Rethinking Optimization: A Systems-Based Approach to Social Externalities

Pegah Nokhiz[1*], **Aravinda Kanchana Ruwanpathirana**[2*], **Helen Nissenbaum**[1]

[1]Cornell University, Cornell Tech
[2]NUS School of Computing, National University of Singapore
{pegah.nokhiz, kanchana.ruwanpathirana}@gmail.com, hn288@cornell.edu

Abstract

Optimization is a widely used tool for decision-making across various domains, valued for its ability to improve efficiency and resource allocation. However, poor implementation practices can lead to unintended consequences, particularly in socioeconomic contexts where externalities (costs or benefits experienced by third parties outside the optimization process) are significant. To propose solutions to use optimization responsibly, it is crucial to first *characterize* the involved stakeholders, their goals, and the types of subpar practices causing unforeseen outcomes. This task is complex because affected stakeholders often fall outside the direct focus of the optimization process. Furthermore, incorporating these externalities into the optimization process requires going beyond traditional economic frameworks, which often focus on describing externalities but fail to address their normative implications or their interconnected nature, and feedback loops.

This paper suggests a framework that combines systems thinking with the economic concept of externalities to tackle these challenges. This approach aims to characterize *what went wrong, who was affected*, and *how (or where) to include them in the optimization process.* Economic externalities, along with their established quantification methods, assist in identifying "who was affected and how" through stakeholder characterization and measurement. Meanwhile, systems thinking which is an analytical approach to comprehending relationships within complex systems, provides a holistic, normative perspective. Systems thinking contributes to an understanding of the interconnections among externalities, feedback loops, and determining "when" to incorporate them in the optimization. Together, these approaches create a comprehensive framework for addressing optimization's unintended consequences, balancing descriptive accuracy with normative objectives. This integration is applied to examine three common types of subpar practices in this paper: ignorance, error, and prioritization of short-term goals, providing actionable insights for a range of use cases.

1 Introduction

Optimization is a key tool for decision-making in various domains, including industrial engineering, urban planning, social systems, and business, enabling efficient resource allocation and driving advancements in areas like artificial intelligence and machine learning (Bottou 2010). However, optimization simplifies the complexity of real-world systems by focusing on specific variables or objectives, often neglecting the broader social, environmental, or systemic effects of decisions. While this approach can provide clear and actionable solutions, it results in poor practices in applying optimization that frequently lead to unintended consequences (Berk et al. 2021; Buolamwini and Gebru 2018; McMillan 2011). Optimization, often carried out as a descriptive process, underpins the creation of many computational tools like machine learning models. However, it also leads to unintended outcomes that inherently carry normative inquiries (Laufer, Gilbert, and Nissenbaum 2023) that bring attention to the impact of actions on involved parties, in particular, within complex and emergent socioeconomic systems.

As a community focused on ML and AI research, we aim to address these unintended consequences. However, before we can propose effective solutions for such unintended consequences, it is essential to recognize that simply labeling all undesired aftereffects as unintended consequences is problematic. This is because: (1) it is overly broad, failing to account for the nuances and specific differences between cases that require tailored approaches, and (2) it implies that all problematic outcomes are uniformly distributed or objectively problematic, which may not always be the case.

To effectively resolve unintended consequences, it is essential to *characterize* their types (e.g., unexpected benefits or drawbacks (Martin 2003)), causes (e.g., optimizers' ignorance or error), and all affected parties. Only with these characterizations established, targeted solutions can be developed. For example, in a company, optimizing water cooler waiting times by assigning time slots may enhance efficiency but unintentionally disrupt informal social interactions (Fayard and Weeks 2007; Sævarsson 2022; McAlpine 2017), diminishing collaboration, ultimately harming both employees and the organization. To address this, it is crucial to identify *who is affected*, *how*, and *when* communal values should be integrated into the optimization. This characterization would also help pinpoint the underlying subpar practice in deploying optimization (like ignorance, mistakes, or a focus on short-term gains instead of long-term goals (Martin 2003; Merton 1936)) that caused unforeseen outcomes.

However, characterizing optimization's unintended consequences is challenging due to the difficulty in determin-

*These authors contributed equally.

ing their extent in social contexts. Traditional models focus on specific objectives, often overlooking indirect effects like broader social values or stakeholder impacts. For example, benefits such as social interactions in the workplace may be excluded from the water cooler optimization. Identifying all stakeholders (those affected by decisions) and quantifying the impacts on them is critical. This requires both *descriptive* (recognizing stakeholders and their goals) and *normative* considerations (addressing the omission of their interests) for characterizations and developing solutions.

Externalities and Internalization. To address these challenges, we suggest characterizing unintended consequences of optimization in complex socioeconomic systems as externalities, i.e., outcomes that affect stakeholders not directly involved in the decision-making process. While not widely discussed in ML and AI communities in optimization-based computations, externalities are well-established phenomena in economics, defined as costs or benefits incurred by third parties due to an activity (Coase 1960; Pigou 1920). Solutions to externalities often involve internalization methods such as cost-benefit analysis and Pigouvian tax (Pigou 1920), which adjust incentives to account for societal costs or benefits. For instance, carbon taxes aim to internalize the social costs of emissions. This procedure aligns individual incentives with broader societal welfare and it ensures that social and communal values are preserved alongside efficiency (Fleurbaey, Kanbur, and Viney 2021).

The externality framework as a descriptive tool is beneficial in two key ways: (1) It enables the characterization of unintended consequences by distinguishing between positive externalities (benefits to third parties) and negative externalities (costs to third parties). This involves identifying stakeholders and their goals which also helps detect the subpar practice causing the externality. (2) It provides well-established, quantifiable solutions for addressing subpar practices by utilizing commonly used methods like cost-benefit analysis, social welfare assessments, taxes, subsidies, and policy interventions to measure and internalize externalities based on their nature (Pigou 1920; Coase 1960; Arrow 1951; Boardman et al. 2018; Weitzman 1974).

However, open complex systems like societies have emergent behaviors, feedback loops, and interconnections among different parts of the system. These significantly influence social externalities (Fleurbaey, Kanbur, and Viney 2021). Feedback loops (positive or negative) can amplify the effects of social externalities, making them far more significant than initially perceived (Fleurbaey, Kanbur, and Viney 2021; Satz 2010; Kanbur 2004; Stern 2014). For example, negative feedback loops, such as declining morale or eroding trust due to poorly designed water cooler interventions, can spiral into long-term company dysfunction. These dynamics make the management of social externalities (e.g., determining where or when to internalize them) difficult. Economic methods as descriptive toolkits are precise in quantification but they are too narrow in scope to address these challenges.

Systems Thinking of Optimization. To overcome these challenges, we propose incorporating systems thinking as an additional perspective. Systems thinking is a problem-solving and analytical approach that emphasizes understanding the interactions and interdependencies among components of a system within the context of a broader whole (Meadows 2008; Sterman 2000). This holistic mindset to decision-making emphasizes interconnections, feedback loops (where targeted actions could amplify positive effects or mitigate negative feedback loops), and emergent properties within complex systems (Meadows 2008; Dörfler et al. 2024). Unlike traditional optimization that isolates variables, systems thinking considers hidden dependencies and long-term ripple effects. Accordingly, applying a systems thinking mindset in practice, involves systems theoretic mathematical architectures with the same purpose. For example, a layered systems theoretic mathematical architecture (Sarjoughian, Zeigler, and Hall 2001; Voros 2005; Jensen 1970; Smith and Sage 1973; Zhu, Wei, and Ji 2016; Matni, Ames, and Doyle 2024; Inc. 2024) would help with thinking about "when or where should be" the level at which externalities are addressed: lower levels focus on localized stakeholder goals. Higher levels analyze broader overarching goals.

Our Position: Why Neither Concept Alone is Sufficient? Neither systems thinking nor economic externalities alone can fully address the complexity of identifying, quantifying, and internalizing unintended consequences in optimization:

- **Externalities for stakeholders:** Who is affected? How?
- **Systems thinking for interconnected dynamics:** When or where do we internalize externalities? How do these impacts propagate, and what secondary or tertiary effects emerge over time? In the water cooler case, externalities identify the direct loss of collaboration (a cost). Systems thinking shows the ripples through the company affecting long-term performance and when to internalize the cost.

