From Pluralistic Ignorance to Common Knowledge with Social Assurance Contracts

By Matthew Cashman*

Draft: August 20, 2025

Societies and organizations often fail to surface latent consensus because individuals fear social censure. A manager might suspect a silent majority would offer a criticism, support a change, report a risk, or endorse a policy—if only it were safe. Likewise, individuals with beliefs they think are rare and controversial might stay quiet for fear of consequences at work or an online mob. In both cases pluralistic ignorance produces a public discourse misaligned with privately-held beliefs. Social assurance contracts unlock latent consensus, making the public discussion accurately reflect the underlying distribution of actual beliefs. They are akin to an open letter that publishes only when a stated threshold number of private signatures is reached. If it is not reached, nothing is revealed and no one is exposed. A single hand raised in dissent might get cut off, but a thousand can safely be raised together. I build a formal model and derive rules for choosing the threshold. The mechanism (i) induces participation from those willing to speak if assured of company, resolving the core coordination problem in pluralistic ignorance; (ii) makes the threshold a transparent policy lever—sponsors can maximize success, maximize informativeness, or hit a desired success probability; and (iii) turns success into information: meeting the threshold publicly reveals hidden agreement and widens the boundary of acceptable speech (the Overton window). I consider robustness to mistrust, organized opposition, and network structure, and outline low-trust implementations like cryptographic escrow. Applications include employee voice, safety and compliance, whistleblowing, depolarization, and civic expression.

JEL: D71, C72, D83, Z13

Keywords: Assurance contracts, public goods, coordination, pluralistic ignorance, depolarization, mechanism design

* Cashman: Warwick Business School, Coventry CV4 7AL, UK and MIT Sloan, 100 Main Street, Cambridge, MA, USA. Email: matthew.cashman@wbs.ac.uk. I wish to thank Adam Bear, Rahul Bhui, Caspar Kaiser, Birger Wernerfelt, Zachary Wojotowicz, and Erez Yoeli for many helpful comments and suggestions. I am especially indebted to Alex Tabarrok for his substantial feedback.

There is often a tyranny of the minority in public discourse, where a majority of discussion is generated by a small number of voices (Kim et al. 2021; McClain et al. 2021). Instead of reflecting the diversity and relative moderation of actual beliefs and behavior, these voices tend to be extreme: news media and social media content is as inflammatory as is profitable (Dey, Lahiri, and Mukherjee 2025; Ahler 2014). Disagreeing with the dominant viewpoint is often perceived to be too dangerous to be worthwhile, so the mere threat of punishment has chilling effects—dissenters opt for silence, blocs form, and an honest, open exchange of ideas eludes us all. The resulting distorted public discourse means the average person wrongly infers that talked-about opinions are commonly held even when it is not the case—a phenomenon referred to as pluralistic ignorance (Katz, Allport, and Jenness 1931; Prentice and Miller 1996).

In many ways, this is an old problem. In authoritarian states, for instance, citizens have often publicly applauded regimes they privately despised (Kuran 1997). Modern information ecologies, however, seem to amplify the problem. Social media platforms allow for greater reach than in-person communication or the telephone, and they greatly lower the barriers for transmission relative to broadcast. Any person can participate in the public conversation at little or no cost and wind up with global reach. This is compounded by the core logic of these platforms: they are driven by algorithms designed to maximize engagement—at the cost of veracity and social harmony, if necessary. These have been shown to systematically amplify emotionally charged and out-group hostile content, which can make users feel worse about their political out-groups (Milli et al. 2025). Under these circumstances it is easy to mistake a vocal minority for a silent majority's proxy. Recent experimental evidence also points to uncertainty about others' true preferences as a key driver of the persistence of inefficient or "bad" social norms, which tend to dissipate when individuals are better informed about their peers' actual views (Smerdon, Offerman, and Gneezy 2020). One of the drivers of the increase in polarization in recent years appears to be this imperfect reflection of privately-held beliefs on to the public sphere (Kuran 1997; Ahler 2014). Take the United States as an example: If we were to try to infer what the average American is like on the basis of news coverage and social media, we would end up with a caricature: this person is Blue or Red, Left or Right, R or D. He eats steak, drives a truck, and goes to church—or he eats vegetarian, drives a Prius, and goes to protests.

However, there may be a way to safely surface beliefs that feel dangerous to express, thereby making public discourse more accurately reflect actual beliefs. To do this, I transport a piece of economic technology into the social domain: the assurance contract. Assurance contracts (also called provision-point mechanisms) are mechanisms that enable the private provision of public (or

club) goods by getting around free-rider and collective action problems. The most well-known examples are crowdfunding platforms like Kickstarter (2025a). At their most basic, these contracts solicit contributions which enable the production of a good if enough is contributed to reach a "provision point". If the provision point is not reached, contributions are returned. For instance, two photographers might design a backpack that handles common camera gear better than anything on the market. They think other camera-junkies will like it, but they lack the capital and risk tolerance to move directly to mass production. So, they set up a campaign on Kickstarter: they solicit $300 payments from interested parties on the condition that the backpack will be manufactured and sent to contributors if 1,000 people make the payment. If they get 1,000 or more $300 payments from contributors, they arrange for the backpack to be manufactured and sent to the contributors. If they don't reach that threshold, the $300 payments are returned. Transported to the social domain, these contracts are something like an open letter—with the exception that one's signature on the letter is secret from both the public and other endorsers, and become public only *after* some assurance conditions (a provision point) are fulfilled. At a university it might look like this:

> We, the undersigned believe [controversial belief] and think the University should take the following steps [1, 2, 3. . .]. Signatures on this letter will become public *only when there are at least [200] signatures from faculty and [800] signatures from students.* Until then, no one except the keeper of this letter can see who has signed.

This serves to make expressing a controversial (or controversial-seeming) belief much less costly: while a single hand raised in dissent might get cut off, a thousand can safely be raised together.

The first general implementation of this mechanism, Spartacus.App, was launched in 2024, and use cases in the social domain abound. In a university department many faculty may privately disagree with an administrative policy but say nothing, each fearing they are alone in their concern. Similarly, within a corporation, employees might silently oppose an unethical practice, or within a political party, party members might privately dissent from the party line but say nothing. In all such cases, an expressive collective action—such as a group letter or joint statement—could shift the public narrative if only individuals could be assured enough other people are with them. The key barrier is one of coordination and assurance: no one wants to be the first (or among the few) to step forward, risking backlash. Once a statement has a critical mass, public revelation is easy. This is a public good provision problem where the "good" is the revelation of shared sentiment and the potential shift in norms of public discussion. With *social* assurance contracts, instead of monetary contributions, individuals commit to publicly express a certain view. Their commitment

(and identity) is held in secret and only revealed if a pre-specified number of others also pledge their support; no one's identity is exposed unless a critical mass joins, mitigating first-mover risk. By converting isolated private sentiments into a collective public statement, such a contract can foster common knowledge of actual, underlying beliefs while protecting individuals against the personal costs of speaking out in isolation.

The primary contributions of this paper are threefold: First, I formalize social assurance contracts as a means of providing expressive public goods. I develop a model tailored to the social context, where payoffs include reputational costs (c_i), esteem (e_i), a benefit from safe public revelation of the controversial belief (v_i), and a "warm-glow" bonus (w) for participation. Second, I analyze the equilibrium participation strategies under both complete and incomplete information. The paper explores the existence, uniqueness, and properties of these equilibria. I deliver comparative statics that map observables and design levers into threshold choice and success probabilities, and I connect these to implementation choices. Third, I formalize the "Overton window" (Lehman 2010), the range of acceptable public discourse, and provide a micro-foundation for how coordinated revelation of suppressed views can shift it to include previously suppressed beliefs. I then use this formalization to derive actionable guidance for setting the assurance threshold T given certain goals. In plain terms, the Overton formalization lets a contract architect translate a target change in what is publicly sayable into a threshold choice. Under incomplete information, the optimal threshold $T^\star$ increases with the dispersion and curvature of idiosyncratic expression costs, decreases with warm-glow and esteem, and aligns with the mass of latent supporters near the acceptability boundary. Section V gives the design and shows how to compute $T^\star$ from primitives one can estimate or elicit. I also discuss how, when pluralistic ignorance suppresses moderate voices *within* otherwise polarized factions, social assurance contracts can offer a path towards depolarization.

The remainder of the paper is organized as follows. Section I reviews relevant literature. Section II formalizes the social assurance contract mechanism and participant incentives in detail. Section III presents the equilibrium analysis under complete and incomplete information, including comparative statics and the case of incomplete trust in the contract administrator. Section IV formalizes the Overton window and considers shifts in the public discussion, and Section V discusses the choice of the assurance threshold T. Section VI discusses future directions for research, and Section VII concludes. The Mathematical Appendix contains detailed proofs and derivations, and the Supplemental Appendix discusses vulnerabilities and their mitigations in B and the consequences of social network topology in C, and offers an outline of a model with strategic opposition as a basis for future research in D.

I. Related Literature

A. Public Goods, Step-Level Games, and Assurance Contracts

This work draws on several strands of literature. First, it is related to the game theory of step-level public goods and threshold games. The concept of the assurance contract has its origins in the work of Dybvig and Spatt (1983), who analyzed the coordination problems inherent in public goods provision and externalities. Building on this foundation, Palfrey and Rosenthal (1984) examined the strategic aspects of threshold public good contribution, showing how binary contribution decisions in a provision point setting can lead to multiple equilibria. The classic problem of voluntary public good provision often leads to multiple equilibria or coordination failures. Bagnoli and Lipman (1989) showed that if contributions are refunded when a funding threshold isn't met (a provision-point mechanism), efficient public good provision can be achieved under certain conditions, breaking the standard free-rider outcome.

Subsequent research has studied variations of threshold public goods games in theory and in laboratory experiments. Croson and Marks (2000) conducted a meta-analysis of step-level returns in threshold public goods experiments, identifying key factors that influence successful provision. In addition, researchers have documented how cooperation persists in public goods experiments despite theoretical predictions of free-riding, with attribution to confusion (Andreoni 1995), altruism, or a warm-glow effect (Andreoni 1990), among other explanations. There are several studies which provide experimental evidence for a warm-glow effect specifically in provision-point mechanisms applied to green electricity markets (Rose et al. 2002; Menges, Schroeder, and Traub 2005; Mitra and Moore 2018), providing evidence of the phenomenon in the specific context of assurance contracts. These games are characterized by an all-or-nothing dynamic: if enough people contribute (or cooperate) the project succeeds, but if too few do, it fails and contributions are wasted (or refunded). This structure creates strategic uncertainty in that each player must guess how others will behave. The assurance contract, by design, tries to resolve some of this uncertainty by protecting contributors in case of failure. Tabarrok (1998) advanced this concept significantly by introducing *dominant* assurance contracts, which add a crucial twist to standard assurance contracts: if the threshold is not met, contributors not only get their money back but also receive a small bonus from the contract administrator. This innovation makes contributing a dominant strategy under certain payoff assumptions—individuals have incentive to contribute regardless of whether they believe others will contribute. This, in theory, solves the coordination problem by eliminating the risk of non-provision equilibria. Further extending the concept of bonuses to ensure

provision, Zubrickas (2014) introduces a refund bonus that is proportional to their contribution if the project is not funded. This competition for potential refund bonuses also drives the equilibrium towards successful provision, potentially achieving Lindahl pricing without bonuses being paid in equilibrium. Cason and Zubrickas (2017) later tested dominant assurance contracts experimentally and found that they do indeed increase the provision of public goods, confirming Tabarrok's theoretical predictions. The social assurance contract builds on this logic, adapting it to non-monetary, expressive "contributions."

This work is also informed by the information escrow literature in law and economics. Information escrows are effectively social assurance contracts, though as yet very narrow in their scope. Ayres and Unkovic (2012) discuss mechanisms whereby information (like allegations of wrongdoing, romantic interest, or a negotiated end to a dispute) is held in escrow by a trusted third party and only released if certain conditions are met. In particular, they propose allegation escrows to encourage reporting of sexual harassment. With these, a victim can confidentially lodge a complaint which remains hidden unless one or more additional complaints against the same perpetrator are filed. Only when a threshold of independent allegations is reached does the system take action. This protects accusers from the risk of retaliation or not being believed if they are alone in their claim. This logic connects to the broader literature on coordination games and strategic complementarities. When individual strategies are complements (one person's willingness to act increases others' willingness), multiple self-fulfilling equilibria can arise. The silence-versus-speaking problem is exactly such a case: if others are silent, you prefer to be silent (silence reinforces silence), but if enough others speak, you would also speak. This is analogous to models of bank runs or currency attacks, where each investor's decision is complementary to others'. Global games analysis (Carlsson and van Damme 1993; Morris and Shin 2002) has been used to study how slight uncertainties can select equilibria in such coordination settings. In our context, strategic complementarities in silence create the risk of a no-expression equilibrium, and the assurance contract can be seen as a mechanism to eliminate the worst equilibrium by changing the payoff structure. As Morris and Shin (2002) discuss in related models, public signals or guarantees can move the game to the efficient outcome by coordinating expectations. Here, the contract serves as that coordinating device. This mechanism differs by focusing on endogenous preference revelation rather than exogenous signals.

B. Information Cascades, Herding, and Opinion Dynamics

Next, I situate our work in the context of informational cascades and herding models. A rich body of theory (e.g., Bikhchandani, Hirshleifer, and Welch 1992, Raafat, Chater, and Frith 2009)

examines how individuals often base their actions on observations of others, sometimes leading to cascades where everyone ignores their private information and simply follows others' previous behavior. In a classic informational cascade, if early individuals act (or fail to act) in a certain way, later individuals will rationally imitate that behavior even if their own private signals differ. Thus, once a cascade of silence starts, social learning is essentially blocked and the false norm persists (Bikhchandani et al. 2024). This yields a fragile equilibrium – an entire group may be acting contrary to its private interests or beliefs simply because each person is waiting for someone else to deviate.

The problem of suppressed beliefs can be seen as a form of such a cascade: everyone remains silent because everyone else is silent. Each individual, observing no one speaking out, assumes silence is the safer strategy. This mechanism aims to break this cascade by coordinating a simultaneous deviation by many individuals at once. By doing so, it forces the outcome to reveal the underlying private information (the true level of dissent) that was previously hidden. In essence, a social assurance contract can short-circuit an information cascade which leads to silence by giving everyone a secure way to express themselves conditional on others doing so. No one has to move first; when the provision point is met, the suppressed consensus is revealed all at once.

C. Social-psychological Theories

The approach here also builds on social-psychological theories of conformity and suppressed expression. The notion of people publicly conforming to norms they privately reject was analyzed by Kuran (1997) as "preference falsification", and a process leading to this by Noelle-Neumann (1974) as the "spiral of silence". The concepts "groupthink", where group members choose consensus over correctness (Esser 1998), and "peer pressure" (Kandel and Lazear 1992) are related in that they describe how outward consensus can come to be different from inner beliefs. There is a significant literature addressing social conformity as a broader set of phenomena as well (see Capuano and Chekroun 2024, for a review), and social desirability bias specific to survey responses (Krumpal 2013). Opposite to pluralistic ignorance is "false consensus", where an agent believes others hold beliefs more like their own than is actually the case (Ross, Greene, and House 1977; Marks and Miller 1987). When a privately-held belief is already within the public discourse, the situation is generally stable: wrongly believing others share your beliefs more than they do poses no risk of the kind we are concerned with here. In the case where a privately-held belief is *not* within the public discourse *and* a person believes it is more widely-held than is the case, the situation will tend to self-correct. Such a person will be more likely to share this controversial belief than is actually warranted,

will face the consequences, and will update their beliefs about how widespread the controversial belief is accordingly. Pluralistic ignorance itself has been documented in numerous studies—on college campuses (e.g. Prentice and Miller 1993), in the workplace (Halbesleben, Wheeler, and Buckley 2007), and in political opinion climates (Ahler 2014). Smerdon, Offerman, and Gneezy (2020) provide direct evidence from a lab experiment that "bad" norms persist not because they are stable equilibria of a coordination game, but rather because players have uncertainty about others' preferences when modeling group behavior. Taken together, these studies suggest that providing credible information about others' true beliefs via a coordination mechanism could dramatically change individuals' willingness to speak or act—exactly what social assurance contracts provide.

D. Alternative Revelation Mechanisms

The question of how to surface true opinions in the public discussion is related to a literature in survey methodology focused on eliciting truthful answers to sensitive questions. While anonymity is a standard approach to reducing social desirability bias, more sophisticated indirect questioning techniques provide plausible deniability to respondents in an effort to get more accurate measurements (Krumpal 2013). The Randomized Response Technique (RRT) uses a randomizing device to add noise to individual answers, shielding the respondent from the researcher (Warner 1965). The Unmatched-Count Technique, or list experiment, uses aggregation to conceal a sensitive response within a statistical estimate (Glynn 2013), while the Crosswise model uses joint-question evaluation (Yu, Tian, and Tang 2008). The distinction between these methods and social assurance contracts lies in the ultimate goal: these survey techniques are designed to reveal an aggregate estimate of true beliefs to a researcher while permanently protecting individual identities. They do not, by design, create common knowledge among the participants themselves or overcome the coordination problem inherent in public expression. The social assurance contract, in contrast, uses temporary anonymity as a means of creating common knowledge through coordinated, public revelation.

