AI for engineering

MDPI

Article

Proper Problem Formulation as a Condition for the Responsible Use of AI in Engineering Design

Mariusz Mazurek [1,*] and Nadia Klimek [2]

[1] Institute of Philosophy and Sociology of the Polish Academy of Sciences, Nowy Świat 72, 00-330 Warsaw, Poland
[2] Faculty of Management, Lublin University of Technology, Nadbystrzycka 38, 20-618 Lublin, Poland; klimeknadiaa@gmail.com
* Correspondence: mariusz.mazurek@ifispan.edu.pl

Abstract

Assessments of responsible artificial intelligence (AI) use in engineering design usually focus on properties of the system, such as accuracy, reliability, transparency, bias and human oversight. The task given to the system (its objective, requirements, constraints and evaluation criteria) results from earlier choices that correct system operation cannot justify. Based on a critical-comparative analysis of the literature on AI ethics, design theory and requirements engineering, we distinguish the assessment of solution correctness from the assessment of the adequacy of problem formulation. In our coding, frameworks such as the NIST AI RMF, IEEE 7000, Value Sensitive Design and requirements validation cover this stage only partially. We identify three roles of AI in the design process: solver, evaluator and co-author of the task. The latter two act while the task is still being formed. We propose the Problem Formulation Assessment (PFA), a conceptual model with seven dimensions, including the validity of proxy measures, stakeholder representation and AI involvement in task formulation; only the last has no counterpart in the frameworks examined. The model is illustrated with generative space planning, a multi-agent specification system, passive vehicle safety tests and a hypothetical footbridge project. It has not been empirically validated, and we outline a validation agenda.

Keywords: artificial intelligence; engineering design; responsible AI; requirements engineering; validation; values in design

1. Introduction

Artificial intelligence is being applied to a growing number of engineering tasks. Machine learning systems detect damage, predict structural behaviour, support material selection, generate shapes, search design spaces and take part in the creation of CAD models. In the AI ethics literature, the related concerns are usually framed as requirements imposed on the system itself or on the way it is used. Those most frequently listed include transparency, justice and fairness, non-maleficence, responsibility and privacy [1,2]. Work on architecture, construction and structural engineering organises the issues in a similar way, adding, among others, reliability, safety and human oversight [3,4].

An AI system that mispredicts load-bearing capacity, detects damage with different effectiveness for different classes of objects, or produces outputs that its user cannot assess creates a direct technical and social risk. The assessment of system properties therefore cannot be omitted. The difficulty arises when the necessity of this assessment is taken to imply, tacitly, that the subject matter of AI ethics is primarily the correctness of the system's

Check for updates

Academic Editor: Yinlun Huang

Received: 25 September 2026
Revised: 1 November 2026
Accepted: 16 November 2026
Published: 23 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

operation with respect to the task it has been given. In engineering design, the task itself must first be formed. Before the system starts generating or evaluating solutions, someone has to decide what is to be achieved, which conditions are to be treated as non-negotiable, which features of a solution are to be compared, and how to reconcile objectives that cannot be maximised simultaneously.

This issue has not been entirely overlooked in AI ethics. On the basis of ethnographic fieldwork with a data science team, Passi and Barocas showed that moving from a general goal to a task that can be solved with data requires a series of decisions and may have different consequences for fairness [5]. Selbst and colleagues argue that formalising a social problem can remove from view elements that matter for evaluating its solution [6]. Ethics by Design and more recent accounts of Value Sensitive AI shift attention to earlier stages of system development and to the values that should be taken into account in the design process [7,8]. It cannot therefore be claimed that AI ethics as a whole treats the purpose of a system as neutral and beyond assessment.

A narrower problem nevertheless remains, concerning AI used in engineering design. The literature on responsible AI in engineering develops requirements for systems, organisations and professional oversight, whereas design theory and requirements engineering have long analysed how objectives and requirements come into being. The two strands partly overlap but start from different points. In the first, the intended purpose of a system is usually documented and treated as a reference point for subsequent assessment rather than as a separate object of assessment. Passi and Barocas observe that normative assessments of data science projects often take the formulation of the problem for granted [5]. In the second strand, establishing the task is itself part of design. Since both strands concern the same process, this difference calls for examination. The gap we address is one of integration rather than of omission: in the literature analysed, no framework for AI in engineering design treats the formulation of the task as a distinct object of assessment, links it to the verification–validation distinction and considers AI as a participant in formulation (Section 3.3).

Our research question is as follows. Can the ethical assessment of AI used in engineering design be limited to assessing how the system performs the task it has been given, if the formulation of that task itself involves choices about objectives, requirements, constraints and solution criteria? The answer we defend is negative. Correct execution of a task by an AI system does not guarantee that the design problem was properly formulated. Responsible use of AI therefore requires assessing both the operation of the system and the earlier process in which it was decided what problem the system is to solve.

A second, practical question follows. If assessment should cover how the problem is formulated, which dimensions should it address, and how does it differ from procedures that already exist in AI risk management, in value-based approaches to design and in requirements engineering?

The thesis requires two qualifications. First, we do not claim that every engineering problem is open to the same degree. In many tasks, requirements are tightly determined by physical laws, standards, regulations, contracts or well-established professional practice. Even then, however, the question remains whether the set of adopted requirements is complete with respect to the purpose of the project. Second, we do not claim that the problem of an improperly formulated task emerged with AI. It predates machine learning. What AI changes is that an increasing share of the path from objective to solution is carried out automatically, and that the system sometimes takes part in producing the problem description itself. For this reason, a problem familiar to design acquires a new technical context.

The article makes three contributions. The first is a distinction between two relations of assessment, the correctness of a solution relative to the task and the adequacy of the

task itself, which we relate to the classic verification–validation pair and extend with a normative dimension that requirements validation does not cover. The second is a typology of three roles of AI in the design process (solver, evaluator, co-author of the task) that shows where the involvement of contemporary AI changes the conditions of assessment. The third is a model for assessing problem formulation, derived from the literature analysis and compared with existing frameworks and illustrated with four examples, one of them a worked hypothetical application. The model is conceptual and has not been validated empirically; Section 7.4 states which of its dimensions consolidate existing concepts and which add to them. We do not claim that the ethical significance of problem formulation is a new discovery; we identify the work that has recognised it and specify what the model adds.

For readers who prefer a compact statement, the argument can be summarised in the following steps. The sections that support each step are given in brackets.

P1. In engineering design, the task handed to a computational system (objective, requirements, constraints and criteria) is the product of earlier choices; it is not simply given (Sections 4.1–4.3).

P2. These choices are constrained but underdetermined by technical knowledge, standards and physical laws, and they involve judgements about values, stakeholders and trade-offs (Sections 4.2 and 4.3).

P3. Therefore, the correctness of a solution relative to the task does not establish the adequacy of the task relative to the situation that the project is meant to change (Section 5).

P4. Contemporary AI raises the importance of P3: as a solver it intensifies the optimisation of whatever criteria it is given, as an evaluator it may implement criteria that are not explicit, and as a co-author it helps to write the task itself (Section 6).

P5. Existing frameworks cover parts of the formulation stage but, in our coding, leave proxy validity, the justification of trade-offs and AI involvement in formulation only partly addressed or not addressed at all (Sections 3 and 7.2).

C. Hence the assessment of responsible AI use in engineering design must include the adequacy of problem formulation. PFA is a structured way of organising this assessment; its usefulness remains to be tested empirically (Sections 7–9).

Section 2 describes the method and the selection of literature. Sections 3 and 4 reconstruct how the object of ethical assessment is framed in the AI literature, in design theory and in requirements engineering, and Section 3.3 states the research gap. Section 5 formulates the distinction between the two relations of assessment, and Section 6 analyses the three roles of AI and what distinguishes the latter two from earlier computational methods. Section 7 presents the assessment model, and compares it with existing frameworks and specifies its boundaries and novelty. Section 8 applies the model to three published examples and one worked hypothetical case, compares existing frameworks on these examples and considers objections, and Section 9 presents conclusions, limitations and a validation agenda.

We adopt the following terminological distinctions. The design problem is the situation that the project is meant to change, together with the needs and consequences that must be taken into account. The task is the formalised representation of this problem that is handed to a computational system. It comprises an objective (for example, an objective function), requirements, constraints and criteria for comparing alternatives. A requirement specifies properties that the solution must have, a constraint delimits the set of admissible solutions, and a criterion serves to compare and select alternatives. Problem formulation (in design theory also referred to as problem framing or problem setting) denotes the process of moving from the design problem to the task, and the formulated problem denotes the result of that process. Responsible AI refers in this article to the field of research and practice

concerned with the responsible development and use of AI, and responsible use of AI to the practice this field addresses.

2. Method and Selection of Literature

The article is conceptual and critical-comparative. It is not a systematic literature review and does not follow the PRISMA procedure. Its aim is to reconstruct how several research traditions treat the stage of problem formulation and to derive an assessment model from this reconstruction. The selection of literature was therefore purposive rather than exhaustive.

We analysed three areas. The first comprises AI ethics and responsible AI, including work on engineering applications and AI risk management frameworks. The second comprises design theory insofar as it addresses problem formulation and the translation of values into requirements. The third comprises requirements engineering and systems engineering, including the distinction between verification and validation and requirements engineering for machine learning systems. Works were included if they met at least one of the following conditions: they addressed normative aspects of formulating the task, objective or requirements for an AI or computational system; they addressed problem formulation or requirements in engineering design, including AI-assisted design; or they proposed responsible AI frameworks for engineering applications. Purely methodological machine learning papers and domain applications in which the formulation of the task was not reflected upon were excluded.

The starting point was the canonical work in each area. The selection was complemented by a documented search of the Crossref database, conducted on 23 September 2026 for publications from 2016–2026 (seven queries; for each, the first 25 results by relevance, screened by title). The queries, the numbers of results and the numbers of works judged relevant are given in Table A1 in Appendix A. This search was supplemented by a targeted search for the works closest to the thesis of the article, in order to assess the extent of its novelty.

The procedure has important limitations. The relevance ranking in Crossref is not very selective, screening was carried out by a single person on the basis of titles, and the Scopus and Web of Science databases were not used. For this reason, statements about gaps in the literature refer only to the analysed body of work. The comparison of frameworks in Section 7.2 relies on publicly available texts and, for standards available only for a fee, on the description of their scope and their tables of contents. The coding in the table is the authors' judgement, not a measurement, the rubric and the evidence for each cell are given in Appendix B.

The dimensions of the PFA model were not obtained by inductive coding of a corpus with an independent coding frame. They were derived analytically, by decomposing the transition from design problem to task (Section 4) into the decisions that shape the task: what is to be achieved (D1), what is inside and outside the problem boundary (D2), whose needs are represented (D3), how a value is operationalised as a measurable criterion (D4), how conflicting criteria are weighed (D5), which of these elements are proposed or implemented by AI (D6) and who approves the result (D7). Table 1 traces each dimension to the works that support it. Because this derivation is interpretive and was carried out by the authors, it is exposed to selection bias. We limited this risk by stating the inclusion criteria explicitly, by documenting the search and supplementing it with a targeted search for works that might pre-empt the thesis, by tracing each dimension to sources, and by treating the strongest objections, including the one on novelty, in Section 8.4. We did not use an independent coder or a second expert panel, either for the derivation or for the coding in Table 2. This is a limitation that we state in Section 9 and address in the validation agenda.

3. The Object of Ethical Assessment of AI in Engineering

3.1. System Properties and Conditions of Use

One of the basic ways of organising AI ethics is to identify classes of problems that arise when data are processed, outputs are produced and decisions are made. Mittelstadt and colleagues distinguish, among others, inconclusive evidence, inscrutable evidence, unfair outcomes and traceability, that is, the difficulty of establishing who is responsible for the effects of an algorithm [1]. The focus of this approach lies in the relationship between the system, the data, the output and the people affected by it. Many sets of ethical principles share a similar structure. The analysis by Jobin, Ienca and Vayena shows that guidelines converge on transparency, justice and fairness, non-maleficence, responsibility and privacy, with substantial differences in how these principles are interpreted [2]. On our reading, most of them answer the question of what properties a system should have or how its use should be organised.

In the engineering literature, this approach is adapted to professional practice. The review by Liang and colleagues of architecture, engineering and construction covers, among other issues, data security, transparency, reliability, trust, legal liability and the impact on employment [3]. Plevris and Hosamo propose a responsible AI framework for structural engineering that combines technical requirements for models with oversight, professional responsibility and risk management [4]. In both cases, the scope is broader than quality control of an algorithm. It includes the organisation, people and the context of use. There are therefore no grounds for the charge that responsible AI in engineering is concerned only with model parameters.

In the approaches analysed, assessment usually begins once the task for which AI is to be used is already known. We ask whether the system is sufficiently accurate, whether the data are appropriate, whether its operation can be explained, whether humans retain control and who is responsible for the outcome. All these questions are necessary, but they presuppose an earlier decision about what the system is supposed to do. If the system is to predict load-bearing capacity, we assess the accuracy of the prediction. If it is to generate structural alternatives, we assess the quality of the generated alternatives against the adopted requirements. If it is to support material selection, we check whether the recommendation is correct according to the established criteria. AI risk management frameworks do reach further back and require, among other things, documenting the intended purposes of a system and an initial decision on whether the system should be designed and deployed at all [9]. They do not, however, make the formulation of the problem a distinct object of assessment with elaborated criteria. We return to this issue in Sections 7 and 8.

