AIMED: Towards a Philosophically Legitimated AI-assisted Iterative Method for Ethical Deliberation

Luca Rivelli*

2025

Abstract

This paper addresses the accelerating crisis of ethical governance in an age of complex socio-technical change, particularly in the domain of Artificial Intelligence. It poses a foundational philosophical question: when, if ever, is AI assistance in ethical deliberation legitimate? An answer is developed through three theses: i) the Ethical No-Free-Lunch (ENFL) principle, which establishes the indispensability of human normative intervention and accountability; ii) the Discovery/Justification Separation inspired by Reichenbach's work, which restricts AI use to the exploratory "context of discovery"; iii) the Algorithmic Mediated Control Framework (AMCF), which mandates that only scrutable, human-vetted deterministic algorithms generated with AI assistance, and not the AI itself, be entrusted with critical societal processes. From these theses, five legitimacy criteria for AI-assisted ethical deliberation are derived. Finally, the paper proposes the "AI-assisted Iterative Method for Ethical Deliberation" (AIMED), an actionable multi-stage workflow that fulfills the exposed criteria for ethical AI-assisted deliberation. This method integrates digital literature analysis, structured human–AI dialogue, human-only verification, and continuous feedback. The paper explicitly addresses several potential objections. It is shown how the AIMED framework aligns with and provides a concrete implementation for major international regulatory guidelines, such as the EU AI Act and the NIST AI Risk Management Framework. By situating the AIMED within traditions of proceduralism, the governance of inductive risk, and human–AI collaboration, the paper argues that this framework offers a philosophically justified, practically implementable model of AI-assisted ethical governance, that can be seen as an actionable instance of Digital Humanism.

*Cultore della Materia, FISPPA Department, University of Padua, Padova, Italy. (luca.rivelli@gmail.com).

1 Introduction: The Crisis of Deliberation in an Age of Acceleration

There is a concern that is central to the current societal condition: the growing crisis of governance in the face of accelerating and increasingly chaotic socio-technical change. Humanity is now confronted with a gallery of scenarios involving complex, interconnected, and potentially catastrophic existential risks: from the escalating climate crisis to the quick and continuous development of progressively advanced forms of Artificial Intelligence. The sheer speed and scale of these challenges risk to consistently outstrip the capacity of our traditional deliberative institutions.

The challenges posed by this new reality are challenges of interpretation and dialectical social negotiation, not mere technical problems to be managed through engineering approaches alone. In this context, a failure of traditional institutions to timely deliberate, decide, and act with both agility and wisdom constitutes itself an existential risk.

This crisis is most acute in the domain of AI, where the pace of innovation is exponential, but it is surfacing in a progression of evidence in other contexts characterized by complex dynamics, global economic crises, geopolitical tensions, destabilization of natural environment on a planetary scale.

The vast scale and the rapid and unpredictable pace of development of these crises quite likely means that ethical governance nowadays can't be reduced to a fixed set of rules. Instead, it needs to be an adaptable, agile learning process able to quickly produce, in a continuous manner, flexible new standards and rules.

Faced with the scale of this crisis of governance, we could wonder if the current availability of advanced AI could help itself sort out the situation. Actually, based on the current ubiquity of forms of AI and the results of their use in daily life, it is reasonable to suppose that Large Language Models (LLMs), with their virtues, such as rapid and large-scale ideative capacity based on the vastness of the AI's implicit knowledge, and memory span in general larger than humans', could help governing bodies produce more agile and adaptable solutions.

This clearly invites a foundational philosophical question: *when, if ever, is AI assistance in ethical deliberation legitimate?* What is offered here as an answer is a philosophically-justified guideline for AI-assisted legitimate ethical deliberation, presented in the the form of a set of conditions under which AI-assisted ethical deliberation is permissible.

The paper will proceed from first principles, arguing that if one accepts these foundational theses, then five legitimacy criteria follow, and that any acceptable AI-assisted procedure with ethical import should satisfy them. A scheme for a practical implementation of an AI-assisted workflow for ethical deliberation and decision-making respecting these criteria is then presented. We will show that

the proposed framework, that is called *AIMED* (*AI-assisted Iterative Method for Ethical Deliberation*) is compatible with many current regulatory guidelines for AI released by the most important political national and supranational bodies, and that it can be considered a scheme, based on a philosophical justification, for the practical implementation of such regulations.

2 Three Foundational Theses for AI-Assisted Deliberation

The account presented here rests on three interlocking philosophical theses that define the proper roles of humans and machines in a normative ethical process.

2.1 Thesis 1. ENFL: The unavoidable Human Normative Source

In Rivelli (2025b) an argument is put forth to the effect that an AI self-building as an ethical subject is logically and computationally impossible. This establishes the necessity of human normative intervention at distinct stages of the machine training and operations. The argument, called the *Ethical No-Free-Lunch principle* (*ENFL*), shows that in a decision stemming from an AI-assisted workflow the AI output *alone* cannot provide justificatory force for a normative claim, and that all normativity and accountability must be authored and owned by the human component.

