Ex-Post Evaluation of Data-Driven Decisions: Conceptualizing Design Objectives

Nada Elgendy[1[0000-0001-6765-017X]], Ahmed Elragal [2[0000-0003-4250-4752]], Markku Ohenoja [3[0000-0001-7577-7218]], and Tero Päivärinta [1[0000-0002-7477-0783]]

[1] M3S, Faculty of Information Technology and Electrical Engineering
University of Oulu, Finland
[2] Department of Computer Science, Electrical and Space Engineering
Lulea University of Technology, Sweden
[3] Environmental and Chemical Engineering, Faculty of Technology
University of Oulu, Finland
nada.sanad@oulu.fi

Abstract. This paper addresses a need for developing ex-post evaluation for data-driven decisions resulting from collaboration between humans and machines. As a first step of a design science project, we propose four design objectives for an ex-post evaluation solution, from the perspectives of both theory (concepts from the literature) and practice (through a case of industrial production planning): (1) incorporate multi-faceted decision evaluation criteria across the levels of environment, organization, and decision itself and (2) acknowledge temporal requirements of the decision contexts at hand, (3) define applicable mode(s) of collaboration between humans and machines to pursue collaborative rationality, and (4) enable a (potentially automated) feedback loop for learning from the (discrete or continuous) evaluations of past decisions. The design objectives contribute by supporting the development of solutions for the observed lack of ex-post methods for evaluating data-driven decisions to enhance human-machine collaboration in decision making. Our future research involves design and implementation efforts through on-going industry-academia cooperation.

Keywords: Data-driven decisions, ex-post evaluation, design objectives, collaborative rationality, human-machine collaboration.

1 Introduction

The ex-post evaluation of data-driven decisions emerges as an increasingly relevant, whilst difficult, topic [7–9]. Its complexity lies in that data-driven decision making involves five interrelated elements: the human decision maker, machine (analytics algorithms), data, decision-making process, and decision outcome [9]. This coexistence of machine learning (ML) and artificial intelligence (AI) systems with human decision makers has ignited interest in augmenting human intelligence and capabilities, resulting in more "intelligent" data analysis and support for decision-making and learning [13, 18, 42]. However, ex-post evaluation is crucial for enabling feedback for experiential learning to improve both organization and machine decisions, and measure the benefits of human-machine collaboration [20, 24, 33, 42]. It can result in organizational and experiential learning [2, 23], rationalization [22], and sensemaking [51] from the decision outcomes and consequences, as well as allow for analysis, benchmarking, and comparison of the results [24]. Understanding how, why, and to what extent ML and AI systems are being used in decision making and their influence on individual and organizational decisions [8] is difficult to assess without a holistic evaluation perspective. This requires a feedback loop between actions and outcomes, and encoding the past into rules and procedures for future learning [23].

Nevertheless, despite its importance as one of the main stages in classical decision-making processes, ex-post evaluation is commonly overlooked in recent data-driven decision making research. Decision evaluation is more complicated than the mere evaluation of a choice [52]. Data-driven decision evaluation is further complicated by the fact that many interrelated factors and metrics affect the evaluation, involving both humans and machines in a constantly changing environment.

Shrestha et al. [36] distinguish further between three categories of human/machine decisions:

1) purely machine decisions (e.g., recommender systems, personalized ads);

2) sequence-based decisions, which can involve two sub-types:

(a) human-to-machine (e.g., sports analytics on which the human expert seeks evidence from data);

(b) machine-to-human (e.g., ideation in innovation);

3) aggregated decisions involving both humans and machines, in peer-like group decision making (e.g., assisted medicine healthcare applications).

This research focuses on the last two categories, in line with Ransbotham et al.'s [33] modes of collaboration where AI recommends and the human decides, or AI generates insights which the human uses in the decision process, or the human generates hypothetical situations and relies on AI to evaluate and assess them.

Such collaboration results in decisions provided by machines, decisions made by humans, and the final data-driven decision which is selected, implemented, and leads to certain outcomes, which all require evaluation. For example, let us consider an AI system used in clinical decision support. Depending on the context of the decision, one performance measure might be more important than another, such as with predicting mortality which requires high accuracy and precision [21]. Nevertheless, erroneous diagnoses can be made based on differences in the training data, and ground truth labels may not always be correct and are subjective to different opinions [19, 21]. Thus, a highly accurate model based on the available training data cannot indicate an accurate or correct decision, nor positive outcomes. Human intervention and monitoring are necessary, yet if their diagnosis conflicts with that of the machine, which one is correct? Accordingly, some outcomes (e.g., correct diagnosis) can only be known after time in order to evaluate whether or not the data-driven decision was, in fact, accurate and better than purely human decisions, which necessitates ex-post evaluation.

Due to a lack of information about how to perform such evaluation and the lack of IT artifacts to employ, organizations rarely conduct ex-post evaluation of their past decisions. Ex-post evaluation is designed to help companies learn from their mistakes in the past, avoid repeating them in the future, and convey their knowledge to others. It does not impede decision-making or promote a no-decision scenario, because the goal is to learn from the decision and improve the quality of future decisions, not to evaluate the decision-maker (human or machine).

