IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 February 2023, accepted 7 March 2023, date of publication 29 March 2023, date of current version 3 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3262984

RESEARCH ARTICLE

Information Disclosure for Increasing User Satisfaction From a Shared Ride

DAVID ZAR, NOAM HAZON, (Member, IEEE), AND AMOS AZARIA
Department of Computer Science, Ariel University, Ariel 4070000, Israel

Corresponding author: Noam Hazon (noamh@ariel.ac.il)

This work was supported in part by the Ministry of Science, Technology and Space, Israel.

ABSTRACT On-demand ridesharing services play a crucial part in the development of modern smart cities. Unfortunately, despite their advantages, not many people opt to use them. We believe that increasing the user satisfaction from the services will cause more people to utilize them. Sometimes, it is possible to increase user satisfaction by providing accurate information related to the alternative modes of transportation, such as a private taxi ride and public transportation. For example, a passenger may be more satisfied with a shared-ride if she is told that a private taxi ride would have cost her 50% more. The challenge is thus to decide which information should be revealed to the user in order to increase the user satisfaction. To address this problem, we model our environment as a signaling game and analyze the perfect Bayesian equilibria for three agents' classes: 1) the honest agent model, in which the agent must only provide truthful information, 2) a no utility for lying model, in which the agent receives no utility if it elects to provide false information, and 3) a penalized false information model, in which the agent is penalized for providing false information. We show that in the honest agent model and in the no utility for lying model, the agent must reveal all the information regarding the possible alternatives to the passenger. However, in the penalized false information model, there are two types of equilibria, one in which she is truthful (but must keep silent sometimes), and the other, in which the agent provides false information. The latter equilibrium type includes equilibria that seem unreasonable. Therefore, we propose a novel criterion to filter out such equilibria, and demonstrate its usefulness in another game.

INDEX TERMS Multi-agent systems, signaling games, information disclosure, perfect Bayesian equilibrium criteria.

I. INTRODUCTION

More than 55% of the world's population are currently living in urban areas, a proportion that is expected to increase up to 68% by 2050 [1]. Sustainable urbanization is a key to successful future development of our society. A key inherent goal of sustainable urbanization is an efficient usage of transportation resources in order to reduce travel costs, avoid congestion, and reduce greenhouse gas emissions.

While traditional services—including buses and taxis—are well established, large potential lies in shared but flexible urban transportation. On-demand ridesharing, where the driver is not a passenger with a specific destination, appears to gain popularity in recent years, and big ride-hailing services such as Uber and Lyft are already offering such services. However, despite the popularity of Uber and Lyft [2], their ridesharing services, which group together multiple passengers (Uber-Pool and Lyft-Line), suffer from low usage [3], [4].

In this paper we propose to increase the user satisfaction from a given shared-ride, in order to encourage her to use the service more often. That is, we attempt to use a form of persuasive technology [5], not in order to convince users to take a shared ride, but to make them feel better with the choice they have already made, and thus improve their attitude towards ridesharing. It is well-known that one of the most influencing factors for driving people to utilize a specific service is to increase their satisfaction from the service (see for example, [6]). Moreover, if people are satisfied and use the service more often it will improve the quality of the

The associate editor coordinating the review of this manuscript and approving it for publication was Sudipta Roy.

service, such as the waiting time, cost, travel time, and service availability, which in turn further increase the user satisfaction.

Sometimes, it is possible to increase user satisfaction by providing accurate information related to the alternative modes of transportation, during the shared ride or immediately after the passenger has completed it. Therefore, we model our environment as a signaling game [7], which models the decision of a rational agent whether to provide the exact price (i.e., the cost or the travel time) of a possible alternative mode of transportation, or not. In this game there are three players: nature, the agent and the passenger. Nature begins by randomly choosing a price from a given distribution; this distribution is known both to the agent and the passenger. The agent observes the price and decides whether to disclose this price to the passenger, provide false information, or keep silent. The passenger then determines her current expectation over the price of the alternative. The goal of the agent is to increase the passenger satisfaction, and thus it would like the passenger to believe that the price of the alternative is higher than the price of the shared-ride as much as possible. We note that the agent may be a human being or a computerized agent.

We use the standard solution concept of Perfect Bayesian Equilibrium (PBE) [8], and analyze three agents' models. In the 'honest agent' (HA) model, the agent is not allowed to report false information. In the 'no utility for lying' (NUFL) model, the agent may provide false information, but she does not receive any utility if she opts to do so. In the third model, 'penalized false information' (PFI), the agent may provide false information, but a penalty is imposed on her for doing so. We show that in the HA and NUFL models, the agent must reveal all the information regarding the price of the possible alternative to the passenger (unless nature selects the minimum possible value, in which the agent may reveal the value, may keep silent, or may use any mixed strategy of the two). However, in the PFI model, there are two types of equilibria, one in which the agent is truthful (but must keep silent for some values of nature), and the other, in which she provides false information. The latter equilibrium type includes equilibria that seem unreasonable. Therefore, we propose a new criterion, the credible belief criterion, to filter out such equilibria. Intuitively, the credible belief criterion states that if the agent deviates, and plays an off-the-path action, the user should not increase her belief (over the prior distribution) in a selection of nature that would cause the agent to lose more by deviating than her belief in a selection of nature that would cause the agent to lose less by deviating. We further demonstrate the usefulness of the credible belief criterion in a signaling game in the context of occupation and education.

The contributions of this paper are twofold:

- We model the information disclosure in the ridesharing domain as a signaling game and determine the unique set of Perfect Bayesian Equilibria (PBE) for three different agent models.
- We introduce the credible belief criterion, which filters unreasonable PBEs.

II. RELATED WORK

A. RIDESHARING

Most work on ridesharing has focused on the assignment of passengers to vehicles. See the comprehensive surveys by Parragh et al. [9], [10], and a recent survey by Psaraftis et al. [11]. In particular, the dial-a-ride problem (DARP) is traditionally distinguished from other problems of ridesharing since transportation cost and user inconvenience must be weighed against each other in order to provide an appropriate solution. Therefore, the DARP typically includes more quality constraints that aim at capturing the user's inconvenience. We refer to a recent survey on DARP by Molenbruch et al. [12], which also makes this distinction. In recent years there is an increasing body of works that concentrate on the passenger's satisfaction during the assignment of passengers to vehicles [13], [14], [15]. Similar to these works we are interested in the satisfaction of the passenger, but instead of developing assignment algorithms (e.g., [16]), we focus on the role of information disclosure as a means to improve user satisfaction.

B. INFORMATION DISCLOSURE

There are other works in which an agent provides information to a human user (in the context of the roads network) for different purposes. For example, Azaria et al. [17], [18], [19] develop agents that provide information or advice to a human user in order to convince her to take a certain route. Several other works have discussed the implications of information disclosure on environmental factors, including traffic and pollution [20], [21].

Bilgic and Mooney [22] present methods for explaining the decisions of a recommendation system to increase the user satisfaction. In their context, user satisfaction is interpreted only as an accurate estimation of the item quality.

Grossman [23] studies markets in which sellers may opt to reveal information to buyers in the form of a set of possible values of their items. The sellers must include the value of their item in the set of values revealed, or they may opt to reveal an empty set. Grossman shows that the buyers will always believe that the item's value is the minimum value in the set revealed by the seller, and only a seller with the least valued item may opt to reveal an empty set. In our work, we model our environment as a signaling game allowing mixed strategies and continuous values, and we analyze it for three agents' classes.

C. SIGNALING GAMES

Signaling games are used to model problems in several domains. For example, Noe [27] models financial decisions of a firm (whether to use equity financing or debt financing) as a signaling game. Bangerter et al. [29] model the job market using signaling games, and analyze relationships between applicants and organizations, among applicants,

TABLE 1. A comparison of related works on signaling games.

Paper	Application	New criterion
Cho and Kreps (1987) [24]	N/A	Intuitive criterion
Cho (1987) [25]	N/A	Forward induction equilibrium
Banks and Sobel (1987) [26]	N/A	Divine criterion
Noe (1998) [27]	Financial decisions of a firm	N/A
Rogers (2001) [28]	Legistratures-court interaction	N/A
Bangerter et al. (2012) [29]	Job market	N/A

and among organizations. Rogers [28] model the interaction between the legislatures and the court as a signaling game. In this work, we use signaling games to model user satisfaction in ridesharing problems, and we use the perfect Bayesian equilibrium as the solution concept [8], [30], [31]. We also consider a refinement of the PBE, the intuitive criterion introduced by Cho and Kreps [24], which filters out PBEs where the user believes that the agent chose an action that would certainly result in a loss. However, there are cases in which this criterion is not adequate, and additional refinements have been suggested. Banks and Sobel define the divine criterion [26], a refinement of the intuitive criterion, that compares the value for the agent with different actions while taking into account the user's actions. Cho suggests [25] the forward induction equilibrium, which is another refinement of the intuitive criterion. In this work, we encounter PBEs that seem unreasonable, yet none of the previously defined criteria filter them. Therefore, we define the credible belief criterion, a novel criterion that filters out these unreasonable equilibria. We further show that this new criterion is useful in other signaling games.

