Part I

Contributions, and Related Work

Chapter 1

What are the Contributions of This Book

In this chapter, I will directly address the core question: what are the key contributions of this book? Together with my colleagues—Dr. Lei Wang, Dr. Wanling Gao, Ph.D. students, Mr. Hongxiao Li, and Mr. Chenxi Wang, as well as Postdocs Dr. Fanda Fan and Dr. Guoxin Kang—we have made several novel contributions that represent significant advancements beyond prior work. Mr. Qian He made contributions to the presentation of several figures.

Throughout the remainder of this book, "I" refers to Dr. Jianfeng Zhan. When referring to my colleague—whether a Ph.D. student or a PostDoc—I will use their name. The term "we" denotes multiple contributors, with details provided in the context.

1.1 Uncovering the Essence of Evaluation

I reveal that the essence of evaluation is to infer the effect of an (evaluated) object.

Gravitation between objects manifests as a universal effect; combustion embodies oxygen's effect; heredity arises from the effect of genetic and other material factors; a criminal act invariably yields a detrimental effect; law implementation generates societal effects; while CPUs and Large Language Models in computing produce objectively quantifiable effects.

To systematically reveal these diverse effects, a universal conceptual framework, model, and methodology can be established. This constitutes the primary motivation for my proposal of Evaluatology as a new discipline.

1.2 Formalization of Evaluation and its Dual and Inverse Problems

Evaluation and design are dual problems. An object and external essential objects induce overall effects on an affected object. The evaluation is to uncover the true effect of a

specific object (a cause) from the overall effect, while the design of an object aims to search for a specific object configuration to achieve the optimal overall effect.

We observe a phenomenon that consists of many objects, showing different quantities. The inverse problem of the evaluation, which I call de-evaluation, is to trace back the objects and their observed quantities to the causes, external essential objects, and their effects on the affected objects.

I will formally define those concepts.

1.3 Formal Introduction of the Discipline of Evaluatology

I coined the term Evaluatology [1] to cover this exciting science. I formally define Evaluatology as the science of uncovering the effects.

I propose the fundamental components of Evaluatology.

1. universal evaluation concepts.
2. five axioms of evaluation.
3. categories of evaluation problems.
4. fundamental issues of Evaluatology.
5. fundamental Evaluatology methodology [2].

1.4 Why the Past Efforts Failed to Establish the Discipline of Evaluation?

I systematically explain why the past efforts failed to establish the discipline of evaluation.

First, the consensus of the evaluation community referred to evaluation as "the process of determining the merit, worth, or value of things" [172, 174, 175, 176, 55, 68, 85, 196, 195]. However, this definition fails to capture the essence of evaluation, as the terms "merit," "worth," and "value" are inherently subjective—their interpretation varies significantly across individuals, which constitutes a fundamental flaw in the definition.

Secondly, past works failed to propose universal evaluation concepts, problem statements, axioms, mathematical formulations, and methodologies. Instead, the community or leading scholars rely on the encyclopedic approach to define several hundred concepts or terms as shown in [174, 118].

[1] I originally coined the term "Evaluationology". Mr Hongxiao Li suggested I change it to Evaluatology after consulting with his English teacher. I adopted his suggestion.

[2] Dr. Wanling Gao, Mr. Chenxi Wang, and Dr. Lei Wang also contributed to this methodology.

1.5 New World Model: What Distinguishes Intelligent Lives from Normal Objects?

I propose foundational assumptions and models of the three worlds, upon which a distinct pathway to achieving strong AI becomes feasible.

An object is a class of entities owning a set of properties. The world consists of objects. I propose that intelligent Lives have two unique properties that distinguish them from other normal objects: *free will* and *interrogation*.

Free will is the capacity and capability to make free and intentional choices. Free will has different degrees. *Interrogation* is the capacity and capability to understand objects and their mutual effects. Interrogation has different complexities.

I formally define the cause and effect. For objects A and B, when measurable or testable differences occur in B depending on the presence or absence of A, A is the *cause*, B is the affected object, and the measurable or testable difference in B is the *effect* on B induced by A. The *effect mechanism* is the way the cause induces the effect on the affected object.

A microscopic object world consists of microscopic objects at the scale of atoms and subatomic particles, which is governed by the principles of quantum mechanics, dominated by probability.

A normal object world consists of normal objects, determined and governed by the principle of cause and effect.

A free will world consists of the normal objects and the intelligent lives, and a free will world is governed not only by the principle of cause and effect but also by free will. That is to say, an intelligent life not only receives the effects of causes, but also has the capacity and capability to make free and intentional choices

1.6 New Interpretation of Fundamental Interrogations: Measurement, Testing, and Reasoning

I posit that comparison constitutes a primitive form of interrogation, upon which intelligent life has built three fundamental interrogations: measurement, testing, and reasoning. I offer a novel interpretation of them with Dr. Lei Wang and Dr. Fanda Fan.

In Metrology, a unit of measurement like the meter was initially defined by referencing a specific Earth-based length, later refined using light speed as the reference for enhanced precision.

