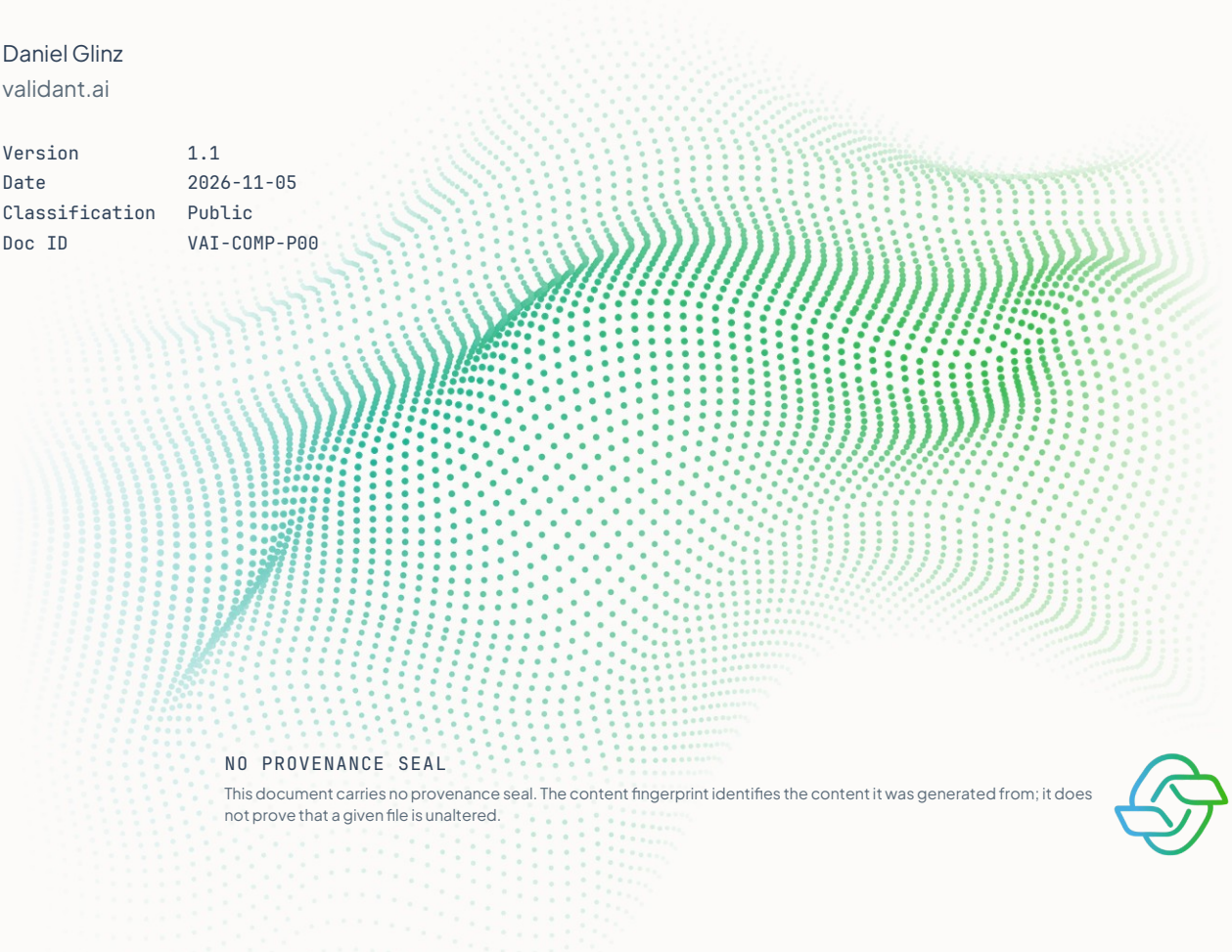
THE AI TRUST & FAIRNESS COMPENDIUM

Front Matter and Master Table of Contents

Front matter of The AI Trust & Fairness Compendium: preface and purpose, executive overview, scope, audience and limitations, scientific and editorial method, how to use the compendium, and the full table of contents of all twelve parts and the three appendix volumes

Daniel Glinz
validant.ai

Version	1.1
Date	2026-11-05
Classification	Public
Doc ID	VAI-COMP-P00



NO PROVENANCE SEAL

This document carries no provenance seal. The content fingerprint identifies the content it was generated from; it does not prove that a given file is unaltered.

ABOUT THIS DOCUMENT

Front Matter and Master Table of Contents

Front matter of The AI Trust & Fairness Compendium: preface and purpose, executive overview, scope, audience and limitations, scientific and editorial method, how to use the compendium, and the full table of contents of all twelve parts and the three appendix volumes

SUGGESTED CITATION

Glinz, D. (2026). *Front Matter and Master Table of Contents: Front matter of The AI Trust & Fairness Compendium: preface and purpose, executive overview, scope, audience and limitations, scientific and editorial method, how to use the compendium, and the full table of contents of all twelve parts and the three appendix volumes* (Compendium · Front matter VAI-COMP-P00, Version 1.1). Glinz & Company GmbH / validant.ai.

AUTHOR

Daniel Glinz, validant.ai

PUBLISHER AND IMPRINT

validant.ai · Glinz & Company GmbH · Zolliker Strasse 98 · CH-8702 Zollikon · Switzerland · hello@validant.ai · www.validant.ai

VERSION HISTORY

1.0 · 2026-11-05 · first edition: compiled from the fact-checked chapters and checked against them

1.1 · 2026-11-06 · sources described as a literature review, practical and industry best practice, and the knowledge base collected by validant.ai on fairness, non-discrimination and explainability

DOCUMENT IDENTIFIER AND VERIFICATION

VAI-COMP-P00

This document carries no provenance seal.

Content fingerprint (SHA-256) 8355b18742976d7eda427ebb16319143e321d6ad37cce5fec1af4c51c6db73f9

COPYRIGHT AND USE

© 2026 Glinz & Company GmbH. All rights reserved.

CLASSIFICATION

Public

CONTENTS

Contents of the front matter

About this part 3

Preface and Purpose 5

Executive Overview 9

Scope, Audience and Limitations 14

Scientific and Editorial Method 18

How to Use This Compendium 23

Full Table of Contents 30

OVERVIEW

About this part

KEY MESSAGE

The compendium makes AI fairness checkable: twelve standalone parts with 115 fixed chapter numbers and three appendix volumes, written from identified sources, reviewed and fact-checked, and read by part, by task or by question.

This document is the front matter of The AI Trust & Fairness Compendium, a reference work in twelve parts and three appendix volumes published by validant.ai. Each part is a standalone document that can be read on its own; this volume introduces the whole work. It contains the Preface and Purpose (F1), the Executive Overview (F2), Scope, Audience and Limitations (F3), the Scientific and Editorial Method (F4) and How to Use This Compendium (F5), followed by the full table of contents of all twelve parts, their 115 chapters and the three appendix volumes. Chapter numbers are fixed across the whole work, so every cross-reference in this volume names the chapter in which the topic is treated in full.

TABLE 1. The compendium and the place of this document in it

PART	TITLE	
Front	Front matter and master table of contents	This document
I	Foundations of Trust, Justice and Non-Discrimination	
II	Formal Foundations of Algorithmic Fairness	
III	Bias, Discrimination and Causality	
IV	Measuring Fairness: Metrics and Statistical Evidence	
V	Fairness Interventions	
VI	Fairness Engineering and Operations	
VII	Fairness in Advanced AI Architectures	
VIII	Explainability, Interpretability and Non-Discrimination	
IX	Legal, Ethical and Regulatory Foundations	
X	AI Assurance, Evidence and Digital Trust	
XI	Fairness in Organizational Practice	
XII	Emerging Challenges and Research Frontiers	
App.	Appendices A to L (three volumes)	

The front matter introduces The AI Trust & Fairness Compendium as a whole. The Preface states why the compendium was written, the positions that run through all parts and who carries editorial responsibility for the text. The Executive Overview summarizes each of the twelve parts in one paragraph, describes the three appendix volumes and draws the conclusions the parts support together. The chapter on scope, audience and limitations states which systems, domains and jurisdictions are covered, for whom the text is written and what it cannot establish: it is not legal advice, certifies no system and states law as of November 2026. The chapter on scientific and editorial method describes the sources, the writing rules, the review and the fact-check, and reports the counts the fact-checks recorded: 2268

reference entries, 2867 claims and 861 formulas checked. The chapter on how to use the compendium gives reading paths by task and by question, and the volume closes with the full table of contents.

TABLE 2. Chapters in this document and their central claims

CHAPTER	TITLE	CENTRAL CLAIM
F1	Preface and Purpose	The compendium exists to make AI fairness checkable: it treats fairness as a set of explicit normative choices, measured evidence and accountable decisions, and it gives each of them a chapter, a method and a source.
F2	Executive Overview	The twelve parts move from why fairness and trust matter, through how unfairness is defined, traced, measured and corrected, to how it is engineered, explained, regulated, assured, governed and researched; each part can be read on its own.
F3	Scope, Audience and Limitations	The compendium covers fairness, non-discrimination, explainability and assurance for AI systems that make or support decisions about people; it informs judgment but is not legal advice, does not certify any system and states law as of November 2026.
F4	Scientific and Editorial Method	Every chapter was written from identified sources under fixed rules, reviewed by a scientific editor and fact-checked: references looked up in Crossref, OpenAlex or DBLP or confirmed from a primary source, and claims and formulas rechecked.
F5	How to Use This Compendium	The compendium can be read by part, by task or by question: fixed chapter numbers, one chapter template, classified statements and four linked assessment methods let a reader enter where needed and follow cross-references to the evidence.
Contents	Full Table of Contents	The compendium has twelve parts with 115 chapters and three appendix volumes; each part is a standalone document, and chapter numbers are fixed across the whole work.

