Part III

The Science of Evaluation

Chapter 8

Basic Evaluation Concepts and Problem Statements

In this chapter, I present the basic evaluation concepts and problem statements.

8.1 Basic Concepts

For objects A and B, when measurable or testable differences occur in B depending on the presence or absence of A, I define A as the *cause object (in short, cause)*, B as the affected object (in short, AO), and the measurable or testable difference in B as the *effect* on B induced by A. Additionally, I define the *effect mechanism* as the way through which the cause object induces the effect on the AO. When I mention *mutual influences between objects* A *and* B, I refer to both A's effect on B and B's effect on A. Figure 8.1 shows the relationships among the cause, AO, effect, and effect mechanism.

Confounding occurs when two independent objects are associated in a manner that makes it challenging to differentiate their specific effects on a third object. In other words, the effects of these independent objects become entangled, making it difficult to attribute specific effects to each one.

In statistics, a *correlation* measures the mutual effect of variables a and b, among which many confounding variables exist. Correlation does not imply a cause-and-effect relationship.

Several Examples on Cause, AO, Effect, and Effect Mechanism.

An Physics Example

Issac Newton's Apple Tree: The *cause* is the earth, the *AO* is the apple, the *effect* is the gravitation, the *effect mechanism* is the Law of Universal Gravitation.

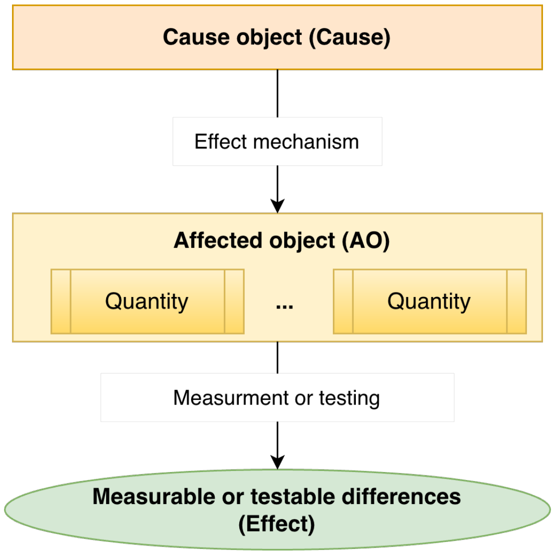

Figure 8.1: The relationships among Cause, AO, Effect, and Effect mechanism.

A Chemistry Example

Lavoisier's Combustion Experiment: The *cause* is oxygen, the *AO* is any substance that can be burned or calcined, the *effect* is combustion, the *effect mechanism* is the chemical reaction equation for combustion.

A Biology Example

Gregor Johann Mendel's Peas: The *cause* is the gene, the *AO* is the pea, the *effect* is the trait of the pea, and the *effect mechanisms* are the laws of segregation and independent assortment.

A Computer Science Example

CPU evaluation: The *cause* is the CPU, the *AO* is the computer system, the *Effect* is the CPU's effect on the measured overall performance metrics, and the *effect mechanism* is the collaboration mechanisms among the CPU and other indispensable components.

An Artificial Intelligence Example

Evaluating an AI algorithm: The *cause* is the AI algorithm, the *AO* includes the dataset labeled with ground truth, the algorithm implementation, the OS, the processor, and the memory. The *effect* is the AI algorithm's effect on the overall algorithmic accuracy, and the *effect mechanism* is the capability of an algorithm to fit input-output relationships.

A Medicine Example

Drug evaluation: The *cause* is the drug, the *AO* is the patients, the *Effect* is the measurable or testable difference in patients, and the effect mechanism is some known or unknown biological mechanism.

A Social Science Example

Policy evaluation: The *cause* is the policy, the *AO* is different participants, the *effect* is the policy effect on overall evaluation outcomes, such as the approval rating, and the *effect mechanism* is embodied in many forms, like political, social, or psychological ones.

8.2 Essence and Problem Statement

Based on the fundamental concepts presented in Chapter 8.1, I will define the concept of evaluation in this section and formally state the problem of evaluation.

When evaluating an object A, we need to identify the set of other objects B_i on which differences can be measured or tested, depending on the presence or absence of A.

We call A, the *evaluated object (in short, EO)* or the *cause object (in short, cause)*, B_i, the AO, and the measurable or testable difference in B_i, the *effect* on B_i induced by A. Essentially, the problem of evaluation can be formally stated as *how to uncover the effect of the EO on the AO.* If the instantiations of an object do not induce different effects, we do not distinguish between an object and its instance.

In addition to the EO A, other objects, A_j, also affect the AO, B_i, and we call those objects, the *essential external objects (in short, EXO).* We call the effect induced by both the EO and the EXO on the AO *the overall effect.*

An EO exerts effects on numerous other objects. We categorize two distinct types: *a direct AO* and *an indirect AO*. A direct AO is *a minimal system on which the effect of the EO can be directly measured or tested.* An AO is either a direct AO or an indirect AO.

I formally call *the effect of the EO on the direct AO*, *the derived EO*, and call *the effect of the derived EO on the indirect AO*, *the derived effect.* In this context, the essence of evaluation is *how to uncover the effect of the EO on the direct AO and its derived*

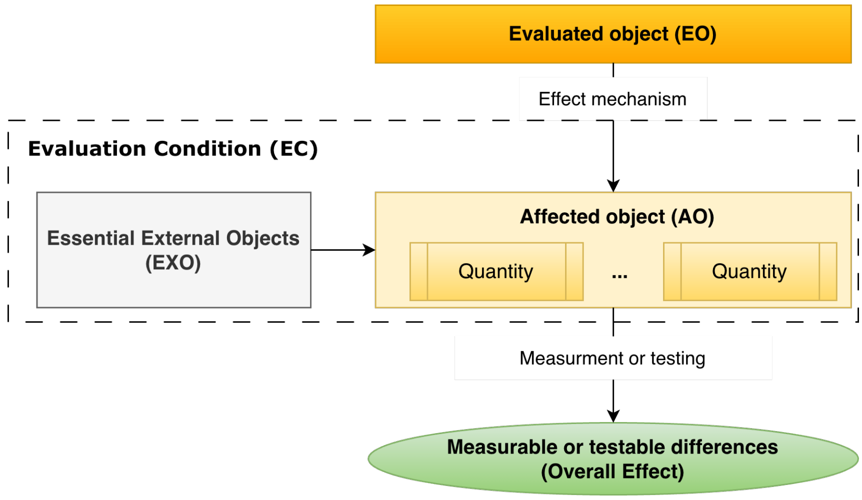

Figure 8.2: The fundamental components of an SES.

effect.

Some Examples on Direct AO and Indirect AO.

A Computer Example

A group of fans aims to snatch up concert tickets for their favorite superstar using computer devices equipped with different CPUs. An evaluation of the CPUs in this scenario is required. Here, the CPU serves as the EO, the computer system running the ticket-snatching program acts as the Direct AO, and the fans attempting to purchase the tickets function as the indirect AO.

A Physics Example

Issac Newton's Apple Tree: The EO is the Earth, the direct AO is the apple, the indirect AO is Issac Newton.

In reality, when measuring or testing the *effect* on B_i induced by A, we can not isolate the effect of the EXO, A_j, from that of the EO, A, on the direct AO, B_i. Instead, we have to take a holistic approach and consider the EO, A, the direct AO, B_i, and EXO, A_j, together as an entire system, under which we infer the true effect of the EO on the direct AO from the overall effect. We formally name this entire system, *a self-contained evaluation system (in short, SES)*. Upon the removal of the EO from the SES, we call the derived system the evaluation conditions (in short, EC). We illustrate the fundamental components of an SES in Figure 8.2.

If an EO has an indirect AO, a derived problem to be addressed is *how to infer the derived effect*. The same methodology mentioned above can be utilized to uncover the derived effect. We call the other objects that impact the derived effect, *the derived*

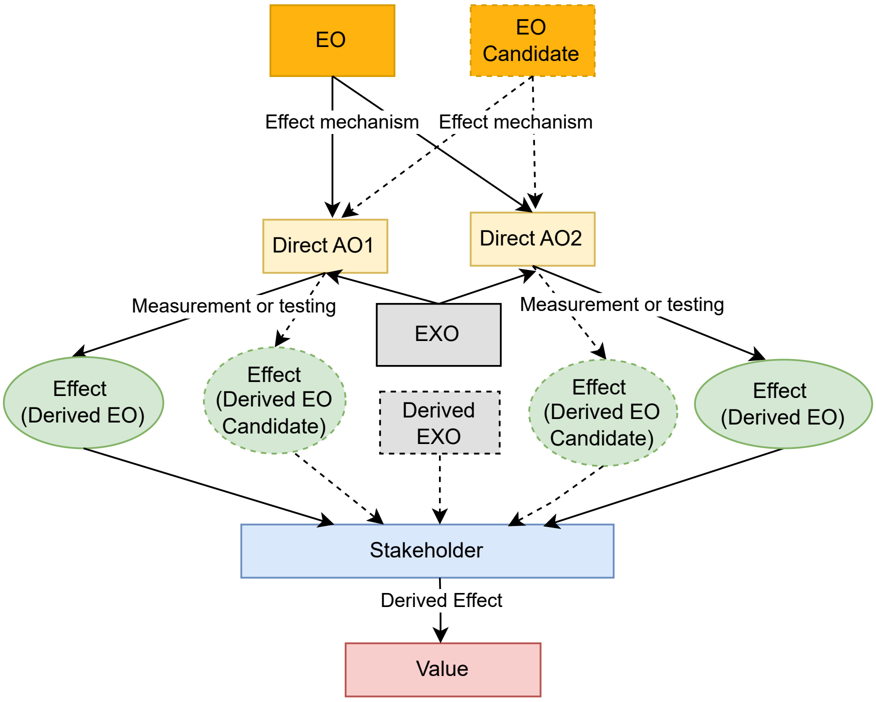

Figure 8.3: Essentially, the value is the derived effect on the stakeholder.

EXO. We isolate a derived SES, which includes the derived EOs, the indirect AO, and the derived EXO. Upon the removal of the derived DO from the derived SES, we call the derived system the derived EC. Within the context of this derived SES, we can infer the derived effect. However, it may include many direct or indirect AOs; the effect of the EO could be passed through several indirect AOs, and the chains could be long.

A *stakeholder* is an intelligent life or a system consisting of intelligent life that holds a stake of responsibility or interest in the EO. If an EO has a stakeholder, I call the effects of the derived EOs on the stakeholder, *the value of the EO*, as a stakeholder could make free choices among different EO candidates. The problem could be further stated as *how to infer the value of the EO*.

I formally call the effects of other EO candidates on the direct AOs *the derived EO candidates* when the stakeholder makes different choices. The same methodology that infers the derived effect still works, albeit with a slight difference, as other indirect AOs differ from the stakeholder in that they cannot make free choices. To infer the value of an EO on a stakeholder, a derived SES needs to include the derived EOs, the derived EO candidates, and the derived EXOs, under which the value could be inferred.

In summary, the essence of the evaluation is to *uncover an EO's effects and derived effects*.

A Typical Example on Drug Evaluation.

The First Step

When evaluating a drug, the first step is to identify a set of objects, e.g., a specific patient or a group of patients, on which differences can be measured or tested, depending on the presence or absence of that drug. The differences could be treatment outcome or cost.

In this case, the drug is the cause, a specific patient or a group of patients is the direct AO, and the measurable or testable difference in a specific patient or a group of patients is the effect induced by the drug. Often, identifying the mechanism of the effect of the drug is very challenging in biology.

For the treatment outcome, the EXO may include the working conditions, which may exert different pressures on the patient, and the living conditions, which may have different air or water quality.

In reality, we can not isolate the effect of the working conditions and living conditions. So, the SES contains the EO, the AO, and the EXO together, under which we reveal the effect of the drug.

The Second Step

In the drug evaluation case, the stakeholders could be patients' families, hospitals, or the government. The second step is to reveal the effect of the derived EO on the stakeholder. To address this issue, we need to isolate a derived SES, which includes the derived EOs, the derived EO candidates, the stakeholder, and the derived EXOs. Within the derived SES, we can infer the value of the EO.

8.3 Summary

This chapter presents the basic concepts of evaluation, including the cause, effect, and effect mechanism. I reveal that the essence of evaluation is how to uncover an EO's effects or derived effects.

Chapter 9

Evaluation Assumptions and Axioms

In this chapter, I present the assumptions and axioms of evaluation.

9.1 Assumption of Three Worlds

In this section, I elaborate on the assumption of the three worlds. The assumption of the three worlds states that there are three worlds: *the microscopic object world*, *the normal object world*, and *the free will world*, which are governed by consistent but different laws. Essentially, a law is a model.

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* [20, 122] [1].

Microscopic objects possess four properties, commonly referred to as *quantum effects*, which are fundamentally different from those of normal objects.

The first property is quantization, that is, energy, momentum, angular momentum, and other countable properties of microscopic objects are not continuous but come in discrete packets called "quanta."

The second property is wave-particle duality, that is, microscopic objects, like electrons and photons, exhibit both particle-like properties (e.g., having a location, colliding) and wave-like properties (e.g., interference, diffraction).

The third property is quantum superposition, that is, a microscopic object, or so-called quantum system, like an electron or photon, can exist in a combination of multiple possible states at the same time until it is measured. Upon measurement, it "collapses" into just one of those possible states.

Schrödinger's cat [169] is a famous example of microscopic objects, as being simultaneously both alive and dead until the box is opened and observed.

The fourth property is quantum entanglement, that is, two or more microscopic objects, like particles, can become linked or entangled in such a way that the quantum

[1] Mr. Chenxi Wang also contributed to the investigation of the principles of quantum mechanics.

state of one particle is instantly correlated with the state of the other(s), no matter how far apart they are. Measuring one instantly determines the state of the other.

Some Examples on the "Strange" Properties of Microscopic Objects

Electrons Jump Between Energy Levels

Electrons in an atom can only occupy specific energy levels (orbits). They jump between these levels by absorbing or emitting a quantum of light (a photon) [61].

Double-Slit Experiment

In the Double-Slit Experiment [30], a single electron fired at two slits behaves like a wave and creates an interference pattern on a screen, as if it passed through both slits simultaneously. Yet, it hits the screen at a single point like a particle.

Photons in an Entangled Pair

When the polarization direction of one photon in an entangled pair [23] is measured and determined, the polarization direction of the other photon immediately becomes correlated in a non-local and instantaneous manner [220].

Quantum Tunneling Effect

In semiconductor devices, electrons can traverse a potential barrier that is higher than their own energy, a phenomenon known as the quantum tunneling effect [216]. Even when an electron's energy is lower than the height of the potential barrier, there is still a certain probability that it will appear on the other side of the barrier.

A normal object is the object at the normal scales without free will. A normal object world consists of normal objects, which are determined and governed by *the principle of cause and effect*. This principle implies that an object as a cause will induce an effect on the other objects that can be objectively measured or tested. Please note that an automatic object, which we defined in Section 3.3, is a normal object, though it owns interrogation capacity and capability with different levels of complexity. Section 8.1 provides many examples of normal objects.

