Humboldt-Universität zu Berlin

Master's Thesis

Lying, Its Detection, and Their Economic Implications

Hashem Amireh

Matriculation number: 594387

for acquiring the degree of

Master of Science (M.Sc.)

in Management Science and Economics

at the School of Business and Economics

of Humboldt-Universität zu Berlin

Examiner: Prof. Georg Weizsäcker

Berlin, 2 November 2020

Abstract

There exists a non-negligible amount of research on lying in economics. However, most of this research is largely self contained and only attempts to answer certain questions about behaviors around lying. This thesis attempts to achieve three things: (1) incorporate lying into a well-established theoretical microeconomic model, (2) create a mathematical formalization for rational lying, and (3) test certain practical hyptheses about lying and detection through an interactive experiment. This thesis incorporates lying into Akerlof (1970)'s adverse selection under asymmetric information model, leading to some interesting insights. The rational lying mathematical model establishes clear rules on when individuals lie, when they believe such lies, and how they react based on potential payoffs. The experiment results provides some limited evidence that those who estimate that their statements are more likely to be believed, are in fact more believed. It also provides some evidence for altruistic and tribal behavior when it comes to lying choices. The experiment also provides insights on lying behavior in an online setting, namely that the "illusion of transparency" and the "self-as-target" bias may not play a role in online lying behavior, unlike real life.

Data and R code:
https://github.com/hashemamireh/masters-thesis

A playable demo of the experiment:
https://masters-thesis.herokuapp.com/demo/

Experiment source code:

https://www.otreehub.com/projects/masters-thesis/

Contents

1 Introduction

Understanding what governs human interactions and human beliefs has always been the topic of discussion and research in many fields from psychology, sociology, anthropology, and even computer science. The economists has also taken interest in understanding how agents (companies, governments, individuals, etc..) behave in the market. However, economists often found themselves under fire for their simplifying assumptions regarding human rationality and behavior when coming up with models. At the most basic level, behavioral economists attempt to measure how much individuals deviate away from presumed rational behavior. Measuring this deviation is key to validating—or invalidating—economic models and theories. They can also be used to improve economic models and make them reflect reality more accurately. It also allows economists to rethink and reevaluate their theories and incorporate deviations from presumed rational behavior into their models.

Lying is an integral part of the human experience and it can indeed be considered rational behavior. As I will demonstrate later, lying is quite common in our lives and serves a very important purpose. Classical economic models generally ignored lying. Even models with information asymmetries—which seem ripe for incorporating lying—did not attempt to do so. In real life, when information asymmetries exist, people will rely on some level of intuition (along with other things) to determine if someone is trying to deceive them. Therefore, it is clear that lying—and its detection—can have profound effects on outcomes. In order to demonstrate this and motivate this thesis, I chose a contract theory model as a use case. By incorporating lying into Akerlof (1970) adverse effects under asymmetric information model, we find that there are theoretical probabilistic welfare gains.

Once we make the case for how impact lying can be and hence highlighting the importance of understanding lying better, I establish a mathematical framework for it. The

framework is based on two economic actors: a sender and a receiver. The sender must determine whether they should lie or not. This choice is based on benefits, costs, and the likelihood that their lie would be believed. We also explore how the likelihood can change based on other factors. The receiver must also determine the probability that they are being lied to. After determining the probability, they would make a choice depending on that probability and the potential payoffs. Once we establish a framework, to facilitate our understanding, we move onto the more practical part of this thesis, which looks into measuring our ability to lie without detection, and our ability to detect lies, as well as other related features.

Measuring lying ability and detection is a complex task and some pitfalls. If we think of the general population as a homogeneous monolith, then by definition, we can either be good at lying without detection or good at lying. By assuming this homogeneity, being good at these two things is mutually exclusive. The past experimental research argues that we are good at lying (without detection) but we are bad at detecting lies. In other words, lies often go undetected. But it does not try to differentiate people based on lying ability.

Much of the experimental research done on this topic simply looks at how the sample performs on average. However, what if we dropped this assumption of homogeneity? Put simply, what if some of us were good at lying while others are not? Or what if some of us are good at *detecting* lies while others are not? The mathematical model will give us some intuition on possible answers to this. Furthermore, the lying experiment tests a different approach towards estimating our ability to detect lies that allows for this homogeneity.

Rather than simply trying to see if, on average, we can accurately predict if our lies are being believed, we try to see if the variability within our estimates (meta beliefs) are a predictor of the variability within the real beliefs. A positive statistically significant

effect would suggest that we can indeed estimate others beliefs with some degree of accuracy.

1.1 Lying, Deception, and Evolution

When thinking about lying, we take humans' ability to lie for granted. However, our ability to lie is dependent on our brains being able to conceive lies to begin with. At the most basic levels, conceiving lies requires us to first conceive an alternate reality, not just the reality we perceive and believe. In order to help conceptualize this, it's useful to think of how the brain deals with object permanence. In Piaget (1977)'s "theory of cognitive development", he demonstrated how children's brains do not develop *object permanence* until a certain age. Before that age, a toddler's brain cannot conceive the existence of an object that she cannot see. When an object is removed from her field of vision, her brain is unable to conceive the "permanence" of this object's existence. Some adult animals' brains cannot conceive *object permanence* either. We therefore, cannot take our ability to do so for granted. And similarly, our brain's ability to even imagine alternate realities and therefore its ability to conceive lies, cannot be taken for granted either. Once we understand that our brains' ability to conceive lies is not something intrinsic, then we can ask the question, how did this ability come about?

To answer that question, it might be illuminating to look at something that is less complex than the human brain. Deception does not need a complex brain or a central nervous system. For instance, take the example of the genus Ophrys, which is a type of orchids. Species in the Ophrys genus have the appearance of a female insect. This lures male insects looking to mate, which in turn helps the orchids get pollinated. This mimicry is a form of deception that is the result of evolution (Schiestl 2005). There is a clear evolutionary advantage to such a physical trait. Similarly to how natural selection allowed these orchids to flourish, we can imagine how evolution selected for brains that

are capable of conceiving lies.

The unparalleled complexity of the human brain makes finding a clear and satisfactory link between evolutionary pressures and developing deception abilities, harder than finding the link for orchids. However, there are theories that attempt to explain how lying and the ability to lie provided an evolutionary advantage. One theory is that early humans who had developed the ability and disposition to strategically lie, were able to use this newfound ability to increase chances of successful mating. The premise here is that humans generally look for high quality individuals when choosing mates. If a mate is able to successfully able to "oversell" themselves through lying and deception, they are more likely to be selected than those who are unable to engage in such behavior. This meant that over time, evolution selected for the ability to lie and deceive strategically (Adenzato and Ardito 1999).

By framing lying as a result of evolution, we can see that it can have a clear strategic advantage. However, one can see how lying can also incur costs. These costs can arise immediately if the lie is detected. They can also arise in the long run when the lie is revealed to be one (after being initially believed). These costs can vary far and wide; from lack of trust in future interactions with the liar to exclusion from one's "tribe"[1] to criminal prosecution (in cases of perjury). This helps us comprehend lying as a choice that has a set of benefits, risks, and costs. We can therefore formalize lying as a rational behavioral choice.

1.2 Lying in Human Thought and Discourse

Lying is a widely studied and discussed topic in psychology and philosophy. Attempting to explain lying motivations and its ethics, is not a novel endeavour. Greek philosophers extensively discussed lying. Much of their focus, however, was on the morality of lying.

[1]Here we are using the more broad definition of "tribe" which includes any social group one may belong to.

Aristotle generally viewed lying as inherently immoral and problematic. However, according to Mahon (2019) Aristotle's view was not as absolutist as that of Socrates, who believed that lying is not permissible even to an enemy in times of war. Plato on the other hand, had a more permissive approach towards lying.

Similarly to Plato, contemporary perspectives on lying generally view lying as permissible in certain situation and even necessary. In fact, lying can be viewed as the most "moral" choice in certain situations. Viewing lying as something that can be morally excusable or even advised, means that the questions "why do we lie?" and "is it okay to lie?" are very intertwined. We will therefore discuss these two questions in tandem.

1.3 The Morality of Lying

While we often regard lying as an immoral behavior, one can easily think of situations in which lying can be moral and even selfless. Someone telling their bedridden dying spouse that they have not been a burden, is "honorable" even if it was a lie. In the opening scene of the 2009 film "Inglorious Bastards" (Tarantino 2009), a French farmer lies to a Nazi SS officer about the fact that he was hiding Jews in his house to protect them from persecution. By today's standards, this behavior is considered moral and appropriate. That said, examples of "non-immoral" lying don't have to be this dramatic or consequential. For instance, when invited over to dinner, one might choose to say that the they enjoyed the food even if they thought otherwise. Another example is coming up with an excuse to avoid a social gathering without offending the host. Dietz (2019) broadly defines this type of lying as "white" and "prosocial" lying. In her book chapter, she also highlights the different types of white and prosocial lies: (1) benevolent lying, (2) lying in self-defense, (3) altruistic lying, and (4) collaborative lying. Vrij (2008) argues that while lying can be a selfish act, it can also be integral to social function, In this context, he refers to lying as "a social lubricant". It is evident that, from a utilitarian

point of view, one can see why one would choose to lie. However, it is important to note that proponents of "radical honesty" such as Kant (1797) would argue that such lies are not as harmless as utilitarians might assert. For the purposes of this thesis, we will not delve into this question.

