A General Quantitative Methodology for the Ethical Auditing of AI Systems

Alessio Donvito[1]

[1]University of Bari "Aldo Moro", Piazza Umberto I, 1, Bari, 70121, Italy

Abstract

High-level principles such as fairness, transparency, accountability, privacy, human oversight, robustness, and safety are now central to AI ethics and governance. However, their practical use in auditing remains difficult: principles are abstract, context-sensitive, internally heterogeneous, and subject to disagreement among stakeholders and experts. This paper proposes a fuzzy framework for computing an Ethical Compliance Index (ECI) for AI systems. The framework separates two layers: the contextual importance of ethical principles and the empirical satisfaction of the corresponding technical requirements. Expert judgements about the importance of ethical sub-dimensions are expressed through linguistic labels, represented as triangular or trapezoidal fuzzy numbers, and aggregated into fuzzy contextual importance weights. Technical metrics associated with each principle are translated into satisfaction scores and then combined with the importance weights to compute the ECI. The result is an audit-oriented procedure that makes ethical compliance assessments more explicit, traceable, and robust under disagreement, while preserving the graded nature of ethical evaluation.

Keywords

AI auditing; ethical compliance; fuzzy logic; multi-criteria decision-making; AI ethics; algorithmic auditing; responsible AI

Introduction

Artificial Intelligence systems are increasingly deployed in contexts where automated decisions may affect human rights, safety, and access to opportunities. This has led to a growing body of work in AI ethics and governance, where a deontological approach – i.e. one based on principles such as fairness, transparency, accountability, privacy, human oversight, robustness, safety, etc. – has gained ground [1, 2]. However, the existence of high-level principles does not by itself provide a sufficiently precise method for assessing AI systems. Ethical principles are vague, abstract, and context-sensitive; moreover, different stakeholders may reasonably disagree about their interpretation, priority, and implications [3]. This creates a methodological challenge for AI auditing: on the one hand, ethical assessment cannot remain at the level of general principles, since auditing requires that ethical principles be translated into operational, measurable dimensions. However, as highlighted in recent work on AI ethics implementation, this translation is not a linear process in which abstract principles are mapped onto technical requirements, but it involves iterative processes of interpretation, negotiation, and adaptation to different organisational, technical or regulatory contexts [4].

On the other hand, ethical evaluation cannot be reduced to a technical checklist. The metrics used to assess fairness, transparency, robustness, or human oversight are not neutral substitutes for ethical principles: they operationalise only specific aspects of those principles and often express a wealth of normative assumptions (for instance, fairness may refer to allocative, representational, or procedural concerns, and different fairness metrics may be mutually incompatible in practice [5, 6, 7]). In many cases, principles are satisfied only by degree, their relevance depends on the specific context of application, and their technical counterparts become meaningful only when appropriately selected and interpreted. This makes binary compliant/non-compliant assessments too coarse for most real-world auditing scenarios [8].

BEWARE-26 (Bias, Risk, Explainability and the role of Logic and Logic Programming), co-located with AIxIA 2026, October 8, 2026, Perugia, Italy.

a.donvito22@phd.uniba.it (A. Donvito)

0009-0007-7546-0342 (A. Donvito)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-17

CEUR-WS.org/Vol-4332/paper10.pdf

This paper proposes a fuzzy framework for computing an Ethical Compliance Index (ECI) for AI systems. The framework combines: (1) expert judgements about the importance of ethical principles within a specific application context; and (2) empirical evidence obtained from ethically relevant technical metrics. Expert judgements are expressed through linguistic labels and represented as triangular or trapezoidal fuzzy numbers, following the use of fuzzy sets to model vagueness and graded membership [9].

The framework proceeds as follows: first, ethical principles are decomposed into sub-dimensions, so that broad normative ideals can be connected to more specific operational properties. Second, a panel of experts assigns linguistic importance labels to these sub-dimensions within a given context of application. These labels are mapped onto fuzzy numbers, aggregated across experts, and translated into fuzzy contextual importance weights. Third, technical metrics associated with each sub-dimension are applied and then normalised into satisfaction scores in the interval $[0, 1]$. Fourth, the resulting importance weights and satisfaction scores are combined through α-cut propagation into a fuzzy Ethical Compliance Index, which summarises the degree to which the system satisfies the ethical requirements considered relevant in the examined context.

The framework also includes a decomposition component and a sensitivity analysis. The first one identifies which principles contribute most to the final score and which account for the largest deficits, while the latter recomputes the index from the standpoint of individual experts, making explicit whether the compliance judgement is stable under disagreement about ethical priorities.

This work extends previous research by offering four key contributions. First, it defines a fuzzy expert-based procedure for deriving contextual ethical importance weights from linguistic judgements over ethical principles. Second, it integrates these weights with technical satisfaction scores into an audit-oriented Ethical Compliance Index. Third, it introduces safeguards for non-negotiable ethical requirements and decomposes the resulting score into principle-level contributions and deficits. Fourth, it proposes an individual-expert analysis to assess whether the compliance judgement remains stable under disagreement about ethical priorities. The framework is designed to support ex-post assessment of AI systems with respect to explicitly selected principles and metrics.

The remainder of the paper is organised as follows: Section 1 discusses related work on AI ethics principles, AI auditing, and machine ethics; Section 2 introduces the framework; Section 3 provides an illustrative application of the proposal.

1. Related Work

A large body of literature has documented the proliferation of AI ethics guidelines [2, 10], along with the problem of their application in real-world scenarios. Mittelstadt, for instance, stresses that AI ethics lacks robust institutional mechanisms and professional norms and procedures comparable to those available in medical ethics [3]. Works on the operationalisation of AI ethics have shown the obstacles to its translation into practices, tools, procedures, and measurable criteria that can be applied during the design, deployment, and assessment of AI systems [11, 12, 13].

Algorithmic auditing emerged as a response to the opacity and social impact of algorithmic systems, with early work focusing on how external or internal audits can detect discriminatory or harmful behaviour [14]. Mökander and colleagues develop the notion of ethics-based auditing as a process for assessing whether automated decision-making systems are consistent with relevant ethical principles and norms [15, 16]. Later work further analyses how such auditing procedures can be operationalised in organisational contexts and how independent audits may support AI safety and governance [17, 18]. The framework proposed here is compatible with this perspective, but focuses on a narrower problem: how to combine context-dependent ethical priorities and heterogeneous technical indicators into an interpretable compliance index.

The need for such a combination becomes evident when considering the literature on specific ethically relevant metrics. In algorithmic fairness, for example, many definitions and metrics have been proposed, but these are not always mutually compatible and may carry different normative assumptions

[6, 7, 19]. Similar issues arise in explainability and transparency, where different methods address different explanatory goals [20]. This literature shows that technical indicators are indispensable for ethical assessment, yet they cannot be treated as neutral measurements of ethical value. Metrics require selection and interpretation in relation to the ethical dimension they operationalise.

This requirement places the paper within the domain of machine ethics, specifically regarding how artificial agents interface with ethical reasoning and action [21, 22, 23]. Classical discussions distinguish between top-down approaches, which encode ethical theories or rules, bottom-up approaches, which learn ethical behaviour from examples or interaction, and hybrid approaches that combine both strategies [24, 23]. More recent work has explored logic- and learning-based approaches to creating ethical agents, including methods for learning ethical principles from user interactions [25]. Other contributions have used fuzzy cognitive maps or fuzzy ethical decision-making models to represent uncertainty, graded risk, and ethical pluralism in artificial moral agents [26, 27]. The proposed framework stems from these approaches, but it does not aim to prescribe which action an autonomous agent should take in a morally charged situation. Rather, it provides an ex-post or audit-oriented methodology for evaluating the ethical profile of an AI system with respect to a set of principles, sub-dimensions, and metrics deemed relevant in a given context.