Quantum effects of microscopic objects reveal a universe far stranger and more probabilistic than that of the normal objects. At the smallest scales, certainty that dominated the normal objects gives way to probability distributions. Microscopic objects exhibit both particle-like and wave-like properties, exist in multiple states at once, and can be mysteriously connected across vast distances [20, 122]. These models of microscopic objects, while counterintuitive, are incredibly successful at explaining and predicting the behavior of matter and energy at the atomic and subatomic level.

A free will world consists of the normal objects and the intelligent lives, and a free will world is governed by *both the principle of cause and effect and 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

Typical examples of free will include love, writing, scientific research, art, voting, commodity purchasing, and college/program application in human society.

9.2 Evaluation Axioms

In this section, I present five evaluation axioms that are derived from the essence of evaluation, which I explain in Chapter 8. They focus on key aspects of evaluation outcomes, including true quantity (The first and second axioms), traceability of discrepancy (the third axiom), comparability (the fourth axiom), and estimate (the fifth axiom).

9.2.1 First Axiom of Evaluation Outcome

This axiom declares that *the essence of the evaluation outcome either uncovers an EO's effect or derived effect that carries inherent physical significance or is solely dictated by the value function.*

In nature, an evaluation metric refers to a base quantity, which indeed possesses physical significance, or other quantities that possess physical significance, or a combined quantity that is constructed using base quantities and other quantities.

There are two categories of evaluation metrics. In the first category, an evaluation metric carries inherent physical significance, revealing the effect of the EO on the direct AO, that is, the change of countable property of the direct AO induced by the EO.

In the second category, an evaluation metric reveals the effect of the derived EOs on the indirect AO or stakeholders. An EO can induce different effects on the same EO; that is, the EO's different quantities are affected. An EO can induce the same or different effects on different AOs. Those effects can be passed to the same indirect AO or stakeholder.

In this context, those effects are the changes in different countable properties of the AOs. The value function serves as a mechanism that maps base quantities and other quantities carrying physical meaning into a composite evaluation metric. The composite evaluation metric does not necessarily carry inherent physical significance. Instead, it is defined as a value function that could be interpreted in a derived SES.

9.2.2 Second Axiom of True Evaluation

This axiom declares that *for an EO, when its SES is known, the effect of the EO on the direct AO possesses true outcomes; If an EO has an indirect AO and its derived SES is known, the derived effect on the indirect AO possesses true outcomes.*

In Section 8.2, I have formally defined what an SES is. Hereby, I further explain what I mean by referring to "when its SES is known." This proposition has a two-fold meaning. First, an SES is known, indicating that an SES and its components are known by the subject other than itself. Second, the effects of the EO and EXO, that is, the crucial components of the SES, on the direct AO can be measured or tested by the subject other than itself.

When revealing the effect of the EO, if we can isolate the SES, all relevant objects that have effects on the AO are identified, and their effects on the direct AO can be measured or tested. In this way, we can derive the true outcome of the effect.

Suppose the EO has an indirect AO. When uncovering the effect of the derived EO on the indirect AO, the same methodology can work.

Otherwise, the effect or the derived effect is undefined, and the subject other than the EO can not infer it.

9.2.3 Third Axiom of Evaluation Traceability

This axiom declares that *for the same (derived) EO, the divergence in the (derived) Effect can be attributed to disparities in (derived) ECs, thereby establishing evaluation traceability.*

This axiom focuses on the traceability of discrepancies in the effect. For the same EO, when the direct AO and the EXO are defined, the effect of the EO has a true outcome according to the First Axiom of Evaluation. The disparities in the overall effect of the EO can be rationalized as the consequence of variations in the ECs. In the absence of this axiom, the differences observed in the overall effect would be inexplicable, contradicting our scientific and engineering intuitions.

For the derived EO, the disparities in derived effects can be rationalized as the consequence of variations in the derived EC. The axiom holds for the same reason.

9.2.4 Fourth Axiom of Comparable Evaluation Outcomes

This axiom declares that *when different pairs of well-defined (derived) EO and direct (indirect) AO are equipped with equivalent (derived) EXO, their (derived) effects are comparable.*

It goes without saying that this axiom is related to the comparability of the evaluation outcomes. The equivalent EXO provides the same reference. When a pair of well-defined

EO and direct AO is equipped with the equivalent EXO, its evaluation outcomes possess true values.

Additionally, when two pairs of well-defined EO and direct AO are subjected to the equivalent EXO, their evaluation outcomes accurately reflect the effects of the EO on the direct AO under the same conditions, making them comparable.

When two pairs of well-defined derived EO and indirect AO are subjected to equivalent derived EXO, their derived evaluation outcomes accurately reflect the effects of the derived EO on the indirect AO under the same conditions, making them comparable.

9.2.5 Fifth Axiom of Consistent Evaluation Outcomes

This axiom asserts that *when uncovering the effect of a known (derived) EO on a known direct (indirect) AO using different samples from a population of (derived) EXOs, their (derived) effect consistently converges towards the true quantities.*

This axiom provides an estimate of the true outcomes of the (derived) effect under the population of (derived) EXOs. According to the Second Axiom of Evaluation, when a pair of the known EO and the direct AO is equipped with a population of known EXOs, the effect possesses the true outcomes. When a sample is taken from a population of EXOs, it serves as an approximation of the entire population. As a result, different samples yield consistent effect outcomes that converge towards the true outcomes of the entire population of EXOs. This convergence is influenced by the sample's ability, which is determined by the chosen sampling policy, to represent the underlying characteristics of the population accurately.

When a pair of a known derived EO and a known indirect AO is equipped with a population of the known derived EXOs, the derived effect possesses the true outcomes. The axiom holds for the same reason.

9.3 Summary

This chapter presented the assumption of three worlds and the five axioms of evaluation.

Chapter 10

Revisiting Interrogations

In this chapter, I revisit and redefine several concepts and clarify the unique positions and interrelationships of four interrogations: measurement, testing, reasoning, and evaluation.

10.1 Revisiting Several Concepts

As the essence of the evaluation is to uncover the effects of objects, this section will extend several definitions in Chapter 5.

A *proposition* is a testable statement about an object or *several objects with mutual influences*.

A *model* is a streamlined representation of an object or *several objects with mutual influences* [221, 224]. A model can manifest as a physical, mathematical, or other construct, e.g., a causal model. A valid model that can be verified objectively by a subject other than itself is a kind of *fact or truth*. A model about an object or *several objects with mutual influences* could be a hypothesis under test, which is defined in Section 3.6.

A *fact or truth* is a proposition or a model about an object and *several objects with mutual influences* that can be proven true or verified objectively by a subject other than itself.

For instance, the fact that water boils at 100 °C under standard atmospheric pressure describes an object (water) and its interaction with another object (air) and can be objectively verified by the observer (the subject) using a thermometer and a barometer.

A *test oracle* is the ground truth or facts about an object or *several objects with mutual influences*. Additionally, for artifacts, a test oracle could be the intended behavior expected for an object or several objects with influences.

For instance, in a system with multiple interacting objects, such as a temperature sensor (Object A) and a cooling fan (Object B), the test oracle defines the expected behavior between them: when the temperature exceeds 50 °C, the fan must switch to high speed ($\geq$3000 RPM) within one second, and when the temperature falls below 30

°C, it must return to low speed (≤1500 RPM).

A *test case* is a predefined interrogation condition for testing that is designed and implemented for a test oracle and is ready for executing an object or *several objects with mutual influences* under test to verify whether the actual outcomes are consistent with the mandated or expected results defined by the test oracle.

For instance, a corresponding test case may be designed for the test oracle by raising the temperature from 25°C to 45°C and expecting the fan to reach at least 3000 RPM within one second.

10.2 Redefinition of Testing and Reasoning

Based on the extended concepts, I redefine testing and reasoning.

Testing is a verification process of running test cases to determine whether a proposition or a model of an object or *several objects with mutual effects* conforms to the test oracle through comparing their outcomes [224, 223].

The proposition or a model states what is expected for an object or several objects with mutual influences under an interrogation condition. The test oracle mandates the reference outcome under the reference interrogation conditions. The essence of testing is to compare the interrogation outcomes to those of the test oracle.

Being testable or testability means a proposition or a model of an object or several objects with mutual influences can be falsifiable through comparing its outcome with that of test oracles.

Reasoning is a capacity, capability, and mental process that generates new propositions about an object or *several objects with mutual influences* based on the propositions, models, and other interrogation outcomes.

The definition of measurement remains unchanged.

10.3 The Unique Position of Four Interrogations

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, or mutual influence between objects; reasoning answers "why" in terms of the the underlying logical mechanisms that connect causes to their effects. Together, these four interrogations form a complete cognitive cycle—from observation, to

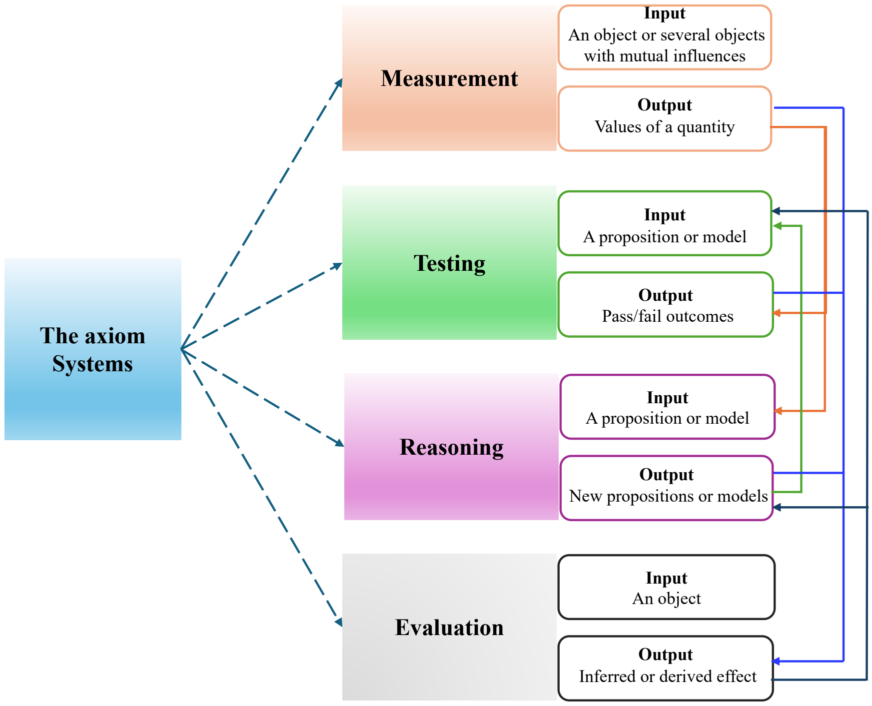

Figure 10.1: Interrelations of the four interrogations.

validation, to explanation [1].

10.4 Interrelations of the Four Interrogations

The interrelations of the four interrogations are depicted in Figure 10.1 [2]. Measurement, testing, reasoning, and evaluation are built on an axiom system that is accepted as the truth without proof.

Measurement attributes the value to the quantity of an object or several objects with mutual influences, based on which the subject generates propositions or models that serve as the input for reasoning, or compares the outcomes of testing cases to those of test oracles to draw a testing conclusion.

Through reasoning, the subject could generate new propositions or models about an object or several objects with mutual influences, which can only be validated by testing.

Evaluation relies upon measurement, testing, and reasoning to infer the true effect of an object. Evaluation reveals the effect of objects, which could generate the propositions or models that serve as the input for testing or reasoning.

[1] Dr. Fanda Fan's work provided a basis for this passage.
[2] Dr. Wanling Gao contributed to this figure based on the discussion with Dr. Lei Wang and me.

Testing could generate new propositions about an object or several objects with mutual influences. From this perspective, testing not only validates reasoning but also provides new input for reasoning.

10.5 Summary

This chapter expands the definitions of testing and reasoning to encompass several objects with mutual influences. Also, I clarify the unique positions and four interrelationships among the four interrogations: measurement, testing, reasoning, and evaluation.

Chapter 11

Universal Evaluation Concepts

In this chapter, I delve deeper into various evaluation cases across diverse disciplines, with the primary aim of enhancing our comprehension of the evaluation. For the sake of convenience, henceforth, within this section, each case study shall be assigned a unique case ID for differentiation purposes.

11.1 A Computer Science Example: The CPU evaluation

In the first case, which we will refer to as Case One, an organization is in the process of acquiring a critical component of the computer, the CPU. To make an informed decision, the organization evaluates various CPU options by executing its applications on a computer built upon the corresponding CPU. During this evaluation, the organization will collect extensive data on performance and energy efficiency.

In any evaluation process, we refer to the object, individual, or system under evaluation as an EO. As discussed in 8.1, when we uncover the effect of the EO, the EO is the cause object or cause. In the context of Evaluatology, we do not distinguish between EO or cause unless explicitly stated.

A direct AO is defined as *a minimal system on which the effect of the EO can be directly measured or tested* in Section 8.2. Whether and how to measure or test the EO's effect will determine the scope of the direct AO. In Case One, the prospect of measuring the mere properties of an object, such as its weight and power consumption, possesses a certain degree of utility. Nevertheless, such measurements fall considerably short of meeting stakeholders' evaluation requirements.

If the effect of the CPU is in terms of the running time of a typical workload, it can only be directly measured or tested on a minimal system built on that CPU, including the OS, memory, and disk, as shown in Figure 11.1, which is the direct AO. Significantly different from the remaining cases we discussed later, the EO is a component of the direct AO in this case.

For an EO and a specific AO instance, the EXO also impacts the overall effect.

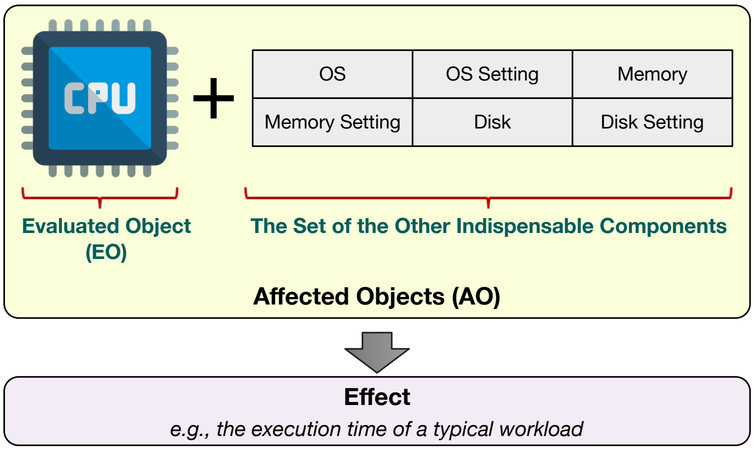

Figure 11.1: For a CPU, its AO is a minimal system built on it, of which the CPU is a component of the AO.

Figure 11.2 showcases how the overall effect is measured on a direct AO instance built on a specific CPU, with the impact of the varying EXO.

After obtaining different dimensions of the effect of the EO, a further question remains: *How does a stakeholder determine different CPU options when they exhibit varying levels of performance across different applications?*

In Case One, the stakeholders could be an organization seeking to acquire the computers, a specific user with different requirements, the designers responsible for creating the CPU specifications, and the producers who manufacture the CPUs.

Based on the collected data, the organization will then formulate an explicit or implicit function that can map different dimensions of the effects of EO on the direct AO to one or several composite evaluation metrics. In the rest of this article, I refer to this function as *a value function*. I interpret the EO's merit or worth [172, 174, 175, 176] as the value this function maps.

