Innovation and Technology Studies
Print ISSN: 3007-6919
Online ISSN: 3007-6927
DOI: https://doi.org/10.61784/its4030

WHO IS ACCOUNTABLE FOR AI DECISIONS? RETHINKING LEGAL LIABILITY AND ETHICAL RESPONSIBILITY THROUGH COUNTERFACTUAL HARM APPORTIONMENT

XiaoGuo Chen
Georgetown University Law Center, Washington, DC 20001, USA.

Abstract: When an automated decision wrongs a person, responsibility is distributed across the provider that built the model, the deployer that configured it and the human overseer who confirmed it. Prevailing attribution rules either leave part of this harm unassigned or concentrate it on the last human in the loop. This paper introduces SHARE (Shapley Harm Apportionment with Risk-controlled Exposure), a framework that models the decision pipeline as a structural causal model in which each actor's actual conduct is contrasted with a legally specified standard of care, apportions every counterfactual harm among the actors with the Shapley value, and certifies the harm attributable to the AI system through group-conditional conformal risk control. We prove that the Shapley rule is the unique apportionment that satisfies four legal desiderata: no responsibility gap, no liability without causal contribution, equal treatment and claim consistency. On 34,711 real human–AI decisions from the HAIID study, 16.4% of erroneous AI-assisted decisions were induced by advice that overturned a correct human judgment, and raising the provider's stated accuracy from 65% to 95% increased the adoption of wrong advice by 21.3 percentage points. In COMPAS, German Credit and Adult pipelines calibrated with these behavioral data, the last-in-loop rule assigns up to 81.5% of the attributable harm to the overseer beyond its causal share, and the but-for test leaves up to 27.5% of it unassigned. Harm-aware routing certified by conformal risk control meets the target exposure while deferring 2.7 times fewer cases than confidence-based routing, and group-conditional calibration removes the excess exposure of the most affected group.
Keywords: Accountability; Legal liability; Responsibility gap; Counterfactual harm; Shapley value; Conformal risk control; Human oversight; Automation bias; AI Act

1 INTRODUCTION

Automated systems now inform decisions on pretrial detention, credit and employment, and they occasionally wrong the people they assess. When such harm occurs, the question that matters to the affected person, namely who is accountable, rarely has a single answer. A risk score is produced by a provider, converted into a recommendation by a deployer that chooses an operating threshold, and confirmed or overridden by a human overseer. This is the problem of many hands that Nissenbaum identified for computerized society [1], sharpened by learning systems whose behavior no single actor fully controls, the so-called responsibility gap [2-3]. Sociotechnical studies show that, in practice, responsibility tends to collapse onto the nearest human operator, who becomes a moral crumple zone that protects the integrity of the technical system [4].

Regulation has begun to address this distribution explicitly. The EU Artificial Intelligence Act assigns distinct obligations to providers (data governance, accuracy and transparency) and to deployers (use in accordance with the instructions and human oversight), and it requires human overseers to remain aware of automation bias [5]. The revised Product Liability Directive treats software, including AI systems, as a product and introduces presumptions of defectiveness and causation [6], and the proposed AI Liability Directive pursues similar presumptions for fault-based claims [7]. These instruments specify who owes which duty, but not how a single harm should be divided when several actors deviate from their duties at the same time. Classical tort doctrine offers the but-for test and its refinement for overdetermined outcomes, the NESS test [8], the reasonable-alternative-design requirement for design defects and negligence standards for the human users of AI [9-10]. These tools were crafted for a small number of tortfeasors. Applied to human–AI pipelines, they can leave a harm without any responsible party or count it twice, and human-oversight mandates may shift accountability to frontline workers whose capacity to exercise it is unproven [11].

Computational accounts of responsibility exist: graded responsibility and blame in structural causal models [12-13], causal definitions of harm [14-15], blame attribution among cooperating agents [16], and bounds on the counterfactual harm that a decision support system can cause to a human expert [17]. What is still missing is an operational framework that (i) maps the legally specified duties of distinct value-chain actors onto counterfactual interventions, (ii) apportions observed harm among those actors with properties that can be argued before a court or a regulator, and (iii) turns the apportionment into an ex ante, statistically certified liability budget. This paper provides such a framework and evaluates it on real data. Its contributions are as follows.

1.1 Counterfactual Model of Accountability

© By the Author(s) 2024, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

We formalize a provider–deployer–overseer pipeline as a structural causal model in which each coalition of actors follows either its actual conduct or its standard of care, with each standard anchored in the AI Act and in tort doctrine (Section 3).

1.2 Axiomatic apportionment

We show that the Shapley value is the only rule that satisfies completeness, no liability without contribution, equal treatment and claim consistency, and we derive diagnostics that quantify the responsibility gap, the double counting and the moral crumple zone created by existing doctrines (Section 4).

1.3 Certified liability exposure

We combine the apportionment with conformal risk control to guarantee, in finite samples and per protected group, that the expected harm attributable to the AI system stays below a target [18], and we introduce a harm-aware routing score that lowers the cost of this guarantee (Section 4.3).

1.4 Empirical evidence

Using 34,711 real human–AI decisions and pipelines built on COMPAS, German Credit and Adult [19], we quantify advice-induced harm, the effect of providers' accuracy claims, the misallocation produced by current doctrines and the price of certification (Sections 5 and 6).
Figure 1 summarizes the framework.

