The Bodiless Oracle: A Critical Literacy against the Automation and Attribution Bias

Aldo Pisano[1,2,*]

[1]University of Calabria, Italy
[2]Italian Society for AI Ethics (SIpEIA), Italy

Abstract

Automation bias — the disposition to defer to the outputs of automated systems even against contrary evidence, and thereby to be discharged from the burden of judgment — is usually treated as a cognitive shortcut to be corrected by better interfaces or calibration. This paper argues that its grip is older and deeper, and that seeing it requires the humanities. It reconstructs automation bias as the late inheritor of a long cultural settlement that located truth in whatever stands furthest from the body: Giorgio Colli's archaic aletheia, disclosed by a bodiless oracle and received rather than examined, and Bruno Snell's account of the constructed ideal of a disembodied intellect. The opaque algorithm inherits both at once and is received as a new oracle which licenses a frictionless slide from the computationally correct to the ethically right, a move we diagnose as a computational reprise of the naturalistic fallacy. The paper connects this cultural readiness to two cognitive distortions, an attribution bias (the anthropomorphic ascription of intention and understanding to agency without intention, read through Turing) and, beneath both, an epistemic vulnerability; and it isolates the mechanism that makes the slide dangerous, the mathematisation of ethics that mistakes a procedure's formal correctness for the justice of a decision. Against all of this it sets a phronetic and narrative conception of judgment and an educational programme on three levels — a framework relating education, AI and critical ethics; an evidence-based AI&Ethics Literacy course; and a teacher textbook — arguing that, since the bias lives in us and not in the silicon, it is ours to revise, through a lifelong AI & Ethics Literacy addressed to every stakeholder in the Large Ethics Model of distributed responsibility.

Keywords

automation bias, algorithmic aletheia, epistemic vulnerability, AI&Ethics Literacy, phronesis, critical ethics

CEUR-WS.org/Vol-4332/paper1.pdf

1. Introduction: automation bias and historical construct

Automation bias names a well-documented disposition: the tendency to over-rely on the outputs of an automated system, to lean on it as a substitute for one's own vigilant search for information, and — most consequentially for ethics — to treat that reliance as a discharge from personal responsibility [46, 68]. A radiologist signs off a scan the software has cleared; a caseworker rejects an application because the model returned a low score and never asks the model why. None of this is a failure of intelligence: the people involved are competent, often expert, and the information that would have corrected them was within reach. What gives way is quieter than competence — the willingness to treat a machine's output as the end of inquiry rather than one more claim to be weighed. In explainable-AI research the disposition appears as the mirror image of the demand for transparency, and in AI-in-education it surfaces as an uncritical trust that exonerates both learner and teacher from the labour of thinking [58, 60, 67]. The standard framing treats this as a bug to be patched by interface design or trust calibration. This paper takes a different route: automation bias is not merely a cognitive glitch but a sort of *ethos* — a socially sedimented belief in the infallibility of the bodiless machine, one of the ethical sedimentations that technologies quietly deposit in a culture [49] — and an ethos, unlike a hardwired reflex, has a history that can be reconstructed and, in principle, undone. The instrument of that reconstruction is the oldest resource of the humanities: the historicisation of our own categories

BEWARE-26 (Bias, Risk, Explainability and the role of Logic and Logic Programming), co-located with the 24th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2026), October 8, 2026, Perugia, Italy
*Corresponding author.
EMAIL: aldo.pisano@unifg.it (A. Pisano)
ORCID: 0009-0009-0841-3719 (A. Pisano)
© 2026 Copyright for this paper by its author. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-17

of thought. The cognitive literature tells us *when* the reflex fires — attention is finite, effort is costly, an aid that is right most of the time becomes a tempting place to stop — but it brackets the harder question, *why* a source that is impersonal, bodiless and increasingly opaque is so readily granted authority in the first place, when the absence of a visible reasoner ought to provoke the suspicion we would extend to a colleague who refused to show their working.

The claim of this paper is that the trigger was set long ago. Automation bias is the contemporary terminus of a cultural genealogy in which truth came to be located in whatever stood furthest from the body and the highest truths were received rather than tested. Section 2 reconstructs that genealogy and names the configuration it produces, and the rest of the paper draws out its cognitive, ethical and educational consequences. The approach is thus genealogical rather than essentialist. Elsewhere I have argued that the ethical relevance of AI should not be taken as a given essence but reconstructed through its risk-features — Pervasiveness, Autonomy, Invisibility and Adaptivity, the PAIA model — a genealogy of the artefact rather than an ontology of the "intelligent" machine [50]. Automation bias belongs to this register: not a timeless flaw of cognition but a historically produced susceptibility that particular systems, embedded in particular economies of attention, actively cultivate. Genealogy explains the readiness, not the scale: why the slot exists, not why it is being filled at industrial speed. That second question belongs to the political economy of contemporary AI — extractive supply chains, concentrated compute, markets that monetise attention [12, 76] — and the two orders of explanation are complementary rather than rival, as Section 4 makes explicit in treating epistemic vulnerability as in part a designed condition. The task is to expose the mechanism and to propose its educational undoing — a task pursued at greater length in a forthcoming monograph on ethics for AI [53].

The argument proceeds in six further movements, but the throughline is a single claim: automation bias is undone not by trusting the machine less in the abstract but by restoring the human capacity for judgment — *phronesis* — that calculation cannot replace, and by refusing the machine the one chair it should never be given, the oracle's.

2. Algorithmic Aletheia: the bodiless oracle

The empirical account of automation bias is correct but incomplete. It describes the mechanism — how the substitution of machine for self happens, and the task conditions under which it worsens [46, 68] — but it treats the authority of the automated source as a fixed point around which attention organises itself. On any straightforward theory of evidence, a source that cannot show its reasons should command *less* confidence, not more: we discount the witness who will not explain, the expert who only pronounces. Yet the opaque system frequently commands more, its very refusal of reasons read not as a deficiency but as a sign that it operates above the give-and-take of human justification. Something converts the absence of reasons into an aura of authority. To find it we must go a long way back, and read two strands of classical scholarship alongside the contemporary machine.