2. Ethical Compliance Index

The quantitative core of the proposal draws on fuzzy logic, which provides a formal apparatus for modelling uncertainty and vagueness, characteristic features of ethical concepts and principles. At the same time, the assessment process relies on qualitative judgements expressed in natural language. A panel of domain experts assesses the relative importance of each ethical principle within a given application scenario. These judgements are then translated into fuzzy weighting coefficients, reflecting the contextual relevance of each principle rather than assuming uniform ethical priorities across domains. The weights are then integrated with empirical evidence obtained from technical assessments of the system, namely its performance on the metrics most closely associated with each ethical principle. Through the sequence of fuzzy transformations and aggregation procedures defined below, this process yields a fuzzy compliance profile, the Ethical Compliance Index, together with a defuzzified value for compact reporting.

2.1. Ethical Principles and Sub-dimensions

Ethical principles are typically introduced as high-level normative ideals and are therefore too abstract to be directly applied to the evaluation of actual AI systems. Each of these principles encompasses a broad array of concerns: in practice, no principle can be represented as a single property, and each one is an umbrella term that bundles multiple meanings.

For this reason, the framework decomposes each principle into sub-dimensions, more specific components that isolate one ethically relevant property. For example, the principle of *fairness* may include *representational fairness* (how groups are portrayed in the data) [28], *allocative fairness* (how resources or opportunities are distributed) [29], *procedural fairness* (how decisions are reached) [30], and *outcome fairness* (whether the model's outputs systematically disadvantage particular groups) [31]. In broader terms, sub-dimensions help reduce the polysemy and vagueness inherent in ethical principles, allowing them to be analysed in parts without collapsing their overall meaning.

Each sub-dimension d is associated with two mathematical objects: a fuzzy *importance weight* $\hat{I}_d$ and a *satisfaction score* T_d. These values correspond to two different kinds of information: (1) the relative importance of the ethical sub-dimension in context, obtained from expert judgements expressed in natural language; and (2) the empirical evidence regarding how well the system satisfies that aspect of the principle, obtained by computing the relevant technical metrics.

Let $P = \{P_1, \dots, P_K\}$ be a set of ethical principles, and for each principle P_i, let $D_i = \{d_{i_1}, \dots, d_{i_k}\}$ denote its set of sub-dimensions. Each sub-dimension $d \in D$ is associated with a corresponding set of metrics $M_{D_i} = \{m_{d_1}, \dots, m_{d_q}\}$. Also, let $E = \{e_1, \dots, e_R\}$ be the panel of experts. Given an AI system $\mathcal{S}$, the goal

of the following procedure is to compute a fuzzy Ethical Compliance Index $\widetilde{\mathrm{ECI}}(\mathcal{S})$, together with a defuzzified representative value $\mathrm{ECI}^{\star}(\mathcal{S}) \in [0,1]$, based on the combination of: (1) the ethical analysis and judgement expressed by E on the importance of each ethical principle P_i within the application context of the system $\mathcal{S}$, and (2) the performance of $\mathcal{S}$ on M.

2.2. Expert Judgements as Fuzzy Numbers

We begin by defining a variable to represent the importance of each principle P_i within its specific context of application. For each sub-dimension $d \in D$ and each expert $e_r \in E$, an individual linguistic judgement is selected from the set $L = \{\mathrm{VL}, \mathrm{L}, \mathrm{M}, \mathrm{I}, \mathrm{VI}, \mathrm{NN}\}$, where VL = Very Low, L = Low, M = Medium, I = Important, VI = Very Important, NN = Non-Negotiable; the label chosen by expert e_r for sub-dimension d is $L_{r,d} \in L$.

Let $L^{\star} = \{VL, L, M, I, VI\}$ denote the set of ordinary fuzzy importance labels. Principles, and their sub-dimensions, that experts qualify as *Non-Negotiable* (NN), must be treated as hard ethical constraints that the system must satisfy before the index is computed. Let $D_{NN} \subseteq D$ be the set of sub-dimensions designated as NN according to the rule adopted by the expert panel. When experts disagree, two rules may be adopted: 1) *Quorum*, where $d \in D_{NN}$ if and only if the share of NN votes reaches an agreed value $\theta_{NN} \in (0,1]$:

$$d \in D_{NN} \iff \frac{1}{R}\sum_{r=1}^{R} \eta_{r,d} \geq \theta_{NN}, \tag{1}$$

or 2) *Veto*, where a single NN vote determines that $d \in D_{NN}$, for example in safety-critical settings:

$$d \in D_{NN} \iff \exists r \in \{1, \ldots, R\} : \eta_{r,d} = 1. \tag{2}$$

For transparency and traceability, experts can record the rule they agreed upon, together with its justification. If rule (1) does not trigger the NN status for d, the sub-dimension can still receive a high importance label (e.g. VI). For each $d \in D_{NN}$, the protocol specifies a minimum acceptable technical satisfaction threshold τ_d and a satisfaction score $T_d(\mathcal{S}) \in [0,1]$. The threshold τ_d may be fixed directly by the audit protocol, so that a safeguard pre-check can be executed according to the rule:

$$\forall d \in D_{NN} : T_d(\mathcal{S}) \geq \tau_d. \tag{3}$$

If this condition is not satisfied, the assessment terminates with a *non-compliance/ethical risk* alert. If all NN constraints are satisfied, the evaluation proceeds with the ordinary fuzzy importance labels $\ell \in L^{\star}$.

Given their qualitative and vague nature, ordinary importance labels $\ell \in L^{\star}$ are represented as fuzzy numbers [32, 33]. For the intermediate labels, the framework uses Triangular Fuzzy Numbers (TFNs). A TFN is a triple $A = (a, b, c)$, with $0 \leq a \leq b \leq c \leq 1$, where a and c define the support of the fuzzy number and b is the modal value, i.e., the value with maximum membership. Its membership function is:

$$\mu_A(x) = \begin{cases} 0, & x \leq a, \\ \dfrac{x-a}{b-a}, & a < x \leq b, \\ \dfrac{c-x}{c-b}, & b < x < c, \\ 0, & x \geq c. \end{cases} \tag{4}$$

For the extreme labels, the framework uses Trapezoidal Fuzzy Numbers (TrFNs), which allow a label to have an interval of equally representative values. A TrFN is a quadruple $B = (a, b, c, d)$, with $0 \leq a \leq b \leq c \leq d \leq 1$, where $[a, d]$ is the support and $[b, c]$ is the core, i.e., the interval on which

membership is equal to 1. For the non-degenerate case, its membership function is:

$$\mu_B(x) = \begin{cases} 0, & x \le a, \\ \dfrac{x-a}{b-a}, & a < x \le b, \\ 1, & b < x \le c, \\ \dfrac{d-x}{d-c}, & c < x < d, \\ 0, & x \ge d. \end{cases} \tag{5}$$

For the sole purpose of computing the contextual importance weights, each individual NN vote is represented by the degenerate trapezoidal fuzzy number

$$w_{\mathrm{NN}} = (1, 1, 1, 1).$$

The qualitative scale $L^\star = \{\mathrm{VL}, \mathrm{L}, \mathrm{M}, \mathrm{I}, \mathrm{VI}\}$ is mapped onto fuzzy numbers in $[0, 1]$. Each label is assigned a prototypical central value:

$$c_{\mathrm{VL}} = 0.1, \quad c_{\mathrm{L}} = 0.3, \quad c_{\mathrm{M}} = 0.5, \quad c_{\mathrm{I}} = 0.7, \quad c_{\mathrm{VI}} = 0.9.$$

For the intermediate labels $\ell_2 = \mathrm{L}$, $\ell_3 = \mathrm{M}$, and $\ell_4 = \mathrm{I}$, we use symmetric triangular fuzzy numbers:

$$w_{\ell_j} = (c_{j-1}, c_j, c_{j+1}), \quad j = 2, 3, 4. \tag{6}$$

Thus, each intermediate label peaks at its central value, and its support is bounded by the central values of the adjacent labels. The extreme labels VL and VI are represented as left- and right-shoulder trapezoidal fuzzy numbers. This reflects the assumption that values sufficiently close to the lower and upper bounds of $[0, 1]$ are equally representative of "very low" and "very high" importance, respectively. With the central values above, the two labels are defined as:

$$w_{\mathrm{VL}} = (0.0, 0.0, 0.1, 0.3), \tag{7}$$

and

$$w_{\mathrm{VI}} = (0.7, 0.9, 1.0, 1.0). \tag{8}$$

Table 1 summarises the fuzzy numbers assigned to each label $\ell \in L^\star$.

