Fairness, Explainability and Robustness in Automated Decision-making

*Survey of systems*

### ABSTRACT

In this research piece we outline three criteria to assess an automated decision-making algorithm: Fairness, Explainability and Robustness (REF). We present each criterion from a conceptual and mathematical standpoint in each of the following sections. Every section starts with the conceptual debate, providing a non-mathematical definition, case studies and current debate. After that, we move to a technical definition of each criterion, closing this with the main techniques and technologies available to assess and avert the lack of REF. In striking this balance, we want to inform readers with different backgrounds and interests: those interested in REF issues and debate, and those looking for the technical tools to improve REF in automated decision-making.

|  |  |  |  |
| --- | --- | --- | --- |
| View \\ Criterion | Fairness | Explainability | Robustness |
| Core principles | Universal Declaration of Human rights | The Enlightenment/ Age of Reason | Scientific Method |
| Law, code of conduct and rules | Equal opportunities act, Civil Rights Law, etc. | GDPR “right to explain” | Consumer rights |
| Abstract, technical or mathematical description | Group level fairness, demographic parity, individual fairness, counterfactual fairness | counterfactual analysis, global and local, model-specific and agnostic explanations | Generalization performance  Adversarial robustness |
| Implementation and technological solutions | Fairness regularizers/constraints, fair empirical risk minimization, causal modelling, Accuracy-Fairness frontiers | Feature importance metrics, weights and coefficients, partial dependency charts, etc. | CV, stress-testing, adversarial training, |
| Regulatory compliance | Group level metrics | Reports, systems in place to provide real time explanations, etc. | Metrics, real time monitoring |

|  |  |  |  |
| --- | --- | --- | --- |
| View \\ Criterion | Fairness | Explainability | Robustness |
| Philosophical principle or movement | Universal Declaration of Human rights | The Enlightenment/ Age of Reason | Scientific Method |
| Law, code of conduct and rules | Equal opportunities act, Civil Rights Law, etc. | GDPR “right to explain” | Consumer rights |
| Abstraction, technical or mathematical description | Group level fairness, demographic parity, individual fairness, counterfactual fairness | counterfactual analysis, global and local, model-specific and agnostic explanations | Generalization performance  Adversarial robustness |
| Implementation, Code/Execution | Fairness regularizers/constraints, fair empirical risk minimization, causal modelling, Accuracy-Fairness frontiers | Feature importance metrics, weights and coefficients, partial dependency charts, etc. | CV, stress-testing, adversarial training, |
| Regulation, Auditing, Enforceability | Group level metrics | Reports, systems in place to provide real time explanations, etc. | Metrics, real time monitoring |

**Contents**

[ABSTRACT 1](#_Toc3456773)

[1. Fairness 3](#_Toc3456774)

[2.1 Core principles 3](#_Toc3456775)

[2.2 Law, code of conduct and rules 3](#_Toc3456776)

[2.3 Abstract, technical or mathematical description 3](#_Toc3456777)

[2.4 Implementation and technological solutions 3](#_Toc3456778)

[2.5 Regulatory compliance 3](#_Toc3456779)

[References 3](#_Toc3456780)

[2. Explainability 3](#_Toc3456781)

[2.1 Conceptual definition 3](#_Toc3456782)

[2.2 Case studies and current debate 4](#_Toc3456783)

[2.3 Technical definition 4](#_Toc3456784)

[2.4 Techniques and technologies 4](#_Toc3456785)

[2.5 References 4](#_Toc3456786)

[3. Robustness 4](#_Toc3456787)

[3.1 Conceptual definition 4](#_Toc3456788)

[3.2 Case studies and current debate 4](#_Toc3456789)

[3.3 Technical definition 4](#_Toc3456790)

[3.4 Techniques and technologies 4](#_Toc3456791)

[3.5 References 5](#_Toc3456792)

[4. BM Canvas // Factsheet for REF 5](#_Toc3456793)

[Acknowledgements 6](#_Toc3456794)

# Fairness

## Core principles

## Law, code of conduct and rules

## Abstract, technical or mathematical description

## Implementation and technological solutions

## Regulatory compliance

## References

# Explainability

## Conceptual definition

Principle, rules/law, abstraction/technical/mathematical, code/execution, regulation/auditing

In epistemology, transparency is a property of epistemic states defined as follows: An epistemic state E is "weakly transparent" to a subject S if and only if when S is in state E, S can know that S is in state E; an epistemic state E is "strongly transparent" to a subject S if and only if when S is in state E, S can know that S is in state E, AND when S is not in state E, S can know S is not in state E.