How to read this volume. Readers new to the compendium can start with the Executive Overview (F2), which summarizes each part in one paragraph, and then follow the reading paths by task in How to Use This Compendium (F5). Before relying on a finding for a decision, read Scope, Audience and Limitations (F3), and, to judge how far a statement can be relied on, the Scientific and Editorial Method (F4). The full table of contents at the end lists every chapter by its fixed number. Statements are classified throughout: established scientific findings, legal or normative requirements, methodological positions of validant.ai and open questions are marked as such in framed boxes.

FRONT MATTER

Preface and Purpose

KEY MESSAGE

The compendium exists to make AI fairness checkable: it treats fairness as a set of explicit normative choices, measured evidence and accountable decisions, and it gives each of them a chapter, a method and a source.

Purpose and scope

This chapter is the preface of The AI Trust & Fairness Compendium. It explains why the compendium was written, what it is for, which positions run through all of its parts, and who is responsible for it. The other front matter chapters complete the picture: the Executive Overview summarizes each part, the chapter on scope, audience and limitations states what the compendium covers and what it cannot establish, the chapter on scientific and editorial method describes how the text was produced and checked, and the chapter on how to use the compendium gives reading paths. The full table of contents of all parts and appendix volumes follows the front matter.

The front matter was compiled from the compendium's own chapters, which were fact-checked; the chapter on scientific and editorial method describes the sources the compendium rests on: a literature review, practical and industry best practice, and the knowledge base on fairness, non-discrimination and explainability collected by validant.ai. Every substantive statement in it names the chapter in which the topic is treated in full. A reader who wants the argument, the evidence or the references behind a statement should follow that chapter number.

Why this compendium

AI systems already support decisions about who is hired, who receives credit and who gains access to services. The same technique faces different evidence, notice, review and audit obligations in credit, employment, insurance, healthcare, education and public services (Chapter 91). When such systems are not built and governed with care, they can carry existing social hierarchies forward: historical discrimination reaches contemporary AI through identifiable mechanisms such as biased targets and labels, proxies and feedback loops (Chapters 6 and 8).

The difficulty is not a lack of fairness metrics. It is that fairness is a family of contested conceptions, that formal criteria encode different ones of them (Chapter 3), and that, when base rates differ between groups and prediction is imperfect, the main criteria cannot all hold at once (Chapter 19). Choosing among them is therefore unavoidable and normative (Chapter 22). A fairness metric computed without a stated question, a fixed analysis plan, validated measurements and quantified uncertainty is a number, not evidence (Chapter 36). And an assurance claim built on such evidence holds only for a stated scope, depth, criteria and date (Chapter 101).

The compendium was written to close the gap between these insights and daily practice. It brings into one place, with one notation and one vocabulary, the knowledge an organization needs to examine an AI system for unfair treatment of people, to decide what to do about it, to document the decision and to show others that the claim it makes is supported.

What the compendium is

The compendium is a reference work in twelve parts with 115 chapters and three appendix volumes. Each part is a standalone document that can be read on its own, and chapter numbers are fixed across the whole work, so that a cross-reference such as "Chapter 19" means the same chapter everywhere (the full table of contents). It is published as a public document by validant.ai. It is also designed to be read section by section: each section is written to make sense when it is read on its own, for example when it is retrieved by a search system.

The work follows one line from foundations to practice. It begins with trust and justice (Part I), gives fairness a precise mathematical language (Part II), explains where bias comes from and how causal models trace it (Part III), turns criteria into statistical evidence (Part IV), describes the interventions available when a criterion is violated (Part V) and makes fairness an engineering and operational discipline (Part VI). It then treats advanced architectures (Part VII), explainability (Part VIII), law and regulation (Part IX) and assurance and digital trust (Part X), before turning to organizational practice (Part XI) and to the open research frontier (Part XII). The Executive Overview summarizes each part in one paragraph.

The appendices collect, from the chapters, the material a practitioner needs to look up rather than read: a controlled glossary, the mathematical notation, a catalog of fairness metrics and one of explainability methods, statistical methods, worked numerical examples, a legal and regulatory matrix, standards and normative references, templates and worksheets for the assessment methods, a taxonomy of evidence and assurance claims, a bibliography and source register, and an index of definitions (the full table of contents).

From diagnosis to treatment to practice

Several parts of the compendium follow one sequence of work on a specific system. As an analogy, it resembles medical practice: an assessment first diagnoses where and how a system treats groups unequally, an intervention then treats the cause, and organizational practice keeps the treatment in place.

- **Diagnosis.** Four methods examine a system in sequence: the Historical Context Assessment Method identifies which documented patterns of discrimination are relevant (Chapter 11), the Fairness Definition Selection Method chooses the criteria for the context (Chapter 22), the Bias Source Identification Method locates where bias enters the life cycle (Chapter 35) and the Fairness Metrics Selection Method links the chosen criteria to specified, validated metrics (Chapter 44). The Fairness Audit Playbook combines the four into one traceable workflow (Chapter 110).
- **Treatment.** Causal analysis shows why a disparity arises and where an intervention can remove it (Chapter 30). The Intervention Strategy Selection Method then chooses among pre-processing, in-processing and post-processing interventions according to where the disparity enters, the target criterion, what can be changed and whether the protected attribute may be used (Chapter 46).
- **Implementation in code.** Fairness libraries, automated checks in development pipelines, monitoring and drift detection keep the measured properties in place after deployment (Chapters 58 to 62).
- **Implementation in organizations.** Delivery practices, roles, documentation, governance and incident handling carry explicit fairness content, so that bias work is planned, owned, recorded, decided and repaired (Part XI, Chapters 102 to 109).

METHODOLOGICAL POSITION · ONE CHAIN FROM HISTORY TO EVIDENCE

A fairness audit is credible only as a connected chain: historical context determines which harms matter, those harms determine the criteria, criteria and bias sources determine the metrics, and every finding, decision and remedy traces back along that chain (Chapter 110).

Positions that run through the compendium

The chapters differ in subject, but several positions recur across them. They are stated here once, each with the chapter that argues for it.

- **Trust must follow trustworthiness.** Trust is warranted only when it tracks the other party's actual trustworthiness, so a trustworthy AI practice must produce evidence rather than ask for trust (Chapter 1). For AI, the goal is not more trust but appropriate reliance, calibrated to a system's actual, group-specific capability (Chapter 2).
- **The choice of criterion is normative.** Each formal criterion encodes one conception of fairness, so the choice of criterion must be made explicitly, for a context, and documented (Chapter 3). The Fairness Definition Selection Method makes that choice a documented, repeatable decision (Chapter 22).
- **Removing a protected attribute does not remove its influence.** Other variables act as proxies, and assessment often needs sensitive data that law restricts (Chapter 24).
- **A metric is evidence only with its denominators and uncertainty.** Every rate must be reported with its denominator, reference group, direction, interval and intersectional breakdown (Chapter 39), and small groups must never be dropped silently (Chapter 43).
- **Interventions have a price that must be measured.** No intervention method is best everywhere, and the price of each must be measured on the specific system (Part V, Chapter 46).
- **Fairness is monitored, not certified once.** A system found fair on historical data can become discriminatory after deployment as populations, base rates and relationships change (Chapter 34).
- **Explanations show how a model works, not that it is fair.** They are evidence only when faithful to the model, robust, tested by group and understood by their audience (Part VIII, Chapters 71 and 81).
- **Law is a mandatory but moving floor.** It names protected grounds and defines direct and indirect discrimination, yet no metric proves compliance and compliance alone does not establish fairness (Part IX, Chapters 83 and 92).
- **Assurance states its own bounds.** No certificate can show that an AI system is fair in general; an honest scheme publishes the scope, depth, criteria and date of its verdict (Chapter 101).

KEY TAKEAWAY · FAIRNESS IS MADE CHECKABLE, NOT DECLARED

Across all parts, the compendium treats a fairness claim as acceptable only when its normative choices are explicit, its evidence is measured with stated uncertainty, its reasoning can be reviewed, and a person or body is accountable for the decision it supports.

Authorship and editorial responsibility

The compendium is written by Daniel Glinz and published by Glinz & Company GmbH under the name validant.ai. The author carries editorial responsibility for the whole text, including the classification of statements and the methodological positions of validant.ai that the chapters mark as such.

Parts of the text were drafted with AI assistance under human editorial responsibility. Every chapter was then reviewed and fact-checked against its sources: references were looked up in bibliographic databases or confirmed from a primary source, and claims and formulas were rechecked. The chapter on scientific and editorial method describes these steps and reports the counts that the fact-checks recorded.

The compendium is a reference work, not legal advice, and not a certification of any system. Legal content is stated as of November 2026 and should be checked against the current law before it is relied on (Part IX); the chapter on scope, audience and limitations states what follows from this.