Table 1
Fuzzy numbers associated with the importance labels

Label ℓ	Linguistic variable	Type	Parameters
VL	Very Low	Trapezoidal	(0.0, 0.0, 0.1, 0.3)
L	Low	Triangular	(0.1, 0.3, 0.5)
M	Medium	Triangular	(0.3, 0.5, 0.7)
I	Important	Triangular	(0.5, 0.7, 0.9)
VI	Very Important	Trapezoidal	(0.7, 0.9, 1.0, 1.0)
NN	Non-Negotiable	Trapezoidal	(1.0, 1.0, 1.0, 1.0)

Having assigned a fuzzy weight w_ℓ to each label $\ell \in L^\star$, and assuming that each expert has provided an evaluation for every sub-dimension d, the next step consists in aggregating the individual judgements into a collective fuzzy representation of the contextual importance of d.

Let $L_{r,d}$ denote the label selected by expert e_r for sub-dimension d. Through the fuzzy scale defined in Table 1, each $L_{r,d}$ is associated with its corresponding fuzzy number $w_{L_{r,d}}$, so that for every $d \in D$ the set of judgements expressed by the panel E becomes:

$$\mathscr{W}_d = \{w_{L_{1,d}}, w_{L_{2,d}}, \dots, w_{L_{R,d}}\}.$$

Since the framework requires a single fuzzy representation of the contextual importance of d, the individual fuzzy numbers $w_{L_{r,d}} \in \mathscr{W}_d$ must be combined into an aggregated fuzzy value $\hat{I}_d$. In order to aggregate triangular and trapezoidal fuzzy numbers, all are first represented in trapezoidal form. A triangular fuzzy number is transformed into its trapezoidal equivalent via the mapping

$$(a, b, c) \mapsto (a, b, b, c).$$

Let

$$\widetilde{w}_{L_{r,d}} = (a_{r,d}, b_{r,d}, c_{r,d}, d_{r,d})$$

denote the trapezoidal representation of the fuzzy number associated with the label selected by expert e_r for sub-dimension d. The aggregated fuzzy importance weight of d is then computed by averaging:

$$\hat{I}_d = \left(\frac{1}{R} \sum_{r=1}^{R} a_{r,d},\ \frac{1}{R} \sum_{r=1}^{R} b_{r,d},\ \frac{1}{R} \sum_{r=1}^{R} c_{r,d},\ \frac{1}{R} \sum_{r=1}^{R} d_{r,d} \right). \tag{9}$$

We can now combine the fuzzy importance weights of the sub-dimensions associated with each principle P_i into an individual fuzzy value $\hat{I}_{P_i}$, representing the overall importance of that principle.

Let $D_i = \{d_{i1}, \ldots, d_{ik_i}\}$ be the set of sub-dimensions associated with principle P_i. The fuzzy importance of principle P_i is the fuzzy mean:

$$\hat{I}_{P_i} = \frac{1}{k_i} \ \odot \bigoplus_{j=1}^{k_i} \hat{I}_{d_{ij}}. \tag{10}$$

2.3. Ethical Performance Metrics and Normalisation

In order to complement the values obtained from the fuzzification and aggregation of the judgements expressed by the panel of experts, the framework incorporates quantitative evidence derived from the technical evaluation of the system under analysis. Each sub-dimension $d \in D$ is thus associated with a set of metrics $M_{D_i} = \{m_{d_1}, \ldots, m_{d_q}\}$ that operationalise it.

The selected metrics are to be interpreted as *partial quantitative proxies* of the ethical property defined by d. No assumption is made that M_{D_i} is complete or exhaustive: additional metrics can be included depending on the desired granularity of the protocol.

For some measures, higher values indicate better performance with respect to the ethical property under consideration. For others, higher values indicate worse performance, as in metrics that quantify risk, bias, error, or disparity. Thus, when a metric yields a raw value, the ethical interpretation of that value depends on the metric itself and its formal structure. Because no single rule for the use and aggregation of the scores M_{D_i} is sufficient, let us start by defining the normalisation conditions of each category of metrics. Given a system S, each metric is a function $m : \mathcal{S} \to \mathbb{R}$ which yields a raw value

$$x_m = m(S) \in X_m, \tag{11}$$

where $X_m \subseteq \mathbb{R}$ is the codomain of the function. Since the metrics differ in scale and directionality, the values x_m might not be directly comparable. To integrate them into the assessment procedure, they must be translated, whenever necessary, into values in the interval $[0, 1]$ via a function:

$$\varphi : \mathbb{R} \to [0, 1],$$

appropriately defined for each class of metrics. The normalised score for each metric is:

$$n(x_m) = \varphi(x_m).$$

For metrics whose values x_m are not already bounded in $[0, 1]$, we specify two thresholds ($L, U \in X_m, \quad L < U$) representing, respectively, the lower (L) and upper (U) bounds of the range of ethically acceptable values. Specifically, L represents the limit below which $\mathcal{S}$ is considered non-compliant, while

U represents the saturation point for the metric, beyond which further technical improvements yield diminishing returns in terms of ethical benefit. For *lower-is-better* metrics, the interpretation is reversed: U denotes the threshold above which $\mathcal{S}$ is considered non-compliant, while L represents the saturation point below which further reductions in the metric yield diminishing ethical benefits. Rather than being fixed a priori, L and U must be calibrated on a per-metric basis, following explicit and documented criteria.

Thus, the following set of normalisation functions is defined. For metrics ($m^{\uparrow}$) where higher values represent better ethical alignment, the function is:

$$\varphi(x_{m^{\uparrow}}) = \begin{cases} 0, & x_m \leq L, \\ \dfrac{x_m - L}{U - L}, & L < x_m < U, \\ 1, & x_m \geq U. \end{cases} \tag{12}$$

The following, instead, is applied to metrics ($m^{\downarrow}$) expressing risk, bias, or error:

$$\varphi(x_{m^{\downarrow}}) = \begin{cases} 1, & x_m \leq L, \\ \dfrac{U - x_m}{U - L}, & L < x_m < U, \\ 0, & x_m \geq U. \end{cases} \tag{13}$$

In some cases, a metric can require an optimal equilibrium point, a target value T. Any deviation from T in either direction reduces ethical compliance. To model this behaviour, the framework adopts a normalisation function centred on an equilibrium point $T \in X_m$, representing the ethically optimal value of the metric. The normalised score is defined as:

$$\varphi(x_m) = \max\left(0,\ 1 - \frac{|x_m - T|}{\delta}\right), \tag{14}$$

where $\delta > 0$ denotes the maximum admissible deviation from the target value T. Equation 14 assigns a compliance score of 1 when the metric exactly matches the equilibrium point ($x_m = T$). As the value deviates from T, the score decreases linearly with the absolute distance $|x_m - T|$. When the deviation reaches δ, the score becomes 0.

For qualitative requirements (e.g., "Is a Human-in-the-loop mechanism present?"), the protocol may employ a binary scoring system:

$$\varphi(x_m) = \begin{cases} 1 & \text{if satisfied} \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

2.4. Aggregation rules

Once all the raw outputs x_m have been normalised, they must be combined into an overall score for each principle P_i. Here, the employment of a naive averaging procedure can result in unwanted *compensatory masking*, by which we denote the occurrence where a high overall score obscures poor performance on a metric of a given sub-dimension or principle. One practical criterion for selecting an aggregation operator can be found in the degree of correlation among the metrics in M_d. If the metrics are strongly correlated across a set of evaluation conditions, they behave as overlapping proxies of the same property, and averaging is less likely to conceal distinct failures. In these cases, a simple arithmetic mean may be appropriate to compute the technical satisfaction score for d:

$$T_d(\mathcal{S}) = \frac{1}{q_d} \sum_{j=1}^{q_d} n_{d,j}(\mathcal{S}). \tag{16}$$

Conversely, if the metrics are weakly correlated, they likely capture complementary aspects of d, so a mean would more easily hide local weaknesses. In the latter case, a less compensatory aggregation strategy is required.