Similarly, in testing, the ground truth of objects constitutes the test oracle. For instance, when an image objectively depicts a cat, algorithms are tested on diverse feline images by comparing outputs against the test oracle to determine their validity. Modern artificial intelligence has evolved precisely through this paradigm.

Reasoning, as an intellectual activity utilizing inference rules that perfectly pass such tests, can supplant physical-world operations—all underpinned by comparison.

I also propose the term Testology to cover the science of testing and its application. The motivation is to present universal testing principles and methodologies across different areas, like verifying a Physics theory, hypothesis testing in statistics, and artifact testing, such as software and hardware testing.

1.7 Revealing the Essence of Value

I interpret the value of an object as its derived effect on the stakeholder. A stakeholder is an intelligent life or an organization that consists of intelligent lives. As a stakeholder can make free and intentional choices on different objects, a value function could be formulated on the basis of different quantities of their derived effects.

1.8 Fundamental Roles and Interrelationship of Four Interrogations

I clarify the unique fundamental roles and interrelationships of four interrogations: measurement, testing, reasoning, and evaluation. There are other interrogations, which I will explain in another book.

In the epistemic hierarchy of Evaluatology, measurement, testing, reasoning, and evaluation represent four fundamental interrogations through which intelligent lives explore the unexplored world and their unknown lives, and build massive knowledge systems.

Measurement answers "how much", attributing values to countable quantities of objects; testing answers "whether", determining conformity to the test oracle through verification and falsification; evaluation answers "why" in terms of how an object influences another one; reasoning answers "why" in terms of the underlying logical mechanisms that connect causes to their effects. Together, these four interrogations form a complete cognitive cycle—from observation, to validation, to explanation [3].

Mr. Chenxi Wang, Dr. Lei Wang, Dr. Wanling Gao, Dr. Fanda Fan, and I provide many examples to support our propositions or models.

1.9 The Summary of Other Fundamental Evaluation Methodology

My collaborators systematically summarize the other fundamental evaluation methodologies. Mr. Chenxi Wang contributed to the design of experiments, randomized control trials. Mr. Hongxiao Li contributed to quasi-experiments, structural causal models, Dr. Lei Wang contributed to structural equation models and instrumental variables, and Dr. Wanling Gao contributed to the potential outcome theory.

[3] Dr. Fanda Fan's work provided a basis for this passage.

I joined with all collaborators to discuss what is the basic concepts, problem statements, assumptions, and principles of each methodology. I contributed to the part of the content.

1.10 Formal Definitions of Benchmarks and Testbed

Benchmarks and testbeds are widely used in engineering without formal definitions and rigorous methodologies. I formally established the operational definitions of benchmarks and testbeds. Dr. Wangling Gao joined me to propose the principles and methodology for the testbed.

1.11 New Possible Paths to Strong Artificial Intelligence

I conceived the core idea of the paths to strong artificial intelligence based on Evaluatology. Dr. Guoxin Kang and Dr. Wanling Gao joined me to elaborate on those ideas.

1.12 Applications of Evaluatology in Different Areas.

Many collaborators actively utilize Evaluatology in different areas, demonstrating the power of Evaluatology. In this book, Dr. Fanda Fan and I showcase how to utilize Evaluatology to evaluate science and technology research institutes.

Chapter 2

Evaluation: Ancient Practice, Undeveloped Discipline

In this chapter, I will present the state-of-the-art and state-of-the-practice of evaluation. Section 2.1 will explain why evaluation is an ancient practice, incorporating the remarks of Dr. Michael Scriven. Section 2.2 presents different evaluation concepts and theories, and ad hoc practices in different domains. Section 2.3 provides an answer to why the past efforts failed to establish the discipline of evaluation. Section 2.4 presents the summary of this chapter.

2.1 Evaluation Is an Ancient Practice

Scriven [174] defined the evaluation as "the process of determining the merit, worth, or value of things, or the result of that process" [172, 174, 175, 176]. Based on this definition, he argued that evaluation is an ancient practice older than other human practices. I incorporate his remarks in this section.

Scriven [174] posited that Logic, closely analogous to reasoning elucidated in Chapter 6, and evaluation constitute "two foundational tool disciplines with pervasive applications across diverse academic domains." He concluded that the practical utilization of informal logic and evaluation predates the establishment of any formal academic disciplines or their early forms.

According to Scriven [174], informal logic and grammar co-evolved with any language that existed before formal academic disciplines, as both disciplines fundamentally rely on linguistic structures as their bedrock for further advancement. Nevertheless, evaluation practices endure even before the emergence of linguistic capabilities, owing to their indispensable role in the creation of early artifacts. For instance, the earliest recorded craftsmen—the stoneworkers—exhibit a consistent trajectory in material quality and design sophistication, a phenomenon discernible not only at individual archaeological sites but also across millennia of human history.

In Parts II and III, I define interrogation as the capacity, capability, and process to understand objects and their mutual effects. I systematically analyze four fundamental interrogations: measurement, testing, reasoning, and evaluation.

