SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Who Is Really Making The Decision? Agency Without A Subject In Algorithmic Decision Systems

Ziheng Wang

Nanjing No. 1 Middle School, 301 South Zhonshan Road, Qinhuai District, Nanjing, Jiangsu, China, 210000
Corresponding Email: 18051645618@163.com

Abstract
Decision systems in which a person selects from options that an algorithm has generated, ranked, and time-stamped strain the model of agency underwriting most of our moral vocabulary: one subject, one intention, one act. This article argues that agency in such systems is best understood not as a property some node possesses but as a structural property instantiated in the allocation of decision-constitutive powers across a system. The argument proceeds by conceptual analysis, supported by a documented case set drawn from the AI Incident Database, regulatory instruments, and published evaluations of deployed models. Distributed agency is defended as a supervenient, multiply realisable, relational property whose bearer is a structure rather than a subject: weaker than collective agency, and requiring no intentions be attributed to artefacts. Four indicators — informational control, option-space shaping, temporal structure, and traceability — make the distribution readable in concrete systems, and are shown to be distinct from adjacent frameworks concerning cognitive integration and meaningful human control. Where retrospective blame fails for structural reasons, responsibility should attach to the power to design and revise the distribution; this is non-equivalent to outcome-based developer liability, and its limits are identified for cases in which no party holds revision power. Coding ten cases yields four morphologies, one of which the framework did not anticipate and which sceptical treatments of the responsibility gap struggle to accommodate.

Keywords: distributed agency; human–AI interdependence; free will; moral responsibility; algorithmic decision-making; responsibility gap.

1. Introduction

A person opens an application, looks at a short list, and taps one item. The finger belongs to the person; the intention behind the tap is the person's; the consequences follow from what was selected. Every element of the ordinary picture of action appears to be in place.

Widen the frame and the picture loses its stability. The list contained five items rather than five hundred because a ranking model scored a candidate pool and truncated it. The pool reflected what a training corpus happened to record, which reflected the behaviour of earlier users under earlier versions of the same interface. Nothing showed what had been excluded, and the moment of choice was structured by a design that rewards speed. Ask who determined the outcome, and no single locus did: the person contributed the final selection among alternatives whose composition, ordering, and framing were fixed elsewhere.

The tempting reply is that the person still decided, since they could have refused or searched independently. That preserves the traditional picture at the cost of emptying it. Formal decision authority — the capacity to be the last node in the causal chain — is cheap. What matters morally about deciding is not being last but exercising judgement over a space of live alternatives. Once the generation of that space migrates elsewhere, retaining the final click is compatible with having lost most of what made the decision one's own.

Philosophical work on artificial intelligence has largely engaged this situation through a different question: whether AI systems themselves possess agency, consciousness, or moral status. That is a dispute about property-attribution, and such disputes obscure structural facts. Whether a ranking model has intentions is orthogonal to the observation that it determines which alternatives a person will ever consider; a system can redistribute the conditions of agency without any component acquiring agency in the traditional sense.

The alternative pursued here reformulates the question: not who possesses agency in an algorithmically mediated decision, but how agency is distributed across it — what shape the distribution takes, and which nodes hold which powers.

Three commitments constrain the argument. Nothing presupposes machine consciousness; the concept is relational, and a system may instantiate a distribution of agency while every artificial component remains mindless. Second, a thesis of this kind incurs a metaphysical debt usually left unpaid: if agency is a structural property, one must say what sort of property that is and what it means for a structure rather than a subject to bear it. Section 4.3 discharges that debt. Third, the distribution must be readable from concrete systems, or the thesis is untestable; Section 4.6 proposes four indicators and Section 5 applies them.

The stakes are practical. If a decision is jointly produced by a person and an algorithmic infrastructure, attributing an error to "the person who used the AI" is either a fiction or a convenience. Section 6 relocates responsibility toward the power to design and revise the distribution — and argues, against an obvious suspicion, that this is not developer liability renamed.

2. Method and Materials

The primary method is conceptual analysis: the central claim concerns what agency is and how the concept should be applied, and such claims are settled by argument and by testing a framework against cases. No original data were generated and no causal claims are advanced. Engagement with prior work is concentrated in Section 3 as positioning and in Section 4.5 as differentiation. Because analysis unsupported by cases risks distinctions that cannot be applied, a documented case set serves a limited purpose: to test whether the indicators can be assigned to real systems from public evidence, and whether the framework's predicted forms exhaust what documented systems display. The case work is diagnostic, not evidential.

Four categories of public source were used. Incident records come from the AI Incident Database (AIID), a curated public catalogue of documented AI harms modelled on aviation incident reporting ([19]). Six numbered incidents were used — 4, 37, 57, 101, 123, and 373 — selected because their published descriptions make the allocation of decision-relevant powers explicit; each is described where it is analysed in Section 5. Regulatory instruments: Article 27 of the EU Digital Services Act obliges platforms to disclose the main parameters of their recommender systems, used here as documentary evidence about where option-space authorship sits, not as a dataset. Regulatory inventories: the published database of FDA-authorised AI/ML medical devices supplies an assistive-to-autonomous spectrum and records that most authorisation announcements did not identify the AI/ML basis of the device ([4]). Published evaluations supply evidence about the epistemic standing of the advice these distributions are built around ([5], [9], [24]).

Ten case types were assembled purposively across two dimensions — high versus low stakes, strong versus weak mediation — deliberately including least-distributed cases as controls. Each was coded on the four indicators using a three-level scale recording displacement: how far a power sits from the human occupying the nominal decision position. On traceability, Low therefore denotes a system whose process can be reconstructed.

Four limitations bear on how this may be read. Coding was performed by a single analyst without replication, so assignments are interpretive rather than measured. The sample is purposive and small, so no claim is made about frequency. AIID records are compiled from public reporting of variable completeness — a limitation the database's editors acknowledge — and several describe allegations rather than adjudicated findings. The recommender row rests on a regulatory obligation rather than any platform's transparency report, and is the weakest-evidenced. These constraints suit the purpose: showing that indicators are assignable and a typology non-vacuous requires documented instances and honest coverage, not representativeness.