The *Ethical No-Free-Lunch principle* rests on a three-pronged philosophical argument:

1. *The Is-Ought Gap*: drawing on Hume[1], it is argued no machine can self-learn as an ethical subject from training on data alone, even if trained on corpora of *normative texts*: barred human normative input under the form of a specific injected choice, the machine would be unable to make a choice among the various possible metaethical frameworks it has encountered in its training dataset. Thus, normative premises must be injected into the machine or in the training data by a human deliberative act: the metaethical decision can come only from the human component.
2. *The No-Free-Lunch (NFL) Theorems*: even if a machine *could* learn a normative stance purely from data, the NFL theorems of machine learning[2] prove that effective learning requires the imposition of a-priori *inductive biases*, a form of computational "pre-understanding" determining which kinds of patterns in the data the machine will prioritize and how it will generalize. For machines learning ethics, these biases are not neutral technical settings, but value-laden choices to be made by accountable human agents. Thus, even in this case a human ethical choice is unavoidable.

[1] Hume (1739).
[2] Wolpert and Macready (1997).

3. *The Ought-Is Gap*: based on Translational Ethics[3], there is an unbridgeable gap for the machine in translating an abstract norm into a concrete, context-sensitive action. The translation requires situational judgment that necessarily falls to human stakeholder subjects participating in the concrete situation.

These three prongs act synergistically: even in the hypothetical scenario in which one or two of them fails, the remaining ones still work to support the conclusion.

The ENFL principle shows that human value-loading is a necessary feature of any ethically-involved AI system.

2.2 Thesis 2. The Discovery/Justification Separation

Along the lines of Reichenbach's famous distinction[4] in philosophy of science between the *context of discovery* and the *context of justification*, I propose to deploy the same distinction in AI ethics. I argue, based on the ENFL (*thesis 1*) that it is legitimate to use LLMs *only* in the *context of discovery* of possible solutions, and illegitimate to treat their outputs as reasons in the *context of justification*, that is, in the context of ethically legitimating these produced solutions, because, according to the ENFL, no AI output can carry any independent justificatory weight. In other words: exploiting LLMs for ethically-laden work is legitimate, as long as their use is *channeled* by human operators in the context of an exploratory phase, without ever, even implicitly, attributing spurious ethical authority to the AI. To enforce this recommendation, some thorough vetting and verification of the AI-produced output has to be argued for and ensured, as better explained in the coming sections.

2.3 Thesis 3. AMCF: The Mediated Control Framework

Beyond the normative argument of *thesis 1*, and more in general when employing AIs for providing solutions to critical societal problems, there is a strong precautionary case for human control. Rivelli (2025a) proposes an *Algorithmic Mediated Control Framework (AMCF)* as a response to the problem of the inherent unreliability of LLMs when they are applied to scenarios of control of critical processes of societal import. Conceived for safety-critical systems to avoid a catastrophic misaligned AI singularity The AMCF, in tune with the discovery/justification dichotomy of *thesis 2* requires that practical adoption of any proposal for a solution that is generated by an AI-assisted process be immediately subjected to a chain of *human-only* verification and vetting, before being subsequently submitted to a phase of human-only deployment decisions. This way, the framework offers an AI governance model operating within humanistic constraints.

[3] Kagarise and Sheldon (2000), Cribb (2010), Bærøe (2014), Sisk et al. (2020).
[4] Reichenbach ([1938] 1938).

2.3.1 The probabilistic nature of LLMs

A preliminary explication is needed before proceeding to describe the structure of the AMCF. LLMs act as probabilistic machines, as their core mechanism is to predict the next word in a sequence by sampling from a probability distribution over possible next tokens. Their behavior emerges from statistical inference, yielding outputs that are creative but also prone to inconsistencies and "hallucinations", while classic algorithms are deterministic, in that they operate on fixed, step-by-step rules, making them verifiable and more predictable. This makes the LLMs inherently unreliable, in the sense that there is a non-zero chance of unexpected outputs, especially with novel inputs. Even with identical prompts, empirical studies report request-to-request variations[5], and slight changes in input can lead to disproportionately different responses.