Furthermore, a comprehensive viewpoint to the multiple, socio-technical, elements involved in data-driven decision making is lacking [25], and there is little agreement in the literature on what and how to evaluate [17, 41]. The aspect of time and the requirements for a longitudinal, or processual evaluation remains ignored, which would be imperative to capture the complex dynamics involving change related to multi-faceted decisions [5]. Accordingly, we set out with the following research question:

"What are the requirements and design objectives for ex-post evaluation of data-driven decisions in organizations?"

This research problematizes ex-post evaluation of data-driven decision making by highlighting the gap in research and the industry need for a more holistic solution. We define design objectives (DOs) for the solution by first extracting the relevant ex-post evaluation concepts from the literature. These concepts are then exemplified through an industrial example of a chemical production plant to foresee how ex-post evaluation of data-driven decisions could be done in practice, and accordingly outline the initial requirements for a design solution.– covering two first steps of a design research program (cf. [29]) with industry.

The remainder of the paper is structured as follows. Section 2 covers the related research and literature analysis. The research method is outlined in section 3. Section 4 describes the results and finding, which are the evaluation requirements and DOs. Finally, section 6 concludes the paper with suggestions for further research.

2 Literature on Evaluating Data-Driven Decisions

2.1 Lack of Ex-Post Evaluation Support

In search of evaluation concepts, criteria, or solutions, we reviewed literature from various streams and disciplines, including decision research, information systems (IS), behavioral sciences, AI, ML, and information technology (IT). In the following, the literature was divided roughly into three streams.

The first stream focuses on decision theories with attention on human rationality [10, 22, 37, 47] and decision making [1, 11, 27] in various fields, such as management, economics, and psychology. In this stream, ex-ante evaluation of alternatives and choices was extensively studied, often focusing on individual metrics and values (e.g., utility). Although ex-post evaluation was included as a stage of suggested decision processes and deemed crucial in some fields (e.g., policy making [50]), less attention was paid on how evaluation was conducted (methods, metrics, time) or how it influenced the remainder of the process. The collaboration between humans and machines and how to evaluate data-driven decisions was non-prevalent in this stream.

The second stream focuses on AI and ML from the technical perspective of computer science and engineering, and the application of algorithms, models, and methods to a dataset to solve a specific problem [18, 35]. Research tends to focus on the use of machines for selecting among ex-ante alternatives [25, 45]. Evaluation covers the performance of the algorithm or model used in decision making, and a limited set of evaluation metrics are prevalent in the field [26, 49]. Different metrics have their strengths and limitations, are dependent on the available data, and may be conflicting (efficiency vs. accuracy vs. cost, etc.) [34]. Moreover, evaluating model performance is not the same as evaluating the resulting decision or its consequences.

The third stream of research involves data-driven decision making, highlighting the sociotechnical aspect and the relationship between the human and machine decision makers, mainly in an organizational context and with an IS perspective [9, 20, 36, 46]. Evaluation is still generally limited to the evaluation of choices and the evaluation of the performance of algorithms and models. Limited sources considered evaluating outcomes [43], let alone with multiple metrics [14]. However, no holistic evaluation solutions were found to consider the data-driven decision as a whole.

Consequently, the interaction between humans and machines and their roles in decision making is still not clear, and further research is necessary to evaluate the resulting decisions and determine the benefit, impact, and learning achieved through human-machine collaboration. Hence, we need new ways to evaluate AI-enabled decisions and benefits of human-machine collaboration in data-driven decision making. [7–9, 20].

2.2 Ex-Post Evaluation Concepts for Data-Driven Decisions

The literature introduces various concepts relevant to evaluating data-driven decisions. These serve as a theoretical basis for the requirements analysis leading to the suggested DOs (cf. the left column of Table 2 in section 4.1). First, there are embedded *contexts* for examining the decision situation to comprehend the factors affecting the decision and its impact [32]. The context pertains to the types of decisions made at different levels, ranging from individual to global, with varying requirements, as well as the decision environment, both internal and external [24]. The *environmental context* is the broadest perspective and includes the external environment and circumstances. The *organizational context* covers the characteristics of the organization in which the decision was made. The *decision context* includes aspects regarding the focus of the decision and the reasons for it, its relationship to other decisions, the complexity of the decision, constraints, etc. [24, 30, 32, 39].

Time highlights the processual nature of evaluation and refers to when and how often the evaluation is conducted, since the outcomes of the decision may vary across time. Decisions should be viewed from the perspective of process science which is concerned with understanding processes and influencing change in the desired directions over time [5]. One of the core requirements is to understand the emergent, situational, and holistic features of the decision, or the decision-making process, in its changing context [30], which adds to the necessity of a multi-faceted, process-oriented decision evaluation.

Data-driven decisions comprise *data-driven decision elements*, which include the *decision maker*, the *decision-making process*, the *data*, the *analytics/machine*, and the *decision outcome* [9]. Nevertheless, identifying decision outcomes is a difficult challenge due to their multi-faceted nature, variability in interpretation, acceptability and accountability to stakeholders, volatility and change, as well as their difficulty to fully grasp or quantitatively measure through indicators of success [27].

Accordingly, the decision outcome may be evaluated through multiple concepts, which extend across the various contexts and vary with time. Of particular importance are the *impact and consequences* of the decision and its perceived gains and losses [4, 27, 48, 52]. Furthermore, there is the *conformance* of the decision to certain criteria, as a decision involves some goals or values, some facts about the environment, and some inferences drawn from the values and facts. It must comply with objectives, criteria, standards, rules, and regulations, not only at the organizational context, but also at the environmental and societal contexts, since the decision may impact each [4, 27, 52].