2 Lying, lying detection, and the "illusion of transparency"

To further our understanding of the of the intricacies of lying, we need to understand lying in a causal fashion. We are primarily concerned with three different junctions: (1) the decision to lie or not, (2) whether the "listener" believes the "speaker", which may inform her actions, and (3) the speaker's belief about whether the listener believed him or not. It is not hard to imagine how each one of these junctions informs the next one. For instance, if someone has a "nervous tick", they might give away the fact that they're lying or not. Similarly, the listener's body language and choice of words may also give away how skeptical they are of the speaker's statement. This can then inform the speaker on whether she believed him or not.

While this sequence seems strictly chronological, the probabilistic nature of these junctions can imply backward causality. For example, if the speaker thinks it is likely that the listener won't believe them, they might choose not to lie. That is especially the case if there is a meaningful cost to getting detected while lying. We generally refer to this as backward induction; the agent looks into the future and works their way back to inform their decision. This intertwining relationship makes grasping these different concepts important for understanding lying, its motivations, and its prevalence.

Before delving into the significance of these questions and how we can answer them, first we must clarify what the questions mean. When we say "are we good at lying" what

we really mean is "are we good at lying without being detected" (i.e. without others being able to tell that we are lying). And when we say "can we detect lies" what we mean is whether we can consistently assess the truthfulness of a statement with an accuracy better than that of random luck.

2.1 Prevalence of Lying

In subsection 1.3, we mentioned that lying can be an important tool for social interactions and is often used as a "social lubricant". Studies have shown that lying is quite pervasive in our daily lives. DePaulo et al. (1996) found that college students told two lies a day, on average. They obtained their results by enrolling students in an experiment in which they had to keep a journal of every social interaction they have and record whether they lied or not. George and Robb (2008) were able to replicate these experiments with comparable results. Interestingly, however, Erat and Gneezy (2012) find that people are less inclined in engage in what they term as "Pareto lies" and lies that are altruistic. Pareto lies are ones that make both the liar and the counterpart better off. Altruistic lying harms the liar but makes the counter part better off. This could be due to the positive psychological benefits of altruism compensating for the the negative psychological cost of lying.

The results of these experiments are based on lies that include "white" social-lubricating lies. What we are more interested in, are deliberate lies meant advance one's interest at the expense of others. The prevalence of this type of lying in everyday life is harder to pin down. Participants are less inclined to report lies that are generally associated with negative moral perceptions.

However, in experiments in which participants were given a very clear monetary incentive to lie, we see that some choose not to lie while some "partially lie" hence increasing their payoff but not to the maximum possible extent (Gneezy 2005; Mazar, Amir, and Ariely 2008; Lundquist et al. 2009; Gneezy, Rockenbach, and Serra-Garcia 2013; Fis-

chbacher and Föllmi-Heusi 2013; Gneezy, Kajackaite, and Sobel 2018). In an experiment by Lundquist et al. (2009), the participants were assigned to different treatment groups. In one treatment group, the participants had the choice between not lying and a pre-specified lie with varying levels of strength. What they found was that, the stronger the magnitude of the lie was, the less likely were the participants to engage in lying. This indicates that participants are more averse to committing a lie the more severe it is. Interestingly, a treatment group given the opportunity to write their own free-response message, engaged in less lying than other pre-specified treatment groups. This seems to indicate that the more agency is involved (i.e. having to formulate the lie), the higher the psychological cost. Therefore, people become more averse to lying. While the link is not explicit, this is consistent with Mazar, Amir, and Ariely (2008)'s "self-concept" hypothesis. They argue that lying can often harm one's "self-concept", which is how well people perceive themselves. They argue that people are averse to lying as to not diminish their self-concept. One could argue that when taking more agency in lying (like writing out the lie), participants would diminish their self-concept more than just clicking a single button.

2.2 Lying Detection

Lying detection is well documented in multiple fields of academia from psychology to criminology and even economics. According to Vrij (2008), generally speaking, people are bad at detecting lies. We also have a tendency to overestimate our ability to detect lies (Elaad 2003). Furthermore, lying detection machines like polygraphs are also unreliable. He also makes the case in his book that humans overestimate their ability to distinguish lies from true statements.

There has been a wide range of research that tries to document the certain behaviors suspected of giving away liars. These include verbal and non-verbal cues. Verbal

cues include things like filler phrases such as "um" or "ah", response length, pause duration, and voice pitch, among others. Non-verbal cues may include eye-contact (or lack thereof), hand gestures, fidgeting, blinking, and others. The evidence for how accurate both verbal and non-verbal cues are, is mixed at best (Vrij 2008; Mann 2019).

Over estimating our ability to detect lies is problematic and leads us to make decisions based on inaccurate information and hence risk not maximizing our respective payoffs.

2.3 The Illusion of Transparency

Gilovich, Savitsky, and Medvec (1998) study the "illusion of transparency," which refers to human tendency to assume that their lies are "leaking." In other words, when lying, humans often assume that whoever they are addressing is picking up on the fact that they are lying. It is called an illusion since, as mentioned earlier, we generally cannot accurately determine whether we are being lied to or not even though a liar might think her lying behavior is leaking. Gilovich, Savitsky, and Medvec (1998) designed and executed several experiments to test whether the illusion of transparency does indeed exist. In their first experiment, they split their subjects into groups of five. Five rounds were played. In each round, all subjects were supposed to answer a question. However, in each round, one subject/player was instructed to lie while the others were instructed to tell the truth. Participants were received their lying or truth telling assignment using a card that was only visible to the player holding it. After each answer, the four "observing" players recorded whether they believed the answer or not. Hence, the probability that one person was lying is 1-in-5. Five rounds are played with five different questions. Following each round, a new player was instructed to be the liar for that round; so every player got to lie once. At the end of the game, each player was asked to estimate how many of the observers picked up on their lie.

Two initial variations of this experiment were conducted. In the first one, the players

knew that each player will have to lie exactly once which effectively makes the players aware of the probability that a given player is a liar. In the second variation, the players were told that the liar in each round is chosen randomly by a computer with replacement—even though that was not actually the case. The results of both variations of the experiment indicated that the illusion of transparency does indeed exist. In the first variation, liars overestimated the percentage of observers that picked up on the lie by about 23 percentage-points (25.6% actual vs 48.8% perceived). In the second variation, the results are similar at 23 percentage-point difference (27% actual vs 50% perceived).

The percentages above predict the overestimation of liars thinking the observer can tell they are lying. However a flip side to this overestimation; when we are telling the truth can we accurately predict whether whoever we are talking to believes us or not? To answer this question, in the first variation the players were also asked to estimate how many players observers mistakenly thought the player was lying when they were indeed telling the truth. This was done to check for another factor that could be conflated with the illusion of transparency effect: the "self-as-target" bias. The *self-as-target* bias refers to our tendency to misjudge situations when we are the subject. Gilovich, Savitsky, and Medvec (1998) were worried that individuals may overestimate the percentage of observers saying they are lying regardless of whether they were lying or not. When looking at the numbers, the researchers found that truth-tellers did indeed overestimate the number of observers predicting a lie by 15 percentage points (as opposed to 23 percentage points for when they lie). This confirms the existence of the "self-as-target" bias but shows that the bias does not capture the full 23 percentage-point difference. This indicates that the illusion of transparency persists. Note that all these results were statistically significant.

A third variation of the experiment was also conducted to compensate for a couple

of other confounding factors that could arise: (1) The curse of knowledge, and (2) inherently detectable lies. The curse of knowledge refers to humans overestimating how detectable their lies are not due to thinking that their behavior is giving away the fact that they are lying, but rather, due to a mental bias caused by a tendency of humans to project their knowledge on others and subconsciously assume that others know what we know even if that is not the case. The other issue is that some lies can be inherently detectable. Certain statements can be flagged as lies simply because they are unlikely to be true, not because the agent making the statement is showing signs that he or she is lying. We can demonstrate this by using an example based in a stereotype. If a man was asked in an experiment to state his favorite color and was also instructed to lie, he may choose to say pink. Pink is more likely to be flagged as a lie since, stereotypically, pink is a color that is preferred by women over men. To address this, Gilovich et al. added a yoke to the third variation. The yoke would be aware of the truthfulness—or the lack thereof—of a statement before it is made. The yoke would then attempt to predict how many of the unaware observers thought the statement was a lie. The results showed that the liars' estimations were 19 percentage-points higher than the yokes. Therefore, it is unlikely that the overestimation is caused by the *curse of knowledge* or inherently detectable lies. This further confirming the existence of an illusion of transparency.

Evidently, understanding lying and its layers is a complex task. In section 4, we attempt to establish a mathematical framework that can facilitate our understanding of these complex relations.

3 Contract Theory and Lying

Information is an integral part of economic models. It informs agents' decisions and choices. In an ideal world, there is complete information. In other words, information is absolutely accurate and it is not costly to access such it. However, when information

is asymmetric or incomplete, the possibility of market failure arises (Akerlof 1970).