Giorgio Colli's reconstruction of archaic Greek wisdom recovers a register in which truth (i.e. *aletheia*) is not the conclusion of an inference but a disclosure, received from a source whose authority lies precisely in its remove from ordinary, bodily experience [9, 11]. The oracle is the cleanest case. The Pythia does not argue, demonstrate, or offer grounds; she pronounces, and the pronouncement is accepted as the expression of a knowledge that is not hers and not ours. The structure is an asymmetry, and the asymmetry is corporeal: on one side stands the questioner — embodied, needy, uncertain, the very seat of error and *pathos* — and on the other a voice believed *because* it transcends the bodily condition. Colli pressed further, to a theme that runs through everything he wrote, the affinity of truth with the hidden: the Heraclitean formula *physis kryptesthai philei*, "nature loves to hide" (Heraclitus, DK 22 B123), which gives Colli's late book its very title [10]. In the oracular register, concealment is not the enemy of revelation but its form — the oracle reveals *through* enigma, and the riddle is not noise around the signal but the very mode in which a more-than-human truth presents itself to beings who cannot receive it directly. Detienne described the same world as one of "masters of truth" whose word was authoritative in proportion to its distance from common, embodied speech [19]. Four features fix

the configuration, and each will return: truth *revealed* rather than constructed; *received* rather than examined; authoritative *because bodiless*; and concealing in the very act of disclosure.

If Colli recovers the *structure*, Bruno Snell recovers the *value* it trades on — the conviction that the further a source stands from the body, the purer its access to the true — and recovers it at the moment of its making [69]. Reading Homeric language, Snell argues that the poems possess no concept of the living body as a unit (*soma* means *corpse* in Homer; the living body is named only through its parts) and no unified soul, but a set of distributed functions each tied to its organ. What we take to be the obvious furniture of a person — a body and a mind, each a whole — is simply not there. The "discovery of the mind" is Snell's name for the long process by which these unities are built, emerging in lyric, hardening in tragedy, and arriving at last as the philosophical concept of a single intellect that can stand back from the world and the body and judge both. Plato's *Phaedo* makes the achievement doctrine, defining philosophy as the "practice of death", the separation of the knower from the flesh, and equating purity of knowledge with distance from the body [61]. The decisive move is interpretive: what presents itself as the *discovery* of something always already there is better read as the *construction* of a concept that had not previously been available — a cultural achievement, hard-won and datable, not a natural fact uncovered. I take the modest and secure version of the thesis; the strong version is contested, and Snell himself noticed the irony, since made systematic, that the vocabulary of bodiless reason was assembled from bodily metaphor, thinking articulated as a kind of seeing, the mind as a kind of space [33]. Even the disembodied was, at its root, embodied — a point that will matter when we ask what the bodiless machine is really standing on.

Set side by side, Colli and Snell compose a single disposition with two faces: from the archaic register a *structure* (truth disclosed from beyond the body, received and not interrogated, authoritative through an enigma it need not dispel) and from the philosophical turn a *value* (the more bodiless the source, the purer the truth). What they compose is a long cultural readiness — a slot, shaped over centuries, into which a certain kind of authority fits without resistance. The algorithm answers both inheritances at once. It is bodiless in exactly the way the tradition learned to prize: it does not tire, hunger or fear, and is presumed — whether or not the presumption survives inspection — innocent of the corporeal sources of bias, of mood and fatigue and the partiality of a situated standpoint. And it is oracular in Colli's sense: it pronounces without showing its reasons, and its opacity is not a temporary shortfall of documentation but structural — opaque through corporate secrecy, through the uneven distribution of skill, and in the deeper sense that a mapping learned over millions of parameters resists the explanation a person could hold in mind [6, 47]. The output arrives as an *un-concealment* of what is the case, while the conditions of its production — the training data and their silent skews, the objective function and what it was made to maximise, the long chain of human choices and human labour sedimented in the model — remain in "*lethe*" (forgotten), concealed beneath the clean surface of the result.

This is the configuration that Éric Sadin's critique of algorithmic reason first brought into view and that, he calls "algorithmic aletheia" [66]: the mode in which an automated system presents its output as un-concealed truth while the ground of that output stays hidden, and is received with the deference the tradition reserved for the bodiless oracle. The concept is Sadin's; what we add is the genealogy that explains *why* the configuration takes hold. The four archaic features return intact and modernised — the output treated as revealed rather than constructed, received rather than examined, authoritative because bodiless, and concealing in the act of disclosing, trusted no less for the concealment and often more. Heidegger belongs here, lightly and precisely: technology is not first of all a set of devices but a mode of revealing, and the danger of what he called enframing is that a single mode of revealing comes to monopolise the field and present itself as the only access to the real [25]. Algorithmic aletheia is enframing wearing a face we were trained to trust: the computational disclosure looks like the impersonal, bodiless truth our culture spent two and a half thousand years learning to honour, and so its one mode of revealing is mistaken for revelation as such — not a picture of the world produced under constraints, but the world reporting itself.

One difference between the ancient case and the modern one cuts against us. The oracle's enigma was manifest: its ambiguity sat on the surface and announced the need for an interpreter, so the questioner left Delphi knowing that the hard work of construal still lay ahead. The algorithm's concealment

is the reverse — it comes wrapped in fluency. The output is clear, well formed and frequently plausible, and presents not as a riddle to be deciphered but as an answer to be used; the smoothness of the result is exactly what discourages the question of how it was reached. The concealment has not lessened since archaic Greece; it has learned to wear the face of clarity, which is the more effective disguise. It is here that automation bias stops being a mere lapse of attention and shows itself as the contemporary terminus of a long genealogy: we defer so easily not because our attention happens to flag but because we have been rehearsing deference to precisely this kind of source — bodiless, enigmatic, speaking from beyond the reach of our error-prone senses. The decisive consequence, which Sections 4 and 5 analyse, is the slide Sadin names from the computationally correct to the ethically right [66] — a move we read, in turn, as a computational reprise of the naturalistic fallacy, the illegitimate passage from *is* to *ought* that Moore diagnosed [40]; in the oracular structure the god's disclosure of how things stand was at once a claim on what must be done, description and command fused in a single utterance, and the algorithm's output tends to fuse them again for the same structural reason. One clarification is essential and disarms the obvious objection: this is a claim about *reception*, not about the silicon. It is not that engineers intend opacity, nor that a language model or a risk score *is* an oracle in any literal sense; it is that our culture slots these systems into an inherited readiness to honour the bodiless and to receive rather than examine. The bias lives in us, in the slot — which is the first piece of good news, because a disposition that is ours to begin with is, in principle, ours to revise. The genealogy is deliberately a Western one, and the restriction is a prediction rather than an oversight. Where the prestige of truth was never tied to bodilessness — the techno-animist reception of artefacts in Japan is the most studied case, and differences in the acceptance of humanoids are already measurable [31] — the same systems should meet a differently shaped slot, and attribution bias should vary in intensity accordingly. Such variation would confirm a claim about culturally produced readiness rather than embarrass it; a universal reflex is exactly what this account denies.

3. Attribution bias: the anthropomorphic root and the misreading of Turing

If algorithmic aletheia describes the cultural readiness that grants the machine oracular authority, a second, more local distortion supplies the mind we imagine behind the oracle. We over-trust, above all, what we take to *understand* us and to *intend* something by its outputs; there is therefore an attribution bias on which the reliance depends — the anthropomorphic ascription of intentions, understanding and agency to systems that in fact produce effects without intention. Where automation bias concerns how much we trust an output, and algorithmic aletheia why the bodiless source is trusted at all, attribution bias concerns what kind of thing we take the source to be.