D. PREVIOUSLY PUBLISHED RESULTS

In our previous work [32], we modeled our environment as a signaling game and analyzed the perfect Bayesian equilibria for only a single agent class, the honest agent model. In this paper, we analyze two additional agents' classes: a no utility for lying model, and a penalized false information model. In addition, in this paper, we propose a novel criterion to filter out unreasonable equilibria, and demonstrate its usefulness in another game.

III. PRELIMINARIES

Recall that we attempt to increase user satisfaction by proving accurate information related to alternative modes of transportation. Specifically, we assume that the passenger has some estimate over the possible prices of the alternative modes of transportation, while the agent has a more accurate knowledge related to the prices. Therefore, we model our setting with the following signaling game. We assume that there is a given random variable X with a prior probability distribution over the possible prices of a given alternative mode of transportation. The possible values of X, denoted by the set χ, are bounded within the range $[min, max]$, where $min > 0$. Without loss of generality, $\forall x \in \chi$, $Pr(X = x) > 0$ for a discrete distribution, and $\forall \epsilon > 0$, $F_X(x+\epsilon) - F_X(x-\epsilon) > 0$ for a continuous distribution. In addition, we assume that $min \in \chi$. For ease of notation, when a distribution is concentrated at a single point, we state that the probability at that point is 1, but do not state that the probability of any other value of the random variable is 0.

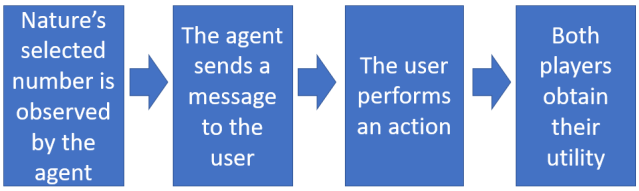

FIGURE 1. A flowchart demonstrating the process of the signaling game.

The game is composed of three players: nature, player 1 (agent) and player 2 (passenger/user). It is assumed that both players are familiar with the prior distribution over X. Nature randomly chooses a number x according to the distribution over X. The agent observes the number x and plays an action $a_1 \in A_1$, where A_1 is the set of possible actions for the agent. We note that A_1 depends on the environment, and it may also depend on nature's choice, x. We denote by $[p, a_1'; (1-p), a_1'']$ a mixed strategy of playing $a_1' \in A_1$ with a probability of p and $a_1'' \in A_1$ with a probability of $(1-p)$, where $0 \leq p \leq 1$. Intuitively, this action is a signal (message) sent to the user. The user observes the agent's action and plays an action $a_2 \in A_2 = [min, max]$. See Figure 1 for a flowchart demonstrating this process, and Table 2 for a list of symbols used throughout the paper. We consider several models for our environment.

IV. HONEST AGENT (HA) MODEL

We begin by considering an agent that is not allowed to provide any false information. That is, the agent's action is either φ (quiet) or x (say), i.e, $A_1 = \{\varphi, x\}$.

That is, we assume that the agent may not provide false information. This is a reasonable assumption, since providing false information is usually prohibited by the law, or may harm the agent's reputation. The user observes the agent's action and her action, denoted a_2, is any number in the range $[min, max]$. The user's action essentially means setting her estimate about the price of the alternative. In our setting, the agent would like the user to think that the price of the alternative is as high as possible, while the user would like to know the real price. Therefore, we set the utility for the agent to a_2 and the utility of the user to $-(a_2 - x)^2$. Note that we did not define the utility of the user to be simply $-|a_2 - x|$, since we want the utility to highly penalize a large deviation from the true value.

We first note that if the agent plays $a_1 \neq \varphi$ then the user knows that a_1 is nature's choice. Thus, a rational user would

TABLE 2. Table of symbols used throughout the paper.

Symbol	Used in Model	Meaning
X	All models	The random variable that represents nature's choice
x	All models	A value of nature's choice
χ	All models	The possible values for X (which have non-zero probability)
A_1	All models	The set of possible actions of the agent
A_2	All models	The set of possible actions of the user
a_1	All models	An action of the agent
a_2	All models	An action of the user
φ	All models	A specific action of the agent that means keep quiet
σ_1	All models	A strategy of the agent
σ_2	All models	A strategy of the user
u_1	All models	The utility of the agent
u_2	All models	The utility of the user
μ_2	All models	The belief of the user
f	PFI	The penalty factor
Y_{a_1}	PFI	A distribution that describes the user belief given a_1
F	PFI	A set of values of x, which the agent lies for
S	PFI	A set of values of x, which the agent is quiet for
T	PFI	A set of values of x, which the agent tells the truth for
$\mathcal{Q}$	PFI	A set of actions that the agent plays when lying
E_F	PFI	The expectation of X, if $a_1 \in \mathcal{Q}$
F_i	PFI	Subsets of F, which form a partition of F
$l()$	PFI	The loss term from the credible belief criterion, when playing an off-the-path action

play $a_2 = a_1$. On the other hand, if the agent plays $a_1 = \varphi$ then the user would have some belief about the real price, which can be the original distribution of nature, or any other distribution. Clearly, the user's best response is to play the expectation of this belief. Formally,

Observation 1: Assume that the agent plays $a_1 = \varphi$, and let Y be a belief over x. That is, Y is a random variable with a distribution over $[min, max]$. Then, $\text{argmax}_{a \in A_2} E[-(a - Y)^2] = E[Y]$.

Proof: Instead of maximizing $E[-(a - Y)^2]$ we can minimize $E[(a - Y)^2]$. In addition, $E[(a - Y)^2] = E[(a)^2] - 2E[aY] + E[Y^2] = (a)^2 - 2aE[Y] + E[Y^2]$. By differentiating we get that

$$\frac{d}{da}\left((a)^2 - 2aE[Y] + E[Y^2]\right) = 2a - 2E[Y].$$

The derivative is 0 when $a = E[Y]$ and the second derivative is positive; this entails that

$$\underset{a \in A_2}{\text{argmin}}\left((a)^2 - 2aE[Y] + E[Y^2]\right) = E[Y].$$

□

Now, informally, if nature chooses a "high" value of x, the agent would like to disclose this value by playing $a_1 = x$. One may think that if nature chooses a "low" value of x, the agent would like to hide this value by playing $a_1 = \varphi$. However, since the user adjusts her belief accordingly, she will play $E[X|a_1 = \varphi]$. Therefore, it would be more beneficial for the agent to reveal also low values that are greater than $E[X|a_1 = \varphi]$, which, in turn, will further reduce the new $E[X|a_1 = \varphi]$. Indeed, Theorem 1 shows that a rational agent should always disclose the true value of x, unless $x = min$. If $x = min$ the agent can play any action, i.e., φ, *min* or any mixture of φ and *min*. We begin by applying the definition of PBE to our signaling game.

Definition 1: A tuple of strategies and a belief, $(\sigma_1, \sigma_2, \mu_2)$, is said to be a perfect Bayesian equilibrium in our setting if the following hold:

1) *The strategy of player 1 is a best response strategy. That is, given σ_2 and x, deviating from σ_1 does not increase player 1's utility.*
2) *The strategy of player 2 is a best response strategy. That is, given a_1, deviating from σ_2 does not increase player 2's expected utility according to her belief.*
3) *μ_2 is a consistent belief. That is, μ_2 is a distribution over x given a_1, which is consistent with σ_1 (following Bayes' rule, where appropriate).*

Theorem 1: A tuple of strategies and a belief, $(\sigma_1, \sigma_2, \mu_2)$, is a PBE if and only if:

- $\sigma_1(x) = \begin{cases} x: & x > min \\ [p, min; (1-p), \varphi], 0 \le p \le 1: & x = min \end{cases}$
- $\sigma_2(a_1) = \begin{cases} a_1: & a_1 \neq \varphi \\ min: & a_1 = \varphi \end{cases}$
- $\mu_2(x = a_1 | a_1 \neq \varphi) = 1$ *and* $\mu_2(x = min | a_1 = \varphi) = 1$.

Proof: ($\Leftarrow$) Such a tuple is a PBE: σ_1 is a best response strategy, since the utility of player 1 is x if $a_1 = x$ and *min* if $a_1 = \varphi$. Thus, playing $a_1 = x$ is a weakly dominating strategy. σ_2 is a best response strategy, since it is the expected value of the belief μ_2, and thus it is a best response according to Observation 1. Finally, μ_2 is consistent: If $a_1 = \varphi$ and according to σ_1 player 1 plays φ with some probability (greater than 0), then according to Bayes' rule $\mu_2(x = min | a_1 = \varphi) = 1$. Otherwise, Bayes' rule cannot be applied (and it is thus not required). If $a_1 \neq \varphi$, then by definition $x = a_1$, and thus $\mu_2(x = a_1 | a_1 \neq \varphi) = 1$.

($\Rightarrow$) Let $(\sigma_1, \sigma_2, \mu_2)$ be a PBE. It holds that $\mu_2(x = a_1 | a_1 \neq \varphi) = 1$ by Bayes' rule, implying that if $a_1 \neq \varphi$, $\sigma_2(a_1) = a_1$. Therefore, when $a_1 = x$ the utility of player 1 is x.