According to the discussion in Section 8.2, *the value of an EO reflects the effects of derived EO on the stakeholder*. Please note that the effects of the EO can be propagated to multiple direct and indirect AOs, ultimately reaching the stakeholder.

Figure 11.3 [1] illustrates a typical CPU stakeholder. Their default direct AO component configuration excludes the CPU. The stakeholder operates a business where application execution speed directly impacts profitability—faster performance yields higher revenue. However, increased CPU speed comes at the cost of higher power consumption.

When a user compares two CPUs, A and B, their performance profiles exhibit distinct trade-offs: A delivers higher computational speed at the expense of increased power consumption, while B prioritizes energy efficiency with reduced speed. Since A outperforms B in speed, it generates greater revenue for stakeholders. Consequently, the economic impact of the derived EO (e.g., revenue vs. cost trade-offs) can be quantified

[1] Mr. Chenxi Wang contributed to this figure based on the discussion with me.

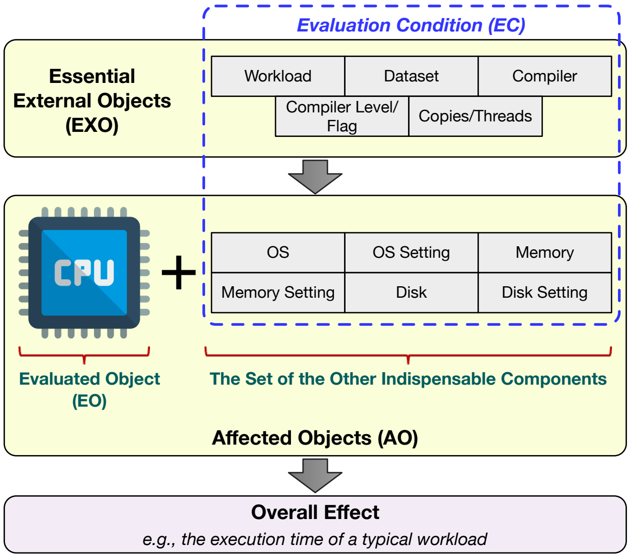

Figure 11.2: A typical CPU example showcases how the evaluation outcome (overall effect) is measured on an AO built on a specific EO (CPU), with the impact of the varying EXO.

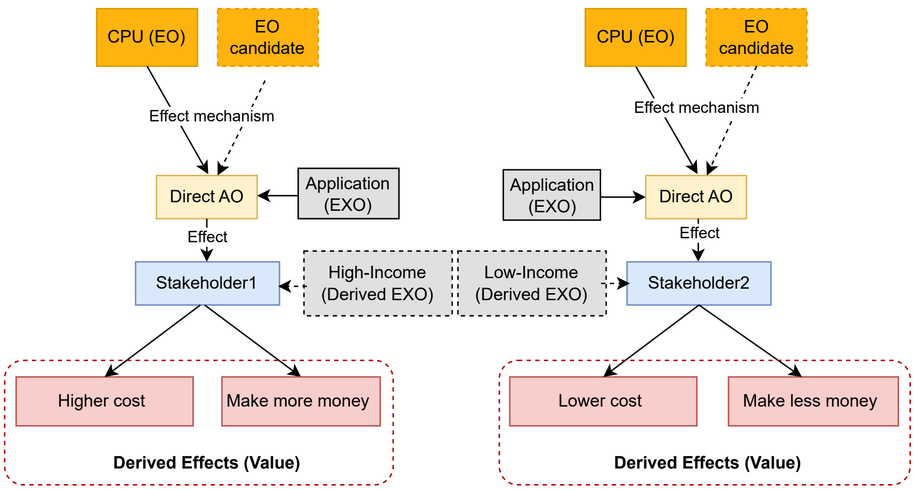

Figure 11.3: For a typical stakeholder, how the value is assigned to the CPU.

for stakeholders.

The stakeholder's valuation of these CPU candidates further depends on the stakeholder's derived EXO, such as the financial context and revenue expectations, which influence how stakeholders assign the value to different CPU options.

11.2 A Physics Example: Issac Newton's Apple Tree

It is a well-known legend that an apple tree at Cambridge University—later immortalized as the "Newton Tree"—inspired Isaac Newton, a once-in-a-millennium genius, to ponder why an apple falls from the tree (Case Two). I argue that Case Two exemplifies the inverse problem of evaluation, which I formally termed "de-evaluation" in Section 12.1.3.

Newton observed an apple falling from the tree and pondered the underlying cause. In Case Two, the direct AO is unambiguously the apple itself. Isaac Newton could measure numerous quantities of the direct AO, which collectively reflect the overall effects induced by myriad objects, including the unknown EOs. However, Newton faced three formidable challenges in addressing this phenomenon.

- What primary EO candidates induced the apple to fall from the tree?
- What is the effect mechanism through which the primary EOs act upon the direct AO?
- How to infer the effect of the primary EO?

Issac Newton could guess and identify several EO candidates and design different SESes, through which he could uncover the primary EO and its effect mechanism that induces the apple to fall from the tree.

Case Two has three unique characteristics. First, it is very challenging to isolate an SES. Even Newton was bold and daring; he could include the Earth, the Sun, the Moon, the wind, and the air into the SES. There could be other invisible or mysterious objects, as shown in Figure 11.4.

Second, it is almost impossible to change different settings in this SES. However, Issac Newton is a genius; he could design many extreme SESes to isolate other objects, which was later perfectly achieved by Henry Cavendish.

For instance, an orange, a pear, or a perch could substitute for the apple, demonstrating that the effect mechanism is independent of the AO. An EC populated with a vacuum would eliminate the influence of wind and air, while an EC populated with a vacuum at different physical locations on Earth would reveal that the effect mechanism varies with geographical location.

Henry Cavendish's experiment [31, 108] to measure the gravitational constant (G) in 1797 provided a very simple SES through many intricate experimental skills.

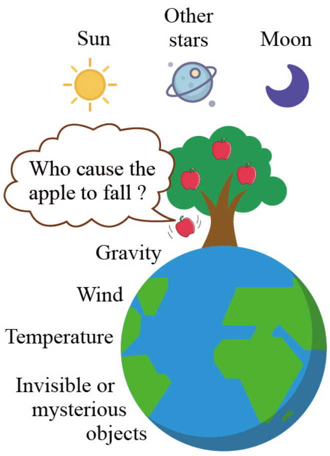

Figure 11.4: For identifying the cause that induces an apple from a tree, it is very challenging to isolate an SES.

How Henry Cavendish Designed a Very Simple SES to Measure the gravitational constant (G)

Apparatus.

- A torsion balance consisting of a six-foot wooden rod suspended from a wire, with two small lead spheres (1.6 pounds each) attached to its ends.
- Two larger lead spheres (350 pounds each) were placed near the smaller ones, separated by about nine inches.
- A mirror attached to the wire reflected a light beam onto a scale, amplifying tiny torsional movements.

Method.

- The gravitational attraction between the large and small spheres caused the rod to twist slightly, deflecting the light beam.
- By measuring this deflection, Cavendish calculated the force between the masses and derived G using Newton's inverse-square law.

Mitigation of External Objects' Effects to Isolate the SEC.

Shielding Earth's Gravity

- Symmetrical setup: The experiment balanced Earth's gravitational pull on both sides of the torsion balance, canceling its net effect.
- Local mass symmetry: Large lead spheres were placed symmetrically around the small spheres to minimize asymmetric gravitational gradients.

Isolation from nearby objects

- Sealed enclosure: The apparatus was housed in a closed room to eliminate air drafts and reduce temperature variations
- Non-magnetic materials: Lead was chosen to avoid magnetic interference from external objects like magnetic fields

key findings

- Cavendish determined the gravitational constant as $G\ \ 6.754\times10^{\ 11}\ \mathrm{N\cdot m^2/kg^2}$, remarkably close to modern values ($6.6732\times10^{\ 11}$).
- From G, he calculated Earth's average density (5.481 times water's density) and mass.

Sources of Error in Cavendish's Experiment

Environmental Disturbances

- Air currents: Even slight air movements could cause false torsion balance deflection.
- Temperature fluctuations: Thermal expansion of the suspension wire or lead spheres might alter the measured torque.

Instrumental Limitations

- Non-ideal masses: The lead spheres were not perfect point masses, introducing deviations from Newton's inverse-square law.
- Torsion wire imperfections: Variations in wire elasticity or manufacturing defects could affect the torsion constant.

Measurement Uncertainties

- Optical alignment: Mirror misalignment or scale parallax errors in tracking light spot displacement.
- Periodic oscillations: Damping incomplete torsional oscillations led to timing errors in period measurements.

Through the design of simple SESes, Issac Newton could speed up the process that drew close to the Law of Universal Gravitation; however, it is still very challenging. Through the work of Issac Newton, we know the EO is the Earth, and the effect mechanism is gravitation. However, if we aim to fully understand the effect of the Earth on the other objects, there are many open issues.

Third, uncovering the effect of Earth's gravity on the apple is a purely Physics problem; discussing its value is inherently meaningless. However, if Earth's influence extends to the apple's sweetness—a property that plant owners or end-users do care about—then it becomes meaningful to evaluate the Earth's value in this context. This is particularly relevant if we consider planting apple trees on the moon or other EO candidates, where the Earth's role as an EO could be changed.

Fourth, even in uncovering the effect of the Earth on the apple, there are many different perspectives. For example, the Earth also provides indispensable living environments for planting an apple tree.

11.3 An Artificial Intelligence Example: Evaluating an AI algorithm

In Case Three, the objective is to evaluate an AI algorithm, specifically focusing on an Image Classification task as a case study. Real-world images are collected and annotated with accurate labels, such as a cat or a dog. A portion of these images is randomly selected to construct the training, validation, and testing datasets based on a designated percentage. To assess the image classification algorithm, that is, the EO in this case, it must be implemented on a computer system utilizing a specific programming framework, such as PyTorch or TensorFlow.

During the evaluation process, the testing data is provided to the algorithm, which generates an output. This output is then compared to the ground truth: the labels.

Additionally, measurements are collected for each run of the evaluation process.

The fundamental evaluation process in Case Three bears a resemblance to that of Case One. However, there exist three notable distinctions. First, the EO, i.e., the algorithms, must be implemented on a computer. This case has two-fold implications. On the one hand, the EO has different implementations, implying that different EO instantiations have a diversifying effect. On the other hand, within a complex system featuring diverse implementations, the EO instantiation is initiated, and these varied implementations significantly influence the effects of the EO instantiation.

Secondly, the EC in Case Three differs. The presence of a dataset labeled with ground truth constitutes a vital aspect of an EC. A dataset labeled with the ground truth represents a specific instance of a problem or task. Regrettably, in Case Three, it is challenging to define a problem or task mathematically. Hence, diverse problem or task instances are devised in an ad-hoc manner, for example, selecting images randomly to form training, validation, and test datasets based on a predetermined percentage. Given that image selection relies on professional expertise, varying professional expertise will inevitably introduce distinctions among ECs.

Thirdly, in Case Three, upon feeding the algorithm with the testing data, it generates an output that is then compared with the ground truth, also known as the test oracle. Apart from measurements, there exist other forms of activities in the evaluation process, namely testing, as expounded upon in Chapter 5.

Fourth, the direct AO and EXO encompass significantly more diverse and complex components. For example, if stakeholders prioritize algorithmic accuracy, the direct AO includes the dataset labeled with ground truth, the algorithm implementation, the operating system, the processor, and the memory, while the EXO includes the compiler, the programming framework, like PyTorch.

11.4 An Chemistry Example: Lavoisier's Combustion Experiment

Case Four [2] is a famous chemistry example: Lavoisier's combustion experiment. Lavoisier's combustion experiment is considered a milestone in the history of chemistry, as it revealed the crucial role of oxygen in the combustion process [109].

Lavoisier's classic Combustion experiment, central to his investigation of the transformation of matter and the composition of air, was both ingeniously designed and meticulously executed.

In this experiment, as shown in Figure 11.5, mercury was placed in a distillation flask and heated, with a mercury trap below to capture the evaporated mercury vapor. The vapor was then guided through a pipe into a separate container, where it cooled and was re-liquefied via a condenser. The experiment was conducted in two phases:

[2] Dr. Lei Wang authored this subsection.

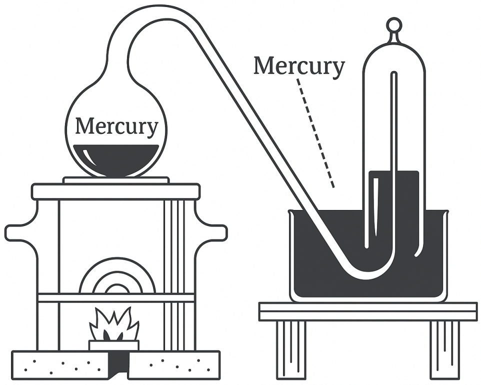

Figure 11.5: Overview of Lavoisier's classic combustion experiment.

In the first phase, heating continued for twelve days, during which mercury reacted with components in the air to form red mercuric oxide. The volume of gas in the system contracted by approximately one-fifth, and the total mass of mercury and mercuric oxide remained unchanged, providing preliminary support for the principle of mass conservation. In the second phase, mercuric oxide was subjected to intense heat in the distillation flask, decomposing into a gas that could support combustion and respiration. Notably, the volume of this gas precisely filled one-fifth of the space previously reduced, while the remaining four-fifths of the gas exhibited no such properties.

Through this experiment, Lavoisier not only demonstrated the reversibility of mercury vaporization and oxidation —confirming that mass was not lost but merely transformed —but also conclusively proved that air is composed of "combustible oxygen" (approximately one-fifth) and "inert nitrogen" (approximately four-fifths).

This groundbreaking discovery effectively tested and invalidated the phlogiston theory, establishing the foundational principles of mass conservation and oxidation reactions and laying the cornerstone for modern chemistry.

The goal of this experiment is to identify the primary cause object that induces combustion. If we treat Lavoisier's Combustion Experiment as an evaluation, the EO is some mysterious substance to be determined.

The direct AOs are any substances that can be burned or calcined, such as mercury, phosphorus, and sulfur, and we can directly measure quantities of direct AOs, such as mass, gas volume, and temperature. Before Lavoisier's combustion experiment, people believed that the mysterious substance was "phlogiston" and the EO (phlogiston) was a part of these direct AOs.

Antoine Laurent Lavoisier designed various SESes centered around the closed space, which isolated the effects of other objects on direct AOs.

He sealed mercury (the direct AO) in a retort and heated it, observing changes in the gas volume within the closed space and collecting the gases produced by combustion for analysis.

The key steps included:

- *Heating mercury*: Sealed in a glass vessel, mercury formed red calx (HgO) while the air volume decreased by roughly one-fifth.
- *Gas analysis*: The remaining gas (later called *azote*, i.e., nitrogen) failed to support life or combustion.
- *Decomposition of calx*: When heated, HgO released a highly respirable gas (oxygen), and the gas volume returned accordingly.