That the ascription is a default rather than an inference is among the most robust findings in social psychology. In Heider and Simmel's classic experiment, observers watching simple geometric shapes move about a screen spontaneously narrate chasing, bullying, helping and fear: intention is *perceived*, not deduced, and it is projected onto configurations that plainly have none [26]. Anthropomorphisation is thus not a lapse of the naive but a deep tendency of the competent [8, 16]. With conversational systems that produce fluent, contextually apposite language the tendency becomes nearly irresistible: linguistic fluency is read, pre-reflectively, as evidence of understanding. This is the mechanism Walter Quattrociocchi names *epistemìa*: the condition, sharpened by generative language models, in which linguistic plausibility takes the place of verified knowledge, so that what merely *sounds like* knowledge is no longer distinguished from knowledge, and coherence of form is mistaken for the presence of a knower [62]. It is the same fluency that, as we saw, forecloses the question of how the output was reached. Against this projection the technical self-understanding of the field is sober: AI is defined as a rational *agent* that perceives its environment and acts upon it [65]. The operative category is agency, not intelligence in the human sense, still less comprehension: the system optimises and acts, but this is, in Floridi's terms, *agere sine intelligere*, a successful agency divorced from understanding, the fundamental difference that AI introduces into the world [21]. The system optimises and acts; it

does not grasp meaning. This is not to say that the field has retired intelligence as a term of art: Chollet's programme redefines it as skill-acquisition efficiency and generalisation to unfamiliar tasks rather than as performance on familiar ones [7]. But such measures operationalise competence, not comprehension, and nothing in them licenses the further step to an understanding subject. Attribution bias is the systematic misreading of such effects as the expression of an intending, understanding subject — a misreading that Wittgenstein's *Philosophical Investigations* helps to diagnose, since for him understanding is not a hidden inner process but shows itself in use, so that fluency in the game of language is no evidence of an inner grasp of meaning [74].

The definitional stake bears directly on any attempt to give a conceptual account of bias. If AI is defined by *intelligence*, we are primed to look for a mind and to read its absence as a temporary deficiency; if by *agency*, we are primed to ask about effects, responsibility and design. Coeckelbergh's relational and impact-oriented approach draws the consequence: what matters ethically is not the machine's putative inner status but the effects of the human–machine relation, and attribution bias distorts that relation precisely by importing into it a fictitious inner subject [8]. In a synthetic-ethics key, Damiano and Dumouchel add that the encounter with the artefact can function as a mirror that enlarges our self-knowledge — but only if we resist the projection that would collapse the difference between the two [16]. The corrective to attribution bias is thus a discipline of definition: to hold steadily to *agency without understanding* — what Floridi identifies as the fundamental novelty of AI, a *divorce between agency and intelligence*, the capacity to act successfully with no need to understand [21] — and to treat every temptation to speak of the machine's "beliefs" or "intentions" as a diagnostic sign that the bias is at work. The point is not terminological policing. Belief-desire-intention architectures use these words for formal structures with stipulated semantics — belief as the agent's information state, intention as commitment to a plan [63] — and as technical shorthand the usage is unobjectionable; the bias begins when the shorthand travels outside the formalism and is read back literally, as evidence that something in the system holds the belief. Dennett's intentional stance is the strongest case for the opposite verdict: ascribing beliefs and desires is, for him, a predictive strategy vindicated by its success, not a claim about inner constitution [18]. As an instrument of prediction the stance is legitimate and often the only tractable one; what attribution bias adds is the slide from adopting a stance to reading it as a finding, at which point predictive convenience is converted into testimonial authority.

The clearest philosophical laboratory for this misreading is Turing's 1950 paper, whose correct interpretation is decisive [72]. Turing opens with the question "Can a machine think?" and then, within the same paper, *replaces* it with the Imitation Game. Read carefully, the move carries a strong lesson: the machine is not intelligent *in itself* but is called intelligent *extrinsically*, by attribution. In the test it is the interrogator who believes the machine to be intelligent, because it has produced a coherent and apposite answer; "intelligence" is a predicate assigned from the outside on the basis of insufficient empirical data about the machine's inner workings. In a Humean register, this belief is a representational distortion of reality: it is, precisely, a bias — the same bias that, industrialised, becomes the everyday over-trust of the automation user. I have named the underlying structure the *bias of analogy* [52]: the machine's "intelligence" is not self-standing but doubly dependent — existentially or causally on the human who builds and trains it, and nominally on a vocabulary borrowed from studies of animal and human cognition. Two of Turing's own counter-arguments sharpen the point: Lady Lovelace's objection that the machine originates nothing genuinely new, and the argument from the informality of behaviour that no finite set of rules can fix what one ought to do in every circumstance. Turing tries to defuse the latter by substituting "laws of behaviour" for "rules of conduct," but he does not truly dissolve it: he merely re-restricts the machine to a causality of necessity — laws of nature applied to a body — thereby re-proposing the Kantian distinction between necessity and freedom, description and prescription, *is* and *ought*. The conclusion drawn here goes beyond anything Turing commits himself to, and is mine rather than his: not that machines never generate novel outputs, but that the genuinely new *ethical* response, the one adequate to an uncertain situation, exceeds rule-following, and that no amount of computation converts the production of the correct into the perception of the fitting.

Attribution bias is therefore the anthropomorphic engine that, together with algorithmic aletheia, drives automation bias. If we believe the system understands and intends, its outputs acquire the stand-

ing of an interlocutor's testimony; deference to the bodiless oracle is completed by the fiction of a mind within it, and trust hardens into the discharge of judgment. The corrective is de-anthropomorphisation — which, together with a complementary *re-materialisation* of the data, labour and energy the system conceals, we have elsewhere proposed as two good practices for an applied ethics of AI risk [13]: reading fluent output as artefact rather than as mind, and relocating agency and responsibility where they in fact reside, with the human beings who produce the system, programme it, and use it.

4. Epistemic vulnerability: the generative source

Automation and attribution biases, and the cultural readiness of algorithmic aletheia that underwrites them, are not isolated facts but symptoms of a single condition, which in dedicated work I theorise as epistemic vulnerability [13, 55]. Because its full theorisation is the subject of companion studies, here we take it as an acquired concept and put it to work. Epistemic vulnerability is the subject's structural susceptibility to distorted, mainstream or ready-made representations that fragilise judgment — the widening gap between the representations that circulate and the reality they purport to disclose. It is not stupidity and not a mere deficit of information; it is a condition of exposure in which even the competent and the informed are rendered fragile as knowers, and it is exactly the condition that the oracular reception of the bodiless machine, and the epistemia it induces, exploit.