E. The Overton Window

The range of policies and ideas considered acceptable for mainstream political discourse is not infinite; it is constrained within what has come to be known as the Overton window (Lehman 2010). The concept posits that at any given moment, a set of public policies exists that a politician can support without being seen as too extreme to gain or keep public office. This window of acceptability is not static. It can be shifted, typically not by politicians themselves, who are constrained by it, but by social and political action that gradually makes formerly radical ideas seem first acceptable,

then sensible, and eventually popular (Youvan 2024). The window describes the political reality that public figures must navigate. When pluralistic ignorance is widespread, the discussion that is possible in public fails to reflect the true, underlying diversity of private beliefs, creating the conditions that a social assurance contract is designed to remedy.

There are several concepts related to that of the Overton window. Hallin's model of media coverage divides the political world into three concentric spheres (Hallin 1989). The inner sphere is the sphere of consensus, containing topics on which journalists assume general agreement and which are not subject to debate, and the outer sphere is the sphere of deviance, containing views so far from the mainstream that they are considered unworthy of airing or are actively ridiculed. Between these two lies the sphere of legitimate controversy, which is where mainstream political debate occurs. Here, the middle and inner spheres correspond to the Overton window. The "opinion corridor" (Swedish: åsiktskorridor), describes a similar narrow range of publicly acceptable viewpoints in Scandinavian political culture (Simons 2021). It highlights a tendency toward self-censorship to preserve a climate of consensus, thereby reinforcing the boundaries of the corridor.

The formalization of the Overton window as a function of expression costs connects this work to a rich literature on opinion dynamics, which models how public discourse can persistently diverge from private beliefs. These models typically generate the stability of misaligned norms as an emergent property of repeated, uncoordinated interactions. Kuran (1997) provides the foundational logic by which social penalties separate private beliefs from public claims, enabling informational cascades. Subsequent formal and empirical work shows how public misrepresentation can persist or even reshape beliefs: models and experiments identify conformity traps that prolong status-quo expression (Duffy and Lafky 2018), agent–based accounts trace how pluralistic ignorance evolves with thresholds, rank, and memory (Lütz and Wardil 2024), and coevolution mechanisms show "fakers becoming believers" and the entrenchment of ideological myths via ex post rationalization (León-Medina, Tena-Sánchez, and Miguel 2020; Apolte and Müller 2022). The novelty here is to make the Overton window itself the construct to be modeled and influenced, rather than having the set of publicly acceptable ideas be downstream outcomes of exogenously specified influence rules. I define it as a set of publicly acceptable ideas whose boundary is an expression frontier—the point at which making a statement ceases to be privately costly for a minimally sustainable share of participants under shared beliefs. This definition is independent of any particular diffusion rule, yields measurable predictions about how the frontier moves with changes in forum sanctions, reputational payoffs, policy, and of course beliefs about others' beliefs, and it supports comparative statics that link window shifts to those primitives. Social assurance contracts enter as one practical

lever that exploits this connection. By creating a public, incentivized signal when the number of signers reaches a known threshold under known terms, they change common knowledge and perceived pivotality in a way that the framework maps into testable threshold effects at success and design results for targeting desired shifts in the frontier.

F. Crowdfunding and Provision-Point Mechanisms in Practice

Social assurance contracts are actively used for practical purposes (though this has mostly escaped the notice of academics). In the economic domain, crowdfunding platforms such as Kickstarter (2025a) generally employ a straightforward provision-point rule ("all-or-nothing" funding): a project is funded only if the pledge threshold is reached, otherwise contributions are refunded. Civic crowdfunding initiatives (e.g., Spacehive) use similar threshold pledges for community projects. In the social domain, there are a handful of implementations of assurance contracts: Spartacus.App and CollAction are more general implementations, offering a very flexible setup for contracts, while Callisto implements the allegation escrow described in Ayres and Unkovic (2012) whereby a victim's report remains confidential until other reports are filed against the same perpetrator. Together We Can Fix Academia (Smout et al. 2021) aims to allow researchers to coordinate to encourage open science practices (which produce public goods costly to the individual researcher, but beneficial to the group). More generally, existing assurance contract implementations incorporate provision-point or quorum rules to ensure a show of sufficient support before a tender goes public or is enacted. It is also possible to think of certain dating apps as providing one-off assurance contracts in the case where Alice's interest in Bob is only revealed to Alice if Bob reciprocates, and vice-versa.

It is worth pausing to reflect on why assurance contracts have seen runaway success in the economic domain (Kickstarter alone boasts approximately 280,000 successfully funded projects, Kickstarter 2025b), but comparatively mild success in the social domain. One explanation is simply time. SellaBand, the first crowdfunding site to use a provision-point mechanism, was founded in 2006 (Agrawal, Catalini, and Goldfarb 2014), while the (now defunct) Stanford Catalyst, perhaps the first attempt at a platform offering something like social assurance contracts, was founded in 2013 (Cheng and Bernstein 2014). The intervening seven years may explain some of the difference, but more of an explanation is called for. The demand for social assurance contracts is likely lower in the social vs. economic domain. It is limited to situations where i) agents have some privately-held belief that is not part of the public discussion already; ii) agents exist under an Overton regime that will punish speaking out alone iii) punishments for voicing those beliefs are severe enough the agents will not speak out alone; iv) agents wish to speak out (it is important to the agent that the belief

be part of the public discussion); v) agents believe there are at least some others with the same privately-held belief who are similarly prevented from speaking out and wish to; vi) agents know of social assurance contracts; and vii) there is some low-friction platform or other means for actually running the contract. These requirements are likely to apply to many fewer situations than the simple desire for novel music or manufactured products, leading us to the conclusion that demand in the social domain would be lower than for crowdfunding mechanisms. It may also be the case that the costs of implementing an effective platform have thus far been too high, existing attempts may be missing some important feature, or it could be that the necessary technology simply has not been available until relatively recently (e.g., cryptographic technology, see Appendix B). Finally, marketing is a key issue. Success of such a technology benefits greatly from network effects, and it is possible that whatever awareness or demand threshold is necessary for runaway growth has simply not been reached yet (see Appendix C for a discussion of the influence of social network topology). Most likely, a combination of these issues are at play.

Indeed, it may be the case that the necessary pieces have come together only recently. Spartacus.App, founded in 2024, is the first general-purpose platform which implements social assurance contracts as they are conceptualized here. Previous efforts such as Stanford Catalyst and CollAction have focused on coordinating noncontroversial collective action: users contribute their *intention* to engage in some activity when enough other people have registered the same intention. This is not strictly an assurance contract in that the contribution of an intention is not equivalent to actually *doing* the costly thing. Signing up for "Let's gather our litter, electronic & bulk waste this month" on CollAction costs a user little or nothing, and the user is at best weakly bound to completing the action. In contrast, a signature contributed to a social assurance contract is itself the action that invites all the relevant costs and benefits. Earlier platforms also do not keep who has signed secret, preferring instead to publicize the list of endorsers—another key difference. Callisto implements most of the desiderata described here and pre-dates Spartacus.App, but it focuses very narrowly on allegations of sexual assault and has the interesting property that the actual content of the assurance contract is not determined until the assurance condition is met. That is, the person against whom complaints have been made is not something potential signers can coordinate on until enough complaints have been made *independently* and the assurance threshold has been reached. Spartacus does not at present offer cryptographic guarantees and so may be limited by endorsers' trust in it (discussed in Appendix B.B1), but in other respects it is the first complete implementation of the social assurance contract mechanism.

II. The Social Assurance Contract Mechanism

I model a *social assurance contract* as a simultaneous one-shot game with no information flow between players[1]. The contract is designed by a contract *architect* and administered by a contract *administrator*, who may be the same person. The administrator may also be a computer system. Let $\mathcal{P}$ be the overall population of individuals who could potentially be aware of or affected by the contract's statement. Each individual $j \in \mathcal{P}$ has an individual-specific valuation v_j for the success of the contract, which can be positive, zero, or negative. The set of *potential endorsers* for the contract, denoted by $\mathcal{N}$, consists of all individuals $i \in \mathcal{P}$ for whom the successful revelation of the statement provides a positive personal benefit, i.e., $\mathcal{N} = \{i \in \mathcal{P} \mid v_i > 0\}$. Let $N = |\mathcal{N}|$ be the number of such potential endorsers. The game is modeled among these N individuals, who privately hold a particular belief or support a controversial statement which is in conflict with the current public norm.

A. Contract Structure

The contract is defined by a specific statement to be endorsed, s_{belief} (which could be a proposition or normative claim, representing an underlying belief $\mathbf{b}_s$), and an assurance threshold T. This threshold T is a positive integer representing the minimum number of endorsements required for the contract's list of endorsers to become public; for success to be possible, T must satisfy $1 \leq T \leq N$. The contract structure is as follows:

1) Individuals $i \in \mathcal{N}$ decide simultaneously their action $a_i \in \{0, 1\}$, where $a_i = 1$ denotes choosing to endorse the contract, Sign, and $a_i = 0$ denotes Not Sign.

2) Let $M = \sum_{j \in \mathcal{N}} a_j$ be the total number of endorsers. If $M \geq T$, the statement is published along with the list of all M endorsers. The contract is then considered "successful."

3) If $M < T$, the identities of the endorsers remain confidential from each other, the public, and potentially the contract administrator. No statement is released, and the contract is considered "failed."

This structure mirrors the core logic of assurance contracts (Bagnoli and Lipman 1989; Tabarrok 1998) while introducing two crucial differences. First, there is complete secrecy about who has signed the contract. Second, participation payoffs include psychological terms that are state-dependent. Warm-glow from participation accrues whenever one signs (and whether the contract

1. Here beliefs are treated as exogenous, but of course what is acceptable to talk about publicly will influence not just publicly stated beliefs but also privately-held ones (Capuano and Chekroun 2024). Indeed, there is a small literature about operationalizing this fact as "Social Norms Marketing" (Nolan and Wallen 2021)

succeeds or fails), reputational esteem materializes on success, and sanctions bite when expression is public. Treating these as utility terms rather than monetary transfers matters for both design and welfare. The equilibrium selection results continue to go through when these terms are replaced by equivalent pecuniary payments that enter additively; however, state-dependence and curvature (e.g., convex sanction costs) break simple subsidy equivalence and shift the optimal threshold T. This paper quantifies those shifts and shows how they inform threshold choice under explicit objectives.

B. The Agent's Decision Problem: Payoffs

The components of utility are:

Individual Benefit (v_i)—If the contract succeeds ($M \geq T$), agent $i \in \mathcal{N}$ receives this individual-specific benefit $v_i > 0$. This represents agent i's personal valuation of the public good created by the revealed consensus (e.g., a policy shift, norm change, or validation of the view). This benefit v_i is received by agent $i \in \mathcal{N}$ if the contract succeeds, regardless of their own action a_i. For $i \in \mathcal{N}$, v_i is heterogeneous. Individuals $k \in \mathcal{P} \setminus \mathcal{N}$ (for whom $v_k \leq 0$) are not considered potential endorsers in this model, though their existence and potential opposition could be an avenue for future research.

Private Cost (c_i)—If an agent $i \in \mathcal{N}$ chooses $a_i = 1$ (Sign) *and* the contract succeeds, they incur a private cost $c_i > 0$. This cost arises from being publicly associated with a controversial statement. c_i is heterogeneous across individuals. A more specific formulation could be $c_i(\delta_i, M)$, where $\delta_i = |\theta_i - \hat{\theta}|$ is the perceived deviation of agent i's true belief θ_i from the pre-existing public norm $\hat{\theta}$, and M is the number of endorsers. Costs might be convex in δ_i, e.g., $c_i = \kappa_i \delta_i^{\gamma_i}$ with $\gamma_i > 1$.

Private Esteem (e_i)—If an agent $i \in \mathcal{N}$ chooses $a_i = 1$ (Sign) *and* the contract succeeds, they may also receive a private esteem payoff $e_i \geq 0$. This can represent social recognition, personal satisfaction, or enhanced reputation. e_i is also heterogeneous.

Note: For convenience, write $x_i \equiv e_i - c_i$; under additive payoffs this reparametrizes individual net private terms without changing the equilibrium form.

Warm-Glow (w)—An agent $i \in \mathcal{N}$ who chooses $a_i = 1$ (Sign) receives a small intrinsic "warm-glow" payoff $w > 0$ (common to all endorsers), *regardless* of whether the contract succeeds or fails. This captures non-consequentialist utility from participation (Andreoni 1990; Menges, Schroeder, and Traub 2005; Rose et al. 2002; Mitra and Moore 2018); the feeling of "Well, at least I've done *something*" experienced by the signer in all states where $a_i = 1$, rather than only on success. The parameter w is crucial, akin to the bonus F in Tabarrok (1998)'s dominant assurance contracts,

though w here is an intrinsic utility from participation experienced by the signer in all states where $a_i = 1$, rather than a contingent monetary transfer from an organizer that is not paid if the contract succeeds.

The payoffs for agent i are summarized in Table 1.

With equivalent monetary transfers substituted for the psychological terms, we can continue to write the net idiosyncratic component as $x_i \equiv e_i - c_i$, interpreting e_i to include any additive, state-independent payoff. Under this reparametrization, the fixed-point structure and cutoff logic are unchanged. From a design perspective, however, two features of esteem and sanctions remain distinct. First, these payoffs are typically *state-contingent*: they are realized only when expression is public, so their marginal effect depends on beliefs about that state, unlike an unconditional transfer of equal expected value. Second, curvature in $c_i(\cdot)$ means that even when expected values are matched, differences in how payoffs interact with participation beliefs can shift the optimal threshold T. These features imply that unconditional monetary transfers may not replicate the incentive or welfare effects of esteem and sanctions.

Table 1—Payoffs for Agent i under a social assurance contract

Action a_i	**Contract Fails** ($M < T$)	**Contract Succeeds** ($M \geq T$)
$a_i = 0$ (Not Sign)	0	v_i
$a_i = 1$ (Sign)	w	$v_i + e_i - c_i + w$

I assume agents are rational expected utility maximizers. The structure of the game and the payoff components (except for individual v_i, e_i, c_i types in the incomplete information setting) are common knowledge.

If agent i chooses $a_i = 1$ (Sign) and the contract succeeds, their utility is $v_i + e_i - c_i + w$. If it fails, their utility is w. If agent i chooses $a_i = 0$ (Not Sign) and the contract succeeds (due to others' actions), their utility is v_i. If it fails, their utility is 0.

A critical condition for choosing $a_i = 1$ emerges from comparing payoffs when the contract succeeds. If it succeeds, $a_i = 1$ gives $v_i + e_i - c_i + w$ and $a_i = 0$ gives v_i. Thus, choosing $a_i = 1$ yields a higher payoff in this state if $e_i - c_i + w \geq 0$. This is a key individual rationality constraint for costly participation when success is assured by one's action (or occurs anyway). Since $w > 0$, even if $e_i - c_i$ is negative (i.e. private cost outweighs private esteem), signing can still be preferable.

Assumption on Trust— For the baseline model developed in Sections III, IV, and V, I assume perfect trust in the contract administrator and the integrity of the mechanism. That is, agents believe with certainty ($\rho = 1$) that their identity will remain secret if the contract does not suc-

ceed ($M < T$). While this may seem like a strong assumption, it is motivated by the fact that modern cryptographic methods can create a "trustless" system, where the mechanism's integrity is guaranteed by code rather than a human administrator. Such an implementation would directly address risks like insider attacks or premature leaking of signatures. Smart contracts, threshold cryptography, and anonymous credential systems, when combined with robust identity verification such as digital certificates provided by an organization (e.g., a university), can provide certainty in the mechanism itself. Cryptographic tools for assurance contracts could also be programmed to delete the contract (to self-destruct) if the threshold is not reached in some pre-specified amount of time—thereby removing any doubts about attacks with tools developed in the future. While the full development of user-friendly platforms incorporating these guarantees is ongoing, these technologies theoretically offer a solution to problems of trust in the contract.

Of course, the theoretical existence of a technological solution to problems of trust does not imply that any given implementation is trustworthy. The technologies for implementing this sort of system are, fortunately, battle-tested in another domain: that of cryptocurrencies. Billions of dollars are held in smart contracts, there for the taking if only a chink in the armor can be found. And, indeed, chinks in the armor are occasionally found (Rezaei et al. 2025). Because of this, the resources expended on securing (and attacking) the technological pieces required will greatly outclass any efforts to compromise social assurance contracts. The technical aspects are discussed further in Appendix B.B2 and B.B3, and the implications of relaxing this trust assumption are explored in Appendix B.B1.