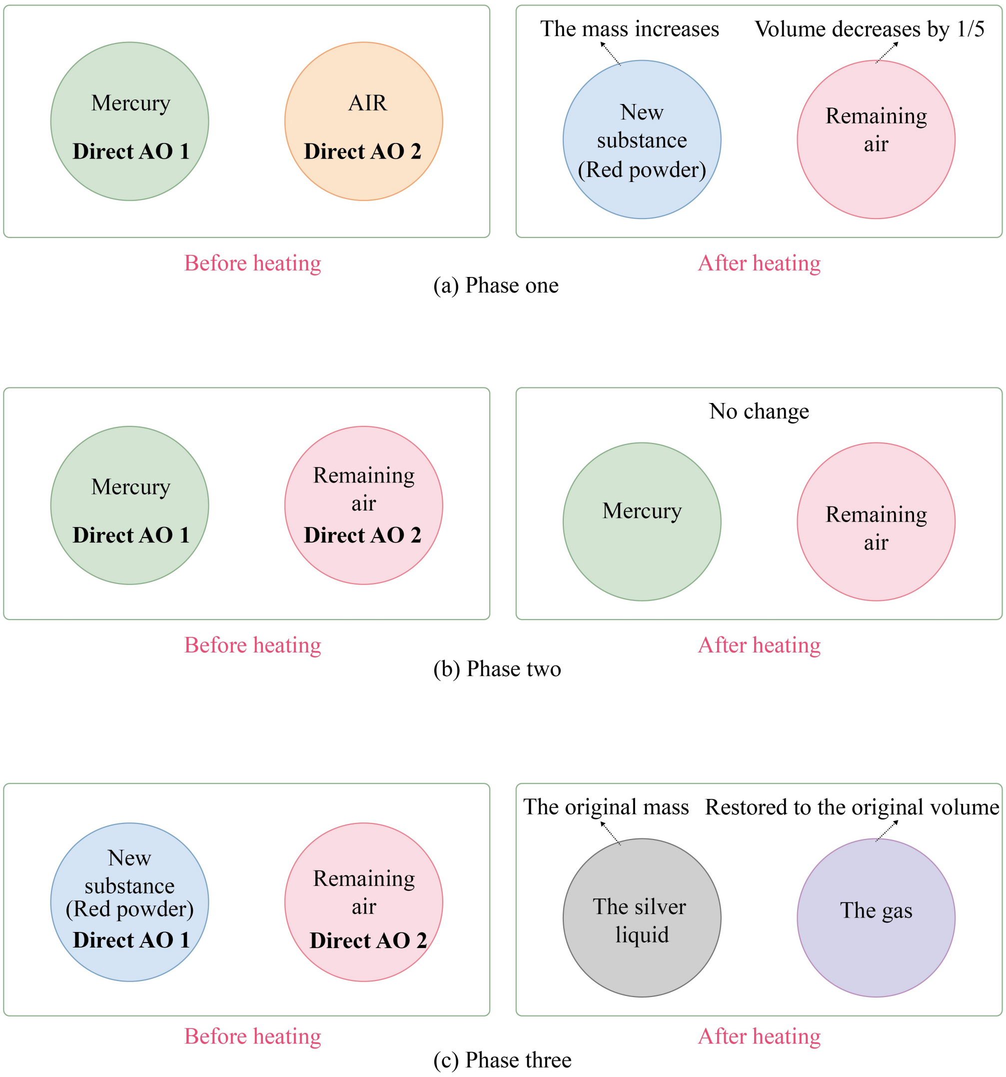

Figure 11.6: Antoine Laurent Lavoisier's experiments: The changes after heating in three phases.

Significantly different from the other cases, Case Four has a unique characteristic: a closed space could perfectly isolate the other objects' effect; however, though in the same closed space, the direct AOs and EXO change in different phases as shown in Figure 11.6.

In the first phase, there are two direct AOs: the mercury (silver liquid) and the air, consisting of unknown objects at his time. After heating, the mercury changes into a new substance (red powder), with the mass increased; the air volume decreased by about one-fifth.

In the second phase, there are two direct AOs: the mercury (silver liquid) and the remaining air from the first step. After heating, there is no change, and no combustion is observed.

In the third phase, there are two direct AOs: the new substance (red powder) and the remaining air produced in Step One. After heating, the red powder changed into a silver liquid with the original mass and released a highly respirable gas. Finally, the gas was restored to the original volume.

By comparing the SESes in different steps while using the closed space to isolate the other objects' effect (the control as we discussed in Part IV), Antoine Laurent Lavoisier induced that combustion depended on a reactive component in air (later named *oxygen*) and that, by volume, it accounted for about one-fifth of air. He ultimately overturned the prevailing phlogiston theory, which claimed combustion was due to the release of "phlogiston" and established the scientific theory that "combustion is the oxidation of substances by oxygen."

Additionally, he used other combustible materials (direct AO), such as phosphorus and sulfur, to verify the universality of the combustion mechanism.

He also revealed the effect of the mechanism in combustion as follows.

$$\text{Combustion: } 2\text{Hg} + \text{O}_2 \xrightarrow{\Delta} 2\text{HgO} \quad (\text{silver liquid} \rightarrow \text{red powder})$$
$$\text{Decomposition: } 2\text{HgO} \xrightarrow{\Delta} 2\text{Hg} + \text{O}_2 \uparrow \quad (\text{red powder} \rightarrow \text{silver liquid} + \text{gas})$$

Lavoisier showed:

- Weight gain in combustion came from oxygen absorption, not "phlogiston" release.
- Air comprised about 20% oxygen (vital for combustion) and about 80% nitrogen.

Despite contemporary debate over priority and interpretation, Lavoisier's use of precise balances and careful mass accounting cemented the oxygen theory of combustion. His experiments demonstrated that combustion is a chemical reaction between substances and oxygen, fundamentally changing our understanding of chemical processes.

His work introduced:

- *Law of Conservation of Mass*: The total mass in a reaction remains constant.
- *Oxidation Theory*: Combustion is the oxidation of substances by oxygen.

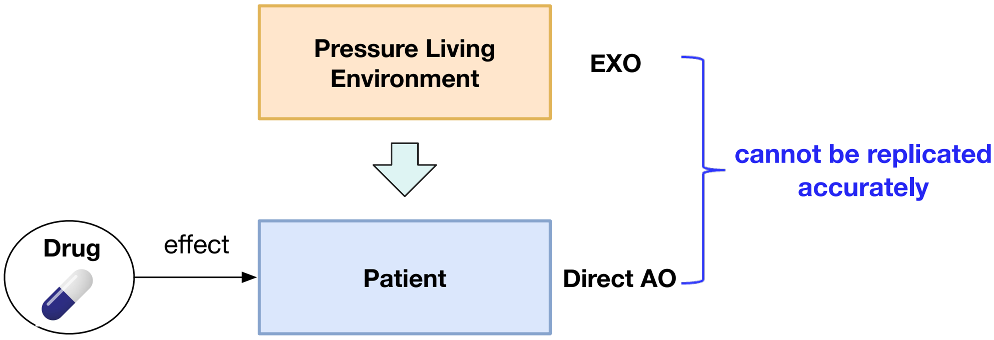

Figure 11.7: In the case of the drug evaluation, we can not replicate the direct AO and EXO very accurately.

11.5 A Pharmacology Example: Drug evaluations

The Fifth Case (Case Five) involves evaluating a drug. In this case, the EO refers to a specific drug, while the direct AO encompasses the participants (human beings) targeted by these interventions. Figure 11.7 shows the composition of the SES.

When comparing Cases Five with Cases One and Three, we have made three observations. First, the direct AO is independent of the EO. Second, the overall effect can be measured or tested on the direct AO, with the contributions of EXOs, e.g, the pressure exerted on the patients, and the environment in which the patients live.

The last but not least observation revolves around the distinctions of the EXO and direct AO found in Cases One, Three, and Five. In Cases One or Three, the direct AO and EXO can be well-defined; that is to say, we can replicate an EXO or direct AO very accurately. However, in Case Five, we can not replicate an EXO or direct AO very accurately. For example, we can not replicate a patient accurately, not to mention other critical factors included in the EXO. Instead, there is substantial variability in EXO and direct AO in Case Three.

11.6 A Social Science Example: Policy evaluation

The Sixth Case (Case Six) involves evaluating a policy aimed at addressing drug addiction within a community. In this case, the EO refers to a policy aimed at addressing drug addiction intervention. When comparing Cases Six with Cases Three, we have made three observations.

Firstly, they have a resemblance. The policy may have different instantiations, like that of the algorithm. The EO instantiation has a significant impact on its effect, as same to that in Case Three. However, in Case Six, the EO instantiations can not be controlled accurately, differing significantly from those of the algorithm. For the latter, the instantiations of an algorithm can be replicated completely and accurately.

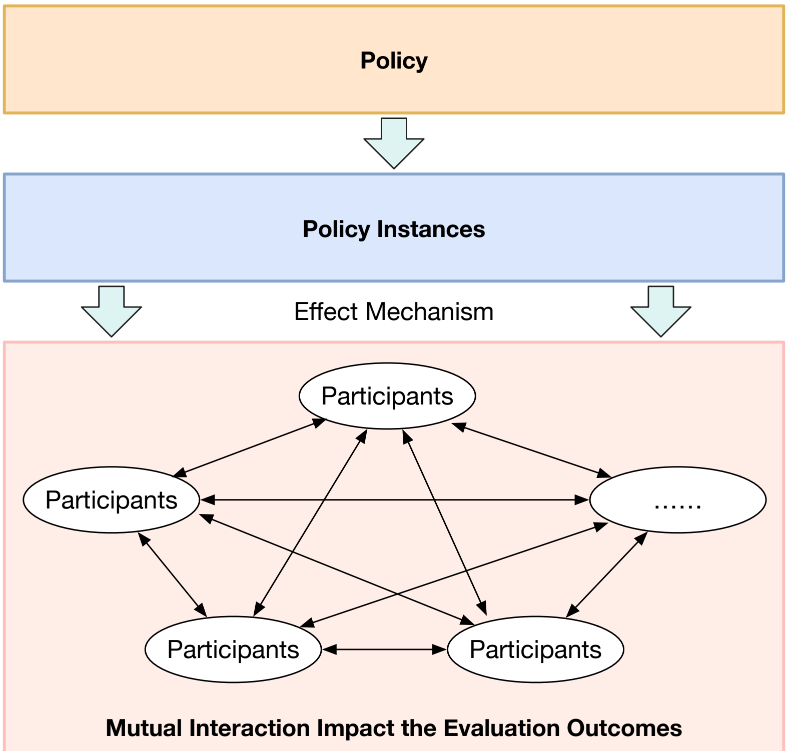

Figure 11.8: In the policy evaluation, the interaction of direct AO significantly impacts the evaluation outcomes.

Second, the EXO, which impacts the overall effect, is more difficult to identify in Case Six than in Cases Five or Three. For example, the participants' attitude towards the policy may dominate the evaluation outcomes.

Third, the interactions between different participants, the direct AOs, also impact the evaluation outcome significantly, as shown in Figure 11.8. In Cases One, Three, and Five, each direct AO is independent from the other; however, in Case Six, each direct AO is dependent on the others.

11.7 An Biology Example: Gregor Johann Mendel's Peas

Biological traits, such as human height and body shape, a cat's straight or curly fur, and the sweetness or waxiness of corn, represent the observable variations among biological individuals. What determines these biological traits? Why do the offspring of a tall-stemmed pea plant and a short-stemmed one all exhibit tall stems? And why do some offspring of tall-stemmed pea plants display short stems? The SES of the inheritance of natural traits is shown in Figure 11.9.

The answers to these questions were provided by Gregor Johann Mendel. Through

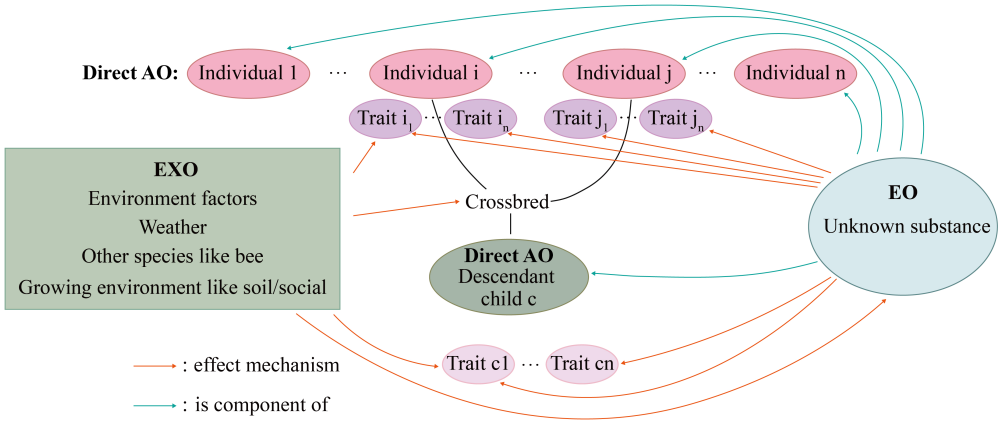

Figure 11.9: SES of the inheritance of natural traits, there is a complex relationship among the EO, the direct AO, and the EXO.

meticulously designed pea experiments (Case Seven [3]), Gregor Johann Mendel formulated the classic laws of segregation and independent assortment in genetics [120].

Different from other cases, Case Seven has a unique characteristic. When two peas are crossbred, both their gene (two EOs) are the cause, while the direct AO are those peas and their descendant. Figure 11.10 shows the complex relationship among the EC, the direct AO, and the EXO.

Mendel's choice of the garden pea (direct AO) was strategic and critical to his success:

- *True-Breeding Strains:* Peas can self-fertilize, and Mendel identified varieties that, when self-fertilized, consistently produced offspring identical to the parent for specific *traits* (e.g., seed shape). These were his *P generation* (parental generation).
- *Distinct, Heritable Traits:* He selected seven traits that exhibited clear-cut, contrasting *traits* (see Table 11.1), such as Round vs. Wrinkled seeds. There were no intermediates.
- *Controlled Crosses:* The physical structure of the pea flower allowed Mendel to easily prevent self-pollination and perform *cross-pollination* (*hybridization*) between selected parents.
- *Short Generation Time:* Multiple generations could be studied within a reasonable time frame.

Mendel initially spent several years allowing peas to undergo self-pollination, thereby selecting strains that could stably inherit specific traits (with consistent offspring phenotypes), known as "homozygous". For instance, the offspring of tall-stemmed pea plants would always be tall-stemmed. His initial experiments focused on a single trait at a time, a *monohybrid cross*.

[3] Mr Chenxi Wang authored this section.