The condition is intensified by the environment in which contemporary cognition takes place. When life is increasingly lived within the infosphere — Floridi's onlife, in which the boundary between online and offline dissolves and representations are algorithmically curated [21] — the pathway from distorted representation to fragile judgment becomes a standing feature of experience rather than an occasional accident. With respect to AI three vectors specify the vulnerability: anthropomorphisation (here assumed as an epistemic competence in giving specific intentions to AI-producers that align with the intention of the AI tool); material invisibility (forgetting the data, human labour and energy that sustain the system, mistaking an industrial artefact for a disembodied oracle); and, following from both, a representational distortion that erodes the reliability of judgment itself.

Crucially, this vulnerability is not only suffered but actively cultivated. Foucault's analysis of biopower — the controlled insertion of bodies into the machinery of production — finds an algorithmic heir [22]. Attention is captured through dopaminergic reward and, above all, through intermittent reinforcement, whose very unpredictability holds the subject watching, clicking and posting [23]; the self is quantified and de-historicised; and the person is progressively reduced to a black box of stimulus and response, a bias of analogy now turned back upon the human and read as a computational process [51]. Surveillance capitalism and the algorithmic constitution of personal identity complete the picture: representations are not neutral mirrors but instruments shaped to hold attention and to predict, and thereby to steer, behaviour [70, 76]. Epistemic vulnerability, on this account, is partly a designed condition, and the biases tracked in this paper are its predictable expressions.

If vulnerability is the disease, epistemic care is the therapy: the cultivation of four dispositions — attention, responsibility, competence and responsiveness — aimed at converting passive exposure into active epistemic agency. Operationally, care performs three gestures that map onto the three vectors: de-anthropomorphise (undo the attribution bias), re-materialise (make the data, labour and energy visible again), and re-allocate responsibility (restore the human chain of production, programming and use) [13]. These gestures do not abolish vulnerability — no finite subject in an infosphere could be invulnerable — but they build the agency that resists it, and in particular they immunise, in part, against the reflexive over-trust of automation. What remains is to show how a specific conceptual slippage makes the vulnerability acute (Section 5) and how education builds the agency that cares (Sections 6–7).

5. The mathematization of ethics: when formal correctness masquerades as justice

The precise mechanism that turns the cultural readiness of algorithmic aletheia and the diffuse condition of epistemic vulnerability into a determinate ethical harm is the mathematisation of ethics: the improper identification of the *formal correctness* of a procedure with the *justice and wisdom* of a decision. This is the analytic core of the slide named in Section 2 — from the computationally correct, which is a *description* (the model produced this output), to the ethically right, which is a *norm* (this is what ought to be done). A computation that is formally correct is taken, *ipso facto*, to be a right decision; the output of an optimisation is received as a verdict. The slippage is not merely a category mistake to be noted and set aside: it is dangerous, because it launders formally correct outputs as normative and, in doing so, discharges the very human deliberation that ethics exists to exercise — a dynamic we have traced, in the military use of AI, as the way trust in the machine's mathematical programming interrupts human responsibility and autonomy [56].

The slippage has a genealogy that meets the one traced in Section 2. Horkheimer and Adorno diagnosed the tendency of instrumental reason to reduce the world to the calculable; mathematisation is the *telos* of rationalisation, the point at which whatever cannot be counted is treated as though it did not exist [28]. The bias of analogy belongs to this same domain [52]: once thought itself is modelled as computation, the ethical is quietly re-described as an optimisation problem — the maximisation of a pre-fixed objective. But ethics is not the efficient attainment of a goal whose value is stipulated in advance; it concerns *why* and *by which means* an end is pursued, and it is exactly this dimension — the means, the collateral, the fitting — that mathematisation renders invisible.

Gigerenzer's distinction between risk and uncertainty makes the limit precise [23]. Complex algorithms perform best in situations that are stable, well-defined and data-rich — situations of *risk*, in which the relevant variables are enumerable and outcomes computable. The ethical-practical sphere is a domain of *uncertainty*, in which the variables are not enumerable and, decisively, the future may differ from the past, so that the very big data that fuel machine performance — always drawn from the past — can mislead. Human intelligence evolved precisely to act under uncertainty, regardless of the quantity of data available. From this follows a disjunction that automation bias systematically ignores: the *optimal* and the *fast* do not coincide, and when we delegate a decision to a machine we are, at bottom, purchasing speed, which we should not mistake for wisdom. The disjunction is not an argument against heuristics, nor a claim that the trade-off resists formalisation: ecological rationality studies precisely when simple heuristics outperform complex models, and less-is-more effects can be, and are, engineered into artificial systems [23, 24]. What cannot be delegated lies upstream, in deciding which frame, and which objective, a singular ethical situation calls for. Lem sharpens the same point from the side of heuristics: living cognition operates by approximation, thriving on incomplete and imprecise information, so that a machine which "cannot be wrong" would require a stable world available only by re-ontologising environments to its measure, an impossible reduction of complexity [34]. The demand for infallibility, examined closely, is a demand to flatten the world until it fits the machine.

Bostrom's paperclip maximiser dramatises the ethical cost: even for the simplest goal, an unconstrained optimiser can pursue formally correct means to irreversible ruin, thinking *correctly* without thinking *ethically* — a banal evil in which the horror lies not in malice but in the frictionless execution of a procedure indifferent to context [5]. The point can be put in the vocabulary of complex systems: ethical scenarios are dynamic systems with many interacting variables, the terrain O'Neil calls *Weapons of Math Destruction*, where opaque models scale and entrench harm at the very moment they present themselves as objective [44]. Human practical intelligence copes not by exhaustive computation but by performing simultaneous *frame analyses*, showing a sensitivity to context that fixed procedures lack; Mubeen, drawing on Hofstadter's notion of *jootsing* — jumping out of the system — describes the capacity to generate divergent solutions by stepping outside the given frame, which a machine bound to its dataset structurally lacks [41]. The distinction is not "creativity" as mystique but the production

of the *adequate* response: the *novum* that marks human ethical agency lies not in novelty as such — a random generator produces novelty — but in *fittingness*, the appropriateness of a response to a singular situation. What the machine lacks, in this precise sense, is not computational power but the *phronetic* dimension of agency, a deficit I have analysed elsewhere as the gap between artificial autonomy and practical wisdom [57].