III. Equilibrium Analysis

This section characterizes equilibrium participation under complete and incomplete information and connects those characterizations to threshold design. The upshot is a cutoff rule: an agent signs if net idiosyncratic payoff $x_i \equiv e_i - c_i$ exceeds a pivotality-adjusted threshold that depends on w and beliefs about others. Incomplete-information equilibria exist and are unique under standard monotonicity conditions; they imply a simple comparative-statics map from w, esteem, cost curvature, and heterogeneity to the success probability $P(M \geq T)$ and the optimal threshold $T^\star$ for each design objective in Section V.

A. Complete Information

In the complete information setting, each agent i's type (v_i, e_i, c_i) is common knowledge, as are w, N, and T. Since the social assurance contract is a mechanism for discovering the distribution of

TABLE 2—SUMMARY OF MODEL ASSUMPTIONS AND KEY IMPLICATIONS

Feature / Implication	**Baseline: Complete Information** (Sec. III.A)	**Incomplete Information** (Sec III.B)
Agent's Knowledge of Others' Types (v_j, e_j, c_j)	Common Knowledge	Drawn from common prior $G(v, x)$
Trust in Admin (ρ)	Perfect ($\rho = 1$)	Perfect ($\rho = 1$)
Payoff if Agent Signs & Contract Fails ($M < T$)	w	w
Primary Signing Determinant(s)	Individual rationality checks (pivotal, dominant for $\mathcal{S}_{strong}$)	Cutoff $x^* = -w/p_{succ}(x^*)$ (using simplified type x_i)
Possibility of "All-Silent" Equilibrium	No (if $w > 0$)	Unlikely (if $w > 0$, x^* typically negative)
Role of Warm-Glow (w)	Breaks "all-silent"; key for $\mathcal{S}_{strong}$	Drives x^* negative, encouraging participation
Impact of v_i (Public Good Value)	Direct impact on pivotal decisions	Indirect via $P(\text{pivotal}_i)v_i$ (assumed small in simplified model)
Impact of $x_i = e_i - c_i$ (Net Idiosyncratic Payoff)	Key for $\mathcal{S}_{strong}$	Central via cutoff x^*
Analytical Complexity	Nash Equilibrium (Relatively Simple)	Bayesian Nash Equilibrium (Cutoff Strategy)
Overall Mechanism Effectiveness / Participation	High if $N_{strong} \geq T$	Depends on $w, F_x(x), T, N$

types, it would not be needed in this setting. However, as a toy example it is useful for developing intuitions.

Consider agent i's decision, $a_i \in \{0, 1\}$. Let $M_{-i} = \sum_{j \neq i, j \in \mathcal{N}} a_j$ be the number of other potential endorsers who choose $a_j = 1$. Agent i compares $U_i(a_i = 1|M_{-i})$ with $U_i(a_i = 0|M_{-i})$.

- If $M_{-i} \geq T$ (contract succeeds regardless of i's action): $U_i(a_i = 1) = v_i + e_i - c_i + w$. $U_i(a_i = 0) = v_i$. Agent i prefers $a_i = 1$ if $e_i - c_i + w \geq 0$.
- If $M_{-i} = T - 1$ (agent i is pivotal for success): $U_i(a_i = 1) = v_i + e_i - c_i + w$ (contract succeeds). $U_i(a_i = 0) = 0$ (contract fails). Agent i prefers $a_i = 1$ if $v_i + e_i - c_i + w \geq 0$.
- If $M_{-i} < T-1$ (contract fails regardless of i's action, i.e., even if i chooses $a_i = 1$, $1+M_{-i} < T$): $U_i(a_i = 1) = w$. $U_i(a_i = 0) = 0$. Agent i prefers $a_i = 1$ if $w > 0$, which is true by assumption.

This structure, especially when an agent expects the contract to fail regardless of their action, highlights the power of the warm-glow $w > 0$. We can categorize agents based on these payoff considerations. Let $\mathcal{S}_{strong} = \{i \in \mathcal{N} \mid e_i - c_i + w \geq 0\}$ be the set of *strong supporters*, for whom the net idiosyncratic payoff from signing (esteem minus cost, plus warm glow) is non-negative if the contract succeeds, irrespective of their pivotality regarding the public good v_i. Let $\mathcal{S}_{pivotal} =$

$\{i \in \mathcal{N} \mid v_i + e_i - c_i + w \geq 0\}$ be the set of agents who prefer to sign if their action is pivotal for the contract's success. Note that $\mathcal{S}_{strong} \subseteq \mathcal{S}_{pivotal}$ since $v_i > 0$ for $i \in \mathcal{N}$.

Proposition 1 (Complete-information baseline)**.** *From this structure, a few statements immediately follow*

1) *Choosing $a_i = 1$ (Sign) is a weakly dominant strategy for any player $i \in \mathcal{S}_{strong}$ (i.e., if $e_i - c_i + w \geq 0$). More generally, if $w > 0$, choosing $a_i = 0$ (Not Sign) is never strictly dominant for any agent $i \in \mathcal{N}$. If the contract fails, $a_i = 1$ gives $w > 0$, while $a_i = 0$ gives 0. If the contract succeeds, choosing $a_i = 1$ gives $v_i + e_i - c_i + w$ versus v_i for $a_i = 0$.*

2) *The profile where all agents choose $a_j = 0$ (all-silent) is not a Nash equilibrium if $w > 0$. Any single player i can deviate to $a_i = 1$ and receive utility $w > 0$ (if the contract still fails and $1 < T$) or $v_i + e_i - c_i + w$ (if i is pivotal, $T = 1$, and this sum is > 0).*

3) *Let $N_{strong} = |\mathcal{S}_{strong}|$. If $N_{strong} \geq T$, there is a Nash equilibrium where all players $i \in \mathcal{S}_{strong}$ choose $a_i = 1$, and players $j \notin \mathcal{S}_{strong}$ (for whom $e_j - c_j + w < 0$) choose $a_j = 0$, provided that for these non-signing j, they are not pivotal or being pivotal does not make them wish to sign (i.e., $j \notin \mathcal{S}_{pivotal}$). The contract succeeds. See Appendix A.A1 for proof.*

The logic of dominant assurance contracts (Tabarrok 1998) is mirrored: the esteem e_i and cost c_i are only paid upon success. The warm-glow $w > 0$ ensures a positive payoff from choosing $a_i = 1$ if the project fails, versus 0 for $a_i = 0$. If $e_i - c_i + w > 0$, choosing $a_i = 1$ weakly dominates $a_i = 0$. The important part is that $w > 0$ ensures that if an agent believes the contract will fail (i.e., $M_{-i} < T - 1$), they strictly prefer $a_i = 1$. This breaks the all $a_j = 0$ equilibrium.

As mentioned, agents in $\mathcal{S}_{strong}$ are the "strong supporters." Let $\mathcal{S}_{weak} = \{i \in \mathcal{N} \mid (v_i + e_i - c_i + w \geq 0)$ and $(e_i - c_i + w < 0)\}$ be the set of "weak supporters." These are agents who choose $a_i = 1$ only if pivotal and their v_i is sufficiently large to offset the negative $e_i - c_i + w$. Note that $\mathcal{S}_{weak} = \mathcal{S}_{pivotal} \setminus \mathcal{S}_{strong}$. Individuals $k \notin \mathcal{S}_{pivotal}$ (for whom $v_k + e_k - c_k + w < 0$) are "non-supporters" who would choose $a_k = 0$ even if pivotal.

If $N_{strong} \geq T$, then a Nash equilibrium exists where all $i \in \mathcal{S}_{strong}$ choose $a_i = 1$. If some $j \in \mathcal{S}_{weak}$ are needed to reach T, they may also sign. Specifically, if there exists a set $A' \subseteq \mathcal{N}$ such that $|A'| \geq T$, all $i \in A'$ find it optimal to choose $a_i = 1$ given others in A' do, and all $k \notin A'$ find it optimal to choose $a_k = 0$, this constitutes a Nash equilibrium. If $N_{strong} < T$: All $i \in \mathcal{S}_{strong}$ choose $a_i = 1$. Other players $j \in \mathcal{S}_{weak}$ will choose $a_j = 1$ if they believe they are pivotal (i.e., their signature makes $M = T$) and $v_j + e_j - c_j + w \geq 0$. If the combined set of strong supporters and pivotal weak supporters is insufficient to reach T (i.e., $M < T$), then the contract fails. In

this scenario, everyone who chose $a_i = 1$ (which could be all N agents if $w > 0$ and they anticipate failure) receives utility w.

B. *Incomplete Information*

Now, agent i's private type includes their individual valuation of success v_i, their private esteem e_i, and private cost c_i. The contract is specified to keep who has signed secret from the public *and each other*, so potential endorsers are by design largely ignorant of the properties of those who have signed. Let $x_i = e_i - c_i$ be the net idiosyncratic component related to the social consequences of the contract's publication, a function of esteem and social censure. The agent's *type* is given by the tuple (v_i, x_i). I assume (v_i, x_i) are drawn independently and identically for each agent i from a known joint cumulative distribution function $G(v, x)$ with support $[\underline{v}, \overline{v}] \times [\underline{x}, \overline{x}]$. The parameters w, T, and N are treated as common knowledge. N can be thought of as the size of a specific group that would benefit from the success of the contract; this is, critically, *not* the same as the group for whom it could be net positive to speak out given the costs associated with doing so. Assuming N is common knowledge is a reasonable approximation for many real-world scenarios (e.g., employee-employer relations). Relaxing this assumption to allow for individual expectations over N would add realism but would also introduce considerable complexity, compromising the symmetry that allows for the derivation of a single equilibrium cutoff. Agents aim to maximize their expected utility.

A full Bayesian Nash equilibrium would require agents to form strategies based on their two-dimensional type (v_i, x_i). Agent i chooses $a_i = 1$ if $E[U_i(a_i = 1)|v_i, x_i] \geq E[U_i(a_i = 0)|v_i, x_i]$. Comparing the payoffs from Table 1, agent i signs if:

$$p_{succ|i}(v_i + x_i + w) + (1 - p_{succ|i})w \geq p_{succ|i,\neg i}v_i$$

where $p_{succ|i}$ is the probability of success if i signs, and $p_{succ|i,\neg i}$ is the probability of success if i does not sign (i.e., success is due to others), and $p_{succ|i} > p_{succ|i,\neg i}$. This can be rewritten considering agent i's pivotality. Let $P(\text{pivotal}_i)$ be the probability that $M_{-i} = T - 1$. Then the condition to sign is:

$$P(\text{pivotal}_i)(v_i + x_i + w) + P(M_{-i} \geq T)(x_i + w) + P(M_{-i} < T - 1)w \geq 0$$

This simplifies to $P(\text{pivotal}_i)v_i + p_{succ}(x_j \geq \sigma^*(v_j, x_j))x_i + w \geq 0$, where $\sigma^*(\cdot)$ denotes the symmetric equilibrium strategy mapping from a player's private information (v_i, x_i) to an action $a_i \in \{0, 1\}$, and p_{succ} is the probability that at least $T-1$ others sign based on their own (potentially

complex) strategies $\sigma^*(v_j, x_j)$. Solving for such a multi-dimensional equilibrium strategy $\sigma^*(v_j, x_j)$ is complex.

For tractability, I adopt a common simplification: I assume that the signing decision primarily hinges on a cutoff x^* for the net idiosyncratic payoff $x_i = e_i - c_i$, rather than a function of both v_i and x_i. This simplification can be justified by several observations: First, particularly as the number of potential endorsers N becomes large (formally, as $N \to \infty$), the probability of any single agent i being exactly pivotal for the realization of the public good v_i (i.e., $M_{-i} = T - 1$) approaches zero. Consequently, the expected utility component $P(\text{pivotal}_i)v_i$ derived from being pivotal becomes negligible. Thus, for settings involving a large number of potential participants this pivotal utility term is approximately zero, simplifying the focus to the more direct impacts of x_i and w. This simplification is particularly relevant for larger communities; in smaller groups where $P(\text{pivotal}_i)$ may not be negligible, the v_i component would likely retain more significance, leading to a more complex decision problem (potentially explaining why such formal contracts might be rarer or structured differently in such settings), though one would also presume that social connections would be stronger in such very small groups, and therefore N known with higher certainty. Second, the decision to participate in similar collective actions, such as voting, often occurs even when the probability of being pivotal is negligible. The "paradox of voting" is the observation that people vote, and in large numbers, despite the fact that they have effectively zero probability of being pivotal (Downs 1957). This suggests that non-instrumental motivations (akin to w in this model) or expressive benefits (related to x_i) can dominate strict pivotality calculations concerning the main outcome (v_i). Given these considerations, I assume agents simplify their decision by focusing on the components of utility they directly control or experience regardless of exact pivotality concerning v_i. The warm-glow w is received for signing irrespective of outcome. The idiosyncratic payoff $x_i = e_i - c_i$ is experienced conditional on success, which occurs with probability p_{succ} (from agent i's perspective, if they sign). Consistent with the typically negligible impact of the pivotal term, I assume that the influence of $P(\text{pivotal}_i)v_i$ does not systematically alter the signing threshold x^* for x_i for the majority of agents, or that this effect is sufficiently small to be abstracted away. The decision to sign is thus primarily driven by comparing the expected net idiosyncratic payoff x_i (conditional on success) and the warm-glow w, versus not signing. Agent i considers signing if the expected utility from x_i and w covers the implicit opportunity cost of 0 from not signing (if the contract fails) or not incurring x_i (if it succeeds anyway).

Under this simplifying assumption, agent i chooses $a_i = 1$ (Sign) if the expected payoff from signing, focusing on the x_i and w components relevant to the act of signing itself, is non-negative.

If agent i signs, they receive w regardless of outcome. If the contract succeeds (which happens with probability $p_{succ}(x^*)$ if i signs and others use cutoff x^*), they also receive x_i. Thus, agent i signs if:

$$w + p_{succ}(x^*)x_i \geq 0$$

This implies a cutoff rule: sign if $x_i \geq -w/p_{succ}(x^*)$. The equilibrium cutoff x^* for the type $x_i = e_i - c_i$ must therefore satisfy:

$$x^* = -\frac{w}{p_{succ}(x^*)} \tag{1}$$

where $p_{succ}(x^*)$ is the probability that at least $T-1$ other agents $j \in \mathcal{N} \setminus \{i\}$ have a type $x_j \geq x^*$ (and thus choose $a_j = 1$). Given that each of the $N-1$ other agents' types x_j is an independent draw from the marginal CDF $F_x(x)$, $p_{succ}(x^*)$ is given by:

$$p_{succ}(x^*) = \sum_{k=T-1}^{N-1} \binom{N-1}{k} (1 - F_x(x^*))^k F_x(x^*)^{N-1-k} \tag{2}$$

. Here $F_x(x)$ is the CDF of x_i. Therefore, the equilibrium cutoff x^* solves (1), with $p_{\text{succ}}(\cdot)$ defined in (2).

Proposition 2 (Existence and Properties of BNE under simplified type)**.** *Under standard regularity conditions on $F_x(\cdot)$ (such as continuity and a support that allows for the existence of a solution), a symmetric Bayesian Nash Equilibrium characterized by a cutoff x^* for the type $x_i = e_i - c_i$ satisfying Eq. 1 exists.*

This equilibrium cutoff x^* is defined by the condition $x^* = -w/p_{succ}(x^*)$. Given that $w > 0$, it follows that x^* must be negative if an equilibrium with $p_{succ}(x^*) > 0$ (i.e., where the contract has some chance of success if agent i signs) exists. This means that even agents who would suffer a net loss from $(e_i - c_i)$ if the contract succeeds (i.e., $x_i < 0$) might still choose $a_i = 1$ if the warm-glow w is sufficiently attractive relative to the probability of success. This highlights how the mechanism encourages participation.

The formal proof of existence, and an analysis of the conditions required for the uniqueness of this equilibrium, are provided in Appendix A.A2.

Robustness and Multiplicity of Equilibria—The existence of $w > 0$ is critical. If $w = 0$, then Eq. 1 becomes $x^* p_{succ}(x^*) = 0$. This implies either $x^* = 0$ (sign if $e_i \geq c_i$) or $p_{succ}(x^*) = 0$. The latter can lead to an equilibrium where no one chooses $a_j = 1$ if $T > 1$. The $w > 0$ term

ensures x^* is typically negative. While the cutoff x^* for Eq. 1 might be unique under some conditions, coordination failures could lead to pessimistic beliefs about $p_{succ}(x^*)$. However, the BNE formulation assumes agents correctly calculate $p_{succ}(x^*)$ based on $F_x(\cdot)$ and x^*.