I defined the evaluation as uncovering the effects and derived effects of an (evaluated) object. If an object has a stakeholder, I interpreted determining the merit, worth, or value of an object [172, 174, 175, 176] as uncovering the derived effect of an evaluated object on the stakeholder. For any intelligent life, not just humans but also animals, evaluation is one of the fundamental interrogation abilities, as most of them can know and predict the effects of some causes. For example, a deer knows the emergence of a lion will endanger its life. For this reason, I concluded that evaluation practices endure much earlier than the period mentioned by Dr. Michael Scriven, even in the period when human beings did not exist.

2.2 Evaluation Concepts, Theories, and Ad Hoc Practices

In this section, I will summarize different evaluation concepts, theories, and ad hoc practices in different domains.

Stufflebeam [196] suggested eight questions to be addressed in any attempt to conceptualize evaluation, based on which Nevo [128, 129] extended to ten major dimensions in a conceptualization of evaluation. I propose nine dimensions, some of which overlap with those of Nevo [128, 129], incorporating the discussions by Scriven [175].

2.2.1 Definitions

Educational evaluation pioneer Ralph Tyler [206] perceives evaluation as "the process of determining to what extent the educational objectives are actually being realized." According to my definition, this definition could be reformulated as *the process of uncovering the effect of any object in the education process to determine whether it meets the intended effect.*

Another widely accepted definition of the evaluation has been that of providing information for decision-making, suggested by various leading evaluators such as Cronbach [44], Stufflebeam [194], and Alkin [2]. According to my definition, this definition could be reformulated as *uncovering and reporting the effects of an object for decision making.* Scriven referred to evaluation as "the process of determining the merit, worth, or value of things, or to the result of that process" [172, 174, 175, 176]. Evaluators and researchers in social sciences reached a considerable consensus regarding the definition of evaluation as the assessment of merit or worth [55, 68, 85, 196], or as an activity comprised of both description and judgment [74, 184]. A joint committee on standards for evaluation, comprised of 17 members representing 12 organizations associated with educational evaluation, recently published their definition of evaluation as "the systematic investigation of the worth or merit of some objects [195]."

According to my definition, anything or the result of any evaluation process could be an evaluated object. The essence of determining the merit, worth, or value of a thing is to uncover its derived effect on the stakeholder. Alternatively, it is feasible to compare the effects or derived effects of different evaluated objects of the same class. As analyzed in Chapter 10, I consider comparison as the most primitive interrogation.

Some groups or scholars rejected the judgmental definition of evaluation. For example, the Stanford Evaluation Consortium group defined evaluation as "systematic examination of events occurring in and consequent of a contemporary program -an examination conducted to assist in improving this program and other programs having the same general purpose" [45]. According to my definition, the events occurring in and consequent to a contemporary program could be an evaluated object. Uncovering its effect naturally helps improve the program.

Cronbach and his associates [45] perceived the evaluation as "an educator [whose] success is to be judged by what others learn", rather than a "referee [for] a basketball game", who is hired to decide who is "right or wrong." According to my definition, what other learn is one of the effects of an educator; the effects of different educators could be compared.

Rossi et al. present the concept framework in their famous book [152]. Throughout the book, the terms "evaluation", "program evaluation", and "evaluation research" are used interchangeably. Although they focus on the evaluation of the social program, they claim that the evaluation research is not limited to that arena [152].

Rossi et al. [152] defined program evaluation as "the application of social research methods to systematically investigate the effectiveness of social intervention programs in ways that are adapted to their political and organizational environments and are designed to inform social action to improve social conditions." In this definition, the social programs, also referred to as social interventions, cover human services programs in the domain of "health, education, employment, housing, community development, poverty, criminal justice, and international development."

According to my definition, the social intervention programs are the evaluated object. After uncovering the effects of different programs, we can investigate their effectiveness.

2.2.2 Essences or Views

In [175], Scriven summarized several different thoughts on the essences or views of evaluations, including A: "strong decision support," B: "weak decision support," C: "relativistic," D: "rich description," E: "social process," F: "constructivist" or "fourth generation." Actually, Scriven used the term "models of evaluation." I would rather use essence or view than a model for two reasons: the essence or view is more accurate; the evaluation model has other uses in Evaluatology.

A: "strong decision support" view

View A was exemplified in, but not made explicit by, the work of Ralph Tyler, and extensively elaborated in the CIPP (Context, Input, Process, and Product) model of evaluation [194]. Unfortunately, this view does not reveal the essence of evaluation. Instead, it focuses on the purpose of the evaluation: "explication of the use of program evaluation as part of the process of rational program management, conceiving of evaluators as doing investigations aimed at arriving at evaluative conclusions designed to assist the decision-maker."

B. The "weak decision support" view

This point of view is represented by evaluation theorists such as Marv Alkin [2], who define evaluation as "factual data gathering in the service of a decision-maker who is to draw all evaluative conclusion." Similarly, this view does not reveal the essence of evaluation. Instead, it focuses on a sub-process of the evaluation: it provides decision-relevant data, and even stops short of drawing evaluative conclusions.