Various *metrics* can be used for evaluating data-driven decisions and decision alternatives. Data-driven decisions are often evaluated with over-reliance or unwarranted dependence upon quantification and quantitative data [32]. Such metrics may potentially be conflicting, and generally focus on evaluating decision alternatives which differs from the ex-post evaluation of the decision after it is made [52].

Errors and biases can affect the outcome of decisions and thus need to be pinpointed and evaluated. Algorithmic predictions, although susceptible to their own types of errors, may influence human decisions. It is also necessary to differentiate between errors and biases that stem from the decision maker, and those which stem from the data, analytics, or machine, since each should be managed differently and require pertinent action [31].

3 Research Method

3.1 Research Design and Process

This research covers the first two steps of a design science research (DSR) process, to identify and motivate the problem, and to define the objectives of a solution [28]. Artifacts and solutions should be based on the relevant business needs from the environment and the applicable knowledge gained from the knowledge base [15]. Accordingly, the relevant concepts for ex-post evaluation were extracted from the literature (section 2.2). These concepts were used to theoretically support the industrial case and categorize the interview questions and thematize the evaluation requirements (section 4.1).

For portraying the practical aspects of our research, a case example of a chemical production plant is utilized. This plant, named ChemML (anonymized), is a simplified abstraction of a larger organization collaborating in an ongoing project, enhanced by the extensive knowledge and expertise of one of the authors experienced in chemical process engineering and decision making in such processes, and knowledgeable of the decisions, roles, data, and processes of ChemML and other chemical production plants.

This example was selected as chemical production plants have a high availability of mission-critical data-driven decisions. In such processes, hundreds or thousands of sensors routinely measure and automatically record data with high frequency. In a short time period, massive volumes of data are collected for process monitoring, evaluation, and control, which requires transforming the data into information for business and operation decision-making, in which ML tools have an important role [6, 44]. Furthermore, ChemML has recurring, operational, data-driven decisions in its production process, long-lasting adoption of systems utilizing ML in supporting decision making, and a desire to further evaluate, automate, and enhance the data-driven decision making processes.

Two expert interviews were held, from viewpoints of both the production planner and process operator roles, to discuss ChemML's data-driven decisions, explore the current evaluation methods, and discuss the need and requirements for a desired evaluation solution. By comparing the current and desired approaches for decision evaluation stated in the interviews, and applying deductive thematic analysis [38], we summarized the results into a set of evaluation requirements for each of the concepts. According to their functional similarities, the requirements were further thematized and mapped to more abstract and implementable DOs for an evaluation solution. The requirements were revised again to ensure that each requirement mapped to at least one DO.

The value of this study resides in the Eval 1 stage of Sonnenberg and vom Brocke's [40] DSR evaluation process for designing artifacts. This initial evaluation is conducted to justify a solution's novelty and importance for practice and to ensure that a meaningful problem has been identified. Accordingly, we attempted to evaluate feasibility, understandability, simplicity, completeness, and level of detail of the evaluation concepts and DOs in future design of an ex-post evaluation solution.

Internal validity was achieved through revision and agreement on the evaluation concepts, requirements, and DOs by each of the authors and expert in the case. The evaluation concepts and DOs were further presented to, and validated by, four experts in external software organization, under a case for utilizing data-driven decision making and AI to predict and prevent customer churn. The interview questions were validated by one of the analytics experts in the organization, and an additional interview was conducted with a customer success expert. The results were found to support the case of ChemML and findings of this paper, thus supporting external validity.

3.2 Case of ChemML

The production planning problem is a typical example of a complex, data-driven decision process with many interrelated factors, constraints, and major impacts. The orders placed by customers put a great pressure on production. Adjusting the production sequence must be done carefully to avoid disruption of production cycles, such as reduction in production rate and shutdowns. Moreover, ramping up the process and recovering from interruptions requires expenditure of energy, thus increasing the environmental load of the plant. Abrupt product sequence changes may cause quality deviations and wear of the machines and equipment.

Figure 1(a) depicts a simplified flowsheet of ChemML's multi-step, multi-product production process. Two critical features complicate the decision making:

(1) production planning is based on make-to-order (MTO) as the production batches cannot be stored for prolonged times, and

(2) routine operation has slow feedback from product quality to operational decisions.

Figure 1(b) illustrates data flows of ChemML. The automation system data includes the sensory measurements such as temperature (TI), material consumption (FI), quality attributes (QI), and energy consumption (EI) from the production process. ML tools infer data to routine operations and predict performance for manual production planning. The data-driven tools are advisory since the final decisions (process operation and resource planning) need to be made by humans due to responsibility issues.

