A Perspective on Incentive Design: Challenges and Opportunities

Lillian J. Ratliff[1], Roy Dong[2], Shreyas Sekar[1], and Tanner Fiez[1]

[1]Department of Electrical Engineering, University of Washington, Seattle, WA, USA, 98115; email: ratliffl@uw.edu
[2]Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, IL, USA, 61801

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–37

https://doi.org/10.1146/((please add article doi))

Keywords

incentive design; control theory; economics; machine learning

Abstract

The increasingly tighter coupling between humans and system operations in domains ranging from intelligent infrastructure to e-commerce has lead to a new challenging class of problems founded on a well-established area of research: incentive design. There is a clear need for a new toolkit for designing mechanisms that aid in coordinating self-interested parties while avoiding unexpected outcomes in the face of information asymmetries, exogenous uncertainties from dynamic environments, and resource constraints. The purpose of this article is to provide a perspective on the current state of the art in incentive design from three core communities—economics, control theory, machine learning—and highlight interesting avenues for future research at the interface of these domains.

Contents

1. Introduction

In recent years, technological advancements have enabled cost-effective deployment of sensors and actuators at scale. This has, in turn, led to the promise of improved performance, efficiency, and reliability in almost all of today's modern systems. Moreover, enabled by such technologies, humans are able to make real-time decisions that dynamically impact the performance of these systems. Thus, as these new technologies reach further, the decisions, interactions, and motivations of human agents that increasingly influence the operations and dynamics of engineered systems need to be considered as an integral part of the design of such systems and their day-to-day operations.

The following now-commonplace examples are demonstrative not only of wide-spread sensor-actuator deployment but also issues that may arise when stakeholder motivations are not properly accounted for:

Smart Grid. Many energy-efficiency programs run by electric utility companies use data collected from households to forecast future energy demand, and some programs issue rewards for curtailing or deferring energy consumption at peak times. However, these *incentive programs* may inadvertently motivate users to use energy storage systems (e.g., batteries) in inefficient ways, and these behaviors are often not observable by the system operators. Furthermore, users can often receive monetary gains by strategically misrepresenting their usage patterns (e.g., baseline inflation) and preferences to the utility companies, and many of the incentive programs in deployment today are not robust to strategic data manipulation (see (1) and the references therein).

Mobility Markets. Disruptive ride-sharing companies rapidly gain market share by providing cheap and convenient rides to users on very short notice. They have been able to achieve this by using smart device applications to allocate portions of the transportation infrastructure that were previously underutilized. Additionally, these companies often issue incentives to both sides of the market. On the passenger side, they offer incentives to encourage increased adoption, more frequent use, and ahead-of-time announcement of travel plans to aid in better resource allocation. Similarly, on the driver side, they offer a variety of monetary incentives for a number of reasons including predictable supply, microscopic and macroscopic redistribution of supply, and more frequent use. However, these allocation algorithms need to account for the utilities and motivations, which are *private information*, of the users (i.e. drivers and passengers) to ensure proper operation. It has also been noted that ride-sharing platforms promote discriminatory behavior toward socio-economically disadvantaged groups (2). Even further, a malicious actor can manipulate the distribution of transportation resources throughout an area using dishonest requests; e.g., in (3), the authors analyze the effects of denial-of-service attacks on mobility-as-a-service systems, and show that supply can be arbitrarily depleted using spoofed ride requests.

Crowdsourcing. Due to recent advancements, machine learning algorithms require increasingly large datasets. Deep learning is a prominent example; given a sufficiently large and representative dataset, deep learning can achieve very low test error without any prior knowledge of the problem space. However, this requires large amounts of data and, to achieve datasets of sufficient scale, much of the data collection is crowdsourced. These crowdsourcing mechanisms do not always incentivize accurate data collection: data sources may not feel motivated to exert sufficient effort to collect quality data, and, further, some malicious data sources may intentionally poison data to induce poor results in the algorithms. Recent research has analyzed the impact of incorrectly aligning incentives of the data sources (4, 5, 6, 7), as well as the sensitivity of many modern algorithms to perturbations in a small fraction of the dataset (see (8) and the references therein).

A common thread throughout all of these examples is that human agents have a significant impact on the output of systems with which they interact. For instance, in traditional infrastructure systems humans were *passive participants*, consuming resources with no real impact on exchange of goods and services. Yet, now in intelligent infrastructure systems, such as the smart grid or intelligent transportation system, humans are *active participants*, having the ability—through intelligent augmentation or through now commonplace Cyber-Physical System (CPS)/Internet of Things (IoT) technologies—to make decisions in real-time that influence market and system operations.

The design of such human-in-the-loop systems requires a careful analysis of the objectives and incentives of the relevant agents not only to promote efficiency but also to avoid unintended consequences. While on the surface this appears to be a long-standing, and perhaps obvious, problem space, there are new challenges due to the tight coupling between humans, system operations, and market exchanges, the multi-time scale nature of decisions and interactions, and the increasing level of automation that has lead to complex, mixed autonomy environments in which mission critical tasks must be executed. Furthermore, new technologies and their supporting market structures are being realized, having been translated from prototypes to production while bypassing the development of robust mechanisms to certify their performance and guarantee avoidance of unexpected outcomes. An example in point is the push for and testing of autonomous vehicles; many companies are attempting to advance the frontier in the autonomous vehicle space and there are numerous examples of partial and full autonomous vehicles on the road despite the lack of guarantees, even probabilistic, for the algorithms and automation they employ.

Returning to the examples above, we note that they each illustrate how a misalignment of incentives can lead to inefficiencies and even cause unexpected or undesirable results. Thus, these new technology-enabled markets and application domains drive the need for an understanding of how to design mechanisms that:

- account for the behavior of human agents, such as competition between users and adversarial decision making;
- maintain desirable economic properties (e.g., incentive compatibility, individual rationality, a balanced budget, and social welfare maximization);
- are able to operate in dynamic, non-stationary environments, which include both physical dynamics as well as coupling in various input distributions;
- are based on limited prior knowledge, yet have performance guarantees; and,
- have explainable and interpretable models that support generalization and policy/regulation design.

We believe there is a gap between the theoretical and computational tools in the state-of the-art and those needed to not only analyze these systems but also to design interventions for shaping them. However, focusing on the problem of incentive design—the design of mechanisms for shaping the behavior of autonomous agents—in these systems, there is a large body of work which we can draw on to build the requisite toolkit.

1.1. Overview of the Current State-of-the-Art

Historically, this problem has been of interest in, but not limited to, three primary communities: economics, control theory, and machine learning. There are promising developments in each of the fields, yet taken alone, they are not sufficient. With this observation, the goal of this article is to provide a perspective on challenges for incentive design in human-in-the-loop systems and to motivate the development of a new set of tools for addressing them by highlighting existing approaches, pitfalls and all, that have traditionally been siloed in the fields of economics, control, and machine learning, respectively, and expose the reader to open problems at their interface. We believe that with the realization of new market structures for resource consumption and production in previously stagnated infrastructure systems along with the increasing availability of data and computational resources, now is the time for a merging of these fields in a deeper, more meaningful way than previously explored.

Incentive design has long been studied within the economics community, and approaches from this domain largely focus on designing incentives in static environments with significant *a priori* information and are very heavily model-based. For instance, prior information typically includes a distribution across preference types of users, or an assumption that the utilities of users belong to a relatively specific class of functions such as monotonic, concave functions. While the model-based approach allows for interpretation and often, generalization, scalability remains a challenge. Moreover, these approaches have lead to the development of *economically motivated* constraints such as incentive compatibility and individual rationality—the former ensures *truthful reporting* and the latter, *voluntary participation*. These approaches usually have very interpretable models which makes them useful for policy or regulatory design.

Similarly, the control theory community has developed a number of approaches to the design of incentives which address some of the desiderata listed above. A notable aspect of these approaches is that they are often capable of accounting for dynamics. Yet, they often fail to consider the economically motivated constraints mentioned above. Moreover, by and large, these approaches presuppose a lot of prior knowledge and structure: the dynamics are often either known or given in a parameterized form, it is commonly assumed that distributions on exogenous uncertainties are known *a priori*, and the system designer typically has access to reliable information that cannot be manipulated by other agents. The latter, in particular, allows the designer to sidestep issues of moral hazard (i.e. lack of visibility into the actions of agents) and adverse selection (i.e. lack of visibility into preferences of agents), which often arise in practical applications. These approaches are generally very model-based, and as such, also benefit from being highly interpretable.

Third, the machine learning community has studied similar problems using online learning methods. These approaches can operate with no prior knowledge, and provide algorithms that are often completely model-agnostic. Despite their optimality when very little underlying structure is assumed, the results and theoretical performance guarantees, which come in the form of regret bounds or worst-case competitive ratios, are often very conservative. Indeed, as an example in point, in many of the applications of interest, systems are interacting with human users, and humans are not completely adversarial in general nor are they completely random (i.e. stochastic). Hence, when either a stochastic or adversarial environment is assumed, as in many machine learning approaches, the theoretically prescribed number of samples required to determine optimal actions are too many to achieve satisfactory performance in practice and are not identifying the true underlying model. Moreover, the approaches tend to assume statistically independent and identically distributed (i.i.d.) observations and stationary environments, both of which are far removed from reality.

More generally, each of these domains has individually developed techniques for addressing the incentive design problem by making assumptions structured to allow for the tools of their field to apply. Yet, in many practical settings these assumptions fail to hold, and this is increasingly the case in human-in-the-loop systems and emerging markets by which we are motivated. None-the-less, we believe that a marriage of these different approaches may lead to new advancements in the theory of incentive design, leading to practically relevant analysis tools and certifiable algorithms.

1.2. Organization

The rest of the article is organized as follows. In Section 2, we provide a high-level description of incentive design problems introduced with a small amount of mathematical formalism as needed. The purpose of this section is to give the reader a formal sense of what an incentive design problem is and what the features of an incentive design problem are.

In Section 3, we provide an overview of the existing work that treats the incentive design problem in the economics, control theory, and machine learning communities. We describe at a high-level the foundation of the incentive design problem, and concepts salient to the approaches taken in engineering and computer science, as it is formulated within the economics community in Section 3.1. Building on this, we introduce and overview techniques applied by the engineering and computer science communities in Sections 3.2 and 3.3, focusing on control theory and machine learning, respectively. Specifically, we shed light on the problems each of the communities has addressed and point out how they complement one another in an attempt to motive new work at the intersection of these domains. Throughout the section, we introduce examples, using the three highlighted examples introduced earlier in this section, in order to facilitate describing different features of the incentive design problem handled by each domain. This section also exposes parts of the incentive design problem not treated by existing techniques in each of the three domains, while also foreshadowing that a combination of approaches from the three domains may lead to advancements in the state-of-the-art.

Such an overview then leads naturally into in Section 4 in which we provide a discussion that illuminates open problems and challenges for which we believe developing tools at the intersection of these domains may lead to solutions. We discuss our perspective on how these approaches can be reconciled to address the new problems of incentive design with desirable economic properties in dynamic settings with limited information. Finally, in Section 5, we make concluding remarks.

2. A Formal Introduction to Incentive Design

We restrict our commentary to a special class of incentive design problems that has a rich history in three core domains: economics, control theory, and machine learning. Specifically, we focus our attention on so-called *principal-agent* problems (9): a class of incentive design problems in which there are two types of participants—i.e. the *principal* and the *agent*. Before diving into the review of incentive design as it has been studied in these three domains, we provide a brief overview of the mathematical formalism used in the remaining sections in support of conveying ideas relevant to the concepts introduced therein.

We use the notation $J_P : U \times V \to \mathbb{R}$ for the principal's utility and $J_A : U \times V \to \mathbb{R}$ for the agent's utility where U and V are the action spaces of the agent and principal, respectively. We note that there may be more than one principal and more than one agent.

To provide an example, consider the mobility market example described in the introduction. It could be abstracted in such a way that the ride-sharing platform is the principal, and there may be many competing platforms and hence, multiple principals. A platform's users (i.e. passengers and drivers) are agents. The ride-sharing platform wants to maximize revenue, say J_P, which is a function of how users interact with the platform. That is, passengers decide when and how often to solicit a ride and drivers decide when and how often to work for the platform by accepting fares. All such possible actions form the set U. One way to maximize revenue via increasing user participation is to offer incentives to the two user groups. On the driver side, e.g., such incentives might be correspondences γ that return a value $v \in V$ for a weekly bonus as a function of the number of fares, say u, accepted during the week. The platform must decide the structure of γ. It does so by noting that given $\gamma : U \to V$, users each have a utility $J_A(u, \gamma(u))$ that associates a value to possible actions U which determines their level of participation. The platform then aims to design γ so as to induce a particular behavior on the part of the agents—that is, encourage each of them through the incentive γ to choose an action u that leads to the platform's utility J_P being maximized. In essence, the platform gets to influence the behavior of the users through γ.

2.1. Formalism

As is illustrated in this example, the agent's and principal's utilities are coupled since they are both functions of pairs $(u, v) \in U \times V$ and, thus there is a game between the principal and agent. However, there is a specific order of play. That is, the principal announces a mapping $\gamma : U \to V$ of the agent's action space into the principal's action space, after which an agent selects its action in response to the announced mechanism. Formally, γ is the incentive mapping and, as noted, it is the goal of the principal to design γ to induce behaviors that lead to their utility J_P being maximized.