3. Theoretical Background and Positioning

Modern philosophy of action inherits a compact schema: an agent acts for a reason, and the reason — a pro-attitude towards an end together with a belief that the action promotes it — both justifies and causes what is done ([3]). Its limitation lies in a presupposition it never states, that intention, belief, and movement belong to

one bounded subject. Refinements deepen rather than dislodge this: hierarchical accounts locate freedom in second-order volitions, so that one who acts on a desire he wants to be effective acts freely while the wanton does not ([8]) — a distinction internal to a single psychology. A well-known objection holds that the causal story omits the agent himself, and that participation rather than internal causation turns an event into an action ([31]). That is instructive for a reason its author did not intend: participation is precisely what becomes hard to locate when the alternatives among which one participates were assembled elsewhere.

Attributing an action conventionally requires that the agent controlled what happened and that the action originated in the agent. Both are stated as thresholds formulated for cases where causal contributions are few and legible. Algorithmic infrastructure violates those assumptions: control becomes graded, since a user may control the final selection while controlling neither the candidate set nor the ranking criteria; sourcehood becomes ambiguous, since the preferences making one option attractive may themselves have been cultivated by prior exposure to the system.

Work on extended and distributed cognition supplies the crucial precedent. The extended mind thesis holds that a reliably available external resource playing the functional role an internal state would otherwise play counts as part of the cognitive system ([1]). Studies of socio-technical work converge from another direction: computation performed by an aircraft cockpit is a property of the assembly of pilots, instruments, and procedures rather than of any pilot ([12]), and cognition is better located in cultural ecosystems of practices and artefacts than in isolated heads ([13]). Integration is a matter of degree along multiple dimensions, so human–artefact systems occupy positions in a space rather than falling on one side of a line ([10]).

The thesis is contested, and the present argument does not presuppose it. The standing objection charges a coupling-constitution fallacy: causal coupling to an external resource does not make the resource a constituent, and a more conservative embedded cognition hypothesis explains the same phenomena ([25]). Section 4 needs only that decision-constitutive powers can be held by parties other than the deciding person — a claim about allocation of authority, not about the boundaries of mind. An embedded-cognition theorist holding that all thinking happens inside the skull can consistently accept that the option set was composed elsewhere. This literature therefore functions as precedent and vocabulary rather than premise. What it does not supply, on either reading, is a treatment of agency: showing that thinking crosses the organism's boundary is not yet showing that acting, and answering for what one does, are similarly distributed.

Philosophy of technology supplies the complement. Technologies do not merely transmit purposes; they co-shape perception and action, so design decisions materialise moral positions ([32]). Actor-network theory goes further, treating nonhumans as participants on symmetrical terms with humans while detaching agency from intentionality ([28]) — a detachment that makes the vocabulary usable, since one can say a ranking model exercises agency in the network-theoretic sense without asserting anything about what it wants.

Ethics offers parallel resources. Artificial systems can be treated as agents at an appropriate level of abstraction and can be accountable for effects without being morally responsible ([7]); and morally significant outcomes frequently aggregate from individually negligible contributions, so morality itself becomes distributed and depends on an infrastructure of non-moral enablers ([6]). Bringing the two literatures together yields a nuanced conclusion: artefacts densely and transparently integrated into a user's processing may form a system with agency, while loosely coupled systems lacking central control may not ([11]). A further strand applies group-agency machinery directly to artificial systems ([16]); Section 4.3 argues that human–AI assemblies fail its conditions, and that the failure is informative.

These resources explain how cognition extends and how moral outcomes aggregate. None reconstructs agency for algorithmic decision-making — specifying what a distribution of agency is, what status it has, and how a given form would be recognised. Existing work maps ethical concerns about evidence quality, unfair outcomes, and traceability ([20]) while treating agency as background. A sceptical literature denies there is a problem at all: responsibility may be a dynamic practice able to absorb new technological entities without leaving anyone unanswerable ([30]; a reply to the pessimism of [18] and [29]), or the gap may reduce to familiar problems once disambiguated ([14]). These bear on Section 6 rather than Sections 4 and 5, and are answered in Section 7.

4. A Conceptual Framework

4.1 Two preliminary stages

At the first stage a person uses an algorithmic system as an instrument: a translation tool renders a phrase, a calculator returns a product. Dependence may be extensive, but agency remains undivided, because the user sets the task, evaluates the output against independently held standards, and can reject it. The artefact is causally necessary and agentially inert.

The second stage begins when the system participates in forming the judgement rather than executing it. A recommender does not answer a question the user has formulated; it shapes which questions seem worth asking. A diagnostic model does not confirm a clinician's hypothesis; it supplies the hypothesis that examination will test. The coupling is bidirectional: outputs shape the user's cognitive state, and user responses become training signal shaping future outputs. Attribution of the judgement to the person becomes an approximation — not because users are weak-willed, but because even a vigilant user cannot evaluate what was never displayed. What integration transfers is not effort but scope.

4.2 Distributed agency

The third stage is a conceptual transition rather than a further increment. At stages one and two, agency remains something nodes possess in varying degrees, and one asks how much the person retains. That question is malformed for tightly coupled systems, because agency in them is a structural property of the relational configuration rather than an intrinsic property of any node.

Take the powers jointly constituting agency in a decision: determining what information is available, what alternatives exist, when the decision occurs, and whether the process can afterwards be reconstructed. In an unmediated decision one subject holds all four, which is why the single-subject model works — it describes the limiting case in which the distribution is maximally concentrated. In an algorithmically mediated decision these powers are held by different parties: the model holds informational control, a design team holds option-space authorship, the interface holds temporal structure, and traceability may be held by no one. Agency has not disappeared, nor migrated into the machine; it has been decomposed and reallocated. The system decides, in that the outcome is a determinate function of the configuration; no component decides, in the sense that would license the traditional attribution.