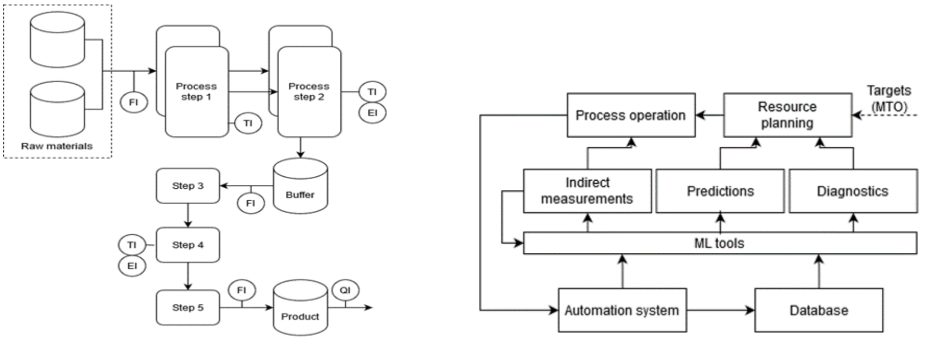

Fig. 1. **(a)** Production process schematic; **(b)** Data flows and decision support architecture

Based on the interviews, the data-driven decisions from the viewpoints of both the production planner and process operator roles, as well as the need for ex-post evaluation are described below in Table 1 (due to confidentiality requirements, some details could not be disclosed).

Table 1. Data-driven decisions at ChemML

Decision Context	Production Planner Viewpoint	Process Operator Viewpoint
Description	• Determine and plan the production targets and capacities for a specific time interval and schedule the production process (weekly).	• Operation decisions (continuous) during the execution of the process. Includes choosing set points for the process (such as feed rates and temperature), steering the process, and avoiding/overcoming fault situations.
Purpose	• Optimize production rate and product portfolio to meet market demand.	• Optimize the process in terms of efficiency (energy, material), avoid faults, and solve possible problems.
Decision Maker(s)	• Production planner (human decision maker) determines objectives and constraints. • ML tool supports decision by simulating scenarios and suggesting alternative schedules.	• ML tool provides outputs, insights, and predictions based on data and process parameters to steer the process and dynamically overcome fault situations. • ML tool provides suggestions of values for optimizing process efficiency.

	• Human selects best schedule to meet designated criteria and makes the final decision. • Human may overlook output of the ML tool and decide not to use it.	• Final decisions are made by the process operator who may use their own knowledge and expertise, along with additional monitoring methods.
Mode of Collaboration	• AI recommends, human decides.	• AI recommends, human decides. • AI generates insights, human uses in decision process.
Additional Requirements (Environment, Organization)	• Meet sales demands and maximize profit. • Conform to safety and quality requirements, meet standards and regulations, and laboratory testing. • Minimize waste and carbon footprint, and conform to pollution limits and the use of hazardous materials.	• Meet production targets in time. • Conform to safety and quality requirements, and professional standards and regulations.
Need for Ex-Post Evaluation	• Assess reliability and effectiveness of ML tool. • Enhance, both human and machine, learning from evaluation feedback. • Evaluate the collaboration between the human and machine. • Evaluating decisions at different time intervals would give indications if the reliability of the ML tool is increasing across time.	• Evaluate ML tool and its value to decision making. • Evaluate ML indicators and their usefulness in decision making. • Evaluate the extent to which ML tool is used and affects the decision. • Evaluate the decision outcome • Evaluate expertise of the process operator, and the monitoring methods used to reach the decision. • Evaluate uncertainties in measurement data and their effect on the decision.

4 Results and Findings

4.1 Analysis of Ex-Post Evaluation Requirements

Table 1 summarizes the requirements and considerations found necessary in the case for ex-post evaluation, in light of the analytical concepts originating in our literature review. Each of these conceptual elements helped to identify the interrelated requirements. Accordingly, we thematically grouped the similar requirements together from which we derived four main DOs, explained below. Each requirement in the table is labelled with the pertaining DO it corresponds to.

Table 2. Data-driven decision evaluation requirements

Evaluation Concepts		**Proposed Evaluation Requirements/ Considerations**
Overall Evaluation		Define the evaluation metrics. Evaluation should not revolve purely around a single metric. (***DO1***) Determine the evaluation process, the evaluators, and their roles. (***DO1, DO2, DO4***) Differentiate between the evaluation of the ML tool output/decision, and the evaluation of the overall decision involving both humans and machines. (***DO1, DO2, DO3***) Simplicity of evaluation, without requiring much time or work. (***DO1, DO2, DO4***) Transparency of the evaluation process to increase trust in the ML tools. (***DO1, DO2, DO3, DO4***) Identify relevant criteria for each data-driven decision evaluation. Some criteria and elements need only be evaluated when triggered by change, or problems arise. (***DO1, DO2, DO4***) Automated/partially automated evaluation. (***DO1, DO2, DO4***) Continuous feedback and learning from past decisions. (***DO2, DO4***)
Evaluation Across Contexts (Decision, Organization, Environment)		Determine the relevant criteria and metrics in each of the contextual levels. (***DO1, DO2***) Determine the interrelationship between the metrics across the contextual levels, and how they affect the data-driven decision and its evaluation. (***DO1, DO4***) Determine what is being evaluated (decision/set of decisions) and by whom. (***DO1, DO2, DO4***)
Evaluation Across Time (Processual)		Determine the time intervals and periods for which certain types of decisions on various levels should be evaluated and/or re-evaluated. (***DO1, DO2, DO4***) Account for changes in decision related concepts and contexts (***DO1, DO2, DO4***
Data-Driven Decision Elements	**Decision Maker**	Include both the human and the machine decision makers and suggest various metrics or criteria for evaluating each type of decision maker. (***DO1, DO3***) Enhance learning of the decision makers based on the results of past decisions. (***DO3, DO4***) Differentiate evaluation according to the mode of collaboration between the human and the machine. This may call for different evaluation methods, metrics, and requirements for different modes of collaboration. (***DO1, DO2, DO3, DO4***) Evaluate decision maker-related aspects; it may be useful in learning from past decisions. (***DO3, DO4***)
	Process	In this case not required, or too difficult to evaluate. (***DO1, DO2***))
	Data	Determine criteria and metrics for evaluating the data. (***DO1***)