This paper advocates for a combination of both perspectives, in particular, in widely descriptive optimization-based computations in AI/ML with numerous unforeseen outcomes (Laufer, Gilbert, and Nissenbaum 2023; Berk et al. 2021; Buolamwini and Gebru 2018). Combining externalities with systems thinking creates a powerful framework to address unintended consequences. Externalities provide the descriptive foundation, highlighting where optimizations (as another descriptive tool) fall short. Systems thinking adds normative reflections of *what should be*, ensuring that solutions align with societal values and account for complexity. Together, they address the issues noted earlier: how to identify, quantify, and internalize unintended outcomes. Thus, in this *position* paper, we present three key messages:

- To effectively resolve the unintended consequences of optimization, it is essential to move beyond treating all unforeseen outcomes as a single category. Instead, we must carefully analyze what went wrong, identify who was impacted, and determine how and where to incorporate these considerations into the optimization process.
- We propose externalities as a descriptive tool to identify impacted stakeholders, classify effects (costs or benefits), and use proper quantification methods for internalization.
- We propose systems thinking and a layered system-theoretic model for a normative perspective on when

to internalize externalities. This also accounts for local and broader goals, interconnections, and feedback loops within complex systems.

2 Related Work

Economic Externalities. Economic Externalities are unintended costs/benefits experienced by third parties outside market transactions. They arise from unaccounted spillover effects within standard markets, such as pollution (negative) or societal benefits of scientific research (positive). These externalities highlight the misalignment between private incentives and social welfare (Pigou 1920; Coase 1960).

To address externalities, several methods have been developed for internalization. Pigouvian taxes and subsidies correct negative and positive externalities by aligning private and social costs (Pigou 1920). Social welfare functions incorporate externalities into utility functions to measure societal well-being (Arrow 1951). Cost-benefit analysis (CBA) compares costs and benefits of externalities to guide decision-making (Boardman et al. 2018). Regulatory mechanisms like emission caps and Cap-and-Trade systems limit harmful activities (Weitzman 1974). Coasian bargaining offers negotiation-based solutions when property rights are clearly defined and transaction costs are low (Coase 1960). Also, recent research has extended externality analysis to social and environmental contexts, including urban planning, education, and algorithmic systems (Arrow 1969; Fleurbaey, Kanbur, and Viney 2021). However, traditional methods face limitations, such as undermining intrinsic motivations in social interactions, prioritizing efficiency over equity, and failing to account for long-term feedback loops and emergent behaviors (Satz 2010; Kanbur 2004; Stern 2014).

Optimization's Unintended Consequences. Optimization, while powerful, frequently simplifies complex real-world scenarios, leading to unintended consequences. Fairness-aware machine learning models, for example, can inadvertently create inequalities among subgroups (Berk et al. 2021; Stinar and Bosch 2022; Nokhiz et al. 2021, 2024, 2025a,b; Nokhiz 2024; Shelby et al. 2023). Algorithms designed to maximize overall success often misclassify underprivileged groups, prioritizing the majority (Buolamwini and Gebru 2018; Hardt 2014). Stakeholders not directly involved in decision-making are frequently excluded, resulting in their neglect in the optimization process (Lopez 2018; Overdorf et al. 2018). Systems trained on specific domains struggle in new environments (McMillan 2011; Rodger and Pendharkar 2004; Sugiyama, Lawrence, and Schwaighofer 2017). Optimization efforts tend to focus on benefiting high-priority users while disregarding others (Huffaker 2016; Tassi 2016). Risks from experimentation and parameter selection are often shifted onto users (Bird et al. 2016). Lastly, research in reinforcement learning also shows how optimization can amplify unintended outcomes (Amodei et al. 2016; Whittlestone, Arulkumaran, and Crosby 2021; Rathnam et al. 2024).

Systems Dynamics. Systems thinking (and theory) offer valuable frameworks for focusing on the study of interconnections and feedback loops (Meadows 2008; Reader et al. 2022). This approach has been influential in fields such as organizational management, environmental sustainability, and policy design, where traditional reductionist methods often fall short (Sterman 2000). It has been applied across domains to address the challenges of unintended consequences, e.g., prior work shows the importance of systems thinking in decision-making under economic uncertainty, in predicting systems, in understanding systemic risks in sociotechnical AI, and in identifying leverage points for sustainable solutions (Reader et al. 2022; Sterman 2000; Meadows 2008; Bertalanffy 1968; de Troya et al. 2025).

3 Preliminaries and Definitions

In this section, we explore three interconnected aspects of optimization: the mathematical foundations that define and solve problems (§3.1), the stakeholders and externalities that influence and are impacted by optimization (§3.2), and the system's dynamics that govern the interactions and feedback mechanisms within complex systems (§3.3).

3.1 Mathematical Optimization

Mathematical optimization finds the best solution from feasible alternatives (Wang and Zhao 2023) and is vital in fields like economics, AI and ML, engineering, and operations research. The process involves defining the *problem scope* that sets the boundaries (domain, temporal, and physical), specifying the *variables* (x) and the *objective function* ($f(x)$) that quantifies utility, determining the *constraints* (C) for feasible solutions, identifying the corresponding *solution space* (X), and selecting the *evaluation metrics* (M) to assess solutions. The optimization problem is then formalized as:

$$\text{Optimize}_{\ x \in X} f(x)$$

Once a solution is found, it is evaluated to verify its optimality, often requiring multiple iterations where previous solutions/evaluations inform refinements in subsequent steps.

3.2 Stakeholders and Externalities

Optimization, however, leads to unintended consequences. That is, in real-world optimization applications, there are generally agents or groups that are either directly interacting with or indirectly affected by optimization. These individuals are known as the *s*takeholders (Brauers 2013) and could be either direct participants of the optimization process (internal stakeholders) or those who are indirectly affected by the optimization (external stakeholders). Internal stakeholders often set objectives and constraints for the optimization and external stakeholders introduce additional considerations which are known as *externalities* in economics (Pigou 1920; Coase 1960). Externalities can be positive (benefits to third parties) or negative (drawbacks to third parties) and are not reflected in the optimization's objective function.[1]

Addressing externalities in optimization involves figuring out how to internalize them (i.e., quantify them), to ensure

[1]Note that although certain stakeholders are considered internal from an organizational outlook, externalities can still impact them: The organization may fail to add their input into the optimization, preventing them from directly interacting with the optimization.

the optimization accounts for the impacts of externalities on the stakeholders (Bejan 2024). Depending on the specific externality, this could involve different measures such as cost-benefit analysis (CBA) (Boardman et al. 2018) which is a comparison of costs and benefits of externalities and using them as an internalization quantitative framework, social welfare functions (SWF) (Arrow 1951) which are an aggregation of utilities (and disutilities) from externalities, or a Cap-and-Trade System (investopedia 2024), which establishes a fixed limit on allowable externalities and enables the trading of permits to encourage compliance.[2]

In sum, this framework helps characterize and internalize unintended outcomes by discerning positive and negative types, external stakeholders, their goals, and relevant quantifications. However, no single economic method might fully capture the complexity of social cases. Other (even non-economic) methods might be more suitable for certain contexts. Even so, externalities offer valuable guidelines for analyzing stakeholders and the *category* of required quantification methods. Despite their shortcomings, tools like CBA, SWF, and Cap-and-Trade still offer key practical utilities.

3.3 Systems Dynamics

Externalities in complex socioeconomic systems are influenced by dynamic behaviors, feedback loops, and interconnections, which can amplify externalities over time (Fleurbaey, Kanbur, and Viney 2021). These interconnected dynamics make managing externalities challenging, especially in determining when and where to internalize them. While economic methods offer precise characterizations, they are too limited to fully address these complexities. Therefore, we suggest adopting *systems thinking* as a broader analytical approach to gain deeper insights into these issues, as its normative view focuses on how systems should be designed and function, rather than simply describing their current state.