In the literature, there are mainly three different types of information asymmetry models: (1) hidden action, (2) hidden information, and (3) "adverse-selection" models. Hidden action and hidden information models describe situations in which the information asymmetry does not arise until *after* a contract is signed. Therefore, lying and lying detection have no role in the contract signing stage. Furthermore, unless a third party with perfect lying detection ability can be serve as an arbiter, lying detection cannot be used effectively to enforce contracts. As we mentioned earlier, even professional lying detection methods like polygraphs are not reliable. We therefore cannot use lying detection as a legally binding method to enforce contracts. What we are rather more interested in, is probabilistic lying detection by the economic actors themselves. Therefore, we will use the "adverse-selection" model to motivate our discussion.

Under the "adverse-selection" model, information asymmetry occurs before a contract is signed. The "agent" has an information advantage over the "principal". Since the information is hidden and uncertain, the principal must rely on the second best thing: probabilistic estimations of information. The crude method, which is used in the standard adverse selection models, would be to consider the population distribution of possible types and assume that the information is randomly selected from this population distribution. For instance, in the example of auto insurance, the insurer assigns an estimated risk equal to the mean population risk, to every customer. However, economic actors can sometimes refine their predictions to assign specific probability distributions to specific cases, rather than simply relying on the population distribution. This allows them to minimize some of the disadvantages of incomplete information, and possibly reduce the over all impact of the market failure.

Spence (1973) and Stiglitz (1975) propose "signaling" and "screening" as methods to help refine probabilistic estimations. Signaling is the process in which the agent

with the complete information takes certain actions that would "signal" her type to the principal. Spence (1973) uses the example of going to college to "signal" to employers certain abilities like being able to learn new things. In turn, employers adjust their probabilistic estimation of candidates' abilities based on this information. Employers still face imperfect information but are able to achieve more accurate predictions. Screening refers to the principal offering a bundle of contracts that are designed in such a way that the agent's choice actually reveals something about their type. A good example of this is auto insurance firms offering different policies (deductibles, co-pays, etc...) at different prices. The high risk individuals would "self-select" and choose the expensive policy that has better terms. Both signaling and screening are costly. But what if the principal can refine their probabilistic estimation and measure how truthful the agent is being about their type?

3.1 The Original Model

As we have seen, lying is discussed fairly often in behavioral economics. However, it is seldom applied to existing economic models. For the purposes of this thesis, I outline a theoretic application of how lying ability, the ability to detect lies, and how well we can accurately estimate our own lying ability. This is done to highlight the there could be real effects on outcomes when lying and lying detection are incorporated into economic models. Information economics models were the obvious candidate for such application. Models with asymmetric information allow for one party to leverage their information advantage to extract more utility at the expense of the other by misrepresenting the facts. Think of the employment market discussed by Spence (1973); any worker can claim that they is smart, hard-working, or a fast learner (cheap talk). However, if the employer is able to somewhat accurately detect who is lying and who is telling the truth, they she make more informed hiring choices.

First let us briefly go over the adverse-selection model described by Mas-Colell, Whinston, and Green (2011).[2] Note that I will express some of the functions differently as they will help us adapt the model in the next section. In the model, we have N workers and one firm. When working for the firm, the workers have varying levels of productivity θ_i where $i \in N$ is a given worker. Workers earn wage $w_i = w(\theta)$. Workers can also generate income r from an alternative (referred to as "home production" in model). This alternative income *can* also be a function of θ. Therefore we will state it as $r_i = r(\theta_i)$. This alternative is the opportunity cost of working for the firm. The worker population's productivity has some distribution $F(\theta)$. For now, the only assumption we will make about this distribution is that $\theta \in [\underline{\theta}, \bar{\theta}]$ where $0 \leq \underline{\theta} < \bar{\theta} < \infty$, holds true.

Workers will work for the firm if $w_i \geq r_i$. We define $\hat{\theta}$ as the cutoff in which workers with $\theta_i > \hat{\theta}$ will prefer to work at the firm while workers with $\theta_i < \hat{\theta}$ will prefer to work from home. Therefore the workers' total revenue is as follows:

$$\underbrace{\int_{\underline{\theta}}^{\hat{\theta}} F(\theta)d\theta \cdot w(\theta)}_{\text{(A)}} + \underbrace{\int_{\hat{\theta}}^{\bar{\theta}} F(\theta)d\theta \cdot r(\theta)}_{\text{(B)}} \tag{1}$$

(A) captures the total revenue for those who work at the firm while (B) captures the total revenue for those who do not. If the following is true:

$$w_i < r_i \quad \forall \quad i \tag{2}$$

then we will always reach the Pareto optimal allocation in which all workers would choose home production.

However, if we consider a situation in which some $w_i \geq r_i$ is true for some i while $w_i \leq r_i$ is true for others, the situation becomes a little more complicated.

[2]For more details, see 13.B in Mas-Colell, Whinston, and Green (2011).

3.1.1 Under Perfect Information

When the firm can fully observe θ_i for every i, they will then assign $w(\theta_i) = \theta_i$ for every i. This is Pareto optimal and maximizes total worker revenue. We can plug $w(\theta) = \theta$ into (1) and get the Pareto optimal allocation below. Since this is the maximum possible workers' revenue, we will use this a a baseline to test efficiency.

$$\int_{\underline{\theta}}^{\hat{\theta}} F(\theta)d\theta \cdot \theta + \int_{\hat{\theta}}^{\bar{\theta}} F(\theta)d\theta \cdot r(\theta) \tag{3}$$

3.1.2 Under "Perfectly Imperfect" Information

"Perfectly imperfect" situations are those in which an economic actor only knows the population distribution without any hints or partial information about specific situations/agents. We create this distinction because, in the next subsection, we will introduce the concept of probabilistic or partial information.

Since we are namely concerned with how lying detection can reduce inefficiency (i.e. achieve a more Pareto efficient outcome), we will focus on how lying detection affects the worst case scenario. The worst case scenario is referred to as "complete market failure." Under complete market failure, the outcome is that no workers end up working at the firm due to to adverse effect.

Consider the following situation:

$$\begin{aligned} &r(\underline{\theta}) = \underline{\theta} \\ &r(\theta) < \theta \quad \text{for all other } \theta \end{aligned} \tag{4}$$

This implies that there exists a lowest productivity level θ at which the worker would be indifferent between working at the firm or not. However, all workers that are more productive would be better off working at the firm if they get wage $w(\theta_i)$. However, due to "perfectly imperfect information" the firm will have to assign constant wage $\hat{w}$ to

all workers. If the firm assumes that all the workers will work for the firm, they would choose:

$$\hat{w} = \frac{\sum_i \theta_i}{N} \tag{5}$$

Then the following situations *could* occur for different ranges of θ:

$$r(\theta_i) \leq \hat{w} \tag{6}$$

$$r(\theta_i) > \hat{w} \tag{7}$$

In every case, some workers will definitely fall into the (6) condition. That is because those with productivity $\underline{\theta}$ are indifferent between the choices (recall $r(\underline{\theta}) = \underline{\theta} \leq \hat{w}$) which falls into (6). The distribution of the rest of the workers between the two conditions relies on two things: (1) the underlying density distribution $F(\theta)$, and the specification of the function $r(\theta)$. We will therefore assume the following:

$$F(\theta) \text{ distribution is symmetric} \tag{8}$$

$$\underline{\theta} = 0 \implies \theta \in [0, \bar{\theta}] \tag{9}$$

$$r(\theta) = \alpha\theta \text{ where } 0 < \alpha < 1 \tag{10}$$

In (10), the condition $0 < \alpha < 1$ ensures that (4) remains true. Based on these assumptions, we can calculate:

$$\hat{w} = \frac{1}{2}\bar{\theta} \tag{11}$$

If we then plug in (10) and (11) into (7) and rearrange, we get:

$$\theta_i > \frac{\hat{w}}{\alpha} = \frac{1}{2}\frac{\bar{\theta}}{\alpha} = \theta^* \tag{12}$$

This means that workers with θ_i larger than the given expression, will choose home

production over working for the firm. The threshold θ^* serves as the cutoff between the workers who would prefer home production and those who prefer working at the firm. Note that θ^* is a function of α. Note that since $\frac{\partial \theta^*}{\partial \alpha} < 0$, the higher α means a larger proportion fo workers would prefer home production. This actually makes intuitive sense because as α gets smaller, the average worker's productivity increases compared to the home production productivity, so it becomes more lucrative to work for the firm.