We now show that $\sigma_2(a_1 = \varphi) = min$. Assume by contradiction that $\sigma_2(a_1 = \varphi) \neq min$ (or $Pr(\sigma_2(a_1 = \varphi) =$

min) < 1), then $E[\sigma_2(\varphi)] = c > min$. We now deduce the strategy of player 1. There are three possible cases: if $x > c$, then $a_1 = x$ is a strictly dominating strategy. If $x < c$, then $a_1 = \varphi$ is a strictly dominating strategy. If $x = c$, there is no advantage for either playing φ or x; both options give player 1 a utility of c, and thus she may use any strategy. That is

$$, \sigma_1(x) = \begin{cases} x: & x > c \\ \varphi: & x < c \\ [p, min; (1-p), \varphi], 0 \le p \le 1: & x = c. \end{cases}$$

Given this strategy, we need to apply Bayes' rule to derive $\mu_2(x|a_1 = \varphi)$. By σ_1, it is possible that $a_1 = \varphi$ only if $x \le c$. That is, $\mu_2(x > c|a_1 = \varphi) = 0$ and $\mu_2(x \le c|a_1 = \varphi) = 1$. Therefore, the expected value of the belief, $c' = E_{X\sim\mu_2(x|a_1=\varphi)}[X]$, and according to Observation 1, $\sigma_2(\varphi) = c'$. However, $c' = E_{X\sim\mu_2(x|a_1=\varphi)}[X] \le E[X|X \le c]$ since player 1 plays φ only when $x < c$ and possibly also when $x = c$. In addition, $E[X|X \le c] < c$, since $c > min$. That is, $E[\sigma_2(\varphi)] = c' < c$, which is a contradiction. Therefore, the strategy for player 2 in every PBE is determined. In addition, since $\sigma_2(\varphi) = E_{X\sim\mu_2(x|a_1=\varphi)}[X]$ according to Observation 1, then $\mu_2(x|a_1 = \varphi) = min$, and the belief of player 2 in every PBE is also determined.

We end the proof by showing that for $x > min$, $\sigma_1(x) = x$. Since σ_2 is determined, the utility of player 1 is *min* if $a_1 = \varphi$ and x if $a_1 = x$. Therefore, when $x > min$, playing $a_1 = x$ is a strictly dominating strategy.

□

V. NO UTILITY FOR LYING (NUFL) MODEL

The following model is identical to the first model, except that it allows the agent to provide false information; however, the agent does not receive any utility if she opts to do so. Formally, the agent's action is either φ or any number in the range $[min, max]$ (which does not necessarily equal x), i.e., $A_1 = \{\varphi\} \cup [min, max]$. In this setting, the utility of the agent is

$$u_1(x, a_1, a_2) = \begin{cases} a_2: & a_1 \in \{\varphi, x\} \\ 0: & otherwise. \end{cases}$$

The analysis of the possible PBE for the HA model (Theorem 1) holds for the current model as well. However, in the current model there are additional perfect Bayesian equilibria. For example,

- $\sigma_1(x) = \varphi$
- $\sigma_2(a_1) = \begin{cases} min: & a_1 \ne \varphi \\ E[X]: & a_1 = \varphi \end{cases}$
- $\mu_2(x = min|a_1 \ne \varphi) = 1$ and $\mu_2(x|a_1 = \varphi) = Pr(X = x)$.

Note that the belief μ_2 is consistent, since the agent plays $a_1 \ne \varphi$ with probability 0, and thus Bayes' rule is not violated. Indeed, the user believes that if the agent deviates and plays $a_1 > min$ she does not provide the truthful value of x. However, this belief is not reasonable, since the agent does not have an incentive to do so, as it would result in the lowest possible utility for her (zero). We thus use the intuitive criterion [24] to filter the equilibria with non-reasonable beliefs.

In order to define the intuitive criterion for our setting, we first define the notion of a seemly deviation action. Informally, an action is considered a *seemly deviation* if there exists a situation in which the agent may expect to gain (or not lose) from this deviation.

Definition 2: For nature's choice x and strategy σ_1, let a_1' be an action such that $Pr(\sigma_1(x) = a_1') = 0$. We say that a_1' is a seemly deviation for the agent, if there exist user actions $w, z \in A_2$ such that $u_1(x, a_1', w) \ge u_1(x, \sigma_1(x), z)$.

We note that in our NUFL model, if the agent's strategy for a given x is either φ or x, providing false information is never a seemly deviation for the agent. The reason is that by deviating, the agent will always receive an outcome of zero, regardless of the user's action, which is certainly less than the agent's payoff had she played her original strategy.

Recall that an action is considered an *off-the-path* action for the agent if, according to a specific strategy, it should never be played (regardless of nature's choice of x). That is, an agent action that the user does not expect to see.

Definition 3: Given a strategy for the agent, σ_1, an agent action, $a \in A_1$ is off-the-path, if $\forall x \in \chi\ Pr(\sigma_1(x) = a) = 0$.

We can now define the intuitive criterion for our setting. Informally, the criterion requires that given an off-the-path action a, the user believes that nature's choice of x is such that a is a seemly deviation (unless a is not a seemly deviation for all x).

Definition 4: A Perfect Bayesian Equilibrium, $(\sigma_1, \sigma_2, \mu_2)$, is said to satisfy the intuitive criterion, if for all off-the-path actions $a \in A_1$, if there exists $x \in X$ such that a is a seemly deviation from $\sigma_1(x)$ then for all $x \in X$ that a is not a seemly deviation from $\sigma_1(x)$, $\mu_2(x|a) = 0$.

Clearly, in our NUFL model, a PBE that satisfies the intuitive criterion cannot consist of a user's belief that the agent provides false information with a probability greater than 0.

Similarly to the HA model, we show that under the NUFL model using the intuitive criterion, a rational agent should always disclose the true value of x (unless $x = min$).

Theorem 2: A tuple of strategies and a belief, $(\sigma_1, \sigma_2, \mu_2)$, is a PBE that satisfies the intuitive criterion if and only if:

- $\sigma_1(x) = \begin{cases} x: & x > min \\ [p, min; (1-p), \varphi], 0 \le p \le 1: & x = min \end{cases}$
- $\sigma_2(a_1) = \begin{cases} a_1: & a_1 \ne \varphi \\ min: & a_1 = \varphi \end{cases}$
- $\mu_2(x = a_1|a_1 \ne \varphi) = 1$ *and* $\mu_2(x = min|a_1 = \varphi) = 1$.

Proof: ($\Leftarrow$) As shown in Theorem 1 such a tuple is a PBE. It also satisfies the intuitive criterion: the only actions that can be off-the-path are φ and *min*. Given each of these actions, the user's belief is that $x = min$. In both cases, if $x = min$, the actions $a = \varphi$ and $a = min$ are seemly deviations.

($\Rightarrow$) In any PBE the agent will never lie, since lying is a strictly dominated strategy. Furthermore, since the PBE satisfies the intuitive criterion, the user never believes that the agent lies. Specifically, given an action $a_1 \ne \varphi$, if it is possible to apply Bayes' rule (i.e., the action is not

off-the-path) then the user will not believe that the agent lies. If the action a_1 is off-the-path then the user can believe that $x = a_1$ (the agent told the truth). This is a seemly deviation, since the user can play $a_2 = max$ (which will result in $u_1 = max$). However, the user cannot believe that $x \neq a_1$, since it is not a seemly deviation. Overall, the agent never lies and the user never believes that the agent lies and thus we are back to the case of Theorem 1. □

VI. PENALIZED FALSE INFORMATION (PFI) MODEL

This model is identical to the NUFL model, except for the utility of the agent when providing false information. Namely, the agent is penalized by a fraction of a_2 when she provides false information. Formally, let $0 < f < 1$, the utility of the agent is

$$u_1(x, a_1, a_2) = \begin{cases} a_2 : & a_1 \in \{\varphi, x\} \\ f \cdot a_2 : & otherwise. \end{cases}$$

Note that this formulation captures situations in which there is a chance that the lie is revealed and then the utility is zero. However, there is also a probability (f) that the lie is not revealed, and thus the agent's expected utility, in case of a lie, is $f \cdot a_2$. We assume that $min < f \cdot max$ (otherwise, the PFI model becomes identical to the NUFL model, because the utility for the agent for providing false information is always lower than her utility for playing $a_1 = x$ or $a_1 = \varphi$).

Interestingly, under the PFI model a rational agent should not always disclose the true value of x. Intuitively, if the user always plays $a_2 = a_1$, the agent is better off by playing a_1 that is higher than x, such that $f \cdot a_1 > x$. We obtain two general PBEs: one in which the agent is truthful (but sometimes plays φ), and one in which the agent lies. Specifically, the strategy of a truthful agent is to play φ on a set S (silent), and otherwise to play x (the truth). In general, the agent will remain silent except for some values that are slightly higher than the expectation on the values in S. S cannot be empty, i.e., the agent must keep silent for some values of x, but S may include all values of x, i.e., the agent may always play φ. The strategy of the non-truthful agent uses a partition of the interval $[min, max]$ to three sets: F (false), S (silent), and T (truth). In general, the agent will lie, and she will say the most beneficial lie, that is, the value that will maximize σ_2. However, in some cases the agent will say the truth. Let E_F be the maximum value of σ_2. If $\sigma_2(x)$ is only slightly lower than E_F, that is $\sigma_2(x) \geq f \cdot E_F$, the agent can play x (the truth), since she will not be penalized. The agent may play φ if $\sigma_2(\varphi)$ equals $f \cdot E_F$. We use $\mathcal{Q}$ to indicate the set of lies used by the agent, that is, the values that the agent uses when $a_1 \neq x$.