2.3.2 The Structure of the AMCF

The AMCF proposes a division of labor between human and machine inside a cyclic process:

1. *Probabilistic* advanced AI language models (like the LLMs) are typically prone to errors, hallucinations, losing sight of the current goal, inconsistencies. Nevertheless, they *excel* as programmers. In the AMCF this ability is exploited by operating the AI, under human guidance, in an initial creative phase, to help produce procedures and guidelines in the form of deterministic, verifiable algorithms. It is these *deterministic* algorithms, *not* the AI itself, that are to be put in *control* of critical technological and societal processes. It is crucial to note that saying "deterministic" procedure and algorithms we include *deterministic decision policies* that can *consume probabilistic signals*. The difference here with the probabilistic nature of the AI is that in these algorithms, there's a well defined flow chart on how to decide based on probabilistic signals: with explicit predefined thresholds, escalation, and logging. While inside the AI the intricate, non-linear, and often unpredictable input-output mappings emerge from the self-organization of complex neural networks during training[6], and are not based on an explicit, verifiable decision flow.
2. The *AI-generated* algorithmic proposal is submitted to several steps of *human* verification, evaluation, and vetting.
3. Human executive bodies produce decisions on the deployment of the resulting deterministic algorithm in high-stakes societal domains.
4. The real-world effects are observed, and are fed back to the first step to improve the deterministic algorithm.

This cyclic governance model aims to prevent a runaway misaligned singularity, while still letting AI improve crucial aspects of societal life in a gradual and controlled manner through the AI-devised deterministic algorithms. This model

[5] Ouyang et al. (2025), Astekin, Hort, and Moonen (2024).
[6] Teehan et al. (2022).

promotes human agents as the sole ethical subjects retaining accountability and ultimate oversight in every critical phase of the process. The key proposal of the AMCF is to use AI solely for creative design of advanced deterministic algorithms. It is these algorithms, and not the AI itself, that, after thorough vetting and verification, are put in control of potentially critical processes. It is this entrusting deterministic algorithms for control, the crucial step intended to prevent the risk of a runaway AI singularity.

3 Five Criteria for AI-Assisted Ethical Deliberation

Based on the above foundational theses, I propose five guiding criteria that any legitimate AI-assisted ethical deliberation should satisfy.

C1. Human Normative Primacy: The decisive normative premises and the final justificatory step are authored and explicitly endorsed by identifiable human agents.

C2. The Assisted-Discovery Restriction: AI assistance is restricted to discovery functions (option generation, drafting) and contributes no independent justificatory weight.

C3. Scrutability of Proposals: AI-assisted outputs, before justification, must be cast as *Deterministic Governance Proposals (DGPs)*. A DGP is a clear, stepwise and verifiable procedure, in other words an algorithm, that human agents can rationally assess and critique.

C4. Accountable Verification: There must be a formal, *human-only verification gate* where the decision to accept, reject, or amend a DGP is made through human-only debate and deliberation, with every reason and any dissent occurring in the debate recorded and attributed to the responsible participants.

C5. Feedback and Revision Obligation: The final real-world application of a DGP entails a binding obligation to monitor its real-world outcomes and to revise or revoke it if those outcomes deviate from the expectations that justified its adoption.

4 The Rationale for AI Assistance: A Response to the Crisis of Speed

The preceding theses establish why human oversight is necessary in any deliberation process involving AI, but not why AI assistance is *desirable*. The rationale for integrating AI into deliberation is a direct response to the crisis of governance outlined in the introduction. The current unpredictable, accelerating technological development requires that ethical governance not be a static imposition of rules, but an adaptive, learning process able to *swiftly* output flexible normative decisions.

Now, it can be argued that AI and algorithms integrated *as tools* in human-centered deliberative tasks will improve *in general* swiftness, adaptability and efficacy of the produced decision, thanks to the enhanced capabilities of the AIs compared to humans: quick and large-scale ideative capacity based on the vastness of the AI's implicit knowledge and large memory span. But, as we know, LLMs also manifest abundant flaws, basically amounting to unpredictability and unreliability of the produced output.

I argue that, provided that the five criteria C1-C5 for a legitimate AI-assisted deliberation are satisfied, we can *safely* employ an AI-assisted deliberation workflows, even for decisions fraught with critical ethical implications.

5 The AIMED Workflow for AI-Assisted Ethical Deliberation and Decision-Making

The five criteria for the legitimacy of AI-assisted ethical deliberation listed in section 3 are not merely an abstract ideal. I propose here a lean, actionable scheme for the practical implementation of AI-assisted ethical deliberation and decision workflows that are ethically sound: the *AI-assisted Iterative Method for Ethical Deliberation (AIMED)*. The scheme offered here is very general and can accommodate much more detailed and complex specific implementations, provided that the criteria listed in section 3 are not violated.

The core of the AIMED is a repeatable, *four-stage* socio-technical cycle, described in the following sections.

5.1 Stage 1: Problem Framing via Digital Literature Analysis (DLA)

This stage is purely within the *context of discovery* (criterium C2 of section 3).

The process begins by ensuring the deliberation is grounded in a robust understanding of the problem space. Instead of relying on anecdotal evidence or the biases of pre-existing conceptual accounts of the problem to treat, AIMED initiates with a systematic *Digital Literature Analysis (DLA)* of a targeted corpus of preexisting literature about the problem to address (e.g., policy documents, technical standards, academic articles). This is not a simple literature review but a computational method to more objectively map the discursive landscape. The analysis is to be based not directly on LLMs, but on well-known algorithms for DLA, ranging from algorithms for natural language processing such as topic modeling, to community and motif detection on citation networks[7].

