ARIEL UNIVERSITY

MASTER THESIS

# Explaining Ridesharing

*Author:*
David ZAR

*Supervisors:*
Amos AZARIA, Noam HAZON

*A thesis proposal submitted in partial fulfillment of the requirements for the degree of Masterof Sciences*

Department of Computer Science

July 13, 2022

# *Acknowledgements*

Ariel University

# *Abstract*

Faculty of Natural Sciences
Department of Computer Science

Master

**Explaining Ridesharing**

by David ZAR

Transportation services play a crucial part in the development of modern smart cities. In particular, on-demand ridesharing services, which group together passengers with similar itineraries, are already operating in several metropolitan areas. These services can be of significant social and environmental benefit, by reducing travel costs, road congestion and $CO_2$ emissions.

Unfortunately, despite their advantages, not many people opt to use these ridesharing services. We believe that increasing the user satisfaction from the service will cause more people to utilize it, which, in turn, will improve the quality of the service, such as the waiting time, cost, travel time, and service availability. One possible way for increasing user satisfaction is by providing appropriate explanations comparing the alternative modes of transportation, such as a private taxi ride and public transportation. For example, a passenger may be more satisfied from a shared-ride if she is told that a private taxi ride would have cost her $50\%$ more. Therefore, the problem is to develop an agent that provides explanations that will increase the user satisfaction.

We first model our environment as a signaling game and analyze 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 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 equilibira, 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.

In the second part of this thesis, we develop a machine learning based agent that, when given a shared-ride along with its possible alternatives, selects the explanations that are most likely to increase user satisfaction. Using feedback from humans, we show that our machine learning based agent outperforms the rational honest agent and an agent that randomly chooses explanations, in terms of user satisfaction.

# Contents

# Chapter 1

# Introduction

## 1.1 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 [41]. 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 [40], their ridesharing services, which group together multiple passengers (UberPool and Lyft-Line), suffer from low usage [33, 16].

In this thesis 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 [26], 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, [53]). Moreover, if people will be satisfied and use the the service more often it will improve the quality of the service, such as the waiting time, cost, travel time, and service availability, which in turn further increase the user satisfaction.

One possible way for increasing user satisfaction is by providing appropriate explanations [14], during the shared ride or immediately after the passenger has completed it. Indeed, in recent years there is a growing body of literature that deals with explaining decisions made by AI systems [30]. In our ridesharing scenario, a typical approach would attempt to explain the entire assignment of all passengers to all vehicles. Clearly, a passenger is not likely to be interested in such an explanation, since she is not interested in the assignment of other passengers to other vehicles. A passenger is likely to only be interested with her own current shared-ride when compared to other alternative modes of transportation, such as a private taxi ride or public transportation.

Comparing the shared-ride to other modes of transportation may provide many different possible explanations. For example, consider a shared-ride that takes $20$ minutes and costs $\$10$. The passenger could have taken a private taxi that would have cost $\$20$. Alternatively, the passenger could have used public transportation, and such a ride would have taken $30$ minutes. A passenger is not likely to be aware of the exact costs and riding times of the other alternatives, but she may have some

estimations. The agent, on the other hand, has access to many sources of information, and it can thus provide the exact values as explanations. The challenge is to design an agent that provides the appropriate explanation in any given scenario.

We first model our environment as a signaling game [54], 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 use the standard solution concept of Perfect Bayesian Equilibrium (PBE) [27], and analyze three agents' models. In the 'honest agent' (HA) model, the agent is not allowed reporting 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.

Interacting with humans and satisfying their expectations is a very complex task. Research into humans' behavior has found that people often deviate from what is thought to be the rational behavior, since they are affected by a variety of (sometimes conflicting) factors: a lack of knowledge of one's own preferences, framing effects, the interplay between emotion and cognition, future discounting, anchoring and many other effects [56, 38, 4, 15]. Therefore, algorithmic approaches that use a pure theoretically analytic objective often perform poorly with real humans [49, 5, 42]. In addition, attempting to provide false information raises ethical concerns and may violate regulations. We thus concentrate on the honest agent model and develop an Automatic eXplainer for Increasing Satisfaction (AXIS) agent, that when given a shared-ride along with its possible alternatives selects the explanations that are most likely to increase user satisfaction.

For example, consider again the setting in which a shared-ride takes 20 minutes and costs \$10. The passenger could have taken a private taxi that would have taken 15 minutes, but would have cost \$20. Alternatively, the passenger could have used public transportation. Such a ride would have taken 30 minutes, but would have cost only \$5. A *human* passenger may be more satisfied from the shared-ride if she

is told that a private taxi would have cost her $100\%$ more. Another reasonable explanation is that a public transportation would have taken her $10$ minutes longer. It may be even better to provide both explanations. However, providing an explanation that public transportation would have cost $50\%$ less than the shared-ride is less likely to increase her satisfaction. Indeed, finding the most appropriate explanation depends on the specific parameters of the scenario. For example, if public transportation still costs $\$5$ but the shared ride costs only $\$6$, providing an explanation that public transportation would have cost only $\$1$ less than the shared-ride may now become an appropriate explanation.

For developing the AXIS agent, we utilize the following approach. We collect data from human subjects on which explanations they believe are most suitable for different scenarios. AXIS then uses a neural network to generalize this data in order to provide appropriate explanations for any given scenario. Using feedback from humans, we show that AXIS outperforms the PBE honest agent and an agent that randomly chooses explanations. That is, human subjects that were faced with shared-ride scenarios, were more satisfied from the ride given the explanations selected by AXIS, than by the same ride when shown all explanations and when the explanations were randomly selected.

The contributions of this thesis are fourfold:

- The thesis introduces the problem of automatic selection of explanations in the ridesharing domain, for increasing user satisfaction. The set of explanations consists of alternative modes of transportation.
- We model the explanation selection problem as a signaling game and determine the unique set of Perfect Bayesian Equilibria (PBE) for three different agent model.
- We introduce the credible belief criterion, which filters unreasonable PBEs.
- We develop the AXIS agent, which learns from how people choose appropriate explanations, and experimentally show that it outperforms the PBE honest agent and an agent that randomly chooses explanations, in terms of user satisfaction.

# Chapter 2

# Related Works

## 2.1 Related Work

Most work on ridesharing has focused on the assignment of passengers to vehicles. See the comprehensive surveys by Parragh et al. [46, 47], and a recent survey by Psaraftis et al. [50]. 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 [21]. 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. [39], 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 [37, 35, 52]. Similar to these works we are interested in the satisfaction of the passenger, but instead of developing assignment algorithms (e.g., [11]), we focus on the role of information disclosure as a mean to improve user satisfaction.