Note that in the current model the intuitive criterion cannot be violated, since for nature's choice x and a deviation a_1', $u_1(x, a_1', max) > u_1(x, \sigma_1(x), min)$. That is, every deviation of the agent is a seemly deviation. To simplify the exposition, we concentrate on PBEs with pure strategies.

Before we formally describe the PBEs under the PFI model, we show two lemmas that provide constraints on the user's strategy, $sigma_2$, in a PBE.

Lemma 1: If $(\sigma_1, \sigma_2, \mu_2)$ is a PBE then $\forall x_1, x_2 \in X$, $\sigma_2(\sigma_1(x_1)) \geq f \cdot \sigma_2(\sigma_1(x_2))$.

Proof: Assume by contradiction that for some x_1, x_2 it holds that $\sigma_2(\sigma_1(x_1)) < f \cdot \sigma_2(\sigma_1(x_2))$. Then, σ_1 is not a strategy of an equilibrium since the agent will benefit from deviating from it and playing $\sigma_1(x_2)$ given x_1. □

As a corollary of Lemma 1 we can deduce that there exists some c such that $\sigma_2(\sigma_1(\cdot)) \in [f \cdot c, c]$.

Lemma 2: $\forall x \in X$, $\sigma_2(\sigma_1(x)) \geq \sigma_2(\varphi)$.

Proof: Assume by contradiction that for some x it holds that $\sigma_2(\sigma_1(x)) < \sigma_2(\varphi)$. Then, σ_1 is not a strategy of an equilibrium since the agent will benefit from deviating from it and playing φ given x. □

We are now ready to formally describe the PBEs under the PFI model.

Theorem 3: A tuple of strategies and a belief, $(\sigma_1, \sigma_2, \mu_2)$, is a PBE if and only if it is one of the following:

1) *(truthful agent) Let $S \subseteq [min, max]$ where S is non-empty, such that if $x \notin S$ then $E[X \mid X \in S] \leq x \leq E[X \mid X \in S]/f$. For $s \in S$ let Y_s be a random variable such that $E[Y_s] \leq E[X \mid X \in S]$.*
   - $\sigma_1(x) = \begin{cases} \varphi : & x \in S \\ x : & otherwise \end{cases}$
   - $\sigma_2(a_1) = \begin{cases} E[X \mid X \in S] : & a_1 = \varphi \\ a_1 : & a_1 \notin S \cup \{\varphi\} \\ E[Y_{a_1}] : & a_1 \in S \end{cases}$
   - $\mu_2(x = a_1 \mid a_1 \notin S \cup \{\varphi\}) = 1$
     $\mu_2(x \mid a_1 = \varphi) = \begin{cases} \frac{Pr(X = x)}{Pr(\sigma_1(X) = \varphi)} : & x \in S \\ 0 : & x \notin S \end{cases}$
     $\mu_2(x \mid a_1 \in S) = Y_{a_1}$.
2) *(non-truthful agent) Let F, S, T be a partition of $[min, max]$ where F is not empty. Let $\mathcal{Q} = \{q_1, \ldots, q_r\}$ for some natural number r, where $q_i \in [min, max]$ and $\forall i \neq j$, $q_i \neq q_j$. Let $E_F = E[X \mid X \in F \cup (\mathcal{Q} \cap T)]$. Let $F_1, F_2, \ldots, F_r$ be a partition of F, such that for all $i \in \{1, 2, \ldots, r\}$ it holds that $E[X \mid X \in F_i \cup (\{q_i\} \cap T)] = E_F$. For each $x \in T$, $f \cdot E_F \leq x \leq E_F$. For $x \notin T \cup \mathcal{Q}$, let Y_x be a random variable such that $E[Y_x] \leq f \cdot E_F$, and let Y_φ be also such a variable. If S is not empty, then $E[X \mid X \in S] = f \cdot E_F$.*
   - $\sigma_1(x) = \begin{cases} q_i : & x \in F_i \text{ for some } i \\ x : & x \in T \\ \varphi : & x \in S \end{cases}$
   - $\sigma_2(a_1) = \begin{cases} a_1 : & a_1 \in T \setminus \mathcal{Q} \\ f \cdot E_F : & a_1 = \varphi \text{ and } S \neq \emptyset \\ E_F : & a_1 \in \mathcal{Q} \\ E[Y_{a_1}] : & otherwise \end{cases}$
   - $\mu_2(x = a_1 \mid a_1 \in T \setminus \mathcal{Q}) = 1$
     $\mu_2(x \mid a_1 = q_i) =$
     $\begin{cases} \frac{Pr(X = x)}{Pr(X \in F_i \cup (\{q_i\} \cap T))} : & x \in F_i \cup (\{q_i\} \cap T) \\ 0 : & otherwise \end{cases}$
     $\mu_2(x \mid a_1 \notin T \cup \mathcal{Q} \text{ or } (a_1 = \varphi \text{ and } S = \emptyset)) = Pr(Y_{a_1} = x)$.

If $S \neq \emptyset$ *then*

$$\mu_2(x \mid a_1 = \varphi) = \begin{cases} \dfrac{Pr(X = x)}{Pr(\sigma_1(X) = \varphi)} : & x \in S \\ 0 : & x \notin S. \end{cases}$$

Proof: We begin with the truthful agent case.

($\Leftarrow$) Let $(\sigma_1, \sigma_2, \mu_2)$ be a tuple of strategy and belief that satisfies the conditions of the truthful agent. μ_2 satisfies Bayes' rule:

- If $a_1 \notin S \cup \{\varphi\}$, according to σ_1, $a_1 = x$; therefore, by Bayes' rule: $\mu_2(x = a_1 \mid a_1 \notin S \cup \{\varphi\}) = 1$.
- If $a_1 = \varphi$, according to σ_1 and Bayes' rule: $\mu_2(x \mid a_1 = \varphi) = \frac{Pr(X=x)Pr(a_1=\varphi|x)}{Pr(\sigma_1(X)=\varphi)} = \begin{cases} \dfrac{Pr(X = x)}{Pr(\sigma_1(X) = \varphi)} : & x \in S \\ 0 : & x \notin S. \end{cases}$
- If $a_1 \in S$ then the agent's action is off-the-path, and thus mu_2 is not required to follow Bayes' rule.

Given σ_1 and μ_2, the strategy of the user, σ_2, is a best response, since it is the expectation over the user's belief regarding x (according to Observation 1). Finally, given σ_2 and μ_2, the agent does not have an incentive to deviate from σ_1:

- If $x \in S$, the agent strategy is $\sigma_1(x) = \varphi$, and the utility is $E[X \mid X \in S]$. If the agent deviates and plays x instead, her utility is $E[Y_x]$ which is at most $E[X \mid X \in S]$. If the agent plays any other action $a_1 \notin \{x, \varphi\}$ then her utility is $f \cdot \sigma_2(a_1)$. However, the maximum value of σ_2 is $E[X \mid X \in S]/f$, which is obtained when $a_1 = max(A_1 \setminus (S \cup \{\varphi\}))$. Therefore, there is no action that provides higher utility for the agent.
- If $x \notin S$, the agent strategy is $\sigma_1(x) = x$, and the utility is x. By definition, $x \geq E[X \mid X \in S]$. If the agent deviates and plays φ instead, her utility is $E[X \mid X \in S]$. If the agent plays any other action her maximal utility is $f \cdot E[X \mid X \in S]/f$. Therefore, there is no action that provides higher utility for the agent.

($\Rightarrow$) Let $(\sigma_1, \sigma_2, \mu_2)$ be a tuple of strategies and belief in PBE, and assume that $\forall x, \sigma_1(x) \in \{\varphi, x\}$. That is, there exists a set $S = \{x : \sigma_1(x) = \varphi\}$, where for $x \notin S$, $\sigma_1(x) = x$. Applying Bayes' rule entails that: $\mu_2(x \mid a_1 = \varphi) = \frac{Pr(\sigma_1(X)=\varphi|X=x) \cdot Pr(X=x)}{Pr(\sigma_1(X)=\varphi)}$. That is, if $x \in S$, $\mu_2(x \mid a_1 = \varphi) = \frac{Pr(X=x)}{Pr(\sigma_1(X)=\varphi)}$, and 0 otherwise. For $s \in S$ define $Y_s = \mu_2(x \mid a_1 = s)$. For any other a_1, $\sigma_1(x) = x$, therefore, (according to Bayes' rule): $\mu_2(x = a_1 \mid a_1 \notin S \cup \{\varphi\}) = 1$. Since the user plays the expectation on her belief, the user's strategy in a PBE must match the σ_2 defined above. It remains to show that for every $s \in S$ it holds that $E[Y_s] \leq E[X \mid X \in S]$. For $x \in S$, $\sigma_1(x) = \varphi$. Therefore, since the strategies are in PBE, $u_1(x, \varphi, \sigma_2(\varphi) \geq u_1(x, a_1, \sigma_2(a_1)$ for every a_1 (otherwise the agent would have an incentive to deviate). Hence, we can set $a_1 = x$, and obtain $E[X \mid X \in S] \geq E[Y_x]$.