		Suggest the effect of the data or changes in the data, on the decision. (***DO1, DO2***) Provide simple, automated methods for evaluating the data. (***DO1, DO4***) Distinguish between the data required for making the decision, and the data required for evaluating the decision. (***DO1, DO2***)
	Analytics/ Machine	Incorporate additional metrics and deal with conflicting metrics. (***DO1***) Suggest the effect of the analytics (and choice of analytics) on the decision, and how to evaluate and incorporate the ML output. (***DO1, DO2, DO3, DO4***) Enhance learning of the ML tool and feed the results back into the training data. (***DO3, DO4***)
	Decision Outcome	Determine metrics and criteria for evaluating the outcome of the decision after it is made (what defines a "good" decision?). (***DO1, DO2***) Observe the effect of the other decision factors, and their changes, on the decision outcome. (***DO1, DO2, DO4***) Determine when the decision should be evaluated. (***DO1, DO2, DO4***) Consider changing outcomes and the temporal factor. (***DO1, DO2, DO4***)
Impact and Consequences		Determine the metrics and criteria for evaluating the impact and consequences of the decision across contexts. (***DO1, DO2***) Determine the timeframe within which the impact should be evaluated. (***DO1, DO2, DO4***)
Conformance		Determine the relevant conformance metrics and criteria across the contextual levels. (***DO1***) Distinguish between short-term and long-term conformance evaluation criteria. Conformance requirements may change across time. (***DO1, DO2, DO4***)
Metrics		Deal with conflicting metrics, goals, and constraints. (***DO1***) Incorporate separate metrics related to the human decision maker and the decision, along with the AI/ML metrics and those related to the machine. (***DO3***) Prioritize the most important/relevant criteria and metrics and providing guidance on weights and selection of metrics. (***DO1, DO4***) Differentiate between short-term and long-term evaluation metrics. Do not include all metrics each time. (***DO1, DO2, DO4***)
Errors and Biases		Differentiate between errors and biases related to human decision makers, machines (analytics), and data. (***DO1, DO3***) Define appropriate metrics and criteria for evaluating errors and biases. (***DO1***) Identify errors to learn from past decisions for future decisions. (***DO1, DO4***)

4.2 Design Objectives for Ex-Post Decision Evaluation

Consequently, four main DOs were concluded for a future solution which responds to the needs of ChemML, as shown in Table 3.

The first DO for an implementable ex-post data-driven decision evaluation method is that it should be comprehensive and incorporate multi-faceted criteria. These criteria may range across the contextual levels priorly discussed, and include some of the proposed concepts as facets. For instance, in the ChemML case, the contextual levels can incorporate environmental impacts, which are governed by averaged or long-term process performance, whereas short-term decisions related to process operation may have positive short-term impacts (on a decision level) but negative long-term impacts. Furthermore, the criteria should differentiate between the data-driven decision elements, such as the evaluation of the machine, the decision outcome, the data, etc., which may potentially be conflicting and otherwise lead to confusion. In ChemML, although the accuracy and evaluation metrics of the ML tool's decision may have been high in a majority of instances, the expert's evaluation generally differed and took into account different aspects and criteria.

Similarly, DO2 encourages performing processual evaluation across different stages in time. This considers the changing contexts and aspects regarding the data-driven decision, which should be captured in the evaluation to understand the longitudinal consequences and impact of the decision. A concrete example related to ChemML would be the performance evaluation of indirect measurements, which can be dependent on factors such as seasonal variability of the raw materials, changes in ambient conditions, and unmodeled changes related to equipment fouling or degradation.

Table 3. Design objectives for ex-post evaluation of data-driven decisions

	Design Objective
1	Incorporate multi-faceted (potentially conflicting) evaluation criteria across contextual levels.
2	Perform processual evaluation across time.
3	Define the applicable mode of collaboration between humans and machines and evaluate its effect on decision-making, decision outcomes, and collaborative rationality.
4	Enable a (potentially automated) feedback loop for learning from the (discrete or continuous) evaluation of past decisions.

DO3 focuses on the relation between the human and the machine in the data-driven decision making process. By incorporating into the evaluation the mode of collaboration between humans and machines and the consequent effect on decision making, the decision outcomes, and achieving a collaborative rationality, we can glean more insights on such a collaboration and how to steer it to make better decisions. In ChemML, the evaluation would require, for example, regular interviews with end-users to assess the utilization degree of the machine, or development of automated logging of the human-machine interaction during the decision-making process. The latter could also facilitate DO4, where a (possibly automated) feedback loop ensues from the evaluation and enables learning through evaluating past decisions, both from an organizational and machine perspective, and consequently updates the training data to enhance ML. Similar to how decisions may be discrete or continuous, the evaluation of decisions and the resulting learning may also be discrete or continuous, depending on the decision type and context. Therefore, a design solution should be developed ingraining these objectives.

