RSAConference2020

San Francisco | February 24 – 28 | Moscone Center

HUMAN ELEMENT

SESSION ID: **LAW-W11**

The Past is Not the Future: Practical Ethics for Algorithms

Behnam Dayanim

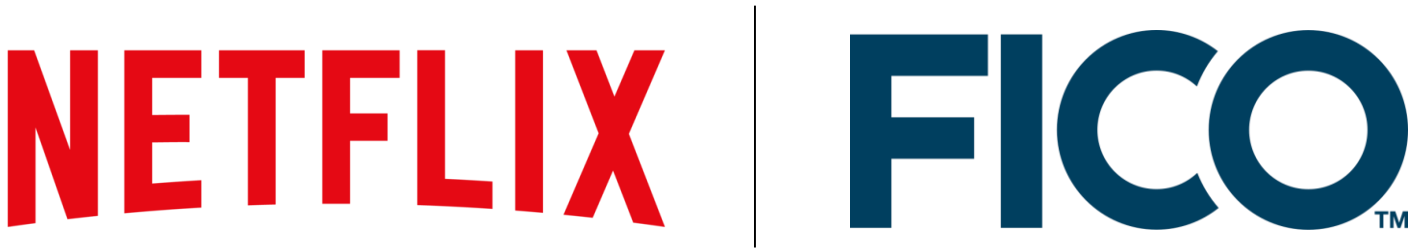
Partner, Global Chair of Privacy & Cybersecurity Practice
Paul Hastings LLP
@BDayanim

Mike Kiser

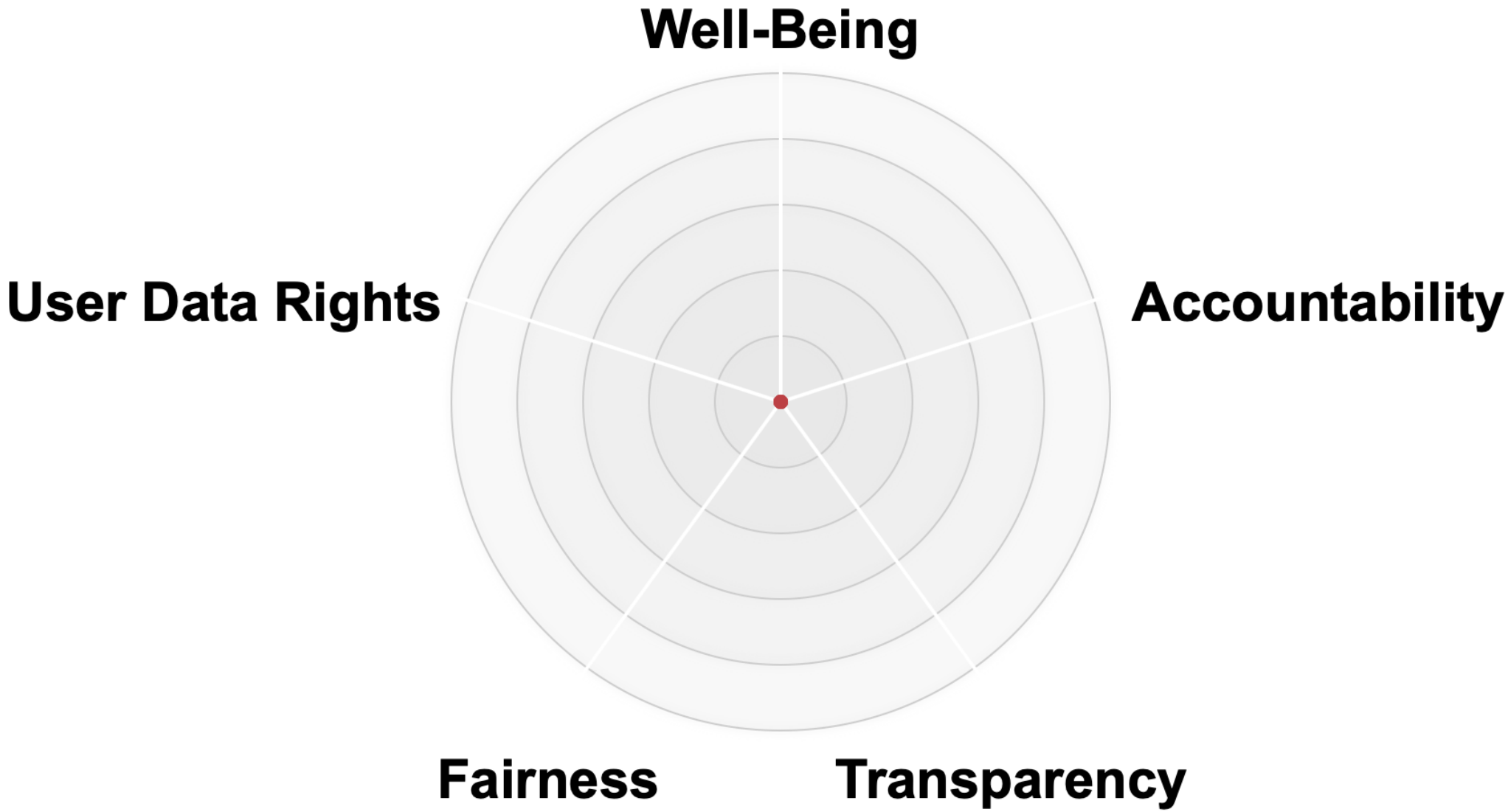
Global Security Advocate, Office of the CTO
SailPoint
@_MikeKiser