AI can be used at this stage as an *interpretive assistant* to help interpret the output of the algorithmic DLA. For example, topic modeling can identify some

[7] Lean, Rivelli, and Pence (2023).

hidden thematic structures in a body of text by providing '*topics*' in the form of sets of co-occurring keywords. AI-assisted semantic analysis can then help human researchers provide meaningful labels for these sets. Citation network analysis can reveal distinct intellectual communities of co-citing authors, and the AI can help identify the common core theoretical tenets of each community. The role of the DLA phase is to produce a concise *problem brief* that highlights ambiguities, value conflicts, or implementation gaps in the current discourse (e.g., the conflicting definitions of "robustness" in legal vs. technical contexts).

This stage respects the *Assisted-Discovery Restriction (C2)* by expanding the informational basis of data for deliberation without assigning any justificatory weight to the AI contribution.

5.2 Stage 2: AI-Assisted Formulation of Actionable Guidelines (Human-LLM Dialogue)

This stage is at the core of the *context of discovery* (criteria C2, C3 of section 3). It proceeds as follows.

The problem brief from Stage 1 serves as the input for a structured, interactive dialogue between a small group of human experts and a Large Language Model. This stage is designed to embody the core principles of our framework:

First, the human experts perform the crucial act of injecting the normative premise, directly satisfying *Human Normative Primacy (C1)*. For instance, an expert might state: *"For this problem, the guiding normative principle is a robust interpretation of the precautionary principle."*

The LLM is then tasked not with a vague discussion, but with a concrete translation exercise of the injected normative principle into possible solutions to the given problem respecting that principle. Guided by human experts and the normative premises they provide, the AI's role is to explore, on the basis of the problem brief, the solution space (C2) and, to cast potential solutions into the required format of Deterministic Governance Proposals (DGPs), thus satisfying the *Scrutability of Proposals (C3)*.

In this phase, the AI uses its ability to:

- Synthesize vast amounts of textual data analysis from the DLA phase, informing the explorative phase.
- Operationalize the injected principle by proposing several solutions.
- Surface Trade-offs by identifying how these draft rules might conflict with other principles based on patterns in the literature.
- Finally, cast the draft solutions produced under human supervision into distinct, precise, solutions in the form of rule-based DGPs.

The human experts always guide, challenge, and direct the dialogue. Their role is to ensure the generated solution space is broad and relevant, while remaining critically aware of the LLM's inherent biases and propensity to error

and hallucinations, providing human correction where needed and steering the AI to let it remain in the relevant space of the discussion.

5.3 Stage 3: Human-Only Deliberation and Verification

This is the first stage falling within the context of justification (criteria C1, C4 of section 3).

The set of candidate DGPs generated in Stage 2 is brought to an interdisciplinary, human-only committee, composed by different subjects than the supervisors of Stage 2. This is the formal, *human-only verification gate (C4)*. The committee's role is not to engage in an open-ended debate, but to perform a rigorous justification function, analogous to the process of the formal review of code in computer programming.

Here, without the direct influence of the LLM, the human experts deliberate on the concrete proposals. Their tasks are to:

1. Critically assess whether the proposed DGPs are faithful implementations of the initial normative principle.
2. Debate the relative merits of the competing DGPs.
3. Formally select and endorse the most justified proposal, potentially amending it in the process.
4. If no solution is accepted, requesting to repeat Stage 2 (or Stage 1+2) in order to get a new set of proposed solutions, providing the human experts operating there with some sensible feedback.
5. Record the whole discussion session.
6. The committee must *argue* for their final decision through a freestanding normative argument, clearly providing in a specific record the underlying reasons for the decision.

This stage ensures that the final recommendation is not just a plausible idea generated by a machine, but a philosophically defended, collectively endorsed, and practically viable guideline that is a product of *accountable human reason*, thus satisfying *Human Normative Primacy (C1)* and *Accountable Verification (C4)*.

5.4 Stage 4: Feedback Loop and Continuous Improvement

This feedback stage falls within the context of justification (criteria C1, C4, C5 of section 3).

The guideline produced by the human committee is the output of one full cycle, but the process does not end there. The adoption of the DGP entails a binding *Feedback and Revision Obligation (C5)*. The methodology includes a crucial feedback mechanism where, after the guideline is applied (even in a simulated or pilot context), data on its effects and stakeholder feedback are collected. This is also the stage where assurance techniques, such as formal methods for

verifying key properties or monitoring and rollback plans for non-deterministic components, are implemented to ensure the system behaves as intended.

