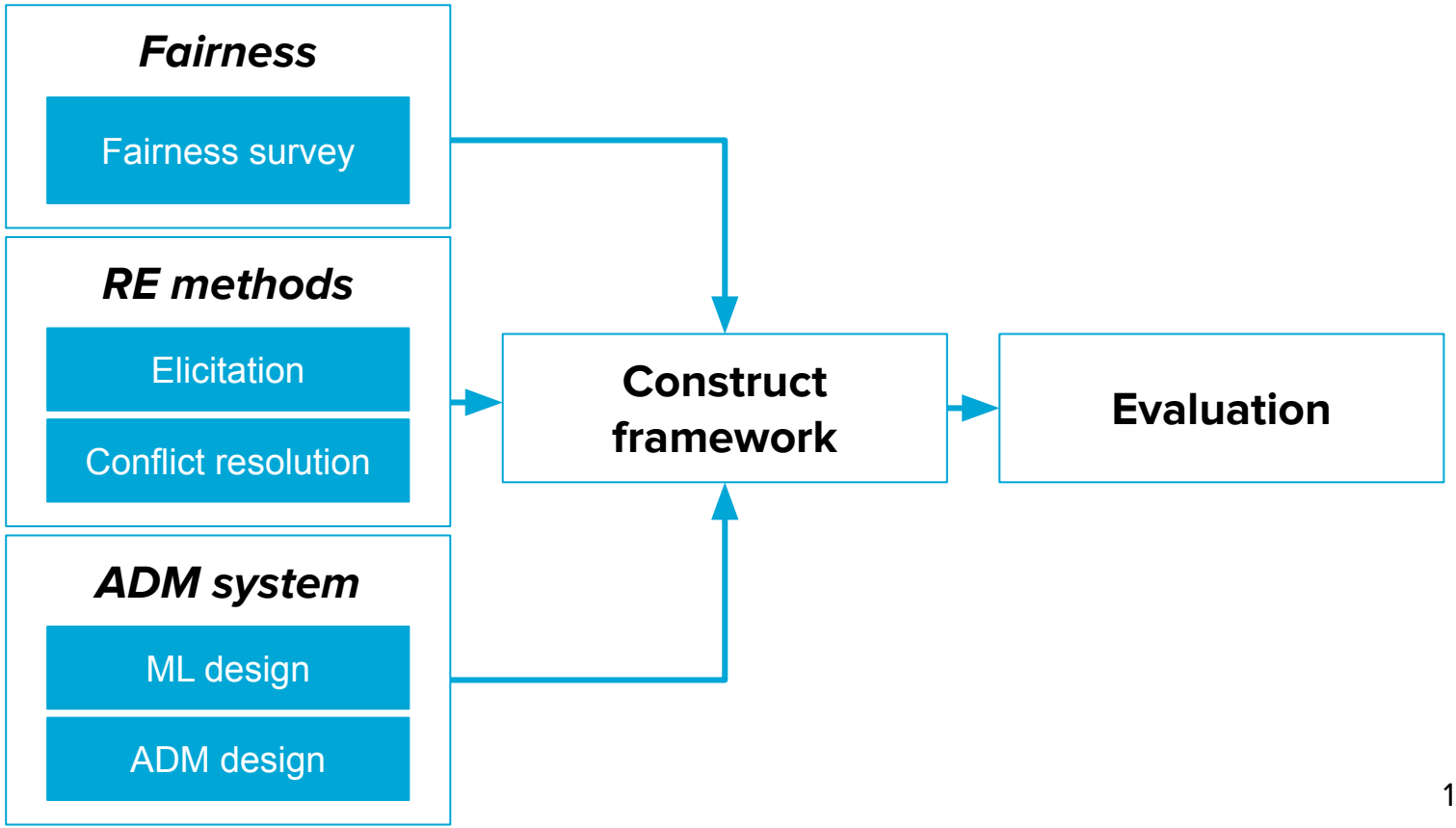
Contributions

1. Integrative fairness specification
2. Ethical Requirements Engineering for data-driven ADM
3. Formalisation of data-driven ADM
4. ADM datasheet