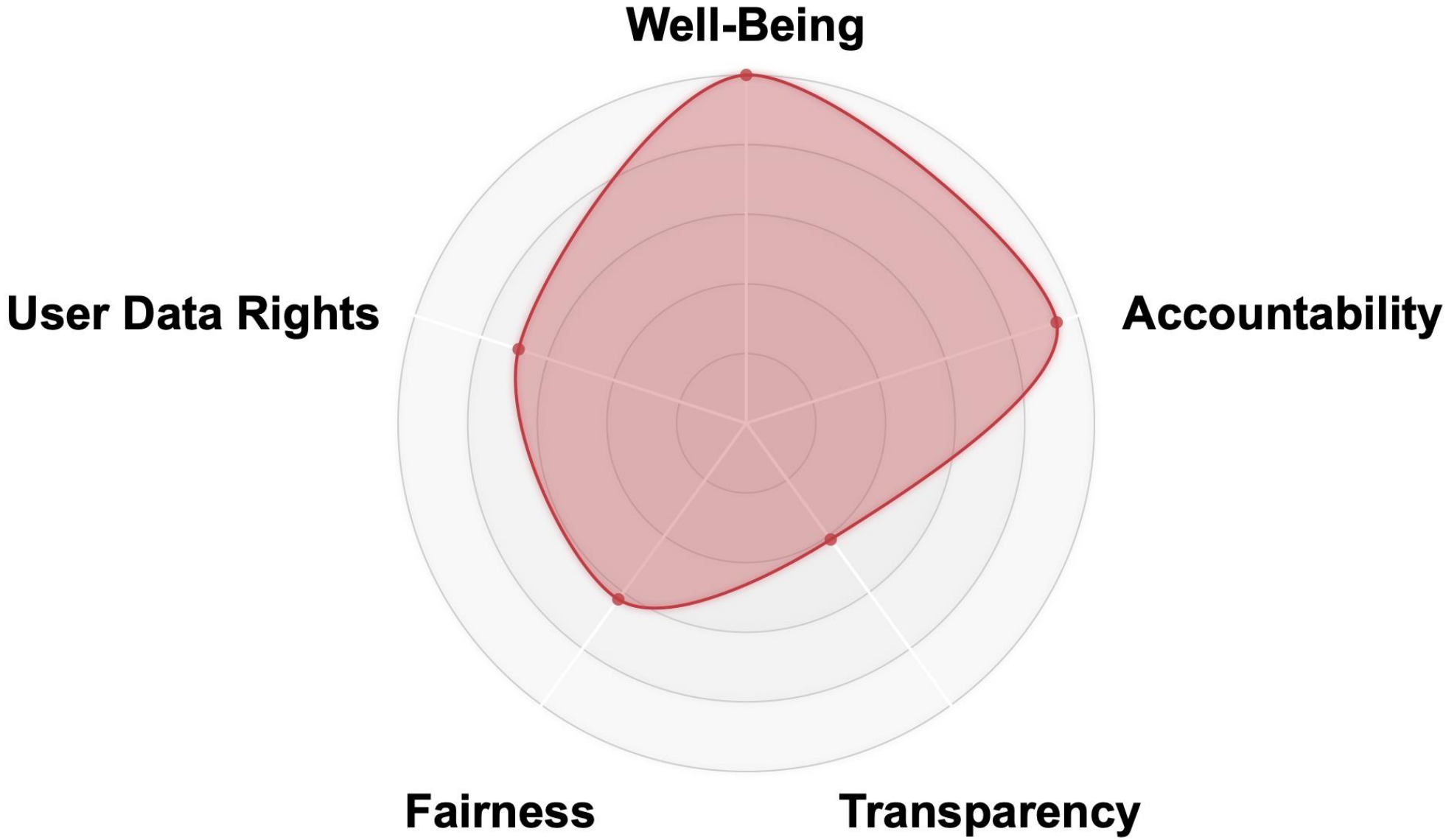

TEMPLE OF JUSTICE

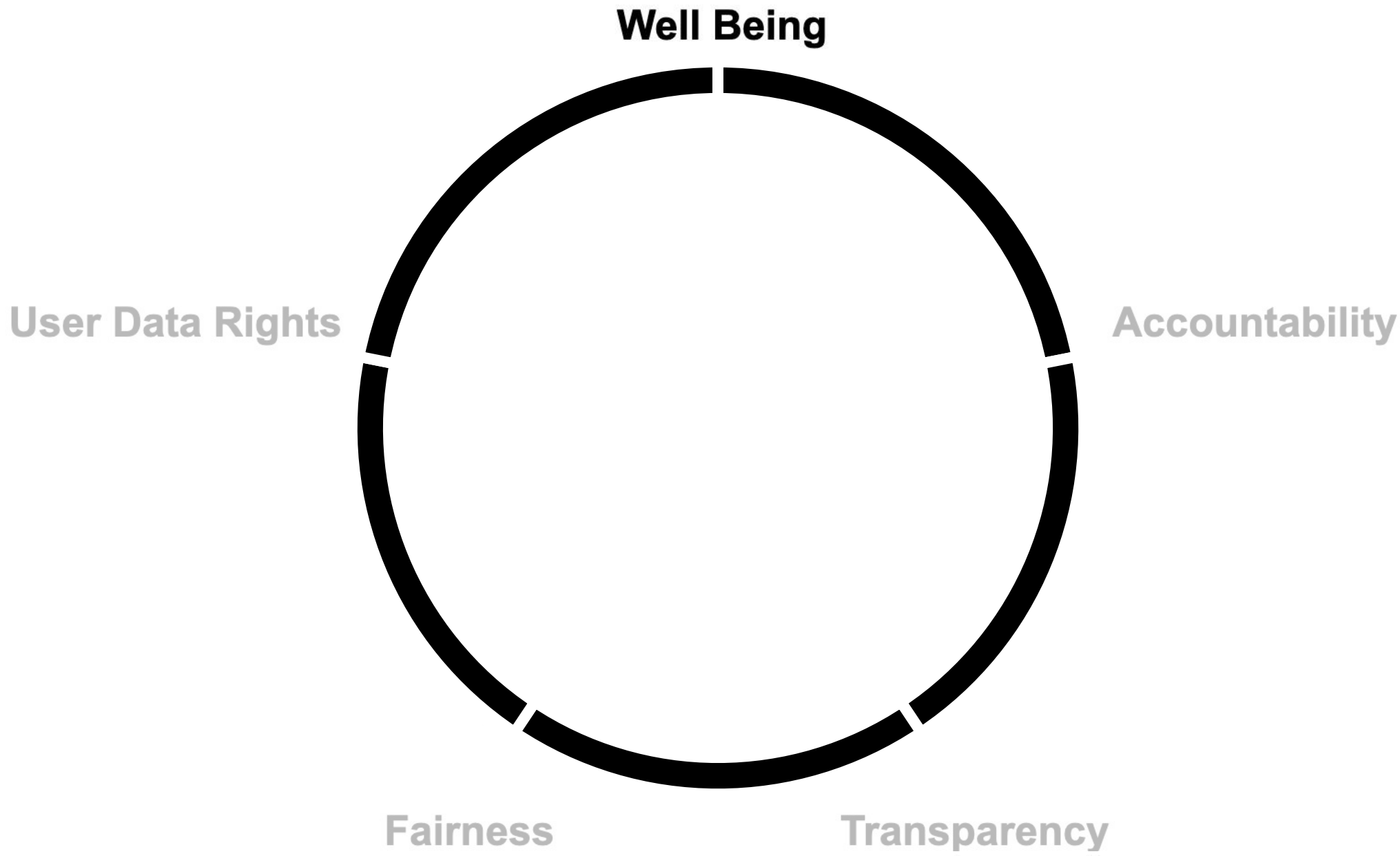

NETFLIX


NETFLIX
FICO
TM


NETFLIX
FICO
Palantir


Well-Being
User Data Rights
Accountability
Fairness
Transparency


Well-Being
User Data Rights
Accountability
Fairness
Transparency


Well Being
User Data Rights
Accountability
Fairness
Transparency

ΙΠΠΟΚΡΑΤΟΥΣ
ΟΡΚΟΣ.

HIPPOCRATIS
IVSIVRANDVM.

΄ΟΜΝΥΜΙ΄ Απόλλωνα ἰητρὸν καὶ Ασκληπιὸν καὶ Υγείαν καὶ Πανά-

ER Apollinem Medicum, & Æsculapium, Hygiamque & Pana-

What is Harm?

ONLINE ETHICS CANVAS

Canvas Title: New canvas

Collaborators (1 active)

Individuals Affected

Add an idea

1

Behaviour

Add an idea

3

Relations

Add an idea

4

What can we do?

Add an idea

9

Worldviews

Add an idea

5

Group Conflicts

Add an idea

6

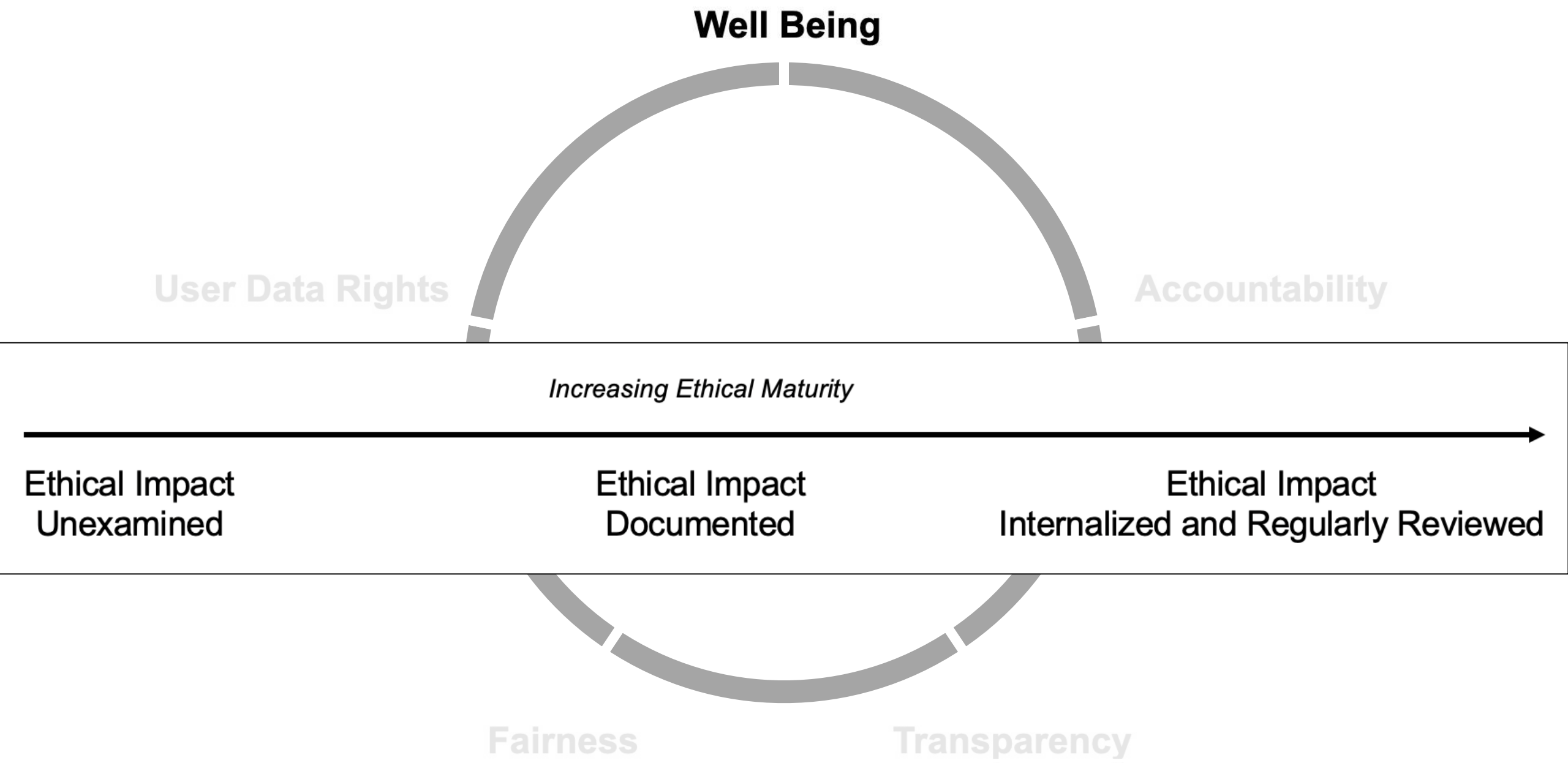
Groups Affected

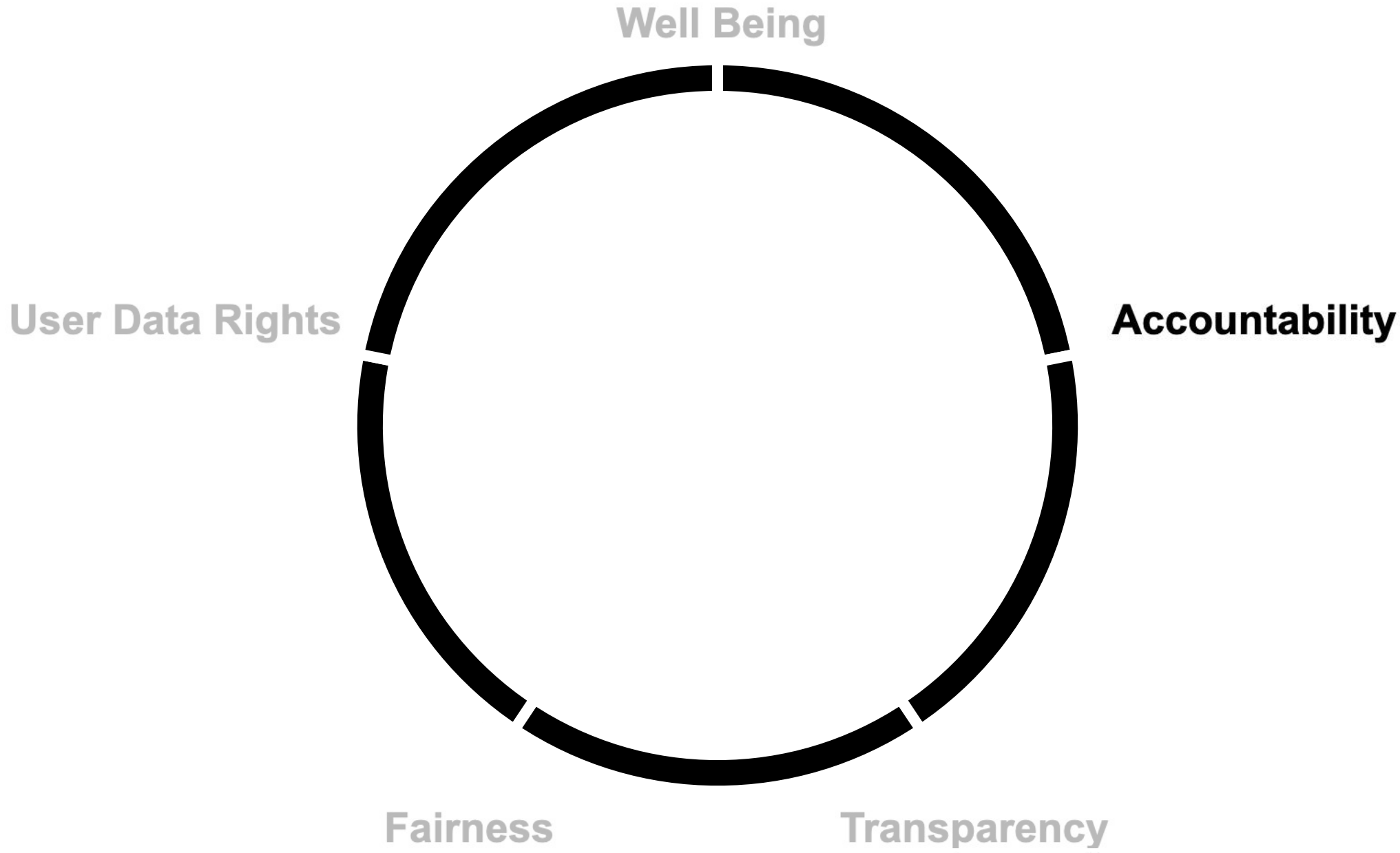
Add an idea

2

Product or Service Failure

Problematic Use of Resources


Well Being
User Data Rights
Accountability
Increasing Ethical Maturity
Ethical Impact
Unexamined
Ethical Impact
Documented
Ethical Impact
Internalized and Regularly Reviewed
Fairness
Transparency