C. Comparative Statics (Incomplete Information)

The equilibrium cutoff x^* for an agent's net idiosyncratic payoff $x_i = e_i - c_i$ responds to changes in the model's parameters. Assuming a unique interior equilibrium x^* exists, we can analyze these responses. A formal derivation of these comparative statics using the Implicit Function Theorem is provided in Appendix A.A3.

An increase in the warm-glow w makes participation more attractive regardless of the outcome. Consequently, the equilibrium cutoff x^* decreases (becomes more negative), implying that agents with a less favorable idiosyncratic payoff x_i are now willing to choose $a_i = 1$. This leads to an overall increase in participation.

Shifts in the underlying distribution of esteem e_i and costs c_i, as captured by the CDF $F_x(x)$, also affect the equilibrium. If the distribution $F_x(x)$ shifts such that $x_i = e_i - c_i$ becomes stochastically higher (e.g., average esteem increases or average costs decrease), the probability of success $p_{succ}(x^*)$ for any given cutoff x^* increases. To maintain the equilibrium condition $x^* p_{succ}(x^*) = -w$, the cutoff x^* typically increases (becomes less negative). While this makes the cutoff itself less stringent, the overall effect on participation $(1 - F_x(x^*))$ depends on the magnitude of the distributional shift relative to the change in x^*; however, a common finding in such threshold models is that better underlying types ease the conditions for coordination.

The assurance threshold T directly influences the difficulty of achieving success. An increase in T makes success harder, reducing $p_{succ}(x^*)$ for any given x^*. To satisfy the equilibrium condition, x^* must decrease (become more negative). This means individuals must be willing to participate even with a less favorable x_i to compensate for the increased difficulty of reaching a higher T.

Changes in the population size N also have an impact. For a fixed T and x^*, an increase in N generally raises $p_{succ}(x^*)$, as there are more potential endorsers. This tends to increase x^* (making it less negative or more positive), meaning the signing condition becomes more stringent for any given individual. While the probability of any single agent signing might decrease, the total number of expected endorsers could still rise due to the larger pool. Finally, if the convexity of costs c_i (e.g., via a parameter γ) increases such that $x_i = e_i - c_i$ becomes stochastically lower (i.e., $F_x(x)$ shifts left), then $p_{succ}(x^*)$ decreases for any given x^*. This requires x^* to decrease (become more negative) to maintain equilibrium, implying individuals with worse idiosyncratic payoffs would need

to be willing to sign. Higher convexity (larger γ) FOSD-shifts F_x left, lowering $p_{\text{succ}}(x^*)$ for fixed x^* and, in equilibrium, inducing $dx^*/d\gamma < 0$.

IV. Information Revelation, Belief Updating, and Shifts in the Overton Window

A successful social assurance contract for a statement s_{belief} (representing an underlying belief $\mathbf{b}_s$) does more than achieve its immediate publication; it functions as a potent information revelation device. By demonstrating that $M \geq T$ individuals were willing to publicly endorse s_{belief} under the contract's specific incentive structure (T, w), it provides new public information that can significantly alter the collective understanding of the group's private views. This section formalizes how such revelations can lead to belief updating and, consequently, shift the boundaries of the Overton window—the range of beliefs considered acceptable for public discourse in a given forum R.

Defining the Overton Window based on Expression Costs—Let $\mathcal{I}$ be a space of beliefs, where each belief $\mathbf{b} \in \mathcal{I}$ can be conceptualized as a point in a d-dimensional real space. For any belief $\mathbf{b}$, let $c_{indiv}(\mathbf{b}, \mathcal{B}_t)$ denote the expected net cost for a representative agent to *unilaterally* (i.e., outside a social assurance contract mechanism) express belief $\mathbf{b}$ publicly at time t. This cost is a function of the prevailing collective beliefs $\mathcal{B}_t$ about the distribution of private support for $\mathbf{b}$ and related beliefs, which in turn shapes perceived reputational risks ($\bar{c}_{\mathbf{b}}(\mathcal{B}_t)$) and potential esteem ($\bar{e}_{\mathbf{b}}(\mathcal{B}_t)$). The Overton window for forum R at time t, $O_R(\mathcal{B}_t, \tau)$, can be defined as the set of beliefs whose individual expression cost (net of any individual esteem from unilateral expression) is below a societal or forum-specific tolerance threshold τ:

$$O_R(\mathcal{B}_t, \tau) = \{\mathbf{b} \in \mathcal{I} \mid c_{indiv}(\mathbf{b}, \mathcal{B}_t) \leq \tau\}$$

A belief $\mathbf{b}_s$ is considered "outside the window" before the contract if $c_{indiv}(\mathbf{b}_s, \mathcal{B}_t) > \tau$. The contract is typically invoked for such a belief precisely because individual expression is too costly.

This definition turns an architect's discourse goal into a design target: choose a threshold T that raises the posterior share of publicly expressible support for the focal statement above the tolerance level τ with probability at least π. In Section V I show how to map an empirical or elicited distribution of net willingness to sign into a choice of T that achieves a specified Overton-window expansion goal with target confidence π.

The Contract Outcome as a Public Signal—The success of a social assurance contract for the statement s_{belief} (representing belief $\mathbf{b}_s$), revealing $M \geq T$ endorsers, provides a public signal $y_s = (M, T, w, \text{characteristics if revealed})$. This signal is generated by individuals $i \in \mathcal{N}$ comparing

their private type $x_i = e_i - c_i$ (related to belief $\mathbf{b}_s$) to an equilibrium cutoff x^* (derived from Equation 1), and also considering their private valuation v_i for $\mathbf{b}_s$'s success. The observation of y_s allows agents to update their beliefs $\mathcal{B}_t$. For example, suppose agents are uncertain about underlying parameters θ_{params} that govern the joint distribution $G(v, x)$ from which individual types (v_i, x_i) for belief $\mathbf{b}_s$ are drawn. Their prior beliefs about these parameters can be denoted $P(\theta_{params}|\mathcal{B}_t)$. After observing y_s, they update to a posterior $P(\theta_{params}|y_s, \mathcal{B}_t)$ using Bayes' rule:

$$P(\theta_{params}|y_s, \mathcal{B}_t) \propto P(y_s|\theta_{params}, T, w, x^*(\theta_{params})) \cdot P(\theta_{params}|\mathcal{B}_t)$$

Here, $P(y_s|\theta_{params}, T, w, x^*(\theta_{params}))$ is the likelihood of observing M endorsers given the true underlying parameters θ_{params} (which determine $G(v, x)$ and thus $F_x(x)$), the contract terms (T, w), and the resulting equilibrium cutoff $x^*(\theta_{params})$ that depends on these underlying parameters. This inference leads to an updated overall belief state $\mathcal{B}_{t+1}$, which primarily reflects a revised perception of the prevalence and intensity of support (i.e., favorable x_i values and high v_i values) for belief $\mathbf{b}_s$ within the population $\mathcal{N}$.

Shifting the Window's Boundaries—The critical step is how this updated belief state $\mathcal{B}_{t+1}$ concerning $\mathbf{b}_s$ affects the expected individual expression costs $c_{indiv}(\mathbf{b}', \mathcal{B}_{t+1})$ for $\mathbf{b}_s$ itself and for other beliefs $\mathbf{b}' \in \mathcal{I}$. First, there is the impact on the expressed belief $\mathbf{b}_s$. The demonstration of significant support for $\mathbf{b}_s$ can directly reduce its future individual expression cost. The perceived risk $\bar{c}_{\mathbf{b}_s}(\mathcal{B}_{t+1})$ may fall, and/or perceived esteem $\bar{e}_{\mathbf{b}_s}(\mathcal{B}_{t+1})$ from unilateral expression may rise, because the belief is now known not to be fringe. If $c_{indiv}(\mathbf{b}_s, \mathcal{B}_{t+1}) \leq \tau$, then $\mathbf{b}_s$ has effectively entered the Overton window for unilateral expression. The contract has served as a beachhead. Second, there are spillover effects on related beliefs $\mathbf{b}'$. The belief update concerning $\mathbf{b}_s$ can spill over to semantically or ideologically related beliefs. Let $d(\mathbf{b}_s, \mathbf{b}')$ be a measure of distance or dissimilarity in the belief space $\mathcal{I}$. The change in perceived support for $\mathbf{b}_s$ might influence perceived support for $\mathbf{b}'$: $\Delta\mathbb{E}[\text{support}_{\mathbf{b}'}] = g(\Delta\mathbb{E}[\text{support}_{\mathbf{b}_s}], d(\mathbf{b}_s, \mathbf{b}'))$, where g is typically decreasing in distance. If $\mathbf{b}'$ is "close" to $\mathbf{b}_s$ (e.g., $d(\mathbf{b}_s, \mathbf{b}') < \epsilon_{neighbor}$), or positively correlated in the public mind, the perceived cost $c_{indiv}(\mathbf{b}', \mathcal{B}_{t+1})$ might decrease. Some such $\mathbf{b}'$ previously outside $O_R(\mathcal{B}_t, \tau)$ may now enter $O_R(\mathcal{B}_{t+1}, \tau)$. The window effectively "widens" or "drags" neighboring beliefs with it. For beliefs $\mathbf{b}''$ that are "oppositional" or negatively correlated with $\mathbf{b}_s$, the increased perceived viability of $\mathbf{b}_s$ might increase the perceived cost $c_{indiv}(\mathbf{b}'', \mathcal{B}_{t+1})$ (e.g., by making $\mathbf{b}''$ seem more isolated or its proponents more deviant). Such beliefs might fall out of the window, representing a "shift" or "tilt" rather than a simple expansion, though this seems less clear. The Overton window at $t+1$ is $O_R(\mathcal{B}_{t+1}, \tau)$. The formal shift is then characterized by the set differences: $O_R(\mathcal{B}_{t+1}, \tau) \setminus O_R(\mathcal{B}_t, \tau)$

(newly permissible beliefs) and $O_R(\mathcal{B}_t, \tau) \setminus O_R(\mathcal{B}_{t+1}, \tau)$ (beliefs no longer as acceptable).

This dynamic—where a coordinated, protected expression act (the contract) alters the landscape of perceived support and subsequently the costs for future individual expression—is how the mechanism can transition a group from an equilibrium of silence (sustained by mis-calibrated beliefs $\mathcal{B}_t$) to one of more open expression (sustained by updated beliefs $\mathcal{B}_{t+1}$). The contract acts as the catalyst for this belief and norm shift. While a full dynamic model of the window's evolution is beyond the current scope, this framework illustrates its core mechanism.

A. Social Assurance Contracts and Depolarization

The phenomenon of political and social polarization is often characterized by societies sorting into two primary, seemingly irreconcilable factions (Duverger 1954; Iyengar, Sood, and Lelkes 2012), with public discourse appearing more extreme and antagonistic than the distribution of private individual beliefs might suggest (Iyengar and Krupenkin 2018; Fiorina and Abrams 2008). Such a divide can be characterized along two dimensions, attitude and affective polarization. People can have differing beliefs which leads to sorting, or have a more holistic group identities that entail dislike of other groups; and, of course, one may cause the other (Iyengar et al. 2019). This section explores how social assurance contracts can enable suppressed majorities or significant minorities *within* currently polarized groups to express themselves—a novel pathway to depolarization. They do this by, first, promoting authenticity and making each faction's public discourse more representative of its members' diverse private views. This, in turn, can create more authentic and potentially more moderate conversation if underlying views are less extreme than the intra-faction public conversation. The expressive goods produced could bear on both attitude and affective polarization: "Actually, I think those other guys aren't all that bad..." is as relevant a target belief as some substantive policy position. This analysis assumes a *de facto* two-party system where the public discussion is dominated by voices that are considerably more extreme than average, both within and between factions.

The Landscape of Polarized Discourse and Intra-Faction Dynamics—Consider a policy or belief space $\mathcal{I}$ (e.g., a unidimensional spectrum $[L, R]$). This society is largely divided into two factions, Blue Team (F_{Blue}) and Red Team (F_{Red}). While the distribution of private beliefs θ_i for individuals $i \in F_{Blue}$ (denoted $f_{Blue}(\theta)$) and $i \in F_{Red}$ (denoted $f_{Red}(\theta)$) may show considerable internal diversity and overlap, their respective *public discourses* are often dominated by voices and positions near the extremes of each faction's range. Let $\bar{\theta}^{public}_{Blue}$ and $\bar{\theta}^{public}_{Red}$ represent the perceived mean or dominant public stance of Blue Team and Red Team, respectively. Often, $\bar{\theta}^{public}_{Blue}$ is sig-

nificantly to the "left" of the mean private Blue Team belief $\int \theta f_{Blue}(\theta)d\theta$, and $\bar{\theta}^{public}_{Red}$ is to the "right" of $\int \theta f_{Red}(\theta)d\theta$. This can occur if extreme voices are louder, more organized, or if internal dissent/moderation is suppressed (Abramowitz and Saunders 2008; Kuran 1997).

Importantly, each faction $F \in \{Blue, Red\}$ cultivates its own internal Overton window, $O_F(\mathcal{B}_F, \tau_F)$ (defined analogously to O_R). For an individual $i \in F_{Blue}$, the expected net cost of unilaterally expressing a belief $\mathbf{b}$, $c_{indiv,Blue}(\mathbf{b}, \mathcal{B}_{Blue})$, is determined by the perceived norms and potential sanctions from *within Blue Team*. Ideas that deviate significantly from $\bar{\theta}^{public}_{Blue}$ (even if privately held by many Blue Team members, such as more moderate or nuanced positions $\mathbf{b}_{mod,Blue}$) may be "outside" $O_{Blue}(\mathcal{B}_{Blue}, \tau_{Blue})$ because expressing them incurs high intra-factional social costs (e.g., accusations of disloyalty, being a "Republican In Name Only" or "Democrat In Name Only"). Thus, a state of pluralistic ignorance can exist *within* each faction, where moderate members stay silent, believing they are in a smaller minority within their faction than they truly are. This internal silence reinforces the faction's extreme public-facing stance and exaggerates the difference between F_{Blue} and F_{Red}.

Social Assurance Contracts for Revealing Intra-Faction Moderate Consensus—A social assurance contract can be deployed *within* a faction (say, Blue Team) to challenge this internal dynamic. Suppose a social assurance contract is initiated for a statement $s_{mod,Blue}$ representing a more moderate belief $\mathbf{b}_{mod,Blue}$ that is privately supported by many in F_{Blue} but currently outside the Overton window $O_{Blue}(\mathcal{B}_{Blue}, \tau_{Blue})$. If this contract is successful (i.e., $M_{Blue} \geq T_{Blue}$ members of Blue Team sign), it serves as a powerful signal $y_{s_{mod,Blue}}$ primarily to other members of Blue Team. This signal leads to an update of intra-factional beliefs from $\mathcal{B}_{Blue}$ to $\mathcal{B}'_{Blue}$. Specifically, members of Blue Team update their perceptions of $f_{Blue}(\theta)$, realizing that support for moderate belief $\mathbf{b}_{mod,Blue}$ is more widespread among their Blue Team peers than previously thought.

This belief update can shift the intra-faction Overton window $O_{Blue}(\mathcal{B}'_{Blue}, \tau_{Blue})$. The expected cost for an individual Blue Team member to unilaterally express $\mathbf{b}_{mod,Blue}$ in the future, $c_{indiv,Blue}(\mathbf{b}_{mod,Blue}, \mathcal{B}'_{Blue})$, is likely to decrease as the perceived risk of intra-factional sanction diminishes. If $c_{indiv,Blue}(\mathbf{b}_{mod,Blue}, \mathcal{B}'_{Blue}) \leq \tau_{Blue}$, the belief $\mathbf{b}_{mod,Blue}$ enters Blue Team's internal Overton window. Similarly, other nearby moderate beliefs $\mathbf{b}'_{mod,Blue}$ may also become more "permissible" within Blue Team, as discussed in Section IV. The faction's internal discourse becomes more tolerant of moderation.

Mechanisms of Depolarization— Moderate expression *within* factions engendered by successful social assurance contracts can contribute to broader societal depolarization through several interconnected channels. One mechanism is the internal moderation of faction stances. As more mod-

erate views become expressible and are demonstrably supported within, for instance, Blue Team due to a contract's success, the faction's aggregate public discourse, $\bar{\theta}^{public}_{Blue}$, may itself begin to shift from its previously more extreme position. This shift would ideally move the public stance closer to the faction's actual private mean, $\int \theta f_{Blue}(\theta) d\theta$. Such a transformation occurs if these newly "permissible" moderate voices gain prominence and subsequently influence the faction's internal deliberations and its external communications.