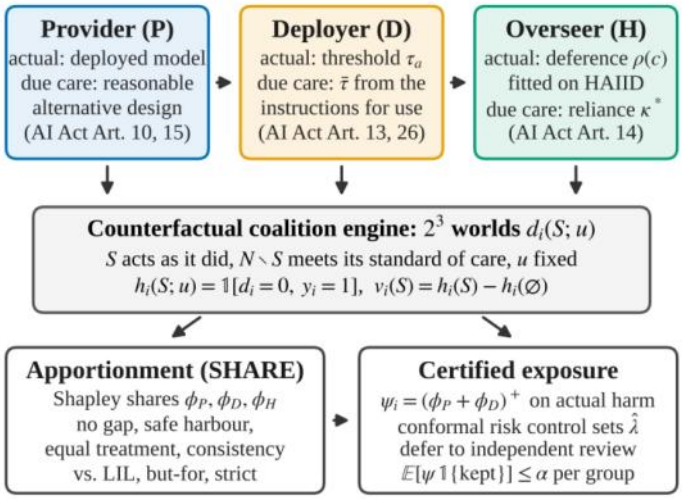

Figure 1 Overview of SHARE

Note: Each actor's actual conduct is contrasted with its standard of care; the coalition engine evaluates all counterfactual worlds; harm is apportioned with the Shapley value, and the exposure attributable to the AI system is certified with conformal risk control.

2 RELATED WORK

2.1 Accountability and Responsibility Gaps

Accountability is commonly analyzed as a relation in which an actor must answer for its conduct to a forum that can impose consequences; for AI, Novelli et al. separate its compliance, reporting, oversight and enforcement functions [20]. Cooper et al. revisit Nissenbaum's four barriers (many hands, bugs, the computer as scapegoat, and ownership without liability) for data-driven systems [21], and Kroll et al. argue for technical mechanisms that make automated decisions verifiable against legal standards [22]. The responsibility-gap debate and the moral crumple zone motivate the central measurement of this paper: how much harm a doctrine assigns to actors whose conduct did not make the difference, and how much it assigns to nobody [2-4].

2.2 Legal Causation and Liability for AI

Factual causation in tort law is usually established through the but-for test, and Wright's NESS test resolves overdetermination, in which several sufficient causes coincide [8]. Design-defect claims in the United States require a reasonable alternative design (RAD) that would have reduced the foreseeable risk [9]. Selbst shows that negligence law struggles when AI introduces unforeseeable errors and capacity limits at the human interface [10], and Green documents that policies requiring human oversight rarely rest on evidence that people can perform it [11]. The European instruments create role-specific duties and evidentiary presumptions [5-7]; SHARE supplies their quantitative counterpart, an apportionment of harm conditional on those duties.

2.3 Formal Responsibility and Harm

Chockler and Halpern define a degree of responsibility within the Halpern–Pearl theory of actual causality [12-13]. Beckers et al. define harm through contrastive causation against a default utility [14], and Richens et al. define counterfactual harm and benefit for decision policies [15]. Triantafyllou et al. compare Shapley- and Banzhaf-based blame in multi-agent sequential decision making and show that no single method satisfies every desirable property [16]. We adopt the Shapley value but redefine the game: players are legally defined actors, coalitions switch conduct between actual behavior and the standard of care, and the payoff is the harm suffered by a specific decision subject [23]. This construction gives each axiom a legal reading and turns the null-player property into a safe harbour for compliant actors.

2.4 Human–AI Decision Making and Risk Control

People frequently over-rely on AI advice [24], and their reliance depends on whether the advice is believed to come from an AI [19]. Learning-to-defer methods route cases between a model and an expert to maximize joint accuracy [25]. Conformal prediction provides distribution-free guarantees under exchangeability [26], and conformal risk control extends them to the expectation of any bounded monotone loss [18]. Straitouri et al. use conformal risk control to limit how often prediction sets cause counterfactual harm to experts [17]. SHARE uses the same statistical engine for a different target, the share of harm that is legally attributable to the provider and the deployer, and adds group-conditional guarantees.

3 ACCOUNTABILITY AS COUNTERFACTUAL CONTRIBUTION

3.1 Value-Chain Model

Consider decision subjects with features x_i, protected attribute a_i and outcome $y_i \in \{0, 1\}$, where $y_i = 1$ means that the subject deserves the favorable decision (no reoffence, repayment or eligibility). The pipeline involves the actor set $N = \{P, D, H\}$. The provider P supplies a score $s(x) \in [0, 1]$; the deployer D converts it into a recommendation $r = \mathbf{1}[s(x) \geq \tau]$; the overseer H, who forms an own judgment b_i with confidence $c_i \in [0, 1]$, issues the final decision d_i. Each actor k has an actual conduct π_k and a standard of care $\bar{\pi_k}$. For a coalition $S \subseteq N$, let s_S, τ_S and H_S denote the actual conduct of the members of S and the standard of care of the non-members. With the exogenous factors u held fixed, the counterfactual decision and harm are

$$d_i(S; u) = H_S(\mathbb{1}[s_S(x_i) \geq \tau_S], b_i, c_i; u) \quad (1)$$

$$h_i(S; u) = \mathbb{1}[d_i(S; u) = 0, y_i = 1] \quad (2)$$

The harm considered is a wrongful adverse decision, the outcome for which a subject would seek redress; harms to third parties, such as wrongful approvals, are handled identically by redefining h. The overseer's actual conduct follows an empirically estimated deference policy: when $r_i \neq b_i$, the overseer adopts the recommendation if $u_i < \rho(c_i)$, where ρ is fitted on real interaction data (Section 5). Its standard of care is appropriate reliance, which adopts the recommendation if and only if $c_i < \kappa^*$, where κ^* minimizes the error of the fully compliant pipeline. Because u only decides whether the overseer defers, each case has two exogenous states with probabilities $\rho(c_i)$ and $1 - \rho(c_i)$, and all expectations below are computed exactly.

3.2 Standards of Care

Table 1 lists the actual conduct, the standard of care and the legal anchor of each actor. The standard of care is a normative input supplied by law or regulation rather than an output of SHARE; the framework computes how much harm each deviation from it causes, so that a change in the legal standard translates directly into a change in liability.