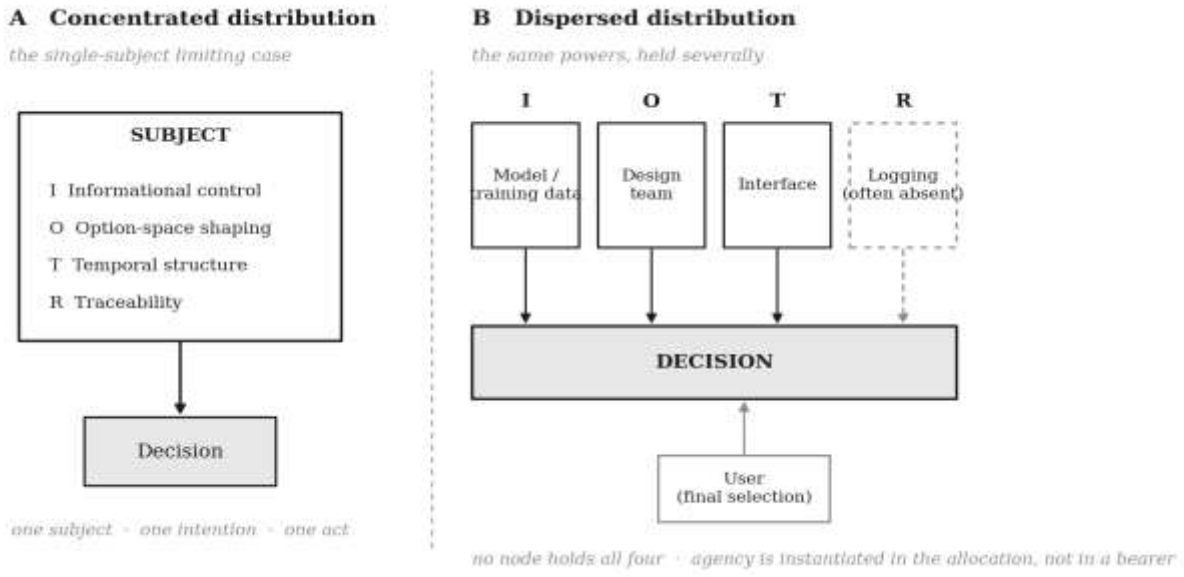

Figure 1. The decomposition of decision-constitutive powers. In Panel A the single-subject model applies without adjustment. In Panel B the same powers are allocated across distinct parties; the decision is a determinate function of the allocation, though no node determines it. Traceability is dashed because it is the one power that may be held by no party at all.

In Figure 1, Panel A shows the concentrated case, where one arrow carries the whole explanation; Panel B the dispersed case, where the user's contribution enters alongside the others rather than governing them.

A misreading must be blocked. To say agency is structural is not to say individuals have none. The person continues to deliberate, to hold reasons, and to be capable of refusal; what is denied is that this deliberation exhausts the decision. Showing that a cockpit computes a landing speed does not show pilots stop thinking; it shows what they think is one contribution to a computation whose other parts are elsewhere ([12]). The individual remains an agent, but is no longer the whole of the agency at work.

4.3 The metaphysics of a structural property

Four questions follow, and leaving them unanswered has been the standing weakness of relational proposals.

What does it supervene on? Distributed agency supervenes on the allocation of the four constitutive powers across a system's components, together with the couplings among them: two systems identical in that allocation cannot differ in their distribution of agency. Supervenience secures determination without reduction, since the property is fixed by the base but is a property of no element of it — allocation facts being irreducibly relational, as a wealth distribution is not identical to any individual's wealth.

Is it multiply realisable? It is, and this bounds the framework's scope. A bureaucratic chain in which a clerk applies a rule to a file prepared by others instantiates the same allocation as a triage nurse acting on a model's alert. Finding one form across algorithmic and pre-algorithmic bureaucracy shows the concept is no artefact of novelty, but entails that algorithmic systems are not sui generis: any claim that they are must rest on degree, speed, or opacity rather than kind.

What is the bearer? Here a natural slide must be resisted. The system does not possess agency as a subject possesses a capacity; the powers belong severally to different parties, and what the system exhibits is a shape into which they fall. Predication divides accordingly: agency is predicated of the structure, responsibility of the parties holding the powers. Nothing is thereby asked to answer that cannot answer — and Section 6 relocates responsibility for precisely this reason.

Is this ontologically robust or merely a useful description? The answer draws on the method of levels of abstraction ([7]). At a level whose observables are decision-constitutive powers, distributed agency is a real property: determinate, variable across systems, and supporting counterfactuals about what would have happened had the allocation differed. At a lower level, whose observables are transistors and neurons, it disappears. This is not anti-realism, because which level is appropriate is fixed by the explanatory question rather than by preference: if the question is who is answerable, the level at which powers are observable is the only one at which it can be posed.

Placing the resulting entity is instructive. Collective agents are functionally organised to form and revise attitudes and can be modelled as rational subjects, though they may lack phenomenal consciousness ([15]). A human–algorithm assembly is weaker: no attitude-forming organisation, no procedure for revising joint intentions, no candidate for a unified perspective. Distributed agency is therefore not a diminished form of collective agency but a distinct category — agency without a subject at any level, rather than agency belonging to a subject constituted out of many. That is a stronger claim than the group-agency literature makes, and it rests on the powers remaining genuinely dispersed rather than aggregated by any procedure.

4.4 What distributed agency is not

Three demarcations prevent vacuity. Distributed agency is not collective agency, for the reason just given, and not machine agency: nothing asserts that algorithmic components have intentions or interests. Most importantly it is not mere causal contribution. Weather contributes causally to countless decisions without holding any decision-relevant power; what distinguishes distribution from causal background is that some node holds a power constitutive of deciding — over information, options, timing, or traceability — and holds it systematically rather than incidentally. Distribution requires a division of decision-relevant authority, not merely a plurality of causes.

4.5 Differentiation from adjacent frameworks

Three existing frameworks describe overlapping territory, and the indicators below would be redundant if they merely renamed one. The differences concern each framework's object.

Cognitive integration. An influential taxonomy characterises human–artefact systems along dimensions including information flow, reliability, trust, transparency, and transformation ([10]). Its object is the degree of integration; the present indicators have a different object, the location of decision-constitutive authority, and the two come apart in both directions. A resource may be densely integrated, trusted, and individualised while holding no option-space authority, as with a well-learned personal notation system; conversely it may be poorly integrated and distrusted while holding complete option-space authority, as with a court-mandated risk

instrument a sceptical judge must consult. Integration measures how close the coupling is; distribution measures who holds what.