Drawing on systems thinking as a mindset, we also require a corresponding mathematical model to formally analyze systems, i.e., the "theory." Systems theory provides a formal approach to examine how different parts of a system interact and influence each other to form a complex cohesive whole. Systems theory includes various frameworks, however, we refer to the layered architectures (Sarjoughian, Zeigler, and Hall 2001; Voros 2005; Jensen 1970; Smith and Sage 1973; Zhu, Wei, and Ji 2016; Matni, Ames, and Doyle 2024; Inc. 2024). They generally have established layers on different levels of abstraction where each layer addresses different aspects of the system's functioning and control. Layers organize components hierarchically and make understanding complex systems easier and more manageable by dividing them into specialized parts. This in turn helps us identify when and where to address externalities: lower levels focus on localized goals and higher levels analyze broader impacts. We use the following layered architecture:

1. **Physical Layer**: This layer is about direct, tangible components of a system (products, resources, infrastructure, or actions that are physically observable and measurable). It focuses on input-output dynamics, like resource allocation, production levels, or environmental changes.
2. **Regulatory Layer**: This layer sets rules, constraints, and enforcement mechanisms to govern behavior within the system. It introduces regulations or levies to influence entities and address externalities directly.
3. **Supervisory Layer**: This layer ensures real-time monitoring, feedback, and adjustment of systems behavior. It uses physical and regulatory layers' data to dynamically optimize and correct deviations from desired outcomes.
4. **Strategic Layer**: This layer encompasses the long-term planning and optimization of the system, integrating systemic goals, predictive modeling, and stakeholder objectives. It evaluates policies holistically, balancing competing priorities and externalities over time.

[2]The quantification method used depends on the specific problem setting as well as stakeholders' goals in the optimization.

4 Systematic Externality Internalization

We can now explore the use of the layered architecture from §3.3 to analyze an optimization function, internalize externalities, examine the roles of various layers, and explain why externalities are addressed at a specific layer.

4.1 Physical Layer: Operational and Direct System Decisions

In the physical layer, the optimization considers internal stakeholders' goals and the externalities impacting external stakeholders. Let G_{int} be the set of internal stakeholder goals, and $E_p^-(x), E_p^+(x)$ be the unexpected drawbacks and benefits at the physical layer. Let $f_{G_{\text{int}}}(x)$ capture the cost of the solution x based on the goals G_{int}. Let $f_p(x) = f_{G_{\text{int}}}(x) + \gamma \cdot E_p^-(x) - \delta \cdot E_p^+(x)$ where $\gamma, \delta > 0$ are penalty weights for externalities. The optimization can be written as,

$$\min_x f_p(x) \quad \text{subject to} \quad x \text{ satisfies } \{R(x) \leq R_{\max}, C_p(x)\}$$

where $C_p(x)$ is the physical layer constraints, and the resources used $(R(x))$ satisfies $R(x) \leq R_{\max}$, with $R_{\max}$ capturing the resource limitations.

Feedback Mechanism The real-time adjustments would modify the solution based on the changes in resource availability, i.e., setting $x \leftarrow x - \eta \nabla R(x)$ where η is a penalty weight and $\nabla R(x)$ is the gradient of $R(.)$ at x. Using the observed externalities, we would also readjust the parameters δ, γ, e.g., set $\gamma \leftarrow \gamma + \hat{\eta} \frac{\partial E_p^-}{\partial x}$ where $\hat{\eta} > 0$ is an adjustment factor based on stakeholder input.

Why Internalize Externalities at the Physical Layer? Externalities related to aspects such as resource consumption arise directly from actions in the physical layer. Internalizing them here ensures that the system accounts for these impacts in its decisions. The physical layer is an ideal location to internalize them as it directly interacts with them.

4.2 Regulatory Layer: Compliance and Enforcement

The regulatory layer would involve regulatory constraints captured by $C_r(x)$. Furthermore, the regulatory layer would

involve taxation (using a tax function $T(x)$) and subsidies (using a subsidy function $S(x)$) to allow for adjusting the externalities. Let $f_r(x) = f_p(x) + T(x) - S(x)$. This results in the following optimization,

$$\min_x f_r(x) \quad \text{subject to} \quad x \text{ satisfies } C_r(x)$$

The objective at this layer depends on the physical layer.

Feedback Mechanism This layer's feedback involves regulatory adjustments, i.e., continuous monitoring of compliance metrics and updating regulatory constraints ($C_r(x)$) as well as readjusting tax functions $T(x)$ and subsidy functions $S(x)$. The regulatory layer also involves policy feedback where stakeholders and supervisory data necessitate regulatory revisions to the constraints ($C_r(x)$).

Why Internalize Externalities at the Regulatory Layer? Regulations impose necessary constraints on externalities. Internalizing them at this layer ensures that the system complies with legal boundaries and respects societal standards. The regulatory layer enables the optimization to formally internalize externalities arising from regulatory oversight.

4.3 Supervisory Layer: Monitoring and Real-Time Adjustments

In this layer, we monitor system deviations. At each time point t, we consider the current solution x and evaluate the solution using $M(x)$ (evaluation metrics in §3.1) and compare against the target M_{target}. We also have dynamic constraints $C_s(x)$ that ensure the feasibility of solutions in real time. There are also positive externalities E_s^+, and negative externalities E_s^-, in this layer. Let $f_s(x) = \|M(x) - M_{\text{target}}\|^2 + \eta \cdot E_s^-(x) - \theta \cdot E_s^+(x)$ where η, θ are penalties for externalities. Utilizing $f_s(x)$, we get the objective,

$$\min f_s(x) \quad \text{subject to} \quad x \text{ satisfies } C_s(x)$$

Feedback Mechanism Here, feedback involves dynamic corrections, using supervisory data from monitoring systems and deviations from the targets to take corrective actions, i.e., update the solution such that $x \leftarrow x - \alpha \cdot (M(x) - M_{\text{target}})^2$ with α being the learning rate. This layer also prompts externality adjustments, where weights are updated based on observed externalities using $\eta \leftarrow \eta + \beta \cdot \Delta E_s^-$ and $\theta \leftarrow \theta + \beta \cdot \Delta E_s^+$ where β is an adjustment factor. Further, the layer gathers continuous feedback from stakeholders and adjusts decision variables or constraints in response.

Why Internalize Externalities at the Supervisory Layer? The Supervisory Layer plays a critical role in monitoring and refining system operations by evaluating performance and outcomes. Addressing externalities at this level enables the real-time identification and resolution of deviations and inefficiencies. By doing so, the system can maintain alignment with stakeholder goals without having to use higher levels to resolve deviations from the goals.

4.4 Strategic Layer: Long-Term Planning and Sustainability

In the strategic layer, we rely on long-term constraints captured by $C_l(x)$. We also evaluate the overarching system-level and long-term goals (G_{system}) such as sustainability, as well as internal and external stakeholder goals (G_{int} and G_{ext}, respectively) and refine the goals as needed. Let $f_{G_{\text{system}}}(x)$ capture system goals' cost and $f_{G_{\text{int}} \cup G_{\text{ext}}}(x)$ be the cost of stakeholder goals at the strategic level. Let $f_l(x) = \alpha f_r(x) + \beta \cdot f_{G_{\text{system}}}(x) + \gamma \cdot f_{G_{\text{int}} \cup G_{\text{ext}}}(x)$ where $\alpha, \beta, \gamma > 0$ are weights for balancing short-term and long-term objectives. In this layer, we get the following optimization,

$$\min f_l(x) \quad \text{subject to} \quad x \text{ satisfies } C_l(x)$$

Feedback Mechanism Here, feedback involves predictive updates where over periods of time, the deviations from expected outcomes are evaluated. This also involves modification of system and stakeholder goals, i.e., modifying system goals G_{system} to include new long-term sustainability goals and the stakeholder goals G_{int} and G_{ext} to include the goals related to stakeholders with externalities observed.

Why Internalize at the Strategic Layer? Internalizing long-term externalities ensures that the system remains sustainable in the future. At the strategic level, the system interacts with high-level goals of optimization that affect its long-term behavior. Therefore, the goals that are intended to be long-term can be internalized in the strategic layer.

5 Putting It All Together and Use Cases

In sum, our framework for understanding and resolving unintended consequences has the following steps: (1) Begin with the original vanilla optimization, which focuses on objectives related to internal stakeholders, leading to unforeseen outcomes (descriptive). (2) Identify all affected stakeholders, including external stakeholders who are not directly involved in the optimization (and their goals) and utilize the economic framework of externalities to categorize the unforeseen outcome as either an unexpected drawback or benefit (descriptive). (3) Identify the cause of the externality, i.e., the subpar practice responsible, using the information from step 2 (descriptive). (4) Apply a pertinent quantification method, such as cost-benefit analysis (CBA), social welfare functions (SWF), Cap-and-Trade, or taxes, to internalize the externality based on its type (descriptive). (5) Incorporate systems thinking (and the layered model) as an added outlook to address feedback and interconnections (normative). (6) Determine the most appropriate layer for internalizing externalities (normative). The logical overview for the entire framework is also depicted in Figure 1.