We now consider the non-truthful agent case.

($\Leftarrow$) Let $(\sigma_1, \sigma_2, \mu_2)$ be a tuple of strategy and belief that satisfies the conditions of the non-truthful agent. μ_2 satisfies Bayes' rule:

- If $a_1 \in T \setminus \mathcal{Q}$ according to σ_1, $a_1 = x$; therefore, by Bayes' rule: $\mu_2(x = a_1 | a_1 \in T \setminus \mathcal{Q}) = 1$.
- If $S \neq \emptyset$ and $a_1 = \varphi$, according to σ_1 and Bayes' rule: $\mu_2(x \mid a_1 = \varphi) = \frac{Pr(X=x)Pr(a_1=\varphi|x)}{Pr(\sigma_1(X)=\varphi)} = \begin{cases} \dfrac{Pr(X = x)}{Pr(\sigma_1(X) = \varphi)} : & x \in S \\ 0 : & x \notin S. \end{cases}$
- If $a_1 = q_i$ (for some i), according to σ_1 and Bayes' rule: $\mu_2(x \mid a_1 = q_i) = \frac{Pr(X=x)Pr(a_1=q_i|x)}{Pr(\sigma_1(X)=q_i)} = \begin{cases} \dfrac{Pr(X = x)}{Pr(X \in F_i \cup (\{q_i\} \cap T))} : & x \in F_i \cup (\{q_i\} \cap T) \\ 0 : & \text{otherwise.} \end{cases}$
- Otherwise (i.e., $a_1 \in (S \cup F) \setminus \mathcal{Q}$), the agent's action is off-the-path, and thus μ_2 is not required to follow Bayes' rule.

Given σ_1 and μ_2, the strategy of the user, σ_2, is a best response, since it is the expectation over the user's belief regarding x. Finally, given σ_2 and μ_2, the agent does not have an incentive to deviate from σ_1:

- If $x \in F_i$ for some i, the agent strategy is $\sigma_1(x) = q_i$, and the utility is $f \cdot E_F$. Note that $\max_x \sigma_2(x) = E_F$; therefore, there is no other non-truthful action that provides higher utility for the agent. In addition, if the agent deviates and plays x instead, her utility is $E[Y_x] \leq f \cdot E_F$. Similarly, playing φ results in a utility of at most $f \cdot E_F$. Therefore, there is no action that provides higher utility for the agent.
- If $x \in T$, the agent strategy is $\sigma_1(x) = x$, and the utility is either E_F or x, which is at least $f \cdot E_F$. If the agent deviates and plays φ instead, her utility is at most $f \cdot E_F$. Any other action is non-truthful and thus results in a utility at most $f \cdot E_F$. Therefore, there is no action that provides higher utility for the agent.
- If $x \in S$, the agent strategy is $\sigma_1(x) = \varphi$, and the utility is $f \cdot E_F$. If the agent deviates and plays x instead, her utility is $E[Y_x]$ which is at most $f \cdot E_F$. Any other action is non-truthful and thus results in a utility at most $f \cdot E_F$. Therefore, there is no action that provides higher utility for the agent.

($\Rightarrow$) Let $(\sigma_1, \sigma_2, \mu_2)$ be a tuple of strategies and belief in PBE, and assume that there exists x such that $\sigma_1(x) \notin \{x, \varphi\}$. Let $F = \{x : \sigma_1(x) \notin \{x, \varphi\}\}$. Let $S = \{x : \sigma_1(x) = \varphi\}$ and $T = \{x : \sigma_1(x) = x\}$. Clearly, F, S and T are a partition of $[min, max]$. Let $\mathcal{Q} = \{\sigma_1(x) : x \in F\}$ and $r = |\mathcal{Q}|$. Denote the members of $\mathcal{Q}$ as $q_1, \ldots, q_r$, and for $i \in [r]$ let $F_i = \{x \in F : \sigma_1(x) = q_i\}$. Assume towards contradiction that there exist $x_1, x_2 \in F$ such that $u_1(x_1, \sigma_1(x_1), \sigma_2(\sigma_1(x_1))) > u_2(x_2, \sigma_1(x_2), \sigma_2(\sigma_1(x_2)))$. Then, the agent should deviate by playing $\sigma_1(x_1)$ when $x = x_2$, which is a contradiction to $(\sigma_1, \sigma_2, \mu_2)$ being a PBE. Therefore, in equilibrium, all $x \in F$ must lead to the same utility for the agent, and the user's action must be the same for any $q \in \mathcal{Q}$; denote this action by E_F. That is, the utility of the agent is $f \cdot E_F$. Similarly, if S is not empty, then

$\sigma_2(\varphi) = f \cdot E_F$, otherwise the agent should deviate and play some $q \in \mathcal{Q}$ if $\sigma_2(\varphi) < f \cdot E_F$, or play φ instead of lying if $\sigma_2(\varphi) > f \cdot E_F$. Following the above arguments regarding σ_1 and since μ_2 must follow Bayes' rule when it is applicable, we obtain that $\mu_2(x = a_1 \mid a_1 \in T \setminus \mathcal{Q}) = 1$, $\mu_2(x \mid a_1 = q_i) = \begin{cases} \frac{Pr(X=x)}{Pr(X \in F_i \cup (\{q_i\} \cap T))} : & x \in F_i \cup (\{q_i\} \cap T) \\ 0 : & \text{otherwise} \end{cases}$, and if $S \neq \emptyset$ then $\mu_2(x \mid a_1 = \varphi) = \begin{cases} \frac{Pr(X=x)}{Pr(\sigma_1(X)=\varphi)} : & x \in S \\ 0 : & x \notin S \end{cases}$. Since the user must play the expected value of her belief, for any q_i, $\sigma_2(q_i) = \sum_{x \in [min,max]} x \cdot \mu_2(x|a_1 = q_i) = \sum_{x \in [min,max]} x \cdot Pr(X = x \mid X \in F_i \cup (\{q_i\} \cap T) = E[X|X \in F_i \cup (\{q_i\} \cap T)] = E_F$. That is, $E_F = E[X \mid X \in F \cup (\mathcal{Q} \cap T)]$. Overall, the strategy of the agent in a PBE must match the σ_1 defined above.

For an off-the-path action a_1, that is $a_1 \notin T \cup \mathcal{Q}$, or $a_1 = \varphi$ and $S = \emptyset$, the belief is a random variable; we denote this variable as Y_{a_1}. Since $\sigma_2(a_1) = E[Y_{a_1}]$, then $E[Y_{a_1}] \leq f \cdot E_F$. Otherwise, if $E[Y_{a_1}] > f \cdot E_F$ the agent will have an incentive to deviate and play a_1. Specifically, if $E[Y_\varphi] > f \cdot E_F$, the agent will benefit from playing φ when $x \in F$, and if for some $a \in [min, max]$ $E[Y_a] > f \cdot E_F$, the agent will benefit from playing a when $x = a$. Overall, the belief of the user and her strategy in a PBE must match μ_2 and σ_2 defined above, respectively.

□

VII. CREDIBLE BELIEF CRITERION

The PBEs in which the agent of the PFI model is non-truthful include equilibria that seem unreasonable. Consider the following PBE: the agent always plays $a_1 = \frac{min+max}{2}$. First note that the agent always lies, unless $x = \frac{min+max}{2}$. Therefore, $E_F = E[X]$ and her utility will be $f \cdot E[X]$ (unless $x = \frac{min+max}{2}$), while a truthful agent obtains a utility of $E[X]$. Suppose that $x = max$, the agent will still play $a_1 = \frac{min+max}{2}$ since playing max or even φ would cause the user to update her belief such that the expectation of X under this belief is less than $f \cdot E_F$, which will result in a lower utility for the agent. However, while the user's belief does not violate Bayes' rule or the intuitive criterion, there is no justification for it, except for allowing this PBE.

We therefore propose a new filtering criterion, by applying a restriction on the belief of the user. Namely, we propose the *credible belief* criterion, which intuitively states that if the agent deviates, and plays an off-the-path action, the user should not increase her belief (over the prior distribution) in a selection of nature that would cause the agent to lose more by deviating than her belief in a selection of nature that would cause the agent to lose less by deviating. For the previous example, suppose that $\sigma_2(max) = min$, which implies that $\mu_2(x = min|a_1 = max) = 1$. However, $u_1(min, max, min) = f \cdot min$ and $u_1(min, \frac{min+max}{2}, E_F) = f \cdot E_F$ so $u_1(min, \frac{min+max}{2}, E_F) - u_1(min, max, min) = f \cdot E_F - f \cdot min$. On the other hand, $u_1(max, \frac{min+max}{2}, E_F) - u_1(max, max, min) = f \cdot E_F - min$; therefore, the agent loses more from deviating and playing $a_1 = max$ when $x = min$ than when $x = max$, but the user increased her belief (over the prior) for $x = min$ and decreased it for $x = max$.