Table 1 Actors, Conduct and Standards of Care

Actor	Actual conduct π_k	Standard of care $\bar{\pi_k}$	Legal anchor
Provider (P)	Vendor score (COMPAS); model trained on non-representative data (German, Adult)	Reasonable alternative design trained on representative data	AI Act Art. 10, 15 [5]; RAD test [9]; PLD [6]
Deployer (D)	Operating threshold $\tau_a = 0.6$	Documented operating point $\bar{\tau} = 0.5$	AI Act Art. 13, 26 [5]
Overseer (H)	Confidence-dependent deference $\rho(c)$ fitted on HAIID	Appropriate reliance with threshold κ^*	AI Act Art. 14(4)(b) [5]; negligence [10]

3.3 Doctrinal Attribution Rules

We compare SHARE with three rules that summarize current practice. Under the last-in-loop (LIL) rule, the human who confirmed a harmful decision is liable for it. Under strict provider liability (SPL), the provider bears every harm in which its system recommended the adverse decision and the overseer bears the rest. Under the individual but-for (BF) test, actor k is liable for a harm event if the harm would not have occurred had k alone complied:

$$\beta_k(i; u) = \mathbb{1}[h_i(N; u) = 1, h_i(N \setminus \{k\}; u) = 0] \quad (3)$$

When two deviations are each sufficient for the harm, neither is a but-for cause and a responsibility gap opens; when two deviations are each necessary, both actors are fully liable under joint and several liability and the harm is counted twice.

4 THE SHARE FRAMEWORK

4.1 Shapley Apportionment of Counterfactual Harm

For every subject and exogenous state, define the harm game $v_i(S; u) = h_i(S; u) - h_i(\varnothing\ ; u)$, the excess harm when the actors in S deviate and the others comply. The responsibility of actor k is its Shapley value

$$\varphi_k(i; u) = \sum_{S \subseteq N \setminus \{k\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [v_i(S \cup \{k\}; u) - v_i(S; u)] \tag{4}$$

Two aggregates serve different legal purposes. The ex post liability share conditions on the attributable harm events $E^+ = \{h_i(N; u) = 1,\ h_i(\varnothing\ ; u) = 0\}$, harms that occurred and would not have occurred under full compliance. The ex ante contribution averages over all cases and corresponds to the risk–utility view of design defects:

$$L_k = \mathbb{E}[\varphi_k(i; u) | E^+], \quad \Phi_k = \mathbb{E}[\varphi_k(i; u)] \tag{5}$$

Harms with $h_i(\varnothing\ ; u) = 1$ are residual: they occur even when every actor meets its standard of care. They are non-negligent and fall on the victim under fault-based liability or on the provider under strict liability.

Proposition 1 (Legal characterization). An apportionment rule L that maps harm games to actor shares satisfies (L1) completeness, $\Sigma_k L_k(v) = v(N)$; (L2) no liability without contribution, $L_k(v) = 0$ whenever $v(S \cup \{k\}) = v(S)$ for every S; (L3) equal treatment of interchangeable actors; and (L4) claim consistency, $L(v + w) = L(v) + L(w)$ for games that describe different victims or harm types, if and only if L is the Shapley value (4).

Proof sketch. (L1)–(L4) are the efficiency, null-player, symmetry and additivity axioms, which characterize the Shapley value on the space of finite cooperative games [23].

Two consequences matter for liability. An actor whose conduct equals its standard of care leaves every coalition unchanged and is therefore a null player, so compliance is an exact safe harbour. Completeness rules out both the gap and the double counting of the BF test, whereas LIL violates (L1) and (L2) by charging the overseer with residual harm and with harm it did not cause, and SPL violates (L2) and (L3).

4.2 Accountability Diagnostics

Within E^+ we measure the gap G and the over-attribution O of the BF test,

$$G = \mathbb{P}(\textstyle\sum_k \beta_k = 0 | E^+), \quad O = \mathbb{E}[(\sum_k \beta_k - 1)^+ | E^+] \tag{6}$$

the crumple-zone index $\mathrm{CZI} = 1 - L_H$, which is the share of attributable harm that LIL assigns to the overseer in excess of its causal contribution, and the misallocation of each rule, defined as half the L_1 distance between its allocation (including an unassigned category) and the SHARE allocation.

4.3 Certified Liability Exposure

Apportionment is retrospective, whereas deployers, conformity assessment bodies and insurers need a prospective guarantee. We define the system exposure of case i as the expected share of an actual harm that is attributable to the provider and the deployer,

$$\psi_i = \mathbb{E}_u[(\varphi_P(i; u) + \varphi_D(i; u))^+ \mathbb{1}\{E_i^+(u)\}] \in [0,1] \tag{7}$$

A routing score $g(x) \in [0, 1]$ sends a case to independent human review, in which the AI output is withheld and the system exposure is zero, whenever $g(x) \geq 1 - \lambda$. The loss $\ell_i(\lambda) = \psi_i\ \mathbf{1}[g(x_i) < 1 - \lambda]$ is bounded by $B = 1$ and non-increasing in λ because the review sets are nested. Given n calibration cases with empirical risk $\hat{R}_n(\lambda)$, conformal risk control selects [18]

$$\hat{\lambda} = \inf\{\lambda \in [0,1] : \frac{n}{n+1} \hat{R}_n(\lambda) + \frac{B}{n+1} \leq \alpha\} \tag{8}$$

Proposition 2 (Certified exposure). If calibration and test cases are exchangeable, then $\mathrm{E}[\ell_{n+1}(\hat{\lambda})] \leq \alpha$. Running (8) separately within each protected group yields $\mathrm{E}[\ell_{n+1}(\hat{\lambda}_a) \mid a_{n+1} = a] \leq \alpha$ for every group a.

Proof sketch. The loss is bounded, right-continuous and monotone in λ, and it vanishes at $\lambda = 1$, so the conformal risk control theorem of [18] applies; the group version applies it to the exchangeable subsample of each group.