FRONT MATTER

Executive Overview

KEY MESSAGE

The twelve parts move from why fairness and trust matter, through how unfairness is defined, traced, measured and corrected, to how it is engineered, explained, regulated, assured, governed and researched; each part can be read on its own.

Purpose and scope

This chapter gives a reader the whole compendium in a few pages. It contains one paragraph for each of the twelve parts and a short description of the three appendix volumes. Each paragraph was compiled from the summary that the part itself carries, so it states what that part establishes in the part's own terms, and it names the chapters in which each topic is treated. A reader can use the overview to decide which part to open first, or as a map to return to; the chapter on how to use the compendium adds reading paths by role and by question.

TABLE 3. The twelve parts of the compendium

PART	TITLE	CHAPTERS
I	Foundations of Trust, Justice and Non-Discrimination	1 to 11
II	Formal Foundations of Algorithmic Fairness	12 to 22
III	Bias, Discrimination and Causality	23 to 35
IV	Measuring Fairness: Metrics and Statistical Evidence	36 to 45
V	Fairness Interventions	46 to 57
VI	Fairness Engineering and Operations	58 to 64
VII	Fairness in Advanced AI Architectures	65 to 69
VIII	Explainability, Interpretability and Non-Discrimination	70 to 82
IX	Legal, Ethical and Regulatory Foundations	83 to 92
X	AI Assurance, Evidence and Digital Trust	93 to 101
XI	Fairness in Organizational Practice	102 to 111
XII	Emerging Challenges and Research Frontiers	112 to 115

The twelve parts

Part I: Foundations of Trust, Justice and Non-Discrimination

Part I sets out the concepts on which every later part relies. It begins with trust, a willingness to accept vulnerability that is warranted only when it tracks the other party's actual trustworthiness, and, for AI, with the goal of appropriate reliance calibrated to a system's actual, group-specific capability (Chapters 1 and 2). It presents fairness as a family of

contested conceptions of justice, the consequentialist, deontological, virtue-based and non-Western ethical frameworks used to evaluate a system, and the distinct ideas of equality, equity and human dignity (Chapters 3 to 5). Three chapters trace how discrimination persists through technology: recurring historical mechanisms, data representation as a social practice, and the technical pathways of biased labels, proxies and feedback loops (Chapters 6 to 8). Two chapters ground fairness in human rights, contestability and remedy, and in accountability and explanations that are faithful for every group (Chapters 9 and 10). The part closes with the Historical Context Assessment Method, which turns knowledge of historical discrimination into a documented, prioritized list of system-specific risks (Chapter 11).

Part II: Formal Foundations of Algorithmic Fairness

Part II gives the compendium its mathematical language. It defines the objects every fairness assessment works with (features, protected attribute, outcome, score, prediction and decision) and the per-group rates computed from them (Chapters 12 and 13). It presents the three families of group fairness criteria, independence (demographic parity, disparate impact), separation (equal opportunity, equalized odds) and sufficiency (predictive parity, calibration), followed by individual fairness and causal and counterfactual criteria (Chapters 14 to 18). The impossibility theorems of Kleinberg, Mullainathan and Raghavan and of Chouldechova show that, with unequal base rates and imperfect prediction, these criteria cannot all hold at once (Chapter 19); the resulting trade-offs are analyzed with frontiers and worked examples (Chapter 20). The criteria are extended to scores, regression and ranking (Chapter 21), and the part closes with the Fairness Definition Selection Method, a documented procedure for choosing criteria in a specific legal and social context (Chapter 22).

Part III: Bias, Discrimination and Causality

Part III explains where unfairness in AI systems comes from. It begins with the legal distinction between direct discrimination (disparate treatment) and indirect discrimination (disparate impact), and with systemic discrimination that reaches a model through targets, labels, samples and features; it shows why removing protected attributes does not remove their influence, because other variables act as proxies (Chapters 23 and 24). It then follows bias through the life cycle: historical, representation and measurement bias in data; objectives, regularization and evaluation protocols that favor the largest group; feedback loops that amplify initial disparities; and deployment contexts that turn a model that tests as fair into unequal outcomes (Chapters 25 to 28). Intersectional analysis shows that disparities can fall on combinations of attributes that single-attribute checks miss (Chapter 29). The causal chapters explain why associations cannot reveal the mechanism behind a disparity and introduce causal graphs, structural equation models, counterfactual fairness and inference from observational data (Chapters 30 to 33). The part closes with distribution shift and the Bias Source Identification Method (Chapters 34 and 35).

Part IV: Measuring Fairness: Metrics and Statistical Evidence

Part IV explains how fairness criteria become measured, defensible evidence. It treats a fairness assessment as an empirical study with a stated question, an analysis plan fixed before outcomes are examined and findings reported in three states: beyond tolerance, within tolerance or inconclusive (Chapter 36). It shows how to audit data for representation, proxy associations, label quality and baseline disparities, and how to detect bias in raw data before training (Chapters 37 and 38). It derives the group fairness metrics from the per-group confusion matrix and shows how to implement them so that empty denominators, missing attributes and sparse intersections are handled deliberately (Chapters 39 and 40). The statistical chapters cover confidence intervals, bootstrap and Bayesian methods, corrections for multiple comparisons, effect sizes, permutation and exact tests, robustness checks and the special care small groups need (Chapters 41 to 43). The Fairness Metrics Selection Method links selected criteria to specified, validated metrics and reports (Chapter 44), and the final chapter states what metrics alone cannot establish (Chapter 45).

Part V: Fairness Interventions

Part V explains what can be changed once an assessment finds that an AI system violates the fairness criterion that matters in its context. It distinguishes three intervention points and opens with a documented method for choosing, combining and evaluating interventions, because comparative studies show that no method is best everywhere (Chapter 46). Pre-processing changes the training data: reweighting and resampling adjust how strongly records count, distribution transformation and feature engineering remove the proxy information that seemingly neutral features carry, and fairness-aware data generation adds synthetic or counterfactual records (Chapters 47 to 50). In-processing changes how the model learns, through constraints enforced by reductions or Lagrangian methods, fairness penalties added to the loss, adversaries that try to recover the protected attribute, and multi-objective optimization over the Pareto frontier (Chapters 51 to 54). Post-processing changes how a trained model's outputs are used: group-specific thresholds, group-wise recalibration, and transformations that align, re-rank or defer predictions to human decision makers (Chapters 55 to 57). Each chapter states which criteria a method can target, what it requires, what it costs and how its effect is verified.

Part VI: Fairness Engineering and Operations

Part VI explains how fairness moves from a one-time analysis to an engineering and operational discipline. It starts with tooling: the main open-source libraries (AI Fairness 360, Fairlearn, Aequitas, the What-If Tool and Fairness Indicators), why the same metric name can yield different numbers, and how a thin abstraction layer composes them (Chapter 58). It then builds fairness into daily development through experiment tracking, pull requests, automated tests and review checklists, and into CI/CD pipelines as versioned tests with thresholds fixed in advance, explicit handling of statistical noise and tiered responses to failures (Chapters 59 and 60). In production, fairness is tracked as a time series per group and per intersection, real drift is separated from noise with sequential tests and adaptive thresholds, and alerts are prioritized and routed (Chapters 61 and 62). Dashboards carry the results to executives, engineers, auditors and affected communities while protecting small groups (Chapter 63), and controlled experiments with multi-objective criteria and subgroup power test whether a fix works and whom it might harm (Chapter 64).

Part VII: Fairness in Advanced AI Architectures

Part VII explains how fairness work changes when systems learn their own representations, rank items, write text, combine modalities and act. Deep networks can encode protected attributes through proxies, amplify them through the layers and inherit them from pretrained components, so assessment probes representations as well as outputs (Chapter 65). Recommenders allocate scarce attention among users and providers, and feedback loops let small biases grow, so fairness is measured as exposure, quality across user groups and diversity over time (Chapter 66). Large language models produce open-ended text, so classification metrics give way to counterfactual consistency, bias benchmarks, red teaming and human evaluation (Chapter 67). Vision and multi-modal systems fail unevenly under capture conditions such as lighting and can compound bias in fusion, so evaluation is stratified by condition and intersection (Chapter 68). Generative and agentic systems add stereotyped content, disparities that compound across steps and failures shared by everyone who uses the same foundation model (Chapter 69). Each chapter includes an illustrative case and the relevant EU requirements.

Part VIII: Explainability, Interpretability and Non-Discrimination

Part VIII treats explainability as a family of distinct demands and examines what explanations do for fairness. It defines interpretability, transparency and fidelity, and shows that explanations can help detect, analyze and contest unfairness but can also be unfaithful for some groups, manipulated to hide bias or mistaken for proof of fairness (Chapters 70 and 71). It classifies the methods and explains feature attribution (Shapley values, LIME, permutation importance, integrated gradients), counterfactual explanations and algorithmic recourse, and causal explanation, including the decomposition of a disparity into direct, indirect and spurious effects (Chapters 72 to 75). It then explains disparities at the level of groups and intersections, measures the fidelity and robustness of explanations, and

tests their effect on human reliance (Chapters 76 to 78). The legal chapter analyzes the GDPR, the SCHUFA and Dun & Bradstreet Austria judgments and the EU AI Act (Chapter 79). The part closes with explanation in generative and agentic systems, explanations as assurance evidence, and their limits and risks (Chapters 80 to 82).