Note on Identifying Subpar Practices. Optimization involves various types of subpar practices. A notable advantage of using the economic framework of externalities is its ability to help us identify the specific subpar practice responsible for an externality. This is achieved by first identifying all stakeholders and their goals, as outlined in step 2 above.

We provide three examples to show how observable externalities relate to a specific subpar practice and how to resolve them using our proposed framework. The formulations presented are not the only possible choices but demonstrate how different subpar practices manifest in terms of externality internalization. In our use cases, we identify the corresponding categories (Martin 2003; Merton 1936) of subpar practices, as: 1. *ignorance* which refers to a lack of complete

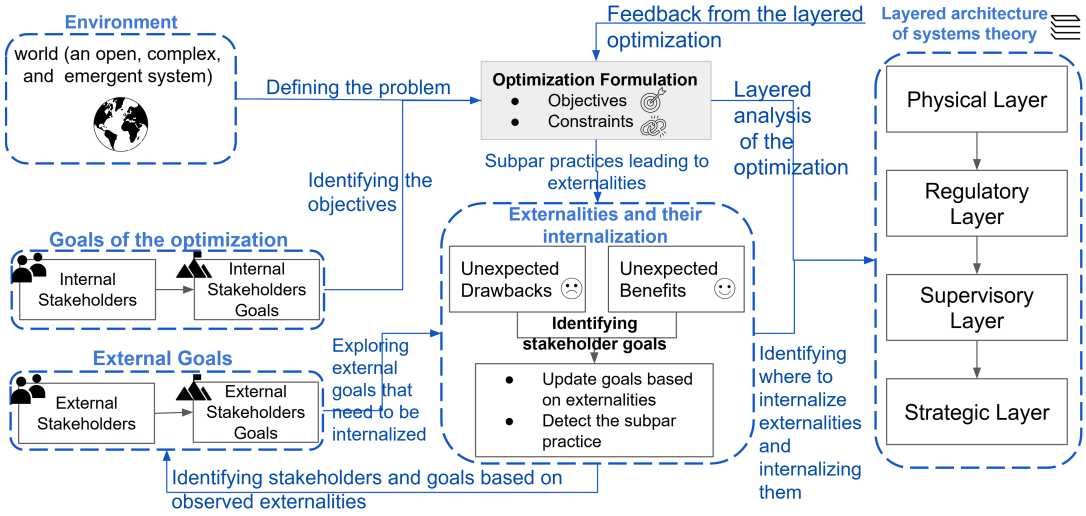

Figure 1: Framework overview: An optimization is defined with internal goals and the core task to optimize for. The optimization has externalities due to subpar practices and unmet external goals. Missed stakeholders, their goals, and the subpar practice are identified to revise goals and internalize externalities within the layered architecture. The optimization is then refined periodically by feedback.

knowledge about the system, variables, or interactions in the optimization process. This often happens when the systems are complex and there is a lack of information to capture a complete picture, 2. *error* which occurs when incorrect assumptions, erroneous models, or incorrect reasoning lead to unintended outcomes. This can be caused by biases, oversights, or relying on incomplete or out-of-date information, and 3. *immediacy of interests* which stems from focusing too narrowly on immediate benefits while neglecting long-term impacts. Decisions made with short-term goals often disregard potential long-term effects or future externalities (Martin 2003; Merton 1936). We can now explore our use cases.

5.1 Office Water Coolers and Social Interactions

The water cooler effect, studied in social sciences (Fayard and Weeks 2007; Sævarsson 2022; McAlpine 2017), explores how informal interactions impact employee productivity and well-being. This example explores a company implementing staggered time slots for water cooler access to enhance productivity while neglecting the water cooler effect and the importance of social interactions. We discuss the unexpected repercussions and how to integrate them into the optimization process to avoid such issues.

Internal Stakeholders and Basic Optimization Formulation. To define the optimization formulation for the water cooler problem, we first need to identify the internal stakeholders of the optimization and their goals. Studies such as (Fayard and Weeks 2007; Sævarsson 2022; McAlpine 2017) investigating the importance of the water cooler effect and articles exploring the concept of stakeholders in a corporate context (Zaichenko 2024), help us identify **company leaders and management** as internal stakeholders. These stakeholders aim to increase productivity, reduce idle time, and ensure employees have sufficient off-task time.

Based on stakeholder goals, we can derive the following vanilla optimization problem. For N employees, assign them to k time slots $T = \{t_1, t_2, \ldots, t_k\}$, each with a capacity Γ_j for $j \in \{1, 2, \ldots, k\}$. Let x_{ij} be a binary variable indicating if employee i is assigned to time slot t_j, and L_j be the loss function for delay cost at t_j. Let $n_j = \sum_{i=1}^{N} x_{ij}$ where $j \in \{1, 2, \ldots, k\}$. The company's objective is:

$$\min \sum_{j=1}^{k} L_j(n_j) \quad \text{subject to} \tag{1}$$

$$\sum_{j=1}^{k} x_{ij} = 1 \; \forall i \in \{1, 2, \ldots, N\},$$

$$n_j \leq \Gamma_j \; \forall j \in \{1, 2, \ldots, k\}, x_{ij} \in \{0, 1\}$$

where $\sum_{j=1}^{k} x_{ij} = 1 \; \forall i \in \{1, 2, \ldots, N\}$ implies each employee is assigned to exactly one timeslot and $n_j \leq \Gamma_j \; \forall j \in \{1, 2, \ldots, k\}$ implies that the number of people assigned to each time slot is within the capacity.

External Stakeholders. We can now use the vanilla optimization formulation to investigate the outcomes of the optimization. The vanilla optimization focuses on minimizing the time wasted at the water cooler, aligning with internal stakeholders' primary goal. However, we need to expand our view to the external stakeholders as well. Referring to the literature on the water cooler behavior (Fayard and Weeks 2007; Sævarsson 2022; McAlpine 2017; Zaichenko 2024), we identify the following external stakeholders (and stakeholder goals): (1) **Employees**, who desire opportunities for social interactions and positive relationships. Note that employees are external stakeholders because their goals were excluded from the optimized objectives, and they had no direct influence on the optimization process, and (2) **Customers**, who seek effective services at an affordable cost.

Layer	Optimization Function	Feedback Loops
Physical	Let T_j be the mean waiting time for an employee in time slot t_j. The congestion cost in this layer is defined as $L_{\text{physical}} = \sum_{j=1}^{k} T_j \cdot n_j$	Adjust the number of time slots k using feedback, based on resource availability.
Regulatory	Ensuring scheduling feasibility: $\min \sum_{j=1}^{k} T_j \cdot n_j$ s.t. $\sum_{j=1}^{k} x_{ij} = 1$ $\forall\, i \in [N], n_j \leq \Gamma_j$ and $x_{ij} \in \{0,1\}$. i is the employee index.	Use feedback to ensure regulatory compliance by checking if $n_j \leq \Gamma_j$ and reassigning individuals to time slots where $n_j > \Gamma_j$.
Supervisory	Adjust policies to optimize productivity: $\min \sum_{j=1}^{k} T_j \cdot n_j + \sum_{i\ell} \text{CB}_{i\ell}(x)$ s.t. $\sum_{j=1}^{k} x_{ij} = 1 \,\forall\, i \in [N], n_j \leq \Gamma_j$, and $x_{ij} \in \{0,1\}$	Use feedback like employee-reported satisfaction and team outputs to modify $C_{i\ell}$ and $B_{i\ell}$.
Strategic	Consider long-term goals like retention cost $R(x)$. The strategic layer aligns goals with satisfaction and productivity: $\min \sum_{j=1}^{k} T_j \cdot n_j + \sum_{i\ell} \text{CB}_{i\ell}(x) + R(x)$ s.t. $\sum_{j=1}^{k} x_{ij} = 1$, $n_j \leq \Gamma_j$ and $x_{ij} \in \{0,1\}$	Periodic evaluation of retention rates/adjustment of long-term policies using supervisory insights.

Table 1: Layered Architecture Layout for the Water Cooler Example

Unintended Consequences. We can see that the vanilla optimization with the sole goal of reducing waiting time is unaware of the impact of the water cooler behavior on the employees or employees' and customers' external goals. Therefore, a solution derived from the vanilla formulation could result in unexpected drawbacks: 1. Strict time slots reduce informal information and idea sharing, which are pivotal for building trust and a positive work culture (McAlpine 2017; Sævarsson 2022). 2. This could harm interpersonal relationships, create tension, increase stress, and reduce performance, potentially leading to employee retention issues and a negative company image, 3. All this could affect customer service quality as well as customer perception of the company. Conversely, there are also unintended benefits, such as potentially enhanced employee privacy due to the reduction of informal interactions (Fayard and Weeks 2007).