Meaningful human control. The tracking and tracing conditions require that a system respond to relevant human moral reasons and that outcomes be traceable to at least one human with appropriate understanding ([27]). The overlap with traceability is real. The difference is modal: tracking and tracing are normative requirements specifying when control is adequate, whereas the indicators are descriptive diagnostics specifying how authority is in fact allocated, including in systems failing every normative requirement. The descriptive layer permits comparison across systems a normative criterion would lump together as failing.

Distributed morality. The claim that morally significant outcomes aggregate across many agents ([6]) concerns the production of moral outcomes; the present claim concerns the constitution of decisions. A strongly distributed decision may produce no morally significant outcome at all. The indicators therefore isolate a variable that adjacent frameworks hold fixed, presuppose, or address only normatively.

4.6 Four diagnostic indicators

Informational control asks who determines what the deciding person sees: high where selection, filtering, and ordering are algorithmically determined and opaque, low where the person can inspect the underlying evidence.

Option-space shaping asks who generates and bounds the alternatives: high where they are produced and truncated upstream, low where the person can construct alternatives outside the system. This does the most work, isolating the difference between influencing a choice and authoring the space within which choosing occurs.

Temporal structure asks how the decision is positioned in time. Time pressure, defaults, and single-pass interfaces compress deliberation; coded high where design allows no revisitation. Its inclusion is not incidental: control conditions on responsibility have an ineliminably temporal dimension, since what an agent could have done depends on when and for how long alternatives were open ([2]).

Traceability asks whether the process can afterwards be reconstructed, coded high where it cannot. It is treated in the ethics-of-algorithms literature as a precondition for holding anyone answerable ([20]), and is the indicator most amenable to deliberate design.

Independence is not merely stipulated: in Section 5 résumé scoring codes high on option shaping yet medium on timing, and supervised driving high on shaping yet low on traceability. Each is assignable from published documentation, and together they cover the constitutive powers.

5. Case Analysis

5.1 Recommendation: generation separated from selection

Consumer and informational recommendation is the low-stakes, high-mediation limiting case, exhibiting the central structure with unusual clarity: the party generating the option set and the party selecting from it are entirely distinct. The Digital Services Act's requirement that platforms disclose the main parameters of their recommenders presupposes exactly this separation — a disclosure obligation about ranking parameters is intelligible only if ranking is an authored artefact serving specified objectives rather than a neutral presentation of what exists. The individual determines which item is chosen; the system determines what "the items" are.

The stakes of any single recommendation are low, which is why the case is instructive: a strongly distributed structure is compatible with trivial consequences, so structure and severity are independent variables. Figure 3 traces the reflexive loop this creates and marks where intervention power exists within it — a distinction Section 6.5 relies on. Preferences formed under a distribution return to it as training signal, so the sourcehood condition erodes across iterations rather than in any single instance. What a disclosure obligation establishes is where authorship sits, not how far any particular platform displaces judgement; that stronger claim would require the transparency reports themselves.

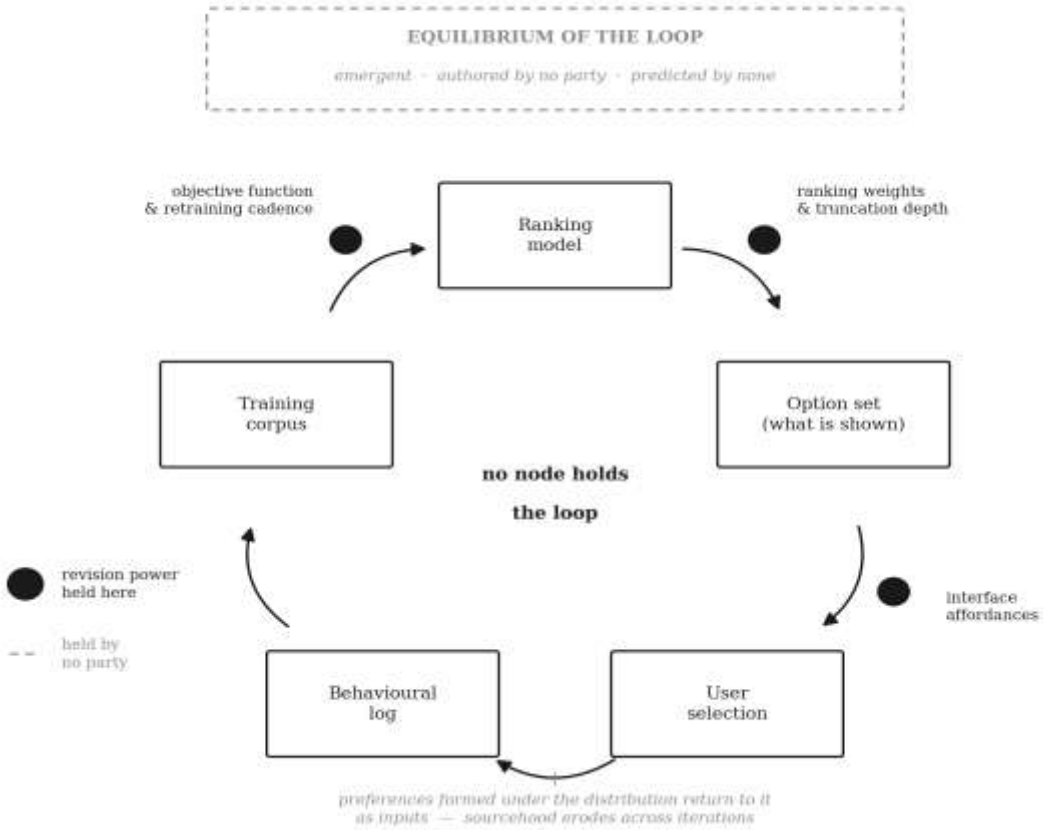

Figure 3. The reflexive loop in recommendation, and the limits of revision power. Solid markers indicate points at which a party holds intervention power. The dashed banner denotes what no party holds: the equilibrium the loop settles into is authored and predicted by none, though each lever remains adjustable.

5.2 Assisted judgement: substantive judgement compressed

Algorithmically assisted judgement inverts the stakes while preserving the structure, and here the divergence between formal and substantive authority is sharpest. The human retains formal final authority — the clinician acts on the alert, the recruiter reviews the ranking, the operator supervises the vehicle — but the documented record shows how little that guarantees.