This does not mean that the objective is treated as ethically neutral. Responsible AI frameworks often require consistency with values, law and the public interest. The difference is more specific. One can require a system to realise social values and still leave unexplained how to move from a general postulate to a concrete design task. In engineering practice, it is precisely at this stage that one must decide what will be measured, what will be a necessary condition, which quantities may be traded off against one another and which needs will be treated as typical cases. Ethical assessment of the system and general declarations of values do not yet resolve this problem.

3.2. Problem Formulation in AI Ethics

The strongest argument against the diagnosis presented above comes from work that deals directly with problem formulation. Passi and Barocas analyse the process in which an organisational goal is transformed into a problem that can be solved with data and computational methods [5]. On the basis of a case study, they show that this is not a purely technical activity. The choice of target variable, data, categories and success criteria

can change the meaning of the original goal. It follows that two teams starting from the same practical problem could create different computational tasks with different ethical consequences. For our purposes, this means that the AI literature already has conceptual tools with which to question the neutrality of the task set for a system.

A similar conclusion follows from the analysis by Selbst and colleagues [6]. The authors describe abstraction traps that arise when social concepts are translated into technical models. Two of them are particularly important here: drawing the boundaries of the system under assessment too narrowly (the framing trap), and reducing a normative concept to a formal definition that does not preserve all of its relevant properties (the formalism trap). If a complex problem is reduced to a measurable objective, formally correct operation of the system may fail to deliver what was expected of the solution in the broader context.

Ethics by Design develops a similar intuition in a different vocabulary. Brey and Dainow argue that ethical issues should be addressed during the design of AI rather than after the system has been built [7]. On the basis of a critical review of design processes, Sadek, Calvo and Mougenot argue in turn that designing value-sensitive AI should involve a broader socio-technical process in which values are identified, prioritised and translated into design requirements [8]. Van de Poel further points out that values are not features added to a finished algorithm but can be embedded in various elements of the socio-technical system to which AI belongs [10]. In their introduction to a Special Issue, Kudina and van de Poel frame AI as part of a broader socio-technical system, which further weakens the view that assessing the algorithm alone is sufficient [11]. These positions clearly shift the focus to a point before technical testing of the system.

The problem has thus been recognised, but it has been worked out mainly within research on fairness, data science, Value Sensitive Design and general AI ethics. In the reviews and frameworks on AI in engineering that we analysed [3,4], it does not appear as a separate issue and is not connected with the theory of how engineers formulate design problems. Whether such a connection is needed therefore depends on the status of problem formulation in design itself.

AI ethics does not overlook the formulation of tasks entirely, it covers organisations, stakeholders, values and consequences, and part of the literature explicitly problematises how tasks are formulated. There is therefore no gap consisting in a complete omission of this stage. In engineering applications, however, a problem of integration remains. The responsible AI frameworks analysed specify the ethical requirements that the system and the process of its use should meet, and risk management frameworks require the intended purpose to be documented. They do not, however, explain in any detail the nature of the transition from an open design situation to a formally specified task. If design theory shows that this transition involves significant choices, it belongs in the ethical assessment of AI in engineering.

3.3. Research Gap

The analysis in Sections 3.1 and 3.2 delimits the gap more narrowly than a claim of omission would. It consists of three missing connections within the corpus analysed. First, work on problem formulation in AI ethics [5,6,12,13] concerns mainly data-driven decision systems and social problems. It shows that formalising a problem is a normative act, and Abebe and colleagues describe computing as a "formalizer" that shapes how social problems are defined [13]. This work does not, however, draw on the account of problem setting developed in design theory, nor does it address AI used to produce technical artefacts. Second, the responsible AI frameworks for engineering [3,4] and the risk management frameworks [9] address system properties and the documentation of intended purpose, but they do not make the formulation of the task a distinct object of assessment with its

own criteria and evidence. Third, design theory and requirements engineering explain how problems are formulated and validated [14–23], but they do not consider AI as a participant in formulation or evaluation, and they treat the circle of stakeholders and the meaning of criteria as inputs to validation.

The gap is therefore one of integration and specification, not of recognition. What is missing is a way of treating problem formulation in AI-assisted engineering design as an object of assessment, linked to verification and validation and covering AI involvement in the formulation itself. The targeted search described in Section 2 found works that each cover part of this ground [5,6,12,24], but none that combines the three connections. This statement is limited to the corpus analysed (Section 2).

4. Problem Formulation in Engineering Design

4.1. Design as a Transition from an Existing to a Preferred Situation

Herbert Simon characterises design by reference to the difference between what exists and what is to be achieved. Everyone designs who devises courses of action aimed at changing existing situations into preferred ones [14]. This definition contains two elements. The first is the current state, which can be described with empirical knowledge. The second is the preferred state. This state does not follow from a description of reality alone. One must first judge that a certain state is better than another in a particular respect.

This difference bears directly on the tasks entrusted to AI. A system may help determine how to move from the current state to an adopted objective, but the fact that the objective has been formalised does not yet explain why this particular preferred state was chosen. For some problems the answer is relatively simple. If a damaged component is to be replaced, the required parameters follow largely from the function of the existing system, standards and safety conditions. In other design situations, there are many admissible ways of specifying the objective. One may minimise cost, energy consumption or manufacturing time, or adopt several objectives at once. The mere technical possibility of optimisation does not indicate which of these objectives should take priority.

Simon's account makes it possible to separate two questions that may be intertwined in practice. The first concerns how to achieve an adopted objective; the second concerns the choice of the objective. Computational methods are particularly effective with respect to the first, because they can search large solution spaces, evaluate alternatives and seek trade-offs. Their effectiveness does not, however, eliminate the second question. On the contrary, the larger the part of the search for solutions delegated to a system, the more important it becomes to decide in advance by what standard the results will count as good.

4.2. The Problem Is Not Simply Given

Donald Schön criticised the image of professional practice in which problems are presented in a form ready for the application of technical methods [15]. In real design situations, the practitioner rather encounters a situation that is unclear, incomplete or conflicting. Before solving the problem, the practitioner must shape it. Schön calls this process problem setting. It involves selecting the elements of the situation that will be treated as relevant and adopting a frame that determines what will count as an improvement.

On this view, solving and formulating a problem are not two entirely independent stages. The way a problem is formulated affects the space of available solutions, and anticipated solutions may prompt changes to the earlier formulation. Engineering work therefore involves iteration. On the basis of protocol studies of experienced designers, Dorst and Cross described it as the co-evolution of problem and solution [16]. The problem is refined as knowledge is gained about constraints, costs, materials, consequences and user

needs. This differs from a situation in which a system receives a ready-made objective function and a closed set of conditions and is expected to find the best result.

Earlier, Rittel and Webber had formulated a similar diagnosis for problems of social planning [17]. Their concept of wicked problems refers to situations in which there is no definitive formulation of the problem and the information needed to understand it depends on one's idea for solving it. Not every engineering problem has this character. Sizing a component for given loads and material may be a well-defined problem. The design of urban infrastructure, a public building or a transport system, by contrast, involves conflicting needs and consequences distributed across different groups. In such cases, the way the problem is formulated may determine which interests become visible in subsequent design.

The conclusion to be drawn from these accounts is not that design is arbitrary. The problem is constrained by technical knowledge, physical laws, standards, budget, available materials and many other conditions. These constraints do not, however, automatically yield a single objective. Different designs can be derived from the same set of conditions depending on what is judged most important. Problem formulation is thus both constrained and underdetermined.

4.3. From Values to Requirements

The issue becomes clearer when we move from a general objective to specific requirements. Van de Poel analyses the translation of values into design requirements by means of a values hierarchy, in which values are specified into norms and norms into requirements for the solution being designed [18]. Safety, accessibility, privacy or well-being are not yet parameters that can be fed directly into a calculation. This translation is not deductive. It requires attention to context and usually admits more than one tenable specification.

This process shows why formal requirements cannot be treated as a transparent record of values. Two teams may agree on the importance of safety and still define safety margins, test scenarios or user groups differently. They may also resolve the conflict between safety and cost, mass or performance differently. Mathematics makes it possible to calculate the consequences of particular alternatives, but it does not by itself produce the hierarchy of values from which the choice of a particular trade-off follows.

A similar problem arises in multi-criteria decisions. If two solutions belong to the Pareto front, that is, the set of non-dominated solutions, mathematical correctness does not indicate which of them to choose without additional information about preferences. Keeney argued that decision thinking should start from values rather than from the available alternatives, since otherwise the problem is easily confined to the criteria already present in a given tool or set of alternatives [25]. This is particularly important in computationally supported design. A system may rank alternatives very effectively according to the adopted criteria, but the quality of the ranking depends on whether those criteria correspond to what has actually been judged important.

Debates in the algorithmic fairness and privacy literature show that such trade-offs cannot be removed by better technique. Kleinberg and colleagues proved that, except in highly constrained special cases, no risk score can satisfy three plausible fairness conditions at the same time [26]. Chouldechova reached a related conclusion for calibration and error-rate balance when base rates differ between groups [27]. Friedler and colleagues argue that different fairness mechanisms rest on different assumptions about the relation between the constructs of interest and what is observed, so that the choice among them expresses a value system [28]. A similar structure appears in the trade-off between privacy and accuracy: Bagdasaryan and colleagues show for differentially private training that the loss of accuracy is not borne equally by all subgroups [29]. In engineering design, analogous conflicts arise between, for instance, safety for different occupant groups, mass, cost and accessibility. In

each case, a formally correct optimiser or evaluator presupposes a prior, normative decision about which criterion is to be satisfied and for whom. This is the content of dimension D5 and, when the choice is embedded in a measurable criterion, of dimension D4.

Value Sensitive Design develops methods for involving values and stakeholders in the design process [30]. IEEE 7000 describes a model process in which ethical values are elicited and prioritised as early as the concept exploration stage and then translated into ethical value requirements [31]. The standard applies to the design of systems in general, including AI systems, which shows that the two strands analysed here already meet. The existence of these methods also supports the earlier conclusion. If separate procedures are needed to identify values and translate them into requirements, it cannot be maintained that technical requirements arise automatically from the task description. They are the product of design work.

4.4. Problem Formulation, Requirements Engineering and Validation

The transition from needs to formalised requirements is the central concern of requirements engineering and systems engineering. ISO/IEC/IEEE 29148 describes processes for defining stakeholder needs and requirements, for requirements validation and for requirements traceability [19], and overviews of the field treat elicitation, modelling, negotiation and validation of requirements as distinct engineering tasks [20]. Zave and Jackson showed that requirements concern phenomena in the environment of a system and that the adequacy of a specification depends on the assumptions made about that environment [21]. If these assumptions are incomplete, a specification may be internally correct and still fail to ensure that the requirements are met.

The distinction between verification and validation is particularly close to the thesis of this article. Boehm captured it in a pair of questions: whether the product is being built right and whether the right product is being built [22]. Goal-oriented requirements engineering goes further and treats goals as the rationale for requirements, uncovered by asking why a given requirement is needed [23]. The distinction developed in Section 5 is therefore not new in engineering. What is new is the question of whether it is sufficient when the task is handed over to an AI system.

We argue that it is not, for three reasons. Validation checks the consistency of requirements with stakeholder needs, but it does not settle whether the circle of stakeholders taken into account is the right one, nor does it assess whether a measurable criterion that will enter the objective function preserves the meaning of the value it is meant to represent. These issues are taken up only by value-based approaches such as IEEE 7000 [31]. Moreover, in AI-assisted design some of the decisions that affect the result, such as criterion weights, thresholds and the way a solution is selected from the Pareto front, are made in the configuration of the tool rather than in the requirements document. Finally, requirements engineering for machine learning is concerned mainly with requirements for the AI system as a product [32,33]. In AI-assisted design, the product is a technical artefact, and AI is a tool that receives the task as input or co-creates it. Early studies on the use of large language models for requirements engineering tasks show that this is an actual practice rather than a mere possibility [34].

The assessment proposed here therefore does not replace requirements engineering. It extends validation with two elements: the normative justification of choices that validation takes as given, and the involvement of AI in the creation of the task itself.

The concepts analysed differ in vocabulary and scope but converge on one finding: an engineering problem is not always a ready-made object that can simply be handed to a solving system. The state to be achieved must be specified, requirements selected, constraints set and a decision made about how trade-offs will be assessed. Some of these

decisions follow from technical knowledge and standards, others depend on which needs, risks and values have been judged relevant. The way the problem is formulated is thus itself a part of design, and requirements engineering provides the procedures that document and validate this part. If assessment within responsible AI relates the operation of a system to its task, and the task results from earlier choices, a full assessment cannot begin only after those choices have been made. This is the problem the article addresses.

5. Correct Solution Versus Proper Problem Formulation

In AI-assisted design, two relations of assessment must be distinguished (Figure 1). The first holds between the output of the system and the task the system received. One can examine whether a generated alternative meets the requirements, whether the model predicts the relevant parameters, whether constraints have been violated and whether the selection procedure works as intended. These are technical questions, even though they sometimes also have ethical consequences.