For the definition of the credible belief criterion, we use the following notation. Given a PBE, let

$$l(x, a_1) = u_1(x, \sigma_1(x), \sigma_2(\sigma_1(x))) - u_1(x, a_1, \sigma_2(a_1)).$$

Intuitively, $l(x, a_1)$ is the loss in utility of the agent when nature chose x and the agent deviates and plays a_1 (instead of $\sigma_1(x)$).

Definition 5: A tuple of strategies and a belief $(\sigma_1, \sigma_2, \mu_2)$ that form a PBE, is said to violate the credible belief criterion if there exists an off-the-path action a_1 and $x_1, x_2 \in [min, max]$ such that $l(x_1, a_1) \leq l(x_2, a_1)$ but $Pr(X = x_2) \cdot \mu_2(x = x_1 \mid a_1) < Pr(X = x_1) \cdot \mu_2(x = x_2 \mid a_1)$.

Intuitively, we would have liked to write the last inequality in Definition 5 as $\frac{\mu_2(x=x_1|a_1)}{\mu_2(x=x_2|a_1)} < \frac{Pr(X=x_1)}{Pr(X=x_2)}$ or $\frac{\mu_2(x=x_1|a_1)}{Pr(X=x_1)} < \frac{\mu_2(x=x_2|a_1)}{Pr(X=x_2)}$; however, since the denominators may be zero, we use the equivalent inequality $Pr(X = x_2) \cdot \mu_2(x = x_1 \mid a_1) < Pr(X = x_1) \cdot \mu_2(x = x_2 \mid a_1)$.

The following theorem describes the PBEs under the PFI model that satisfy the credible belief criterion (based on the PBEs that appear in Theorem 3).

Theorem 4: A tuple of strategies and a belief $(\sigma_1, \sigma_2, \mu_2)$ is a PBE that satisfies the credible belief criterion, if it takes the form of case (1) in Theorem 3 (truthful agent) with the following restrictions on $\mu_2(x \mid a_1)$ for an off-the-path action a_1, which, in turn, restrict Y_{a_1}:

1) $\forall x_1, x_2 \in S \setminus \{a_1\}, Pr(X = x_2) \cdot \mu_2(X = x_1 \mid a_1) = Pr(X = x_1) \cdot \mu_2(X = x_2 \mid a_1)$.
2) $\forall x_1 \in S, x_2 \notin S, Pr(X = x_2) \cdot \mu_2(X = x_1 \mid a_1) \geq Pr(X = x_1) \cdot \mu_2(X = x_2 \mid a_1)$.
3) $\forall x_1, x_2 \notin S$, *where* $x_1 < x_2$, $Pr(X = x_2) \cdot \mu_2(X = x_1 \mid a_1) \geq Pr(X = x_1) \cdot \mu_2(X = x_2 \mid a_1)$.
4) $\forall x \in S, Pr(X = x) \cdot \mu_2(X = a_1 \mid a_1) \geq Pr(X = a_1) \cdot \mu_2(X = x \mid a_1)$,

or if it takes the form of case (2) in Theorem 3 (non-truthful agent) with the following restrictions on $\mu_2(x \mid a_1)$ for an off-the-path action a_1, which, in turn, restrict Y_{a_1}:

1) $\forall x \in F \cup S \cup T \setminus \{a_1\}, Pr(X = x) \cdot \mu_2(X = a_1 \mid a_1) \geq Pr(X = a_1) \cdot \mu_2(X = x \mid a_1)$.
2) $\forall x_1, x_2 \in F \cup S \setminus \{a_1\}, Pr(X = x_2) \cdot \mu_2(X = x_1 \mid a_1) = Pr(X = x_1) \cdot \mu_2(X = x_2 \mid a_1)$.
3) $\forall x_1 \in F \cup S, x_2 \in T, Pr(X = x_2) \cdot \mu_2(X = x_1 \mid a_1) \geq Pr(X = x_1) \cdot \mu_2(X = x_2 \mid a_1)$.
4) $\forall x_1, x_2 \in T \setminus \mathcal{Q}$, *where* $x_1 < x_2$, $Pr(X = x_2) \cdot \mu_2(X = x_1 \mid a_1) \geq Pr(X = x_1) \cdot \mu_2(X = x_2 \mid a_1)$.
5) $\forall x_1 \in T \setminus \mathcal{Q}, x_2 \in \mathcal{Q}, Pr(X = x_2) \cdot \mu_2(X = x_1 \mid a_1) \geq Pr(X = x_1) \cdot \mu_2(X = x_2 \mid a_1)$.
6) $\forall x_1, x_2 \in \mathcal{Q}, Pr(X = x_2) \cdot \mu_2(X = x_1 \mid a_1) = Pr(X = x_1) \cdot \mu_2(X = x_2 \mid a_1)$.

Proof: We begin by showing that there exists at least one instance that follows the form of case (1) in Theorem 3 that satisfies the above restrictions. Specifically, $\forall x \notin S$, we may set $\mu_2(X = x \mid a_1) = 0$ and $\forall x \in S$, we may set $\mu_2(X = x \mid a_1) = \frac{Pr(X=x)}{Pr(X \in S)}$. By doing so all the above restrictions are satisfied. Furthermore, in this case

$E[Y_{a_1}] = E[X \mid X \in S]$, which satisfies the restriction on Y_{a_1} in Theorem 3. This implies that the additional set of restrictions on $\mu_2(x \mid a_1)$ does not nullify the PBE of the form of case (1) in Theorem 3.

Next, we show that any PBE that takes the form of case (1) in Theorem 3 and satisfies the above restrictions, satisfies the credible belief criterion. We note that the credible belief criterion is only applicable to the user's belief for the agent's off-the-path actions, i.e., $\mu_2(x \mid a_1)$. Therefore, we only consider the case that $a_1 \in S$. We consider the following different cases for x: $x = a_1, x \in S \setminus \{a_1\}$, and $x \notin S$. We note the following:

- $l(x = a_1, a_1) < l(x \in S, a_1) < l(x \notin S, a_1)$, since $E[X \mid X \in S] - E[Y_{a_1}] < E[X \mid X \in S] - f \cdot E[Y_{a_1}]$ and for all $x \notin S$, $E[X \mid X \in S] - f \cdot E[Y_{a_1}] < x - f \cdot E[Y_{a_1}]$.
- for $x_1, x_2 \notin S$, where $x_1 < x_2$, $l(x_1, a_1) < l(x_2, a_1)$.

We show that for any x_1, x_2, if $l(x_1, a_1) \leq l(x_2, a_1)$ then $Pr(X = x_2) \cdot \mu_2(x = x_1 \mid a_1) \geq Pr(X = x_1) \cdot \mu_2(x = x_2 \mid a_1)$. There are five possible cases:

- $x_1, x_2 \in S \setminus \{a_1\}$, the credible belief criterion is satisfied by restriction (1).
- $x_1 \in S \setminus \{a_1\}, x_2 \notin S$, the credible belief criterion is satisfied by restriction (2).
- $x_1, x_2 \notin S$ and $x_1 < x_2$, the credible belief criterion is satisfied by restriction (3).
- $x_1 = a_1, x_2 \in S$, the credible belief criterion is satisfied by restriction (4).
- $x_1 = a_1, x_2 \notin S$, the credible belief criterion is satisfied by restriction (2).

Next, we show that any PBE that takes the form of case (2) in Theorem 3 and satisfies the above restrictions, satisfies the credible belief criterion. Recall that since a_1 is an off-the-path action, $a_1 \in F \cup S$. We show that for any x_1, x_2, if $l(x_1, a_1) \leq l(x_2, a_1)$ then $Pr(X = x_2) \cdot \mu_2(x = x_1 \mid a_1) \geq Pr(X = x_1) \cdot \mu_2(x = x_2 \mid a_1)$. There are six possible cases:

- $x_1 = a_1, x_2 \in F \cup S \cup T \setminus \{a_1\}$, the credible belief criterion is satisfied by restriction (1).
- $x_1, x_2 \in F \cup S \setminus \{a_1\}$, the credible belief criterion is satisfied by restriction (2).
- $x_1 \in F \cup S \setminus \{a_1\}, x_2 \in T$, the credible belief criterion is satisfied by restriction (3).
- $x_1, x_2 \in T \setminus \mathcal{Q}$, the credible belief criterion is satisfied by restriction (4).
- $x_1 \in T \setminus \mathcal{Q}, x_2 \in \mathcal{Q}$, the credible belief criterion is satisfied by restriction (5).
- $x_1, x_2 \in \mathcal{Q}$, the credible belief criterion is satisfied by restriction (6).

We proceed by proving that the credible belief criterion is not satisfied in any other case. We first show that in case (1) of Theorem 3 (truthful agent) where the above restrictions are violated, the credible belief criterion does not hold.