In our work we build an agent that attempts to influence the attitude of the user towards ridesharing. Our agent is thus a form of persuasive technology [44]. Persuasion of humans by computers or technology has raised great interest in the literature. In his book [26], Fogg surveyed many technologies to be successful. One example of such a persuasion technology (pg. 50) is a bicycle connected to a TV; as one pedals at a higher rate, the image on the TV becomes clearer, encouraging humans to exercise at higher rates. Another example is the Banana-Rama slot machine, which has characters that celebrate every time the gambler wins. Overall, Fogg describes 40 persuasive strategies. Other social scientists proposed various classes of persuasive strategies: Kellermann and Tim provided over 100 groups [32], and Cialdini proposed six principles of influence [20]. More specifically, Anagnostopoulou et al. [3] survey persuasive technologies for sustainable mobility, some of which consider ridesharing. The methods mentioned by Anagnostopoulou et al. include several persuasive strategies such as self-monitoring, challenges & goal setting, social comparison, gamification, tailoring, suggestions, and rewards. Overall, unlike most of the works on persuasive technology, our approach is to selectively disclose information, available to the agent, regarding alternative options. This information aims at increasing the user satisfaction from her action, in order to change her attitude towards the service.

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. [7, 6, 5] develop agents that provide information or advice to a human user in order to convince her to take a certain route. Bilgic and Mooney [10] 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 [29] 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 minimal 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.

Signaling games are used to model problems in several domains. For example, Noe [43] models financial decisions of a firm (whether to use equity financing or debt financing) as a signaling game. Bangerter et al. [8] model the job market using signaling games, and analyze relationships between applicants and organizations, among applicants, and among organizations. Rogers [51] 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 [27, 28, 24]. We also consider a refinement of the PBE, the intuitive criterion introduced by Cho and Kreps [19], 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 [9], 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 [18] the forward induction equilibrium, which is another refinement of the intuitive criterion. In this work, we encounter PBEs that seems 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.

Explainable AI (XAI) is another domain related to our work [22, 30, 17]. In a typical XAI setting, the goal is to explain the output of the AI system to a human. This explanation is important for allowing the human to trust the system, better understand, and to allow transparency of the system's output [1]. Other XAI systems are designed to provide explanations, comprehensible by humans, for legal or ethical reasons [23]. For example, an AI system for the medical domain might be required to explain its choice for recommending the prescription of a specific drug [31]. Despite the fact that our agent is required to provide explanations to a human, our work does not belong to the XAI settings. In our work the explanations do not attempt to explain the output of the system to a passenger but to provide additional information that is likely to increase the user's satisfaction from the system. Therefore, our work can be seen as one of the first instances of x-MASE [34], explainable systems for multi-agent environments.

# Chapter 3

# Theoretical Model

## 3.1 Equilibrium-Based Agents

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 $\chi$, 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 $0$.

The game is composed of three players: nature, player 1 (agent) and player 2 (passenger). 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 $0 \leq p \leq 1$. The passenger observes the agent's action and plays an action $a_2 \in A_2 = [min, max]$. We consider several models for our environment.

### 3.1.1 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 $\varphi$ (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 passenger observes the agent's action and her action, denoted $a_2$, is any number in the range $[min, max]$. The passenger's action essentially means setting her estimate about the price of the alternative. In our setting the agent would like the passenger to think that the price of the alternative is as high as possible, while the passenger would like to know the real price. Therefore, we set the utility for the agent to $a_2$ and the utility of the passenger to $-(a_2 - x)^2$. Note that we did not define the utility of the passenger 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 passenger knows that $a_1$ is nature's choice. Thus, a rational passenger would play $a_2 = a_1$. On the other hand, if the agent plays $a_1 = \varphi$ then the passenger would have some belief about

the real price, which can be the original distribution of nature, or any other distribution. Clearly, the passenger'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 3.1.1 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., $\varphi$, $min$ or any mixture of $\varphi$ 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 $\sigma_2$ and $x$, deviating from $\sigma_1$ 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 $\sigma_2$ does not increase player 2's expected utility according to her belief.*
3. *$\mu_2$ is a consistent belief. That is, $\mu_2$ is a distribution over $x$ given $a_1$, which is consistent with $\sigma_1$ (following Bayes' rule, where appropriate).*

**Theorem 3.1.1.** *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: $\sigma_1$ 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. $\sigma_2$ is a best response strategy, since it is the expected value of the belief

$\mu_2$, and thus it is a best response according to Observation 1. Finally, $\mu_2$ is consistent: If $a_1 = \varphi$ and according to $\sigma_1$ player 1 plays $\varphi$ 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 $\varphi$ or $x$; both options give player 1 a utility of $c$, and thus she may use any strategy. That is,

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

Given this strategy, we need to apply Bayes' rule to derive $\mu_2(x|a_1 = \varphi)$. By $\sigma_1$, it is possible that $a_1 = \varphi$ only if $x \leq c$. That is, $\mu_2(x > c|a_1 = \varphi) = 0$ and $\mu_2(x \leq c|a_1 = \varphi) = 1$. 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, $c' = E_{X \sim \mu_2(x|a_1=\varphi)}[X] \leq E[X|X \leq c]$ since player 1 plays $\varphi$ only when $x < c$ and possibly also when $x = c$. In addition, $E[X|X \leq c] < c$, 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 $\sigma_2$ 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.

□

### 3.1.2 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 $\varphi$ 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 3.1.1) holds for the current model as well. However, in the current model there are additional perfect Bayesian equilibria. For example,

- $\sigma_1(x) = \varphi$
- $\sigma_2(a_1) = \begin{cases} min: & a_1 \neq \varphi \\ E[X]: & a_1 = \varphi \end{cases}$

- $\mu_2(x = min|a_1 \neq \varphi) = 1$ and $\mu_2(x|a_1 = \varphi) = Pr(X = x)$.

Note that the belief $\mu_2$ is consistent, since the agent plays $a_1 \neq \varphi$ 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 with the lowest possible utility for her (zero). We thus use the intuitive criterion [19] 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 $\sigma_1$, 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 $u_1(x, a_1', w) \geq u_1(x, \sigma_1(x), z)$.*

We note that in our NUFL model, if the agent's strategy for a given $x$ is either $\varphi$ 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, $\sigma_1$, 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 3.1.2.** *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:*

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

- $\mu_2(x = a_1 | a_1 \neq \varphi) = 1$ *and* $\mu_2(x = min | a_1 = \varphi) = 1$.