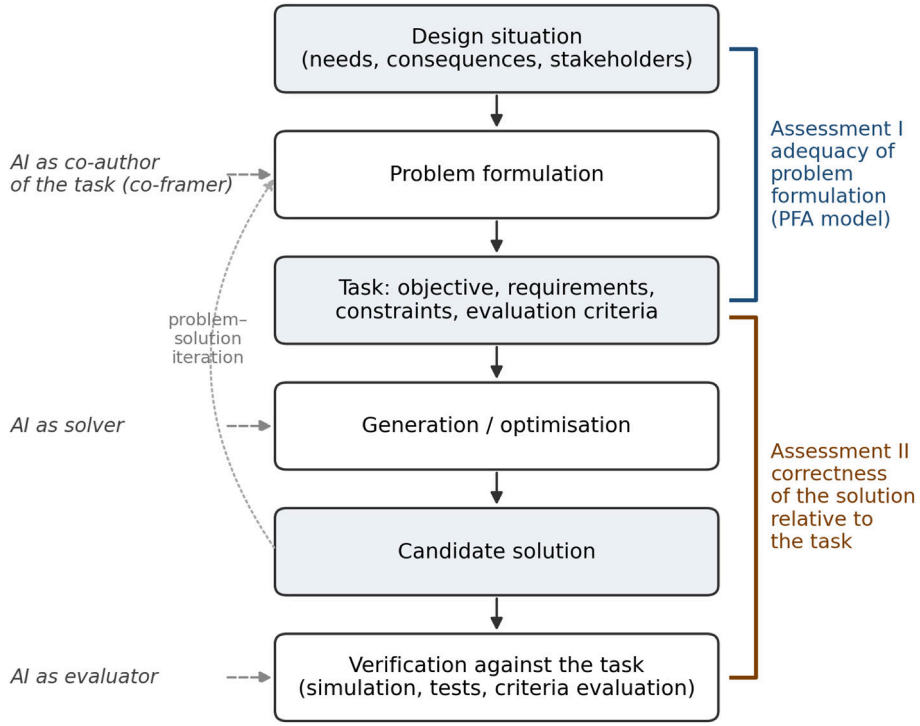

Figure 1. Two relations of assessment in AI-assisted design and the points at which AI can take part in the process. Assessment I concerns the adequacy of problem formulation; Assessment II concerns the correctness of the solution relative to the adopted task. Solid arrows indicate the sequence of design steps, dashed arrows the points at which AI can enter the process, and the dotted curved arrow problem–solution iteration; the blue and brown brackets mark Assessment I and Assessment II, respectively. Source: own elaboration.

The second relation holds between the formulation of the problem itself and the situation that the project is meant to change. Here we ask whether the right objective was chosen, whether the requirements omit a relevant group of users, whether the adopted criterion actually represents the value it was meant to represent, and whether the boundaries of the problem have been drawn too narrowly. The answers require technical knowledge but cannot be reduced to it. They also involve judgements about acceptable risk, the distribution of benefits and costs, user needs and the consequences for people who do not take direct part in the design process. In systems engineering terms, the first relation corresponds to verification and the second is close to validation [22]. It differs from validation in that it

also covers the justification of choices that validation takes as given, such as the circle of users represented.

The two relations may coincide: a well-formulated problem may be solved correctly. They may, however, diverge. A system may solve a properly formulated problem incorrectly; this is the classic case of model or algorithm failure. It may also correctly solve an improperly formulated problem. In the latter case, improving the AI model itself does not remove the difficulty. The system may become more accurate, faster and more transparent and yet pursue more consistently an objective that should have been formulated differently.

This possibility matters because technical assessment detects the first kind of problem more easily. If an alternative violates a strength requirement, fails a simulation or falls outside the admissible geometry, the error becomes visible during verification. If, by contrast, an important requirement was omitted, the result may pass all tests, because the tests were built on the same incomplete formulation. A missing criterion does not generate a message that the criterion has been violated.

A well-documented engineering example, which does not involve AI, shows this asymmetry. The London Millennium Footbridge was opened in June 2000, and on the opening day unexpected lateral movements occurred as pedestrians crossed it [35]. Lateral crowd synchronisation was not a loading effect anticipated in the design [35]. A verification against the load model of the day would therefore have passed, because the missing element lay in the formulation of that model and in its assumptions about the environment of the structure (cf. [21]). The case is not about AI, and we do not treat it as an ethical failure in itself. It illustrates the structure at stake: an analysis that is correct relative to its inputs cannot reveal what the inputs omit. If such a load model were the fixed input of an automated generative search, the search would have no basis for detecting the omission.

One might reply that the difference between solving and formulating a problem is obvious and requires no separate ethical analysis, since in engineering practice requirements are agreed, approved and revised. The objection would be apt if the article claimed to have discovered a new design activity. It does not. The issue concerns the place of this activity in the assessment of responsible AI use.

If the quality of an AI application is assessed only from the moment when objectives and requirements have already been adopted, a system that operates correctly within the limits of a poorly formed task may receive a positive assessment. From an engineering perspective, the design process may still be subject to control. From the perspective of responsible AI, however, the question arises whether this earlier control is part of the assessment of AI or lies outside it. Our thesis is that in design applications it should not lie outside it, because the formulation of the problem determines what the AI will search for and what it will have no reason to consider.

The point is not to change the meaning of technical correctness. A solution can be correct with respect to the adopted requirements. The point is to limit the conclusion that can be drawn from this correctness. It does not follow that the adopted requirements include everything that a responsible design should take into account.

The phrase “properly formulated problem” might suggest that every project has a single objectively correct formulation. We make no such assumption. In many projects, different reasonably justified objectives and different trade-offs are possible. Ethical assessment is not meant to find the only admissible formulation.

By proper problem formulation we mean a weaker condition. The objective and requirements must be justified in the light of available technical knowledge, standards and professional duties, and the process of establishing them should not omit foreseeably relevant needs and consequences merely because they are difficult to express as a measurable

criterion. Where several admissible options exist, it should be possible to state the reasons for choosing one of them.

This understanding does not eliminate normative disagreement. It does, however, make it possible to distinguish a pluralism of justified solutions from a situation in which a relevant element of the problem has not been recognised at all. Proper problem formulation does not mean complete knowledge of all future consequences. It means, rather, that the design team can explain why it chose a given objective, which needs it took into account, which constraints it adopted and which trade-offs it considered acceptable.

6. Three Roles of AI and the Significance of Problem Formulation

The problem of an improperly formulated task is not new. Engineers have always been able to optimise the wrong objective, omit a group of users or draw the boundaries of a problem too narrowly. It would therefore be unjustified to claim that AI creates a new category of design error. We defend the weaker thesis that AI can change the scale and course of a familiar problem. How it does so depends on the role AI plays in the process. We distinguish three such roles, marked in Figure 1.

6.1. AI as Solver

In the first role, AI searches the solution space according to the adopted criteria, and what is new is above all the scale of this search. A review of deep generative models in engineering design documents the rapid development of methods for generating shapes, structures and other design solutions [36]. Attempts have also been made to incorporate neural networks into topology optimisation, although a critical review of these attempts points to a lack of convincing results [37]. Language and multimodal models, in turn, are being tested at successive stages of design and manufacturing [38,39]. In such applications, the system does more than compute a property of a previously created design. It can generate many alternatives and rank them according to criteria.

The effectiveness of such a process is an advantage when the criteria correspond well to the purpose of the project. The same effectiveness, however, can amplify the consequences of an incomplete criterion. According to the generalised formulation of Goodhart's law popularised by Strathern, when a measure becomes a target, it ceases to be a good measure [40]. Manheim and Garrabrant distinguish several mechanisms of this phenomenon and indicate that their importance grows with the strength of optimisation applied to the proxy [41]. Thomas and Uminsky relate the problem directly to AI and argue that over-reliance on measurable targets leads to real-world harms [42]. Zhuang and Hadfield-Menell showed in a formal model that, under certain assumptions, unconstrained pursuit of an incomplete proxy objective that captures only some of the relevant attributes can lead to arbitrarily poor outcomes [43].

It does not follow that every AI system will "game" its objective function, or that the better the optimisation, the worse the social outcome. What follows is only that, if a criterion is incomplete, a system that searches effectively according to that criterion has no basis of its own for correcting the earlier omission. A value not reflected in the objective, the constraints or an additional assessment does not directly affect how a given optimisation mechanism ranks alternatives.

The same role raises the problem of models trained on existing designs. Historical data contain design knowledge, but they may also contain entrenched conventions and the limitations of past solutions. It cannot be claimed that a model automatically adopts all the values present in the data. One can, however, point to the possibility of reproducing statistical patterns without distinguishing those that are technically desirable from those that result from accidental or historical practices.

Regenwetter and colleagues argue that metrics based on statistical similarity may not capture the requirements of engineering applications, and they propose metrics for constraint satisfaction, functional performance, novelty and conditioning of generative models [44]. In design, what matters is not only similarity to the training set but also the fulfilment of functions and constraints, feasibility and novelty. This conclusion can be cautiously extended. Since the quality of a design cannot be reduced to similarity to historical data, responsible use of a generative model requires criteria external to the data distribution itself.

This point links the problem of data with the problem of task formulation. If the external criteria are incomplete, the model may receive strong information from historical data and weak information about the values the project should realise today. Improving the quality of the generative model alone therefore does not solve the problem. What is needed is a prior decision about the standard by which the generated solutions will be evaluated.

The solver role is not specific to AI. Classical optimisation and evolutionary algorithms have the same structure, as the example discussed in Section 8 shows. What AI mainly increases here is the scale of search and the strength of optimisation, and hence the importance of the validity of the adopted criteria.

6.2. AI as Evaluator

In the second role, AI does not generate solutions but evaluates them: it compares alternatives, estimates feasibility, detects defects or selects a concept. Picard and colleagues examined vision–language models in such tasks, including assessing design similarity, selecting concepts with Pugh charts, analysing engineering drawings and assessing manufacturability [39]. The results were mixed. In similarity assessment, the model was more self-consistent than humans performing the same task, but in concept selection it often failed to fill in the Pugh chart or produced only an imitation of one [39].

This role is particularly significant for problem formulation. When an evaluation criterion is an explicit formula, one can ask whether it represents the value it is meant to represent. When the criterion is implemented by a learned model, it becomes part of the model and is difficult to inspect directly. The question of proxy validity then takes the form of a question of construct validity, which Jacobs and Wallach analysed in the context of fairness [45]. Regenwetter and colleagues note that even the metrics used to evaluate generative models themselves in engineering design may not capture the requirements of engineering applications [44]. If AI acts as an evaluator, the assessment of problem formulation must include the question of which criteria the model actually implements and how they have been checked.

6.3. AI as Co-Author of the Task

The third role concerns the point at which AI enters the process. In earlier computational applications, humans often defined the problem, and the tool performed calculations or searched the space of alternatives. These tools should not, however, be contrasted with contemporary AI as if earlier methods never generated solutions. Evolutionary algorithms and topology optimisation have long been able to create forms that the designer did not construct directly. What distinguishes contemporary systems is rather the broader range of activities in which they can take part.

Obieke, Bridgeman and Han describe a framework for AI collaboration in engineering design in which a multi-agent system based on language models generates, among other things, problem definitions and product design specification documents [46]. Clay and Sha describe a shift from design thinking in which the solution space and the objective space co-evolve in the designer's work to design thinking that also uses generative AI to

define objectives [47]. Language models are also used to transform verbal descriptions into specifications and more formal design representations, such as CAD code [38]. Norheim and colleagues analyse the challenges of using large language models for requirements engineering tasks [34]. If a system helps not only to find a solution but also to write down the task itself, the boundary between formulating and solving the problem becomes less distinct.

The available empirical evidence does not support the assumption that such assistance improves the quality of problem framing. In an experiment with 280 participants, Shin and colleagues found no evidence that using language models improved the quality of design problem reframing; it did, however, widen the gap between experienced and inexperienced designers [48]. This is an argument for subjecting AI-generated task proposals to assessment at least as careful as that applied to proposed solutions.

This does not mean that AI takes over responsibility. The system's proposal must be accepted, modified or rejected within the design process. What changes, however, is the condition for responsible oversight. Humans cannot limit themselves to checking the finished solution. They must also understand which assumptions have been built into the problem description, especially when part of that description was suggested by the tool.

Generative tools can influence the formation of requirements in several ways, which we state as hypotheses grounded in the literature rather than as established findings. First, a drafted specification supplies a complete-looking list of requirements and stakeholders, so that omissions are harder to see than in an empty template. Second, once a draft exists, later revision tends to adjust it rather than to reconsider the framing, which is the familiar anchoring effect [49]. Third, fluent text suggests completeness, and the evidence on automation bias shows that users of imperfect decision aids make both omission and commission errors [50]. The experiment of Shin and colleagues [48] adds that fluent proposals do not by themselves improve the quality of reframing.

Over-reliance on AI-generated objectives and constraints carries three risks. The first is the silent loss of requirements: items that the tool does not propose are never discussed (D2, D3). The second is the transfer of authority: a proposal presented as a technical output may be adopted without an identifiable decision-maker (D7). The third is circularity: if the same or a similar model proposes the requirements and later evaluates the design against them, the evaluation cannot reveal the model's own omissions (D6).