Part IX: Legal, Ethical and Regulatory Foundations

Part IX explains what the law requires of AI systems that make or support decisions about people, and where the law stops. It begins with anti-discrimination law: which attributes are protected, how disparate treatment and disparate impact (direct and indirect discrimination) are established and defended, and why no fairness metric on its own establishes compliance (Chapter 83). International human rights law supplies the deeper layer, defining discrimination by purpose or effect and permitting only objectively and reasonably justified distinctions (Chapter 84). The part then maps the AI-specific regulation: the global landscape (Chapter 85), the EU AI Act as amended by the Digital Omnibus and GDPR Article 22 (Chapter 86), the Swiss constitutional, statutory and data protection framework (Chapter 87), United States sector statutes and the federal policy changes of 2025 and 2026 (Chapter 88), and the OECD, UNESCO, NIST and ISO/IEC instruments (Chapter 89). Two chapters show how risk classification and sectoral rules decide which obligations apply (Chapters 90 and 91). The final chapter treats compliance as a moving floor and sets out how to exceed it and track legal change (Chapter 92).

Part X: AI Assurance, Evidence and Digital Trust

Part X explains how fairness measurement becomes assurance that others can rely on. It begins with the four gaps between a metric and a claim (construct, inference, time and context), the roles of first, second and third parties, and the way the EU AI Act, the NIST AI RMF and the New York City bias audit rule turn measurement into assurance obligations (Chapter 93). Assurance cases in GSN and CAE make the reasoning from evidence to claim reviewable, with defeaters exposing doubt (Chapter 94). Evidence is planned from the requirements, traced to each obligation and kept in tamper-evident audit trails, and independence and reproducibility let a party without a stake check a finding (Chapters 95 and 96). The digital trust chapters cover verifiable credentials with selective disclosure, identity, provenance and integrity, validity periods, revocation and status lists, and the registries that decide who may vouch for an assessment, with the Swiss swiyu infrastructure as a case (Chapters 97 to 100). The part closes with eight limits of certification and a method for assurance that states its own bounds (Chapter 101).

Part XI: Fairness in Organizational Practice

Part XI shows how fairness is carried through the way organizations build, govern and repair AI systems. It first extends agile delivery: Fair AI Scrum adds fairness to values, events, artifacts and roles; fairness user stories and acceptance criteria name the groups, the definition, the metric, the tolerance and the test; Sprint Planning reserves capacity and orders work by harm as well as value; and reviews, retrospectives and mid-Sprint checkpoints keep fairness in view (Chapters 102 to 105). It then sets out organizational structures: fairness roles with one accountable owner per decision, decision records and layered communication, and governance mechanisms with decision tiers, escalation, gates and a fairness council scaled to risk (Chapters 106 to 108). An eight-stage incident life cycle covers detection, containment, root cause analysis, remediation and redress for affected people (Chapter 109). The Fairness Audit Playbook combines four assessment methods into one traceable workflow (Chapter 110), and the case studies separate documented cases from constructed scenarios across eight domains (Chapter 111).

Part XII: Emerging Challenges and Research Frontiers

Part XII treats the questions on which the science of AI fairness is still moving. It first examines large language models and generative AI: the representational and allocational harms they can cause, the formal desiderata and three families of bias metrics proposed for them, the evidence that many of these metrics do not predict harm in real use, and the EU AI Act obligations for general-purpose models (Chapter 112). It then turns to autonomous agents that act for people over many steps, where disparities can compound along a trajectory, third parties are affected, and

responsibility, human control and the ability to stop an agent become central (Chapter 113). Adaptive systems and collective impacts follow: performative prediction, strategic classification, delayed impact, systems that retrain on their own effects, algorithmic monoculture and the collective and societal harms addressed by the Digital Services Act and the AI Act (Chapter 114). The closing chapter integrates the compendium's open questions into a research agenda with six cross-cutting tensions, ten research programs, priorities and methodological standards for credible evidence (Chapter 115).

The appendix volumes

The three appendix volumes compile reference material from the chapters; they add no new findings. Appendix Volume 1 holds the controlled glossary, with one consolidated definition for each term defined across the 115 chapters (Appendix A), the mathematical notation fixed in Chapters 12 and 13 (Appendix B), and the index of definitions and cross-references, which records where each of the 802 defined terms is defined (Appendix L). Appendix Volume 2 holds the fairness metrics catalog (Appendix C), the explainability methods catalog (Appendix D), the statistical methods (Appendix E) and the worked numerical examples (Appendix H). Appendix Volume 3 holds the legal and regulatory matrix, compiled by jurisdiction as of November 2026 (Appendix F), the standards and normative references (Appendix G), the templates, checklists and method worksheets in the order of the fairness audit workflow (Appendix I), the evidence and assurance-claim taxonomy (Appendix J), and the bibliography and source register of the 1173 distinct sources cited in the chapters (Appendix K).

What the parts establish together

Read together, the parts support a small number of conclusions that no single chapter states alone. Fairness in AI is a normative and historical question before a technical one (Part I). Fairness criteria are precise statistical statements, and the mathematics shows which of them can hold together, but choosing among them remains a normative decision (Part II). Group metrics show that a disparity exists, while causal models explain why and where an intervention can remove it (Part III). A metric becomes evidence only when it is computed correctly, reported with its denominators and uncertainty and scoped to what quantitative measurement can establish (Part IV). The right intervention point follows from where a disparity enters (Part V), and fairness becomes dependable only when it is executable and monitored in production (Part VI).

In advanced architectures, unfairness lives in learned representations, exposure, generated content, fused modalities and multi-step actions (Part VII), and explanations show how a model works, not that it is fair (Part VIII). Law sets a mandatory but moving floor (Part IX), and a metric becomes assurance only inside a scoped, argued claim backed by preserved evidence and independent checking (Part X). Fairness becomes organizational practice only when delivery, roles, documentation, governance and incident handling carry explicit fairness content (Part XI). At the frontier, valid methods for producing evidence about open-ended outputs, delegated action and systems acting over time are still being built (Part XII).

KEY TAKEAWAY · ONE LINE OF ARGUMENT

Each part answers one question in a sequence: why fairness matters and what it means, how it is defined, where unfairness comes from, how it is measured, how it is corrected, how it is kept in place, how it is explained, what the law requires, how claims about it are assured, how organizations carry it, and what remains unknown.

FRONT MATTER

Scope, Audience and Limitations

KEY MESSAGE

The compendium covers fairness, non-discrimination, explainability and assurance for AI systems that make or support decisions about people; it informs judgment but is not legal advice, does not certify any system and states law as of November 2026.

Purpose and scope

This chapter sets the boundaries of the compendium. It describes the subject matter and the kinds of systems, domains and jurisdictions covered, the readers for whom the text is written, and the limitations a reader must keep in mind when relying on any chapter. It was compiled from the compendium's chapters, each of which states its own limitations in a section on competing interpretations and limitations; the statements below name the chapter where each limitation is argued in full. A reader should use this chapter before relying on a finding for a decision, and return to the cited chapter for the detail.

Subject matter

The compendium treats fairness and non-discrimination in AI systems that make or support decisions about people, together with the explainability, assurance and governance needed to make fairness claims checkable. Four themes run through it.

- **Fairness and non-discrimination.** The normative conceptions of fairness and justice (Chapters 3 to 5), the formal criteria and their conflicts (Chapters 12 to 22), the legal forms of discrimination and the role of proxies (Chapters 23 and 24), the sources of bias across the life cycle (Chapters 25 to 35), measurement and statistical evidence (Chapters 36 to 45) and interventions (Chapters 46 to 57).
- **Trust and explainability.** Trust and appropriate reliance (Chapters 1 and 2), and explanations as a means of detecting, contesting and remedying unfairness, with their limits (Chapters 70 to 82).
- **Law and assurance.** Anti-discrimination law, human rights law and AI-specific regulation (Chapters 83 to 92), and the step from evidence to claims that others can rely on, including verifiable credentials and trust registries (Chapters 93 to 101).
- **Practice and research.** Engineering and operations (Chapters 58 to 64), organizational practice (Chapters 102 to 111) and open research questions (Chapters 112 to 115).

Systems and life cycle

The compendium covers binary classifiers, scores, regression and rankings (Chapter 21), deep learning systems, recommendation and ranking systems, large language models, multi-modal and vision models, and generative and agentic systems (Chapters 65 to 69), as well as autonomous agents and adaptive systems that reshape the

populations they act on (Chapters 113 and 114). It follows a system from data collection and design through training, evaluation and deployment to monitoring, drift detection and incident handling (Chapters 25 to 28, 61, 62 and 109).

Domains

Examples and case studies come from credit and lending, employment, insurance, healthcare, education and admissions, criminal justice, public services and benefits, platforms and pricing, and vision systems. The same technique faces different obligations in each sector (Chapter 91), and the applied case studies are organized across eight domains (Chapter 111).

Jurisdictions

The legal chapters treat anti-discrimination law in general terms (Chapter 83), international human rights law (Chapter 84), the global regulatory landscape (Chapter 85), the European Union, in particular the AI Act and GDPR Article 22 (Chapter 86), Switzerland (Chapter 87), the United States (Chapter 88) and international standards and governance principles (Chapter 89). Other jurisdictions, among them Canada, China and South Korea, are treated in the survey of the global landscape (Chapter 85).