5 Discussion

The two main contributions of our research are:

1) the extraction of evaluation concepts from the literature, and

2) building upon them in the case of ChemML to define the requirements and DOs for data-driven decision evaluation in practice.

The concepts to be considered in ex-post decision evaluation are particularly of interest due to their ability to capture the multi-faceted and changing nature of data-driven decisions, rather than focus on individual or static evaluation concepts (e.g., at the level of the decision itself), as is mainly done in current studies. These concepts theoretically support, and are supported by, the practical example of ChemML. Its contemporary evaluation methods did not consider multiple criteria or contexts, although a comprehensive, ex-post evaluation method was desired to enable learning, as well as to enhance future decision making and increase adoption of the ML tool.

The DOs further contribute to theory and practice, and emphasize the need for a comprehensive evaluation method which incorporates multi-faceted evaluation criteria across the levels of the decision itself, organization, environment, and time. By reflecting on ChemML, we can see that multiple interrelated factors are present in each decision, and individual evaluation metrics on a single level remain insufficient in terms of ex-post learning. This challenges current research, which focuses on ML evaluation metrics, such as confidence, uncertainty, specificity, sensitivity, accuracy, area under the curve (AUC), etc. [26, 49]. Whereas such measures are necessary for evaluating the ML model performance as such, our paper argues that they are insufficient for ex-post evaluation and learning about the decisions.

This argument is in line with Lebovitz et al. [19], which show the limitations of primary performance measures used by managers to evaluate AI tools and their output. Contrarily, the actual results and knowledge of the experts, in many instances, conflict with the reported measures of the tools [16], which in the ChemML case decreased trust in the tool. Furthermore, the machine ignores certain important variables only human experts are capable of considering [14, 19], which was also the case with ChemML. This emphasizes the need for additionally accounting for the modes of collaboration between humans and machines in the evaluation of data-driven decisions.

Depending on the use case and level of analysis, one performance measure may be more important than another, and the mathematically optimal may become ethically problematic. Decision outcomes are thus the ultimate indicators of success and multiple factors should be considered in the evaluation, along with long-term follow up [19]. Accordingly, our first and third DOs support, and are supported by, such claims in recent research and endeavor to provide a solution to the evaluation paradox. Although some papers do consider evaluation of the algorithms, along with evaluation of the impact of the decision [14], their research focuses on a particular approach for data-driven decision making in a domain-specific decision, and they do not aim to provide ex-post evaluation solutions.

The second DO highlights the importance of a process science perspective and capturing the changes in contexts, concepts, and consequences, as well as understanding how they evolve, interact, and unfold, through a processual evaluation across time [5]. The set of decisions and concepts involved in the evaluation, as well as the evaluation method, may differ according to the stage in time when the evaluation is made. For example, in ChemML, production was evaluated within a shorter time frame based on whether the production targets were met. However, the environmental impact is used to evaluate a set of decisions at a later stage in time. Thus, it is crucial to know what to evaluate when.

The fourth DO builds on the traditional claim that ex-post evaluation enables learning from past decisions. This accentuates the need for designing a feedback loop which performs an evaluation based on the first two DOs, and feeds the results of the evaluation back into the process to enable a combination of both organizational and machine learning. This feedback loop is part of a prospective solution for monitoring how data-driven decisions are taken, cultivating criteria to evaluate such decisions, and reflecting through double-loop learning for the continuous evaluation and improvement of human-machine collaboration [7, 8, 20, 42]. While this feedback loop (or parts of it) could potentially be automated to simplify the task, we still support Grønsund and Aanestad's [12] claims that necessitate the human-in-the-loop configuration for ensuring that performance of the algorithm meets the organization's requirements.

Utilizing the knowledge presented by these DOs, a theory-ingrained and practically feasible solution to the ex-post evaluation of data-driven decisions can be developed. This further contributes to practice by enabling the evaluation and understanding of data-driven decisions, enhancing learning usage of AI and ML tools, and adding insights to the collaboration between humans and machines and the impact on decision making. Accordingly, decision makers, developers, and collaborators in the data-driven decision making process can benefit from the results. Finally, the development of a data-driven decision evaluation solution following the determined DOs may potentially address the data-driven decision making challenges faced by ChemML and many other organizations.

6 Conclusion and Future Work

In this paper, we aimed to problematize the ex-post evaluation of collaborative data-driven decisions, from the perspectives of theory and practice, and determine the DOs for a solution. Accordingly, by perusing the literature we determined the need for ex-post evaluation and a variety of concepts and factors to consider in the evaluation. From a practical perspective, ChemML exemplified the need for data-driven decision evaluation in industry, and was used to identify the necessary requirements and considerations for a proposed ex-post evaluation solution.

From these requirements, four DOs for a solution were proposed: (1) the existence of an implementable, comprehensive method incorporating multi-faceted (potentially conflicting) evaluation criteria across contextual levels (decision, organization, environment), (2) accounting for the changes in concepts, contexts, and outcomes across time and supporting a processual evaluation, (3) incorporating into the evaluation the mode of collaboration between humans and machines and its effect on decision-making, decision outcomes, and achieving a collaborative rationality, and (4) enabling a (potentially automated) feedback loop for learning from the evaluation of past decisions.