To formalize the *incentive design problem* that the principal faces, there are often restrictions on the structure of γ. For example, consider a demand response scenario in which the principal is an electric utility company and the agent is an energy consumer. Due to regulatory mandates, it may be very likely that the structure of incentives that the electric utility company can offer is pre-specified or the value capped. We use the notation $\Gamma = \{\gamma : U \to V\}$ for the admissible set of such mappings from which the principal can choose. Following the example, the mappings in Γ may have a particular structure—e.g., Γ may be defined to be the set of continuous linear maps with a specified upper and lower bound and, as noted, it may be practically motivated such as a tariff structure imposed by regulation.

The order of events is as follows: the principal designs γ knowing the agent has utility J_A. Then, it announces γ, after which the agent responds by selecting $u \in \arg\max J_A(u, \gamma(u))$. In particular, supposing the agent is a rational decision-maker, given an announced $\gamma \in \Gamma$, the agent aims to select an action that maximizes their utility—i.e. $u^*(\gamma) \in \arg\max_{u \in U} J_A(u, \gamma(u))$ where we denote the dependence of u^* on γ. In this setting, if the principal is also a rational, utility maximizing decision-maker, then their goal is to choose $\gamma \in \Gamma$ such that the agent chooses an action that leads to the maximization of the principal's utility—i.e. the principal seeks to find γ such that $\gamma(u^d) = v^d$ and $u^d = u^*$ where $(u^d, v^d) \in \arg\max J_P(u, v)$. This is to say that the principal wants to *incentivize* the agent to play according to what is 'best' for the principal. In this way, γ realigns the preferences of the agent with those of the principal.

While there is a misalignment of objectives between the principal and the agent, if there exists a γ such that $\gamma(u^d) = v^d$ and u^d is a maximizer of $J_A(u, \gamma(u))$, then both the principal and agent are doing what is in their best interest: the agent is 'compensated' via γ to play u^d and $\gamma(u^d) = v^d$ ensuring the principal's utility is maximized.

2.2. Challenges

Finding such a mapping is not as simple as it may seem since, in practice, there are *information asymmetries* between the principal and the agent. That is, in reality the principal and the agent make their decisions based on some *information set* that is available to them. For instance, returning to the ride-sharing example, the platform may not precisely know the drivers' or passengers' utilities J_A. It is fairly intuitive that how individuals value different features that would impact their utility such as time-money tradeoffs would not be publicly known. In fact, making things even more challenging, the users themselves may be unaware of the precise representation of J_A and may be learning their valuation/preferences for services over time. Analogously, platform users do not have clear insight into the motivations of the platform. The information that is available to the platform and users alike plays a role in how they make decisions. How such challenges are treated by the economics, engineering, and computer science approaches to incentive design will be formalized in Section 3, and we specifically note in that section and Section 4, that a number of interesting and practically relevant questions remain open.

In particular, how this information set is conceived and mathematically modeled is a large part of what distinguishes the different approaches taken in the three domains of economics, control, and machine learning. In the treatment of information asymmetries, different communities start by making some assumptions on the abstraction of the partial information—e.g., encoded in a prior distribution or revealed over time through sampling—which then inform the approach that is taken. There are many forms which partial information can take depending on what is observable by the principal and the agent and *when* it is revealed to them. The treatment of these *informational asymmetries* is varied from field-to-field. Yet, as we elude to in Section 4, there are ample research opportunities in combining them in an effort to derive theoretically sound and practically meaningful solutions to the class of incentive design problems for human-in-the-loop systems.

Beyond information asymmetries, other features may arise making the problem formulation closer to reality while at the same time making solutions more elusive. For instance, the principal and the agent may also face constraints due to the physical system or environment in which they operate, the market structure which constrains their economic exchanges, or even due to other economic considerations—e.g., ensuring voluntary participation (i.e. agents do not opt for alternative services) or truthfulness (i.e. agents respond in accordance with their true preferences), concepts we formalize in Section 3.1. It also may be the case that the incentive design problem occurs repeatedly in time or is in fact dynamic, where the actions are time dependent and the state of the environment evolves in time. Again, how these features are formalised and treated often depends on the domain application and the community. In the next section, we overview such approaches with the goal of highlighting both benefits and detriments and suggesting that a merger of domains may lead to new and interesting solution approaches.

3. A Review of Approaches to Incentive Design

In the following sections on each of the core areas (economics, control, machine learning), we will introduce such features as they arise in the treatment of the incentive design problem. We describe at a high-level the problems each of the communities has addressed and point out how they complement one another in an attempt to motive new work at the intersection of these domains. The works in these fields are too numerous to cover all of them in this short perspective[1], and hence, our approach is to highlight fundamental contributions from these domains that apply to the types of systems—e.g., human-in-the-loop systems spanning intelligent infrastructure to e-commerce—that we are interested in.

Specifically, from economics, we focus our review on the classical treatment of information asymmetries. We note that the approach from economics, being the first community to formalize the incentive design problem, lays out the conceptual building blocks on which the other approaches are founded. Hence, the subsection on economics is written in such a way to provide the reader with a cursory introduction to these concepts. The subsections that follow refer back to these concepts.

From control, we focus on dynamics and the introduction of auxiliary *state* variables that encode information about the evolution of the environment as it depends on agent choices. From machine learning, we focus on adaptation and online learning. In economics and control, models are key and they shape the flavor of a large portion of the results, while in machine learning, the approaches are largely model-agnostic enabling scalability.

[1] In each of the subsections, we point the interested reader to relevant texts that summarize or cover large portions of the work.

Bringing these domains closer together by leveraging their successes is a great opportunity for future research as we highlight in Section 4.

In each of the sections, we provide at least one running example, accompanied by several smaller examples, to guide the reader through the material. These examples align with the three examples introduced in Section 1.

3.1. Economics

The class of problems outlined in Section 2 was first studied by economists as a mathematical formalisim for understanding and designing contracts between differently invested parties each potentially possessing some *private information*. This *asymmetric information* between the two entities is really the crux of these principal-agent problems. As we will see, the strategic decision-making of agents in these classical settings can cause certain efficient and desirable outcomes to be unattainable.

We remark that there is a significant body of work from the economics community on the issue of asymmetric information and on the class of incentive design problems we described at a high-level in the preceding section, so much so that it is impossible to review it all. We point the reader to a handful of useful textbooks (9, 10), including one from the control perspective (11), for that purpose.

In this section, we review the specific approaches, assumptions, and flavor of results for the conceptualization of two core information asymmetry representations, adverse selection and moral hazard, and their their treatment via screening and monitoring, respectively. The purpose of selecting the particular approaches we discuss is that they complement approaches taken in control theory and machine learning, which we discuss in the sections that follow, and we believe the particular approaches give insight into the challenging problems that remain open (see Section 4) in the development of a broader systems theory for human-in-the-loop systems at scale.

To facilitate the introduction of core concepts, let us begin with an illustrative example. Numerous applications in which economics techniques have been applied to the design of incentives can be found in the wide body of literature. One engineering application where there has been significant cross-over of economics approaches is in the energy systems area.

Example 1 (Demand Response.) *In demand response programs, an energy utility company issues incentives to energy consumers to change their energy consumption patterns. In this setting, the energy utility company is the principal and the energy consumer the agent. The action $u \in U$ chosen by the agent is the energy consumption and the incentive program—designed by the utility company to reward the consumer for timely curtailment—is denoted by $\gamma \in \Gamma$.*

In this case, if a user's energy consumption profile is u, then the utility company gives incentive $\gamma(u) = v \in V$ to the user—that is, v is the realized reward for the behavior u. This may come in the form of cash-back rewards, raffled prizes, or discounted energy rates. The value of this incentive to the consumer is captured in the energy consumer's utility, $J_A(u, \gamma(u))$, which models their satisfaction with the energy consumption patterns associated with u, and the trade-off when the offered incentive is $\gamma(u)$. Put another way, under this model, when $J_A(u_1, \gamma(u_1)) = J_A(u_2, \gamma(u_2))$, the energy consumer is indifferent between receiving incentive $v_1 = \gamma(u_1)$ for energy consumption u_1 and receiving incentive $v_2 = \gamma(u_2)$ for energy consumption u_2.

Analogously, the utility company's utility, $J_P(u, v)$, models the operational costs of pro-

viding u to the energy consumer, as well as the cost of offering incentive $\gamma(u) = v$. Oftentimes these incentives γ are chosen to induce a consumption u with more energy-efficient behaviors, or to curtail or shift some energy demand from peak hours to off-peak hours.

In practice, there is information asymmetries that cause the design of demand response programs to be challenging. The first information asymmetry that arises is the principal's lack of knowledge of J_A. In this example, the utility company does not know the consumption preferences of the consumer. For example, does the consumer work from home, do they have a medical condition that requires the temperature of the house to be higher than normal, what energy-consuming devices does the consumer own, or are they particularly environmentally conscious and hence, open to extreme curtailment. All of these questions are examples of things the utility company does not know a priori. *Another information asymmetry that may arise, and is in fact common in many developing countries and parts of Europe, is the observation of u. The consumer may spoof their energy signal in an attempt to pay less.*

In many practical demand response programs, the utility company uses the historical energy consumption as a baseline, and then issues incentives during, e.g., peak times. The baseline is used to determine the value of the incentive, meaning users are paid based on how much they curtail relative to their baseline. In these situations, energy consumers can use their private knowledge of J_A to their advantage. For instance, an energy consumer may artificially inflate their baseline just prior to a demand response program event in order to receive larger payouts under the program. Examples of this behavior have been noted in practice (see (1), and references therein). Ideally, the utility company would like to design incentive-based demand response prorgrams that are robust to strategic manipulation.

As illustrated in the above example, *market failures*—such as the incentive to artifically inflate a baseline—due to information asymmetries can broadly fall into two categories: *adverse selection* and *moral hazard*. Referring to the example, the utility company's lack of knowledge of J_A is leads to the former while the lack of precise knowledge of consumption u due to the agent having the ability to lie leads to the latter. Generally speaking, adverse selection arises when the *preferences* of the agent are not known to the principal—i.e. the principal does not have full knowledge of J_A. On the other hand, if J_A is known, but the principal is unable to observe the action $u \in U$ chosen by the agent, then *moral hazard* arises. These two issues, and the information asymmetry scenarios under which they arise, are key in categorizing inefficiencies that arise in problems of incentive design and the approach that is taken. Hence, we dedicate in the remainder of this subsection to formalizing each of them and then conclude with a short description of limitations of a purely economic approach and desiderata for alternative approaches that build on the base economic formulation.

3.1.1. Adverse selection. As noted, adverse selection arises precisely in situations where the principal is unable to identify the preferences of the agent. For example, as we pointed out in the previous section, within the class of problems we consider this could be realized as the agent's utility being dependent on some parameter $\theta \in \Theta$ representing the agent's *type*—that is, $J_A(u, v; \theta)$ where we use the notation $J_A(\cdot, \cdot; \theta)$ to indicate that J_A is parameterized by θ. The agent's type θ can abstractly encode the agent's preferences or even their internal state, and θ is *private information*. Adverse selection arises when the type is unknown *a priori* to the principal.

One of the earliest and most famous papers on the topic is the 1970 paper 'The Market for Lemons' by George Akerlof (12) in which the economic consequences when a used car buyer cannot distinguish between a *good used car* and a *lemon* are considered. In particular, conditions are identified in which no used car sales will occur and the market will shut down. This market shutdown can occur even when there are good used cars that sellers are willing to sell to buyers at mutually beneficial prices. Adverse selection has been extensively studied since this seminal work (see, e.g., (9, 10, 13, 14, 15) and references therein).

Referring back to the demand response example, as in the 'Market for Lemons', a utility company may not be able to distinguish between energy conscientious users, frugal customers, traditional users, and potential uninformed users when designing the demand response program and issue incentives under that program. Furthermore, these users might have something to gain by misrepresenting their type. In this case, it is entirely possible for demand-response programs to be inefficient, just as the used-car market can unravel.

When decision-relevant information is held privately by an agent, the uninformed principal may be able to elicit credible revelation of this private information by designing an appropriate *screening mechanism* which is *incentive-compatible*—i.e. under the mechanism, an agent achieves the best outcome by acting according to their true preferences. The idea for the design of a screening mechanism is that the principal proposes a *menu of contracts* containing variations of the instrument, from which the agent is expected to select the one that aligns with its preferences. That is, the principal designs a correspondence that relates to each possible agent type an action-value pair with an action u that the agent should take and a payout v it will receive.

While the principal does not know the type θ, in the design of such a menu, it is typically assumed that the principal has *a priori* information in the form of a prior distribution ρ over the type space Θ which encodes their beliefs about the agent's type. Besides the assumption of *a priori* information in the form of a distribution, it is also very typical to assume that the agent's utility J_A is concave in its actions and monotonically increasing in the preferences. These characteristics capture the diminishing marginal utility property and ensure that the problem is computationally tractable—in many cases, such assumptions lead to simple analytical solutions that are easily interpretable.