Well Being
User Data Rights
Accountability
Fairness
Transparency

+1-163-555-0168

+1-700-555-5408

+1-415-555-0289

+1-785-555-7537

+1-715-555-0699

+1-505-555-9777

+1-155-555-4727

+1-614-555-0037

+1-755-555-5931

+1-555-867-5309

+1-703-555-2912

1-337-555-5568

1-512-555-4430

+1-775-555-3565

+1-281-555-8201

+1-405-555-3621

+1-925-555-2801

+1-409-555-8132

+1-163-555-0168

+1-700-555-5408

+1-785-555-7537

+1-614-555-0037

+1-755-555-5931

+1-555-867-5309

+1-405-555-3621

+1-925-555-2801

+1-409-555-8132

+1-700-555-5408

+1-755-555-5931

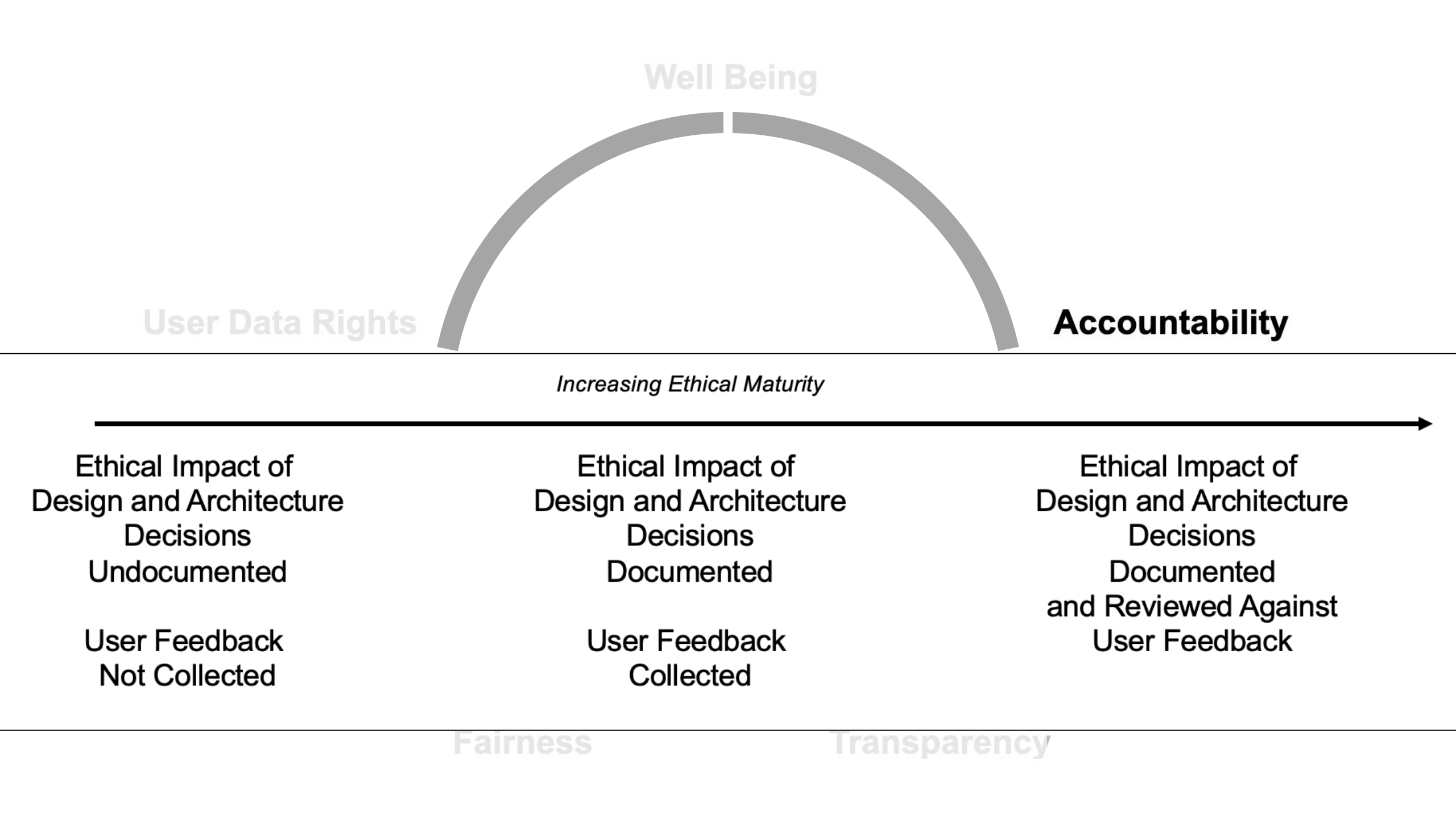
+1-555-867-5309

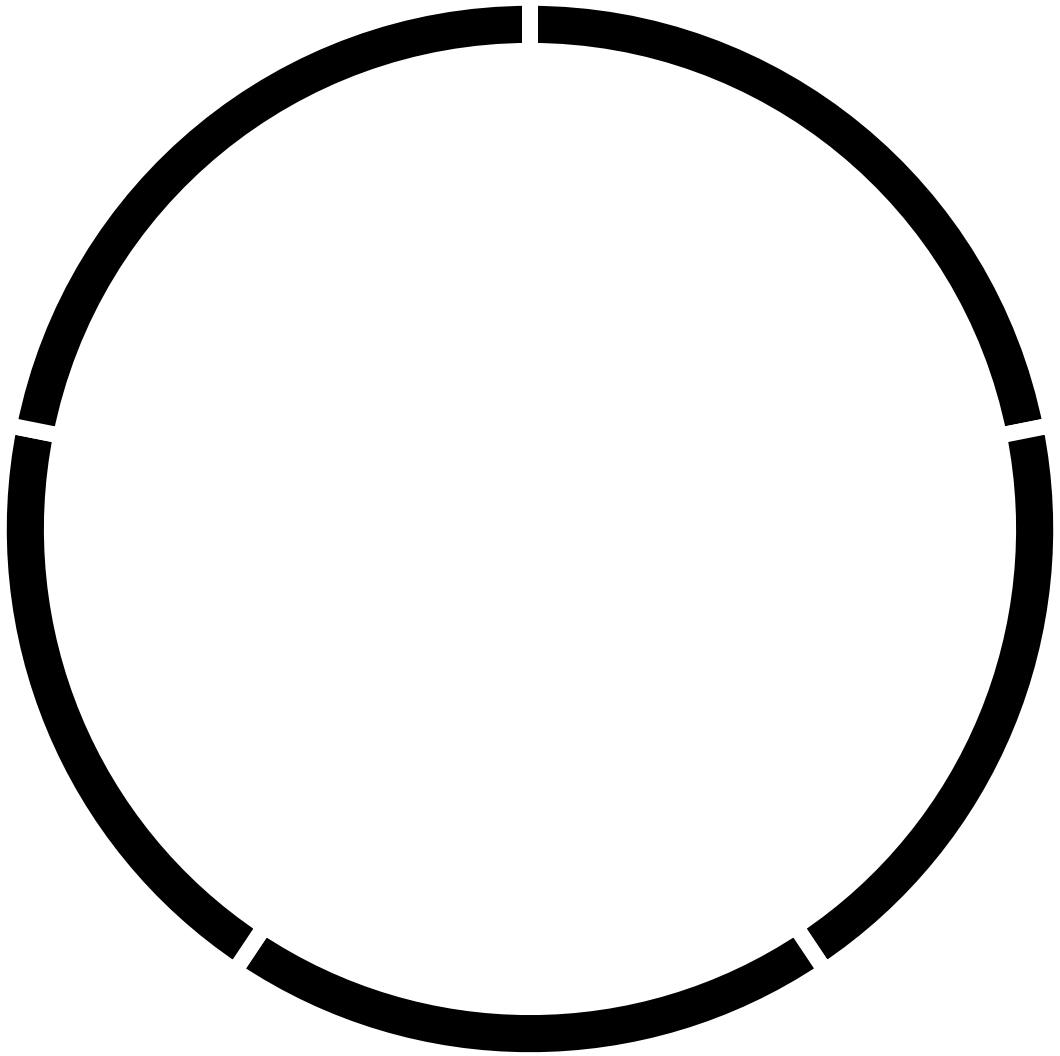
+1-925-555-2801

+1-409-555-8132

+1-555-867-5309


5
4
3
2
1
6
7
8
9
0
DIAL
999
FIRE
POLICE
AMBULANCE
100
OPERATOR


Well Being
User Data Rights
Accountability
Increasing Ethical Maturity
Ethical Impact of Design and Architecture Decisions Undocumented
User Feedback Not Collected
Ethical Impact of Design and Architecture Decisions Documented
User Feedback Collected
Ethical Impact of Design and Architecture Decisions Documented and Reviewed Against User Feedback
Fairness
Transparency


Well Being
User Data Rights
Accountability
Fairness
Transparency

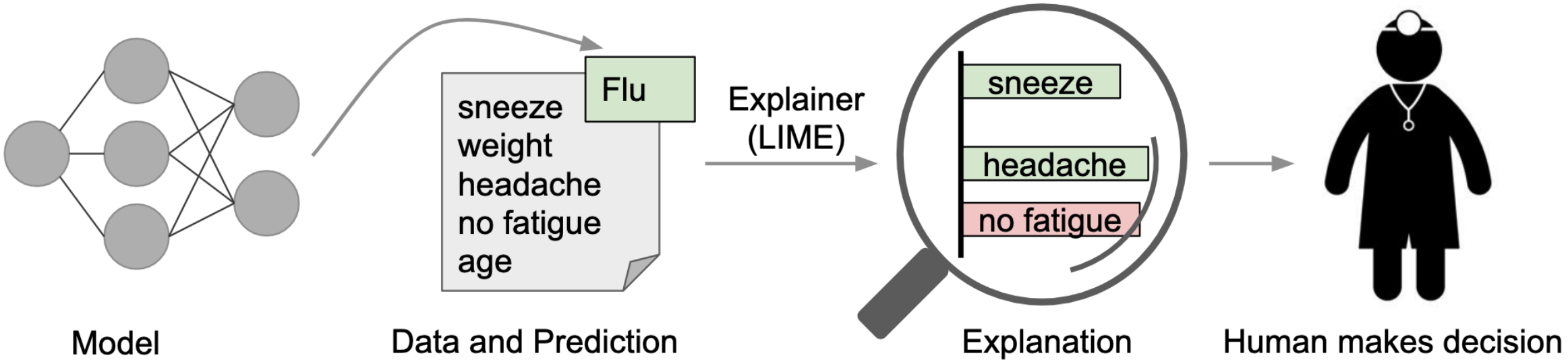
"Why Should I Trust You?" Explaining the Predictions of Any Classifier