- If restriction (1) is violated, then there exist $x_1, x_2 \in S \setminus \{a_1\}$ such that $Pr(X = x_2) \cdot \mu_2(X = x_1 \mid a_1) < Pr(X = x_1) \cdot \mu_2(X = x_2 \mid a_1)$. But since $l(x_1, a_1) = l(x_2, a_1)$, this violates the credible belief criterion.
- If restriction (2) is violated, then there exist $x_1 \in S, x_2 \notin S$ such that $Pr(X = x_2) \cdot \mu_2(X = x_1 \mid a_1) < Pr(X = x_1) \cdot \mu_2(X = x_2 \mid a_1)$. But since $l(x_1, a_1) < l(x_2, a_1)$, this violates the credible belief criterion.
- If restriction (3) is violated, then there exist $x_1, x_2 \notin S$, such that $Pr(X = x_2) \cdot \mu_2(X = x_1 \mid a_1) < Pr(X = x_1) \cdot \mu_2(X = x_2 \mid a_1)$. But since $l(x_1, a_1) < l(x_2, a_1)$, this violates the credible belief criterion.
- If restriction (4) is violated, then there exist $x \in S$ such that $Pr(X = x) \cdot \mu_2(X = a_1 \mid a_1) < Pr(X = a_1) \cdot \mu_2(X = x \mid a_1)$. But since $l(x, a_1) < l(a_1, a_1)$, this violates the credible belief criterion.

Finally, we show that in case (2) of Theorem 3 (non-truthful agent), where the above restrictions are violated, the credible belief criterion does not hold.

- If restriction (1) is violated, then there exist $x \in F \cup S \cup T \setminus \{a_1\}$ such that $Pr(X = x) \cdot \mu_2(X = a_1 \mid a_1) < Pr(X = a_1) \cdot \mu_2(X = x \mid a_1)$. But since $l(x, a_1) < l(a_1, a_1)$, this violates the credible belief criterion.
- If restriction (2) is violated, then there exist $x_1, x_2 \in F \cup S \setminus \{a_1\}$ such that $Pr(X = x_2) \cdot \mu_2(X = x_1 \mid a_1) < Pr(X = x_1) \cdot \mu_2(X = x_2 \mid a_1)$. But since $l(x_1, a_1) = l(x_2, a_1)$, this violates the credible belief criterion.
- If restriction (3) is violated, then there exist $x_1 \in F \cup S \setminus \{a_1\}, x_2 \in T$ such that $Pr(X = x_2) \cdot \mu_2(X = x_1 \mid a_1) < Pr(X = x_1) \cdot \mu_2(X = x_2 \mid a_1)$. But since $l(x_1, a_1) < l(x_2, a_1)$, this violates the credible belief criterion.
- If restriction (4) is violated, then there exist $x_1, x_2 \in T \setminus \mathcal{Q}$, where $x_1 < x_2$, such that $Pr(X = x_2) \cdot \mu_2(X = x_1 \mid a_1) < Pr(X = x_1) \cdot \mu_2(X = x_2 \mid a_1)$. But since $l(x_1, a_1) < l(x_2, a_1)$, this violates the credible belief criterion.
- If restriction (5) is violated, then there exist $x_1 \in T \setminus \mathcal{Q}, x_2 \in \mathcal{Q}$ such that $Pr(X = x_2) \cdot \mu_2(X = x_1 \mid a_1) < Pr(X = x_1) \cdot \mu_2(X = x_2 \mid a_1)$. But since $l(x_1, a_1) < l(x_2, a_1)$, this violates the credible belief criterion.
- If restriction (6) is violated, then there exist $x_1, x_2 \in \mathcal{Q}$ such that $Pr(X = x_2) \cdot \mu_2(X = x_1 \mid a_1) < Pr(X = x_1) \cdot \mu_2(X = x_2 \mid a_1)$. But since $l(x_1, a_1) = l(x_2, a_1)$, this violates the credible belief criterion.

□

Finally, we show another signaling game in which the credible belief criterion is useful. In this game there are two players: a worker and an employer. There are three types of workers: spiritual, social, and analytical. The worker type is drawn from a uniform distribution known to the employer; the worker is familiar with her type. The worker has to choose which education to acquire: spiritual education, social education or analytical education. The education is visible to the employer and thus, serves as a signal. Education that matches the worker's type is obtained for free, but she must pay 1 for education that does not match her type. After acquiring her education, the worker is assigned, by the employer, to one of three jobs: spiritual job, social job, or analytical job. The worker obtains a reward of 1 for spiritual job, 2 for social job, and 3 for analytical job, regardless of her type and education. The employer's utility is 1 if the worker's job matches her type, and -1 otherwise. Formally, the game is defined as follows:

TABLE 3. A comparison of our agent models.

Model	Penalty for Lying	Solution Concept
Honest Agent (HA)	Not possible	PBE
No Utility For Lying (NUFL)	High	PBE + Intuitive criterion
Penalized False Information (PFI)	Low	PBE + Credible belief criterion

- $Types = \{sp, so, an\}$ where $\forall x \in Types, Pr(X = x) = 1/3$
- $A_1 = \{sp_{ed}, so_{ed}, an_{ed}\}$
- $A_2 = \{sp_j, so_j, an_j\}$
- $u_1(x, a_1, a_2) = reward(a_2) - payment(x, a_1)$, where:
  - $reward(a_2) = \begin{cases} 1: & a_2 = sp_j \\ 2: & a_2 = so_j \\ 3: & a_2 = an_j \end{cases}$
  - $payment(x, a_1) = \begin{cases} 0: & x = a_1 \\ 1: & x \neq a_1 \end{cases}$
- $u_2(x, a_1, a_2) = \begin{cases} 1: & x = a_2 \\ 0: & x \neq a_2 \end{cases}$

One of the PBEs in this game is the following:

- $\sigma_1(x) = sp_{ed}$
- $\sigma_2(a_1) = \begin{cases} so_j: & a_1 = sp_{ed} \\ sp_j: & \text{otherwise} \end{cases}$
- $\mu_2(X \mid a_1 = sp_{ed}) = \begin{cases} 1/3: & x = sp \\ 1/3: & x = so \\ 1/3: & x = an \end{cases}$
- $\mu_2(X \mid a_1 \neq sp_{ed}) = \begin{cases} 1: & x = sp \\ 0: & x = so \\ 0: & x = an \end{cases}$

This tuple is a PBE. The worker does not benefit from deviating: if the worker is of a spiritual type, she will only lose from choosing any other education. If the worker is of a social or analytical type, and she chooses any other education, the employer will assign her to a spiritual job, which will result in a lower or equal utility. The employer also does not benefit from deviating: if the worker played sp_{ed}, according to the employer's belief, all types are equally likely, so the employer does not benefit from deviating. If the worker played so_{ed} or an_{ed}, according to the employer's belief, the worker's type is sp, so she must play sp_j. Finally, the belief is consistent: for $a_1 = sp_{ed}$ the belief is same as the original distribution, which is consistent with Bayes' rule since $\sigma_1(X) = sp_{ed}$ with probability of 1. For $a_1 \neq sp_{ed}$, which is off-the-path, any belief is consistent.

Indeed, this PBE is unreasonable. For example, if the worker chose to acquire analytical education, it is more likely that her type is analytical, but the employer believes that the worker is of a spiritual type. The intuitive criterion does not filter this PBE, because it is always possible for the employer to play $a_2 = an_j$, in which case the worker will not lose.

However, the credible belief criterion filters this PBE: for the off-the-path action an_{ed}, the worker loses more if her type is an than if her type were sp; however, the employer increases her belief over the prior more for $x = sp$ than for $x = an$. More formally, if $a_1 = an_{ed}$, and $x_1 = an, x_2 = sp$, it holds that $l(x_1, a_1) < l(x_2, a_1)$, but $\frac{\mu_2(x_1|a_1)}{Pr(X=x_1)} < \frac{\mu_2(x_2|a_1)}{Pr(X=x_2)}$.

We note that there is a PBE in this game that satisfies the credible belief criterion:

- $\sigma_1(x) = \begin{cases} so_{ed}: & x = so \\ an_{ed}: & \text{otherwise} \end{cases}$
- $\sigma_2(a_1) = \begin{cases} so_j: & a_1 = so_{ed} \\ an_j: & a_1 = an_{ed} \\ sp_j: & a_1 = sp_{ed} \end{cases}$
- $\mu_2(X = sp \mid a_1 = sp_{ed}) = 1$
- $\mu_2(X = so \mid a_1 = so_{ed}) = 1$
- $\mu_2(X = an \mid a_1 = an_{ed}) = \begin{cases} 1/2: & x = sp \\ 0: & x = so \\ 1/2: & x = an \end{cases}$

This is a PBE since no player can benefit from deviating and the employer's belief is consistent. Moreover, the credible belief is satisfied since for the only off-the-path action $a_1 = sp_{ed}$, the belief is higher than the prior only for $x = sp$, and as required, this is the x with the lowest loss: $l(sp, sp_{ed}) = 1$, $l(so, sp_{ed}) = 2$ and $l(an, sp_{ed}) = 3$.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we took a first step towards the analysis of information disclosure for increasing user satisfaction from a shared ride. We modeled our environment as a signaling game and analyzed the perfect Bayesian equilibria for three agents' classes: an honest agent model, a no utility for lying model, and a penalized false information model. We showed that in the honest agent model and in the no utility for lying model, the agent must reveal all the information regarding the possible alternatives to the passenger. However, in the penalized false information model, there are two types of equilibria, one in which she is truthful (but must keep silent sometimes), and the other, in which the agent provides false information. The latter equilibrium type includes equilibria that seem unreasonable. Therefore, we proposed a novel criterion to filter out such equilibria. After filtering out the unreasonable equilibria, we can conclude from the theoretical analysis that in all three agent models, the agent should never provide any false information. Table 3 summarizes the properties of each agent model and the solution concepts that we use.

