# Conforming and Signaling Conformity

William Radaic*

December 11, 2024

**Abstract**

People often act under social pressure, choosing whether to conform to popular tastes. I identify two distinct perceived sources of social pressure: *action-based* judgment, in which approval depends on choosing an action that matches the audience's opinions; and *taste-based* judgment, in which audiences approve of people who are believed to share their private tastes. While action-based judgment creates a simple choice under uncertainty, taste-based judgment induces a signaling game. I formally define how these two perceived sources of social judgment translate into utility functions, and using a simple static game, draw out the implications of each assumption for behavior and welfare. While the results are structurally similar, they are driven by distinct mechanisms and generate some contrasting results. Action-based judgment leads to unique predictions, with unwavering approval from the share of players who were pleased with the observed action. In contrast, taste-based judgment generates multiple equilibria, and can leave observers uncertain even after interacting under an uninformative pooling equilibrium. I discuss these results in light of the current literature on conformity and social norms, and intuitively consider extensions in which each source of social judgment generates distinct predictions.

**Keywords:** conformity, social norms, signaling.
**JEL:** D83, D85, Z13.

## 1 Introduction

People often have to decide how to behave under social pressure. Over the course of a regular day, one may feel compelled to laugh at an unfunny joke from his boss, to say that he agrees with a co-worker's opinion on the latest socially-charged trending topic, to accept boring social plans with his friends, and to introduce his long-time boyfriend as a friend to his grandparents in order to hide his sexual orientation. Even though all of these conforming behaviors are driven by a fear of potential social repercussions, upon closer inspection, it is possible to identify two distinct sources of social pressure between them. Specifically, social judgment may be directed at one's *actions* or *tastes*.

*Pre-Doctoral Fellow, Department of Economics, Harvard University (wradaic@fas.harvard.edu). I thank Ophir Averbuch, Leon Eliezer, Jared Fang, Chris Liao, Gabriel Montaño, João Ramos, and participants at the Harvard Pre-Doc Workshop for helpful comments and suggestions, and especially Daniel Monte for thoughtful advising and Matthew Rabin for guidance throughout the project.

To make this distinction clear, consider the following motivating example. Ian is a closeted gay man, and is considering whether to come out to his potentially intolerant family. Ian would like to express his sexuality at ease and to be truthful to his family but, of course, fears that his relatives may react badly. The potential negative reaction of (say) intolerant grandparents may be driven by two distinct feelings. Maybe the grandparents don't mind that gay people exist, or even that Ian is gay in abstract, as long as he does not *act gay* in their presence. This first possibility illustrates *action-based* social judgment. Another possibility, which I call *taste-based* social judgment, is that the intolerant grandparents have an intrinsic distaste for LGBT people, *regardless* of how they act, dress, or speak.

Both alternatives are unpleasant for Ian; however, these different sources have potentially distinct implications. In particular, the second case seems more nuanced and insidious. If all the grandparents care about is Ian's public behavior, he can still come out and "act straight", changing nothing about how he's treated by his family. On the other hand, if he comes out and the grandparents are intrinsically homophobic, there is nothing Ian can do to avoid the rude treatment he for sure will receive. The grandparents learned that Ian has an immutable trait they inherently dislike; no matter how much he tries to obfuscate his sexual orientation by acting straight, he will still be judged negatively by the grandparents.

In this paper, I formalize these two distinct perceived sources of social pressure in a simple game-theoretic environment and draw out their main implications for behavior and welfare. I have in mind a setting in which people in a large group are simultaneously asked their opinions on a socially-charged topic—for instance, Ian's family is celebrating Thanksgiving and talking about LGBT rights over dinner. Each family member is either progressive or conservative, and can state a progressive or a conservative stance. Ian would like to come out as gay, but fears the potential backlash of revealing his sexual orientation to the family. Similarly, Ian's grandparents fear that their intolerance will be met with vigorous criticism from the rest of the family. Indeed, every family member faces a similar dilemma, varying with their privately-held opinions. While every relative has an accurate estimate over the *size* of the majority, they are uncertain about the majority opinion in the family. Formally, this environment is represented by a one-shot simultaneous game with binary actions, types, and states of the world, and a continuum of agents who are certain about majority size but uncertain about majority type.

I use this underlying structure to compare two different extreme cases, in which every family member believes in the same source of social judgment, and puts the same relative weight on social judgment relative to the intrinsic utility of being honest. In the case of action-based social judgment, an agent is positively judged whenever their public *action* matches an audience member's *type*. For instance, if Ian decides to express his sexual orientation, his progressive aunt will judge him positively and his intolerant grandfather will judge him negatively, based purely on how he acts in front of the family. Note that in this case, social judgment is not driven by matching private tastes, but rather by pleasing the audience through public actions. When Ian is deciding how to act during dinner, he simply considers his beliefs about the tastes of each family member, and acts in accordance to the type he thinks is most likely to be dominant (conditional on the expected social judgment outweighing his intrinsic utility). Crucially, since social judgment is directly derived from his action, Ian does not worry about what his action signals about his tastes. This case is similar to the situation described above in which intolerant grandparents do not care if Ian is gay, as long as he doesn't act gay in front of them. All he has to do to be fully accepted by his grandparents in this scenario is to act straight.

In contrast, with *taste-based* social judgment, family members are effectively signaling their types to each other. Ian's relatives now update their beliefs about his tastes after observing how *everyone* acted, and base their judgment of him on the probability that he shares their type. This means that Ian now has to consider, on top of the composition of types in his family, also his beliefs about how his entire family is going to act, since the observed strategy profile directly affects people's inferences over others' types (and thus payoffs). For instance, if his family is widely believed to be mostly conservative and everyone pools on conservative actions, then even if he acts straight in front of them, his relatives may still have second thoughts about his sexuality. This is driven by the fact that in a pooling equilibrium, actions are completely uninformative, leaving players with unchanged beliefs about each agent's types after interacting. If, on the other hand, the family is believed to be open and truthful with each other, then Ian's public action will be taken as a reliable and diagnostic signal of his sexuality.

Both sources of social judgment yield results with similar structures in this one-shot setting. Depending on the combination of the common prior over majority type, and the commonly known majority size, agents play either pooling or separating equilibria. Considering first an individual's incentives, he will only consider acting against his taste if the common prior is *unfavorable* to his type. Intuitively, this means that conformity is only appealing when an agent faces a trade-off between deriving utility from honesty and (expected) utility from social judgment. If the common prior is favorable to his type, both of these components increase when acts honestly. Furthermore, this trade-off needs to be sufficiently binding for pooling equilibria to be stable. If the threat of social judgment is small relative to the intrinsic reward of acting honestly—which can happen either when the prior over majority type is not sufficiently unbalanced, or when the group is perceived to be roughly equally divided—then all agents will choose to reveal their types, leading to a separating equilibrium. Conformity, here represented by pooling equilibria, depends on a combination of a sufficiently unbalanced prior and a sufficiently dominant majority.

Furthermore, an increase in majority size has two opposite effects on the stability of pooling equilibria. First, a larger majority means that your privately-known type is a more informative signal of the state of the world. This *information effect* pushes the player in the direction of speaking his mind. On the other hand, increasing majority size increases the threat of social judgment, since the group is more starkly divided—which increases the wedge in judgment between each available action. The *judgment effect* raises the importance of social judgment relative to intrinsic utility, making the agent more likely to conform. These effects operate with different magnitudes as majority size grows, inducing a quadratic shape for the space of pooling equilibria.

Even in this simple environment, however, some results are different across both sources of social pressure. Whereas action-based social judgment implies generally unique equilibria, the signaling dynamics inherent to taste-based social pressure give rise to multiple equilibria. This difference is due to the fact that posited strategy profiles affect expected payoffs for all available actions in the signaling game. In contrast, the decision rule under action-based social pressure depends purely on the agent's beliefs over others' types and the relative sensitivity to social judgment. Furthermore, even if ex-ante expected payoffs are similarly structured, *ex-post payoffs* can be notably different in each of these worlds. In a pooling equilibria, an agent's ex-post social utility is composed of *certain* positive judgments from the *share* of players that match the pooled action. For instance, if Ian pools on the conservative action, only the conservative members of his family will approve of him, and do so unwaveringly. On the other hand, the very nature of a pooling equilibrium in a signaling game is that it is uninformative. This leaves all players

with *uncertain* social judgments from every player after interacting. If Ian's family believes in taste-based social judgment and Ian pools on the conservative action, all family members will remain unsure about his true tastes. The privately progressive members who also hid their opinions will have some remaining faith that he's one of them, and the conservatives will have some second thoughts about Ian, not fully trusting that he's a true conservative.

This paper is motivated by an extensive literature in Economics and other social sciences on conformity. Timur Kuran's book *"Private Truths, Public Lies"* (Kuran 1995) is a prime example of this literature, discussing compellingly the prevalence of what he calls *preference falsification* in our daily lives, its potential implications for economically-relevant settings, and proposing a formal model of such phenomenon. However, while the main examples and motivations he presents in the book are closer in spirit to taste-based social pressure, the model he writes down actually assumes action-based social pressure. Signaling-based models have also been proposed in the literature, such as Bernheim (1994)'s seminal model of conformity, and more recently (with a similar structure to mine) Bursztyn, Egorov, and Fiorin (2020)'s motivating model for their experimental design. Aside from Kuran (1989, 1991, 1995), other models of conformity with action-based social judgment have similarly been proposed (Michaeli and Spiro 2017; Duffy and Lafky 2021; Smerdon, Offerman, and Gneezy 2020).