Local Interpretable Model-agnostic Explanations (LIME)

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu



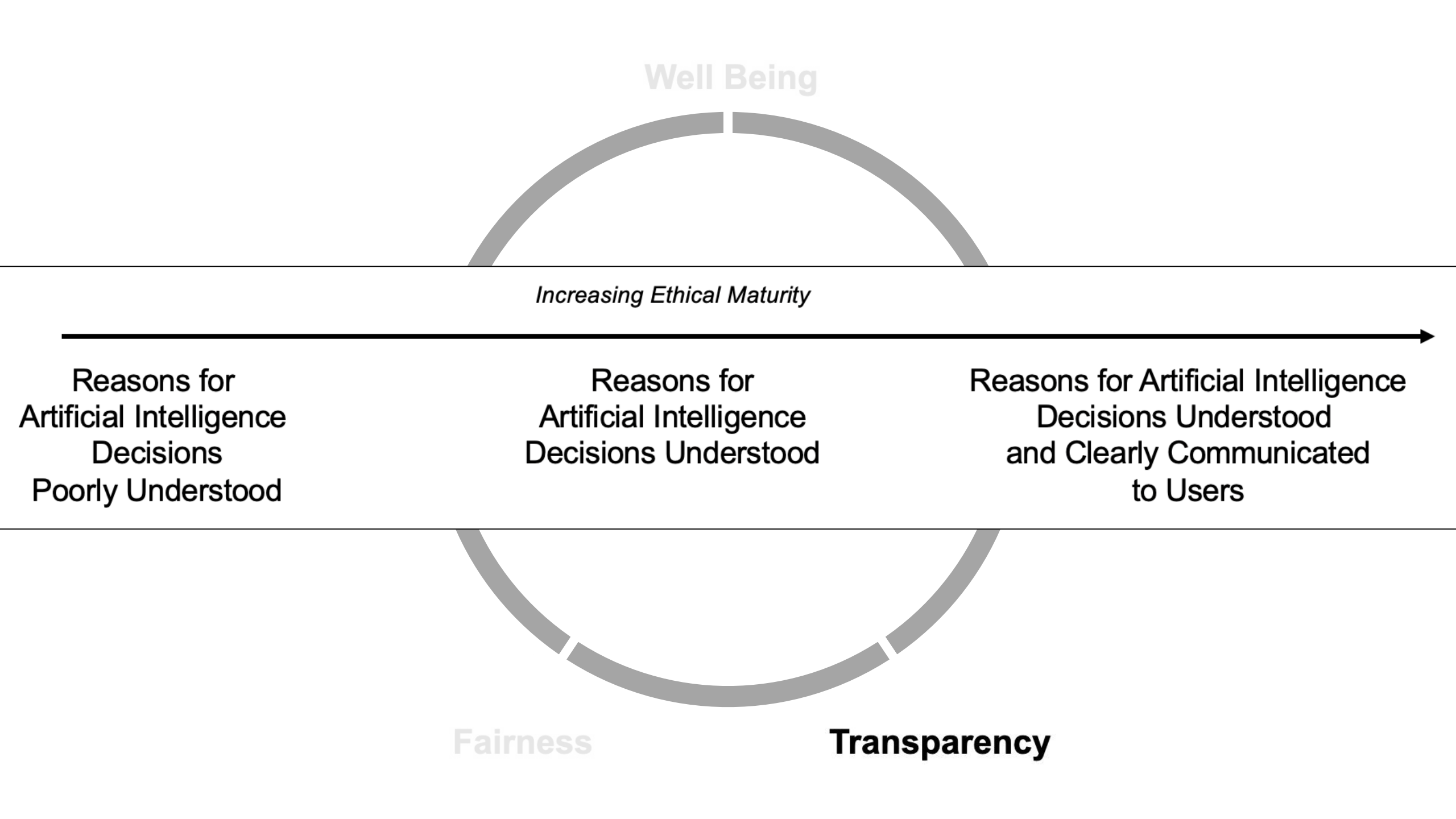
http://arxiv.org/pdf/1602.04938.pdf

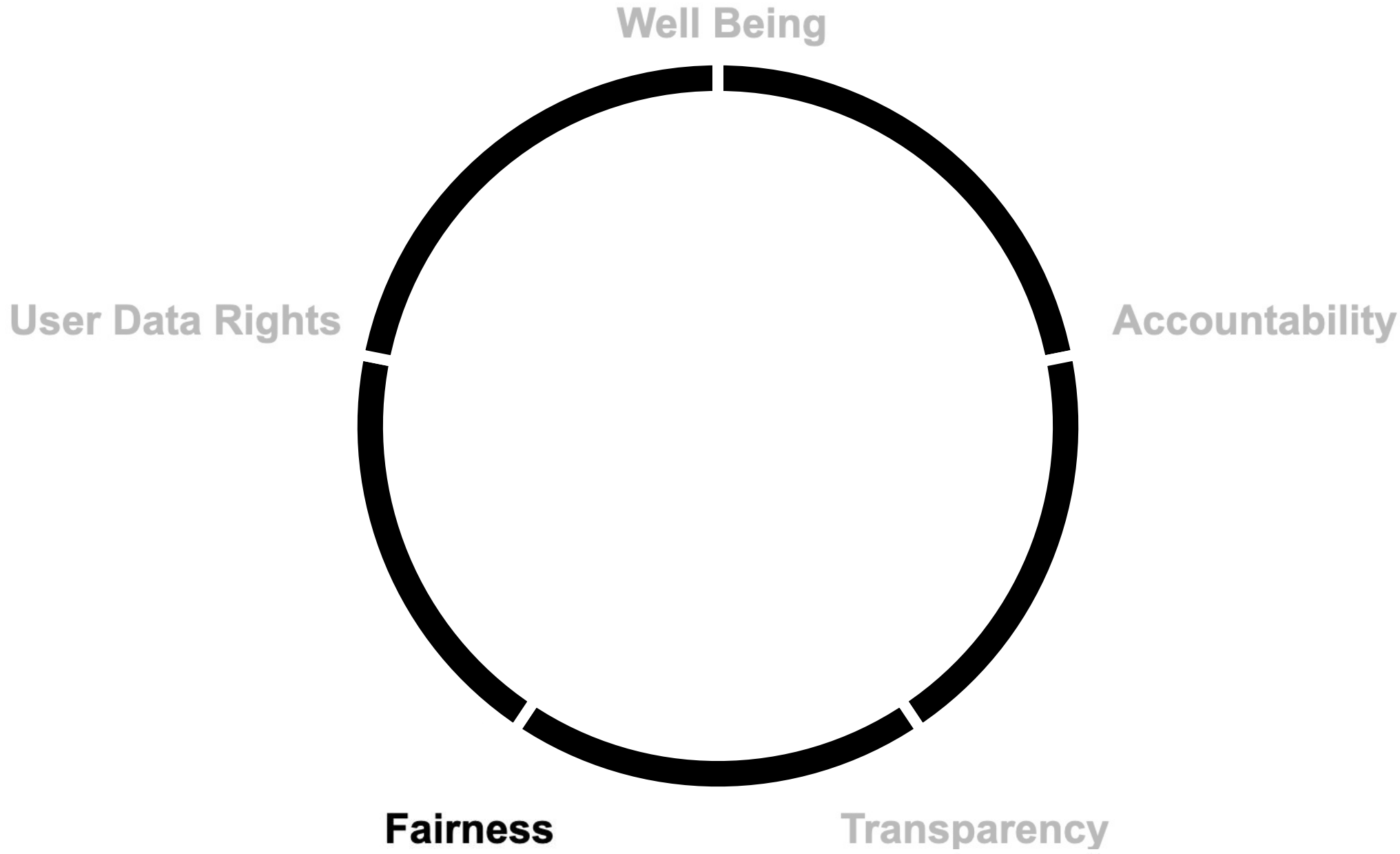
Why are bubbles round?

Consider a smooth surface in $\mathbb{R}^{n+1}$ representing the graph of a function $x_{n+1} = u(x_1, \ldots, x_n)$ defined on a bounded open set Ω in $\mathbb{R}^n$. Assuming that u is sufficiently smooth, the area of the surface is given by the nonlinear functional

$$\mathcal{A}(u) = \int_\Omega \left(1 + |\nabla u|^2\right)^{1/2} \mathrm{d}x_1 \ldots \mathrm{d}x_n, \tag{1}$$

where ∇u is the gradient vector $(\partial u/\partial x_1, \ldots, \partial u/\partial x_n)$ and $|\nabla u|^2 = (\nabla u) \cdot (\nabla u)$.