The menu of contracts is designed by the principal so as to maximize their expected utility given the prior distribution ρ. For example, when $\Theta = \{\theta_i\}_{i=1}^m$ is a finite set[2], the principal attempts to design an assignment of actions $u \in U$ (e.g., amount of energy a consumer curtails) and $v \in V$ (e.g., reward for curtailment) to type θ via γ—i.e. $\gamma(u(\theta)) = v(\theta)$—so as to maximize $\sum_{i=1}^m \rho(\theta_i) J_P(u(\theta_i), v(\theta_i))$. These assignments are referred to as *contracts* and the fact that there is one contract for each of the types θ_i is why the term *menu of contracts* is used.

This optimization problem is subjected to two fundamental constraints, *incentive compatibility* and *individual rationality*, which we have casually mentioned in the introduction and more formally define here. Incentive compatibility constraints ensure that the agent selects the contract that corresponds to his true type—i.e. if the agent's true type is $\bar{\theta} \in \Theta$, then his expected utility is highest for the contract $\gamma(u(\bar{\theta})) = v(\bar{\theta})$. When $\Theta = \{\theta_i\}_{i=1}^m$ is a finite set, incentive compatibility constraints are given by

$$J_A(u(\theta_i), v(\theta_i); \theta_i) \geq J_A(u(\theta_j), v(\theta_j); \theta_i), \quad \forall\, i, j \in \{1, \ldots, m\}. \qquad 1.$$

That is, for an agent of type θ_i, the contract $(u(\theta_i), v(\theta_i))$ should be preferable to any other contract $(u(\theta_j), v(\theta_j))$.

[2] The type space does not need to be finite dimensional and the treatment of the more general case, with has the same essential formulation and features, can be found in textbooks such as (10).

Individual rationality—also referred to as *voluntary participation*—constraints ensure that the agent participates. That is, relative to an outside option, say $\bar{J}_A$, the expected utility under the contract designed for each agent type is greater than the $\bar{J}_A$. Again, when $\Theta = \{\theta_i\}_{i=1}^m$ is a finite set, the individual rationality constraints takes the form

$$J_A(u(\theta_i), v(\theta_i); \theta_i) \geq \bar{J}_A, \quad \forall\, i \in \{1, \ldots, m\}. \qquad 2.$$

In the previously discussed demand response example, the menu of contracts would represent different available plans for a demand-response program. Individual rationality ensures that energy consumers choose to participate in the incentive program. Incentive compatibility ensures that energy consumers select the contract that is designed for their type—e.g., if the consumer is an 'energy conscientious' user, then the contract designed for such users is preferable to them. In other words, energy consumers are best off when they choose the option designed for them, and deviation only increases their cost.

One of the challenges with the incentive compatibility constraints are their combinatorial nature: supposing that Θ has m elements, there are $m(m-1)$ constraints. Issues with scalability arise frequently in these settings, and much of the work in this area is focused on identifying assumptions which can effectively reduce the number of constraints. As noted above, concavity and monotonicity assumptions on J_A and its derivatives aid in the reduction of constraints. For example, the Spence-Mirrlees single-crossing condition (16) is one such assumption. In the case where $\Theta = \{\theta_i\}_{i=1}^m$ is a finite set, the Spence-Mirrlees condition states that $J_A(\cdot, v; \theta_{i+1}) - J_A(\cdot, v; \theta_i)$ is monotonically increasing for every fixed v and every $i \in \{1, \ldots, m-1\}$. Intuitively, this means that the marginal utility of consumption is increasing with respect to the type. Under this assumption, the number of constraints is reduced from $m(m-1)$ to merely $2(m-1)$ constraints. More generally, a common thread in the treatment of the principal-agent problem with adverse selection is to identify broad conditions that allow the system designer to pinpoint conditions on the agents type under which it would select one contract over another.

On the other hand, in cases when the principal is unable or unwilling to create a screening mechanism, it may be at least partially in the agent's best interest to credibly *signal* their private information to the principal. Signaling mechanisms are also commonly studied in the context of adverse selection (10). In this case, it is assumed that the agent has available to them a set of signaling mechanisms from which it selects a signal according to its preferences. The goal of the principal is to again design an incentive mapping γ that elicits truthful reporting and participation. For example, in a demand response setting, environmentally conscientious users will likely gain much more satisfaction from buying an eco-friendly thermostat than a traditional user. In an economic sense, buying an eco-friendly thermostat costs the environmentally conscientious user less than it costs a user of another type. Furthermore, the utility company can use this information as a signal of the energy preferences of the consumer. If a utility company wishes to only recruit environmentally conscientious users for an incentive program (e.g., if they expect this user group to be more responsive and thus more lucrative to engage with), they can require an eco-friendly thermostat. Then, they must design their rewards so that the rewards are positive for environmentally conscientious users, but participation is not worthwhile for other users in consideration of the cost of the eco-friendly thermostat.

3.1.2. Moral hazard.

When the agent's actions are hidden from the principal, then this form of information asymmetry gives rise to the so-called problem of *moral hazard*. The term 'moral hazard' originated from the study and design of insurance contracts. For example, people are likely to take more risky actions once they have insurance coverage and, due to the coverage, do not bear the full burden of the risk. Common solutions to the problem of moral hazard include the introduction of mechanisms for monitoring the agent's actions (17) or sharing compensation with the agent (18).

Formally, moral hazard arises when the principal is unable to observe u, the agent's actions. In the formulation of solutions to this type of information asymmetry, it is typically assumed that the principal is able to observe some event $s \in \Sigma$ where Σ is the space of *observable events*. The event s is a random variable which is a function of the agent's action u and some random, unknown state of nature z. In particular, the principal observes $s(u, z) \in \Sigma$, and does not observe u—that is, the only knowledge the principal has of u is through the observation $s(u, z)$. The principal's goal is to design a mapping $\gamma : \Sigma \to V$ such that the agent is induced to select the desired action which is desirable from the point of view of the principal.

Consider the demand-response setting described in Example 1 in which the utility company (principal) wishes to motivate the energy consumer (agent) to reduce their energy consumption. Recall that v represents the reward given to the energy consumer for curtailing their consumption by u under the incentive mapping γ. In this setting, we can model the baseline consumption without any curtailment as z. The utility company does not know how much energy the consumer would have used in the absence of any incentives, but rather only observes the realized energy consumption, say $s(u, z)$, which depends on the baseline level of consumption z and the amount of curtailment u. If not properly incentivized, a user may try to falsely claim that even though the realized energy consumption, $s(u, z)$ is high, that several factors caused their baseline energy consumption z to be extremely high, and that, in fact, they curtailed a lot of energy consumption—i.e. u is high.

The order of events are as follows. First, the principal offers a contract $\gamma : \Sigma \to V$ which commits to an action $v = \gamma(s)$ for each observable signal s. The agent either accepts or rejects the contract. If the agent rejects the contract, their payoff is the value of their outside option, $\bar{J}_A$. Alternatively, if the agent accepts, then they choose an action $u^* \in \arg\max_u \mathbb{E}_z[J_A(u, \gamma(s(u, z)))]$ and nature subsequently draws the random variable z determining $\gamma(s(u, z))$. Then, the principal observes $s(u^*, z)$ and the agent's realized utility is $J_A(u^*, \gamma(s(u^*, z)))$.

The principal designs $\gamma(s)$ to maximizes $\mathbb{E}_z[J_P(u, \gamma(s(u, z)))]$ and does so by formulating an optimization problem in (u, γ) given the objective $\mathbb{E}_z[J_P(u, \gamma(s(u, z)))]$. As in the adverse selection problem, this optimization problem for the design of γ is subjected to two key constraints. First, there is the *individual rationality* constraint which is given by $\mathbb{E}_z[J_A(u, \gamma(s(u, z)))] \geq \bar{J}_A$ and, as we noted, ensures the agent does not opt-out. Second, there is the incentive compatibility constraint which is given by $u \in \arg\max_{u'} \mathbb{E}_z[J_A(u', \gamma(s(u', z)))]$ and ensures the agent chooses an action in accordance with its true preferences given the prior the principal has on the environment—i.e. a prior distribution over z. Note that the key issue is that the contract γ cannot depend on u; as a consequence, rather than perfect risk sharing, there is an analysis of the *incentives-insurance trade-off*. Similar to the case of adverse selection, we find that assumptions are often motivated by the scale and intractability of the original problem; for example, the *first-order approach* makes strong assumptions to replace the incentive compatibility constraint with its first-order optimality conditions. There is a rich literature on the analysis of moral hazard problems (see (9, 10) and the references therein) which seeks to solve this difficult problem, sometimes with further constraints.

3.1.3. Desiderata and Limitations. Fundamentally the models assume users are rational and have significant amount of prior information, even if faced with very stylized, but meaningful information asymmetries. They also make fairly restrictive assumptions such as concavity and monotonicity on the form of utilities—adopted because they capture diminishing returns properties while also remaining extremely computationally tractable—which would most certainly be violated if behavioral decision models—discussed in Section 4—such as prospect theoretic value functions or satisficing, both of which can introduce non-smoothness, are used in their place. The economic approach broadly allows for quite a bit of interpretation, explanation, and generalization due to the heavy model-based tools, yet at the same time this also causes them to not be scalable. Recent work has considered dynamic contracts which handle time-varying user preferences and environments (see (19, 20, 21, 22, 23, 24, 25, 26) and the references therein), but the assumptions are often too restrictive to be applied to the dynamics of an underlying state that corresponds to a physical system[3]. This may be in large part due to the motivating applications that are considered by economists such as labor or insurance markets that may not necessarily have these features.

3.2. Control Theory

The control approach to the incentive design problem builds on the economic foundation described in the previous subsection by offering an approach to handling the notion of an exogenous state variable which summarizes the environment as well as dynamics. In particular, the incentive design problem directly embodies the spirit and form of a control problem: the principle is the controller and the agent is the plant. It is even common in control to design the controller using some objective function—i.e. optimal control or policy. However, unlike the typical plant structure, the agent also chooses actions by optimizing some criteria—that is, the agent or plant is strategic itself. Models from control that capture this sort of behavior fall under the category of leader-follower decision problems or, synonymously, Stackelberg games (29). There is a quite a large body of literature drawing on classical control tools to solve Stackelberg games and hierarchical decision problems, sometimes even using the moniker ‘incentive design’ (see, e.g., (30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42)).

Before we dive into the details of these approaches, we note that as with the economic literature review, there is too large a body of work to review here[4]. Hence, we focus our review on elements arising in control that either complement the approaches from economics already mentioned or introduce new, and interesting model features that are relevant for human-in-the-loop systems. Specifically, we discuss how the control theoretic approach allows naturally for dynamics and enables the introduction of a *state* which encodes axillary environment information and itself may be dynamic.

[3] There are a few application domain specific works that do model physical dynamics; for example, in power economics, there is work on the design of pricing mechanisms, largely in the form of tariff structures or auctions, given some time-varying exogenous signal such as wind (27, 28).

[4] The series of papers (36, 37) by Olsder provides an overview of much of the work in this area up to 2009.

3.2.1. Example. In many of the example applications we mention in the introduction, there is some natural environment feature that can be treated as the state—e.g., demand response as described in Example 1, a natural abstraction of 'state' is the temperature of a consumer's home which evolves dynamically in time and impacts their energy consumption, and hence their utility. In the following example, we present two motivating abstractions of ride-sharing markets which not only highlight control theoretic models that allow for useful exogenous state characterizations, but also illustrate some open problems and challenges in incentive design problems in dynamic, uncertain contexts which we will touch on in Section 4. As noted in the introduction, in ride-sharing markets, platform providers provide incentives to both drivers and passengers. In this scenario, the platform serves as the principal and there are two types of agents: drivers and passengers.

Example 2 (Incentivizing Drivers in Ride-Sharing Markets.) *On one side of the coin, passengers can be modeled as forming queues at different nodes on a graph abstracting locations in the city. For instance, passengers willing to accept a ride arrive to nodes according to a Poisson process and once they accept they are* in the queue *associated with their arrival node—i.e. they wait to be matched to a driver and then wait for that driver to arrive. Once they are picked up, they are* in service*.*

In this model, the state x_t represents a vector of queue lengths at each node. These queues have their own dynamics, $x_t = f(x_t, u_t, v_t, t)$ which depend on external arrivals, an abstraction of the actions of the drivers u_t (i.e. their decisions of which node to be circling near and which fares they accept at a given time t), and an abstraction of the incentives offered to the drivers $v_t = \gamma(u_t)$ (e.g., higher prices for certain nodes or end-of-day incentives for visiting a node more than once). One goal of the platform might be to minimize average user wait time across nodes by incentivizing drivers to be near locations of high demand which change dynamically throughout the day. The drivers have their own utility function which depends on the information available to them—e.g., $\mathbb{E}_{x,u}[J_{A_i}(x_t, u_t, v_t)]$ where the expectation is taken with respect to driver i's beliefs about the state of the system and the strategies of the other drivers.

The challenge in designing incentives $v_t = \gamma(u_t)$ is that not only does the platform have uncertainty regarding the dynamics of the network of queues but also most certainly lacks knowledge of the drivers' utility functions. Moreover, drivers are strategic. For example, they may work for multiple platforms. Websites even exist that offer strategies for drivers to take advantage of bonus programs offered by ride-sharing platforms.