Why ignorance? Given the set of unforeseen drawbacks and external goals, an intriguing question to explore is what causes these unexpected issues. In the water cooler example, the company aimed to reduce the time employees spent at the water cooler, with the sole objective of minimizing time. However, for the employees, social interactions were an essential aspect of their objectives. Because employees were excluded from the decision-making process, despite being directly affected by it, their goals were overlooked in the optimization. This disregard for employees' objectives and community dynamics led to unintended drawbacks. Thus, this example highlights the flawed practice of ignorance.

Internalizing Externalities: A Systems Theory Framework. With the externalities and their causes identified, the focus now is to make the optimization more resilient by internalizing these externalities. In the water cooler case, internalizing externalities requires incorporating social interactions. Since the goal is to improve productivity and reduce time spent on unproductive activities, we need to determine which social interactions contribute positively to productivity. We can consider the interactions between any two individuals, recognizing that these interactions come with costs, such as time taken away from work and logistical expenses needed to facilitate the interaction. At the same time, these interactions can bring benefits to the company, such as fostering idea generation. To incorporate social interactions, we would ideally need to ensure that benefits outweigh costs. One approach to achieve this is using cost-benefit analysis (CBA) to evaluate both the costs and benefits of interactions and adjust the objective by penalizing it per interaction costs in the solution. We can formally define the CBA-based internalization as: for individuals $i, \ell \in \{1, 2, \ldots, N\}$, let $C_{i\ell}$ be the cost and $B_{i\ell}$ be the benefit of their interactions. The cost-benefit for employees i and ℓ is $\text{CB}_{i\ell}(x) = C_{i\ell} \cdot \mathbf{1}\left(\exists j \in \{1, 2, \ldots, k\} x_{ij} \neq x_{\ell j}\right) - B_{i\ell} \cdot \mathbf{1}\left(x_{ij} = x_{\ell j}\ \forall\, j \in \{1, 2, \ldots, k\}\right)$. This captures the interactions' benefits occurring in the same time slot and the cost of missing interactions when they are in different time slots. Using CBA, we internalize the externalities with $CB_{i\ell}$ by modifying the objective as,

$$\min \sum_{j=1}^{k} L_j(n_j) + \sum_{i\ell} \text{CB}_{i\ell}(x)$$

While CBA helps, it does not specify when internalization should occur during optimization. Thus, we use the layered architecture, which has a detailed view of connections and feedback in the process. A simple overview of the layered architecture for the water cooler example is in Table 1. In this stack, the supervisory layer internalizes the loss of collaboration by considering the total cost-benefit, $\sum_{i\ell} \text{CB}_{i\ell}(x)$.

Remark. The use of a supervisory layer to internalize externalities in this example is based on the fact that the externalities in the question stem from ignorance. To internalize them, we must be able to observe and track deviations from the desired outcomes (which were initially aimed at increasing productivity but shifted away from that goal due to fewer social interactions) and address the externalities caused by this ignorance leading to the deviation. The supervisory layer monitors deviations, making it ideal for incorporating ignorance-related externalities.

5.2 Using AI to Select Candidates for Hiring

With AI becoming a key tool in decision-making, its use in corporate recruitment has gained significant attention (Gordon 2023; Ahmed 2024; Reuters 2018; Chen 2023). AI recruitment relies on training data and parameters to decide which candidates move forward. This example considers a company using an AI tool based on historical data, with an assumption that the features derived from historical data represent diverse and best-fit candidates. We explore the unexpected optimization outcomes and how to address them.

Internal Stakeholders and Basic Optimization. Our first goal is to formally define the basic vanilla optimization for the hiring problem. Before defining the formulation itself, we need to identify the internal stakeholders and their goals. Referring to prior research on AI in hiring (Gordon 2023; Ahmed 2024; Reuters 2018; Chen 2023), we identify the following internal stakeholders (and goals): (1) **The organization, HR, and hiring managers** who aim to select the best candidates for interviews and hiring and to ensure fair representation in the candidate pool, eliminate discrimination, and avoid practices that may lead to litigation, and (2) **Teams** who seek candidates that best fit their team.

Our objective is to select the best set of candidates given an AI trained on prior data. To formally define the optimization formulation, we first define a scoring function $\text{Score}(x) = \sum_{i=1}^{n} w_i \cdot f_i(x)$, where $f_i(x)$ represents feature i of candidate x, and w_i represents weights derived from historical data. Let C represent the set of all candidates. The scoring function calculates how well a candidate fits the observations from the training data. The goal is to select the top k candidates with the highest scores for interviews:

$$\max_{X \subseteq C, |X|=k} \sum_{x \in X} \text{Score}(x).$$

External Stakeholders. The goal of the vanilla optimization is to select top k candidates that best fit the selection criteria based on past data. However, noting prior work on AI hiring (Gordon 2023; Ahmed 2024; Reuters 2018; Chen 2023), we identify the following external stakeholders (and goals): (1) **Candidates** who want to ensure fair and equal opportunities for selection, and (2) **Regulatory agencies** who want to ensure fairness and bias-free hiring practices.

Unintended Consequences. The unexpected drawbacks affecting external stakeholders are: 1. AI tools may introduce bias based on factors like gender, age, and disabilities (Reuters 2018; Ahmed 2024). 2. AI could select unqualified candidates over more qualified individuals (Ahmed 2024), resulting in perceived biases and potential litigation. 3. Consequently, a lack of diversity in selected candidates could hinder innovation in the organization.

Why Error? Here, the optimization assumed that past training data accurately represented diverse subgroups, apt candidates, and was free from biases. However, this assumption was flawed, as past data may contain subtle biases that AI tools can amplify with the training, leading to biased results. A notable example is Amazon's automated recruitment system, which was found to be biased against female applicants (Reuters 2018). Despite the company's claims of commitment to equal representations, erroneous data assumptions led to unexpected drawbacks in the optimization, aligning with the "error" category in subpar practices.

Internalizing Externalities: A Systems Theory Framework. Having identified the externalities arising from the vanilla optimization, the next step is to internalize them. In this case, the optimization produces biased solutions due to a flawed assumption. To address this, a method is needed to ensure the optimization avoids such biases. We propose the use of Cap-and-Trade systems. A Cap-and-Trade system enables the optimizer to introduce a threshold on the diversity required in the solution. This is motivated by rules such as the EEOC 4/5 rule which flags potential bias if a group's hiring rate is under 80% of the highest group's rate (Equal Employment Opportunity Commission 1978). In this use case, we use the Cap-and-Trade system to meet a specific threshold of diversity in hiring decisions to comply with the 4/5 rule. This "cap" ensures that hiring practices do not disproportionately disadvantage protected groups. We use a lower-bound cap and no trading. Assume we have a set of k groups with hiring rates $h_1(X), h_2(X), \ldots, h_k(X)$ for the solution X, and a minimum threshold hiring rate $h_{\text{threshold}}(X)$, e.g., $h_{\text{threshold}}(X) = 0.8 \max_i h_i(X)$. This introduces the constraint:

$$h_i(X) \geq h_{\text{threshold}}(X) \; \forall i \in \{1, 2, \ldots, k\}$$

To determine when internalization occurs, like before, we use the layered architecture. A simple overview of the layered architecture for the AI hiring example is provided in Table 2. In this use case, the regulatory layer internalizes diversity needs by incorporating Cap-and-Trade bounds.

Remark. The use of this layer to internalize error-related externalities is grounded in the understanding that errors arise from factors like flawed assumptions and biases. The regulatory layer is responsible for managing regulations and ensuring solutions comply with accepted social standards and legal boundaries. For instance, adherence to regulations like the EEOC 4/5 rule should be managed at this layer.

5.3 Aggressive Campaigns for Quarterly Sales

Aggressive marketing campaigns are commonly used to maximize short-term profits, despite potential long-term consequences (Phua 2014; Pauwels et al. 2004). This example explores a company using aggressive marketing to boost quarterly sales. We examine the unexpected outcomes of this scenario and how to incorporate them into the optimization.