TU Delft

Thesis overview



Research phase 1

Understanding fairness

Finding a fairness specification

Field	What makes a decision fair?
Economics	Optimal resource distribution/social *welfare* function
Game theory	Envy-freeness, Shapley value & Rabin-fairness
(Political) philosophy	Fairness found in universal rules (Veil of Ignorance)
(Business) ethics	Moral standards for decisions that affect others
Psychology	Perception and judgment
Law	Non-discrimination, Equality, right to *due process*
Social justice	Non-discrimination, equal social opportunity
Organizational justice	Fairness judgments of managerial decisions

Integrative fairness specification

Fairness			
Distributive	**Procedural**	**Informational**	**Interpersonal**
Distributive norm	Control	Adequacy of explanation	Respect
Preferential treatments	Consistency	Truthfulness	Judgment
	Bias suppression		Privacy
	Accuracy		Dangers
	Correctability		
	Representation		
	Ethicality		

Fair-ML community

Fairness measures in ML

- Statistical measures
  - Statistical parity
  - False negative error rate
- Economical measures
  - Inequality indices
- Similarity-based
  - Individual fairness
- Causal reasoning
  - No proxy discrimination

Bias suppression

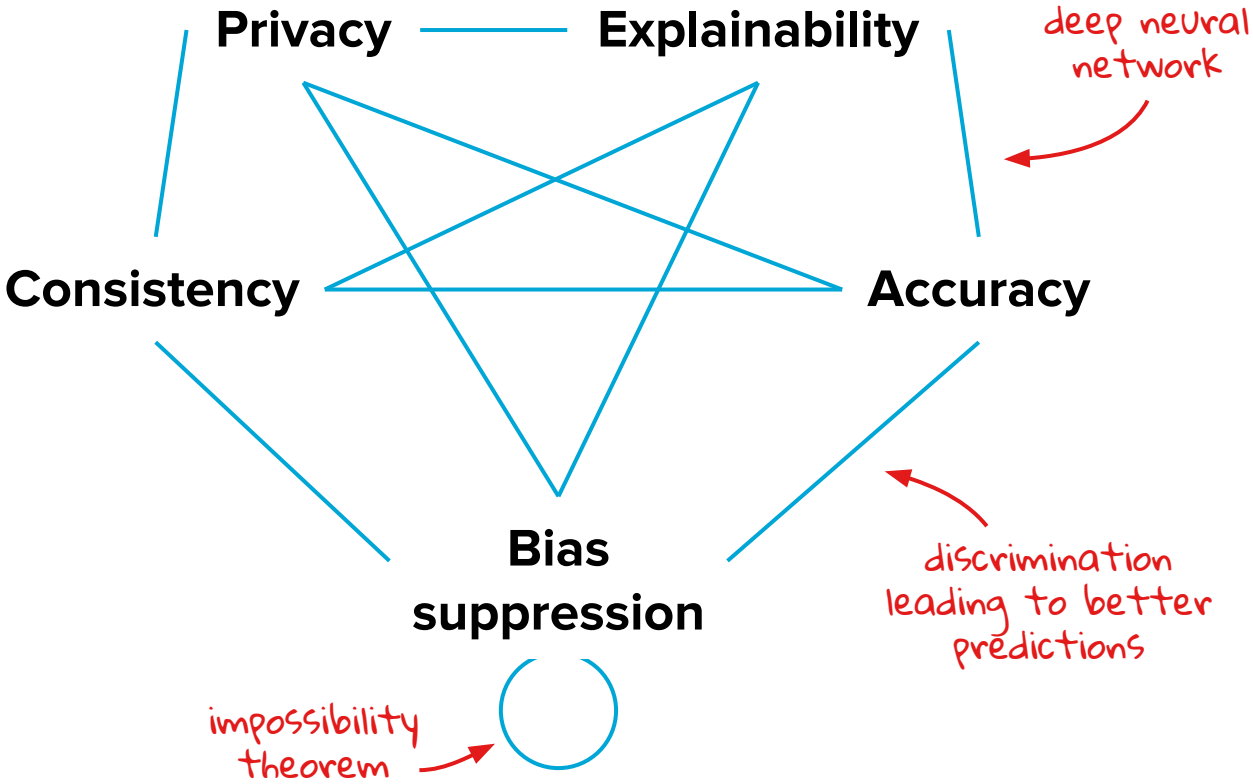
Consistency

Model validity

Verma & Rubin, 2018
Speicher et al, 2018

TU Delft

Trade-offs

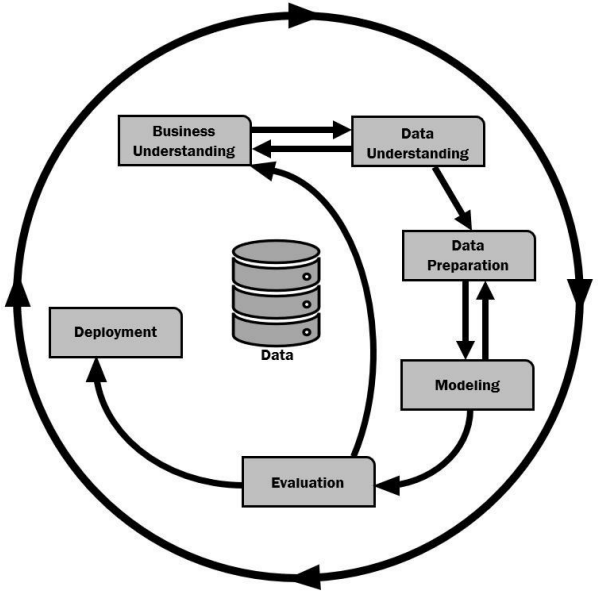

Privacy
Explainability
deep neural network
Consistency
Accuracy
Bias suppression
discrimination leading to better predictions
impossibility theorem