Audience

The compendium is written for readers who build, assess, govern or study AI systems that affect people:

- **Data scientists and machine learning engineers**, who compute fairness metrics, implement interventions and build them into pipelines.
- **Auditors and assessors**, internal or independent, who examine a system and report findings that others rely on.
- **Risk and compliance officers and legal counsel**, who decide which obligations apply and how they are evidenced.
- **Product owners and managers**, who decide which trade-offs are acceptable and own fairness decisions in delivery.
- **Researchers**, who need a consolidated statement of what is established and what remains open.

TABLE 4. Where each audience can start

READER	START WITH	THEN
Data scientist	Chapters 12, 39 and 40	Chapters 41 to 43, 46, 58 to 60
Auditor or assessor	Chapters 36 and 110	Chapters 11, 22, 35, 44, 93 to 96
Risk or compliance officer	Chapters 83, 86 and 90	Chapters 87, 88, 91, 95, 108
Product owner	Chapters 20 and 22	Chapters 103, 104, 106, 108
Researcher	Chapter 115	Open questions of each chapter

Readers new to the formal treatment can begin with Part II, which introduces the notation and probability tools that every later chapter uses (Chapters 12 and 13); readers new to the law can begin with Part IX, which introduces the legal vocabulary that the legal chapters use (Chapter 83). The chapter on how to use the compendium gives fuller reading paths.

What the compendium does not do

METHODOLOGICAL POSITION · A REFERENCE WORK, NOT ADVICE OR CERTIFICATION

The compendium informs professional judgment. It is not legal advice for a specific case, it does not certify or approve any AI system, and it does not prescribe one fairness criterion for all contexts. Applying it to a system remains the responsibility of the people who build, deploy and assess that system.

It does not prescribe a single criterion because, when base rates differ and prediction is imperfect, the main criteria cannot all hold at once, so the choice must be made for each decision context (Chapters 19 and 22). It does not certify because no certificate can show that an AI system is fair in general; assurance supports only a bounded claim (Chapter 101). Privacy, data protection and security are treated where they bear on fairness, for example the lawful use of sensitive data for bias detection (Chapters 24 and 86) and the leakage of private information through explanations (Chapter 82), not as subjects in their own right.

Limitations

What metrics cannot establish

Fairness metrics measure statistical properties of recorded variables; they cannot alone establish valid measurement of constructs, whether a disparity is unjust, why it arose or how a system behaves over time, so every quantitative finding must be scoped and complemented (Chapter 45). A finding drawn from a sample is an estimate, and small groups are where fairness harms are most likely to hide and hardest to measure (Chapters 41 and 43).

Causal claims rest on assumptions

Causal fairness criteria gain their power only from an assumed causal model, and designating fair pathways remains a normative choice (Chapter 18). Causal conclusions almost always rest on observational data and incomplete causal knowledge; they are trustworthy only when each method's identifying assumptions are stated, several methods are compared, sensitivity to unmeasured confounding is quantified and uncertainty is reported as bounds (Chapter 33).

Explanations do not prove fairness

Explanations can be unfaithful for some groups, manipulated to hide bias or mistaken for proof of fairness (Chapter 71). They show how a model works, not that it is fair (Chapter 81), and they have hard epistemic limits (Chapter 82).

Assurance is bounded

An assurance claim holds for a stated subject, at a stated time, until a stated date (Chapter 99). A verifiable credential proves who said what about whom and that nothing changed, not that the statement is true (Chapter 97). The eight limits of certification, from scope and depth of access to incentives and an immature audit ecosystem, apply to any scheme (Chapter 101).

Scenarios are constructed, cases are documented

Many chapters use worked examples and scenarios to show how a method responds. The applied case studies distinguish documented cases, checked against court and agency records, peer-reviewed papers and official

statements, from constructed scenarios whose outcomes are illustrative (Chapter 111). A reader must always know which of the two is being read; a constructed scenario is not evidence that the outcome it describes occurred.

The research frontier is unsettled

For large language models and generative AI, there is evidence that many proposed bias metrics do not predict harm in real use (Chapter 112). Methods for agents, adaptive systems and collective impacts are still being built (Chapters 113 and 114), and the integrated research agenda separates what is established from what rests on working assumptions (Chapter 115). Statements in these areas are marked as open where they are open.

The law changes

LEGAL OR NORMATIVE REQUIREMENT · LEGAL STATUS AS OF NOVEMBER 2026

Legal content is stated as of November 2026 and should be checked against the current law before it is relied on. Regulation (EU) 2026/1744, the Digital Omnibus on AI, published in the Official Journal on 24 July 2026, amended the AI Act: it moved the legal basis for processing special categories of personal data for bias detection and correction from Article 10(5) into a new Article 4a, replaced the AI literacy obligation of Article 4, and moved the application dates of the high-risk rules (Chapter 86).

AI fairness regulation is diverse, fast-changing and in places moving in opposite directions (Chapter 85). In the United States, federal policy since 2025 opposes disparate impact while state and city laws fill gaps (Chapter 88), and Switzerland plans targeted laws implementing the Council of Europe AI Convention instead of a Swiss AI Act (Chapter 87). Legal compliance is a moving floor, not a measure of fairness (Chapter 92).

Limits of the editorial process

The compendium was drafted in part with AI assistance under human editorial responsibility, and every chapter was fact-checked against its sources. A fact-check reduces errors; it does not exclude them. Some fact-check runs recorded open issues, and some could not complete the automated database lookup and confirmed the affected references by hand from primary sources instead (the chapter on scientific and editorial method). Corrections are recorded in the version history of each part.

FRONT MATTER

Scientific and Editorial Method

KEY MESSAGE

Every chapter was written from identified sources under fixed rules, reviewed by a scientific editor and fact-checked: references looked up in Crossref, OpenAlex or DBLP or confirmed from a primary source, and claims and formulas rechecked.

Purpose and scope

This chapter describes, factually, how the compendium was produced and checked: the sources it draws on, the rules every chapter was written under, the review and fact-check each chapter went through, the classification of statements, and the versioning of the published parts. It was compiled from the compendium's master structure and writing guide and from the records that the fact-checks of the 115 chapters left. Readers who want to judge how far a statement can be relied on should read this chapter together with the limitations in the chapter on scope, audience and limitations.

Sources

Three kinds of sources

The compendium rests on three kinds of sources.

1. **A review of the scientific literature** on algorithmic fairness, non-discrimination, causal inference, statistical evidence and explainability: the papers, books and surveys that define the criteria, prove the results and report the empirical findings the chapters present.
2. **Practical and industry best practice**: documented assessment and audit methods, open-source fairness and explainability toolkits and their documentation, standards and frameworks (for example ISO/IEC, NIST, OECD and UNESCO), legislation and case law, regulatory guidance, and published practice in organizations that build, deploy and audit AI systems.
3. **The knowledge base collected by validant.ai** on fairness, non-discrimination and explainability, together with its knowledge bases on trustworthy AI, digital trust and self-sovereign identity, including structured notes on research papers.

A master structure fixes every chapter number and assigns each topic to exactly one chapter as its primary home, which treats it in full; other chapters refer to that chapter instead of repeating it. Nothing from these sources was taken on trust: every formula, theorem statement, attribution and number was checked, corrected where wrong and dropped where it could not be verified, and each such case was logged. Primary sources are preferred over secondary summaries: the paper itself, the publisher, official legislation portals such as EUR-Lex and fedlex.admin.ch, the U.S. Code and Federal Register, and the standards and intergovernmental bodies themselves.

Coverage ledgers

Each writer kept a coverage ledger for its chapters. It lists every source unit used, with every concept, example and question in it and the chapter and section where it now lives, every factual or formula correction with its reason, every removed claim or citation with its reason, and every reference with how it was verified. The ledgers make it possible to check that no source content was silently dropped.

Writing rules

All chapters were written under one writing guide. Its content rules are the following.

1. **Carry the full content of each topic.** Every concept, definition, formula, argument, example, case study, question, table and list that belongs to a chapter's topic appears in that chapter, restructured and in the compendium's own wording.
2. **Same science, own words.** Sentences are rephrased, never copied. Fixed technical terms are kept exactly (for example demographic parity, equalized odds, calibration within groups, counterfactual fairness, base rate, protected attribute, proxy variable), and definitions, formulas and the strength of claims are neither weakened nor altered.
3. **Grounding.** Every concrete fact (a number, a statistic, a date, a finding, a court case and its outcome, a statute and article, a standard and clause, a quotation) must come from a source opened while writing or checking the chapter, and is cited. A fact that no opened source confirms is left out. No statistic, author, year, page range, DOI or quotation is estimated.
4. **Real citations.** Every cited work is verified for author, year, title and venue, and every reference carries a DOI or a stable URL. References follow APA 7.
5. **One notation.** All chapters use the notation fixed in Chapter 12: X for features, A for the protected attribute, Y for the true outcome, R for a score, $\hat{Y}$ for a prediction, D for a decision, groups a and b, and per-group quantities written with a subscript, such as TPR_a and p_a. Appendix B lists every symbol.
6. **Methods of the compendium.** Methods are named as methods of the compendium (for example the Fairness Definition Selection Method), and cross-references point to compendium chapters.
7. **Voice and form.** American English, third person, short paragraphs, terms defined before use, concrete examples with their domain named, and analogies marked as analogies.