Human oversight does not reliably correct algorithmic error. Experimental work on algorithm-in-the-loop decision-making found that people evaluate risk predictions poorly, do not reliably distinguish accurate from inaccurate advice, and are influenced by predictions in ways that do not track their quality ([9]). This is long recognised rather than novel: the automation literature describes misuse from over-reliance alongside disuse from distrust, with reliance governed by trust, workload, and perceived reliability rather than actual reliability ([22]). AIID Incident 4 is the limiting illustration. A vehicle operating autonomously struck and killed a pedestrian while a human safety operator occupied the driver's seat; the federal investigation established that the perception system classified the pedestrian inconsistently and that automatic emergency braking had been disabled during autonomous operation, leaving the operator as the sole intended mitigation. Formal authority had been retained at the very point where the configuration removed the means of exercising it.

Epistemic authority also outruns demonstrated performance. Reanalysis of a widely deployed recidivism instrument found accuracy comparable to untrained non-experts ([5]) — a comparison since contested on the artificiality of the crowdworker task and its generalisability, and not to be treated as settled. The weaker claim needed here is independently supported: for such tasks, interpretable models of a few variables perform on a par with proprietary black boxes, so opacity purchases no accuracy justifying the deference it receives ([24]). AIID Incident 123 shows the pattern clinically, external evaluation of a proprietary sepsis model finding alerting behaviour markedly worse than vendor claims.

AIID Incident 37 adds a dimension the philosophical literature tends to miss. The deploying firm's position was that recruiters never relied on the scoring engine alone — simultaneously true and beside the point. Where a model ranks a candidate pool, non-exclusive reliance is compatible with complete option-space displacement, because independent judgement operates on whatever survived scoring. "Not relying on it alone" describes informational control and says nothing about option-space shaping; the indicators separate what this defence conflates.

5.3 Evacuated authority

Two cases display a structure the framework did not predict, and reporting this is more useful than smoothing it over. AIID Incident 373 records a state unemployment-insurance system that adjudicated fraud determinations without human oversight at a reported error rate near 85%, imposing penalties and wage garnishments on tens of thousands. AIID Incident 57 records an automated debt-assessment programme generating individualised notices from averaged income data, later held unlawful. AIID Incident 101 sits adjacent: a benefits fraud model treating second nationality as a risk factor, contributing to wrongful accusations at a scale culminating in governmental resignation.

These are not cases of compressed human judgement. The human decision position was eliminated while the institutional form of individualised adjudication was retained: the output carried the legal force of a determination about a particular person, but no person determined it. The framework anticipated a spectrum from concentrated to dispersed authority, not a vacant locus behind which an institution goes on speaking as though it were occupied. A framework that had merely been read into the cases would not have produced a category its own construction excluded.

5.4 Cross-case comparison

High denotes substantial displacement away from the nominal human decision-maker; Low denotes retention.

Case	Anchor	Information control	Option shaping	Timing	Traceability
Content feed ranking	DSA Art. 27	High	High	High	High
Résumé scoring engine	AIID 37	High	High	Medium	High
Benefits fraud classification	AIID 101	High	High	Medium	High
Automated debt assessment	AIID 57	High	High	High	Medium
Unemployment fraud adjudication	AIID 373	High	High	High	High
Sepsis prediction alerting	AIID 123	Medium	High	High	Medium
Supervised autonomous driving	AIID 4	High	High	High	Low
Pretrial risk assessment	[5], [9], [24]	Medium	High	Medium	Low
Assistive imaging triage	FDA inventory ([4])	Medium	High	High	Medium
Calculator, grammar check	Everyday tooling	Low	Low	Low	Low

Selected rationales. Supervised autonomous driving and pretrial risk assessment code traceability Low because a federal investigative body and court records respectively permit reconstruction — preserved by external institutions rather than by the system. Unemployment fraud adjudication codes High throughout, the deploying agency being reportedly unable to produce evidence for individual determinations. Résumé scoring codes timing Medium because review was not time-critical, showing the indicators vary independently. The final row is a control: a framework that could not return Low here would be vacuous.

5.5 Four morphologies

The instrumental form (final row) codes low throughout; the traditional model applies unchanged. The option-authored form (feed ranking, résumé scoring, benefits classification) displaces agency at the front end: judgement operates normally on a space someone else composed. The oversight-shell form (sepsis alerting, supervised driving, pretrial assessment, imaging triage) codes high on option shaping while informational control remains partial and traceability is often preserved by external institutions rather than by the system;

formal authority is retained and substantive judgement compressed. The authority-evacuated form (debt assessment, fraud adjudication) differs in kind: no human node holds any decision-constitutive power, yet institutional outputs retain the form of individualised determination.