An analogous model can be constructed for incentivizing passengers in ride-share markets where drivers in the system form an exogenous state process.

Example 3 (Incentivizing Passengers in Ride-Share Markets.) *On the other side of the coin, the drivers can be modeled as forming queues at nodes in a graph representing different locations in a city. For instance, drivers in a particular neighborhood waiting for fares can be abstracted as a queue which gets served based on some priority rule set by the platform—e.g., a first-come-first-serve basis such as is the case at airports.*

In this framework, existing works have modeled passengers as one-off users of the platform which decide to participate based on the immediate price (43). Expanding on this model, passengers are in fact repeat customers that not only make choices about participation and usage based on the immediate price shown to them but also incentives offered to them over time—e.g., discounts for taking a ride with a particular platform at a particular location during an expected high demand event or for planning/scheduling a ride ahead of *time. In such a model, the platform again acts as the principal with cost $J_P(x_t, u_t, v_t, t)$ at time t where x_t is a vector of the driver queue lengths at each node (i.e. neighborhood) which has its own dynamics $x_{t+1} = f(x_t, u_t, v_t, t)$, u_t is a vector of choices by each users (e.g., a zero-one vector indicating if users accepted a ride or not in a location), and v_t is a vector containing both the price at different nodes for different passengers as well as the realized values of incentives under γ currently targeted at passengers taking actions u_t.*

The challenges here are similar; the platform faces uncertainties about the dynamics and does not directly observe the passengers' preferences. Moreover, passengers are strategic—e.g., they may have an incentive to price shop*, both by looking at other platforms' prices/offers as well as* juking *the system by searching for lower cost rides at nearby blocks.*

Of course both of these models are very abstract and in fact it might be the case that the platform tries to simultaneously match drivers and passengers in the system which are modeled both as strategic market participants, a model that invites many more interesting challenges which we discuss in Section 4. None the less these examples illustrate how the notion of state along with state dynamics can be used to abstract some exogenous process—e.g., queue length—impacting the decision of the principal who's efforts are focused on incentivizing a particular user group. Such exogenous environment information and its dynamics are captured in the modeling approaches taken by the control community.

3.2.2. Overview of Literature and Techniques. The control-theoretic approach in most cases is first to determine what the principal can achieve with respect to its objective and both choice variables (u, v). Then, to try and find a strategy γ that lets the principal reach this goal by inducing the agent to play a particular strategy. In repeated or dynamic settings, it can then be treated as a control tracking problem by formulating an auxiliary tracking cost. This philosophy is also core to many control problems: characterize what performance is at once desirable and achievable for a plant and then design a controller (sometimes optimal for a given objective) that induces the plant to meet this performance objective. On the other hand, if one does not have a sense of what the principal can achieve in terms of their utility, very little is known (36), although the machine learning community has developed techniques for designing algorithmic strategies for this problem in repeated or sequential settings with limited-to-no feedback from the agent or the environment as will be discussed in Section 3.3.

In the dynamic setting, the principal and agent both have time varying utilities and the underlying model of the environment dynamics is a differential/difference equation. For instance, as alluded to in Examples 2 and 3, in a discrete time setting[5], the agent's utility is modeled as $J_A(x_t, u_t, v_t)$ where $x_{t+1} = f(x_t, u_t, v_t, t)$ is the state dynamics and (u_t, v_t) are the decisions of the agent and principal, respectively at time t. The principal's utility is similarly formulated as $J_P(x_t, u_t, v_t)$. The principal and the agent both face problems of maximizing their utility over some horizon (e.g., it could be time-averaged or discounted and the horizon finite or infinite, both are treated in the literature).

There are two typical approaches to the leader-follower type problem: forward (alternatively, bilevel optimization) and reverse Stackelberg games. In the former, the principal

[5] There are analogous continuous time models, however, for the sake of brevity we do not detail them.

tries to optimize their utility subject to the constraint that the agent is selecting an optimal action at each time given v_t and x_t and the subject to the dynamics. There are many works in the control community addressing this type of problem, however the latter more directly captures the class of incentive design problems we consider and hence, we focus our review on existing approaches to it.

In a reverse Stackelberg game, the order of play is as follows. A principal (referred to as *leader* in this body of work) announces a mapping γ of the agent's (*follower's*) decision space into the principal's decision space. Then, the agent determines its response. In this case, the principal first determines a set of $\{(u_t^d, v_t^d)\}_t$ pairs that optimize its expected utility over the horizon. Then, they find a mapping $\gamma_t : U \to V$ that induces the agent to choose action u_t^d at each time t. For example, consider a T horizon problem in which the principal and the agent both seek to maximize their expected utilities $\sum_{t=0}^T J_P(x_t, u_t, v_t)$ and $\sum_{t=0}^T J_A(x_t, u_t, v_t)$, respectively, subject to the dynamics $x_{t+1} = f(x_t, u_t, v_t, t)$. Then, the principal selects $\{(u_t^d, v_t^d)\}_t \in \arg\max \sum_{t=0}^T J_P(x_t, u_t, v_t)$, after which they select a γ in the following set:

$$\begin{aligned}\mathcal{M}(T) = \Big\{\gamma \in \Gamma \;\Big|\; & \gamma(\{u_t^d\}_t) = \{v_t^d\}_t, \\ & \{u_t^d\}_t \in \arg\max \big\{ \textstyle\sum_{t=0}^T J_A(x_t, u_t, v_t) | \{v_t\}_t = \gamma(\{u_t\}_t), x_{t+1} = f(x_t, u_t, v_t)\big\}\Big\}. \quad \text{P-1.}\end{aligned}$$

One such mechanism might be, e.g., a sequence $\{\gamma_t\}_t$ such that $\gamma_t(u_t) = v_t$.

The reverse Stackelberg structure of play, as compared to the forward Stackelberg game, allows for the principal to design a mapping $\gamma : u \mapsto v$ as opposed to simply just the response v and hence, affords the principal more influence over the behavior of the agent. This revelation lead to defining the term *incentive controllability* (33), a loosely related concept to *incentive compatibility* in the sense that the objective is to characterize when it is possible to *control* the agent to a desired choice. This structure of play also allows for the introduction of multiple, non-cooperative agents where the principal's objective is to coordinate them around a set of choices which is 'best' from its point of view (35, 36, 37, 44, 45). In dynamic case, a significant number of works from the control community address the problems of incentive controllability and multiple agents within a very specific class of system dynamics and costs (i.e. linear quadratic) that are well-explored. For instance, assuming linear dynamics and quadratic costs, there have been several efforts focused on characterizing the solution (e.g., existence and uniqueness) to the reverse Stackelberg game and reducing the problem of finding it a convex optimization problem (31, 32, 33, 35, 37). Other efforts have relaxed the linear assumption on dynamics and similarly seek to characterize local equilibria (44).

The reverse Stackelberg structure is also amenable to situations of partial information where, e.g., the principal or the agents lack information about the state, others' actions, or even the utility functions of others. For instance, referring back to Example 2, the platform may not know drivers' preferences regarding which node they would like to finish working at or how long they intend to work. And, in either Example 2 or 3, they may also not know the arrival rates of drivers or passengers in the respective queue models, and hence, have partial information about the state dynamics.

Efforts have also been made to address the *partial information* case (see, e.g., (33, 34, 46)). With the exception of a few recent works[6], these approaches tend to not identify the

[6] As with the economics literature, from the control community there are application driven

lack of information as adverse selection and moral hazard despite the fact that the form of information asymmetry is the same and the approach that is taken in the event of partial information is often very different. In particular, given the dynamics, in the face of partial information, agents can form estimates and propagate priors using the observations they get over time. Some recent approaches have begun to develop learning algorithms leveraging techniques from adaptive control, game-theoretic learning, and reinforcement learning to design incentives in the face of partial information (48, 49, 50). These approaches take the view that the principal lacks some information about the decision-making process of the agents, impose a model structure on the aspect of the decision-making process they lack, and then try to make inferences about this model structure.

For instance, in a repeated, one-shot game scenario in the absence of an auxiliary state, (50) treats the case of adverse selection in which the principal does not know the agent's utility function $J_A(x, u, v; \theta_A)$, but knows that it belongs to some class of functions $\mathcal{F}(\theta)$. Specifically, the principal finds $(u^d, v^d) \in \arg\max J_P(u, v)$ and seeks to induce the agent to play u^d by repeatedly offering incentives to them. Since θ_A is unknown, instead of designing a menu of contracts with respect to a prior, the approach is to maintain and update an estimate of θ_A which is then used to adaptively design a sequence $\{\gamma_t\}_t$ with the goal of ensuring the agent's action asymptotically approaches the desired action—i.e. $u_t(\gamma_t) \to u^d$. Under the assumptions of zero-mean, finite-variance, i.i.d. noise and stable and persistently exciting dynamics—the latter of which is very difficult to verify—such results can be obtained. Relaxing the conditions, it is also possible to obtain asymptotic guarantees ensuring that $u_t \in B_\epsilon(u^d)$—i.e. an ϵ–neighborhood around the desired action.

In general, the typical control theoretic approach in the case of partial information is to assume a model structure, construct an estimator or inference method, and design γ based on its output. And, the typical analysis and results have the flavor of almost sure, asymptotic guarantees. In practice, this may be limiting as systems with many agents and non-stationary environments may not reach a *steady-state* very quickly or at all. Moreover, while the efforts from the control community form a very rich set of tools that address a number of the challenges which motivate this article including dynamics in the decision-making process, the inclusion of an auxiliary state, and partial information, the techniques are very heavily model-based, they assume very significant problem structure and the results, particularly in the partial information case, are often limited to very specific problem classes such as linear-quadratic problems with stabilizable, detectable dynamics and Gaussian noise. It is also the case that when there are uncertainties or partial information, the distributions are assumed known *a priori* thereby making the estimation problem much more tractable to solve, when in practice this information is rarely available.

3.3. Machine Learning

Approaches from the machine learning community tend to be less model-based than those in the control or even economics communities, and hence the results and techniques are complementary. Indeed, in recent years, there has been an increasing interest in studying adaptive incentive design problems through the lens of online learning. This line of inquiry looks at repeated principal-agent interactions where the principal faces some uncertainty

works—e.g., in the area of power systems/smart grid—that have been looking at contract design where there is adverse selection and moral hazard (see, e.g., (47)).

regarding the preferences or actions of the agent. The objective of the principal in the problem is to design a policy γ that determines the best action to play at each interaction with the agent using only the information that has been amassed prior to each respective interaction. The assumptions on the information available to the principal *a priori* and what is revealed to them over time informs the algorithm design; in fact, the mathematical formalism encoding what feedback is received by the decision-maker after an action is taken is a key attribute of how methods are devised in sequential decision-making problems more broadly.

As noted, in comparison with the approaches taken in the economics and control theory literature, the methods developed in online learning lean more towards model-agnostic than model-based. That being said, predominantly, the literature on online learning has focused on *direct optimization* problems that do not capture economically motivated constraints such as incentive compatibility, individual rationality, or preferences that evolve in time.

As a prelude to discussing how the online learning lens can be used for adaptive incentive design, we first describe the traditional framework under which such problems have been studied in the literature. The canonical online learning model considers a sequential game between a decision-maker and nature over a finite time horizon T. At each round t of the game, the decision-maker selects a move $v_t \in V$ and simultaneously nature takes an action $z_t \in \mathcal{Z}$, following which the decision-maker receives utility $J(z_t, v_t)$. The decision-maker seeks to maximize the utility at each round so that the cumulative regret over the horizon, defined as

$$R_T = \sup_{v\in V} \sum_{t=1}^{T} \mathbb{E}[J(z_t, v)] - \sum_{t=1}^{T} \mathbb{E}[J(z_t, v_t)], \tag{3.}$$

is minimized. Note that the per-round regret compares the action taken by the decision-maker with the best action that could have been taken in hindsight.

The literature and techniques developed for this problem can be broadly classified on the basis of the feedback observed by the principal after selecting an action. In traditional online learning, the underlying assumption is that the decision-maker is able to observe nature's move (z_t) and hence, the utility $J(z_t, v_t)$ for all $v_t \in V$, even those actions in V not selected by the decision-maker. In contrast, a parallel stream of literature has studied online learning in the presence of *bandit feedback*, where the decision-maker only observes the utility $J(z_t, v_t)$ for the action taken (v_t) and uses this information to shape future actions. The need for limited feedback can arise in many applications such as online ad placement where the decision-maker only observes whether or not the user clicked on an advertisement (i.e. $J(z_t, v_t) \in \{0, 1\}$), and not the user's underlying features (i.e. z_t). Finally, an alternative distinction in the literature stems from the source of the action z_t adopted by nature: in the stochastic model, z_t is drawn i.i.d. from a distribution, whereas, in the adversarial model, z_t is arbitrarily chosen.