Below $\alpha = B/(n + 1)$ the only certified policy routes every case to review, which ties the smallest certifiable budget to the size of the calibration sample. We compare a confidence score $g = 1 - |s(x) - \tau_a|$ with a harm-aware score $g(x) \approx \mathrm{E}[\psi \mid x]$, fitted by gradient-boosted regression on an audit split that is disjoint from the calibration data. Algorithm 1 summarizes the procedure; each case requires $2^{|N|} = 8$ counterfactual evaluations, so the cost is linear in the number of cases.

Algorithm 1: SHARE apportionment and certification
Input: conduct (π_k, $\bar{\pi}_k$) for $k \in N$; audit and calibration data; budget α; protected groups
1: for each case i and exogenous state u do
2: evaluate $h_i(S; u)$ for all $S \subseteq N$ by (1)–(2)
3: compute $\varphi_k(i; u)$ by (4); update L_k, Φ_k, G, O
4: compute ψ_i by (7); fit $g(x) \approx \mathrm{E}[\psi \mid x]$ on the audit split
5: for each protected group a do compute $\hat{\lambda}_a$ by (8)
6: route x to independent review iff $g(x) \geq 1 - \hat{\lambda}_{a(x)}$
Output: liability shares; certified budget (α, $\hat{\lambda}$, review rate)

5 EXPERIMENTAL SETUP

5.1 Data

HAIID. The Human–AI Interactions Dataset records [19], for five binary tasks (art dating, city recognition, sarcasm detection, census income and dermatology), an initial response given before any advice, the advice, labeled as coming either from an AI or from other people, and a final response; responses are signed confidences in [−1, 1]. After removing undecided responses, 34,711 decisions from 1,125 participants remain, 17,521 of them with AI-labeled advice. The initial response is a directly observed counterfactual for the world without AI. In the art task, the stated accuracy of the advice was experimentally set to 65%, 80% or 95%, while its realized accuracy stayed close to 78%.
Decision pipelines. COMPAS contains Broward County defendants with the vendor's decile risk scores and two-year recidivism outcomes [27]; after the standard filters 6,172 defendants remain, and African-American defendants (51.4%) form the protected group. German Credit (1,000 applicants; protected group: age ≤ 25, 19.0%) and Adult (48,842 persons; protected group: women, 33.2%) are taken from the UCI repository [28]. Each dataset is split into training (50%), audit (15%, or 10% for Adult) and a pool used for calibration and testing.

5.2 Pipeline Instantiation

Provider. In COMPAS, the actual product is the vendor's decile score, mapped to probabilities by isotonic regression on the training split; this mapping plays the role of the calibration table that a provider documents. Its RAD is an L_2-regularized logistic regression on age, prior counts, juvenile counts, charge degree, sex and jail stay, without race, following evidence that simple models match the vendor tool [29]. In German and Adult, the actual product is trained on a non-representative sample from which a fraction $\beta = 0.5$ of the favorable outcomes of the protected group is missing, a breach of the representativeness requirement of Art. 10(3) of the AI Act [5]; the RAD is the same learner (logistic regression for German, histogram gradient boosting for Adult) trained on the representative sample.
Deployer. The standard of care is the documented operating point $\bar{\tau} = 0.5$, under which an adverse decision requires that the subject be more likely undeserving than deserving. The actual deployer applies $\tau_a = 0.6$, the threshold that minimizes its own cost when a wrongful favorable decision costs 1.5 times as much as a wrongful adverse one.
Overseer. The overseer's own judgment is a logistic model fitted on 300 cases using the salient features shown to laypeople in [29] or their analogs, plus inconsistency noise in the logit. The noise scale is set so that the overseer's discrimination relative to the AI matches real human–AI comparisons, $(\mathrm{AUC}_H - 0.5) = r(\mathrm{AUC}_{\mathrm{RAD}} - 0.5)$, with $r = 0.796$ for COMPAS (62.1% versus 65.2% accuracy) [29], $r = 0.809$ for Adult (HAIID census task, which mirrors Adult) and $r = 0.673$ for German (pooled HAIID tasks). Confidence is calibrated by Platt scaling on the audit split, in line with HAIID, where the accuracy of initial answers rises from 55.0% to 80.5% across confidence bins. The actual deference policy $\rho(c) = \sigma(1.017 - 2.033c)$ is fitted on 5,355 HAIID disagreements with AI-labeled advice, and κ^* is selected on the audit split. Table 2 summarizes the resulting pipelines; the simulated COMPAS overseer reaches 62.8% accuracy, close to the 62.1% measured for real participants [29].

Table 2 Pipeline Instantiation on Held-Out Cases

Dataset	n	Prot.	$\mathrm{AUC_P}$	$\mathrm{AUC_{RAD}}$	$\mathrm{AUC_H}$	$\mathrm{Acc_H}$	κ^*	$\bar{\rho}$
COMPAS	6,172	0.514	0.712	0.729	0.672	0.628	0.38	0.62
German	1,000	0.190	0.773	0.766	0.652	0.688	0.01	0.53
Adult	48,842	0.332	0.925	0.927	0.851	0.810	0.98	0.44

Note: Prot.: share of the protected group; $\mathrm{AUC_P}$: actual product; $\mathrm{AUC_H}$, $\mathrm{Acc_H}$: overseer's own judgment; $\bar{\rho}$: mean deference on disagreement.

5.3 Protocol and Metrics

Attribution metrics are computed on the audit and pool cases. For certification, the pool is split into equal calibration and test halves 500 times, and α is expressed as a fraction of the unmitigated exposure $\mathrm{E}[\psi]$. We compare naive selection, which chooses the smallest λ with $\hat{R}_n(\lambda) \le \alpha$ without finite-sample correction, with marginal and group-conditional CRC, each using either the confidence or the harm-aware routing score. All results are means over the repetitions; the full pipeline runs on a single CPU core in about 70 s.