Safeguards should therefore operate upstream of the solver. We propose five, each tied to a dimension of the PFA model: (1) logging the provenance of task elements, so that AI-proposed elements are marked (D6); (2) independent review of AI-generated stakeholder lists and requirements, preferably by a person who first listed them without seeing the draft, in order to reduce anchoring (D2, D3); (3) separation of the proposer from the evaluator, so that the model that drafted a requirement is not the only check on it (D6); (4) approval of the final task by a named role, with a record of what was changed relative to the AI proposal (D7); and (5) a justification of target values and weights that does not rely on the model (D4, D5). These safeguards do not make AI proposals reliable. They make the status of the proposals visible and keep the decision with identifiable people. They have not been tested.

The three roles differ in which element of Figure 1 AI affects. As a solver, it operates within Assessment II, but the strength of optimisation amplifies the effects of errors made earlier. As an evaluator, it implements criteria that belong to the task, although they are not always explicitly stated. As a co-author of the task, it takes part directly in the stage covered by Assessment I. The latter two roles are characteristic of contemporary AI systems, particularly language and vision–language models, and it is these roles that most

strongly justify including the assessment of problem formulation in the assessment of AI use. Section 6.4 discusses how they differ from earlier computational methods.

6.4. What Distinguishes Dimension D6 from Earlier Computational Methods?

Computational tools have long embedded choices about task formulation. Optimisation packages ship with default objectives, weights and constraint handling, and expert systems and parametric CAD templates encode a designer's earlier framing. Dimension D6 is therefore not wholly new in logic, and the corresponding defaults are covered by D5 and D7 of the model. We claim a narrower difference, in three respects. The first concerns the translation of needs into a task. Classical tools accept a formal task written by a person, so the step from stakeholder needs expressed in natural language to a formal task was performed, and could be inspected, by that person. Language models can perform this step themselves, from verbal descriptions to specifications and CAD code [38,46], so that the translation that validation must check is now partly produced by the tool. The second concerns the form of the criteria. A formula-based objective can be read, whereas a learned evaluator implements criteria that are implicit in its parameters and training data and can be examined mainly through its behaviour on test cases [39,45]. The third concerns the form of the output. Proposals from generative models are fluent and complete-looking, which, together with the evidence on automation bias [50], makes uncritical adoption more likely than with the tabular output of a solver. These are differences of degree and of the kind of translation involved, not a new moral category. For tasks in which AI takes no part in formulation or evaluation, D6 is simply recorded as not applicable, as for Project Discover in Section 8.1.

7. A Model for Assessing Problem Formulation

7.1. Dimensions of Assessment

From the analysis we derive a model for assessing problem formulation, which we call PFA (Problem Formulation Assessment). The model is not a checklist to be ticked off mechanically. It organises the questions that a design team should be able to answer before adopting a task as the basis for the work of an AI system, and it indicates what evidence should document the answers. Each dimension derives from a specific part of the analysis, as shown in the last column of Table 1. Figure 2 shows where the dimensions apply in the process that leads from the design problem to a solution. PFA is a conceptual model: a structured set of questions and evidence requirements, not a validated measurement instrument (Section 8.4).

Table 1. Dimensions of the Problem Formulation Assessment (PFA) model. Source: own elaboration.

Dimension	Guiding Question	Expected Evidence	Warning Sign	Basis in the Analysis
D1. Justification of the objective	Why were this preferred state and this objective adopted rather than another admissible one?	Record of the reasons for choosing the objective and of the alternatives considered	Objective taken over from the default tool configuration or from a previous project without justification	[5,14,23,25]; Sections 4.1 and 4.4

Table 1. *Cont.*

Dimension	Guiding Question	Expected Evidence	Warning Sign	Basis in the Analysis
D2. Problem boundaries and completeness of requirements	What was left outside the problem boundaries, and where are foreseeably relevant requirements that are absent from the objective function controlled?	List of requirements with the assigned point of control: objective function, constraint or separate verification	A requirement absent from every stage of the process	[6,17,21]; Section 5
D3. Representation of stakeholders and users	Whose needs, bodies and use situations does the task represent, and who is missing from it?	Stakeholder map; description of the population represented in scenarios and tests	Evaluation criteria built on a single reference group	[30,51,52]; Section 8.1
D4. Validity of proxy measures	Does the measurable criterion preserve the meaning of the value it is meant to represent, also under strong optimisation?	Justification of the operationalisation; analysis of what the measure does not capture	A criterion chosen because it is easy to compute	[6,18,40–43,45]; Sections 6.1 and 6.2
D5. Justification of trade-offs	Who adopted the weights, thresholds and the choice of a design from the Pareto front, and on what basis?	Record of decisions on weights, thresholds and design selection, including the decision-makers	Default tool weights and thresholds; design selected without justification	[18,25]; Sections 4.3 and 4.4
D6. AI involvement in formulating and evaluating the task	Which objectives, requirements, stakeholder groups or evaluation criteria were proposed, or are implemented, by an AI system?	Provenance of task elements; description of the criteria implemented by an evaluator model and of how they were checked	Model-generated specification adopted without review; evaluator model with unchecked criteria	[34,38,39,46,48]; Sections 6.2 and 6.3
D7. Approval and traceability	Who approved the task as the basis of the design, and can the path from value to criterion be traced?	Record of approval; value–requirement–criterion linkage	No owner of the decision to adopt the task	[19,31,53]; Section 7.3

Dimensions D1–D5 concern the content of the task, and D6 and D7 the process by which it is created. The dimensions partly overlap: the lack of representation of a user group (D3) usually shows up as a missing requirement (D2). We keep them separate because they call for different evidence. For each dimension, we propose a qualitative assessment on three levels: documented and justified; documented but not justified; undocumented. We deliberately do not propose an aggregate score. A summary score would become yet another proxy and would be subject to the same distortions that the model is meant to reveal [40,41]. The overlap is deliberate where it reflects the way problems appear in practice, but each dimension has its own object of judgement and its own evidence; rules for assigning an issue to a dimension are given in Section 7.4 (Table 3).

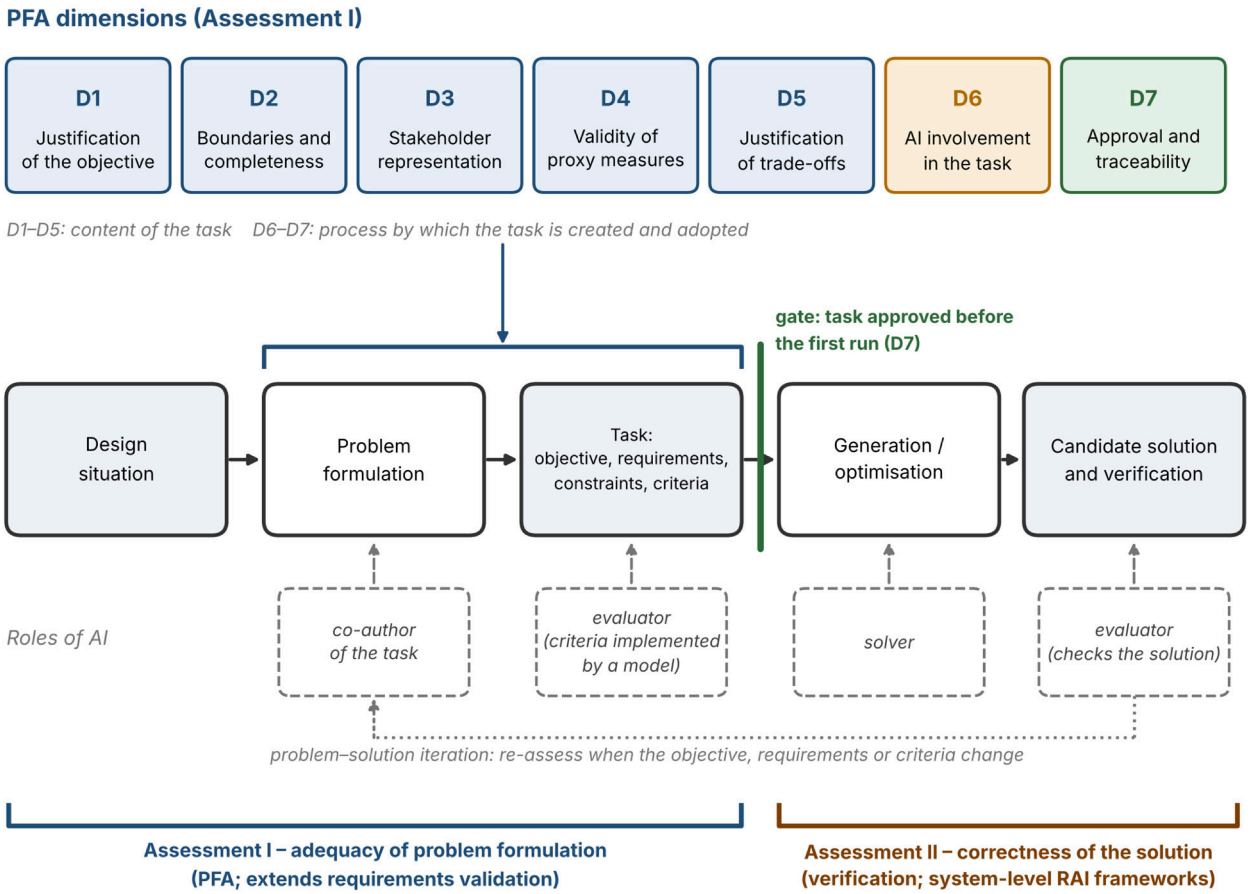

Figure 2. The PFA model in the design process. Dimensions D1–D5 concern the content of the task and D6–D7 the process by which it is created and adopted. They are applied while the task is formed and before it is adopted; the assessment is completed, and the task approved, before the AI system is first run. AI can take part as co-author of the task, as evaluator and as solver. The dotted loop indicates that the assessment is repeated when the objective, requirements or criteria change. The blue and brown brackets mark Assessment I and Assessment II, respectively; the green vertical line marks the approval gate (D7); dashed boxes indicate the roles of AI. Source: own elaboration.

Not all answers need to be recorded as numerical values. This matters because part of the problem stems precisely from equating what is important with what is easy to measure. A requirement may be qualitative or may be checked at a separate stage. What matters is not that every value be placed in the objective function, but that it not disappear from the process merely because it has not been formally encoded.

The assessment should be carried out before the system is first run and repeated whenever the objective, requirements or criteria change significantly. In many organisations, the necessary information already exists in design documentation, requirements analysis, risk analysis and records of agreements, and in organisations that apply AI risk management frameworks also in the documentation of the intended purpose and context of the system [9]. The model therefore does not require a new elaborate procedure but rather the linking of this information with the assessment of AI use. The assessment of the model, data, explainability and human oversight remains necessary, but it is preceded by the question of the reasons behind the task.

7.2. Relation to Existing Frameworks

Table 2 compares the model with the frameworks that Sections 3 and 4 identified as the closest. The coding indicates whether a given dimension is addressed explicitly

(E), partially or indirectly (P), or not at all (N) in each framework. This is the authors' judgement, based on the texts indicated in Section 2 and made with the rubric in Appendix B. It is not a measurement and was not checked by an independent coder. The table describes the coverage of the formulation stage; it does not rank the frameworks. Appendix B gives the textual evidence for each cell, so that the coding can be checked and challenged.

Table 2. Coverage of the PFA dimensions by existing frameworks. E—addressed explicitly; P—addressed partially or indirectly; N—not addressed. Own coding based on [4,9,19,23,30,31].

Dimension	NIST AI RMF	IEEE 7000	VSD	Requirements Engineering (Validation, GORE)	RAI Framework in Structural Engineering	PFA Model
D1. Justification of the objective	P (MAP 1.3–1.4; go/no-go decision)	P (values at the concept exploration stage)	P	P (goals as rationale for requirements)	N	E
D2. Boundaries and completeness	P (MAP 1.1: context and assumptions)	P	P	E (validation, completeness)	N	E
D3. Stakeholders	E (MAP 1.1–1.2)	E	E (direct and indirect stakeholders)	E (stakeholder needs)	P	E
D4. Proxy validity	P (MAP 2.3: construct validation)	?	N	P (verifiability of requirements)	N	E
D5. Trade-offs	P (trade-offs among trustworthiness characteristics)	P (value prioritisation)	P (value tensions)	P (requirements negotiation)	N	E
D6. AI involvement in formulating and evaluating the task	N	N	N	N	N	E
D7. Approval and traceability	P (documentation, accountability)	E (value traceability)	N	E (requirements traceability)	P (professional responsibility, oversight)	E
Assessment II: operation of the AI system	E	P	N	P	E	N

?—not determined from the publicly available description of the standard. The coding of IEEE 7000 and ISO/IEC/IEEE 29148 is based on the description of their scope and their tables of contents.

The comparison yields three observations. The dimensions concerning stakeholders and traceability are well covered by existing frameworks, so the model should draw on them rather than duplicate them. The validity of proxy measures (D4) and the justification of trade-offs (D5) appear in them only partially; the NIST AI RMF, for example, mentions construct validation among methodological considerations [9] but does not link it to the objective function of a design tool. AI involvement in formulating and evaluating the task (D6) is not addressed by any of the frameworks analysed. The model, for its part, does not assess the operation of the AI system itself (Assessment II), which remains the subject of existing responsible AI frameworks.