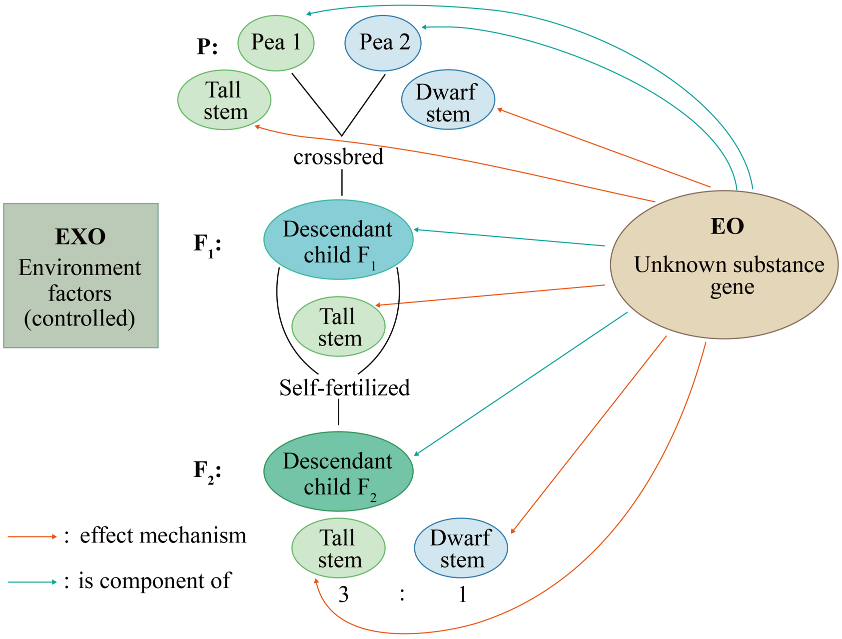

Figure 11.10: SES of Gregor Johann Mendel's Peas experiments [120], there is a complex relationship among the EO, the direct AO, and the EXO.

Trait	Dominant Trait	Recessive Trait
Seed Shape	Round	Wrinkled
Seed Color	Yellow	Green
Flower Color	Purple	White
Pod Shape	Inflated	Constricted
Pod Color	Green	Yellow
Flower Position	Axial	Terminal
Stem Length	Tall	Dwarf

Table 11.1: The seven pea plant traits studied by Mendel

- He crossbred two true-breeding (homozygous) pea plants that differed in one trait (e.g., Tall vs. Dwarf).
- He collected the seeds and grew the plants—this was the *first filial generation (F_1)*.

Observation One: The F_1 offspring always resembled only one of the parents. For example, the cross between a pure Tall and a pure Dwarf plant produced *all* Tall offspring. The trait that appeared in the F_1 generation he called the *dominant* trait (Tall). The trait that seemed to vanish was the *recessive* trait (Dwarf).

Mendel then allowed the F_1 plants to self-fertilize and produce the *second filial generation (F_2)*.

Observation Two: The recessive trait, absent in the F_1 generation, reappeared in the F_2 generation! Furthermore, Mendel counted the numbers of each type and found a consistent ratio of approximately *3 Dominant: 1 Recessive.*

To explain these results, Mendel proposed a revolutionary model:

- *Particulate Factors:* Inheritance is controlled by discrete "factors" (now called *genes*) that are passed unchanged from parents to offspring.
- *Alleles:* For each trait, an organism inherits two factors, one from each parent. These different forms of a gene are called *alleles* (e.g., a Tall allele and a Dwarf allele).
- *Dominance/Recessiveness:* In a heterozygote (with two different alleles), the *dominant* allele determines the organism's appearance, while the *recessive* allele has no noticeable effect.
- *The Law of Segregation:* The two alleles for a heritable trait segregate (separate) from each other during the formation of gametes (sex cells), so that each gamete carries only one allele for each trait.

Mendel then asked: What happens when two traits are inherited simultaneously? He performed a *dihybrid cross* between parents that differed in two traits—for example, a plant true-breeding for Round Yellow seeds ($RRYY$) and one true-breeding for Wrinkled Green seeds ($rryy$).

- *F_1 Generation:* All offspring were $RrYy$ and had Round Yellow seeds.
- *F_2 Generation:* When the F_1 plants self-fertilized, the F_2 generation showed *four phenotypes* in a consistent ratio:

9 Round Yellow (1 RRYY + 2 RRYy + 4 RrYy + 2 RrYY) : 3 Round Green (1 RRyy + 2Rryy) : 3 Wrinkled Yellow (1 rrYY + 2 rrYy): 1 Wrinkled Green (1rryy)

The 9:3:3:1 ratio was a dihybrid ratio that could be derived from two independent monohybrid 3:1 ratios ($(3{:}1) \times (3{:}1) = 9{:}3{:}3{:}1$). This led to his *Law of Independent Assortment.*

- *The Law of Independent Assortment:* Genes for different traits assort independently of one another during gamete formation. The allele a gamete receives for one gene does not influence the allele received for another gene.

In Mendel's pea experiments, the EO refers to genes, the direct AO denotes peas, which can be substituted with any biological individuals, the effect refers to the traits exhibited by the peas, and the effect mechanism pertains to the laws of segregation and independent assortment.

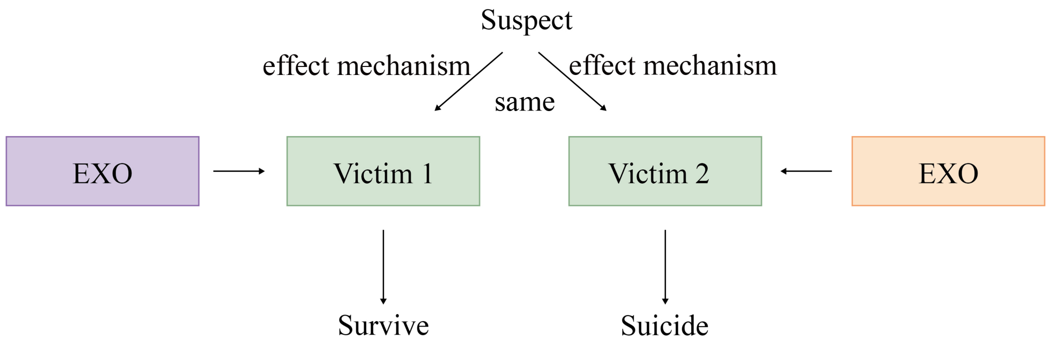

Figure 11.11: A suspect uses violent language with the same intensity to insistently intimidate two victims. The overall effects on the two victims are significantly varied.

11.8 Scientific and Technology Achievement Evaluations

Case Eight involves evaluating a specific scientific or technological achievement. The EO is obvious, including but not limited to a specific idea, an article, a patent, or a technology product.

Compared with other cases, Case Eight has two significant distinctions. First, in Case Eight, the direct AO is not limited to a single kind of object. For example, it could be the other scientists or engineers, other achievements, the industry, or even the society. Instead, in the other cases, the direct AOs are basically limited to a single kind of object. Second, it is almost impossible to perform an experiment to investigate the effect of the EO. We have to take a retrospective instead of a prospective approach.

11.9 A Legal Liability Example: Determining Accountability for Suspects.

In Case Nine, we presume there is a suspect who uses violent language with the same intensity to insistently intimidate two victims, and hence induce the same effects. One victim has strong family support and lives an unpressurized life, while the other victim lives a poverty-stricken life by himself, having economic or other pressures. The first one survived, while unfortunately, the other one chose to commit suicide.

In Case Nine, several unique characteristics are present. First, the effect mechanism is significantly different from the other case. The EO or cause uses violent language, but with no force to insistently intimidate the affected objects. Second, the differences of EXOs dominated the overall effect. For two victims, though the EO and its effect are the same, the overall effect is totally different, with strikingly different consequences, as shown in Figure 11.11.

11.10 An Education Example: Evaluating a Learning Intervention

Case Ten [4] concerns the evaluation of an educational intervention, such as a new digital learning platform or pedagogical method, aimed at improving students' learning outcomes. In this case, the EO refers to the educational intervention itself: a particular curriculum design, AI-assisted tutoring system, or instructional strategy introduced to enhance teaching and learning effectiveness.

Here, the AO is the individual learner, the minimal observable system in which the overall effect is measured. However, as in previous cases, the observed evaluation outcome, or the overall effect, is not determined by the EO and AO alone. The EXO includes a wide range of contextual factors such as teacher expertise, classroom environment, socioeconomic background, parental support, institutional resources, and even cultural norms surrounding education. These EXOs interact with both the learning process and its outcomes, often confounding attempts to isolate the true effect of the EO.

Importantly, the EO, AO, and EXO in this case are not static entities. Instead, they form an evolving system that unfolds over time and can naturally be described as a time series, a sequence of observations reflecting the dynamics of learning. In this system, the EO, representing a particular educational intervention, exerts its influence directly on the AO, which is the individual learner. The AO "receives" this influence through modifications in its learning process, for example, changes in cognitive engagement, interaction patterns, or responsiveness to feedback, which then propagate to observable learning outcomes such as comprehension, retention, or transfer. In other words, the measured effect emerges from the interaction between EO, EXO, and AO, rather than from either component in isolation.

11.11 Adversarial Exercise Evaluation: A Case of Systemic Confrontation

In a systemic confrontation scenario (Case Eleven [5]), such as a defensive exercise between a defending force and an attacking force, the fundamental question of evaluation is not merely to determine who wins or loses. Most importantly of all, it is to uncover the effect of the EO, e.g., a defensive command and control system, from the overall outcome of the exercise.

The challenge lies in that the observed outcome (e.g., the number of successfully intercepted attacks, response latency, or system survivability) is the result of multiple intertwined influences, not limited to the EO alone. Thus, to perform a scientifically

[4] Dr. Wanling Gao authored this section.
[5] Dr Fanda Fan authored this section.

valid evaluation, one must build an SES that captures all relevant interacting objects and allows the attribution of the true effect of the EO.

In such a confrontation scenario, different from the other cases, the SES embodies a dynamic and interdependent environment: each object not only responds to the others but may also alter its own subsequent states. Therefore, the evaluation must not treat the effect as a static or linear response but as a systemic phenomenon that emerges from the mutual interactions among EO, AO, and EXO in time series.

Also, unlike static laboratory experiments, the SES in a confrontation exercise is *dynamic, adaptive, and reciprocal*: every decision made by the EO induces a counter-reaction from the EXOs, which in turn modifies the operational state of the AO. This cyclical causation challenges the traditional linear model of cause–and–effect analysis and highlights the necessity of viewing evaluation as a *systemic causal inference process* rather than a unidirectional measurement.

The adversarial confrontation scenario provides a rich and realistic context in which the fundamental principles of Evaluatology can be examined and extended.

In adversarial or competitive systems, the role of Evaluatology thus expands from verifying outcomes to uncovering the *structure of influence* that generates them. Through explicit modeling of EO–AO–EXO interactions, evaluators can isolate the true effect of a system even under continuous adaptation and opposition, thereby fulfilling the ultimate goal of Evaluatology: to scientifically uncover the effects of causes within a self-contained, yet dynamically interacting, world.

11.12 Another Physics Example: Verification of a Theory or Model

Case Twelve involves verifying a theory of a model of surface physics, which could be extended to any discipline. Dr. Michael Scriven [175] exemplified this case as the fundamental role of evaluation science, as it is witnessed in almost all disciplines. However, I do not intend to consider this as an evaluation. Instead, it should be categorized into testing, as the aim of this practice is to confirm whether the theory or model conforms to the test oracle. Please see our discussion in 5.8 in Chapter 5.

11.13 Difference from the Other Definitions of Evaluation

In Chapters 1 and 2, I have summarized the unique differences of Evaluatology concepts from those of other works.

In this section, I repeat two points. First, I define the evaluation as uncovering an EO's effects and derived effects, contrasted with determining the merit, worth, or value in [172, 174, 175, 176, 55, 68, 85, 196, 195].

Second, I propose a simple and concise concept system. Instead, the others relied on the encyclopedic approach to define several hundred concepts or terms in [174, 118].

11.14 Summary

This chapter demonstrated that it is feasible to propose a universal and concise concept framework that can be applied across various evaluation scenarios in different disciplines.

Chapter 12

Categories of Evaluation Problems

Per discussion in Section 8.2, the essence of evaluation is to *uncover an EO's effects and derived effects.*

This chapter analyzes the structure of different evaluation problems, which could be based on the category theory introduced in Section 3.7. First, I formalize the evaluation problem and its dual problem, and the inverse problem. Then I discuss the categories of evaluation problems

12.1 Formalization of Evaluation Problem and Its Dual and Inverse Problems

In this section, I formally define the evaluation problem, then introduce the dual problem of evaluation, that is, design, and the inverse problem of evaluation, that is, de-evaluation. Figure 12.1 [1] shows the relationships among evaluation, design, and de-evaluation problems.

12.1.1 Formalization of Evaluation Problem

In this section, I formulate the evaluation problem [2]. I note an EO as O and its element as o. O has n (n is a constant) components F_i, noted as $O = F_1 \times \cdots \times F_i \times \cdots \times F_n$. Each component F_i has $\sigma_i \geq 1$ configurations, where σ_i is a constant. I note the element of F_i as f_{ij_i}, where $j_i \geq 1$ and $j_i \leq \sigma_i$. $o = (f_{1j_1}, \cdots, f_{ij_i}, \cdots, f_{nj_n})$.

I note an EC as C and its element as c. C has m (m is a constant) components G_i, noted as $C = G_1 \times \cdots \times G_i \times \cdots \times G_m$. Each component G_i has $\gamma_i \geq 1$ configurations, where γ_i is a constant. I note the element of G_i as g_{ij_i}, where $j_i \geq 1$ and $j_i \leq \gamma_i$.

[1] Mr. Chenxi Wang contributed to this figure based on the discussion with me.

[2] Mr. Chenxi Wang and Mr. Hongxiao Li also contributed to this section.

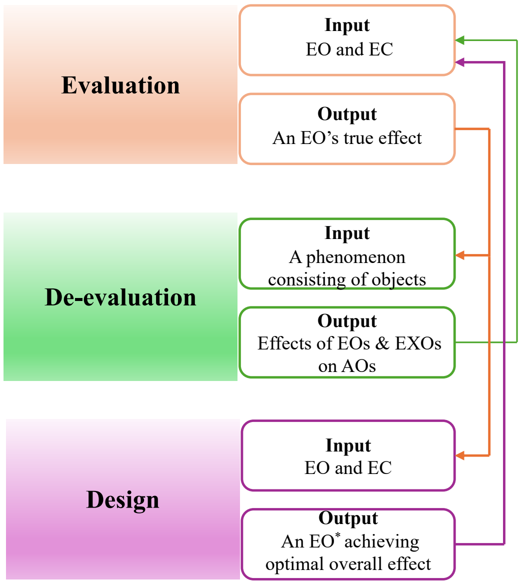

Figure 12.1: The relationship among evaluation, de-evaluation, and design problems.

$c = (g_{1j_1}, \cdots, g_{ij_i}, \cdots, g_{mj_m})$. We note the distribution of a specific configuration of c as $h_{EC}(c)$ where $\sum h_{EC}(c) = 1$.

I note that the overall effect observed on the AO is oe, and the true effect of EO is to be inferred, noted as e. In this context, attributing the overall effect to an EO is formulated as follows.

For a specific EO configuration o, given the overall effect, $oe(o|c)$, under the EC where c follows a distribution $h_{EC}(c)$, how to infer the true effect of the EO, $e(o)$?

12.1.2 Design: The Dual Problem of Evaluation

In this section, I explain why a design problem is the dual problem of an evaluation [3].

Evaluation and design are dual problems. The overall effect is induced by both EO and EXOs on the AO. Evaluation aims to uncover the true effect of an EO from the overall effect, while design seeks a specific EO configuration within the design space to achieve optimal overall effect.

12.1.3 De-evaluation: The Inverse Problem of Evaluation

Based on the discussions in Chapter 8, we can easily define the inverse problem of evaluation, which we call de-evaluation.

[3] This idea is inspired by my colleague, Dr. Chunjie Luo, in a group meeting.