This real-world data (or simulated real-world data) is then used to inform the next iteration of the cycle. This data can serve as a new input for the DLA in Stage 1 or, if the initial problem landscape is already well understood, it can be introduced directly into the Human-LLM dialogue in Stage 2.

This Stage 4 transforms the framework from a static model for making a single decision into an adaptive, learning system for ongoing ethical governance.

6 Possible Objections and Replies

Any proposal to integrate a powerful but opaque AI technology like an LLM into a sensitive human process like ethical deliberation will rightly face serious objections. In this section, I address four of the most significant challenges to our account, arguing that our five legitimacy criteria (C1-C5) and the AIMED workflow are specifically designed to mitigate these risks.

6.1 Objection 1: The Problem of Automation Bias and 'Authority Laundering'

Objection: The AIMED framework is philosophically tidy, but psychologically naive. Decades of research on automation bias show that humans tend to over-trust and uncritically accept the outputs of automated systems. Won't the human committee in Stage 3 simply rubber-stamp the 'solution' proposed by the AI in Stage 2? The human-only verification gate will become a mere formality, and the AI's output will be laundered into a human decision, smuggling in its spurious authority.

Reply: This is the most serious practical objection, and our framework is designed with this risk at its core. Three specific features of our account serve as a direct defense against automation bias:

- *The Presentation of Multiple, Competing Options:* importantly, Stage 2 does not deliver a single "AI recommendation", but it is tasked with generating *several distinct, competing DGPs*, each designed to operationalize a *different* value trade-off. This immediately breaks the dynamic of a simple accept/reject decision: the human committee is *forced* into a comparative, critical mode of thinking, asking which of the proposed solution is best, and why, rather than asking if the AI's answer is correct. Moreover, the committee can reject all proposed solutions, provide feedback on the reasons for rejection and request a new set of proposals from Stage 2, or even Stage 1+2.
- *The Requirement of Scrutability (C3):* the DGP format is a powerful cognitive debiasing tool. The committee is not evaluating a free-form, narrative suggestion from an "oracle". They are evaluating a *clear, stepwise,*

algorithmic procedure. This format invites critique, testing of edge cases, and demands a different, more analytical kind of reasoning than simply agreeing with a persuasive output coming out of Stage 2. It is easier to find a flaw in an algorithm than to argue with a seemingly wise pronouncement.

- *The Burden of Justification (C4):* the accountable verification criterion requires the human committee not just to *choose*, but to *author and record the reasons for their choice.* They cannot simply say, "We chose DGP-A because the model suggested it was the most equitable." They must construct a freestanding normative argument for why DGP-A is the most equitable and why that value ought to be prioritized in this context. This formal requirement to author a justification forces the shift of reason and accountability back onto the human agents.

6.2 Objection 2: The Epistemic Status of LLM-produced Reasons

Objection: The claim that LLM outputs carry no justificatory weight seems too strong. Modern LLMs can produce sophisticated, well-structured arguments, complete with citations and logical steps. If an LLM produces a text that is, by all appearances, a good reason, why is it a category mistake to treat it as one?

Reply: This objection rests on a confusion between the *form* of a reason and the *act* of bringing forth reasons. The ENFL principle is not a claim about the syntactic or semantic *properties of the text* an LLM produces. Rather, it is a claim about the *source of normative authority.*

An LLM is a stochastic model that generates text by predicting the most probable next token. It can produce a sequence of words that we recognize as a valid argument, but it does so without any understanding, intentionality, or commitment to the truth of its premises. A human agent, in contrast, when offering a reason, is performing a specific normative act: they are *endorsing a claim and taking responsibility for its truth and its relevance.* Therefore, when an LLM outputs a text that looks like a reason, that text is best understood as a *candidate consideration* or a *scaffold for human production of reasons.* Its epistemic status is that of a *found object.* Its justificatory force, that is, its status *as a reason* in the deliberation, is only conferred upon it when a human agent critically assesses it, understands its implications, adopts it as their own, and becomes accountable for it. The LLM can *display* an argument, while only a human can *assert* it.

6.3 Objection 3: The Challenge of Public Reason and Inclusiveness

Objection: The AIMED model, with its emphasis on expert committees and AI, appears technocratic and elitist. Where are the citizens? How does this framework align with the democratic ideals of public reason, which require that

justifications be accessible and acceptable to all members of the public, not just a small group of experts?

Reply: This is a crucial point. Our framework, as presented, is *procedurally thin* and is intended to be a *modular component* that is fully compatible with, and indeed strengthened by, broader participatory processes. The five legitimacy criteria are the "inner kernel" of legitimate AI assistance, but this kernel can and *should* be embedded within larger democratic institutions.