Future work includes utilizing the DOs towards building and testing a design artifact, in collaboration with industry, which could be used in an organizational context for the purpose of data-driven decision evaluation. This artifact should support "how" (process, metrics, and criteria) and "when" (which stages in time, if at all) to evaluate data-driven decisions. Additionally, we aim for a longitudinal case study in order to understand the organizational context surrounding data-driven decisions prior to the introduction of the evaluation, during the implementation, and post- implementation, following Bailey and Barley's [3] approach to studying intelligent systems in organizational contexts. Finally, we intend to research the concept of collaborative rationality further, and how to enhance the collaboration between humans and machines in decision making.

Acknowledgment

This research has been partially funded by the ITEA3 project Oxilate (https://itea3.org/project/ oxilate.html).

References

1. Ajzen, I.: The Social Psychology of Decision Making. In: Social Psychology: handbook of basic principles. pp. 297–325 (1996).

2. Argyris, C., Schön, D.A.: Organizational Learning: A Theory of Action Perspective. Addison-Wesley. 77/78, 345 (1997). https://doi.org/10.2307/40183951.

3. Bailey, D.E., Barley, S.R.: Beyond design and use: How scholars should study intelligent technologies. Information and Organization. 30, 2, 100286 (2020). https://doi.org/10.1016/j.infoandorg.2019.100286.

4. Bouyssou, D. ed: Evaluation and decision models: a critical perspective. Kluwer Academic Publishers, Boston (2000).

5. vom Brocke, J. et al.: Process Science: The Interdisciplinary Study of Continuous Change. Social Science Research Network, Rochester, NY (2021). https://doi.org/10.2139/ssrn.3916817.

6. Chiang, L. et al.: Big Data Analytics in Chemical Engineering. Annual Review of Chemical and Biomolecular Engineering. 8, 1, 63–85 (2017). https://doi.org/10.1146/annurev-chembioeng-060816-101555.

7. Duan, Y. et al.: Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. International Journal of Information Management. 48, 63–71 (2019). https://doi.org/10.1016/j.ijinfomgt.2019.01.021.

8. Dwivedi, Y.K. et al.: Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management. 57, 101994 (2021). https://doi.org/10.1016/j.ijinfomgt.2019.08.002.

9. Elgendy, N. et al.: DECAS: a modern data-driven decision theory for big data and analytics. Journal of Decision Systems. 1–37 (2021). https://doi.org/10.1080/12460125.2021.1894674.

10. Gigerenzer, G., Gaissmaier, W.: Decision Making: Nonrational Theories. In: International Encyclopedia of the Social & Behavioral Sciences. pp. 911–916 Elsevier (2015). https://doi.org/10.1016/B978-0-08-097086-8.26017-0.

11. Gigerenzer, G., Gaissmaier, W.: Heuristic Decision Making. Annual Review of Psychology. 62, 1, 451–482 (2011). https://doi.org/10.1146/annurev-psych-120709-145346.

12. Grønsund, T., Aanestad, M.: Augmenting the algorithm: Emerging human-in-the-loop work configurations. The Journal of Strategic Information Systems. 29, 2, 101614 (2020). https://doi.org/10.1016/j.jsis.2020.101614.

13. Grover, V. et al.: The Perils and Promises of Big Data Research in Information Systems. Journal of the Association for Information Systems. 21, 2, (2020). https://doi.org/10.17705/1jais.00601.

14. Herm-Stapelberg, N., Rothlauf, F.: The crowd against the few: Measuring the impact of expert recommendations. Decision Support Systems. 138, 113345 (2020). https://doi.org/10.1016/j.dss.2020.113345.

15. Hevner, A.R. et al.: Design Science in Information Systems Research. MIS Quarterly. 28, 1, 75–105 (2004). https://doi.org/10.2307/25148625.

16. Ioannidis, J.P.A. et al.: Forecasting for COVID-19 has failed. Int J Forecast. (2020). https://doi.org/10.1016/j.ijforecast.2020.08.004.

17. Klecun, E., Cornford, T.: A critical approach to evaluation. European Journal of Information Systems. 14, 3, 229–243 (2005). https://doi.org/10.1057/palgrave.ejis.3000540.

18. Kotsiantis, S.B. et al.: Machine learning: a review of classification and combining techniques. Artif Intell Rev. 26, 3, 159–190 (2007). https://doi.org/10.1007/s10462-007-9052-3.

19. Lebovitz, S. et al.: Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What. MISQ. 45, 3, 1501–1526 (2021). https://doi.org/10.25300/MISQ/2021/16564.

20. Lyytinen, K. et al.: Metahuman systems = humans + machines that learn. Journal of Information Technology. 36, 4, 427–445 (2020). https://doi.org/10.1177/0268396220915917.

21. Magrabi, F. et al.: Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications: A Position Paper from the IMIA Technology Assessment & Quality Development in Health Informatics Working Group

and the EFMI Working Group for Assessment of Health Information Systems. Yearb Med Inform. 28, 01, 128–134 (2019). https://doi.org/10.1055/s-0039-1677903.

22. March, J.G.: Bounded Rationality, Ambiguity, and the Engineering of Choice. The Bell Journal of Economics. 9, 2, 587–608 (1978). https://doi.org/10.2307/3003600.