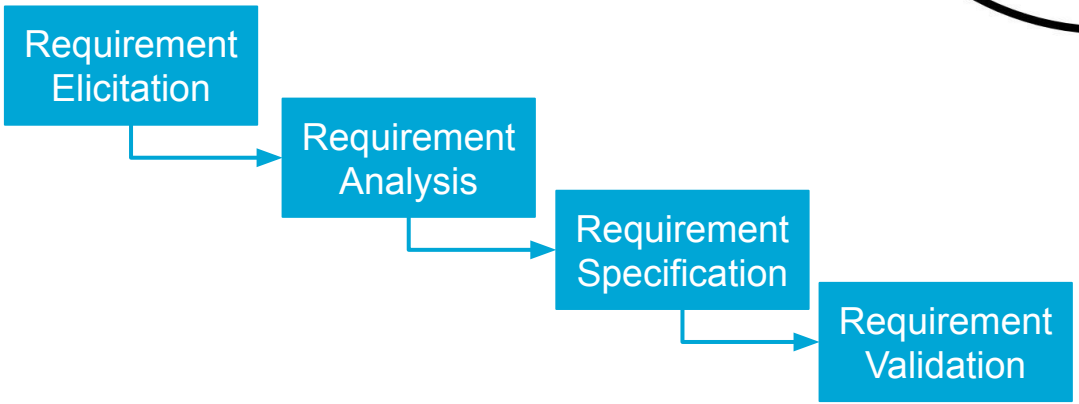
Research phase 2

The framework

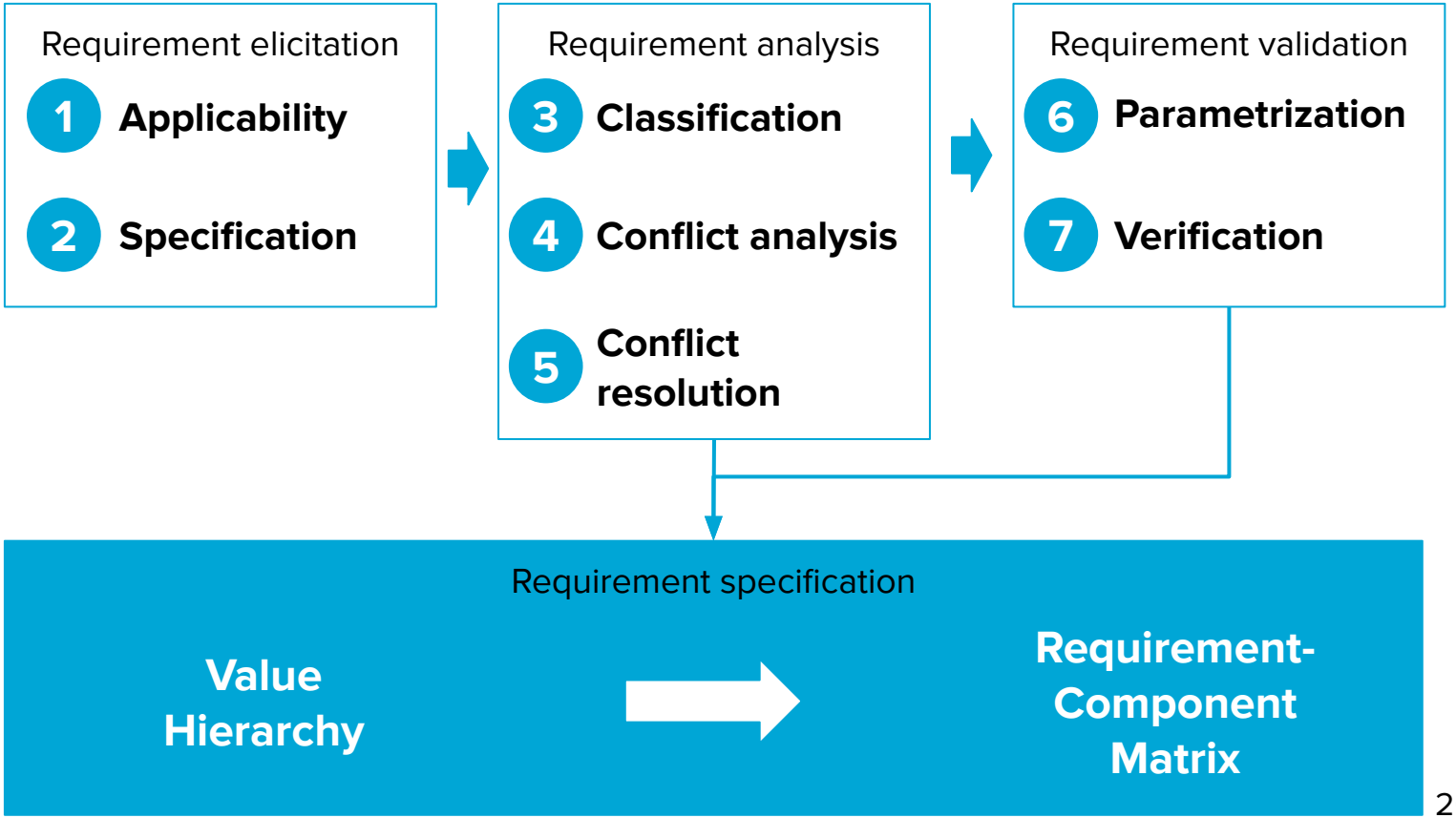
Framework

- Semi-iterative
- Guiding process
- Canvas
- Integrated in RE and DM





Framework

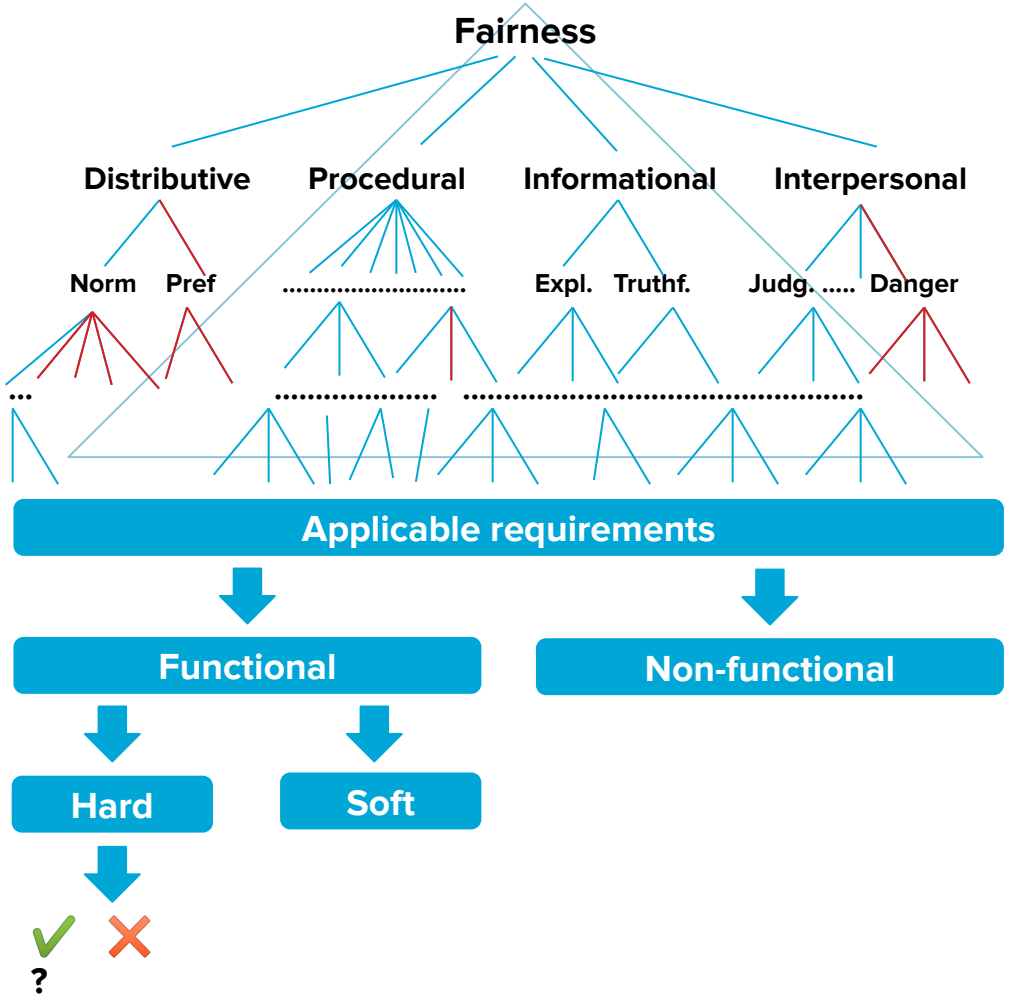

Requirement elicitation
1 Applicability
2 Specification
Requirement analysis
3 Classification
4 Conflict analysis
5 Conflict resolution
Requirement validation
6 Parametrization
7 Verification
Requirement specification
Value Hierarchy
Requirement-Component Matrix