AIMED is not a *replacement* for public deliberation, but a tool to make that deliberation *more effective*. Citizens' panels or stakeholder workshops could be integrated directly into Stage 3. The DGPs drafted in Stage 2 could be the specific, structured proposals that are put to a citizens' jury for evaluation, possibly under the guidance of experts if needed. The DGPs themselves could be made public for an *open comment* period before the final verification gate. The requirement for explicit, recorded reasons (C4) directly serves the ideal of public reason by creating a transparent record that can be communicated and debated in the broader public sphere.

The proposed criteria do not solve the problem of democratic participation, but they ensure that *however* that participation is structured, the specific role of AI within it is constrained in a way that preserves human authorship and accountability.

6.4 Objection 4: The Theory-Dependence of the Framework

Objection: The entire AIMED framework rests on the ENFL principle. What if one rejects this principle? There are, for instance, forms of ethical naturalism or certain machine-centric ethical theories that might argue that a sufficiently advanced AI *could*, in fact, *discover* objective moral truths.

Reply: The ENFL principle is indeed the normative foundation of the proposed account, but it is intentionally modest in its commitments. It is compatible with a very wide range of meta-ethical views (constructivism, contractualism, virtue ethics, deontology). It only makes one core claim: that the act of *taking responsibility for a normative claim* is a uniquely human one.

To reject this minimal principle, a theorist must accept a very heavy philosophical burden. They would need to argue for one of two radical positions: (1) that the concept of accountability is no longer central to our ethical practices, or (2) that we can and should attribute *genuine moral agency and accountability* to current or near-future AI systems. While these are interesting philosophical positions to debate in the long term, they are not a viable basis for practical governance in the here and now. Our framework is designed for the world we actually live in, a world where humans are, and, I argue, *should* remain, the authors of their normative commitments. Any other post-human scenario could certainly be coming, but such a different world would fall outside the scope of the AIMED

proposal, which is an implementation of *Digital Humanism*[8]. For any theorist who accepts this minimal condition of human accountability, the ENFL and the legitimacy criteria that follow from it should hold.

7 Discussion

7.1 Situating the Framework in the Scholarly Landscape

The account of legitimate AI-assisted deliberation presented here offers a specific contribution to at least three major streams of contemporary thought: proceduralism in political philosophy, the governance of inductive risk, and the emerging field of human-AI collaboration.

7.1.1 Proceduralism and Public Reason

The AIMED framework shares the core ambition of proceduralist theories of justice, grounding the legitimacy of a norm not in its substantive content alone, but in the fairness and accountability of the procedure by which it was generated. It specifically draws on the traditions of Jürgen Habermas's discourse ethics and John Rawls's public reason. The framework's insistence on a human-only verification gate (Stage 3) operationalizes Habermas's ideal of a protected space for rational discourse, where legitimacy is forged through uncoerced human dialogue. Similarly, the requirement for an explicitly recorded, freestanding normative argument (C4) is a direct instantiation of Rawls's concept of public reason, ensuring the final justification is transparent and based on reasons all can accept.

However, while classical proceduralism is concerned with the ideal conditions for human-to-human deliberation, the AIMED framework addresses a novel and urgent problem: how to maintain the integrity of that deliberation when one of the participants is a powerful, non-human, and non necessarily completely rational cognitive tool, namely, the LLM or a similarly probabilistic AI. The five proposed legitimacy criteria (C1-C5), grouped under the distinct phases of discovery and justification, can therefore be understood as an AI-specific amendment to proceduralist theory. They are the necessary guardrails for ensuring that the integration of AI *enhances the discovery* of options, rather than *corrupting the justificatory process* of public reason.

7.1.2 The Governance of Inductive Risk

The philosophical literature on *inductive risk*[9] argues that non-epistemic ethical and social values are a necessary and legitimate part of scientific and policy-related decision-making under uncertainty. When the evidence is not conclusive, the choice of which hypothesis to accept or which policy to enact

[8] See VA (2025).
[9] See Douglas (2000).

inevitably involves a value judgment about which *kinds of errors* we are more willing to risk. For example, we could ask if it is worse to wrongly approve an unsafe drug or to wrongly reject a safe one.

The AIMED framework provides a concrete, procedural answer to the challenge of managing inductive risk. The entire proposed workflow is a device for making these value-laden choices, made under uncertainty, choices that are transparent, deliberate, and accountable. Specifically, Stage 2 is designed to systematically make the inductive risks become evident by generating multiple Deterministic Governance Proposals (DGPs), each of which embodies a different value trade-off. After this, Stage 3, based on human-only verification, is precisely the moment where the value judgments are made. The requirement to record the reasons for choosing one DGP over another (C4) is a formal mechanism for documenting and taking responsibility for the management of inductive risk.

7.1.3 Human-AI Collaboration and Deliberation

An emerging field within Human-Computer Interaction (HCI) and AI research has begun to study the empirical dynamics of human-AI teams in deliberative or creative tasks[10]. This research often focuses on the empirical questions of whether and how AI assistance can improve the speed, creativity, or quality of group decisions.