Put simply, all the workers with $\theta_i > \theta^*$ will also have $\hat{w} < r_i$. They therefore choose not to work at the firm. Knowing this would end up happening, the firm would readjust its $\hat{w}$ formula (5) to exclude those with $\theta_i > \theta^*$. This would, in turn, make $\hat{w}$ in (11) go down leading to θ^* to also go down. This creates a feedback cycle until it reaches $\theta^* = \underline{\theta}$. This leads to only least productive workers being hired. This equilibrium is clearly not Pareto optimal due to the coordination problem involved. In fact, this equilibrium is the worst possible outcome as it minimizes the workers' total revenue. If we back track a bit, we can see that if α is sufficiently low (at least as low as 1/2 in this case), then we can see that $\theta^* = \bar{\theta}$. This means that all workers would prefer to work for the firm and receive wage $\hat{w}$. Therefore, the initial state that triggers the feedback cycle is achieved. Pareto optimal outcome is achieved and total worker revenue is achieved. Figure 1 shows for what values of α, a Pareto optimal outcome is achieved. When $0 < \alpha \leq 1/2$, the feedback cycle is

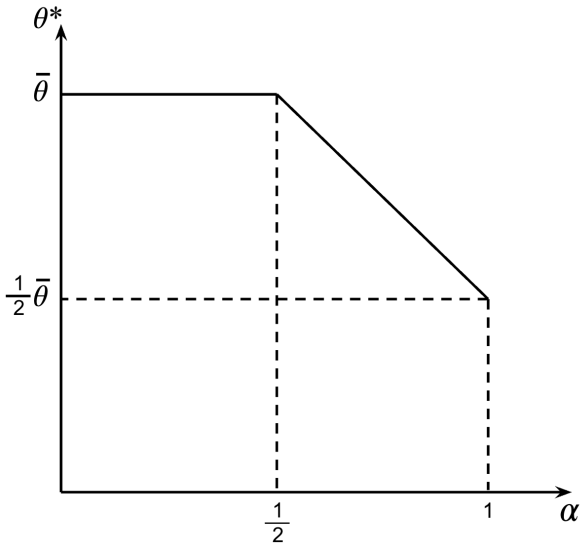

Figure 1: The relationship between α and θ^*

averted and a Pareto optimal equilibrium is reached. However, when $1/2 < \alpha < 1$, a value $\theta^* < \bar{\theta}$ is reached, triggering the feedback cycle and an adverse effect, and leading to the worst possible outcome.

3.2 Incorporating Lying (Under Partial Information)

Now consider a situation in which the firm interviews the workers and then assigns each worker an estimated productivity level $\tilde{\theta}_i$. Let the overall accuracy of these estimations be $p \in [0, 1]$. The exact formula of p depends on the underlying distribution of distribution of $F(\theta)$. Let us assume $\theta \sim U(\underline{\theta}, \bar{\theta})$ (i.e uniformly distributed). Then the accuracy equation for a given worker i would be as follows:

$$p_i = 1 - \frac{\sqrt{(\tilde{\theta}_i - \theta_i)}^2}{\bar{\theta} - \underline{\theta}} \tag{13}$$

The overall accuracy equation would be as follows:

$$p = 1 - \frac{\sum_i \sqrt{(\tilde{\theta}_i - \theta_i)}^2}{N(\bar{\theta} - \underline{\theta})} \tag{14}$$

Due to the uniform distribution and the implied accuracy equation, assigning all workers the same $\tilde{\theta}_i$ leads to $p = 1/2$. This is the most crude way to assign productivity and is effectively the same as just using the population mean. Any method to estimate productivity with $0 \leq p < 1/2$ is worse than just random chance. Therefore, any method with $1/2 < p \leq 1$, is considered to me an improvement over random chance.

Now let us assume that the firm assigns every worker wage $w_i = w(\tilde{\theta}_i) = \tilde{\theta}_i$. Let us also assume that (9) and (10) hold true. In other words, we're using the same same setup as we did in the previous section with the addition of the uniform distribution. If $p < 1$ (i.e. if the firm does not accurately predict θ_i for every worker), then we should

expect that the following situations could occur:

$$r(\theta_i) \leq \tilde{\theta}_i \tag{15}$$

$$r(\theta_i) > \tilde{\theta}_i \tag{16}$$

Similarly to the previous section, if any worker ends up with a situation similar to (16), they will choose home production. This triggers the feedback cycle and moves us away from the Pareto optimal outcome. Therefore, we need to check under which conditions (16) cannot hold true for any workers. If we re-arrange (13) and solve for the squared term, we get:

$$\tilde{\theta}_i = \begin{cases} (\text{1-p}_i) - \theta_i & \text{for} \quad \theta_i < \tilde{\theta}_i \quad \text{(17a)} \\ \theta_i - \bar{\theta}(1 - p_i) & \text{for} \quad \theta_i > \tilde{\theta}_i \quad \text{(17b)} \end{cases}$$

Cases in which the estimated productivity is higher than the actual productivity ($\theta_i < \tilde{\theta}_i$) are not compatible with (16). We therefore only need work with situations in which the productivity of the worker was underestimated ($\theta_i > \tilde{\theta}_i$). If we plug in (10) and (17b) into (16), we get the following expression:

$$\alpha > 1 - \frac{\hat{\theta}}{\theta_i}(1 - p_i) = \alpha^* \tag{18}$$

For the expression above, α^* represents the threshold for which has to be bigger than to trigger the feedback loop. In other words, for $\alpha \in (0, \alpha^*]$, the feedback loop in not triggered and a Pareto optimal outcome is achieved. For $\alpha \in (\alpha^*, 1)$, the feedback loop is triggered and the equilibrium is not Pareto optimal. Now if we take the partial

derivative of α^* with respect to p_i, we get the following:

$$\frac{\partial \alpha^*}{\partial p_i} = \frac{\hat{\theta}}{\theta_i} > 0 \tag{19}$$

The partial derivative is strictly positive since both both $\hat{\theta}$ and θ_i are strictly positive. This means that as the accuracy of estimation of productivity by the firm p_i goes up, the range of α that would trigger the feedback loop and lead to a non-Pareto optimal outcome $(0, \alpha^*]$, would shrink. Put simply, this means that as the firm is more able to accurately assess workers' productivity, more values of α would lead to a Pareto optimal outcome.

We did not discuss the details of how the firm can assign estimated productivities to workers. One can imagine a wide range techniques that the firm can employ. They can do trial tasks or have the workers take an exam. Alternatively, if the interviewer can detect lies well, they might be able to detect when someone is over stating their abilities/productivity. This is meant to demonstrate that when lying and lying detection are introduced into economic models, they can have some profound impacts. However, this is only theoretical. As we discussed earlier, the literature seems to suggest that humans are generally bad at lying detection. This would imply that that the data would not support this theoretical impact of lying and detection. However, what if some of us are good at detecting lies while others are not? In section 5, we will discuss the experiment and its results.

4 A Mathematical Model for Rational Lying

In order to aid in our understanding of the rational behavior of lying, I will outline a simple mathematical model. The model relies on the rational behavior of both the "liar" and the person being lied to. However, this terminology it problematic for our model.

In our model, we need to allow for both true and false statements (lies). In other words, the model needs to allow for the possibility of of the speaker telling the truth, from the perspective of the one being (possibly) lied to. To more accurately capture the two roles in the model while also simplifying the terminology, we shall refer to the individual who makes the statement x as the "sender" i, while the one the receiving end of the statement will be referred to as the "receiver" j. The statement—which may be a true one or a lie—will simply be referred to as the "message". Let's define $t_i^x \in \{0, 1\}$ as the (binary) truthfulness of message x (1 = truth, 0 = lie).

4.1 The sender's utility framework

In our model, the sender i has the choice to either lie or tell the truth to "receiver" j. First, we will define their utility function when they tell truth and when they lie. Then, we can write the utility maximization equation.

4.1.1 When the sender lies:

When $(t_i^x = 0)$, the expected utility function is as follows:

$$\mathbf{E}[u_i] = \underbrace{u_i(x|b_j^x = 1) \cdot \Pr[b_j^x = 1]}_{\textbf{(A)}} + \underbrace{u_i(x|b_j^x = 0) \cdot \Pr[b_j^x = 0]}_{\textbf{(B)}} - \underbrace{C_i^f(x)}_{\textbf{(C)}} \tag{20}$$

where,

x	is a given message
$b_j^x \in \{0, 1\}$	indicates[3] whether j believed message x
	(0 = did not believe, 1 = believed)
$u_i(x\|b_j^x)$	is the utility extracted by i from message x given b_j^x
$\Pr[.] \in [0, 1]$	is a probability function (discussed in detail below)

Note that the premise of this model is that if the sender i is lying, the receiver j would either detect the lie or believe it. This would then inform j' actions which affect i's realized utility. The first term (A) captures the "belief component", which is the utility extracted if the receiver believed the lie. The second term (B) captures the "detection component", which can be thought of as the cost of the lie being detected. However, it does not actually have to be a cost (i.e. a negative value). It can be a (positive) opportunity cost. We only need to assume that the following holds true:

$$u_i(x|b_j^x = 1) > u_i(x|b_j^x = 0) \tag{21}$$

This assumption means that the liar extracts more utility (or less disutility) if their lie is believed than when their lie is detected. In other words, the liar would prefer for their lie to go undetected.

The third term (C) captures any inevitable costs associated with lying. These are costs that are incurred *regardless* of whether the lie is detected or not. This can capture a wide variety of costs. For instance, if the lie is not initially detected but will eventually be revealed to be a lie, this term captures such future costs. If someone is playing poker and she attempts to "bluff" the other players and they believe her, she would gain some value. However, immediately after the round, her deception is revealed which might make the other players trust her less in future rounds. This lack of trust, would be captured by this term. This term can also internalize harm done to others through lying (i.e. it allows us to incorporate altruistic behavior in the model). Mazar, Amir, and Ariely (2008)'s "Self-Concept" (mentioned in section 2.1) costs would also be captured by this term, which would compensate for this seemingly "irrational" behavior. Note subscript i on the term, which allows for this cost to vary across different individuals. For instance, different people may have different levels of altruism.

[3]This definition of b_j^x is too simplistic and will be developed further later on. See section 4.2 for more details.