This internal moderation can, in turn, lead to reduced inter-faction animosity and perceived distance. If Blue Team's public stance $\bar{\theta}^{public}_{Blue}(\mathcal{B}'_{Blue})$ becomes more moderate, it may appear less threatening or alien to members of Red Team, and vice-versa if Red Team also undergoes a similar internal shift. Such changes have the potential to reduce negative partisanship and affective polarization. Even if only one major faction moderates its public stance, the overall perceived "gap" or "extremity" in the political landscape can diminish. Formally, if initial polarization is measured by a distance metric $P_{initial} = d(\bar{\theta}^{public}_{Blue}(\mathcal{B}_{Blue}), \bar{\theta}^{public}_{Red}(\mathcal{B}_{Red}))$, and after internal moderation via social assurance contracts the new stances are $\bar{\theta}^{public}_{Blue}(\mathcal{B}'_{Blue})$ and $\bar{\theta}^{public}_{Red}(\mathcal{B}'_{Red})$, then depolarization can be quantified by observing $P_{final} = d(\bar{\theta}^{public}_{Blue}(\mathcal{B}'_{Blue}), \bar{\theta}^{public}_{Red}(\mathcal{B}'_{Red})) < P_{initial}$.

Furthermore, the successful operation of social assurance contracts within factions could contribute to weakening the influence of extreme flanks. The demonstrated presence of moderates could dilute the disproportionate influence previously wielded by more extreme, albeit highly vocal, elements. This could render the faction more resilient to internal capture by such elements and foster greater openness to nuanced policy positions that were previously difficult to surface. These processes could culminate in improved common knowledge across faction lines. If the internal moderation of Blue Team, for example, were to become evident to Red Team, it could help to correct Red Team members' potentially exaggerated stereotypes about Blue Team's inherent extremism, and vice-versa. This could even mean an increase in voting *for* and a decrease in voting *against*; fewer protest votes would suggest leaders who more accurately reflect their constituents' beliefs and desires. Finally, more accurate inter-group understanding—a form of common knowledge at the societal level—could foster more constructive dialogue and problem-solving. It remains to be seen how this would interact with the tendency to form two factions (Duverger 1954; Downs 1957). It is possible that pursuing highly polarized public discourse is actually the most effective strategy for a faction, meaning that revelation mechanisms like the social assurance contract, when deployed, would result in less competitive factions. Similarly, elites' perceptions of the distribution of opinion within the electorate is key for the policy and campaign decisions they make (Furnas and LaPira 2024). When there is false consensus among elites about underlying beliefs, successful assurance

contracts could offer corrective information.

B. Normative Concerns when Expanding the Overton Window

The contract mechanism is, of course, content-neutral. While it promotes truthful and democratic discourse by allowing any sufficiently supported view to surface, it also means that social assurance contracts could be used to reveal support for views that, from certain normative standpoints, seem harmful or regressive. This aspect does not undermine the mechanism's usefulness for better-aligning public discussion with privately held views, but it does highlight a potential normative challenge if such revealed preferences then gain influence. While increased transparency in opinion expression may be normatively desirable because truth-seeking is normatively desirable, balanced understanding of such contracts acknowledges this possibility. The hope is that by reducing misperceptions about where others stand, genuine disagreements can be addressed more openly and constructively, depolarizing public discourse by weakening the false consensus component of polarization that thrives on pluralistic ignorance.

V. Choosing the Assurance Threshold T

The contract architect chooses the assurance threshold T, a critical design parameter. This decision can be framed as the architect selecting T to maximize their utility function, u_a. The arguments of this utility function, $u_a(T; G(v,x), w, N, \mathcal{P} \setminus \mathcal{N}, O_R(\mathcal{B}_t, \tau), \dots)$, reflect what the architect values. For instance, u_a might depend on the probability of the contract succeeding ($P(M \geq T|T)$), the expected number of endorsers ($E[M|T]$), the likelihood of achieving a durable shift in the Overton window (as discussed in Section IV), or broader social welfare implications.

The choice of T involves inherent trade-offs. A T set too low (*e.g.*, $T = 1$) might result in publication with minimal participation, thereby limiting its impact on public discourse or its ability to confer substantial public good benefits (v_i). Conversely, a T set too high relative to the actual distribution of potential endorser types $x_i = e_i - c_i$ (characterized by $F_x(x)$) and their resulting equilibrium participation (influenced by $p_{succ}(x^*; T)$) makes contract failure more probable. An optimal T also considers the broader societal context, including the characteristics and potential reactions of individuals in $\mathcal{P} \setminus \mathcal{N}$ (for whom $v_i \leq 0$) and the existing state of the Overton window, $O_R(\mathcal{B}_t, \tau)$, as these factors can influence the ultimate effectiveness of revealing M signatures.

I now consider several natural objectives over choices of T, each tied to a different notion of success. One objective is to maximize the probability the contract succeeds, $P(M \geq T)$. This prioritizes revelation itself when success unlocks non-linear reputational or coordination gains. A

second objective is to maximize the probability of expanding the Overton window by at least a prespecified amount around the focal statement; this anchors T to a discourse-change target. A third objective is welfare-oriented: maximize expected social surplus from moving from silence to expression, net of participation costs and the externalities reflected in esteem and sanction terms. Finally, a risk-sensitive objective chooses T to meet a target success probability π at minimum expected cost, yielding a quantile-style rule.

For each objective, T trades off pivotality against risk. Higher T raises informational value at the cost of more coordination risk; lower T does the reverse. Under the incomplete-information equilibrium characterized above, the optimal T is characterized by a cutoff condition that equates the marginal gain from raising the required mass of signatures to the marginal loss in success probability. The comparative statics are simple: $T^\star$ falls with larger warm-glow w and baseline esteem, and rises with thicker sanction tails and greater convexity of individual costs.

Maximize Probability of Success $P(M \geq T)$— This objective, if pursued naïvely, may often suggest a low T (e.g., $T = 1$). Such a low assurance threshold might be met if there is a single person or a very small group for whom c_i is exceptionally low or $e_i + w$ is very high. While the contract would technically "succeed," this may not be meaningful in the context of social assurance contracts aimed at broader social change. With a very low T, the Overton window might not be meaningfully expanded, prevailing social norms may remain unshifted, and any desired practical policy or organizational change might not occur.

Maximize Probability of Expanding the Overton Window— A primary goal of a social assurance contract may be to ensure that the endorsed belief $\mathbf{b}_s$ not only gets published but also durably enters the Overton window. Following Section IV, $\mathbf{b}_s$ enters the window if, after the belief update $\mathcal{B}_t \to \mathcal{B}_{t+1}$ triggered by the contract's success, the expected net cost of *unilateral* expression falls: $c_{indiv}(\mathbf{b}_s, \mathcal{B}_{t+1}) \leq \tau$. The architect might believe that this durable shift requires the number of revealed endorsers M to meet or exceed a certain critical impact level, $M_{Overton}$. This $M_{Overton}$ reflects the number of endorsements deemed sufficient to significantly alter collective second-order beliefs ($\mathcal{B}_{t+1}$) and thereby reduce the perceived risks or increase the perceived esteem associated with unilaterally expressing $\mathbf{b}_s$ in the future. The architect's problem is then to choose the mechanism's assurance threshold T to maximize the probability $P(M \geq M_{Overton})$. This involves a careful trade-off. Setting $T = M_{Overton}$ directly aims the mechanism at the impact threshold. However, if $M_{Overton}$ is very high, this might lead to a low probability of success $P(M \geq M_{Overton})$ due to coordination challenges. Alternatively, the architect might set $T < M_{Overton}$ to increase the likelihood of the contract succeeding ($P(M \geq T)$ being higher). In this case, the architect relies

on the equilibrium dynamics—specifically, a sufficiently negative $x^*(T)$ resulting from the chosen T—to draw in a number of endorsers M that significantly exceeds T and reaches $M_{Overton}$. The optimal T would balance the directness of aiming for $M_{Overton}$ with the need to ensure a high enough probability of overall success and sufficient participation. Estimating $M_{Overton}$ itself requires an understanding of the social context and how beliefs about public opinion translate into individual expression costs. The problem is framed in Appendix D.

Maximize the Number of Endorsers M, Conditional on the Contract Succeeding— The architect might aim to set T in such a way that, if the contract succeeds, it does so with the largest possible number of endorsers. This involves a trade-off: a higher T directly sets a higher bar for M. However, a very high T drastically reduces the probability of success $p_{succ}(x^*)$. While it would make the cutoff x^* more negative (inducing a higher proportion of types to be willing to sign if success were perceived as likely), the overwhelming risk of failure might mean the expected number of endorsers in a successful outcome is lower than if a more moderate, achievable T had been chosen, or it could lead to outright contract failure. Conversely, a T that is too low might lead to success with few endorsers, not maximizing M. The architect would need to find a T that optimally balances the likelihood of success with the number of participants drawn in. This often means setting T at a level that is ambitious but perceived as achievable by a substantial portion of the target population, encouraging widespread participation rather than just meeting a minimal threshold. This objective seeks to maximize the visible display of support once the contract triggers.

Maximize the Architect's Own Expected Utility— An architect who is also a member of $\mathcal{N}$ (a participant-organizer) might choose T to maximize their own expected utility. This architect has personal stakes (v_a, e_a, c_a, w_a), which may differ from the average (e.g., higher e_a for leadership, potentially higher c_a due to visibility, or a distinct warm-glow w_a related to their efforts), and also incurs an organizing cost $o_a > 0$. The architect's first decision is their own signing strategy, $a_a^*(T) \in \{0, 1\}$, for any given T. If the architect signs ($a_a = 1$), their expected utility is

$$E[U_a(T)|a_a = 1] = P(M_{-a} \geq T - 1|T) \cdot (v_a + e_a - c_a) + w_a - o_a$$

Here, M_{-a} is the number of other participants who sign. The probability $P(M_{-a} \geq T - 1|T)$ that at least $T - 1$ others sign (making the contract succeed if the architect signs) is equivalent to $p_{succ}(x^*(T); T)$ as defined in Eq. (2), where $x^*(T)$ is the equilibrium cutoff for the other $N - 1$ potential participants. If the architect does not sign ($a_a = 0$), their expected utility is

$$E[U_a(T)|a_a = 0] = P(M_{-a} \geq T|T) \cdot v_a - o_a$$

Here, $P(M_{-a} \geq T|T) = \sum_{k=T}^{N-1} \binom{N-1}{k}(1 - F_x(x^*(T)))^k F_x(x^*(T))^{N-1-k}$ is the probability that at least T of the other $N-1$ participants sign (making the contract succeed without the architect's signature). The architect would choose $a_a^*(T)$ by comparing these expected utilities for a given T. Then, the architect chooses T to maximize $E[U_a(T)|a_a^*(T)]$. This objective highlights how an organizer's personal incentives, specific type characteristics, and the costs of organization can directly shape the contract's design.

Maximize Expected Social Welfare— In the most basic case, the architect aims to maximize $E[W|T] = P(M \geq T|T)E[W_{succ}|M \geq T, T] + P(M < T|T)E[W_{fail}|M < T, T]$, where $W_{succ} = \sum_{k:a_k=0} v_k + \sum_{j:a_j=1}(v_j + e_j - c_j + w)$ and $W_{fail} = (\sum a_j)w$. This requires knowledge of the joint distribution of (v_i, e_i, c_i). The architect chooses T to maximize this expression. The probabilities $P(M \geq T|T)$ and other expectations depend on the equilibrium cutoff $x^*(T)$, which is determined by $x^* = -w/p_{succ}(x^*; T)$ from Equation (1). Note that this strategy requires the architect to first define the population over which welfare is to be maximized. This could be the population $\mathcal{P}$ as a whole, which would typically involve extending the summation for v_k in W_{succ} to include all individuals in $\mathcal{P} \setminus \mathcal{N}$ who are affected by the contract's success (and whose v_k may be non-positive). This could also be limited to the set of potential endorsers $\mathcal{N}$ (in which case the sums in W_{succ} and W_{fail} apply directly to members of $\mathcal{N}$ based on their actions). It seems quite natural that an architect would focus on the welfare of potential endorsers, or a subset of potential endorsers.

The optimal T is likely context-dependent. If many potential endorsers have high v_i but moderately negative $x_i = e_i - c_i$, a T that is achievable (given $x^* < 0$ due to $w > 0$) might be optimal if it unlocks substantial $\sum v_i$. In practice: elicit or estimate the distribution of net willingness to sign, choose an objective (probability, discourse shift, surplus, or risk), and compute the implied cutoff rule for $T^\star$; Section IV provides the link from T to observable discourse change.

VI. Future Directions

A. Theoretical Extensions and Refinements

Further theoretical work could extend this model in several important directions, offering a richer understanding of the strategic environment:

Truth and Competing Contracts—One may wonder whether an environment with social assurance contracts is more likely to guide the public discussion towards truth relative to one without. Investigating scenarios with multiple social assurance contracts, possibly with opposed or complementary objectives, vs. none at all is an important topic for future research and may shed some

light on this question.

Endogenous Network Effects and Information Diffusion—A key extension is to more formally model how network topology influences outcomes. This includes how it affects the perceived probability of success $p_{succ}(x^*)$, individual cost-benefit calculations (v_i, e_i, c_i), and the diffusion of information about the contract's existence and its accumulating support. This could involve analyzing how knowledge (or rumors) about who has already chosen $a_j = 1$ might leak—perhaps deliberately by endorsers themselves—and how such dynamics could affect success. Appendix C discusses these possibilities in more detail.

The Role of Time—The present model does not explore how timing may influence the success and failure of social assurance contracts. A person's opinion may change with time, and so time limits on the validity of a signature may be reasonable. Similarly, a time limit on the contract itself may be a good idea. A contract that does not succeed quickly, or at least within some reasonable timeframe, may outlive its usefulness and become a liability to endorsers. Contracts that are destroyed, or which are programmed to self-destruct, may be used to handle this contingency. It may also be a good idea under some circumstances to let a contract run after T is reached, either having revealed the T endorsers or not. In the case where signatures are coming in quicker than expected, more of an effect may be achieved by increasing T or allowing signatures after publication.

Dynamic and Iterative Models—Exploring the implications of repeated interactions, such as iterative assurance contracts where outcomes of one contract influence subsequent ones, or sequential decision-making by agents, would be a valuable extension. This could also involve analyzing how parameters like individual costs c_i, esteem e_i, public good valuations v_i, or the type distribution $F_x(x)$ evolve based on past contract successes or failures.

Richer Information Structures—Incorporating more complex belief systems, such as higher-order beliefs, or introducing asymmetric information that extends beyond payoffs could yield new insights. Various signaling aspects during the signing process, where early actions might convey information to later potential endorsers, also warrant investigation.

Variations in Contract Design—Exploring more flexible contract designs, such as staged contracts where a vanguard group might have a lower initial assurance threshold $T_{vanguard}$ to help publicize the contract and increase the potential pool of endorsers N for a main, higher assurance threshold T_{main}. It could also be the case that different classes of endorsers have different assurance thresholds: a contract might require 100 tenured professors, 200 untenured faculty, and 400 graduate students, for instance. Another design, similar to ideas in Ayres and Unkovic (2012), could allow each individual endorser i to specify their own assurance threshold T_i, with tranches of signatures becoming public

as these individual thresholds are met. The number of current endorsers (but not their identities) might also be made public before the contract meets its assurance threshold T. This could create momentum effects, strategic waiting, or herding/anti-herding behavior. Analyzing the case where the order of endorsers is revealed upon success could be interesting as well, since this might affect an individual's perceived esteem e_i or cost c_i depending on their position in the revealed list (e.g., being an early versus a late endorser).

Broader Population Dynamics—Expanding the model to include the decisions and influence of individuals in the wider population $\mathcal{P}\backslash\mathcal{N}$ (for whom $v_i \leq 0$). This would involve exploring how those who are indifferent or actively opposed to the contract's success might interact with the mechanism, for example, through counter-mobilization or attempts to discredit the contract, beyond the simple adversarial attacks currently considered. The problem is framed in Appendix D. Similarly, the model could be extended from two-faction to multi-faction settings.

B. Empirical Investigation and Testable Predictions

The model elaborated here makes several predictions that are in principle empirically testable. Data from platforms facilitating social assurance contracts, laboratory and field experiments would be allow testing and refining this theoretical framework. Some predictions include:

Impact of Warm-Glow (w)—Contracts or platforms offering stronger explicit or implicit warm-glow incentives for participation (e.g., forms of acknowledgment even if the contract fails) should, *ceteris paribus*, observe higher participation rates or a lower (more negative) effective equilibrium cutoff x^*.

Impact of the Assurance Threshold (T)—The model predicts that a higher assurance threshold T leads to a lower (more negative) equilibrium cutoff x^*. While this implies that, conditional on success, a larger proportion of the relevant population might be willing to sign (as the individual bar $x_i \geq x^*$ is lower), the overall probability of success is expected to decrease with very high T. The relationship between T and the actual number of endorsers M in successful contracts, as well as success rates, can be empirically investigated.