*Proof.* ($\Leftarrow$) As shown in Theorem 3.1.1 such a tuple is a PBE. It also satisfies the intuitive criterion: the only actions that can be off-the-path are $\varphi$ 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 $a_1 \neq \varphi$, 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 3.1.1. □

### 3.1.3 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 $\varphi$), and one in which the agent lies. Specifically, the strategy of a truthful agent is to play $\varphi$ 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 $\varphi$. 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 $\sigma_2$. However, in some cases the agent will say the truth. Let $E_F$ be the maximum value of $\sigma_2$. 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 $\varphi$ 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, $\sigma_1$ 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, $\sigma_1$ is not a strategy of an equilibrium since the agent will benefit from deviating from it and playing $\varphi$ given $x$. □

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

**Theorem 3.1.3.** *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 $Y_\varphi$ be also such a variable. If $S$ is not empty, then $E[X \mid X \in S] = f \cdot E_F$.*

   - $\sigma_1(x) = \begin{cases} q_i : & x \in F_i \text{ for some } i \\ x : & x \in T \\ \varphi : & x \in S \end{cases}$
   - $\sigma_2(a_1) = \begin{cases} a_1 : & a_1 \in T \setminus \mathcal{Q} \\ f \cdot E_F : & a_1 = \varphi \text{ and } S \neq \emptyset \\ E_F : & a_1 \in \mathcal{Q} \\ E[Y_{a_1}] : & otherwise \end{cases}$

- $\mu_2(x = a_1 \mid a_1 \in T \setminus \mathcal{Q}) = 1$
  $\mu_2(x \mid a_1 = q_i) =$
  $\begin{cases} \frac{Pr(X=x)}{Pr(X \in F_i \cup (\{q_i\} \cap T))} : & x \in F_i \cup (\{q_i\} \cap T) \\ 0 : & \textit{otherwise} \end{cases}$
  $\mu_2(x \mid a_1 \notin T \cup \mathcal{Q}$ *or* $(a_1 = \varphi$ *and* $S = \emptyset)) = Pr(Y_{a_1} = x)$.
  *If* $S \neq \emptyset$ *then*
  $\mu_2(x \mid a_1 = \varphi) = \begin{cases} \frac{Pr(X=x)}{Pr(\sigma_1(X)=\varphi)} : & x \in S \\ 0 : & x \notin S. \end{cases}$

*Proof.* We begin with the truthful agent case.
($\Leftarrow$) Let $(\sigma_1, \sigma_2, \mu_2)$ be a tuple of strategy and belief that satisfy the conditions of the truthful agent. $\mu_2$ satisfies Bayes' rule:

- If $a_1 \notin S \cup \{\varphi\}$, according to $\sigma_1$, $a_1 = x$; therefore, by Bayes' rule: $\mu_2(x = a_1 \mid a_1 \notin S \cup \{\varphi\}) = 1$.
- If $a_1 = \varphi$, according to $\sigma_1$ and Bayes' rule:
  $\mu_2(x \mid a_1 = \varphi) = \frac{Pr(X=x)Pr(a_1=\varphi|x)}{Pr(\sigma_1(X)=\varphi)} = \begin{cases} \frac{Pr(X=x)}{Pr(\sigma_1(X)=\varphi)} : & x \in S \\ 0 : & x \notin S. \end{cases}$
- 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 $\sigma_1$ and $\mu_2$, the strategy of the user, $\sigma_2$, is a best response, since it is the expectation over the user's belief regarding $x$ (according to Observation 1). Finally, given $\sigma_2$ and $\mu_2$, the agent does not have an incentive to deviate from $\sigma_1$:

- 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 $\sigma_2$ 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 $\varphi$ 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 $\sigma_2$ 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 satisfy the conditions of the non-truthful agent. $\mu_2$ satisfies Bayes' rule:

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

Given $\sigma_1$ and $\mu_2$, the strategy of the user, $\sigma_2$, is a best response, since it is the expectation over the user's belief regarding $x$. Finally, given $\sigma_2$ and $\mu_2$, the agent does not have an incentive to deviate from $\sigma_1$:

- 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 $\varphi$ 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 $\varphi$ 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 $\varphi$ instead of lying if $\sigma_2(\varphi) > f \cdot E_F$. Following the above arguments regarding $\sigma_1$ and since $\mu_2$ 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 $\sigma_1$ 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 $\varphi$ 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 $\mu_2$ and $\sigma_2$ defined above, respectively.

□

The PBEs in which the agent 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 $\varphi$ 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 3.1.3).

**Theorem 3.1.4.** *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 3.1.3 (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. $\forall x_1, x_2 \in S \backslash \{a_1\}, Pr(X = x_2) \cdot \mu_2(X = x_1 \mid a_1) = Pr(X = x_1) \cdot \mu_2(X = x_2 \mid a_1)$.
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 3.1.3 (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. $\forall x \in F \cup S \cup T \backslash \{a_1\}, Pr(X = x) \cdot \mu_2(X = a_1 \mid a_1) \geq Pr(X = a_1) \cdot \mu_2(X = x \mid a_1)$.
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 3.1.3 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 3.1.3. 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 3.1.3.

Next, we show that any PBE that takes the form of case (1) in Theorem 3.1.3 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 3.1.3 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 3.1.3 (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 3.1.3 (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:

- $Types = \{sp, so, an\}$ 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 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:

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

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 $l(an, sp_{ed}) = 3$.

# Chapter 4

# Experiments Approach

## 4.1 The AXIS Agent

The analysis in the previous section is theoretical in nature. However, attempting to provide false information raises ethical concerns and may violate regulations. We thus concentrate on the honest agent model and note that the provided analysis can be applied independently to any alternative mode of transportation and to any type of price (e.g. travel-time or cost). Thus, the PBE honest agent must provide all of the possible explanations. Unfortunately, several studies have shown that algorithmic approaches that use a pure theoretically analytic objective often perform poorly with real humans. We conjecture that an agent that selects a subset of explanations for a given scenario will perform better than the PBE honest agent. In this section, we introduce our Automatic eXplainer for Increasing Satisfaction (AXIS) agent. The AXIS agent has a set of possible explanations, and the agent needs to choose the most appropriate explanations for each scenario. Note that we do not limit the number of explanations to present for each scenario, and thus AXIS needs also to choose how many explanations to present. AXIS was built in $3$ stages.

First, an initial set of possible explanations needs to be defined. We thus consider the following possible classes of factors of an explanation. Each explanation is a combination of one factor from each class:

1. Mode of alternative transportation: a private taxi ride or public transportation.
2. Comparison criterion: time or cost.
3. Visualization of the difference: absolute or relative difference.
4. Anchoring: the shared ride or the alternative mode of transportation perspective.

For example, a possible explanation would consist of a private taxi for class $1$, cost for class $2$, relative for class $3$, and an alternative mode of transportation perspective for class $4$. That is, the explanation would be "a private taxi would have cost $50\%$ more than a shared ride". Another possible explanation would consist of public transportation for class $1$, time for class $2$, absolute for class $3$, and a shared ride perspective for class $4$. That is, the explanation would be "the shared ride saved $10$ minutes over public transportation". Overall, there are $2^4 = 16$ possible combinations. In addition, we added an explanation regarding the saving of $CO_2$ emission of the shared ride, so there will be an alternative explanation for the case where the other options are not reasonable. Note that the first two classes determine which information is given to the passenger, while the later two classes determine how the information is presented. We denote each possible combination of choosing form

the first two classes as an *information setting*. We denote each possible combination of choosing form the latter two classes as a *presentation setting*.