Fortunately, there are well developed, near-optimal learning strategies in each of these environments. In the stochastic model, upper confidence bound (UCB) index policies (51) are ordinarily adopted, while in the adversarial model, multiplicative-weights-based policies (52, 53) are employed. The index policy stores a UCB index—i.e. the sum of the empirical mean of rewards experienced and the confidence width—on the empirical mean utility of each available action and plays the action with the maximum index. The crucial philosophy underlying these policies is that of balancing *exploration* and *exploitation*—i.e. continuing to learn about the utility of each action to minimize long-term regret while simultaneously focusing on the most promising actions to minimize short-term regret. On the other hand in multiplicative weights, a probability distribution over actions is maintained and updated using a multiplicative weights update rule based on the observed utility each time an action is taken. At each round, the action to play is sampled at random from this distribution. For more a comprehensive coverage of such online learning approaches, see (54, 55, 56).

Many of the applications mentioned in the introduction as well as in other domains such as *digital marketplaces*—e.g., crowdsourced systems, recommendation engines—are characterized by repeated principal-agent interactions where the principal must design a policy to induce strategic agents to coordinate around actions that ultimately maximize the principal's own utility and do so in the face of environmental uncertainties and informational asymmetries. The traditional model of online learning model does not directly capture principal-agent interactions where individual agents act based on their own self-interest. However, there are very promising attempts at extending it to take into account the agency available to individual agents (see, e.g., (57, 58, 59, 60, 61)). Indeed, the online learning model above can be extended to a multi-round principal-agent, in which the decision-maker is the principal, by allowing the principal's reward at each round to not only depend on their action and the realization of the state of nature, but also on the action u_t of an agent.

Formally, in the most basic formulation, a multi-round principal-agent problem can be described as follows. At round t, the principal selects an action $v_t \in V$, z_t is realized, and then the agent selects $u_t \in \arg\max_{u \in U} \mathbb{E}[J_A(z_t, u, v_t)]$. The principal then receives per-round utility $J_P(z_t, u_t, v_t)$. It is typically assumed that the principal is not aware of the agent's private type and utility function, and sometimes, not even the agent's selected action u_t[7]. Hence, one can imagine either issue, *adverse selection* and *moral hazard*, being addressed via online learning approaches that leverage only the information known *a priori* and the feedback that has been obtained. The goal of the principal is to find a policy γ (usually an algorithm) that minimizes a regret notion over a finite horizon by determining the best action from V at each round, given information up to that round. Given this setting, the goal of many works in this area, is to provide finite time bounds on regret.

3.3.1. Example.

As previously mentioned, problems arising within the realm of digital marketplaces are increasingly being modeled as repeated principal-agent interactions. A prominent example in recent years is crowdsourcing. In general, crowdsourcing is the practice of soliciting contributions in the form of services, content, etc. from willing participants of the online community. In the example that follows, we describe an application of crowdsourcing involving two self-interested parties that captures many of the salient aspects of the repeated principal-agent problem—strategic interactions, adverse selection, and moral hazard—and demonstrates how the problem can be solved using techniques from online learning.

Example 4 (Crowdsourcing.) *Crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) are designed to match available workers with tasks to complete. The functionality is simple: a requester posts a task as well as the amount they will pay for the completion of the task; a worker can then choose to do the work and upon completion is paid the specified amount. While the advent of these systems has provided an inexpensive and on-demand workforce that was once unavailable, it has been well documented that the quality of the crowdsourced work can be highly variable (62, 63, 64). Taking this into ac-*

[7] A notable exception is the work on contextual bandit approaches for online decision making when in addition to the per-round reward, the decision-maker gets some additional contextual information—e.g., observation of some auxiliary state or type information.

count, we model the task of incentivizing high quality contributions from workers using the framework of online principal-agent interactions.

We consider a principal (requester) who wants a set of tasks completed via a pool of agents (workers) crowdsourced through (say) MTurk with maximum quality at minimum cost. To incentivize high quality contributions, the principal seeks to design a policy for sequentially selecting a payment mechanism, consisting of a base payment and a quality-contingent bonus payment, to offer an agent for completing a task. The principal-agent interaction at each time $t \in T$ *is as follows: the principal selects a payment mechanism* $\gamma_t : Q \to V$ *from a finite set* Γ *that is a mapping from a quality to a payment for a task, an agent of type* z_t *is matched to the task, and this agent completes the task with effort level* $u_t \in U$*. The type of an agent is modeled to be drawn from a stochastic distribution at each time and may represent attributes such as skill, dedication, etc. Moreover, the amount of effort an agent expends is strategically chosen to maximize the expected value of the utility function given by the payment from the principal minus the cost of the effort exerted. In our notation, the utility function of the agent is* $J_A(z_t, u_t, v_t) = v_t - c_t$*, where* $v_t = \gamma_t(q_t)$ *is the realized payment from the mechanism,* $q_t = f(u_t, z_t)$ *represents the quality of the work, and* $c_t = g(u_t, z_t)$ *denotes the cost the agent incurs completing the work in terms of time spent, energy loss, etc.*

Following the principal-agent interaction, the principal observes (only) the quality of the work, pays out the realization of the payment mechanism, and can use the information obtained to adjust the policy for selecting payment mechanisms. The utility that the principal receives from an interaction with an agent is the value of the work minus the payment made to the agent. That is, $J_P(z_t, u_t, v_t) = r_t - v_t$*, where* $r_t = h(q_t)$ *denotes the value derived from the quality of the work and recall* $v_t = \gamma_t(q_t)$ *is the realized payment from the mechanism. The goal of the principal is to maximize the expected utility obtained from a policy over a finite horizon. Equivalently, the principal seeks to learn a policy that can minimize the cumulative regret defined as follows:*

$$R_T = \sup_{v \in V} \textstyle\sum_{t=1}^{T} \mathbb{E}[J_P(z_t, u_t, v)] - \sum_{t=1}^{T} \mathbb{E}[J_P(z_t, u_t, v_t)].$$

Each payment mechanism available to the principal has a stochastic distribution on the utility that the principal will obtain. This means for each $\gamma \in \Gamma$*, there exists a mean* μ *such that* $\mathbb{E}[J_P(z_t, u_t, v)] = \mu$*. This is since each agent selects actions to maximize the expected utility and the agent type is drawn i.i.d. from a stochastic distribution. Hence, the problem of learning a regret minimizing policy for incentivizing high quality contributions in crowdsourcing can be reduced to a stochastic multi-armed bandit problem. The UCB policy is a near-optimal strategy that could be applied to solve the problem. In short, the principal would select at each time the payment mechanism which had the maximum UCB on the empirical mean utility obtained from the payment mechanism being offered to agents in the past.*

The crowdsourcing example captures important aspects of online principal-agent decision problems including strategic behavior, adverse selection, and moral hazard. For instance, adverse selection and moral hazard can occur as the principal does not directly observe the type (utility function) or action (effort level) of the agent but only the final quality of the work. Under the assumptions on the users and the environment as presented, a near optimal strategy could be directly obtained using a well known multi-armed bandit algorithm.

The above example does not completely capture reality, however. It assumes that workers are one-off participants in the system—that is, they only interact with the system once. Moreover, worker types are assumed to be drawn i.i.d. from a stationary distribution, when in reality the agent would likely have memory and their responses in each round would depend on the previous actions of the principal. Several recent works, many in the crowdsourcing context, have started to address some issues related to the incentive design problem (5, 6, 59, 60) For example, the crowdsourcing problem presented above is closely related to the work of (5) on bandit algorithms for repeated principal-agent problems, but the authors of that paper extend the formulation presented so that the set of payment mechanisms being considered can be extremely large or even infinite while obtaining similar performance guarantees. However, there is still many open problems related to handling repeat users whose preferences—and hence, behavior—depend on the actions taken by the principal. The available tools from online learning will need to either be extended or integrated with existing approaches from the economics and control theory literature as we discuss further in Section 4.5.

3.3.2. Overview of Literature and Techniques. Going beyond the most basic formulation, as earlier noted, the principal may also face constraints on their budget (this could be a per-round budget or coupled over time) or desire that agents participate (individual rationality) and be *truthful* (incentive compatibility). There have been a handful of approaches that address one or more of these constraints for the principal-agent problem in the online learning setting (65, 66). In the crowdsourcing example discussed earlier, the set of feasible payment mechanisms that can be offered by the principal may be limited by the principal's initial monetary endowment—i.e. total budget. Incentive compatibility, for the same example, could refer to the notion that the agent's utility is maximized when their effort level is aligned with maximizing the quality of the work subject to costs incurred—e.g., see (6, 59).

A prototypical example of incentive-compatible online learning can be seen in the works pertaining to two-sided markets with sellers (principal) and buyers (agents). The literature in this domain can be divided into two distinct streams (67): (i) online posted pricing mechanisms where the principal seeks to learn an optimal set of prices for each good, and (ii) truthful online auction design (68, 69, 70) which could involve complex interactions between the entities (e.g., multi-round bids). We focus on the former as it falls under the broad umbrella of incentive design where the prices serve as incentives to guide buyer decisions. Early work in this area (71) concentrates on single-item markets with limited supply and developed dynamic pricing algorithms that extended traditional work in online learning to the pricing problem by discretizing the action space and proposing new index policies based on greedy selection. In follow-up works, many of these regret bounds were extended to settings involving fixed budgets (6) and multiple goods (72, 73). The latter works exploit the correlation across goods to limit the exploration phase, which could potentially be large owing to the exponential size description of agents' utility functions. Although these papers focus only on markets, their techniques have yielded new insights on online learning where the principal's actions are coupled across time, e.g., due to a finite budget.

While markets wield prices in order to influence the behavior of myopic agents, most digital platforms pursue alternative means to incentivize agents to explore unknown actions without sacrificing incentive-compatibility. In this regard, a line of research has focused on designing both monetary (74, 75, 76) and non-monetary incentives (61, 77, 78) in an online fashion to promote exploration. Particularly notable is the design of signaling strategies as in (61, 77) that offer information as an incentive to converge to welfare-maximizing outcomes. Although it is typical to consider asymmetric information structures in favor of the principal, a few works have looked at settings where the agent possesses an informational advantage (59, 79, 80). Here, the goal is to incentivize agents to reveal their private information or beliefs in a truthful manner. Broadly speaking, the incentives proposed in this line of work can be classified as *dynamic contracts*, that extend the techniques from Section 3.1 to an online environment. We refer the reader to (5) for a detailed exposition on the subject.

These works, however, tend to make the strong assumption that agent behavior is independent of time—that is, their preferences are static and not influenced by, e.g., the incentives offered by the principal. Moreover, it is assumed that the behavior of each individual agent is independent of the behavior of all other agents. It is worth mentioning several works that have considered dynamic agents or eschewed the independence assumption. Dynamic agents were considered in (81); a repeated principal-agent interaction was constructed to model the problem of a seller learning auction prices to maximize long-term revenue while a buyer strategically attempts to maximize their own long-term profit. An analogous principal-agent approach was taken in (82) with the extension to several agents, each of which when selected receives utility that it can strategically share with the principal in order to maximize their utility over the horizon, while the principal concurrently attempts to maximize utility received from the agents. Recently, a variant of the principal-agent problem was considered where the agent's preferences evolve in time according to a Markov chain and the principal's actions impact the evolution dynamics (57); this is one of the first attempts to address non-stationary environments in principal-agent interactions in that the same agent repeatedly interacts with the principal and the principal's actions influence that agents behavior so that from the point of view of the principal, the environment is non-stationary. In particular, there is a single stochastic process that evolves and hence, the actions are dependent on one another. This work was further extended to the combinatorial setting (58), where at each round the principal must match incentives to agents given budget constraints.

We remark that, at present, the online learning literature with dynamic agents or sources of dependence is relatively unfocused with many important open problems. Accordingly, there has been a limited amount of work considering incentive-compatibility when agent behavior is correlated with time or dependent on the actions of the principal. For example, in the crowdsourcing setting mentioned earlier, an agent who interacts with the principal for multiple rounds may seek to benefit by resorting to low effort levels if it can influence the payment mechanism offered in subsequent rounds. Clearly, there is potential to extend work in the online principal-agent domain to capture richer agent behavior and dynamics in future work.

One particular feature of the online learning literature that differentiates it from the adaptive control and learning techniques briefly mentioned in Section 3.2 is that most works—in particular, those providing solutions to a variant of the principal-agent problem—assume the action space of the principal is a finite set. And, very often these works create benchmarks based on the single best action in the set independent of time as in Equation 3. This is largely due to the fact that in the online learning literature, the view of incentive design that tends to be formed is a repeated interaction between the principal and the agent versus a dynamic or sequential interaction where the utilities are dependent on time (e.g., through some exogenous state variable or time dependent components of the utilities).

None-the-less, the techniques allow for the design of algorithms with performance guarantees for adaptively designing incentives given very little *a priori* information and feedback over time. This motivates, perhaps, a rapprochement between online learning techniques and those from adaptive control.

4. Open Questions and Research Opportunities

Having overviewed the various approaches to different formulations of the incentive design problem from the communities of economics, control, and machine learning, we now provide our perspective on a number of interesting open problems which have not been completely solved by any one of the individual communities, but through an interdisciplinary approach we may find solutions.