Impact of Perceived Success Probability ($p_{succ}(x^*)$)—Factors hypothesized to increase agents' subjective $p_{succ}(x^*)$ (e.g., clear evidence of broad underlying support for the statement, effective network communication about the contract, a lower T relative to the estimated N) should correlate with a higher (less negative) x^*, making it easier for marginal individuals to choose $a_i = 1$. This could be proxied by comparing contracts on topics with different levels of perceived underlying support.

Impact of Population Size (N)—The model suggests that for a fixed T, a larger N leads to a higher (less negative) x^*. This implies that while more potential endorsers are available, the condition for any one individual to sign becomes more stringent. Empirical settings where N varies for similar types of contracts could be used to test this.

Impact of Trust (ρ)—As discussed in Section B.B1, higher perceived trustworthiness and security of the platform or contract administrator (ρ closer to 1) should lead to higher participation rates, particularly when the warm-glow w is low or average costs c_i are perceived to be high. This could be tested by observing changes in participation after trust-enhancing features are implemented or including survey questions about trust, including bets on the likelihood of malfeasance.

VII. Conclusion

This paper has proposed and formally analyzed social assurance contracts as a mechanism to overcome the coordination failures and information scarcity that lead to pluralistic ignorance and self-censorship. By adapting the logic of economic assurance contracts to the domain of social expression, where contributions are reputational and payoffs involve social costs and esteem, this mechanism enables individuals to pledge support for a controversial statement with safety. The game-theoretic analysis under both complete and incomplete information demonstrates that such contracts can shift behavior towards truthful revelation of private beliefs. The introduction of a warm-glow bonus w for signing plays a crucial role in making participation more attractive, potentially making signing a dominant strategy under certain conditions and lowering the effective threshold for participation in Bayesian Nash Equilibria. The model provides a framework for understanding how individuals weigh the benefits of collective expression (v_i, e_i) against personal risks (c_i) and the intrinsic value of participation (w). A key contribution of this work lies in formally modeling how such a mechanism can directly impact pluralistic ignorance and, by revealing hidden consensus, potentially shift the Overton window. This provides a micro-founded explanation for how public norms can realign with private truths, potentially reducing polarization that stems from mis-perceptions. The practical success of social assurance contracts demands trust in the mechanism. Robust safeguards against vulnerabilities like Sybil attacks and malicious organizers, using modern identity verification systems, cryptographic methods, and decentralized trust mechanisms, can provide reliable platforms for such contracts. Potential applications range from enabling faculty and students in academic communities to challenge dominant norms, to facilitating collective whistleblowing or feedback in corporate settings. Furthermore, this mechanism could empower moderates in political organizations, allow for unofficial no-confidence votes, or be inte-

grated into online platforms to mitigate chilling effects and reveal latent support for non-dominant views. For practitioners, the contribution is operational: the Overton formalization, paired with the equilibrium cutoff, turns discourse goals into a concrete choice of T with clear comparative statics and measurable inputs. As societies grapple with polarization and the challenges to free expression, tools that enable conditional collective voice are increasingly valuable. Social assurance contracts offer a promising way to foster more honest, inclusive, and ultimately informed public conversations—in the end, building genuine common knowledge.

References

Abramowitz, Alan I., and Kyle L. Saunders. 2008. "Is Polarization a Myth?" *The Journal of Politics* 70, no. 2 (April): 542–555. ISSN: 0022-3816. https://doi.org/10.1017/S0022381608080493.

Agrawal, Ajay, Christian Catalini, and Avi Goldfarb. 2014. "Some Simple Economics of Crowdfunding." *Innovation Policy and the Economy* 14 (January 1, 2014): 63–97. ISSN: 1531-3468. https://doi.org/10.1086/674021.

Ahler, Douglas J. 2014. "Self-Fulfilling Misperceptions of Public Polarization." *The Journal of Politics* 76, no. 3 (July): 607–620. ISSN: 0022-3816. https://doi.org/10.1017/S0022381614000085.

Andreoni, James. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving." *The Economic Journal* 100, no. 401 (June 1, 1990): 464–477. ISSN: 0013-0133. https://doi.org/10.2307/2234133.

———. 1995. "Cooperation in Public-Goods Experiments: Kindness or Confusion?" *The American Economic Review* 85 (4): 891–904. ISSN: 0002-8282, accessed May 2, 2025. JSTOR: 2118238. https://www.jstor.org/stable/2118238.

Apolte, Thomas, and Julia Müller. 2022. "The Persistence of Political Myths and Ideologies." *European Journal of Political Economy* 71 (January 1, 2022): 102076. ISSN: 0176-2680. https://doi.org/10.1016/j.ejpoleco.2021.102076.

Ayres, Ian, and Cait Unkovic. 2012. "Information Escrows." *Michigan Law Review* 111:145. https://heinonline.org/HOL/Page?handle=hein.journals/mlr111&id=165&div=&collection=.

Bagnoli, Mark, and Barton L. Lipman. 1989. "Provision of Public Goods: Fully Implementing the Core through Private Contributions." *The Review of Economic Studies* 56, no. 4 (October 1, 1989): 583–601. ISSN: 0034-6527. https://doi.org/10.2307/2297502.

Bikhchandani, Sushil, David Hirshleifer, Omer Tamuz, and Ivo Welch. 2024. "Information Cascades and Social Learning." *Journal of Economic Literature* 62, no. 3 (September): 1040–1093. ISSN: 0022-0515. https://doi.org/10.1257/jel.20241472.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. 1992. "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." *Journal of Political Economy* 100, no. 5 (October): 992–1026. ISSN: 0022-3808, 1537-534X. https://doi.org/10.1086/261849.

Callisto. "Callisto." Callisto. Accessed April 29, 2025. https://www.projectcallisto.org.

Camenisch, Jan, and Anna Lysyanskaya. 2001. "An Efficient System for Non-transferable Anonymous Credentials with Optional Anonymity Revocation." In *Advances in Cryptology — EUROCRYPT 2001,* edited by Birgit Pfitzmann, redacted by Gerhard Goos, Juris Hartmanis, and Jan Van Leeuwen, 2045:93–118. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-540-42070-5. https://doi.org/10.1007/3-540-44987-6_7.

Capuano, Carla, and Peggy Chekroun. 2024. "A Systematic Review of Research on Conformity." *International Review of Social Psychology* 37, no. 1 (July). https://doi.org/10.5334/irsp.874.

Carlsson, Hans, and Eric van Damme. 1993. "Global Games and Equilibrium Selection." *Econometrica* 61 (5): 989–1018. ISSN: 0012-9682. https://doi.org/10.2307/2951491. JSTOR: 2951491.

Cason, Timothy N., and Robertas Zubrickas. 2017. "Enhancing Fundraising with Refund Bonuses." *Games and Economic Behavior* 101 (January): 218–233. ISSN: 08998256. https://doi.org/10.1016/j.geb.2015.11.001.

Catalyst, Stanford. "Stanford Catalyst." Stanford Catalyst. Accessed April 29, 2025. https://web.archive.org/web/20150323122726/http://catalyst.stanford.edu/.

Cheng, Justin, and Michael Bernstein. 2014. "Catalyst: Triggering Collective Action with Thresholds." In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing,* 1211–1221. CSCW'14: Computer Supported Cooperative Work. Baltimore Maryland USA: ACM, February 15, 2014. ISBN: 978-1-4503-2540-0. https://doi.org/10.1145/2531602.2531635.

CollAction. "CollAction." CollAction. Accessed April 29, 2025. https://www.collaction.org/.

Croson, Rachel T. A., and Melanie Beth Marks. 2000. "Step Returns in Threshold Public Goods: A Meta- and Experimental Analysis." *Experimental Economics* 2, no. 3 (March 1, 2000): 239–259. ISSN: 1386-4157, 1573-6938. https://doi.org/10.1007/BF01669198.

Dey, Debabrata, Atanu Lahiri, and Rajiv Mukherjee. 2025. "Polarization or Bias: Take Your Click on Social Media." *Journal of the Association for Information Systems* 26, no. 3 (January 1, 2025): 850–878. ISSN: 1536-9323. https://doi.org/10.17705/1jais.00925.

Douceur, John R. 2002. "The Sybil Attack." In *Peer-to-Peer Systems,* edited by Peter Druschel, Frans Kaashoek, and Antony Rowstron, 251–260. Berlin, Heidelberg: Springer. ISBN: 978-3-540-45748-0. https://doi.org/10.1007/3-540-45748-8_24.

Downs, Anthony. 1957. *An Economic Theory of Democracy.* In collaboration with Internet Archive. New York : Harper.

Duffy, John, and Jonathan Lafky. 2018. "Living a Lie: Theory and Evidence on Public Preference Falsification." *Department of Economics Working Paper Series* (September 1, 2018). https://digitalcommons.carleton.edu/econ_repec/15.

Duverger, Maurice. 1954. *Political Parties: Their Organisation and Activity in the Modern State.* 1st English language ed. London: Methuen.

Dybvig, Philip H., and Chester S. Spatt. 1983. "Adoption Externalities as Public Goods." *Journal of Public Economics* 20, no. 2 (March 1, 1983): 231–247. ISSN: 0047-2727. https://doi.org/10.1016/0047-2727(83)90012-9.

Esser, James K. 1998. "Alive and Well after 25 Years: A Review of Groupthink Research." *Organizational Behavior and Human Decision Processes* 73, nos. 2–3 (February): 116–141. ISSN: 07495978. https://doi.org/10.1006/obhd.1998.2758.

Fiorina, Morris P., and Samuel J. Abrams. 2008. "Political Polarization in the American Public." *Annual Review of Political Science* 11 (Volume 11, 2008 2008): 563–588. ISSN: 1094-2939, 1545-1577. https://doi.org/10.1146/annurev.polisci.11.053106.153836.

Furnas, Alexander C., and Timothy M. LaPira. 2024. "The People Think What I Think: False Consensus and Unelected Elite Misperception of Public Opinion." *American Journal of Political Science* 68 (3): 958–971. ISSN: 1540-5907. https://doi.org/10.1111/ajps.12833.

Glynn, Adam N. 2013. "What Can We Learn with Statistical Truth Serum?: Design and Analysis of the List Experiment." *Public Opinion Quarterly* 77, no. S1 (January 1, 2013): 159–172. ISSN: 0033-362X. https://doi.org/10.1093/poq/nfs070.

Goldreich, Oded, and Yair Oren. 1994. "Definitions and Properties of Zero-Knowledge Proof Systems." *Journal of Cryptology* 7, no. 1 (December 1, 1994): 1–32. ISSN: 1432-1378. https://doi.org/10.1007/BF00195207.

Halbesleben, Jonathon R.B., Anthony R. Wheeler, and M. Ronald Buckley. 2007. "Understanding Pluralistic Ignorance in Organizations: Application and Theory." *Journal of Managerial Psychology* 22, no. 1 (January 1, 2007): 65–83. ISSN: 0268-3946. https://doi.org/10.1108/02683940710721947.

Hallin, Daniel C. 1989. *The Uncensored War: The Media and Vietnam.* Berkeley: University of California Press. ISBN: 978-0-520-06543-7.

Iyengar, Shanto, and Masha Krupenkin. 2018. "The Strengthening of Partisan Affect." *Political Psychology* 39 (S1): 201–218. ISSN: 1467-9221. https://doi.org/10.1111/pops.12487.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. 2019. "The Origins and Consequences of Affective Polarization in the United States." *Annual Review of Political Science* 22 (Volume 22, 2019 2019): 129–146. ISSN: 1094-2939, 1545-1577. https://doi.org/10.1146/annurev-polisci-051117-073034.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. "Affect, Not Ideology: A Social Identity Perspective on Polarization." *Public Opinion Quarterly* 76, no. 3 (January 1, 2012): 405–431. ISSN: 0033-362X. https://doi.org/10.1093/poq/nfs038.

Kandel, Eugene, and Edward P. Lazear. 1992. "Peer Pressure and Partnerships." *Journal of Political Economy* 100, no. 4 (August): 801–817. ISSN: 0022-3808. https://doi.org/10.1086/261840.

Katz, D., F. H. Allport, and M. B. Jenness. 1931. *Students' Attitudes; a Report of the Syracuse University Reaction Study.* xxviii, 408. Students' Attitudes; a Report of the Syracuse University Reaction Study. Oxford, England: Craftsman Press.

Kickstarter. 2025a. Kickstarter. Accessed May 2, 2025. https://www.kickstarter.com/about.

———. 2025b. Kickstarter Stats, May 23, 2025. Accessed May 23, 2025. https://www.kickstarter.com/help/stats.

Kim, Jin Woo, Andrew Guess, Brendan Nyhan, and Jason Reifler. 2021. "The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity." *Journal of Communication* 71, no. 6 (December 1, 2021): 922–946. ISSN: 0021-9916. https://doi.org/10.1093/joc/jqab034.

Knowledge, Free Our. "Together We Can Fix Academia." Free Our Knowledge. Accessed April 30, 2025. https://freeourknowledge.org/.

Krumpal, Ivar. 2013. "Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review." *Quality & Quantity* 47, no. 4 (June 1, 2013): 2025–2047. ISSN: 1573-7845. https://doi.org/10.1007/s11135-011-9640-9.

Kuran, Timur. 1997. *Private Truths, Public Lies: The Social Consequences of Preference Falsification.* Cambridge, Massachusetts London, England: Harvard University Press. ISBN: 978-0-674-70758-0.

Lehman, Joseph G. 2010. "'The Overton Window': Made in Michigan." *Viewpoint on public issues,* July 5, 2010, no. 19, 2. ISSN: 1093-2240.

León-Medina, Francisco J., Jordi Tena-Sánchez, and Francisco J. Miguel. 2020. "Fakers Becoming Believers: How Opinion Dynamics Are Shaped by Preference Falsification, Impression Management and Coherence Heuristics." *Quality & Quantity* 54, no. 2 (April 1, 2020): 385–412. ISSN: 1573-7845. https://doi.org/10.1007/s11135-019-00909-2.

Lütz, Alessandra F., and Lucas Wardil. 2024. "The Evolution of Pluralistic Ignorance." *Physica A: Statistical Mechanics and its Applications* 647 (August 1, 2024): 129920. ISSN: 0378-4371. https://doi.org/10.1016/j.physa.2024.129920.

Marks, Gary, and Norman Miller. 1987. "Ten Years of Research on the False-Consensus Effect: An Empirical and Theoretical Review." *Psychological Bulletin* 102 (1): 72–90.

McClain, Colleen, Regina Widjaya, Gonzalo Rivero, and Aaron Smith. 2021. *The Behaviors and Attitudes of U.S. Adults on Twitter: A Minority of Twitter Users Produce a Majority of Tweets from U.S. Adults, and the Most Active Tweeters Are Less Likely to View the Tone or Civility of Discussions as a Major Problem on the Site.* Pew Research Center. Accessed August 19, 2025. JSTOR: resrep63496. http://www.jstor.org/stable/resrep63496.

Menges, Roland, Carsten Schroeder, and Stefan Traub. 2005. "Altruism, Warm Glow and the Willingness-to-Donate for Green Electricity: An Artefactual Field Experiment." *Environmental and Resource Economics* 31, no. 4 (August 1, 2005): 431–458. ISSN: 0924-6460, 1573-1502. https://doi.org/10.1007/s10640-005-3365-y.

Milli, Smitha, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D Dragan. 2025. "Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media." *PNAS Nexus* 4, no. 3 (March 1, 2025): pgaf062. ISSN: 2752-6542. https://doi.org/10.1093/pnasnexus/pgaf062.

Mitra, Arnab, and Michael R. Moore. 2018. "Green Electricity Markets as Mechanisms of Public-Goods Provision: Theory and Experimental Evidence." *Environmental and Resource Economics* 71, no. 1 (September 1, 2018): 45–71. ISSN: 1573-1502. https://doi.org/10.1007/s10640-017-0136-5.

Morris, Stephen, and Hyun Song Shin. 2002. "Social Value of Public Information." *American Economic Review* 92, no. 5 (November 1, 2002): 1521–1534. ISSN: 0002-8282. https://doi.org/10.1257/000282802762024610.

Noelle-Neumann, Elisabeth. 1974. "The Spiral of Silence a Theory of Public Opinion." *Journal of Communication* 24, no. 2 (June 1, 1974): 43–51. ISSN: 0021-9916, 1460-2466. https://doi.org/10.1111/j.1460-2466.1974.tb00367.x.

Nolan, Jessica M., and Kenneth E. Wallen. 2021. "Social Norms and Persuasion." In *The Cambridge Handbook of Compliance,* 1st ed., 404–421. Cambridge University Press, May 31, 2021. ISBN: 978-1-108-75945-8. https://doi.org/10.1017/9781108759458.028.

Palfrey, Thomas R., and Howard Rosenthal. 1984. "Participation and the Provision of Discrete Public Goods: A Strategic Analysis." *Journal of Public Economics* 24, no. 2 (July 1, 1984): 171–193. ISSN: 0047-2727. https://doi.org/10.1016/0047-2727(84)90023-9.