My view is that this literature has overlooked this subtle but potentially relevant distinction in the sources of social judgment, which vary in importance across settings and can lead to distinct predictions. For instance, *pluralistic ignorance*—a situation in which most people believe that they hold the minority opinion about a socially charged topic, leading them to conform to the false social norm—is documented by Bursztyn, González, and Yanagizawa-Drott (2020), who show that men misperceive support for women working outside their homes in Saudi Arabia. In these environments, taste-based judgment seems to be the main driver of social pressure.[1] In fact, the model proposed by Fernández-Duque (2022) to study pluralistic ignorance assumes taste-based social pressure. On the other hand, social situations in which organized collective action is the driver of social pressure, such as political parties campaigning or congressmen whipping votes, seem to be more suited to be theoretically analyzed by a model that assumes action-based social pressure.

This paper is best interpreted as an initial exercise that attempts to clearly delineate these two sources of social judgment, and draw out some implications for observed behavior and welfare. While limited by the simple and coarse game structure assumed here, this exercise highlights clearly some specific implications and sheds light on the distinct mechanisms through which these results arise, and can serve as a guide for more extensive and in-depth work on social conformity.

The paper is also related to a few broader topics in economic theory. Aside from a signaling game, taste-based social judgment can be interpreted as a *psychological game*, first proposed and formalized by Geanakoplos, Pearce, and Stacchetti (1989), which models situations where beliefs about others' strategies influence payoffs. It also relates to a growing literature in belief-based utility, which rests on the observation that beliefs can affect people's hedonic experiences directly (see Loewenstein and Molnar 2018 for a brief review).

The rest of the paper will proceed as follows. Section 2 presents the game structure and formally defines the utility functions that represent each source of social judgment. Section 3

[1] Other documented instances of pluralistic ignorance include support for racial segregation (O'Gorman 1975), the soviet regime (Kuran 1991), female genital cutting in Somalia (Gulesci, Jindani, Ferrara, Smerdon, Sulaiman, and Young 2023), and alcohol consumption in college campuses (Prentice and Miller 1993).

then briefly goes over the relevant belief-updating processes for the game. The main results are presented in Section 4, and further discussions and extensions are presented in Section 5. The final section concludes the paper.

# 2 Setup

## 2.1 Game Structure

In this model, social interaction is represented by a simple static, one-shot game, with binary actions, types, and states of the world.

Formally, the society is represented by a continuum of Bayesian agents, indexed by $i \in I$ over the interval $[0, 1]$. Each agent has a privately known type $t_i \in \Theta \equiv \{h, l\}$. For example, we can interpret $h-$types as progressives, and $l-$types as conservatives. There is uncertainty over the state of the world, denoted by $w \in \Omega \equiv \{H, L\}$, which determines the majority type. For instance, in state $w = H$, most people are progressives, whereas if $w = L$, then the majority is conservative. Agents have common priors over the state of the world, and are denoted by $p_H \equiv \mathbb{P}(w = H)$ and $p_L \equiv \mathbb{P}(w = L) = 1 - \mathbb{P}(w = H)$. Agents' types are independent conditional on the state realization $w$. Action sets are binary: each player chooses $a_i \in A_i \equiv \{h, l\}$. Following our example, agent $i$ can either state a progressive ($a_i = h$) or a conservative ($a_i = l$) opinion.

Majority size is formally defined as $\mu \equiv \mathbb{P}(t_i = \theta \,|\, w = \theta) \in (1/2, 1)$.[2] In more Bayesian terms, $\mu$ can be interpreted as the *precision* of the signal contained in an agent's private type. Following Fernández-Duque (2022), I assume that $\mu$ is symmetric across states and common knowledge. Intuitively, this means that all players know the size of the majority with certainty, but are unsure about the most prevalent type. For example, if $\mu = 0.7$, agents are uncertain between the group being composed of 70% progressives and 30% conservatives, or 30% progressives and 70% conservatives.

While I recognize that the combination of certainty over majority *size* and uncertainty over majority *type* may not be the most natural formal representation of the informational structure of many relevant settings, I take these assumptions for a couple of reasons. First, it unbundles a player's beliefs regarding others' types into two distinct components—majority *type* and majority *size*—which allows for comparative statics on two distinct dimensions of uncertainty. Second, common knowledge over majority size makes Bayesian inference more tractable, especially for computing second-order beliefs—which are crucial to analyzing the implications of taste-based social pressure.

Figure 1 summarizes the timeline of the game. First, nature draws the state of the world $w$ according to the common prior $(p_H, p_L)$, which determines agents' types according to majority size $\mu$. Players observe their types and use this information to update their beliefs over the state of the world with Bayes' Rule. Then, agents play simultaneously, choosing between stating $a_i = h$ or $a_i = l$. Finally, payoffs are revealed.

[2]Throughout this paper, I slightly abuse notation by treating equalities in probabilities as case-insensitive. Specifically, if the state is $w = H$ and agent $i$ has type $t_i = h$, I will interpret $t_i = w$ as true. This approach simplifies notation and allows for more concise general statements.

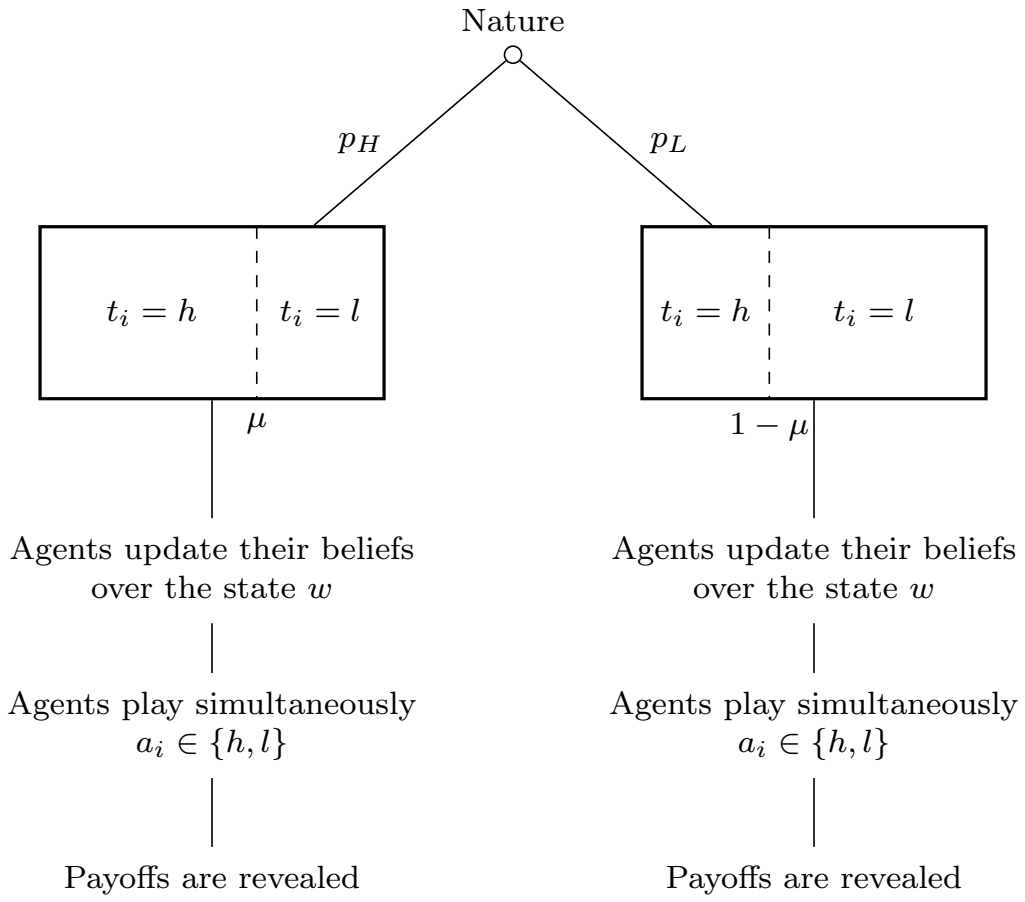



Figure 1: Game timeline

## 2.2 Preferences

The distinction between sources of conformity-inducing social pressure will be represented formally with different utility functions. I propose below two utility functions that share the same hedonic trade-off between truth-telling and conforming to the majority sentiment, but differ in the perceived foundations of social pressure. Throughout the paper, I analyze cases in which all agents share the same utility function—heterogeneity in people's perceived sources of social pressure is an extension for future work.

For clarity in exposition, it will be often useful to frame these distinctions in terms of the *audience*—i.e., the $-i$ players who observe agent $i$'s statement and form their judgments about $i$. Note that the audience is *not* a distinct group of players, since all agents act and form beliefs about others simultaneously.

In this model, the audience's judgment can be interpreted as a reduced-form representation of all hedonically-relevant social repercussions of a public action or a belief about one's type. For instance, it is capturing situations such as a gay couple being mistreated at a restaurant by an intolerant waiter; a worker being fired because the manager learned discovered his socialist affiliations; a family being torn apart by political divergences; a movie star losing future roles following allegations of sexual assault; or a producer being shunned from the film industry after a

wave of reports on his abusive behavior.