Presenting all 17 possible explanations with the additional option of "none of the above" requires a lot of effort from the human subjects to choose the most appropriate option for each scenario. Thus, in the second stage we collected data from human subjects regarding the most appropriate explanations, in order to build a limited subset of explanations. Recall that there are 4 possible information settings and 4 possible presentation settings. We selected for each information setting the corresponding presentation setting that was chosen (in total) by the largest number of people. We also selected the second most chosen presentation setting for the information setting that was chosen by the largest number of people. Adding the explanation regarding the $CO_2$ emissions we ended up with 6 possible explanations.

In the final stage we collected data from people again, but we presented only the 6 explanations to choose from. This data was used by AXIS to learn which explanations are appropriate for each scenario. AXIS receives the following 7 features as an input: the cost and time of the shared ride, the differences between the cost and time of the shared ride and the alternatives (i.e., the private ride and the public transportation), and the amount of $CO_2$ emission saved when compared to a private ride. AXIS uses a neural network with two hidden layers, one with 8 neurons and the other one with 7 neurons, and the logistic activation function (implemented using Scikit-learn [48]). The number of neurons and hidden layers was determined based on the performance of the network. AXIS used 10% of the input as a validation set (used for early stopping) and 40% as the test set. AXIS predicts which explanations were selected by the humans (and which explanations were not selected) for any given scenario.

## 4.2 Experimental Design

In this section we describe the design of our experiments. Since AXIS generates explanations for a given assignment of passengers to vehicles, we need to generate assignments as an input to AXIS. To generate the assignments, we first need a data-set of ride requests.

To generate the ride requests we use the New York city taxi trip data-set [1], which was also used by other works that evaluate ridesharing algorithms (see for example, [36, 12]). We use the data-set from 2016, since it contains the exact GPS locations for every ride.

We note that the data-set contains requests for taxi rides, but it does not contain data regarding shared-rides. We thus need to generate assignments of passengers to taxis, based on the requests from the data-set. Now, if the assignments are randomly generated, it may be hard to provide reasonable explanations, and thus the evaluation of AXIS in these settings is problematic. We thus concentrate on requests that depart from a single origin but have different destinations, since a brute force algorithm can find the optimal assignment of passengers to taxis in this setting.

We use the following brute force assignment algorithm. The algorithm receives 12 passengers and outputs the assignment of each passenger to a vehicle that minimizes the overall travel distance. We assume that every vehicle can hold up-to four passengers. The brute force assignment algorithm recursively considers all options to partition the group of 12 passengers to subsets of up to four passengers. We

[1] https://data.cityofnewyork.us/Transportation/2016-Green-Taxi-Trip-Data/hvrh-b6nb

note that there are $3,305,017$ such possible partitions. The algorithm then solves the Travel Salesman Problem (TSP) in each group, by exhaustive search, to find the cheapest assignment. Solving the TSP problem on 4 destinations (or less) is possible using exhaustive search since there are only $4! = 24$ combinations. The shortest path between each combination is solved using a shortest distance matrix between all locations. In order to compute this matrix we downloaded the graph that represents the area of New York from Open Street Map (using OSMnx [13]), and ran the Floyd-Warshall's algorithm.

We set the origin location to JFK Station, Sutphin Blvd-Archer Av, and the departing time to 11:00am. See Figure 4.1 where the green location is the origin, and the blue locations are the destinations.

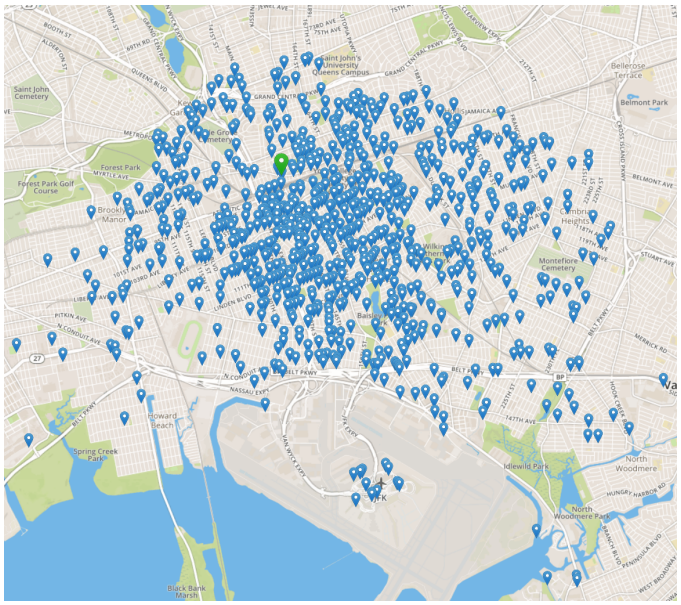


FIGURE 4.1: A map depicting the origin (in green) and destinations (in blue) of all rides considered.

In order to calculate the duration of the rides we use Google Maps (through Google Maps API). Specifically, the duration of the private taxi ride was obtained using "driving" mode, and the duration of the public transportation was obtained using "transit" mode. The duration of the shared-ride was set as the duration of the ride to the destination of the last passenger (using "driving" mode) with the destinations of the other passengers as way-points.

In order to calculate the cost of the private ride we use Taxi Fare Finder (through their API)[2]. The cost for public transportation was calculated by the number of buses required (as obtained through Google Maps API), multiplied by $\$2.5$ (the bus fare). The cost for the shared-ride was obtained from Taxi Fare Finder. Since this service does not support a ride with way-points, we obtained the cost of multiple taxi rides, but we included the base price only once. Note that this is the total cost of the shared-ride. The cost for a specific passenger was determined by the proportional sharing pricing function [25], which works as follows. Let $c_{p_i}$ be the cost of a private ride for passenger $i$, and let $total_s$ be the total cost of the shared ride. In addition, let $f = \frac{total_s}{\sum_i c_{p_i}}$. The cost for each passenger is thus $f \cdot c_{pi}$.

We ran 4 experiments in total. Two experiments were used to compose AXIS (see Section 4.1), and the third and fourth experiments compared the performance

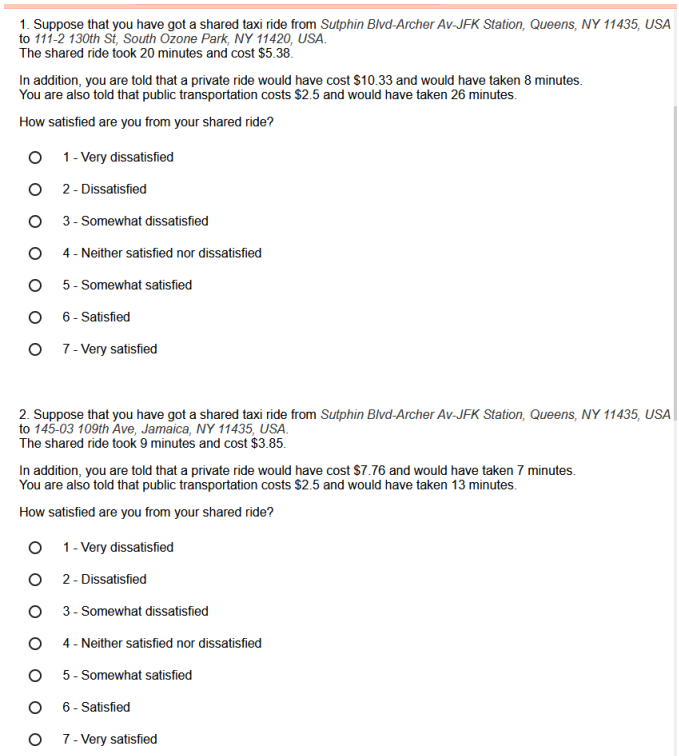
[2] https://www.taxifarefinder.com/

of AXIS with that of non-data-driven agents (see below). All experiments used the Mechanical Turk platform, a crowd-sourcing platform that is widely used for running experiments with human subjects [2, 45], and all human subjects agreed to participate in the experiment.