C. The "relativistic" view

This view is from two social scientists and essentially represents this approach [153]. It holds that evaluation should be done by using the client's values as a framework, without any judgment by the evaluator about those values or any reference to other values. Unfortunately, it does not explain what value is. In Evaluatology, if an evaluated object has a stakeholder, the value is interpreted as the derived effect of the evaluated object on the stakeholder. Please note that we have a formal definition of a stakeholder. Additionally, the client is only one stakeholder, which can not justify excluding the other clients.

D. The "rich description" approach

This view has been very widely supported–by Bob Stake [186], the North Dakota School, many of the UK theorists, and others. It claims that evaluation can be done as "a kind of ethnographic or journalistic enterprise, in which the evaluators report what they see without trying to make evaluative statements or infer to evaluative conclusions– not even in terms of the client's values (as the relativist can)."

I would like to interpret "done as a kind of ethnographic or journalistic enterprise" as a kind of interrogation. However, this view is very vague without explaining what the valid interrogation methodologies are, and how measurements or testing are performed under different interrogation conditions.

E. The "social process" school

By a group of Stanford academics led by Lee Cronbach, referred to here as C&C (for Cronbach and Colleagues [45], this view denied the importance of functions of evalua-

tion, "(i) as providing support for external decisions about programs, or (ii) to ensure accountability."

This view emphasizes denying the importance of the functions of evaluation; however, it fails to reveal the essence of the evaluation.

F. The "constructivist" or "fourth generation" approach

By Egon Guba and Yvonna Lincoln [75], it rejects evaluation as a search for quality, merit, worth, etc. Instead, they claim the evaluation outcome is "the result of construction by individuals and negotiation by groups."

This view ignores that the effect of any object is objective. It seems that they try to explain how a value (quality, merit, worth, etc) function is assigned to an object. In Evaluatology, the effect and the derived effect are both objective. The value is interpreted as the derived effect on the stakeholder.

G: A "transdisciplinary" view

Scriven held a transdisciplinary view [174] to treat evaluation as a tool discipline.

Scriven claimed that this view has four characteristics that distinguish it from the previous works [175].

First, it is "an objectivist view of evaluation, like A, holding that the evaluation is to determine the merit or worth of, for example, programs, personnel, or products." Unfortunately, when it comes to worth, merit, or value without an objective interpretation, it's easy to fall into the quagmire of subjectivity.

Second, the approach here is "a consumer-oriented view rather than a management-oriented (or mediator-oriented, or therapist-oriented) approach to program evaluation– and correspondingly to personnel and product evaluation, etc." From the perspectives of Evaluatology, the consumer, management, meditator, or therapist are different stakeholders; different views do not contradict. Instead, the effects or derived effects on different stakeholders could be inferred using the same methodologies.

Third, the approach here is "a generalized view." It is "not just a general view; it involves generalizing the concepts of evaluation across the whole range of human knowledge and practice." Unfortunately, Scriven did not propose general concepts, terminology, problem statements, axioms, mathematical notation, formulations, and methodologies. Instead, he still relies on the encyclopedic approach to define several hundred concepts or terms in [174], just as other evaluation communities did in [118].

Fourth, the transdisciplinary view is "a technical one." Unfortunately, Scriven fails to find a suitable technique language, like mathematics, to define the different categories of evaluation problems, and propose universal methodologies to address different evaluation problems.

2.2.3 Function

Scriven [172] was the first to suggest the distinction between "formative evaluation" and "summative evaluation," referring to two major roles or functions of evaluation, although he was not the first to realize the importance of such a distinction.

Referring to the same two functions, Stufflebeam [193] suggested "the distinction between proactive evaluation intended to serve decision-making, and a retroactive evaluation to serve accountability." Thus, evaluation can serve two functions, the "formative" and the "summative."

Robert E. Stake [184] distinguished the distinctions between formative and summative evaluation in an analogical manner from perspectives of different stakeholders: "When the cook tastes the soup, that is formative; when the guests taste the soup, that is summative." In its formative function, evaluation is used for the improvement and development of an ongoing activity (or program, person, product, etc.). In its summative function, evaluation is used for accountability, certification, or selection.

There are other discussions on the functions of evaluation from different perspectives.

Process evaluation focuses on "the activities and events during a program or intervention, investigating why and how a program or intervention achieves its results through documenting and collecting data [107, 118]."

Impact evaluation or assessment focuses on "the outcomes or impacts of an evaluand, e.g., a program, intervention, policy, organization, or technology, aiming to make a causal inference that connects the evaluand (the evaluated object) with the outcomes [107, 118]."

From the perspective of Evaluatology, there is only one unique function of evaluation: revealing the effect or derived effect of an (evaluated) object. From this angle, for formative evaluation or process evaluation, the evaluated object is the intermediate one in the different phases of creating an artifact, or the whole process of creating an artifact. While for summative evaluation or impact evaluation, the evaluated object is the object delivered. Though the evaluated objects differ, the methodology remains the same.

2.2.4 Role

In [174], Scriven considered evaluation as one of the most powerful and versatile of the "transdisciplines"—tool disciplines such as logic, design, and statistics. He claimed "Science itself is only distinguishable from pseudoscience by means of evaluation, by evaluation of the quality of evidence, research designs, instruments, interpretations, and so on [174]."