I take the diagnosis of traffic jams as an example to showcase a de-evaluation problem. Traffic jams are widely found in many cities. But the traffic jam is a composite phenomenon that reflects the quantities of many objects. The purpose of the diagnosis of traffic jams is to attribute the composite phenomenon to the effects of EOs and EXOs on AOs. In nature, this essential behind the phenomenon is the inverse problem of the evaluation, that is, *how to accurately attribute the effects to their causes*.

Let me explain the inverse problem of the evaluation according to the definition of cause and effect in Section 8.1. We observe a phenomenon P that consists of many objects, O_i, showing different quantities, Q_{ij}, and the changes of quantities, $\mathrm{d}Q_{ij}$ that can be measured or tested. The inverse problem of the evaluation is to trace back the Q_{ij}, and the changes of quantities, $\mathrm{d}Q_{ij}$ of objects, O_i, to the EOs (causes), EXOs, and their effects on the AOs.

12.2 Categories of Evaluation Problems

This section categorizes evaluation problems from different perspectives.

12.2.1 Categories of Evaluation Problems According to the Nature of EOs

According to Section 3.1, an object is a class of entities owning a set of *properties*. Every object could be an EO, so it is also reasonable to classify the evaluation problems according to the nature of the EO.

The first kind of evaluation problem is when the EO is a natural object. I classify it into several sub-categories. The first sub-category is that its properties can not yet be interrogated by a subject. The Second sub-category is that its properties are partially known by the subject. The third sub-category is that the EO can be well-defined, that is, it can be replicated accurately. The fourth sub-category is that the EO can be well-defined, and its components can be freely modified at the subject's discretion.

The second kind of evaluation problem is when the EO is an artifact. According to the definition 3.3, an *artifact* refers to an object that demonstrates intentional conjecture, design, and/or fabrication by a subject. An artifact could be an idea, an imagination of something that does not exist, fake news, or a new design and implementation.

As they are created by a subject, which could be an automatic object or an intelligent life, the unique characteristics of this kind of problem depend on how the artifact induces the effect on the subject, which could be significantly different from that of a natural object.

12.2.2 Categories of Evaluation Problems According to the SESes

According to the definition in Sections 3.5, 10.1, a *model* is a streamlined representation of an object or several objects with mutual effects [221, 224]. I formally call a streamlined representation of an evaluation system *an evaluation model*. Essentially, an SES or a derived SES is an evaluation model. So it is reasonable to classify different evaluation problems according to the nature of the SES.

The first kind of evaluation problem is when an SES is unknown. For example, when we evaluate the parallel universes or the soul or dark matter to reveal their effect, we do not even know what the AOs are.

The second kind is when an SES is only partially known, e.g., when evaluating an object in cosmology and astronomy. The exact number of AOs remains unknown, and it is unclear whether other EXOs exist.

The third kind is when an SES is known. In this case, there are three different subcategories. The first subcategory is when an SES is very complex and cannot be *well-defined*. Being not well-defined indicates that we can not replicate the SES accurately. The second subcategory is when an SES is known and well-defined but not subject to arbitrary manipulation for different reasons, such as realization limitations, unaffordable costs, unaffordable consequences, or ethical reasons. "If a system can be modeled in a function, arbitrary manipulation entails setting its independent variables to any arbitrary number within its domain [222]." The third subcategory is when an SES is known, well-defined, and subject to arbitrary manipulation. For example, a computer nearly falls into this category [222].

12.2.3 Categories of Evaluation Problems According to the Effect Mechanisms

According to the definition in Section 8.1, the effect mechanism is the way through which the cause object induces the effect on the affected objects. Another perspective on analyzing categories of evaluation problems could be through their effect mechanism.

The first dimension is the relationship between the EO and the AO. There is a class of evaluation problems whose common characteristics are worth considering; the EO is a component of the AO. For example, in Case One discussed in Section 11.1, the EO is a CPU, which is the component of the direct AO, a computer system upon which the CPU has an effect.

In Section 11.10, the relationships between the EOs and AOs are much more complex. The genes (components), that is, the EO, of the parent peas (AOs) induce different effects on the descendant peas (AOs). While in the other case, the EO is independent of the AOs. For example, when an EO is a drug, it will affect a specific patient or a population of patients. The EO is independent from the AO.

The second dimension is different effect mechanisms. For example, the Earth could induce gravity on an apple. A CPU works together with several indispensable components and culminates as a computer system; the effect mechanism is the collaboration. A criminal suspect could hurt a victim through intimidation, through insulting or mean language.

The third dimension is whether the interactions of EOs or AOs will impact the effect or not. This case is often found in the policy evaluation, as discussed in Section 11.6. For example, in evaluating a policy to curb drug addiction, the interactions among the AOs, especially their attitude, will impact the evaluation outcome significantly. In the cases discussed in Section 11.10, the interactions of EOs or AOs also impact the evaluation outcome significantly.

12.3 Summary

This chapter formally defines the evaluation problem and its inverse problem and dual problem, while categorizing different types of evaluation problems.

Chapter 13

Fundamental Issues in Evaluatology

In this chapter, I present four fundamental issues in Evaluatology.

13.1 What Evaluation Problems Yield a True or Undefined Evaluation Outcome?

This is the most fundamental issue in Evaluatology. For any evaluation problem that yields an undefined evaluation outcome, we can not test the evaluation approach against the test oracle. Unfortunately, this fundamental issue was never discussed until I formally formulated it in [224].

I address this issue in the proposed evaluation axiom in 9.2. I claimed that when an SES is known, the effect of the EO has the true outcome, which indirectly states what evaluation problem yields a true or undefined evaluation outcome. Other fundamental evaluation approaches, which we will discuss in Part IV, did not explicitly consider this issue.

13.2 Meta-evaluation: Which Evaluation Methodology Yields a Valid Outcome?

As the name of meta-evaluation implies, the purpose of meta-evaluation is to evaluate different evaluation approaches. As I discussed in Chapter 8, for this task, a specific evaluation approach is the EO; the direct AOs are the *inferred effects* of different kinds of EOs that the specific evaluation approach targets. The aim is to reveal the effect of a specific evaluation approach on the inferred effects of different kinds of EOs that the specific evaluation approach targets.

According to the discussions in Chapter 12, how to classify different evaluation problems depends on the nature of the EOs, SESes, and effect mechanisms. I can roughly classify the evaluation problem into two categories of sub-problems.

First, when an evaluation problem has a true outcome, which evaluation methodology can yield the inferred effect approaching the true outcome?

Second, when an evaluation problem has an undefined evaluation outcome, are there any valid evaluation methodologies?

These two fundamental meta-evaluation problems are rarely discussed.

13.3 What Are The types and Natures of EOs, AOs, and Their Effect Mechanisms?

As discussed in Section 8, the essence of the evaluation is to uncover the effect of the EO on the AO. A fundamental question remains: What are the types and nature of the EOs, AOs, and their effect mechanisms?

This is not a trivial issue, as how to uncover the effect of an object is the fundamental issue in almost any discipline. Through a systematic understanding of different EOs, AOs, and their effect mechanisms, we can gain a deeper understanding of the evaluation problems.

13.4 How to Propose Effective and Efficient Evaluation Models?

This section is based on our previous work [224]. I have made significant simplifications.

Effectiveness in evaluation means achieving the desired evaluation outcomes that closely align with the true outcomes, while efficiency refers to achieving these outcomes with minimal resource expenditure.

The key to the effectiveness and efficiency of evaluations in different scenarios is to establish a series of evaluation models that ensure the transitivity of the primary characteristics. I define *a perfect evaluation model* as one under which we can infer the true effect of the EO, e_t.

The *discrepancy threshold* ϵ is defined as *the difference between the true effect* e_t *and the inferred effect* e_i . In reality, how to perform an efficient and effective evaluation with controlled discrepancies is the most important engineering issue. That is how to strike a balance between ensuring the discrepancy threshold and managing the associated costs.

By disregarding the accuracy of an evaluation model, conducting evaluations solely through fully exploring the evaluation model space may indeed result in maximum confidence. However, this approach also comes with a significant drawback - the exorbitant cost it entails.

When creating an evaluation model, the discrepancy threshold ϵ, which is a discrepancy limit that can be tolerated in an evaluation scenario, holds the potential to exert a profound influence on the evaluation outcome and, in certain instances, it would give rise to grave concerns, particularly in the context of safety-critical tasks where failure could lead to detrimental side effects such as harm, loss of life, or significant environmental damage. So, after thoroughly understanding the stakeholders' evaluation requirements, a risk function $\gamma(\cdot)$ could be predefined. When the stakes are high and there is a greater risk associated with the evaluation outcomes, it becomes imperative to have a lower discrepancy threshold between the inferred effect under an evaluation model and the true effect. This is because the potential consequences of making a wrong decision or drawing inaccurate conclusions become more significant.

In creating an evaluation model, we use the notation $m(\cdot)$ to represent this modeling process. The accuracy of an evaluation model is decided by $m(\cdot)$.

The discrepancy function of the evaluation outcomes disc$(\cdot)$ between M_g under a specific evaluation model, g, and M_p under a perfect evaluation model, p, is defined as follows:

$$\begin{cases} \text{discrepancy threshold } \epsilon = |\mathrm{e_t} - \mathrm{e_i}|, \\ \text{discrepancy threshold } \epsilon = \gamma^{-1}(\cdot), \\ \mathrm{M_p} = \mathrm{M_p}(\mathrm{k_1}, \cdots, \mathrm{k_n}), \\ \mathrm{disc}(\mathrm{M_g}, \mathrm{M_p}) = \mathrm{disc}\Big(\rho\big(\mathrm{m}(\mathrm{M_p})\big), \rho(\mathrm{M_p})\Big), \\ \mathrm{accuracy}(\mathrm{M_g}) = \mathrm{accuracy}\big(\mathrm{m}(\mathrm{M_p})\big), \\ \mathrm{cost} = \psi(\|\mathrm{M_g}\|). \end{cases} \tag{13.1}$$

In the formulation, $\rho(\cdot)$ is a measurement function. Besides, we define the evaluation cost as the product of a constant ψ and the space capacity of M_g. This cost factor allows us to incorporate the resource constraints and practical considerations associated with the evaluation process. Based on the above formulation, the evaluation issue of balancing evaluation cost and the discrepancies in the evaluation outcomes can be framed as an optimization problem. The objective is to minimize the evaluation cost, represented by cost, while ensuring the discrepancies in the evaluation outcomes, denoted as disc(M_g, M_p), do not exceed a predefined discrepancy threshold ϵ.

The optimization problem can be formulated as follows:

$$\arg\min \mathrm{cost}(M_g) \quad \text{subject to} \quad \mathrm{disc}(M_g, M_p) < \epsilon, \tag{13.2}$$

or

$$\arg\max \mathrm{accuracy}(M_g) \quad \text{subject to} \quad \mathrm{cost}(M_g) < \varepsilon. \tag{13.3}$$

Another related issue is how to ensure evaluation traceability, which involves attributing any divergence in evaluation outcomes to disparities in the underlying ECs, thereby establishing clear and transparent traceability.

According to the third axiom of evaluation, for the same (derived) EO, the divergence in the (derived) Effect can be attributed to disparities in (derived) ECs, thereby establishing evaluation traceability, establishing evaluation traceability.

Conceptually, traceability asks for a quantified mapping between the differences in the input and output of the measurement function $\rho(\cdot)$ through the evaluation process, described by the mathematical model we formulated above. Our previous work [224] reveals that this concept aligns well with the mathematical notation of the gradients of a function, which gives the rate of change in the output for each input variable.

In the context of evaluation, the gradient of evaluation outcomes can be written as follows, which is a matrix or tensor:

$$\nabla\rho(m(M_p)) = \nabla\rho\Big(m\big(M_p(k_1, \cdots, k_n)\big)\Big) = \left(\frac{\partial\rho}{\partial m}\frac{\partial m}{\partial M_p}\frac{\partial M_p}{\partial k_1}, \cdots, \frac{\partial\rho}{\partial m}\frac{\partial m}{\partial M_p}\frac{\partial M_p}{\partial k_n}\right). \quad (13.4)$$

"∇" is the gradient operator, which in this context represents the vector of partial derivatives of a scalar function with respect to its variables. "∂" is the partial derivative operator, indicating the rate of change of a function with respect to one of its variables while holding others constant.

The closed-form mathematical expression is not always available for various EC components in evaluation. Nevertheless, we can follow the method of acquiring gradients in numerical methods by creating perturbations in the ECs for various input variables and observing the differences in the composite evaluation outcomes, thus approximating the gradients.

13.5 Summary

This chapter presented four fundamental evaluation issues regarding true or undefined evaluation outcomes, meta-evaluation, types and natures of EO, AO, and effect mechanisms, and effective and efficient evaluation models.

Chapter 14

Fundamental Evaluatology Methodology

In this chapter, I present the fundamental Evaluatology methodology.

14.1 Challenges in Real-world Evaluations

We refer to the entire population of real-world systems that are used to evaluate a specific EO as the *real-world evaluation system*. Assuming no safety concerns are present, the real-world evaluation system seems like a reasonable choice for the prime candidate for evaluating the EO. Unfortunately, there are three significant obstacles to consider when assessing diverse EOs with a real-world evaluation system.

Firstly, it is formidable or impossible to isolate an SES from the real-world evaluation system. It is very difficult to recognize EOs, AOs, and EXOs. Also, the presence of irrelevant objects in the real-world evaluation system poses a considerable challenge. It is often difficult, if not impossible, to eliminate these confounders. They can complicate the evaluation process by introducing factors that make it challenging to isolate the effects of other irrelevant objects. Also,

Secondly, manipulating the real-world evaluation system is a formidable task, making it virtually impossible to establish controlled environments for evaluating objects. Even if it is possible to isolate an SES from a real-world evaluation, it is very formidable or impossible to change the state space of the EO, AO, and EXOs. Additionally, the interconnected nature of the components of SES and other irrelevant objects further complicates this challenge.

Thirdly, regardless of the nature of the evaluation problems, there is a tendency for the EOs, AOs, and EXOs to exhibit bias towards certain object instances or their state subspace. This bias towards specific groups can limit our ability to fully explore and understand the entire range of possibilities available to us.

In this context, it is very challenging to propose an effective and efficient evaluation model to address the above issue.

14.2 Universal Evaluation Methodology in Complex Scenarios

In reality, there are evaluation scenarios with different levels of complexity.

For a specific evaluation problem that has the true evaluation outcome, I define the *accuracy* as the absolute value of the difference between the true effect and the inferred effect under an evaluation model divided by the absolute value of the true effect. A lower ratio, closer to 0%, indicates a better accuracy.

For clarity, it is important to distinguish among the *overall effect*, *inferred effect*, and *true effect*. The overall effect refers to the observed evaluation outcome obtained directly from the direct AO, encompassing the joint influence of the EO and EXOs. From this overall effect, the evaluator derives the inferred effect—an estimation of the EO's true effect.

However, the inferred effect does not necessarily equal the true effect. In practice, due to incomplete control or imperfect knowledge of the system, the inferred effect serves only as an approximation of the true effect, and the accuracy metric quantifies the deviation between them.