Pain is usually considered to be strongly transparent: when someone is in pain, he knows immediately that he is in pain, and if he is not in pain, he will know he is not. Transparency is important in the study of self-knowledge and meta-knowledge.

Transparency as an ideal

Transparency, as an ideal, can be traced through many histories of practice. From philosophers concerned with the epistemological production of truth, through activists striving for government accountability, transparency has offered a way to see inside the truth of a system.

Transparency concerns are commonly driven by a certain chain of logic: observation produces insights which create the knowledge required to govern and hold systems accountable. This logic rests on an epistemological assumption that “truth is correspondence to, or with, a fact” (David, 2015: n.p.). The more facts revealed, the more truth that can be known through a logic of accumulation. Observation is understood as a diagnostic for ethical action, as observers with more access to the facts describing a system will be better able to judge whether a system is working as intended and what changes are required. The more that is known about a system’s inner workings, the more defensibly it can be governed and held accountable.

This chain of logic entails “a rejection of established representations” in order to realize “a dream of moving outside representation understood as bias and distortion” toward “representations [that] are more intrinsically true than others.” It lets observers “uncover the true essence” of a system (Christensen and Cheney, 2015: 77). The hope to “uncover” a singular truth was a hallmark of The Enlightenment, part of what Daston (1992: 607) calls the attempt to escape the idiosyncrasies of perspective: a “transcendence of individual viewpoints in deliberation and action [that] seemed a precondition for a just and harmonious society.”

Several historians point to early Enlightenment practices around scientific evidence and social engineering as sites where transparency first emerged in a modern form (Crary, 1990; Daston, 1992; Daston and Galison, 2007; Hood, 2006).