Lists of questions for the early stages of AI development have been proposed before, for example in fairness checklists [54] and in tools that support deliberation on public-sector AI proposals [55]. The PFA model differs from them in its object and context. It concerns design tasks whose outcome is a technical artefact, it covers multi-criteria decisions and safety requirements, and it treats AI involvement in task formulation as a separate dimension of assessment.

Table 2 should not be read as showing that the other frameworks are weaker. Each was developed for a different purpose: the NIST AI RMF and IEEE 7000 for the governance and ethical design of systems in general, Value Sensitive Design for the elicitation of values, and requirements engineering for the specification of any system. They address questions that PFA does not, as the last row of Table 2 shows. The claim is narrower: for AI-assisted engineering design, the questions on proxy validity, trade-offs and AI involvement are not made explicit in them, and PFA is meant to be used together with these frameworks, not instead of them (Section 7.6).

7.3. Responsibility and Oversight

In AI-assisted design, responsibility for the outcome is distributed among many actors, as analysed by Douglas and colleagues [53,56]. The thesis of this article does not require settling the entire debate on the responsibility gap [53]. A weaker finding suffices. An AI system does not morally justify an objective merely because it can pursue it effectively. Responsibility for adopting the objective and requirements remains with the people and organisations who formulate or approve them or who accept the system's proposal. This matters especially when AI takes part in defining the problem. A model-generated description of requirements should not be treated as a neutral technical output. It is a proposal that requires assessment in the same sense as a proposed geometric or material solution. Humans remain responsible not because they must personally devise every requirement, but because they, or the relevant institution, decide to adopt the task as the basis for further design.

The model does not add a new catalogue of principles to existing sets of responsible AI principles. It expresses a single principle that orders the scope of assessment. Responsible use of AI in engineering design requires assessing not only whether the system performs the task correctly but also whether the problem it is to solve has been properly formulated. This principle does not compete with safety, fairness, transparency or accountability. It specifies the point from which they should be applied. The assessment of fairness should not be limited to the distribution of outcomes after the system has been run if the problem arose already in the choice of represented users. Safety should not be limited to checking whether a generated alternative passes the adopted test if the test itself is inadequate for the population. Human oversight should not begin only when the finished result is approved if the system has earlier co-created the requirements.

7.4. Boundaries, Origins and Novelty of the Dimensions

The dimensions are not independent, and we do not claim that they partition the problem space. They are separated by the object of judgement and by the evidence they call for. Table 3 states, for each dimension, what it assesses, how it is separated from its nearest neighbour, where it comes from and what the model adds. Four pairs raise the question of overlap most often. D1 and D5: D1 concerns whether the objectives adopted were the right ones to adopt, whereas D5 presupposes at least two admitted objectives or criteria and concerns how they are weighed against each other. D2 and D7: D2 concerns whether a foreseeably relevant requirement exists and has a point of control, whereas D7 concerns whether the resulting set was approved by an identifiable owner and can be traced to values;

a task can have complete requirements and no accountable approver, or the reverse. D4 and D5: D4 asks about the validity of one criterion as a representation of a value, whereas D5 asks about weights, thresholds and selection across criteria, so a task can use valid proxies and still hide an unjustified trade-off. D2 and D3: a missing user group usually appears as missing requirements, but it is a D3 issue when the group is absent from scenarios, tests or criteria, and a D2 issue when the group is represented but a requirement concerning it has no point of control. Where an issue legitimately falls under more than one dimension, it is recorded under each, with the evidence specific to that dimension.

Table 3 also answers the question of what is new. Dimensions D1, D3 and D7 largely consolidate established concepts (goal rationale, stakeholder analysis, traceability), and they are included to make the model self-contained. Dimensions D2, D4 and D5 re-specify known concepts for a setting in which part of the search and evaluation is automated: D2 requires an explicit point of control for each requirement, D4 asks about proxy validity under strong optimisation, and D5 locates trade-off decisions in tool configuration (weights, thresholds, defaults) as well as in the requirements document. Dimension D6 has no counterpart in the frameworks examined (Table 2). The contribution of PFA is therefore the selection and integration of these questions around the formulation stage, one new dimension, and the recasting of the older ones for AI-assisted design. It is not seven new concepts.

The overlap with requirements validation is considerable and intended. Validation asks whether requirements correspond to stakeholder needs, and D2 and D7 use the same evidence (requirement lists, traceability records). PFA goes beyond validation in three ways that can be separated in practice: (a) it asks for the justification of choices that validation takes as given, namely the circle of stakeholders (D3) and the meaning of the measurable criterion (D4); (b) it covers decisions made in tool configuration rather than in requirements documents (D5); and (c) it covers task elements generated or implemented by AI (D6). Where a project already has a mature validation process, the PFA can be performed as an extension of it, and dimensions D2 and D7 may be satisfied by its records.

Table 3. Distinctness, origin and contribution of the PFA dimensions. Source: own elaboration.

Dimension	Object of Judgement	Boundary Rule	Origin	What PFA Adds
D1. Justification of the objective	The end: why this preferred state and objective were adopted.	Precedes the criteria. Differs from D5, which presupposes at least two admitted objectives or criteria.	Goal rationale in requirements engineering [23]; design as change towards a preferred state [14]; value-focused thinking [25].	Largely consolidates. Makes the record of the choice of objective an explicit item in the assessment of AI use.
D2. Problem boundaries and completeness	Coverage: what lies inside and outside the boundary, and where each requirement is controlled.	Asks whether a requirement exists and has a point of control. Differs from D7 (approval) and D3 (who is represented).	Abstraction traps [6]; assumptions about the environment [21]; wicked problems [17].	Re-specifies. Requires a point of control (objective, constraint or separate verification) for each requirement.
D3. Representation of stakeholders and users	Representation: whose needs, bodies and use situations the task encodes.	Concerns presence in scenarios, tests and criteria. D2 concerns the control of requirements.	Value Sensitive Design [30]; user representation in tests [51,52].	Consolidates. Applies the question to the evaluation scenarios of AI-assisted tools.

Table 3. *Cont.*

Dimension	Object of Judgement	Boundary Rule	Origin	What PFA Adds
D4. Validity of proxy measures	Validity: the relation between one criterion and the value it stands for.	Concerns a single criterion. D5 concerns weighting across criteria.	Construct validity [45]; Goodhart effects [40–43]; abstraction [6]; values hierarchy [18].	Re-specifies. Asks about validity under strong optimisation and for criteria implemented by learned evaluators.
D5. Justification of trade-offs	Trade-offs: who set weights, thresholds and the choice from the Pareto front, and why.	Presupposes at least two admitted criteria. D1 concerns the objectives themselves.	Values hierarchy [18]; value-focused thinking [25]; incompatibility of fairness criteria [26–28].	Re-specifies. Locates trade-off decisions in tool configuration and defaults.
D6. AI involvement in formulating and evaluating the task	Provenance: which task elements are proposed or implemented by AI.	Process dimension; applies on top of D1–D5 to each AI-proposed element.	Language models in requirements engineering [34]; AI co-authoring [46,47]; evaluator models [39]; evidence on reframing [48].	New. No counterpart in the frameworks examined (Table 2).
D7. Approval and traceability	Accountability: who approved the task and whether the path from value to criterion can be traced.	Concerns authorisation and traceability, not the completeness of content (D2).	IEEE 7000 [31]; requirements traceability [19]; responsibility for computationally designed products [53].	Consolidates. Extends traceability to AI provenance and tool configuration.

7.5. Using PFA in the Design Process

The dimensions are neither strictly sequential nor independent, and Figure 2 shows the intended use. Dimensions D1–D5 are addressed iteratively while the task is being formed, because problem and solution co-evolve [16]. A first objective (D1) leads to requirements and boundaries (D2, D3), criteria (D4) and weights (D5), and exploratory results may send the team back to any of them. Dimension D6 runs in parallel, as a provenance check on every task element, whichever dimension it falls under. Dimension D7 acts as a gate: the assessment is completed, and the task approved, before the AI system is first run. It is repeated when the objective, requirements or criteria change significantly, and the earlier approval is then superseded.

In a project that already has requirements elicitation, validation and risk analysis, PFA is performed over their records. Elicitation supplies most of the evidence for D1–D3, validation supplies parts of D2 and D7, and the documentation required by AI risk management frameworks supplies the intended purpose and context [9]. What PFA adds are the questions in Table 3 that these records do not normally answer: the justification of the stakeholder circle, proxy validity, weights and thresholds, and AI provenance. The result for each dimension is one of three levels (documented and justified; documented but not justified; undocumented), with the evidence cited and without an aggregate score. A worked application is given in Section 8.2.

7.6. Benefits, Costs and Proportionality

Compared with the use of system-level frameworks alone, PFA offers three benefits. First, it makes the formulation of the task an object of assessment before the system is run,

so that omissions cannot pass as successful tests (Section 5). Second, it states evidence requirements that can be audited: records of the objective, the point of control of each requirement, the population represented, the operationalisation of each criterion, weights and thresholds, AI provenance and approval. Third, it covers AI involvement in the formulation stage, which is not addressed in the frameworks compared. Compared with requirements validation alone, it adds the three extensions described in Section 7.4. Section 8.3 compares how the frameworks handle the examples. We do not claim that PFA is superior to these frameworks. Its claimed benefit is complementarity at a point that they leave open, and this benefit is hypothetical until it is tested (Section 8.4).

Any additional assessment adds effort, and three measures keep it proportionate. First, PFA reuses existing documentation (Section 7.5), so that the main new work consists in answering questions for which no record exists. Second, the rigour can be scaled to the openness of the task. A short screening asks three questions: Is the objective chosen by the team or fixed by a standard? Does the task affect people who are not represented in the requirements? Does AI propose or implement any element of the task? If the objective is fixed by a standard, the affected people are all represented and AI plays no part, the record may reduce to the approval (D7). Third, the effort is best spent on the dimensions in which an error does not trigger a failed test (D2–D4). We have not measured the effort required; it is part of the validation agenda (Section 8.4).

8. Testing the Argument

8.1. Examples

The first two examples correspond to the roles of solver and co-author of the task. The third concerns evaluation criteria, that is, the part of the task that an AI model could implement in the evaluator role. Table 4 shows what applying the PFA model to them reveals. A fourth, hypothetical example is worked through step by step in Section 8.2. None of the examples constitutes an assessment of the quality of the projects described.

Table 4. Application of the PFA model to the three examples. The entries indicate questions and observations, not assessments of the quality of the projects described. Source: own elaboration.

Dimension	Project Discover (Solver)	Pro-Explora (Co-Author of the Task)	Passive Safety Tests (Evaluation Criteria; Thought Experiment)
D1	Goals set by the team on the basis of an analysis of project goals and discussions with managers and workers–documented	The problem definition is generated by the system; a justification independent of the model is needed	Objective “improve performance in the existing tests”–is this the right objective for the design?
D2	The account does not indicate where accessibility, evacuation and the needs of facilities staff were addressed	Functional requirements and constraints are proposed by an agent; completeness requires review	The tests do not cover some scenarios, e.g., the average female in rear impact tests
D3	Preferences collected through employee surveys; no information on other building users	The list of users and stakeholders is generated by an agent	Choice of dummy percentile as a decision about whom the task represents

Table 4. *Cont.*

Dimension	Project Discover (Solver)	Pro-Explora (Co-Author of the Task)	Passive Safety Tests (Evaluation Criteria; Thought Experiment)
D4	Six metrics as the operationalisation of goals, e.g., productivity measured via sight lines at desks–the meaning of the measure requires justification	Target values in the specification require justification	Test result as a proxy for the protection of the population
D5	Selection from the Pareto front, analysis by the designer and discussion with stakeholders–described	Not described in the available evaluation	Trade-offs between the protection of different groups are invisible if one group is not evaluated
D6	Not applicable: AI does not take part in formulating the task	Key dimension: the task is co-created by the model	Not applicable in the experiment; an evaluator model trained on such tests would inherit their scope (hypothesis)
D7	Decisions of the design team; details of approval not described	The designer is responsible for validating and taking ownership of the output (AICED framework assumption)	Criteria approved outside the design team (regulations, protocols)

Nagy and colleagues describe Project Discover, an application of generative design to the space planning of Autodesk's new office in Toronto [57]. The system used a multi-objective genetic algorithm which generated 10,000 designs over one hundred generations. No machine learning was used; the authors mention it only as a direction for future work [57]. The example therefore illustrates the broader class of automated generative search rather than AI in the narrow sense. The six goals (adjacency preference, work style preference, buzz, productivity, daylight and views to the outside) were defined by the design team on the basis of an analysis of the project goals and discussions with managers and workers, and the data for them were collected through surveys [57]. The authors also describe selecting designs from the Pareto front, their analysis by the designer and discussion with stakeholders [57].

From the perspective of the PFA model, the justification of the goals (D1) and of the trade-offs (D5) is relatively well documented in this account. Less is known about the boundaries of the problem and the completeness of the requirements (D2). The account does not say whether or how accessibility for people with disabilities, evacuation safety or the needs of facilities staff were taken into account, and it provides no grounds for claiming that they were omitted. They may have been handled outside the evaluation function, for instance as hard constraints or at another stage of the project. The model's question is therefore not "does the algorithm rank the alternatives well?" but "where were the requirements that are absent from the six criteria addressed?" This is not a criticism of Project Discover. The example shows that the logical independence of the two assessments is visible without any system failure or blatantly poor design.