6 RESULTS

6.1 Evidence from Real Human–AI Decisions

Table 3 decomposes the AI-assisted decisions. Advice raised accuracy from 68.7% to 76.1%, yet in 3.9% of all decisions it overturned a correct initial judgment, compared with 11.3% in which it corrected an error. Of the erroneous final decisions, 16.4% were induced by the advice, 45.9% were overdetermined (both the participant and the advice were wrong) and 37.7% ignored correct advice. An LIL rule blames the participant for all three categories, although in the first one the advice is a but-for cause and the participant's own judgment was correct. When a correct participant received wrong advice, the participant adopted it in 35.7% of cases.

Table 3 Decomposition of AI-Assisted Decisions in HAIID

Task	n	Acc_0	Acc_{adv}	Acc_1	Switch	Induced	Correct.	Ind./err.
Art	8,691	.664	.789	.759	.532	.044	.139	.183
Census	1,352	.713	.774	.749	.557	.033	.069	.132
Cities	2,448	.722	.775	.787	.538	.053	.118	.249
Derm.	462	.703	.835	.764	.157	.004	.065	.018
Sarcasm	4,568	.703	.785	.755	.379	.028	.080	.113
All	17,521	.687	.786	.761	.488	.039	.113	.164

Note: Acc_0, Acc_1: accuracy before and after advice; Switch: adoption rate on disagreement; Induced, Correct.: advice-induced errors and corrected errors per decision; Ind./err.: share of final errors that were advice-induced.

Figure 2(a) shows that deference on disagreement falls from 68.4% at low confidence to 29.3% at high confidence, and that identical advice is adopted more often when it is labeled as AI (48.8%, 95% CI [46.1, 51.6]) than when it is labeled as human (39.3%, [36.7, 42.2]). Figure 2(b) isolates the provider's representation. With the realized accuracy held near 78%, raising the stated accuracy from 65% to 95% raised the switching rate by 25.1 points [14.1, 36.5] and the adoption of wrong advice by 21.3 points [8.5, 35.7], and advice-induced harm more than doubled, from 2.4% to 5.2% of decisions (difference 2.8 points [1.1, 4.7]). The accurate statement of 80% produced the highest net benefit (11.4%), whereas the overstated 95% claim lowered it by 4.1 points [0.8, 7.1]. The accuracy level that Art. 13 of the AI Act obliges providers to declare is therefore not neutral information: it shifts harm, and misstating it is a provider-side deviation that SHARE can price.

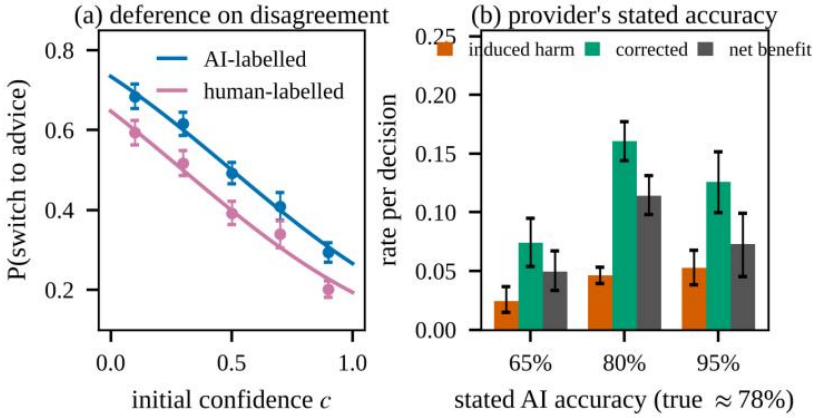

Figure 2 Real Human–AI Decisions (HAIID)

Note: (a) Probability of switching to the advice on disagreement versus initial confidence; points are binned rates with 95% intervals, lines are logistic fits. (b) Art task with AI-labeled advice: advice-induced harm, corrected errors and net benefit per decision under three stated accuracies (participant-clustered bootstrap 95% CIs).

6.2 Who Bears the Harm?

Table 4 and Figure 3 compare the allocation of attributable harm. In COMPAS, wrongful adverse decisions affect 20.3% of defendants; 43.7% of these harms are residual, so 11.4% of all decisions carry harm that is attributable to some deviation. SHARE assigns 34.3% of it to the provider, 47.2% to the deployer and 18.5% to the overseer. LIL assigns the full amount to the overseer (CZI = 0.815) and additionally charges it with every residual harm; SPL assigns 88.7% to the provider; and the BF test leaves 27.5% of the attributable harm without any liable actor while counting 26.0% twice. Relative to SHARE, the misallocation is 0.815 for LIL, 0.544 for SPL and 0.165 for BF.

Table 4 Apportionment of Attributable Wrongful Adverse Decisions

Data	Harm	Res.	Attr.	L_P	L_D	L_H	G	O	CZI	TV_{SPL}	TV_{BF}
COMPAS	.203	.437	.114	.343	.472	.185	.275	.260	.815	.544	.165
AA	.232	.460	.125	.377	.476	.146	.333	.207	.854	.547	.229
German	.136	.407	.081	.060	.080	.859	.000	.307	.141	.927	.153
Age ≤ 25	.142	.217	.111	.109	.112	.779	.000	.571	.221	.891	.285
Adult	.122	.667	.040	.112	.286	.602	.201	.060	.398	.347	.171
Women	.073	.612	.028	.317	.198	.485	.345	.127	.515	.285	.282

Note: Harm: wrongful adverse decisions per decision; Res.: residual share of harm; Attr.: attributable harm per decision; L: SHARE shares; G, O: BF gap and over-attribution; TV: misallocation relative to SHARE (for LIL it equals CZI); AA: African-American defendants.