Prentice, Deborah A, and Dale T Miller. 1993. "Pluralistic Ignorance and Alcohol Use on Campus: Some Consequences of Misperceiving the Social Norm." *Journal of Personality and Social Psychology* 64 (2): 243–256.

———. 1996. "Pluralistic Ignorance and the Perpetuation of Social Norms by Unwitting Actors." In *Advances in Experimental Social Psychology,* edited by Mark P. Zanna, 28:161–209. Academic Press, January 1, 1996. https://doi.org/10.1016/S0065-2601(08)60238-5.

Raafat, Ramsey M., Nick Chater, and Chris Frith. 2009. "Herding in Humans." *Trends in Cognitive Sciences* 13, no. 10 (October 1, 2009): 420–428. ISSN: 1364-6613, 1879-307X. https://doi.org/10.1016/j.tics.2009.08.002. PMID: 19748818.

Rezaei, Hadis, Mojtaba Eshghie, Karl Anderesson, and Francesco Palmieri. 2025. "SoK: Root Cause of $1 Billion Loss in Smart Contract Real-World Attacks via a Systematic Literature Review of Vulnerabilities." Pre-published, July 27, 2025. https://doi.org/10.48550/arXiv.2507.20175. arXiv: 2507.20175 [cs].

Rose, Steven K, Jeremy Clark, Gregory L Poe, Daniel Rondeau, and William D Schulze. 2002. "The Private Provision of Public Goods: Tests of a Provision Point Mechanism for Funding Green Power Programs." *Resource and Energy Economics,* Economist and Editor George Tolley - A Special Issue in His Honour, 24, no. 1 (February 15, 2002): 131–155. ISSN: 0928-7655. https://doi.org/10.1016/S0928-7655(01)00048-3.

Ross, Lee, David Greene, and Pamela House. 1977. "The "False Consensus Effect": An Egocentric Bias in Social Perception and Attribution Processes." *Journal of Experimental Social Psychology* 13, no. 3 (May 1, 1977): 279–301. ISSN: 0022-1031. https://doi.org/10.1016/0022-1031(77)90049-X.

SellaBand. "SellaBand." SellaBand. Accessed May 23, 2025. https://web.archive.org/web/20061206113014/http://www.sellaband.com/.

Shoup, Victor. 2000. "Practical Threshold Signatures." In *Advances in Cryptology — EUROCRYPT 2000,* edited by Bart Preneel, redacted by Gerhard Goos, Juris Hartmanis, and Jan Van Leeuwen, 1807:207–220. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-540-67517-4. https://doi.org/10.1007/3-540-45539-6_15.

Simons, Greg. 2021. "Swedish Media, Fundamental Values and the Opinion Corridor in the 2018 Election." *Etkileşim,* no. 8 (8 2021): 12–34. ISSN: 2636-7955, 2636-8242. https://doi.org/10.32739/etkilesim.2021.4.8.136.

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. ISSN: 1573-6938. https://doi.org/10.1007/s10683-019-09616-z.

Smout, Cooper, Dawn Liu Holford, Kelly Garner, Ruddy Manuel Illanes Beyuma, Paula Andrea Martinez, Megan Ethel Janine Campbell, Ibrahim Khormi, et al. 2021. "An Open Code Pledge for the Neuroscience Community." Pre-published, December 7, 2021. https://doi.org/10.31222/osf.io/vrwm7.

Spacehive. "Spacehive: The Home of Community Fundraising." Spacehive. Accessed May 2, 2025. https://www.spacehive.com.

Spartacus.app. "Spartacus.App." Spartacus.app. Accessed May 2, 2025. https://spartacus.app.

Steiner, Jennifer G., B. Clifford Neuman, and Jeffrey I. Schiller. 1988. "Kerberos: An Authentication Service for Open Network Systems." *Usenix Winter,* 191–202.

Szabo, Nick. 1997. "Formalizing and Securing Relationships on Public Networks." *First Monday* 2, no. 9 (September 1, 1997). ISSN: 13960466. https://doi.org/10.5210/fm.v2i9.548.

Tabarrok, Alexander. 1998. "The Private Provision of Public Goods via Dominant Assurance Contracts." *Public Choice* 96, no. 3 (September 1, 1998): 345–362. ISSN: 1573-7101. https://doi.org/10.1023/A:1004957109535.

Vacca, Anna, Andrea Di Sorbo, Corrado A. Visaggio, and Gerardo Canfora. 2021. "A Systematic Literature Review of Blockchain and Smart Contract Development: Techniques, Tools, and Open Challenges." *Journal of Systems and Software* 174 (April): 110891. ISSN: 01641212. https://doi.org/10.1016/j.jss.2020.110891.

Warner, Stanley L. 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60, no. 309 (March 1, 1965): 63–69. ISSN: 0162-1459. https://doi.org/10.1080/01621459.1965.10480775. PMID: 12261830.

Watts, Duncan J. 2002. "A Simple Model of Global Cascades on Random Networks." *Proceedings of the National Academy of Sciences* 99, no. 9 (April 30, 2002): 5766–5771. https://doi.org/10.1073/pnas.082090499.

Youvan, Douglas C. 2024. "Shifting Boundaries of Acceptability: Examining the Overton Window and Its Modern Manipulators in U.S. Discourse," https://doi.org/10.13140/RG.2.2.29924.39044.

Yu, Jun-Wu, Guo-Liang Tian, and Man-Lai Tang. 2008. "Two New Models for Survey Sampling with Sensitive Characteristic: Design and Analysis." *Metrika* 67, no. 3 (April 1, 2008): 251–263. ISSN: 1435-926X. https://doi.org/10.1007/s00184-007-0131-x.

Zubrickas, Robertas. 2014. "The Provision Point Mechanism with Refund Bonuses." *Journal of Public Economics* 120 (December 1, 2014): 231–234. ISSN: 0047-2727. https://doi.org/10.1016/j.jpubeco.2014.10.006.

Mathematical Appendices

A1. Proof of Proposition 1

The proposition states that if $N_{strong} = |\mathcal{S}_{strong}| \geq T$, a Nash equilibrium exists where:

- Strategy $\mathbf{a}^*$: All $i \in \mathcal{S}_{strong}$ choose $a_i^* = 1$. All $j \notin \mathcal{S}_{strong}$ (for whom $e_j - c_j + w < 0$) choose $a_j^* = 0$.
- Condition: Agents $j \notin \mathcal{S}_{strong}$ are not pivotal or, if pivotal, they do not prefer signing (i.e., $j \notin \mathcal{S}_{pivotal}$).

Under strategy $\mathbf{a}^*$, since $N_{strong} \geq T$, the total number of endorsers is $M^* = N_{strong} \geq T$, thus the contract succeeds. We verify no unilateral deviation exists:

1. Agent $i \in \mathcal{S}_{strong}$ (chooses $a_i^* = 1$): Their equilibrium utility is $U_i(a_i^* = 1, \text{Success}) = v_i + e_i - c_i + w$.

If i deviates to $a_i' = 0$:

- If $N_{strong} - 1 \geq T$, the contract succeeds anyway: $U_i(a_i' = 0, \text{Success}) = v_i$. Agent i does not deviate if

$$v_i + e_i - c_i + w \geq v_i \quad \Rightarrow \quad e_i - c_i + w \geq 0,$$

  which holds by definition for all $i \in \mathcal{S}_{strong}$.

- If $N_{strong} - 1 < T$ (implying $N_{strong} = T$, agent i pivotal), the contract fails if i deviates: $U_i(a_i' = 0, \text{Fail}) = 0$. No deviation occurs if

$$v_i + e_i - c_i + w \geq 0,$$

  which trivially holds given $e_i - c_i + w \geq 0$ and $v_i > 0$.

Thus, no $i \in \mathcal{S}_{strong}$ deviates.

2. Agent $j \notin \mathcal{S}_{strong}$ (chooses $a_j^* = 0$): Their equilibrium utility is $U_j(a_j^* = 0, \text{Success}) = v_j$ (contract succeeds due to $\mathcal{S}_{strong}$).

If j deviates to $a_j' = 1$:

- Contract success is unaffected (total signers: $N_{strong} + 1$). Utility becomes $U_j(a_j' = 1, \text{Success}) = v_j + e_j - c_j + w$. Agent j does not deviate if

$$v_j \geq v_j + e_j - c_j + w \quad \Rightarrow \quad e_j - c_j + w \leq 0,$$

which is true since $j \notin \mathcal{S}_{strong}$ implies $e_j - c_j + w < 0$.

- **Clarification on pivotality:** If $N_{strong} = T$, exactly T agents from $\mathcal{S}_{strong}$ sign, thus ensuring contract success. Hence, no agent outside $\mathcal{S}_{strong}$ is pivotal under equilibrium. If pivotality were hypothetically considered, $j \notin \mathcal{S}_{pivotal}$ by assumption ensures j has no profitable incentive to deviate.

Thus, no $j \notin \mathcal{S}_{strong}$ deviates either.

Therefore, strategy profile $\mathbf{a}^*$ constitutes a Nash equilibrium when $N_{strong} \geq T$.

A2. Proof of Proposition 2

Proof.— Fix the distribution F_x and define

$$H(x) \;\equiv\; x\, p_{succ}(x) + w,$$

where $p_{succ}(x)$ is given by Eq. (2) with $q(x) \equiv 1 - F_x(x)$:

$$p_{succ}(x) \;=\; \sum_{k=T-1}^{N-1} \binom{N-1}{k} q(x)^k \left[1 - q(x)\right]^{N-1-k}.$$

Continuity and monotonicity. Since F_x is continuous, so is $q(\cdot)$, hence $p_{succ}(\cdot)$ is a finite sum of continuous functions and therefore continuous in x. Moreover, p_{succ} is nonincreasing in x because it is increasing in q and $q'(x) = -f_x(x) \leq 0$. Thus H is continuous on $\mathbb{R}$.

Boundary signs. For any $x > \overline{x}$, $q(x) = 1 - F_x(x) \to 0$ so $p_{succ}(x) \to 0$ and $H(x) \to w > 0$. For $x < \underline{x}$ we have $F_x(x) = 0$, so $q(x) = 1$ and $p_{succ}(x) = 1$, hence $H(x) = x + w$. Choosing x sufficiently negative yields $H(x) < 0$.

Existence. By the intermediate value theorem there exists $x^* \in \mathbb{R}$ with $H(x^*) = 0$, i.e.

$$x^* \;=\; -\frac{w}{p_{succ}(x^*)},$$

which is Eq. (1). Finally, $x^* < 0$ because for any $x \geq 0$, $H(x) = x\, p_{succ}(x) + w \geq w > 0$, so no root can lie at or to the right of 0. This establishes existence of a symmetric cutoff equilibrium and the sign property stated in Proposition 2.

A3. Formal Derivations of Comparative Statics (Incomplete Information)

Equilibrium Definition and Assumptions: For $T > 1$, the equilibrium cutoff x^* is implicitly defined by:

$$H(x^*, \alpha) = x^* p_{succ}(x^*; \alpha) + w = 0$$

where α represents parameters of interest (e.g., w, distribution parameters of F_x, threshold T, or population size N).

We apply the Implicit Function Theorem (IFT):

$$\frac{dx^*}{d\alpha} = -\frac{\partial H/\partial \alpha}{\partial H/\partial x^*}.$$

Assumption (Stability/Uniqueness): We assume (see Appendix A.A2):

$$\frac{\partial H}{\partial x^*} = p_{succ}(x^*) + x^* p'_{succ}(x^*) > 0.$$

Derivation of Comparative Statics:

1) **Warm-glow (w):** We have $\frac{\partial H}{\partial w} = 1$, hence

$$\frac{dx^*}{dw} = -\frac{1}{p_{succ}(x^*) + x^* p'_{succ}(x^*)} < 0.$$

Thus, increasing w decreases x^*, increasing participation.

2) **FOSD Improvement in Distribution F_x:** Let an increase in α_{shift} represent an FOSD improvement (stochastically higher types x_i). We have $\frac{\partial p_{succ}}{\partial \alpha_{shift}} > 0$. Thus,

$$\frac{dx^*}{d\alpha_{shift}} = -\frac{x^*(\partial p_{succ}/\partial \alpha_{shift})}{p_{succ}(x^*) + x^* p'_{succ}(x^*)} > 0.$$

An improvement in distribution raises the equilibrium cutoff.

3) **Threshold (T):** Increasing T makes success harder ($\frac{\partial p_{succ}}{\partial T} < 0$). Thus,

$$\frac{dx^*}{dT} = -\frac{x^*(\partial p_{succ}/\partial T)}{p_{succ}(x^*) + x^* p'_{succ}(x^*)} < 0.$$

Higher T lowers cutoff, increasing participation.

4) **Population Size (N):** Increasing N makes success easier ($\frac{\partial p_{succ}}{\partial N} > 0$). Thus,

$$\frac{dx^*}{dN} = -\frac{x^*(\partial p_{succ}/\partial N)}{p_{succ}(x^*) + x^* p'_{succ}(x^*)} > 0.$$

Larger N raises cutoff, tightening individual condition.

5) **Cost Convexity (Increase in Costs):** Let α_{cvx} represent increased costs, causing a stochastic decrease in types ($\frac{\partial p_{succ}}{\partial \alpha_{cvx}} < 0$). Thus,

$$\frac{dx^*}{d\alpha_{cvx}} = -\frac{x^*(\partial p_{succ}/\partial \alpha_{cvx})}{p_{succ}(x^*) + x^* p'_{succ}(x^*)} < 0.$$

Higher costs lower cutoff, relaxing individual participation criteria.

Summary: I establish rigorous comparative statics results clearly aligned with intuition under the maintained assumption of a stable and unique equilibrium ($\partial H/\partial x^* > 0$).

Vulnerabilities and Mitigations

Any mechanism relying on the conditional revelation of sensitive information must consider potential adversarial scenarios and practical implementation challenges. This section outlines key vulnerabilities and suggests corresponding mitigation strategies.

B1. Imperfect Trust in the Administrator and Mechanism Integrity

A crucial assumption underpinning an agent's willingness to participate in a social assurance contract, especially when w is small or c_i is large, is their trust in the contract administrator and the overall mechanism to uphold confidentiality if the assurance threshold T is not met. Let ρ be the subjective probability an agent assigns to the administrator being trustworthy and the system functioning as intended (i.e., secrecy is maintained upon failure). Consequently, $1-\rho$ is the probability of a "betrayal", where a signature is revealed despite $M < T$.

This imperfect trust ($\rho < 1$) alters the expected payoff for an agent i who chooses $a_i = 1$ (Sign) if the contract fails. Instead of receiving w with certainty, their expected utility in this scenario becomes:

$$E[U_i(a_i = 1, M < T)] = \rho \cdot w + (1-\rho)(w - c_i) = w - (1-\rho)c_i$$

The agent still receives the warm-glow w for the act of signing, but now faces an expected additional cost of $(1-\rho)c_i$ due to the risk of premature, unsanctioned exposure. The revised payoffs are shown in Table B1:

TABLE B1—PAYOFFS FOR AGENT i WITH IMPERFECT TRUST ($\rho < 1$)

Action a_i	**Contract Fails** ($M < T$)	**Contract Succeeds** ($M \geq T$)
$a_i = 0$ (Not Sign)	0	v_i
$a_i = 1$ (Sign)	$w - (1-\rho)c_i$	$v_i + e_i - c_i + w$

Impact on Equilibrium Behavior with Complete Information— The condition for agent i to prefer $a_i = 1$ when they expect the contract to fail (i.e., $M_{-i} < T-1$) changes from $w > 0$ to $w-(1-\rho)c_i > 0$, or $w > (1-\rho)c_i$. If $w \leq (1-\rho)c_i$ for all agents, the "all-silent" profile (all $a_j = 0$) can re-emerge as a Nash equilibrium, even if $w > 0$. The power of the warm-glow to ensure some participation is thus diminished by distrust. Participation becomes less likely if trust ρ is low, expected cost of exposure c_i is high, or warm-glow w is low.

Impact on Equilibrium Behavior with Incomplete Information— Imperfect trust makes signing less attractive. The simplified equilibrium condition $x^* = -w/p_{succ}(x^*)$ (where $x_i = e_i - c_i$) was based on the premise that the net payoff from signing if the contract fails was w. With imperfect trust, this payoff is now $w - (1-\rho)c_i$.