14.2.1 Concepts of Perfect and Imperfect SESes

I defined *a perfect SES* as *an evaluation model that can fully and accurately identify the EXO that impacts the AO, isolate the irrelevant objects, and infer the true effect of an EO.*

A perfect SES has four unique characteristics that differ from those of real-world evaluation systems. First, it can correctly recognize the EO, AOs, and EXOs. Second, it can completely isolate irrelevant objects. Third, under a perfect SES, we can infer the true effect of the EO. Fourth, we can freely manipulate the full space of SES. This flexibility would enable researchers to explore various scenarios and assess EOs under a range of conditions, enhancing the depth and breadth of the evaluation process.

However, in reality, it's challenging or even impossible to isolate all irrelevant objects or identify all EXO to establish a perfect SES under which we can infer the true effect. Even in the worst cases, we only partially know the SES, as we discussed in Section 14.

I define *the imperfect SES* as *an evaluation model that partially and inaccurately recognizes AOs, EXOs, and isolates the irrelevant objects.* Under an imperfect SES, the true effect of an EO can only be approximately achieved.

While we can strive to create more accurate evaluation models, it is crucial to recognize the inherent limitations and constraints that exist. Due to the limits of our capacity and capability in understanding objects, in reality, it is much more possible that we can only establish an imperfect SES than a perfect SES.

14.2.2 Simple SESes

A perfect or imperfect SES encompasses a vast number of objects, their respective quantities, as well as a large population of EOs, AOs, and EXOs, which leads to significant evaluation costs. However, to address this challenge, it is important to simplify the perfect or imperfect SES in two key ways.

Firstly, to reduce the evaluation costs associated with a large number of objects and their respective quantities, it is crucial to identify and focus on the objects and their quantities that have a significant impact on the evaluation outcomes.

By identifying and prioritizing these targets, researchers can streamline the evaluation models and allocate resources more efficiently. Negligible objects and their quantities, which have a minimal effect, can be excluded or controlled for, thereby reducing complexity and costs. It is worth emphasizing that the simplification will inevitably lead to a decrease in the accuracy of the evaluation model.

Secondly, sampling techniques can be employed to manage the extensive populations of the EO, AO, and EXO. Rather than evaluating every single possibility, researchers can select representative samples that capture the diversity and range of the population. This approach allows for a more manageable evaluation model while still maintaining a good level of coverage and representation.

After these steps, we obtain *a simple SES*, which we define as *the evaluation model that simplifies the perfect or imperfect SES*. A simple SES reduces the objects and their quantities that have little impact on the effect and samples the state space of the perfect or imperfect SES. Accordingly, a simple EC is the simple SES upon the removal of the EO.

A simple SES can be considered a sample of a perfect or imperfect SES. In order to measure the extent to which the statistics of a simple SES can infer the parameters of a perfect SES, we employ the criterion of confidence level and interval.

The confidence level provides us with the probability that the estimated parameters of a perfect or imperfect SES fall within a specific range of values. Meanwhile, the confidence intervals establish a range of values within which we can reasonably expect the true parameters of a perfect or imperfect SES to fall. By utilizing these statistical measures, we can assess the degree of alignment between the statistics of a simple SES and the parameters of a perfect or imperfect SES. This allows us to gauge the efficiency and effectiveness of the simple SES.

By implementing these simplifications in a simple SES, researchers can strike a balance between comprehensiveness and feasibility. The simple SES allows for a more effective and efficient evaluation of EO within the EC framework, mitigating the challenges posed by evaluation costs and the complexity of the perfect or imperfect SES.

14.2.3 Mathematical Formulations of Different SESes

This subsection presents the formulations of different SESes.

First, a sample of the perfect or imperfect SES, M_{ps}, is taken from the perfect or imperfect SES, M_p, and I introduce the notation $s(\cdot)$ to represent the sampling process. Additionally, the Simple SES M_s reduces the independent variables within the sample of the perfect or imperfect SES, M_{ps}, by excluding those that have minimal impact. To formalize this process, I introduce the notation $m(\cdot)$ to represent the modeling process consistent with that in Section 13.4.

When transforming a perfect or imperfect SES into a simple SES, it is essential to maintain the transitivity of the following characteristics:

$$M_{ps} = s(M_p). \tag{14.1}$$

The sample of the perfect or imperfect SES M_{ps} is obtained through a sampling process $s(\cdot)$ on the perfect or imperfect SES M_p.

$$M_{ps} \subset M_p. \tag{14.2}$$

The sample of the perfect or imperfect SES M_{ps} is a subset of the perfect or imperfect SES M_p.

$$M_s = m(M_{ps}). \tag{14.3}$$

The simple SES M_s is obtained through a modeling process $m(\cdot)$ on the sample of the perfect or imperfect SES, M_{ps}.

For each element m_s in the simple SES M_s, we denote the corresponding element in M_{ps} as $m_{p's} \in M_{ps}$.

$$m_s = (k_1, \cdots, k_{n'}) \in M_s. \tag{14.4}$$

Each element m_s in the simple SES M_s consists of a set of independent variables $(k_1, \cdots, k_{n'})$.

$$m_{p's} = (k_1, \cdots, k_{n''}) \in M_p, \text{ where } n' \leq n''. \tag{14.5}$$

The corresponding element $m_{p's}$ in the sample of the perfect or imperfect SES M_{ps} consists of a set of independent variables $(k_1, \cdots, k_{n''})$. The n'' is greater than n', ensuring that the corresponding element in the sample of the perfect or imperfect SES includes at least as many independent variables as that of the element in the simple SES.

14.3 What is a Benchmark?

Benchmarks are extensively employed across various disciplines, albeit lacking a formal definition. Based on the science of evaluation, we propose a precise delineation of a benchmark as *the EXOs in a simple EC* according to the definition in Sections 14.2.1 [1].

[1] This definition is based on the discussion with Dr. Lei Wang and Mr. Chenxi Wang.

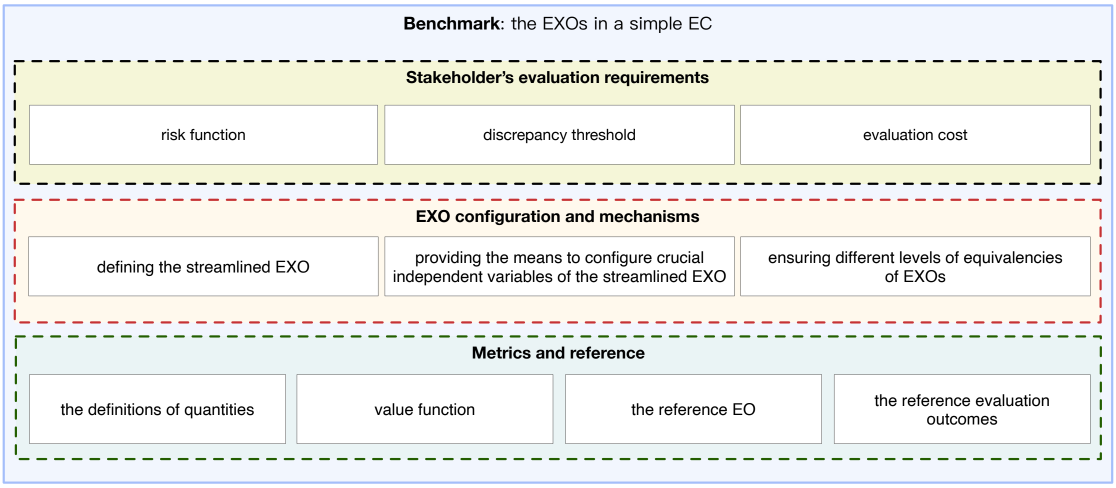

Figure 14.1: A benchmark comprises three essential constituents.

Specifically, a benchmark refers to the EXOs in a simple EC, where the combined effect of the EXOs and the EO is exerted on the AO.

Within the framework of this definition, a benchmark comprises three essential constituents, as shown in Figure 14.1. The first constituent is the *stakeholder's evaluation requirements*, which encompass various factors. These include the risk function, which evaluates the potential risks associated with the benchmark. Additionally, the discrepancy threshold, which determines the acceptable level of deviation from the true evaluation outcomes, is considered. Lastly, the evaluation cost is taken into account, and the resources required for conducting the evaluation are assessed. By considering these elements, the benchmark can effectively address the evaluation requirements of stakeholders.

The second constituent of the benchmark framework is the *EXO configuration and mechanisms*. This includes several elements crucial for the benchmark's effectiveness. It involves defining the streamlined EXOs; it provides the means to configure crucial independent variables of the streamlined EXOs; it ensures different levels of equivalencies of EXOs. By considering these EXO configurations and mechanisms, the benchmark can provide a comprehensive and standardized approach to different evaluation issues.

The third constituent is the *metrics and reference*, which includes the definitions of quantities, the value function, the reference EO, and the reference evaluation outcomes.

14.4 Fundamental Methodology Under a Perfect SES

As shown in Figure 14.2, this section proposes a fundamental Evaluatology methodology under ideal conditions where we can isolate and manipulate a perfect SES [2].

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

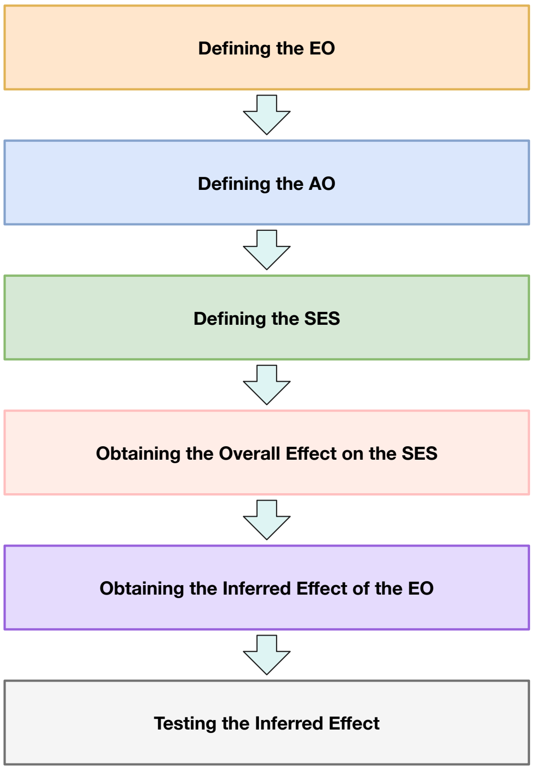

Figure 14.2: A fundamental Evaluatology methodology under ideal conditions where we can isolate and manipulate a perfect SES.

14.4.1 Defining the EO

The first and most fundamental step in Evaluatology is to clearly characterize and define the EO. This process goes far beyond merely naming or labeling the EO; this requires rigorous specification of its boundaries and configurations, including the internal components and their interconnections. Without such a precise definition, subsequent comparisons across different EOs or even across different instances of what is assumed to be the "same" object are rendered ambiguous and potentially invalid. The EO is the *cause object* in the causal structure of evaluation—it is the entity whose effect we aim to uncover or quantify.

A central motivation for this step lies in the principle of comparability. In any evaluation, the validity of results depends on ensuring that the EOs under study are indeed instances of the same class according to the definition of object in Section 3.1. For example, two entities may both be labeled as "processors", but if one refers to a central processing unit (CPU) and the other to a graphics processing unit (GPU), then their fundamental architectures, functions, and operational contexts differ to such an extent that direct comparison would yield misleading or nonsensical conclusions. Even within a narrower category, such as CPUs, the EO may not be monolithic: the same processor can operate under multiple configurations (e.g., with or without turbo boost enabled, or under different thermal or power management settings). Without explicitly specifying which configuration is being treated as the EO, the evaluation outcomes may vary widely and unpredictably.

Following the notation defined in Chapter 12 Formally, we can represent an element of the configuration space of a well-defined EO as: $o_i = \{o_{i1}, o_{i2}, ..., o_{ix}, ...\}$. Here, i denotes the instance of the EO (e.g., "CPU model X"), while x enumerates the possible configurations (e.g., "default mode", "turbo boost enabled", "low-power mode", etc.). Each element o_{ix} thus corresponds to a specific operational configuration of the i_{th} component. Selecting which configurations are to be evaluated is not merely a technical choice but a scientific one, since it determines the interpretability and comparability of the evaluation results.

From a practical perspective, the configuration space O_i is often extremely large, and it is usually neither feasible nor necessary to evaluate the object under every possible configuration. Consequently, evaluators often reduce this space by selecting a representative configuration o_{ij}. Such a configuration is typically chosen because it achieves high or stable performance under a broad set of ECs, or because it reflects the most common real-world usage scenario. For example, in CPU evaluation, one might fix the processor to its default settings or select the configuration that yields the best performance on a majority of standard workloads.

However, this strategy comes with a critical caveat: no single configuration is guaranteed to perform optimally across all ECs. A CPU setting that excels in compute-intensive workloads may underperform in energy-constrained scenarios. Similarly, a software algorithm tuned for accuracy may sacrifice efficiency when faced with real-time constraints. Therefore, while adopting a representative configuration can greatly reduce the evaluation cost and complexity, it also risks oversimplification, potentially leading to biased or incomplete conclusions.

This tension—between comprehensiveness (evaluating all possible configurations) and feasibility (reducing the configuration space)—is at the heart of defining the EO. Striking the right balance requires a careful understanding of both the object's intrinsic variability and the intended scope of the evaluation. If the goal is a broad scientific characterization, more configurations may need to be included. If the goal is practical deployment guidance, a smaller number of well-chosen configurations may suffice.

In summary, defining the EO is not a trivial preliminary step but a foundational process that determines the validity and interpretability of the entire evaluation. It ensures that comparisons are meaningful, outcomes are reproducible, and conclusions can be properly attributed to the EO itself rather than to hidden or uncontrolled variations in its configurations.

14.4.2 Defining the AO

Once the EO is clearly defined, the second step in Evaluatology is to specify the AO—the minimal system on which the effects of the EO can be directly measured or tested, as we defined in Section 8.2.

The AO serves as the *effect-receiving system*. The definition of AO is critical, as it determines the scope of measurement or testing.

Formally, the AO can be conceptualized as a composed system:

$$a_j = \{c_{j1}, c_{j2}, ..., c_{jk}\}, \tag{14.6}$$

where c_{jk} represents the k_{th} configuration of j_{th} *indispensable components*—the minimal set of functional elements that collectively allow the EO's effects to manifest and be empirically measured or tested. The AO must be sufficiently complete to support the EO's operation, yet minimal enough to avoid the introduction of unnecessary confounding factors. This principle—*completeness without redundancy*—ensures that every observed effect is causally meaningful and traceable to the EO within the defined boundaries of the system.

In one category of evaluation problems, the EO cannot function independently; it must be embedded within a broader system that provides essential infrastructure for its operation. For example, when the EO is a central processing unit (CPU), it cannot be evaluated in isolation, as it depends on memory, storage, and input/output subsystems to execute any meaningful workload. Here, the AO is the *computer system* itself—a minimal operational environment that includes all indispensable components required to sustain computation. Specifically, such an AO typically consists of:

- *Primary computational unit (CPU):* the EO whose performance or characteristics are under evaluation;

- *Memory subsystem:* provides working storage for instruction and data execution, directly influencing latency and throughput;

- *Persistent storage:* enables data loading, checkpointing, and I/O operations necessary for full program execution;
- *Interconnect and power subsystem:* ensures communication and energy delivery among components, maintaining operational stability.