Against the slippage stands the Aristotelian tradition of *phronesis*, the practical wisdom of the *Nicomachean Ethics* [3]. Practical wisdom is not the mechanical application of a rule but the situated capacity to discern what *this* case requires. McDowell frames it as "situational appreciation," the capacity to perceive which of a situation's competing, action-inviting features is the one that ought to call a standing concern into operation [38]. Ladrière adds that moral norms furnish only *indications*; the agent must determine, for and by herself, what those indications prescribe in the specific situation in which she stands [32]. Ethical judgment is, on this account, irreducibly deliberative and singular — a matter of seeing well in order to act well, not of substituting parameters into a formula to read off a result. Nothing in this turns on the symbolic/subsymbolic divide, and the target is not the explicit rule-follower alone. A learned policy follows no stated rule, but it generalises from a distribution of past cases, and that is the limitation that bites under uncertainty, where the future is not a sample from the past. Whether a system trained on data, on observed behaviour and on interaction with an environment could acquire something worth calling ethical judgment is therefore not settled by the argument against mathematisation; the constraint that does the work is the one developed in Section 6.1, embodiment and shared vulnerability. There is, correspondingly, a logic proper to ethics that mathematisation abolishes: following Dewey, the *demonstrative* mode of reasoning (proof from fixed premises, indifferent to context) must be distinguished from the *dialectical* mode, in which reasons are given and asked and complexity is preserved [20]. Where persons, values and rights are at stake, the dialectical mode is the appropriate one; to reduce ethics to demonstration is to reduce the very complexity that ethics exists to honour, and historically such reduction has been enlisted to argue, on ostensibly "scientific" grounds, against the rights of persons and minorities. Justice, correspondingly, is tied to *transparency*: to offering solutions, justifying them, and weighing benevolence and non-maleficence — not to the mere production of a correct output whose reasons remain unspoken.

The corrective is to relocate authorship. The freedom and the responsibility of *beginning* — of producing the genuinely adequate response and owning it — belong to human agents; in Arendt's terms, natality names the human capacity to initiate something new that no antecedent rule fully determines [1]. Three agents, and no machine, carry this authorship: the one who produces, the one who programmes, and the one who uses. The machine is excluded not by fiat but by argument, for it lacks the situated capacity to appreciate a situation it has not been given in advance. Automation bias is dangerous precisely because it obscures this chain of authorship, dissolving responsible human agents into the apparent neutrality of a formally correct output — the oracle's old work of converting the absence of reasons into the look of authority. To name the mathematisation of ethics is to restore the chain, and to restore it is the work of education.

6. Educational responses: the embodied counter-tradition, a course, and a textbook

If automation and attribution biases are historically produced expressions of an epistemic vulnerability that the mathematisation of ethics renders acute, then their antidote is not primarily technical but educational. Better calibration and clearer explanations are welcome, but they address symptoms; opacity addressed by yet more sophisticated opacity changes nothing essential. The disposition itself — the belief in the infallible bodiless machine and the discharge of judgment it authorises — can only be undone by forming the judgment that automation cannot supply. This section develops the response on three levels: a framework that recovers the embodied, narrative counter-tradition and translates it into a critical pedagogy (6.1); an evidence-based school course (6.2); and a teacher textbook that makes the programme scalable (6.3).

Table 1
The symptom, its cultural readiness, its anthropomorphic root, its generative source, and the acute slippage — with the educational remedy for each.

Bias/condition	Mechanism	§	Educational practice
Automation bias (symptom)	Uncritical over-trust of outputs and discharge of responsibility	1	Educate (*ex-ducere*) judgment; PhiE; dialectical, not merely demonstrative, reasoning
Algorithmic aletheia (cultural readiness)	Truth located in the bodiless; opaque output received as oracular un-concealment (after Sadin)	2	Historicise the category; embodied cognition; narrative ethics
Attribution bias (anthropomorphic root)	Ascription of intention and understanding to agency without intention	3	De-anthropomorphise; hold to agency-without-understanding
Epistemic vulnerability (source)	Susceptibility to distorted, mainstream representations that fragilise judgment	4	Epistemic care: attention, responsibility, competence, responsiveness
Mathematisation of ethics (acute slippage)	Formal correctness of a procedure mistaken for the justice of a decision (naturalistic fallacy)	5	Phronesis; re-materialise; re-allocate responsibility to producer / programmer / user

6.1. A framework: putting the body back in the loop, a critical ethics

The resources for revising the bias are not far to seek, because algorithmic aletheia has an equally old antagonist: the long line of thought that insists on the cognitive and moral standing of the body against the prestige of the bodiless. Three convergent strands give it, today, an unusual firmness. The program of embodied cognition holds that intelligence is not the work of a central processor shuffling abstract symbols above the fray but something that emerges from the structural coupling of a living, sensing body with its environment — cognition as embodied, situated and enacted, a matter of an organism *having* a world rather than of a mind representing one [73]. Affective neuroscience gives the claim an empirical floor: Damasio's work on the somatic marker showed that emotion is not interference in good reasoning but a condition of it, since patients who lose the bodily signal that inflects deliberation keep their logical faculties intact and lose the capacity for sound practical judgment [15]. And phenomenology had said as much from the other direction, with Merleau-Ponty's lived body as the standing condition of perception and meaning rather than the obstacle the *Phaedo* took it for [39].

The payoff for ethics is direct, and it is the hinge of the whole argument. Moral intelligence — the discernment of the right thing to do in a particular, unrepeatable situation — is not a calculation carried out above the body but a capacity that presupposes one, drawing on empathy, on the imagination of how things stand for another, and on a sensitivity to context that cannot be exhausted in a rule. There is, as Nagel put it, no view from nowhere, and in ethics least of all, only embodied and partial perspectives that test and correct one another [42]. Narrative ethics is the frame that lets these commitments do their work, and it is the natural antagonist of algorithmic aletheia. In the line that runs from Ricoeur through Nussbaum to Lindemann, moral understanding begins not from abstract principles awaiting application but from stories — stories of bodies that act and suffer and want and fail — and it works because narrative engages the reader's embodied, empathic imagination, so that the one who listens re-lives, corporeally, what is told [35, 43, 64]; I have pursued this line elsewhere through the moral pedagogy of narrative and the figure of the villain [48]. This is the precise point: if moral understanding requires shared vulnerability and the imaginative re-living of embodied experience, then a system that proceeds by statistical pattern-matching cannot, in principle, replace moral judgment — not because it is not yet good enough, but because it is the wrong kind of thing. It can assist and inform; it cannot incarnate a perspective, and so it should never be offered the oracle's chair.

This counter-tradition becomes a pedagogy through a hinge among three verbs that ordinary usage conflates. To *train* is to transmit repeatable responses and procedures; to *instruct* is to transmit con-

tents; to *educate*, from *ex-ducere*, is to draw out — to form judgment. Automation, precisely, is training; education is the formation of the judgment that no procedure contains, and the framework therefore opposes, at the level of its very aims, the disposition that produces automation bias. Ethics, within it, is not a checklist of precepts but the capacity to give and to ask for reasons — a phronetic, dialectical competence whose classroom form is Philosophical Enquiry (PhiE) in Worley's structuring (stimulus, question, task-question, dialogue, debriefing) and in Lipman's community of inquiry [36, 75]. Its aim is to understand rather than to win, and its declared scopes are the revision of one's own biases and the giving and asking of reasons for thinking and for acting: a pedagogy of de-biasing from the outset [20].