## Case studies and current debate

## Technical definition

## Techniques and technologies

## References

# Robustness

## Conceptual definition

## Case studies and current debate

## Technical definition

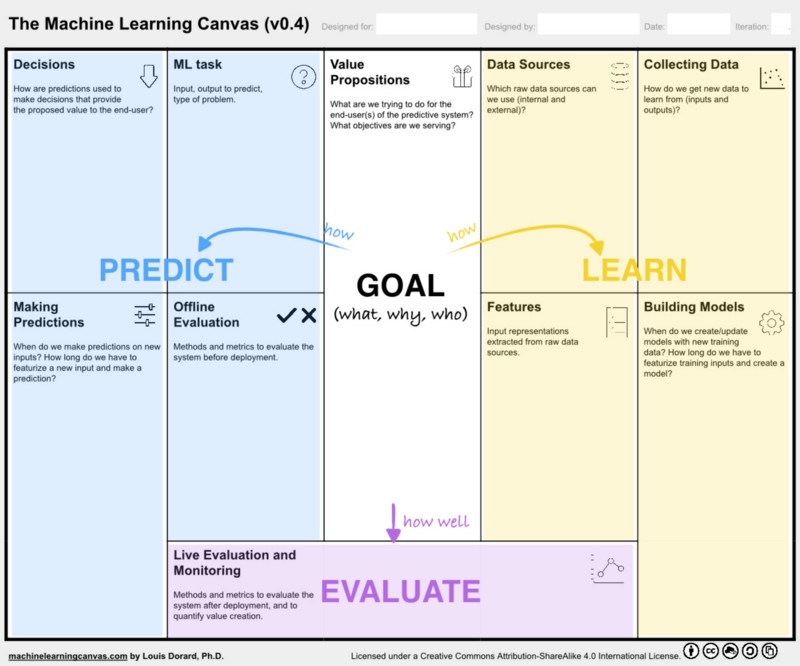
Generalization performance

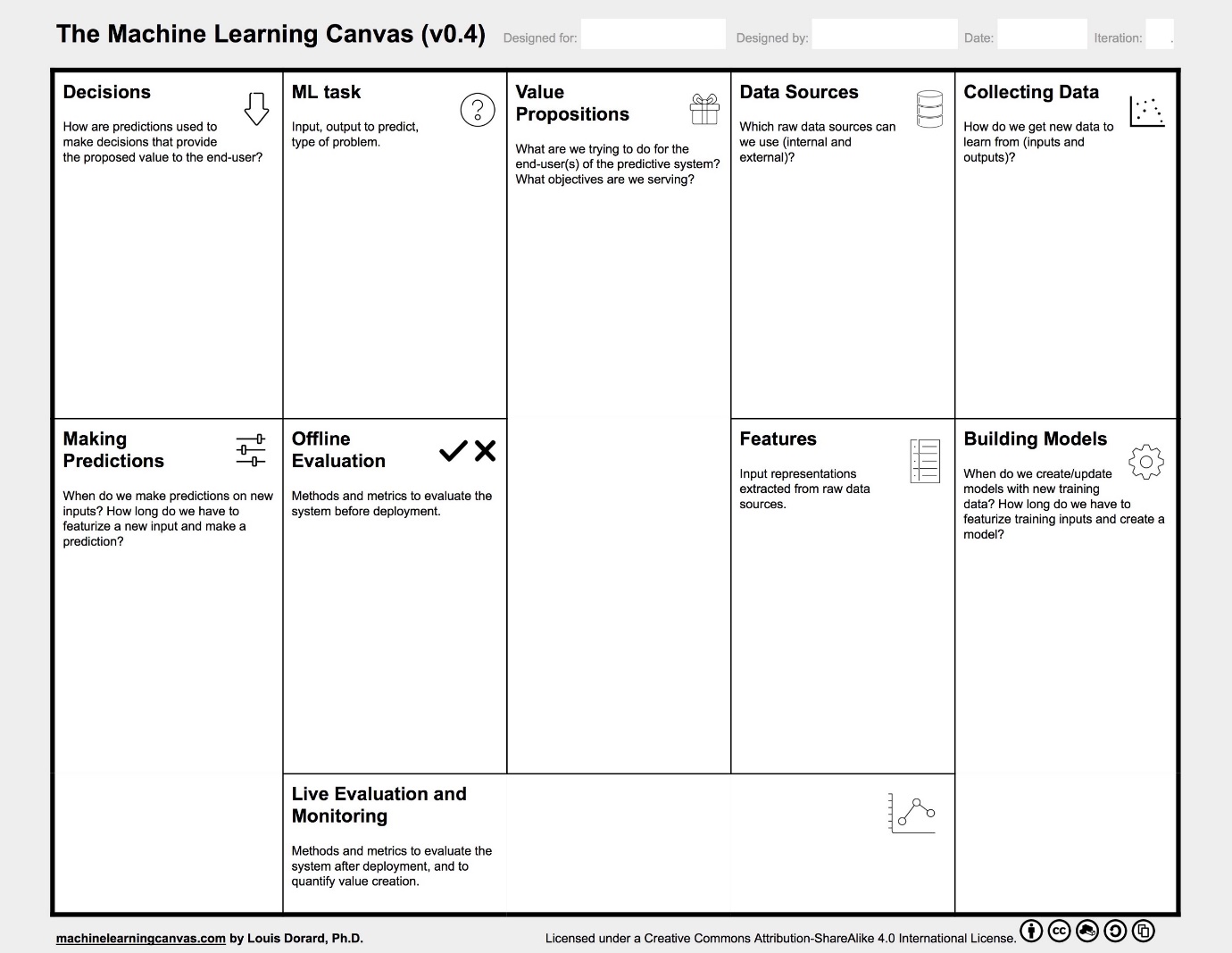
Adversarial robustness

## Techniques and technologies

## References

# BM Canvas // Factsheet for REF





# Acknowledgements

The authors would like to thank the Reviewers for their comments and ideas that have helped us reshape and improve our paper.