Now we will define the probability component of the function as follows:

$$\Pr[b_j^x | B_x, a_i^L, a_j^D] \tag{22}$$

where,

B_x	is an exogenous variable describing the believability of message x
$a_i^L \in [0,1]$	i's ability to lie without being detected
	(0 = 100% of lies are detected, 1 = 100% of lies go undetected)
$a_j^D \in [0,1]$	j's ability to correctly identify the truthfulness of a statement
	0 = inaccurately identifies the truthfulness of 100% of statements
	1 = accurately identifies the truthfulness of 100% of statements

Some lies are inherently less believable than others. For instance, when someone claims that "the sky is green and has always been green", that lie is inherently unbelievable because of everyone's knowledge about the state of the world. This an example of what is being captured by B_x. The ability variables a_i^L and a_j^D only capture an individual's general ability regardless of the lie itself. We could also add another variable to capture j's information that is relevant to statement x. This would capture j's ability to detect specific lies based on some information they know. However, we will leave that out of the model for simplicity's sake.

Note that the probability that a lie is detected and the probability that it is not, are a function of one another as shown below:

$$\Pr[b_j^x = 1] = 1 - \Pr[b_j^x = 0] \tag{23}$$

Also note the impact of B_x, a_i^L, and a_j^D on the probabilities above. The probability

that a lie is believed goes up with lie message believability B_x and lying ability a_i^L and goes down with detection ability a_j^D:

$$\begin{aligned} \frac{\partial \Pr[b_j^x = 1]}{\partial B_x} &> 0, \\ \frac{\partial \Pr[b_j^x = 1]}{\partial a_i^L} &> 0, \\ \frac{\partial \Pr[b_j^x = 1]}{\partial a_j^D} &< 0 \end{aligned} \tag{24}$$

Note that the above equations use the probability of the lie being believed. The same can be of course done for the probability of the lie being not believed, we would just have to reverse which is larger than zero and which is smaller than zero.

Sender i's knowledge about j's detection ability a_j^D is limited. Therefore, they can only estimate it. We also want the model to allow for the possibility that the sender inaccurately estimate their own lying ability a_i^L. We will therefore replace a_i^L and a_j^D with $\mathbf{E}_i[a_i^L, a_j^D]$. We can then plug it into the probability equation (22) and get:

$$\Pr_i[b_j^x | B_x, \mathbf{E}_i[a_i^L, a_j^D]] \tag{25}$$

Going back to equation (20), the terms (A) and (B) can be condensed by using a summation term to get a more concise notation. We can also plug in the updated probability function:

$$\mathbf{E}_i[u_i] = \sum_{b_j^x=0}^{1} (u_i(x|b_j^x) \cdot \Pr_i[b_j^x]) - C_i^f(x) \tag{26}$$

4.1.2 When the sender tells the truth:

When $(t_i^x = 1)$, we will define the utility function simply as a "reserve utility" $\bar{U}_i$. Note that when compared to the utilities when lying, we assume that the reserve utility $\bar{U}_i$ is less than the utility of lying and **not** getting detected, and greater than or equal to the

utility of lying and not getting detected. Put simply, if we ignore the inevitable costs term $C_i^f(x)$ (or assume is is equal to zero) for now, the sender always prefers to lie when they know their lie wouldn't be detected. However, they would rather tell the truth in their lie were to be caught (or in the case of equality . We can therefore update (21) as follows:

$$u_i(x|b_j^x = 1) > \bar{U}_i \geq u_i(x|b_j^x = 0) \tag{27}$$

The reason I chose to make the first inequality strict, is because we are talking about situations in which there is an some incentive to lie. Otherwise, this entire discussion would be moot. However, the second inequality is not strict. This is to allow situations for which the sender has a weakly dominant strategy to lie (again, assuming $C_i^f(x) = 0$), which is the case in our experiment as we will see later.

4.1.3 The sender's lying condition

Based on the expected utilities for the sender outlined in the previous sections, we can determine the condition in which they would choose to lie is:

$$t_i^x = 0 \implies \sum_{b_j^x=0}^{1} (u_i(x|b_j^x) \cdot \Pr_i[b_j^x]) - C_i^f(x) \geq \bar{U}_i \tag{28}$$

4.2 The receiver's utility framework

From the perspective of the receiver j, they have a choice to either believe the sender or not. This would then inform j's behavior. Therefore, their utility is a function of whether they believe the sender's statement but also whether the sender was actually telling the truth or lying. Similarly to the sender, we need to establish utility functions based on j's choice variable; whether they believe the message x or not. This variable is b_j^x, which we already defined in section 4.1.1. The utility function for both states is

as follows:

$$\mathbf{E}[u_j] = \underbrace{u_j(x|b_j^x, t_i^x = 0) \cdot \Pr[t_i^x = 0]}_{\textbf{(A)}} + \underbrace{u_j(x|b_j^x, t_i^x = 1) \cdot \Pr[t_i^x = 1]}_{\textbf{(B)}} \tag{29}$$

where $b_j^x \in \{0, 1\}$ is the choice variable (0 = don't believe, 1 = believe). (A) is the "lie component", which pertains to the possibility that message x turns out to be a lie, while (B) is the "truth component".

We assume the following statements to be true:

$$u_j(x|b_j^x = 1, t_i^x = 1) > u_j(x|b_j^x = 0, t_i^x = 1) \tag{30}$$

$$u_j(x|b_j^x = 1, t_i^x = 0) < u_j(x|b_j^x = 0, t_i^x = 0) \tag{31}$$

Equations (30) and (31) above indicate that the receiver j is always better off when they correctly determine the truthfulness of message x than when they don't. We need these assumptions because we are trying to model situations in which there is an incentive to lie and have this lie believed. If, for instance, the assumption in (31) did not hold true, then i would not need to lie to begin with. Let's discuss an example to in order to perceive this more concretely. Consider a game of poker, the sender here is a player says they have a "royal flush" and "goes all in". There are two possibilities: (a) the sender is lying/bluffing, in which case the other player(s) (i.e. receivers) are better off "calling his bluff" and matching his bet to stay in the game and win. The other possibility is that (b) the sender is actually telling the truth, and therefore, the other player(s) are better off folding instead of betting more money to stay in the game. The receiver(s) are clearly better off correctly predicting whether the sender is lying or not. In other words, the interests of the sender and the receiver are incompatible. The inequalities in (27), (30), and (31) are what establish this incompatibility of interests. Note that the

inequalities are strict to prevent any weakly dominant strategies.

Similarly to (23), we know that:

$$\Pr[t_i^x = 1] = 1 - \Pr[t_i^x = 0] \tag{32}$$

While we earlier explained the variable b_j^x as whether the receiver j believes the message x, that is not in fact accurate. What b_j^x actually captures is whether the receiver j will act as if they believe x or not. To put things more concretely, it might be tempting to assume that if the receiver encounters a $\Pr_j[t_i^x = 1] < 0.5$ (i.e. they think that it is more likely than not they are being lied to), they would act as if they are being lied to. That line of thinking, however, does not take into account the potential rewards and costs. One can envision a situation in which we think it is more likely than not that we are being lied to, but we still act as if we believe the message simply because the benefit if we acted as if we believed x and it was in fact true is so much higher than the cost of believing when it is in fact false. Therefore any experiment that tries to estimate the perceived $\Pr_j[t_i^x = 1]$, it needs to balance the costs and benefits (or at least compensate for the imbalance in the quantitative analysis).

For j to chooses $b_j^x = 1$ (act as if they believe x), the following must be true:

$$\mathbf{E}[u_j | b_j^x = 1] \geq \mathbf{E}[u_j | b_j^x = 0] \tag{33}$$

Before expanding the inequality above, first let's simplify the notation in (29) as follows:

$$\text{utility:} \qquad u_j(x | b_j^x = n, t_i^x = m) \equiv u_j^{n,m} \tag{34}$$

$$\text{probability } x \text{ is a lie:} \qquad \Pr[t_i^x = 0] \equiv P_L \tag{35}$$

$$\text{probability } x \text{ is true:} \qquad \Pr[t_i^x = 1] \equiv P_T \tag{36}$$

Now we can expand (33) and also use the simplified notation above:

$$u_j^{1,0} P_L + u_j^{1,1} P_T \geq u_j^{0,0} P_L + u_j^{0,1} P_T \tag{37}$$

Using (32) we can rearrange the the above inequality to get the expression:

$$u_j^{1,1} - u_j^{0,1} \geq P_L \cdot (u_j^{1,1} + u_j^{0,0} - u_j^{1,0} - u_j^{0,1}) \tag{38}$$

From (30) and (31), we know that $(u_j^{1,1} + u_j^{0,0} - u_j^{1,0} - u_j^{0,1}) > 0$, Therefore we can divide it over without reversing the inequality to get the following expression:

$$b_j^x = 1 \implies P_L \leq \frac{u_j^{1,1} - u_j^{0,1}}{u_j^{1,1} + u_j^{0,0} - u_j^{1,0} - u_j^{0,1}} \equiv \widehat{P_L} \tag{39}$$

This is equivalent to the following expression:

$$b_j^x = 1 \implies P_T \geq \frac{u_j^{0,0} - u_j^{1,0}}{u_j^{1,1} + u_j^{0,0} - u_j^{1,0} - u_j^{0,1}} \equiv \widehat{P_T} \tag{40}$$

While both are equivalent, the notation in (40) is slightly more intuitive. It can be expressed as: j will choose to believe i' message x, if the probability that x is in fact true is higher than threshold $\widehat{P_T}$.