While there has been a lot of work addressing different formulations of and aspects of the incentive design problem, it is still an open problem to solve incentive design with repeatedly returning agents whose decision-making processes evolve with time and are functions of the principal's actions—thus, making the environment the principal interacts with non-stationary—and where the principal faces moral hazard and adverse selection type information asymmetries, is subjected to constraints (e.g., on their budget over time or due to a surrounding market structure), or is exposed to some external context (i.e. physical system dynamics or exogenous observations of the environment). The agents may also compete or have a more complex interaction structure amongst themselves. There may be more than one principal, adding an additional layer of complexity. These are all challenges in practical realizations of the incentive design problem which have yet to be sufficiently addressed. In the remainder, we discuss opportunities and additional challenges where we we believe potential solutions are on the horizon for research in this area.

4.1. Bounded Rationality and Risk-Sensitivity

A common thread across the disciplines cited previously is their supposition that the principal and the agents are rational entities that unambiguously favor strategies that maximize their expected utility. In reality, it is well understood that human decision making is bound by various cognitive limitations. Indeed, the rise of digital marketplaces has led to a renewed focus on the field of behavioral economics, pioneered by Nobel laureates such as Kahneman (and his collaborator Tversky) (83) and Thaler.

The interaction between human cognitive biases and incentives aimed at rational agents has led to the emergence of *perverse incentives* that achieve unintended, often adverse consequences. For example, in the domain of urban transportation, city officials who enforce *zone-based congestion pricing* in a bid to ease traffic may observe that these incentives often have only limited or even negative impact on overall congestion (84, 85). This occurs because the congestion pricing tariffs do not take into account the time-money trade-offs among users and secondly, drivers get acclimated to the increased prices (e.g., due to *anchoring bias* (86)). Further, such schemes may achieve the unintended effect of raising home prices inside the congestion zone as residents pay higher prices to avoid road taxes (87), e.g., due to *loss aversion* (86).

Many works, too many to cite, have sought to address these issues by introducing more realistic utility functions that capture several aspects of human behavior—this could include risk sensitivity, loss aversion, and reference point dependence, among other pertinent behavioral decision-making features. Such non-linear utilities are a core component of the famed *prospect theory* (86, 88). Alternatively, other decision-theoretic models such as *satisficing* (89) capture myopic behavior such as choosing the option that first meets an agent's minimal criteria. These works provide strong preliminary support. They tend to be rather simplistic and their empirical validation has largely been limited to static decision-making problems between two outcomes. There is still significant work to be done in extending and integrating these models (or at least the salient features that well-model human decision-making) in an incentive design framework, particularly in large-scale systems with many agents and dynamics.

With this in mind, a promising direction for future work involves leveraging recent advances in neural networks, deep learning, and classical results from inverse learning to infer (potentially) non-linear models of how humans respond to various incentives under a repeated interaction model (90, 91, 92, 93, 94, 95). A significant challenge is to develop techniques for *model agnostic*, scalable learning which results in explainable and interpretable outcomes. An alternative approach to tackling the problem of bounded rationality is in the design of *robust incentives* that achieve desirable outcomes irrespective of how agents behave. Although such approaches are preferable to model-specific incentives, they are, predictably, limited by their efficacy and they tend to result in very conservative strategies.

4.2. Information Design: Leveraging Uncertainty for Good

Uncertainty is an unavoidable aspect of not just physical systems but also digital systems involving human behavior. Almost all of the works on human decision making under uncertainty pertaining to incentive design consider uncertainty as an adverse phenomenon—indeed, it is intuitive to believe that suboptimal decisions are an obvious by-product of uncertainty. This raises a very natural question: *are there situations in which one can design incentives that perform better under uncertainty when compared to more deterministic environments?*

Surprisingly, in a number of settings, uncertainty does help in the design of more effective incentives; e.g., in transportation networks, the overall congestion can be decreased when a principal carefully calibrates the level of information available to each user (96, 97, 98). The intuition for this phenomenon comes from the fact that, in certain cases, incentives exist that close the gap between user selected equilibrium and social welfare maximizing equilibrium (e.g., the existence of optimal tolls in routing games is demonstrative); consequently, there must exist settings in which uncertainties cause users to behave more like the socially optimal solution. Indeed, in (96), the authors cast the classic *Braess paradox* (99)—which says that, under certain conditions, adding links to a network can increase the total congestion felt by users when they behave in a self-interested way—in light of informational uncertainties and highlight that in many networks, the average travel time could decrease when users are only aware of some routes (as opposed to having perfect information about all of the routes). More generally, in the face of uncertainty, a conservative user tends to over-estimate the delay on some paths—this could lead to 'less crowding' on popular routes and a balanced distribution of traffic (97). The surprising effects of uncertainty can also be seen in security allocation in airports (100), energy markets (**?**), and recommendation systems (61).

In light of these counterintuitive results, there are a number of important avenues including:

(Leverage Uncertainty in Incentive Design). The positive effects of uncertainty as observed in some scenarios motivates the development of a new theory of incentive design that deviates from the norm by explicitly leveraging uncertainty as a positive effect in decentralized systems.

(Information as an Incentive). Information or uncertainty can itself be thought of as a design feature thereby motivating the development of methods for using information as an incentive (61, 77, 101, 102) which enables a principal to control the level of uncertainty of the various agents to achieve a more desirable outcome.

(Co-Design of Incentives and Information). In many cases, what is achievable with incentives may not be achievable with information shaping and vice versa. This motivates deriving a theoretical and computational framework for the *co-design of incentives and information* that lead to a quantifiable improvement in performance while mitigating unintended consequences.

Central in each of these avenues is the design of information is some form. Yet, information design leads to the technically challenging question of whether information design can be achieved without unfair discrimination.

4.3. Fairness

In online learning, as with most of incentive design, it is typical to focus solely on algorithms that maximize social welfare over a finite horizon (e.g., in terms of regret). A notable exception involves the work on mean-variance optimization in online learning (103, 104). In systems comprising of multiple independent entities (principal, agents, etc.), it is well-known that maximizing the utilitarian welfare does not necessarily lead to egalitarian or equitable outcomes. These implications are exacerbated in multi-agent incentive design problems where a principal may offer vastly different incentives (or information) to different agents leading to contentions about unfair treatment by individual users or communities, e.g., dynamic pricing of parking and other public facilities can systematically disenfranchise populations in high-demand environments (105, 106).

Motivated by this, an impressive body of work in recent years has looked at online algorithms that learn the preferences of agents without sacrificing fairness—according to one or more metrics such as envy-freeness (107, 108), statistical parity (109), individual fairness (110), maximin fairness (111). A possibility raised by many of these works is that achieving fair outcomes may be intrinsically misaligned with maximizing social welfare. Despite these constraints, a number of promising research directions warrant exploration:

(Approximations and Trade-Offs). Given that achieving fairness may be incompatible with maximizing welfare, a reasonable compromise is to approximately maximizing efficiency while retaining fairness (112). Such an approach could then naturally segue into a thorough characterization of the efficiency-fairness Pareto frontier (113).

(Long-Term Fairness). While fairness may be harder to guarantee in a one-time interaction between a principal and agents, repeated interactions provide an opportunity for the designer to implement solutions that are equitable over a longer horizon (e.g., the average amount of perceived unfairness approaches zero over many interactions). An important open question is to identify algorithms that satisfy this property. Preliminary results support the hypothesis that long-term fairness may be easier to achieve without compromising social efficiency (114, 115).

(Model-Based Fairness). Almost all of the works mentioned above consider a typical design or optimization problem and add fairness as an external constraint. In many settings, it may be more natural to embed fairness directly into the model as in (116). e.g., a sequential game where self-interested agents maintain fidelity levels for various principals based on the perceived unfairness of the incentive received.

The ubiquity of incentives in society and the adverse socio-economic implications of *algorithmic discrimination* make it imperative that researchers include fairness in the design process and not simply as an after-thought. Fortunately, healthy discussions by a diverse range of academic communities and industry practitioners provide an encouraging sign that fairness-based constraints will play a key role in developing learning policies in the future (109, 117, 118, 119). Inherent in the quest for fairness in online learning is a trade-off with efficiency, which has been shown to be quite costly (113). In some problems with certain fairness criteria the steep loss in efficiency is unavoidable, it remains to be seen whether new learning approaches and fairness metrics can be developed to mitigate the cost of such a trade-off.

4.4. Interaction Between Markets: Cooperation to Competition

In the principal-agent problem, it is typical to consider settings where a single principal interacts with self-interested agents or multiple principals interact with different agents in isolation. Incentive design for such systems often relies crucially on the assumption that either there is no external option available to the agents, or that the external option does not interact or compete with the offers the principal is making. On other hand, in the case of digital marketplaces, it is more often the norm that agents have a choice between multiple principals particularly in repeated interaction settings; e.g., drivers and passengers selecting between different ride-sharing platforms or customers switching between ticket booking portals. It is customary to expect each principal to design independent incentives for its users to increase adoption. This begs the question: *how robust are current mechanisms to the presence of external competition and how does one redesign incentives to take into account competing principals or even platforms*?

On the one hand, there exists a considerable body of literature that explores competition in the field of market design, industrial organization, and game theory. For example, economists have long studied the problem of *competition versus innovation* (see (120) and references therein)—i.e. how does the level of competition in the market affect the type of incentives received by the agent? On the other hand, in repeated interaction settings featuring multiple principals, our understanding of how competing incentives and externalities affect agent behavior is rather limited.

An urgent need, therefore, is to gravitate towards a broader theory of incentive design via online learning that is cognizant of competition between providers—perhaps leveraging techniques from economics and control theory to model multi-agent interaction—without being too sensitive to the strategies adopted by other principals (121). At the same time, it is imperative to understand how current learning approaches perform as more participants enter the market (4). For example, preliminary results (122, 123) indicate that in the presence of competition, markets could 'get stuck' at a bad equilibrium where all of the principals play greedy strategies without performing sufficient exploration. A key research direction therefore is the design of *upstream* incentives that motivate principals to pursue policies that are aligned with the social good; e.g., in ride-sharing markets, a regulatory authority could impose upper caps on the price paid by consumers and lower caps on the revenue guaranteed to drivers.

A closely related issue that has raised concerns from anti-trust policy makers (124) and algorithm designers alike is that of *algorithmic collusion* (125)—scenarios where multiple algorithms representing independent principals (sometimes unintentionally) interact with each other to yield socially undesirable solutions. The problem is particularly acute in the field of automated pricing, where competing algorithms could engage in concurrent price increases resulting in poor social welfare. In light of these serious risks, it is critical that designers reexamine the classic approaches for developing incentives to identify which algorithms are more susceptible to collusive behavior (e.g., see (61)).

4.5. Integrating Model-Based Approaches into a Model-Agnostic Regime

The economic and control theoretic incentive design approaches that have been discussed are overwhelmingly model-based. Owing to this paradigm, several advantageous properties exist including strong performance guarantees and explainable outcomes. However, these techniques often do not scale well and may not be applicable in problems for which significant *a priori* information is unavailable.

On the contrary, online learning methods are predominantly model agnostic in that, from the point of view of the principal, very little is assumed about the agent. Moreover, for each of these cases, algorithms exist which are near-optimal under the limited assumptions. Yet, since correlation or structure is not being exploited, the near optimal guarantees may still be relatively weak or unattainable in large-scale environments. To give a concrete example, the standard upper confidence based and near-greedy algorithms in online learning (51) require the principal to take each possible action before any learning begins. In problems with many possible actions (e.g., selecting advertisements and item recommendation), it is clear that such an approach would be unrealistic.

To overcome such deficiencies, standard online learning techniques have been augmented with stronger assumptions and endowed with model-like structure, consequently improving sample efficiency and the ability to generalize. Despite the exciting progress in this area, by and large these methods have not extended to the online incentive design problem which has several further challenges including information asymmetries between the principal and the agent and non-stationarity in repeated interactions owing to agent behavior. In the rest of this section, we present models that have been imposed on the traditional online learning framework and consider how they may be promising in future work towards online incentive design.

A prominent example is that of online stochastic linear optimization with bandit feedback (126, 127), which models the cost of the principal to be a linear function of the actions taken with initially unknown parameters. Such an approach is advantageous as the decision maker can learn the cost of each action by solely learning the parameters of the linear function. Although this problem is more difficult to analyze technically, due to the loss of the standard independence assumption, the ability to leverage correlations between actions and the structure of the model makes this method interpretable and scalable (128). There is a related line of work examining how *a priori* knowledge of a similarity structure between actions can be leveraged in the online learning setting (129, 130, 131, 132). Certainly, considering the principal's cost in the incentive design problem to be a linear function of a selected agent would raise compelling questions. It would also be intriguing to investigate how knowing that groups of similar agents existed could be leveraged to speed up incentivizing agents.

As opposed to a purely optimization approach, probabilistic online learning methods which leverage priors on the distribution of costs, such as Thompson sampling (133, 134) and Gaussian process optimization (135, 136), have received increased attention in recent years and have shown to be empirically effective. These methods could be used for incentive design in several ways, including the principal maintaining distributions over parameters modeling agents' behavior, and agents updating priors on the principal's behavior for strategic purposes. Connecting back to Section 4.1, we note that maintaining layers of beliefs also allow for bounded rationality interpretations of the behavior exhibited by agents.