If we adapt the simplified style of Eq. 1, an agent i with type $x_i = e_i - c_i$ and specific cost c_i might sign if $x_i \geq \frac{-(w-(1-\rho)c_i)}{p_{succ}(x^*)}$. This can be rewritten as $e_i p_{succ}(x^*) + w \geq c_i(1 - \rho + p_{succ}(x^*))$. This condition now depends on e_i and c_i separately, not just their difference x_i. This means that the agent's type for the signing decision effectively becomes multi-dimensional (e.g., depending on (e_i, c_i) or even (v_i, e_i, c_i) when considering pivotality more broadly), and a single cutoff x^* for $x_i = e_i - c_i$ no longer fully characterizes the equilibrium strategy in a simple way. A formal analysis of the BNE with imperfect trust would require solving for an equilibrium strategy over a multi-dimensional type space, which is an interesting avenue for future work. However, the clear qualitative implication is that lower trust ρ (i.e., a higher perceived probability $1 - \rho$ of premature exposure) acts as an additional expected cost for signing, leading to lower participation rates and a reduced likelihood of the social assurance contract achieving its assurance threshold T. This underscores the importance of the insider attack mitigation strategies discussed in Appendix B.B3 (particularly cryptographic guarantees) and the overall need for contract organizers and platforms to establish and maintain a high degree of trustworthiness (i.e., ensure $\rho \approx 1$). Without sufficient trust, the assurance mechanism itself is undermined. A summary of the implications of relaxing the trust assumption in comparison to the other two cases examined is found in Table 2.

B2. Sybil Attacks and Ballot-Stuffing

A primary risk is the *ballot-stuffing* or *Sybil* attack (Douceur 2002), where malicious actors generate fake or insincere signatures to falsely achieve the assurance threshold T, leading to premature or misleading publication. Robust identity verification is the key defense. Requiring participants to authenticate via verifiable methods (e.g., organizational email, unique institutional credentials, or digital signatures linked to known identities within population $\mathcal{N}$) significantly hinders the creation of multiple fraudulent entries. Organizational authentication systems like Kerberos offer a model for such verification (Steiner, Neuman, and Schiller 1988). Additionally, a small, non-monetary stake or commitment, such as a verifiable pledge, could deter frivolous sign-ups.

A distinct challenge arises if legitimate members of $\mathcal{N}$ sign insincerely merely to trigger publication and expose sincere endorsers. Such insincere endorsers might then publicly disavow the statement, claiming they signed only to "out" others. This could be an effective strategy if it is wholly implausible that the insincere endorser could possibly have been sincere. In this case, drawing up a list of likely suspects and excluding them from the contract in the first place would be an effective defense. In the case where it is ambiguous whether a endorser who claims to have signed only to cause the contract to succeed actually did so, this endorser now has their signature indelibly affixed to an allegedly repugnant statement. Whether a *post hoc* claim of insincerity will be remembered as long as an indelible signature on a contract is quite questionable, and should provide a disincentive.

B3. Insider Attacks and Mechanism Integrity

The integrity of the contract administrator (system or person) is paramount. An insider attack (such as prematurely leaking signatures, falsely claiming T is met, or succumbing to external coercion to reveal identities) or a technical compromise of the platform severely undermines the contract's assurance. Mitigating these risks often involves removing single points of failure and employing robust technical and procedural safeguards.

Distributing trust, for instance, by using multiple independent escrow holders requiring a quorum for action, is one approach. Cryptographic methods, however, should be considered the gold standard for ensuring both secrecy and integrity. Decentralized mechanisms like smart contracts on a blockchain could automate conditional release based on verifiable cryptographic conditions (e.g., T valid digital signatures from institutional certificates), minimizing human intervention (cf. Szabo 1997, for foundational concepts of smart contracts; Vacca et al. 2021, for a review). Threshold cryptography offers solutions where signatures are encrypted such that decryption requires either

T participants to cooperatively reveal their keys or a quorum of trustees to combine shares (Shoup 2000). Furthermore, anonymous credential systems can allow users to prove eligibility without revealing identity to the administrator until publication (Camenisch and Lysyanskaya 2001), and zero-knowledge proofs could enable verification that $M \geq T$ valid signatures exist without revealing which specific signatures they are (Goldreich and Oren 1994). While the full development of user-friendly platforms incorporating such advanced cryptographic guarantees for general-purpose social assurance contracts is ongoing, these technologies can offer strong protection against both premature exposure and administrator malfeasance.

Beyond purely technical measures, legal agreements imposing significant penalties for malfeasance can provide a deterrent. Centralized platforms must also cultivate a strong reputation for integrity, potentially bolstered by independent third-party audits and open-source software development. Ultimately, participants' trust ($\rho \approx 1$, as discussed in Section B.B1) is essential, and this may rely on a combination of technical assurances, reputational trust in the administrator, and clear governance regarding the contract's terms, including the immutability of the statement being endorsed.

Network Topology and the Success of Assurance Contracts

The effectiveness of a social assurance contract can also depend on the underlying social network structure of the N potential endorsers. While formal modeling of these network effects is left for future research, this section qualitatively discusses several key influences.

The presence of cohesive subgroups or dense clusters within the network may facilitate coordination. Pre-existing trust within such a subgroup can bolster individuals' confidence in the integrity of the contract process and their belief that peers within the subgroup will also choose $a_j = 1$ (Sign). This enhanced trust could translate into a higher subjective probability of success, $p_{succ}(x^*)$, or a lower perceived individual cost, c_i. Furthermore, an organizer could leverage existing trust links to quietly recruit a core group of endorsers. If this core group is sufficiently large to approach or meet the assurance threshold T, it could catalyze wider participation from the periphery. Strong intra-group ties, therefore, can be instrumental in overcoming the initial coordination hurdle, suggesting that $p_{succ}(x^*)$ might be influenced not only by the global distribution $F_x(x)$ but also by local network characteristics that shape beliefs about others' actions.

Conversely, coordination becomes more challenging if potential supporters are dispersed across a fragmented network, for instance, in different departments, distinct social circles, or otherwise structurally isolated parts of the community. Such fragmentation can create information barriers, leaving individuals unaware of the extent of shared sentiment or even of the contract's existence.

In these scenarios, the role of trusted individuals or communication channels that act as bridges between disconnected segments, or as central hubs, becomes crucial. These network bridges are essential for disseminating information about the contract and for recruiting a broad base of endorsers; without them, $p_{succ}(x^*)$ may be severely underestimated by many potential participants. In highly fragmented networks, it might even be necessary to employ multiple, parallel assurance contracts, perhaps one for each cluster, potentially with an overarching mechanism to combine successes if each meets a local assurance threshold. The design of the assurance threshold T itself could be adapted, for instance, by requiring a vector of participation $T = (T_1, \ldots, T_k)$ from k different groups for publication, ensuring broad representation.

The social insularity of a group also plays a role, which may come down to the question: who has the power to hurt me? In an insular group, the opinions of outsiders may be of little importance. This could mean the people who might potentially levy costs c_i upon contract success are limited in number and therefore easier to reason about. In the case where a group is not insular, and has more links to society at large rather than to each other, we might expect that the assurance threshold T must be very high since "the rest of society" provides a rather large number of people who could contribute to c_i. In insular groups with strong, monolithic cultures or those that heavily censor dissent, individuals who disagree with the prevailing norm may be completely isolated, with no easy way to find like-minded others or gauge latent support. Launching an assurance contract in such environments might necessitate external intervention or anonymous broadcasts, as any known organizer could be targeted before reaching T. In contrast, in more open insular groups where social ties cut across clusters of beliefs, information about a contract could diffuse more freely. A successful contract in such an environment may be more likely to draw endorsers from a diverse cross-section of beliefs, lending greater credibility and broader impact to the revealed consensus and subsequent belief updates.

From a network diffusion perspective, as explored in models like Watts (2002), a social assurance contract essentially seeks to trigger a cascade of endorsement once the critical mass T is achieved. One might consider the role of "k-cores" within the network: if a subset of at least T mutually connected or trusting individuals can be convinced to sign (perhaps due to highly favorable x_i or a strong response to w), their collective commitment, held in escrow, can then persuade others on their periphery to join. However, such cascades might be halted at the boundaries of network modules if bridging ties are weak. Overall, a network exhibiting moderate connectivity—neither overly fragmented nor excessively centralized around a few vulnerable hubs—is likely most conducive to the success of a social assurance contract. Such a structure supports the formation of an initial

critical mass and the subsequent widespread propagation of the revealed stance, potentially leading to a more significant and group-wide update of beliefs. The perceived $p_{succ}(x^*)$ for any individual would then be influenced by their network position and the local signals of support they observe.

Outline of a Model Extension with Strategic Opposition

The core model analyzed in this paper provides foundational insights into the mechanics of social assurance contracts by focusing on the strategic decisions of potential endorsers under exogenous (though possibly heterogeneous) cost and benefit parameters. This appendix outlines a more comprehensive model of the interation between the Overton window and a social assurance contract. It incorporates additional realistic features, primarily by endogenizing social costs through the strategic actions of individuals who may oppose the success of the contract, and by allowing for uncertainty regarding the total number of potential supporters. Formal analysis of this extension is left for future research.

D1. Model Components in the Extension Framework

1. Population Segmentation and Preferences— Let $\mathcal{P}$ be the total underlying population. Each individual $j \in \mathcal{P}$ possesses an individual-specific valuation v_j concerning the public revelation and potential normalization of the statement s_{belief} endorsed by the contract. This valuation v_j can be positive, zero, or negative, leading to a natural segmentation of $\mathcal{P}$:

- **Supporters ($\mathcal{N}$):** The set of individuals for whom $v_i > 0$. Let $N = |\mathcal{N}|$ be the true number of such potential endorsers. In this extension, unlike the main model, N is not assumed to be common knowledge. Instead, each agent $k \in \mathcal{P}$ (particularly potential supporters $i \in \mathcal{N}$) forms a subjective belief or estimate, $\hat{N}_k$ (which could be a point estimate or a probability distribution), about the true value of N. Each supporter $i \in \mathcal{N}$ also has heterogeneous private esteem e_i, cost c_i (which will now be endogenous), and warm-glow w_i (which I assume is heterogeneous here) associated with choosing $a_i = 1$ (Endorse).

- **Indifferents ($\mathcal{I}_0$):** The set of individuals for whom $v_j = 0$. The size of this set, $|\mathcal{I}_0|$, may also be uncertain.

- **Opposers ($\mathcal{O}$):** The set of individuals for whom $v_j < 0$. These individuals experience disutility if s_{belief} is successfully publicized and enters the Overton window. The size of this set, $|\mathcal{O}|$, may also be uncertain.

2. Endogenous Social Costs (c_i) for Supporters— For a supporter $i \in \mathcal{N}$, the private cost c_i incurred if they endorse ($a_i = 1$) and the contract succeeds ($M \geq T$) is now explicitly a function of the aggregate punishment enacted by opposers. Let P_{total} be the total punishment enacted by the set of opposers $\mathcal{O}$. Then $c_i = C_i(P_{total}, M, \text{own characteristics}_i)$, where a higher P_{total} or a lower M (number of endorsers sharing the burden) would typically increase c_i.

3. Opposers' Characteristics and Strategic Behavior— Each opposer $j \in \mathcal{O}$ is characterized by their negative valuation $v_j < 0$ and an "influence" or "power" parameter $i_j \geq 0$. Their capacity to contribute to the punishment pool is $p_j(v_j, i_j)$, where higher $|v_j|$ (stronger opposition) and higher i_j lead to greater p_j. Let $\mathbf{p} = \{p_j\}_{j \in \mathcal{O}}$ be the vector of punishment potentials. If the contract succeeds ($M \geq T$) and s_{belief} is published, opposers then decide whether to enact punishment. Crucially, I assume enacting punishment is *costless* to the opposers themselves ($k_p = 0$). However, they only choose to punish if they believe their collective punishment will be *effective*. Punishment is effective when it prevents s_{belief} from durably entering or remaining in the Overton window. If the total punishment $\sum p_j$ is sufficiently high relative to the number of endorsers M (i.e., if T was set too low by the administrator, leading to a small M), the resulting c_k for endorsers might be so large that expressing s_{belief} remains prohibitively costly overall, negating the contract's impact on the Overton window. In this case, ex-post utility for a significant number of revealed endorsers is negative: $v_k + e_k - c_k(M, \sum p_j) + w_k < 0$ for many $k \in \{i \mid a_i = 1\}$. Opposers would not punish if they estimate that M is so large (potentially influenced by their own estimate of N and of how supporters react to that) that their pooled punishment $\sum p_j$ would be too diluted to effectively increase c_k to a deterrent level, meaning the underlying belief $\mathbf{b}_s$ (represented by s_{belief}) would enter the Overton window regardless of their efforts.

4. Supporters' Decision with Endogenous Costs— Supporters $i \in \mathcal{N}$ must now form expectations that are conditional on their individual estimate of the total number of supporters, $\hat{N}_i$. These expectations cover: (a) the likely number of actual endorsers M who will sign (which depends on $\hat{N}_i$, the threshold T, the distribution of private types (v_j, e_j, c_j, w_j) among the estimated $\hat{N}_i$ supporters, and their strategic responses); (b) the potential collective punishment P_{total} from $\mathcal{O}$ if the contract succeeds with M endorsers; and (c) how this punishment translates into their individual $c_i(M, P_{total})$. Their decision $a_i = 1$ would depend on whether their subjective expected utility $E_i[v_i + e_i - c_i(M, P_{total}) + w_i \mid \hat{N}_i]$ (if $M \geq T$) is sufficiently high compared to $E_i[w_i \mid \hat{N}_i]$ (if $M < T$) and $E_i[v_i \mid \hat{N}_i]$ (if $a_i = 0, M \geq T$). The calculation of whether $v_i + e_i - c_i(M, P_{total}) + w_i$ is positive in expectation, considering the punishment subgame and the uncertainty over N, becomes central.

5. Architect's Objective— The contract architect sets T. Their objective might be to maximize the probability that $\mathbf{b}_s$ *durably* enters the Overton window. This now explicitly involves choosing a T considering not only the likely distribution of types but also the fact that potential supporters (and opposers) operate with potentially heterogeneous estimates $\hat{N}_k$ of the true number of supporters N. An optimal T would need to be robust enough or be perceived as achievable given these beliefs, such that if $M = T$ endorsers are revealed, this number is sufficient to signal to opposers that punishment would be futile, thus ensuring c_i remains manageable for supporters.

D2. Straightforward Implications and Complexities of the Model with Strategic Opposition

Introducing strategic, (personally) costless but efficacy-dependent punishment by opposers, coupled with uncertainty about the total number of supporters N, has several key implications for the social assurance contract model. Firstly, the concept of "safety" for endorsers becomes endogenously determined and subject to more layers of uncertainty. It is no longer solely about being one of $M \geq T$ individuals, but more critically about M being sufficiently large (relative to individual beliefs $\hat{N}_i$ and the true N) to deter or significantly dilute punishment from the opposing group $\mathcal{O}$. If the administrator sets the mechanism's threshold T too low, a social assurance contract might formally "succeed" (i.e., achieve $M \geq T$ endorsements), yet this M could be insufficient to deter opposers. In such a scenario, the resulting costs c_i imposed on endorsers might become so high as to render their ex-post utility negative. Consequently, the endorsed belief $\mathbf{b}_s$ (represented by the endorsed statement s_{belief}) might fail to durably enter the Overton window because association with it remains prohibitively costly. This implies the existence of an effective impact threshold, $M_{Overton}$, representing the level of observed endorsement M at which opposers would likely deem punishment ineffective and stand down. The administrator's optimal choice for T would then ideally be at or above this $M_{Overton}$, or at least set to maximize the probability $P(M \geq M_{Overton})$, considering the distribution of beliefs about N. Second, this framework significantly increases the complexity of strategic calculations for all involved. Supporters must now forecast not only the participation decisions of other potential supporters (from a pool whose size $\hat{N}_i$ they estimate) but also the strategic response of the entire group of opposers, which influences their expected c_i. Opposers, in turn, must forecast the likely M resulting from the social assurance contract to determine if their potential punishment would be worthwhile. Third, this introduces new potential failure modes for the contract. Beyond the simple failure of $M < T$ (resulting in no publication), a social assurance contract could achieve $M \geq T$ only to be met with effective punishment. Such an outcome would make endorsers regret their decision and could prevent the intended shift in the

Overton window, representing a pyrrhic victory for the initial publication. Finally, the success of a social assurance contract in durably shifting norms and achieving its objectives will depend on the characteristics and coordination of opposers, and now also on the collective beliefs and estimation accuracy regarding N. The aggregate punishment potential, $\sum p_j$, and the ability of opposers to coordinate their punishment strategy become paramount factors. Analyzing these rich interactions would indeed require advanced game-theoretic tools.

A full analysis of this extension is a substantial undertaking well beyond the scope of the present paper. However, outlining it serves to highlight the broader context in which social assurance contracts operate. It emphasizes that social costs (c_i) faced by supporters are not merely exogenous parameters but can be the result of active, strategic opposition, and that fundamental parameters like the number of potential allies (N) may themselves be subject to uncertainty. The simpler model analyzed in the main body of this paper provides a foundation by elucidating the core coordination mechanism among supporters. On this foundation, future research can model these more complex inter-group dynamics and informational and strategic uncertainties.