SHARE does not exonerate humans by construction. In German, the overseer's judgment is more reliable than the recommendation on disagreements ($\kappa^* \approx 0$), so deferring is itself the deviation, and SHARE assigns 85.9% of the attributable harm to the overseer (CZI = 0.141). SPL, which places 98.7% of this harm on the provider, is the most

misallocating rule in this pipeline (0.927). In Adult, 66.7% of the harms are residual, and the attributable remainder splits into 11.2%, 28.6% and 60.2% for provider, deployer and overseer, with a BF gap of 20.1%.
The apportionment also identifies who causes disparate harm. Among African-American defendants, attributable harm is higher (12.5% versus 10.2% for the other defendants), the provider's share rises from 29.7% to 37.7% and the overseer's share falls from 23.7% to 14.6%, so the crumple zone deepens to 0.854. In Adult, the data-governance defect concentrates on women, for whom the provider's share is 31.7%, against 5.0% for men.

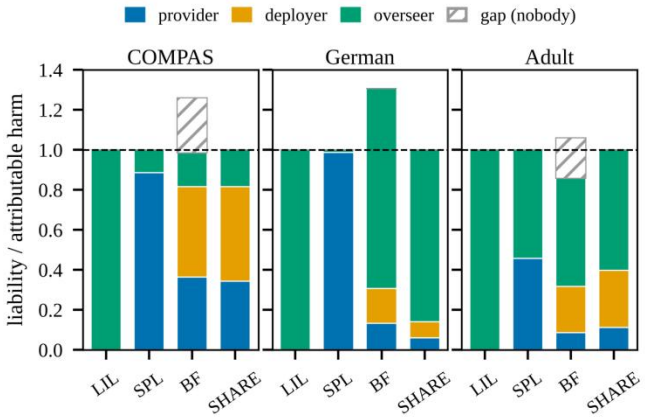

Figure 3 Allocation of Attributable Harm under the Last-In-Loop Rule (LIL), Strict Provider Liability (SPL), the But-For Test (BF) and SHARE

Note: Totals above the dashed line indicate double counting; hatched segments are harms assigned to nobody.

6.3 Reliance, Defects and Safe Harbours

Figure 4 varies the overseer's deference while everything else is held fixed. In COMPAS, moving from no deference to full deference raises attributable harm from 80 to 136 per 1,000 decisions and moves it from the overseer (58 to 6) to the provider (5 to 57) and the deployer (18 to 73), whereas LIL charges the overseer with 163 to 227 harms per 1,000 throughout. In Adult, where due care means following the stronger model, the overseer's attributable harm falls from 56 per 1,000 to zero at full deference. Responsibility therefore tracks whether reliance was appropriate, not whether a human signed the decision.

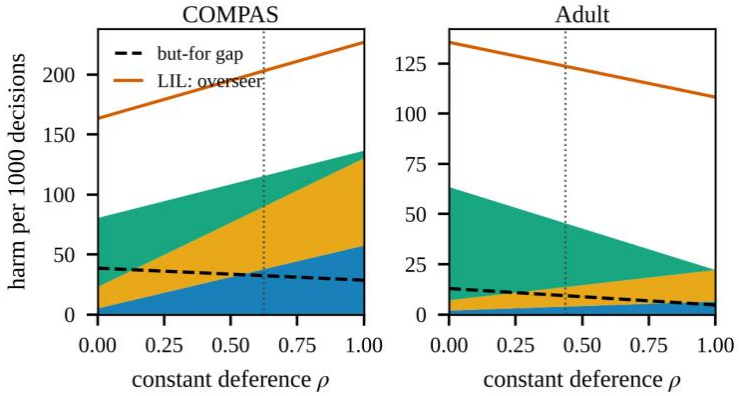

Figure 4 Attributable Harm per 1,000 Decisions under a Constant Deference Rate ρ, Stacked for Provider (Blue), Deployer (Orange) and Overseer (Green)

Note: The solid red line is the overseer's liability under LIL, the dashed line is the BF gap, and the dotted line marks the HAIID-calibrated mean deference.

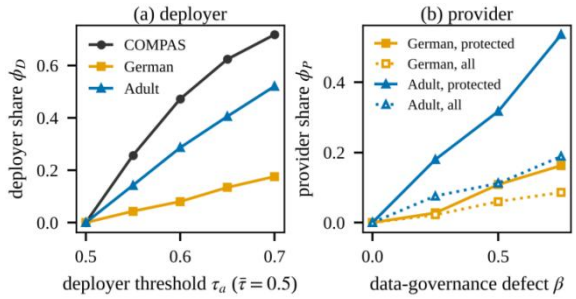

Figure 5 Safe Harbours

Note: (a) Deployer share versus its operating threshold (standard of care $\bar{\tau} = 0.5$). (b) Provider share versus the severity β of the data-governance defect, for all subjects and for the protected group.

Figure 5 confirms the safe harbour empirically. When the deployer adopts the documented threshold ($\tau_a = \bar{\tau}$), its share is exactly zero in all three pipelines, and it rises monotonically with the deviation, up to 71.8% in COMPAS at $\tau_a = 0.7$. Replacing the vendor tool by its RAD sets the provider's share to exactly zero, as does $\beta = 0$ in German and Adult. As the defect grows from $\beta = 0.25$ to 0.75, the provider's share of the harm suffered by women in Adult grows from 18.0% to 53.6%.

6.4 Ex Ante versus Ex Post Responsibility

Table 5 reports the ex ante contributions Φ_k. In COMPAS, the vendor tool is ex ante neutral (−0.28 points of excess subject harm, against +9.03 points for the deployer), yet it carries 34.3% of the ex post attributable harm. The two views diverge because the vendor score and its RAD reach similar accuracy but err on different defendants, an instance of predictive multiplicity [30]: a risk–utility analysis finds no defect, while individual victims can show that the alternative design would have spared them. The deployer's strict threshold also prevents wrongful releases and lowers third-party harm by 7.76 points. Because the Shapley value is additive (L4), each harm type can be apportioned separately and combined with whatever weights a court or a legislature adopts.