### 2.2.1 Action-based social judgment

We first consider Stanley, an agent who believes that the audience judges him purely based on the opinion he publicly states. Stanley's utility function is given by $U^A : A_i \to \mathbb{R}$

$$U^A(a_i \,|\, t_i) := \mathbb{1}\{a_i = t_i\} + \beta_A \int_I \mathbb{E}_i[\mathbb{1}\{a_i = t_j\} \,|\, t_i] \; dj \tag{2.1}$$

with $\beta_A > 1$ fixed and homogenous for all agents, and common knowledge.[3] Stanley has two dimensions of utility. Using the term coined by Kuran (1995), *intrinsic utility* represents the payoff the agent receives from publicly stating their true taste. The last term in Equation 2.1 represents Stanley's *social-judgment utility*—i.e., the payoff he receives when others judge him positively. Following our example, Stanley's intolerant grandparents will only judge him negatively if he *acts* gay, and positively if Stanley acts straight—even if they are aware of his sexual orientation. Accordingly, Stanley would judge his grandparents positively if the grandparents treated him and other gay people well, regardless of their private attitudes regarding the LGBT population. The parameter $\beta_A$ represents the *relative weight* Stanley puts on social judgment compared to intrinsic utility. As $\beta_A$ grows, Stanley values social approval more relative to speaking his mind.

Crucially, note that an agent's type is irrelevant for the audience's judgment; the only directly relevant factor for social judgment is the agent's publicly-observed action. Furthermore, an agent's expected payoff *does not* depend on the strategy profile being considered; it is driven only by the agent's own action, and his beliefs regarding others' *types*. This rules out all signaling dynamics and eliminates second-order beliefs from the agent's decisionmaking.

We can think of Stanley's environment as analogous to an expected-utility maximizer choosing between risky prospects. Lottery $a_i = h$ yields some positive social-judgment utility if $h$ is the majority, which happens with probability $\mathbb{P}(w = h \,|\, t_i)$, and a lower social-judgment utility if $l$ is the majority, with associated probability $1 - \mathbb{P}(w = h \,|\, t_i)$. Accordingly, lottery $a_i = l$ returns the larger social-judgment payoff with probability $\mathbb{P}(w = l \,|\, t_i)$ and the smaller payoff with probability $\mathbb{P}(w = h \,|\, t_i)$.

### 2.2.2 Taste-based social judgment

We now turn our attention to Sophie, who believes that social pressure comes from congruence in private attitudes. From the perspective of the audience, this means that Sophie is judged positively when the audience member believes they share the same tastes.

Formally, Sophie's utility function is given by $U^T : A \to \mathbb{R}$ such that

$$U^T(a_i \,|\, t_i) := \mathbb{1}\{a_i = t_i\} + \beta_T \int_I \mathbb{E}_i \left[ \mathbb{P}_j \left( t_i = t_j \,|\, t_j, a_i, a_{-i} \right) \,\middle|\, t_i \right] \; dj \tag{2.2}$$

in which $\beta_T > 1$ is fixed and homogenous for all agents, and common knowledge. Subscripts in expectations and probabilities refer to which agent holds these beliefs. For example, $\mathbb{P}_j \left( t_i = t_j \,|\, t_j, a_i, a_{-i} \right)$ refers to the probability agent $j$ attributes to sharing a type with $i$ given $t_j$ and the observed action profile.

[3] This condition rules out cases in which intrinsic utility strictly dominates social-judgment utility.

In Sophie's model of the world, an audience member approves her whenever they think Sophie shares their type. Again following our opening example, Sophie's intolerant grandparents will judge her positively if they believe Sophie is straight, and negatively if they believe Sophie is a lesbian. Crucially, the grandparents do not *intrinsically* care about how Sophie acts, dresses, and speaks, or if she has been sexually active with same-sex partners. Instead, the grandparents have an inherent and immutable distaste for anyone who identifies as LGBT, regardless of how they act. Similarly to Stanley's case, note that $\beta_T$ represents the relative weight Sophie puts on social judgment relative to intrinsic utility.

The above example illustrates the key role played by actions in this environment. In contrast to Stanley's world, where actions directly affect judgments—and hence social-judgment utility—Sophie believes that actions serve to *signal* her type. Whereas Stanley's decision is similar to a choice over risky prospects, Sophie is playing a signaling game—one in which she's both a messenger for her own type, and a receiver of others' messages about their types. Furthermore, since the audience's beliefs are unobserved to Sophie, she must rely on her second-order beliefs to guide her decisionmaking. Given the informational structure of the game, with a common prior and a common-knowledge majority size, all agents are able to correctly compute not only their beliefs, but also the beliefs of agents who have the opposite type. Thus, Sophie's uncertainty regarding others' beliefs is pinned down by her Bayesian expectations given the primitives of the model. Note that in this scenario, the agent's beliefs over the strategy profile is directly relevant for behavior and welfare, connecting this setting to psychological games (Geanakoplos, Pearce, and Stacchetti 1989; Battigalli and Dufwenberg 2022).

In closing the presentation of these two alternative utility functions, it is helpful to intuitively consider some of their distinct implications for behavior and highlight the forces at play in each case. For example, suppose that in addition to their own type, agents also observe all players' types. This allows players to learn the state of the world and the share of agents of each type with certainty. Now consider the choice of a minority agent who is extremely susceptible to social pressure ($\beta \to \infty$), such that social-image concerns are the only relevant factor for decisionmaking. With complete knowledge over the composition of the group, Stanley is able to always please the majority, guaranteeing the largest possible (ex post) payoff. Following our main example, if Stanley found out that most of his family is intolerant, he will refrain from acting gay near them, and enjoy social approval from the intolerant relatives—even when they know for a fact he is gay.[4]

On the other hand, Sophie cannot affect others' judgment of her, since her type was revealed to every other member of society, and signaling devices become ineffective in the absence of uncertainty. Social judgment thus becomes unavoidable, and entirely unaffected by the action she takes. Sophie will thus take the minority action, since it maximizes her intrinsic utility and does not affect the judgment she receives from the audience. Going back to the opening example, conditional on finding out about Sophie's sexual orientation, the intolerant grandparents will forever judge her negatively, even if, seeking their approval, she chooses to "act straight" with her family, or even if she chooses to cease all same-sex interactions. To the grandparents, what matters is that Sophie is not straight, and nothing could ever change their beliefs and the negative judgments they have against her.[5]

[4]For example, following the empirical example from Bursztyn, Egorov, and Fiorin (2020), if Stanley learns that most citizens in their neighborhood voted for Trump, then he will conform and publicly support Trump as well, receiving social approval from the majority.

[5]Of course, these statements all hinge on the assumption that types are fixed for all agents. In the example above, this translates to assuming that (i) sexual orientation is not a choice, but rather an intrinsic trait of an

# 3 Belief Updating

Before presenting the main results for each source of social-judgment utility, I walk through the belief-updating process in the game, as it provides useful objects and intuitions to think through the results.

Recall that an agent's private type can be interpreted as a *signal* regarding the underlying state of the world. Thus, when agents observe their types, they go through a round of Bayesian updating. For a given common prior $p_h$ and majority size $\mu$, agent $i$ with type $\theta$ believes that the state of the world is $\theta$ with probability

$$\begin{aligned}\pi_\theta \equiv \mathbb{P}_i(w=\theta \,|\, t_i=\theta) &= \frac{\mathbb{P}(t_i=\theta \,|\, w=\theta)\cdot\mathbb{P}(w=\theta)}{\mathbb{P}(t_i=\theta \,|\, w=\theta)\cdot\mathbb{P}(w=\theta)+\mathbb{P}(t_i=\theta \,|\, w\neq\theta)\cdot\mathbb{P}(w\neq\theta)}\\ \pi_\theta &= \frac{\mu\cdot p_\theta}{\mu\cdot p_\theta+(1-\mu)\cdot(1-p_\theta)},\end{aligned} \tag{3.1}$$

in which $p_\theta$ represents the common prior that $\theta$ is the majority type. In other words, $\pi_\theta$ denotes the posterior probability that an agent of type $\theta$ assigns to being in the majority. Given the binary structure of the game, an agent of type $\theta$ attributes probability $(1-\pi_\theta)$ to being in the minority.

Furthermore, in deciding which action to state, agent $i$ considers the probability that a member of the majority $j$ shares her type. Formally, agent $i$ with type $\theta$ believes that agent $j$ also has type $\theta$ with probability

$$\begin{aligned}q_\theta \equiv \mathbb{P}_i(t_j=\theta \,|\, t_i=\theta) &= \mathbb{P}_i(t_j=\theta \,|\, w=\theta, t_i=\theta)\cdot\mathbb{P}_i(w=\theta \,|\, t_i=\theta)\\ &\quad+\mathbb{P}_i(t_j=\theta \,|\, w=\tilde{\theta}, t_i=\theta)\cdot\mathbb{P}_i(w=\tilde{\theta} \,|\, t_i=\theta)\\ q_\theta &= \pi_\theta\mu+(1-\pi_\theta)(1-\mu),\end{aligned} \tag{3.2}$$

where $\widetilde{\theta}$ denotes the opposite of $\theta$.[6] Given the binary structure of the game, agent $i$ with type $\theta$ attributes probability $(1-q_\theta)$ to agent $j$ not sharing her type.

Given the combination of a common prior and commonly known majority size, every agent is able to compute posteriors about the state of the world and the probability of a member of the audience sharing her type. That is, an agent of type $t_i=h$ knows not only the precise value of $\pi_h$ and $q_h$, but also of $\pi_l$ and $q_l$. This will be instrumental in working through agents' second-order beliefs when we analyze taste-based social pressure.

# 4 Results

Unless otherwise mentioned, I will present the results by considering the decisionmaking process of an agent with type $t_i=h$. Analogous conditions can be obtained for $t_i=l$ by symmetry. I also assume that agents reveal their types ($a_i=t_i$) when indifferent. All proofs are presented in the Appendix.

individual; (ii) people cannot be persuaded to change their views regarding the LGBT population.

[6] Formally, $\widetilde{\theta}$ is the (unique) element of the set $\Theta\backslash\{\theta\}$.