Unfortunately, since participation is anonymous and linked to monetary incentives, experiments on a crowd-sourcing platform can attract participants who do not fully engage in the requested tasks [55]. Therefore, the subjects were required to have at least $99\%$ acceptance rate and were required to have previously completed at least $500$ Mechanical Turk Tasks (HITs). In addition, we added an attention check question for each experiment, which can be found in the Appendix.

In the first two experiments, which were designed for AXIS to learn what people believe are good explanations, the subjects were given several scenarios for a shared-ride. The subjects were told that they are representatives of a ridesharing service, and that they need to select a set of explanations that they believe will increase the customer's satisfaction. Each scenario consists of a shared-ride with a given duration and cost.

In the third experiment we evaluate the performance of AXIS against the PBE honest agent. The subjects were given $2$ scenarios. Each scenario consists of a shared-ride with a given duration and cost and it also contains either the explanations that are chosen by AXIS or the information that the PBE honest agent provides: the cost and duration a private ride would take, and the cost and the duration that public transportation would have taken. The subjects were asked to rank their satisfaction from each ride on a scale from 1 to 7. See Figure 4.2 for a snapshot of a scenario shown to a worker on Mechanical Turk, along with the information provided by the PBE honest agent.


1. Suppose that you have got a shared taxi ride from *Sutphin Blvd-Archer Av-JFK Station, Queens, NY 11435, USA* to *111-2 130th St, South Ozone Park, NY 11420, USA*.
The shared ride took 20 minutes and cost $5.38.

In addition, you are told that a private ride would have cost $10.33 and would have taken 8 minutes.
You are also told that public transportation costs $2.5 and would have taken 26 minutes.

How satisfied are you from your shared ride?

- 1 - Very dissatisfied
- 2 - Dissatisfied
- 3 - Somewhat dissatisfied
- 4 - Neither satisfied nor dissatisfied
- 5 - Somewhat satisfied
- 6 - Satisfied
- 7 - Very satisfied

2. Suppose that you have got a shared taxi ride from *Sutphin Blvd-Archer Av-JFK Station, Queens, NY 11435, USA* to *145-03 109th Ave, Jamaica, NY 11435, USA*.
The shared ride took 9 minutes and cost $3.85.

In addition, you are told that a private ride would have cost $7.76 and would have taken 7 minutes.
You are also told that public transportation costs $2.5 and would have taken 13 minutes.

How satisfied are you from your shared ride?

- 1 - Very dissatisfied
- 2 - Dissatisfied
- 3 - Somewhat dissatisfied
- 4 - Neither satisfied nor dissatisfied
- 5 - Somewhat satisfied
- 6 - Satisfied
- 7 - Very satisfied

FIGURE 4.2: A snapshot of a scenario shown to a worker, along with the information provided by the PBE honest agent.

In the forth experiment we evaluate the performance of AXIS against a random

| | #1 | #2 | #3 | #4 | Total |
|---|---|---|---|---|---|
| Number of subjects | 343 | 180 | 156 | 274 | 953 |
| Scenarios per subject | 2 | 4 | 2 | 2 | - |
| Total scenarios | 686 | 720 | 312 | 548 | 3266 |

TABLE 4.1: Number of subjects and scenarios in each of the experiments.

| | #1 | #2 | #3 | #4 | Total |
|---|---|---|---|---|---|
| Male | 157 | 66 | 52 | 117 | 392 |
| Female | 183 | 109 | 104 | 153 | 549 |
| Other or refused | 3 | 5 | 0 | 4 | 12 |

TABLE 4.2: Gender distribution for each of the experiments.

baseline agent. The random explanations were chosen as follows: first, a number between $1$ and $4$ was uniformly sampled. This number determined how many explanations will be given by the random agent. This range was chosen since over $93\%$ of the subjects selected between $1$ and $4$ explanations in the second experiment. Recall that there are 4 classes of factors that define an explanation, where the fourth class is the anchoring perspective (see Section 4.1). The random agent sampled explanations uniformly, but it did not present two explanations that differ only by their anchoring perspective. The subjects were again given $2$ scenarios. Each scenario consists of a shared-ride with a given duration and cost, and it also contains either the explanations that are chosen by AXIS or the explanations selected by the random agent. The subjects were asked to rank their satisfaction from each ride. The exact wording of the instructions for the experiments can be found in the Appendix.

$953$ subjects participated in total, all from the USA. The number of subjects in each experiment and the number of scenarios appear in Table 4.1. Tables 4.2 and 4.3 include additional demographic information on the subjects in each of the experiments. The average age of the subjects was $39$.

## 4.3 Results

Recall that the first experiment was designed to select the most appropriate explanations (out of the initial $17$ possible explanations). The results of this experiment are depicted in Figure 4.3. The x-axis describes the possible explanations according to the $4$ classes. Specifically, the factor from the anchoring class is denoted by s-p or p-s; s-p means that the explanation is from the shared-ride perspective, while p-s means that it is from the alternative (private/public) mode of transportation. The

| | #1 | #2 | #3 | #4 | Total |
|---|---|---|---|---|---|
| High-school | 72 | 39 | 38 | 80 | 229 |
| Bachelor | 183 | 86 | 84 | 131 | 484 |
| Master | 60 | 29 | 37 | 46 | 172 |
| PhD | 15 | 2 | 0 | 10 | 27 |
| Trade-school | 8 | 4 | 5 | 10 | 27 |
| Refused or did not respond | 5 | 3 | 0 | 6 | 14 |

TABLE 4.3: Education level for each of the experiments.