This paper is a *normative complement* to this empirical work. While empirical studies can tell us what *is* effective, the philosophical explication proposed here is concerned with what *is legitimate*. We are not primarily making an empirical claim that AIMED is faster or more creative (though we hypothesize that it is). We are making the philosophical claim that any AI-assisted deliberative process, to be legitimate, must satisfy our five criteria.

Our framework can therefore serve a dual purpose: it provides a normative "gold standard" against which empirical researchers can evaluate their systems, and it offers a philosophically grounded and structurally sound model for them to test and build upon. What is put forth here is not a contribution to the empirical study, but to the philosophical foundations of AI-human collaboration.

7.2 The AIMED Framework and the Current International Regulatory Landscape

The AIMED framework's core principles align with the current global regulatory and standards landscape for AI, showing its potential practical relevance beyond the realm of philosophical justification. The concepts of *meaningful human oversight*, *documented reasoning*, and *ongoing monitoring* are foundational to our approach and are also at the heart of major policy initiatives. Here are some of the more prominent ones.

[10] See for example Memmert and Bittner (2022).

- The *EU AI Act*[11] mandates strict *human oversight* and requires deployers of high-risk AI systems to conduct *Fundamental Rights Impact Assessments (FRIAs)* before deployment.
  - A FRIA is a systematic process to proactively identify, assess, and mitigate the potential negative impacts of a high-risk AI system on human rights such as non-discrimination, privacy, and due process. The AIMED framework provides a structured workflow to operationalize such an assessment. Specifically, the problem-framing of Stage 1 can map the landscape of relevant rights and the risk that endanger them, while the human-LLM dialogue in Stage 2 can generate concrete mitigation strategies in the form of scrutable DGPs. The human-only verification gate of Stage 3 then serves as the formal deliberative forum for assessing these impacts and justifying the chosen mitigation, with Stage 4 ensuring the required ongoing monitoring and revision. Our criteria C1-C5 are already fully consistent with these requirements, as they ensure that human agents retain accountability and that the process is auditable. While the stages of the AIMED are very general and do not make specific prescriptions, a more specific implementation for high-risk contexts formalizing the FRIA as a core artifact of the human-only deliberation gate could be easily derived from the AIMED. The EU AI Act's focus on the entire value chain is also relevant, as AIMED could be a tool for organizations to document their own accountability and demonstrate due diligence.
- The *NIST AI Risk Management Framework (AI RMF 1.0)*[12] and its *Generative AI Profile* propose a continuous governance cycle of "MAP, MEASURE, MANAGE, and GOVERN". This directly parallels our AIMED loop of discovery, human verification, and continuous feedback. The NIST framework's emphasis on documented human oversight and risk treatment echoes our requirement for a clear, human-only verification gate and recorded decision-making.
- The *ISO/IEC 42001 standard*[13] for AI management systems also stresses governance, accountability, and iterative improvement. Our AIMED workflow can be viewed as a concrete, procedurally-defined method for organizations to implement and satisfy the controls and requirements of such a management system.
- In the public sector, mandates like the *US OMB M-24-10*[14] also emphasize risk triage, impact assessments, transparency, and ongoing review. The AIMED framework provides a structured process for meeting these

[11] EU (2025).
[12] Tabassi (2023).
[13] ISO (2023).
[14] USgov (2024).

mandates, particularly with its emphasis on creating explicit, scrutable proposals (DGPs) and recording the rationale for their adoption (C4).

The procedural thinness of our framework, which was noted in the reply to Objection 3, is precisely what makes it a versatile, modular component that can be embedded within these broader regulatory structures. It provides the inner kernel for legitimate AI assistance that, when fleshed out in more specific implementations, can be made to fully address the demands of the more extensive regulatory frameworks mentioned above.

8 Conclusion

The ubiquity of LLMs requires a clear philosophical account of their legitimate use in the sensitive domain of ethical deliberation. The paper has argued that this legitimacy hinges on maintaining a strict separation between the context of discovery, where AI assistance can be invaluable, and the context of justification, which must remain the exclusive domain of accountable human verification based on human reason. The paper formalized this position into five criteria (C1–C5) of philosophical legitimation and showed their practical feasibility with the AIMED workflow. If the proposed account is correct, then, possibly, many current, unstructured uses of LLMs in deliberation will show themselves as failing to meet these conditions. The point is not to reject the powerful tools current AI makes available to us, but to discipline their use in a philosophically informed way within a framework that preserves the integrity of human moral agency, along the lines of Digital Humanism.

References

Astekin, Merve, Max Hort, and Leon Moonen. 2024. "An Exploratory Study on How Non-Determinism in Large Language Models Affects Log Parsing." In *Proceedings of the ACM/IEEE 2nd International Workshop on Interpretability, Robustness, and Benchmarking in Neural Software Engineering*, 13–18. InteNSE '24. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3643661.3643952.