## 4.1 The Stanley Parable

I begin by assuming that all agents in the group believe in action-based social pressure—i.e., that all agents are Stanleys. Given the simple structure of the game and the lack of signaling dynamics for Stanley, each combination of majority size $\mu$, common prior over majority type $p_h$, and relative sensitivity to social pressure $\beta_A$ uniquely determine the Bayes-Nash Equilibrium of the game. This result is presented below.

**Proposition 1** (Characterizing Equilibria for Stanley)**.** *If all agents have Stanley's utility function (2.1), for any $\beta_A > 1$, majority size $\mu \in (1/2, 1)$ and common prior over majority type $(p_H, p_L)$, there exists a unique Bayes-Nash Equilibrium.*

*(i) If $\pi_h \geq \frac{1}{2} - \frac{1}{2\beta_A(2\mu-1)}$ and $\pi_l < \frac{1}{2} - \frac{1}{2\beta_A(2\mu-1)}$, then agents pool on h: for all $i \in I$, $a_i = h$;*

*(ii) If $\pi_h \geq \frac{1}{2} - \frac{1}{2\beta_A(2\mu-1)}$ and $\pi_l \geq \frac{1}{2} - \frac{1}{2\beta_A(2\mu-1)}$, then agents truthfully reveal their types: for all $i \in I$, $a_i = t_i$;*

*(iii) If $\pi_h < \frac{1}{2} - \frac{1}{2\beta_A(2\mu-1)}$ and $\pi_l \geq \frac{1}{2} - \frac{1}{2\beta_A(2\mu-1)}$, then agents pool on l: for all $i \in I$, $a_i = l$.*

Figure 2 illustrates Proposition 1 visually, plotting majority size $\mu$ on the $x$ axis, and the common prior over majority type on the $y$ axes. Given that the state of the world is binary, any point on the coordinate system pins down not only $p_H$ but also $p_L = 1 - p_H$, with $p_L$ growing in the opposite direction as $p_H$.

First, note that if agent $i$'s type is favored by the prior, she prefers to reveal her type. This result actually holds more generally: under an unbalanced prior, the dominant strategy of the favored type is to tell the truth.

**Corollary 1.** *If $p_\theta > 1/2$ for some $\theta \in \Omega$, then the dominant strategy for all Stanleys with type $t_i = \theta$ is to reveal their type, $a_i = \theta$.*

This result is intuitive and highlights the key trade-off an individual potentially faces in the model. If the prior favors $i$'s type, then both intrinsic utility and expected social-judgment are higher if $i$ reveals her type. The trade-off between speaking one's mind and conforming under social pressure is only binding when there is a reasonably good chance that $i$ is part of the minority.

Now consider the comparative statics of an increase in majority size. Note that the spaces of pooling equilibria have a roughly quadratic shape. That is, as $\mu$ grows, the space of pooling equilibria first increases and then decreases. We can interpret this shape by considering a conceptual decoupling of the effect of increasing $\mu$. First, following the Bayesian interpretation of $\mu$ as the *precision* of the private signal, an increase in $\mu$ means that agents' types are more informative of the underlying state of the world. I call this first force the *information effect*. Second, an increase in $\mu$ raises the stakes of potentially acting against the majority relative to intrinsic utility, since the wedge in (ex post) social-judgment utility between actions grows with $\mu$. I call this second component the *judgment effect*. In the lower range of $\mu$, the judgment effect dominates the information effect, and the opposite is observed if $\mu$ is high enough.

To see this two effects at play, it is instructive to consider the limit cases and an intermediate value of $\mu$. If $\mu \to 0.5$, then: (i) the information effect is weak, since the private signals are barely informative of the underlying state; (ii) the judgment effect is also weak, given that a slim

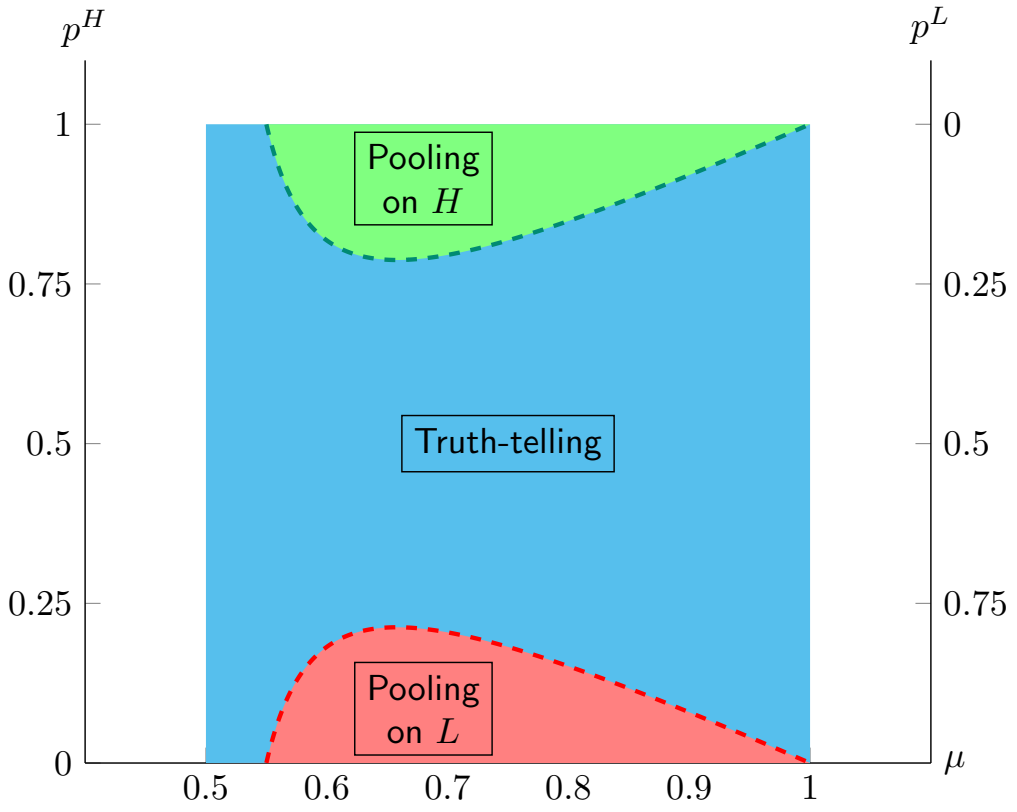



Figure 2: Equilibria for Stanley, $\beta_A = 10$.

majority implies only a slight wedge in social-judgment utility between both actions. In this environment, intrinsic utility dominates social pressure, and agents prefer to tell the truth. In an intermediate value of $\mu$ (say, $\mu = 0.65$ for $\beta = 10$, as seen in Figure 2), we observe an increase in the likelihood of conformity, since the private type is not terribly informative of the state of the world while the risk of losing social approval grew at a faster pace. Here, given the dominance of the judgment effect, the agent becomes more likely to conform. A further increase in $\mu$ flips the dominance of these effects. Finally, if $\mu \to 1$, then the private signal is borderline diagnostic of the state of the world: an $h-$type is extremely confident that the world is $H$, and an $l-$type is extremely confident of the opposite. Even though the threat of losing social approval is at its highest here, Stanley's beliefs grow so confident that they are ex ante better off by telling the truth.

Furthermore, note that an increase in $p_H$ reduces monotonically the prevalence of pooling on $a_i = l$. This is intuitive and readily apparent when considering the conditions outlined in Proposition 1: from Bayes' Rule, we know that an increase in the base rate for hypothesis $\theta$ increases its posterior after observing a signal for $\theta$. Finally, as expected, an increase in the social-judgment scalar $\beta_A$ increases the space of pooling equilibria, since it increases the likelihood that social-judgment utility dominates intrinsic utility. For comparison, Figure 4 in Appendix B replicates Figure 2 for a higher value of $\beta_A$.

Going back to the coming-out example, the model predicts the possible outcomes that may be observed from Stanley's interaction with his intolerant grandparents (who also believe in action-based social pressure). First, if they perceive the family as a starkly divided group on

LGBT issues ($\mu \approx 1/2$), then Stanley will come out and act gay, and the grandparents will judge him negatively, revealing their intolerance. If they perceive the family to be substantially divided, but with a clear majority, and much more likely to be conservative than progressive, then Stanley will act straight, and the grandparents will feel free to act intolerant. On the other hand, if the prior heavily favors progressive attitudes, Stanley will come out, and the intolerant grandparents will feel compelled to hide their intolerance.

We can similarly frame the predictions of the model in light of the experimental evidence in Bursztyn, Egorov, and Fiorin (2020), who show that a perceived increase in community support for Trump in the 2016 election increases people's willingness to contribute to xenophobic causes.[7] In the context of my model, labeling $w = H$ as a Democrat majority and $w = L$ as a Republican majority, the information treatment amounts to an increase in $p_L$ (or equivalently a decrease in $p_H$), which pushes the equilibrium away from pooling on progressive actions into truth-telling, since xenophobic Republicans now feel more comfortable acting in accordance to their true beliefs.

We can also consider the model's predictions for *pluralistic ignorance* (e.g., Bursztyn, González, and Yanagizawa-Drott 2020). In the context of this model, pluralistic ignorance happens whenever agents pool on some action, but nature had drawn the opposite state. For instance, the family may believe that their members are much more likely to be conservative and hence pool on intolerant actions, when in fact most of the family is actually progressive. In this scenario, the whole family would be collectively acting intolerant out of fear of their relatives' reactions, which is in fact unwarranted.