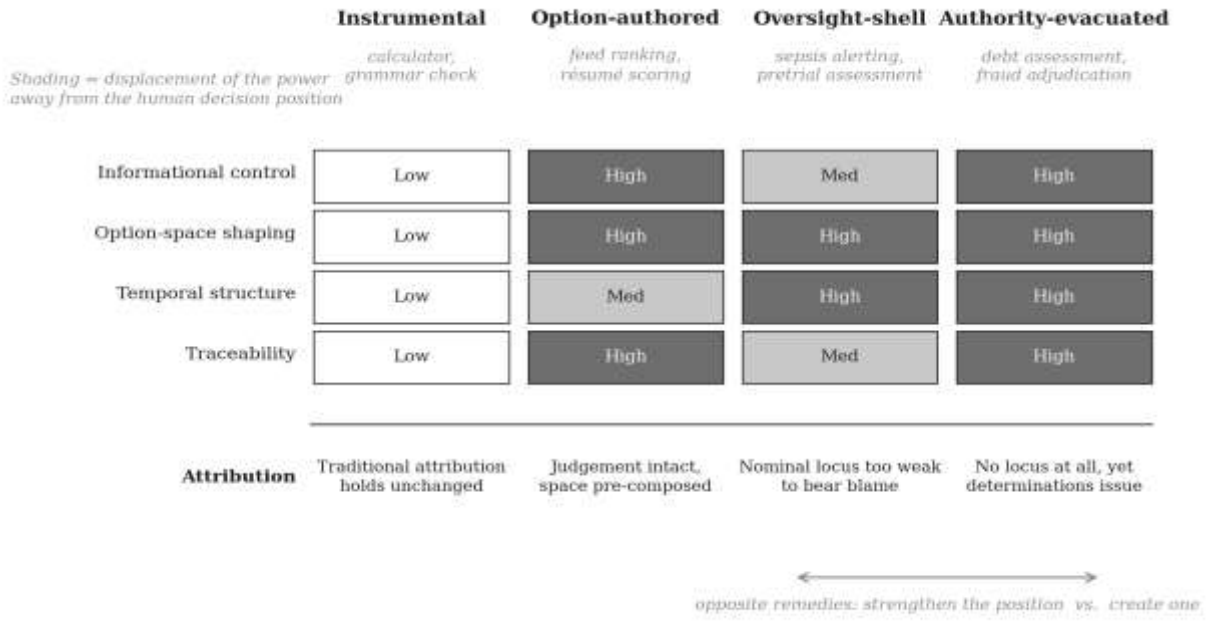

Figure 2. Four distributional morphologies and their attribution consequences. Shading indicates displacement of each power away from the human decision position. The oversight-shell and authority-evacuated forms code similarly in aggregate yet require opposite remedies.

Figure 2 makes visible a pattern the table obscures. The oversight-shell and authority-evacuated columns are close in aggregate shading, which is why an approach treating distribution as a quantity would group them; yet their attribution consequences are not merely different but opposed. Strengthening a position and creating one are different interventions, and no measure of overall displacement distinguishes them. Aggregate "amount" of distributed agency predicts little; the shape predicts which attributions fail and which interventions could succeed.

6. Reconstructing Responsibility

6.1 The gap and its anatomy

Responsibility is not a further stage of coupling but a consequence of the analysis. Attributions presuppose a locus of control and a source; if both have been decomposed across a network, attributions directed at any single node inherit the decomposition as a defect — the structural core of the responsibility gap first identified for learning systems whose behaviour operators cannot predict ([18]). An error produced by an oversight-shell system therefore cannot be attributed straightforwardly to the person who used it: their control was partial and their access to alternatives constrained, yet no other node satisfies the traditional conditions either. The model has no intentions; the designers did not decide this case; the institution did not act.

The gap is not singular. Culpability, moral accountability, public accountability, and active responsibility fail separately and for distinct technical, organisational, and legal reasons ([26]), so remedies are not interchangeable: an explainability requirement may close a public accountability gap while leaving culpability untouched. The "many hands" of organisational ethics is compounded by a problem of "many things", and by a temporal dimension in which the actions shaping the decision occurred long before it ([2]).

6.2 Three candidate allocations

Assign responsibility to the user. This preserves a determinate locus and matches institutional practice, but holds people answerable for powers they did not hold. Where overseers cannot reliably evaluate the advice

they receive ([9], [22]), the rule converts the human into an accountability sink. The mechanism has been named agency laundering — deflection of moral responsibility onto a process or artefact by a party who continues to exercise the underlying discretion ([17]) — and the present framework specifies what makes it possible: laundering succeeds precisely where the nominal position has been stripped of option-space authority while retaining formal authority, which is the oversight-shell signature. AIID Incident 4 is the paradigm, criminal proceedings having followed against the safety operator whose intervention capacity the vehicle's configuration had itself curtailed. Against the authority-evacuated form the rule does not merely misfire; it has no target.

Assign responsibility to designers and platforms for outcomes. This tracks the power to shape the option space and creates incentives where change is possible. But it over-attributes, since designers cannot foresee particular cases and general foresight is a weak basis for particular culpability; and it under-attributes, since locating responsibility entirely upstream removes any reason for the person in the loop to exercise such judgement as remains.

Distribute responsibility in proportion to causal contribution. This has the appearance of rigour and fails on details: contributions are not commensurable — a training corpus and a click share no scale — and any proportion assigned would encode a normative judgement while presenting itself as a measurement.

6.3 Structural responsibility

Each candidate seeks a party satisfying the traditional conditions, when the analysis says those conditions are jointly satisfied by no one. The alternative is to change what responsibility attaches to.

Where agency is distributed, responsibility attaches to the power to design and revise the distribution: to whoever could exercise authority over how informational control, option-space shaping, temporal structure, and traceability are allocated. This relocates responsibility rather than weakening it, restoring the connection between responsibility and control that traditional attribution has already lost. A party who can determine what alternatives a decision-maker will ever see holds a decision-constitutive power, and holding it grounds answerability.

Structural responsibility presupposes retained individual responsibility: the person in the loop remains responsible for exercising the powers they hold, and ceases to be responsible for powers they do not. Work on meaningful human control converges on this shape from system design ([27]); in the present vocabulary, tracking and tracing are requirements on the distribution rather than on any node.

6.4 Why this is not developer liability renamed

The obvious suspicion is that this reinstates the second candidate under a new description, since the parties holding design and revision power are largely the designers and platforms just rejected. The two views come apart in both directions.

Consider two hospitals deploying the same sepsis-prediction model with identical predictive performance. Hospital A displays each alert with a confidence band and the features driving it, permits override with a single action, logs every alert, override, and outcome, reviews the log monthly, and lets clinicians adjust the threshold. Hospital B displays a binary alert with no features, requires written justification for override, logs nothing, and fixes the threshold centrally with no local review. Suppose that over one year both record the same number of adverse outcomes attributable to model error.

Outcome-based developer liability must treat these hospitals identically: same product, same performance, same harm. Structural responsibility does not. B has allocated informational control, temporal structure, and traceability so as to make substantive judgement practically unavailable and retrospective reconstruction impossible; A has allocated them so as to preserve both. On the present account B is answerable and A is not, notwithstanding identical outcomes. The divergence runs the other way too: if in a second year A suffers more adverse outcomes through statistical variation alone, outcome-liability makes A more culpable while structural responsibility still makes B more culpable, responsibility here being indexed to the distribution rather than the result.