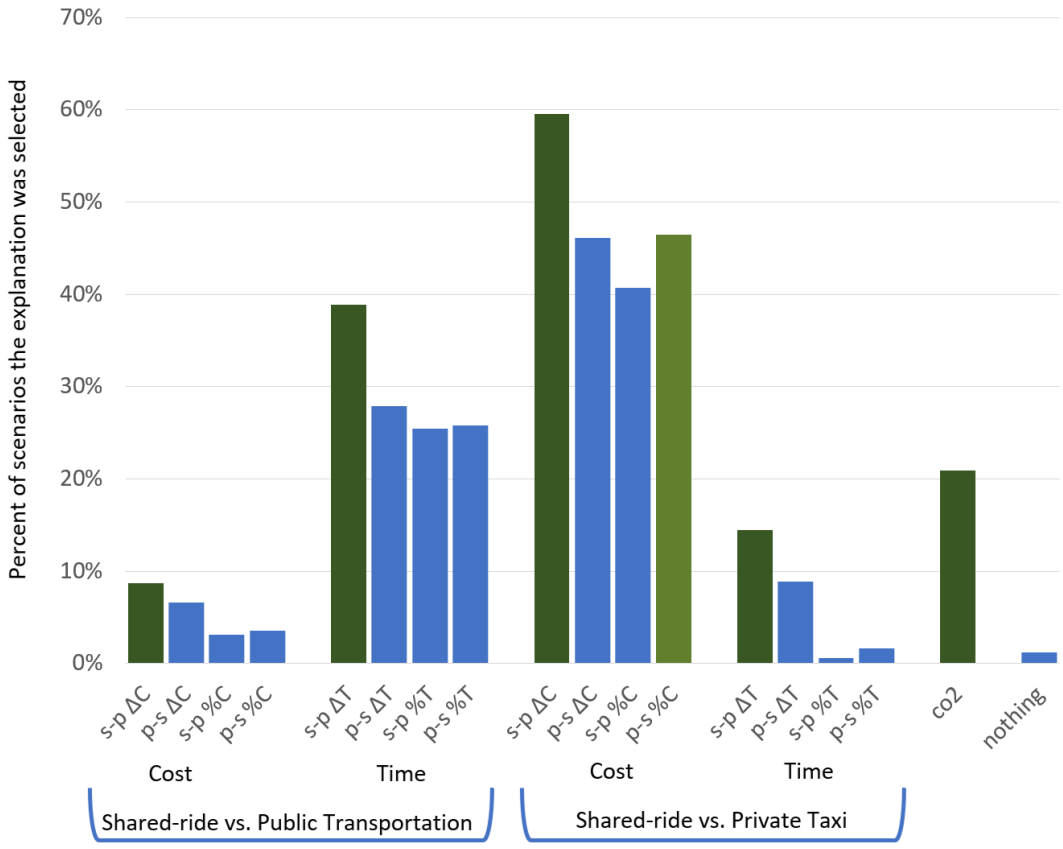



FIGURE 4.3: The percent of scenarios that every explanation was selected in the first experiment. The explanations marked in green were selected for the second experiment.

factor from the comparison criterion class is denoted by $\Delta$ or $\%$; $\Delta$ means that the explanation presents an absolute difference while $\%$ means that a relative difference is presented. We chose $6$ explanations for the next experiment, which are marked in green.

As depicted by Figure 4.3, the subjects chose explanations that compare the ride with a private taxi more often than those comparing the ride with public transportation. We believe that this is because from a human perspective a shared-ride resembles a private taxi more than public transportation. Furthermore, when comparing with a private taxi, the subjects preferred to compare the shared-ride with the *cost* of a private taxi, while when comparing to public transportation, the subjects preferred to compare it with the travel time. This is expected, since the travel time by a private taxi is less than the travel time by a shared ride, so comparing the travel time to a private taxi is less likely to increase user satisfaction. We also notice that with absolute difference the subjects preferred the shared ride perspective, while with relative difference the subjects preferred the alternative mode of transportation perspective. We conjecture that this is due to the higher percentages when using the alternative mode prospective. For example, if the shared ride saves $20\%$ of the cost when compared to a private ride, the subjects preferred the explanation that a private ride costs $25\%$ more.

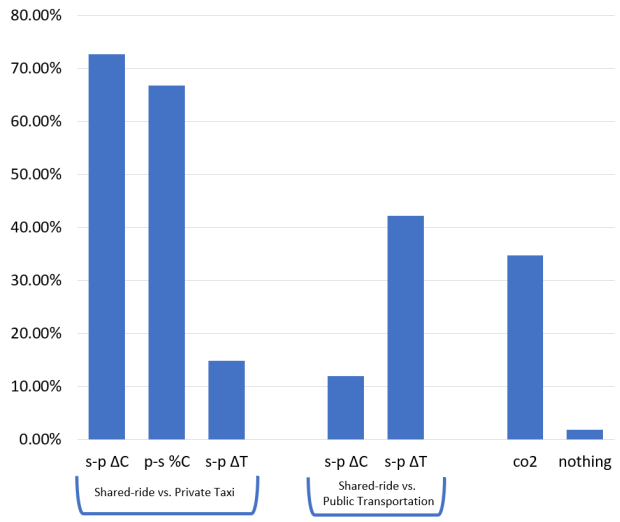



FIGURE 4.4: The percent of scenarios that every explanation was selected in the second experiment. The obtained data-set was used to train AXIS.

The second experiment was designed to collect data from humans on the most appropriate explanations (out of the $6$ chosen explanations) for each scenario. The results are depicted in Figure 4.4. This data was used to train AXIS. The accuracy of the neural network on the test-set is $74.9\%$. That is, the model correctly predicts whether to provide a given explanation in a given scenario in almost $75\%$ of the cases.

The third experiment was designed to evaluate AXIS against the PBE honest agent; the results are depicted in Figure 4.5. AXIS outperforms the PBE honest agent; the difference is statistically significant ($p < 10^{-5}$), using the student t-test. We note that achieving such a difference is non-trivial since the ride scenarios are identical and only differ by the information that is provided to the user.

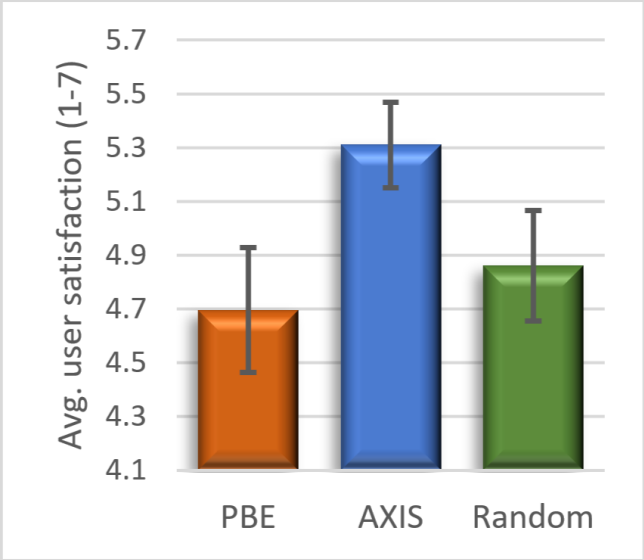



FIGURE 4.5: A comparison between the performance of AXIS, the PBE honest agent and the random agent. The bars indicate the $95\%$ confidence interval. AXIS significantly outperformed both baseline agents ($p < 0.001$).

The forth experiment was designed to evaluate AXIS against the random baseline agent; the results are depicted in Figure 4.5. AXIS outperforms the random agent; the difference is statistically significant ($p < 0.001$), using the student t-test. We note that AXIS and the random agent provided a similar number of explanations on average ($2.551$ and $2.51$, respectively). That is, AXIS performed well not because of the number of explanations it provided, but since it provided appropriate explanations for the given scenarios.