## 4.2 Sophie's World

I now consider an alternative environment in which all agents believe in taste-based social pressure—i.e., all agents are Sophies. First, recall that the main difference between assuming action-based and taste-based social pressure in game-theoretic terms is that Sophie faces a signaling game, whereas Stanley's decision environment is closer in spirit to a simple choice between risky prospects.

The induced signaling dynamics in the game have substantial implications for analyzing equilibria. Since Sophie's expected utility depends on her second-order beliefs *conditional on the strategy profile*, and given that all agents play and update their beliefs simultaneously, considering a deviation for agent $i$ potentially affects *all players'* beliefs, and thus payoffs. This implies that further assumptions on off-equilibrium-path beliefs are necessary to find the equilibrium of this game. I choose the D1 criterion (Banks and Sobel 1987; Cho and Kreps 1987), a standard refinement for Perfect Bayesian Equilibria in the signaling-games literature frequently used in models of conformity (e.g., Bursztyn, Egorov, and Fiorin 2020). The D1 criterion restricts receivers' beliefs after observing an off-equilibrium message to placing positive probability only on types who could benefit in general from the deviation compared to the equilibrium payoff. In this scenario, the D1 refinement implies that if agents are pooling on $h$, and agent $i$ plays $l$, then all players place probability 1 on $t_i = l$; and if agents are pooling on $l$ and observe $a_i = h$, then they all infer that $t_i = h$ with probability 1. Furthermore, I focus on pure-strategy equilibria. With these restrictions in mind, I am ready to present the main results for Sophie.

**Proposition 2** (Characterizing Equilibria for Sophie)**.** *If all agents have Sophie's utility function*

[7] It is worth pointing out that the model proposed by Bursztyn, Egorov, and Fiorin (2020) to motivate its experimental design is closer in spirit to taste-based social pressure, which I analyze in the following section.

*(2.2), for $\beta_T > 1$, majority size $\mu \in (1/2, 1)$ and common prior over majority type $(p_H, p_L)$, pure-strategy equilibria are given by:*

*(i) If* $(\pi_h - \pi_l)\Big(\mu - \pi_l(2\mu - 1)\Big) > \frac{1}{\beta_T(2\mu-1)}$, *there exists a pooling equilibrium on h: for all* $i \in I$, $a_i = h$;

*(ii) If* $\pi_h \geq \frac{1}{2} - \frac{1}{2\beta_T(2\mu-1)}$ *and* $\pi_l \geq \frac{1}{2} - \frac{1}{2\beta_T(2\mu-1)}$, *then there exists a separating equilibrium: for all* $i \in I$, $a_i = t_i$;

*(iii) If* $(\pi_l - \pi_h)\Big(\mu - \pi_h(2\mu - 1)\Big) > \frac{1}{\beta_T(2\mu-1)}$, *there exists a pooling equilibrium on l: for all* $i \in I$, $a_i = l$.

Figure 3 visually presents Proposition 2, again plotting majority size $\mu$ on the $x$ axis and $p_H, p_L$ on the $y$ axes. The regions for each type of equilibrium and their corresponding shapes are similar to those seen in Stanley's world. As in Stanley's case—albeit with structurally different constraints and mechanisms—the main driver of pooling equilibria is the wedge between posterior beliefs about the state of the world, $\pi_h$ and $\pi_l$. A pooling equilibrium on $h$ is more likely to be observed when an $h-$type is much more confident that he's part of the majority compared to the probability an $l-$type attributes to having the majority type. This condition also reflects a familiar conclusion regarding the common prior: due to Bayes' Rule, this wedge between $\pi_h$ and $\pi_l$ exists when the common prior is unbalanced in favor of one state. This implies that a pooling equilibrium on $l$ can never arise when $p_h > p_l$, and similarly, pooling on $h$ is never a stable equilibrium if $p_l > p_h$. Furthermore, the pooling regions exhibit the same quadratic shape, reflecting the varying strengths of the information and judgment effects discussed previously.

The truth-telling region in Sophie's world is even more similar to Stanley's case—indeed, the stated conditions for truth-telling in Propositions 1*(ii)* and 2*(ii)* are equivalent in structure. This is not surprising: in a separating equilibrium, all players (credibly) take other actions at face value, which reduces the second-order beliefs in Sophie's utility function to a simple belief over the share of players who share her revealed type. Thus, if Stanley and Sophie have the same relative weight on social-judgment utility, $\beta_A = \beta_T$, the range of parameters for which a world of Stanleys or Sophies tell the truth will be exactly equal. The comparative statics with respect to $\beta_T$ are also similar to Stanley's case. As the relative weight on social-judgment utility grows, the bounds on stability of separating equilibria become tighter, and pooling equilibria are stable for a wider range of parameters. Figure 5 in Appendix B illustrates the equilibrium regions for a higher value of $\beta_T$.

This brings us to the two main differences in the results we obtained for each source of social judgment. First, note that whereas equilibria for Stanley were generally unique given a common prior $p_H$ and a majority size $\mu$, we get a substantial range of parameters that admit multiple equilibria in Sophie's world. Furthermore, Figure 3 also shows a small range of parameters for which there does not exist any pure-strategy equilibria. These results are driven by the more complex signaling structure that Sophie faces, since the expected payoffs of each possible action are determined not only by her beliefs about others' types (as was the case for Stanley), but also by her beliefs about others' chosen actions. The equilibrium is no longer uniquely determined by the common prior $p_H$ and the majority size $\mu$, because varying players' beliefs over others' strategy profiles directly affects the expected utility of each available alternative. For instance, consider the lower multiple-equilibria region, which admits either a separating equilibrium or a pooling equilibrium on $l$. Both types of equilibria may be stable because the expected payoff from

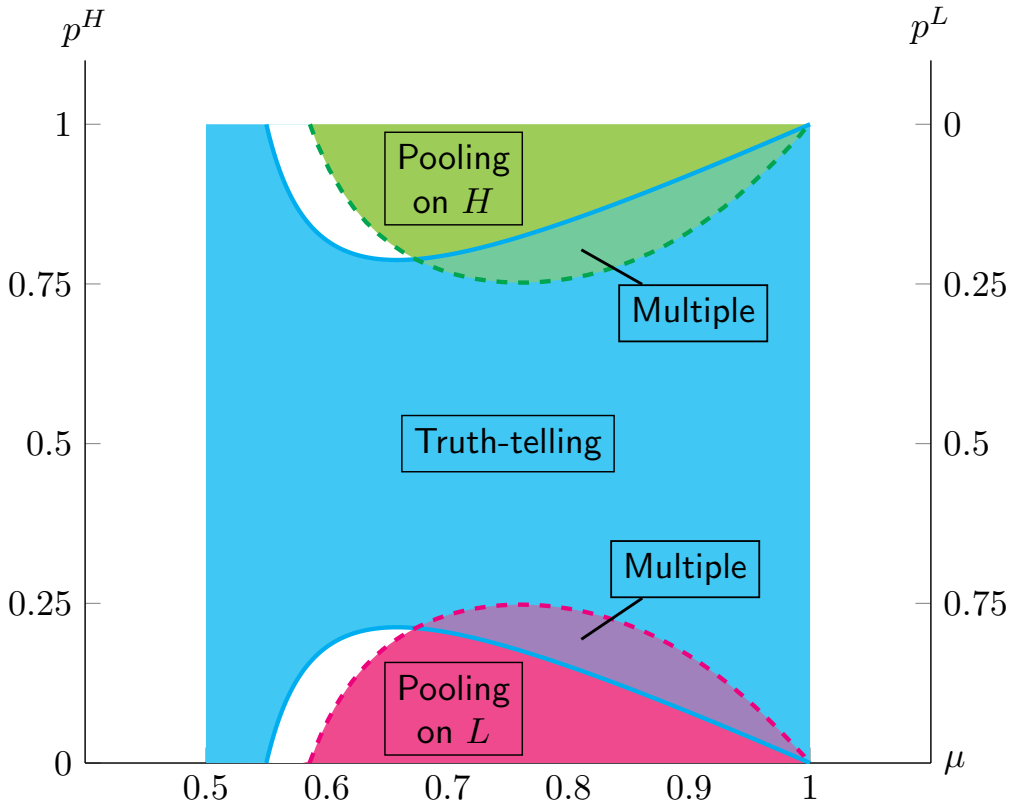



Figure 3: Equilibria for Sophie, $\beta_T = 10$.

a deviation in each case—which are generally distinct—are unprofitable relative to the expected payoff of following the equilibrium action.

For example, fixing a point in the multiple-equilibria region described above, if Sophie with type $t_i = h$ thinks everyone is speaking their minds, she faces the trade-off determined directly by her beliefs about the composition of the group, exactly like Stanley. On the other hand, if Sophie believes everyone is playing $l$, she compares the payoff of a certain judgment caused by "coming out" as $h$ (weighted by her beliefs about the composition of the group), against sticking to an *uncertain* ex-post judgment from all types of players. This average uncertain judgment is attainable for Sophie precisely because she's signaling that her type conforms to the majority, instead of acting in conformity. In a pooling equilibrium, messages are uninformative, which implies that audiences are forced to stick to their prior judgments of each player. In contrast, Stanley does not have the option to collect an uncertain ex-post judgment. This outcome is unavailable to him precisely due to the nature of the social judgment he perceives. Since Stanley's observed action is directly connected to the judgment he receives, pleasing one type and offending another is *unavoidable*.

This example highlights the main distinction between action-based and taste-based social judgment. Whereas judging on actions makes players decide how to act purely based on their beliefs about others' types, taste-based social judgment makes the strategy profile relevant for expected payoffs. The distinction can be made particularly clear by considering the setting analyzed by Bursztyn, Egorov, and Fiorin (2020), who show that people are more likely to reveal their support for xenophobic policies after learning that Trump won most votes in their

community in the 2016 election. Whereas a xenophobic Stanley would be comfortable speaking his mind simply because he grew more confident about the political leanings of his community, Sophie will also consider how her action will be interpreted in light of other people's behavior.