Table 5 Ex Ante Excess Harm per Decision (Percentage Points)

Dataset	Harm type	Excess	Φ_P	Φ_D	Φ_H
COMPAS	Subject	8.66	−0.28	9.03	−0.09
	Third party	−5.81	0.96	−7.76	0.99
German	Subject	5.43	0.48	0.77	4.19
	Third party	−6.88	−0.20	−0.63	−6.05
Adult	Subject	3.52	0.29	1.47	1.77
	Third party	−0.65	−0.19	−1.18	0.72

6.5 Certified Liability Exposure

Table 6 and Figure 6 evaluate certification over 500 calibration/test splits. CRC keeps the mean realized exposure at or below α in every setting (for example, 0.0477 against $\alpha = 0.0479$ in COMPAS and 0.0080 against 0.0081 in Adult), whereas naive selection exceeds α in expectation on the small German sample (0.0080 against 0.0075). The harm-aware score makes certification much cheaper: at $\alpha = 0.5\mathrm{E}[\psi]$, it routes 11.7% of COMPAS cases to independent review instead of 31.6%, a 2.7-fold reduction, and 2.2% of Adult cases instead of 4.9%. Because routed cases receive independent review, the total rate of wrongful adverse decisions also falls, from 20.8% to 17.5% in COMPAS and from 12.2% to 11.8% in Adult.

Table 6 Certified System Exposure over 500 Calibration/Test Splits

Routing / selection	Exposure	Review	Max group	Harm
COMPAS ($\alpha = 0.0479 = 0.5\mathrm{E}[\psi]$)				
Naive, confidence	0.0405	30.8%	0.0558	18.6%
CRC, confidence	0.0403	31.6%	0.0555	18.6%
Naive, harm-aware	0.0486	11.5%	0.0625	17.6%
CRC, harm-aware	0.0477	11.7%	0.0611	17.5%
Group CRC, harm-aware	0.0470	11.9%	0.0475	17.7%
German ($\alpha = 0.0075 = 0.7\mathrm{E}[\psi]$)				
Naive, confidence	0.0080	5.5%	0.0083	15.1%
CRC, confidence	0.0026	16.9%	0.0026	14.3%
Naive, harm-aware	0.0081	1.4%	0.0083	14.6%
CRC, harm-aware	0.0024	12.2%	0.0029	13.0%
Group CRC, harm-aware	0.0013	28.9%	0.0016	12.5%
Adult ($\alpha = 0.0081 = 0.5\mathrm{E}[\psi]$)				
Naive, confidence	0.0080	4.8%	0.0115	12.3%
CRC, confidence	0.0079	4.9%	0.0115	12.3%
Naive, harm-aware	0.0081	2.1%	0.0086	11.8%
CRC, harm-aware	0.0080	2.2%	0.0085	11.8%
Group CRC, harm-aware	0.0079	2.2%	0.0080	11.8%

Note: Exposure: mean realized system-attributable harm per decision; Review: share of cases routed to independent review; Max group: larger of the two group exposures; Harm: total wrongful adverse decisions per decision.

Marginal certification hides group disparities: African-American defendants in COMPAS reach 1.28α and men in Adult 1.05α. Group-conditional CRC brings every group to at most α (0.99α in both cases) for an additional review rate of only 0.2 points in COMPAS. In German, the 175 calibration cases make budgets below $1/176 \approx 0.0057$ infeasible, so CRC routes every case to review at $\alpha = 0.0054$. The smallest certifiable budget is thus an explicit function of the audit sample size, which regulators can require providers and deployers to document.

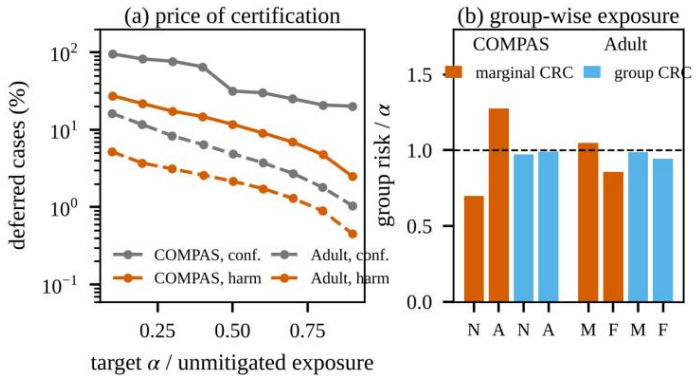

Figure 6 Certification

Note: (a) Share of cases routed to independent review that is needed to meet a target α, for confidence-based and harm-aware routing (log scale). (b) Realized group exposure relative to α at α = 0.5E[ψ] under marginal and group-conditional CRC (N: non-African-American, A: African-American, M: men, F: women).

7 DISCUSSION: RETHINKING LIABILITY AND RESPONSIBILITY

7.1 From Proximity to Contribution

The results support replacing proximity-based attribution (who touched the decision last) with contribution-based attribution (whose deviation from a duty made the difference). LIL converts the deviations of providers and deployers into the overseer's liability and charges the overseer with harms that no one could have prevented. SHARE implements proportionate liability with an exact safe harbour, which aligns incentives: each actor can reduce its exposure to zero by meeting its own standard of care, and only by doing so.

7.2 Evidence for Presumptions and Disclosure

The revised Product Liability Directive and the proposed AI Liability Directive ease the claimant's burden through disclosure duties and rebuttable presumptions [6-7]. SHARE specifies what the disclosed evidence should contain: the actors' actual conduct, the documented standards of care and enough data to replay the counterfactual worlds of (1)–(2). The HAIID results add that accuracy statements are part of a system's safety-relevant presentation, because overstated claims increase advice-induced harm.