5 The Experiment

Since this experiment attempts to answer questions regarding lying detection and the "illusion of transparency", it should ideally be done with in-person interactions. However, due to the Novel Coronavirus pandemic, the experiment was done online instead. Using the online format, paired participants can interact with one another using a chat box. This allows us to simulate a real-life back-and-forth conversation as closely as pos-

sible. The experiment was coded and deployed using oTree; an open-source platform for interactive games (Chen, Schonger, and Wickens 2016). Participants were recruited and paid through Amazon mTurk (see subsection 5.2 for details).

While doing the experiment using an online format may seem like a drawback since we generally assume that lying becomes obvious through body language, it actually presents an opportunity. Previous research on lying detection has been largely done using an in-person format. However, assuming that lying detection can only be done using an in-person format implies that lying "leakage" can be only through body language. That might not be the case. Leakage could, in theory, take place through other routes such as word choice, syntax, what we say, and how we say it. Therefore, conducting an experiment using an online format (combined with results from previous in-person experiments) allows us to delineate the various ways in which things like leakage, detection, the illusion of transparency, and the "self-as-target" bias, can take place. Furthermore, since I am employing a novel experimental design, running it online first could be a less costly test for the design itself. Additionally, comparing the results between the online experiment and a future in-person experiment with a similar design, would provide a more robust delineation than when comparing to the results of other in-person experiments that have different designs.

The experiment was done in batches of about 20 players each (some drop out and are sometimes replaced). Each player is assigned to a maximum of six identical games (except for the treatment effect). In each game, two players are randomly assigned to each other. Each player is randomly assigned either the role of a potential liar or the one who has to detect the lies. Players who have already played together once may be randomly assigned to each other again but only if they have the opposite roles. In other words no exact assignment is repeated. Players know that they are randomly assigned a new counterpart after every game.

5.1 Experimental Design

The premise of the experiment is that a player assumes either the role of a "buyer" or a "seller". The buyer is initially endowed with \$2 while the seller is endowed with \$0. The seller and buyer are informed that the seller had satisfactorily provided a service to the buyer for a fee of \$1. A computer randomly chooses with a 50/50 chance to transfer the amount due. Both players are informed of this. However, only the seller is told whether the automatic transfer was successful or not (i.e. whether they received the amount). The buyer is left in the dark but is told that the seller knows for a fact whether the automatic transfer was successful or not. Both are informed that the buyer would get the chance later on in the game to transfer \$1 regardless of the result of the random transfer.

If the random transfer succeeds and the buyer chooses to transfer the \$1 (i.e the amount gets transferred twice), the seller ends up with \$2 while the buyer ends up with \$0. In the case where the random transfer succeeds and the buyer chooses not to transfer the \$1, and in the opposite case (where random transfer fails and the buyer transfers the \$1), both would end up with \$1 each. Both are also informed that in the case that the initial random transfer fails and the buyer still does not transfer the \$1, a \$2 penalty is levied on the buyer. In that case, both would end up with \$0 payout for that round. This is done so the buyer does not have a weakly dominant strategy not to transfer (i.e. expected payout of transferring and not transferring are both equal). In other words, we are trying to create a situation in which the buyer has an incentive to always try to accurately predict whether the transfer was indeed successful or not. Figure 2 shows the extensive form representation of the game, which summarizes the possible situations described above.

Since the buyer plays the role of the "receiver" in the lying model we established, if we plug in her utilities in each situation into equation (40), we would get $P_T > \widehat{P_T} = 1/2$.

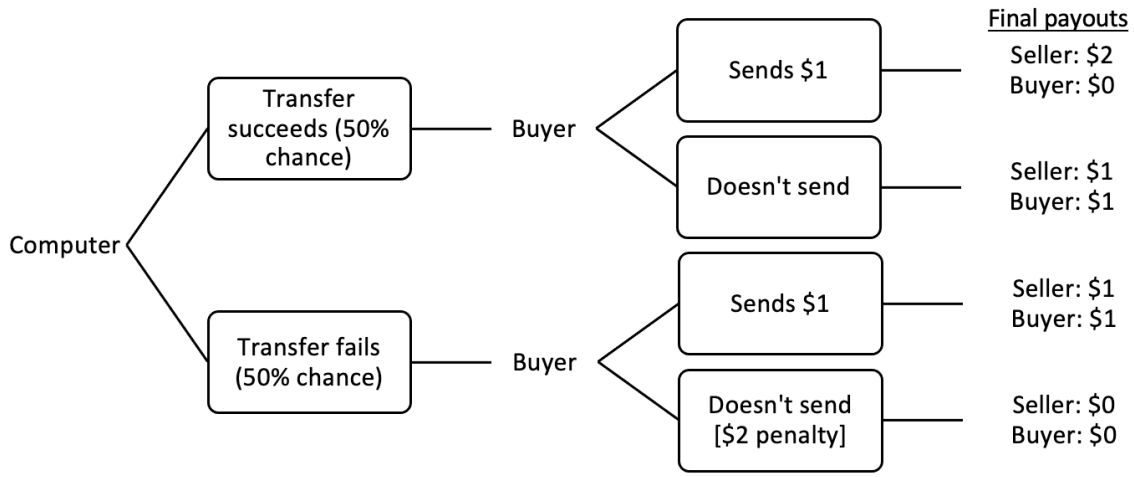

Figure 2: Extensive Form Representation of Game

This means the seller should act as if they believe the buyer (i.e. transfer the $1) if their estimate of the probability that the buyer is telling the truth is higher than 1/2. This is precisely because we do not have a weakly dominant strategy. Note that, on the other hand, the seller is always better off if the buyer sends the $1. Therefore, based on the payouts alone, the seller should have a strictly dominant strategy to lie.

Before the buyer gets the chance to choose whether to transfer the $1, a real-time chatroom between the two players is initiated. The buyer is told that they should use the chatroom to try to determine whether the seller actually received the $1 or not (i.e. whether the random transfer for was successful or not). The seller is told that they "may use" the chatroom to convince the buyer that they have not received the amount due. In the case the random transfer succeeded, the seller is not explicitly asked to lie, they are only being given a monetary incentive to do so (when the random transfer succeeds). This was a conscious choice in the design. Since the experiment aims to measure lying ability and lying detection, participants might be less apprehensive about lying since they are not making the "immoral choice" to lie themselves. As we discussed earlier, liars may "leak" the fact that they are lying. This "leakage" is generally assumed

because their behavior (body language, word-choice, etc...) unintentionally changes to to the trepidation. If there is less of this apprehension, then there might be less leakage, which would in turn hamper our ability to measure lying and detection ability. This is in contrast to Gilovich, Savitsky, and Medvec (1998) in which participants where either explicitly instructed to lie or tell the truth.

The players must spend 2 minutes and 30 seconds on the chat page (they cannot click next) in order to force them to communicate. Following the chat page, the buyer is presented with a page with two questions whose order is randomized. First, they must estimate the probability that the random transfer was successful or not, on a scale of [0-1]. By subtracting that value from one,[4] we get buyer's estimate of the probability that the seller is telling the truth. The second question is a radio button question asking the player to decide whether they would like to transfer the amount due. Both of these questions serve as a measure of whether the the buyer believes the seller not. However, one is binary while the other is a continuous probability. Additionally, the belief probability questions had no consequences on the game itself while the transfer question on whether to transfer affects both players' payoffs. Therefore, it is not unreasonable to think that some players may not give a well-thought out estimate of belief since it has no impact on their payout.

Simultaneously, the seller is presented with a page with one question: to estimate the probability that the buyer believes that the the random transfer was successful. By subtracting that value from one,[4] we get seller's meta belief about the probability that the seller is telling the truth.

Note that in Figure 2 there is an imbalance between the expected total payoffs (for both players) when the the buyer transfers ($2) and when she does not ($1). This could lead to a tendency to transfer more than not transfer in order to maximize total welfare. This altruistic behavior could be expected especially if the buyers find themselves unable

[4]Only for cases where the seller actually said that the transfer was not successful in the chat.

to decide whether they were being lied to or not. In other words, their altruism would tip the scale. That said, there is another force that could tip the scale in the other direction. That is the tendency towards loss aversion (Kahneman and Tversky 1979), in which participants add an additional psychological cost of losing. The buyer may prefer to not transfer than to transfer to avoid this psychological cost. Since these forces point in opposite directions, we will assume for now that they cancel out. Once we get to the results, we will discuss whether we have evidence if either of them is stronger.

Before beginning the games, players were presented with the instructions and then two scale questions: to rate their ability to lie without being detected and their ability to accurately detect others' lies. These questioned were asked at the beginning to avoid being anchored by the performance in the games. Following the games, the particpants were asked a series of demographic questions.

5.2 The Use of Amazon Mechanical Turk (mTurk)

Amazon Mechanical Turk (mTurk) is a platform that allows "workers" to perform tasks offered by "requesters" for financial reward. Requesters choose a flat fee to be paid to the workers for the task. Additionally, requesters may offer additional bonuses. Using this feature, requesters can create tasks that have variable payouts depending on a given worker's performance. This allows for creating behavioral experiments that involve monetary incentives.