I agree with Scriven on the fundamental role of evaluation. I thought evaluation is one of the fundamental interrogation methodologies, just like measurement, testing, and reasoning. However, Scriven falsely classifies testing into the evaluation. Distinguishing science from pseudoscience is the fundamental role of testing, as we discussed in Chapter 5.

2.2.5 Collected Data

Previous work has extensively discussed how to collect data without deeply thinking about the different natures of the evaluated object, and hence, their thoughts are ad

hoc.

For example, Stufflebeam's CIPP Model [193] suggested that evaluation focuses on "four variables for each evaluated object: (a) its goals, (b) its design, (c) its process of implementation, and (d) its outcomes." According to this approach, an evaluation of an educational project, for example, would be "an assessment of (a) the merit of its goals, (b) the quality of its plans, (c) the extent to which those plans are being carried out, and (d) the worth of its outcomes."

Stake [184] in his Countenance Model suggested that two sets of information be collected regarding the evaluated object: descriptive and judgmental. The descriptive set should focus on "intents and observations regarding prior conditions that may affect outcomes, transactions, process of implementation, and outcomes." The judgmental set of information comprises "standards and judgments regarding the same prior conditions that may affect outcomes, transactions, and outcomes."

Guba and Lincoln [74], expanding Stake's Responsive Education Model [185] and applying the naturalistic paradigm. Guba and Lincoln [74] suggest that the evaluator focused on "five kinds of information: (a) descriptive information regarding the evaluation object, its setting, and its surrounding conditions, (b) information responsive to concerns of relevant audiences, (c) information about relevant issues, (d) information about values, and (e) information about standards relevant to worth and merit assessments."

2.2.6 Standards and Criteria

Having to choose the standards and criteria to judge the merit and worth of an evaluated object is one of the root reasons why evaluation is considered subjective.

Some scholars went straight to ignore the judgmental nature of evaluation. Those who defined evaluation as an information collection activity to serve decision-making or other purposes [2, 44, 192] did not have to deal with the problem of choosing evaluation criteria.

Other scholars used "goal achievement" as the evaluation criterion without having justified its being an appropriate criterion [206, 145]. They ignored the issue of evaluation criteria.

Several attempts have been made in recent years to develop standards and criteria for evaluations of educational and social programs [39, 195, 194, 198, 136]. Even though some scholars [45, 187] have criticized the rationale for the whole standard-setting effort as being premature at the present state of the art in evaluation, there seems to be a great deal of agreement regarding their scope and content.

Boruch and Cordray [22] analyzed six sets of such standards. They concluded that there has been a large degree of overlap and similarity among them. The Joint Committee on Standards for Educational Evaluation [195] developed and published the most elaborate and comprehensive set. Chaired by Dr. Daniel Stufflebeam, these standards committees consist of a committee of 17 members, representing 12 professional organizations associated with educational evaluation. The proposed 30 standards were divided into four major groups: "utility standards ensure that the evaluation serves practical

information needs; feasibility standards ensure that the evaluation is realistic and prudent; propriety standards ensure that the evaluation is conducted legally and ethically; accuracy standards ensure that the evaluation reveals and conveys technically adequate information."

Most evaluation experts seem to agree that the criterion (or criteria) to be used for the assessment of a specific object must be determined within the specific context of the object and the function of its evaluation. With different levels of acceptance, the evaluation criteria suggested by the literature include: identified needs of actual and potential clients [195, 138, 172], ideals or social values [74, 85], known standards set by experts or other relevant groups [74, 184], or the quality of alternative objects [85, 172].

Rossi et al. [152] discussed the criteria or standard for program performance, which may manifest in different forms for various dimensions of program performance. The criteria or standards could be "the needs or wants of the target population, stated program goals and objectives, professional standards, customary practice, norms for other programs, legal requirements, ethical or moral values, social justice, equity, past performance, historical data, targets set by program managers, expert opinions, pre-intervention baseline levels for the target population, conditions expected in the absence of the program (counterfacutal), cost or relative cost." The effectiveness of a social program is gauged by the change it produces in outcomes that represent the intended improvements in the social conditions it addresses.

In Chapter 10, I posit that comparison is the most primitive interrogation, on which the principle and methodology of measurement and testing rely. From this perspective, comparing has nothing to do with subjectivity, as measurement and testing are considered objective. According to Evaluatology, we can create reference evaluated objects and compare the evaluated object to the reference one.

2.2.7 Process

The evaluation process is ad hoc and differs according to the different views of evaluation.

A theoretical approach perceiving evaluation as an activity intended to determine whether goals have been achieved [206] might recommend the following evaluation process: "(a) stating goals in behavioral terms, (b) developing measurement instruments, (c) collecting data, (d) interpreting findings, and (e) making recommendations."

According to Stake's Countenance Model [184], the evaluation process should include: "(a) describing a program, (b) reporting the description to relevant audiences, (c) obtaining and analyzing their judgments, and (d) reporting the analyzed judgments back to the audiences."

Later on, in his Responsive Evaluation Model Stake [185] suggested a continuing "conversation" between the evaluator and all other parties associated with the evaluand. He specified 12 steps of dynamic interaction between the evaluator and his audience in the process of conducting an evaluation.