This is subtle but important. Suppose we observe both Stanley's and Sophie's worlds with communities that heavily voted for Trump. Stanley now knows how to maximize his utility, will act conservative, and crucially, enjoy the assured positive judgment of other conservatives. On the other hand, if Sophie pools, people will still be unsure about her true political leaning. Even if Sophie is truly a conservative, the fact that everyone is acting like Trump supporters makes them uncertain about her type.[8] Whereas in Stanley's world, a conservative audience member will be unwaveringly positive towards him if he acts conservative, in Sophie's world a conservative audience member will still have some cold feet towards her. Similarly, a progressive audience member won't completely shun her, since maybe she's just a hidden progressive trying to signal her conformity. Bursztyn, Egorov, and Fiorin (2020) actually provide some suggestive evidence of this latter hypothesis, showing that people are less willing to punish those who supported donations to xenophobic organizations if they know the donor lived in a community that particularly supported Trump in the election.

In summary, while results are similarly structured for Stanley and Sophie, introducing social signaling as a motivating factor changes the mechanisms for observed results, suggesting distinct implications for welfare and behavior. First, the fact that expected payoffs now depend on the strategy profile being considered opens the door for multiple equilibria. Furthermore, ex-post welfare can be remarkably different between the two worlds. In a pooling equilibrium, Stanley always gets *complete* positive approval from a (strict) share of the group. On the other hand, in a pooling equilibrium, Sophie receives *partial* approval of varying degrees (depending on their types) from *every* member of a group. While this distinction does not lead to distinct predictions and comparative statics in this simple, symmetric and static environment, it may drive distinctions in more complex settings, which I briefly discuss below.

# 5 Discussion and Extensions

In many ways, the approach taken in this paper so far is a first step towards understanding the potentially distinct implications of action-based and taste-based social judgment. While the simple game structure adopted in this paper allows us to clearly understand the specific mechanisms driving predictions for each source of social judgment, it restricts our ability to generate stronger testable distinctions between these two alternatives. This caveat notwithstanding, considering a few extensions to the model suggests that these two sources of social judgment have broader distinct implications for behavior and welfare.

First, as discussed above, ex post payoffs often have generally distinct patterns in Stanley's and Sophie's world. While this distinction does not lead to distinct predictions and comparative statics in this simple environment, one can think of situations in which it could generate relevant differences. For instance, if the social-judgment function is concave with respect to the receiver's perceived probability that the sender matches their type, pooling may become more appealing to Sophie without affecting Stanley's decision due to the binary action and type spaces.

[8] This intuition highlights the importance of extending this model into a more continuous structure, which is not covered in this paper.

Furthermore, extending the timeline to include a second period of social interaction may generate distinct observed behavior. Similar to the full-information case discussed above, if a separating equilibrium is played in the first period, behavior in the second period will be remarkably distinct between these two worlds. If sensitivity to social judgment is high enough, a minority Stanley will conform and enjoy the confident approval of the majority. A minority Sophie, on the other hand, will be unavoidably shunned by the majority, and will decide to speak her mind since her action ceases to affect others' impressions of her. One can speculate whether extending the structure towards a repeated or infinite game generates more distinct patterns of behavior over time, such as more sustained pooling for one source of social pressure or another.

Stanley and Sophie represent two extremes for how perceived social judgment affects behavior and welfare. Interacting these two sources of social pressure in the same environment, both within and across individuals, may also lead to interesting implications. For instance, if agents perceive both types of social pressure, the interplay between receiving approval from your actions and signaling your tastes could generate interesting predictions. Furthermore, analyzing a world composed of both Sophies and Stanleys could generate distinct predictions for collective behavior, especially in regards to information transmission in society. It is possible that the heterogeneity in perceived sources of social judgment generates and equilibrium in which (say) Stanleys reveal their types and Sophies conform, inducing more nuanced patterns of communication and social inference.

Finally, we can consider possible mechanisms for which Sophie and Stanley are more strongly connected. It turns out that by incorporating *second-order naive inference* makes Sophie behaviorally equivalent to Stanley, given a social-judgment sensitivity of $\beta_T$. More concretely, consider Sophie's brother, Newey. He believes in taste-based social judgment, but mistakenly believes that all other members of society take all actions at face value. Given this bias in statistical reasoning, Newey will be behaviorally equivalent to Stanley.

**Proposition 3** (Newey and Stanley are behaviorally equivalent)**.** *For any $\beta_T > 1$, majority size $\mu \in (1/2, 1)$ and common prior over majority type $(p_H, p_L)$, if agent $i$ with type $t_i = \theta$ has Sophie's utility function (2.2) and believes that for all $j \neq i$,*

$$\mathbb{P}_j(t_i = a_i \,|\, t_j, a_i, a_{-i}) = 1, \quad \forall i \neq j, \tag{5.1}$$

*then agent $i$ will be behaviorally equivalent to an agent with Stanley's utility function (2.1) and relative weight on social-judgment utility equal to $\beta_T$.*

This result is intuitively appealing: if one believes that all actions are taken at face value, their perceived inference problem reduces to considering their beliefs about the composition of types in society. In other words, second-order naiveté rules out signaling by assumption, transforming it in a simpler action-matching exercise that Stanley faces. Thus, in contexts where Stanley and Sophie would potentially behave differently—such as the ones intuitively discussed above—, naive probabilistic reasoning may lead us to misidentify action-based social judgment.

Of course, this result hinges on a strong and extreme assumption. First, people may be only *partially* naive, which leads them to attenuate the signaling forces at play. In reality, biases in statistical reasoning are unlikely to be widespread to the point of reaching *all* observed actions, instead of more constricted situations. In any case, considering this hypothesis fleshes out the distinction and the possible connections between the forces at play for each source of social judgment.

# 6 Conclusion

In this paper, I point to a relevant distinction in perceived sources of social judgment, formalizing the two proposed sources and drawing out initial implications with a simple game. The simple one-shot structure of the game clearly limits the extent to which I can draw out broader comparisons between action-based and taste-based social pressure. This paper is thus best interpreted as an exercise in considering how to translate intuitions about a relevant social phenomenon to a formal environment. While these alternatives for formalization often lead to qualitatively similar predictions—as is the case for many of the results presented here—their structure and mechanisms reveals different forces at play, which suggest further distinctions in observed behavior and welfare.

This paper can also serve as a starting point towards considering more complex structures that drive decisionmaking under social pressure. For instance, extending the action and type spaces to a more continuous structure would allow us to study the extent to which people engage in more extreme behaviors for signaling purposes—even with full awareness that actions reflect much more an effort to signal than an honest and accurate portrayal of one's preferences. This could be similar to the signaling dynamics of job markets, in which job seekers are induced to study in excess of their own preferences in order to more effectively signal their intrinsic ability (e.g., Spence 1973). Similarly, if agents can more precisely tune their messages regarding a socially-charged topic, this could potentially generate incentives for people to exaggerate the degree to which they support a given policy or a political group more generally.

Another potentially fruitful avenue relates to more nuanced forms of conformity. For instance, people may have a desire to fit in, while at the same time hiding that objective from the public. If people can clearly tell that you're trying to act cool, then you will never be considered cool. A related hypothesis is that the audience values honesty, in the sense that it punishes a messenger who gets caught hiding their types. For instance, a conservative audience may not be fond of gay men, but may treat him especially badly upon learning that he's been hiding his sexuality all along. Further investigations of such phenomena are important given the prevalence of decisionmaking under social pressure and the relevance of many of these settings to social norms, political movements, and welfare more generally.

# References


Banks, Jeffrey S., and Joel Sobel. 1987. "Equilibrium Selection in Signaling Games." *Econometrica* 55 (3): 647–661.

Battigalli, Pierpaolo, and Martin Dufwenberg. 2022. "Belief-Dependent Motivations and Psychological Game Theory." *Journal of Economic Literature* 60, no. 3 (September): 833–82. https://doi.org/10.1257/jel.20201378.

Bernheim, B. Douglas. 1994. "A Theory of Conformity." *Journal of Political Economy* 102 (5): 841–877. https://doi.org/10.1086/261957.

Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin. 2020. "From Extreme to Mainstream: The Erosion of Social Norms." *American Economic Review* 110, no. 11 (November): 3522–48. https://doi.org/10.1257/aer.20171175.

Bursztyn, Leonardo, Alessandra L. González, and David Yanagizawa-Drott. 2020. "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia." *American Economic Review* 110, no. 10 (October): 2997–3029. https://doi.org/10.1257/aer.20180975.

Cho, In-Koo, and David M. Kreps. 1987. "Signaling Games and Stable Equilibria." *The Quarterly Journal of Economics* 102, no. 2 (May): 179–221. https://doi.org/10.2307/1885060.

Duffy, John, and Jonathan Lafky. 2021. "Social conformity under evolving private preferences." *Games and Economic Behavior* 128:104–124. https://doi.org/10.1016/j.geb.2021.04.005.

Fernández-Duque, Mauricio. 2022. "The probability of pluralistic ignorance." *Journal of Economic Theory* 202:105449. https://doi.org/10.1016/j.jet.2022.105449.

Geanakoplos, John, David Pearce, and Ennio Stacchetti. 1989. "Psychological games and sequential rationality." *Games and Economic Behavior* 1 (1): 60–79. https://doi.org/10.1016/0899-8256(89)90005-5.