To determine whether the metrics associated with a given sub-dimension provide redundant information, consider a finite collection of evaluation instances

$$\mathscr{E} = \{\mathcal{S}_1, \dots, \mathcal{S}_N\}, \tag{17}$$

where each $\mathcal{S}_n$ may correspond, for example, to different systems or versions of the same system, dataset slices, or perturbations of the input data (e.g., through bootstrapping [34]). To compute the correlation, construct a matrix whose columns represent the values $n(x_m)$ for each metric and whose rows represent the evaluation instances $\mathcal{S}_t$:

$$\mathbf{X} = \begin{pmatrix} n(x_{1,1}) & n(x_{2,1}) & \dots & n(x_{M,1}) \\ n(x_{1,2}) & n(x_{2,2}) & \dots & n(x_{M,2}) \\ \vdots & \vdots & \ddots & \vdots \\ n(x_{1,N}) & n(x_{2,N}) & \dots & n(x_{M,N}) \end{pmatrix} \tag{18}$$

Given the matrix $\mathbf{X}$, correlation is computed via Spearman's rank correlation [35, 36, 37, 38]. The resulting correlation coefficient $\bar{\rho}_d$ provides an indicator of the degree of agreement among the metrics associated with sub-dimension d. In particular, a high positive value of $\bar{\rho}_d$ (typically, $\bar{\rho}_d \geq 0.7$) [39] indicates that the metrics tend to correlate strongly, which supports the assumption that they capture closely related aspects of the same underlying ethical principle.

If the metrics associated with d are strongly correlated ($\bar{\rho}_d \geq 0.7$), a simple arithmetic mean of the normalised values $n(x_m)$ is sufficient. Conversely, when the correlation is weak or inverse ($\bar{\rho}_d < 0.7$), they plausibly capture distinct aspects of the same sub-dimension. In such cases, it would be safer to use a conservative aggregation operator in order to reduce masking. This can be achieved by using a *soft minimum* based on log-sum-exp (LSE) [40]:

$$T_d(\mathcal{S}) = -\frac{1}{\beta_d} \log\left(\frac{1}{q_d} \sum_{j=1}^{q_d} e^{-\beta_d n(x_j)}\right). \tag{19}$$

Here, $T_d(\mathcal{S}) \in [0, 1]$ denotes the aggregated score of system $\mathcal{S}$ on ethical dimension d, and $q_d \in \mathbb{N}$ is the number of metrics selected to operationalise that dimension. $n(x_j) \in [0, 1]$ is the normalised value of the j-th metric. The parameter $\beta_d > 0$ determines how strongly the operator concentrates on the worst-performing values. The expression inside the logarithm changes the scale of the metric values in order to make low scores more influential. Each normalised score $n(x_j) \in [0, 1]$ is transformed into $e^{-\beta_d n(x_j)}$. When $n(x_j)$ is close to 0, the resulting value is close to 1, whereas when $n(x_j)$ is close to 1, the value is much smaller. This gives lower scores greater influence in the aggregation. The operator adds up these transformed values across all metrics through the sum $\sum_{j=1}^{q_d} e^{-\beta_d n(x_j)}$, then the total is divided by q_d, i.e., $\frac{1}{q_d}\sum_{j=1}^{q_d}(\cdot)$. Finally, the logarithm and the outer factor $-1/\beta_d$ rescale the averaged quantity so that the final result lies again in the interval $[0, 1]$. By construction, if all metrics take the value $n(x_j) = 1$, then $T_d(\mathcal{S}) = 1$; if they all take the value $n(x_j) = 0$, then $T_d(\mathcal{S}) = 0$.

Equation 19 follows from the standard log-sum-exp (LSE) function: $\mathrm{LSE}_\beta(z_1, \dots, z_q) = \frac{1}{\beta}\log\left(\frac{1}{q}\sum_{j=1}^{q} e^{\beta z_j}\right)$ is an approximation of $\max_j z_j$, in the sense that it converges to $\max_j z_j$ as $\beta \to +\infty$. A minimum can be rewritten as a negated maximum over negated arguments, $\min_j a_j = -\max_j(-a_j)$.

Combining Eqs. 16 and 19, the framework can employ an adaptive aggregation rule:

$$T_d(\mathcal{S}) = \begin{cases} \dfrac{1}{q_d}\displaystyle\sum_{j=1}^{q_d} n_{d,j}(\mathcal{S}), & \text{if } \bar{\rho}_d \geq \theta, \\ -\dfrac{1}{\beta_d}\log\left(\dfrac{1}{q_d}\displaystyle\sum_{j=1}^{q_d} e^{-\beta_d n(x_j)}\right), & \text{if } \bar{\rho}_d < \theta, \end{cases} \tag{20}$$

where θ defines the correlation threshold set at 0.7 in the illustrative application below.

Once the technical satisfaction scores $T_{P_i}(S) \in [0,1]$ have been computed for each ethical principle $P_i \in \mathcal{P}$, and the corresponding contextual importance weights $\hat{I}_{P_i} \in [0,1]$ have been obtained, the final step consists in combining these quantities.

2.5. Ethical Compliance Index

The *Ethical Compliance Index* (ECI) of a system $\mathcal{S}$ is defined as:

$$\widetilde{\mathrm{ECI}}(\mathcal{S}) = \frac{\sum_{i=1}^{K} \hat{I}_{P_i} T_{P_i}(\mathcal{S})}{\sum_{i=1}^{K} \hat{I}_{P_i}}. \tag{21}$$

The numerator aggregates the technical satisfaction scores obtained by $\mathcal{S}$, weighting each contribution by the contextual importance assigned by the expert panel. As a result, insufficient performance on principles that are considered highly important has a stronger negative impact on the overall index. The denominator rescales the aggregation so that the resulting fuzzy number is defined over the interval $[0,1]$.

Let

$$[\hat{I}_{P_i}]_\alpha = [\underline{I}_{i,\alpha}, \overline{I}_{i,\alpha}]$$

be the α-cut of the fuzzy importance weight assigned to P_i. For each $\alpha \in [0,1]$, the corresponding α-cut of $\widetilde{\mathrm{ECI}}(\mathcal{S})$ is given by

$$[\widetilde{\mathrm{ECI}}(\mathcal{S})]_\alpha = [\underline{E}_\alpha(\mathcal{S}), \overline{E}_\alpha(\mathcal{S})], \tag{22}$$

where

$$\underline{E}_\alpha(\mathcal{S}) = \min_{w_i \in [\underline{I}_{i,\alpha}, \overline{I}_{i,\alpha}]} \frac{\sum_{i=1}^{K} w_i T_{P_i}(\mathcal{S})}{\sum_{i=1}^{K} w_i}, \tag{23}$$

and

$$\overline{E}_\alpha(\mathcal{S}) = \max_{w_i \in [\underline{I}_{i,\alpha}, \overline{I}_{i,\alpha}]} \frac{\sum_{i=1}^{K} w_i T_{P_i}(\mathcal{S})}{\sum_{i=1}^{K} w_i}. \tag{24}$$

The resulting output is a fuzzy ethical compliance profile.