The use of mTurk for behavioral experiments carries similar external validity concerns as conducting experiments in university research labs with students. There are certainly certain self-selection biases that could affect the results of academic experiments (Hauser, Paolacci, and Chandler 2018). Those who choose to be workers on mTurk are not necessarily representative of the overall population. For instance, it is unlikely for an upper middle class individual with a stable job to be a worker on mTurk.

While there are "qualifications" such as demographics that allow requester to filter the worker pool assigned to their task, these can often be unreliable. For instance, while requesters can specify which countries they would like to restrict their offer to, workers can use VPNs to spoof their location.

I chose the United States as the location and decided against including other English speaking countries like the United Kingdom or Australia. The reason behind this decision is related to the fact the workers were going to communicating together on a chatbox. Differences in spelling, slang, linguistic register, and so on, could create unnecessary noise in the data. For instance, one may be more trusting of someone with typing in British English leading to an increased likelihood of believing them. As mentioned earlier, this is not a foolproof method as some workers use VPNs to spoof their locations. There were a few instances in our experiment in which the participant clearly did not even have an elementary proficiency in English. However, I reviewed each game's chat logs and those who were unable to communicate due to language impairment or a a clear lack of understanding of the task, were excluded.[5]

The use mTurk in general has been criticized. A study concluded that "workers earned a median hourly wage of only $2/h, and only 4% earned more than $7.25/h" (Hara et al. 2017). However, academic tasks on mTurk generally pay well above $7.25 per hour (the federal minimum wage in the United States). The experiment in this thesis had a participation fee of $3 and a potential bonus of up to $12. Based on my calculations, all participants made at least $10 per hour.

[5]In the data set published for this thesis, the variable "game.useful_convo" indicates the games that were excluded based on this criteria. The actual chat logs for each game are under the variable "game.chat_log".

5.3 Results and Discussion

In the experiment, 314 games were played in six different sessions. In total, 138 people participated in the game with an average of 4.5 games per participant. Of the 112 who reported their gender identity, 61% identified themselves as men and 39% as women. None identified themselves as "other/diverse".

Of the 314 games, 108 games were excluded based in the analysis because the chat logs indicated at least one of the following: (1) at least one player did not send any messages, (2) at least one player was unable to communicate effectively, or (3) at least one player showed a complete lack of understanding of the game and couldn't follow instructions.[6]

Of the remaining 206 games, 90 were randomly assigned the truth treatment (random transfer failed so sellers were given the incentive to tell the truth). The remaining 116 were assigned the lying treatment (random transfer succeeded but sellers were given the incentive to lie to double their payout). In 44 of those 116 games, the seller chose to tell the truth even though they had a monetary incentive to lie. This is in line with Gneezy (2005), Mazar, Amir, and Ariely (2008), Lundquist et al. (2009), Gneezy, Rockenbach, and Serra-Garcia (2013), Fischbacher and Föllmi-Heusi (2013), and Gneezy, Kajackaite, and Sobel (2018). These 44 observations are not useful for certain analyses. This leaves us with 72 games in which the seller lied and 90 games in which the seller told the truth.

5.3.1 Preconceived Impressions of Own Lying and Detection Abilities

We find a sadistically significant correlation between the participants' own lying ability rating and own lying detection rating $R^2 = 0.13$ ($p < 0.001$). This relationship is show in Figure 3. This is in line with our expectations that the ability ratings may be a result of overall confidence. Interestingly, however, we find no statistically significant effect of

[6]The chat logs are available in the data set provided.

own lying ability rating on actual lying ability scores. Similarly, we find no statistically significant effect of own detection ability rating on actual detection ability scores. This suggests that players are unable to accurately estimate their lying and detection abilities.

5.3.2 Altruistic Behavior vs. Loss Aversion

Recall from subsection 5.1 that based on the buyers possible payoffs, they should transfer the amount if they believe the probability that the seller is telling the truth to be $P_T > \widehat{P_T} = 1/2$. Therefore, we expect that—assuming rational behavior and a good understanding on the experiment—buyers will transfer the amount if they indicate that the likelihood that the seller is telling the truth to be higher than 1/2 and do the opposite if less than 1/2 (i.e. P_T is effectively the equivalent of the reported beliefs by the buyer). However, the data shows that only 67% of those who reported a $P_T > 1/2$ actually transferred the amount ("home economicus" would transfer 100% given $P_T > 1/2$). Even more concerning, only 52% of those who reported a $P_T < 1/2$ refused to transfer the amount (again, we'd expect 100% to do so).

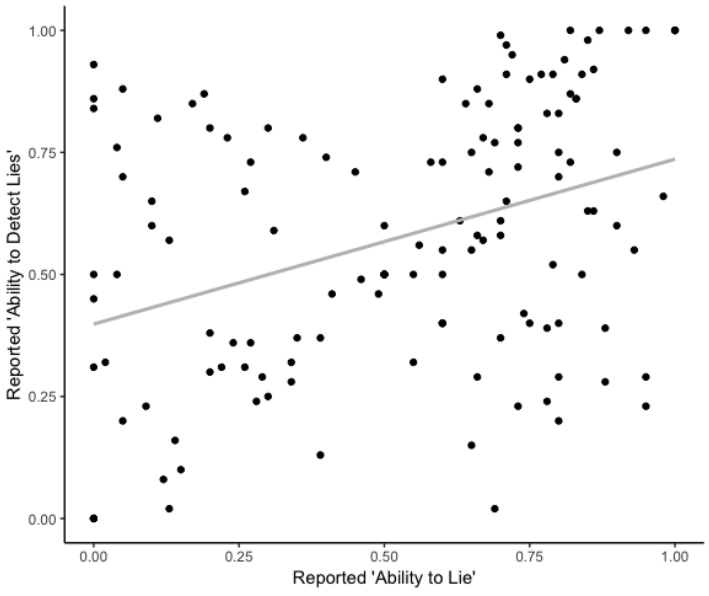

Figure 3: Self Rating on Lying and Detection Abilities

It appears that players have a tendency to lean more towards transferring the amount than what we expect to be rational. This suggests that they may be acting altruistically since, on average, transferring the amount increases total reward. Another explanation

would be that some players are not reading the questions carefully, are not grasping the basic premise of the game, or are simply not paying attention to belief question because it does not affect their payout. However, we would expect such behavior be random rather than systemically biased as we see here. When checking if these numbers change for those who received were asked the probability/belief question before the transfer question and vice versa, we get similar results: (66% vs. 48%) when the transfer question came first, and (68% vs 56%) when the probability/belief question came first. We see indications of this altruistic behavior among sellers as well, since they sometimes choose not to lie when they have a financial incentive to do so.

This altruistic behavior is consistent with Gneezy and Kajackaite (2020) who found that participants were more willing to lie to help increase a fellow player's payoff at the expense of the experimenter and themselves when the stakes are low (comparable to our experiment's stakes). While this possible altruistic behavior may be described as inherent, it might be actually enhanced in our experiment by the fact that it took place on mTurk. Since mTurk has a very clear structure of workers and requesters (who are sometimes criticized for taking advantage workers), workers may view themselves as a "tribe". They may have a "we're in this together" attitude, in which they help each other maximize their payoffs. There are even hints of such fraternal behavior in the chat logs where players will ask each other how their "turking" (i.e. if they're earnings are good that day so far). They also sometimes share their opinions about the experiment and wish each other good luck. This behavior is especially noticeable after the seller altruistically tells the chooses to tell the truth even when it's in their best interest to lie.

5.3.3 Lying Detection and the Illusion of Transparency

In order to check if the overall sample can detect, I performed T-tests to check if buyers beliefs actually change depending on whether the seller was lying or telling the truth. However, I find no statistically significant difference. This is inline with Elaad (2003)

and Vrij (2008) findings.

When I performed analyses similar to those by Gilovich, Savitsky, and Medvec (1998) on our data, I was unable to reach the same results. T-tests similar to those done by Gilovich, Savitsky, and Medvec (1998) do not show any statistically significant differences. This is the case for comparing the sellers' meta beliefs variable to both the buyers' beliefs and transfer variables. This seems to indicate that in the online format, individuals may not experience the illusion of transparency that they do in real life. Perhaps the limited nature of online interactions limits one's ability to feel that their nervousness is "leaking". This may be loosely related to the "online disinhibition effect" in which people are more willing to say things online that they would not be willing to say in real life—even after controlling for anonymity (Lapidot-Lefler and Barak 2012). This disinhibition effect is theorized to be a result of feeling that the internet as a medium offers some level of psychological protection. Similarly, one can imagine how online lying may be associated with less feeling of leakage.

5.3.4 Those who think others believe them, are they actually more likely to be believed?

As mentioned earlier, the buyers' transfer choice actually affects their payoffs while their indicated belief doe not. Therefore, I will use the transfer choice as the primary variable for analysis and the belief choice will be used for robustness checks.

We will primarily regress the buyers' transfer variable on the sellers meta belief variable. A statistically significant positive result would indicate that the instances in which sellers were more confident about their statements being believed, are actually associated with higher rates of being believed. This implies that even if the illusion of transparency does exist, we are still able to accurately predict whether we are being lied to or not after we compensate for the existing bias. In other words, there might exist a case in which we underestimate our meta belief measurements, but the the variability within

these beliefs is actually predictive of the variability within the actual beliefs.