Gulesci, Selim, Sam Jindani, Eliana La Ferrara, David Smerdon, Munshi Sulaiman, and H. Young. 2023. "A Stepping Stone Approach to Norm Transitions." *SSRN Electronic Journal,* https://doi.org/10.2139/ssrn.4425503.

Kuran, Timur. 1989. "Sparks and prairie fires: A theory of unanticipated political revolution." *Public Choice* 61, no. 1 (April): 41–74. https://doi.org/10.1007/bf00116762.

Kuran, Timur. 1991. "Now Out of Never: The Element of Surprise in the East European Revolution of 1989." *World Politics* 44 (1): 7–48.

Kuran, Timur. 1995. *Private Truths, Public Lies.* Harvard University Press.

Loewenstein, George, and Andras Molnar. 2018. "The renaissance of belief-based utility in economics." *Nature Human Behaviour* 2, no. 3 (February): 166–167. https://doi.org/10.1038/s41562-018-0301-z.

Michaeli, Moti, and Daniel Spiro. 2017. "From Peer Pressure to Biased Norms." *American Economic Journal: Microeconomics* 9, no. 1 (February): 152–216. https://doi.org/10.1257/mic.20150151.

O'Gorman, Hubert J. 1975. "Pluralistic ignorance and White estimates of White support for racial segregation." Place: United Kingdom Publisher: Oxford University Press, *Public Opinion Quarterly* 39:313–330. https://doi.org/10.1086/268231.

Prentice, Deborah A., and Dale T. Miller. 1993. "Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm." Place: US Publisher: American Psychological Association, *Journal of Personality and Social Psychology* 64:243–256. https://doi.org/10.1037/0022-3514.64.2.243.

Smerdon, David, Theo Offerman, and Uri Gneezy. 2020. "'Everybody's doing it': on the persistence of bad social norms." *Experimental Economics* 23, no. 2 (June 1, 2020): 392–420. https://doi.org/10.1007/s10683-019-09616-z.

Spence, Michael. 1973. "Job Market Signaling." *The Quarterly Journal of Economics* 87 (3): 355–374.

# A Proofs

## A.1 Stanley

*Proof of Proposition 1 (Characterizing Equilibria for Stanley).*

Fix $\beta_A = \beta > 1$, majority size $\mu \in (1/2, 1)$, an agent $i \in I$ with type $\theta \in \Theta$, and a common prior $p_H \in (0, 1)$ with $p_L = 1 - p_H$. Since the expectation of an indicator variable for an event is the probability of that event, we can rewrite agent $i$'s expectation of his action matching agent $j$'s type as

$$\mathbb{E}\left[\mathbb{1}\{a_i = t_j\} \,|\, t_i\right] = \mathbb{P}(t_j = \theta \,|\, t_i = \theta) = q_\theta$$

This condition simplifies Stanley's expected utility for each action in his choice set. If Stanley chooses to tell the truth, $a_i = \theta$, he gets expected utility $\mathbb{E}U^A(a_i = \theta \,|\, t_i = \theta) = 1 + \beta q_\theta$. If he chooses to state the opposite type, $a_i = \widetilde{\theta}$, Stanley gets expected utility $\mathbb{E}U^A(a_i = \widetilde{\theta} \,|\, t_i = \theta) = \beta(1 - q_\theta)$. Thus, Stanley will reveal his type if

$$1 + \beta q_\theta \geq \beta(1 - q_\theta) \iff q_\theta \geq \frac{1}{2} - \frac{1}{2\beta}, \tag{A.1}$$

or, in terms of agent $i$'s belief that they are part of the majority,

$$\pi_\theta \geq \frac{1}{2} - \frac{1}{2\beta(2\mu - 1)}. \tag{A.2}$$

If Equation A.2 holds for both $\pi_h$ and $\pi_l$, then both types tell the truth and we have a separating equilibrium. If it holds only for $\pi_h$, then both types play $a_i = h$ and we have a pooling equilibrium on $h$. Analogously, if A.2 holds only for $\pi_l$, then both types play $a_i = l$ and we have a pooling equilibrium on $l$.

To finish the proof, note that the inequality cannot be invalid for both types simultaneously. By way of contradiction, suppose that $\pi_l < \frac{1}{2} - \frac{1}{2\beta(2\mu-1)}$ and $\pi_h < \frac{1}{2} - \frac{1}{2\beta(2\mu-1)}$. This implies that

$$\mathbb{P}(w = L \,|\, t_i = h) = 1 - \pi_h > \frac{1}{2} + \frac{1}{2\beta(2\mu - 1)},$$

and by Bayes' Rule, $\pi_l \geq 1 - \pi_h$. Combining these results gives us

$$\frac{1}{2} + \frac{1}{2\beta(2\mu - 1)} \leq \pi_l < \frac{1}{2} - \frac{1}{2\beta(2\mu - 1)},$$

which is a contradiction. □

*Proof of Corollary 1.* Fix $\mu, \beta_A$, and $\theta \in \Omega$ such that $p_\theta > 1/2$. Choose $i \in I$ such that $t_i = \theta$. Equation A.2 implies that Stanley's dominant strategy is to tell the truth ($a_i = \theta$) if $\pi_\theta > 1/2$, which, from Bayes' Rule, always holds when $p_\theta > 1/2$.[9] □

[9] Recall that

$$\mathbb{P}(w = \theta \,|\, t_i = \theta) = \frac{\mathbb{P}(t_i = \theta \,|\, w = \theta) \cdot \mathbb{P}(w = \theta)}{\mathbb{P}(t_i = \theta)} \geq P(w = \theta) \iff \mathbb{P}(t_i = \theta \,|\, w = \theta) \geq \mathbb{P}(t_i = \theta)$$

## A.2 Sophie

I begin by stating and proving two Lemmas that will be useful for deriving most of the results for Sophie.

**Lemma 1.** *If for all $i \in I$ and all $j \neq i$, $\mathbb{P}_j(a_i = t_i \,|\, t_j, a_i, a_{-i}) = 1$, then*

$$\mathbb{E}_i\left[\mathbb{P}_j\left(t_i = t_j \,|\, t_j, a_i, a_{-i}\right) \,\middle|\, t_i\right] = \mathbb{P}_i(t_j = a_i \,|\, t_i) = \mathbb{E}_i[\mathbb{1}\{a_i = t_j\} \,|\, t_i].$$

*Proof.* Fix $a_i = \theta$. Using the Law of Total Probability, we have

$$\begin{aligned}\mathbb{P}_j(t_i = t_j \,|\, t_j, a_{-i}, a_i = \theta) &= \mathbb{P}_j(t_i = t_j \,|\, t_j, a_{-i}, a_i = \theta, t_i = a_i) \cdot \overbrace{\mathbb{P}_j(a_i = t_i \,|\, a_i = \theta, t_j, a_{-i})}^{\text{1 by assumption}} \\ &\quad + \mathbb{P}_j(t_i = t_j \,|\, t_j, a_{-i}, a_i = \theta, t_i \neq a_i) \cdot \underbrace{\mathbb{P}_j(a_i \neq t_i \,|\, a_i = \theta, t_j, a_{-i})}_{\text{0 by assumption}} \\ &= \mathbb{P}_j(t_i = t_j \,|\, t_j, a_{-i}, a_i = \theta, t_i = a_i). \end{aligned} \tag{A.3}$$

Taking $i$'s ex-ante expectations over A.3 and applying the Law of Iterated Expectations, we have

$$\begin{aligned}&\mathbb{E}_i\left[\mathbb{P}_j(t_i = t_j \,|\, t_j, a_{-i}, a_i = \theta, t_i = a_i) \,\middle|\, t_i\right] = \\ &\quad \underbrace{\mathbb{E}_i\left[\mathbb{P}_j(t_i = t_j \,|\, t_j, a_{-i}, a_i = \theta, t_i = a_i) \,\middle|\, t_i, t_j = \theta\right]}_{1} \cdot \mathbb{P}_i(t_j = \theta \,|\, t_i) \\ &\quad + \underbrace{\mathbb{E}_i\left[\mathbb{P}_j(t_i = t_j \,|\, t_j, a_{-i}, a_i = \theta, t_i = a_i) \,\middle|\, t_i, t_j \neq \theta\right]}_{0} \cdot \mathbb{P}_i(t_j \neq \theta \,|\, t_i) \\ &\quad = \mathbb{P}_i(t_j = \theta \,|\, t_i) = \mathbb{P}_i(a_i = t_j \,|\, t_i) = \mathbb{E}_i[\mathbb{1}\{a_i = t_j\} \,|\, t_i], \end{aligned} \tag{A.4}$$

which completes the proof.