Well Being
Increasing Ethical Maturity
Reasons for Artificial Intelligence Decisions Poorly Understood
Reasons for Artificial Intelligence Decisions Understood
Reasons for Artificial Intelligence Decisions Understood and Clearly Communicated to Users
Fairness
Transparency


Well Being
User Data Rights
Accountability
Fairness
Transparency

Bias

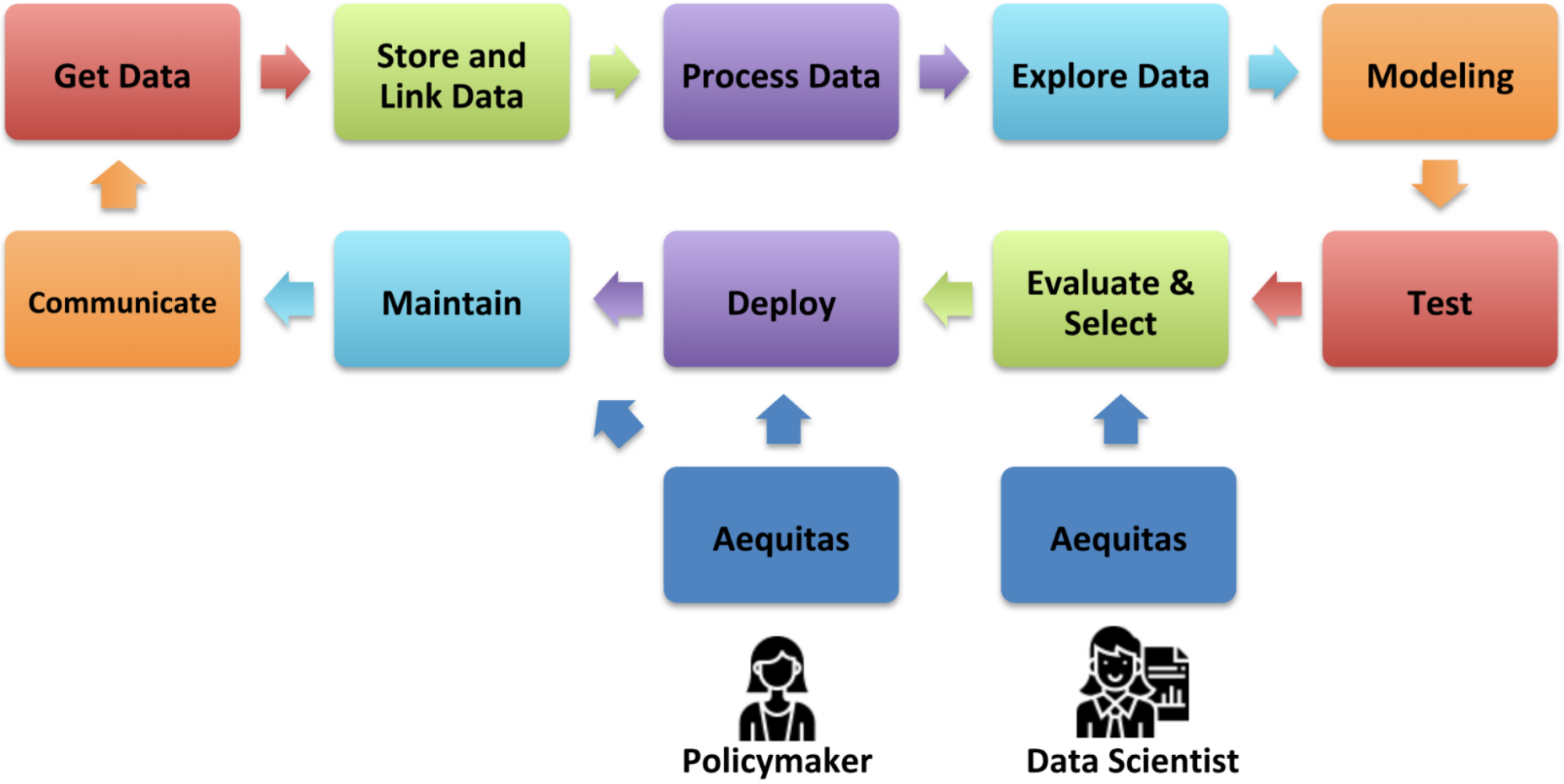
Shortcut

Shortcut | Impartiality

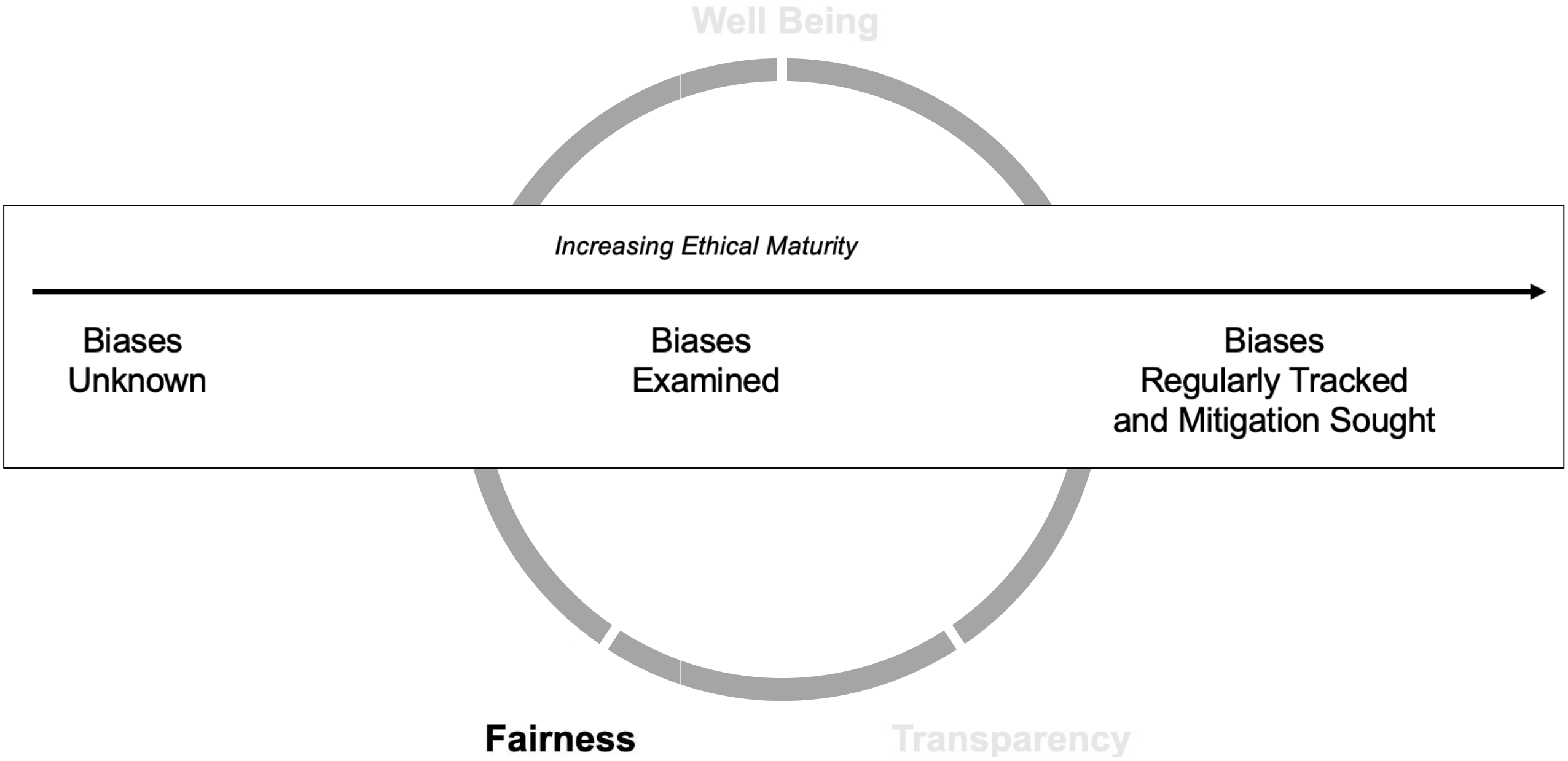
Shortcut | Impartiality | Self-Interest

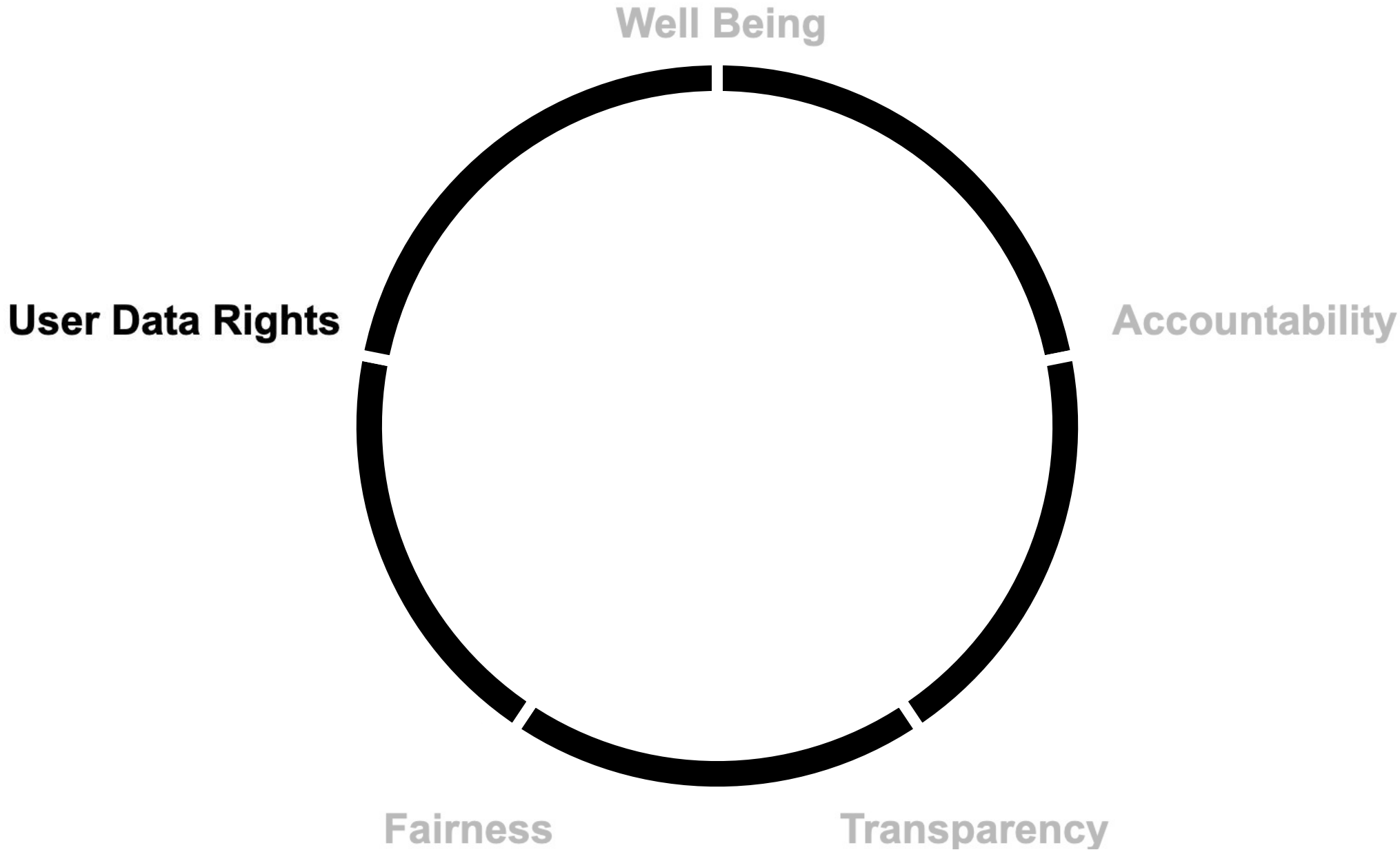
Tarra Simmons

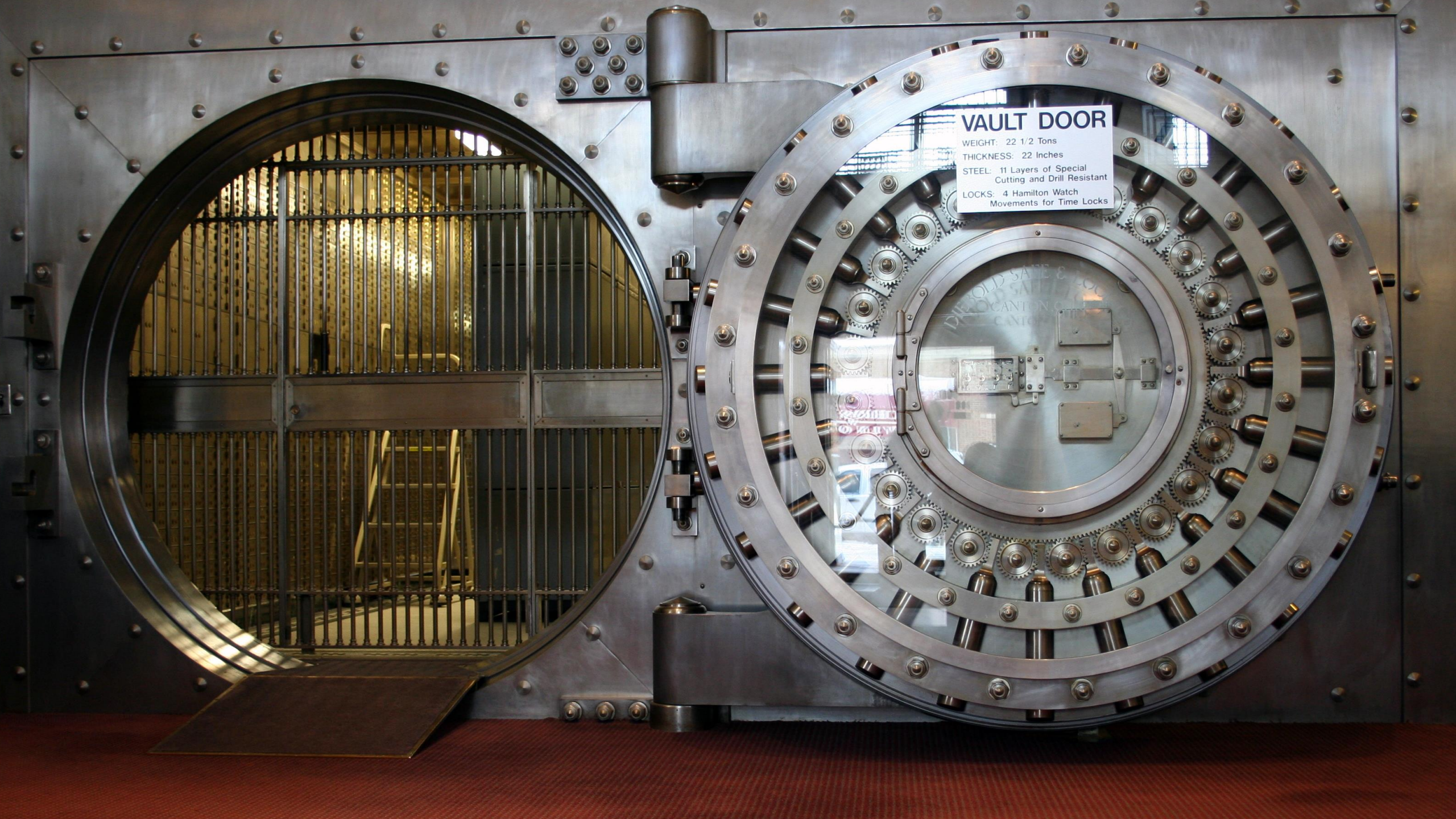
Aequitas: A Bias and Fairness Audit Toolkit