Internal Stakeholders and Basic Optimization. We first need to identify the internal stakeholders and their goals to define the vanilla optimization. Referring to the literature on aggressive marketing campaigns (Phua 2014; Pauwels et al. 2004; Kessler 2024; Mela, Gupta, and Lehmann 1997; Mela, Jedidi, and Bowman 1998), we identify internal stakeholders as the **company leaders** who aim to increase quarterly profits. These stakeholders want to increase quarterly sales through aggressive marketing and discounting. Thus, the basic optimization focuses on maximizing short-term sales by allocating marketing budget and discounts efficiently, subject to operational and budget constraints. Formally, let x represent the decision variable that captures pricing, marketing cost, and target sales. The optimization problem is,

$$\max_x \sum_{i=1}^{n} p_i(x) \cdot s_i(x) \text{ subject to}$$

$$\left\{ \sum_{i=1}^{n} m_i(x) \leq B, s_i(x) \leq \text{Inventory}_i \; \forall i, p_i(x) \geq p_{\min} \; \forall i \right\}$$

where $m_i(x)$ is the marketing cost, $p_i(x)$ is the product price, $s_i(x)$ is the sales volume, Inventory_i is the inventory size for product i, $p_{\min}$ is the minimum price, and B is the marketing budget.

Layer	Optimization Function	Feedback Loops
Physical	The AI assigns a score $\text{Score}(x) = \sum_{i=1}^{n} w_i \cdot f_i(x)$ to each candidate, where $f_i(x)$ is the value of feature i for candidate x, and w_i is the weight for feature i derived from historical hiring data. Let C represent the set of all candidates. The objective is to select a subset X of size k such that: $\max_{X \subseteq C, \lvert X \rvert = k} \sum_{x \in X} \text{Score}(x)$.	The weights w_i are updated after analyzing the set of selected candidates.
Regulatory	The lack of diversity is mitigated. Here, we introduce Cap-and-Trade bounds, leading to optimization $\max_{X \subseteq C, \lvert X \rvert = k} \sum_{x \in X} \text{Score}(x)$ s.t. $h_i(X) \geq h_{\text{threshold}}(X)\ \forall i \in \{1, 2, \ldots, k\}$.	Feedback should be used to adjust the $h_{\text{threshold}}(X)$.
Supervisory	Monitor the diversity and optimality of the solution. Let $\text{Div}(X)$ be the diversity of solution X and $\text{Div}_{\text{target}}$ be the target diversity. The optimizer evaluates $\lVert \text{Div}(X) - \text{Div}_{\text{target}} \rVert$ to measure deviation.	Feedback entails modifying solution X based on the deviation from $\text{Div}_{\text{target}}$ according to organizational and regulatory requirements.
Strategic	Optimizes hiring for long-term goals like diversity (G_{Div}), creativity (G_{Create}), and retention (G_{Reten}): $G = \omega_1 \cdot G_{\text{Div}} + \omega_2 \cdot G_{\text{Create}} + \omega_3 \cdot G_{\text{Reten}}$, with $\omega_1, \omega_2, \omega_3$ representing the weights for each goal.	Evaluate the alignment between hiring outcomes and stakeholder goals by using and adjusting $\omega_1, \omega_2, \omega_3$ or objectives based on the deviations.

Table 2: Layered Architecture Layout for the AI Hiring Example

External Stakeholders. Using prior work on aggressive campaigns (Phua 2014; Pauwels et al. 2004; Kessler 2024; Mela, Gupta, and Lehmann 1997; Mela, Jedidi, and Bowman 1998), we note external stakeholders (and goals) as: (1) **Customers** who seek quality items at reasonable prices and desire long-term reliability from the company and its products, (2) **Competitors** who seek to maintain a competitive market and prevent market monopolization by the company.

Unintended Consequences. Maximizing short-term profits while ignoring long-term impacts may result in unexpected drawbacks for external stakeholders: 1. Aggressive marketing encourages discount-seeking and stockpiling behavior (Mela, Jedidi, and Bowman 1998), while reducing regular consumption in anticipation of future promotions (Kessler 2024). 2. It may obscure product issues and their societal impact, especially in sensitive industries like healthcare (Phua 2014). 3. Aggressive campaigns can also lead to market instability and unhealthy competition. Conversely, aggressive campaigns do not always have drawbacks. With quality products, well-planned marketing can foster customer loyalty (unexpected benefit) (Kessler 2024).

Why Immediacy of Interests? Aggressive campaigns to boost profits inherently prioritize short-term gains over long-term sustainability. These campaigns focus on immediate objectives, often neglecting long-term profitability and sustainability. This, by definition, makes them a clear example of subpar practices driven by the immediacy of interests.

Internalizing Externalities: A Systems Theory Framework. To internalize externalities from the immediacy of interests, we suggest a social welfare function $W(x)$. SWF is based on the utilities gained by internal and external parties and the disutility from externalities. These utilities and disutilities could be aggregated over time. Thus, SWF allows heeding the long-term effects on the company, consumers, and market, and reducing the impact of short-term interests.

Let $U_c(x) = \sum_{i=1}^{n} (p_i(x) \cdot s_i(x) - m_i(x))$ represent the company's utility, $U_u(x) = \sum_{i=1}^{n} (v_i(x) - p_i(x))$ (where $v_i(x)$ is the intrinsic value of the product i in regards to the consumer utility), and $U_D(x)$ be the disutility from externalities. The SWF is defined as $W(x) = U_c(x) + U_u(x) - U_D(x)$. Then the objective function becomes,

$$\max_x \sum_{i=1}^{n} p_i(x) \cdot s_i(x) + \lambda W(x)$$

Here, λ is a weight parameter. A simple overview of the layered architecture for aggressive marketing is listed in Table 3. In this example, the strategic layer internalizes externalities using SWF, which allows evaluation of short-term and long-term goals based on the loss of utility in SWF.

Remark. This layer is suited to internalize *immediacy of interest* externalities due to its long-term planning to balance short-term and long-term gains and ensuring short-term goals do not compromise a company's long-term success.

6 Discussions and Limitations

In sum, we examined how to better characterize unintended consequences in optimization, identify impacted stakeholders, and determine when and how to internalize externalities. We proposed integrating externalities with systems thinking to develop a comprehensive framework for navigating these challenges in complex, interconnected systems.

6.1 Normative Directions and Value Conflicts

Externalities can present value trade-offs. So we first discuss some points on research on values in AI that cover three aspects: embedding value analysis through *Value Sensitive Design (VSD)* (Friedman 1996; Friedman, Kahn, and Borning 2002; Friedman and Hendry 2019; Umbrello and van de Poel 2021; Sadek, Calvo, and Mougenot 2023; Hopkins 2023), *managing stakeholder tensions* in sociotechnical systems (Kallina and Singh 2024; Heymans and Heyman 2024; Katsikeas, Papachristopoulos, and Gounaris 2023), and addressing *trade-offs in value conflicts* like fairness versus efficiency (Victor and Bélisle-Pipon 2024; Fan et al. 2024; Research 2025). Together, these strands emphasize the need for designs integrating values, stakeholder outlooks, and normative trade-off reasoning into optimization. E.g., *how do we decide whether to trade a one-hour loss of road construction efficiency over the displacement of a household?*

Layer	Optimization Function	Feedback Loops
Physical	$\max_x \sum_{i=1}^{n} p_i(x) \cdot s_i(x)$ subjected to $s_i(x) \leq \text{Inventory}_i \ \forall i$ and $\sum_{i=1}^{n} m_i(x) \leq B$ where physical constraints, including inventory limits (Inventory_i) and a budget constraint (B), are incorporated to ensure feasible solutions while maximizing the objective.	Adjust inventory and budget allocations based on the solution.
Regulatory	$\max_x \sum_{i=1}^{n} p_i(x) \cdot s_i(x)$ s.t. $s_i(x) \leq \text{Inventory}_i \ \forall i$, $\sum_{i=1}^{n} m_i(x) \leq B$, and $p_i(x) \geq p_{\min} \ \forall i$ where regulatory constraints are introduced to prevent aggressive low pricing that could lead to price gouging.	Adjust the price thresholds by updating $p_{\min}$ as follows: $p_{\min} \leftarrow p_{\min} + \eta \sum_{i=1}^{n} \frac{\partial p_i(x)}{\partial x}$ where η is a learning rate.
Supervisory	Evaluate solutions by balancing profit and cost through $\max_x \sum_{i=1}^{n} \big(p_i(x) \cdot s_i(x) - m_i(x)\big)$, subjected to $s_i(x) \leq \text{Inventory}_i \ \forall i$, $\sum_{i=1}^{n} m_i(x) \leq B$, and $p_i(x) \geq p_{\min}; \forall i$.	Feedback uses sales data to assess optimization outcomes and adjust solutions to address deviations.
Strategic	Maximize the SWF: $W(x)$.	Feedback updates $U_D(x), U_c(x)$ and $U_u(x)$ using long-term observations.