In practice, it is often the case that any of the aforementioned structures are combined with side information or context that is available to the principal when making decisions (128, 137, 138, 139). Relating back to the control perspective, one may relate context in an incentive problem to be some observation of the state of the environment. In this way, the principal can leverage the extra information to learn more fine-grained policies. Drawing on the ideas of context and information exchange from online learning is ripe for exploration in the incentive design problem.

In the online learning community, performance is often analyzed using the metric of *competitive ratio* (140, 141, 142), which gives the ratio of the online learning optimum to the offline full information optimum. Future work in incentive design may benefit from assimilating such analysis. Essentially, in the incentive design problem, a competitive ratio would inform the value of *a priori* information. If this were known, it may give insights into cases where acquiring information and applying model based methods may be preferable to model agnostic methods or vice versa.

As standard online learning frameworks are endowed with increasingly complex assumptions and structures, they begin to edge closer to and obtain the favorable aspects of the model-based methods in economics and control, while at the same time maintaining scalability and the ability to learn in a sample efficient manner. However, only a select few works (60, 143, 144, 145) have focused on applying these richer methods to the incentive design problem. Owing to this, there is significant opportunity to leverage the prosperous online learning literature to these problems.

4.6. Causal and Counterfactual Reasoning

In both physical as well as digital ecosystems, the rapid pace of evolution of the underlying environment necessitates that the principal constantly test new incentives aimed at better aligning the agents' objective with its own. Traditionally, firms have preferred to employ methodologies such as A/B testing (146) to evaluate how proposed treatments compare against existing incentives. However, in many cases such an approach may be infeasible, e.g., in online marketplaces, frequent A/B testing could adversely affect revenues or result in claims of unfair treatment (147). A powerful technique in the field of learning theory that allows the designer to circumvent these issues is that of *counterfactual reasoning*—using observations about a past treatment to infer about the effectives of an alternative intervention.

The field of counterfactual inference features a rich set of tools both in online and offline learning (147, 148, 149) to evaluate the performance of untested incentives and solve for optimal incentives, thereby allowing a designer to 'make the most out of limited data samples'. At the same time, almost all of this work has focused on static systems without economic constraints such as individual rationality or incentive compatibility. Therefore, the design of incentives for multi-agent systems with self-interested users whose behavior may evolve with time remains uncharted territory.

Extending classical theories of counterfactual learning to game-theoretic models is a non-trivial task due to the presence of confounding variables (e.g., see (149, 150)) and hidden dependencies. That is, unobserved system variables or externalities that correlate positively with one incentive may fail to do for another. For instance, digital incentives that are deployed via mobile applications may correlate with the age of the recipient and the results may fail to replicate for more traditional incentives. This calls for a more holistic approach towards counterfactual learning for designing incentives that take into account a *causal graph of relationships* between different variables that could potentially affect agents' response in direct and indirect ways (151). Understanding how traditional approaches in online learning via causal inference extend (151, 152, 153) to principal-agent or Stackelberg models remains an important open questions.

The multi-armed bandit approaches in online learning briefly discussed in Section 3.3 represent interesting solutions to incentive design via exploration-exploitation strategies for assessing the performance of a set of incentives when the principal has no *a priori* information and receives limited feedback. The classic multi-armed bandit and contextual bandit models can be expressed as special cases of the more general framework for causal inference (149, 153). A promising direction of future work is in drawing on more general casual learning techniques to develop algorithms for incentive design that exploit casual feedback to make inferences about the performance of incentives without having to explore all possibilities.

5. Closing Remarks

Motivated by applications in which there are technology-enabled, largely self-interested humans interacting and consuming resources in a constrained physical system, the purpose of this article is to provide a perspective on challenges and opportunities in the development of a toolkit for designing of incentives. We review work from economics, control theory, and machine learning which we believe to be building blocks for this new toolkit. Incentive design has long been studied in economic and control and is a more recent venture for machine learning. Each of these fields contributes a unique perspective on the design of incentives, and we try to articulate open questions and expose avenues for future research which bridge these domains by leveraging existing contributions to move the theoretical and computational frontier for incentive design forward.

LITERATURE CITED

1. Campaigne C, Balandat M, Ratliff LJ. 2016. Welfare effects of dynamic electricity pricing. Working paper, University of California, Berkeley
2. Ge Y, Knittel CR, MacKenzie D, Zoepf S. 2016. Racial and gender discrimination in transportation network companies. Working Paper 22776, National Bureau of Economic Research
3. Yuan C, Thai J, Bayen AM. 2016. Zubers against zlyfts apocalypse: An analysis framework for dos attacks on mobility-as-a-service systems. In *Proceedings of the ACM/IEEE 7th International Conference on Cyber-Physical Systems*. 1–10 pp.
4. Westenbroek T, Dong R, Ratliff LJ, Sastry SS. 2017. Statistical estimation in competitive

settings with strategic data sources. In *Proceedings of the 56th IEEE Conference on Decision and Control*. 4994–4999 pp.

5. Ho CJ, Slivkins A, Vaughan JW. 2016. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. *J. Artificial Intelligence Research* 55:317–359
6. Singla A, Krause A. 2013. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *Proceedings of the 22nd International World Wide Web Conference*. 1167–1178 pp.
7. Cai Y, Daskalakis C, Papadimitriou CH. 2015. Optimum statistical estimation with strategic data sources. In *Proceedings of The 28th Conference on Learning Theory*. 280–296 pp.
8. Goodfellow IJ, Shlens J, Szegedy C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*
9. Laffont JJ, Martimort D. 2002. *The Theory of Incentives: The Principal-Agent Model*. Princeton university press
10. Bolton P, Dewatripont M. 2005. *Contract theory*. MIT Press
11. Weber T. 2011. *Optimal Control Theory with Applications in Economics*. MIT Press
12. Akerlof GA. 1970. The market for ”lemons”: Quality uncertainty and the market mechanism. *The Quarterly J. Economics* 84:488–500
13. Mussa M, Rosen S. 1978. Monopoly and product quality. *J. Economic Theory* 18:301 – 317
14. Maskin E, Riley J. 1984. Monopoly with incomplete information. *The RAND J. Economics* 15:171–196
15. Rothschild M, Stiglitz J. 1976. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *The Quarterly J. Economics* 90:629–649
16. Spence A. 1974. *Market signaling: informational transfer in hiring and related screening processes*. Harvard economic studies. Harvard University Press
17. Alchian AA, Demsetz H. 1972. Production, information costs, and economic organization. *The American Economic Review* 62:777–795
18. Hlmstrom B. 1979. Moral hazard and observability. *The Bell Journal of Economics* 10:74–91
19. Dirk B, Juuso V. 2010. The dynamic pivot mechanism. *Econometrica* 78:771–789
20. Susan A, Ilya S. 2013. An efficient dynamic mechanism. *Econometrica* 81:2463–2485
21. Courty P, Hao L. 2000. Sequential screening. *The Review of Economic Studies* 67:697–717
22. Battaglini M. 2005. Long-term contracting with markovian consumers. *The American Economic Review* 95:637–658
23. Esö P, Szentes B. 2007. Optimal information disclosure in auctions and the handicap auction. *The Review of Economic Studies* 74:705–731
24. Board S. 2007. Selling options. *Journal of Economic Theory* 136:324 – 340
25. Kakade SM, Lobel I, Nazerzadeh H. 2013. Optimal dynamic mechanism design and the virtual-pivot mechanism. *Operations Research* 61:837–854
26. Alessandro P, Ilya S, Juuso T. 2014. Dynamic mechanism design: A myersonian approach. *Econometrica* 82:601–653
27. Borenstein S. 2005. The long-run efficiency of real-time electricity pricing. *The Energy Journal* :93–116
28. Jónsson T, Pinson P, Madsen H. 2010. On the market impact of wind energy forecasts. *Energy Economics* 32:313–320
29. Başar T, Olsder GJ, Clsder G, Basar T, Baser T, Olsder GJ. 1995. *Dynamic noncooperative game theory*, vol. 200. SIAM
30. Ho YCH, Luh PB, Olsder GJ. 1980. A control-theoretic view on incentives. In *Proceedings of the 19th IEEE Conference on Decision and Control*. 1160–1170 pp.
31. Groot N, Schutter BD, Hellendoorn H. 2012. Reverse stackelberg games, part i: Basic framework. *IEEE International Conference on Control Applications* :421–426
32. Zheng YP, Basar T, Cruz JB. 1984. Stackelberg strategies and incentives in multiperson de-

terministic decision problems. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-14:10–24

33. Ho YC, Luh P, Muralidharan R. 1981. Information structure, stackelberg games, and incentive controllability. *IEEE Transactions on Automatic Control* 26:454–460
34. Zheng YP, Başar T. 1982. Existence and derivation of optimal affine incentive schemes for stackelberg games with partial information: a geometric approach. *International J. Control* 35:997–1011
35. Ratliff L, Coogan S, Calderone D, Sastry S. 2012. Pricing in linear-quadratic dynamic games. In *Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing*. 1798–1805 pp.
36. Olsder GJ. 2009. Phenomena in inverse stackelberg games, part 1: Static problems. *J. Optimization Theory and Applications* 143
37. Olsder GJ. 2009. Phenomena in inverse stackelberg games, part 2: Dynamic problems. *J. Optimization Theory and Applications* 143
38. Liu X, Zhang S. 1992. Optimal incentive strategy for leader-follower games. *IEEE Transactions on Automatic Control* 37:1957–1961
39. Ho YC. 1983. On incentive problems. *Systems & Control Letters* 3:62–68
40. Cruz J. 1978. Leader-follower strategies for multilevel systems. *IEEE Transactions on Automatic Control* 23:244–255
41. Basar T, Selbuz H. 1979. Closed-loop stackelberg strategies with applications in optimal control or multilevel systems. *IEEE Transactions on Automatic Control* 24:166–179
42. Tolwinski B. 1981. Closed-loop stackelberg solution to a multistage linear-quadratic game. *J. Optimization Theory and Applications* 34:485–501
43. Banerjee S, Johari R, Riquelme C. 2015. Pricing in ride-sharing platforms: A queueing-theoretic approach. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, EC '15. New York, NY, USA: ACM, 639–639 pp.
44. Calderone D, Ratliff LJ, Sastry SS. 2014. Pricing for coordination in open–loop differential games. *Proceedings of the International Federation of Automatic Control* 47:9001 – 9006
45. Jing Yw, Zhang Sy. 1988. The solution to a kind of stackelberg game systems with multi-follower: Coordinative and incentive. In *Analysis and Optimization of Systems*, ed. A Bensoussan, JL Lions, pp. 593–602. Berlin, Heidelberg: Springer Berlin Heidelberg, 593–602 pp.
46. Zhang SY. 1987. A nonlinear incentive strategy for multi-stage stackelberg games with partial information. In *Proceedings of the 25th IEEE Conference on Decision and Control*. 1352–1357 pp.
47. Dobakhshari DG, Gupta V. 2016. A contract design approach for phantom demand response. *arXiv preprint arXiv:1611.09788*
48. Vamvoudakis KG, Lewis FL, Dixon WE. 2017. Openloop stackelberg learning solution for hierarchical control problems. *International J. Adaptive Control and Signal Processing* :1–15
49. Ratliff LJ. 2015. Incentivizing efficiency in societal-scale cyber-physical systems. Thesis, University of California, Berkeley
50. Ratliff LJ, Fiez T. 2018. Adaptive incentive design. *arXiv preprint arxiv:1806.05749*
51. Auer P, Cesa-Bianchi N, Fischer P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47:235–256
52. Arora S, Hazan E, Kale S. 2012. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing* 8:121–164
53. Auer P, Cesa-Bianchi N, Freund Y, Schapire RE. 2002. The nonstochastic multiarmed bandit problem. *J. Computing* 32:48–77
54. Bubeck S, Cesa-Bianchi N, et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5:1–122
55. Bubeck S. 2011. Introduction to online optimization. *Lecture Notes [Available Online: `http://sbubeck.com/BubeckLectureNotes.pdf`]* :1–86