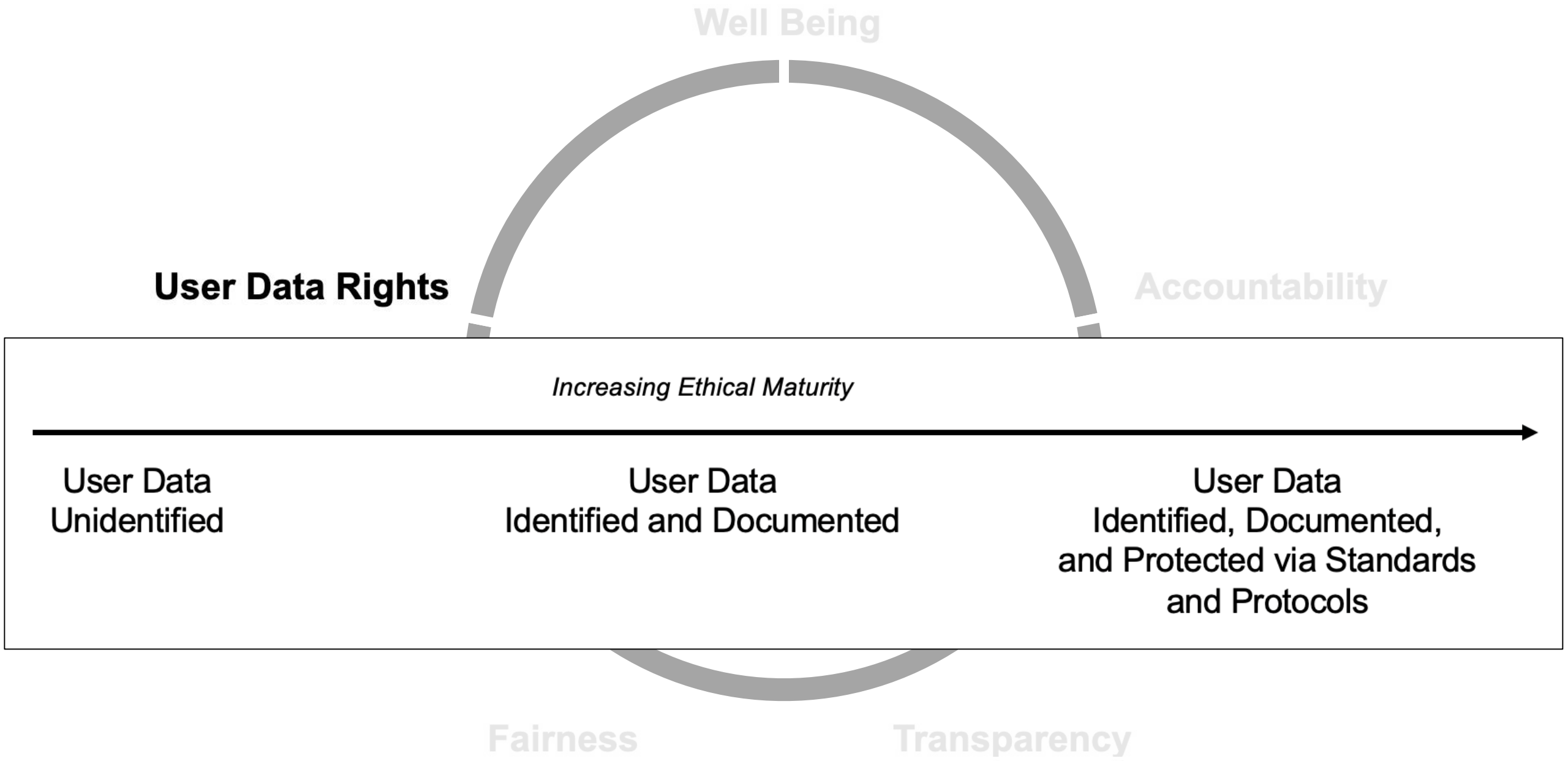


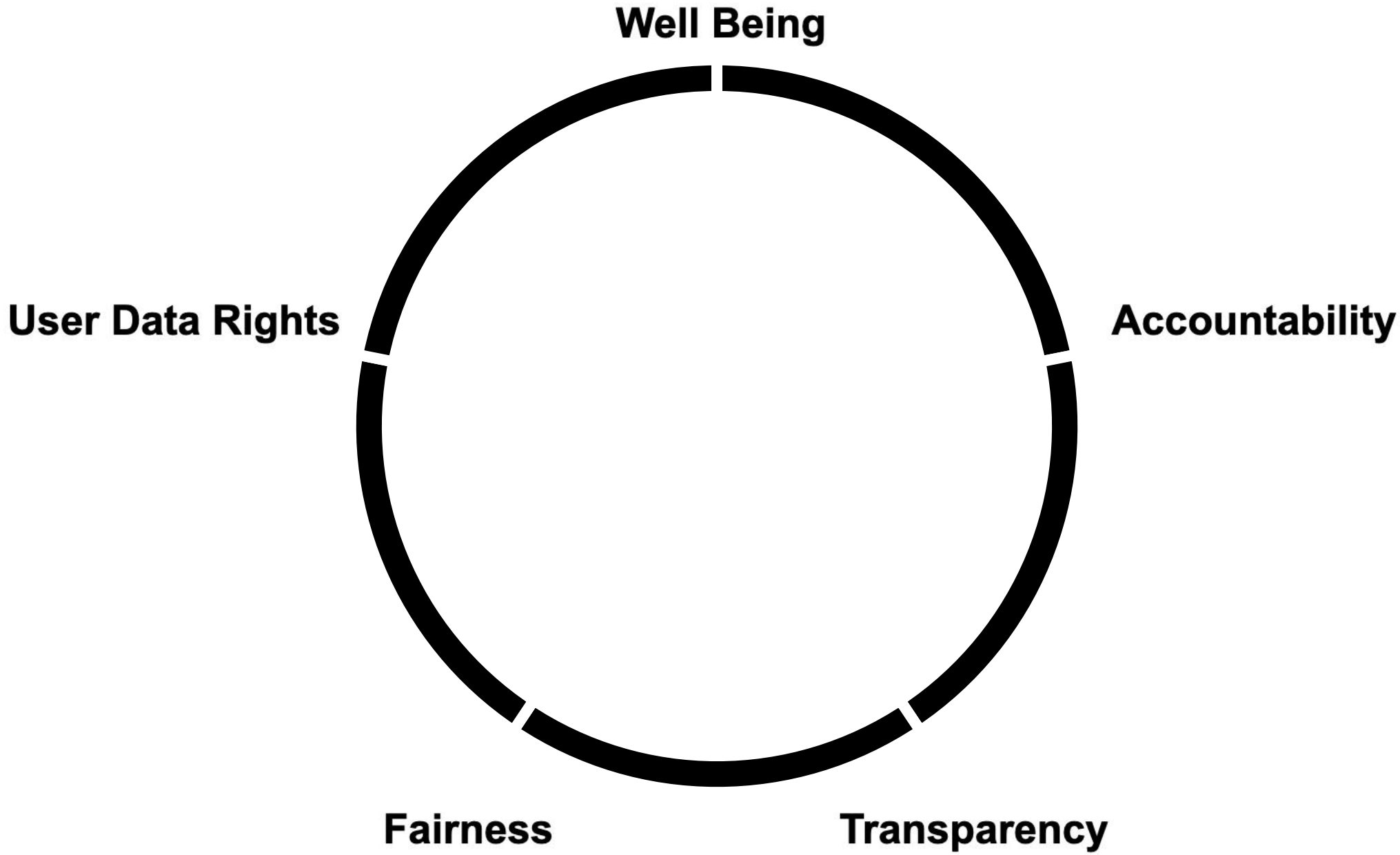
https://arxiv.org/pdf/1811.05577.pdf

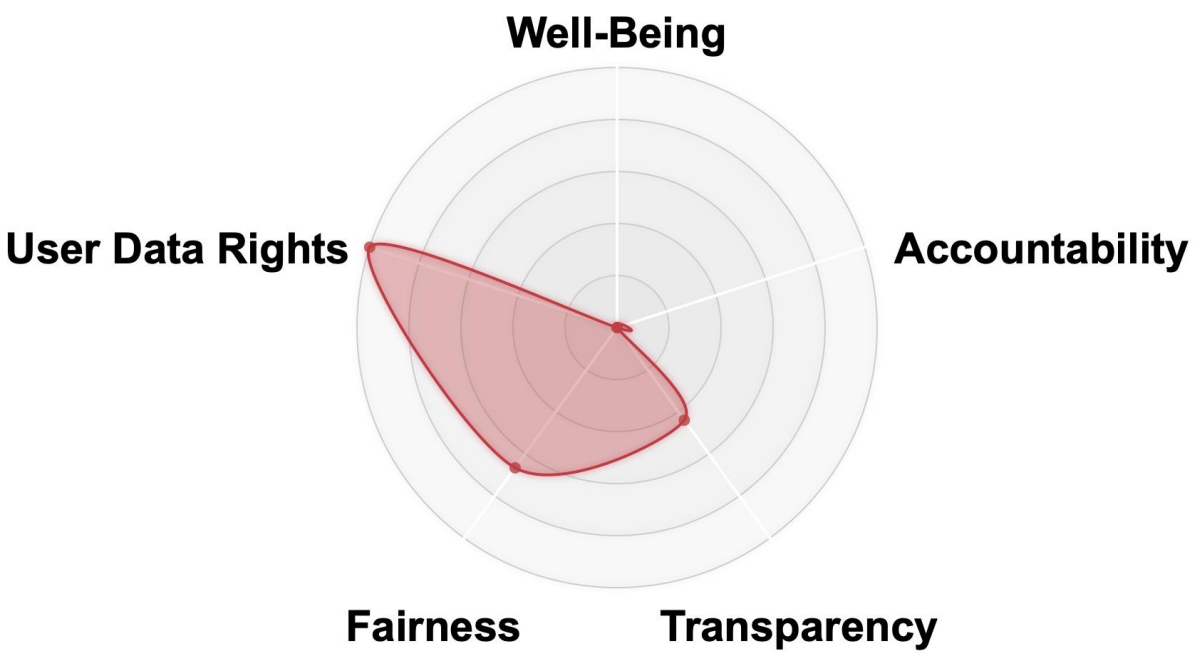

Well Being
Increasing Ethical Maturity
Biases
Unknown
Biases
Examined
Biases
Regularly Tracked
and Mitigation Sought
Fairness
Transparency

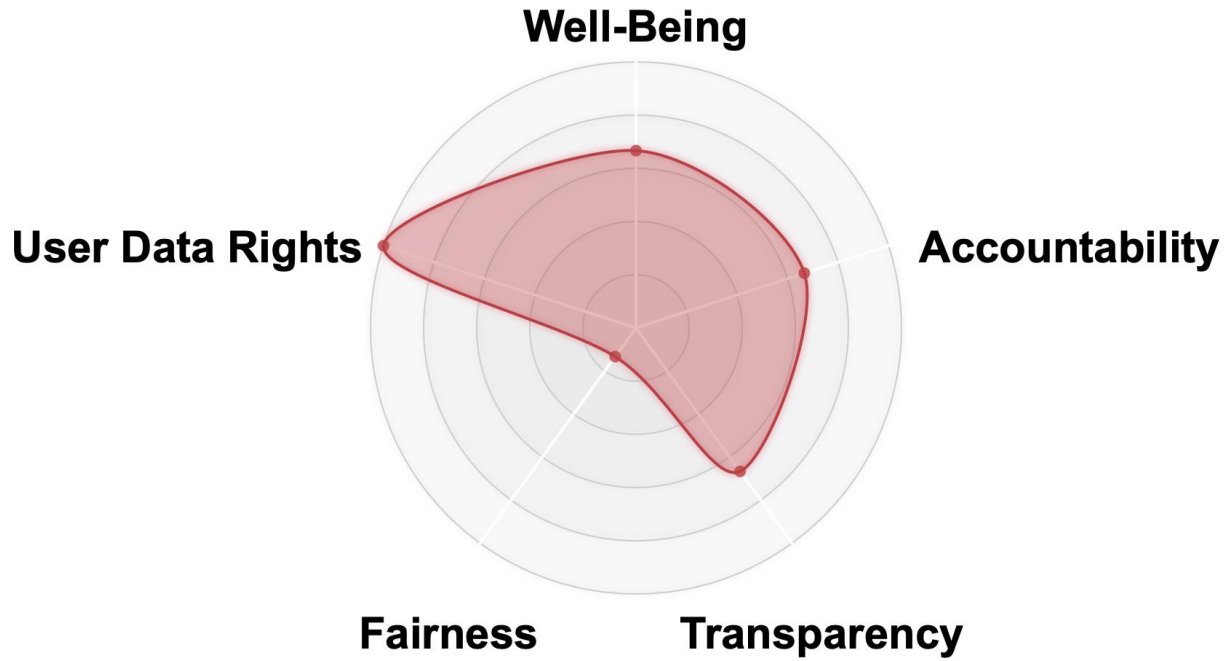

Well Being
User Data Rights
Accountability
Fairness
Transparency


VAULT DOOR