2.6. Score Decomposition

The value returned by $\widetilde{\mathrm{ECI}}(\mathcal{S})$ provides a summary of the ethical profile of a system. To enhance its analytical power, we might want to make explicit how much each principle contributes to the compliance score, both positively and negatively. For each $\alpha \in [0,1]$, let

$$\mathcal{W}_\alpha = \prod_{i=1}^{K} [\underline{I}_{i,\alpha}, \overline{I}_{i,\alpha}]$$

be the set of admissible crisp weight vectors induced by the α-cuts of the fuzzy importance weights. The absolute contribution of principle P_i to the ECI can therefore be defined through its α-cuts as

$$[\tilde{C}_i(\mathcal{S})]_\alpha = \left[\min_{w \in \mathcal{W}_\alpha} \frac{w_i T_{P_i}(\mathcal{S})}{\sum_{j=1}^{K} w_j}, \ \max_{w \in \mathcal{W}_\alpha} \frac{w_i T_{P_i}(\mathcal{S})}{\sum_{j=1}^{K} w_j} \right]. \tag{25}$$

For reporting purposes, it may also be useful to express each contribution as a share of the ECI score in percentage terms. When the total positive score is greater than zero, the percentage contribution of principle P_i can be defined as the proportion of the score that is due to principle P_i:

$$[\tilde{C}_i^{\%}(\mathcal{S})]_\alpha = \left[100 \cdot \min_{w \in \mathcal{W}_\alpha} \frac{w_i T_{P_i}(\mathcal{S})}{\sum_{j=1}^{K} w_j T_{P_j}(\mathcal{S})}, \ 100 \cdot \max_{w \in \mathcal{W}_\alpha} \frac{w_i T_{P_i}(\mathcal{S})}{\sum_{j=1}^{K} w_j T_{P_j}(\mathcal{S})} \right]. \tag{26}$$

For every admissible weight vector $w \in \mathscr{W}_\alpha$, the corresponding percentage contributions sum to 100.

Since a principle may contribute little to the ECI either because its weight $\hat{I}_{P_i}$, thus its importance in the CoA, is low or because its technical satisfaction score $T_{P_i}(\mathcal{S})$ is low, the evaluation procedure should also compute a *deficit contribution* for each principle. For each P_i, the deficit contribution is defined through its α-cuts as

$$[\widetilde{D}_i(\mathcal{S})]_\alpha = \left[\min_{w \in \mathscr{W}_\alpha} \frac{w_i(1 - T_{P_i}(\mathcal{S}))}{\sum_{j=1}^{K} w_j},\ \max_{w \in \mathscr{W}_\alpha} \frac{w_i(1 - T_{P_i}(\mathcal{S}))}{\sum_{j=1}^{K} w_j} \right]. \tag{27}$$

$\widetilde{D}_i(\mathcal{S})$ measures how much P_i contributes to the *deficit* of system $\mathcal{S}$ on the ECI. Pointwise, for each admissible weight vector w, the deficit of the corresponding crisp index is given by

$$1 - \mathrm{ECI}^{(w)}(\mathcal{S}) = \sum_{i=1}^{K} D_i^{(w)}(\mathcal{S}).$$

A principle has a high deficit contribution when it is highly important and is either poorly satisfied or not satisfied by the system. Analogously to Eq. 25, it can be useful to express the deficit measure on a percentage scale as well. When the total deficit is greater than zero:

$$[\widetilde{D}_i^{\%}(\mathcal{S})]_\alpha = \left[100 \cdot \min_{w \in \mathscr{W}_\alpha} \frac{w_i(1 - T_{P_i}(\mathcal{S}))}{\sum_{j=1}^{K} w_j(1 - T_{P_j}(\mathcal{S}))},\ 100 \cdot \max_{w \in \mathscr{W}_\alpha} \frac{w_i(1 - T_{P_i}(\mathcal{S}))}{\sum_{j=1}^{K} w_j(1 - T_{P_j}(\mathcal{S}))} \right]. \tag{28}$$

For every admissible weight vector $w \in \mathscr{W}_\alpha$, the corresponding pointwise percentage deficit contributions sum to 100.

The resulting vectors $\widetilde{\mathbf{C}}(\mathcal{S})$ and $\widetilde{\mathbf{D}}(\mathcal{S})$ allow the audit procedure to identify which principles explain the largest share of both ethical compliance and non-compliance as computed with the ECI. For compact presentation, each fuzzy contribution and deficit can additionally be reported through its defuzzified value and its support interval.

2.7. Sensitivity of the Ethical Compliance Index to Expert Disagreement

The fuzzy Ethical Compliance Index is computed from a single set of fuzzy weights $\{\tilde{I}_{P_i}\}_{i=1}^{K}$, obtained by aggregating the judgements expressed by the panel of experts. This aggregation compresses the individual positions of the experts into a collective representation, a loss of information that may matter in an ethical auditing process. In many cases, disagreement among experts may reflect some form of legitimate *pluralism*: within a rational community, different moral orientations may coexist without being reducible to a single doctrine [41, 42, 43, 44]. The defuzzified value $\mathrm{ECI}^{\star}(\mathcal{S})$ does not, by itself, show whether the same compliance judgement would remain stable if the system were evaluated from the standpoint of each individual expert. Thus, after computing the standard fuzzy ECI, the index is recomputed separately for each expert, using only that expert's own weights. The resulting set of individual fuzzy ECI values measures how sensitive the compliance judgement is to variation in the ethical priorities represented in the panel.

For each principle P_i, the individual fuzzy importance assigned by expert e_r is obtained by aggregating the fuzzy numbers associated with the sub-dimensions of P_i:

$$\hat{I}_{P_i}^{(r)} = \frac{1}{k_i} \odot \bigoplus_{d \in D_i} w_{L_{r,d}}, \tag{29}$$

where D_i is the set of sub-dimensions of P_i and $k_i = |D_i|$.

Let

$$[\hat{I}_{P_i}^{(r)}]_\alpha = [\underline{I}_{i,\alpha}^{(r)}, \overline{I}_{i,\alpha}^{(r)}]$$

be the α-cut of the individual fuzzy importance assigned by expert e_r to principle P_i. The individual Ethical Compliance Index $\widetilde{\mathrm{ECI}}^{(r)}(\mathcal{S})$ associated with expert e_r is then defined by applying equation 21 with the individual fuzzy weights in place of the aggregated ones:

$$[\widetilde{\mathrm{ECI}}^{(r)}(\mathcal{S})]_\alpha = \left[\min_{w_i \in [\underline{I}_{i,\alpha}^{(r)}, \overline{I}_{i,\alpha}^{(r)}]} \frac{\sum_{i=1}^{K} w_i\, T_{P_i}(\mathcal{S})}{\sum_{i=1}^{K} w_i}, \max_{w_i \in [\underline{I}_{i,\alpha}^{(r)}, \overline{I}_{i,\alpha}^{(r)}]} \frac{\sum_{i=1}^{K} w_i\, T_{P_i}(\mathcal{S})}{\sum_{i=1}^{K} w_i} \right], \qquad r = 1, \dots, R. \tag{30}$$

The technical satisfaction scores $T_{P_i}(\mathcal{S})$ remain the same for all experts. For reporting purposes, each individual fuzzy ECI can be represented by a defuzzified value:

$$\mathrm{ECI}^{(r)\star}(\mathcal{S}) = \mathrm{Defuzz}\left(\widetilde{\mathrm{ECI}}^{(r)}(\mathcal{S})\right). \tag{31}$$

The R individual scores can then be collected in the individual compliance vector:

$$\vec{\mathrm{ECI}}^{\star}(\mathcal{S}) = \left(\mathrm{ECI}^{(1)\star}(\mathcal{S}), \dots, \mathrm{ECI}^{(R)\star}(\mathcal{S})\right) \in [0,1]^R, \tag{32}$$

which is the basis for the indicators defined below.

Mean individual ECI. The first indicator is the arithmetic mean of the defuzzified individual ECI values:

$$\overline{\mathrm{ECI}}^{\star}(\mathcal{S}) = \frac{1}{R} \sum_{r=1}^{R} \mathrm{ECI}^{(r)\star}(\mathcal{S}). \tag{33}$$

This summarises the average compliance score obtained for each individual expert, useful for comparison with the standard defuzzified ECI.