The second example concerns AI as co-author of the task. Obieke, Bridgeman and Han describe the Pro-Explora tool, in which a multi-agent system based on language models generates a problem definition, a product design specification document and concept images [46]. The individual agents are assigned roles: one provides the problem context, design constraints and functional requirements, another advises on manufacturing and

sustainability, and yet another identifies users and stakeholders [46]. The designer may perform a task personally or delegate it to the agents, while remaining responsible for validating and taking ownership of the output [46]. The evaluation was a pilot: it involved a single design task and the opinions of two experienced engineers [46].

The example is important because elements corresponding to dimensions D2 and D3, that is, the requirements and the circle of stakeholders, are produced here as proposals of the system. The PFA model does not assess whether Pro-Explora does this well; the available evaluation does not settle that. It does, however, indicate what questions should be asked before a generated specification becomes the basis of a design. Which requirements and which stakeholder groups were proposed by the system (D6)? Do the target values in the specification have a justification independent of the model (D4)? Who reviewed the list of stakeholders for omitted groups (D3), and who approved the specification (D7)? The findings of Shin and colleagues [48] suggest that answering these questions is particularly important when the tool is used by a less experienced designer.

The third example concerns a problem that predates contemporary AI: evaluation criteria in which not all user groups are represented. Linder and Svedberg analysed European regulations on passive safety assessment as they stood in 2019 and showed that adult occupants are represented in them by the average-sized male, while the average-sized female is not included [51]. This does not mean that female dummies are absent altogether. UN Regulation No. 137 provides for a Hybrid III 5th percentile female dummy in the passenger seat [58], and the Euro NCAP protocol applicable from 2026 uses such a dummy in selected frontal impact tests, including in the driver's seat [59]. A 5th percentile dummy, however, represents small women rather than the average woman. In whiplash protection tests, Euro NCAP still uses a BioRID dummy corresponding to the 50th percentile male [60], while the average female dummy (BioRID P50F) remains a prototype. In a 2026 study, NIC values in the protraction phase for the female dummy in the same seat were almost twice those for the male dummy [52].

Using US crash data from 1998–2008, Bose, Segui-Gomez and Crandall found that in comparable crashes the odds of severe injury for a belt-restrained female driver were 47% higher than for a belt-restrained male driver [61]. The study does not establish whether, or to what extent, the difference results from how the body is represented in tests. Moreover, the two works concern different jurisdictions. Taken together, they therefore do not warrant a causal conclusion, but only the weaker claim that the choice of a representative user is an element of safety requirements and may have consequences for less well represented groups.

This example shows that the decision about user representation is not binary. Including some female dummy does not settle which part of the population the task represents. The choice of percentile is a normative choice that belongs to dimension D3.

Let us consider a thought experiment on this basis. A hypothetical generative system is to optimise the geometry of selected passive safety components against a set of tests in which one group of users is represented much better than another. Assume further that the system meets all technical requirements concerning numerical stability, repeatability and constraint handling. If the objective is to improve performance in the existing tests, the system may perform the task correctly. This correctness does not, however, settle whether the tests adequately represent the population that the design is meant to protect.

The example does not suggest that the safety differences described are related to AI. The studies cited do not concern AI and do not examine any such relationship. What the thought experiment shows is that an error in user representation may lie upstream of optimisation. In such a case, a more accurate system does not remove the error, because the success criterion was defined on its basis. What is needed is a change in how the problem is formulated, not merely a better algorithm.

The case also shows a limitation of the vocabulary of fairness when it is used only to assess model outputs. If a design were evaluated for several representative groups, differences in outcomes between them could be examined. If, however, one group is absent from the evaluation conditions, the problem arises earlier. The first question to ask is who should be represented in the design process.

A system may operate correctly with respect to the adopted objective and requirements while an ethical problem arises from what the objective and requirements left out. It is precisely this possibility that the assessment of responsible AI use in design must cover.

8.2. Worked Example: Step-by-Step Application of the PFA Model

To show how the dimensions are applied in sequence, we use a hypothetical case. It was constructed by the authors for illustration and does not describe a real project; the findings in Table 5 are the questions that the model would raise, not the results of a study. A municipality commissions a pedestrian footbridge over a river. The design team uses a generative design tool to search for lightweight deck and truss layouts, and an LLM assistant drafts the project brief from the tender text and from earlier projects.

Table 5. Step-by-step application of the PFA model to a hypothetical footbridge project. The findings are illustrative and constructed by the authors. Source: own elaboration.

Dimension	PFA Question	Finding in the Draft Brief	Action Taken and Evidence	Level (Before → After)
D1	Why was mass minimisation adopted as the objective?	Objective proposed by the LLM assistant; no alternatives recorded.	Alternatives (whole-life cost, maintainability) recorded; mass kept as a proxy for material cost and embodied carbon, with reasons.	Undocumented → documented and justified
D2	What lies outside the problem boundary, and where is each missing requirement controlled?	Static load, deflection, wind and fatigue listed; lateral crowd response, ice on the deck, accessible gradients and maintenance clearance absent.	Lateral dynamic response added as a separate verification [35]; gradients and clearance added as constraints; slip resistance verified separately.	Undocumented → documented and justified
D3	Whose needs and use situations does the task represent?	Stakeholder list names the municipality and pedestrians in general.	Users with visual impairments, wheelchair users, maintenance staff and cyclists added after review by an accessibility officer; their scenarios added to the load cases.	Documented but not justified → documented and justified
D4	Does each criterion preserve the meaning of the value it represents, also under strong optimisation?	Comfort represented only by a vertical acceleration limit.	Lateral acceleration limit added; the record states what the comfort measures do not capture.	Documented but not justified → documented and justified
D5	Who set the weights, thresholds and the choice from the Pareto front, and why?	Default tool weights; selection not recorded.	Weights and thresholds set by the team and recorded; selection made by the chief engineer after consultation, with reasons.	Undocumented → documented and justified

Table 5. *Cont.*

Dimension	PFA Question	Finding in the Draft Brief	Action Taken and Evidence	Level (Before → After)
D6	Which task elements were proposed or are implemented by AI?	Objective, stakeholder list and first requirements list proposed by the assistant; a learned scorer ranks concepts for aesthetic fit.	Provenance of each element logged; scorer used only after constraints are met and checked against expert ratings on a sample.	Undocumented → documented and justified
D7	Who approved the task, and can the path from value to criterion be traced?	No named approver.	Project lead signs the specification and the list of changes to the AI draft before the first run; value–requirement–criterion table kept; D1–D6 re-checked when the stiffness limit changes.	Undocumented → documented and justified

The assessment starts from the draft brief (Table 5). The assistant proposes the objective "minimise structural mass subject to static load and deflection limits" (D1) with no record of alternatives. The list of requirements covers static load, deflection, wind load and fatigue but not lateral response under crowds, ice on the deck, accessible gradients or maintenance clearance (D2), and the stakeholder list names the municipality and pedestrians in general (D3). Comfort appears only as a vertical acceleration limit (D4), the tool's default weights are used and the selection from the Pareto front is not recorded (D5). The objective, the stakeholder list and the first list of requirements were proposed by the assistant, and a learned scorer for aesthetic fit is used to rank concepts (D6). No one is named as approver (D7).

The team then acts on each finding. It records the alternatives to the objective and the reason for keeping mass as a proxy for material cost and embodied carbon. It adds lateral dynamic response under crowd loading as a separate verification, drawing on the lesson of the Millennium Footbridge [35], and adds gradients and maintenance clearance as constraints. A review by an accessibility officer adds users with visual impairments, wheelchair users, maintenance staff and cyclists, and their scenarios enter the load cases. The team adds a lateral acceleration limit and records what the comfort measures do not capture. It records who set the weights and thresholds, and why the chief engineer chose a particular design from the Pareto front. It logs the provenance of each task element and restricts the learned scorer to ranking concepts that already satisfy the constraints, with a sample check against expert ratings. The project lead signs the task specification, together with the list of changes to the AI draft, before the first optimisation run. When the first results show a very slender deck and the stiffness limit is changed, D1–D6 are re-checked for the changed elements.

The example shows the sequence of the dimensions, the iteration loop and the kind of evidence that each dimension calls for. It does not show that PFA detects problems that other procedures miss. That claim needs the empirical test described in Section 9.

8.3. How Existing Frameworks Would Handle the Examples

To show where PFA complements the frameworks compared in Table 2, Table 6 sets out how each of them would typically be applied to the four examples. The entries are our reading of the published descriptions of the frameworks. They indicate what each frame-

work would prompt and where the questions of PFA remain open; they are not an empirical test of the frameworks, and they do not describe what happened in the projects concerned.

Table 6. Typical handling of the examples by existing frameworks and the questions that remain open. Own reading of published descriptions; not an empirical test. Source: own elaboration.

Example	NIST AI RMF	IEEE 7000 and Value Sensitive Design	Requirements Engineering (Validation, GORE)	Questions Left Open (PFA Dimension)
Project Discover (solver)	Prompts documentation of intended purpose, users and context (MAP 1.1, 1.6) and of organisational goals. Construct validation is mentioned (MAP 2.3) but not linked to the objective function.	Strong on eliciting the values of workers and managers (cf. the surveys) and on indirect stakeholders such as visitors and facilities staff. Weaker on how a value such as productivity becomes a metric.	Validation would check the six goals against stakeholder needs; completeness checks could find accessibility or evacuation if a checklist lists them. Tool-level weights and Pareto selection are outside the usual documents.	D4 (meaning of the metrics, e.g., productivity); D2 (where unlisted requirements are controlled).
Pro-Explora (co-author of the task)	Applies to the AI system and its documentation; does not by itself ask which requirements or stakeholders were proposed by the model (D6 coded N).	Could be applied to the generated specification as an input to value elicitation; does not ask which values or stakeholders came from the model.	Reviews, inspections and traceability apply to a generated specification as to any other, which is a mature strength. They do not ask for provenance or for the model's tendency to omit.	D6 (provenance); D3 (who reviews the generated stakeholder list); D7 (approval).
Passive safety tests (evaluation criteria)	Impact on groups and diversity of AI actors are covered (MAP 1.1–1.2), but test protocols are external inputs, not system properties.	The attention of Value Sensitive Design to direct and indirect stakeholders and to value tensions addresses user representation directly; the best fit among the frameworks compared.	Requirements come from regulation and protocols; validation against stakeholder needs tends to take the protocol as given.	D4 (test result as a proxy for the protection of the population); D5; D7 (criteria approved outside the team).
Footbridge (hypothetical)	Prompts documentation of purpose, context and impacts on affected groups; does not prompt a completeness check of an LLM-drafted requirements list.	Would elicit accessibility and safety early as values and derive requirements; translating comfort into a criterion is left to the team.	Domain checklists, codes and validation would catch many omissions, but cannot reveal loading effects that lie outside the assumed environment [21,35].	D4 under strong optimisation; D5 (default weights); D6 (provenance).

In all four examples the existing frameworks supply a starting point (stakeholder identification, value elicitation, validation of requirements) that PFA reuses. The questions that remain open are the same ones that Table 2 marks as partly addressed or absent: proxy validity, trade-offs in tool configuration and AI provenance.

8.4. Objections

The most serious objection concerns novelty. Passi and Barocas analyse the relationship between problem formulation and fairness [5], Selbst and colleagues describe errors resulting from inappropriate abstraction [6], and Dobbe, Gilbert and Mintz argue that the vagueness surrounding AI safety cannot be resolved by formal means alone [12]. In design research, Menold and colleagues state that in design with AI, designers remain responsible for setting objectives and resolving trade-offs, and that automated generation of alternatives should reflect stakeholder priorities rather than optimise narrow proxies [24]. We agree that these works recognise the problem. The article does not compete with them in recognising it. It adds three elements that we did not find in them: the link to the distinction between verification and validation, the typology of AI roles, and an assessment model with a dimension for AI involvement in task formulation. Section 7.4 and Table 3 state which dimensions consolidate existing concepts and which re-specify or add to them.

A second objection holds that this stage is already covered by Value Sensitive Design, IEEE 7000 and AI risk management frameworks. The NIST AI RMF requires documenting the intended purposes of the system, the context of its use and system requirements and, after completing the MAP function, an initial go/no-go decision on whether to design, develop or deploy the system [9]. The EU AI Act uses the concept of the intended purpose of a system, and it ties human oversight of high-risk systems to the period in which the system is in use [62]. Table 2 shows that the objection is valid with respect to stakeholders and traceability, and weaker with respect to proxy validity and AI involvement in task formulation. For this reason, the model draws on these frameworks rather than proposing a competing procedure. Section 8.3 compares how these frameworks would handle the examples.

A third objection comes from systems engineering: the distinction between a correct solution and a properly formulated problem is simply verification and validation. We accept this genealogy (Section 4.4). Validation, however, takes the circle of stakeholders and the meaning of the criteria as given, and requirements engineering for machine learning concerns mainly the requirements for the AI system as a product, not situations in which AI co-creates requirements for a technical artefact. The PFA model extends validation precisely at these two points. Section 7.4 sets out the boundary between PFA and requirements validation.

Another objection concerns the specificity of the problem. Engineers were responsible for objectives and requirements long before machine learning appeared. Faulty requirements can lead to poor designs without AI as well. One might therefore conclude that we are dealing with classical engineering ethics and that the connection with AI is incidental.