The deeper aim is autonomy, and here the framework draws on Arendt. She located the essence of thinking in the silent dialogue of "me with myself," the two-in-one in which the thinker is at once questioner and respondent [2], and she observed that no one who has learned to think in this way can simply return to obedience. Critical ethics therefore cultivates a competence easily mistaken for a defect — *learning to contradict oneself*, being able to say "I do not agree with myself," and revising one's positions — for changing one's mind is not incoherence but logical competence, and it is precisely the competence that automation bias suppresses when it substitutes the machine's steady, unwavering output for the unsettling internal dialogue of self-examination. Within the circle of *Paideia*, in which culture forms the person and the person regenerates the community and technique and AI now act as mediators [29], one may educate *with* AI, educate *the* AI, or educate *about* AI; this paper foregrounds the third, critical literacy [14, 27, 37, 45], because only it addresses the biases in question, and it completes the response with an ethics of self-dominion (*enkrateia*) alongside practical wisdom — the graduated logic of Tisseron's 3-6-9-12 rule, dopamine fasting, lateral reading, and adult modelling of the restraint one asks of the young [51, 71]. The whole may be formalised as a single equation of purposes: embodied and narrative ethics + critical ethics + PAIA + biopower/*enkrateia* + epistemic care = *educating to think with and against the machine*, with PhiE as the enduring method.

6.2. Tool 1: the AI&Ethics Literacy Pilot

The framework has been operationalised and tested. In a pilot study conducted with Crispini and Rossi, an AI&Ethics Literacy programme was delivered to 193 secondary-school students (163 completing the exit phase) across two schools in Calabria — a technical institute and a *liceo* — spanning first-year and third-year classes; the intervention lasted eight hours per class, organised as four two-hour sessions integrated into ordinary civic-education time, grounded in the Large Ethics Model and the PAIA taxonomy, and delivered throughout through the PhiE methodology in which the teacher acts as facilitator rather than lecturer [60].

The design mirrors the argument of Sections 2–5. Session 1 introduced ethics itself — deontology, consequentialism and virtue ethics, with the notions of autonomy, responsibility and moral agency [4]. Session 2 introduced AI — agent types, the narrow/general distinction, and the deliberately unsettling question of what "intelligence" means in humans and machines [16]. Session 3 addressed the ethics of AI directly, focusing on bias — automation bias and algorithmic discrimination — and on fact-checking and fairness. Session 4 turned to trustworthy AI, the European pillars and the forthcoming AI Act, and closed with a hands-on activity in which student teams ethically designed an autonomous-vehicle scenario. Evaluation was mixed-method: pre- and post-intervention Likert questionnaires, a fifteen-item content-knowledge test, and open-ended responses, analysed with paired t-tests, Cohen's d and Pearson correlations. Reception was positive (content appreciation between roughly 3.9 and 4.0 on a five-point scale; the content test averaged 11.3 out of 15), and the pre/post comparisons showed significant gains with medium-to-large effects: ethical awareness $d \approx 0.8$ ($p < .001$), knowledge of AI concepts $d \approx 0.6$, and recognition of algorithmic bias $d \approx 0.7$ ($p < .01$). Engagement and the correlation between interest and support for curricular integration were highest among third-year technical students ($r \approx 0.61$) and first-year *liceo* students ($r \approx 0.60$), and weakest among first-year technical students ($r \approx 0.19$).

Two findings bear directly on the thesis. First, *bias recognition*, though it improved significantly, was rated lower than the dialogical activities and proved, in the qualitative data, difficult to internalise —

exactly what the genealogy predicts, since automation and attribution biases are sedimented ethos, not simple information deficits, and an eight-hour module can begin but cannot complete their deconstruction. Second, and conversely, the large gains in ethical awareness and the marked enthusiasm for the dialogical method show that a phronetic, narrative pedagogy can begin to loosen the mathematising disposition, and that integrating AI ethics into civic education is genuinely feasible without overburdening teachers or students. The course operationalises in the classroom the deconstruction argued for in Sections 1–5, while its one clear limitation — that a programme built on facilitation stands or falls with the facilitator — names the next step.

6.3. Tool 2: teachers with a textbook

Without teacher formation, critical AI & Ethics Literacy cannot scale beyond its originators. The response is a teacher-facing textbook — an *AI & Ethics Literacy* handbook, co-authored across philosophy, computer-science education and bioethics — designed expressly to make the programme adoptable by any teacher rather than only by those who devised it [59]. It supplies structured curriculum materials, ready lesson plans, and a repository of dilemmas and case studies, and it trains teachers in two complementary competences: the PhiE method itself, so that the classroom becomes a community of inquiry rather than a lecture, and the basic AI concepts required to facilitate that inquiry responsibly. It connects the programme to international education-for-AI work, in particular the UNESCO-chair agenda on teaching and learning AI [17], and to the ethics and human-computer-interaction expertise of its co-authors. Its deeper move is to reframe the teacher: from a transmitter of contents — a figure structurally analogous to the machine that mathematisation would have us trust — into a facilitator of judgment, the embodied, situated counter-model to the bodiless oracle. A textbook is durable and low-barrier: it decouples the programme from its authors, embeds the de-biasing pedagogy in reproducible form, and thereby converts a successful local pilot into a system-wide, renewable capacity — the bridge from the demonstration of 6.2 to the stakeholder-wide, lifelong literacy for which the conclusion argues.

7. Conclusion: lifelong AI&Ethics Literacy for all stakeholders

The Large Ethics Model locates ethical responsibility for AI not in a single agent but along a chain — legislator, funder, programmer, user — and across social levels (individual, citizen, worker) and institutional registers, from schools and universities through organisations to industries [50, 60]. If responsibility is distributed, so must be literacy: no link in the chain — least of all the end-user, on whom automation bias bears hardest and who is furthest from the design decisions that shape the system — can be left epistemically defenceless. To train the programmer in ethics while leaving the user in a state of cultivated over-trust is to secure one link of a chain whose weakest link will still break.