□

**Lemma 2** (Equivalence for separating equilibria)**.** *For any majority size $\mu \in (1/2, 1)$ and common prior over majority type $(p_H, p_L)$, if $\beta_A = \beta_T = \beta$, a separating equilibrium exists for Sophie if and only if it exists for Stanley.*

*Proof of Lemma 2 (Equivalence for separating equilibria).* Fix $\beta_A = \beta_T = \beta > 1$, $p_H \in (0, 1)$, $\mu \in (1/2, 1)$ and $i \in I$ with $t_i = \theta$. Furthermore, assume that the other $-i$ agents are playing a separating equilibrium: for all $j \in I \backslash \{i\}$, $a_j = t_j$. Now consider Sophie's second-order beliefs about her type, $\mathbb{E}_i\left[\mathbb{P}_j\left(t_i = t_j \,|\, t_j, a_i, a_{-i}\right) \,\middle|\, t_i\right]$. In a separating equilibrium, every player $j \neq i$ interprets $i$'s action as fully diagnostic of $t_i$. Formally, $\mathbb{P}_j(t_i = \theta \,|\, a_i = \theta) = 1$. Thus, from Lemma 1, $i$'s second-order beliefs are equivalent to $i$'s expectation over an indicator variable for $i$'s action matching $j$'s type,

$$\mathbb{E}_i\left[\mathbb{P}_j\left(t_i = t_j \,|\, t_j, a_i, a_{-i}\right) \,\middle|\, t_i\right] = \mathbb{E}_i[\mathbb{1}\{a_i = t_j\} \,|\, t_i],$$

which implies that Sophie faces Stanley's payoffs conditional on a separating equilibrium. □

*Proof of Proposition 2 (Characterizing Equilibria for Sophie).* Fix $\beta_T = \beta > 1$, $p_H \in (0, 1)$, $\mu \in (1/2, 1)$, and suppose that all agents have utility function 2.2. Given the subtleties of the signaling

The last condition is always satisfied in the model since $\mu = \mathbb{P}(t_i = \theta \,|\, w = \theta) \in (1/2, 1)$.

game, I proceed by positing each possible equilibrium, and deriving the conditions under which it is stable.

**Parts *(i)* and *(iii)*.** I begin by considering pooling equilibria on $a_i = l$ (part *iii*). The conditions for pooling on $a_i = h$ follow from symmetry.

Fix $i \in I$ with $t_i = h$, and assume that for all $j \in I \backslash \{i\}$, $a_j = l$. That is, we fix all other agents $(-i)$'s strategies, and consider $i$'s expected payoff from conforming, $a_i = l$, or telling the truth, $a_i = h$. By the D1 criterion, if agent $i$ plays $a_i = h$, all receivers infer that $t_i = h$ with certainty:

$$\mathbb{P}_j(t_i = h \,|\, a_i = h, a_{-i} = l) = 1, \quad \forall j \neq i. \tag{A.5}$$

Thus, $i$'s expected utility from playing $a_i = h$ is

$$\begin{aligned}
\mathbb{E}U^T(a_i = h \,|\, t_i = h) = 1 + \beta \Big[ & \overbrace{\mathbb{P}_j(t_i = t_j \,|\, a_i = h, a_{-i} = l, t_j = h)}^{1, \text{ by A.5}} \cdot \overbrace{\mathbb{P}_i(t_j = h \,|\, t_i = h)}^{q_h} \\
& + \underbrace{\mathbb{P}_j(t_i = t_j \,|\, a_i = h, a_{-i} = l, t_j = l)}_{0, \text{ by A.5}} \cdot \underbrace{\mathbb{P}_i(t_j = l \,|\, t_i = h)}_{1-q_h} \Big] \\
\mathbb{E}U^T(a_i = h \,|\, t_i = h) = 1 + \beta q_h. &
\end{aligned} \tag{A.6}$$

Note that if $i$ follows along with the pooling equilibria and plays $a_i = l$, the entire observed action profile is uninformative. Crucially, this allows us to ignore the action profile in agents' beliefs:

$$\mathbb{P}_j(t_i = t_j \,|\, a_i, a_{-i}, t_j) = \mathbb{P}_j(t_i = t_j \,|\, t_j) = q_{t_j}, \quad \forall i, j \in I \tag{A.7}$$

We can then rewrite $i$'s expected payoff from conforming as

$$\begin{aligned}
\mathbb{E}U^T(a_i = l \,|\, t_i = h) = \beta \Big[ & \overbrace{\mathbb{P}_j(t_i = t_j \,|\, t_j = h)}^{q_h} \cdot \overbrace{\mathbb{P}_i(t_j = h \,|\, t_i = h)}^{q_h} \\
& + \underbrace{\mathbb{P}_j(t_i = t_j \,|\, t_j = l)}_{q_l} \cdot \underbrace{\mathbb{P}_i(t_j = l \,|\, t_i = h)}_{1-q_h} \Big] \\
\mathbb{E}U^T(a_i = l \,|\, t_i = h) = \beta \left( q_h^2 + (1 - q_h) q_l \right). &
\end{aligned} \tag{A.8}$$

Combining A.6 and A.8, agent $i$ will play $a_i = l$ iff

$$(1 - q_h)(q_l - q_h) > \frac{1}{\beta}, \tag{A.9}$$

or, in terms of $\pi_h$ and $\pi_l$,

$$(\pi_l - \pi_h)\Big(\mu - \pi_h(2\mu - 1)\Big) > \frac{1}{\beta(2\mu - 1)}. \tag{A.10}$$

To complete the proof, I now show that if agents with type $h$ play $l$ in a pooling equilibrium, it is optimal for $l-$types to also play $l$. By way of contradiction, assume that A.9 holds and that $\mathbb{E}U^T(a_i = h \,|\, t_i = l) > \mathbb{E}U^T(a_i = l \,|\, t_i = l)$. The latter assumption implies that

$$\beta(1 - q_l) > 1 + \beta \left( q_l^2 + q_h(1 - q_l) \right) \iff (1 - q_h)(1 - q_l) - q_l^2 > \frac{1}{\beta}$$

Combining this inequality with A.9, we get

$$(1-q_h)(q_h-q_l) < \frac{1}{\beta} < (1-q_h)(1-q_l) - q_l^2 \implies (1-q_h)^2 > q_l^2$$

Since $(1-q_h)$ and $q_l$ are strictly positive, the inequality also holds if we drop the squares. Thus,

$$(1-q_h) > q_l \iff \mathbb{P}_i(t_j = l \,|\, t_i = h) > \mathbb{P}_i(t_j = l \,|\, t_i = l),$$

which is a contradiction given that agents update their beliefs using Bayes' Rule and $\mu \in (0.5, 1)$.
**Part *(ii)*.** Follows directly from Lemma 2 for a given $\beta_T$. □

*Proof of Proposition 3 (Newey and Stanley are behaviorally equivalent).* This result follows directly from Lemma 1. Newey's defining trait is that he believes that others take all actions at face value, i.e., for all $i \in I$ and all $j \neq i$,

$$\mathbb{P}_j(t_i = a_i \,|\, t_j, a_i, a_{-i}) = 1,$$

which is exactly the assumption defined in Lemma 1. Thus, we have that Newey's second-order beliefs are equivalent to $i$'s expectation over an indicator variable for $i$'s action matching $j$'s type,

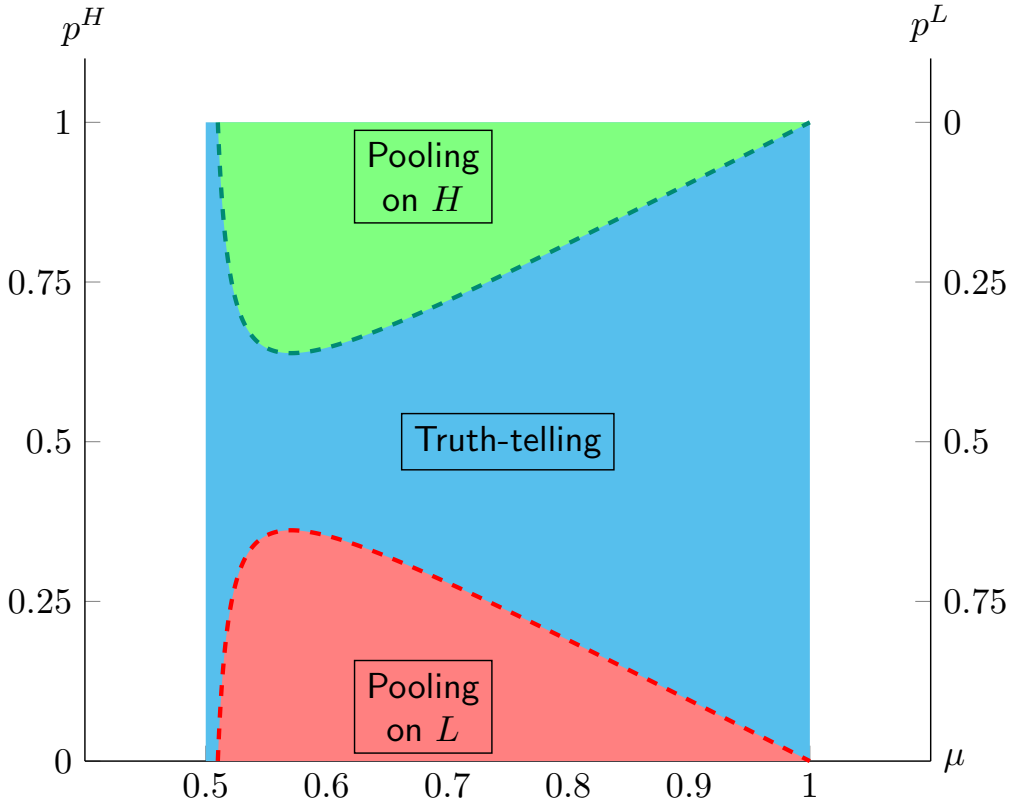
$$\mathbb{E}_i\left[\mathbb{P}_j\left(t_i = t_j \,|\, t_j, a_i, a_{-i}\right) \,\middle|\, t_i\right] = \mathbb{E}_i[\mathbb{1}\{a_i = t_j\} \,|\, t_i].$$

This implies that Newey's perceived utility function is equivalent to Stanley's (Equation 2.1), and that Newey acts in accordance to Stanley's decision rule (Equation A.2) for a given $\beta_T$. □

# B Additional Figures

## B.1 Stanley

Figure 4: Equilibria for Stanley, $\beta = 50$



## B.2 Sophie

Figure 5: Equilibria for Sophie, $\beta = 50$