Table 3: Layered Architecture Layout for the Aggressive Marketing Example

We recognize that such value conflicts demand normative judgment. Our framework, however, does not aim to resolve these tensions outright. Instead, it seeks to ensure they are acknowledged and methodically incorporated into the optimization process, rather than being overlooked. Also, our approach may not offer definitive resolutions, but externality quantification tools like Cost-Benefit Analysis help surface and partially (and structurally) navigate trade-offs. In detail:

The Normative Dimension is Central. Systems thinking may not resolve conflicts, but 1) it reframes optimization problems beyond efficiency maximization by broadening what is considered relevant in optimization. This reframing is inherently normative, 2) it normatively asks where and how externalities must be incorporated into the optimization process. Even if a choice is made to prioritize household displacement over efficiency, a key question remains: at which level of abstraction should this externality be addressed? Systems thinking navigates this normative decision.

Recent Normative Research. Thus, our work aligns directly with recent normative conversations. Recent key work on responsible AI, AI governance, and algorithmic fairness like Chan et al. (2023); Hutchinson et al. (2022); Donia (2022); Laufer, Gilbert, and Nissenbaum (2023); Dobbe, Krendl Gilbert, and Mintz (2021); Gyevnár and Kasirzadeh (2025) examine normative directions in algorithmic, ML, and optimization decision settings to limit harms. Our work contributes to these debates by structuring externalities, feedback loops, and value commitments in optimization.

Value Conflicts, Traditional Stakeholder Analysis, and Value-Sensitive Design (VSD). Our framework does not ignore trade-offs but explicitly makes them visible/actionable within a broader system. It extends VSD and stakeholder tension frameworks by adding systemic and procedural dimensions often missing in VSD and stakeholder analyses. It offers a formal lens on systemic impacts that stakeholder analysis often overlooks: Externalities are unintended, system-wide consequences of pursuing specific goals, often invisible in stakeholder frameworks when actors cannot foresee or articulate them. Unlike traditional stakeholder analysis focused on direct relationships, externalities with systems thinking emphasize second and third-order system effects, including long-term social and institutional spillovers. In AI/ML, focusing solely on stakeholder tensions captures direct harms but misses broader effects harming groups not initially framed as stakeholders.

Our work (and math) complement VSD by: 1) focusing on systemic/indirect externalities beyond stakeholder engagement, 2) embedding externality analysis within optimization itself, not as an afterthought. These points show another *extra* aspect of our contributions pertaining to these lines of work. Thus, since our proposition does not replace VSD but complements it, comparisons should be made with vanilla optimizations without any systematic externality considerations (not with VSD or trade-off analyses that work alongside our framework, not in competition with it).

No framework may perfectly resolve deep value conflicts. It matters, however, how well frameworks: 1) surface critical questions and tensions, 2) ensure recognition of externalities and indirect harms, and 3) prevent optimization with narrow goals. The six-step method does all of these. It: 1) makes externalities explicit (Steps 1–4), 2) facilitates strategic and regulatory engagement (Steps 4-6), and 3) expands the optimization scope to include systemic consequences.

6.2 Limitations

First, subpar practices can also stem from data collection and sampling (not discussed in our work), leading to unmet goals (Dörfler et al. 2024). Also, our framework helps identify causes such as error or ignorance, but comprehensively capturing all causes, stakeholders, and goals remains a challenge. Externalities provide structure but not a full methodology for this. Moreover, categorizing issues (ignorance, error, or short-termism) is context-dependent; e.g., a single scenario may be interpreted differently based on organizational priorities. Opting for parameters (like objective weights, short- vs long-term goals, or feedback terms) also depends on context. Moreover, abstraction of the optimization problem and the goals can obscure key sociotechnical factors (Selbst et al. 2019). Lastly, tools like CBA or SWF help quantify externalities, but they may sometimes serve better as guiding perspectives and require adaptations to fit specific socioeconomic contexts. These decisions should be informed by stakeholder goals and require expertise to align with individual, organizational, and societal resolutions.

References

Ahmed, S. 2024. Navigating the Pitfalls of AI in Hiring: Unveiling Algorithmic Bias.

Amodei, D.; et al. 2016. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565*.

Arrow, K. J. 1951. *Social Choice and Individual Values*. New York, NY, USA: Wiley: New York.

Arrow, K. J. 1969. The organization of economic activity: Issues pertinent to the choice of market versus non-market allocation. *The Analysis and Evaluation of Public Expenditures: The PBB-System*, 47–64.

Bejan, C. 2024. On the shareholders versus stakeholders debate. *Journal of Economic Behavior and Organization*, 218: 68–88.

Berk, R.; et al. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods and Research*.

Bertalanffy, L. v. 1968. *General System Theory: Foundations, Development, Applications*. George Braziller.

Bird, S.; Barocas, S.; Crawford, K.; and Wallach, H. 2016. Exploring or exploiting? Social and ethical implications of autonomous experimentation in AI. In *FAT Conference*.

Boardman, A. E.; Greenberg, D. H.; Vining, A. R.; and Weimer, D. L. 2018. *Cost-Benefit Analysis: Concepts and Practice*. Cambridge University Press, Cambridge: Cambridge University Press, 5 edition.

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT*, 177–186.

Brauers, W. K. 2013. *Optimization methods for a stakeholder society: a revolution in economic thinking by multi-objective optimization*, volume 73. Springer Science & Business Media.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability, and Transparency*.

Chan, A.; Salganik, R.; Markelius, A.; Pang, C.; Rajkumar, N.; Krasheninnikov, D.; Langosco, L.; He, Z.; Duan, Y.; Carroll, M.; et al. 2023. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 651–666.

Chen, Z. 2023. Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications*, 10(1): 567.

Coase, R. H. 1960. The problem of social cost. *Journal of Law and Economics*, 3: 1–44.

de Troya, I. n.; Kernahan, J.; Doorn, N.; Dignum, V.; and Dobbe, R. 2025. Misabstraction in Sociotechnical Systems. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 1829–1842. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.

Dobbe, R.; Krendl Gilbert, T.; and Mintz, Y. 2021. Hard choices in artificial intelligence. *Artificial Intelligence*, 300: 103555.

Donia, J. 2022. Normative Logics of Algorithmic Accountability. In *FAccT*, 598.

Dörfler, F.; He, Z.; Belgioioso, G.; Bolognani, S.; Lygeros, J.; and Muehlebach, M. 2024. Towards a systems theory of algorithms. *IEEE Control Systems Letters*.

Equal Employment Opportunity Commission. 1978. Uniform Guidelines on Employee Selection Procedures. Federal Register, 43 FR 38290. 29 CFR Part 1607.

Fan, J.; Nguyen, T.; Lin, Y.; and Amershi, S. 2024. Minion: Mediating Human-AI Value Conflicts Through User and Expert Feedback. *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Fayard, A.-L.; and Weeks, J. 2007. Photocopiers and Watercoolers: The Affordances of Informal Interaction. *Organization Studies*, 28(5): 605–634.

Fleurbaey, M.; Kanbur, R.; and Viney, B. 2021. Social externalities and economic analysis. *Social Research: An International Quarterly*, 88(1): 171–202.

Friedman, B. 1996. Value-sensitive design. *interactions*, 3(6): 16–23.

Friedman, B.; and Hendry, D. G. 2019. *Value sensitive design: Shaping technology with moral imagination*. Mit Press.

Friedman, B.; Kahn, P.; and Borning, A. 2002. Value sensitive design: Theory and methods. *University of Washington technical report*, 2(8): 1–8.

Gordon, C. 2023. AI Recruiting Tools Are Rich With Data Bias And CHROs Must Wake Up.

Gyevnár, B.; and Kasirzadeh, A. 2025. AI safety for everyone. *Nature Machine Intelligence*, 7(4): 531–542.

Hardt, M. 2014. How big data is unfair? *Medium*.

Heymans, F.; and Heyman, R. 2024. Identifying stakeholder motivations in normative AI governance: a systematic literature review for research guidance. *Data & Policy*, 6: e58.

Hopkins, P. 2023. Value sensitive design and AI: A reconsideration. *Mind the Product*.

Huffaker, C. 2016. There are fewer Pokémon Go locations in black neighborhoods, but why? *The Miami Herald*.