Every chapter follows one template: an overview with learning objectives, key questions, definitions and terminology, the body, a section on competing interpretations and limitations, common questions with answers, conclusions, open questions and references. Part II served as the approved model for depth, structure and tone.

Review and fact-check

Scientific editor review

Each chapter was read in full by a scientific editor against the writing guide, the master structure, the coverage ledger and the notation of Chapter 12. The editor spot-checked items of the ledger against the sources it names, recomputed worked examples, and recorded each change with its location and reason in a review report for each group of chapters.

Independent fact-check

Every chapter was then fact-checked in a separate run. The fact-check has three parts.

- **References.** Each entry in a chapter's reference list was looked up by a script. A DOI in the entry is resolved at Crossref, and the title and year must match. Otherwise Crossref, OpenAlex and DBLP are searched, and a hit counts only when the title matches with a similarity of at least 0.88, the year is within one year and the first author's surname appears among the hit's authors. The script never marks a reference verified on a weak match. An entry the script cannot match, and any book, statute, judgment, standard or guideline, was confirmed from its primary source (publisher, legislation portal, court, standards body) or removed. The script also checks that every in-text citation has a reference entry with that surname and year, and that every entry is cited.
- **Claims.** Concrete claims were checked against the cited work or an official source opened in the run, and corrected or removed where the source did not support them.
- **Formulas.** Display formulas were rechecked, and worked examples were recomputed.

When an external database rate-limited the automated lookup, the affected run completed the database step with the same matching rules where it could and confirmed the remaining entries by hand from primary sources; the run records say so.

Counts recorded by the fact-checks

The fact-check of the 115 chapters ran in 37 runs, each covering two to four chapters. The table below adds up the counts that the runs recorded. A reference entry is counted once for each chapter that lists it, so a work cited in several chapters is counted several times; the bibliography and source register lists 1173 distinct sources (Appendix K).

TABLE 5. Totals of the counts recorded by the 37 fact-check runs

ITEM	CHECKED	OUTCOME
Reference entries	2268	1154 database-verified, 1109 confirmed from a primary source, 5 removed
Concrete claims	2867	337 corrected or removed
Formulas	861	29 corrected

Source: sums of the counts in the completion records of the fact-check runs, Chapters 1 to 115, 5 November 2026.

Four runs recorded a total of nine open issues that they could not resolve within the run; the run reports list them.

KEY TAKEAWAY · WHAT THE COUNTS SHOW AND WHAT THEY DO NOT

The counts show that every reference entry was either matched in a bibliographic database, confirmed from a primary source or removed, and that 337 of 2867 claims (about 12 percent) and 29 of 861 formulas (about 3 percent) were corrected or removed in the process. They do not show that the text is free of error; a fact-check reduces errors, it does not exclude them.

Format and build checks

Each part is built from its chapters by a script that rejects headings below the third level, inline dollar mathematics, unclosed display equations, tables without a caption, diagrams without a caption, unknown box types, em dashes and double hyphens. A part builds only when every chapter passes. Charts are drawn from the exact data or function the chapter states, and diagrams and illustrations carry a caption and their source.

Classification of statements

The compendium distinguishes four kinds of statements and marks important ones in framed boxes, so that a reader can see at a glance what kind of authority a statement claims.

TABLE 6. Classes of statements and their boxes

BOX	CLASS	WHAT IT MARKS
Established finding	ESTABLISHED	A proven result or a replicated scientific finding
Legal or normative requirement	LEGAL	A requirement of a named legal or normative instrument, with its article
Methodological position	METHOD	A methodological position or recommended practice of validant.ai
Open question	OPEN	An unresolved or contested question

Three further boxes structure the text without claiming authority of their own: a definition box (DEFINITION), an example box (EXAMPLE) and a key takeaway box (KEY). Boxes are used sparingly, about one to four per major section. Statements outside boxes carry the authority of their citations. The classes are meant to be kept apart: a methodological position of validant.ai is not an established finding or a legal requirement.

METHODOLOGICAL POSITION · KEEP THE CLASSES APART

An assessor who relies on the compendium should keep the four classes apart in their own reports: a theorem does not become a legal requirement, a legal requirement does not settle what is fair, and a recommended practice remains a choice that must be justified for the system at hand.

Versioning

Each of the twelve parts is published as a standalone document with its own identifier (VAI-COMP-P01 to VAI-COMP-P12), a version number, a date and a version history. A part that would exceed 5 MB as a PDF or DOCX file is split into volumes, each a standalone document; Parts III, V and VIII are published in two volumes. The version history records the changes made to a part after its first edition. For example, Part II records an independent fact-check in its version 1.1 and, in its version 1.4, the citation of the AI Act's special-category data permission for bias detection as Article 4a, inserted by Regulation (EU) 2026/1744 in place of the former Article 10(5). Legal content carries the date of the law it states, November 2026 (Part IX).

Chapter numbers are fixed across all versions and parts, so that a citation of a chapter remains valid when a part is revised.

Use of AI assistance and editorial responsibility

Parts of the text were drafted with AI assistance under human editorial responsibility. No draft was published as drafted: every chapter was reviewed and fact-checked against its sources as described above, references were looked up in Crossref, OpenAlex or DBLP or confirmed from a primary source, and claims and formulas were

rechecked. The author, Daniel Glinz, carries editorial responsibility for the whole text. The compendium is published by Glinz & Company GmbH under the name validant.ai.

Limitations of the method

The method has limits that a reader should know. A database match confirms that a work exists with that title, year and first author; it does not confirm that the work supports the sentence that cites it, which is why claims were checked separately. Claims were checked against the sources a run could open, and some references were confirmed from the publisher's page (title, authors, venue) rather than from the full text. Although every source was checked section by section, a reader should treat any statement without a citation as the compendium's own synthesis. Legal content is current only to its stated date.

FRONT MATTER

How to Use This Compendium

KEY MESSAGE

The compendium can be read by part, by task or by question: fixed chapter numbers, one chapter template, classified statements and four linked assessment methods let a reader enter where needed and follow cross-references to the evidence.

Purpose and scope

This chapter is a practical guide to the compendium. It explains how the parts and chapters are organized, offers reading paths for common tasks and readers, maps the foundational questions of fairness practice to the chapters that answer them, and explains the conventions a reader meets in every chapter. It was compiled from the reading guides that open each part and from the compendium's chapters. A reader who knows the task at hand can go straight to the reading paths; a reader new to the subject can start with the Executive Overview and the first chapters of Parts I and II.

How the compendium is organized

The compendium has twelve parts with 115 chapters and three appendix volumes. Each part is a standalone document that opens with a short account of the part, its position in the whole and a reading guide, and each can be read on its own. Parts III, V and VIII are each published in two volumes because of their length. Chapter numbers run continuously from 1 to 115 and are fixed across all parts and versions, so a cross-reference such as "see Chapter 32" always names the same chapter. The full table of contents follows the front matter.

The chapter template

Every chapter follows one template, so that a reader always knows where to look.

TABLE 7. Sections of a chapter and their use

SECTION	WHAT IT GIVES THE READER
Overview	What the chapter covers, why it matters, its neighbors, learning objectives
Key questions	The foundational questions the chapter answers
Definitions and terminology	The terms the chapter uses, defined before use
Body sections	Foundations, formal treatment, methods, worked examples, practice
Competing interpretations and limitations	Where the field disagrees and what the chapter's claims do not cover
Common questions	Questions practitioners ask, with answers
Conclusions and open questions	Normative and methodological conclusions, and research gaps
References	Every cited work in APA 7, with a DOI or stable URL

Each section is written to make sense when read on its own, so a section found by search can be used without reading the whole chapter; its definitions and cross-references point to the rest.

Reading paths by task

TABLE 8. Where to start for common tasks

TASK	START WITH	CONTINUE WITH
Start a fairness assessment of a system	Chapter 36	Chapter 110 and its four methods
Identify historical risks for a use case	Chapter 11	Chapters 6 to 8
Choose fairness criteria	Chapter 22	Chapters 14 to 20
Locate where bias enters	Chapter 35	Chapters 25 to 29
Audit data before training	Chapters 37 and 38	Chapters 24 and 43
Choose and specify metrics	Chapter 44	Chapters 39 to 43, Appendix C
Report a quantitative finding	Chapter 45	Chapters 41 and 42
Choose an intervention	Chapter 46	Chapters 47 to 57 by intervention point
Fix a model that cannot be retrained	Chapters 55 to 57	Chapter 46
Automate fairness checks	Chapters 59 and 60	Chapter 58
Monitor a deployed system	Chapters 61 and 62	Chapters 63 and 64
Assess a large language model or agent	Chapters 67 and 69	Chapters 112 and 113
Use explanations in an assessment	Chapters 76 and 77	Chapters 71 and 81
Meet a legal duty to explain	Chapter 79	Chapter 86
Decide which obligations apply	Chapters 90 and 91	Chapters 85 to 88
Make a claim others can rely on	Chapter 93	Chapters 94 to 96

TASK	START WITH	CONTINUE WITH
Judge a certificate or audit report	Chapter 101	Chapters 97 to 100
Organize fairness work in a team	Chapter 102	Chapters 103 to 105
Set up fairness governance	Chapter 106	Chapters 107 to 109

Reading orders within parts

Some chapters build on one another and are best read in sequence; the reading guide of each part states which. Chapters 12 and 13 fix the notation that every later chapter uses. The causal chapters, Chapters 30 to 33, build on one another and on Part II (Part III). The statistical chapters, Chapters 41 to 43, are best read in order, and Chapter 43 is essential wherever groups or intersections are small (Part IV). Chapter 62 assumes the tracking set up in Chapter 61 (Part VI). Chapter 69 assumes Chapters 67 and 68 (Part VII). Chapters 70 and 71 are best read before the rest of Part VIII, Chapters 97 to 100 form one sequence on digital trust infrastructure (Part X), Chapters 102 to 105 form one sequence on agile delivery (Part XI), and Chapter 115 is best read last (Part XII).