The first part of the objection is true. The problem is older than AI, which is precisely why it should not be presented as a new moral harm that emerged with generative design. The connection with AI is nonetheless significant for the reasons given in Section 6. When a system evaluates alternatives or takes part in producing the task description, control over how the problem is formulated becomes part of control over the use of AI itself. The mechanism described for the solver role does apply to optimisation and automated decision systems in general. The evaluator and co-author roles, however, are characteristic of contemporary AI and differ from earlier tools in the way the task is translated and the criteria are implemented (Section 6.4), and for the rest the argument does not require specificity. It requires only that the assessment of AI use cover this mechanism.

It may also be argued that the problem is exaggerated, because engineering does not operate in a normative vacuum. Designs are subject to regulations, standards, safety procedures, certification and professional liability. In many fields, these requirements largely determine how the problem is formulated.

This is an important limitation of the thesis. The more strongly a field is regulated and the more unambiguous its technical and functional requirements, the smaller the room for discretion in formulating the task. It does not follow, however, that this room disappears altogether. Standards usually set minimum levels, limits of admissibility or verification procedures. They do not determine all the objectives of a project or all the trade-offs between cost, durability, accessibility, environmental impact and other properties.

The Statement of Ethical Principles of the UK Engineering Council and the Royal Academy of Engineering likewise emphasises responsibility for social and environmental consequences and the need for sound professional judgement [63]. Compliance with a minimum technical requirement is thus not always a sufficient justification for a design choice. The same holds for AI. Compliance of the system and the solution with standards is necessary, but it does not replace the assessment of whether the problem has been formulated adequately to the purpose of the project and the groups affected by the solution.

A further objection holds that an additional layer of assessment adds effort that tightly constrained projects cannot justify. Section 7.6 answers it in two steps. PFA reuses existing documentation, and a three-question screening scales the rigour to the openness of the task, so that for a task fixed by standards the record may reduce to the approval. We have not measured the effort involved, and we list this as an open question.

The last objection concerns the status of the model. PFA has not been validated empirically; its dimensions were derived from the literature analysis and illustrated with published and hypothetical examples. Why, then, is it proposed as an assessment model? The conceptual argument, that the formulation stage is a distinct object of assessment, does not depend on the empirical performance of the model, and a structured set of questions is a precondition for studying that performance. We therefore present PFA as a conceptual model. We do not claim that it predicts failures, that design teams can apply it reliably, or that it detects problems that other procedures miss. The validation agenda has four parts: (i) independent coding, in which two or more coders apply the rubric of Appendix B to the same frameworks and documented projects, with inter-rater agreement reported (for example, Cohen's kappa or Krippendorff's alpha); (ii) retrospective application to documented AI-assisted design projects, comparing the problems found by PFA with those found by requirements validation and by AI risk management documentation; (iii) studies with design teams on feasibility and effort; and (iv) revision of the dimensions, including the merging of those that coders cannot separate.

Having considered these objections, we state the thesis more cautiously. We do not claim that contemporary AI ethics ignores the problem of task formulation, that AI governance frameworks neglect the stage at which the purpose of a system is determined, that engineering lacks methods for identifying values, or that AI creates a new type of responsibility. We claim that the assessment of AI in design is incomplete if it begins only after the objective and requirements of the task have been set and treats them merely as inputs. Correct operation of the system with respect to these assumptions does not justify the assumptions themselves. The formulation of the problem should therefore belong to the object of assessment of responsible AI use in engineering.

9. Conclusions

The starting point of the article was the question of whether the ethical assessment of AI in engineering design can be limited to checking how the system performs the task it has been given. The analysis of three bodies of literature leads to a negative answer, which nonetheless requires qualification. The AI ethics literature does not entirely overlook the problem of task formulation entirely, work on fairness, abstraction traps, Ethics by Design and Value Sensitive AI argues that the choice of objectives, representations and criteria may

itself be ethically significant, and AI risk management frameworks require the intended purpose of a system to be documented. In the analysed work on AI in engineering, however, this problem is not linked with design theory. The ideas of Simon, Schön, and Rittel and Webber, together with analyses of the translation of values into requirements, show in turn that a design task is not simply given. It is created through the choice of a preferred state, the specification of requirements and a decision on how trade-offs will be assessed.

Combining these findings makes it possible to distinguish a correct solution to a task from a properly formulated problem. The first assessment concerns the relation between the system's output and the adopted requirements. The second concerns the reasons behind the objective and the requirements. An AI system may satisfy the first criterion without settling the second. The examples of generative space planning, a multi-agent system generating a product specification, user representation in vehicle safety tests and a hypothetical footbridge project illustrate how a problem can arise before optimisation begins.

This distinction has its genealogy in the classic verification–validation pair. Validation, however, takes the circle of stakeholders and the meaning of the criteria as given, while contemporary AI increasingly takes part in the stage at which these elements are created. The solver role is not specific to AI. The evaluator and co-author roles are characteristic of contemporary AI because they move into the tool the translation of needs into a formal task and the implementation of criteria (Section 6.4). They justify including the assessment of problem formulation in the assessment of AI use. The empirical support for this claim is so far limited to pilot evaluations and one experiment on problem reframing [46,48].

The contribution of the article does not consist in discovering the ethical significance of problem formulation. It consists of three elements: linking this theme with design theory and with requirements validation, a typology of AI roles, and the PFA model, which specifies what should be assessed and with what evidence the assessment should be documented. The novelty of the model is limited. One dimension (D6, AI involvement) has no counterpart in the frameworks examined, three (D2, D4, D5) re-specify existing concepts for AI-assisted design, and three (D1, D3, D7) largely consolidate them (Table 3). In our coding, existing frameworks address the formulation stage partly, and AI involvement in formulating and evaluating the task not at all (Table 2). Table 7 summarises what each conclusion rests on and how far the evidence goes.

Table 7. Conclusions, supporting evidence and limits of the support. Source: own elaboration.

Conclusion	Supported by	Limits of the Support
Ethical assessment of AI in engineering design cannot be limited to the correctness of the execution of a given task.	Sections 4 and 5 (design theory, requirements engineering, verification and validation); the Millennium Footbridge case; the examples in Section 8.1.	Conceptual argument. The examples illustrate it and do not test it.
Problem formulation is recognised in AI ethics but is not integrated in the frameworks for AI in engineering.	Sections 3.2 and 3.3; the targeted search in Section 2.	Limited to the corpus analysed (Crossref, title screening, a single screener).
The evaluator and co-author roles change the conditions of assessment.	Sections 6 and 6.4; [39,45,46,48].	Evidence is partial: pilot evaluations and one experiment on reframing. The differences from earlier tools are of degree and of the kind of translation.
Existing frameworks cover the formulation stage only partly, and do not cover AI involvement in it.	Table 2 and Appendix B.	Authors' coding without a second coder; one cell (IEEE 7000, D4) undetermined.

Table 7. *Cont.*

Conclusion	Supported by	Limits of the Support
PFA organises the assessment of problem formulation.	Sections 7 and 8; Tables 3–6.	Conceptual. Published and hypothetical examples only; usefulness, reliability and effort are untested.

The argument is conceptual and relies on purposively selected literature rather than a systematic review. It applies most strongly to tasks in which the objective and criteria remain open, and less strongly in fields where requirements are largely determined by standards and regulations. The coding in Table 2 is the authors' own and was not checked by an independent coder, and the examples illustrate the model rather than test it. The results therefore do not show that PFA is reliable, that it requires little effort, or that it reveals problems that requirements validation and AI risk management miss. The next steps are those set out in Section 8.4: independent coding of Table 2 and of documented projects with inter-rater agreement reported, retrospective application to documented AI-assisted projects, studies with design teams on feasibility and effort, and revision of the dimensions that coders cannot separate.

Author Contributions: Conceptualization, M.M.; Methodology, M.M.; Formal analysis, M.M. and N.K.; Resources, N.K.; Writing – original draft, M.M.; Writing – review & editing, N.K.; Supervision, M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article (including Appendices A and B). Further inquiries can be directed to the corresponding author(s).

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (GPT-4.5, OpenAI, San Francisco, CA, USA) and DeepL Translator (DeepL SE, web version, Cologne, Germany) for translation support, language editing, and academic style refinement. The authors have reviewed and edited all AI-assisted outputs and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Literature Search

Table A1 documents the supplementary search described in Section 2. For each query, the first 25 results by relevance were screened by title against the inclusion criteria given in Section 2. Some of the works judged relevant were used directly in the analysis, in particular work on requirements engineering for machine learning and on the use of large language models for requirements engineering tasks [32,34].

Table A1. Queries and results of the supplementary search.

Query (Crossref, 2016–2026, 23 September 2026)	Number of Results	Screened	Judged Relevant
"AI ethics engineering design"	5,065,411	25	4
"problem formulation machine learning ethics"	2,735,425	25	1
"problem framing generative AI design"	2,220,270	25	1
"requirements engineering machine learning"	5,372,683	25	10
"value sensitive design artificial intelligence requirements"	2,888,415	25	6
"large language models requirements engineering design specification"	6,102,660	25	7
"objective misspecification proxy optimization AI"	1,227,253	25	1
Total	–	175	30

The large numbers of results arise because Crossref matches any word of the query and orders results by relevance; what matters is therefore the ranking, not the number of results.

Appendix B. Coding Rubric and Evidence for Table 2

Table A2 gives the rubric used to code Table 2, and Table A3 gives the evidence and the reasons for the coding, framework by framework. For each framework, the text examined was read for provisions that match the guiding question of each dimension (Table 1); the provision was recorded with its location and a code was assigned according to the rubric. The coding was carried out by the authors. No independent second coder was used, and therefore no inter-rater agreement can be reported. This is a limitation, and independent recoding is the first step of the validation agenda (Section 8.4). The evidence is given so that the coding can be checked and challenged.

Table A2. Coding rubric.

Code	Criterion
E	The text names the question or decision and calls for a record, evidence or procedure concerning it in the context of formulating the task or requirements.
P	The text addresses the topic indirectly or in general terms (for example, documentation of purpose, or a general call for valid measurement) without specifying the question or the evidence at the stage of task formulation.
N	No provision relating to the dimension was found in the text examined.
?	The relevant text was not accessible (standard available only for a fee); the cell is not coded.

Table A3. Evidence and reasons for the coding in Table 2.

Framework	Source Examined	Evidence and Coding Decisions
NIST AI RMF 1.0 [9]	AI 100-1 core (GOVERN, MAP, MEASURE, MANAGE), in particular subcategories MAP 1.1–1.6 and 2.1–2.3.	D1 P: MAP 1.3–1.4 require goals and business value to be documented, and a go/no-go decision follows the MAP function, but there is no requirement to justify the choice of an objective among admissible ones. D2 P: MAP 1.1 (assumptions and limitations), MAP 1.6 (system requirements elicited) and MAP 2.1 (task defined); no requirement to state where each requirement is controlled. D3 E: MAP 1.1 (types of users, their expectations, impacts on individuals and communities) and MAP 1.2 (diverse AI actors). D4 P: MAP 2.3 names construct validation among TEVV considerations, not in relation to the objective function of a design tool. D5 P: trade-offs among trustworthiness characteristics are discussed, but not weights or selection in design tools. D6 N: no provision on AI-proposed task elements. D7 P: documentation and accountability, but no traceability from value to criterion. Assessment II E: MEASURE function.
IEEE 7000-2021 [31]	Description of scope and table of contents (standard available only for a fee): concept of operations and context exploration, ethical values elicitation and prioritisation, ethical requirements definition, ethical risk-based design, transparency management.	D1 P: values are elicited at the concept exploration stage. D2 P. D3 E: elicitation of stakeholders and values. D4 ?: not determined from the public description whether the validity of operationalised criteria is addressed. D5 P: value prioritisation. D6 N. D7 E: traceability from values to requirements. Assessment II P: ethical risk-based design of the system.
Value Sensitive Design [30]	Friedman and Hendry (2019): conceptual, empirical and technical investigations; direct and indirect stakeholders; value tensions.	D1 P. D2 P. D3 E: direct and indirect stakeholders. D4 N. D5 P: value tensions. D6 N. D7 N. Assessment II N.
Requirements engineering [19,23]	ISO/IEC/IEEE 29148 (stakeholder needs, requirements validation, traceability) and goal-oriented requirements engineering (goals as the rationale for requirements).	D1 P: goals as rationale for requirements. D2 E: validation and completeness. D3 E: stakeholder needs. D4 P: verifiability of requirements. D5 P: requirements negotiation. D6 N. D7 E: requirements traceability. Assessment II P.
RAI framework in structural engineering [4]	Plevris and Hosamo (2025): technical requirements for models, oversight, professional responsibility, risk management.	D1 N. D2 N. D3 P. D4 N. D5 N. D6 N. D7 P: professional responsibility and oversight. Assessment II E.