Hutchinson, B.; Rostamzadeh, N.; Greer, C.; Heller, K.; and Prabhakaran, V. 2022. Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 1859–1876.

Inc., E. A. S. 2024. What Is the Five Layer Automation Pyramid? — empoweredautomation.com. https://www.empoweredautomation.com/what-is-the-five-layer-automation-pyramid.

investopedia. 2024. Cap and Trade Basics: What It Is, How It Works, Pros & Cons — investopedia.com. https://www.investopedia.com/terms/c/cap-and-trade.asp.

Jensen, O. G. 1970. *Linear systems theory applied to a horizontally layered crust*. Ph.D. thesis, University of British Columbia.

Kallina, E.; and Singh, J. 2024. Stakeholder Involvement for Responsible AI Development: A Process Framework. *EAAMO '24: Equity and Access in Algorithms, Mechanisms, and Optimization*.

Kanbur, R. 2004. On Obnoxious Markets. *Globalization, Culture, and the Limits of the Market: Essays in Economics and Philosophy*, 39–61.

Katsikeas, C. S.; Papachristopoulos, D.; and Gounaris, S. 2023. Artificial Intelligence and Stakeholder Engagement in Innovation Management. *Journal of Business Research*, 160: 113778.

Kessler, L. 2024. Promotion effectiveness in MENA: How do promotions change shopper behavior in the long term?

Laufer, B.; Gilbert, T.; and Nissenbaum, H. 2023. Optimization's neglected normative commitments. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 50–63.

Lopez, S. 2018. On one of L.A.'s steepest streets, an app-driven frenzy of spinouts, confusion and crashes. *Los Angeles Times*.

Martin, D. 2003. Robert K. Merton, Versatile Sociologist and Father of the Focus Group, Dies at 92. *The New York Times*.

Matni, N.; Ames, A. D.; and Doyle, J. C. 2024. Towards a Theory of Control Architecture: A quantitative framework for layered multi-rate control. arXiv:2401.15185.

McAlpine, K. L. 2017. *Don't Abandon the Water Cooler Yet: Flexible Work Arrangements and the Unique Effect of Face-to-Face Informal Communication on Idea Generation and Innovation*. Ph.D. thesis, Cornell University.

McMillan, G. 2011. It's not you, it's it: Voice recognition doesn't recognize women. *TIME*.

Meadows, D. H. 2008. *Thinking in Systems: A Primer*. Chelsea Green Publishing.

Mela, C. F.; Gupta, S.; and Lehmann, D. R. 1997. The Long-Term Impact of Promotion and Advertising on Consumer Brand Choice. *Journal of Marketing Research*, 34(2): 248–261.

Mela, C. F.; Jedidi, K.; and Bowman, D. 1998. The Long-Term Impact of Promotions on Consumer Stockpiling Behavior. *Journal of Marketing Research*, 35(2): 250–262.

Merton, R. K. 1936. The unanticipated consequences of purposive social action. *American Sociological Review*, 1(6): 894–904.

Nokhiz, P. 2024. *Modeling and Simulation of Artificial Societies to Study Precarity and Inequity*. Ph.D. thesis, Brown University.

Nokhiz, P.; Ruwanpathirana, A. K.; Bhaskara, A.; and Venkatasubramanian, S. 2025a. Counting Hours, Counting Losses: The Toll of Unpredictable Work Schedules on Financial Security. *arXiv preprint arXiv:2504.07719*.

Nokhiz, P.; Ruwanpathirana, A. K.; Bhaskara, A.; and Venkatasubramanian, S. 2025b. Counting Hours, Counting Losses: The Toll of Unpredictable Work Schedules on Financial Security. *Transactions on Machine Learning Research*.

Nokhiz, P.; Ruwanpathirana, A. K.; Patwari, N.; and Venkatasubramanian, S. 2021. Precarity: Modeling the Long Term Effects of Compounded Decisions on Individual Instability. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, 199–208. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384735.

Nokhiz, P.; Ruwanpathirana, A. K.; Patwari, N.; and Venkatasubramanian, S. 2024. Agent-Based Simulation of Decision-Making Under Uncertainty to Study Financial Precarity. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 43–56. Springer.

Overdorf, R.; Kulynych, B.; Balsa, E.; Troncoso, C.; and Gürses, S. 2018. Questioning the assumptions behind fairness solutions. *arXiv preprint arXiv:1811.11293*.

Pauwels, K.; Silva-Risso, J.; Srinivasan, S.; and Hanssens, D. M. 2004. New Products, Sales Promotions, and Firm Value: The Case of the Automobile Industry. *Journal of Marketing*, 68(4): 142–156.

Phua, K. 2014. Harm to the Health of the Public Arising from Aggressive Marketing and Sales of Health-Related Products and Services. *Research in the sociology of health care*, 32: 199–212.

Pigou, A. C. 1920. *The Economics of Welfare*. Macmillan and Co.

Rathnam, S.; Parbhoo, S.; Swaroop, S.; Pan, W.; Murphy, S. A.; and Doshi-Velez, F. 2024. Rethinking Discount Regularization: New Interpretations, Unintended Consequences, and Solutions for Regularization in Reinforcement Learning. *Journal of Machine Learning Research*, 25(255): 1–48.

Reader, L.; Nokhiz, P.; Power, C.; Patwari, N.; Venkatasubramanian, S.; and Friedler, S. 2022. Models for understanding and quantifying feedback in societal systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1765–1775.

Research, M. 2025. Best Practices for Constrained Optimization in AI Systems. White Paper.

Reuters. 2018. Insight - Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/.

Rodger, J. A.; and Pendharkar, P. C. 2004. A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. *International Journal of Human-Computer Studies*.

Sadek, M.; Calvo, R. A.; and Mougenot, C. 2023. Designing value-sensitive AI: a critical review and recommendations for socio-technical design processes. *AI and Ethics*, 4: 949–967.

Sævarsson, S. Ö. 2022. *The post COVID water cooler effect: the meaning of interpersonal interaction of peers for the future of work*. Ph.D. thesis, Reykjavik University.

Sarjoughian, H. S.; Zeigler, B. P.; and Hall, S. B. 2001. A layered modeling and simulation architecture for agent-based system development. *Proceedings of the IEEE*, 89(2): 201–213.

Satz, D. 2010. *Why Some Things Should Not Be for Sale*. Oxford University Press.

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 59–68. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; and Virk, G. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 723–741. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Smith, N.; and Sage, A. 1973. An introduction to hierarchical systems theory. *Computers & Electrical Engineering*, 1(1): 55–71.

Sterman, J. D. 2000. *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Irwin/McGraw-Hill.

Stern, N. 2014. Ethics, Equity and the Economics of Climate Change: Science, Economics and Politics. *Economics and Philosophy*, 30(3): 397–501.

Stinar, F.; and Bosch, N. 2022. Algorithmic unfairness mitigation in student models: When fairer methods lead to unintended results. In *Proceedings of the 15th International Conference on Educational Data Mining*, 606.

Sugiyama, M.; Lawrence, N. D.; and Schwaighofer, A. 2017. *Dataset shift in machine learning*. The MIT Press.

Tassi, P. 2016. I am now a rural 'Pokémon GO' player and it's the worst. *Forbes*.

Umbrello, S.; and van de Poel, I. 2021. Mapping value sensitive design onto AI for social good principles. *AI and Ethics*, 1: 283–296.

Victor, G.; and Bélisle-Pipon, J.-C. 2024. Medical AI, Categories of Value Conflict, and Conflict Bypasses. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1482–1489.

Voros, J. 2005. A generalised "layered methodology" framework. *Foresight*, 7(2): 28–40.

Wang, L.; and Zhao, J. 2023. *Mathematical Optimization*, 87–119. Berkeley, CA: Apress. ISBN 978-1-4842-8853-5.

Weitzman, M. L. 1974. Prices vs. Quantities. *The Review of Economic Studies*, 41(4): 477–491.

Whittlestone, J.; Arulkumaran, K.; and Crosby, M. 2021. The societal implications of deep reinforcement learning. *Journal of Artificial Intelligence Research*, 70: 1003–1030.

Zaichenko, M. 2024. Internal and External Stakeholders in IT.

Zhu, Q.; Wei, D.; and Ji, K. 2016. Hierarchical architectures of resilient control systems: concepts, metrics, and design principles. In *Cyber Security for Industrial Control Systems*, 151–182. CRC Press.