Bærøe, Kristine. 2014. "Translational Ethics: An Analytical Framework of Translational Movements Between Theory and Practice and a Sketch of a Comprehensive Approach." *BMC Medical Ethics* 15 (1): 71. https://doi.org/10.1186/1472-6939-15-71.

Cribb, Alan. 2010. "Translational Ethics? The Theory–Practice Gap in Medical Ethics." *Journal of Medical Ethics* 36 (4): 207–10. https://doi.org/10.1136/jme.2009.029785.

Douglas, Heather. 2000. "Inductive Risk and Values in Science." *Philosophy of Science* 67 (4): 559–79. https://doi.org/10.1086/392855.

EU. 2025. "AI Act | Shaping Europe's Digital Future." September 16, 2025. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

Hume, David. 1739. *A Treatise of Human Nature (2003 Edition)*. Courier Corporation.

ISO. 2023. "ISO/IEC 42001:2023." ISO. 2023. https://www.iso.org/standard/42001.

Kagarise, Mary Jane, and George F. Sheldon. 2000. "Translational Ethics: A Perspective for the New Millennium." *Archives of Surgery* 135 (1): 39–45. https://jamanetwork.com/journals/jamasurgery/article-abstract/390485.

Lean, Oliver M., Luca Rivelli, and Charles H. Pence. 2023. "Digital Literature Analysis for Empirical Philosophy of Science." *The British Journal for the Philosophy of Science* 74 (4): 875–98. https://doi.org/10.1086/715049.

Memmert, Lucas, and Eva Bittner. 2022. "Complex Problem Solving Through Human-AI Collaboration: Literature Review on Research Contexts." *Hawaii International Conference on System Sciences 2022 (HICSS-55)*, January. https://aisel.aisnet.org/hicss-55/cl/machines_as_teammates/3.

Ouyang, Shuyin, Jie M. Zhang, Mark Harman, and Meng Wang. 2025. "An Empirical Study of the Non-Determinism of ChatGPT in Code Generation." *ACM Transactions on Software Engineering and Methodology* 34 (2): 1–28. https://doi.org/10.1145/3697010.

Reichenbach, Hans. (1938) 1938. *Experience and prediction: an analysis of the foundations and the structure of knowledge*. Chicago, Ill.,: The University of Chicago Press. http://archive.org/details/experiencepredic00reic.

Rivelli, Luca. 2025a. "A Soft Landing into the Singularity: Mediated Control Through AGI-Produced Algorithmic Solutions." Preprint. Philsci-Archive. https://philsci-archive.pitt.edu/24870/.

———. 2025b. "The Ethical No-Free-Lunch Principle: Fundamental Limits to Purely Data-Driven AI Ethics." In *Proceedings of 0th Moral and Legal AI Alignment Symposium, Joint IACAP/AISB 2025 Conference on Philosophy of Computing and AI, July 2025, University of Twente, Enschede, the Netherlands*, edited by Daniel D. Hromada and Bertram Lomfeld, 25–37. University of Twente, Enschede, the Netherlands.: International Association for Computing and Philosophy (IACAP) and the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB). https://udk.ai/alignment_symposium_0.pdf.

Sisk, Bryan A., Jessica Mozersky, Alison L. Antes, and James M. DuBois. 2020. "The "Ought-Is" Problem: An Implementation Science Framework for Translating Ethical Norms Into Practice." *The American Journal of Bioethics* 20 (4): 62–70. https://doi.org/10.1080/15265161.2020.1730483.

Tabassi, Elham. 2023. "Artificial Intelligence Risk Management Framework (AI RMF 1.0)." *NIST*, January. https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10.

Teehan, Ryan, Miruna Clinciu, Oleg Serikov, Eliza Szczechla, Natasha Seelam, Shachar Mirkin, and Aaron Gokaslan. 2022. "Emergent Structures and Training Dynamics in Large Language Models." In *Proceedings of BigScience Episode #5 -- Workshop on Challenges & Perspectives in Creating Large Language Models*, 146–59. virtual+Dublin: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.bigscience-1.11.

USgov. 2024. "Memorandum on Advancing the United States' Leadership in Artificial Intelligence; Harnessing Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security, and Trustworthiness of Artificial Intelligence." October 24, 2024. https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2024/10/24/memorandum-on-advancing-the-united-states-leadership-in-artificial-intelligence-harnessing-artificial-intelligence-to-fulfill-national-security-objectives-and-fostering-the-safety-security/.

VA. 2025. "Vienna Manifesto on Digital Humanism." March 3, 2025. https://digitalhumanism.at/en/2024/11/21/vienna-manifest-on-digital-humanism/.

Wolpert, David H., and William G. Macready. 1997. "No Free Lunch Theorems for Optimization." *IEEE Transactions on Evolutionary Computation* 1 (1): 67–82. https://doi.org/10.1109/4235.585893.