Standard Deviation. The dispersion of the individual $\mathrm{ECI}^{(r)\star}$ values around their mean can be measured via standard deviation [45]:

$$\sigma_{\mathrm{ECI}}(\mathcal{S}) = \sqrt{\frac{1}{R} \sum_{r=1}^{R} \left(\mathrm{ECI}^{(r)\star}(\mathcal{S}) - \overline{\mathrm{ECI}}^{\star}(\mathcal{S})\right)^2}. \tag{34}$$

Since each $\mathrm{ECI}^{(r)\star}(\mathcal{S})$ lies in $[0,1]$, the standard deviation is bounded by $1/2$ by Popoviciu's inequality[1] [46, 47]. High values of $\sigma_{\mathrm{ECI}}(\mathcal{S})$ (tentatively, $\sigma_{ECI} \geq 0.10$) can function as indicators of a compliance judgement that changes much depending on which expert's priorities are assumed. For descriptive purposes, values below 0.05 may be read as low dispersion, values between 0.05 and 0.10 as moderate dispersion, and values above 0.10 as high dispersion.

Individual ECI Spread. The maximum divergence between defuzzified individual ECI values is measured by:

$$\Sigma(\mathcal{S}) = \max_r \mathrm{ECI}^{(r)\star}(\mathcal{S}) - \min_r \mathrm{ECI}^{(r)\star}(\mathcal{S}). \tag{35}$$

The spread $\Sigma(\mathcal{S}) \in [0,1]$ measures the largest difference between any two individual compliance scores in the panel. A high spread indicates that the compliance score is affected by which expert's priorities are adopted. In this case, the aggregated ECI should be interpreted with caution, and the individual ECI vector should be inspected when

$$\Sigma(\mathcal{S}) > \Sigma^\star, \tag{36}$$

where $\Sigma^\star \in (0,1)$ can be fixed by the audit protocol. A conventional value such as $\Sigma^\star = 0.20$ can be used as a starting point, although the threshold should be adjusted to the risk level of the application domain, the size and composition of the panel of experts, and the tolerance for disagreement assumed by the audit procedure.

[1] For any random variable X taking values in $[a,b]$, $\mathrm{Var}(X) \leq (b-a)^2/4$ by the Popoviciu inequality [46], so $\sigma \leq (b-a)/2$. With $a=0$ and $b=1$, $\sigma_{\mathrm{ECI}} \leq 1/2$.

3. Illustrative Application

To illustrate the evaluation process, this section introduces a synthetic scenario. Consider an AI system $\mathcal{S}$ for automated patient triage based on clinical urgency. The primary objective of $\mathcal{S}$ is to process incoming patient data (e.g. heart rate, blood pressure, oxygen saturation, age, sex, etc.), predict the urgency of a treatment and establish a priority list. The panel $E = \{e_1, \dots, e_5\}$ includes a clinician, a machine-learning engineer, a legal/regulatory expert, an AI ethics expert, and a patient representative. Each e_r assigns one label per sub-dimension (Table 3); protocol parameters are fixed *ex ante*. The panel selects four principles, each decomposed into two sub-dimensions (Table 2).

Table 2
Principles and sub-dimensions.

Principle	Sub-dim.	Label	Operationalises
P_1 Fairness	$d_{1.1}$	Demographic parity	Disparity in priority-rate allocation
	$d_{1.2}$	Equal opportunity	Disparity in error rates across groups
P_2 Transparency	$d_{2.1}$	Explanation quality	Availability and fidelity of explanations
	$d_{2.2}$	Documentation	Completeness of model/data documentation
P_3 Oversight	$d_{3.1}$	Human review	Availability and logging of human review
	$d_{3.2}$	Contestability	Availability of appeal mechanisms
P_4 Safety	$d_{4.1}$	Clinical reliability	Predictive performance and calibration
	$d_{4.2}$	Robustness	Stability and absence of unsafe recommendations

Then, each expert assesses the importance of principles $P_1, \dots, P_4$ and their relative sub-dimensions by choosing one label in L for each d.

Table 3
Expert labels and label counts.

d	e_1	e_2	e_3	e_4	e_5	VL	L	M	I	VI	NN
$d_{1.1}$	VI	I	VI	I	VI	0	0	0	2	3	0
$d_{1.2}$	I	I	VI	I	I	0	0	0	4	1	0
$d_{2.1}$	I	M	I	M	I	0	0	2	3	0	0
$d_{2.2}$	M	I	M	I	M	0	0	3	2	0	0
$d_{3.1}$	I	I	VI	I	I	0	0	0	4	1	0
$d_{3.2}$	I	M	I	I	M	0	0	2	3	0	0
$d_{4.1}$	VI	I	VI	I	VI	0	0	0	2	3	0
$d_{4.2}$	NN	NN	VI	NN	VI	0	0	0	0	2	3

Sub-dimension $d_{4.2}$ receives three NN votes out of five: $3/5 = 0.60 \geq \theta_{NN} \Rightarrow D_{NN} = \{d_{4.2}\}$, thus it meets the Quorum threshold to be qualified as Non-Negotiable.
Based on the choices made by the experts, each ethical dimension is associated with a fuzzy weight $\hat{I}_d = (a, b, c, d)$ and its defuzzified counterpart ($I_d^\star$, Table 4).
Table 5 shows the list of metrics selected to operationalise each sub-dimension, along with their raw and normalised values. When a metric lies in [0,1], it is either left unaltered (identity rule) if it has positive polarity (if "higher-is-better") or inverted (if "lower-is-better").

Table 4
Sub-dimension fuzzy weights and defuzzified values.

d	$\hat{I}_d = (a, b, c, d)$	$I_d^\star$
$d_{1.1}$	$(0.620, 0.820, 0.880, 0.960)$	0.813
$d_{1.2}$	$(0.540, 0.740, 0.760, 0.920)$	0.737
$d_{2.1}$	$(0.420, 0.620, 0.620, 0.820)$	0.620
$d_{2.2}$	$(0.380, 0.580, 0.580, 0.780)$	0.580
$d_{3.1}$	$(0.540, 0.740, 0.760, 0.920)$	0.737
$d_{3.2}$	$(0.420, 0.620, 0.620, 0.820)$	0.620
$d_{4.1}$	$(0.620, 0.820, 0.880, 0.960)$	0.813
$d_{4.2}$	$(0.880, 0.960, 1.000, 1.000)$	0.957

Table 5
Raw values and normalised scores.

d	Metric	x_m	$n(x_m)$	Normalisation rule
$d_{1.1}$	Demographic Parity Difference	−0.08	0.680	Eq. 14, $T = 0, \delta = 0.25$
	Disparate Impact Ratio	0.88	0.427	Eq. 14 on $\ln x$, $\delta = \ln 1.25$
$d_{1.2}$	Equal Opportunity Difference	−0.14	0.440	Eq. 14, $T = 0, \delta = 0.25$
	Average Odds Difference	−0.11	0.560	Eq. 14, $T = 0, \delta = 0.25$
$d_{2.1}$	Explanation Coverage	0.90	0.900	identity
	Local Explanation Fidelity	0.74	0.740	identity
$d_{2.2}$	Model-card Completion	0.80	0.800	identity
	Datasheet Completion	0.70	0.700	identity
$d_{3.1}$	Human-review Mechanism	binary	1.000	binary
	Logged Review Coverage	0.82	0.820	identity
$d_{3.2}$	Appeal Mechanism	binary	1.000	binary
$d_{4.1}$	AUROC	0.86	0.800	$(L, U) = (0.70, 0.90)$
	Expected Calibration Error	0.06	0.940	inverted
	Brier Score	0.10	0.900	inverted
$d_{4.2}$	High-acuity False-negative Rate	0.03	0.970	inverted
	Performance Drop under Shift	0.06	0.940	inverted
	Unsafe-recommendation Rate	0.015	0.985	inverted

When applicable, we compute the Spearman correlation between the metrics of each sub-dimension to determine the aggregation rule. We then calculate the technical satisfaction scores for each sub-dimension and principle (Table 7).