These indispensable components form the minimal closed system in which the EO's effects (e.g., execution time, power consumption, cache efficiency) can be observed, measured, and meaningfully compared. Adding or omitting any of these components would fundamentally alter the measurement context and, therefore, the interpretation of the results.

In other categories of evaluation problems, the AO does not necessarily have to include the EO itself. In many domains, particularly outside of engineering, the AO may exist as an independent system that responds to the EO's intervention or influence. In pharmacological evaluation, for instance, the EO is the pharmaceutical compound, while the AO is the human body, a complete biological system within which the EO's biochemical effects manifest. The body already constitutes a self-sufficient measurable system, and the EO acts upon it externally. This case contrasts with some engineering systems, where the EO is typically an internal component of the AO.

Thus, depending on the domain and causal structure, the EO–AO relationship can be classified into two general categories:

- *Inclusive Relation:* The AO includes the EO as one of its indispensable internal components, such that the EO operates as a subsystem within a larger measurable whole (e.g., CPU within a computing system, or an engine within a vehicle).
- *External Relation:* The AO exists independently of the EO, but serves as the environment or organism upon which the EO acts (e.g., a human body receiving a drug, or an ecosystem exposed to a pollutant).

Regardless of the relation type, defining the AO requires balancing two fundamental criteria: *operational completeness* and *causal clarity.* An AO that is too narrow will fail to capture the full manifestation of the EO's effect. At the same time, one that is too broad may dilute the EO's contribution with unrelated interactions. For instance, evaluating a CPU within a large-scale data center might introduce network latency, scheduling, and resource-sharing effects unrelated to the CPU's intrinsic behavior. Conversely, evaluating it in a minimal system (with controlled memory and I/O subsystems) provides a clearer mapping from cause (EO) to effect (AO).

In practice, defining a suitable AO involves iterative abstraction and reduction, which depends on the expertise. Evaluators begin by identifying all components required for the EO to function, then systematically remove non-essential elements until a stable and measurable system remains. This minimal AO represents the tightest causal interface through which the EO's effect can be empirically observed without external confounding interference.

In summary, the AO is not merely the recipient of the EO's influence but the operational bridge between the EO's intrinsic behavior and the measurable or testable outcomes that form the basis of evaluation. A well-defined AO ensures that evaluation results are both *causally valid*—accurately reflecting the EO's true effect—and *empirically grounded*—derived from a physically coherent and reproducible system. Defining the AO thus transforms evaluation from a descriptive exercise into a structured causal analysis, enabling consistent interpretation across domains and systems.

14.4.3 Defining the SES

After defining both the EO and the AO, the next step in Evaluatology is to establish the SES. The SES serves as a self-contained and isolable environment within which the causal effect of the EO can be meaningfully observed, measured, and analyzed. "Self-contained" indicates that the SES incorporates all EXOs that might affect the measurable or testable outcome. It is "isolable" in that other objects can be disentangled from those objects included in the SES.

Formally, we can express the SES as:

$$SES = \{EO, AO, EXO\}, \tag{14.7}$$

where the *EO* represents the causal source under investigation, the *AO* is the measurable system on which the EO's effects manifest, and the *EXO* (Essential External Objects) encompasses all additional objects that can influence the AO alongside the EO.

(1) The structural role of SES. The SES defines the boundary of what is under evaluation. In practice, both EO and EXO exert influence on the AO, and the measured or tested outcome represents the *overall effect*—the joint consequence of EO and EXO on AO. The central objective of evaluation, therefore, is not simply to measure this overall effect, but to *isolate and quantify the true effect of the EO* within it. Only by distinguishing the EO's true effect from that of EXO can one make valid causal claims and ensure that observed differences truly reflect changes in the EO rather than uncontrolled variation in external conditions.

(2) Composition and interactions. The SES captures the structural and causal relationships among the three components:

$$EO \rightarrow AO \leftarrow EXO. \tag{14.8}$$

The arrow from EO to AO represents the causal path of interest—the true effect we seek to uncover—while the influence from EXO to AO represents external interference, modulation, or context. In many cases, EO and EXO may also interact, forming compound effects (e.g., a CPU's performance interacting with the compiler or workload). The design of the SES, therefore, requires careful selection and control of EXO to make EO's effect identifiable.

In biomedical evaluation, if the EO is a pharmaceutical compound and the AO is the human body, the SES also includes EXOs such as environmental conditions, lifestyle factors, or psychological stressors. These EXOs likewise influence the measured outcomes (e.g., blood concentration, physiological response) and must be controlled or randomized to separate the drug's intrinsic effect from external perturbations.

(3) Minimal completeness and causal isolation. The definition of SES follows two guiding principles:

1. *Minimal completeness:* The SES must include all indispensable entities necessary for the EO's effect to manifest on the AO—nothing more, nothing less.
2. *Causal isolability:* The SES must be structured such that the EO's effect can be distinguished, either through experimental control (fixing EXO) or inferential modeling (adjusting for EXO).

In biomedical or social systems, minimal completeness may consist of the EO (e.g., drug, policy), the AO (e.g., patient, population), and EXOs (e.g., environment, behavioral factors). Across both cases, isolating the EO effect requires that EXOs be explicitly identified, documented, and either held constant, randomized, or modeled.

(4) Practical implications. Defining a proper SES is not merely a formal step—it has practical implications for experimental design, reproducibility, and interpretation. An ill-defined SES can lead to confounded results where observed performance differences stem not from the EO but from uncontrolled EXO variations. Conversely, an overly restricted SES may eliminate meaningful variability, leading to results that lack external validity. The art of evaluation lies in finding the correct balance between environmental control and representational realism.

(5) Summary. In summary, the SES represents the minimal evaluation environment that integrates:

- the *EO* —the causal source under investigation,
- the *AO* —the measurable or testable system receiving the effect, and
- the *EXO* —the essential external factors jointly influencing the AO.

The SES establishes the causal stage upon which evaluation occurs: EO and EXO jointly act on AO to produce observable outcomes. Evaluation, in turn, seeks to disentangle these overall effects to reveal the EO's true effect. Without a rigorously defined SES, evaluation becomes ambiguous—its conclusions may reflect uncontrolled context rather than true causality. By grounding evaluation in the structured triplet $\{EO, AO, EXO\}$, Evaluatology transforms evaluation from a descriptive practice into a scientifically interpretable causal analysis.

14.4.4 Obtaining the Overall Effect on the SES

Once the SES is clearly constructed, comprising the EO, the AO, and the EXO, the next stage is to obtain the *overall evaluation outcome* on the SES, which is the overall effect on the AO induced by the EO and the EXO.

This process represents the empirical phase of Evaluatology, where the theoretical structure of the SES is instantiated into measurable experiments or simulations. The overall effect is not the true effect of the EO alone, but the observable result that arises from the combined interaction of EO, AO, and EXO under a wide range of ECs.

14.4.5 Obtain the Inferred Effect of the EO

After obtaining the overall evaluation outcomes on the SES, the next and most critical stage of Evaluatology is to *infer the effect of the EO* on the AO, which we call *the inferred effect.* Please note the difference between the inferred effect and the true effect: the true effect is the target that the inferred effect approaches.

This step transforms the descriptive phase of evaluation—where the overall outcomes are observed—into a causal phase, where the effect of the EO is isolated and quantified. In essence, the goal is to distinguish the portion of the overall effect that is truly attributable to the EO itself, as opposed to the contributions of the EXOs or random variations within the ECs.

14.4.6 Testing the Inferred Effect

After the inferred effect of the EO has been obtained, the final step of Evaluatology is to validate this inferred effect. This stage treats the inferred effect of the EO as a statistical hypothesis and examines whether it is consistent with empirical evidence collected from additional ECs or repeated measurements. The goal is to determine whether the inferred effect genuinely reflects the true effect of the EO, rather than random fluctuations or residual confounding. Please refer to Section 5.9.

14.5 Summary

This chapter presented three evaluation models: perfect, imperfect, and simple SESes, formally defined the benchmark, and proposed a fundamental evaluation methodology under a perfect SES.

Chapter 15

Hierarchical Formalizations of Well-defined SESes

Based on our previous work [224], this chapter presents hierarchical formalizations of well-defined SESes, which work for a class of evaluation problems where EC can be hierarchically defined. This kind of evaluation problem often manifests in the field of computer science.

15.1 Hierarchical Definition of SESes

In Chapter 8, I presented the definition of an SES. There are many ways to define an SES. In this section, we propose a hierarchy definition of an SES. This approach works quite well for a class of evaluation problems for which the EC and the EO can be well-defined. That is to say, we can replicate an SES fully accurately.

We begin by defining an SES from the problems or task spaces that the EO's stakeholders need to address. The reason is as follows. First, the concerns and interests of the relevant stakeholders are at the core of the evaluation. These concerns and interests are best reflected through the problems or tasks they must address, which provide a reliable means to define an SES. Second, utilizing the same problem or task can ensure the comparability of evaluation outcomes.

Taking Case One, presented in Chapter 8, as an example, we observe that in Case One, an SES encompasses numerous constituents. Notably, we identify five primary components of an SES from top to bottom. Figure 15.1 shows a typical example.

The first top component is a set of equivalent definitions of problems or tasks (in short, problems). While the problem itself serves as the foundation for the evaluation process, it cannot solely serve as the evaluation itself because the problem is often abstract and requires further instantiation to determine its specific parameters. The second component is a set of collective problem instances, each of which is instantiated from an element of the first component. Unlike the first component, an equivalent problem instance is specific and can serve as the evaluation directly.

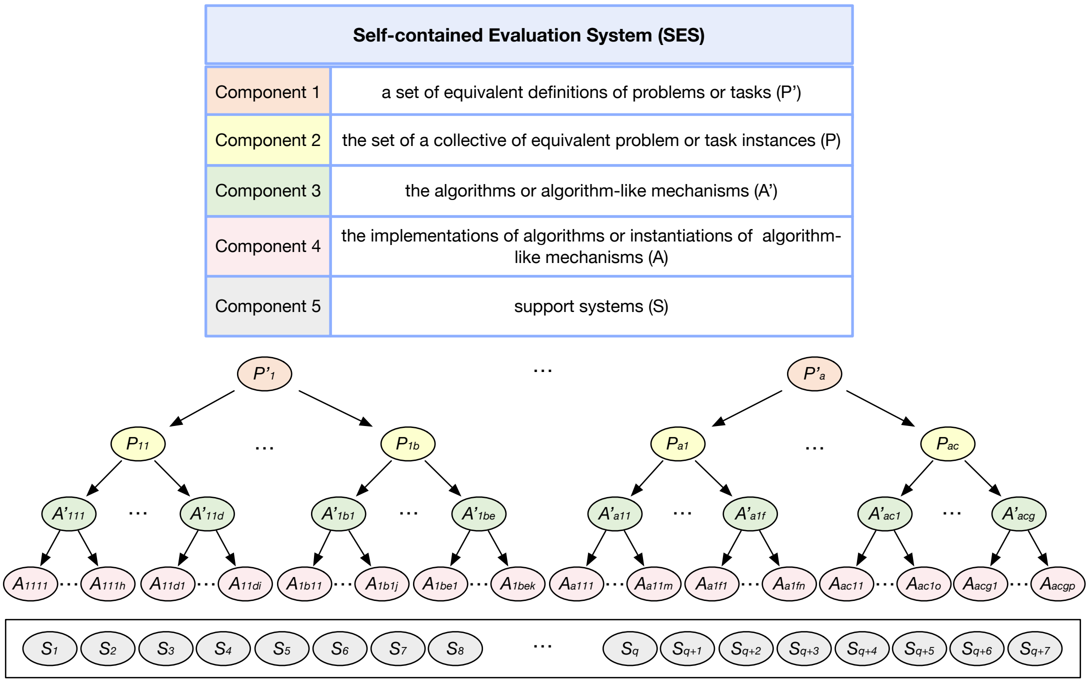

Figure 15.1: The hierarchical definition of an SES.

After a problem instance is proposed, it is necessary to figure out a solution. The third component consists of the algorithm, each of which provides the solution to a specific problem instance.

The fourth component encompasses the implementation of an algorithm. Its implementation involves understanding the algorithm and implementing it in a specific system, which I call *the support system*. This process ensures that the algorithm can effectively and efficiently solve the intended problem instance within the given context.

15.2 Formalization of SESes

To lay the groundwork for the formalization of an SES, it is essential to establish a clear understanding of some crucial notations. The notations P' and S represent two crucial components. P' represents the problem space. S represents the support system space. p', s, o is an element of P', and S, respectively. We note $p' \in P'$, $s \in S$.

In addition to the aforementioned notations, we also define several other fundamental notations. For each problem, $p'_i \in P'$, there is a set of problem instances noted as P_i. For all problems in P', there is an entire collection of problem instances, which can be noted as (p'_i, P_i). The entire set of problem instances P can be defined as the union of all P_i.

We introduce the notation A' to represent the algorithm space. for each problem instance p_{ij} in the problem instance space P_i, the algorithm space A' could be noted as

(p'_i, p_{ij}, A'_{ij}). A' is a union of all A'_{ij}.

We introduce the notation A to represent the instantiations of the algorithm space. This space, A, consists of instantiations of algorithms that are associated with each problem instance, algorithms, and support systems. Specifically, for a given problem p'_i in the problem space P', for each instance p_{ij} in the corresponding instance space P_i, for each algorithm a'_{ijk} in the algorithm space A'_{ij} on each support system s_l in the support system space S, we define the set of instantiations of algorithms as $(p'_i, p_{ij}, a'_{ijk}, s_l, A_{ijkl})$. A is a union of all A_{ijkl}.

By introducing these notations, we establish a comprehensive framework that allows us to delineate the various components of an SES and their respective roles. This formalization enhances our understanding of the key components and their relationships within the EC framework.

We note the EC space as C and the EO space as O. Based on the concepts of the SES defined in Chapter 14, we formalize an SES as $M = C \times O = P' \times P \times A' \times A \times S$.

Likewise, we use symbols to denote a real-world evaluation system (M_r), a perfect or imperfect SES (M_p), and a simple SES (M_s). These are represented as:

$$M_r = P'_r \times P_r \times A'_r \times A_r \times S_r, \tag{15.1}$$

$$M_p = P'_p \times P_p \times A'_p \times A_p \times S_p, \tag{15.2}$$

$$M_s = P'_s \times P_s \times A'_s \times A_s \times S_s. \tag{15.3}$$

These symbols help us distinguish and calculate the various components of ECs and SESes in different contexts.

In the realm of EC spaces, the concept of equivalent ECs plays a significant role. Two EC spaces, denoted as C_1 and C_2, are considered to be equivalent ECs if and only if there exists a bijection, denoted as β, between the two spaces: $\beta : C_1 \mapsto C_2; \beta^{-1} : C_2 \mapsto C_1$. A bijection is a function that is both injective (one-to-one, no two inputs map to the same output) and surjective (onto, every output has a corresponding input), creating a perfect one-to-one correspondence between two sets. This equivalence is denoted as $C_1 \sim C_2$.

15.3 Summary

This chapter presents a hierarchical definition of a well-defined SES that consists of problems, problem instances, algorithms, algorithm implementations, and support systems, which are often found in computer science and engineering. Finally, we formally formalize SESes mathematically.