We conclude this section by showing an example of a ride scenario presented to some of the subjects, along with the information provided by the PBE honest agent, and the explanations selected by the random agent and by AXIS. In this scenario the subject is assumed to travel by a shared ride from JFK Station to 102-3 188th St, Jamaica, NY. The shared ride took $13$ minutes and cost $\$7.53$. The PBE honest agent provided the following information:

- "A private ride would have cost $\$13.83$ and would have taken $12$ minutes".
- "Public transportation costs $2.5 and would have taken 26 minutes".

The random agent provided the following explanations:

- "A private taxi would have cost $\$6.3$ more".
- "A ride by public transportation would have saved you only $\$5.03$".

Instead, AXIS selected the following explanations:

- "The shared ride had saved you $\$6.3$ over a private taxi".
- "A private taxi would have cost $83\%$ more".
- "The shared ride saved you $4$ minutes over public transportation".

Clearly, the explanations provided by AXIS seem much more compelling.

# Chapter 5

# Conclusions

In this thesis, we took a first step towards the development of agents that provide explanations in a multi-agent system with a goal of increasing user satisfaction. We first 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.

We then presented AXIS, an agent that, when given a shared-ride along with its possible alternatives, selects the explanations that are most likely to increase user satisfaction. We ran four experiments with humans. The first experiment was used to narrow the set of possible explanations, the second experiment collected data for the neural network to train on, the third experiment was used to evaluate the performance of AXIS against that of the PBE honest agent, and the fourth experiment was used to evaluate the performance of AXIS against that of an agent that randomly chooses explanations. We showed that AXIS outperforms the other agents in terms of user satisfaction.

In future work, we will consider natural language generation methods for generating explanations that are likely to increase user satisfaction. We also plan to extend the set of possible explanations, and to implement user modeling in order to provide explanations that are appropriate not only for a given scenario but also for a given specific user. We also intend to extend our theoretical analysis to additional domains for demonstrating the usefulness of the credible belief criterion.

## Appendix

### Attention Check Question

Which of the following claims do you agree with?

- A shared ride may take longer than a private ride.
- A shared ride is supposed to be more expensive than a private ride.
- The cost of public transportation is usually less than the cost of a private ride.

- In a private ride there are at least 3 passengers.
- Public transportation usually takes longer than a private ride.

### The Text for the First Two Experiments

In this survey we would like to learn which explanations increase the satisfaction from a ridesharing service. Suppose that you are a representative of a ridesharing service. This service assigns multiple passengers with different destinations to a shared taxi and divides the cost among them. Assume that the customer of your service has just completed the shared ride. Below are given few scenarios for a shared ride. For each of the scenarios you should choose one or more suitable explanation(s) that you believe will increase the customer's satisfaction.

### The Text for the Third and Fourth Experiments

In this survey we would like to evaluate your satisfaction from using shared taxi services. The following questions contain description of some shared rides. Please provide your rate of satisfaction from each ride on a scale from 1 to 7. Please read the details carefully and try to evaluate your satisfaction in each scenario as accurate as possible. Good luck!

# Bibliography


[1] Amina Adadi and Mohammed Berrada. "Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160.

[2] Ofra Amir, David G Rand, and Ya'akov Kobi Gal. "Economic games on the internet: The effect of $1 stakes". In: *PloS one* 7.2 (2012), e31461.

[3] Evangelia Anagnostopoulou, Efthimios Bothos, Babis Magoutas, Johann Schrammel, and Gregoris Mentzas. "Persuasive technologies for sustainable mobility: State of the art and emerging trends". In: *Sustainability* 10.7 (2018), p. 2128.

[4] Dan Ariely, George Loewenstein, and Drazen Prelec. ""Coherent arbitrariness": Stable demand curves without stable preferences". In: *The Quarterly Journal of Economics* 118.1 (2003), pp. 73–106.

[5] Amos Azaria, Zinovi Rabinovich, Claudia V Goldman, and Sarit Kraus. "Strategic information disclosure to people with multiple alternatives". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 5.4 (2015), p. 64.

[6] Amos Azaria, Zinovi Rabinovich, Sarit Kraus, Claudia V Goldman, and Ya'akov Gal. "Strategic advice provision in repeated human-agent interactions". In: *Twenty-Sixth AAAI Conference on Artificial Intelligence*. 2012.

[7] Amos Azaria, Zinovi Rabinovich, Sarit Kraus, Claudia V Goldman, and Omer Tsimhoni. "Giving advice to people in path selection problems". In: *AAMAS*. 2012, pp. 459–466.

[8] Adrian Bangerter, Nicolas Roulin, and Cornelius J König. "Personnel selection as a signaling game." In: *Journal of Applied Psychology* 97.4 (2012), p. 719.

[9] Jeffrey S Banks and Joel Sobel. "Equilibrium selection in signaling games". In: *Econometrica: Journal of the Econometric Society* (1987), pp. 647–661.

[10] Mustafa Bilgic and Raymond J Mooney. "Explaining recommendations: Satisfaction vs. promotion". In: *Beyond Personalization Workshop, IUI*. 2005, pp. 13–18.

[11] Filippo Bistaffa, Alessandro Farinelli, Georgios Chalkiadakis, and Sarvapali D Ramchurn. "A cooperative game-theoretic approach to the social ridesharing problem". In: *Artificial Intelligence* 246 (2017), pp. 86–117.

[12] Arpita Biswas, Ragavendran Gopalakrishnan, Theja Tulabandhula, Koyel Mukherjee, Asmita Metrewar, and Raja Subramaniam Thangaraj. "Profit optimization in commercial ridesharing". In: *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems. 2017, pp. 1481–1483.

[13] Geoff Boeing. "OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks". In: *Computers, Environment and Urban Systems* 65 (2017), pp. 126–139.

[14] Graham L Bradley and Beverley A Sparks. "Dealing with service failures: The use of explanations". In: *Journal of Travel & Tourism Marketing* 26.2 (2009), pp. 129–143.

[15] C. F. Camerer. "Behavioral Game Theory. Experiments in Strategic Interaction". In: Princeton University Press, 2003. Chap. 2, pp. 43–118.

[16] Harry Campbell. *Seven Reasons Why Rideshare Drivers Hate UberPOOL & Lyft Line*. https://maximumridesharingprofits.com/7-reasons-rideshare-drivers-hate-uberpool-lyft-line/. 2017.

[17] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. "ML Interpretability: A Survey on Methods and Metrics". In: *Electronics* 8.8 (2019), p. 832.

[18] In-Koo Cho. "A refinement of sequential equilibrium". In: *Econometrica: Journal of the Econometric Society* (1987), pp. 1367–1389.

[19] In-Koo Cho and David M Kreps. "Signaling games and stable equilibria". In: *The Quarterly Journal of Economics* 102.2 (1987), pp. 179–221.