Three further differences follow. The object differs: candidate two evaluates results, the present view allocations. The temporality differs: candidate two is retrospective and harm-triggered, whereas structural responsibility is assessable in advance, so a system can be criticised before it has injured anyone — something outcome-based allocation cannot do. The bearers differ: design and revision power is held not only by developers but by procuring institutions, deploying managers, interface teams, and regulators setting disclosure obligations.

6.5 The limiting case: distributions no one holds

A harder objection targets the framework's own paradigm. Recommender distributions are shaped by feedback between platform objectives and aggregate user behaviour; the resulting equilibrium was designed by no one and may be intelligible to no one. If no party holds revision power, structural responsibility has nothing to attach to — and the difficulty arises precisely where the account claimed its clearest application.

Three responses, of decreasing strength. First, de facto unpredictability must be distinguished from de jure absence of power. As Figure 3 shows, platforms can ordinarily alter objective functions, ranking weights, retraining cadence, and interface affordances; what they cannot do is predict the equilibrium those changes produce. Inability to predict is not inability to intervene. Structural responsibility attaches to intervention power, not predictive power — which is why it survives conditions under which retrospective blame fails, the gap having been generated by unpredictability in the first place ([18]).

Second, where a distribution is genuinely emergent and currently unownable, responsibility shifts order rather than vanishing: it becomes a second-order duty to create the conditions under which the distribution becomes revisable — instrumentation adequate to detect what it has become, monitoring of drift, capacity to roll back. Failure to build revisability is itself a design choice with an identifiable owner, and a party who has made a distribution unownable cannot invoke its unownability as an excuse.

Third, a residual class survives both responses: distributions shaped by many platforms, several regulators, and aggregate behaviour, where no single party's intervention would be decisive. Here individual structural responsibility genuinely runs out. The conclusion is not that the account fails but that it has located a boundary — the point at which allocation questions become properly collective, requiring negotiation at the level of the algorithmic social contract ([23]). An account marking where individual responsibility ends is more useful than one pretending it extends everywhere.

6.6 Normative implications

Two design requirements follow. Traceability must be built in rather than retrofitted: a system that does not record what was recommended, seen, and overridden makes retrospective allocation impossible in principle. The corpus shows how contingent this is — supplied by an external investigative body in the driving case, absent altogether in the fraud-adjudication case. Since traceability is the indicator most fully under designers' control, its absence is a distributional choice for which structural responsibility can be assigned. Revisability must likewise be preserved: a distribution no party can alter is one to which structural responsibility cannot attach, so revisability is a precondition of there being any responsible party at all. Both are moral demands on design rather than remedies applied after harm — the consequence of the view that engineering decisions materialise moral positions ([32]).

7. Objections and Replies

This is ordinary tool use; hammers shape behaviour too. If mediation sufficed, the concept would apply to all technology. But a hammer changes what is easy; it does not compose the set of things its user will consider doing, nor update that set from the user's prior behaviour. That the indicators code the final table row Low throughout shows the concept discriminates rather than expands.

Distributed agency excuses individuals. This assumes responsibility is conserved, so any relocation must be a reduction; but structural responsibility enlarges the set of answerable parties. The polarity is also reversed: it is the user-liability rule that functions as an alibi, since blaming the person at the terminal is what allows those who shaped the decision to remain unexamined.

Agency cannot be discussed without settling whether AI is conscious. The framework uses a relational concept for which consciousness is not necessary — the move permitting participation to be attributed to nonhumans without attributing intentions ([28]), and artificial agents to be accountable without being morally responsible ([7]). The bearer is a structure, not a subject, so the question of what it is like to be that bearer does not arise.

Better described as human–machine collaboration. On this view automated systems exercise a supervised and initiated agency within joint activity humans direct, so responsibility gaps are narrower than they appear ([21]). The reply grants the description for the systems motivating it and denies its generality. Where a human initiates, supervises, and can terminate a machine's activity, the account is apt and the framework codes such cases as weakly distributed. The oversight-shell form differs: the person does not initiate the option set, cannot inspect the grounds of ranking, and supervises only by occupying the final position. Collaboration presupposes the human party knows what is being coordinated, and where option-space authorship sits upstream and

invisible, that fails — while the authority-evacuated form defeats the account outright, there being no human party to collaborate.

There is no responsibility gap to begin with. Two sceptical lines converge: that responsibility is a dynamic social practice, repeatedly extended to new kinds of entity and able to absorb artificial systems without leaving anyone unanswerable ([30]); and that the alleged gap, once disambiguated, either dissolves into the problem of many hands or rests on an unmotivated demand that someone be culpable whenever harm occurs, when compensation, answerability, and reform may be all the situation requires ([14]).

Both deserve concession, because both are substantially right about their targets. The present argument does not claim no one can be held responsible; it claims the traditional locus fails, which is compatible with practices adapting — 6.3 proposes an adaptation. Nor does it require culpability for every harm: structural responsibility is prospective and non-outcome-indexed, closer to the answerability the sceptics recommend than to the culpability they reject. Where they are too quick is the authority-evacuated form. A practice can extend responsibility along a relation only if there is a relation, and in the fraud-adjudication case there was no human decision position and no record from which one could be reconstructed — not a hard case for a flexible practice but an absence of the material such a practice works on. Reduction to many hands likewise understates the problem: the classical case has many agents each contributing something, whereas here a decision-constitutive power was held by none. A framework saying practices will adapt owes an account of what they adapt to, and the indicators supply one.

The framework is explanatorily idle. Automation bias, sociotechnical systems, and many hands can describe everything reported here, so the framework merely renames. This is the strongest objection, and the answer is the typology. The existing vocabulary cannot distinguish the oversight-shell form from the authority-evacuated form: both are “sociotechnical systems exhibiting automation-related failure with diffuse responsibility”. Yet they require opposite remedies — the first that the human position be given powers commensurate with its nominal authority, the second that such a position be created before determinations acquire legal force. A vocabulary unable to mark a distinction carrying opposite policy implications is not adequate to the phenomenon.

8. Conclusion

The question worth pursuing is not whether artificial systems possess agency, but whether agency has ceased to be something individuals possess at all. Section 4.3 specifies that property as supervenient, multiply realisable, and borne by a structure rather than a subject; the four indicators make it readable in particular systems. Coding ten cases yielded four morphologies, one of which — authority-evacuated adjudication — the framework had not anticipated and which the existing vocabulary cannot distinguish from ordinary oversight failure.