Provus [145] proposed "a five-step evaluation process, including (a) clarification of the program design, (b) assessing the implementation of the program, (c) assessing

its in-term results, (d) assessing its long-term results, and (e) evaluating its costs and benefits."

The Phi Delta Kappa Study Committee on evaluation [194] presented a three-step evaluation process. It included "(a) delineating information requirements through interaction with the decision-making audiences,(b) obtaining the needed information through formal data collection and analysis procedures, and (c) providing the information to decision-makers in a communicable format."

Scriven [170] has suggested nine steps in his Pathway Comparison Model. Guba and Lincoln [74] suggest that a naturalistic responsive evaluation should be implemented through a process including the following four stages: "(a) initiating and organizing the evaluation, (b) identifying key issues and concerns,(c) gathering useful information, and (d) reporting results and making recommendations."

Rossi et al. [152] considered that the evaluation of a program "generally involves assessing five domains: the need for the program; its design and theory; its implementation and service delivery; its outcome and impact; and its efficiency."

The evaluation process in Evaluatology has a universal process, regardless of the evaluated objects.

2.2.8 Methodologies

In Part IV, we summarize other widely used evaluation methodologies. They include Design of Experiments (DoE), RCTs, instrumental variables, potential outcomes, quasi experiments, structural causal models, and structural equation models. Uninformatively, there lack of a systematic evaluation of those evaluation methodologies, which belong to the category of meta-evaluation.

2.2.9 Ad Hoc Practices in Different Domains

This subsection presents a concise overview of ad hoc evaluation practices in different domains, partially based on our previous work [224].

In the Field of Business

Camp [28] defines benchmarking as "the search for those best practices that will lead to the superior performance of a company." Benchmarking consists of two primary steps [28]: (1) establishes operation targets based on industry best practices; (2)"a positive, proactive, structured process leads to changing operations and eventually achieving superior performance and competitive advantage." In the study conducted by Andersen et al. [4], the essence of benchmarking is summarized as "the quest for knowledge and learning from others."

From the perspectives of Evaluatology, the process, the individual, policies, or the intermediate objects in business could be evaluated. The so-called best practices are also an evaluated object. The targets of the operations are some forms of the effects

of the evaluated objects. "A positive, proactive, structured process" is essentially the process of uncovering the effects of and performing trials on different evaluated objects, e.g., policies, to "change operations and eventually achieve superior performance and competitive advantage."

In the Field of Finance

In the fields of finance and education, indices are widely used as benchmarks to assess the overall performance of individuals or systems under study. These indices are derived by calculating the weighted average of a selected group of individuals or systems [58].

For example, stock market indices are used as benchmarks to assess the stock market's performance in the finance field. These indices are derived by calculating the weighted average of a selected group of representative stocks [58]. Some widely recognized stock market indices include the Dow Jones Industrial Average, the S&P 500, the NASDAQ Composite, and the Shanghai Stock Exchange Composite Index. Different indices employ varying calculation methods. The most common approach is the weighted average method, which determines the index value based on the weighted average of the constituent stock prices. Another method is the geometric mean method, which calculates the geometric average of the stock prices and adjusts it using a base period price. Typically, stock market indices are published at the close of each trading day. Some index providers offer real-time index data, enabling investors to stay informed about the latest market conditions.

The Brent benchmark is used to determine the price of Brent crude oil [168]. Brent crude oil is a type of light and low-sulfur crude oil produced from oil fields in the North Sea region. Due to its relatively stable supply and high quality, Brent crude oil has become a significant benchmark in the international oil market. Traders, investors, and industry participants worldwide reference the Brent benchmark to track and evaluate the price of Brent crude oil.

In finance, indexes or benchmarks are essentially the reference objects through which we compare, as we discussed in Chapter 10.

In the Field of Social Sciences:

According to Rossi et al., [152], at the earliest, Thomas Hobbes and his contemporaries tried to "use numerical measures to assess social conditions and identify the cause of mortality, morbidity, and social disorganization in the discipline of social science."

Rossi et al. [152] define program evaluation as the process of using social research methods to systematically assess programs aimed at "improving social conditions and our individual and collective well-being," to provide answers to the stakeholders. Rossi et al. [152] summarize the five domains of evaluation questions and methods that exhibit strong interplays: "(1) the need for the programs, (2) program theory and design, (3) program process, (4) program impacts, and (5) program efficiency."

From the perspective of Evaluatology, the essence of program evaluation is to uncover

the effect of the social program on the social conditions, individuals, and collective well-being.

However, from the perspectives of Evaluatology, the five domains of evaluation questions should be addressed with different interrogations, not a single evaluation. The need for the program could be interrogated by measurement. The program theory needs to be tested before implementation. The program process or its components could be evaluated to uncover their effects.

In the Field of Computer Science

Within the computer science field, there are varying viewpoints and perspectives. For example, Hennessy et al.[80] highlight the significance of benchmarks and define them as "programs specifically selected for measuring computer performance." On the other hand, John et al.[94] compile a book on performance evaluation and benchmarking without providing formal definitions for these concepts.