After checking that the safety gate is satisfied, $T_{d_{4.2}}(\mathcal{S}) = 0.964 \geq \tau_{d_{4.2}} = 0.85$, we compute $\widetilde{\mathrm{ECI}}(\mathcal{S})$. Aggregating the sub-dimension fuzzy weights at principle level gives:

$$\tilde{I}_{P_1} = (0.580, 0.780, 0.820, 0.940), \quad \tilde{I}_{P_2} = (0.400, 0.600, 0.600, 0.800),$$

$$\tilde{I}_{P_3} = (0.480, 0.680, 0.690, 0.870), \quad \tilde{I}_{P_4} = (0.750, 0.890, 0.940, 0.980).$$

The fuzzy ECI is then computed through α-cuts (Table 6).

Table 6
Fuzzy ECI by α-cuts.

α	$\underline{E}_\alpha(\mathcal{S})$	$\overline{E}_\alpha(\mathcal{S})$
0.00	0.750	0.820
0.25	0.758	0.812
0.50	0.767	0.803
0.75	0.774	0.795
1.00	0.781	0.788

The defuzzified value is:

$$\mathrm{ECI}^{\star}(\mathcal{S}) = \mathrm{Defuzz}\left(\widetilde{\mathrm{ECI}}(\mathcal{S})\right) = 0.785. \tag{37}$$

Finally, the result can be decomposed to evaluate the contribution and deficit coefficients of each principle (Table 8), and recomputed for each individual expert to assess whether the index is stable across the panel (Table 9).

In our example, Fairness remains the dominant driver of deficit ($D_1^{\star} = 0.130$, $\approx 60\%$ of $1 - \mathrm{ECI}^{\star}$), mainly due to $d_{1.1}$, while:

$$\overline{\mathrm{ECI}}^{\star}(\mathcal{S}) = 0.785, \qquad \sigma_{\mathrm{ECI}}(\mathcal{S}) = 0.005, \qquad \Sigma(\mathcal{S}) = 0.014,$$

with both indicators far below the alert thresholds introduced in Section 2.7, meaning that the compliance judgement is stable across the panel.

Table 7
Sub-dimension and principle-level aggregation.

Level	$\bar{\rho}$	Rule	Score
$d_{1.1}$	0.42	soft-min	$T_{d_{1.1}} = 0.516$
$d_{1.2}$	0.81	mean	$T_{d_{1.2}} = 0.500$
$d_{2.1}$	0.55	soft-min	$T_{d_{2.1}} = 0.804$
$d_{2.2}$	0.88	mean	$T_{d_{2.2}} = 0.750$
$d_{3.1}$	n/a	soft-min	$T_{d_{3.1}} = 0.890$
$d_{3.2}$	n/a	$\lvert M_d \rvert$=1	$T_{d_{3.2}} = 1.000$
$d_{4.1}$	0.77	mean	$T_{d_{4.1}} = 0.880$
$d_{4.2}$	0.49	soft-min	$T_{d_{4.2}} = 0.964$
P_1	–	soft-min (fixed)	$T_{P_1} = 0.508$
P_2	–	soft-min (fixed)	$T_{P_2} = 0.775$
P_3	–	soft-min (fixed)	$T_{P_3} = 0.938$
P_4	–	soft-min (fixed)	$T_{P_4} = 0.918$

Table 8
Principle contribution and deficit.

Principle	$I_{P_i}^{\star}$	T_{P_i}	$C_i^{\star}$	$C_i^{\star\%}$	$D_i^{\star}$	$D_i^{\star\%}$
P_1 Fairness	0.775	0.508	0.134	17.1	0.130	60.4
P_2 Transparency	0.600	0.775	0.158	20.2	0.046	21.4
P_3 Oversight	0.678	0.938	0.217	27.6	0.014	6.7
P_4 Safety	0.885	0.918	0.276	35.2	0.025	11.5
Total	2.938	–	0.785	100.0	0.215	100.0

Table 9
Individual defuzzified weights and ECI values.

Expert	$I_{P_1}^{(r)\star}$	$I_{P_2}^{(r)\star}$	$I_{P_3}^{(r)\star}$	$I_{P_4}^{(r)\star}$	$\mathrm{ECI}^{(r)\star}$
e_1	0.794	0.600	0.700	0.946	0.787
e_2	0.700	0.600	0.600	0.850	0.787
e_3	0.892	0.600	0.794	0.892	0.780
e_4	0.700	0.600	0.700	0.850	0.792
e_5	0.794	0.600	0.600	0.892	0.778

4. Limitations

The proposed framework has several limitations. The first concerns the role of human normative judgement, which the Ethical Compliance Index does not remove. The selection of principles, their decomposition into sub-dimensions, the choice of the most appropriate metrics, the definition of each threshold, all remain part of the audit protocol. These decisions must be made explicit with respect to the specific context of application. The index is, therefore, not an objective measure of ethical value, but, more modestly, a tool to make a given evaluation protocol for AI systems explicit and reproducible.

A second limitation, related to the previous, pertains to the definition of the importance weights, which still depends on expert elicitation. Fuzzy linguistic labels help represent disagreement and vagueness, but they are not designed to eliminate subjectivity. Different compositions of the same panel or elicitation procedures may lead to different weights.

Similarly, the technical satisfaction scores depend on the choice and calibration of the metrics. In particular, normalisation thresholds and target values must be defined on a per-metric basis, and poorly chosen thresholds and parameters may distort the resulting score. The selection of the metrics themselves is not a neutral procedure. The same holds for the choice between the two proposed aggregation rules. The correlation-based criterion proposed here should be read as a practical decision rule, and not as a universally valid solution. Alternative criteria may be equally viable, and evaluating how they could be integrated into the framework is a compelling direction for future research.

A further limitation is that the illustrative application is entirely synthetic. It shows that the procedure is coherent, but it does not amount to an empirical validation of the framework in a real audit setting. A stronger validation would require application to real systems, independently elicited expert judgements, documented audit protocols, and comparison with external assessment outcomes.

Finally, future work should verify the robustness of the procedure under variation in the protocol-dependent choices made by auditors, expert panels, and data scientists involved in the evaluation process. This can be done by extending the present sensitivity analysis to threshold calibration, aggregation parameters, metric normalisation settings, and expert-weight elicitation. Such an extension would help distinguish cases in which ethical compliance judgements are stable from cases in which the ECI is fragile under reasonable changes in the audit protocol.

5. Conclusion

This paper has introduced a fuzzy multi-criteria framework for computing an Ethical Compliance Index for AI system auditing. The ECI evaluates the degree to which an AI system satisfies a set of explicitly selected ethical principles in a given context of application. Far from attempting to replace the complexity of human ethical deliberation in real-world contexts, the ECI starts from the assumption that such replacement may be neither possible nor desirable. Ethical assessment remains a human-centred process, since the evaluation of an AI system always involves assumptions, priorities, values, and often divergent ethical standpoints; these elements do not disappear when the assessment is translated into the technical realm. They either remain implicit, or they are excluded from the analysis altogether. In both cases, these approaches risk operating at a level where some of the most relevant factors in ethical judgement are not adequately represented.

The contribution of the proposed framework is to address this gap by providing formal means for making some of these factors explicit and reproducible. This point is also relevant for the broader debate on technical approaches to AI ethics. As long as such approaches remain confined to a reductionist perspective, they are likely to appear insufficient to those who regard ethical assessment as a complex, socially embedded practice. Developing formal methods that can incorporate these concerns, without pretending to eliminate their complexity, may therefore be an important step toward more credible and philosophically informed forms of AI ethics assessment.

Declaration on Generative AI

During the preparation of this work, the author used GPT-5.6 Sol in order to: Grammar and spelling check, Paraphrase and reword. After using this tool, the author reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Floridi, J. Cowls, A unified framework of five principles for ai in society, Harvard Data Science Review 1 (2019).