First we run this regression on the entire data set (excluding the non-useful 108 games). Note that this data set includes sellers who chose to altruistically tell the truth. Therefore, for those cases, we coded the belief and transfer variables in the reverse direction. Unsurprisingly, we found a statistically significant positive relationship for both the buyers' transfer variable and their belief variable. However, these results are unremarkable since they include sellers to who admitted to the buyers (against their own interest) that the transfer had already passed and that don't need to transfer again. In these cases, buyers have no reason to be suspicious.

The cases we are interested in are the ones in which the sellers tell the buyers that they have not received the automatic transfer. When we re-run the regression after excluding altruistic truth-tellers, the relationship is not statistically significant ($p = 0.103$ for the buyers' transfer variable. Interestingly however, if we ran the regression only on games in which the initial transfer failed and the sellers still had to convince the buyers that it failed, we find a statistically significant positive effect with a regression co-efficient of 0.0467 ($p = 0.0467$). This is possibly the most striking finding from the experiment. The implication is that when people are telling a self-serving truth, they are able to predict whether the listener believes them or not. Note that the robustness check fails to find a statistically significant result.

6 Conclusion

Incorporating lying into the adverse selection model showed that probabilistic estimation of lying can actually reduce the possibility of a runaway adverse effect that causes a deviation from the Pareto optimal outcome. This shows that incorporating lying into classical economic models can add value. This allows us to have a better understanding of economic interactions and brings us closer to representing reality.

The experimental part of this thesis brought about some intriguing results. However, the results fell short of a breakthrough. However, as far as I know, this is the first lying experiment to be conducted on mTurk. This is a new endeavour and there are definitely lessons learned from this experiment. For instance, in future mTurk experiments, I would recommend reducing the complexity. mTukers tasks are often mindless tasks and therefore asking them to invest a lot of energy into understanding a complex task might be a tall order.

There are also improvements to the experiment itself that I would recommend. First, it would be useful to create a version in which players play more rounds so we can get robust lying and detection accuracy scores for each player. Furthermore, it would be interesting to see if adding an financial incentive on the sellers meta belief estimation could increase that accuracy.

Conducting a similar set up of this experiment in person can be very illuminating. However, I still worry that it might be too complex.

References

Adenzato, Mauro, and Rita Ardito. 1999. “The Role of Theory of Mind and Deontic Reasoning in the Evolution of Deception”.

Akerlof, George A. 1970. “The Market for "Lemons": Quality Uncertainty and the Market Mechanism”. *The Quarterly Journal of Economics* 84, no. 3 (): 488–500.

Chen, Daniel L., Martin Schonger, and Chris Wickens. 2016. “oTree—An open-source platform for laboratory, online, and field experiments”. *Journal of Behavioral and Experimental Finance* 9:88–97. ISSN: 2214-6350. doi:https://doi.org/10.1016/j.jbef.2015.12.001. http://www.sciencedirect.com/science/article/pii/S2214635016000101.

DePaulo, Bella, et al. 1996. “Lying in Everyday Life”. *Journal of personality and social psychology* 70 (): 979–95. doi:10.1037/0022-3514.70.5.979.

Dietz, Simone. 2019. “White and Prosocial Lies”. In *The Oxford handbook of lying*, First Edition, ed. by Jörg Meibauer. Oxford handbooks in linguistics. OCLC: on1079003139. Oxford, United Kingdom: Oxford University Press. ISBN: 978-0-19-873657-8.

Elaad, Eitan. 2003. “Effects of feedback on the overestimated capacity to detect lies and the underestimated ability to tell lies”. *Applied Cognitive Psychology* 17 (3): 349–363. doi:10.1002/acp.871. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.871. https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.871.

Erat, Sanjiv, and Uri Gneezy. 2012. “White Lies”. *Management Science* 58, no. 4 (): 723–733. ISSN: 0025-1909, 1526-5501, visited on 11/01/2020. doi:10.1287/mnsc.1110.1449. http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1110.1449.

Fischbacher, Urs, and Franziska Föllmi-Heusi. 2013. “Lies in Disguise-An Experimental Study on Cheating”. *Journal of the European Economic Association* 11, no. 3 ():

525–547. ISSN: 15424766, visited on 11/01/2020. doi:10.1111/jeea.12014. https://academic.oup.com/jeea/article-lookup/doi/10.1111/jeea.12014.

George, Joey F., and Alastair Robb. 2008. "Deception and Computer-Mediated Communication in Daily Life". *Communication Reports* 21 (2): 92–103. doi:10.1080/08934210802298108.

Gilovich, Thomas, Kenneth Savitsky, and Victoria Husted Medvec. 1998. "The illusion of transparency: biased assessments of others' ability to read one's emotional states." *Journal of personality and social psychology* 75 (2): 332.

Gneezy, Uri. 2005. "Deception: The Role of Consequences". *American Economic Review* 95, no. 1 (): 384–394. ISSN: 0002-8282, visited on 11/01/2020. doi:10.1257/0002828053828662. https://pubs.aeaweb.org/doi/10.1257/0002828053828662.

Gneezy, Uri, and Agne Kajackaite. 2020. "Externalities, stakes, and lying". *Journal of Economic Behavior & Organization* 178 (): 629–643. ISSN: 01672681, visited on 11/01/2020. doi:10.1016/j.jebo.2020.08.020. https://linkinghub.elsevier.com/retrieve/pii/S0167268120302900.

Gneezy, Uri, Agne Kajackaite, and Joel Sobel. 2018. "Lying Aversion and the Size of the Lie". *American Economic Review* 108, no. 2 (): 419–453. ISSN: 0002-8282, visited on 11/01/2020. doi:10.1257/aer.20161553. https://pubs.aeaweb.org/doi/10.1257/aer.20161553.

Gneezy, Uri, Bettina Rockenbach, and Marta Serra-Garcia. 2013. "Measuring lying aversion". *Journal of Economic Behavior & Organization* 93 (): 293–300. ISSN: 01672681, visited on 11/01/2020. doi:10.1016/j.jebo.2013.03.025. https://linkinghub.elsevier.com/retrieve/pii/S016726811300070X.

Hara, Kotaro, et al. 2017. *A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk*. arXiv: 1712.05796 [cs.CY].

Hauser, David, Gabriele Paolacci, and Jesse J. Chandler. 2018. *Common Concerns with MTurk as a Participant Pool: Evidence and Solutions*. Preprint. PsyArXiv. Visited on 09/26/2020. doi:10.31234/osf.io/uq45c. https://osf.io/uq45c.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk". *Econometrica* 47, no. 2 (): 263. doi:10.2307/1914185. https://doi.org/10.2307/1914185.

Kant, Immanuel. 1797. *On a Supposed Right to Tell Lies from Benevolent Motives*.

Lapidot-Lefler, Noam, and Azy Barak. 2012. "Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition". *Computers in Human Behavior* 28 (2): 434–443. ISSN: 0747-5632. doi:https://doi.org/10.1016/j.chb.2011.10.014. http://www.sciencedirect.com/science/article/pii/S0747563211002317.

Lundquist, Tobias, et al. 2009. "The aversion to lying". *Journal of Economic Behavior Organization* 70 (1-2): 81–92.

Mahon, James Edwin. 2019. "Classic philosophical approaches to lying and deception". In *The Oxford handbook of lying*, First Edition, ed. by Jörg Meibauer. Oxford handbooks in linguistics. OCLC: on1079003139. Oxford, United Kingdom: Oxford University Press. ISBN: 978-0-19-873657-8.

Mann, Samantha. 2019. "Lying and Lie Detection". In *The Oxford handbook of lying*, First Edition, ed. by Jörg Meibauer. Oxford handbooks in linguistics. OCLC: on1079003139. Oxford, United Kingdom: Oxford University Press. ISBN: 978-0-19-873657-8.

Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green. 2011. *Microeconomic theory*. Oxford Univ. Press.

Mazar, Nina, On Amir, and Dan Ariely. 2008. “The Dishonesty of Honest People: A Theory of Self-Concept Maintenance”. *Journal of Marketing Research* 45, no. 6 (): 633–644. ISSN: 0022-2437, 1547-7193, visited on 10/16/2020. doi:10.1509/jmkr.45.6.633. http://journals.sagepub.com/doi/10.1509/jmkr.45.6.633.

Piaget, Jean. 1977. *The essential Piaget*. London: Routledge / K. Paul. ISBN: 978-0710087782.

Schiestl, Florian P. 2005. “On the success of a swindle: pollination by deception in orchids”. *Naturwissenschaften* 92, no. 6 (): 255–264. doi:10.1007/s00114-005-0636-y. https://doi.org/10.1007/s00114-005-0636-y.

Spence, Michael. 1973. “Job Market Signaling”. *The Quarterly Journal of Economics* 87, no. 3 (): 355. ISSN: 00335533, visited on 10/27/2020. doi:10.2307/1882010. https://academic.oup.com/qje/article-lookup/doi/10.2307/1882010.

Stiglitz, Joseph E. 1975. “The Theory of "Screening," Education, and the Distribution of Income”. *The American Economic Review* 65, no. 3 (): 283–300.

Tarantino, Quentin. 2009. *Inglorious Basterds*. Universal.

Vrij, Aldert. 2008. *Detecting lies and deceit: pitfalls and opportunities*. 2. ed. Wiley series in the psychology of crime, policing and law. OCLC: 254185238. Chichester: Wiley. ISBN: 978-0470516256.