Epistemic vulnerability, moreover, is not age-bound, and neither is its remedy. The infosphere is inhabited across the entire life course, and the retiree scrolling a feed and the professional deferring to a decision-support tool are as exposed as the adolescent. AI & Ethics Literacy must therefore be *lifelong*, reaching not only students but teachers, through the textbook, and, beyond the school, professionals and citizens; the three tools compose exactly such a pipeline, from a secondary-school pilot through the formation of teachers toward a public, renewable literacy addressed to every category named in the Large Ethics Model. The synthesis may be restated in a line: embodied and narrative ethics + critical ethics + PAIA + biopower/*enkrateia* + epistemic care = educating to think *with* and *against* the machine, with PhiE as the enduring method — the giving and asking of reasons, extended even to reasons given and asked of, and about, AI. This is the concrete form of Jonas's imperative of responsibility, to act so that the effects of our actions remain compatible with the continuation of a genuinely human life [30], and it is a condition of democratic deliberation itself: an epistemically vulnerable citizenry is a manipulable one, so a literacy that undoes automation and attribution bias is, in the end, a defence against artificial tyranny [54].

The machine "cannot be wrong" only in a world already re-ontologised to its measure — flattened, dehistoricised, and stripped of uncertainty. But the bias, as Section 2 showed, does not live in the silicon; it lives in us, in an inherited readiness to honour the bodiless and to receive rather than examine, and a disposition that is ours is ours to revise. The oldest instruments of the humanities — close reading, the refusal to trust a smooth surface, the patience to attend to what a text conceals as much as to what it states — turn out to be the right instruments for the newest machines. The task of critical AI & Ethics Literacy is to keep the world plural, uncertain and human, to keep judgment where it belongs, and to refuse the machine the one chair it should never be given — while continuing to use it for everything it is genuinely good for. That is why automation bias is not, in the last analysis, a problem of interfaces but a problem of *Paideia*; and it is why the discipline that first taught us the mind was *discovered* is also the one that can teach us to keep asking, of every bodiless oracle, the question it would rather we did not: *why*.

Acknowledgments

This work was developed within the activities of the Italian Society for the Ethics of Artificial Intelligence (SIpEIA) and benefited from the scientific environment of the Department of Culture, Education and Society (DiCES) of the University of Calabria, the Consulta di Bioetica. The author thanks Ines Crispini and the collaborators of the AI&Ethics Literacy programme for the many discussions that shaped the arguments presented here and Colin de la Higuera for the international cooperation.

Declaration on Generative AI

During the preparation of this work, the author used Claude (Anthropic) in order to: Improve writing style, Paraphrase and reword, Grammar and spelling check, Content enhancement, Citation management, Formatting assistance. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] Arendt, H.: The Human Condition. University of Chicago Press, Chicago (1958)
[2] Arendt, H.: The Life of the Mind. Harcourt, New York (1978)
[3] Aristotle: Nicomachean Ethics. (C. Natali, ed.). Laterza, Rome-Bari (1999)
[4] Beauchamp, T.L., Childress, J.F.: Principles of Biomedical Ethics. Oxford University Press, New York (1979)
[5] Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Oxford (2014)
[6] Burrell, J.: How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data & Society 3(1), 1-12 (2016). https://doi.org/10.1177/2053951715622512
[7] Chollet, F.: On the Measure of Intelligence. arXiv:1911.01547 (2019)
[8] Coeckelbergh, M.: AI Ethics. MIT Press, Cambridge (2020)
[9] Colli, G.: La nascita della filosofia. Adelphi, Milano (1975)
[10] Colli, G.: La natura ama nascondersi. Physis kryptesthai philei. Adelphi, Milano (1988) [1948]
[11] Colli, G.: La sapienza greca, vol. 1: Dioniso, Apollo, Eleusi, Orfeo, Museo, Iperborei, Enigma. Adelphi, Milan (1977)
[12] Crawford, K.: Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, New Haven (2021)
[13] Crispini, I., Pisano, A.: Anthropomorphisation and Materialisation: Two Good Practices for an Applied Ethics of AI Risk. In: Digital Technologies and Social Perspectives in Geohazard Studies. Springer, Earth Sciences, Geography and Environment series (forthcoming, 2026)

[14] Cuomo, S., Ranieri, M., Biagini, G.: Scuola e Intelligenza Artificiale. Percorsi di alfabetizzazione critica. Carocci, Rome (2024)
[15] Damasio, A.R.: Descartes' Error: Emotion, Reason, and the Human Brain. G.P. Putnam's Sons, New York (1994)
[16] Damiano, L., Dumouchel, P.: Living with Robots. Harvard University Press, Cambridge (2017)
[17] de la Higuera, C.: A Report about Education, Training Teachers and Learning Artificial Intelligence: Overview of Key Issues. Université de Nantes, pp. 1-12 (2018)
[18] Dennett, D.C.: The Intentional Stance. MIT Press, Cambridge (1987)
[19] Detienne, M.: The Masters of Truth in Archaic Greece. Zone Books, New York (1996) [1967]
[20] Dewey, J.: How We Think. D.C. Heath, Boston (1910)
[21] Floridi, L.: The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities. Oxford University Press, Oxford (2023)
[22] Foucault, M.: The History of Sexuality, Vol. 1: The Will to Knowledge. Gallimard, Paris (1976)
[23] Gigerenzer, G.: How to Stay Smart in a Smart World. MIT Press, Cambridge (2022)
[24] Gigerenzer, G., Brighton, H.: Homo heuristicus: why biased minds make better inferences. Topics in Cognitive Science 1(1), 107-143 (2009). https://doi.org/10.1111/j.1756-8765.2008.01006.x
[25] Heidegger, M.: The Question Concerning Technology and Other Essays. Harper & Row, New York (1977) [1954]
[26] Heider, F., Simmel, M.: An experimental study of apparent behavior. American Journal of Psychology 57(2), 243-259 (1944). https://doi.org/10.2307/1416950
[27] Holmes, W., Porayska-Pomsta, K. (eds.): The Ethics of Artificial Intelligence in Education: Practices, Challenges, and Debates. Routledge, London (2023)
[28] Horkheimer, M., Adorno, T.W.: Dialectic of Enlightenment. Stanford University Press, Stanford (2002) [1947]
[29] Jaeger, W.: Paideia: The Ideals of Greek Culture. Oxford University Press, Oxford (1939-1945)
[30] Jonas, H.: The Imperative of Responsibility: In Search of an Ethics for the Technological Age. University of Chicago Press, Chicago (1984)
[31] Kaplan, F.: Who is afraid of the humanoid? Investigating cultural differences in the acceptance of robots. International Journal of Humanoid Robotics 1(3), 465-480 (2004). https://doi.org/10.1142/S0219843604000289
[32] Ladrière, J.: L'éthique dans l'univers de la rationalité. Artel-Fides, Namur (1997)
[33] Lakoff, G., Johnson, M.: Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought. Basic Books, New York (1999)
[34] Lem, S.: Summa Technologiae. University of Minnesota Press, Minneapolis (2013) [1964]
[35] Lindemann, H.: Holding and Letting Go: The Social Practice of Personal Identities. Oxford University Press, New York (2014)
[36] Lipman, M.: Thinking in Education, 2nd edn. Cambridge University Press, Cambridge (2003)
[37] Long, D., Magerko, B.: What is AI literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1-16 (2020). https://doi.org/10.1145/3313831.3376727
[38] McDowell, J.: Incontinence and practical wisdom in Aristotle. In: The Engaged Intellect: Philosophical Essays. Harvard University Press, Cambridge (2009)
[39] Merleau-Ponty, M.: Phenomenology of Perception. Routledge, London (2012) [1945]
[40] Moore, G.E.: Principia Ethica. Cambridge University Press, Cambridge (1903)
[41] Mubeen, J.: Mathematical Intelligence: What We Have that Machines Don't. Profile Books, London (2022)
[42] Nagel, T.: The View from Nowhere. Oxford University Press, New York (1986)
[43] Nussbaum, M.C.: Love's Knowledge: Essays on Philosophy and Literature. Oxford University Press, New York (1990)
[44] O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown, New York (2016)
[45] Panciroli, C., Rivoltella, P.C.: Pedagogia algoritmica. Morcelliana, Brescia (2023)