Kounev et al.[104] present a formal definition of benchmarks as "tools coupled with methodologies for evaluating and comparing systems or components based on specific characteristics such as performance, reliability, or security." The ACM SIGMETRICS group[26, 102] considers performance evaluation as "the generation of data that displays the frequency and execution times of computer system components, with a preceding orderly and well-defined set of analysis and definition steps."

The SPEC CPU benchmark, known as SPEC CPU [181], is widely recognized as the most renowned benchmark suite for CPU performance evaluation. Throughout its history, six versions of the SPEC CPU benchmark suite have been released, with the latest version being SPEC CPU2017 [182]. The SPEC CPU workloads cover a broad range of compute-intensive tasks.

The performance evaluation metric used in SPEC CPU is based on the execution time. The reported score of SPEC CPU represents the ratio of its execution time compared to that of a reference machine. To ensure the credibility of the results, the overall metrics are calculated as the geometric mean of each respective ratio. Each ratio is based on the median execution time from three runs or the slower of the two runs [182].

Dongarra et al. [51] proposed the LINPACK benchmark for evaluating high-performance computing (HPC) systems. The LINPACK Benchmark is designed to solve dense linear systems of equations of order n, represented by the equation $Ax = b$. It originated from the development of the LINPACK software package in the 1970s.

From the perspectives of Evaluatology, the benchmarks in the discipline of computer science are a component of simple evaluation conditions, that is, essential external objects (EXOs), in Chapter 14.3.

In the Field of Artificial Intelligence

As shown in Figure 2.1 [1], ImageNet is a significant benchmark in the field of computer vision, consisting of 14,197,122 high-resolution images manually annotated across 21,841

[1] Dr. Chunjie Luo contributed to this figure. He is one of the authors of our previous work [224].

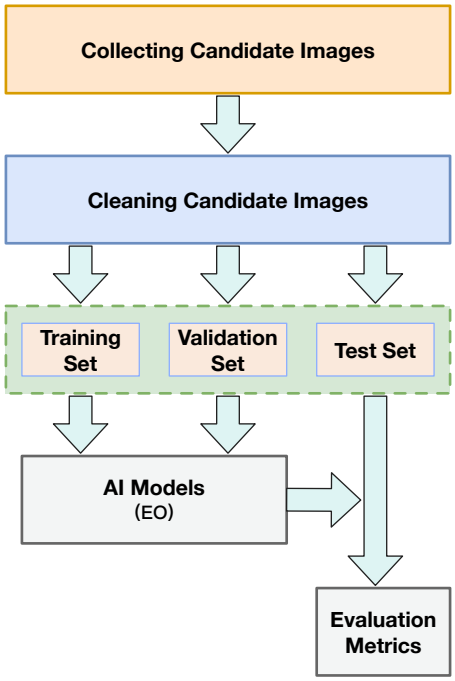

Figure 2.1: The ImageNet evaluation working process.

distinct categories, commonly known as ImageNet-21K [47]. These categories encompass a wide range of objects, animals, and scenes.

The ILSVRC (ImageNet Large Scale Visual Recognition Challenge) is an annual computer vision competition that focuses on a subset of ImageNet-21K called ImageNet-1K [164]. It aims to evaluate the performance of deep learning models in tasks such as image classification and object detection, providing specific task configurations and evaluation criteria.

ImageNet-1K is primarily used for image classification tasks and consists of 1,281,167 training images, 50,000 validation images, and 100,000 test images. The evaluation metrics commonly used in ILSVRC include Top-1 accuracy, which measures the match between the predicted category and the true category of the image, and Top-5 accuracy, which indicates if the true category of the image is among the top five predicted categories by the model.

From the perspectives of Evaluatology, the benchmarks in the discipline of artificial intelligence are a component of simple evaluation conditions, similar to those of the discipline of computer science.

In the Field of Medicine

The evaluation in the field of medicine can be traced back to the early medical eras, although there are no documented records. A rigorous modern medical evaluation methodology and system were established as early as 1938 [32]. Clinical trials, with a history

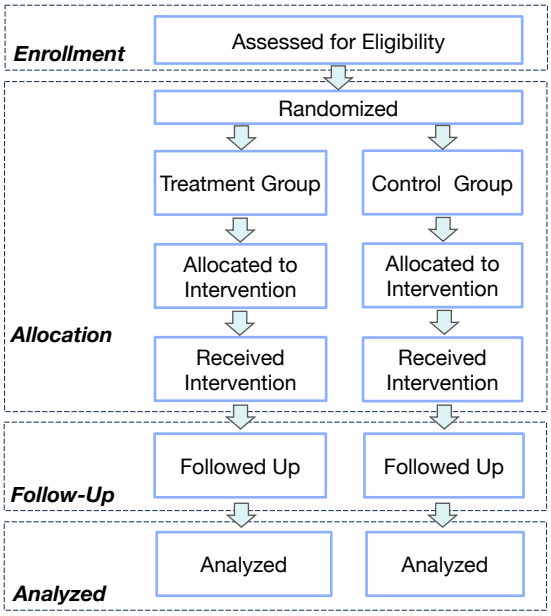

Figure 2.2: The randomized controlled trials (RCTs) evaluation process. [123]

spanning over 250 years, are the primary and widely recognized method for medical evaluation. They are defined as experimental designs to evaluate the potential impact of medical interventions on human subjects [93].