56. Hazan E, et al. 2016. Introduction to online convex optimization. *Foundations and Trends® in Optimization* 2:157–325
57. Ratliff LJ, Sekar S, Zheng L, Fiez T. 2018. Incentives in the dark: Multi-armed bandits for evolving users with unknown type. *arXiv preprint arXiv:1803.04008*
58. Fiez T, Sekar S, Zheng L, Ratliff LJ. 2018. Combinatorial bandits for incentivizing agents with dynamic preferences
59. Jain S, Gujar S, Bhat S, Zoeter O, Narahari Y. 2018. A quality assuring, cost optimal multi-armed bandit mechanism for expert sourcing. *Artificial Intelligence* 254:44–63
60. Jain S, Narayanaswamy B, Narahari Y. 2014. A multiarmed bandit incentive mechanism for crowdsourcing demand response in smart grids. In *AAAI*. 721–727 pp.
61. Mansour Y, Slivkins A, Syrgkanis V, Wu ZS. 2016. Bayesian exploration: Incentivizing exploration in bayesian games. In *Proceedings of the 2016 ACM Conference on Economics and Computation*. 661 pp.
62. Wais P, Lingamneni S, Cook D, Fennell J, Goldenberg B, et al. 2010. Towards building a high-quality workforce with mechanical turk. *Proceedings of computational social science and the wisdom of crowds (NIPS)* :1–5
63. Huang JL, Liu M, Bowling NA. 2015. Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology* 100:828
64. Lovett M, Bajaba S, Lovett M, Simmering MJ. 2018. Data quality from crowdsourced surveys: A mixed method inquiry into perceptions of amazon's mechanical turk masters. *Applied Psychology* 67:339–366
65. Guha S, Munagala K. 2007. Approximation algorithms for budgeted learning problems. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM, 104–113 pp.
66. Babaioff M, Sharma Y, Slivkins A. 2014. Characterizing truthful multi-armed bandit mechanisms. *SIAM J. Comput.* 43:194–230
67. Einav L, Farronato C, Levin J, Sundaresan N. 2018. Auctions versus posted prices in online markets. *J. Political Economy* 126:178–215
68. Blum A, Hartline JD. 2005. Near-optimal online auctions. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*. 1156–1163 pp.
69. Hajiaghayi MT, Kleinberg RD, Parkes DC. 2004. Adaptive limited-supply online auctions. In *Proceedings 5th ACM Conference on Electronic Commerce*. 71–80 pp.
70. Cesa-Bianchi N, Gentile C, Mansour Y. 2015. Regret minimization for reserve prices in second-price auctions. *IEEE Transactions on Information Theory* 61:549–564
71. Babaioff M, Dughmi S, Kleinberg R, Slivkins A. 2015. Dynamic pricing with limited supply. *ACM Transactions on Economics and Computation* 3:4
72. Badanidiyuru A, Kleinberg R, Slivkins A. 2018. Bandits with knapsacks. *J. ACM* 65:13
73. Roth A, Ullman J, Wu ZS. 2016. Watch and learn: optimizing from revealed preferences feedback. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*. 949–962 pp.
74. Frazier PI, Kempe D, Kleinberg JM, Kleinberg R. 2014. Incentivizing exploration. In *Proceedings of the ACM Conference on Economics and Computation*. 5–22 pp.
75. Han L, Kempe D, Qiang R. 2015. Incentivizing exploration with heterogeneous value of money. In *Proceedings of the 11th International Conference on Web and Internet Economics*. 370–383 pp.
76. Singla A, Santoni M, Bartók G, Mukerji P, Meenen M, Krause A. 2015. Incentivizing users for balancing bike sharing systems. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 723–729 pp.
77. Papanastasiou Y, Bimpikis K, Savva N. 2017. Crowdsourcing exploration. *Management Science*
78. Liu Y, Ho C. 2018. Incentivizing high quality user contributions: New arm generation in bandit

learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*

79. Liu Y, Chen Y. 2016. A bandit framework for strategic regression. In *Advances in Neural Information Processing Systems*. 1813–1821 pp.
80. Roughgarden T, Schrijvers O. 2017. Online prediction with selfish experts. In *Advances in Neural Information Processing Systems*. 1300–1310 pp.
81. Amin K, Rostamizadeh A, Syed U. 2013. Learning prices for repeated auctions with strategic buyers. In *Advances in Neural Information Processing Systems*. 1169–1177 pp.
82. Braverman M, Mao J, Schneider J, Weinberg SM. 2017. Multi-armed bandit problems with strategic arms. *arXiv preprint arXiv:1706.09060*
83. Tversky A, Kahneman D. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185:1124–1131
84. Brown PN, Marden JR. 2017. Studies on robust social influence mechanisms: Incentives for efficient network routing in uncertain settings. *IEEE Control Systems* 37:98–115
85. Croci E. 2016. Urban road pricing: a comparative study on the experiences of london, stockholm and milan. *Transportation Research Procedia* 14:253–262
86. Kahneman D, Tversky A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47:263–291
87. Tang CK, et al. 2016. Traffic externalities and housing prices: evidence from the london congestion charge. *Spatial Economics Research Centre Discussion Paper* 205
88. Tversky A, Kahneman D. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertainty* 5:297–323
89. Simon HA. 1955. A behavioral model of rational choice. *The Quarterly Journal of Economics* 69:99–118
90. Cohen A, Einav L. 2007. Estimating risk preferences from deductible choice. *American Economic Review* 97:745–788
91. Gershman SJ, Horvitz EJ, Tenenbaum JB. 2015. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* 349:273–278
92. Ratliff LJ, Mazumdar E. 2017. Inverse risk-sensitive reinforcement learning. *arXiv:1703.09842v3*
93. Mazumdar E, Ratliff LJ, Fiez T, Sastry SS. 2017. Gradient–based inverse risk-sensitive reinforcement learning with applications. In *Proceedings of the 56th IEEE Conference on Decision and Control*. 5796–5801 pp.
94. Shen Y, Tobia MJ, Sommer T, Obermayer K. 2013. Risk-sensitive reinforcement learning. *arXiv preprint arxiv:1311.2097*
95. Majumdar A, Singh S, Mandlekar A, Provone M. 2017. Risk-sensitive inverse reinforcement learning via coherent risk models. In *Robotics: Science and Systems*
96. Acemoglu D, Makhdoumi A, Malekian A, Ozdaglar A. 2018. Informational braess' paradox: The effect of information on traffic congestion. *Operations Research*
97. Sekar S, Zheng L, Ratliff LJ, Zhang B. 2017. Uncertainty in multi-commodity routing networks: When does it help? *arXiv preprint arxiv:1709.08441)*
98. Wu M, Liu J, Amin S. 2017. Informational aspects in a class of Bayesian congestion games. In *American Control Conference*. IEEE, 3650–3657 pp.
99. Braess D, Nagurney A, Wakolbinger T. 2005. On a paradox of traffic planning. *Transportation science* 39:446–450
100. Lo C. 2012. Game theory: introducing randomness to airport security [available online: `http://www.airport-technology.com/features/featuregame-theory-airport-security-teamcore-stackelberg/`. *Airport Technology*
101. Kamenica E, Gentzkow M. 2011. Bayesian persuasion. *American Economic Review* 101:2590–2615
102. Bergemann D, Morris S. 2018. Information design: A unified perspective. *J. Economic Literature, Forthcoming*

103. Vakili S, Zhao Q. 2016. Risk-averse multi-armed bandit problems under mean-variance measure. *J. Selected Topics in Signal Processing* 10:1093–1111
104. Even-Dar E, Kearns MJ, Wortman J. 2006. Risk-sensitive online learning. In *Algorithmic Learning Theory*. 199–213 pp.
105. Haws KL, Bearden WO. 2006. Dynamic pricing and consumer fairness perceptions. *J. Consumer Research* 33:304–311
106. Irwin N. 2017. Why surge prices make us so mad: What springsteen, home depot and a nobel winner know [available online: `https://www.nytimes.com/2017/10/14/upshot/why-surge-prices-make-us-so-mad-what-springsteen-home-depot-and-a-nobel-winner-know.html`. *New York Times*
107. Varian HR. 1974. Equity, envy, and efficiency. *J. Economic Theory* 9:63 – 91
108. Berliant M, Thomson W, Dunz K. 1992. On the fair division of a heterogeneous commodity. *J. Mathematical Economics* 21:201 – 216
109. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 214–226 pp.
110. Joseph M, Kearns M, Morgenstern JH, Roth A. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*. 325–333 pp.
111. Nace D, Pioro M. 2008. Max-min fairness and its applications to routing and load-balancing in communication networks: a tutorial. *IEEE Communications Surveys Tutorials* 10:5–17
112. Gillen S, Jung C, Kearns M, Roth A. 2018. Online learning with an unknown fairness metric. *arXiv preprint arXiv:1802.06936*
113. Bertsimas D, Farias VF, Trichakis N. 2012. On the efficiency-fairness trade-off. *Management Science* 58:2234–2250
114. Hu L, Chen Y. 2018. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. 1389–1398 pp.
115. Benade G, Kazachkov AM, Procaccia AD, Psomas CA. 2018. How to make envy vanish over time. In *Proceedings of the Nineteenth ACM Conference on Economics and Computation*
116. Lorini E, Mühlenbernd R. 2015. The long-term benefits of following fairness norms: A game-theoretic analysis. In *International Conference on Principles and Practice of Multi-Agent Systems*. Springer, 301–318 pp.
117. Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 797–806 pp.
118. Friedler SA, Scheidegger C, Venkatasubramanian S. 2016. On the (im)possibility of fairness. *arXiv preprint arxiv:1609.07236* abs/1609.07236
119. Smith M, Patil D, Munoz C. 2016. *Big data: A report on algorithmic systems, opportunity, and civil rights*. Executive Office of the President
120. Aghion P, Bloom N, Blundell R, Griffith R, Howitt P. 2005. Competition and innovation: An inverted-u relationship. *The Quarterly J. Economics* 120:701–728
121. Anandkumar A, Michael N, Tang A. 2010. Opportunistic spectrum access with multiple users: Learning under competition. In *Proceedings of the IEEE International Conference on Computer Communications*. IEEE, 1–9 pp.
122. Mansour Y, Slivkins A, Wu ZS. 2018. Competing bandits: Learning under competition. In *Proceedings of the 9th Innovations in Theoretical Computer Science Conference*. 48:1–48:27 pp.
123. Ben-Porat O, Tennenholtz M. 2018. Competing prediction algorithms. *arXiv preprint arXiv:1806.01703*
124. Federal Trade Commission. 2017. Algorithms and collusion. Note from the united states submitted for item 10 of the 127th oecd competition committee on 21-23 june 2017, Federal Trade Commission

125. Mehra SK. 2015. Antitrust and the robo-seller: Competition in the time of algorithms. *Minnesota Law Review* 100:1323

126. Dani V, Hayes TP, Kakade SM. 2008. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*

127. Abbasi-Yadkori Y, Pál D, Szepesvári C. 2011. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*. 2312–2320 pp.

128. Li L, Chu W, Langford J, Schapire RE. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, 661–670 pp.

129. Slivkins A. 2011. Multi-armed bandits on implicit metric spaces. In *Advances in Neural Information Processing Systems*. 1602–1610 pp.

130. Kleinberg R, Slivkins A, Upfal E. 2013. Bandits and experts in metric spaces. *arXiv preprint arXiv:1312.1277*

131. Slivkins A, Radlinski F, Gollapudi S. 2013. Ranked bandits in metric spaces: learning diverse rankings over large document collections. *J. Machine Learning Research* 14:399–436

132. Slivkins A. 2014. Contextual bandits with similarity information. *J. Machine Learning Research* 15:2533–2568

133. Chapelle O, Li L. 2011. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems*. 2249–2257 pp.

134. Agrawal S, Goyal N. 2012. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*. 39–1 pp.

135. Srinivas N, Krause A, Kakade SM, Seeger M. 2009. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*

136. Srinivas N, Krause A, Kakade SM, Seeger MW. 2012. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory* 58:3250–3265

137. Langford J, Zhang T. 2008. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*. 817–824 pp.

138. Chu W, Li L, Reyzin L, Schapire R. 2011. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*. 208–214 pp.

139. Agrawal S, Goyal N. 2013. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*. 127–135 pp.

140. Eghbali R, Fazel M. 2016. Designing smoothing functions for improved worst-case competitive ratio in online optimization. In *Advances in Neural Information Processing Systems*. 3287–3295 pp.

141. Eghbali R, Fazel M, Mesbahi M. 2016. Worst case competitive analysis for online conic optimization. In *Proceedings of the IEEE Conference on Decision and Control*. IEEE, 1945–1950 pp.

142. Eghbali R, Saunderson J, Fazel M. 2018. Competitive online algorithms for resource allocation over the positive semidefinite cone. *arXiv preprint arXiv:1802.01312*

143. Mansour Y, Slivkins A, Syrgkanis V. 2015. Bayesian incentive-compatible bandit exploration. In *Proceedings of the 16th ACM Conference on Economics and Computation*. ACM, 565–582 pp.

144. Kannan S, Kearns M, Morgenstern J, Pai M, Roth A, et al. 2017. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*. ACM, 369–386 pp.

145. Ghalme G, Jain S, Gujar S, Narahari Y. 2017. Thompson sampling based mechanisms for stochastic multi-armed bandit problems. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 87–95 pp.

146. Kohavi R, Longbotham R, Sommerfield D, Henne RM. 2009. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery* 18:140–181
147. Swaminathan A, Krishnamurthy A, Agarwal A, Dudík M, Langford J, et al. 2017. Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems*. 3635–3645 pp.
148. Strehl AL, Langford J, Li L, Kakade S. 2010. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems*. 2217–2225 pp.
149. Bottou L, Peters J, Candela JQ, Charles DX, Chickering M, et al. 2013. Counterfactual reasoning and learning systems: the example of computational advertising. *J. Machine Learning Research* 14:3207–3260
150. Bareinboim E, Forney A, Pearl J. 2015. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*. 1342–1350 pp.
151. Alon N, Cesa-Bianchi N, Dekel O, Koren T. 2015. Online learning with feedback graphs: Beyond bandits. In *Proceedings of the 28th Conference on Learning Theory*. 23–35 pp.
152. Hu H, Li Z, Vetta AR. 2014. Randomized experimental design for causal graph discovery. In *Advances in Neural Information Processing Systems*. 2339–2347 pp.
153. Lattimore F, Lattimore T, Reid MD. 2016. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems*. 1181–1189 pp.