23. March, J.G.: Primer on Decision Making: How Decisions Happen. Simon and Schuster (1994).

24. Masha, E.M.: The Case for Data Driven Strategic Decision Making. European Journal of Business and Management. 10 (2014).

25. Namvar, M., Intezari, A.: Wise Data-Driven Decision-Making. In: Dennehy, D. et al. (eds.) Responsible AI and Analytics for an Ethical and Inclusive Digitized Society. pp. 109–119 Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-85447-8_10.

26. Nasir, M. et al.: Developing a decision support system to detect material weaknesses in internal control. Decision Support Systems. 151, 113631 (2021). https://doi.org/10.1016/j.dss.2021.113631.

27. Nutt, P.C., Wilson, D.C. eds: Handbook of decision making. John Wiley, Chichester, West Sussex, U.K. ; Hoboken, N.J (2010).

28. Peffers, K. et al.: A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems. 24, 3, 45–77 (2007). https://doi.org/10.2753/MIS0742-1222240302.

29. Peffers, K. et al. eds: Design Science Research in Information Systems. Advances in Theory and Practice. Springer Berlin Heidelberg, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29863-9.

30. Pettigrew, A.M.: Contextualist Research and the Study of Organizational Change Processes. 20 (1985).

31. Phillips-Wren, G. et al.: Cognitive bias, decision styles, and risk attitudes in decision making and DSS. Journal of Decision Systems. 28, 2, 63–66 (2019). https://doi.org/10.1080/12460125.2019.1646509.

32. Power, D.J. et al.: Analytics, bias, and evidence: the quest for rational decision making. Journal of Decision Systems. 28, 2, 120–137 (2019). https://doi.org/10.1080/12460125.2019.1623534.

33. Ransbotham, S. et al.: Expanding AI's Impact With Organizational Learning, https://sloanreview.mit.edu/projects/expanding-ais-impact-with-organizational-learning/, last accessed 2021/12/22.

34. Raschka, S.: Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, http://arxiv.org/abs/1811.12808, (2020).

35. Russell, S., Norvig, P.: Artificial intelligence: a modern approach. Pearson, Hoboken (2021).

36. Shrestha, Y.R. et al.: Organizational Decision-Making Structures in the Age of Artificial Intelligence. California Management Review. 61, 4, 66–83 (2019). https://doi.org/10.1177/0008125619862257.

37. Simon, H.A.: A Behavioral Model of Rational Choice. The Quarterly Journal of Economics. 69, 1, 99 (1955). https://doi.org/10.2307/1884852.

38. Smith, J.A.: Qualitative Psychology: A Practical Guide to Research Methods. SAGE (2015).

39. Snowden, D.J., Boone, M.E.: A Leader's Framework for Decision Making. harvard business review. 10 (2007).

40. Sonnenberg, C., vom Brocke, J.: Evaluations in the Science of the Artificial – Reconsidering the Build-Evaluate Pattern in Design Science Research. In: Peffers, K. et al. (eds.) Design Science Research in Information Systems. Advances in Theory and Practice. pp. 381–397 Springer Berlin Heidelberg, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29863-9_28.

41. Stockdale, R., Standing, C.: An interpretive approach to evaluating information systems: A content, context, process framework. European Journal of Operational Research. 173, 3, 1090–1102 (2006). https://doi.org/10.1016/j.ejor.2005.07.006.

42. Sturm, T. et al.: Coordinating Human and Machine Learning for Effective Organization Learning. MISQ. 45, 3, 1581–1602 (2021). https://doi.org/10.25300/MISQ/2021/16543.

43. Sturm, T. et al.: The Case of Human-Machine Trading as Bilateral Organizational Learning. 18 (2021).

44. Tomperi, J. et al.: Mass-balance based soft sensor for monitoring ash content at two-ply paperboard manufacturing. Nordic Pulp & Paper Research Journal. 37, 1, 175–183 (2022). https://doi.org/10.1515/npprj-2021-0046.

45. Troisi, O. et al.: Growth hacking: Insights on data-driven decision-making from three firms. Industrial Marketing Management. 90, 538–557 (2020). https://doi.org/10.1016/j.indmarman.2019.08.005.

46. Trunk, A. et al.: On the current state of combining human and artificial intelligence for strategic organizational decision making. Bus Res. 13, 3, 875–919 (2020). https://doi.org/10.1007/s40685-020-00133-x.

47. Tversky, A., Kahneman, D.: Rational Choice and the Framing of Decisions. The Journal of Business. 59, 4, S251–S278 (1986).

48. Tversky, A., Kahneman, D.: The Framing of Decisions and the Psychology of Choice. Science. 211, 4481, 453–458 (1981). https://doi.org/10.1126/science.7455683.

49. Vo, N.N.Y. et al.: Deep learning for decision making and the optimization of socially responsible investments and portfolio. Decision Support Systems. 124, 113097 (2019). https://doi.org/10.1016/j.dss.2019.113097.

50. van Voorst, S., Zwaan, P.: The (non-)use of ex post legislative evaluations by the European Commission. Journal of European Public Policy. 26, 3, 366–385 (2019). https://doi.org/10.1080/13501763.2018.1449235.

51. Weick, K.E.: Sensemaking in organizations. Sage Publications, Thousand Oaks (1995).

52. Weirich, P.: Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances. Oxford University Press (2004).