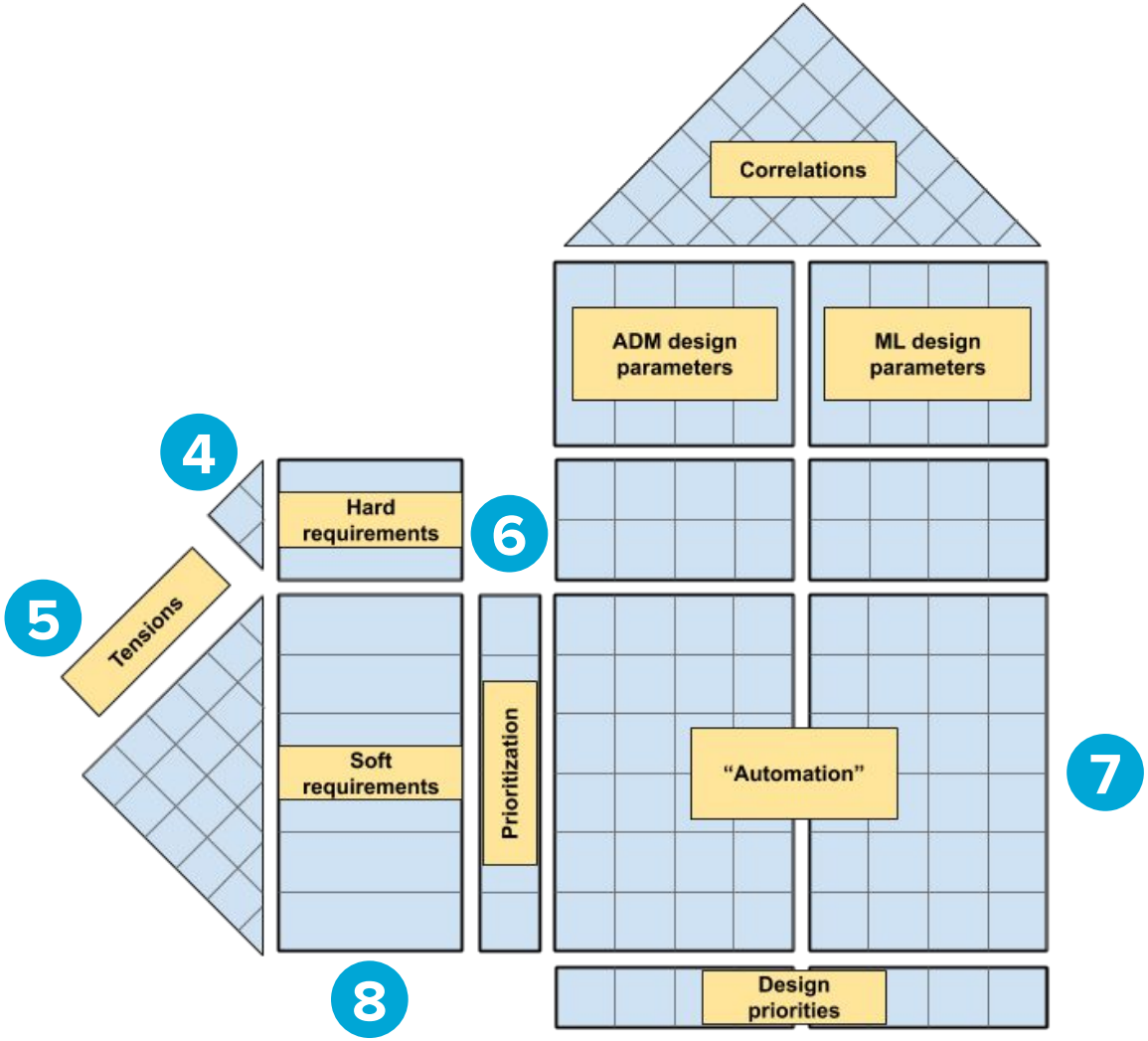
1 Applicability

2 Specification

3 Classification

4 Feasibility study




Correlations
ADM design parameters
ML design parameters
4
Hard requirements
6
5
Tensions
Soft requirements
Prioritization
"Automation"
7
8
Design priorities

Research phase 3

Evaluation

Case study

- Participants:
  - One or more dev team members
- Pre-questionnaire
- Activity with framework
  - Real life case
  - Illustrative scenario
  - Examine previous case
- Post-questionnaire
- Interview

TU Delft

ING collaboration

- Fair-by-design?
- Interesting use cases?
- Participants?
- Assess the framework?
- By april 2020