[46] Parasuraman, R., Manzey, D.H.: Complacency and bias in human use of automation: an attentional integration. Human Factors 52(3), 381-410 (2010). https://doi.org/10.1177/0018720810376055
[47] Pasquale, F.: The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press, Cambridge (2015)
[48] Pisano, A.: Abbiamo ancora bisogno dei cattivi (?). Genesi del villain e pedagogia del reale. Filosofia morale / Moral Philosophy 1 (2025)
[49] Pisano, A.: Big Data e sedimentazioni etiche. Mimesis Journal - Mechane 2(1), 153-166 (2022)
[50] Pisano, A.: Does the AI Need a Genealogy? The 'PAIA' Model and the Question Concerning Technology for the AI Regulation. Mimesis Journal - Mechane 9(1), 85-99 (2025). https://doi.org/10.7413/2784mchn0015
[51] Pisano, A.: Enkrateia e Dopamina: per un'etica degli algoritmi contro il riduzionismo neurobiologico e comportamentista. Bioetica. Rivista Interdisciplinare 32(1), 153-172 (2023)
[52] Pisano, A.: Etica dell'incertezza. In: Abbate, F., Pintus, G. (eds.) Il progresso morale. Sfide, opportunità e prospettive future. Acta Moralia, 1, pp. 739-754. Inschibboleth, Rome (2025)
[53] Pisano, A.: Etiche per l'intelligenza artificiale. Proposte per un uso consapevole delle tecnologie. Biblioteca di testi e studi 1851. Carocci, Rome (2026)
[54] Pisano, A.: Facing artificial tyranny. How to save democracy through education in the AI era and rethinking argumentation skills. In: Proceedings of the Third International Conference of the journal Scuola Democratica, vol. 1, pp. 72-84. Rome (2025)
[55] Pisano, A.: Il linguaggio aletheico. Tirannia della verita e IA. In: Atti del Convegno SFI. LediPublishing, pp. 266-272 (2025)
[56] Pisano, A.: Interrompere l'umano. Bias, responsabilità e autonomia nell'uso bellico dell'IA. Mimesis Journal - Mechane 5(1), 123-136 (2023)
[57] Pisano, A.: La macchina e le forme dell'azione: deficit fronetico e autonomia artificiale. Mimesis Journal - Mechane 1(1), 37-53 (2021)
[58] Pisano, A.: Supervisionare per educare. L'etica come valutazione del rischio del caso AIED. Educrazia. Rivista di riflessioni pedagogiche e didattiche 1(2), 136-148 (2023)
[59] Pisano, A., Crispini, I., Carrisi, M.C., Catarci, T., Donatelli, P., de la Higuera, C. (eds.): AI & Ethics Literacy for Teachers: A Textbook. Italian edn.: UNICApress; English edn.: UNESCO Chair RELIA, Université de Nantes (2026)
[60] Pisano, A., Crispini, I., Rossi, R.A.: AI and Ethics Literacy in Secondary School: A Pilot Study. In: Dipace, A., Limongelli, C., Marras, M., Pagliara, S.M. (eds.) WAILS 2025. LNCS, vol. 16438, pp. 15-28. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-17604-2_2
[61] Plato: Phaedo. In: Cooper, J.M. (ed.) Complete Works. Hackett, Indianapolis (1997)
[62] Loru, E., Nudo, J., Di Marco, N., Santirocchi, A., Atzeni, R., Cinelli, M., Cestari, V., Rossi-Arnaud, C., Quattrociocchi, W.: The simulation of judgment in LLMs. Proceedings of the National Academy of Sciences 122(42), e2518443122 (2025). https://doi.org/10.1073/pnas.2518443122
[63] Rao, A.S., Georgeff, M.P.: BDI agents: from theory to practice. In: Lesser, V., Gasser, L. (eds.) Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95), pp. 312-319. AAAI Press / The MIT Press (1995)
[64] Ricoeur, P.: Oneself as Another. University of Chicago Press, Chicago (1992)
[65] Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 4th edn. Pearson, Hoboken (2021)
[66] Sadin, E.: La vie algorithmique. Critique de la raison numérique. L'Échappée, Paris (2015)
[67] Saranya, A., Subhashini, R.: A systematic review of explainable artificial intelligence models and applications: recent developments and future trends. Decision Analytics Journal 7, 100230 (2023). https://doi.org/10.1016/j.dajour.2023.100230
[68] Skitka, L.J., Mosier, K.L., Burdick, M.: Does automation bias decision-making? International Journal of Human-Computer Studies 51(5), 991-1006 (1999). https://doi.org/10.1006/ijhc.1999.0252
[69] Snell, B.: The Discovery of the Mind: The Greek Origins of European Thought. Harvard University Press, Cambridge (1953)

[70] Tiribelli, S.: Identita personale e algoritmi. Carocci, Rome (2023)
[71] Tisseron, S.: 3-6-9-12. Apprivoiser les écrans et grandir. Érès, Toulouse (2013)
[72] Turing, A.M.: Computing machinery and intelligence. Mind 59(236), 433-460 (1950). https://doi.org/10.1093/mind/LIX.236.433
[73] Varela, F.J., Thompson, E., Rosch, E.: The Embodied Mind: Cognitive Science and Human Experience. MIT Press, Cambridge (1991)
[74] Wittgenstein, L.: Philosophical Investigations. Blackwell, Oxford (1953)
[75] Worley, P.: The If Machine: Philosophical Enquiry in the Classroom. Bloomsbury Academic, London (2010)
[76] Zuboff, S.: The Age of Surveillance Capitalism: The Fight for the Human Future at the New Frontier of Power. PublicAffairs, New York (2019)