Shared cases

Several parts follow one case through their chapters, so that assessments can be compared. Chapters 65, 67 and 68 use a university admissions case and Chapter 66 a university course recommender, and Chapter 69 uses an agent in first-stage hiring (Part VII). Chapters 102 to 105 share a university admissions case, which Chapters 106 to 108 continue at the level of the organization, and Chapter 110 follows a loan application system through every phase of an audit (Part XI). Chapter 111 lists the case studies used across the compendium with the chapter that treats each in full.

The questions the compendium answers

The questions below are the foundational questions of fairness practice, from diagnosis through treatment to organizational and engineering implementation. Each is answered in full in the chapters named.

TABLE 9. Foundational questions and where they are answered

QUESTION	CHAPTERS
How do historical patterns of discrimination appear in technology, and which apply to a given system?	6 to 8, 11
How are abstract ideas of fairness turned into precise definitions, and why do they conflict?	3, 12 to 20
Where and how does bias enter the machine learning life cycle?	25 to 29, 35
How is fairness quantified, validated and communicated?	39 to 44
How do fairness definitions translate into technical interventions?	30 to 32, 46
Where in the pipeline should one intervene?	46
Which technical approaches address which types of bias?	47 to 57
How are fairness improvements balanced against performance?	20, 52, 54
How do existing fairness libraries differ, and how are they combined?	58
Why are fairness metrics hard to compute correctly at scale?	40, 43

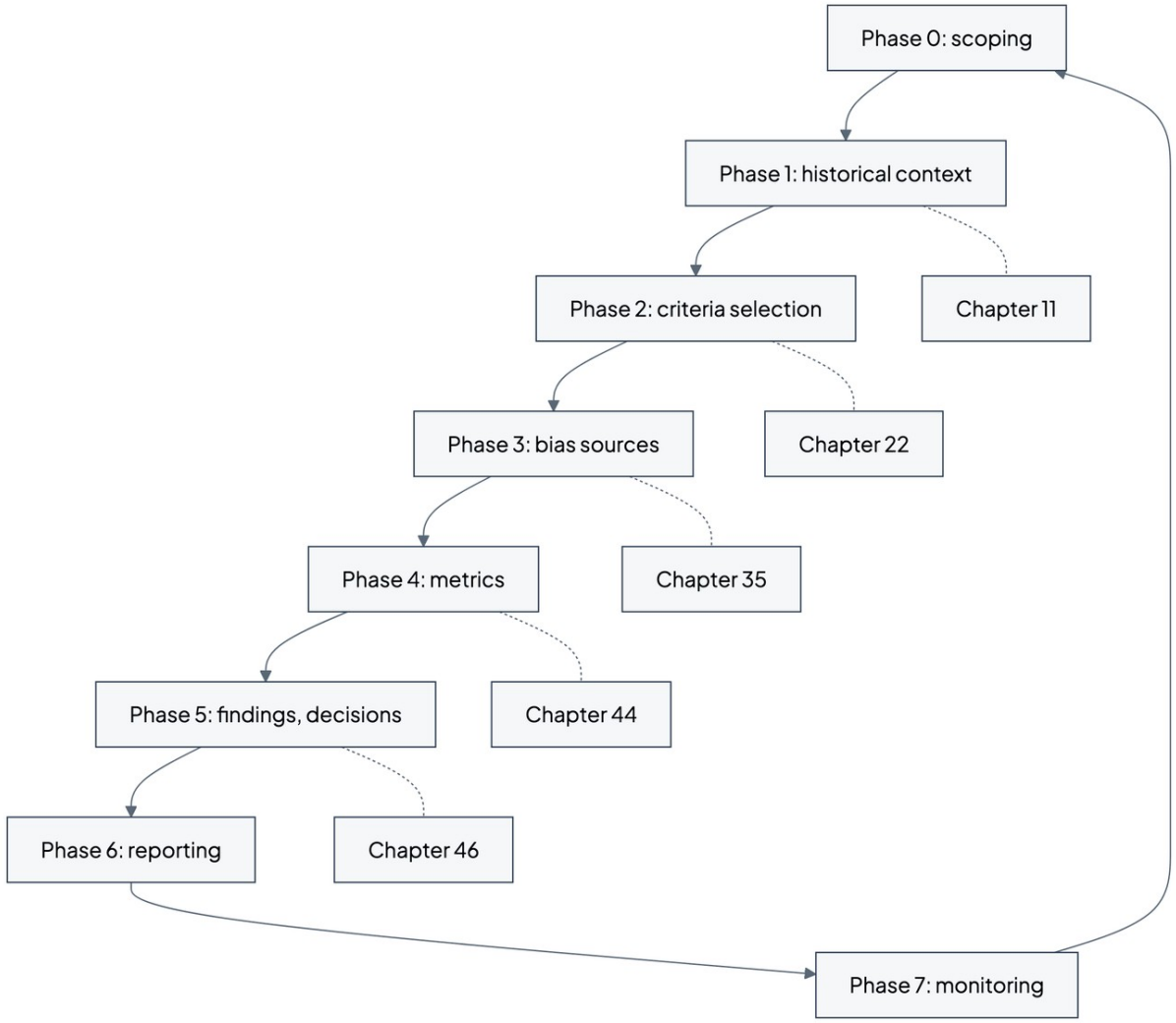
QUESTION	CHAPTERS
Where in data pipelines should bias detection and mitigation occur?	38, 47 to 49, 60
What threatens fairness after deployment, and how is it monitored?	34, 61 to 64
What do complex architectures such as language, ranking and vision models require?	65 to 69
How are agile practices redesigned to carry fairness work?	102 to 105
Who owns fairness outcomes, and how is progress tracked?	63, 106 to 108
How do regulatory requirements become development and governance practice?	85, 86, 90, 95

Interventions are tied to the target criterion, though less tightly than is often claimed: group-specific thresholds, for example, can be chosen to equalize selection rates (demographic parity) or true positive rates (equal opportunity) (Chapters 46 and 55). The intervention point should follow a causal diagnosis of where a disparity enters, because without that diagnosis an intervention may harm the group it is meant to protect or create new problems while addressing the symptom (Chapters 30 and 46). And the trade-off between fairness and predictive performance should be measured on the system rather than assumed (Chapters 20 and 54).

The assessment methods and the audit workflow

Four methods of the compendium carry an assessment from context to evidence, and the Fairness Audit Playbook combines them into one workflow with scoping, decision, reporting and monitoring phases (Chapter 110). A fifth method selects interventions once the audit has found a violation (Chapter 46). The worksheets for the methods are collected in Appendix I in the order of the workflow.

FIGURE 1. The phases of the Fairness Audit Playbook and the chapters behind them



Source: validant.ai.

TABLE 10. The methods of the compendium

METHOD	CHAPTER	WHAT IT PRODUCES
Historical Context Assessment Method	11	A documented, prioritized list of system-specific risks
Fairness Definition Selection Method	22	A documented choice of fairness criteria for the context
Bias Source Identification Method	35	A ranked map of the points where bias enters
Fairness Metrics Selection Method	44	Specified, validated metrics and reports for the chosen criteria
Intervention Strategy Selection Method	46	A chosen intervention point and method
Fairness Audit Playbook	110	One traceable workflow combining the four assessment methods

METHODOLOGICAL POSITION · ENTER ANYWHERE, BUT KEEP THE CHAIN

A reader may enter the workflow at any phase, for example to choose metrics for a system whose criteria were fixed earlier. The documentation should still record the earlier links, because every finding, decision and remedy must trace back to the historical context, the criteria and the bias sources it rests on (Chapter 110).

Conventions in every chapter

Framed boxes

Framed boxes mark the kind of authority a statement claims: an established finding, a legal or normative requirement, a methodological position of validant.ai or an open question. Definition, example and key takeaway boxes structure the text. The chapter on scientific and editorial method explains each class. A reader who needs to know whether a statement is a theorem, a legal duty or a recommended practice should look first for its box.

Notation and terms

All chapters use one notation, fixed in Chapter 12: X for features, A for the protected attribute, Y for the true outcome, R for a score, $\hat{Y}$ for a prediction and D for a decision, with per-group quantities written with a subscript, such as TPR_a for the true positive rate of group a. Appendix B lists every symbol with the chapter that defines it. Fixed technical terms keep one meaning throughout; Appendix A gives the consolidated definition of each and Appendix L shows where each defined term is defined.