WEIGHT: 22 1/2 Tons
THICKNESS: 22 Inches
STEEL: 11 Layers of Special Cutting and Drill Resistant
LOCKS: 4 Hamilton Watch Movements for Time Locks

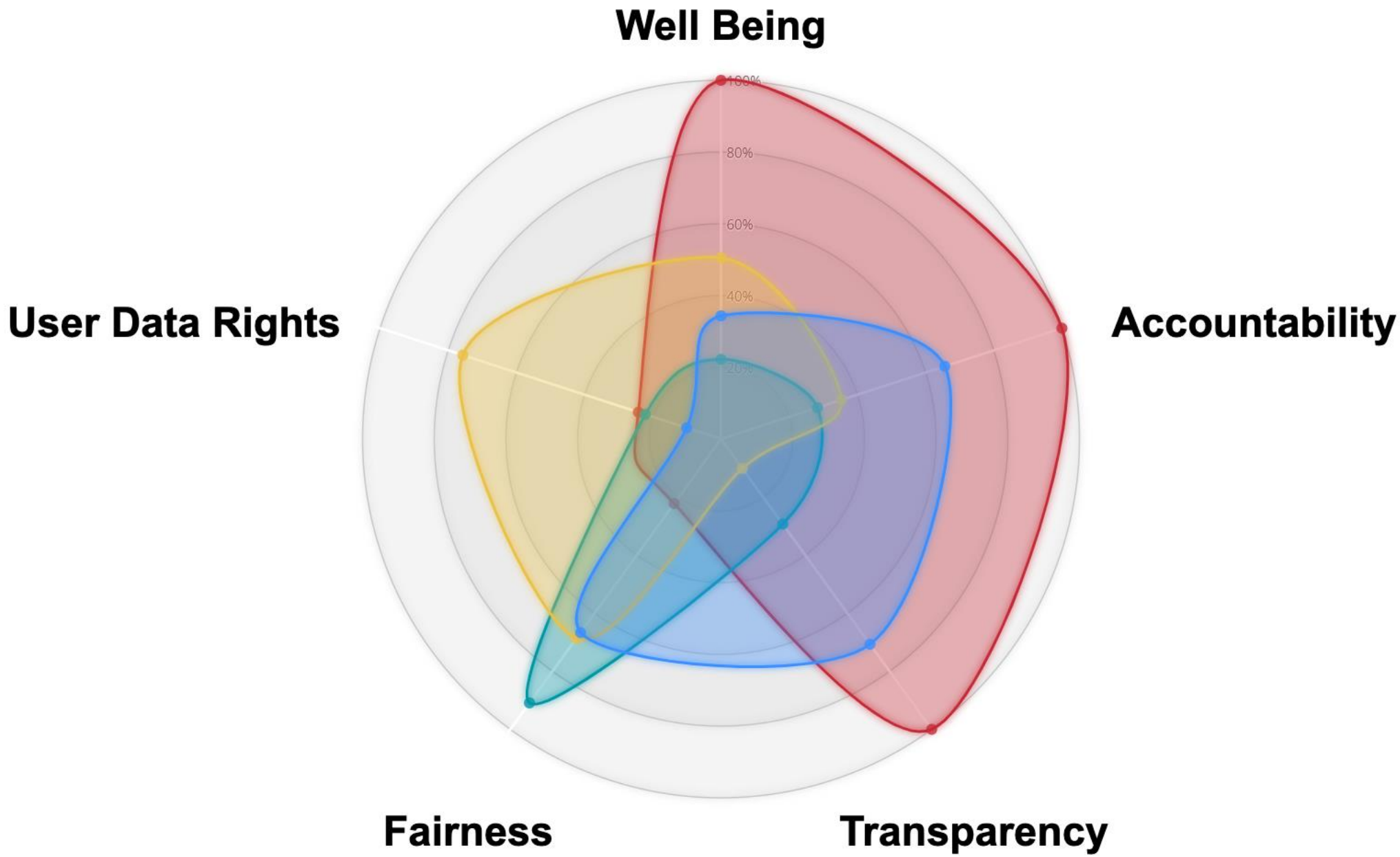

Well Being
User Data Rights
Accountability
Increasing Ethical Maturity
User Data
Unidentified
User Data
Identified and Documented
User Data
Identified, Documented,
and Protected via Standards
and Protocols
Fairness
Transparency