In future work, we intend to extend our theoretical analysis to additional domains for demonstrating the usefulness of the credible belief criterion. In addition, we would like to gather data of humans interacting with each other in the ridesharing scenario described in this paper and according to each of the agent models studied. It will be interesting to investigate (for each agent model) whether humans play according to the PBE strategies.

REFERENCES

[1] *2018 Revision of World Urbanization Prospects*, United Nations, New York, NY, USA, 2018.
[2] R. Molla. (2018). *Americans Seem to Like Ride-Sharing Services Like Uber and Lyft*. [Online]. Available: https://www.vox.com/2018/6/24/17493338/ride-sharing-services-uber-lyft-how-many-people-use
[3] J. Koebler. (2016). *Why Everyone Hates UberPOOL?* [Online]. Available: https://motherboard.vice.com/en_us/article/4xaa5d/why-drivers-and-riders-hate-uberpool-and-lyft-line
[4] H. Campbell. (2017). *Seven Reasons Why Rideshare Drivers Hate UberPOOL & Lyft Line*. [Online]. Available: https://maximumridesharingprofits.com/7-reasons-rideshare-drivers-hate-uberpool-lyft-line/
[5] B. J. Fogg, "Persuasive technology: Using computers to change what we think and do," *Ubiquity*, vol. 2002, p. 2, Dec. 2002.
[6] H. Singh, "The importance of customer satisfaction in relation to customer loyalty and retention," *Acad. Marketing Sci.*, vol. 60, nos. 193–225, p. 46, 2006.
[7] A. M. Spence, *Market Signaling: Informational Transfer in Hiring and Related Screening Processes*, vol. 143. Cambridge, MA, USA: Harvard Univ. Press, 1974.
[8] D. Fudenberg and J. Tirole, "Perfect Bayesian equilibrium and sequential equilibrium," *J. Econ. Theory*, vol. 53, no. 2, pp. 236–260, 1991.
[9] S. N. Parragh, K. F. Doerner, and R. F. Hartl, "A survey on pickup and delivery problems: Part I: Transportation between customers and depot," *J. Betriebswirtschaft*, vol. 58, no. 1, pp. 21–51, Apr. 2008.
[10] S. N. Parragh, K. F. Doerner, and R. F. Hartl, "A survey on pickup and delivery problems. Part II: Transportation between pickup and delivery locations," *J. Betriebswirtschaft*, vol. 58, no. 2, pp. 81–117, 2008.
[11] H. N. Psaraftis, M. Wen, and C. A. Kontovas, "Dynamic vehicle routing problems: Three decades and counting," *Networks*, vol. 67, no. 1, pp. 3–31, 2016.
[12] Y. Molenbruch, K. Braekers, and A. Caris, "Typology and literature review for dial-a-ride problems," *Ann. Oper. Res.*, vol. 259, nos. 1–2, pp. 295–325, Dec. 2017.
[13] Y. Lin, W. Li, F. Qiu, and H. Xu, "Research on optimization of vehicle routing problem for ride-sharing taxi," *Proc. Social Behav. Sci.*, vol. 43, pp. 494–502, Jan. 2012.
[14] C. Levinger, N. Hazon, and A. Azaria, "Human satisfaction as the ultimate goal in ridesharing," *Future Gener. Comput. Syst.*, vol. 112, pp. 176–184, Nov. 2020.
[15] S. Schleibaum and J. P. Müller, "Human-centric ridesharing on large scale by explaining AI-generated assignments," in *Proc. 6th EAI Int. Conf. Smart Objects Technol. Social Good*, Sep. 2020, pp. 222–225.
[16] F. Bistaffa, A. Farinelli, G. Chalkiadakis, and S. D. Ramchurn, "A cooperative game-theoretic approach to the social ridesharing problem," *Artif. Intell.*, vol. 246, pp. 86–117, May 2017.
[17] A. Azaria, Z. Rabinovich, S. Kraus, C. V. Goldman, and O. Tsimhoni, "Giving advice to people in path selection problems," in *Proc. Int. Conf. Auto. Agents Multiagent Syst.*, 2012, pp. 459–466.
[18] A. Azaria, Z. Rabinovich, S. Kraus, C. V. Goldman, and Y. Gal, "Strategic advice provision in repeated human-agent interactions," in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012, pp. 1522–1528.
[19] A. Azaria, Z. Rabinovich, C. V. Goldman, and S. Kraus, "Strategic information disclosure to people with multiple alternatives," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 4, p. 64, 2015.
[20] D. Shi, C. Bu, and H. Xue, "Deterrence effects of disclosure: The impact of environmental information disclosure on emission reduction of firms," *Energy Econ.*, vol. 104, Dec. 2021, Art. no. 105680.
[21] J. Schleich and S. Alsheimer, "How much are individuals willing to pay to offset their carbon footprint? The role of information disclosure and social norms," Sustain. Innov., Karlsruhe, Germany, Tech. Rep. S10/2022, 2022. [Online]. Available: http://hdl.handle.net/10419/264192
[22] M. Bilgic and R. J. Mooney, "Explaining recommendations: Satisfaction vs. promotion," in *Proc. Beyond Pers. Workshop, IUI*, 2005, pp. 13–18.
[23] S. J. Grossman, "The informational role of warranties and private disclosure about product quality," *J. Law Econ.*, vol. 24, no. 3, pp. 461–483, 1981.
[24] I.-K. Cho and D. M. Kreps, "Signaling games and stable equilibria," *Quart. J. Econ.*, vol. 102, no. 2, pp. 179–221, 1987.
[25] I.-K. Cho, "A refinement of sequential equilibrium," *Econometrica, J. Econ. Soc.*, vol. 55, no. 6, pp. 1367–1389, 1987.
[26] J. S. Banks and J. Sobel, "Equilibrium selection in signaling games," *Econometrica, J. Econ. Soc.*, vol. 55, no. 3, pp. 647–661, 1987.
[27] T. H. Noe, "Capital structure and signaling game equilibria," *Rev. Financial Stud.*, vol. 1, no. 4, pp. 331–355, Oct. 1988.
[28] J. R. Rogers, "Information and judicial review: A signaling game of legislative-judicial interaction," *Amer. J. Political Sci.*, vol. 45, no. 1, pp. 84–99, 2001.
[29] A. Bangerter, N. Roulin, and C. J. König, "Personnel selection as a signaling game," *J. Appl. Psychol.*, vol. 97, no. 4, pp. 719–738, 2012.
[30] X. Gao and Y.-F. Zhu, "DDoS defense mechanism analysis based on signaling game model," in *Proc. 5th Int. Conf. Intell. Hum.-Mach. Syst. Cybern.*, vol. 1, Aug. 2013, pp. 414–417.
[31] M. Estiri and A. Khademzadeh, "A theoretical signaling game model for intrusion detection in wireless sensor networks," in *Proc. 14th Int. Telecommun. Netw. Strategy Planning Symp. (NETWORKS)*, Sep. 2010, pp. 1–6.
[32] D. Zar, N. Hazon, and A. Azaria, "Explaining ridesharing: Selection of explanations for increasing user satisfaction," in *Proc. Eur. Conf. Multi-Agent Syst.* Berlin, Germany: Springer, 2021, pp. 89–107.

DAVID ZAR received the B.Sc. degree in computer science from Ariel University, Israel, in 2016, where he is currently pursuing the M.Sc. degree.

NOAM HAZON (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer science from Bar Ilan University, Israel, in 2003, 2005, and 2010, respectively. He was a Postdoctoral Fellow with CMU, Pittsburgh, PA, USA. During his M.Sc. degree, he was a Software Engineer in applied materials in Israel. He is currently a Senior Lecturer with Ariel University. He won an ISF Grant, in 2014, a Planning Grant from Volkswagen Foundation, in 2018, and a grant from the Ministry of Science, Technology and Space, Israel, in 2019. He has coauthored more than 50 articles. His research interests include computational social choice, human–computer interactions, coalitional game theory, stochastic search, influence maximization, and multirobot coverage. He serves as a senior program committee member and a program committee member for leading AI conferences.

AMOS AZARIA received the B.A. degree in computer science from the Technion Institute of Technology, Haifa, Israel, in 2004, and the Ph.D. degree from Bar Ilan University, Ramat Gan, Israel, in 2015. He was a Postdoctoral Fellow with CMU, Pittsburgh, PA, USA. After completing his bachelor's degree, he spent several years in the industry, some of which included working with Microsoft Research and Development, Haifa. He is currently an Associate Professor with Ariel University, Israel. He has coauthored more than 70 articles. His research interests include human–agent interaction, deep learning, human-aided machine learning, and natural language processing. He was a member of the Winning Team of the DARPA SMISC Competition, in 2015, on bot detection. He received the Victor Lesser Distinguished Dissertation Award, in 2015.