7.3 Ex Ante and Ex Post Duties

Because the two views can disagree, as in COMPAS, a duty should state whether it refers to risk or to individual outcomes. Conformity assessment and insurance pricing are ex ante questions, answered by Φ_k and by the certified budget α; compensation of a specific victim is an ex post question, answered by $\varphi_k(i)$. Both answers come from the same additive game and therefore remain consistent.

7.4 Designing Human Oversight

Human oversight cannot serve as a liability shield. The German pipeline shows that inappropriate deference is a genuine fault, while COMPAS shows that much of the harm confirmed by humans originates upstream. Oversight duties under Art. 14 should be paired with interface measures that support appropriate reliance, such as eliciting an independent judgment before revealing the recommendation, as in the HAIID protocol, and cognitive forcing functions [24], and with routing rules such as Algorithm 1 that send the cases with the highest attributable risk to genuinely independent review.

7.5 Scope and Extensions

SHARE takes standards of care as normative inputs and computes their consequences. The same machinery applies to any number of actors, including data suppliers, integrators and providers of general-purpose models, at a cost of $2^{|N|}$ counterfactual evaluations per case. Extensions to sequential and agentic pipelines, in which actors interact over time [16], and to field data from professional decision makers are natural next steps.

8 CONCLUSION

Who is accountable for an AI decision is a question of counterfactual contribution relative to legal duties. SHARE answers it with the only apportionment rule that leaves no gap, attaches no liability without contribution, treats equals equally and stays consistent across claims, and it turns the answer into a certified, group-wise bound on the harm attributable to the AI system. Real human–AI data show that advice induces a measurable share of human errors and that providers' accuracy claims move that share. Across three decision domains, the last-in-loop rule and strict provider liability misallocate up to 81.5% and 92.7% of the attributable harm, respectively, whereas harm-aware certification bounds the system's exposure

at a modest review cost. These results support liability rules that follow causal contribution rather than proximity, and they give providers, deployers and regulators a common, auditable language for responsibility.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Nissenbaum H. Accountability in a computerized society. Science and Engineering Ethics, 1996, 2(1): 25-42.
[2] Matthias A. The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 2004, 6(3): 175-183.
[3] Santoni de Sio F, Mecacci G. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. Philosophy & Technology, 2021, 34(4): 1057-1084.
[4] Elish M C. Moral crumple zones: Cautionary tales in human-robot interaction. Engaging Science, Technology, and Society, 2019, 5: 40-60.
[5] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689, 2024.
[6] Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC. Official Journal of the European Union, L 2024/2853, 2024.
[7] European Commission. Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive). COM(2022) 496 final, Brussels, 2022.
[8] Wright R W. Causation in tort law. California Law Review, 1985, 73(6): 1735-1828.
[9] American Law Institute. Restatement (Third) of Torts: Products Liability, § 2, 1998.
[10] Selbst A D. Negligence and AI's human users. Boston University Law Review, 2020, 100(4): 1315-1376.
[11] Green B. The flaws of policies requiring human oversight of government algorithms. Computer Law & Security Review, 2022, 45: 105681.
[12] Chockler H, Halpern J Y. Responsibility and blame: A structural-model approach. Journal of Artificial Intelligence Research, 2004, 22: 93-115.
[13] Halpern J Y. Actual Causality. Cambridge: MIT Press, 2016.
[14] Beckers S, Chockler H, Halpern J Y. A causal analysis of harm. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2022, 35.
[15] Richens J G, Beard R, Thompson D H. Counterfactual harm. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2022, 35.
[16] Triantafyllou S, Singla A, Radanovic G. On blame attribution for accountable multi-agent sequential decision making. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2021, 34.
[17] Straitouri E, Thejaswi S, Gomez Rodriguez M. Controlling counterfactual harm in decision support systems based on prediction sets. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2024, 37.
[18] Angelopoulos A N, Bates S, Fisch A, et al. Conformal risk control. In: Proceedings of the International Conference on Learning Representations (ICLR), 2024.
[19] Vodrahalli K, Daneshjou R, Gerstenberg T, et al. Do humans trust advice more if it comes from AI? An analysis of human-AI interactions. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), 2022: 763-777.
[20] Novelli C, Taddeo M, Floridi L. Accountability in artificial intelligence: What it is and how it works. AI & Society, 2024, 39(4): 1871-1882.
[21] Cooper A F, Moss E, Laufer B, et al. Accountability in an algorithmic society: Relationality, responsibility, and robustness in machine learning. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2022: 864-876.
[22] Kroll J A, Huey J, Barocas S, et al. Accountable algorithms. University of Pennsylvania Law Review, 2017, 165(3): 633-705.
[23] Shapley L S. A value for n-person games. In: Contributions to the Theory of Games II, Kuhn H W, Tucker A W, eds. Princeton: Princeton University Press, 1953: 307-317.
[24] Buçinca Z, Malaya M B, Gajos K Z. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 2021, 5(CSCW1): Art. 188.
[25] Mozannar H, Sontag D. Consistent estimators for learning to defer to an expert. In: Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR, 2020, 119: 7076-7087.
[26] Vovk V, Gammerman A, Shafer G. Algorithmic Learning in a Random World. New York: Springer, 2005.
[27] Angwin J, Larson J, Mattu S, et al. Machine bias. ProPublica, 2016.
[28] Dua D, Graff C. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, 2019.

[29] Dressel J, Farid H. The accuracy, fairness, and limits of predicting recidivism. Science Advances, 2018, 4(1): eaao5580.
[30] Marx C T, du Pin Calmon F, Ustun B. Predictive multiplicity in classification. In: Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR, 2020, 119: 6765-6774.