Cross-references and references

Cross-references name chapters by number ("Chapter 19"), figures and tables by their subject. Each chapter has its own reference list in APA 7, with a DOI or stable URL for every entry, so a reader can open the source of any claim. Appendix K collects all cited sources in one register with the chapters that cite each.

Legal content

Legal content is stated as of November 2026 and should be checked against the current law before it is relied on (Part IX). In particular, the AI Act provisions on bias detection with special-category data are cited as Article 4a, inserted by Regulation (EU) 2026/1744, and not as the former Article 10(5) (Chapter 86). Appendix F gives the legal and regulatory matrix by jurisdiction.

The appendices

The appendices are for looking up rather than reading. Appendix Volume 1 holds the controlled glossary (A), the mathematical notation (B) and the index of definitions and cross-references (L). Appendix Volume 2 holds the fairness metrics catalog (C), the explainability methods catalog (D), the statistical methods (E) and the worked numerical examples (H). Appendix Volume 3 holds the legal and regulatory matrix (F), the standards and normative references (G), the templates, checklists and method worksheets (I), the evidence and assurance-claim taxonomy (J) and the bibliography and source register (K). The appendix entries name the chapters they are taken from.

Citing the compendium

A part is cited with its author, year, title, part and publisher, and a chapter by its number within the work, for example: Glinz, D. (2026). *The AI Trust & Fairness Compendium, Part II: Formal Foundations of Algorithmic Fairness*. Glinz & Company GmbH (validant.ai), Chapter 19. Each part carries its document identifier (VAI-COMP-P01 to VAI-COMP-P12) and version number; a citation should name the version read, because later versions may correct the text.

FRONT MATTER

Full Table of Contents

KEY MESSAGE

The compendium has twelve parts with 115 chapters and three appendix volumes; each part is a standalone document, and chapter numbers are fixed across the whole work.

Part I: Foundations of Trust, Justice and Non-Discrimination

TABLE 11. Chapters of Part I

CHAPTER	TITLE
1	The Nature of Trust
2	Trust in Artificial Intelligence
3	Conceptions of Fairness and Justice
4	Ethical Frameworks for Fairness Evaluation
5	Equality, Equity and Human Dignity
6	Historical Patterns of Discrimination in Technology
7	Data Representation and Social Context
8	Modern Manifestations of Historical Biases
9	Human Rights, Human Agency and Contestability
10	Fairness, Explainability and Accountability
11	The Historical Context Assessment Method

Part II: Formal Foundations of Algorithmic Fairness

TABLE 12. Chapters of Part II

CHAPTER	TITLE
12	Formalizing Algorithmic Decision-Making
13	Probability and Statistical Foundations
14	Independence: Demographic Parity and Related Criteria
15	Separation: Equal Opportunity and Equalized Odds
16	Sufficiency: Predictive Parity and Calibration
17	Individual Fairness
18	Causal and Counterfactual Fairness Criteria
19	Impossibility Theorems
20	Fairness Trade-offs
21	Fairness Beyond Binary Classification: Scores, Regression and Ranking
22	Contextual Selection of Fairness Criteria

Part III: Bias, Discrimination and Causality

Published in 2 volumes: Volume 1, Chapters 23 to 29; Volume 2, Chapters 30 to 35.

TABLE 13. Chapters of Part III

CHAPTER	TITLE
23	Direct, Indirect and Systemic Discrimination
24	Protected Characteristics, Sensitive Attributes and Proxies
25	Data Collection and Representation Biases
26	Algorithm Design and Implementation Biases
27	Feedback Loops and Amplification Effects
28	System Interaction and Deployment Biases
29	Intersectionality and Compound Discrimination
30	From Correlation to Causation
31	Building Causal Models for Bias Detection
32	The Counterfactual Fairness Framework
33	Causal Inference with Limited Data
34	Distribution Shift and Emerging Discrimination
35	The Bias Source Identification Method

Part IV: Measuring Fairness: Metrics and Statistical Evidence

TABLE 14. Chapters of Part IV

CHAPTER	TITLE
36	Fairness Assessment Methodology
37	Comprehensive Data Auditing
38	Bias Detection in Raw Data
39	Group Fairness Metrics
40	Implementing Fairness Metrics
41	Statistical Validation and Confidence Intervals
42	Statistical Significance and Robustness
43	Small Samples and Sparse Subgroups
44	The Fairness Metrics Selection Method
45	Limits of Quantitative Fairness Assessment

Part V: Fairness Interventions

Published in 2 volumes: Volume 1, Chapters 46 to 52; Volume 2, Chapters 53 to 57.

TABLE 15. Chapters of Part V

CHAPTER	TITLE
46	Choosing an Intervention Strategy
47	Reweighting and Resampling
48	Distribution Transformation
49	Feature Engineering and Transformation for Fairness
50	Fairness-Aware Data Generation
51	Fairness Objectives and Constraints
52	Fairness-Aware Loss Functions and Regularization
53	Adversarial Debiasing
54	Multi-Objective Optimization
55	Threshold Optimization
56	Calibration Across Groups
57	Prediction Transformation Methods

Part VI: Fairness Engineering and Operations

TABLE 16. Chapters of Part VI

CHAPTER	TITLE
58	The Fairness Tooling Ecosystem
59	Fairness in the Development Workflow
60	Automated Bias Checks in CI/CD
61	Metric Tracking and Real-Time Bias Detection
62	Drift Detection and Alerting
63	Fairness Dashboards and Reporting
64	A/B Testing for Fairness

Part VII: Fairness in Advanced AI Architectures

TABLE 17. Chapters of Part VII

CHAPTER	TITLE
65	Deep Learning Systems
66	Recommendation and Ranking Systems
67	Large Language Models
68	Multi-Modal and Vision Models
69	Generative and Agentic Systems

Part VIII: Explainability, Interpretability and Non-Discrimination

Published in 2 volumes: Volume 1, Chapters 70 to 76; Volume 2, Chapters 77 to 82.

TABLE 18. Chapters of Part VIII

CHAPTER	TITLE
70	Foundations of AI Explainability
71	Explainability and Fairness
72	Taxonomy of Explainability Methods
73	Feature Attribution Methods
74	Counterfactual Explanations and Algorithmic Recourse
75	Causal Explainability
76	Group-Level and Intersectional Explanations

CHAPTER	TITLE
77	Explanation Quality, Fidelity and Robustness
78	Human-Centered Explainability
79	Explainability and Legal Rights
80	Explainability in Generative and Agentic AI
81	Explainability as Assurance Evidence
82	Limits, Risks and Fairness of Explanations

Part IX: Legal, Ethical and Regulatory Foundations

TABLE 19. Chapters of Part IX

CHAPTER	TITLE
83	Legal Standards for Algorithmic Fairness
84	International Human Rights and Non-Discrimination Law
85	The Global Regulatory Landscape
86	The EU AI Act and GDPR Article 22
87	Swiss Legal and Regulatory Framework
88	United States Anti-Discrimination Frameworks
89	International Standards and Governance Principles
90	Risk Classification and Assessment Systems
91	Sector-Specific Obligations
92	Ethics Beyond Compliance and Regulatory Change

Part X: AI Assurance, Evidence and Digital Trust

TABLE 20. Chapters of Part X

CHAPTER	TITLE
93	From Fairness Measurement to Assurance
94	Claims, Evidence and Assurance Arguments
95	Evidence Collection and Audit Trails
96	Independent Assessment and Reproducibility
97	Verifiable Credentials and Digital Attestations
98	Identity, Provenance and Evidence Integrity
99	Validity, Revocation and Continuous Assurance
100	Trust Registries and Distributed Trust Infrastructures
101	Limits of Certification and Assurance

Part XI: Fairness in Organizational Practice

TABLE 21. Chapters of Part XI

CHAPTER	TITLE
102	Integrating Fairness into Agile Principles
103	Fairness User Stories and Acceptance Criteria
104	Sprint Planning and Execution for Fairness
105	Fairness Ceremonies and Checkpoints
106	Roles and Responsibilities for Fairness
107	Documentation and Communication Frameworks
108	Governance Mechanisms and Decision Processes
109	Incident Management, Remediation and Redress
110	The Fairness Audit Playbook: an Integrated Workflow
111	Applied Case Studies

Part XII: Emerging Challenges and Research Frontiers

TABLE 22. Chapters of Part XII

CHAPTER	TITLE
112	Fairness in Large Language Models and Generative AI
113	Autonomous Agents and Delegated Decisions
114	Adaptive Systems and Collective Impacts
115	An Integrated Research Agenda

Appendices

TABLE 23. Appendix volumes

VOLUME	APPENDICES
Appendix Volume 1	A Controlled glossary; B Mathematical notation; L Index of definitions and cross-references
Appendix Volume 2	C Fairness metrics catalog; D Explainability methods catalog; E Statistical methods; H Worked numerical examples
Appendix Volume 3	F Legal and regulatory matrix; G Standards and normative references; I Templates, checklists and method worksheets; J Evidence and assurance-claim taxonomy; K Bibliography and source register

END OF FRONT MATTER

Trust is assessed, not asserted.

validant.ai is the AI fairness and digital trust assessment platform of Glinz & Company GmbH.

Glinz & Company GmbH
Zolliker Strasse 98 · CH-8702 Zollikon · Switzerland
hello@validant.ai · www.validant.ai

NO PROVENANCE SEAL