[20] Robert B Cialdini. "Harnessing the science of persuasion". In: *Harvard business review* 79.9 (2001), pp. 72–81.

[21] Jean-François Cordeau and Gilbert Laporte. "A tabu search heuristic for the static multi-vehicle dial-a-ride problem". In: *Transportation Research Part B: Methodological* 37.6 (2003), pp. 579–594.

[22] Mark G. Core, H. Chad Lane, Michael van Lent, Dave Gomboc, Steve Solomon, and Milton Rosenberg. "Building Explainable Artificial Intelligence Systems". In: *Proceedings of the 18th Conference on Innovative Applications of Artificial Intelligence*. 2006, pp. 1766–1773.

[23] Derek Doran, Sarah Schulz, and Tarek R Besold. "What does explainable AI really mean? A new conceptualization of perspectives". In: *arXiv preprint arXiv:1710.00794* (2017).

[24] Mohsen Estiri and Ahmad Khademzadeh. "A theoretical signaling game model for intrusion detection in wireless sensor networks". In: *2010 14th International Telecommunications Network Strategy and Planning Symposium (NETWORKS)*. IEEE. 2010, pp. 1–6.

[25] PC Fishburn and HO Pollak. "Fixed-route cost allocation". In: *The American Mathematical Monthly* 90.6 (1983), pp. 366–378.

[26] Brian J Fogg. "Persuasive technology: using computers to change what we think and do". In: *Ubiquity* 2002.December (2002), p. 2.

[27] Drew Fudenberg and Jean Tirole. "Perfect Bayesian equilibrium and sequential equilibrium". In: *journal of Economic Theory* 53.2 (1991), pp. 236–260.

[28] Xiang Gao and Yue-Fei Zhu. "DDoS defense mechanism analysis based on signaling game model". In: *2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics*. Vol. 1. IEEE. 2013, pp. 414–417.

[29] Sanford J Grossman. "The informational role of warranties and private disclosure about product quality". In: *The Journal of Law and Economics* 24.3 (1981), pp. 461–483.

[30] David Gunning. "Explainable artificial intelligence (xai)". In: *Defense Advanced Research Projects Agency (DARPA), nd Web* 2 (2017).

[31] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. "What do we need to build explainable AI systems for the medical domain?" In: *arXiv preprint arXiv:1712.09923* (2017).

[32] Kathy Kellermann and Tim Cole. "Classifying compliance gaining messages: Taxonomic disorder and strategic confusion". In: *Communication Theory* 4.1 (1994), pp. 3–60.

[33] Jason Koebler. *Why Everyone Hates UberPOOL?* https://motherboard.vice.com/en_us/article/4xaa5d/why-drivers-and-riders-hate-uberpool-and-lyft-line. 2016.

[34] Sarit Kraus, Amos Azaria, Jelena Fiosina, Maike Greve, Noam Hazon, Lutz Kolbe, Tim-Benjamin Lembcke, Jörg P Müller, Sören Schleibaum, and Mark Vollrath. "AI for Explaining Decisions in Multi-Agent Environments". In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence*. 2019, pp. 13534–13538.

[35] Chaya Levinger, Noam Hazon, and Amos Azaria. "Human satisfaction as the ultimate goal in ridesharing". In: *Future Generation Computer Systems* (2020).

[36] Jane Lin, Sandeep Sasidharan, Shuo Ma, and Ouri Wolfson. "A model of multimodal ridesharing and its analysis". In: *2016 17th IEEE International Conference on Mobile Data Management (MDM)*. Vol. 1. IEEE. 2016, pp. 164–173.

[37] Yeqian Lin, Wenquan Li, Feng Qiu, and He Xu. "Research on optimization of vehicle routing problem for ride-sharing taxi". In: *Procedia-Social and Behavioral Sciences* 43 (2012), pp. 494–502.

[38] George Loewenstein. "Willpower: A Decision-theorist's Perspective". In: *Law and Philosophy* 19 (1 2000), pp. 51–76. ISSN: 0167-5249.

[39] Yves Molenbruch, Kris Braekers, and An Caris. "Typology and literature review for dial-a-ride problems". In: *Annals of Operations Research* (2017).

[40] Rani Molla. *Americans seem to like ride-sharing services like Uber and Lyft*. https://www.vox.com/2018/6/24/17493338/ride-sharing-services-uber-lyft-how-many-people-use. 2018.

[41] United Nations. *2018 revision of world urbanization prospects*. 2018.

[42] John J Nay and Yevgeniy Vorobeychik. "Predicting human cooperation". In: *PloS one* 11.5 (2016), e0155656.

[43] Thomas H Noe. "Capital structure and signaling game equilibria". In: *The Review of Financial Studies* 1.4 (1988), pp. 331–355.

[44] Harri Oinas-Kukkonen and Marja Harjumaa. "A systematic framework for designing and evaluating persuasive systems". In: *International conference on persuasive technology*. Springer. 2008, pp. 164–176.

[45] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. "Running experiments on amazon mechanical turk". In: *Judgment and Decision making* 5.5 (2010), pp. 411–419.

[46] Sophie N. Parragh, Karl F. Doerner, and Richard F. Hartl. "A survey on pickup and delivery problems. Part I: Transportation between customers and depot". In: *Journal für Betriebswirtschaft* 58.1 (2008), pp. 21–51.

[47] Sophie N. Parragh, Karl F. Doerner, and Richard F. Hartl. "A survey on pickup and delivery problems. Part II: Transportation between pickup and delivery locations". In: *Journal für Betriebswirtschaft* 58.1 (2008), pp. 81–117.

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[49] Noam Peled, Ya'akov Kobi Gal, and Sarit Kraus. "A study of computational and human strategies in revelation games". In: *AAMAS*. 2011, pp. 345–352.

[50] Harilaos N Psaraftis, Min Wen, and Christos A Kontovas. "Dynamic vehicle routing problems: Three decades and counting". In: *Networks* 67.1 (2016), pp. 3–31.

[51] James R Rogers. "Information and judicial review: A signaling game of legislative-judicial interaction". In: *American Journal of Political Science* (2001), pp. 84–99.

[52] Sören Schleibaum and Jörg P Müller. "Human-Centric Ridesharing on Large Scale by Explaining AI-Generated Assignments". In: *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good*. 2020, pp. 222–225.

[53] Harkiranpal Singh. "The importance of customer satisfaction in relation to customer loyalty and retention". In: *Academy of Marketing Science* 60.193-225 (2006), p. 46.

[54] Andrew Michael Spence. *Market signaling: Informational transfer in hiring and related screening processes*. Vol. 143. Harvard Univ Pr, 1974.

[55] Anne M Turner, Katrin Kirchhoff, and Daniel Capurro. "Using crowdsourcing technology for testing multilingual public health promotion materials". In: *Journal of medical Internet research* 14.3 (2012), e79.

[56] Amos Tversky and Danliel Kahneman. "The Framing of Decisions and the Psychology of Choice". In: *Science* 211.4481 (1981), pp. 453–458.