[2] A. Jobin, M. Ienca, E. Vayena, The global landscape of ai ethics guidelines, Nature Machine intelligence 1 (2019) 389–399.

[3] B. Mittelstadt, Principles alone cannot guarantee ethical ai, Nature Machine Intelligence 1 (2019) 501–507. doi:10.1038/s42256-019-0114-4.

[4] D. Moats, S. Trifuljesko, Valuable actions and actionable values: Tinkering with principles and practices in AI ethics, The Sociological Review 73 (2025) 1368–1387. URL: https://doi.org/10.1177/00380261251316661. doi:10.1177/00380261251316661.

[5] H. Suresh, J. Guttag, A framework for understanding sources of harm in machine learning, Communications of the ACM 64 (2021) 54–62. doi:10.1145/3470515.

[6] S. Verma, J. Rubin, Fairness definitions explained, in: Proceedings of the International Workshop on Software Fairness, ACM, 2018, pp. 1–7. doi:10.1145/3194770.3194776.

[7] J. Kleinberg, S. Mullainathan, M. Raghavan, Inherent trade-offs in the fair determination of risk scores, in: 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2017, pp. 43:1–43:23. doi:10.4230/LIPIcs.ITCS.2017.43.

[8] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, P. Barnes, Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 33–44. doi:10.1145/3351095.3372873.

[9] L. A. Zadeh, Fuzzy sets, Information and control 8 (1965) 338–353.

[10] T. Hagendorff, The ethics of ai ethics: An evaluation of guidelines, Minds and Machines 30 (2020) 99–120. doi:10.1007/s11023-020-09517-8.

[11] J. Morley, L. Floridi, L. Kinsey, A. Elhalal, From what to how: An initial review of publicly available ai ethics tools, methods and research to translate principles into practices, Science and Engineering Ethics 26 (2020) 2141–2168. doi:10.1007/s11948-019-00165-5.

[12] J. Morley, A. Elhalal, F. Garcia, L. Kinsey, J. Mökander, L. Floridi, Ethics as a service: A pragmatic operationalisation of ai ethics, Minds and Machines 31 (2021) 239–256. doi:10.1007/s11023-021-09563-w.

[13] J. Morley, L. Kinsey, A. Elhalal, F. Garcia, M. Ziosi, L. Floridi, Operationalising ai ethics: Barriers, enablers and next steps, AI & Society 38 (2023) 411–423. doi:10.1007/s00146-021-01308-8.

[14] D. Metaxa, J. S. Park, R. E. Robertson, K. Karahalios, C. Wilson, J. Hancock, C. Sandvig, Auditing algorithms: Understanding algorithmic systems from the outside in, Foundations and Trends in Human–Computer Interaction 14 (2021) 272–344. doi:10.1561/1100000083.

[15] J. Mökander, L. Floridi, Ethics-based auditing to develop trustworthy ai, Minds and Machines 31 (2021) 323–327. doi:10.1007/s11023-021-09557-8.

[16] J. Mökander, J. Morley, M. Taddeo, L. Floridi, Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations, Science and Engineering Ethics 27 (2021) 44. doi:10.1007/s11948-021-00319-4.

[17] G. Falco, B. Shneiderman, J. Badger, R. Carrier, A. Dahbura, D. Danks, M. Eling, A. Goodloe, J. Gupta, C. Hart, M. Jirotka, H. Johnson, C. LaPointe, A. J. Llorens, A. K. Mackworth, C. Maple, S. E. Pálsson, F. Pasquale, A. Winfield, Z. K. Yeong, Governing ai safety through independent audits, Nature Machine Intelligence 3 (2021) 566–571. doi:10.1038/s42256-021-00370-7.

[18] J. Mökander, L. Floridi, Operationalising ai governance through ethics-based auditing: An industry case study, AI and Ethics 3 (2023) 451–468. doi:10.1007/s43681-022-00171-7.

[19] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys 54 (2021) 1–35. doi:10.1145/3457607.

[20] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computing Surveys 51 (2018) 1–42. doi:10.1145/3236009.

[21] J. H. Moor, The nature, importance, and difficulty of machine ethics, IEEE Intelligent Systems 21 (2006) 18–21. doi:10.1109/MIS.2006.80.
[22] M. Anderson, S. L. Anderson, Machine ethics: Creating an ethical intelligent agent, AI Magazine 28 (2007) 15–26. doi:10.1609/aimag.v28i4.2065.
[23] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen, A. Bernstein, Implementations in machine ethics: A survey, ACM Computing Surveys 53 (2020) 1–38. doi:10.1145/3419633.
[24] W. Wallach, C. Allen, Moral Machines: Teaching Robots Right from Wrong, Oxford University Press, 2008. Google-Books-ID: _r3N82ETng4C.
[25] A. Dyoub, S. Costantini, F. A. Lisi, Learning domain ethical principles from interactions with users, Digital Society 1 (2022) 28. doi:10.1007/s44206-022-00026-y.
[26] R. M. Omari, M. Mohammadian, Rule-based fuzzy cognitive maps and natural language processing in machine ethics, Journal of Information, Communication and Ethics in Society 14 (2016) 231–253. doi:10.1108/JICES-10-2015-0034.
[27] A. Dyoub, F. A. Lisi, fEDM+: A risk-based fuzzy ethical decision making framework with principle-level explainability and pluralistic validation, 2026. doi:10.48550/arXiv.2602.21746. arXiv:2602.21746.
[28] M. Abbasi, S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, Fairness in representation: quantifying stereotyping as a representational harm, in: Proceedings of the 2019 SIAM International Conference on Data Mining, SIAM, 2019, pp. 801–809.
[29] G. Hersch, T. Rowe, Allocative fairness, Philosophy Compass 20 (2025) e70042.
[30] Z. Wang, C. Huang, X. Yao, Procedural fairness in machine learning, arXiv preprint arXiv:2404.01877 (2024).
[31] R. T. Rabonato, L. Berton, A systematic review of fairness in machine learning, AI and Ethics 5 (2025) 1943–1954.
[32] L. A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning—i, Information sciences 8 (1975) 199–249.
[33] G. Klir, Fuzzy sets and fuzzy logic, 1995.
[34] B. Efron, Bootstrap methods: another look at the jackknife, in: Breakthroughs in statistics: Methodology and distribution, Springer, 1992, pp. 569–593.
[35] C. Wissler, The spearman correlation formula, Science 22 (1905) 309–311.
[36] M. G. Kendall, Rank correlation methods., Griffin, 1948.
[37] C. Spearman, The proof and measurement of association between two things, The American journal of psychology 100 (1987) 441–471.
[38] W. J. Conover, Practical nonparametric statistics, john wiley & sons, 1999.
[39] S. Prion, K. A. Haerling, Making sense of methods and measurement: Spearman-rho ranked-order correlation coefficient, Clinical Simulation in Nursing 10 (2014) 535–536.
[40] P. Blanchard, D. J. Higham, N. J. Higham, Accurately computing the log-sum-exp and softmax functions, IMA Journal of Numerical Analysis 41 (2021) 2311–2330.
[41] J. Rawls, The idea of an overlapping consensus, Oxford J. Legal Stud. 7 (1987) 1.
[42] M. MacAskill, K. Bykvist, T. Ord, Moral uncertainty, Oxford University Press, 2020.
[43] I. Gabriel, Artificial intelligence, values, and alignment: I. gabriel, Minds and machines 30 (2020) 411–437.
[44] C. Rudschies, I. Schneider, J. Simon, Value pluralism in the ai ethics debate–different actors, different priorities, The International Review of Information Ethics 29 (2020).
[45] D. G. Altman, J. M. Bland, Standard deviations and standard errors, Bmj 331 (2005) 903.
[46] T. Popoviciu, Sur les équations algébriques ayant toutes leurs racines réelles, Mathematica 9 (1935) 20.
[47] R. Bhatia, C. Davis, A better bound on the variance, The american mathematical monthly 107 (2000) 353–357.