Currently, clinical trials based on experimental designs can be categorized into various types, including randomized trials, double-blind trials, prospective trials, and retrospective trials [178].

As illustrated in Figure 2.2, Randomized Controlled Trials (RCTs), considered the gold standard for medical evaluation, possess a rigorous and reliable theoretical framework [124]. However, their high time and financial costs limit their application. In addition, RCTs are difficult, if not impossible, to apply in Physics, Chemistry, or Biology to trace the causes of many effects [2].

To compensate for the shortcomings of RCTs, emerging clinical evaluation methods, such as Real-World Data assessment and digital clinical trials, have been proposed [90, 146]. These novel medical assessments are still in their early stages and have noticeable deficiencies in their theoretical foundations, such as a lack of rigor and reliability.

In the field of Psychology

In the field of psychology, social and personality psychologists often rely on scales, such as psychological inventories, tests, or questionnaires [65], to evaluate psychometric variables [65]. These variables include attitudes, traits, self-concept, self-evaluation, beliefs,

[2] This comment is inspired by Dr. Lei Wang in a talk after we had lunch together on a day in the year of 2025.

abilities, motivations, goals, social perceptions, and more [65]. Essentially, the essence of the scale is a component of simple evaluation conditions.

While these tools are commonly used, it is important to recognize that they rely on virtual assessments and self-report-style evaluations, which may introduce potential distortions.

To overcome this limitation, I suggest implementing a physical application of an EC to the evaluated object, supplemented with a variety of measurement instruments. This approach aims to provide a more objective and accurate assessment of various aspects, including attitudes, traits, self-concept, self-evaluation, beliefs, abilities, motivations, goals, and social perceptions [65], by incorporating tangible and observable data.

2.3 Why Past Efforts Failed to Establish the Discipline of Evaluation?

In [172], dated back to 1966, Scriven thought evaluation is "a logical activity which is essentially similar whether we are trying to evaluate a coffee machine or teaching machines, plan for a house or plan for a curriculum. The activity consists simply of the gathering and combining of performance data with a weighted set of goal scales to yield either comparative or numerical ratings." That is the first rudimentary idea on the discipline of evaluation. Since then, Scriven has published many articles towards the goal of establishing the discipline of evaluation from the 1960s to the 2010s [172, 170, 171, 173, 174, 175, 176, 177].

In their 2010 book [56], Chinese Scholars Junping Qiu et al. discuss what scientific evaluation means in Chinese. They used a Chinese Term similar to Evaluatology. Overall, they discussed and summarized many concepts and ad-hoc methods in social sciences, education, and bibliometrics. Unfortunately, they overlooked Scriven's work and failed to notice many rigorous evaluation methodologies which we discussed in Part IV.

In a summary article, dated back to 2016, Scriven [177] claimed the discipline of evaluation is established; however, I think he overstates the situation.

In [212, 224], Wang et. al showed that the state-of-the-art and state-of-the-practice evaluation methods cannot even achieve a true evaluation outcome for a specific artifact, a computer component, like a CPU. Instead, Different areas are still using ad hoc empirical methodologies as we have discussed in Section 2.2.9.

The good news is that many stringent methodologies have been developed, like DOE, RCTs, or the potential outcome framework, which we will summarize in Part IV; however, no previous work has systematically evaluated those evaluation methodologies. In Chapter 13, we consider the meta-evaluation of different evaluation methodologies as one of the most important issues in Evaluatology.

I believe two fundamental reasons contribute to the failure of past efforts to establish the discipline of evaluation.

First, the consensus of the evaluation community referred to evaluation as "the process of determining the merit, worth, or value of things" [172, 174, 175, 176, 55, 68, 85,

196, 195]. Scriven [175] claimed that it is the value-free doctrine that prevents the development of the discipline of evaluation, that is, the science and engineering community rejected to introduce the word of value into their territory.

However, the current definition of evaluation, which is the consensus of the evaluation community, fails to uncover the essence of evaluation. The words of "merit, worth, or value" in nature are subjective—varying from different people—that is one of the root reasons that contribute to the failure.

Secondly, Scriven [174, 175, 176] envisioned that there is a core subject in the discipline of evaluation. However, the core subject of the evaluation was never articulated. The past work [174, 175, 176, 56] failed to propose universal concepts, terminology, problem statements, axioms, mathematical notation, formulations, and methodologies. For example, without those universal ones, Michael Scriven relies on the encyclopedic approach to define several hundred concepts or terms in [174], just as other evaluation communities did in [118].

2.4 Summary

This chapter overviews evaluation's current state, delving into its historical roots and theoretical foundations. It examines evaluation as an ancient practice, but an undeveloped discipline, and explores various evaluation concepts, theories, and ad hoc practices in different domains. Finally, it addresses the reasons behind past failures in establishing evaluation as a discipline.