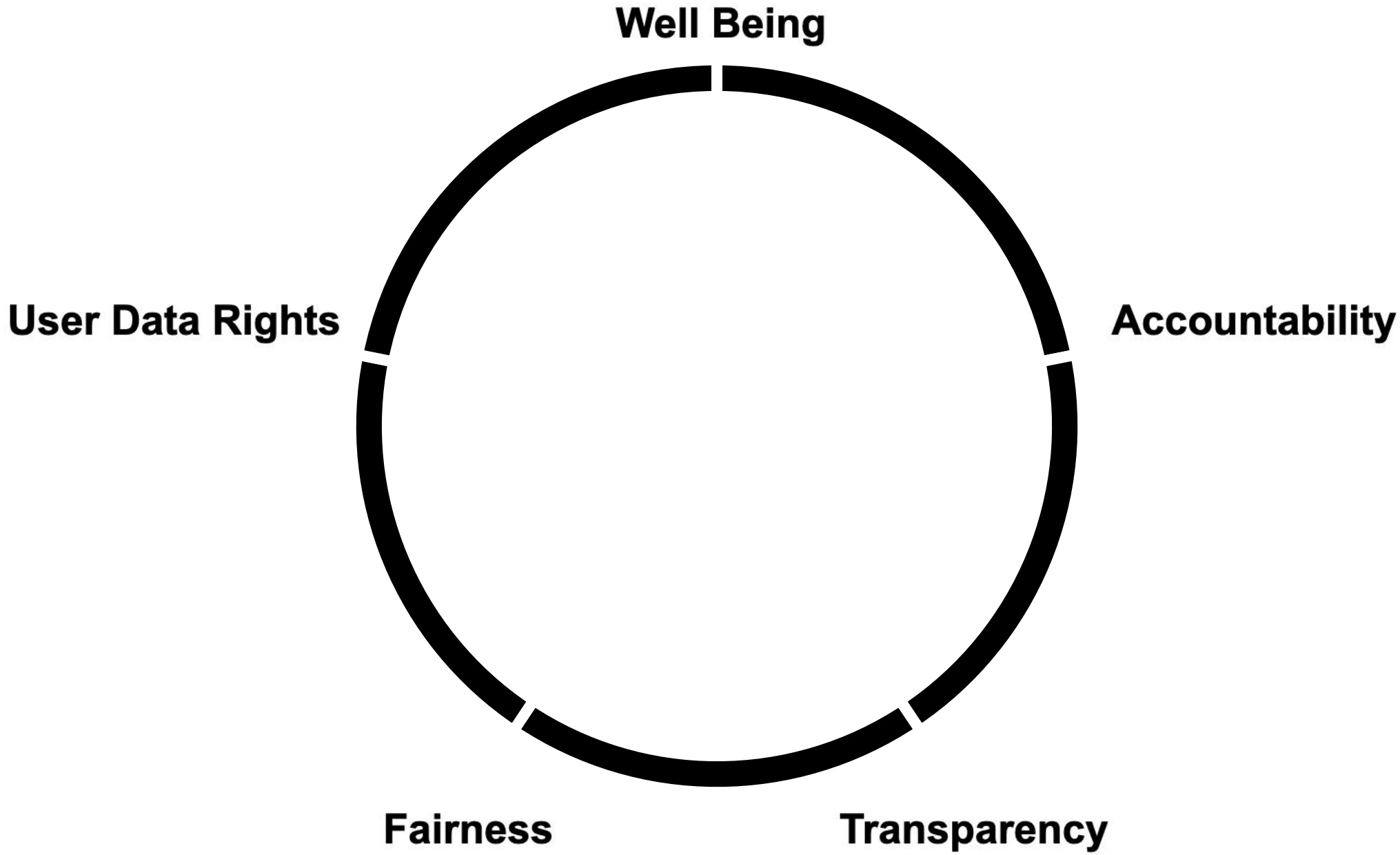

Well Being
User Data Rights
Accountability
Fairness
Transparency





Increasing Ethical Maturity


Well Being
User Data Rights
Accountability
Fairness
Transparency
100%
80%
60%
40%
20%


Well Being
User Data Rights
Accountability
Fairness
Transparency

Comparison of Ethical Frameworks

Ethical Maturity Graph	IEEE	European Commission	IBM
Well-being	Human Rights Well-being Awareness of Misuse	Prevention of Harm Societal and environmental well-being	Value Alignment
Accountability	Accountability	Accountability Human Agency and Oversight Human Autonomy	Accountability
Transparency	Transparency	Transparency Explicability	Explainability
Fairness	(Included in Transparency)	Fairness Diversity, non-discrimination, and fairness	Fairness
User Data Rights	Data Agency	Privacy and Data Governance	User Data Rights
	Effectiveness Competence	Technical Robustness and Safety	

Comparison of Transparency Tools

Local Interpretable Model-agnostic Explanations (LIME)

- Local Model Approximation
- Less Intuitive to Human Reasoning
- Relatively Fast

https://github.com/marcotcr/lime

SHAP

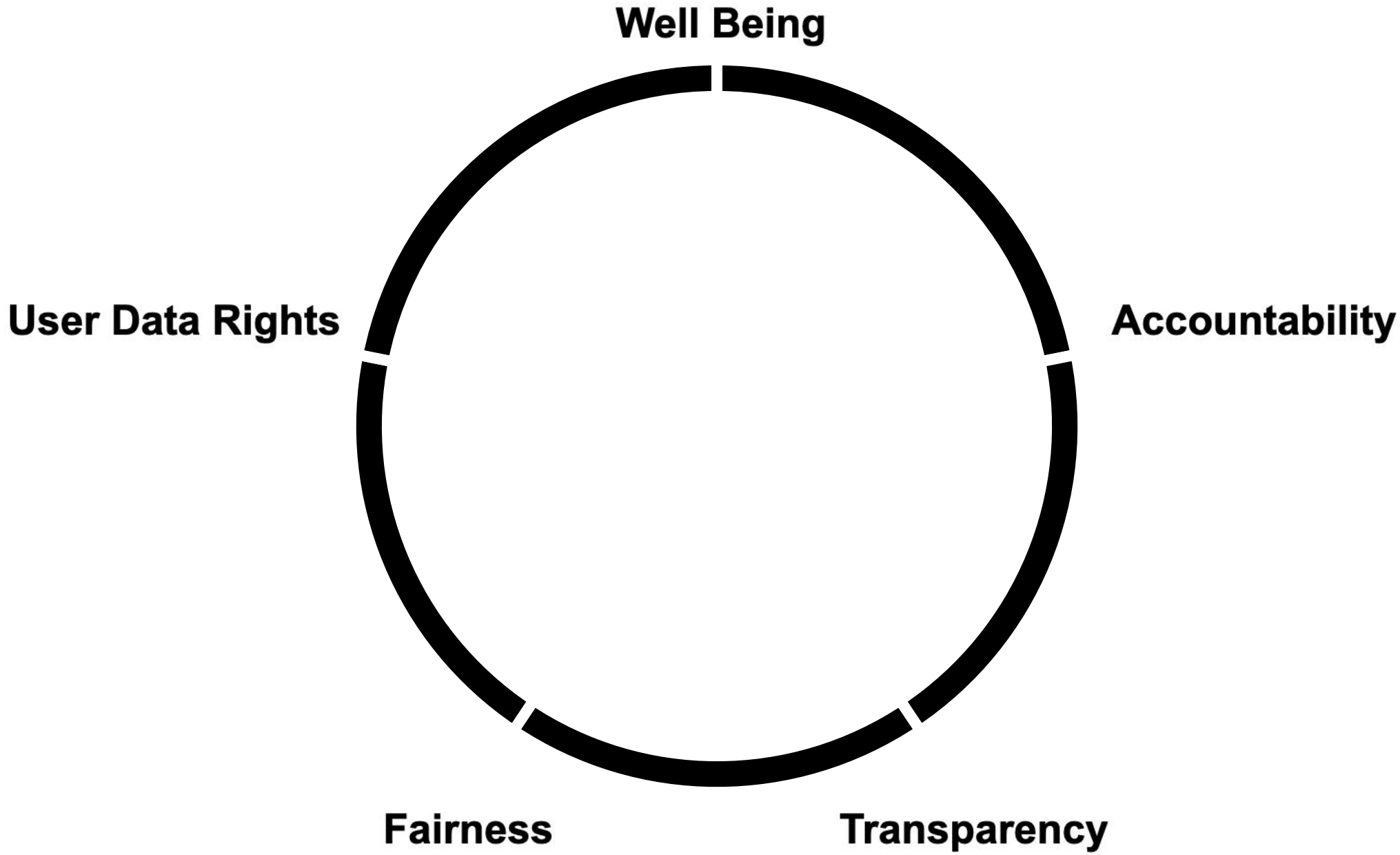
- Game Theory
- More Intuitive to Human Reasoning
- Relatively Slow
  (Faster on tree structures, slow on k-nearest neighbor)

https://github.com/slundberg/shap

Explore, Explain and Examine Predictive Models: https://pbiecek.github.io/ema/

Comparison of Fairness Tools

Fairness Comparison	Benchmarks fairness of algorithms	https://github.com/algofairness/fairness-comparison
AI Fairness 360	Fairness Metrics Embeddable code for bias/mitigation Developer-focused Input cleansing	https://github.com/IBM/aif360
Audit-AI	Equal Employment Opportunity Commission Standards Fairness in hiring driven	https://github.com/pymetrics/audit-ai
Thesis-ML	Fairness-aware ML algorithms	https://github.com/cosmicBboy/themis-ml
Aequitas	Toolkit (libraries, command line, web interface) Usable by policymakers and auditors	https://github.com/dssg/aequitas


Well Being
User Data Rights
Accountability
Fairness
Transparency

So What Now?

- **For the flight home:**
  - Podcast: https://www.npr.org/programs/invisibilia/597779069/the-pattern-problem
  - Academic paper from ISAAI Symposium (p. 55): https://digitaleweltmagazin.de/d/0p6djg8a/DW_0120.pdf
- **Next week:**
  - Catalog the use of algorithms within your organization
- **Three months from now:**
  - Construct an Ethical Maturity Graph for each use of algorithms/AI
  - Complete an ethics canvas for each algorithm
  - Examine potential tools to mature along each axis (transparency / accountability / fairness / user data rights)
- **Six months from now:**
  - Implement appropriate open-source tools
  - Target minimum levels of maturity for each axis
  - Conduct review of progress through self-evaluation


TEMPLE OF JUSTICE

We affirm this court's long history of recognizing that one's past does not dictate one's future.