The theoretical yield is an interface. Free will, agency, and responsibility have been discussed largely in isolation from the systems conditioning their exercise; treating agency as a relational structure connects those debates to algorithmic architecture without requiring either side to abandon its vocabulary. The reconstruction of responsibility follows from the same move: 6.4 established that it is not developer liability renamed, and 6.5 identified where it stops.

The limitations stated in Section 2 bound these claims. What the case set establishes is that the indicators are assignable and the typology non-vacuous, not that any morphology is common. Multi-coder replication on a larger, randomly drawn incident sample would convert the typology from a demonstration into a finding; formal work on multi-agent systems could give the allocation of decision-relevant powers a precision qualitative coding cannot; and comparative institutional research could test whether regulatory environments produce systematically different morphologies — the question on which the framework’s practical value ultimately depends.

Primary Sources

AI Incident Database. Incident 4, Uber AV Killed Pedestrian in Arizona. https://incidentdatabase.ai/cite/4

AI Incident Database. Incident 37, Amazon’s Experimental Hiring Tool Allegedly Displayed Gender Bias in Candidate Rankings. https://incidentdatabase.ai/cite/37

AI Incident Database. Incident 57, Australian Automated Debt Assessment System Issued False Notices to Thousands. https://incidentdatabase.ai/cite/57

AI Incident Database. Incident 101, Dutch Families Wrongfully Accused of Tax Fraud Due to Discriminatory Algorithm. https://incidentdatabase.ai/cite/101
AI Incident Database. Incident 123, Epic Systems's Sepsis Prediction Algorithms Revealed to Have High Error Rates on Seriously Ill Patients. https://incidentdatabase.ai/cite/123
AI Incident Database. Incident 373, Michigan's Unemployment Benefits Algorithm MiDAS Issued False Fraud Claims to Thousands of People. https://incidentdatabase.ai/cite/373
European Union. Regulation (EU) 2022/2065 (Digital Services Act), Article 27 (Recommender system transparency).
National Transportation Safety Board. Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018. Highway Accident Report NTSB/HAR-19/03.

References

1. Clark, A., & Chalmers, D. (1998). The extended mind. Analysis, 58(1), 7–19. https://doi.org/10.1093/analys/58.1.7
2. Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. Science and Engineering Ethics, 26(4), 2051–2068. https://doi.org/10.1007/s11948-019-00146-8
3. Davidson, D. (1963). Actions, reasons, and causes. The Journal of Philosophy, 60(23), 685–700. https://doi.org/10.2307/2023177
4. Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. npj Digital Medicine, 3, 118. https://doi.org/10.1038/s41746-020-00324-0
5. Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580
6. Floridi, L. (2013). Distributed morality in an information society. Science and Engineering Ethics, 19(3), 727–743. https://doi.org/10.1007/s11948-012-9413-4
7. Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. Minds and Machines, 14(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d
8. Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. The Journal of Philosophy, 68(1), 5–20. https://doi.org/10.2307/2024717
9. Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), Article 50. https://doi.org/10.1145/3359152
10. Heersmink, R. (2015). Dimensions of integration in embedded and extended cognitive systems. Phenomenology and the Cognitive Sciences, 14(3), 577–598. https://doi.org/10.1007/s11097-014-9355-1
11. Heersmink, R. (2017). Distributed cognition and distributed morality: Agency, artifacts and systems. Science and Engineering Ethics, 23(2), 431–448. https://doi.org/10.1007/s11948-016-9802-1
12. Hutchins, E. (1995). How a cockpit remembers its speeds. Cognitive Science, 19(3), 265–288. https://doi.org/10.1207/s15516709cog1903_1
13. Hutchins, E. (2014). The cultural ecosystem of human cognition. Philosophical Psychology, 27(1), 34–49. https://doi.org/10.1080/09515089.2013.830548
14. Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? Ethics and Information Technology, 24(3), 36. https://doi.org/10.1007/s10676-022-09643-0
15. List, C. (2018). What is it like to be a group agent? Noûs, 52(2), 295–319. https://doi.org/10.1111/nous.12162
16. List, C. (2021). Group agency and artificial intelligence. Philosophy & Technology, 34(4), 1213–1242. https://doi.org/10.1007/s13347-021-00454-7
17. Rubel, A., Castro, C., & Pham, A. (2019). Agency laundering and information technologies. Ethical Theory and Moral Practice, 22(4), 1017–1041. https://doi.org/10.1007/s10677-019-10030-w
18. Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 6(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

19. McGregor, S. (2021). Preventing repeated real world AI failures by cataloging incidents: The AI Incident Database. Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 15458–15463. https://doi.org/10.1609/aaai.v35i17.17817
20. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2). https://doi.org/10.1177/2053951716679679
21. Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. Science and Engineering Ethics, 24(4), 1201–1219. https://doi.org/10.1007/s11948-017-9943-x
22. Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253. https://doi.org/10.1518/001872097778543886
23. Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. Ethics and Information Technology, 20(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8
24. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
25. Rupert, R. D. (2004). Challenges to the hypothesis of extended cognition. The Journal of Philosophy, 101(8), 389–428. https://doi.org/10.5840/jphil2004101826
26. Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. Philosophy & Technology, 34(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x
27. Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. Frontiers in Robotics and AI, 5, 15. https://doi.org/10.3389/frobt.2018.00015
28. Sayes, E. (2014). Actor–Network Theory and methodology: Just what does it mean to say that nonhumans have agency? Social Studies of Science, 44(1), 134–149. https://doi.org/10.1177/0306312713511867
29. Sparrow, R. (2007). Killer robots. Journal of Applied Philosophy, 24(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x
30. Tigard, D. W. (2021). There is no techno-responsibility gap. Philosophy & Technology, 34(3), 589–607. https://doi.org/10.1007/s13347-020-00414-7
31. Velleman, J. D. (1992). What happens when someone acts? Mind, 101(403), 461–481. https://doi.org/10.1093/mind/101.403.461
32. Verbeek, P.-P. (2006). Materializing morality: Design ethics and technological mediation. Science, Technology, & Human Values, 31(3), 361–380. https://doi.org/10.1177/0162243905285847