References

1. Mittelstadt, B.D.; Allo, P.; Taddeo, M.; Wachter, S.; Floridi, L. The ethics of algorithms: Mapping the debate. *Big Data Soc.* **2016**, *3*, 2053951716679679. [CrossRef]
2. Jobin, A.; Ienca, M.; Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
3. Liang, C.-J.; Le, T.-H.; Ham, Y.; Mantha, B.R.K.; Cheng, M.H.; Lin, J.J. Ethics of artificial intelligence and robotics in the architecture, engineering, and construction industry. *Autom. Constr.* **2024**, *162*, 105369. [CrossRef]

4. Plevris, V.; Hosamo, H. Responsible AI in structural engineering: A framework for ethical use. *Front. Built Environ.* **2025**, *11*, 1612575. [CrossRef]
5. Passi, S.; Barocas, S. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019*; ACM: New York, NY, USA, 2019; pp. 39–48. [CrossRef]
6. Selbst, A.D.; Boyd, D.; Friedler, S.A.; Venkatasubramanian, S.; Vertesi, J. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019*; ACM: New York, NY, USA, 2019; pp. 59–68. [CrossRef]
7. Brey, P.; Dainow, B. Ethics by design for artificial intelligence. *AI Ethics* **2024**, *4*, 1265–1277. [CrossRef]
8. Sadek, M.; Calvo, R.A.; Mougenot, C. Designing value-sensitive AI: A critical review and recommendations for socio-technical design processes. *AI Ethics* **2024**, *4*, 949–967. [CrossRef]
9. Tabassi, E. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*; NIST AI 100-1; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023. [CrossRef]
10. Van de Poel, I. Embedding values in artificial intelligence (AI) systems. *Minds Mach.* **2020**, *30*, 385–409. [CrossRef]
11. Kudina, O.; van de Poel, I. A sociotechnical system perspective on AI. *Minds Mach.* **2024**, *34*, 21. [CrossRef]
12. Dobbe, R.; Krendl Gilbert, T.; Mintz, Y. Hard choices in artificial intelligence. *Artif. Intell.* **2021**, *300*, 103555. [CrossRef]
13. Abebe, R.; Barocas, S.; Kleinberg, J.; Levy, K.; Raghavan, M.; Robinson, D.G. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020*; ACM: New York, NY, USA, 2020; pp. 252–260. [CrossRef]
14. Simon, H.A. *The Sciences of the Artificial*, 3rd ed.; MIT Press: Cambridge, MA, USA, 1996.
15. Schön, D.A. *The Reflective Practitioner: How Professionals Think in Action*; Basic Books: New York, NY, USA, 1983.
16. Dorst, K.; Cross, N. Creativity in the design process: Co-evolution of problem–solution. *Des. Stud.* **2001**, *22*, 425–437. [CrossRef]
17. Rittel, H.W.J.; Webber, M.M. Dilemmas in a general theory of planning. *Policy Sci.* **1973**, *4*, 155–169. [CrossRef]
18. Van de Poel, I. Translating values into design requirements. In *Philosophy and Engineering: Reflections on Practice, Principles and Process*; Michelfelder, D.P., McCarthy, N., Goldberg, D.E., Eds.; Springer: Dordrecht, The Netherlands, 2013; pp. 253–266. [CrossRef]
19. *ISO/IEC/IEEE 29148:2018*; Systems and Software Engineering—Life Cycle Processes—Requirements Engineering, 2nd ed. ISO: Geneva, Switzerland, 2018.
20. Nuseibeh, B.; Easterbrook, S. Requirements engineering: A roadmap. In *Proceedings of the Conference on the Future of Software Engineering (ICSE 2000), Limerick, Ireland, 4–11 June 2000*; ACM: New York, NY, USA, 2000; pp. 35–46. [CrossRef]
21. Zave, P.; Jackson, M. Four dark corners of requirements engineering. *ACM Trans. Softw. Eng. Methodol.* **1997**, *6*, 1–30. [CrossRef]
22. Boehm, B.W. Verifying and validating software requirements and design specifications. *IEEE Softw.* **1984**, *1*, 75–88. [CrossRef]
23. van Lamsweerde, A. Goal-oriented requirements engineering: A guided tour. In *Proceedings of the Fifth IEEE International Symposium on Requirements Engineering (RE'01), Toronto, ON, Canada, 27–31 August 2001*; IEEE Computer Society: Los Alamitos, CA, USA, 2001; pp. 249–262. [CrossRef]
24. Menold, J.; McComb, C.; Tucker, C.; Papalambros, P. Greater than the sum of our parts: The new division of labor in design with AI. *PNAS Nexus* **2026**, *5*, pgag208. [CrossRef] [PubMed]
25. Keeney, R.L. *Value-Focused Thinking: A Path to Creative Decisionmaking*; Harvard University Press: Cambridge, MA, USA, 1992.
26. Kleinberg, J.; Mullainathan, S.; Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), Berkeley, CA, USA, 9–11 January 2017*; LIPIcs 67; Schloss Dagstuhl–Leibniz-Zentrum für Informatik: Dagstuhl, Germany, 2017; pp. 43:1–43:23. [CrossRef]
27. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* **2017**, *5*, 153–163. [CrossRef] [PubMed]
28. Friedler, S.A.; Scheidegger, C.; Venkatasubramanian, S. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM* **2021**, *64*, 136–143. [CrossRef]
29. Bagdasaryan, E.; Poursaeed, O.; Shmatikov, V. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 15479–15488.
30. Friedman, B.; Hendry, D.G. *Value Sensitive Design: Shaping Technology with Moral Imagination*; MIT Press: Cambridge, MA, USA, 2019. [CrossRef]
31. *IEEE Std 7000-2021*; IEEE Standard Model Process for Addressing Ethical Concerns During System Design. IEEE: New York, NY, USA, 2021. [CrossRef]
32. Vogelsang, A.; Borg, M. Requirements engineering for machine learning: Perspectives from data scientists. In Proceedings of the 2019 IEEE 27th International Requirements Engineering Conference Workshops (REW), Jeju Island, Republic of Korea, 23–27 September 2019; pp. 245–251. [CrossRef]
33. Ahmad, K.; Abdelrazek, M.; Arora, C.; Bano, M.; Grundy, J. Requirements engineering for artificial intelligence systems: A systematic mapping study. *Inf. Softw. Technol.* **2023**, *158*, 107176. [CrossRef]

34. Norheim, J.J.; Rebentisch, E.; Xiao, D.; Draeger, L.; Kerbrat, A.; de Weck, O.L. Challenges in applying large language models to requirements engineering tasks. *Des. Sci.* **2024**, *10*, e16. [CrossRef]
35. Dallard, P.; Fitzpatrick, A.J.; Flint, A.; Le Bourva, S.; Low, A.; Ridsdill Smith, R.M.; Willford, M. The London Millennium Footbridge. *Struct. Eng.* **2001**, *79*, 17–33. [CrossRef]
36. Regenwetter, L.; Nobari, A.H.; Ahmed, F. Deep generative models in engineering design: A review. *J. Mech. Des.* **2022**, *144*, 071704. [CrossRef]
37. Woldseth, R.V.; Aage, N.; Bærentzen, J.A.; Sigmund, O. On the use of artificial neural networks in topology optimisation. *Struct. Multidiscip. Optim.* **2022**, *65*, 294. [CrossRef]
38. Makatura, L.; Foshey, M.; Wang, B.; Hähnlein, F.; Ma, P.; Deng, B.; Tjandrasuwita, M.; Spielberg, A.; Owens, C.E.; Chen, P.Y.; et al. How can large language models help humans in design and manufacturing? Part 1: Elements of the LLM-enabled computational design and manufacturing pipeline. *Harv. Data Sci. Rev.* **2024**. Special Issue 5. [CrossRef]
39. Picard, C.; Edwards, K.M.; Doris, A.C.; Man, B.; Giannone, G.; Alam, M.F.; Ahmed, F. From concept to manufacturing: Evaluating vision-language models for engineering design. *Artif. Intell. Rev.* **2025**, *58*, 288. [CrossRef]
40. Strathern, M. 'Improving ratings': Audit in the British University system. *Eur. Rev.* **1997**, *5*, 305–321. [CrossRef]
41. Manheim, D.; Garrabrant, S. Categorizing variants of Goodhart's law. *arXiv* **2018**, arXiv:1803.04585. [CrossRef]
42. Thomas, R.L.; Uminsky, D. Reliance on metrics is a fundamental challenge for AI. *Patterns* **2022**, *3*, 100476. [CrossRef] [PubMed]
43. Zhuang, S.; Hadfield-Menell, D. Consequences of misaligned AI. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; pp. 15763–15773.
44. Regenwetter, L.; Srivastava, A.; Gutfreund, D.; Ahmed, F. Beyond statistical similarity: Rethinking metrics for deep generative models in engineering design. *Comput.-Aided Des.* **2023**, *165*, 103609. [CrossRef]
45. Jacobs, A.Z.; Wallach, H. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT'21), Virtual Event, Canada, 3–10 March 2021*; ACM: New York, NY, USA, 2021; pp. 375–385. [CrossRef]
46. Obieke, C.C.; Bridgeman, J.; Han, J. A framework of AI collaboration in engineering design (AICED). *Proc. Des. Soc.* **2025**, *5*, 91–100. [CrossRef]
47. Clay, J.; Sha, Z. Paradigmatic design thinking: How generative AI changes the role of human designers. *Proc. Des. Soc.* **2025**, *5*, 2571–2579. [CrossRef]
48. Shin, J.; Polyanskaya, A.; Lucero, A.; Oulasvirta, A. No evidence for LLMs being useful in problem reframing. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 26 April–1 May 2025*; ACM: New York, NY, USA, 2025; p. 243. [CrossRef]
49. Tversky, A.; Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science* **1974**, *185*, 1124–1131. [CrossRef] [PubMed]
50. Parasuraman, R.; Manzey, D.H. Complacency and bias in human use of automation: An attentional integration. *Hum. Factors* **2010**, *52*, 381–410. [CrossRef] [PubMed]
51. Linder, A.; Svedberg, W. Review of average sized male and female occupant models in European regulatory safety assessment tests and European laws: Gaps and bridging suggestions. *Accid. Anal. Prev.* **2019**, *127*, 156–162. [CrossRef] [PubMed]
52. Carlsson, A.; Kullgren, A.; Svensson, M.Y. Rear impact responses of 50th percentile female and male dummies in a seat with different whiplash injury risk reduction for women and men. *Front. Bioeng. Biotechnol.* **2026**, *14*, 1837076. [CrossRef] [PubMed]
53. Douglas, D.M.; Howard, D.; Lacey, J. Moral responsibility for computationally designed products. *AI Ethics* **2021**, *1*, 273–281. [CrossRef]
54. Madaio, M.A.; Stark, L.; Wortman Vaughan, J.; Wallach, H. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020*; ACM: New York, NY, USA, 2020; pp. 1–14. [CrossRef]
55. Kawakami, A.; Coston, A.; Zhu, H.; Heidari, H.; Holstein, K. The Situate AI Guidebook: Co-designing a toolkit to support multi-stakeholder, early-stage deliberations around public sector AI proposals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'24), Honolulu, HI, USA, 11–16 May 2024*; ACM: New York, NY, USA, 2024; p. 749. [CrossRef]
56. Douglas, D.M.; Lacey, J.; Howard, D. Ethical risks of AI-designed products: Bespoke surgical tools as a case study. *AI Ethics* **2023**, *3*, 1117–1133. [CrossRef]
57. Nagy, D.; Lau, D.; Locke, J.; Stoddart, J.; Villaggi, L.; Wang, R.; Zhao, D.; Benjamin, D. Project Discover: An application of generative design for architectural space planning. In *Proceedings of the Symposium on Simulation for Architecture and Urban Design, Toronto, ON, Canada, 22–24 May 2017*; Society for Modeling and Simulation International: San Diego, CA, USA, 2017. [CrossRef]
58. UNECE. *UN Regulation No. 137—Uniform Provisions Concerning the Approval of Passenger Cars in the Event of a Frontal Collision with Focus on the Restraint System, Revision 2*; United Nations Economic Commission for Europe: Geneva, Switzerland, 2023. Available online: https://unece.org/sites/default/files/2024-07/R137r2e.pdf (accessed on 15 November 2026).
59. Euro NCAP. *Crash Protection—Frontal Impact Protocol, Version 1.1*; European New Car Assessment Programme: Leuven, Belgium, 2025. Available online: https://cdn.euroncap.com/cars/assets/euro_ncap_protocol_crash_protection_frontal_impact_v11_9659fff24e.pdf (accessed on 15 November 2026).

60. Euro NCAP. *Crash Protection—Rear Impact Protocol, Version 1.2*; European New Car Assessment Programme: Leuven, Belgium, 2026. Available online: https://cdn.euroncap.com/cars/assets/Euro_NCAP_Protocol_Crash_Protection_Rear_Impact_v1_2_47a249dae9.pdf (accessed on 15 November 2026).
61. Bose, D.; Segui-Gomez, M.; Crandall, J.R. Vulnerability of female drivers involved in motor vehicle crashes: An analysis of US population at risk. *Am. J. Public Health* **2011**, *101*, 2368–2373. [CrossRef] [PubMed]
62. European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). *Off. J. Eur. Union* **2024**, L 2024/1689. Available online: http://data.europa.eu/eli/reg/2024/1689/oj (accessed on 15 November 2026).
63. Engineering Council; Royal Academy of Engineering. Statement of Ethical Principles for the Engineering Profession; Updated 23 February 2026. Available online: https://www.engc.org.uk/resources-and-guidance/guidance-for-the-profession/ethical-principles (accessed on 15 November 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.